2014 KDI Journal of Economic Policy Conference

Job Creation and Business Investment as Pathways to a Creative Economy

Hosted by

KDI KAEA

KDI

2014 KDI Journal of Economic Policy Conference

Job Creation and Business Investment as Pathways to a Creative Economy

Hosted by
KDI
KAEA

ⓒ January 2015
Korea Development Institute
263, Namsejong-ro, Sejong-si
30149, Korea

ISBN 978-89-8063-970-0 (93320)

Contents

CHAPTER 1
New-to-market Product Innovation and Firm Performance: Using Innovation Survey from Japan
(Daiya Isogawa, Kohei Nishikawa, Hiroshi Ohashi)

1. Introduction 2
2. Approaches to Evaluating Innovation Outcomes 3
3. Hypotheses on New-to-Market Product Innovation 5
4. Econometric Analysis 16
5. Conclusion 35

References 36

CHAPTER 2
The Contribution of Research and Innovation to Productivity and Economic Growth
(Amani Elnasri, Kevin J. Fox.)

1. Introduction 41
2. Investment in Knowledge Capital and Economic & Productivity Growth 44
3. Intangible Investment 47
4 Growth accounting with intangible capital 61
5. Government spending on science and innovation 68
6. The relation between public support for R&D and market sector MFP growth 76
7. Econometric analysis 78
8. Conclusion 96

References 100
Appendix 104

CHAPTER 3
Assessing an Efficiency Defense:
The Case of Intel's Marketing Campaign
(Hwa Ryung Lee, Andras Pechy and Michelle Sovinsky)
1. Introduction 111
2. Test of Advertising Predation (TAP) 113
3. Application of TAP to "Intel Inside" Campaign 118
4. Demand for CPUs 126
5. Marginal Marketing Revenue 143
6. Marginal Marketing Cost 144
7. Robustness and Other Considerations 154
8. Policy Implications 158

References 160
Appendix 163

CHAPTER 4
Entrepreneurship, Small Businesses, and Economic Growth in Cities
(YONG SUK LEE)
1. Introduction 166
2. A Simple Theory of Entrepreneurship and Urban Growth 170
3. Data and Variables 174
4. The Impact of Entrepreneurship on Urban Growth 177
5. The Impact of Government-backed Entrepreneurship on Urban Growth 199
6. Conclusion 209

References 211
Appendix 214

CHAPTER 5
Changes in Competition of Small vs. Large Firms Resulting from International Trade
(Changwoo Nam, Jiyoon Oh)
1. Introduction 220
2. Theoretical Background 223
3. Estimation 229
4. Empirical Results 240
5. Conclusions 256

References 258
Appendix 261

CHAPTER 6
Endogenous Product Characteristics in Merger Simulation: A Study of the U.S. Airline Industry
(Jinkook Lee)
1. Introduction 269
2. The Model 274
3. Data 284
4. Estimation 288
5. Merger Simulations 300
6. Conclusion 326

References 329

CHAPTER 7
Knowledge, Entrepreneurship and Creation of New Competence: Foundations of the Creative Economy
(Hong Y. Park)
1. Introduction 333
2. The Nature of Knowledge 335
3. Entrepreneurship Theories of Kirzner, Schumpeter and Knight 340
4. Knowledge and Entrepreneurship and Uncertainty 349
5. Corporate Entrepreneurship: The Dow Chemical Company 354

6. Discussion 363
7. Policy Implications 369
8. Conclusion 373

References 374

CHAPTER 8
Canary in a Coal Mine – Analysis of Systemic Risk
(Gabjin Oh, Hyeongsop Shim and Yong-Cheol Kim)
1. Introduction 382
2. Network analysis and Literature Review 386
3. Measurement of Systemic Risk 392
4. Empirical results 414
5. Conclusions 424

References 434

CHAPTER 9
Monetary Policy Transmission Via Risk-taking Channel in the Mortgage Market
(Min-Ho Nam)
1. Introduction 438
2. Risk-taking Channel and Mortgage Lending 442
3. Empirical Estimation of Risk-taking Channel 447
4. Developing a DSGE Model 456
5. Analysis of Monetary Policy Transmission in Baseline Model 463
6. Effects of Risk-taking Channel in Mortgage Market 467
7. Conclusion 476

References 479
Appendix 482

CHAPTER 10
Enhancing the Link between Higher Education and Employment
(Kye Woo Lee, Miyeon Chung)
1. Introduction 488
2. Review of Literature and Fiscal Assistance Progams 491
3. Method and Data 498
4. Analysis Results 502
5. Conclusion 511

References 514

CHAPTER 11
Effect of College Major on Earnings and Gender Gap in Labor Markets: Evidence from Young Adults in South Korea
(Sungjin Cho, Jihye Kam, Soohyung Lee)
1. Introduction 517
2. Institutional Background and Data 520
3. Empirical Framework 529
4. Findings 530
5. Robustness Checks 539
6. Conclusion 541

References 542

CHAPTER 12
Economic Growth and Labor Market Institutions in East Asian Structural Transformation
(Seungjoon Oh, Seung-Gyu Sim)
1. Introduction 545
2. The Baseline Model 550
3. Numerical Analysis 564
4. Counterfactual Experiments 573
5. Conclusion 579

References 581
Appendix 584

CHAPTER 13
Wage Dynamics with Private Learning-by-doing and On-the-job Search
(Seung-Gyu Sim)
1. Introduction 594
2. The Model 598
3. Data 610
4. Estimation 613
5. Conclusion 619

References 620
Appendix 622

List of Tables

CHAPTER 1

Table 1	Classification of Industries in JNIS	5
Table 2	Summary Statistics	23
Table 3	Estimation Results of Equation (1)	25
Table 4	Estimation Results of Equation (2)	27
Table 5	Estimation Results of Equations (3) and (4)	29
Table 6	Estimation Results of Equation (5)	32
Table 7	Estimation Results of Equation (6)	33
Table 8	Estimation Results of Equation (7)	34

CHAPTER 2

Table 1	Definitions of Intangibles, CHS	49
Table 2	Estimates of nominal intangible investment in the Australian market sector	54
Table 3	Depreciation rate assumptions	58
Table 4	Capital and labour income shares, market sector 1974-75 to 2012-13	67
Table 5	Spillovers from intangible investment (1993-94 to 2012-13)	83
Table 6	Spillovers from software (1993-94 to 2012-13)	85
Table 7	Spillovers from innovative property (1993-94 to 2012-13)	85
Table 8	Spillovers from economic competencies (1993-94 to 2012-13)	87
Table 9	Spillovers from total public support (1993-94 to 2012-13)	89
Table 10	Spillovers from public support (1993-94 to 2012-13): Research agencies	91
Table 11	Spillovers from public support (1993-94 to 2012-13): Research agencies – breakdown	92
Table 12	Spillovers from public support (1993-94 to 2012-13): Higher education sector	93

Table 13 Spillovers from public support (1993-94 to 2012-13): Business enterprise sector 94
Table 14 Spillovers from public support (1993-94 to 2012-13): Multisector/Civil 95

CHAPTER 3

Table 1 Market Shares and Advertising Expenditures by PC Firms 120
Table 2 CPU price and CPU marginal cost of Intel-based Dell PCs 122
Table 3 Percent and Market Share of PCs by CPU Type 123
Table 4 Descriptive Statistics 124
Table 5 Descriptive Statistics by CPU Manufacturer (feb 19) 125
Table 6 Probit Regressions of PC Purchase (in 2002) 136
Table 7 PC purchase Probability on PC Characteristics and Demographics 137
Table 8 CPU Purchase Estimates 138
Table 9 Multinomial Logit Estimates of PC/CPU Demand (feb 19) 140
Table 10 Rebate amounts paid by Intel to Dell 146
Table 11 TAP Results of Intel's Marketing Campaign for Dell's Advertising of Intel-Powered PCs 149
Table 12 TAP Results for HP and Toshiba 152
Table 13 TAP Results for Gateway 153

CHAPTER 4

Table 1 Summary Statistics 176
Table 2 Impact of Entrepreneurship on Urban Growth (10 year growth): OLS Estimates 178
Table 3 Impact of Entrepreneurship on Urban Growth (5 year growth): OLS Estimates 183
Table 4 Impact of Entrepreneurship on Urban Economic Growth: First-difference Estimates 185
Table 5 Homestead Exemption in 1975 and Year Interstate Banking was Permitted by State 187
Table 6 Impact of Entrepreneurship on Urban Economic

	Growth: 2SLS Estimates	190
Table 7	Robustness Tests	197
Table 8	Impact of Government-backed Entrepreneurship on Urban Economic Growth: OLS and First-difference Estimates	200
Table 9	Impact of Government-backed Entrepreneurship on Urban Economic Growth: 2SLS Estimates	204
Table 10	Crowd-out of Market Entrepreneurship by Government-backed Entrepreneurship	206

CHAPTER 5

Table 1	Data Statistics	241
Table 2	Statistics of Input Variables	241
Table 3	Statistics of TFPs and Markups	242
Table 4	Means and Differences of Markups	245
Table 5	Market Share and Export Effects on LCK Markups in Unbalanced Panel	251
Table 6	Market Share and Export Effects on LCK Markups in Dynamic Unbalanced Panel	253
Table 7	Import Penetration Effect on Dispersion of Markups	255

CHAPTER 6

Table 1	Data Sources	285
Table 2	Variable Definitions and Summary Statistics for the Estimation Sample	287
Table 3	Instrumental Variables for Endogenous Product Characteristics	291
Table 4	Estimation Results on Model Parameters	294
Table 5	Operating Cost per Available Seat Mile (CASM, in cents)	298
Table 6	Description of Simulation Sample	301
Table 7	Changes in Price and Characteristics of Merged Firm's Products in Oligopoly Markets	306
Table 8	Changes in Price and Characteristics of Merged Firm's Products in Monopoly Markets	314

Table 9 Change in Consumer Surplus (CS) after the Delta and Northwest Airlines Merger 316
Table 10 Change in Producer Surplus (PS) after the Delta and Northwest Airlines Merger 320
Table 11 Change in Total Surplus (TS) after the Delta and Northwest Airlines Merger (unit: $100K) 321
Table 12 Market Competitiveness of QI and QD markets Pre-merger 322
Table 13 Comparison of Average Market Frequency (AMF): Pre-merger vs. Post-merger (Simulated) vs. Post-merger (Actual) 323

CHAPTER 7
Table 1 Lam's Dimensions of Knowledge 339
Table 2 Alvarez & Barney's Central Assumptions of Discovery and Creation Theories of Entrepreneurial Action 347

CHAPTER 8
Table 1 Interactions of SDC and ASDC 403
Table 2 The 48 industry group classification by Fama French 1997 412
Table 3 Descriptive Statistics 413
Table 4 The rank order of systemic risk in 48 industry sectors in U.S. market 417
Table 5 The relationship between the measure of systemic risk (alpha=0.1) and macroeconomic variables 420
Table 6 Predictive power of our systemic risk measures. 422

CHAPTER 9
Table 1 Estimation of LTV Equation using Level Data 451
Table 2 Estimation of LTV Equation using Demeaned Data 453
Table 3 Calibrated Parameters 464
Table 4 Average LTV Ratio for Home Mortgage 465
Table 5 Parameter Values for Backward-looking LTV Decision Rule 469

Table 6 Parameter Values for Forward-looking LTV Decision Rule 472

CHAPTER 10

Table 1 Fiscal Assistance Programs: Management Indicators and Their Weights Used 489
Table 2 Selective Fiscal Assistance Programs 493
Table 3 PEUEC Indicators (2011) 496
Table 4 Explanations of the Independent Variables 500
Table 5 Higher Educational Institutions Classified by Type and Location 502
Table 6 Descriptive Statistics of the Regression Analysis 502
Table 7 Employment Rates by Type and Location of Higher Educational Institutions 503
Table 8 The Results of the OLS Estimation of Equation (1) 504
Table 9 Results of the OLS Estimation of Equation (2) with Squared and Interactive Terms 507

CHAPTER 11

Table 1 Summary Statistics 524
Table 2 CSAT Proxies 525
Table 3 Correlations 526
Table 4 Spearman's Rank Correlation Coefficients 526
Table 5 Summary Statistics 527
Table 6 Gender Gap in Employment: Initial Survey 531
Table 7 Gender Gap in Employment: Follow-up Survey 532
Table 8 Gender Gap: Earnings 534
Table 9 Gender Gap in Job Mobility 536
Table 10 Gender Gap in Employment by High School Track 537
Table 11 Gender Gap: Earnings by High School Track 538
Table 12 Gender Gap in Job Mobility by High School Track 539

CHAPTER 12

Table 1 Parameter Values: Exogenously Assigned (Japan) 568
Table 2 Parameter Values: Endogenously Targeted (Japan) 568
Table 3 Parameter Values: Exogenously Assigned (South Korea) 572

Table 4 Parameter Values: Endogenously Targeted (South Korea) 573
Table 5 Counterfactual Experiments 578

CHAPTER 13

Table 1 Auxiliary Moments 616
Table 2 Parameter Estimation 617
Table 3 Counter Factual Analysis 618

List of Figures

CHAPTER 1

Figure 1 Product Innovation and Firm Sales 6
Figure 2 Novelty and the Sales of New Products 8
Figure 3 Sales of New and Existing Products 9
Figure 4 Technology Acquisition and Provision 11
Figure 5 Information Sources 13
Figure 6 Protection Measures for the Innovation Benefit 14
Figure 7 Novelty and Public Financial Support Classified by Firm Size 15
Figure 8 Overview of the Model 19

CHAPTER 2

Figure 1 Market sector real tangible and intangible investment (1974-75 to 2012-13) 2011-12 dollars, chain volume measures 56
Figure 2 Shares of nominal total intangible investment, by asset type (1974-75 to 2012-13) Percent 57
Figure 3 Tangible, intangible and total capital stock, market sector, 1974-75 to 2012-13 2011-12 dollars, chain volume measures 59
Figure 4 Internal rate of return (IRR) for the market sector, all intangibles treated as capital 60
Figure 5 Market sector gross value added, 1974-75 to 2012-13 2011-12 dollars, chain volume measures 65
Figure 6 Capital services, market sector, 1974-75 to 2012-13 Index 1974-75= 100 66
Figure 7 Multifactor productivity, market sector, 1974-75 to 2012-13 (Index 1974-75= 100) 68
Figure 8 Expenditure on business R&D: relative significance of public support, 1993-94 to 2012-13 (2011-12 dollars) 71
Figure 9 The ratio of public support for business enterprise to business R&D spending, 1993-94 to 2012-13 71

Figure 10 Australian Government spending on research and innovation 2012-13 73
Figure 11 Total Australian Government support for research and innovation 1993-94 to 2012-13 74
Figure 12 Breakdown of underpinning research funded by the Commonwealth and State/territory by socio-economic objective, 1992-93 and 2011-12 75
Figure 13 Commonwealth support for R&D, by type of activity, 1992-93 and 2008-09 76
Figure 14 Market sector MFP growth and public support for research agencies, higher education and the business enterprise sector (1993-94 to 2012-13) 77

CHAPTER 4

Figure 1 Scatterplot of MSA employment growth (1993-2002) and small business births (1993) 180
Figure 2 Scatterplot of MSA payroll growth (1993-2002) and small business births (1993) 180
Figure 3 Scatterplot of MSA wage growth (1993-2002) and small business births (1993) 181

CHAPTER 5

Figure 1 Distributions of Markups According to Estimation Models 243
Figure 2 Distributions of Markups over Time 244
Figure 3 Distributions of Exporters and Non-Exporters' Markups 246
Figure 4 Distributions of Small and Large Plants' Markups 247
Figure 5 Distributions of Markups of Plants sorted by Size and Globalisation 248

CHAPTER 6

Figure 1 An Illustration of Market and Product 275
Figure 2 Operating Cost per Available Seat Mile (CASM) 299
Figure 3 Simulation Design for Decomposing Sources of Price Change 301

Figure 4 Measures for the Extent of Product Differentiation: Within-firm distance and Within-market distance 304
Figure 5 Quality Changes of Merged Firm's Products in All Markets By market power 305
Figure 6 Quality Changes of Merged Firm's Products in Oligopoly Markets By product group 308
Figure 7 Quality Changes of Merged Firm's Products in Monopoly Markets 311
Figure 8 Distribution of Market Frequency in QI and QD Markets: Pre-merger vs. Post-merger (Simulated) vs. Post-merger (Actual) 324
Figure 9 Distribution of Market Frequency in All Markets: Pre-merger vs. Post-merger (Simulated) vs. Post-merger (Actual) 325

CHAPTER 7

Figure 1 Screening process 357

CHAPTER 8

Figure 1 Network structure of symmetric information flows and asymmetric information flows. 393
Figure 2 Network structure of SDC that are significant at the certain threshold value among the daily return of the 48 industry sectors over 2000 to 2001 (a) and over 2007 to 2008 (b). (c) and (d) displays the ASDC network for normal period (2000-2001) and financial crisis (2009-2010), respectively. 397
Figure 3 Time evolution of average approximate entropy (ApEn) measure for whole economy, financial sectors, and real economy sectors 400
Figure 4 Relationship between the measure of systemic risk and asymmetric of information flow 409
Figure 5 Evolution of the strength of directed connectedness and its asymmetry: (a) 1 year rolling estimates of the total summa-tion of strength of directed connectedness; (b) Total asymmetry of SDC; (c) linear correlation between SDC and ASDC 411

Figure 6 Time evolution of systemic risk measure: (a) 1 year rolling estimates of the systemic risk; (b) systemic risk of industry sectors such as finance, real economy, real estate, and trading sectors, and it is normalized by the average of systemic risk of 48 industries; (c) ratio of difference of systemic risk between financial and real economy sectors. 416
Figure 7 Time evolution of t-value of beta coefficient 423
Figure 8 The correlation between systemic risk and essential quantity, such as SDC, ASDC according to various α value from 0 to 1. 427
Figure 9 Time evolution of both systemic risk and 3 month T-bill rate. 427
Figure 10 Correlation between SR and 3 month T-bill rate over different periods of time. 428
Figure 11 Time evolution of both systemic risk and 10 year T-bill rate. 428
Figure 12 Correlation between SR and 10 year T-bill rate over different periods of time. 429
Figure 13 Time evolution of both systemic risk and Libor rate 429
Figure 14 Correlation between SR and Libor rate over different periods of time. 430
Figure 15 Time evolution of both systemic risk and S&P500 index return 430
Figure 16 Correlation between SR and S&P500 index return over different periods of time. 431
Figure 17 Time evolution of both systemic risk and volatility of S&P500 index. 431
Figure 18 Correlation between SR and volatility of S&P500 index over different periods of time. 432
Figure 19 Time evolution of both systemic risk and unemployment rate 432
Figure 20 Correlation between SR and unemployment rate over different periods of time. 433

CHAPTER 9

Figure 1	Risk-taking Channel in Mortgage Market	446
Figure 2	Auto Loan and Home Mortgage LTV Ratio in U.S.	448
Figure 3	Trend Components of Auto Loan and LTV Ratio	449
Figure 4	Mortgage LTV Ratio and House Price Growth in U.S.	450
Figure 5	Mortgage LTV Ratio and Federal Fund Rates in U.S.	452
Figure 6	Impulse Response from VAR Analysis	455
Figure 7	Impulse Responses to a Monetary Policy Shock	465
Figure 8	Comparison of Impulse Responses of Consumption	468
Figure 9	Impulse Responses to a Monetary Policy Shock	470
Figure 10	Impulse Response of LTV Ratio	472
Figure 11	Impulse Responses by Forward-looking LTV Decision Rule	474
Figure 12	Impulse Responses to LTV Shock	475

CHAPTER 10

Figure 1	Relationship between per-student expense and the number of students per faculty with the employment rate	509
Figure 2	Interactive effects of per-student expense and the number of students per instructor	510

CHAPTER 12

Figure 1	Calibration Results (Japan)	569
Figure 2	Calibration Results (South Korea)	571
Figure 3	Counterfactual Results: Labor Market Institutions and Labor Productivity	576
Figure 4	Counterfactual Results: Labor Market Institutions and Economic Growth	577

CHAPTER 13

Figure 1		606
Figure 2		609

List of Appendix

CHAPTER 3

Appendix Table 1 Product Cross-Reference from Processor Core to Brand Name (i.e. Marketing Name) in Sample (Q1:2002 - Q4:2005) 163

CHAPTER 4

Appendix Table 1 Impact of Firm Expansion on Urban Growth 214
Appendix Table 4 Homestead Exemption and Firm Expansion 216
Appendix Table 3 2SLS Estimates Using Both Homestead Exemption Variables as Instruments 217

CHAPTER 5

Appendix Table 1 Market Share and Export Effects on DLW Markups in Unbalanced Panel 264
Appendix Table 2 Market Share and Export Effects on DLW Markups in Dynamic Unbalanced Panel 265
Appendix Table 3 Market Share and Export Effects on OLS Markups in Unbalanced Panel 266
Appendix Table 4 Market Share and Export Effects on OLS Markups in Dynamic Unbalanced Panel 267

CHAPTER 1

New-to-Market Product Innovation and Firm Performance: Using Innovation Survey from Japan

by
Daiya Isogawa
(University of Tokyo)
Kohei Nishikawa
(Setsunan University)
Hiroshi Ohashi
(University of Tokyo)

Abstract

This study evaluates the economic impact of new-to-market product innovation in Japan by using firm-level data obtained from the Community Innovation Survey conducted in Japan. It accounts for possible technological spillovers from innovation activities, and examines the extent to which new-to-market product innovation contributes to firm performance. Casual observations from the data reveals that new-to-market product innovation are likely to contribute higher sales for the firm with less cannibalization with existing products, generate higher degree of technological spillovers to other innovations, and be brought by firms that corroborate with universities and other academic institutions. An econometric analysis on simultaneous equations confirms these observations.

1. Introduction

Product innovation is, by definition, deemed novel, but the degree of novelty differs by product (e.g., Arundel and Hollanders, 2005). The OECD (1992, 1996, and 2005) classifies a firm's product innovation into two types: the introduction of a product only new to the firm, and the introduction of a product new to the market. The latter type should be newer and more drastic than the former (OECD, 2009), and thus is called new-to-market product innovation in this paper. The concept of new-to-market product innovation sheds new insights on the existing literature in two folds. First, new-to-market product innovation may contribute in a greater extent to firm performance, as it can provide a firm with temporary market power (Petrin, 2002). Second, new-to-market product innovation may exhibit technological spillovers for innovation activities of other firms, a research topic which has attracted much attention in both theoretical and empirical studies.[1] For example, recent studies of endogenous growth theory (e.g., Grossman and Helpman, 1991, Aghion and Howitt, 1992, Klette and Kortum, 2004) indicate that spillovers from firms at the technological frontier play an important role. If new-to-market product innovation results in significantly positive spillovers, a policy to promote such innovation can be beneficial from a social welfare point of view (Spence, 1984).

This study quantitatively examines the nature of new-to-market product innovation, in order for us to understand its contribution to firm performance, and its possible need for public policy. We use firm-level data obtained from the Japanese National Innovation Survey (JNIS). We propose an econometric model that comprises technological spillovers, legal protection measures, and other important variables relevant to new-to-market product innovation. Our model is reminiscent of that proposed by Crépon et al. (1998) (CDM); however, we address possible endogeneity in our estimation, the issue that is largely neglected in CDM.

1 Arrow (1962) points out that an innovating firm cannot appropriate the outcome of its innovation activities owing to inherent technological spillovers. Accordingly, many researchers have tried to quantify spillovers, especially in terms of the social rate of return on R&D investment, as discussed in Griliches (1992).

Despite its economic importance, few empirical studies focus on the novelty of a firm's product innovation. To the best of our knowledge, Duguet (2006) is an exception. The present study differs from Duguet (2006) in three important ways. First, Duguet (2006) crudely lumps together product and process innovations into one basket, even though the underlying economics between these innovations work differently (e.g., Klepper, 1996). Rather we focus solely on product innovation to make our analysis and its interpretation clear.

Second, we use sales as a measure of firm performance, rather than productivity. It has been argued that productivity may not be an appropriate measure to assess product innovation (e.g., Van Leeuwen and Klomp, 2006, De Loecker, 2011). Third, we utilize technology outflow, as well as inflow, in order to capture the influence of technological spillovers, whereas existing studies including that by Duguet (2006) focus only on the inflow of technology. Incorporating technology outflow provides us with an unbiased picture of technological spillovers in the context of JNIS.

The rest of this paper is organized as follows. Section 2 provides an overview of the various approaches to the measurement of the economic outcomes of innovation, and describes the data set used in the study. Section 3 proposes eight hypotheses related to new-to-market product innovation. Section 3.1 is dedicated to firm performance, Section 3.2 to technological spillovers, and Section 3.3 to the characteristics of a novel innovator. Section 4 formulates and estimates an econometric model to test the formulated hypotheses. Section 5 concludes.

2. Approaches to Evaluating Innovation Outcomes

Demand estimation, patent data and R&D data analyses are three primary quantitative approaches commonly used to measure economic outcomes accrued by product innovation. Demand estimation has been performed on various new products such as computed tomography scanners (Trajtenberg, 1989) and minivans in the automobile market (Petrin, 2002). This method has an advantage of evaluating the impact of product innovation on consumer surplus. Under the category of the

second approach that uses patent data, studies such as those presented by Pakes (1986) and Schankerman (1998) estimate the value of patents, an intermediate input in the overall innovation process. Finally, among the works that have utilized R&D data are studies that estimate the social rate of return on R&D investment (Griliches, 1992).[2] However, while R&D data have an advantage of quantifying the economic impact, R&D investment only accounts for a proportion of a firm's innovation activities (Mairesse and Mohnen, 2010). As pointed by Arundel et al. (2008), many firms conduct innovation activities without reporting their R&D expenditures.[3]

Turning to inputs, the program evaluation technique is popular for exploring innovation policy (see, for example, Almus and Czarnitzki (2003) and González et al. (2005), which focus on R&D subsidies). While this technique helps solve the endogeneity problem of subsidy assignment, however, it ignores how these subsidies affect other firms through spillovers. This shortcoming has fostered a recent trend of empirical studies using innovation surveys based on the Oslo Manual (Mairesse and Mohnen, 2010), which offer a wide range of information on a firm's innovation activities and their outcomes including innovation novelty.

JNIS follows the Oslo Manual, and has a basis on the survey conducted in the period from April 1, 2006 to March 31, 2009. By using the stratified sampling, surveyed firms are selected among those listed in the Establishment and Enterprise Census 2006, which is conducted by the Statistics Bureau, Ministry of Internal Affairs and Communications. They are further restricted to firms with more than 10 employees that operate in the industries in Table 1. The response rate is 30.3%, corresponding to a sample of 4,579 firms. Note that questions about a firm's product innovation are asked for the market in which the firm supplies its staple goods.

---

2 Recent examples of R&D data analyses include those by Bloom et al. (2013), which focuses on the identification of spillovers, and Xu (2006), which captures the dynamic properties of R&D investment.

3 The results of JNIS show that 47.3% of firms conducting innovation activities do not report any R&D expenditures. This percentage is similar to that of Arundel et al. (2008) and that observed in other countries.

Table 1 Classification of Industries in JNIS

Industry	Japan Standard Industrial Classification (Rev. 12)
Agriculture and forestry	A01-02
Fisheries	B03-04
Mining and quarrying of stones and gravel	C05
Construction	D06-08
Manufacturing	E09-32
Electricity, gas, heat supply, and water	F33-36
Information and communications	G37-41
Transport and postal activities	H42-49
Wholesale and retail trade	I50-60
Finance and insurance	J62, 64-67
Real estate and goods rental and leasing	K68-70
Scientific research, professional and technical service	L71-74
Accommodation, eating and drinking services	M75-77
Compound services	Q86
Services, n.e.c.	R89

3. Hypotheses on New-to-Market Product Innovation

In this section, we propose eight hypotheses on new-to-market product innovation from three aspects:. In Section 3.1, we formulate hypotheses on how new-to-market product innovation influences firm performance. Section 3.2 is dedicated to hypotheses on technological spillovers in innovation activities. Lastly, Section 3.3 is associated with innovation policy, where we propose hypotheses on the characteristics of a novel innovator.

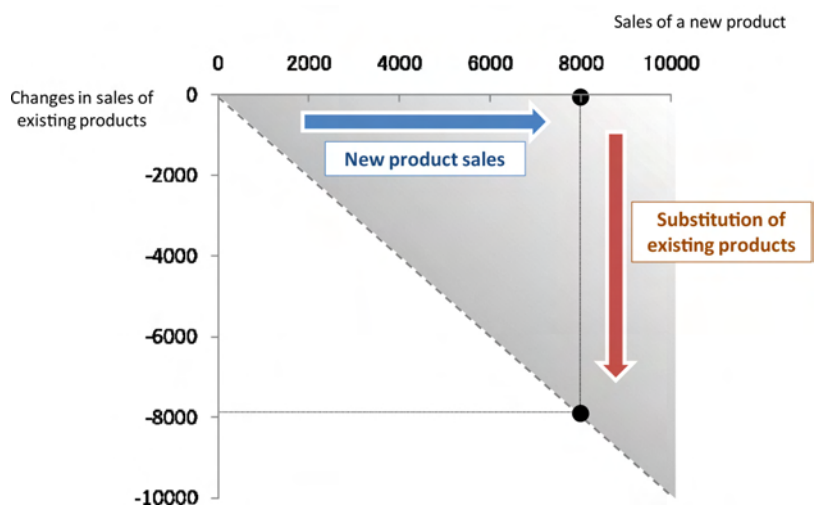
3.1. Firm Performance

To analyze French manufacturing firms, CDM conduct a three-step regression analysis that estimates the research equations, innovation equations, and productivity equations. Their results indicate that firm productivity is positively correlated with innovation output such as number of patents or share of innovation-related sales. Many researchers

in several countries have since taken a CDM-like approach to analyze firm innovation, including Griffith et al. (2006) for France, Germany, Spain, and the UK and Chudnovsky et al. (2006) for Argentina. Furthermore, CDM's approach has been extended in various directions. Jefferson et al. (2006), Lööf and Heshmati (2006), and Van Leeuwen and Klomp (2006) use measures other than productivity to capture firm performance, while Duguet (2006) classifies innovations into radical and incremental ones.

In the present study, we use the sales of a new product and of existing products to measure firm performance in order to decompose product innovation into two affects. Figure 1 summarizes the economic impact of product innovation on firm's sales. The horizontal axis represents the effect on a new product, which is measured by its sales. The vertical axis captures the cannibalization effect, which means that the new product competes with the firm's existing products and thus reduces their sales. The net effect of product innovation on a firm's total sales, which is determined by the difference in the degree of these two effects, is shown gradationally in the figure. The net effect is zero on the 45-degree line, but this becomes positive with lighter graduation.

Figure 1 Product Innovation and Firm Sales



The sales of a new product are considered to be affected by those of existing products in the market. For new-to-firm product innovation (i.e., the introduction of a product already provided by other firms), the firm faces severe competition particularly in the market of a homogeneous product. As a result, the price of the new product will drop and its sales will decrease. Consistent with this view, Duguet (2006) shows that only radical innovations can improve firm performance. Barlet et al. (1998) also indicate that innovation novelty can increase the share of innovation-related sales in situations where technology is important. Therefore, we propose the following hypothesis.

Hypothesis 1: The sales of a new product are larger for a firm with new-to-market product innovation than for one with new-to-firm product innovation on average.

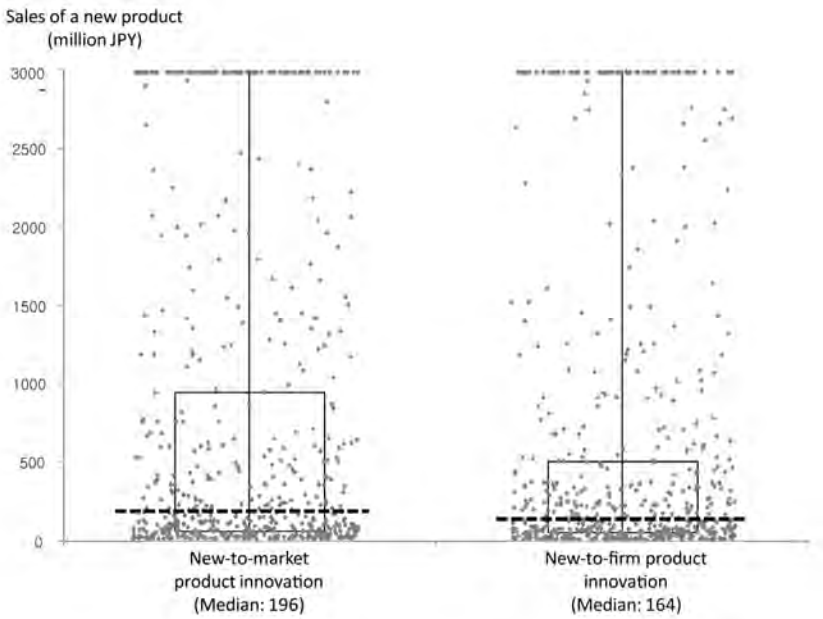
According to the sales information in JNIS,[4] the average sales of a new product in FY2008 were 5,586 million JPY for a firm with new-to-market product innovation and 3,004 million JPY for other firms, which is consistent with Hypothesis 1. In addition, Figure 2 boxplots the sales of a new product for a firm with new-to-market product innovation and for one with new-to-firm product innovation. The rectangle in this figure represents the interval between the 25th and 75th percentiles of sales and the dashed line represents the median. Median sales are 185 million JPY for a novel innovator and 165 million JPY for other innovators. Furthermore, the 75th percentile of sales for a novel innovator is much larger than that for other firms, which implies that some novel product innovations generate huge sales.

Next, we turn to the sales of an innovator's existing products. Jefferson et al. (2006) point out that innovation does not necessarily improve firm performance, and suggest that cannibalization with the firm's existing products may severely deteriorate the firm's profitability. This is expressed by the following two hypotheses.

---

4 To be exact, JNIS asks a firm about the share of its new product sales. However, we can recover the sales amount of the new product by multiplying the share by the firm's total sales in FY2008. Since the question about share is an interval one, we assign the intermediate value for each interval.

Figure 2 | Novelty and the Sales of New Products



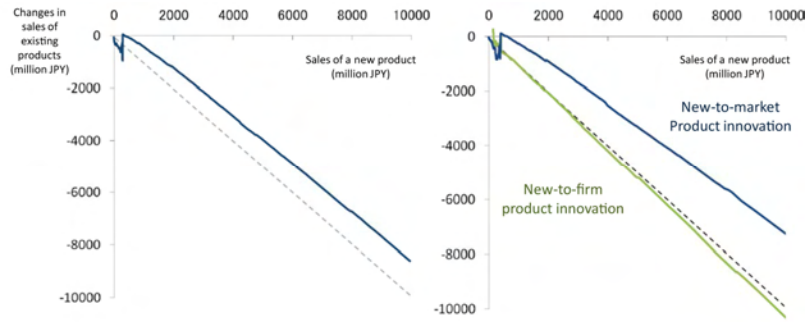
Hypothesis 2: The larger sales of a new product decrease the sales of a firm's existing products on average.

Hypothesis 3: The decrease in the sales of a firm's existing products due to the larger sales of a new product is less for a firm with new-to-market product innovation than for one with new-to-firm product innovation on average.

To test Hypotheses 2 and 3, we need to capture the effect of product innovation on the sales of a firm's existing products. For this purpose, we calculate changes in the sales of existing products during FY2006-FY2008. The left side of Figure 3 plots the relationship between the sales of a new product and changes in the sales of existing products.[5] Larger sales of a new product decrease those of existing products, which is consistent with Hypothesis 2. Changes in total sales, which are

5 We use LOWESS (Locally Weighted Scatterplot Smoothing) as the smoothing algorithm.

Figure 3 | Sales of New and Existing Products



represented gradationally in Figure 1, are around 1,500 million JPY regardless of the new product sales.

The right side of Figure 3 plots the same relationship separately for a firm with new-to-market product innovation and for one with new-to-firm product innovation. There exists a significant difference between them. The curve for a firm with new-to-firm product innovation is almost on the 45-degree line, which indicates that the sales of a new-to-firm product are nearly offset by the decrease in the sales of existing products. On the contrary, the curve for a firm with new-to-market product innovation lies above the 45-degree line, which means that the sales of a new-to-market product lead to an increase in the firm's total sales. These observations are consistent with Hypothesis 3. Combining the observations in Figure 2 and Figure 3 suggests that novel product innovation can increase the sales of a new product and reduce the loss of existing product sales, both of which increase a firm's total sales.

3.2. Technological Spillovers

Many researchers including Arrow (1962) have pointed out that an innovating firm cannot appropriate the outcomes of its innovation activities because of the existence of technological spillovers. In contrast to some previous studies (e.g., Bloom et al., 2013), we 3directly identify technological spillovers from information on a firm's technology acquisition (i.e., inflow) and provision (i.e., outflow). For outflow, of special importance is the technology provision through channels that are

less likely to be accompanied by monetary compensation such as open sourcing and participation in consortia. If a firm does not consider this type of spillover when making decisions on innovation activities, innovation in the private sector could be undersupplied.

Furthermore, some recent studies of endogenous growth theory (e.g., Grossman and Helpman, 1991, Aghion and Howitt, 1992, Klette and Kortum, 2004) and those on dynamic estimation (e.g., Xu, 2006) assume technological spillovers from firms at the technological frontier through nonmonetary channels. Since a firm with new-to-market product innovation is likely to lie near the frontier, we propose the following hypothesis.

Hypothesis 4: A firm with new-to-market product innovation is more likely to provide its technology through open sourcing or participation in consortia than one with new-to-firm product innovation on average.

Of the empirical studies that focus on inflow, Kaiser (2002) considers incoming spillover effects to analyze the relationship between research cooperation and research expenditures. His results indicate that horizontal spillovers lead to a firm's aggressive innovation investment through research cooperation. Similarly, Branstetter and Sakakibara (2002) examine research consortia based on the approach taken by Katz (1986), finding that spillover effects in research consortia have a positive impact on a firm's outcomes. These findings suggest the following hypothesis.

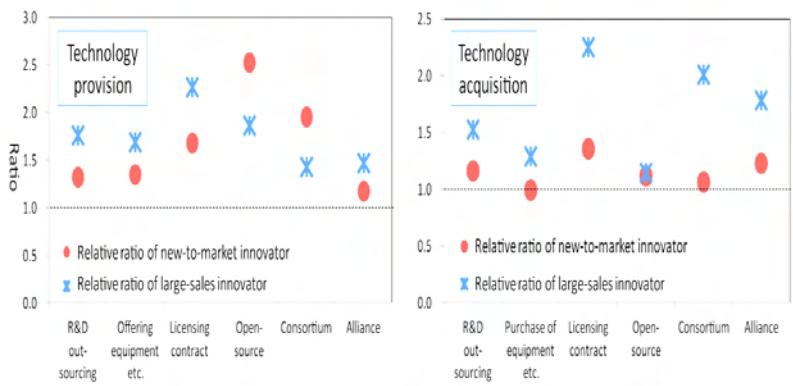
Hypothesis 5: The sales of a new product are larger for a firm that acquires technology through consortia than for one that does not on average.

Figure 4 summarizes a firm's technology acquisition and provision based on the information provided by JNIS. The circles in the figure represent the relative ratio of technology acquisition (or provision) for a firm with new-to-market product innovation to one with new-to-firm innovation by channel, while the snowflakes are the relative ratio of technology acquisition (or provision) for a firm with above-median

product innovation to one with below-median product innovation by channel.[6] The left side of the figure is for a firm's technology provision. While above-median product innovation is associated with channels that are more likely to be accompanied by monetary compensation (e.g., licensing), new-to-market product innovation is linked to nonmonetary channels such as open sourcing and participation in consortia, which is consistent with Hypothesis 4.

The right side of the figure is for a firm's technology acquisition. Innovation novelty seems to have little association with technology acquisition, whereas a firm with above-median product innovation tends to acquire technology through licensing and participation in consortia, which is consistent with Hypothesis 5. By combining this observation with the results on the left side of the figure, we can suggest that participation in consortia plays a significant role in technological spillovers. Indeed, Figure 4 indicates that a firm with new-to-market product innovation provides its technology to other firms through consortia and that the spilled over technology contributes to their high sales of a new product.

Figure 4 | Technology Acquisition and Provision



6 The median amount here relates to sales (i.e., 168 million JPY).

3.3. Other Characteristics of New-to-Market Product Innovation

In this subsection, we focus on the characteristics of a firm with new-to-market product innovation from three aspects: information sources, means of protecting the innovation benefit, and public financial support. Since the previous subsections have implied that novel product innovation leads to significant improvements in firm performance and technological spillovers, public policy for encouraging a firm's new-to-market innovation should work well. To implement such a policy effectively, however, we need to know what types of firms are novel product innovators.

3.3.1. Information Sources

Previous studies have examined the relationship between information sources and innovation novelty. Belderbos et al. (2004) examine the relationship between cooperative R&D and firm performance, showing that information from consumers or universities has a positive impact on the sales of a new product and that cooperation with universities leads to novelty. Similarly, Mohnen and Hoareau (2003) study the relationship between contact with universities and innovation novelty, but their results suggest that such contact does not necessarily result in cooperation. With a few exceptions,[7] most studies imply that information from universities positively affects innovation novelty, and thus we propose the following hypothesis.

Hypothesis 6: A firm with new-to-market product innovation is more likely to obtain information from universities for its innovation activities than one with new-to-firm product innovation on average.

Figure 5 shows the utilization ratio of information sources for innovation activities. As explained above, the circles are for innovation

7 Monjon and Waelbroeck (2003) suggest that information from universities encourages innovation that is not radical. In this regard, since they also show that cooperation with foreign universities leads to radicalness, there may be some link between contact with universities and radicalness.

Figure 5 Information Sources

novelty and snowflakes are for the sales of a new product. Whereas a firm with above-median product innovation aggressively uses various sources, one with new-to-market product innovation tends to obtain information from universities or patents held by other firms, which is consistent with Hypothesis 6.

3.3.2. Means of Protecting the Innovation Benefit

While we have already noted the difficulty of appropriating the innovation benefit, a firm can protect it by legal means such as patent protection and non-legal means such as trade secrets. Theoretically, legal means can serve to encourage a firm's innovation activities by giving it a premium for creating innovations. Among recent empirical studies, Duguet and Lelarge (2006), for example, show the effectiveness of patent protection for defending a firm's product innovation. However, legal protection cannot necessarily prevent firms from circumventing inventions (Levin et al., 1987). In particular, considering the possible positive spillovers to novel product innovation, legal means may not effectively protect the profit from the new-to-market product, which leads to the following hypothesis.

Hypothesis 7: A firm with new-to-market product innovation is no more likely to use legal protection relative to non-legal protection than one with new-to-firm product innovation on average.

Figure 6 | Protection Measures for the Innovation Benefit

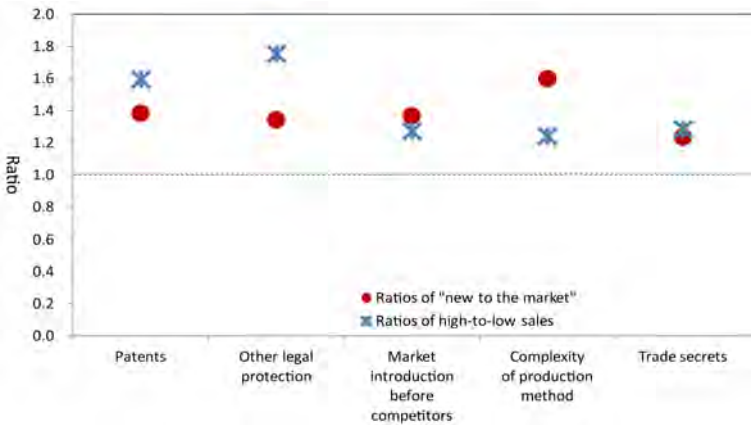


Figure 6 summarizes means of protecting a firm's innovation benefit. As before, circles are for innovation novelty and snowflakes sales of a new product. While a firm with above-median innovation tends to use legal protection actively, one with new-to-market product innovation uses it only as frequently as it applies non-legal means. This finding is consistent with Hypothesis 7 and indicates that legal means may find it difficult to protect novel product innovation relative to above-median product innovation.

3.3.3. Public Financial Support

Lastly, we select public financial support for a firm's innovation activities. This topic has mainly been studied in terms of the relationship between R&D subsidies and a firm's R&D investment. For example, Almus and Czarnitzki (2003) use a matching method and show that R&D subsidies stimulate firms' innovation activities. González et al. (2005) also indicate that some firms would not invest into R&D without subsidies, and that there exists no crowding out of private R&D investment. In addition, some recent studies have drawn attention to public financial support other than subsidies. Finger (2008), for instance, examines the effect of an R&D tax credit by considering the interdependence of firms' R&D investment and shows that such a tax credit encourages a firm's R&D investment in a limited way.

Meanwhile, of the few studies of the relationship between public financial support and innovation novelty, Mohnen and Hoareau (2003) raise the possibility that contact with public institutions, by using them as information sources, leads to novel product innovation. If contact with public institutions through channels other than information sources were also to encourage novel product innovation, public financial support could have a positive effect on novelty. Hence, we propose the final hypothesis.

Hypothesis 8: A firm with new-to-market product innovation is more likely to receive public financial support than one with new-to-firm product innovation on average.

Figure 7 plots the relationship between the ratio of firms with new-to-firm product innovation and public financial support[8] by firm size.[9] While this ratio is higher for middle-sized and large firms, this is not the case for small firms. Hence, Hypothesis 8 is not necessarily confirmed, perhaps because nonfinancial bottlenecks exist for small firms that are

Figure 7 Novelty and Public Financial Support Classified by Firm Size

8 Financial support includes tax credits, subsidies, loan guarantees, and so on.

9 Small firms have fewer than 50 employees, middle-sized firms have 50-249 employees, and large firms have more than 250 employees.

attempting to be novel innovators. In particular, small firms are less likely to use information from universities (Nishikawa et al., 2010), which discourages new-to-market innovation as implied in Figure 5. Therefore, policy intervention intended to increase contact between firms and universities may work well.

4. Econometric Analysis

In the previous section, we proposed hypotheses on a firm's novel product innovation and confirmed that most observations in JNIS seem to be consistent with them. However, concluding only from such data descriptions is inadequate for two reasons. First, omitted variable bias could be present. A firm's innovation activities and their outcomes are affected by a number of factors, which if correlated with an object of interest and not controlled for properly would lead us to draw the wrong conclusions. Second, endogeneity biases are also of concern. Since ignoring endogeneity in variables can distort the estimation results, we must overcome this issue by using techniques such as the instrumental variable method.

In this section, we present an econometric analysis that deals with these problems (Section 4.1) and thus allows us to test the hypotheses previously formulated (Section 4.2). In particular, we construct and estimate a comprehensive econometric model that is a variant of the one proposed by CDM.

4.1. Econometric Model and Estimation

The proposed model is represented by a system of three sets of equations. The first is for a firm's R&D investment. As is widely known, R&D expenditures are endogenously determined and any analyses ignoring this endogeneity may suffer from biased estimates. CDM deal with this issue by formulating research equations. In line with CDM and other empirical studies, we consider several factors that may affect a firm's R&D investment. The first factor is related to consumer demand, expressed by market size herein. Demand structure is considered to be a

major determinant of a firm's innovation activities (e.g., Levin and Reiss, 1984), which is often called *demand pull*. While CDM base their analysis on the influence of market demand, we control for the market size effect by using industry dummies[10] and a dummy variable that indicates whether the market expanded during the survey period.

Another factor considered a fundamental determinant of innovation activities is technological opportunity (e.g., Rosenberg, 1974, Levin and Reiss, 1984) or *technology push*. To capture this effect, we focus on a firm's technology acquisition, which is also interpreted as the inflow of technological spillovers as noted in Section 3. Specifically, we create technology acquisition dummy variables based on the answers to the JNIS question on through which channels a respondent acquires its technology (see the right side of Figure 4).[11]

In addition, we take into account information sources. Some past studies such as Belderbos et al. (2004) have focused on information sources to measure the inflow of technological spillovers. Again, JNIS asks a respondent which information sources it uses (Figure 5), the answers to which we use to create information dummy variables. Besides, demand pull or technology push, CDM also explore those factors involved in the so-called "Schumpeterian hypotheses" on the effect of firm size and market power.[12] Similar to them, we use firm size dummies, the number of competitors in the domestic market,[13] and a dummy variable that indicates whether the market experienced product diversification during the survey period.

Lastly, we consider public financial support for a firm's innovation

---

10 The industry classification used is the same as that defined in Section 2.1.

11 CDM use the answers to a question about the influence of technological developments.

12 There exists a long history of dispute over whether market concentration encourages a firm's innovation activities. It is said that a firm's innovation has two distinct effects. One is the replacement effect (Arrow, 1962), which encourages firms' innovation activities in more competitive situations, and the other is the efficiency effect, also called the Schumpeterian effect (Schumpeter, 1943; see also Gilbert and Newbury, 1982, Reinganum, 1983), which encourages such activities in more concentrated situations. Several empirical studies, including Aghion et al. (2005), have tried to quantify the net effect of these two types of effects.

13 This number is for FY2008. Since the corresponding question is an interval one, we assign the intermediate value to each interval.

activities, the issue which are not covered in CDM. As described in Section 2, a number of studies have sought to identify the effect of public aid on firm innovation. We thus create a dummy variable that indicates whether a firm receives any public financial support from local public agencies or the central government (coded 1) or not (coded 0).

The second set of equations captures a firm's innovation output. As the measure of output, we focus on innovation novelty, which is analyzed by Duguet (2006) but not by CDM, and the protection of the innovation benefit, which CDM proxy for by using number of patent applications. However, for the latter, we do not restrict our attention to patents since firms use various means of protecting their innovation benefits including both legal protection and nonlegal protection such as market introduction before competitors, the complexity of production methods, and trade secrets. We therefore identify whether a firm uses legal or nonlegal protection methods and create a dummy variable for each. For the regressors, we use a similar set of variables as that adopted in the first step. We add a firm's R&D expenditures, which are endogenously determined in the first stage since many empirical studies including CDM consider a firm's R&D investment to be an innovation input. Moreover, we omit the number of competitors in the domestic market from this stage, just as CDM omit market share from their second one. In addition to these variables, we use innovation novelty as a regressor to explain the protection of the innovation benefit (Hypothesis 7).

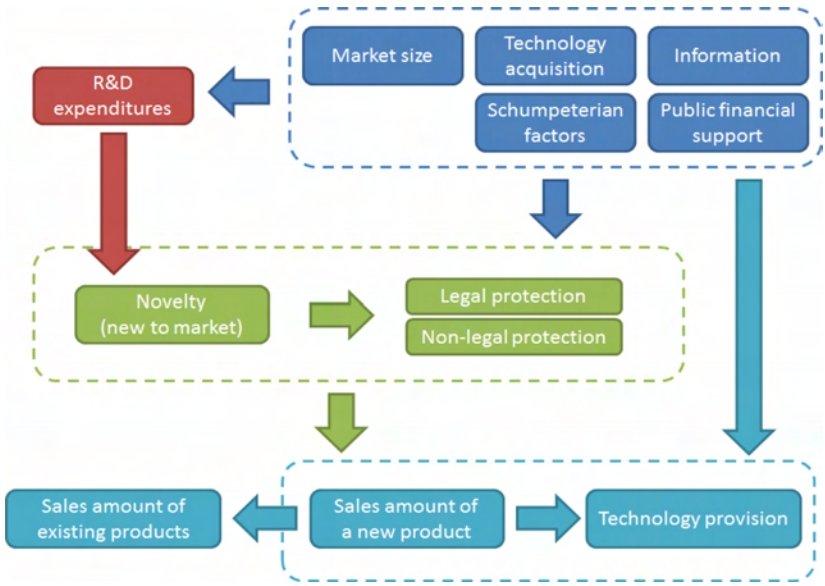
The third set of equations is for a firm's sales and its technology provision. For the former, we consider not only the sales of a new product but also those of existing products, which are important for examining the economic outcomes of product innovation because they can capture the cannibalization effect. For a firm's technology provision, we focus on channels that are less likely to accompany monetary compensation. In particular, we create a dummy variable that takes one if the firm provides its technology through the channels of open sourcing and participation in consortia.

We include three types of regressors for the equations that determine the sales of a new product and technology provision. First, we include innovation novelty and the protection of the innovation benefit, which are endogenously determined it the first stage as above. Following CDM

and Duguet (2006), these innovation outputs may have a positive impact on firm performance. Second, we use the same regressors as adopted in the second stage as control variables. As a result, we control for the effect of demand and technological conditions, firm size (number of employees), and product diversification. Third, we consider the acquisition of tangible fixed assets and number of R&D workers, which correspond to regressors in the third stage of CDM.[14] On the contrary, for the regressors in the equation that determines the sales of existing products, we consider innovation novelty, the sales of a new product and the control variables of a firm's total sales in FY2006 and the firm size and industry dummies. With this equation, we aim to quantify the degree of cannibalization and examine how innovation novelty affects this degree.

The structure of the model is summarized in Figure 8. We can statistically test all of the hypotheses in Section 3 based on this model.

Figure 8 Overview of the Model



14 CDM include physical capital and the shares of engineers and administrators in the total number of employees.

4.1.1. Comparison with CDM

Although our model is a variant of CDM, there are four significant differences other than the practical measures of the model variables. First, we incorporate innovation novelty into the model. As stated in Section 1, it is important to discuss product innovation in terms of its novelty since new-to-market product innovation could affect firm performance strongly and be associated with technological spillovers. Second, we consider both legal and nonlegal means of protecting the innovation benefit given the insufficiency of using patent protection only (Levin et al., 1987). Third, we separately consider the firm's sales of both new and existing products as measures of firm performance. While CDM consider the percentage of a firm's innovation-related sales in their second stage, which is the combination of a firm's sales of new and existing products, only examining a firm's sales of a new product and that of existing ones separately may not be able to capture the cannibalization effect. Fourth, we consider both the inflow and the outflow of technology by using information on a firm's acquisition and provision of technology. In particular, most studies including CDM have not included outflow in their analyses.

4.1.2. Estimation Equations

We propose estimation equations for firm *i* based on the presented model. Equation (1) corresponds to the first part of the model, the determination of a firm's R&D expenditures. Because there are many firms with zero R&D expenditures, our choice is a Tobit model.

$$R\&D_i^* = x_{1,i}\,\beta_1 + u_{1,i}, \tag{1}$$

$$R\&D_i = \begin{cases} R\&D_i^* \; if \; R\&D_i^* > 0, \\ 0 \; otherwise \end{cases}$$

where $R\&D_i$ represents a firm's R&D expenditures and $x_{1,i}$ includes the industry, market expansion, technology acquisition, information, firm size, product differentiation, and public financial support dummies as

well as the number of competitors in the domestic market.

Equations (2), (3), and (4) correspond to the second part. Since all of the dependent variables are binary, we choose a probit model.

$$Novelty_i = \alpha_2 R\&D_i + x_{2,i}\beta_2 + u_{2,i},$$
$$where\ u_{2,i}, \sim N(0,1)\ and\ Novelty_i = \begin{cases} 1\ if\ Novelty_i^* > 0, \\ 0\ otherwise. \end{cases} \quad (2)$$

$$Legal_i = \gamma_3\ Novelty_i + x_{2,i}\beta_3 + u_{3,i},$$
$$where\ u_{3,i}, \sim N(0,1)\ and\ Legal_i = \begin{cases} 1\ if\ Legal_i^* > 0, \\ 0\ otherwise. \end{cases} \quad (3)$$

$$Nonlegal_i = \gamma_4\ Novelty_i + x_{2,i}\beta_4 + u_{4,i},$$
$$where\ u_{4,i}, \sim N(0,1)\ and\ Novelty_i = \begin{cases} 1\ if\ Nonlegal_i^* > 0, \\ 0\ otherwise. \end{cases} \quad (4)$$

where $Novelty_i$ represents innovation novelty, $legal_i$ is the legal protection dummy, $Nonlegal_i$ is the nonlegal protection dummy, and $x_{2,i}$ is similar to $x_{1,i}$ except that it does not include the number of competetors in the domestic market. [15]

Equations (5) to (7) correspond to the third part. For the technology provision equation, we estimate its parameters based on a probit model.

$$log(Newsales_i) = \alpha_5 R\&D_i + [Novelty_i, Legal_i, Nonlegal_i]\eta_5 + x_{5,i}\beta_5 + u_{5,i}, \quad (5)$$

$$log(Existingsales_i) = [Novelty_i, Newsales_i, Novelty_i * Newsales_i]\rho_6 + x_{6,i}\beta_5 + u_{6,i}, \quad (6)$$

$$Provision_i^* = \alpha_7 R\&D_i + [Novelty_i, Legal_i, Nonlegal_i]\eta_7 + x_{5,i}\beta_7 + u_{7,i},$$
$$where\ u_{7,i}, \sim N(0,1)\ and\ Provision_i = \begin{cases} 1\ if\ Provision_i^* > 0, \\ 0\ otherwise. \end{cases} \quad (7)$$

15 We omit a firm's R&D expenditures from Equations (3) and (4) to avoid a convergence problem.

where $Newsales_i$ represents the sales of a new product, Existingsales$_i$ the sales of existing products, Provision$_i$ is the dummy for capturing technology provision through open sourcing or participation in consortia, $x_{5,i}$ includes $x_{2,i}$, purchased tangible fixed assets, and the number of workers in R&D, and $x_{6,i}$ includes the logarithm of a firm's total sales and the firm size and industry dummies.

4.1.3. Methodology and Estimation Sample

We estimate the parameters in this system by using maximum likelihood estimation. Estimation samples are restricted to firms that conduct innovation activities and achieve product innovation, which reflects our interest in innovation output including the novelty of product innovation. This restriction does not become a problem as long as we focus on the economic impact of product innovation *conditional* on a firm conducting innovation activities and achieving product innovation. Similarly, CDM's main estimates are obtained with firms that achieve some kind of innovation.

Furthermore, we omit observations with missing values for the model variables. The characteristics of the omitted firms are similar to those without such missing values.[16] The resulting sample size is 539.[17] Table 2 presents the summary statistics of the model variables.

Regarding the sample selection issue, we attempt to correct for possible sampling biases by a following method. First, for all firms in JNIS, we regress a dummy variable that indicates whether the firm is included in our estimation samples on certain control variables, including a firm's total sales, sales cost, total wages, and the firm size and industry dummies. Then, we calculate the residual for each firm and include them in Equations (1) to (7) as an additional regressor. The estimation results differ little from the baseline results reported next.

---

16 There is little difference in the average size, age, and industry of the sample firms. However, we cannot reject the hypothesis that there is no difference in the average sales and age of the two subsamples based on the t-test results. Similarly, we cannot reject the hypothesis that the existence of missing values and the firm's industry classification are independent based on the results of Pearson's chi-square test.

17 Before omitting observations with missing values, the sample size was 1,224.

Table 2 Summary Statistics

		Mean	Std. Dev.
Novelty		47.40%	50.00%
Sales of a new product	(million JPY)	5148.1	53945.3
Sales of existing products	(million JPY)	42354.8	188152.8
R&D expenditure	(million JPY)	4508	41395.2
Firm size			
	Middle-sized	24.90%	43.30%
	Large	62.80%	48.40%
Number of competitors		10.2	7.64
Product differentiation		61.97%	48.57%
Acquisition of tangible fixed assets	(million JPY)	7179.3	47235.0
No. of workers in R&D		202.2	1374.6
Information			
	Enterprise group etc.	77.50%	41.80%
	Suppliers	57.90%	49.40%
	Customers or clients	68.50%	46.50%
	Competitors	36.40%	36.40%
	Private research institutes etc.	24.20%	48.20%
	Universities etc.	34.20%	47.50%
	Public research institutes	28.60%	45.20%
	Academic conference etc.	36.40%	48.20%
	Professional publications etc.	43.20%	49.60%
	Exhibitions etc.	53.70%	49.90%
	Patent information	37.50%	48.50%
Technology acquisition			
	Buyout	9.70%	29.60%
	R&D outsourcing	37.00%	48.30%
	Purchase of equipment etc.	51.30%	50.00%
	Company split-up	5.30%	22.40%
	Licensing contract	20.50%	40.40%
	Open sourcing	13.40%	34.10%
	Consortium	11.70%	32.20%
	Alliance	16.30%	37.00%
	Accepting researchers etc.	16.30%	37.00%

Table 2 (Continue)

		Mean	Std. Dev.
Technology provision			
	Open sourcing or consortia	11.70%	32.20%
Public financial support		26.20%	44.00%
Protection			
	Legal means	53.80%	49.90%
	Nonlegal means	72.00%	45.00%
Observations		539	

4.2. Estimation Results

Table 3 shows the estimates for Equation (1). Specification (1-a) includes all the regressors discussed in Section 4.1. For the demand side, market expansion is significantly estimated to increase R&D expenditures. On the contrary, few dummies for technology push are significant, except that technology acquisition through corporate reorganization (e.g., buyout, split-up) or open sourcing positively affects a firm's R&D investment. Schumpeterian factors are estimated to have little effect on a firm's R&D investment, which implies that they do not directly determine a firm's innovation activities once we control for demand pull and technology push. The coefficient of public financial support is significantly positive.

Specifications (1-b) and (1-c) omit the industry dummies and technological factors whose coefficients are estimated as insignificant in (1-a). The results are similar to those in (1-a), with the only difference that the coefficient of the large-firm dummy is estimated to be significantly positive. Our results are consistent with the findings of Cohen and Klepper (1996) and Klepper (1996), who argue that firm size has a positive impact on innovation activities.

Table 3 Estimation Results of Equation (1)

		Tobit model Dependent variable: R&D expenditures (million JPY)		
		(1-a)	(1-b)	(1-c)
Market expansion		8275.22 **	8124.01 **	8135.44 **
	(s.e.)	(4020.59)	(4012.51)	(3965.68)
Technology acquisition	Buyout	15914.05 **	16204.31 **	19139.71***
	(s.e.)	(7053.88)	(6984.60)	(6625.08)
	R&D outsourcing	-2149.15	-2395.67	
	(s.e.)	(4546.19)	(4529.89)	
	Purchase of equipment etc.	-2119.86	-1931.71	
	(s.e.)	(4211.06)	(4182.13)	
	Company split-up	39097.56***	39021.40***	40387.06***
	(s.e.)	(9164.63)	(9152.60)	(8811.41)
	Licensing contract	828.84	848.65	
	(s.e.)	(5234.19)	(5219.32)	
	Open sourcing	13447.71**	13000.43**	14746.31***
	(s.e.)	(5648.86)	(5619.70)	(5167.44)
	Consortium	5190.82	5197.15	
	(s.e.)	(6238.81)	(6204.72)	
	Alliance	7539.55	7107.43	
	(s.e.)	(5582.68)	(5529.69)	
	Accepting researchers etc.	2857.23	2606.04	
	(s.e.)	(5195.53)	(5184.03)	
	Enterprise group etc.	-185.12	-609.43	
Information	(s.e.)	(4735.60)	(4720.39)	
	Suppliers	-2704.37	-3352.89	
	(s.e.)	(4016.86)	(3949.60)	
	Consumers or clients	2703.18	3474.55	
	(s.e.)	(4467.36)	(4417.88)	
	Competitors	1218.17	1059.49	
	(s.e.)	(4205.58)	(4188.76)	
	Private research institutes etc.	1655.63	1186.53	
	(s.e.)	(4536.11)	(4480.14)	
	Universities etc.	1234.78	1885.10	
	(s.e.)	(5068.91)	(5022.86)	

Table 3 (Continue)

		Tobit model Dependent variable: R&D expenditures (million JPY)		
		(1-a)	(1-b)	(1-c)
	Public research institutes	3732.63	3876.83	
	(s.e.)	(5142.44)	(5120.27)	
	Academic conference etc.	-5991.11	-5729.08	
	(s.e.)	(5087.50)	(5045.53)	
	Professional publications etc.	2075.06	1701.04	
	(s.e.)	(4976.04)	(4932.46)	
	Exhibitions etc.	-5902.77	-5369.79	
	(s.e.)	(4606.41)	(4568.37)	
	Patent information	5822.03	6718.57	
	(s.e.)	(4691.64)	(4613.64)	
Firm size	Middle-sized	5153.42	6686.65	5862.78
	(s.e.)	(7529.56)	(7370.43)	(7303.05)
	Large	9945.24	11271.57*	12464.83*
	(s.e.)	(6957.73)	(6783.30)	(6600.65)
Number of competitors		179.30	123.38	116.50
	(s.e.)	(248.80)	(243.08)	(241.18)
Product differentiation		-1118.27	-1771.30	-2960.48
	(s.e.)	(4078.83)	(4049.63)	(3957.21)
Public financial support		7638.40*	7543.09*	9736.94**
	(s.e.)	(4554.47)	(4488.56)	(4027.41)
Industry dummies		Yes	No	No

Notes: ***, **, and * indicate that the estimate is significant at 1%, 5%, and 10%, respectively.

Table 4 shows the estimates for Equation (2). Specification (2-a) includes all the regressors discussed in Section 4.1. Interestingly, R&D expenditures have no significant impact on the achievement of new-to-market product innovation. This result contrasts with that put forward by Duguet (2006), who finds a positive impact of a firm's formal R&D on radicalness. One reason for this discrepancy is that Duguet (2006) does not fully control for the effect of demand and technological opportunity as we do in the presented analysis; hence, the estimated coefficient of R&D in Duguet (2006) might be confounded by the effects of other factors. While we find no positive impact of market

expansion on innovation novelty, some of the coefficients of technology acquisition and information are significant. In particular, technology acquisition through accepting new researchers and information from universities seems to positively affect innovation novelty, the latter of which is consistent with Hypothesis 6. Similar to the results of previous studies, universities seem to be influential information sources for novel innovations. Lastly, public financial support has no significant impact on novel innovators, which rejects Hypothesis 8. This finding might be partly because nonfinancial factors, including the utilization of information from universities, are essential for new-to-market innovation, as described in Section 3.3.3.

Specifications (2-b) and (2-c) omit the industry dummies and technological factors whose coefficients are estimated as insignificant in (2-a). The basic implications of the results are the same as those from (2-a).

Table 4 Estimation Results of Equation (2)

		Probit model Dependent variable: Innovation novelty		
		(2-a)	(2-b)	(2-c)
R&D expenditures		5.04E-06	5.46E-06	8.07E-06
	(s.e.)	(5.24E-06)	(5.19E-06)	(4.97E-06)
Market expansion		0.01	-0.02	0.03
	(s.e.)	(0.13)	(0.13)	(0.12)
Technology acquisition	Buyout	0.39	0.37	
	(s.e.)	(0.24)	(0.24)	
	R&D outsourcing	0.13	0.12	
	(s.e.)	(0.14)	(0.14)	
	Purchase of equipment etc.	-0.05	-0.07	
	(s.e.)	(0.13)	(0.13)	
	Company split-up	-0.46	-0.49	
	(s.e.)	(0.34)	(0.34)	
	Licensing contract	0.19	0.17	
	(s.e.)	(0.17)	(0.16)	
	Open sourcing	0.06	0.07	
	(s.e.)	(0.19)	(0.19)	
	Consortium	0.28	0.25	
	(s.e.)	(0.20)	(0.20)	

Table 4 (Continue)

		Probit model Dependent variable: Innovation novelty		
		(2-a)	(2-b)	(2-c)
	Alliance	0.18	0.14	
	(s.e.)	(0.18)	(0.18)	
	Accepting researchers etc.	0.29*	0.28*	0.33**
	(s.e.)	(0.17)	(0.16)	(0.16)
Information	Enterprise group etc.	0.24	0.21	
	(s.e.)	(0.15)	(0.15)	
	Suppliers	-0.11	-0.07	
	(s.e.)	(0.13)	(0.12)	
	Consumers or clients	0.12	0.09	
	(s.e.)	(0.14)	(0.14)	
	Competitors	-0.16	-0.17	
	(s.e.)	(0.13)	(0.13)	
	Private research institutes etc.	-0.09	-0.15	
	(s.e.)	(0.15)	(0.14)	
	Universities etc.	0.39**	0.34**	0.32**
	(s.e.)	(0.16)	(0.16)	(0.15)
	Public research institutes	-0.40**	-0.34**	-0.33**
	(s.e.)	(0.16)	(0.16)	(0.15)
	Academic conference etc.	-0.15	-0.11	
	(s.e.)	(0.16)	(0.16)	
	Professional publications etc.	-0.25	-0.26*	-0.26*
	(s.e.)	(0.16)	(0.16)	(0.14)
	Exhibitions etc.	0.02	0.02	
	(s.e.)	(0.15)	(0.14)	
	Patent information	0.28*	0.30**	0.29**
	(s.e.)	(0.15)	(0.15)	(0.14)
	Middle-sized	-0.08	-0.02	-0.02
Firm size	(s.e.)	(0.23)	(0.23)	(0.22)
	Large	-0.35	-0.25	-0.19
	(s.e.)	(0.22)	(0.21)	(0.20)
Product differentiation		0.18	0.14	0.13
	(s.e.)	(0.13)	(0.13)	(0.12)
Public financial support		-0.11	-0.02	0.00
	(s.e.)	(0.15)	(0.14)	(0.14)
Industry dummies		Yes	No	No
Exogeneity test	(Wald)	0.01	0.02	0.29

Notes: ** and * indicate that the estimate is significant at 5% and 10%, respectively.

Table 5 reports the estimated coefficients for Equations (3) and (4).[18] Specifications (3-a) and (4-a) include all the regressors discussed in Section 4.1 except for a firm's R&D expenditures and industry dummies.[19] On the contrary, specifications (3-b) and (4-b) additionally omit the technological factors whose coefficients are estimated as insignificant.

First, we find that innovation novelty has a significant positive impact on both legal and nonlegal protection. Their similar estimated coefficients suggest that a firm with novel product innovation is no more likely to use legal protection relative to nonlegal protection. Hence,

Table 5 Estimation Results of Equations (3) and (4)

	Dependent variable:	Probit model			
		Legal protection		Nonlegal protection	
		(3-a)	(3-b)	(4-a)	(4-b)
Innovation novelty		2.10***	2.07***	2.11***	2.09***
	(s.e.)	(0.07)	(0.07)	(0.09)	(0.08)
Market expansion		0.00	-0.03	0.00	0.01
	(s.e.)	(0.10)	(0.10)	(0.10)	(0.11)
Technology acquisition	Buyout	-0.29		-0.30*	-0.20
	(s.e.)	(0.17)		(0.18)	(0.20)
	R&D outsourcing	-0.09		-0.09	
	(s.e.)	(0.11)		(0.11)	
	Purchase of equipment etc.	0.05		0.08	
	(s.e.)	(0.10)		(0.11)	
	Company split-up	0.28		0.34	
	(s.e.)	(0.23)		(0.24)	
	Licensing contract	-0.11		-0.11	
	(s.e.)	(0.13)		(0.15)	
	Open sourcing	-0.10		-0.08	
	(s.e.)	(0.14)		(0.14)	
	Consortium	-0.18		-0.21	
	(s.e.)	(0.15)		(0.16)	
	Alliance	-0.09		-0.07	

18 Unfortunately, the effectiveness of the instruments is rejected for (3-a), (3-b), and (4-b), which remains a future issue to resolve.

19 We omit these variables in order to avoid a convergence problem.

Table 5 (Continue)

	Dependent variable:	Probit model			
		Legal protection		Nonlegal protection	
		(3-a)	(3-b)	(4-a)	(4-b)
	(s.e.)	(0.14)		(0.20)	
	Accepting researchers etc.	-0.18		-0.22	-0.20
	(s.e.)	(0.14)		(-0.13)	(0.14)
Information	Enterprise group etc.	-0.17		-0.13	
	(s.e.)	(0.12)		(0.14)	
	Suppliers	0. 05		0.04	
	(s.e)	(0.10)		(0.10)	
	Consumers or clients	-0.05		-0.04	
	(s.e.)	(0.11)		(0.14)	
	Competitors	0. 11		0.09	
	(s.e.)	(0.10)		(0.13)	
	Private research institutes etc.	0.09		0.10	
	(s.e.)	(0.11)		(0.12)	
	Universities etc.	-0.20		-0.24	
	(s.e.)	(0.14)		(0.15)	
	Public research institutes	0.26**	0.17	0.31*	0.29*
	(s.e.)	(0.13)	(0.12)	(0.17)	(0.15)
	Academic conference etc.	0.12		0.10	
	(s.e.)	(0.12)		(0.12)	
	Professional publications etc.	0.22*	0.15	0.22*	0.18
	(s.e.)	(0.12)	(0.11)	(0.13)	(0.12)
	Exhibitions etc.	0.01		0.01	
	(s.e.)	(0.11)		(0.11)	
	Patent information	-0.16		-0.22	
	(s.e.)	(0.14)		(0.13)	
	Middle-sized	0.10	0.16	0.01	-0.03
	(s.e.)	(0.20)	(0.19)	(0.18)	(0.18)
Firm size	Large	0.30	0.33*	0.19	0.13
	(s.e.)	(0.20)	(0.20)	(0.16)	(0.17)
		-0.11	-0.11	-0.10	-0.04
Product differentiation	(s.e.)	(0.10)	(0.10)	(0.10)	(0.11)
		0.01	-0.03	0.01	-0.03
Public financial support	(s.e.)	(0.11)	(0.11)	(0.11)	(0.11)
Industry dummies		No	No	No	No
Exogeneity test	(Wald)	8.54***	31.34***	1.58	9.17***

Notes: ***, **, and * indicate that the estimate is significant at 1%, 5%, and 10%, respectively.

we cannot reject Hypothesis 7. For the other variables, some technological factors positively affect both legal and nonlegal protection. Looking at (3-a), technology acquisition through public research institutes and professional publications seems to have a positive impact. However, this is not robust compared with (3-b). These variables are estimated to be positive in (4-a), too, while technology acquisition through public research institutes is also significant in (4-b).

Table 6 reports the estimates for Equation (5). We omit the technological variables because otherwise they would all be estimated as being insignificant.[20] Specifications (5-a) and (5-b) include the logarithms of the acquisition of tangible fixed assets and of the number of workers in R&D with and without the industry dummies, whereas specifications (5-c) and (5-d) do not take the logarithms of these variables.

Looking at (5-a), novel product innovation has a significant positive effect on the sales of a new product, which is consistent with Hypothesis 1 and implies that new-to-market product innovation helps a firm avoid severe competition when dealing with homogeneous products. On the contrary, the coefficient of legal protection is estimated to be negative. Legal means of protecting the innovation benefit are not associated with firm performance in terms of innovation-related sales here. The other estimates show that a firm with many employees, R&D workers, and tangible fixed assets tends to achieve above-median product innovation.

(5-b) is similar to (5-a) except that the coefficient of public financial support is estimated as significantly negative. However, it is likely that this estimate captures the differences in the market environment because (5-b) omits the industry dummies. Finally, (5-c) and (5-d) are similar to (5-a) but the results of the Sargan tests do not support them.

20 Hence, Hypothesis 5 would not be supported here, in that we find little evidence that technology acquired through consortia directly affects the sales of a new product.

Table 6 Estimation Results of Equation (5)

		Linear model Dependent variable: Sales of a new product (logarithm)			
		(5-a)	(5-b)	(5-c)	(5-d)
Innovation novelty		1.26*	1.26	0.95	0.94
	(s.e.)	(0.73)	(0.78)	(0.72)	(0.77)
Legal protection		-2.13***	-2.19***	-0.28	-0.28
	(s.e.)	(0.82)	(0.83)	(0.74)	(0.73)
Nonlegal protection		1.10	1.47	1.49	1.78*
	(s.e.)	(0.95)	(1.01)	(0.92)	(0.98)
Market expansion		0.21	0.21	0.53***	0.54***
	(s.e.)	(0.19)	(0.19)	(0.18)	(0.18)
Firm size	Middle-sized	1.20***	1.13***	1.73***	1.71***
	(s.e.)	(0.38)	(0.38)	(0.37)	(0.38)
	Large	2.04***	2.00***	3.47***	3.45***
	(s.e.)	(0.42)	(0.41)	(0.40)	(0.40)
Product differentiation		0.04	0.06	-0.08	-0.09
	(s.e.)	(0.19)	(0.19)	(0.18)	(0.19)
Public financial support		-0.22	-0.34*	-0.33*	-0.44**
	(s.e.)	(0.20)	(0.20)	(0.20)	(0.20)
Acquisition of tangible	[logarithm]	0.28***	0.31***		
fixed assets	(s.e.)	(0.06)	(0.06)		
	[nonlogarithm]			1.07E-05***	1.09E-05***
	(s.e.)			(2.78E-06)	(2.83E-06)
No. of workers in R&D	[logarithm]	0.58***	0.55***		
	(s.e.)	(0.09)	(0.09)		
	[nonlogarithm]			1.14E-04*	1.12E-04
	(s.e.)			(6.75E-05)	(6.95E-05)
Industry dummies		Yes	No	Yes	No
Exogeneity test	(Sargan)	26.04	24.32	35.80**	32.16**

Notes: ***, **, and * indicate that the estimate is significant at 1%, 5%, and 10%, respectively.

Table 7 shows the estimates for Equation (6). Specifications (6-a) and (6-b) adopt the specification in Section 4.1.2 with and without the industry dummies, while specifications (6-c) and (6-d) take the logarithm

Table 7 Estimation Results of Equation (6)

		Linear model Dependent variable: Sales of a new product (logarithm)			
		(6-a)	(6-b)	(6-c)	(6-d)
Innovation novelty		-0.03	-0.05	-0.09	-0.11
	(s.e.)	(0.09)	(0.09)	(0.35)	(0.36)
Sales of a new product		-1.12E-05**	-1.21E-05**		
	(s.e.)	(5.55E-06)	(5.72E-06)		
	[logarithm]			-0.07	-0.08
	(s.e.)			(0.05)	(0.05)
Innovation novelty*	sales of a new product	1.14E-05**	1.23E-05**		
	(s.e.)	(5.74E-06)	(5.94E-06)		
	[logarithm]			0.02	0.02
Total sales	(s.e.)			(0.06)	(0.06)
	[logarithm]	0.99***	1.00***	1.02***	1.03***
	(s.e.)	(0.02)	(0.02)	(0.03)	(0.03)
Firm size	Middle-sized	0.04	0.03	0.07	0.07
	(s.e.)	(0.06)	(0.06)	(0.06)	(0.06)
	Large	0.03	0.02	0.10	0.09
	(s.e.)	(0.08)	(0.08)	(0.07)	(0.07)
Industry dummies		Yes	No	Yes	No
Exogeneity test	(Sargan)	24.38	22.17	29.51	27.09

Notes: *** and ** indicate that the estimate is significant at 1% and 5%, respectively.

of the sales of a new product.

For (6-a), the sales of a new product have a significant negative effect on those of existing products. This result is consistent with the view that a new product *cannibalizes* a part of the firm's existing products, which is consistent with Hypothesis 2. By contrast, the coefficient of the cross term of innovation novelty with the sales of a new product is significantly positive and nearly cancels out the cannibalization term. Hence, we can interpret this finding as saying that the cannibalization effect is reversed with innovation novelty, which is consistent with Hypothesis 3.

(6-b) is similar to (6-a). For (6-c) and (6-d), the coefficients of the sales of a new product and the cross term are estimated as insignificant, although their signs are the same as those of (6-a).

Finally, Table 8 shows the estimates for Equation (7). We omit the technological variables because otherwise they would all be estimated as being insignificant as before. Specifications (7-a) and (7-b) include the logarithms of the acquisition of tangible fixed assets and of the number of workers in R&D with and without the industry dummies, while specifications (7-c) and (7-d) do not take the logarithms of these.

Table 8 Estimation Results of Equation (7)

		Linear model Dependent variable: Technology provision through open sourcing or consortia			
		(7-a)	(7-b)	(7-c)	(7-d)
Innovation novelty		2.29 **	2.09 **	2.52 **	2.25 **
	(s.e.)	(0.93)	(0.82)	(1.23)	(1.04)
Legal protection		-1.11	-1.01	-1.17	-1.05
	(s.e.)	(1.06)	(0.97)	(1.12)	(1.00)
Nonlegal protection		0.28	0.58	0.28	0.63
	(s.e.)	(0.98)	(0.98)	(1.07)	(1.08)
Market expansion		-0.04	-0.03	-0.03	-0.02
	(s.e.)	(0.12)	(0.11)	(0.13)	(0.12)
Firm size	Middle-sized	0.19	0.13	0.20	0.16
	(s.e.)	(0.30)	(0.27)	(0.35)	(0.33)
	Large	0.53	0.41	0.60	0.48
	(s.e.)	(0.38)	(0.31)	(0.50)	(0.43)
Product differentiation		-0.10	-0.09	-0.12	-0.10
	(s.e.)	(0.12)	(0.12)	(0.14)	(0.13)
Public financial support		0.17	0.08	0.19	0.09
	(s.e.)	(0.15)	(0.12)	(0.17)	(0.14)
Acquisition of tangible	[logarithm]	-0.02	0.00		
fixed assets	(s.e.)	(0.04)	(0.04)		
	[nonlogarithm]			-8.14E-07	-2.77E-07
	(s.e.)			(2.16E-06)	(1.91E-06)
No. of workers in R&D	[logarithm]	0.05	0.02		
	(s.e.)	(0.08)	(0.08)		
	[nonlogarithm]			6.20E-06	2.59E-07
	(s.e.)			(4.74E-05)	(4.40E-05)
Industry dummies		Yes	No	Yes	No
Exogeneity test	(Sargan)	7.65	9.20	6.30	8.06

Notes: ** indicates that the estimate is significant at 5%.

For all specifications, the coefficient of innovation novelty is estimated as significantly positive. This estimation implies that a firm with new-to-market product innovation is more likely to provide its technology through open sourcing or consortia, which is consistent with Hypothesis 4. Hence, we can say novel product innovation is associated with technological spillovers through channels that are less likely to accompany monetary compensation.

5. Conclusion

The empirical analyses presented in this study focused on the degree to which new-to-market product innovation influences firm performance (i.e., sales of new and existing products), technological spillovers, and other characteristics of new-to-market product innovation. We proposed eight hypotheses and tested them by use of JNIS for the study period of April 2006 to March 2009.

Overall, our results are consistent with the hypotheses presented. We found that innovators tend to achieve higher sales from new-to-market product innovation, and are less likely to suffer from the cannibalization effect. New-to-market product innovation tends to spill its knowledge over to other firms' innovations through channels that are less likely to be accompanied by monetary compensation.

Considering that new-to-market product innovation significantly improves firm performance and is associated with technological spillovers, policy intervention promoting such innovation may be beneficial. Our empirical results show that firms with new-to-market product innovation are more likely to use information from universities, and less likely to rely on legal protection. In addition, public financial support does not necessarily stimulate new-to-market product innovation, especially for small firms. Rather, for creating such innovation, nonfinancial policies such as using personnel to improve the interrela-tions between firms and universities are rather important.

References

Aghion, P. and P. Howitt (1992), "A Model of Growth through Creative Destruction," Econometrica, 60: 323-351.

Aghion, P., N. Bloom, R. Blundell, R. Griffith and P. Howitt (2005), "Competition and Innovation: An Inverted-U Relationship," *Quarterly Journal of Economics*, 120: 701-728.

Almus, M. and D. Czarnitzki (2003), "The Effects of Public R&D Subsidies on Firms' Innovation Activities: The Case of Eastern Germany," *Journal of Business & Economic Statistics*, 21: 226-36.

Arrow, K. (1962), "Economic Welfare and the Allocation of Resources for Invention," *in The Rate and Direction of Inventive Activity*: *Economic and Social Factors*, Universities-National Bureau, (eds.). UMI, Princeton.

Arundel, A., C. Bordoy and M. Kanerva (2008), "Neglected Innovators: How Do Innovative Firms that Do Not Perform R&D Innovate?" INNO-Metrics Thematic Paper.

Arundel, A. and H. Hollanders (2005), "EXIS: An Exploratory Approach to Innovation Scoreboards," Brussels, European Commission, DG Enterprise.

Barlet, C., E. Duguet, D. Encaoua and J. Pradel (1998), "The Commercial Success of Innovation: an Econometric Analysis at the Firm Level in French Manufacturing," *Annales d'Economie et de Statistique*, 49-50: 457-78.

Belderbos, R., M. Carree and B. Lokshin (2004), "Cooperative R&D and Firm Performance," *Research Policy*, 33: 1477-1492.

Bloom, N., M. Schankerman and J. V. Reenen (2013), "Identifying Technology Spillovers and Product Market Rivalry," Econometrica 81(4): 1347-93.

Branstetter, L. G. and M. Sakakibara (2002), "When Do Research Consortia Work Well and Why? Evidence from Japanese Panel Data," *American Economic Review*, 92, pp. 143-159.

Chudnovsky, D., A. Lopez and G. Pupato (2006), "Innovation and Productivity in Developing Countries: A Study of Argentine Manufacturing Firms' Behavior (1992-2001)," *Research Policy*, 35, pp. 266-288.

Cohen, W. M. and S. Klepper (1996), "Firm Size and the Nature of Innovation within Industries: The Case of Process and Product R&D," *Review of Economics and Statistics*, 78: 232-43.

Crépon, B., E. Duguet and J. Mairesse (1998), "Research, Innovation and Productivity: An Econometric Analysis at the Firm Level," *Economics of Innovation and New Technology*, 7: 115-58.

De Loecker, J. (2011), "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity," *Econometrica*, 79: 1407-51.

Duguet, E. (2006), "Innovation Height, Spillovers and TFP Growth at the Firm Level: Evidence from French Manufacturing," *Economics of Innovation and New Technology*, 15: 415-42.

Duguet, E. and C. Lelarge (2006), "Does Patenting Increase the Private Incentives to Innovate? A Microeconometric Analysis," Working Papers 2006-09, Centre de Recherche en Economie et Statistique.

Finger, S. R. (2008), "An Empirical Analysis of R&D Competition in the Chemicals Industry," University of South Carolina.

Gilbert, R. and D. Newbery (1982), "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, 72: 514-26.

González, X., J. Jaumandreu and C. Pazó (2005), "Barriers to Innovation and Subsidy Effectiveness," *Rand Journal of Economics*, 36: 930-950.

Griffith, R., E. Huergo, J. Mairesse and B. Peters (2006), "Innovation and Productivity across Four European Countries," *Oxford Review of Economic Policy*, 22: 483-98 .

Griliches, Z. (1992), "The Search for R&D Spillovers," *Scandinavian Journal of Economics*, 94: 29-47.

Grossman, G. M. and E. Helpman (1991), "Quality Ladders in the Theory of Growth," *Review of Economic Studies*, 58: 43-61.

Jefferson, G. H., B. Huamao, G. Xiaojing and Y. Xiaoyun (2006), "R&D Performance in Chinese Industry," *Economics of Innovation and New Technology*, 15: 345-66.

Kaiser, U. (2002), "An Empirical Test of Models Explaining Research Expenditures and Research Cooperation: Evidence for the German Service Sector," *International Journal of Industrial Organization*, 20: 747-74.

Katz, M. L. (1986), "An Analysis of Cooperative Research and Development," *RAND Journal of Economics*, 17: 527-43.

Klepper, S. (1996), "Entry, Exit, Growth, and Innovation over the Product Life Cycle," *American Economic Review*, 86: 562-83.

Klette, T. J. and S. Kortum (2004), "Innovating Firms and Aggregate Innovation,"

Journal of Political Economy, 112: 986-1018.

Levin, R. C., A. K. Klevorick, R. R. Nelson and S. G. Winter (1987), "Appropriating the Returns from Industrial Research and Development," *Brookings Papers on Economic Activity*, 18: 783-832.

Levin, R and P. C. Reiss (1984), "Tests of a Schumpeterian Model of R&D and Market Structure," in *R & D, Patents, and Productivity*, NBER Chapters, National Bureau of Economic Research, Inc.

Lööf, H. and A. Heshmati (2006), "On the Relationship between Innovation and Performance: A Sensitivity Analysis," *Economics of Innovation and New Technology*, 15: 317-44.

Mairesse, J. and P. Mohnen (2010), "Using innovations surveys for econometric analysis," UNUMERIT Working Paper 2010-023.

Mohnen, P. and C. Hoareau (2003), "What Type of Enterprise Forges Close Links with Universities and Government Labs? Evidence from CIS 2," *Managerial and Decision Economics*, 24: 133-45.

Monjon, S. and P. Waelbroeck (2003), "Assessing Spillovers from Universities to Firms: Evidence from French Firm-Level Data," *International Journal of Industrial Organization*, 21: 1255-70.

Nishikawa, K., D. Isogawa and H. Ohashi (2010), "*Wagakuni niokeru Product Innovation no Genjou: Dai 2-kai Innovation Tyousa wo Motiita Bunseki*," NISTEP DISCUSSION PAPER 70.

NISTEP (2010), "*Report on Japanese National Innovation Survey 2009*," NISTEP REPORT 144.

OECD (1992, 1996, 2005), "Oslo Manual," Paris, 1st, 2nd, 3rd editions.

OECD (2009), "Innovation in Firms," OECD Publishing.

Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54: 755-84.

Petrin, A. (2002), "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110: 705-29.

Reinganum, J. (1983), "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, 73: 61-6.

Rosenberg, N. (1974), "Science, Invention and Economic Growth," *Economic Journal*, 84: 90-108.

Schankerman, M. (1998), "How Valuable is Patent Protection? Estimates by Technology Field," *RAND Journal of Economics*, 29: 77-107.

Schumpeter, J. (1943), "Capitalism, Socialism and Democracy," Allen Unwin, London.

Spence, M. (1984), “Cost Reduction, Competition, and Industry Performance,” *Econometrica*, 52: 101-22.

Trajtenberg, M. (1989), “The Welfare Analysis of Product Innovation, with an Application to Computed Tomography Scanners,” *Journal of Political Economy*, 97: 444-79.

Van Leeuwen, G. and L. Klomp (2006), “On the Contribution of Innovation to Multi-Factor Productivity Growth,” *Economics of Innovation and New Technology*, 15: 367-90.

Xu, Y. (2006), “Structural Empirical Model of R&D, Firm Heterogeneity, and Industry Evolution,” Pennsylvania State University.

CHAPTER 2

The Contribution of Research and Innovation to Productivity and Economic Growth*

by
Amani Elnasri
(University of New South Wales)
*Kevin J. Fox***
(University of New South Wales)

Abstract

This paper examines the impact of investment in research and innovation on Australian market sector productivity. While previous studies have largely focused on a narrow class of private sector intangible assets as a source of productivity gains, this paper shows that there is a broad range of other business sector intangible assets that can significantly affect productivity. Moreover, the paper pays special attention to the role played by public support for research and innovation in the economy. The empirical results suggest that there are

* We thank the Productivity Commission and Melbourne Institute for providing us with their data on intangible investment. Financial support from the Australian Research Council (LP0884095) is gratefully acknowledged, as are helpful comments from Paula Barnes, Erwin Diewert, Dean Parham, Joonghae Suh and participants at the 2014 KDI Journal of Economic Policy Conference. The views expressed in this paper are those of the authors. Any errors are our responsibility.

** Corresponding Author: Kevin J. Fox, School of Economics & CAER, University of New South Wales, Sydney 2052, Australia. E-mail: K.Fox@unsw.edu.au, Tel: +61-2-9385-3320. This paper is a contribution to a series of projects undertaken by the Australian Council of the Learned Academies to examine 'The Role of Science, Research and Technology in Lifting Australia's Productivity'.

significant spillovers to productivity from public sector R&D spending on research agencies and higher education. No evidence is found for productivity spillovers from indirect public support for the business enterprise sector, civil sector or defence R&D. These findings could have implications for government innovation policy as they provide insights into possible productivity gains from government funding reallocations.

1. Introduction

Research and innovation are widely agreed to be major driving forces behind long-term productivity and economic growth. It is now well recognised that the productivity benefits from research and successful innovations are not fully absorbed by the innovating entities but, rather, they diffuse through the rest of the economy leading to positive externalities in growth and the productivity performance of the other using entities.

This paper attempts to have a closer look into some aspects of the Australian innovation system and its impact on the economy. Specifically, the objectives of the paper are three-fold. First, to extend a staff working paper of the Productivity Commission conducted by Barnes and McClure (2009), henceforth referred to as the 'the PC report', on the spending on a broad range of intangible assets for the Australian market sector. These intangibles are incorporated into the Australia National Accounts to provide estimates for recent years.[1] By recognising the additional investment in the economy when 'new' intangible expenditure is treated as investment, the paper adjusts the measures of the market sector gross value added (GVA), capital stock and factor income shares.[2] Using the growth accounting framework,

1 As far as can be ascertained, the PC report and an extension by de Rassenfosse (2012) are the only two previous attempts made to apply the Corrado, Hulten and Sichel (2005, 2006) methodology to measure a set of intangible assets beyond those currently capitalised in the Australian National Accounts.

2 'New' intangibles refer to those intangibles which are currently not included in the National Accounts.

these 'new' measures are employed to construct adjusted estimates of the market sector multifactor productivity (MFP) growth. Furthermore, in line with the method of the PC report, the paper develops two additional sets of the growth accounting components to assess the impact on these com-ponents when either a 'sub-group' of or 'all' intangibles are capitalised.[3] To construct the first set, only those intangibles which are currently capitalised by the ABS (computer software, artistic originals, and mineral exploration and scientific R&D) are included while in the other set the growth accounting components are estimated under the assumption that all intangibles are treated as intermediate inputs.

Second, the paper examines whether there are any productivity spillovers/excess returns from the investment in intangibles or if the returns for these intangibles are restricted to those firms producing or consuming them. Although there are a number of studies that have examined the impact of R&D on Australia's productivity, these studies did not examine the impact of other intangible assets nor did they adjust MFP growth for the capitalisation of knowledge and other intangibles. As in Haskel and Wallis (2010, 2013), henceforth collectively referred to as HW, outline, adjusting MFP growth to include intangibles is helpful in isolating private from social returns.[4] If MFP growth used in a regression model is not adjusted, then the ensuing estimates of the returns on the knowledge assets will suffer from measurement errors.

The third, and most important objective of the paper, is to examine the impact of public support for research and innovation on market sector productivity. Building on HW, the paper aims to investigate spillovers to productivity from various sources of public funding. More Specifically, it is to answer the question of whether or not public support for research and innovation should focus on direct spending on public research institutions (such as Commonwealth Scientific and Industrial Research Organisation, CSIRO, and the Defence Science and

3 A 'sub-group' of intangibles refers to those assets which are currently capitalised in the National Accounts while 'all' intangibles refers to a combination of National Accounts and 'new' intangibles.

4 Haskel and Wallis (2013) is an updated and condensed version of the more comprehensive discussion paper, Haskel and Wallis (2010).

the Technology Organisation, DSTO); funding of higher education (e.g., Australia Research Council, ARC); or provide indirect support to the business sector (for example, through tax incentives such as expanding the R&D Tax Concession to a broad range of intangibles).[5] Answering this question is crucial to informing and designing effective policy. Because governments are constrained by tight fiscal budgets, efficient innovation policies should focus on areas with higher expected social returns in order to maximize the benefits from public spending. For example, for the U.K. HW found strong evidence of spillovers from public R&D expenditure on research councils as opposed to other areas. Accordingly, their findings suggest that for maximum productivity impact in the U.K., government innovation policy should support direct spending on research councils rather than tax breaks, such as the R&D tax credit, to firms.

The paper proceeds as follows: Section 2 briefly provides the theoretical background on how investment in knowledge capital is linked to productivity and economic growth. Section 3 provides estimates of the Australian market sector intangible investment, intangible capital stock, and discusses their trends over the period 1974-75 to 2012-13. Section 4 presents the impact of capitalising intangibles in the growth accounting components by discussing three different definitions of capital (when all intangibles are capitalised, when only National Accounts intangibles are capitalised, and when all intangibles are treated as intermediate goods). Section 5 presents definitions and trends of Australian government spending on research and innovation. A simple analysis of the relationship between public support for R&D and market sector MFP is presented in Section 6. A more comprehensive analysis using econometric techniques is presented in Section 7. Section 8 concludes.

---

5 Public support for the business sector is delivered through a range of programs: The R&D Tax Concession (which accounts for about 50 % of total business support); Rural Research and Development Corporations; grant funding under the Commercial Ready Program; and the Automotive Competitiveness and Investment Scheme.

2. Investment in Knowledge Capital and Economic & Productivity Growth

The New Growth Theory literature (e.g., Arrow 1962 and Romer 1990) has emphasised two points. First, the accumulation of knowledge, innovation or human capital by economic agents is the principal source of technological change (a key source of productivity growth) and hence economic growth. Second, the positive externalities and spillover effects of a knowledge-based economy can reduce the diminishing returns to capital accumulation and hence lead to economic development.

In the context of this literature, the existence of knowledge spillovers are explained by the distinctive characteristics of knowledge: non-rivalry and non-excludability. Knowledge is considered to be a good which is non-rival in nature because it can be made available to a number of users simultaneously without extra costs to the supplier. Unlike a tangible asset, knowledge is not 'consumed' by those who use it. It can be used at multiple times and by multiple users. On the other hand, the non-excludability means that if the knowledge is provided at all, it is available to everyone and its users cannot be denied access to it. The non-rivalry and non-excludability properties of knowledge are the attributes that drive economic growth. Accumulation of more ideas will enable the economy to develop further. Ideas are not subject to diminishing returns; rather, the increasing returns to knowledge boost economic growth.

In general, economic growth can be decomposed into two components: growth of factor inputs (such as capital, labour and land) and growth of productivity. Productivity is a measure of how efficiently an economy utilises finite resources to produce goods and services. Thus, it is a ratio of output to input. Total output can be increased by either increasing the utilisation of resources or by improving the efficiency with which resources are employed. In the long term, contributions through increased utilisation of resources will be limited by the finite endowment of resources. Thus, sustained economic growth will have to come mainly from productivity increases. There are several ways to improve productivity but knowledge capital (through new technology, skills, R&D and efficient services and production processes) is the most significant factor. Due to new technology,

the same level of output can be produced with fewer inputs. Also, technology diffiusion reduces inefficiencies because it enables firms to reach, or come closer, to the production frontier.

The effect of knowledge capital on productivity may work through various channels depending of the source of the knowledge. For example, R&D, a major component of knowledge capital, can be performed either by the business sector, public sector or beyond the borders of a country. Each of these types of R&D performers can be a source of significant domestic technological change. R&D performed by the business sector results in new goods and services, higher quality of output, and new production processes. These are sources of productivity growth at the firm and national levels. Many empirical studies confirm the positive impact of business R&D on productivity; see e.g. Griliches (1998) and Nadiri (1993). Business-performed R&D may be funded by business itself or by the government. Accordingly, business R&D may have a different effect on productivity, depending on its source of funding (which affects the research agenda and the incentive structure). For example, Lichtenberg (1993) tests whether government-funded R&D performed by firms had a different impact than business-funded R&D. The author's evidence suggests that while privately-funded R&D investment has a significant positive effect on productivity, government support for business R&D has a negative impact.

Besides their support for business R&D, governments are major R&D performers through government research agencies or through funding higher education R&D. Research agencies and university R&D are seen to have a strong effect on scientific, basic knowledge and on public missions. Basic research performed by universities enhances the stock of knowledge available for the society. It may open new opportunities for business research, which in turn might improve productivity. Nevertheless, there have been few attempts to measure the impact of public R&D on productivity. In a group of studies only some components of public research have been used in empirical frameworks. For example, Adams (1990) examines the contribution of fundamental stocks of knowledge, proxied by accumulated academic scientific papers and finds significant contributions to productivity growth in the U.S. manufacturing industries. Another example is Poole and Bernard (1992) who

examine military innovations and find a negative impact on Canadian total factor productivity. Among the small number of studies that examined a broader definition of public sector R&D are Park (1995) and HW; Park (1995) conducts a panel data analysis of 10 OECD countries and finds that the public R&D effect on productivity growth becomes insignificant when business R&D is incorporated as an additional regressor.

The knowledge originating from abroad is a third source of new technology for any national economy. Evidence demonstrates many avenues through which knowledge can cross the boarders of a given country and, depending on the absorptive capacity, it may improve other countries' productivity (Mohnen 2001).

The Australian literature has a limited number of studies that have quantitatively examined Australia's innovation system and its impact. Most of these studies have focused on the link between productivity and R&D, ignoring the effect of the other types of intangible capital. The R&D measures employed by these studies largely relates to business R&D (e.g., Shanks and Zheng 2006 and Louca 2003). Moreover, the empirical evidence obtained by these studies was mixed or generally not supportive of the productive role of business R&D. For example, Shanks and Zheng (2006) outline that despite the advances in data collection and methods applied in the study, the research was unable to find a consistently robust measure of the impact of R&D on productivity and the estimated effect of R&D was implausibly large.[6]

There are a small number of cases in which the role of higher education R&D is assessed. One example is a study by Burgio-Ficca (2004) who finds evidence of a positive relationship between higher education R&D and gross state product. With the exception of PC (2007) there is no study which has explicitly scrutinised the effects of publicly funded R&D.[7] Although the findings of PC (2007) suggest significant aggregate economic, social and environmental benefits from publicly supported science and innovation, the quantitative estimates are found to

6 For a concise summary and discussion of this and related work, see Parham (2006).

7 There are a small number of studies which might have partially addressed this question by employing data on the gross expenditure on R&D (GERD, an aggregate measure of business, government and higher education R&D). However, using GERD as a measure will not isolate the effects of government or higher education R&D.

be unreliable.

As mentioned above, the existing body of the Australian literature does not extend to a search of the contribution of the other types of knowledge assets beyond R&D. Despite its importance, R&D is not the only source of new technology. Innovation can result from the contribution made by other types of intangible capital, and extends beyond physical capital accumulation. Some recent studies suggest new methods for defining and measuring intangible capital by measuring investment in innovation-related assets such as skills development, non-scientific R&D, design, organisational improvements and so forth. More discussion of these intangibles is provided in the next section.

3. Intangible Investment

Despite the increase in their prominence, research and innovation, among a large set of 'intangible' assets, are largely ignored in National Accounts and corporate financial reports because they are hard to understand and measure. Two recent studies by Corrado, Hulten and Sichel (2005, 2006), henceforth collectively referred to as CHS, have drawn the attention of researchers to the importance of measuring and capitalising these intangibles. Using U.S. data, CHS have developed a methodology to capitalise a broad range of intangibles and, by applying a growth accounting framework, demonstrated how the conventional growth rates of inputs, output and productivity measures changed as a consequence. Following CHS, researchers in a number of other advanced countries (e.g., United Kingdom, Japan, Netherlands, Canada and Australia) have conducted similar studies, and found results similar to those of CHS.

Following the recommendations of the System of National Accounts (SNA) 1993, some statistical agencies have begun to change the treatment of intangible assets in their National Accounts.[8] Australia was

8 Until recently, expenditures on intangible assets were not recorded as final expenditures in the calculation of gross domestic product. Rather, they were classified as intermediate inputs. The new treatment recognises expenditure on intangible assets as fixed investment and the depreciation of these assets in the consumption of fixed capital.

one of the first countries to capitalise computer software, artistic originals and mineral exploration in 1993. In addition, as part of the revisions to implement the recommendations contained in SNA 2008, Australia started to capitalise scientific R&D in 2009. Nevertheless, intangible assets are not restricted to these four elements. Firms also invest in other types of intangible assets which may represent a source of economic growth; however, these investments are treated in the National Accounts as current expenses. Excluding investment in intangibles underestimates total investment, which in turn may misrepresent the measures of output, capital services, factors income shares and consequently productivity.

CHS classify intangibles into three categories: computerised information, innovative property and economic competencies. Each of these categories is composed of several specific intangibles which are reported in Table 1.

CHS construct measures of these intangibles for the U.S. and use them to examine their contribution to labour productivity growth. They find that the U.S. invests substantially in intangible assets (12.1% of GDP in intangible assets in 2003, CHS 2005). In addition, they find that capitalising intangibles has considerably increased labour productivity growth. In particular, they find it increased by 0.8 % from 1995 to 2003. The work of CHS has motivated studies in other advanced countries where authors find that these countries have significantly invested in intangibles. For example, Marrano and Haskel (2006) find that the private sector in the U.K. invested 10.1% of GDP on intangibles in 2004. In Finland, the private sector invested 9.1% of GDP in intangible assets (Jalava et al. 2007). The Netherlands invested 8.4% of GDP between 2001 and 2004 (van Rooijen-Horsten et al. 2008). Fukao et al. 2009 find that Japan invested 7.5% of GDP from 1995 to 2002 while Baldwin et al. (2012) find that the Canadian business sector invested 13.2% of GDP in intangible assets in 2008. Hao et al. (2009) conducted an international comparison between France, Germany and Italy and found that the shares of intangible investment in GDP in these three countries are 8.3%, 7.1% and 5.2% respectively in 2004. Finally, the PC report suggests that Australia has invested 5.9% of GDP in 2005-06.

Table 1 Definitions of Intangibles, CHS

1. Computerised information
Computer software
Computer databases
2. Innovative property
Scientific R&D; Social sciences R&D (Business R&D)
Mineral exploration
Copyright and licence costs (Artistic originals)
Other product development, design and research
New product development in financial industry
New architectural and engineering designs
3. Economic competencies
Brand equity
Advertising
Market research
Firm-specific human capital
Organisational capital
Purchased
Own account

3.1 Measuring Intangibles

For Australia, the PC report and de Rassenfosse (2012) are the only existing studies that have applied the methodology of CHS to measure and classify a range of 'new' intangibles.[9] The PC report provides estimates over the period 1974-75 to 2005-06 and de Rassenfosse (2012) extends these estimates to 2010-11. However, due to measurement challenges and difficulties in obtaining adequate information, the authors of these studies were required to make a number of assumptions to enable them to obtain measures over time: 'Given the experimental nature of the methodology, the assumptions required, measurement challenges and data limitations, the estimates should be interpreted as only indicative' (Barnes and McClure 2009, p. XIII).

This paper depends on different sources to collect data on investment in intangibles. For those assets which are already capitalised in the

9 A third relevant study is Barnes (2010) in which the author extends the estimates of the PC report to a sectoral level.

National Accounts, the data is sourced from the ABS website. For investment in 'new' intangibles, the estimates of both the PC report and de Rassenfosse (2012) are reconciled to form a series over the period 1974-75 to 2012-13.[10] While a more detailed discussion of the definitions and sources of all data used in the paper is available in the Appendix, a brief description of the investment data in intangibles is provided below.

3.1.1. Computerised Information

Computer software is already treated as investment in the Australian National Accounts. A time series for gross fixed capital formation and capital stock is available for the full period 1974-75 to 2012-13. The ABS computer software implicit price deflator (IPD) is used to obtain the real investment series.

3.1.2. Innovative Property

CHS define four types of innovative property which are:

(i) Business expenditure on R&D

Australian business expenditure on R&D (BERD) is available from the ABS Research and Experimental Development, Businesses (Cat. no. 8104.0). A consistent series for the market sector (excluding Agriculture, forestry & fishing) was compiled for 1974-75 to 2005-06 by Shanks and Zheng (2006) and the PC report.[11] For this paper, it is updated and extended to 2012-13 using revised and updated data from the ABS Cat. no. 8104.0. The ABS's IPD for R&D is used to obtain the real investment series.

---

10 The caveat expressed by the authors of the earlier studies about the experimental nature of the estimates is also applied in this paper. One of the original objectives of this paper was to contribute to the Australian literature by providing improvements and refinements to the existing estimates of intangible assets. Unfortunately, due to severe data limitations and measurement challenges, the endeavours made to improve measurement of new intangibles were of little avail.

11 The ABS did not directly survey farms and other businesses in this industry until 2005-06. Agriculture has been excluded from the 2005-06 to 2012-13 data to maintain comparability over time.

(ii) Mineral exploration and (iii) Artistic originals

Mineral exploration and Artistic originals are already treated as investment in the Australian National Accounts. Time series for gross fixed capital formation and capital stock are available for the full period 1974-75 to 2012-13. The ABS's IPDs for Mineral exploration and Artistic originals are used to obtain the respective real investment series.

(iv) Other product development, design and research

This type of 'non-scientific' R&D is currently treated as intermediate expenditure in the National Accounts. According to CHS, it consists of:

New product development in the financial industry

The PC report has constructed a series for 20% of total intermediate purchases by the financial services industries to cover the period 1974-75 to 2005-06. de Rassenfosse (2012) has extended the series to 2010-11 by applying a relevant growth rate to the year 2004-05 data point from the PC report.[12] Assuming linear growth in recent years, this paper extends the series to 2012-13. The ABS's IPD for the Finance & Insurance industry is used to obtain the real investment series.

New architectural and engineering designs

The PC report has constructed a series for 50% of the revenue of architectural and engineering industries to cover the period 1974-75 to 2005-06. de Rassenfosse (2012) has extended the series to 2010-11 by using turnover data from the ABS Counts of Australian Businesses, including Entries and Exits for classes of Architectural Services and Engineering Design and Engineering Consulting Services. Assuming linear growth in the recent years, this paper extends the series to 2012-13. The ABS's IPD for the market sector GVA is used to obtain the real investment series.

3.1.3. Economic Competencies

The components of economic competencies defined by CHS are

12 See the Appendix for details.

currently treated as intermediate expenditure in the National Accounts. They fall into three categories:

(i) Brand equity

Spending on brand development is measured by the spending on advertising and market research:

Advertising

This type of expenditure is available from an annual survey of the industry conducted by the Commercial Economic Advisory Service of Australia (CEASA). The ABS's IPD for the market sector GVA is used to obtain the real investment series.

Market research

The PC report has constructed a series as the double of the revenue of the market research industry. Interpolation and backdating were performed to construct a series to cover the period 1974-75 to 2005-06. de Rassenfosse (2012) has extended the series to 2010-11 by using turnover data from the ABS Counts of Australian Businesses, including Entries and Exits for the class Market Research and Statistical Services.[13] Assuming linear growth for recent years, the paper extends the series to 2012-13. The ABS's IPD for the market sector GVA is used to obtain the real investment series.

(ii) Firm-specific human capital

No single data source provides a time series of Australian employer-provided training expenditure. The PC report constructed a series to cover the period 1974-75 to 2005-06 using different data sources with a number of assumptions. The main source was the direct costs and wage costs of employee time in training for market sector industries (excluding agriculture) from the ABS Training surveys. de Rassenfosse (2012) has extended the series to 2010-11 by forecasting. Assuming linear growth for recent years, the paper extends the series to 2012-13.

---

13 Due to a change in ANZSIC classification, there is a break in the series in 2006-07 vs. 2007-08. See the Appendix for more details.

The average weekly fulltime ordinary earnings deflator is used to obtain the real investment series.

(iii) Organisational capital

The investment in organisational capital as suggested by CHS is made up of two components, purchased and own account:

Purchased

Using the ABS Industry Survey, the PC report has constructed the series as 77% of sales of all business management services to cover the period 1974-75 to 2005-06. de Rassenfosse (2012) has extended the series to 2010-11 by using turnover data from the ABS Counts of Australian Businesses, including Entries and Exits for the class 6962 Management Advice and Related Consulting Services. Assuming linear growth in the recent years, the paper extends the series to 2012-13. The ABS's IPD for the market sector GVA is used to obtain the real investment series.

Own account

The PC report constructed the series as 20% of salaries of Managers and Administrators (excluding farm managers and IT managers) in the market sector to cover the period 1974-75 to 2005-06. de Rassenfosse (2012) has extended the series to 2010-11 by using the ABS data on employee earnings, benefits and trade union membership. Assuming linear growth in the recent years, the paper extends the series to 2012-13. The ABS's IPD for the market sector GVA is used to obtain the real investment series.

3.2 Trends in Australian Intangibles

Table 2 presents estimates of nominal intangible investment in the market sector for some selected years of the study period: 1974-75, 1984-85, 1994-95, 2004-05 and 2012-13. As seen from the table, investment in intangibles has increased over time and reached about $80 billion in 2012-13, constituting 28% of market sector total investment in that year. With the exception of the last few years, total investment in

Table 2 Estimates of nominal intangible investment in the Australian market sector

Categories	1974-75	1984-85	1994-95	2004-05	2012-13
			Millions of dollars		
Computerised information	26	627	3,512	7,262	9,948
Innovative property	917	3,857	9,342	19,414	38,624
Scientific R&D; Social sciences R&D (Business R&D)	199	614	2,782	7,010	14,483
Mineral exploration	230	1,271	1,567	2,704	7,849
Copyright and licence costs (Artistic originals)	35	172	256	1,045	2,450
Other product development, design and research	480	1,800	4,737	9,286	13,841
New product development in financial industry	342	1,310	3,133	5,311	8.338
New architectural and engineering designs	137	490	1,604	3,975	5,504
Economic competencies	1,259	4,926	11,276	23,374	33,428
Brand equity	653	2,830	4,679	8,365	10,362
Advertising	648	2,774	4,420	7,391	9,463
Market research	5	56	260	974	899
Firm-specific human capital	301	1,024	2,669	3,870	5,791
Organisational capital	306	1,073	3,927	11,138	17,276
Purchased	21	232	1,944	7,058	9,143
Own account	284	840	1,983	4,081	8,133
Total intangibles investment	2,202	9,410	24,130	50,050	82,000
New intangibles	1,739	6,726	16,013	32,659	47,270
National Accounts intangibles	463	2,684	8,118	17,391	34,730
Tangibles	9251	32,333	54,984	106,195	227,751
Total investment	11,453	41,743	79,114	156,245	309,751

Table 2 (Continue)

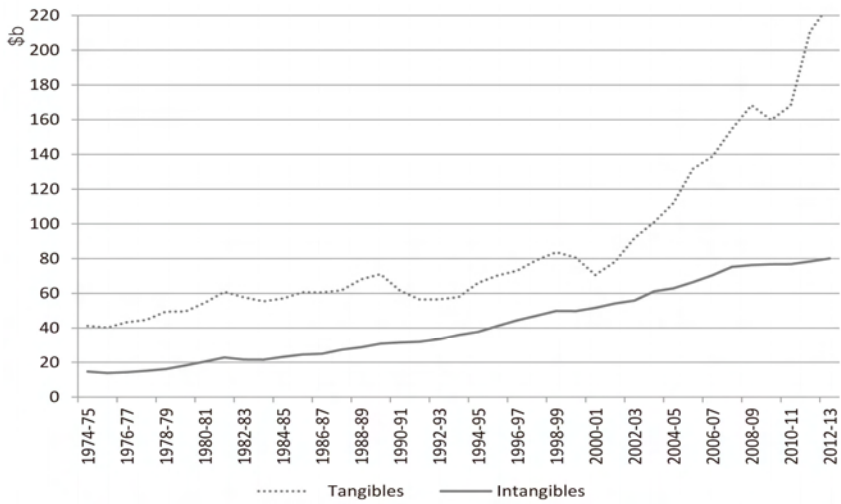
Categories	1974-75	1984-85	1994-95	2004-05	2012-13
Share of computerised information %	1	7	15	15	12
Share of innovative property %	42	41	39	39	47
Share of economic competencies %	57	52	47	47	41
Share of intangible investment%	19	23	31	32	26
Share of tangible investment%	81	77	69	68	74
Ratio intangible to tangible investment	0.24	0.29	0.44	0.47	0.36

The share of tangible (intangible) investment is the ratio of tangibles (intangibles) to total investment.
The shares of computerised information, innovative property, and economic competencies are calculated relative to all intangibles.

intangibles grew more rapidly than investment in tangibles; see Figure 1. The ratio of intangibles to tangibles increased continuously from 0.29 in 1974-75 to 0.53 in 2004-05; however, it decreased to 0.38 by 2012-13. Of all intangibles only computer software, artistic originals, mineral exploration and R&D have been capitalised in the Australian System of National Accounts. As shown in the table the investment in these four intangibles constitutes less than half of total intangible investment. In 2012-13, National Accounts intangibles accounted for 41% of total intangible investment while the new intangibles accounted for 59%.

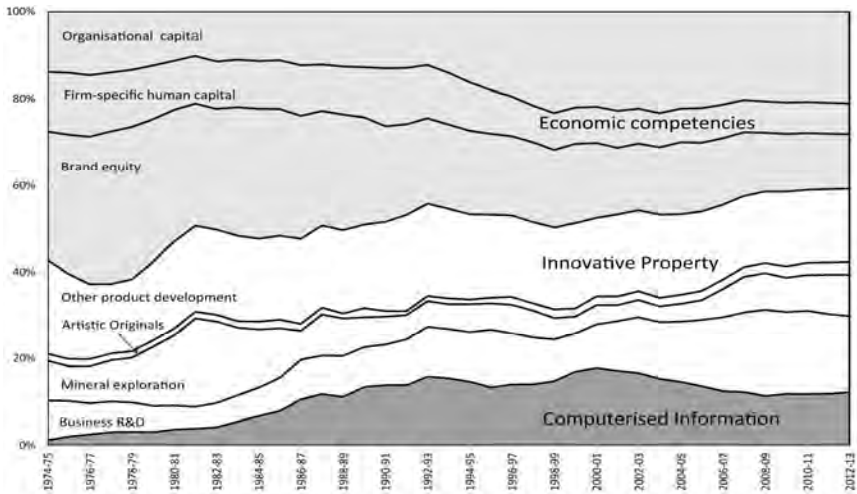
Table 2 and Figure 2 show that the composition of the intangible investment has changed considerably over the last three and half decades. For the first four years presented in Table 2, the economic competencies category is the largest component of intangible investment with an average share of 51%. The second component was the innovative property with an average share of 40%. However, by 2012-13, these two categories of intangibles had reversed their contribution

Figure 1 Market sector real tangible and intangible investment (1974-75 to 2012-13) 2011-12 dollars, chain volume measures



Source: Authors' estimates using the ABS national accounts and the PC report data.

Figure 2 Shares of nominal total intangible investment, by asset type (1974-75 to 2012-13) Percent



Source: Authors' estimates using the ABS national accounts and the PC report data.

ranking; economic competencies decreased to 41% while the share of innovative property increased to 47%. Investment in computerised information has dramatically increased over time, although remaining the smallest component of intangibles. Figure 2 illustrates the extent of the shift towards investment in computerised information and organisational capital over time. The share of organisational capital has increased, while that of economic competencies as a group has decreased, influenced by the decrease in brand equity and firm specific human capital. The share of innovative property decreased slightly; however, it started to recover by the end of the period as the involvement of firms in business R&D has increased noticeably during the recent years.

3.3 Intangible Capital Stocks

The paper uses the ABS stock estimates of software, mineral exploration and artistic originals. To estimate the end-of-period t stock of 'new' intangibles, $R(t)$, the perpetual inventory method (PIM) is used:

$$R(t) = N(t) + (1 - \delta)R(t - 1), \quad (1)$$

where $N(t)$ is the period t investment in intangible capital, $R(t-1)$ is the period $t-1$ real intangible capital stock and δ is the depreciation rate. The implementation of the PIM for estimating intangible capital requires an estimate of initial period 0 capital stock, R^0. Different assumptions were made in previous studies to estimate R^0. For example, CHS (2006) assumed an initial stock of zero in a specific year for each asset while others, such as the PC report, have assumed a constant rate of investment growth for the period prior to the first data point for investment and applied the formula $R^0 = N^0/(\delta + g)$, where g is equal to the average annual growth rate of intangible investment over the period of the study.

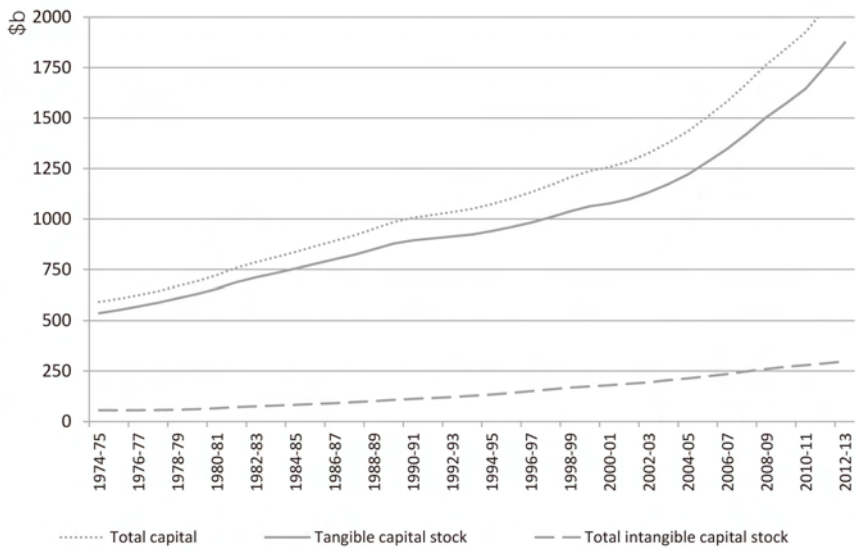
The depreciation rate, δ, used for each intangible in the PIM is reported in Table 3. The depreciation rates for software, mineral exploration and artistic originals are the average rates of the ABS for those assets. Others are the rates suggested by CHS.

Between 1974-75 and 2012-13, the total stock of intangibles grew from $50 billion to $276 billion in real terms with an average annual growth rate of 5%; see Figure 3. The real tangible capital stock increased

Table 3 Depreciation rate assumptions

Intangible	Rate (%)
Computer software	20
Innovative property	
Business R&D	20
Mineral exploration	10
Artistic originals	60
Other product development, design and research	20
Economic competencies	
Brand equity	60
Firm-specific human capital	40
Organisational capital	40

Figure 3 Tangible, intangible and total capital stock, market sector, 1974-75 to 2012-13 2011-12 dollars, chain volume measures



Source: Authors' estimates using the ABS national accounts and the PC report data.

from $540 billion to $1,732 billion over the same period - an average annual growth rate of 3%. Intangible investment increased in importance relative to tangible investment over this period. The percentage of intangible capital in total capital grew from 9% in 1974-75 to 14% in 2012-13, around 55% of which is currently capitalised.

3.3.1 The Rental Price of Intangible Capital

An estimate of the rental price (user cost) of intangible capital is required for the purpose of calculating MFP. The formula used in this paper is based on the ABS standard methodology for measuring capital services (ABS 2013):

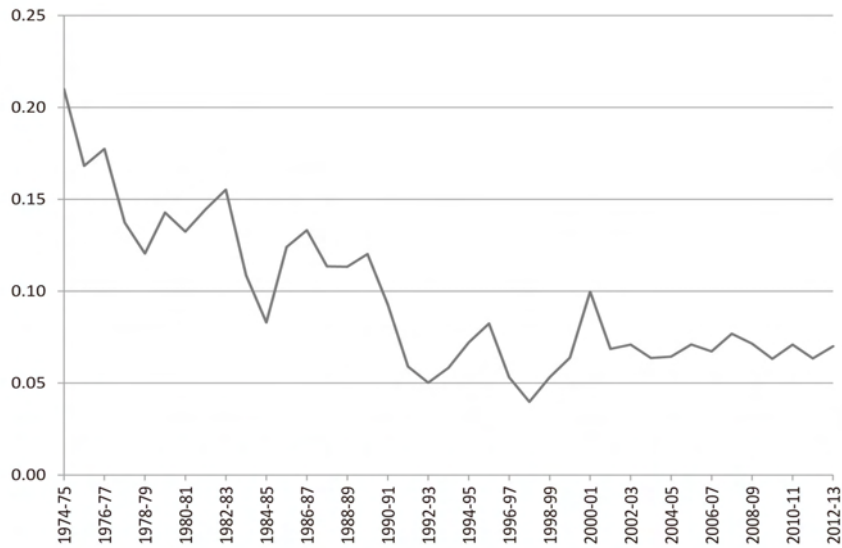
$$r_j = T_j(i.p_j + \delta_j.p_j - p_j + p_{j(t-1)}) + p_j x, \quad (2)$$

where j is the asset type, r_j is the rental price, T_j is the income tax rate, i

is the internal rate of return, p_j is the price deflator, δ_j is the depreciation rate and x is a non-income tax parameter, which is assumed to be the same for all assets types.

Two approaches have been used in previous studies for estimating the nominal rate of return on capital for intangible assets: endogenous rates of return calculated from capital income and exogenous rates of return chosen from observed market rates such as the interest rate on government bonds.[14] CHS (2006) and Marrano, Haskel and Wallis (2009) used endogenous rates of return to calculate the user cost of

Figure 4 Internal rate of return (IRR) for the market sector, all intangibles treated as capital



Source: ABS unpublished National Accounts data (IRR floor rate: CPI growth plus 4%). Estimates of the year 2012-13 is forecasted using the growth rate over the last three years of the sample.

14 An endogenous rate of return, i, can be calculated by assuming that capital income, Q, is equal to capital rent, $\sum_j r_j K_j$ where r_j is the user cost and K_j is the real capital stock. Reorganising (2) to include Q and K_j, i can be calculated as $i = Q - \sum_j K_j\,(T_j(\delta_j.p_j - p_j + p_{j(t-1)})) + p_j x/(\sum_j K_j\,T_j p_j)$.

intangible capital for the U.S. and the U.K. respectively. Van Rooijen-Horsten et al. (2008) used exogenous rates of return for the Netherlands. The PC report used the ABS hybrid methodology.[15] This paper treats the intangibles like any other fixed asset in the growth accounting; thus, the ABS exogenous interest rate is used; see Figure 4.

4. Growth Accounting with Intangible Capital

CHS (2006) demonstrate the effect of treating intangibles expenditure as investment (rather than as an intermediate input) on the National Accounts measures. Their model is based on three goods: a consumption good with real output volume in period *t* of $C(t)$ with price $P^C(t)$; a tangible investment good $I(t)$ with price $P^I(t)$; and an intangible good $N(t)$ with price $P^N(t)$. When intangibles are regarded as being intermediate goods, labour *L* and tangible capital *K* are allocated to the production of all three goods, and *N* is an input to *C* and *I*. The production functions, $F^i(\cdot)$ and flow accounts for each of the three sectors, *i = N, I, C*, are then as follows:

$$\begin{aligned}
Intangible\ sector \quad & N(t) = F^N\left(L_N(t), K_N(t), t\right); \\
& P^N(t)N(t) = P^L(t)L_N(t) + P^K(t)K_N(t), \quad (3) \\
Tangible\ sector \quad & I(t) = F^I(L_I(t), K_I(t), N_I(t), t); \\
& P^I(t)I(t) = P^L(t)L_I(t) + P^K(t)K_I(t) \\
& \qquad + P^N(t)N_I(t), \quad (4) \\
Consumption\ sector \quad & C(t) = F^C\left(L_C(t), K_C(t), N_C(t), t\right); \\
& P^C(t)C(t) = P^L(t)L_C(t) + P^K(t)K_C(t) \\
& \qquad + P^N(t)N_C(t), \quad (5)
\end{aligned}$$

15 The ABS methodology uses an endogenous rate of return unless the endogenous rate falls below the level of consumer price index (CPI) growth plus 4%. If the rate falls below this level, CPI growth plus 4% is used as the rate of return. In practice, the rate of return rarely rises above this mark and can therefore be considered to be an exogenous rate of return for most years.

where tangible capital accumulates according to the PIM. The production functions in these equations are linked to the accounting identities by the assumption that each input is paid the value of its marginal product. In this formulation, $N(t)$ is both an output and an immediate input to the production of the other products, and therefore nets out in the aggregate. Thus,

$N(t)$ does not appear in the total output, $\acute{Y}_t$, identity:

$$P^Y(t)\acute{Y}(t) = P^C(t)C(t) + P^I(t)I(t) = P^L(t)L(t) + P^K(t)K(t), \quad (6)$$

where $L = L_N + L_I + L_C$ *a*nd $K = L_N + K_I + L_C$.

If intangibles are treated as capital, a different model applies. The output of the intangible, $N(t)$, now appears in the production functions of the consumption and tangible investment sectors as a cumulative stock. In the same way as tangible capital, the intangible capital stock $R(t)$ accumulates according to the PIM. The sectoral equations become:

$$\begin{aligned}
Intangible\ sector \quad & N(t) = F^N\,(L_N(t), K_N(t), R_N(t), t); \\
& P^N(t)N(t) = P^L(t)L_N(t) + P^K(t)K_N(t) \\
& \qquad + P^R(t)R_N(t), \qquad (7) \\
Tangible\ sector \quad & I(t)\ = F^I(L_I(t), K_I(t), R_I(t), t)\,; \\
& P^I(t)I(t) = P^L(t)L_I(t) + P^K(t)K_I(t) \\
& \qquad + P^R(t)R_I(t), \qquad (8) \\
Consumption\ sector \quad & C(t)\ = F^C\,(L_C(t), K_C(t), R_C(t), t); \\
& P^C(t)C(t) = P^L(t)L_C(t) + P^K(t)K_C(t) \\
& \qquad + P^R(t)R_C(t), \qquad (9)
\end{aligned}$$

where $P^R(t)$ is the rental price associated with the services of the intangible stock. The total output, Y_t, identity must be expanded to included the flow of new intangibles on the product side and the flow of services from the intangible stock on the income side:

$$P^Y(t)Y(t) = P^C(t)C(t) + P^I(t)I(t) + P^N(t)N(t)$$
$$= P^L(t)L(t) + P^K(t)K_t(t) + P^R(t)R_t(t), \quad (10)$$

where $N = N_N + N_I + N_C$ *a*nd $R = R_N + R_I + R_C$.

Further, CHS (2006) modify the standard Solow (1957) Multifactor Productivity (MFP) growth definition to include investment in intangibles. When treated as intermediate input, intangibles expenditure does not appear in the $M\acute{F}P$ growth equation:

$$\Delta \ln M\acute{F}P = \Delta \ln \acute{Y} - \acute{s}_K \Delta \ln K - \acute{s}_L \Delta \ln L\,, \quad (11)$$

where $\acute{s}_L = P^L L/(P^L L + P^K K)$ and $\acute{s}_K = P^K K/(P^L L + P^K K)$. When treated as capital, intangibles appear as an additional input in the revised MFP growth equation, which becomes:

$$\Delta \ln MFP = \Delta \ln Y - s_K \Delta \ln K - s_L \Delta \ln L - s_R \Delta \ln R\,, \quad (12)$$

where $s_L = P^L L/(P^L L + P^K K + P^R R)$,
$s_K = P^K K/(P^L L + P^K K + P^R R)$ and
$s_R = P^R R/(P^L L + P^K K + P^R R)$.

A comparison of (11) and (12) reveals that capitalising intangibles can change the National Accounts and productivity growth in many ways. The level of aggregate output increases because it includes the value of output of the intangible goods. The share of labour income in GDP declines, while the share of capital income increases due to the expanded total capital stock. In addition, the growth of output is higher because investment in intangibles is typically expected to increase at a rate higher than that of tangible capital. The effect on MFP growth is unclear, depending on the change in output growth relative to the change in input growth. MFP may rise (fall) if capitalising intangibles raises the output growth rate by less (more) than it raises the growth in inputs.

This section presents the impact of capitalising intangibles on the components of the production function and MFP. It uses the estimates of

intangible investment presented here along with the ABS's National Accounts data on market sector GVA, labour input and input income shares.[16] Following the PC report, three different definitions of capital are used to analyse the impact of intangibles on the growth accounting estimates: (i) including tangible and all intangible assets, (ii) including tangible assets and National Accounts intangible assets only, and (iii) including tangible assets only.[17]

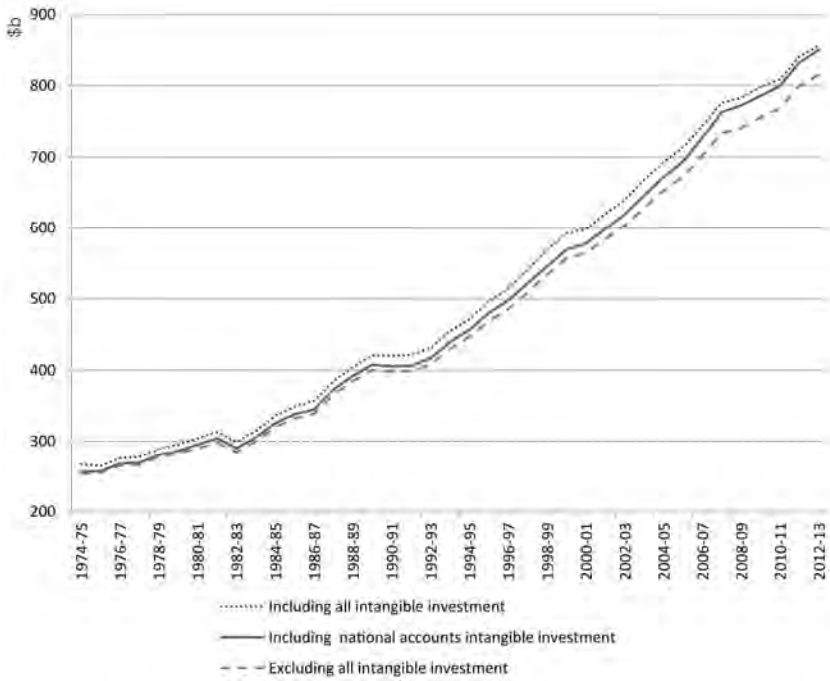
4.1. Output

Figure 5 compares market sector GVA for each of the three definitions of capital. It has been shown in Table 2 that over the period 1974-75 to 2012-13 investment in new intangibles is larger than investment in National Accounts intangibles. Thus, new intangibles make a larger contribution to the total GVA than National Accounts intangibles.

16 A detailed description of data sources for these variables is provided in the Appendix.

17 The estimates of National Accounts intangibles, and thus the ensuing MFP indexes, developed in this paper are not identical to the ABS official estimates. Several factors may explain this. (i) There is a difference in the level of aggregation at which the estimates are constructed. Due to data limitations, the paper aggregates all assets in all industries in a single stage then uses rental prices to construct capital services. On the other hand, the ABS constructs capital services indexes for each of the twelve market sector industries separately then aggregates these indexes together using relevant weights, (ii) The ABS BERD data includes some R&D related to financial services and architectural/engineering services. The scope of these types of R&D as discussed in CHS is broader than those activities that may be covered by the BERD survey. Thus, separate estimates for these types of R&D are developed and the ABS-based BERD estimates were reduced to avoid double counting. (iii) The rental prices and the PIM version used by the ABS to construct capital stock is more complex than the method used in this paper.

Figure 5 Market sector gross value added, 1974-75 to 2012-13 2011-12 dollars, chain volume measures

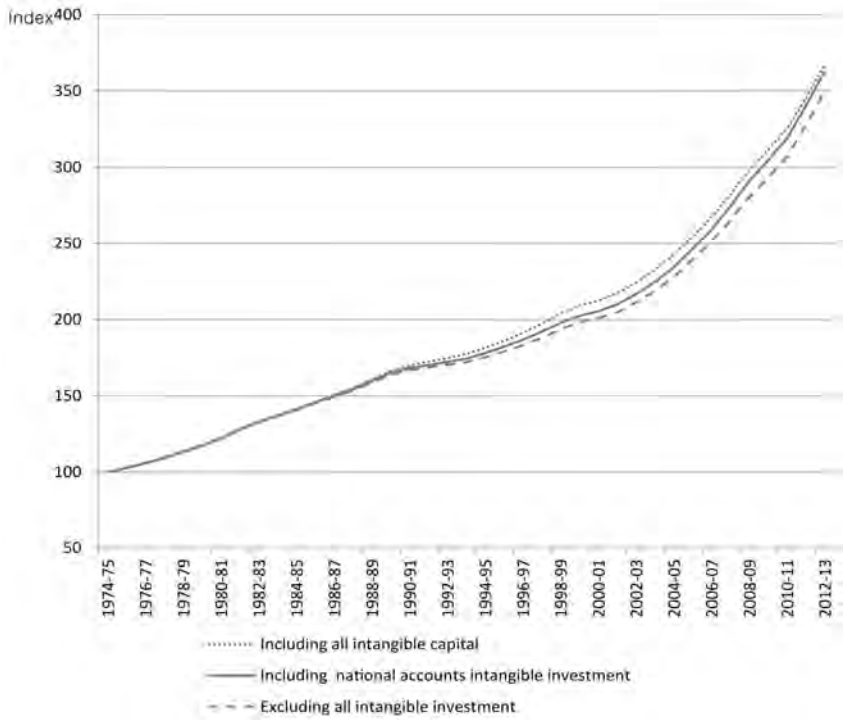


Source: Authors' estimates using the ABS national accounts and the PC report data.

4.2. Capital Services

Aggregate capital services indexes are constructed using the volume index of capital stock of each asset weighted by its rental price weight. Figure 6 presents the total capital services indexes for each of the three definitions. As shown in the figure, capitalising intangibles has increased the growth in capital services. This indicates that growth in capital services from intangibles was faster than growth in capital services from tangibles.

Figure 6 Capital services, market sector, 1974-75 to 2012-13
Index 1974-75=100



Source: Authors' estimates using the ABS national accounts and the PC report data.

4.3. Factor Income Shares

Capitalising intangibles has noticeably changed the factor income shares over the period 1974-75 to 2012-13. Table 4 shows the upward (downward) trend in the capital (labour) share of total factor income. Investment in new intangible assets increases the capital income share by a greater percentage than the National Accounts intangible assets. This is because investment in new intangibles represents a larger proportion of total investment than investment in the National Accounts intangibles.

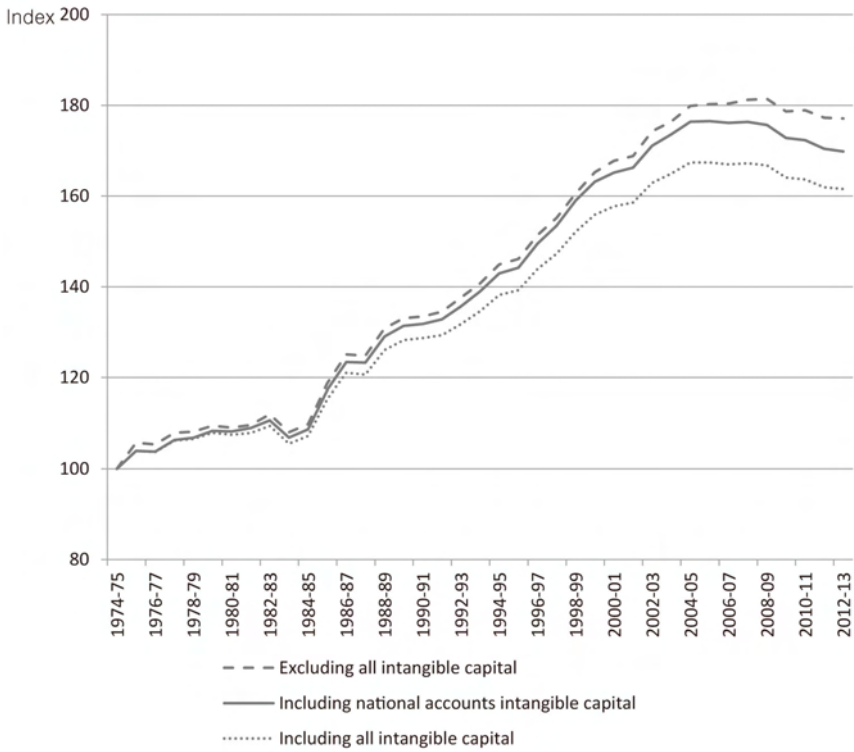
Table 4 Capital and labour income shares, market sector 1974-75 to 2012-13

	1974-75 to 1984-85	1984-85 to 1994-95	1994-95 to 2004-05	2004-05 to 2012-13
Including all intangibles				
New Intangibles	0.044	0.050	0.059	0.055
The ABS Intangibles	0.015	0.022	0.030	0.037
Intangibles	0.059	0.072	0.089	0.092
Tangibles	0.327	0.370	0.379	0.402
Total capital	0.386	0.442	0.468	0.494
Labour	0.614	0.558	0.532	0.506
Including national accounts intangibles				
The ABS Intangibles	0.016	0.024	0.032	0.039
Tangibles	0.342	0.389	0.402	0.425
Total capital	0.358	0.413	0.434	0.464
Labour	0.642	0.587	0.566	0.536
Excluding all intangibles				
Capital	0.348	0.399	0.416	0.443
Labour	0.652	0.601	0.584	0.557

4.4. Multifactor Productivity

Figure 7 shows that capitalising intangibles expenditure has changed the rate of MFP growth. In particular, it indicates that MFP growth has decreased. This can be explained by the fact that the inclusion of intangibles has raised output growth by a lower rate than it has raised the growth in inputs. Although, the rate of MFP growth has decreased across the period, the pattern of the growth remains unchanged. Specifically, the improvement in productivity during the productivity growth cycle of 1998-99 to 2003-04 and the overall decline during the recent productivity growth cycle is still present after capitalising intangibles.

Figure 7 Multifactor productivity, market sector, 1974-75 to 2012-13
Index 1974-75= 100



Source: Authors' estimates using the ABS national accounts and the PC report data.

5. Government Spending on Science and Innovation

Besides fulfilling public needs (such as improving the products and services offered or better delivery of functions), the economic rationale for governmental involvement in the area of research and innovation is the existence of market failure associated with research and innovation. This market failure is typically due to the diffusion of knowledge beyond the control of the inventor, which implies that the private rate of return to research and innovation is lower than its social return. Thus, governments intervene to eliminate this wedge between private and social returns.

Another reason for the provision of public support is that governments may want to stimulate research and innovation performed by the business sector. This is likely to be below the socially optimal level as firms are often discouraged from engaging in research activities by the inherently high risk of research (Arrow 1962). Therefore, governments intervene to assist firms either by mitigating their private costs or by raising awareness of the technological opportunities that are available to reduce both the cost and uncertainty of research and innovation.

Similar to many other OECD-member governments, the Australian government devotes a considerable amount of funding to promote research and innovation in the country. At present there are two main sources of data on public support for R&D and innovation: the Science, Research and Innovation Budget Tables (SRIBTs) and the ABS survey on R&D. With each Federal Budget, the Australian government publishes SRIBTs which provide an overview of government support for science, research and innovation over a period of ten years. The SRIBTs summarise the total of Australian Government support by sectors of performance as well as providing a decomposition of the total expenditure by program and socio-economic objectives. On the other hand, the ABS survey on public spending on R&D captures R&D expenditure at the points at which R&D is performed.

Several technical challenges make the outlays data from the SRIBTs not strictly comparable with the R&D expenditure data captured by the ABS; see Matthews and Howard 2000 for more discussion on this issue. For the purpose of econometric investigation, this paper focuses on the SRIBTs data because the breakdown of spending is more relevant to our research question. Nevertheless, extra interesting information is available from the ABS survey data, which may shed more light on Australia's innovation system. Therefore, a brief snapshot of the ABS survey data will also be presented.

The SRIBTs classify government support for research and innovation into four sectors of performance: Commonwealth research agencies, the higher education sector, the business enterprise sector, and a “multisector”. Figure 10 presents public spending estimated for the year 2012-13.

As shown in the figure, the higher education sector is the most important direct recipient of science and innovation funding from the Australian Government, receiving around 32% of total public support followed by the business enterprise sector and the multisector (or "civil" sector) which respectively received 25% and 23% of the total support. The research agencies sector has received the smallest portion of support which is equivalent to 20% of total support.

The public funds devoted to each of these sectors is allocated to different areas. An analysis of the $8.9 billion outlay by the Australian Government for R&D and innovation in 2012-13 shows the following:

5.1. Higher Education Research

The Performance Based Block Funding (PBBF) accounts for 67% of total funding to the higher education sector. The PBBF is provided through a number of 'performance based' arrangements such as the Research Training Scheme (RTS), the Institutional Grants Scheme (IGS), the Research Infrastructure Block Grants scheme (RIBG) and the Australian Postgraduate Awards scheme (APA). [18] The Australian Research Council (ARC) funding accounts for 31% of total funding to higher education. Other R&D Support accounts for 2%.

5.2. Business Enterprise Sector

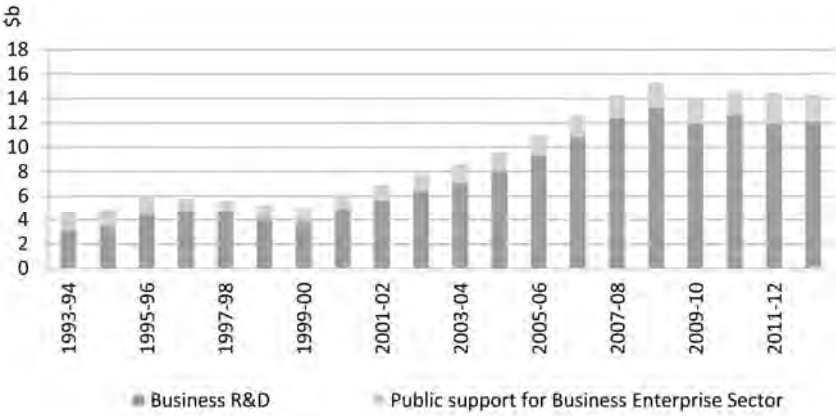
Government support for business sector science and innovation activity is delivered through a range of programs. The main program is the R&D Tax Concession which accounts for about 81% of total business support in 2012-13. Other Innovation Support and Other R&D Support account for 18% and 1% respectively.

To assess the extent to which public support, in the forms in which it is provided, contributes in financing business R&D, Figure 8 presents a

---

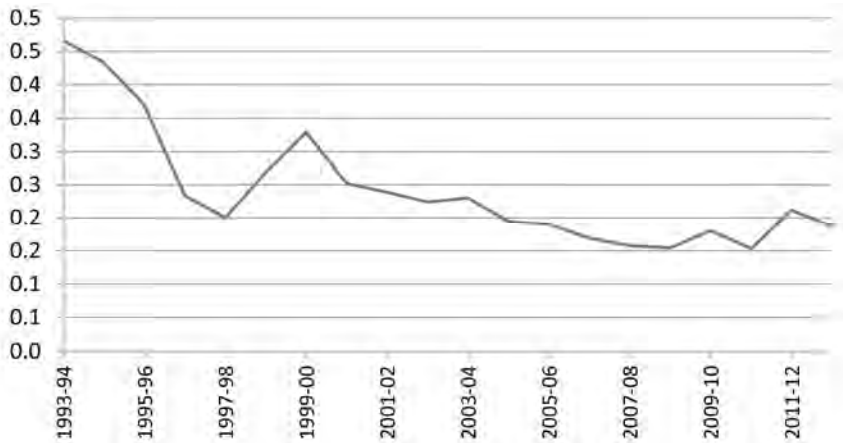
18 These arrangements are known as 'performance based' because allocations to each institution depend on its past 'performance' as assessed by various formulae administered through the Department of Education, Employment and Workplace Relations.

Figure 8 Expenditure on business R&D: relative significance of public support, 1993-94 to 2012-13 2011-12 dollars



Source: Authors' estimates using the ABS survey data on BERD and Science, Research and Innovation Budget Tables. Estimates of business R&D are calculated as the difference between BERD and public support for business enterprise spending.

Figure 9 The ratio of public support for business enterprise to business R&D spending, 1993-94 to 2012-13



Source: Authors' estimates using the ABS survey data on BERD and Science, Research and Innovation Budget Tables. Estimates of business R&D are calculated as the difference between BERD and public support for business enterprise spending.

comparison between business R&D (BERD) spending against government support for business sector. As seen in the Figure, public sector contributed

by less than 25% over the period 1993-94 to 2012-13 in funding business enterprise sector. Despite a long-run increase in the absolute amount of public support, spending has not kept up with business R&D growth, so that the ratio of public to business spending has fallen over this period (Figure 9).

5.3. Research Agencies

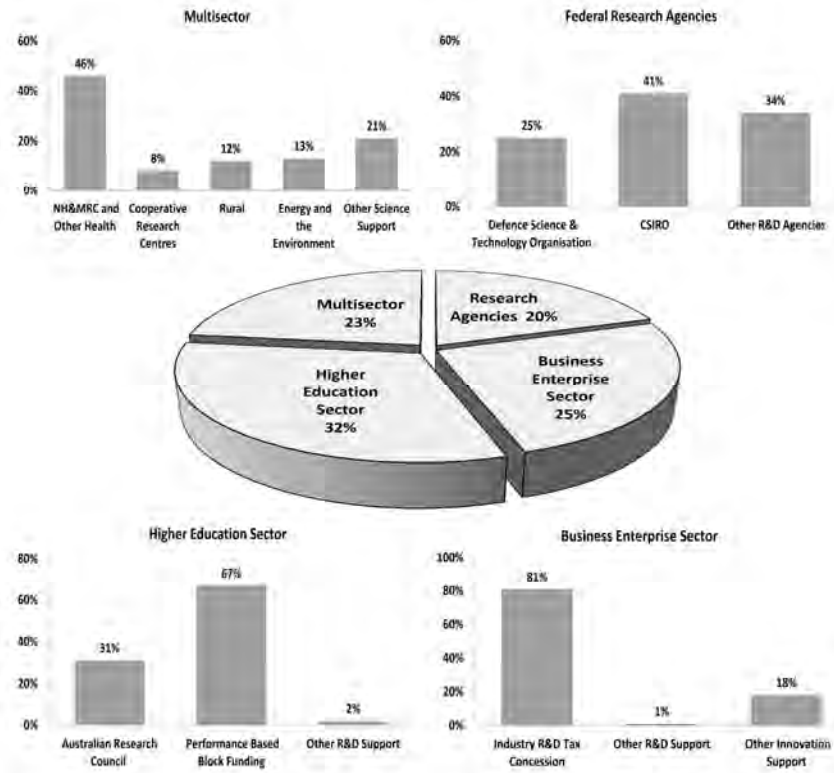
Two main organisations — the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the Defence Science and Technology Organisation (DSTO) — dominate the research funding allocated to public sector research agencies. In 2012-13, the CSIRO accounted for 41% of the total public sector research agency funding while the DSTO accounts for 25%. Other public R&D agencies account for 34%.[19]

5.4. Multisector

About 46% of the multisector funding is devoted to the National Health and Medical Research Council (NHMRC) and Other Health grants, which predominantly go to universities and private nonprofit Medical Research Institutes (MRIs). The Cooperative Research Centres (CRCs) and Rural Funds also have strong university components and they constitute around 8% and 12% of the multisector outlays respectively. Energy and the Environment has a share of 13% and the Other Science Support is 21%.

19 Other public R&D agencies include the Australian Nuclear Science and Technology Organisation (ANSTO); Geoscience Australia; Antarctic Division; Australian Institute of Marine Science (AIMS); Bureau of Meteorology Research Centre; Environmental Research Institute of the Supervising Scientist; Australian Animal Health Laboratory; Great Barrier Reef Marine Park Authority; and the Anglo-Australian Telescope.

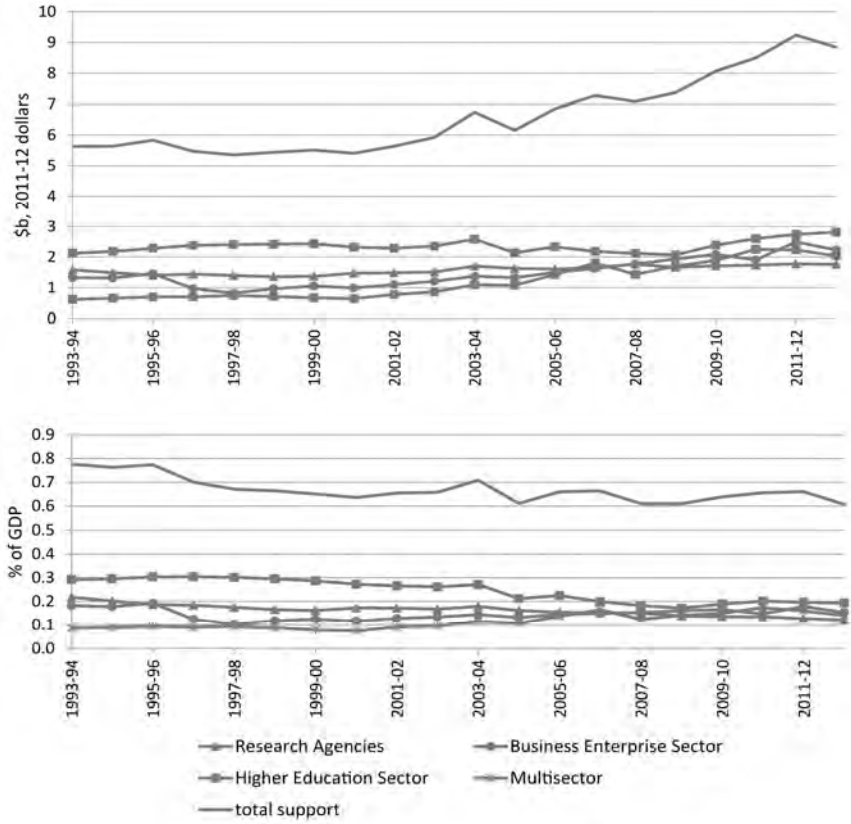
Figure 10 Australian Government spending on research and innovation 2012-13



Source: Authors' estimates using Science, Research and Innovation Budget Tables.

Figure 11 depicts a long-term perspective of the Australian Government support for research and innovation and its components. The total support has increased in real terms over the past two decades; however, it has fallen as a share of GDP. There have been noticeable changes in the role of the government support across its four components of funding. In particular, indirect public support for the business enterprise sector and the multisector has grown in real terms during the past two decades. However, support to higher education and direct support to research agencies has barely grown. This has meant that the share of public support to the multisector has roughly doubled between 1993-94 and 2012-13 while support to the higher education has halved. A number of factors can account for this changing pattern in government

Figure 11 Total Australian Government support for research and innovation 1993-94 to 2012-13



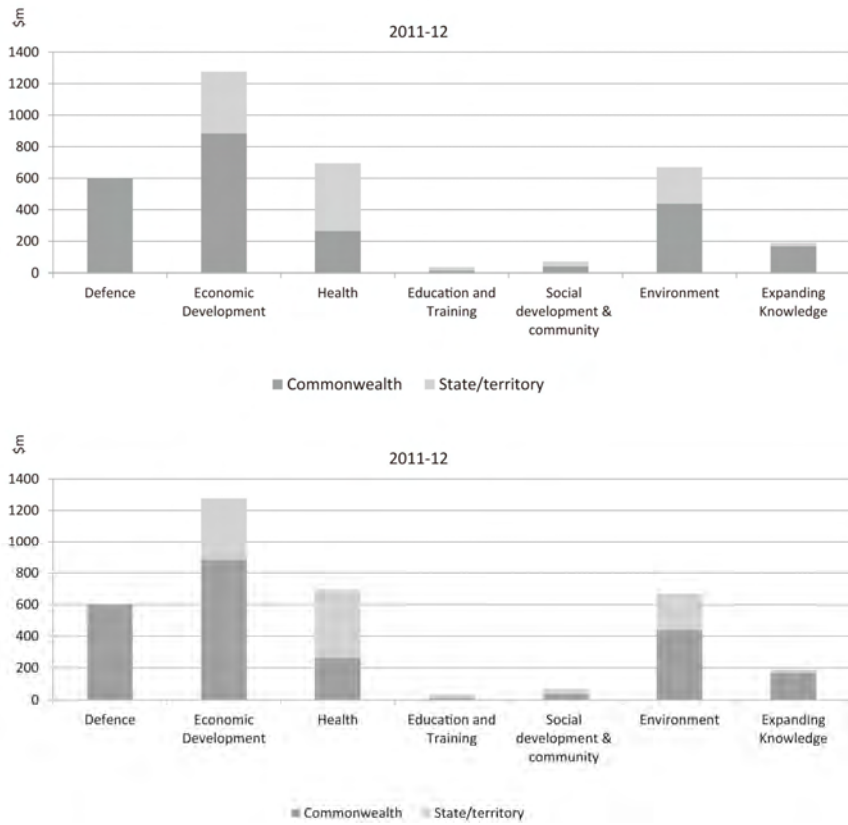
Source: Authors' estimates using Science, Research and Innovation Budget Tables.

investment including, an increased focus on collaboration in the multisector and progressive increases in claims on the R&D Tax Concession in the business enterprise sector.

To explore how public R&D resources are allocated according to the intended purpose or outcome of the research, we employ the ABS survey data on public R&D.[20] Figure 12 presents a comparison between 1992-93 and 2011-12 in breaking down expenditure on R&D by socioeconomic objective. As seen in the figure, the largest share of

20 Note that the ABS surveys have been conducted every two years.

Figure 12 Breakdown of underpinning research funded by the Commonwealth and State/territory by socio-economic objective, 1992-93 and 2011-12

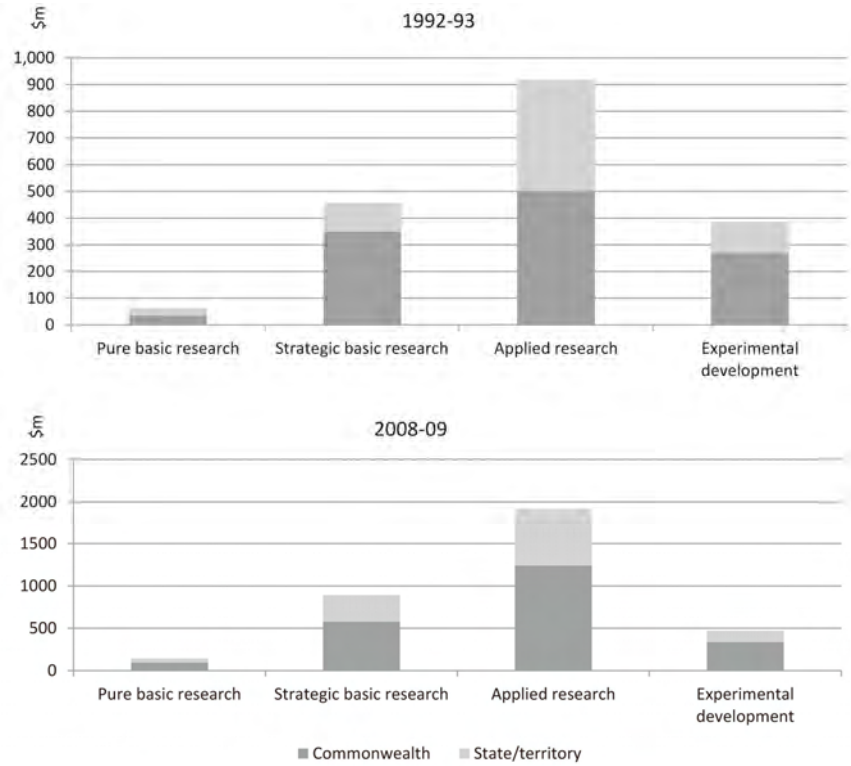


Source: Authors' estimates using the ABS survey data on public R&D.

government R&D expenditure was directed towards economic activities followed by defence and environment activities. However, social activities such as education and training, and social development and community activities receive a small share of government R&D expenditure.

The ABS data also breaks down Commonwealth expenditure on R&D by the type of activities: basic research, applied research and experimental development. Basic research is further broken into two types, pure and strategic basic research. Applied research is a critical input to the innovation system and is often seen to be more immediately

Figure 13 Commonwealth support for R&D, by type of activity, 1992-93 and 2008-09



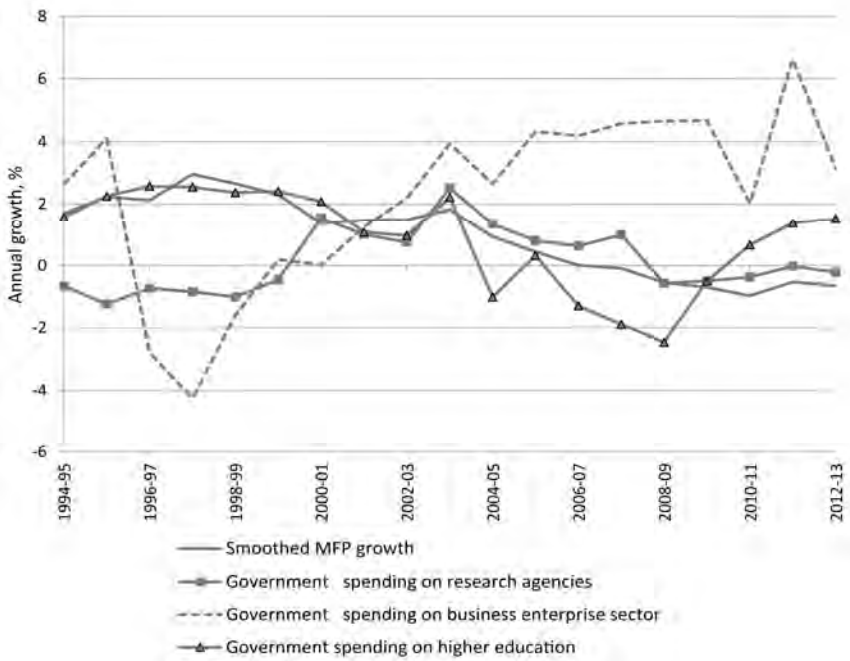
Source: Authors' estimates using the ABS survey data on public R&D.

relevant and applicable for end-users, specifically industry, than basic research. In Figure 13 it is shown that the Commonwealth and State governments focus more on applied research and strategic basic research at the expense of pure basic and experimental development research.

6. The Relation between Public Support for R&D and Market Sector MFP Growth

Trends in Australia's MFP (adjusted for the inclusion of intangibles) were outlined in Section 4. Section 5 has highlighted the key trends in

Figure 14 Market sector MFP growth and public support for research agencies, higher education and the business enterprise sector (1993-94 to 2012-13)



Note: MFP growth rates smoothed by a three year centred moving average.

Source: Authors' estimates using Science, Research and Innovation Budget Tables and the ABS national accounts data.

public spending on R&D. This section investigates the bi-variate relationship between productivity growth and the growth of the public R&D stock.[21]

Figure 14 plots MFP growth, smoothed by a three year centred moving average, against the capital stock growth of public support for

21 Most of the previous studies that examined the relationship between R&D and economic or productivity growth have avoided the problem of obtaining an estimate of R&D capital stock by employing a measure of R&D intensity (i.e. a ratio of R&D expenditures to the value of production). However, this method implicitly assumes that the depreciation rate of R&D is zero which is not necessarily a realistic assumption. The approach here is to use the stock of public sector R&D estimated by using PIM and assuming a depreciation rate identical to the business sector R&D presented in Table 3.

research agencies, higher education, and business enterprise. Productivity and public support for higher education activities are moving together throughout the period, which gives the appearance of a strong relationship. Similarly, with the exception of the early years, there is a co-movement between productivity and research agencies' activities, again suggesting a positive correlation between them. Conversely, the divergent trends in productivity and the public support for business enterprise which are seen to dominate the whole period suggest a negative relationship. However, this casual analysis presup-poses a contemporaneous relationship between R&D and productivity presented in Figure 14; it is more likely that there are lagged effects of R&D expenditure on productivity since knowledge typically takes time to disseminate. The correlations suggested by the bi-variate plots may therefore represent an overly simplistic analysis. There might also be other potential influences on productivity which could be obscuring actual causal relationships. Therefore, to provide stronger evidence on the relationship between productivity and public knowledge, a detailed econometric analysis accounting for other influences is required.

7. Econometric Analysis

7.1. The Model

In line with HW, consider the following production function:

$$Y_t = A_t F(L_t, K_t, N_t^{PRV}, N_t^{PUB}), \tag{13}$$

where Y_t, L_t, K_t, N_t^{PRV} denote value-added output, labour input, tangible and private intangible capital respectively. N_t^{PUB} is capitalised public support for research and innovation, and A_t is any increase in output not accounted for by the increase in the above four factors of production.

Assuming a general Cobb-Douglas representation for the production technology, the production function $F(\cdot)$ can be written in terms of

natural logarithms as follows:

$$\ln Y_t = \ln A_t + \sum_{X=L,K,N^{PRV}} \varepsilon_X \ln X_t + \varepsilon_{N^{PUB}} \ln N_t^{PUB}, \tag{14}$$

where $X = L, K, N^{PRV}$, and ε_X is the elasticity of output with respect to each input.

For estimation purposes, assume:

$$\ln A_t = \alpha_0 + v_t \tag{15}$$

where v_t is an iid error term. In addition, we assume that the elasticity ε_X equals the factor income share, s_X, plus a term, d_x, to account for either deviations from perfect competition or spillovers due to that factor:

$$\varepsilon_X = s_X + d_x \qquad \forall X. \tag{16}$$

Taking N_t^{PUB} to be *freely available* public support for research and innovation, it does not have an observable factor share and hence the output elasticity, $\varepsilon_{N^{PUB}}$ must be econometrically estimated.

Again following the standard Solow (1957) growth accounting approach, an expression for the level of MFP can be obtained from (14) as follows:[22]

$$\ln MFP_t \equiv \ln Y_t - \sum_{X=L,K,N^{PRV}} s_X \ln X_t. \tag{17}$$

By substituting (14) and (15), and allowing for the effect of other factors, Z_t, which are not accounted for in the calculation of MFP, (17)

22 Solow's (1957) approach is based on two simplifying assumptions: (i) competitive markets in which factors are rewarded according to their marginal products — so that the output elasticities can be represented by factor shares in total factor income, and (ii) constant returns to scale — so that factor shares sum to unity.

can be rewritten as:[23]

$$\ln MFP_t \equiv \alpha_o - \sum_{X=L,K,N^{PRV}} d_X \ln X_t + \varepsilon_{N^{PUB}} \ln N_t^{PUB} + \alpha_1 Z_t + v_t, \quad (18)$$

where α_1 is a vector of coefficients. Equation (18) is used to examine two issues of interest: First, the existence of any spillovers from the market sector investment in intangibles. This issue can be addressed by estimating the coefficient on private intangible capital. Further, by breaking intangible capital into classes (software, innovative property and economic competencies), individual effects of these classes can be estimated. Second, (18) is used to examine the impact of public support for R&D through the estimation of the elasticity parameter $\varepsilon_{N^{PUB}}$. In addition, N^{PUB} can be broken into components to assess the impact of direct versus indirect public support for R&D.

Equation (18) involves the use of time-series data in (log) levels. Thus, there is a risk that some of these series show non-stationary behaviour. Previous studies of productivity analysis have extensively discussed the issue of spurious regression which might emerge when nonstationary series are used (e.g., Tatom 1993). Often, the best practice to address the issue of spurious regression is to formally test for unit roots and cointegration. Performing these tests, we find evidence of cointegration; however, the result of this test may not be legitimate due to the small size of the sample.[24] When the series is suspected to be non-stationary, a possible solution suggested in literature is to estimate the

23 Note that there is an important difference between the regression model of this paper represented by (17) and that of HW. Instead of using the stock of public sector R&D as in this paper, HW lagged the ratio of public sector R&D expenditure to GDP and used it as a regressor, assuming a zero depreciation rate of public sector R&D.

24 The augmented Dickey-Fuller (ADF) test (Dickey and Fuller (1979)) is applied to the OLS residuals to test for the null hypothesis of no cointegration. However, a major and widely cited setback with the ADF test is the inherent low power when it is applied to short data series. The power of the test is the ability to reject the null of non-stationarity when it is false; because the ADF test has low power, it may suggest that a series has a unit root while it is stationary.

production function using first differences. Generally speaking, estimation in first differences is problematic in that it may destroy the long-run relationship between the variables of interest, since any common long-run stochastic trends in the data are removed by differencing. However, as with HW, we are also interested in the determinants of productivity growth, so (18) is rewritten in first differences (i.e. growth terms) as follows:

$$\Delta \ln MFP_t \equiv \tilde{\alpha}_o + \sum_{X=L,K,N^{PRV}} \tilde{d}_X \Delta \ln X_t + \tilde{\varepsilon}_{N^{PUB}} \Delta \ln N_t^{PUB} + \tilde{\alpha}_1 Z_t + \eta_t. \quad (19)$$

7.2. Control Variables

A number of studies suggest some external factors as possible determinants of Australia's productivity growth (e.g., Connolly and Fox 2006, Shanks and Zheng 2006). This paper focuses on a group of factors that are generally seen as important in explaining productivity. For some of these factors there is no consensus on either the size or the direction of the effects on productivity.

7.2.1. Public Infrastructure

The significant role of public infrastructure on productivity in Australia has been cited in a number of studies such as Otto and Voss (1994) and Shanks and Barnes (2008). The majority of these studies have used the official measure of the public net capital stock published by the ABS as a measure of public infrastructure. This paper constructs a more suitable measure for public economic infrastructure by utilising the engineering construction data published by the ABS. These data better represent the spending on economic infrastructure in comparison to other available data (Elnasri 2013).

7.2.2. Business Cycle

Connolly and Fox (2006) and Shanks and Zheng (2006), among others, have included a business cycle variable to control for the pro-cyclical nature of MFP. An output gap measure provides one of the methods for controlling for the effects of the business cycle. Here, following a standard approach, the output gap is measured as the difference between the natural logarithm of output and its trend. The trend is calculated using the Hodrick-Prescott filter (Hodrick and Prescott 1997) with the smoothing parameter equal to 1,600.

7.2.3. Trade Openness

Some evidence suggests that in relatively unregulated economies, such as Australia, an increase in trade openness is associated with an increase in GDP per capita (Bolaky and Freund 2004). Thus, a measure of trade openness is constructed as the sum of imports and exports over GDP.

7.2.4. Terms of Trade

Similarly to Madden and Savage (1998), a terms of trade variable (the ratio of export prices to import prices) is used to represent international competitiveness.[25]

7.3. Results

7.3.1. Spillovers from Market Sector Intangible Investment

The first column of results in Table 5 presents the effects of the aggregate market sector's intangible capital on MFP. Results are

25 In addition to the above four control variables, and in line with Connolly and Fox (2006), the paper has attempted to include West Texas crude oil prices as a proxy to control for price shocks that might have a direct impact on the world energy market. However, this variable was dropped from the regressions as it has shown small and insignificant estimates.

Table 5 Spillovers from intangible investment (1993-94 to 2012-13)

	$\ln MFP$	$\Delta\ln MFP^a$	$\ln MFP$	$\Delta\ln MFP^a$
Tangible capital	-0.175	-0.434**	-0.129*	-0.024
	(0.165)	(0.137)	(0.068)	(0.105)
Labour	-0.663***	-0.136	-0.579***	-0.098
	(0.137)	(0.096)	(0.074)	(0.057)
Intangible capital	0.579***	0.329		
	(0.062)	(0.226)		
Software			0.134***	0.100
			(0.007)	(0.059)
Innovative property			0.117*	-0.107
			(0.055)	(0.112)
Economic competencies			0.112***	0.256***
			(0.028)	(0.026)
Business cycle	0.734***	-0.094	0.826***	-0.026
	(0.181)	(0.122)	(0.118)	(0.041)
Public infrastructure	0.194	-0.149	0.038	-0.140
	(0.216)	(0.302)	(0.123)	(0.149)
Openness	0.012**	-0.001	0-.001	0.006*
	(0.005)	(0.001)	(0.000)	(0.003)
Terms of Trade (t-1)	-0.106**	-0.022	-0.022*	-0.022
	(0.044)	(0.050)	(0.011)	(0.050)
$\bar{R}^2$	0.99	0.85	0.99	0.74
Durbin-Watson	1.66	1.15	2.73	3.03
Jarque-Bera test	0.624	0.285	0.467	0.083
Number of Observations	19	18	19	18

Note: $\Delta\ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

reported for both the (log) levels and growth of MFP from estimating (18) and (19), with the exclusion of the public sector R&D terms, i.e. $\varepsilon_{N^{PUB}}\ln N_t^{PUB}$ and $\tilde{\varepsilon}_{N^{PUB}}\Delta\ln N_t^{PUB}$ from these two equations respectively. Focusing first on the (log) levels model, there is strong evidence for the positive impact of intangible capital on the market sector MFP. In particular, the elasticity of MFP relative to intangibles is 0.58 which means an increase of 1% of intangible capital can increase MFP by

0.58%. Bearing in mind that the intangibles variable is an aggregate measure that includes all types of intangible assets, the magnitude of the estimated elasticity is not entirely surprising. As mentioned earlier, the direct effects of the primary inputs (labour, tangible and intangible capital) are accounted for in the calculation of the MFP index. Thus, the estimated coefficients on these variables, presented in (18) and (19), represent indirect effects that may arise due to deviations from perfect competition and CRS assumptions, imposed on the construction of MFP, or spillovers due to those factors.

While intangible capital has positive spillovers, the negatively signed and highly statistically significant coefficients on the tangible capital and labour inputs suggest negative spillovers or decreasing returns. This is not an unusual result for the Australian market sector, as previous studies such as Shanks and Zheng (2006) and Industry Commission (1995) have also found evidence on decreasing returns to labour and capital.

Results on the coefficients of the control variables are broadly acceptable and consistent with theory and prior empirical findings. In particular, business cycle and trade openness suggest positive and statistically significant impacts on productivity. The estimate of public infrastructure coefficient has the expected positive sign, nevertheless, it is statistically insignificant.[26] Finally, while Australia has experienced a marked improvement in its terms of trade during the past decade, the significantly negative coefficient of a one-lag terms of trade variable may indicate the decline in mining productivity which has contributed substantially to a slowdown in market sector productivity growth.[27] The

---

26 A sizeable number of previous studies have found implausible results when regressing productivity on public infrastructure. Elnasri (2013) has suggested that one possible cause of such implausible results is due to the shortcomings of aggregate time-series analysis that make it unsuitable for the examination of the infrastructure spillovers to productivity. Approaches such as panel regressions and spatial econometrics are found to produce more acceptable results.

27 For more discussion on how long lead times between investment in new capacity in mining sector and the corresponding output can lead to short term movement in mining productivity and how the depletion of Australia's natural resource had significant adverse effect on long-term mining productivity see Topp, Soames, Parham and Bloch (2008).

overall fit of the model is good with the $\bar{R}^2$ being equal to 0.99. Results from Jarque-Bera (JB) test suggest that the population from which the sample is drawn follows normal distribution. This indicates the validity of hypothesis testing.[28] Because there is evidence of serial correlations detected by the Durbin-Watson statistic test, Newey-West standard errors are applied in drawing statistical inferences.[29]

The second model describes the effects on MFP growth which is, as in HW (2013), smoothed by a three-year moving average. The result on all variables turn to be statistically insignificant in this specification.

To explore which components of intangible capital are the major drivers of the positive relationship found above, the second column of Table 5 presents the breakdown of intangibles into its three components, i.e. computerised information (software), innovative property, and economic competencies. When the (log) levels model is estimated, the three components of intangibles have positive and significant coefficients with reasonable magnitudes. However, in the differences model, only the result for economic competencies remains robust while the impacts of software and innovative property appear insignificant.

Including the three components of intangibles simultaneously in the same model may raise a problem of multicollinearity which can produce large standard errors and hence have an impact on the statistical inferences. To address this problem, three individual regressions are performed to separately examine the effects of software, innovative property and economic competencies and the results are reported in Tables 6, 7 and 8 respectively.

---

28 Jarque-Bera test presents a test for normality based on skewness and kurtosis. It tests the null hypothesis H_0: normal distribution, skewness is zero and excess kurtosis is zero; against the alternative hypothesis H_1: non-normal distribution.

29 Durbin-Watson statistic (d) tests the null hypothesis H_0 that the errors are uncorrelated against the alternative hypothesis H_1 that the errors are autocorrelated. If the errors are white noise, d will be close to 2. If the errors are strongly autocorrelated, d will be far from 2.

Table 6 Spillovers from software (1993-94 to 2012-13)

	$\ln MFP$	$\Delta \ln MFP^a$
Tangible capital	-0.153	-0.549*
	(0.142)	(0.246)
Labour	-0.604***	-0.036
	(0.089)	(0.135)
Software	0.192***	0.076
	(0.016)	(0.106)
Business cycle	1.206***	-0.061
	(0.111)	(0.083)
Public infrastructure	0.205	-0.595
	(0.221)	(0.353)
Openness	-0.003	0.007
	(0.003)	(0.003)
Terms of Trade (t-1)	-0.001**	0.015
	(0.045)	(0.044)
$\bar{R}^2$	0.99	0.66
Durbin-Watson	1.11	1.07
Jarque-Bera test	0.206	0.914
Number of Observations	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

Table 7 Spillovers from innovative property (1993-94 to 2012-13)

	$\ln MFP$	$\Delta \ln MFP^a$
Tangible capital	-0.372	-0.641**
	(0.233)	(0.264)
Labour	-0.993***	0.004
	(0.157)	(0.150)
Innovative property	0.672***	-0.117
	(0.065)	(0.154)
Business cycle	1.283***	-0.01
	(0.188)	(0.095)
Public infrastructure	0.772**	-0.443
	(0.310)	(0.441)

Table 7 (Continue)

	$\ln MFP$	$\Delta \ln MFP^a$
Openness	0.016	0.007
	(0.010)	(0.004)
Terms of Trade (t-1)	-0.174**	0.024
	(0.075)	(0.037)
$\bar{R}^2$	0.98	0.70
Durbin-Watson	2.13	1.12
Jarque-Bera test	0.846	0.936
Number of Observations	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

Table 8 Spillovers from economic competencies (1993-94 to 2012-13)

	$\ln MFP$	$\Delta \ln MFP^a$
Tangible capital	0.348	-0.369***
	(0.264)	(0.098)
Labour	-0.373	-0.092
	(0.243)	(0.057)
Economic competencies	0.357***	0.239***
	(0.080)	(0.028)
Business cycle	0.279	0.054
	(0.387)	(0.073)
Public infrastructure	-0.149	-0.055
	(0.433)	(0.183)
Openness	0.020*	0.005**
	(0.009)	(0.002)
Terms of Trade (t-1)	-0.137	-0.001
	(0.083)	(0.025)
$\bar{R}^2$	0.98	0.92
Durbin-Watson	1.31	2.00
Jarque-Bera test	0.692	0.035
Number of Observations	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

While in the (log) levels model the three components of intangibles have positive and statically significant effects on productivity, in the differences model, only economic competencies persists in having a significant result. This finding is consistent with the one reported in Table 5. Nevertheless, there is a noticeable increase in the magnitude of the coefficients of innovative property and economic competencies which show a higher impact on productivity. The significant increase of these coefficients may result from bias due to the omission of some variables or a specification problem caused by the small number of observations. In terms of the goodness of the fit, the values of $\bar{R}^2$ remain large and results of Jarque-Bera test suggest normality. Again due to the presence of serial correlation, Newey-West standard errors are reported.

To summarise, the interpretation of the results obtained from the above regression analysis is that there is strong evidence of a positive impact from private sector investment in intangibles on MFP at the market sector level. In line with the findings of the New Growth Theory literature, the significant coefficient on intangibles can be interpreted as productivity gains from increasing returns due to the development of 'know-how' to do business, or knowledge spillovers beyond firms' borders. Because the MFP index used in the regressions takes into account of the treatment of intangibles as capital assets (a necessary adjustment to ensure that the private return to intangibles is not captured in MFP) the obtained estimates of the intangible coefficients may only reflect knowledge spillovers or increasing returns. These results contrast with those of HW, who did not find evidence for significant spillovers from private sector intangibles.

7.3.2. Spillovers from Public Support for R&D

This part discusses the impact of public support for R&D on the market sector MFP, measured by the coefficients $\varepsilon_{N^{PUB}}$ in (18) and $\tilde{\varepsilon}_{N^{PUB}}$ in (19). Because public R&D is not included in the construction of MFP, the estimates of these coefficients will reflect social returns from public R&D. Results from regressing both the level and growth of MFP on an aggregate measure of the public sector stock of R&D are presented in the first column of Table 9. To avoid possible omitted

Table 9 Spillovers from total public support (1993-94 to 2012-13)

	$\ln MFP$	$\Delta \ln MFP^{a}$	$\ln MFP$	$\Delta \ln MFP^{a}$
Tangible capital	-0.423**	-0.523**	-0.168	-0.223*
	(0.173)	(0.179)	(0.118)	(0.103)
Labour	-0.758***	-0.025	-0.547***	-0.008
	(0.159)	(0.108)	(0.117)	(0.094)
Intangible capital	0.461***	0.328	0.440***	0.339*
	(0.066)	(0.217)	(0.067)	(0.166)
Total public support	0.399**	-0.235		
	(0.143)	(0.243)		
Research agencies			0.349***	0.007
			(0.064)	(0.141)
Higher education			0.175*	0.324*
			(0.076)	(0.157)
Business enterprise			-0.056	-0.056
			(0.060)	(0.066)
Multisector			-0.032	-0.021
			(0.035)	(0.051)
Business cycle	1.188***	-0.073	0.876***	0.083
	(0.202)	(0.119)	(0.156)	(0.134)
Public infrastructure	0.521**	-0.328	0.308	-0.432
	(0.222)	(0.319)	(0.211)	(0.404)
Terms of Trade (t-1)	-0.106**	-0.018	-0.037	-0.009
	(0.044)	(0.050)	(0.027)	(0.022)
$\bar{R}^2$	0.99	0.75	0.99	0.87
Durbin-Watson	1.41	1.42	2.33	2.33
Jarque-Bera test	0.732	0.514	0.167	0.320
Number of Observations	19	18	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

variable bias, the two models include the business cycle, public infrastructure and a one-lag terms of trade control variables.

As indicated by the results from the levels model, there are significant spillovers from total public R&D stock to private MFP, with an estimated elasticity of 0.40. While this result suggests a beneficial effect from governmental involvement in the area of research and innovation, it is not informative whether all, or only some, types of the public support are effective as the estimated coefficient is associated

with an aggregate measure of public R&D. Before looking into this issue in more detail, it is noted that the model has good fit, as suggested by the large value of $\bar{R}^2$, the coefficients of the three control variables possess the expected signs and they are statistically significant. The estimated coefficient for the private sector intangibles supports the findings of Table 5 of positive spillovers on productivity. Moreover, there is again evidence of decreasing returns from labour and tangible capital. However, the findings of the levels model are not robust to first differencing.

As outlined in Section 5, public sector R&D is divided into four classes: research agencies, higher education sector, business enterprise sector, and multisector. Breaking down the aggregate stock of public sector R&D into four classes and running the two regressions results in the estimates in the second column of Table 9. They suggest that the observed positive spillovers from public sector R&D are mainly driven by the spending on research agencies and higher education sectors, while the insignificant coefficients on business enterprise sector and multisector variables suggest no impact on productivity.

The robustness of the findings presented in Table 9 is examined by performing several additional regressions to examine the individual effects of each of the four classes. Accordingly, eight models are developed by replacing aggregate public sector R&D in (18) and (19) by a measure of each of these four classes. The first column of Table 10 shows that the result of the impact of spending on research agencies remains robust in the levels model. However, one concern with the contemporaneous relationship presented above is that the significance of the results is sensitive to the dating of the public sector R&D variable. Even though our capital measure is constructed using all previous investments, there might be some lagged effects of public sector R&D that are not captured in the contemporaneous model, which mask the relationship between MFP and public R&D. Thus, to allow for these lagged effects, another set of regressions is performed by replacing the contemporaneous research agencies stock variable by a one-lag stock measure. Results presented in the second column of Table 10 provide

support for all the findings of the contemporaneous model.[30] It can be noted that this strong positive relationship between MFP and government pending on research agencies is consistent with the what is observed in Figure 14.

Table 10 Spillovers from public support (1993-94 to 2012-13): Research agencies

	$\ln MFP$	$\Delta \ln MFP^a$	$\ln MFP$	$\Delta \ln MFP^a$
Tangible capital	-0.381***	-0.586***	-0.419***	-0.505**
	(0.111)	(0.163)	(0.120)	(0.130)
Labour	-0.789***	-0.043	-0.845***	-0.007
	(0.097)	(0.103)	(0.133)	(0.121)
Intangible capital	0.477***	0.289	0.449***	0.078
	(0.038)	(0.276)	(0.053)	(0.268)
Research agencies	0.295***	-0.1376		
	(0.052)	(0.193)		
Research agencies (t-1)			0.358***	-0.490*
			(0.048)	(0.163)
Business cycle	1.163***	-0.006	1.274***	-0.033
	(0.130)	(0.135)	(0.182)	(0.086)
Public infrastructure	0.778***	-0.373	1.025***	-1.023**
	(0.197)	(0.365)	(0.249)	(0.388)
Terms of Trade (t-1)	-0.066**	-0.003	-0.092**	0.022
	(0.034)	(0.050)	(0.038)	(0.032)
$\bar{R}^2$	0.99	0.75	0.99	0.85
Durbin-Watson	2.02	1.42	2.17	1.85
Jarque-Bera test	0.656	0.514	0.876	0.517
Number of Observations	19	18	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms*,**,*** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

30 With a small number of available observations, only one lag is used. HW did not use capital stock of public sector R&D in their regressions. Instead, they used two and three lags of the ratio of spending on public R&D relative to output, assuming by this a zero depreciation rate. Recalling that the PIM employed in this paper for constructing public knowledge capital includes all previous investment expenditures in the accumulation process, it is somewhat equivalent to the model of HW but with more lags.

An extra investigation is made to examine the source of positive spillovers from research agencies. A replication of the above regressions is performed using a breakdown of research agencies capital stock into defence (i.e., DSTO) and non-defence research agencies (i.e., other R&D agencies such as CSIRO, ANSTO, AIMS and so forth). Results of the breakdown, shown in Table 11, indicate that the source of the spillovers is driven mainly by non-defence R&D agencies.

Table 11 Spillovers from public support (1993-94 to 2012-13): Research agencies - breakdown

	$\ln MFP$	$\Delta\ln MFP^a$	$\ln MFP$	$\Delta\ln MFP^a$
Tangible capital	-0.347***	-0.543**	-0.390***	-0.509***
	(0.072)	(0.217)	(0.063)	(0.147)
Labour	-0.696***	-0.007	-0.527***	0.027
	(0.097)	(0.143)	(0.110)	(0.119)
Intangible capital	0.414***	0.308	0.224***	0.095
	(0.053)	(0.263)	(0.027)	(0.293)
Research agencies (x defence)	0.256***	-0.038		
	(0.040)	(0.145)		
Defence	-0.065	-0.140		
	(0.075)	(0.196)		
Research agencies (x defence)(t-1)			0.295***	-0.339
			(0.031)	(0.204)
Defence (t-1)			-0.417	-0.176
			(0.073)	(0.178)
Business cycle	1.120***	-0.024	1.206***	-0.028
	(0.107)	(0.145)	(0.097)	(0.108)
Public infrastructure	0.698***	-0.304	0.857***	-0.953
	(0.127)	(0.361)	(0.104)	(0.748)
Terms of Trade (t-1)	-0.045	-0.002	0.026	0.029
	(0.027)	(0.043)	(0.028)	(0.033)
$\bar{R}^2$	0.99	0.72	0.99	0.84
Durbin-Watson	1.80	1.27	2.70	1.82
Jarque-Bera test	0.702	0.777	0.837	0.618
Number of Observations	19	18	19	18

Note: $\Delta\ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

Similar regressions to those presented above are performed to examine the impact of public support for higher education. The results, presented in Table 12, suggest significant positive spillovers from higher education R&D, which remain robust across both the levels model, growth model and the incorporation of the lagged effects. This finding is consistent with the relationship shown in Figure 14.

Next, market sector MFP is regressed on a stock measure of government support for business sector science and innovation to assess whether or not there were benefits to the business enterprise sector from

Table 12 Spillovers from public support (1993-94 to 2012-13): Higher education sector

	$\ln MFP$	$\Delta\ln MFP^a$	$\ln MFP$	$\Delta\ln MFP^a$
Tangible capital	-0.162	-0.263**	-0.074	-0.254**
	(0.197)	(0.115)	(0.155)	(0.177)
Labour	-0.375	-0.046	-0.426***	-0.044
	(0.221)	(0.093)	(0.101)	(0.064)
Itangible capital	0.535***	0.357**	0.412***	0.460***
	(0.061)	(0.127)	(0.062)	(0.097)
Higher education	0.305**	0.409***		
	(0.120)	(0.123)		
Higher education (t-1)			0.352***	0.378***
			(0.068)	(0.116)
Business cycle	0.519***	0.148**	0.594***	0.075
	(0.230)	(0.076)	(0.075)	(0.070)
Public infrastructure	-0.288	-0.378	-0.098	-0.170
	(0.362)	(0.324)	(0.184)	(0.262)
Terms of Trade (t-1)	-0.019	0.002	-0.045	-0.015
	(0.044)	(0.022)	(0.025)	(0.028)
$\bar{R}^2$	0.99	0.89	0.99	0.89
Durbin-Watson	1.10	2.44	1.92	2.54
Jarque-Bera test	0.656	0.422	0.534	0.810
Number of Observations	19	18	19	18

Note: $\Delta\ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

the R&D Tax Concession and other sources of innovation and R&D support. Results for the four models are reported in Table 13. There is no evidence of contemporaneous gains from government indirect spending on R&D to the market sector MFP. These findings are consistent in both the levels and growth models. When a lagged effect is allowed for, some evidence of negative spillovers from public funding appears, as the coefficients are negatively signed and statistically significant. Note that this negative relationship between MFP and government spending on the business enterprise sector was observed earlier in Figure 14.

Table 13 Spillovers from public support (1993-94 to 2012-13): Business enterprise sector

	$\ln MFP$	$\Delta \ln MFP^a$	$\ln MFP$	$\Delta \ln MFP^a$
Tangible capital	-0.445***	-0.306**	-0.259	-0.480**
	(0.218)	(0.129)	(0.197)	(0.158)
Labour	-0.795***	0.078	-0.522***	-0.018
	(0.205)	(0.090)	(0.154)	(0.118)
Intangible capital	0.604***	0.228	0.589***	0.328*
	(0.050)	(0.166)	(0.063)	(0.097)
Business enterprise	0.108	-0.179**		
	(0.073)	(0.061)		
Business enterprise (t-1)			-0.077	-0.150*
			(0.052)	(0.068)
Business cycle	1.100***	-0.106	0.684***	-0.062
	(0.257)	(0.090)	(0.168)	(0.099)
Public infrastructure	0.558	-0.534	0.199	-0.205
	(0.330)	(0.377)	(0.282)	(0.338)
Terms of Trade (t-1)	-0.087	-0.033	-0.071	0.015
	(0.058)	(0.026)	(0.063)	(0.036)
$\bar{R}^2$	0.99	0.84	0.99	0.82
Durbin-Watson	1.26	1.84	1.23	1.87
Jarque-Bera test	0.896	0.396	0.564	0.888
Number of Observations	19	18	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

Finally, the fourth category of public sector spending on R&D — multisector — is examined. Results reported in Table 14 suggest no significant contemporaneous spillovers to market sector MFP. As in the case of the business enterprise sector, when a lagged effect is allowed for there is some evidence of negative spillovers.

As seen from the above results, Australian government support for research and innovation has different impacts on market sector productivity depending on the spending component. Specifically, spending on Government research agencies - other than defence - and higher education institutions yields productivity gains, while no evidence

Table 14 Spillovers from public support (1993-94 to 2012-13): Multisector/Civil

	$\ln MFP$	$\Delta \ln MFP^a$	$\ln MFP$	$\Delta \ln MFP^a$
Tangible capital	-0.336	-0.510**	-0.252	-0.538**
	(0.204)	(0.177)	(0.251)	(0.199)
Labour	-0.680***	-0.065	-0.584***	-0.047
	(0.159)	(0.096)	(0.158)	(0.110)
Intangible capital	0.530***	0.455**	0.630***	0.429*
	(0.084)	(0.166)	(0.074)	(0.206)
Multisector	0.048	-0.084		
	(0.038)	(0.077)		
Multisector (t-1)			-0.046	-0.047
			(0.099)	(0.059)
Business cycle	0.945***	-0.029	0.814***	-0.024
	(0.165)	(0.105)	(0.165)	(0.122)
Public infrastructure	0.425*	-0.267	0.248	-0.224
	(0.236)	(0.238)	(0.277)	(0.306)
Terms of Trade (t-1)	-0.100	-0.009	-0.084	-0.005
	(0.070)	(0.040)	(0.061)	(0.046)
$\bar{R}^2$	0.99	0.78	0.98	0.74
Durbin-Watson	1.16	1.45	2.70	1.18
Jarque-Bera test	0.957	0.956	0.645	0.586
Number of Observations	19	18	19	18

Note: $\Delta \ln MFP$ smoothed by three-year moving average. Terms in brackets are heteroskedasticity and autocorrelation robust Newey-West standard errors. Terms *, **, *** denote significance at the 10%, 5% and 1% levels respectively. Figures corresponding to Jarque-Bera test are the probabilities (p-values) of rejecting the null hypothesis H_0: residuals are normally distributed against H_1: residuals are not normally distributed.

of gains are observed from spending on business enterprise and civil sectors. These findings are consistent with those of HW who find strong evidence of market sector productivity benefits from public spending on research councils and no evidence of market sector spillovers from public spending on civil or defence R&D.

8. Conclusion

In a world defined by a finite endowment of resources, the contribution to economic growth through resources utilisation is limited; therefore, sustained economic growth in the long term will have to come from productivity enhancements. Investments in research and innovation (such as information technology, R&D, skills development, design and organisational improvements and other types of intangible assets) are central drivers of productivity. They create more efficient services and production processes, more effective workplace organisation and open up new markets.

Despite the prospects of research and innovation in boosting productivity and economic growth, they remain an area that is surrounded by many complexities and challenges, such as difficulties in measurement and inefficiency of provision.

To better understand and improve the functioning of the innovation systems of an economy, it is essential to track, or benchmark, performance. However, developing robust and relevant measures of research and innovation is hard. The intangible nature of research and innovation poses problems for the measurement of spending and the depreciation of spending in defining capital stocks. As such, research and innovation are largely ignored in the National Accounts and corporate financial reports of many countries where they have been only treated as intermediate expenditure. However, excluding investment in these intangible assets means that investment is underestimated, and this may distort measures of growth in capital services and consequently, productivity.

Another complication is the provision of a socially optimal level of research and innovation. It is commonly argued that there are major market failures in the provision of a sufficient amount of knowledge

capital because knowledge diffuses beyond the control of the inventor, which implies that the private rate of return for research and innovation is lower than its social return. Additionally the high risks involved discourage firms from engaging in such activities. For both reasons, the amount invested by firms in research activities in a competitive framework is likely to be below the socially optimal level. This justifies intervention by governments to directly make their own investments in knowledge capital or to indirectly support the private sector to reduce its costs. However, governments face the stumbling block of a large number of projects competing for tight budgets. This study has attempted to make a contribution to the advancement of knowledge about spillovers from innovation through the explicit inclusion of both privately and publically funded intangibles in the analysis.

Recent literature defines a range of private sector intangible assets. This paper extends, and makes more current, Australian studies on three classes of intangible assets (computerised information, innovative property and economic competencies). Next, the paper incorporates these intangibles in the National Accounts to examine their effects on growth accounting components. The paper finds that: (i) investment in intangibles has increased at a faster rate than investment in tangibles over time. (ii) The share of labour income in GDP has declined while the share of capital income has increased due to the expanded total capital stock. (iii) A comparison of the new estimates of MFP (adjusted for the capitalisation of all intangible assets) with the traditional MFP, reveals a noticeable reduction in MFP growth.

Further, the paper utilises the new estimates of MFP in performing a range of regressions for investigating possible spillovers from private and public sectors' knowledge capital. Unlike estimates of the traditional MFP, adjusted estimates have the advantage of isolating social from private returns to knowledge capital. Insufficient attention has been paid to this refinement in the measured rates of return in past studies.

The results provide some policy-relevant insights. First, measuring research and innovation by only focusing on the set of assets which are currently capitalised in the System of National Accounts seems unreliable. Total investment has been found to be under-reported and this has distorted measures of growth in capital services and consequently, productivity.

Second, accumulation of private sector knowledge capital is a source of positive spillovers to market sector productivity. Third, given the pressures on public finances, it is appealing to direct the innovation budget to areas with higher social benefits. The empirical findings suggest that government research agencies and higher education are areas with more potential gains.

Specifically, the paper examines the impact of four classes of public support for research and innovation — Commonwealth research agencies, higher education sector, business enterprise sector and multisector — on market sector MFP growth. The paper finds strong evidence of productivity benefits from public spending on Commonwealth research agencies and higher education. However, the results suggest no evidence of spillover effects on private productivity from public support to the business enterprise sector, multisector or defence R&D. Some reasons for this can be postulated. Health research funding makes up almost 50% of the public expenditure on the multisector in 2012-13; see Figure 10. Its output is not part of market sector value added, and any productivity effects are likely to be very long-run, through improvements in the health of the workforce and population more generally; hence there is a bias against finding a positive significant result. Similarly, it is expected that while some select components of the expenditure on defence may result in innovations with commercial value that appear in the market sector, defence services will not, again biasing the results against finding a positive relationship. The main public support for the business enterprise sector research and innovation is the industry R&D tax concession, comprising 81% of support to the sector in 2012-13. Unlike much of the funding to higher education and research agencies, allocation of support is based on expenditure rather than being performance based. Obviously there are strong financial incentives for firms to maximise the expenditures classified as being related to R&D, potentially biasing the results. In addition, there may be other policy goals of the R&D tax concession than raising productivity. Indeed, providing incentives for the establishment of small innovative firms may actually lower productivity as new entrants often initially have lower productivity compared with incumbents; see e.g. Baldwin (1995) and Aw et al. (2001).

There remain several areas of improvement for future research. The main source of the data on intangibles used in this paper is from the PC report (Barnes and McClure 2009). As indicated by the authors of that report, due to data limitations and measurement challenges these estimates need further refinements and improvement. In addition, a longer time series is important for having more confidence in the regression analysis and the precision of the results. For example, a longer time series can allow for appropriate lag structures for the purpose of forming knowledge stocks. Another issue not taken into account, but one able to provide further insights, is to allow for heterogeneity and variability by modelling for firm or sector behaviour. However, this requires appropriate data at the level of firms or industry. Flexible functional forms are widely recommended to address the complex relationship between output, primary inputs and external factors. With more observations, a more flexible form for modelling the production technology, such as the translog functional form can be adopted. Last but not least, an interesting extension to the presented analysis would be to assess excess returns to private and public sectors' knowledge capital, a useful technique for identifying areas where there is either an excess of, or lack of provision.

References

ABS (2011), 'Research and experimental development, businesses'. Australian Bureau of Statistics, Cat. no. 8104.0.

ABS (2013), 'Australian system of national accounts concepts, sources and methods'. Australian Bureau of Statistics, Cat. no. 5216.0.

Adams, J. (1990), 'Fundamental stocks of knowledge and productivity growth', *Journal of Political Economy* 98(4), 673-702.

Arrow, K. (1962), 'The economic implications of learning by doing', *The Review of Economic Studies* 29, 155-173.

Aw, B.Y., X. Chen and M.J. Roberts (2001), 'Firm level evidence on productivity and turnover in Taiwanese manufacturing', *Journal of Development Economics* 66, 51-86.

Baldwin, J.R. (1995), *The dynamics of industrial competition: a North American perspective*, Cambridge University Press, Cambridge.

Baldwin, J.R., W. Gu and R. Macdonald (2012), 'Intangible capital and productivity growth in Canada'. Research paper, the Canadian Productivity Review, Statistics Canada.

Barnes, P. (2010), 'Investments in intangible assets and Australia's productivity growth: Sectoral estimates'. Productivity Commission, Staff Working Paper.

Barnes, P. and A. McClure (2009), 'Investments in intangible assets and Australia's productivity growth'. Productivity Commission, Staff Working Paper.

Bolaky, F. and C. Freund (2004), 'Trade, regulations and growth'. The World Bank, Policy Research Working Paper Series no. 3255.

Burgio-Ficca, C. (2004), 'The impact of higher education research and development on Australian gross state product'. Deakin University, Faculty of Business and Law, School of Accounting, Economics and Finance, Economics Series 01/2004.

Connolly, E. and K. Fox (2006), 'The impact of high-tech capital on productivity: Evidence from Australia', *Economic Inquiry* 44(1), 50-68.

Corrado, C., C. Hulten and D. Sichel (2005), Measuring capital and technology: An expanded framework, *in* H. J. Corrado, C. and D.Sichel, eds, 'Measuring Capital in the New Economy'. Studies in Income and Wealth, vol. 65, the University of

Chicago Press.

Corrado, C., C. Hulten and D. Sichel (2006), 'Intangible capital and economic growth', *NBER Working Paper No. 11948*. National Bureau of Economic Research, Cambridge, Massachusetts.

de Rassenfosse, G. (2012), Intangible assets and productivity growth - an update. Melbourne Institute of Applied Economic and Social Research University of Melbourne, commissioned by the Department of Industry, Innovation, Science, Research and Tertiary Education (DIISRTE).

Dickey, D. and W. Fuller (1979), 'Distributions of the estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* 74, 427-431.

Elnasri, Amani (2013), 'Three essays on infrastructure investment: the Australian experience'. PhD thesis, School of Economics, UNSW.

Fukao, K., S. Hamagata, T. Miyagawa and K. Tonogi (2009), 'Intangible investment in Japan: Measurement and contribution to economic growth', *The Review of Income and Wealth* 55, 717-736.

Griliches, Z. (1998), *R&D and productivity: the econometric evidence*, The University of Chicago Press, Chicago.

Hao, J., V. Manole and B. van Ark (2009), 'Intangible capital and growth - an international comparison'. Economics Program Working Paper Series No. 08 - 14, The Conference Board, New York, December.

Haskel, J. and G. Wallis (2010), 'Public support for innovation, intangible investment and productivity growth in the UK market sector'. IZA, Discussion paper series No. 4772.

Haskel, J. and G. Wallis (2013), 'Public support for innovation, intangible investment and productivity growth in the UK market sector', *Economics Letters* 119, 195-198.

Hodrick, R. and E. Prescott (1997), 'Postwar U.S. business cycles: An empirical investigation', *Journal of Money, Credit and Banking* 29, 1-16.

Industry Commission (1995), 'Research and development'. Inquiry Report no. 44, AGPS, Canberra.

Jalava, J., P. Aulin-Ahmavaara and A. Alanen (2007), 'Intangible capital in the Finnish business sector, 1975-2005'. ETLA Discussion Paper. Helsinki, Finland. The Research Institute of the Finnish Economy. No. 1103.

Lichtenberg, F. (1993), 'R&D investment and international productivity differences', *NBER Working Paper No. 4161* . Cambridge, MA.

Louca, J. (2003), Multifactor productivity and innovation in Australia and its states, in C. Williams, M. Draca and C. Smith, eds, 'Productivity and Regional Economic Performance in Australia'. Collection of papers prepared by the Office of Economic

and Statistical Research, Queensland Treasury.

Madden, G. and S. Savage (1998), 'Sources of Australian labour productivity change 1950-1994', *The Economic Record* 74(227), 362-372.

Marrano, M.G. and J. Haskel (2006), 'How much does the UK invest in intangible assets?'. Department of Economics Working Paper No. 578, Queen Mary, University of London, London.

Marrano, M.G., J. Haskel and G. Wallis (2009), 'What happened to the knowledge economy? ICT, intangible investment, and Britain's productivity record revisited', *The Review of Income and Wealth* 55, 686-716.

Matthews, M. and J. Howard (2000), 'A study of government R&D expenditure by sector and technolog'. Emerging Industries Occasional Paper 3, Department of Industry, Science and Resources, Canberra.

Mohnen, P. (2001), International R&D spillovers and economic growth, *in* M. Pohjola, ed., 'Information Technology, Productivity, and Economic Growth: International Evidence and Implications for Economic Development', pp. 50-71. Oxford University Press, London.

Nadiri, I. (1993), 'Innovations and technological spillovers', *NBER Working Paper No. 4423* . Cambridge, MA.

Otto, G. and G. Voss (1994), 'Public capital and private sector productivity', *Economic Record* 70(209).

Parham, D. (2006), 'Empirical analysis of the effects of R&D on productivity: Implications for productivity measurement?'. OECD Workshop on Productivity Measurement and Analysis, Bern Switzerland.

Park, W. (1995), 'International R&D spillovers and OECD economic growth', *Economic Inquiry* 33, 571-591.

PC (2007), 'Public support for science and innovation'. Productivity Commission, Research Report.

Poole, E. and J. Bernard (1992), 'Defence innovation stock and total factor productivity growth', *Canadian Journal of Economics* 25(2), 438-452.

Romer, P. (1990), 'Endogenous technological change', *Journal of Political Economy* 98, S71-S102.

Shanks, S. and P. Barnes (2008), 'Econometric modelling of infrastructure and Australia's productivity'. Productivity Commission, Internal Research Memorandum, cat no. 08-01.

Shanks, S. and S. Zheng (2006), 'Econometric modelling of R&D and Australia's productivity'. Productivity Commission, Staff Working Paper.

Solow, R. (1957), 'Technical change and the aggregate production function', *Review of Economics and Statistics* 39(3), 312-320.

Tatom, J. (1993), 'The spurious effect of public capital formation on private sector productivity', *Policy Studies Journal* 21(2), 391-395.

Topp, V., L. Soames, D. Parham and H. Bloch (2008), 'Productivity in the mining industry: Measurement and interpretation'. Productivity Commission Staff Working Paper.

van Rooijen-Horsten, M., D. van den Bergen and M. Tanriseven (2008), 'Intangible capital in the Netherlands: A benchmark'. Statistics Netherlands, Discussion paper 08001.

Appendix

1. Description and Data Sources for Intangibles Data

Data on investment in 'national accounts' intangibles covering the period 1974-75 to 2012-13 is sourced from the ABS website. Data on 'new' intangibles is taken from the PC report and de Rassenfosse (2012). A brief description of the data sources, methodologies and assumptions made by these two studies to construct complete time series over the period 1974-75 to 2010-11 is provided below. A more detailed description is provided in the appendixes of these studies. This paper has extended these series to 2012-2103 by assuming linear growth rates in the recent years.

National Accounts Intangibles:

Computerised Information Time series for gross fixed capital formation and capital stock are available for the full period 1974-75 to 2012-13 (ABS, Cat. no. 5204.0). The ABS defines computer software as: '... computer programs, program descriptions and supporting materials for both systems and applications software. Included are purchased software and, if the expenditure is large, software developed on own-account. It also includes the purchase or development of large databases that the enterprise expects to use in production over a period of more than one year. The ASNA does not separately identify databases from computer software as recommended by the 2008 SNA' (ABS 2013, p.653).

Business Expenditure on R&D (BERD) Shanks and Zheng (2006) have constructed series for Australian BERD for the market sector

(excluding Agriculture, forestry & fishing) covering the period 1968-69 to 2002-03. The PC report has updated and extended their series to 2005-06. For this paper it is extended to 2012-13 using revised and updated data from the ABS Research and Experimental Development, Businesses. The ABS survey of BERD covers scientific R&D and R&D in social sciences and humanities. It also covers some aspects of product development costs in the financial industry and new architectural and engineering designs. The ABS defines R&D activity as 'creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications' (ABS 2011, p.33). It should be noted that CHS (2005) have included R&D related to financial services and architectural/engineering services in sperate broader categories. Thus, as this paper abides by their methodology in defining and categorizing intangibles, R&D related to these services are deducted from the ABS estimates of BERD to avoid double counting.

Mineral Exploration Time series for gross fixed capital formation and capital stock are available for the full period 1974-75 to 2012-13 (ABS, Cat. no. 5204.0). The ABS defines mineral exploration as: '... the value of expenditures on exploration for petroleum and natural gas and for non-petroleum mineral deposits. These expenditures include prelicence costs, licence and acquisition costs, appraisal costs and the costs of actual test drilling and boring, as well as the costs of aerial and other surveys, transportation costs etc., incurred to make it possible to carry out the tests' (ABS 2013, p.662).

Artistic Originals Time series for gross fixed capital formation and capital stock are available for the full period 1974-75 to 2012-13 (ABS, Cat. no. 5204.0). The ABS defines entertainment, literary or artistic originals as: 'original films, sound recordings, manuscripts, tapes, models, etc., on which drama performances, radio and television programming, musical performances, sporting events, literary and artistic output, etc. are recorded or embodied. Included are works produced on own-account. In some cases there may be multiple originals (e.g. films)'(ABS 2013, p.655).

'New' Intangibles:

New Product Development in the Financial Industry The PC report has constructed expenditure series for total intermediate purchases by the financial services industries covering the period 1974-75 to 2005-2006 using the ABS data from input-output (IO) and supply use (SU) tables. The SU industry codes equivalent to ANZSIC 73 and 75 are 380 Finance and 400 Services to finance, investment and insurance. The PC report estimated the investment series as 20% of total intermediate purchases.[31] To extend the series, de Rassenfosse (2012) has collected investment data for the period 2004-08 from Input-Output Tables (ABS Cat. no. 5215.0.55.001). The series is then used to estimate a growth rate in investment, which is applied to the 2004-05 data point from the PC series to obtain an investment series for 2005-06 to 2010-11. [32]

New Architectural and Engineering Designs The PC report has constructed expenditure series of the revenue of architectural and engineering industries using different data sources, this includes data on the ANZSIC 7821 Architectural services industry and ANZSIC 7823 Consulting engineering industry which are sourced from the ABS Industry survey (Cat. no. 8155.0), the ABS data from input-output (IO) and supply use (SU) tables in addition to unpublished data. Several assumption were made to estimate data for the missing years and backcast the aggregate series so that to cover the period 1974-75 to 2005-06. Investment is estimated by the PC report as 50% of the revenue of architectural and engineering industries. de Rassenfosse (2012) has extended the series by using turnover data from the ABS Counts of Australian Businesses including Entries and Exits (Cat. no. 8165.0) for the classes of Architectural Services and Engineering Design and Engineering Consulting Services.

Advertising Annual data on Australian advertising expenditure is available from Commercial Economic Advisory Service of Australia

---

31 Refer to Appendix A of the PC reports for more details on the assumptions made to backcast and forecast sections of the data series.

32 Refer to de Rassenfosse (2012) for a discussion on how a beak in the series caused by the change in ANZSIC classification is treated.

(CEASA). Based on data for the United States and United Kingdom about the proportion of production costs in total advertising costs, the PC report suggests doubling CEASA series to account for production costs. [33]

Market Research Following CHS who estimate this intangible as twice the revenue of the market and consumer research industry, the PC report has constructed a series as the double of the revenue of the market research industry using different data sources, mainly the ANZSIC 7853 Market research services (ABS, Cat. no. 8155.0). Interpolation and backcast were made by the PC report to complete the series.

Firm-specific Human Capital CHS suggest that spending on firm-specific human capital can be measured by the costs of employer-provided workforce training, which consists of two types: direct firm expenses (outlays on in house and external training courses) and wage and salary costs of employee time spent in informal and formal training. Unfortunately, there is no single data source provides a time series of Australian employer-provided training expenditure. The PC report emphasises that the estimated series of firm-specific human capital is only 'indicative' because several data sources were used and a number of assumptions were made to construct this series. The main data source used by the report was from the ABS (Cat. no. 6278.0) Education and Training Experience, which has been updated only once since 2005 (in 2009). Because the ABS has made a significant definitional changes to the work-related training module in 2007, the data on work-related training became not comparable over time. To overcome this problem, de Rassenfosse (2012) has used the share of investment in firm-specific human capital to to forecast observations over the period 2006-07 to 2010-11.

Purchased The PC report has constructed a series as 77% of the

33 Refer to the PC report explanation on how these estimates for the total Australian economy have been scaled down to market sector estimates.

revenue for ANZSIC 7855 Business management services available from Australian Industry (ABS Cat. no. 8155.0) by using a few available data sources (1998-99 to 2004-05). To backcast the series from 1998-99 to 1974-75, the report uses the growth in Market and business consultancy services from the product details of the ABS IO tables (Cat. no. 5215.0). de Rassenfosse (2012) has extended the series to 2010-11 by using turnover data from the ABS (Cat. no. 8165.0) Counts of Australian Businesses, including Entries and Exits for the class 6962 Management Advice and Related Consulting Services.

Own Account The PC report constructed the series to cover the period 1974-75 to 2005-06 as 20 % of salaries of Managers and Administrators (excluding farm managers and IT managers) using the ABS data on average weekly earnings and number of Managers & administrators, available from the Employee Earnings, Benefits and Trade Union Membership (EEBTUM) survey (Cat. no. 6310.0). de Rassenfosse (2012) has extended the series to 2010-11 by using the ABS data on employee earnings, benefits and trade union membership (Cat. no. 6310.0).

2. Growth Accounting Data

Output Gross value-added (ABS, Cat. no. 5204.0).

Labour Total hours worked (ABS, Cat. no 6291.0.55.003).

Tangible Capital Stocks on machinery and equipment and non-dwelling construction sourced from the ABS (Cat. no. 5204.0).

Labour's Income Share (compensation of employees + labour component in mixed income)/total income (ABS, Cat. no. 5204.0).

All Tangibles and Intangibles Capital's Income Share (gross operating surplus+the capital component in mixed income+all investment in new intangibles)/total income (ABS, Cat. no. 5204.0).

National Accounts Capital's Income Share (gross operating surplus + the capital component in mixed income)/total income (ABS, Cat. no. 5204.0).

Tangibles Capital's Income Share (gross operating surplus + the capital component in mixed income - GFCF in the national accounts intangible assets)/total income (ABS, Cat. no. 5204.0).

MFP Indexes are constructed as the ratio of output index, *Y*, over input index, $Q_{L,K}$, where L is labour input, K is capital stock (with three different defintions: including all tangible assets and all intangible assets, including all tangible assets and National Accounts intangible assets, or including tangible assets only). $Q_{L,K}$ is constructed using the Tornqvist index number approach described by the following formula:

$$\mathrm{Q}_{\mathrm{L,K}}^{0,1} = \prod_{i=l,k} \left(\frac{q_i^1}{q_i^0} \right)^{\frac{1}{2}(s_i^0 + s_i^1)},$$

where q_i^t is the quantity of labour and capital input at period *t*; s_i^t is the labour and capital income share at period *t*; and $t = 0, 1$.

CHAPTER 3

Assessing an Efficiency Defense: The Case of Intel's Marketing Campaign*

by
Hwa Ryung Lee
(Korea Development Institute)
Andras Pechy
(University of Zurich)
Michelle Sovinsky
*(University of Zurich and CEPR)***

Abstract

Antitrust authorities typically try to establish exclusivity and the anticompetitiveness of loyalty rebates through pricing, but do not address the strategic use of advertising and, more generally, marketing campaigns. In this paper we focus on non-price anticompetitive behavior arising from marketing. We propose a Test of Advertising Predation (TAP) that can be used to detect non-price predatory behavior. The TAP test is based on structural approach and allows us to disentangle the potential positive impact of a marketing program from the anticompetitive predatory effect. We apply the TAP test to the Intel case, but it can be used to guide antitrust authorities in future cases, as it provides a more general framework for testing for the anticompetitive use of

* We thank Chloe Michel for excellent research assistance. We wish to thank Heski Bar Isaac, David Byrne, Alon Eizenberg, Yang Li, Carlos Noton, Minjae Song, Steve Stern, and Otto Toivanen for helpful comments and suggestions and seminar participants at Conference on the Economics of Advertising and Marketing (Tel Aviv), ETH Zurich, Korea Development Institute, Korea University, MaCCI Summer Institute, Toulouse, the University of Washington, the University of Warwick, and the University of Zurich.
** Corresponding author is Michelle Sovinsky (michelle.sovinsky@gmail.com).

marketing campaigns. We use the test to examine whether Intel's choice of processor marketing via PC firms is consistent with predatory behavior, and find evidence that the "Intel Inside" marketing campaign had predatory effects.

1. Introduction

Generally, predatory pricing is a price reduction that is profitable only because of the additional market power gained from excluding or otherwise inhibiting the rival from competing. However, the predator may also induce rivals to exit the industry via non-price predation. Predatory investments could be made in excessive capacity, product differentiation, advertising, etc. For example, excessive investments that have the objective and likely effect of weakening or eliminating competitors can be predatory. Indeed in many predatory situations, pricing is only one aspect of anticompetitive behavior.

We present a framework to test if firms are using marketing/ advertising campaigns in an anticompetitive fashion. Our "Test of Advertising Predation (TAP)" is based on the presumption that, if a firm's marketing campaign is not predatory, marketing expenses should be efficient (i.e., profit maximizing) and so should result in sufficient increased product demand to justify costs.[1] The TAP test examines if the return on advertising (i.e., how it impacts demand) is high enough to justify marketing expenditures (as these are directed at increasing demand). The test is based on a structural approach that allows us to disentangle the potential positive impact of a marketing program from potential anticompetitive predatory effects. Constructing the TAP test does not require any more "fancy" econometrics than that needed to estimate a model of demand. We use this test to examine if the Intel Inside marketing program, which provided marketing support for firms that sold Intel processors, was used in a predatory fashion (during 2002-2005). TAP results suggest short-term profit sacrifice by Intel over this

1 Our test is related to advertising but as advertising is (a very important) marketing tool, we use the terms marketing and advertising interchangeably, while realizing that marketing can involve more than just advertising (e.g., corporate training).

period, which indicates that the Intel Inside campaign was used for predation.

While there is a vast theoretical literature on predation there are relatively few empirical studies, and these focus exclusively on pricing predation. Related papers in the price predation literature include: Weiman and Levin (1994) who examine predatory behaviour by Southern Bell Telephone Company between 1894 to 1912 when independent phone companies were trying to enter the market. Granitz and Klein (1996) provide evidence that Standard Oil engaged in predatory behavior by threatening to withhold inputs from railroads that were not in the railroad cartel. Genesove and Mullin (2006) estimate the price-cost margin in the sugar industry. They find that the price-cost margin was negative during price wars and predation was profitable in that it established a reputation as a tough competitor. Ellison and Ellison (2011) examine entry deterrence behavior in the pharmaceutical market prior to patent expiration by focusing on the asymmetry in detailing activities in markets of different size. Similarly, Chen and Tan (2007) focus on whether detailing in the pharmaceutical industry is consistent with predation incentives. Finally, Snider (2009) and Besenko, Doraszelski, and Kryukov (2010) estimate dynamic models of predatory pricing.

Our work contributes to the recent stream of research that uses structural models to study strategic behavior in the market for CPUs and PCs. These include papers by Salgado (2008a), Song (2007, 2010), Gordon (2009), and Goettler and Gordon (2009) who study the upstream CPU market. This literature mostly abstracts from PC manufacturer and PC characteristics when estimating CPU demand, and assumes that final consumers buy CPUs directly. Instead we model consumer's choice of a PC and use it to infer CPU demand. An advantage of our approach is that we can more easily estimate the effect of advertising by a PC firm on demand. Given that the Intel Inside program is the marketing subsidy from Intel to PC firms, this will allow us to estimate the effect of the Intel Inside on demand more directly. As a result our work is related to the literature on estimating demand in the PC industry. Papers in this area include Eizenberg (2011), Sovinsky Goeree (2008), Prince (2008), and Gowrisankaran and Rysman (2007). Finally, we estimate the impact of advertising on PC demand, which is related to work by Sovinsky

Goeree (2008) and Salgado (2008b).

This paper is structured as follows. We describe the TAP test and explain how to implement it in section 2. We discuss why the Intel Inside campaign is a useful application of the TAP test and discuss the data we will use to conduct the test in section 3. In section 4, we develop and estimate a model of demand and present the parameter estimates. In section 5, we compute the marginal revenue of the market campaign, and use this to construct the tests presented in section 6. In section 7, we provide robustness tests and model specifications to address limitations of the TAP test. In the final sections we note the policy implications of this test and provide our concluding thoughts.

2. Test of Advertising Predation (TAP)

Predation is not a sensible business strategy if it cannot drive a rival out of a market, discipline a rival not to compete against a predator, or if the predator cannot maintain market power for a sufficient period of time after predation. Predation can be seen as an investment in long-run market power, and, as such we propose that marketing/advertising predation has two components: short-term profit sacrifice and long-term recoupment. We develop a Test of Advertising Predation (TAP) based on the presumption that, if the marketing/advertising campaign is not predatory, marketing expenses should be efficient (i.e., profit maxim-izing), and, so should result in sufficient increased product demand to justify costs.

The first step to construct TAP is to estimate product demand as a function of marketing variables. The second step is to use the demand estimates to compute firm marginal revenue derived from the marketing campaign. That is, one should compute the marginal revenue of marketing/advertising dollars spent on the marketing campaign (at the firm or product level). Notice that the marginal revenue of the marketing program depends on the parameters of consumer utility (including advertising/marketing variables), price and marginal manufacturing cost. Given data on marketing costs, the final step of TAP is to compare the marginal marketing revenue to the observed marketing costs. When observed marketing marginal costs are above the 95 percent confidence

interval for the estimated marketing marginal revenue, we conclude that the marketing program is not consistent with profit maximization and, more specifically, that there was an excessive marketing subsidy. In this case the TAP test result is "positive."

Alternatively, the efficiency of the marketing program could be estimated by explicitly modeling how the marketing program works and estimating the parameters of the profit function. However, such an approach would have some non-trivial complications given that the way marketing programs function differs across firms and due to the fact that joint estimation of demand, supply, and optimal marketing choices would quickly complicate the econometrics. Furthermore, the resulting test would be program specific, and hence not applicable in many situations. A benefit of the TAP efficiency test is that it allows us to circumvent modeling a firm's profit maximization problem (and hence their marketing program) while still allowing for an easy to compute efficiency test.

2.1. Step 1: Estimate Demand

The first part of the test requires the researcher to specify a model of demand as a function of marketing variables. Notice that the demand specification can take any form as long as the marketing variables may have an impact on demand. Some common demand specifications include logit based (e.g., SOURCE) perhaps with different levels of nests (e.g., Goldberg, 1995) or random coefficients (e.g., Berry, Levinsohn, Pakes, 1995; Nevo, YEAR). The benefits and drawbacks of these various specifications are well documented in the empirical IO literature (SOURCE). The demand specification chosen depends on a number of considerations including, the data that are available, the time available to conduct the TAP test, as well as the nature of the market being considered.

In addition to choosing a product demand system, it is important to consider how the marketing variables may impact demand. Some items to consider include the nature of the impact of marketing on demand. For example, is advertising assumed to influence the choice set of consumers (e.g., Sovinsky Goeree, 2008), provide additional utility to

the consumed good (e.g., SOURCE), OTHER EXAMPLES from marketing literature.

For the sake of illustration, we assume that only aggregate data are available and specify product specific market shares (this would arise from a random coefficients BLP type demand, for example). Specifically, let $s_j(p, m \mid \Theta)$ represent the predicted market share for product j, as a function of a vector of all product prices (p) and a vector of all product marketing variables (m), where Θ represents the parameters to be estimated. If the market is one with multiproduct firms then firm f would sell the subset of products denoted by $\mathcal{J}_f$.

2.2. Step 2: Compute Marginal Marketing Revenue

The second step uses the demand side estimates to compute firm f's marginal revenue from the marketing campaign. In other words, this step computes the marginal revenue of marketing/advertising dollars spent on the campaign (this can be done at the firm or product level). The marginal revenue of the marketing program depends on the parameters of consumer utility (including marketing, price and marginal manufacturing cost (denoted mc_j).

Suppressing time notation, the total marketing/ad revenue (TMR) of firm f is given by

$$TMR_f = \sum_{j \in \mathcal{J}_f} (p_j - mc_j)\, \mathcal{M} s_j(p, m),$$

where $\mathcal{J}_f$ is the set of firm f's products; $\mathcal{M}$ is the market size; $s_j(p, m)$ is market share, which depends on product prices in the market (p) and marketing (m). Firm f's marginal revenue from the marketing campaign (*MMR*) is given by

$$MMR_f = \mathcal{M} \sum_{r, j \in \mathcal{J}_f} (p_j - mc_j) \frac{\partial s_j(p, m)}{\partial m_r}$$

where a multiproduct firm will consider the impact on its complete

product line arising from a change in marketing for its product *r*. One important caveat: to compute the marginal marketing revenue it is necessary to have data on (or an estimate of) the marginal cost of production.

2.3. Step 3: Compare to Marginal Cost of Marketing Campaign

The final step of TAP is to compare the marginal marketing revenue to the observed marketing costs. However, in most situations the researcher will not observe marginal marketing costs. To overcome this data restriction a solution implemented in the predatory pricing literature is to use average variable costs to proxy for marginal costs.[2] In this setting, the researcher does not require a proxy for marginal costs of production, but rather a proxy for marginal marketing costs. However, the same solution can be applied to use average variable marketing costs as a proxy for marginal marketing costs.

There may be a concern whether the average cost of the marketing program is a proper measure of the marginal cost of the marketing program. This issue arises in the predatory pricing literature as there is a practical difficulty in determining the nature of production cost: whether it is variable or fixed, or whether it is a sunk cost. However, this is less of a concern as it relates to marketing as there are few fixed or sunk components to advertising expenditures. Aside from practical difficulties, average variable cost may not be a good proxy for marginal marketing cost in the presence of returns to scale of the marketing program. The TAP results will reflect the assumption that firms' responsiveness to the

2 There are other proposals for how to calculate marginal costs associated with predatory pricing behavior. For example, Bolton, Brodley and Riordan (2000, 2001) suggest that the relevant cost is not average variable cost but the long run average incremental cost. This is measured by the per unit cost of producing the predatory increment of output where all costs that were incurred (regardless of when they were incurred) are considered. Specifically, it is calculated as the firm's total production cost less what the firm's total cost would have been had it not produced the predatory units divided by the quantity of the product produced. There is no analogy for the advertising predation measure that we could construct without measuring the output produced under the predatory behavior.

marketing program is constant.

2.4. Step 4: Consider TAP Limitations

There are few issues to consider when using the TAP efficiency test. First, TAP is fundamentally a test of profit sacrifice, hence, a case would need to be made that long-term profit recoupment is possible. Second, short-term profit sacrifice can be rationalized by potential dynamic efficiency reasons such as learning-by-doing, promotional purposes, or network externalities. As the TAP test examines if the return on advertising (i.e., how it impacts demand) is high enough to justify marketing expenditures (as these are directed at increasing demand) the model includes only the current, short-term effect of advertising. Hence the potential long-run benefit of the marketing program is not taken into account. Note this is less of a concern when considering firms that have been operating for a while: just as the efficiency reasons for pricing below marginal cost are not usually applicable to an already dominant, incumbent firm with a large customer base, in this setting too an unprofitable advertising/marketing subsidy is not easily justified by efficiency reasons.

A third issue relates to the brand-loyalty-building effect of marketing programs. Advertising is generally believed to build goodwill and this may be a reason to invest in marketing at the expense of short-term profits. Notice that this critique is less important if there is some time element during which the marketing program behavior is being investigated. The reason being that the incentive to build goodwill is constant across all periods but the predatory motive would only be present during certain periods. That is, we can see whether the TAP results are different between potentially anticompetitive periods and competitive periods. Similarly, if an anticompetitive case does not involve all firms in a market so that we have candidate firms and competitive firms, then we can compare the TAP results for candidate firms with those for competitive firms if not all firms in a market are involved in the anticompetitive case. In general cases, one way to examine the dynamic efficiency argument is to include lagged marketing variables in the demand specification. Another way to address this concern is to

determine how many future periods of MR would need to be considered by firm to rationalize current period MCs.

3. Application of TAP to "Intel Inside" Campaign

3.1. Background on "Intel Inside" Campaign

Intel has been investigated for predatory (pricing), exclusionary behavior, and the abuse of a dominant position in the market for central processing units (CPU). According to U.S. lawsuits, Intel used narketing loyalty rebates, payments, and threats to persuade computer manufacturers, including Dell and Hewlett-Packard (HP), to limit their use of AMD (Intel's main rival) processors. In their investigations, U.S. antitrust authorities focused on whether the loyalty rebates used by Intel were a predatory device in violation of the Sherman Act. The European Commission (EC) brought similar charges and imposed a 1.06 billion Euro fine on Intel for abuse of a dominant position. South Korean and Japanese antitrust authorities also imposed fines on Intel for breach of antitrust regulations.

In the case of Intel, an important component to the case involved their marketing campaign, "Intel Inside," which provided marketing support for firms that sold Intel CPU chips. Specifically, it is a cooperative advertising program in which Intel contributes a percentage of the purchase price of processors to a pool for PC firms to use to market Intel-based computers. According to the rules of the program PC firms can receive a rebate of their marketing expenditures if they include the Intel logo in their advertising. By the end of the 1990s, Intel had spent more than $7 billion on the marketing campaign (Moon and Darwall, 2005).

Intel was accused of using the marketing program to attempt to prevent computer makers from offering machines with non-Intel computer chips. It became clear through correspondence that Intel was trying to circumvent antitrust laws by using non-price predatory avenues. For example, a 2002 Dell document states that the "original basis for the [Intel marketing] fund is … Dell's loyalty to Intel." The document explains

that this means "no AMD processors."[3] The beginning of the alleged anticompetitive use of the Intel Inside program coincides with the introduction by their main rival AMD's Athlon chip (in 1999). Antitrust documentation shows that Intel issued "conditional rebates" from December 2002 to December 2005, whereby they would give rebates to some PC firms (Dell in particular) under the condition that the PC firm buy exclusively from Intel.[4] Otherwise, Intel would retract the marketing rebate and instead use the market development money to fund competitors. An internal Dell presentation (in 2003) noted that if Dell switched to AMD, Intel's retaliation "could be severe and prolonged with impact to all LOBs [Lines of Business]." Intel allegedly treated HP, Lenovo, and Acer similarly. For example, Intel rebates were conditional on HP buying 95% of its microprocessors for business desktops from Intel. In 2002, an HP executive wrote "PLEASE DO NOT ... communicate to the regions, your team members or AMD that we are constrained to 5% AMD by pursuing the Intel agreement."

We focus on Intel's marketing subsidy to Dell during the 2002-2005 period to take advantage of antitrust documentation on marketing rebate payments made to Dell. Although the data are not as extensive for other PC firms, we evaluate the TAP test for firms involved the Intel antitrust case (HP and Toshiba) and a firm that was not involved (Gateway).

3.2. Data

We use three main data sources for our analysis: PC and CPU sales are from Gartner Group, advertising data are from Kantar Media Group, and CPU price and cost data are from In-Stat. All data are available from the first quarter of 2002 through the last quarter of 2005. We discuss each in turn.

Quarterly PC and CPU sales are at the product level, where a product is defined as PC vendor (i.e., Acer), PC vendor brand (i.e., Aspire),

---

3 US District Court for the District of Delaware Complaint. 2009.

4 U.S. District of Court for District of Columbia; SEC (Securities and Exchange Commission) vs. Dell, pp. 10-11 and U.S. District of Court for District of Delaware; State of New York, by Attorney General Andrew M. Cuomo vs. Intel Corporation, p.6.

platform type (i.e., Notebook), CPU vendor (i.e., Intel), CPU family (i.e., Pentium 4), CPU speed (i.e., 1600/1799 MHz) combination. We focus on the market for US home consumers for two primary reasons. First, businesses make multiple purchases at a time, which would greatly complicate the empirical model, and, second, we don't have access to advertising data for each sector separately.

Table 1 Market Shares and Advertising Expenditures by PC Firms

PC firm	Num. obs	Market share (% shipment)	Quarterly Average Total PC-related advertising (M$)
Acer	428	0.31%	9.45*
Apple	223	4.80%	-
Averatec	37	0.42%	0.00
Dell	1020	32.44%	1.58
emachines	59	3.86%	9.45*
Fujitsu	193	0.30%	3.06
Gateway	487	11.62%	9.45*
HP	1438	29.17%	65.59
IBM/Lenovo	535	0.23%	22.81
Sony	360	2.93%	5.08
Systemax	507	0.36%	0.19
Toshiba	294	3.54%	4.13
Other	1867	10.03%	-
Total	7,448	100%	

Notes: Market share is total firm PC Shipments/total industry PC shipments. *Our measure of advertising includes emachines and Gateway together with Acer so we can't separate the three.

Advertising data consist of PC advertising expenditures.[5] The advertising data are quite detailed, sometimes even at the level of a specific product/model (e.g., Acer Aspire AS5735 Notebook Computer). However, it is difficult to match with the data from Gartner Group because the definition of products/models varies between the two datasets. Kantar Media Group uses a model name, such as Aspire AS5735; whereas, Gartner Group defines PC models as a combination of PC vendor, PC

5 PC firms advertise printers and other computer accessories. We do not include these advertising expenditures.

brand, platform type, CPU vendor, CPU family, and CPU speed. We match the two data sets based on brand and platform type. Table 1 shows the market share and total PC advertising expenditures by PC firm in the entire sample.

In-Stat provides data on CPU prices and manufacturing costs for selected processors and time periods. We need to match our PC data (where a CPU is in a PC) from Gartner group with CPU prices and manufacturing costs from In-Stat. CPU prices are available by processor core on a quarterly basis.[6] In our PC data, we know the CPU family (that is, the marketing name, e.g., Pentium 4) and speed (frequency) of the CPU. The same processor core is often used to make processors that are marketed under different family names with different sets of features enabled, and the processor core used in a processor changes over time as technology advances. For instance, processor core "Williamette" was used for processors families marketed as Pentium 4 and as Celeron for desktop computers, while in later periods the same CPU families switched to the next-generation processor core "Northwood." We match the data based on platform group (whether desktop or mobile), type (whether mainstream, value, or ultraportable), family/marketing name of a CPU, CPU speed, year, and quarter.[7] We provide the product cross-reference in Table A1[8] in the appendix.

CPU manufacturing cost estimate data are more limited in that the

---

6 CPU prices are available at several different levels of detail. The most detailed information is list prices of specific processors (e.g. Pentium M 1.40GHz). These prices are available for selected processors from July 2004 to July 2005, mostly on a monthly basis. Although it would be ideal to have list prices for all processors for all time periods, these detailed data cover only a subset of our sample.

7 This process (and a slight generalization of it described below) generates a high match. For example, among Dell PCs, we have 78% match. For the CPUs not matched at first attempt, we drop type, then we have 83% match. When unmatched, the data are matched based on family/marketing name of a CPU, CPU speed, year, and quarter, ignoring platform group. Then we obtain a 96% match. When the data are not matched, we try matching based on platform group, family/marketing name of a CPU, CPU speed, ignoring time, and then we have 99% match. For observations still not matched, we take the averages of prices and cost estimates of CPUs of the same marketing name, year and quarter.

8 The cross-reference table is constructed based on In-Stat's document and an website specialized in CPU information, www.cpu-world.com.

cost estimates are available by CPU processor core for a broader definition as of 2005. For processor core Willamette, we have cost estimates for different years, but not throughout the data period.

Intel has constructed fabs and changed the use of existing fabs, which affected cost levels over time. Also, learning-by-doing drives the cost level downward over time and so cost depends on how mature the manufacturing process is. We use two approaches to construct our CPU marginal cost variable. First, we use In-Stat cost estimates matched with PC data using the cross-reference Table A1.[9] Table 2 presents the summary statistics for the price and cost estimates of Intel CPUs in the sample. Recall that the cost estimates are not time-varying for almost all processor cores. As a second approach, instead of using the same cost estimates for all periods, we estimate marginal manufacturing costs for each processor by regressing the available cost estimates on cost shifters such as processor features, the number of fabs used for processors, size of the fabs (square feet), wafer size, and IC process.

Table 2 | CPU price and CPU marginal cost of Intel-based Dell PCs

	Num. obs	Mean	Std. Dev.	Min	Max
CPU price	1020	146.98	57.14	49	317
CPU marginal cost	1020	36.64	5.77	26	57

Table 3 shows the percentage of PCs sold by PC firm and CPU vendor as well as the overall market share that the firm contributes toward the CPU manufacturers product over the sample period. As the Table illustrates Dell, IBM/Lenovo, and Toshiba used Intel based CPUs exclusively. Dell's purchases of Intel based CPUs contributed the most (32%) of any PC firm to Intel's market share, with HP following a close second (23%). These contributions to Intel's market share are significantly higher than the next closest PC firm, which is Gateway at 9%.

9 As for reliability of the cost estimates, In-Stat document states "Equations to calculate the number of die sites per wafer, yield, and cost per good die are well known throughout the industry. Important physical parameters, such as package type and die size, are generally published by the vendor and are verifiable through destructive analysis. The key areas of uncertainty are in estimating wafer cost, defect density, testing cost, and package cost."

Table 3 Percent and Market Share of PCs by CPU Type

	Intel		AMD		Total
	% PCs	Market share	% PCs	Market share	Num. obs
Acer	89.72%	0.26%	10.28%	0.05%	428
Averatec	40.54%	0.13%	59.46%	0.30%	37
Dell	100.00%	32.44%	0.00%	0.00%	1020
emachines	57.63%	1.75%	42.37%	2.11%	59
Fujitsu	88.82%	0.28%	11.18%	0.01%	170
Gateway	93.84%	8.60%	6.16%	3.23%	487
HP	78.13%	23.03%	21.87%	6.14%	1436
IBM/Lenovo	100.00%	0.23%	0.00%	0.00%	535
Sony	90.40%	2.77%	9.60%	0.14%	354
Systemax	74.16%	0.29%	25.84%	0.07%	507
Toshiba	100.00%	3.54%	0.00%	0.00%	294
Total	88.38%		11.62%		5327

The summary statistics in Table 4 indicate, 88% of the PCs have an Intel CPU, where the average price of a PC is $1,250. Over half of the PCs are mobile as opposed to desk-based. Approximately 19% of the PCs were shipped in the first quarter of the data period, i.e., 2002:Q1 (these are denoted *Older PC*) and about 36% of CPUs where used in these PCs.[10] If the PC is not an *Older PC* then *PC age* indicates a mean of 1.6 quarters since the first shipment of the PC and 2.5 quarters since the first sales record of a CPU. The age variables (*Older PC*, *PC age*, *Older CPU*, and *CPU age*) are intended to capture the quality (how obsolete the production technology is), popularity and consumer awareness of a product (how long it has been on the market). CPU benchmark is a (continuous) quality benchmark that compares the relative speeds of different CPUs (collected by PassMark[11]). CPU

10 If there is a shipment record of a CPU in the first quarter of the data period, we can assume that the CPU has been introduced in that quarter or earlier.

11 CPU Benchmark results were gathered from users' submissions to the PassMark web site (http://www.cpubenchmark.net/cpu_list.php) as well as from internal testing. Performance Test conducts eight different tests and then averages the results to determine the CPU Mark for a system. To ensure that the full CPU power of a PC system is realized, Performance Test runs each CPU test on all available CPUs. Specifically, Performance Test runs one simultaneous CPU test for every logical CPU (Hyper-threaded); physical CPU core (dual core) or physical CPU package

Table 4 Descriptive Statistics

Variable	Mean	Min	Max
Price (1000$2000)	1.25	0.38	3.52
Intel CPU	0.88	0	1
Mobile PC	0.57	0	1
CPU benchmark	0.34	0.13	0.9
CPU speed (100mhz)	2.08	0.65	3.8
Older PC	0.19	0	1
PC age	1.6	0	10
Older CPU	0.36	0	1
CPU age	2.52	0	10
Chip ads (10 mil.$2000)	2.39	0.02	4.66
PC brand ads (10 mil.$2000)	0.44	0	6.49
Num. obs.	5327		

Notes: Price (unit: 1000$) and advertising (unit: 10 million $) variables are adjusted to 2000 dollars using Consumer Price Index (CPI) data from the U.S. Department of Labor Bureau of Labor Statistics.

manufacturers spent an average 23 million dollars on general firm promotions and chip advertising while PC firms spend an average 4.4 million dollars for PC brand advertising. For the summary statistics we present the market share weighted measure of PC brand advertising.[12]

(multiple CPU chips). So hypothetically if you have a PC that has two CPUs, each with dual cores that use hyper-threading then Performance Test will run eight simultaneous tests. Since PassMark point is not available for some models, we use a linear interpolation for those missing data, based on CPU model and CPU speed.

12 We construct a brand-platform advertising variable with a market share weight. It is the sum of brand-platform advertising, brand advertising weighted by market share of a platform within the brand, brand combination advertising weighted by market share of brand-platform within the brand combination, platform advertising weighted by market share of brand within the platform, and firm advertising weighted by market share of brand-platform within the firm. Market shares are computed based on shipments. For model j of brand b, platform p, and firm f, we have:

$$ad_i^{\text{brand}} = a_{b,p}^{\text{brand,platform}} + \frac{\sum_{i\in\mathcal{J}_{fpb}} s_i}{\sum_{i\in\mathcal{J}_{fp}} s_i} a_{b,f}^{\text{brand}} + \frac{\sum_{i\in\mathcal{J}_{fpb}} s_i}{\sum_{b'\in\mathcal{J}_{fpc}} \sum_{i\in\mathcal{J}_{fpb'}} s_i} a_{c,p}^{\text{brand comb}} + \frac{\sum_{i\in\mathcal{J}_{fpb}} s_i}{\sum_{i\in\mathcal{J}_{fp}} s_i} a_{f,p}^{\text{platform}} \frac{\sum_{i\in\mathcal{J}_{fb}} s_i}{\sum_{i\in\mathcal{J}_{f}} s_i} a_f^{\text{firm}}$$

where $\mathcal{J}_*$ denotes the set of models in category *, $a_{b,p}^{\text{brand,platform}}$ is brand-platform advertising, $a_{b,f}^{\text{brand}}$ is brand advertising; $a_{c,f}^{\text{brand comb}}$ is brand combination advertisin, $a_{f,p}^{\text{platform}}$ is firm-platform advertising, and a_f^{firm} is firm advertising. This

Table 5 Descriptive Statistics by CPU Manufacturer (feb 19)

	Mean Intel	Mean AMD	Difference	*t*-value
Price (1000$2000)	1.28	0.98	0.30	17.10
Mobile	0.58	0.52	0.06	2.63
CPU benchmark	0.33	0.46	-0.14	-26.01
CPU speed (1000mhz)	2.09	1.93	0.16	5.20
Older PC	0.19	0.21	-0.02	-1.34
PC age	1.65	1.25	0.40	4.74
Older CPU	0.37	0.33	0.04	1.87
CPU age	2.60	1.93	0.67	5.64
Chip ads (10 mil.$2000)	2.65	0.36	2.29	45.79
PC brand ads (10 mil.$2000)	0.41	0.69	-0.28	-8.35
Num. obs.	4708	619		

Table 5 presents descriptive statistics for the two main CPU manufacturers: Intel and AMD. As the table indicates, over the data period, average AMD chips used in PCs performed significantly better than average Intel chips (see *CPU benchmark*). On average, we can see that the PC models with Intel chips are priced higher and advertised significantly more at the CPU level. In many cases, the same PC brand has models with AMD chips and models with Intel chips. Thus PC brand ads do not accurately capture the difference between PCs with Intel chips versus those with AMD chips. Although PC brand ads are larger for the brands of PCs with AMD chip on average, it is because some heavily advertised brands have models with AMD and Intel chips.

Finally, we use surveys on PC purchases from Forrester Research from 2002 through 2005. These data have information about individual consumers' PC and CPU choices, although they are not detailed at the product level. For example, we observe whether a survey respondent bought a PC in the last year, some characteristics of the PC such as PC firm and CPU manufacturer (Intel, AMD, Apple[13], Other, or Don't know) if purchased.

is a similar methodology that is used in the literature for quantity weighted average prices (see for example, Song (2007).

13 During the data time period, Apple PCs used only IBM chips.

4. Demand for CPUs

Following Berry, Levinsohn, and Pakes (1995) (BLP) and Sovinsky Goeree (2008) we model the demand for PCs as a random-coefficient logit. The demand for CPUs can be inferred from the demand for a PC model as a PC comes equipped with a single CPU. When consumers purchase computers, they choose a combination of PC firm and CPU type.[14] There are T markets, indexed by $t = 1, 2, \cdots, T$, each with I_t consumers, indexed by i. A home market consumer chooses from J products, indexed $j = 1, \cdots, J$, where a product is a PC vendor (i.e., Acer), PC vendor brand (i.e., Aspire), platform type (i.e., Notebook), CPU vendor (i.e., Intel), CPU family (i.e., Pentium 4), CPU speed (i.e., 1600/1799 MHz) quarter combination.

Product j characteristics are price of the PC (p) and non-price observed attributes of the PC (x), which include the platform and PC vendor dummy variables, dummy variables for whether the processor is manufactured by Intel, processor speed, the age of the CPU and the CPU benchmark score.

Advertising is an additional characteristic that may impact consumer demand.[15] Given the difference in advertising campaigns across firms and CPU suppliers, we allow advertising by PC firms (a_{jt}^{pc}) to have different effect on consumer utility than advertising done by CPU vendors (a_{jt}^{cpu}). Finally, attributes unobserved to the researcher but known to consumers and producers (ξ) may influence utility. The indirect utility consumer i obtains from j at time t is

---

14 Many websites which provide CPU performance comparisons, such as CPU ScoreCard.com, categorize CPUs depending on computer type (i.e., desktops or laptops), and they compare CPUs within the same computer type. Considering that CPU chips intended to be used in desktops and laptops are different due to different requirements, we can think that the choice of CPU type includes the choice of computer type. Song (2007) and Salgado (2008b) modeled consumers' choice of CPU type although consumers more often buy computers, rather than CPU chips without computers.

15 It is reasonable to conjecture that all consumers know of the existence of Intel processors, therefore we do not model Intel advertising as impacting the consumer's choice set, but rather as impacting utility directly. We assume that consumers know all firms and processor types when making a purchase decision.

$$u_{ijt} = \delta_{jt} + \mu_{ijt} + \epsilon_{ijt}, \quad (1)$$

where

$$\delta_{jt} = x_{jt}'\beta + a_{jt}^{\text{pc}}\gamma + a_{jt}^{\text{cpu}}\lambda + \xi_{jt}$$

captures the base utility every consumer derives from product j and mean preferences for x_j are captured by β. The composite random shock, $\mu_{ijt} + \epsilon_{ijt}$,[16] captures heterogeneity in consumers' tastes for product attributes, and ϵ_{ijt} is a mean zero stochastic term distributed i.i.d. type I extreme value across products and consumers.

We observe various levels of aggregation of advertising expenditures by PC firms. These include firm-specific advertising (i.e., advertising for Dell), firm-brand specific advertising (i.e., Dell Presario), firm-platform advertising (i.e., Dell Notebooks), and firm-brand-platform advertising (i.e., Dell Presario Notebooks). CPU vendor advertising is at the firm level (i.e., Intel) and at the CPU level. We allow each type of PC firm advertising and CPU advertising to have a different impact on utility, as captured by the vectors γ and λ.[17]

The a_{jt}^{pc} is a vector of PC advertising variables aggregated at different levels and a_{jt}^{cpu} is a vector of CPU advertising variables aggregated at different levels. For sake of exposition, we define $a_{jt} = \{a_{jt}^{pc\prime} a_{jt}^{\text{cpu}'}\}$.

The μ_{ijt} term includes the interactions between observed consumer attributes (D_{it}), unobserved (to the econometrician) random consumer tastes (v_i), and observed product attributes (x_j), and the interactions between observed consumer attributes ($\widetilde{D}_{it}$) and advertising variables, where $\widetilde{D}_{it}$ is a subset of D_{it}. Specifically,

---

16 Choices of an individual are invariant to multiplication of utility by a person-specific constant, so we fix the standard deviation of the ϵ_{ijt}.

17 We also consider a specification in which we allow for nonlinear effects of advertising.

$$\mu_{ijt} = \alpha \ln(y_{it} - p_{jt}) + x_{jt}^{'}(\Omega D_{it} + \Sigma v_i) + a_{jt}' \Upsilon \widetilde{D}_{it}$$

$$v_i \sim N(0, I_k). \qquad (2)$$

where y_{it} is income of individual *i* (in market *t*). The Ω matrix measures how tastes vary with x_j. We assume that v_i are independently and normally distributed with a variance to be estimated. Σ is a scaling matrix. ϒ matrix captures how advertising's impact varies by observed consumer characteristics.

Consumers have an "outside" option, which includes purchase of a computer with non-Intel or non-AMD processor (such as IBM chips exclusively used by Apple computers during the sample period)[18], a PC manufactured by a small firms[19], and self-assembled PCs. Normalizing p_{0t} to zero,[20] the indirect utility from the outside option is

$$u_{i0t} = \alpha \ln(y_{it}) + \xi_{0t} + \epsilon_{i0t.}$$

We normalize ξ_{0t} to zero, because we cannot identify relative utility levels. The (conditional) probability that consumer *i* purchases product *j* is

$$s_{ijt} = \frac{\exp\{\delta_{jt} + \mu_{ijt}\}}{{y_{it}}^{\alpha} + \sum_r \exp\{\delta_{rt} + \mu_{irt}\}}. \qquad (3)$$

The ${y_{it}}^{\alpha}$ term in the denominator is from the presence of the outside good. Let $\zeta_i = (y_{it}, D_{it}, v_i)$ be the vector of individual characteristics. We assume that the consumer purchases at most one good per period,[21]

---

18 Apple Computer is included in the outside option for CPU's because they used IBM processors during the sample period (2002:Q2 -2005:Q4).

19 These include Everex, Medion, Micro Electronics, Motion Computing, MPC, NEC, Sharp, and Velocity Micro.

20 The effect of changes over time in prices of the outside option is captured by the relative attractiveness of goods to the outside option.

21 This assumption may be unwarranted for some products for which multiple purchase is common. However it is not unreasonable to restrict a consumer to purchase one computer per quarter. Hendel (1999) examines purchases of PCs by businesses and presents a multiple-choice model of PC purchases.

that which provides the highest utility, U. Let $R_j \equiv \{\zeta: U(\zeta, p_j, x_j, a_j \xi_j, \epsilon_{ij}) \geq U(\zeta, p_r, x_r, a_r, \xi_r, \epsilon_{ir}) \forall r \neq j\}$ define the set of variables that results in the purchase of j given the parameters of the model. The home market share of product j is

$$s_{jt} = \int_{R_j} dG(y, D, v, \epsilon) = \int_{R_j} s_{ijt} dG_{y,D}(y, D) dG_v(v) \quad (4)$$

where $G(\cdot)$ denotes the respective distribution functions. The second equality follows from independence assumptions. The conditional probability that i purchases j, s_{ij}, is given in (3).

Note that this implies that the market share for firm f of processor type c is given by

$$\begin{aligned} s_{fct} &= \sum_{j \in \mathcal{J}_c \cap \mathcal{J}_f} \int_{R_j} \frac{\exp\{\delta_{jt} + \mu_{ijt}\}}{{y_{it}}^{\alpha} + \sum_r \exp\{\delta_{rt} + \mu_{irt}\}} dG_{y,D}(y, D) dG_v(v) \quad (5) \\ &= \sum_{j \in \mathcal{J}_c \cap \mathcal{J}_f} s_{it}(p, a) \end{aligned}$$

where $\mathcal{J}_f$ are the set of products produced by firm f and $\mathcal{J}_c$ are the set of products with a CPU of type c. Note that processor market share is a function of PC prices and advertising of all PC products. Demand of firm f for CPU processor c at time t is $\mathcal{M}_t s_{fct}$, where $\mathcal{M}_t$ is the market size given by the total number of PCs sold each quarter. The observed market share of a processor is given by the number of units sold of that processor divided by the total number of processors sold. The total number of observations is 5,327.

4.1. Estimation Technique

We implement the econometric technique found in many studies of differentiated products, such as BLP (1995, 1998) and Nevo (2000). The parameters are $\theta = \{\alpha, \beta, \gamma, \lambda, \Sigma, \Omega, \Upsilon\}$. Under the assumption that the observed data are the equilibrium outcomes, we estimate the parameters simultaneously using generalized method of moments (GMM).

Following BLP, we restrict the model predictions for j's market

share to match observed shares. We solve for $\delta(S,\theta)$ that is the implicit solution to

$$S_t^{obs} - \mathrm{s}_t(\delta,\theta) = 0 \tag{6}$$

where S_t^{obs} and s_t are vectors of observed and predicted shares respectively. We substitute $\delta(S,\theta)$ for δ when calculating the moments.[22] The first moment unobservable is

$$\xi_{jt} = \delta_{jt}(s,\theta) - x_j'\beta. \tag{7}$$

We use the Forrester data to construct CPU manufacturer choice micro moments. Petrin (2002) shows how to combine macro data with data that links average consumer attributes to product attributes to obtain more precise estimates. We augment market share data with data relating consumers to product characteristics as in Sovinsky Goeree (2008). The micro data we have connect consumers to processor manufacturer. We combine the processor firm choice data with product level data to obtain more precise estimates of the parameters of the taste distribution (Ω) and the parameters of advertising effectiveness (Υ). The demographic characteristics for these moments (denoted D_{it}) are given by the Forester data, which are linked directly to purchases and advertising exposure.

Let B_i be a $R \times 1$ vector of processor manufacturer choices for individual i. Let b_i be a realization of B_i where b_{ir} =1 if a CPU produced by $r = \{\text{Intel}, \text{AMD}\}$ was chosen.

Define the residual as the difference between the vector of observed choices and the model prediction given (δ, θ):

$$\mathcal{B}_i(\delta,\theta) = b_i - E_v E[B_i|D_{it},\delta,\theta]. \tag{8}$$

For example, the element of $E_v E[B_i|D_{it},\delta,\theta]$ corresponding to Intel

22 As discussed in Dube et al (2011) we use a precise tolerance level for the contraction mapping. For more details see section 7.2.

for consumer i is

$$\sum_{c\in\mathcal{J}_{intel}}\sum_{j\in\mathcal{J}_c}\int_{R_j}\frac{\exp\{\delta_{jt}+\mu_{ijt}\}}{y_{it}^{\alpha}+\sum_r\exp\{\delta_{rt}+\mu_{irt}\}}dG_v(v),$$

where the summand is over products sold by Intel, the integral is over the assumed distribution of v. The population restriction for the micro moment is $E[\mathcal{B}_i(\delta,\theta)|(x,\xi)]=0$. Let $\mathcal{B}(\delta,\theta)$ be the vector formed by stacking the residuals $\mathcal{B}_i(\delta,\theta)$ over individuals.

We use GMM to find the parameter values that minimize the objective function, $\Lambda' Z A^{-1} Z'\Lambda$, where A is an appropriate weighting matrix which is a consistent estimate of $E[Z'\Lambda\Lambda' Z]$ and Z are instruments orthogonal to the composite error term Λ. Specifically, if Z_ξ, $Z_\mathcal{B}$ are the respective instruments for each disturbance/residual, the sample moments are

$$Z'\Lambda=\begin{bmatrix}\frac{1}{J}\sum_{j=1}^{J}Z_{\xi,j},\xi_j(\delta,\beta)\\ \frac{1}{N}\sum_{i=1}^{N}Z_{\mathcal{B},i},\mathcal{B}_i(\delta,\theta)\end{bmatrix}$$

where $Z_{\xi,j}$ is column j of Z_ξ. Joint estimation takes into account the cross-equation restrictions on the parameters that affect both the macro and micro moments, which yields more efficient estimates.

The market shares in (4) must be simulated. As in BLP, the distribution of consumer demographics is an empirical one. As a result there is no analytical solution for predicted shares, making simulation necessary. The simulator for the market share is the average over individuals of choice probabilities. An outline of the technique follows. We sample a set of "individuals" where each consists of $(v_{i1},\cdots,v_{ik})$ taste parameters drawn from a multivariate normal; demographic characteristics, $(y_i,D_{i1},\cdots,D_{id})$, drawn from the Forrester data in the case of the macro moments and data in the case of the micro moments. To construct the market share constraints we draw J uniform random variables for each individual. For a given value of the parameters, we

compute the probability she would buy each product. The market share simulator is the average over individuals of the choice probabilities. The process is similar for the micro moment constraints, but we take R draws for each product-individual pair. The simulator for individual product choice probabilities is the average over the R draws. Individual firm choice probabilities are the sum over the products offered by each firm.

Using the results of Pakes and Pollard (1989), this estimator is consistent and asymptotically normal. As the number of pseudo random draws used in simulation $R \to \infty$ the method of simulated moments covariance matrix approaches the method of moments covariance matrix. To reduce the variance due to simulation, we employ antithetic acceleration (for an overview of simulation techniques see Stern, 1997 and 2000). Geweke (1988) shows if antithetic acceleration is implemented during simulation, then the loss in precision is of order $1/N$ (where N are the number of observations), which requires no adjustment to the asymptotic covariance matrix. The reported (asymptotic) standard errors are derived from the inverse of the simulated information matrix which allows for possible heteroskedasticity.[23]

4.2. Identification

Following the literature, we assume that the demand unobservables (evaluated at the true value of the parameters, Θ_0) are mean independent of a vector of observable product characteristics, (x):

$$E\left[\xi_{jt}(\Theta_0)x_{jt}\right] = 0. \tag{9}$$

We do not observe ξ_{jt}, but market participants do. This may lead to endogeneity of price and advertising. For example, some products may have higher quality, which is unobserved by researchers, and PC firms may set higher prices and/or determine their ad expenditures based on quality. Also, a CPU manufacturer may advertise more when the PCs

23 The reported standard errors do not included additional variance due to simulation error.

with its CPU are of higher quality. To account for the potential endogeneity of price and advertising, we employ instrumental variables.

BLP show that variables that shift markups are valid instruments for price in differentiated products models and Sovinsky Goeree (2008) shows that variables that shift markups are valid instruments for advertising.[24]

Given (9) and regularity conditions, the optimal instrument for any disturbance-parameter pair is the expected value of the derivative of the disturbance with respect to the parameter (evaluated at Θ_0) (Chamberlain, 1987). Optimal instruments are functions of advertising and prices. We form approximations to the optimal instruments as in Sovinsky Goeree (2008)[25], by evaluating the derivatives at the expected value of the unobservables ($\xi = \omega = 0$). The instruments will be biased since the derivatives evaluated at the expected values are not the expected value of the derivatives. However, the approximations are functions of exogenous data and are constructed such that they are highly correlated with the relevant functions of prices and advertising. Hence the exogenous instruments will be consistent estimates of the optimal instruments.[26]

One set of instruments we use is chosen to reflect competitive pressure. Competitive pressure for a model is likely to affect price and advertising choices (via their first-order conditions) but does not impact consumer utility from purchasing the model. We also use the characteristics of other products of the same PC firm and those of other PC firms as instruments for price and advertising. In particular, we use the sum of the values of the characteristics - mobile platform, performance benchmark (PassMark point), CPU speed - of other products offered by the same PC

---

24 Products which face more competition (due to many rivals offering similar products) will tend to have lower markups relative to more differentiated products. Advertising for *j* depends on *j*'s markup. Pricing FOCs show the optimal price (and hence markup) for *j* depends upon characteristics of all of the products offered. Therefore, the optimal price and advertising depends upon the characteristics, prices, and advertising of all products offered. Thus optimal instruments will be functions of attributes and cost shifters of all other products.

25 This was introduced in BLP(1999). See Reynaert and Verboven (2012) for a discussion of the benefits of using optimal instruments.

26 One could also use a series approximation as in BLP to construct exogenous instruments.

firm and the sum of the values of the characteristics of all the products offered by other PC firms. In addition, we construct variables that capture technological change. Technological change affects production cost and hence would be related to the level of price and advertising but unrelated to consumer utility. In particular, we use a series of interactions between CPU cohort[27] dummy variables and time (year - quarter) dummy variables and a series of interactions between PC cohort[28] dummy variables and time (year-quarter) dummy variables. These variables are intended to proxy the change in production cost over time (e.g., declining production cost due to learning by doing).

Using optimal instruments for micro moments.

An informal identification argument follows. Associated with each PC is a mean utility, which is chosen to match observed and predicted market shares. All variation in sales would be driven by variation in product attributes if consumers were identical. Variation in product market shares corresponding to variation in the observable attributes of those products (such as CPU speed) is used to identify the parameters of mean utility (β).

A PC may have attributes that provide more utility to certain types of consumers. For instance, if young male adults prefer to use their PC to play games, then young male consumers may place a higher valuation on CPU speed relative to other cohorts. Identification of the taste distribution parameters (Σ, Ω) relies on information on how consumers substitute (see equation (2)). There are two issues that merit attention. First, new product introductions are common in the PC industry. Variation of this sort is helpful for identification of Σ. The distribution of unobserved tastes, v_i, is fixed over time, but the choice set of available products is changing over time. Variation in sales patterns over time as the choice sets change allows for identification of Σ. Second, we augment the market level data with micro data on CPU manufacturer choice. The extra information in the micro data allows

27 CPU cohort is the set of CPU models which have the first sales record in the same quarter.

28 PC cohort is the set of PC models which have the first sales record in the same quarter.

variation in choices to mirror variation in tastes for product attributes. Correlation between $x_j D_i$ and choices identifies the Ω parameters. The individual-level data contain useful information on ad exposure across households. Variation in advertising exposure and variation in advertising expenditures corresponding to variation in observable consumer characteristics ($\widetilde{D}_i$) identifies Υ.

4.3. Preliminary Demand Estimates

First, we present results from a series of probit regressions of the probability of purchasing a PC in 2002 for use at home as given in the Forrester data. The estimates are presented in Table 6 and illustrate the importance of observed product heterogeneity in PC purchases. The results indicate that the form factor of the PC is important as well as firm fixed effects. Finally, the opportunity to add-on certain items (such as a flat-screen monitor) increases the probability of purchase. As columns (4)-(6) illustrate, consumers value PCs with a Intel or AMD processor even after controlling for other product characteristics. Furthermore, the results suggest the valuation of processor type is not constant over time (Column 6).

Table 7 presents probit estimates of purchase probabilities that illustrate the importance of product and individual observed heterogeneity for PC purchases. The independent variable is again whether the individual bought a PC in the past year. We start by including demographic variables (Column 1). We can see all the estimates on the coefficients of demographic variables are statistically significant. Once we include a laptop dummy and CPU type (Intel dummy and AMD dummy), some demographic variables become less significant (Column 2). Not surprisingly, the results indicate individuals are more likely to purchase a PC with an Intel or AMD processor than non-branded processors. Whether the computer is a laptop is also a significant factor in the purchase decision. After controlling for PC add-ons (Column 3) and Operating System dummy variables (Column 4), preference for a laptop and a PC with Intel or AMD processor have less impact on PC purchase probability. Overall, the results suggest that PC characteristics

Table 6 Probit Regressions of PC Purchase (in 2002)

Variables		Dependent Variable: Whether Bought a PC in the past year (1)	(2)	(3)	(4)	(5)	(6)
Laptop		0.535***	0.473***	0.395***	0.527***	0.473***	0.474***
		(0.010)	(0.011)	(0.011)	(0.010)	(0.011)	(0.011)
Processor Manufacturer	Intel				0.139***	0.021***	0.071***
					(0.008)	(0.008)	(0.015)
	Intel and 2003						-0.056**
							(0.022)
	Intel and 2004						-0.085***
							(0.022)
	Intel and 2005						-0.061***
							(0.020)
	AMD				0.511***	0.344***	0.484***
					(0.016)	(0.016)	(0.028)
	AMD and 2003						-0.232***
							(0.044)
	AMD and 2004						-0.139***
							(0.045)
	AMD and 2005						-0.230***
							(0.041)
PC Manufacturer	Acer	-0.416***	-0.367***	-0.202***	-0.396***	-0.346***	-0.346***
		(0.042)	(0.043)	(0.045)	(0.042)	(0.043)	(0.043)
	Apple	-0.035*	-0.107***	-0.363***	0.093***	-0.055***	-0.050**
		(0.019)	(0.019)	(0.022)	(0.019)	(0.020)	(0.020)
	AST	-1.052***	-1.018***	-0.868***	-1.024***	-0.988***	-0.988***
		(0.166)	(0.167)	(0.174)	(0.167)	(0.168)	(0.168)
	Compaq	-0.056***	-0.089***	-0.125***	-0.045***	-0.078***	-0.078***
		(0.012)	(0.012)	(0.013)	(0.012)	(0.013)	(0.013)
	Dell	0.457***	0.350***	0.131***	0.486***	0.387***	0.386***
		(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
	Emachines	0.348***	0.311***	0.128***	0.371***	0.332***	0.334***
		(0.017)	(0.018)	(0.019)	(0.017)	(0.018)	(0.018)
	Fujitsu	-0.526**	-0.623***	-0.625**	-0.491**	-0.589**	-0.593**
		(0.231)	(0.238)	(0.255)	(0.231)	(0.238)	(0.238)
	Gateway	-0.163***	-0.253***	-0.317***	-0.135***	-0.222***	-0.223***
		(0.013)	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)
	HP	0.089***	0.021*	-0.093***	0.112***	0.046***	0.046***
		(0.011)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
	IBM	-0.279***	-0.259***	-0.169***	-0.247***	-0.232***	-0.232***
		(0.019)	(0.020)	(0.021)	(0.020)	(0.020)	(0.020)
	Sony	0.358***	0.096***	-0.114***	0.366***	0.120***	0.119***
		(0.025)	(0.026)	(0.027)	(0.025)	(0.026)	(0.026)
	Toshiba	0.195***	0.149***	0.055*	0.215***	0.173***	0.172***
		(0.027)	(0.028)	(0.030)	(0.027)	(0.028)	(0.028)

Table 6 (Continue)

Variables		Dependent Variable: Whether Bought a PC in the past year (1)	(2)	(3)	(4)	(5)	(6)
PC Add-ons	Broadband		0.229***	0.157***		0.223***	0.220***
	Adapter		(0.012)	(0.012)		(0.012)	(0.012)
	DVD		0.447***	0.266***		0.435***	0.436***
			(0.008)	(0.008)		(0.008)	(0.008)
	CD Rom		0.044***	0.015		0.029***	0.029***
			(0.010)	(0.011)		(0.010)	(0.010)
	Flat Panel		0.581***	0.421***		0.580***	0.582***
	Monitor		(0.010)	(0.011)		(0.010)	(0.010)
	Webcam		-0.128***	-0.184***		-0.140***	-0.136***
			(0.014)	(0.015)		(0.014)	(0.014)
Operating System	Windows 95 or 98			-0.918*** 0.014)			
	Windows ME			-0.195*** 0.016)			
	Windows 2000			-0.236*** 0.016)			
	Windows NT			-0.064* 0.035)			
	Windows XP			0.442*** 0.014)			
Observations		167221	167221	167221	166246	166246	166246

Notes: Standard errors are in parenthesis. *** indicates significant at 1%; ** at 5% and * at 10%. All regressions include time dummies.

Table 7 PC purchase Probability on PC Characteristics and Demographics

Dependent Variable: Whether Bought a PC in the past year				
Variables	(1)	(2)	(3)	(4)
Laptop		0.518***	0.477***	0.409***
		(0.011)	(0.011)	(0.011)
Intel Processor		0.118***	0.024***	-0.018**
		(0.008)	(0.008)	(0.009)
AMD Processor		0.491***	0.351***	0.225***
		(0.016)	(0.017)	(0.018)
Demographics				
Age	-0.005***	-0.004***	-0.002***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.046***	0.014*	0.006	0.016**
	(0.007)	(0.007)	(0.007)	(0.008)
White	0.031**	0.028**	0.031**	0.003
	(0.012)	(0.013)	(0.013)	(0.014)
Married	0.037***	0.052***	0.039***	0.025**
	(0.009)	(0.009)	(0.009)	(0.010)

Table 7 (Continue)

Dependent Variable: Whether Bought a PC in the past year				
Variables	(1)	(2)	(3)	(4)
Presence of Teenagers	0.145***	0.156***	0.134***	0.113***
	(0.009)	(0.009)	(0.010)	(0.010)
Income > $100,000	0.176***	0.100***	0.047***	0.024**
	(0.009)	(0.010)	(0.010)	(0.010)
Income < $25,000	-0.083***	-0.003	0.046***	0.073***
	(0.008)	(0.008)	(0.008)	(0.009)
Fixed Effects				
Firm		Included	Included	Included
Year	Included	Included	Included	Included
PC-Add ons			Included	Included
Operating System				Included
Observations	163,187	160,982	160,982	160,982

Notes: Standard errors are in parenthesis. *** indicates significant at 1%; ** at 5% and * at 10%.

Table 8 CPU Purchase Estimates

Dependent Variable: CPU in a PC purchased in the past year (Base: Non-Intel/AMD/Mac or unknown)			
Variables	Intel	AMD	Apple CPU
Age	-0.018***	-0.030***	-0.001
	(0.001)	(0.002)	(0.004)
Male	0.702***	1.141***	0.543***
	(0.025)	(0.046)	(0.123)
White	0.166***	0.161**	0.344
	(0.043)	(0.082)	(0.224)
Married	-0.139***	-0.237***	-0.404**
	(0.032)	(0.058)	(0.158)
Presence of Teenagers	-0.110***	-0.041	-0.445***
	(0.031)	(0.055)	(0.162)
Income > $100,000	0.371***	0.353***	0.297*
	(0.035)	(0.061)	(0.156)
Income < $25,000	-0.649***	-0.705***	-0.422***
	(0.028)	(0.052)	(0.146)
Observations	39,536		

Notes: Laptop dummy, Firm dummies, and year dummies included
SE in parenthesis. *** significant at 1%, ** at 5%, * at 10%

affect the purchase decision after controlling for demographics and that certain demographics are more important than others in the purchase decision.

Table 8 presents multinomial logit estimates of the choice of CPU manufacturer. The results show that there is significant consumer observed heterogeneity with respect to CPU choices. Given that a consumer purchased a PC in the past year, brand CPUs are preferred by a male, single, or higher-income household to non-brand CPUs. Among brand CPUs, the choice probability of a particular CPU varies with demographics. The presence of teenagers in the household makes purchase of an Intel or Apple less likely, but doesn't significantly impact the purchase of an AMD processor. Being young and being white are significantly and positively associated with the choice of Intel or AMD processors.

4.4. Structural Demand Estimates

In Table 9 we present results from a Multinomial Logit PC-CPU demand model with aggregate market data, where only product characteristics are included as explanatory variables.[29]

Advertising variables are aggregated at the brand-platform level. How we allocate advertising expenditure to PC brand-platform level is detailed in data section 4. We also include the squared value of the sum of general promotions and brand combination advertising that can be attributed to a given brand to allow for decreasing or increasing return to advertising. In every model specification, PC-firm-specific dummy variables and time-specific dummy variables are included.

The first specification (Column 1) illustrates the importance of observed product characteristics for PC/CPU choice. The results imply that processor speed and performance (benchmark) have a significant and positive effect on utility. This is reasonable as it suggests consumers prefer faster and better-performing CPU chips. However, the age of the

29 Estimates from a nested logit model (with laptop and desk-based PC nests) indicate that purchases do not take place in this nested structure in the sense that the estimated coefficient on the substitutability between the nests is often larger than one.

Table 9 Multinomial Logit Estimates of PC/CPU Demand (feb 19)

Variables		Dependent Variable: ln(Market Share) ln(Share of Outside Goods) (1)	(2)	(3)	(4)	(5)	(6)
Price		-0.286***	-1.047*	-0.515***	-0.337***	-1.765***	-2.228***
		(0.0855)	(0.609)	(0.0671)	(0.0647)	(0.541)	(0.643)
PC	Laptop	-0.178**	0.146			0.410*	0.497*
Characteristics		(0.0711)	(0.266)			(0.236)	(0.292)
	Older PC	0.906***	0.858***			0.655***	0.582***
		(0.122)	(0.128)			(0.114)	(0.145)
	PC Age	0.126***	0.138***			0.0940***	0.0904***
		(0.0222)	(0.0241)			(0.0215)	(0.0238)
CPU	Intel	-0.510***	-0.233	-0.0104	-0.000200	0.502	-2.060
Characteristics		(0.103)	(0.244)	(0.231)	(0.221)	(0.319)	(1.308)
	ln(Benchmark)	0.371**	0.851**	0.575***	0.415***	1.315***	1.362***
		(0.155)	(0.414)	(0.128)	(0.120)	(0.371)	(0.434)
	ln(Speed)	1.594***	1.731***	1.381***	1.108***	1.625***	1.724***
		(0.106)	(0.149)	(0.0883)	(0.0847)	(0.135)	(0.166)
	Older CPU	-0.0906	-0.0641	0.487***	0.494***	0.237**	0.192
		(0.113)	(0.116)	(0.0922)	(0.0832)	(0.107)	(0.129)
	CPU Age	0.00524	-0.0101	0.0994***	0.0960***	0.0360*	0.0226
		(0.0174)	(0.0206)	(0.0135)	(0.0131)	(0.0186)	(0.0229)
Advertising	CPU ad			-0.0259	0.0945	-0.139	4.186***
				(0.187)	(0.174)	(0.193)	(1.360)
	CPU ad squared			-0.00803	-0.0370	0.0167	-0.868***
				(0.0306)	(0.0278)	(0.0321)	(0.253)
	PC ad			2.352***		2.363***	3.496***
				(0.0830)		(0.0862)	(0.302)
	PC ad squared			-0.294***		-0.304***	-0.581***
				(0.0216)		(0.0223)	(0.133)
Included Controls	IV for Price		Included			Included	Included
	IV for Advertising						Included
	PC Advertising				Included		
	PC Advertising Squared				Included		
	PC Firm Fixed Effects	Included	Included	Included	Included	Included	Included
Observations		5,327	5,327	5,327	5,327	5,327	5,327
R-squared		0.498	0.491	0.597	0.649	0.575	0.470
Median (Own) Price Elasticity		-0.336	-1.231	-0.605	-0.396	-2.075	-2.619

Notes: Standard errors are in parenthesis. *** indicates significant at 1%; ** at 5% and * at 10%. All regressions include time dummies.

CPU does not impact consumers valuations of the PC, while the age of the PC does. A likely explanation is that age captures the popularity or consumer awareness of a PC model or a CPU. If a product has been on the market for a while, it may imply either that consumers like the product, due to attractive qualities unobserved by the researcher, or that the product is well-known to potential buyers. Surprisingly, consumers place a lower valuation on the PC if it is a laptop, but this finding is not robust across specifications. Finally, conditional on CPU speed and the performance benchmark, consumers place a lower value on CPUs manufactured by Intel.

As we discussed earlier, prices may be correlated with the structural error term and hence endogenous. In Column (2) we include instruments for price. The impact of price on demand becomes much more negative, which is consistent with price being a proxy for higher quality. The rest of the estimates do not change with the exception of the valuation of Intel and the valuation of laptop, which are no longer significant.

In specification (3), PC characteristics are taken out and advertising variables are added. As compared with the price coefficient estimate from specification (1), the price coefficient is estimated to be smaller (i.e. more negative). This suggests that advertising variables may be correlated with unobserved product attributes and we need to correct for the possible correlation between advertising and unobserved high quality. We also find that consumers marginal valuation of PC advertising is positive, while the valuation of firm level CPU advertising is negative and statistically insignificant.

In specification (4) we allow the effect of advertising to be different across PC firms. The results suggest some firms are more effective at advertising their products than others, but otherwise the coefficient estimates do not change much. Specifications (5) and (6) include all explanatory variables (PC characteristics, CPU characteristics, and advertising variables), but specification (5) instruments only for price while specification (6) instruments for both price and advertising. We have more elastic demand when using specification (6). Again, this suggests that advertising is correlated with unobserved product attributes.

The parameter on the Intel dummy variable is negative or insignificant in all specifications after other PC/CPU characteristics and CPU advertising amount are taken into account. This result is consistent with the preliminary regression result in Table 6 that shows that the purchase probability is actually higher when a PC is powered by AMD chip than by an Intel chip.

Advertising at the CPU level is not valued by consumers in most specifications but has a positive and significant impact in specification (6) after instrumenting for CPU ads. Advertising by a PC firm at the brand level is always positive and significant. In particular, when evaluated at the average value, a one unit increase in PC advertising at the brand level leads to a 0.99 increase in the utility level (specification (5)). When advertising variables are instrumented, there is a larger positive impact of advertising on demand at the PC level: when evaluated at the average value, one unit increase in PC advertising at brand level leads to 1.43 increase in the utility level (specification (6)). Note that a more positive advertising effect after instrumenting for advertising suggests that advertising may be negatively correlated with unobserved product qualities. That is, firms may engage in more advertising when they have lower-quality products. This result appears reasonable given that the low-end processor of Intel, Celeron was heavily advertised during the period. The positive impact of PC advertising implies that CPU firms can enhance CPU sales by inducing PC firms to advertise PCs that contain their chip. This provides us with the pro-competitive justifications for Intel's marketing campaign: promoting CPU sales by subsidizing PC advertising. We will test whether the benefit of promoting CPU sales by subsidizing PC advertising exceeds the cost of the marketing subsidy. If not, the marketing campaign may be driven by anticompetitive motives.

5. Marginal Marketing Revenue

Intel's marketing campaign provided support to PC firms that advertised PCs with Intel chips. One of the benefits of the TAP approach is it allows us to circumvent modeling Intel's profit maximization problem. This is beneficial both because the test (and model) would become very complicated and the data necessary to estimate such a model do not exist, in particular firm-specific rebate rates are not publicly available. Suppressing time notation, the total marketing/ad revenue (TMR) of Intel from PC sales is

$$TMR = \sum_{c \in \mathcal{J}_{intel}} \sum_{f} (p_c^{CPU} - mc_c)\, \mathcal{M} s_{fc}(p, a), \tag{10}$$

where $\mathcal{J}_{intel}$ is the set of products with an Intel CPU; p_c^{CPU} is the price of CPU c; mc_c is the marginal production cost of CPU c and $s_{fc}(p, a)$ is market share of processor c sold by firm f given in equation (5), which depends on the product price (i.e., PC price, p) and advertising (a).[30]

One issue to address in the PC advertising data is that advertising may involve more than one product. For example, PC firms often engage in general promotions both by platform type and across all platforms (e.g., Acer Laptop Computer; Acer Various Computers) or PC brands may be jointly advertised (e.g., Acer Veriton & Travelmate Computers Combo). We will require a composite measure of advertising expenditures by product that includes all advertising done for that product (so it should include all group advertising). Following Sovinsky Goeree (2008) we compute product advertising expenditures as a weighted average of group advertising for that product where the weights are a function of the number of products in that group. Specifically, suppressing the time subscript, let $\mathcal{G}_j$ be the set of all product groups that include product j with group $\mathcal{H} \in \mathcal{G}_j$. Then

30 The CPU price is the listed price. In practice, many PC firms paid less than listed price as Intel granted discounts on CPU prices to selected firms. As a results our measure of the TMR will be larger than what it would be if we had the purchase price. Thus our measure of TMR makes the TAP results more stringent than what they would be if we had the purchase price.

composite product ad expenditures for product *j* are given by

$$a_j^{\text{total pc}} = \sum_{\mathcal{H} \in \mathcal{G}_j} \frac{\lambda_{\mathcal{H}} \alpha_{j\mathcal{H}}^{Pc}}{|\mathcal{H}|}.$$

where the sum is over the different groups that include product *j*. We also estimate a nonlinear specification to allow for increasing or decreasing returns to group advertising.

Intel's marginal revenue from the marketing campaign for firm *f* is given by

$$MMR_f = \mathcal{M} \sum_{\substack{j,r \in \mathcal{J}_f \cap \\ c \in \mathcal{J}_{intel}}} (p_c^{CPU} - mc_c) \frac{\partial s_j(p,a)}{\partial a_r^{\text{total pc}}}. \tag{11}$$

As we discussed previously, we have data on p_c^{CPU} and mc_c. We can use these data together with the demand side estimates to compute the marginal revenue of advertising dollars spent by PC firm *f* on Intel chips.

6. Marginal Marketing Cost

We focus primarily on Intel and its agreements with Dell. This is for two reasons. First, both Intel and Dell were examined separately by antitrust authorities for related antitrust violations. Hence, we have a wealth of information on the amount of Intel's advertising subsidy to Dell, relative to other PC firms, especially during the period 2002-2005. Second, antitrust investigations have produced written evidence that Intel's agreements with Dell were intended to exclude their main rival (AMD) from the market, which provides a good test for our model.

We also apply the TAP test to Intel's rebates to HP and Toshiba, who were also investigated by antitrust authorities as part of the Intel case. In addition, we examine the predatory nature of Intel's rebates with Gateway. This latter application serves as a robustness check of our test as there is no evidence that Gateway was involved in anticompetitive

behavior with Intel over this period.[31]

6.1. Measuring Costs of "Intel Inside" Campaign

AVC of marketing campaign is total marketing subsidy paid by Intel/total advertising by PC firm for PCs with Intel chip. Suppose that PC firms increase advertising expenditures more as Intel increases spending on the marketing program, that is, there are increasing return to scale of the marketing program. Then average variable cost would be larger than marginal cost and the test may lead to a false positive of predation. In contrast, if PC firms tend to be less responsive to Intel's marketing subsidy as the amount of subsidy increases, average variable cost would be smaller than marginal cost and the test would be lenient.

We use a variety of measures of the average variable cost (AVC) of the marketing/ad campaign as a proxy for the marginal cost of the marketing program, which is in the same spirit as using average variable production cost as a proxy for the marginal cost in Areeda-Turner (1975) test of predatory pricing. The AVC of the marketing/ad campaign is computed by dividing the total dollar amount of the ad subsidy that Intel paid to a PC firm by total PC firm advertising for PCs powered by an Intel chip.

We construct four measures of the observed marginal cost of the marketing/ad campaign based on either actual expenditures paid by Intel (as shown in Table 11) or on an assumed percentage rebate rate.

(MC_1) This measure is constructed as the total payment of Intel to Dell as given in the case files divided by total advertising by Dell for PCs powered by an Intel chip. Table 10 (column 4) provides the total payment of Intel's rebates to Dell. Basically Intel provided discounts on CPU prices and sometimes provided a lump-sum payment. At the end of 2001, Intel began a program in which it agreed to give Dell a six percent rebate on all of Dell's CPU purchases (this came to be called the "Meet Competition Program

31 Most large PC firms were involved in the Intel case. Gateway is the largest PC firm that was not under investigation in the Intel case.

Table 10 Rebate amounts paid by Intel to Dell

Quarter	Meet the Competition Program (MCP)					
	Rebate Rate on CPU purchases (1)	Percentage Rebate Payment (2)	Lump Sum Payment (3)	Total MCP Payments (4)	Dell's Operating Income (5)	MCP % of Operating Income (6)
2002Q1	6%	$61m	-	61m	$590m	10%
2002Q2	6%	$57m	$3m	$60m	$677m	9%
2002Q3	6%	$59m	$12m	$71m	$758m	9%
2002Q4	6.3%	$77m	$7m	$84m	$819m	10%
2003Q1	6.3%	$91m	$8m	$99m	$811m	12%
2003Q2	6.3%	$106m	$6m	$112m	$840m	13%
2003Q3	6.3%	$105m	$40m	$145m	$912m	16%
2003Q4	7%	$118m	$82m	$200m	$981m	20%
2004Q1	8.7%	$137m	$70m	$207m	$966m	21%
2004Q2	12% + var.	$210m	-	210m	$1,006m	21%
2004Q3	12% + var.	$250m	-	250m	$1,095m	23%
2004Q4	12% + var.	$293m	$75m	$368m	$1,187m	31%
2005Q1	12% + var.	$307m	$81m	$388m	$1,174m	33%
2005Q2	12% + var.	$313m	$119m	$432m	$1,173m	37%
2005Q3	12% + var.	$339m	-	339m	$754m	45%
2005Q4	12% + var.	$423m	$60m	$483m	$1,246m	39%
2006Q1	12% + var.	$405m	$318m	$723m	$949m	76%

Source: p.14 MCP table from "Securities and Exchange Commission vs. Dell", U.S. District Court of Columbia Summary Statement

(MCP)).”[32] These rebates are treated as a reduction in marketing expenses in accounting by Dell. We can see that rebate rates as well as lump-sum payments have significantly increased between 2002 and 2005. Relative to the operating income, the total amount of rebates were over 70% by 2006 (column 6).

(MC_2) The second measure is the total rebate amount as a percentage of sales of PCs (across all segments sold by Dell) given in column 6 divided by total advertising by Dell for PCs powered by an Intel chip.[33] The rebates to Dell can be used for advertising in all

32 It was orginally called the “Mother of All Programs (MOAP).”

33 The second measure is the total rebate amount multiplied by a percentage of revenue from PC sales (across all segments - that is, PCs and servers - sold by Dell) divided by total Dell advertising on Intel powered CPUs.

segments including servers. Ours is a model of household demand (due to data limitations) so when we construct MC_2 we assume that the rebates are used in each sector in proportion to sector market shares. Hence MC_2 is MC_1 multiplied by the percentage of Dell's revenue from PC sector in each quarter.

(MC_3) This measure is a proxy for the marginal cost that we would have in the absence of antitrust documents. The publicly announced subsidy takes the form of a fixed percentage rebate of CPU purchases made by the PC firm with zero lump-sum payments. To compute this measure we assume six percent of all Dell's CPU purchases from Intel are rebated. We recompute the total expenditure of the marketing program to Dell (i.e., we compute column 4 assuming column 1 is always six percent and column 3 is zero). We then divide the total expenditure by total advertising by Dell for PCs powered by an Intel chip.

(MC_4) This measure provides a further lower bound on marginal costs. According to the Intel Inside website, Intel would provide a three percent rebate on purchases if marketing featured the Intel logo. To compute this measure we assume the rebate rate is three percent and there are no lump sum payments. Hence MC_4 is half of MC_3.

Notice that the last two measures of marginal cost (MC_3 and MC_4) do not rely on information obtained by antitrust officials but are based only on publicly available information. Therefore, MC_3 and MC_4 can be used to construct the TAP test for other PC manufacturers. In addition, these two measures provide a benchmark as they are computed based on the assumption that the marketing program has been executed as in normal periods or as Intel describes on their website. The other two measures (MC_1 and MC_2) are computed based on actual payments, which reflect any potential anticompetitive cost increase.

We would like to point out a limitation of applying TAP to Intel. The main issue concerns situations where a predator operates in multiple markets, which makes it difficult to determine how to allocate costs

across markets. PC firms are active in markets for home consumers, but also in markets for education, business, and government consumers. Also they sell servers as well as PCs. Fortunately, our advertising data allow us to differentiate PC advertising from advertising for non-PC products. However, we include general promotions at the firm-level as an advertising expenditure in the home segment. If general promotions affect sales in every market segment, then they should be allocated across all segments. In this case, our measure of the average cost of the marketing program is likely to underestimate the actual average cost. The marginal revenue of the marketing program would likewise be underestimated, as we cannot allow for spillovers of advertising across segments (due to data restrictions).

6.2. TAP Results

We apply the TAP test to the case of Intel. Specifically, we consider the potential predatory nature of Intel's marketing program over the period 2002 to 2005. We run TAP based on preliminary demand estimation results from specification (6). Intel's marginal revenue from the marketing campaign for firm *f* is given by

$$MMR_f = \mathcal{M} \sum_{\substack{j,r \in \mathcal{J}_f \cap \\ c \in \mathcal{J}_{intel}}} (p_c^{CPU} - mc_c) \frac{\partial s_j(p,a)}{\partial a_r^{\text{total pc}}}.$$

where $\mathcal{J}_{intel}$ is the set of Intel CPUs and $\mathcal{J}_f$ are the set of products produced by firm *f*. We compute the marginal revenue from the advertising campaign earned in that time period (i.e., ours is not a dynamic model). Estimating a dynamic model of demand and computing the associated marginal revenue that arises from a dynamic profit function would introduce considerable difficulties with respect to estimation and data requirements. However, we conduct robustness checks of our TAP results that consider the potential brand building effect of advertising. We discuss this in more detail in section 7.2.

Table 11 TAP Results of Intel's Marketing Campaign for Dell's Advertising of Intel-Powered PCs

	Computed Marginal Revenue			Observed Marginal Cost				TAP Result	# Years	# Years
Time	Est.	95% conf.	Interval	(1)	(2)	(3)	(4)	Positive in	(1)	(4)
2002Q1	1.45	(1.22	1.69)	34.54	31.12	3.33	1.66	1,2,3	5.9	0.3
2002Q2	0.88	(0.73	1.02)	45.62	40.70	3.73	1.87	all	13.0	0.5
2002Q3	1.23	(1.03	1.43)	53.31	48.05	5.90	2.95	all	10.9	0.6
2002Q4	2.22	(1.85	2.58)	91.86	82.59	13.02	6.51	all	10.4	0.7
								all		
2003Q1	2.95	(2.46	3.45)	352.14	315.44	44.78	22.39	all	29.8	1.9
2003Q2	2.38	(1.98	2.77)	192.68	171.66	17.50	8.75	all	20.3	0.9
2003Q3	3.02	(2.52	3.52)	140.80	126.81	13.61	6.80	all	11.6	0.6
2003Q4	2.92	(2.45	3.40)	93.19	82.75	7.57	3.79	all	8.0	0.3
								all		
2004Q1	3.35	(2.79	3.91)	682.73	607.55	56.66	28.33	all	51.0	2.1
2004Q2	1.95	(1.62	2.27)	345.53	306.09	18.57	9.28	all	44.3	1.2
2004Q3	1.72	(1.44	2.00)	133.28	118.52	7.39	3.69	all	19.4	0.5
2004Q4	2.05	(1.71	2.39)	192.46	171.46	8.55	4.27	all	23.4	0.5
								all		
2005Q1	6.30	(5.24	7.35)	390.02	340.19	15.51	7.75	all	15.5	0.3
2005Q2	4.38	(3.65	5.11)	370.16	323.16	11.29	5.65	all	21.1	0.3
2005Q3	4.43	(3.68	5.17)	348.36	307.69	15.68	7.84	all	19.7	0.4
2005Q4	5.85	(4.87	6.83)	340.69	298.21	13.24	6.62	1,2,3	14.56	0.28

Notes: Unit: $ (inflation adjusted - base: 2000)

The estimated marginal revenue of Intel's marketing subsidy to Dell and its 95 percent confidence interval is given in Table 12, along with the marginal cost of the marketing campaign computed in four different ways as described in section (2).[34] When the marginal cost measure is above the 95 percent confidence interval for marginal revenue, we conclude that the marketing program is not consistent with profit maximization and, more specifically, that there was an excessive marketing subsidy. Then the test result is "positive". The final column indicates for which of the four measures of marginal cost the TAP result is positive.

It is worth to note the following two points. First, the marginal revenue estimates imply that 1$ PC advertising tends to increase Intel's revenue by more than 1$ (except in the second quarter of 2000). Thus, if all of Intel's payment to Dell are used for marketing PCs powered by Intel CPU, then the results suggest that Intel gains more than its cost by subsidizing Dell. However, our measures of marginal cost are much larger than 1, that is, PC advertising expenditure falls short of Intel's marketing subsidy to Dell in reality.[35]

This implies that Intel needs to incur more than 1$ to induce Dell to spend 1$ more on advertising. Second, MC_3 and MC_4 are significantly smaller than MC_1 and MC_2. Considering that MC_1 and MC_2 are computed based on the actual payment whereas MC_3 and MC_4 are computed by assuming the rebate rate applied in the normal periods, our

---

34 Since the marginal revenue of the marketing program is a linear combination of utility parameters, the confidence interval around the estimate of the marginal revenue can be computed easily. MC_1 is Intel's total payment to Dell divided by the total advertising expenditure of Dell on PCs powered by Intel CPU; MC_2 is MC_1 multiplied by the fraction of revenue from PC sales (= PC sales/(PC sales + server sales); MC_3 is six percent of Intel's revenue from selling CPUs to Dell (=0.06 · $\mathcal{M} \sum_{c \in \mathcal{J}_{intel}} \sum_{j \in \mathcal{J}_c \cap \mathcal{J}_{Dell}} s_j p_c^{CPU}$)$MC_4$ is three percent of Intel's revenue from selling CPUs to Dell (so $MC_4 = MC_3/2$).

35 There are a few reasons why marginal costs may be larger than one. First, while, in principle, the marketing subsidy should have been used only for marketing, case files indicate that stock analysts had doubt about the use of the subsidy. The subsidy was actually a huge amount that accounts for a significant portion of operating profits. Second, marketing may be broadly defined to include more than advertising. For example, it may encompass training employees in how to sell PCs.

results suggest that Intel has paid much more to Dell than described in their Intel Inside marketing campaign.

The TAP results show that marginal cost, for all measures, is above the 95 percent confidence interval of marginal revenue estimates in most periods. Exceptions are when MC_4 is compared to marginal revenue in the beginning of the anticompetitive period and at the end of the anticompetitive period. Even in these cases, MC_4 is larger than the estimated marginal revenue. Comparisons of marginal revenue to the actual-payment-based marginal cost is consistent with predatory marketing.

We do not have the data on Intel's actual payment to other firms and thus cannot compute MC_1 and MC_2 for firms other than Dell. Although, we can compute MC_3 and MC_4 if we assume fixed rebate rates. The anticompetitive charges were mainly about Intel's payment to Dell, but other PC firms such as HP and Toshiba were also involved in the case. Tables 13 presents the TAP results for HP and Toshiba, respectively. Both the marginal revenue and marginal cost measures tend to be lower in the cases of HP and Toshiba than in the case of Dell. Our TAP results show evidence of predation with Toshiba in 2003 and 2004. During this period the TAP results are positive for both measures of marginal cost.

We do not have a dynamic part of demand, that is, long-term effect of advertising, in our model. As a simple check, we can use the static MR in each period, which we already have. The concern is that the firm is considering the MR in many periods in the future and this is what should be compared to MC. We can get a "static" version of the dynamic MR by summing the *MR* over many periods and comparing this to one period MC. For instance, we could say in order for the MC_1 for Dell to not be predatory we would have to believe that Intel is choosing that amount of advertising taking into account the impact of MR 10 periods in advance. In the last two columns of Table 12, we added the number of periods (in years) over which *MR* should be summer to rationalize the cost of advertising. The first of the two is based on MC_1-which is the highest *MC*estimate -and the second one is based on MC_4-which is the lowest *MC* estimate.

Specifically, we take a very generous assumption about Intel's advertising campaign that Intel is very optimistic and thinks that current

Table 12 TAP Results for HP and Toshiba

Unit: $ (inflation adjusted - base: 2000)

PC firm	HP						Toshiba					
	Computed MR			Observed MC		TAP Result	Computed MR			Observed MC		TAP Result
Time	Est.	95% conf.	Interval	(3)	(4)	Positive in	Est.	95% conf.	Interval	(3)	(4)	Positive in
2002Q1	-0.02	(-0.32	0.28)	0.11	0.05		0.13	(0.11	0.15)	0.14	0.07	
2002Q2	1.04	(0.89	1.19)	0.26	0.13		0.26	(0.22	0.29)	0.09	0.04	
2002Q3	1.20	(1.01	1.39)	0.27	0.13		0.66	(0.56	0.76)	0.16	0.08	
2002Q4	-5.68	(-9.93	-1.43)	0.12	0.06	3,4	0.62	(0.53	0.72)	0.45	0.22	
2003Q1	-0.42	(-1.30	0.46)	0.09	0.05		0.34	(0.29	0.40)	1.37	0.68	3,4
2003Q2	0.15	(-0.34	0.64)	0.11	0.05		0.21	(0.18	0.25)	0.82	0.41	3,4
2003Q3	0.36	(-0.25	0.97)	0.17	0.09		0.31	(0.26	0.36)	1.47	0.74	3,4
2003Q4	0.59	(-0.45	1.64)	0.30	0.15		0.44	(0.37	0.51)	1.45	0.72	3,4
2004Q1	-0.07	(-0.80	0.66)	0.14	0.07		0.34	(0.28	0.40)	0.99	0.50	3,4
2004Q2	0.32	(-0.01	0.66)	0.13	0.07		0.36	(0.30	0.42)	0.99	0.50	3,4
2004Q3	0.59	(0.31	0.87)	0.20	0.10		0.43	(0.36	0.49)	0.48	0.24	3
2004Q4	0.37	(-0.04	0.77)	0.17	0.09		0.99	(0.84	1.14)	0.29	0.15	
2005Q1	2.47	(0.80	4.15)	0.17	0.08		1.96	(1.65	2.27)	0.35	0.17	
2005Q2	4.90	(4.12	5.68)	0.33	0.17		1.99	(1.67	2.32)	0.42	0.21	
2005Q3	6.91	(5.90	7.92)	0.57	0.29		2.94	(2.45	3.42)	0.72	0.36	
2005Q4	6.31	(5.18	7.43)	0.41	0.21		2.97	(2.49	3.44)	0.49	0.24	

advertising would have the same effect (that is, the same *MR*) in the future periods too (for some while). Then we can just compute: $T = MC/MR$ where *MC* is our computed *MC* and *MR* is our estimate. Then, we can say that, if Intel thinks that advertising will have the same effect on future sales as in the current period for *T* time periods, then the advertising expenditure can be justified (that is, the present value of *MR* and *MC* are equalized). Since it may be easier to catch the idea about how long it would take for advertising to be rationalized by dynamic optimization (even if we are very generous, thinking that Intel is allowed to expect the same *MR* as the current *MR* for some future periods). In the table, we computed calculated $T/4=(MC/MR)/4$ so that we can see how many years we need to rationalize Intel's advertising choice.

Table 13 | TAP Results for Gateway

PC firm	Gateway					
	Computed Marginal Revenue			Observed Marginal Cost		TAP Result
Time	Est.	95% conf.	Interval	(3)	(4)	Positive in
2002Q1	0.56	(0.37	0.76)	0.14	0.07	
2002Q2	0.38	(0.28	0.49)	0.10	0.05	
2002Q3	0.47	(0.38	0.55)	0.16	0.08	
2002Q4	0.78	(0.66	0.90)	0.26	0.13	
2003Q1	0.87	(0.75	1.00)	0.20	0.10	
2003Q2	0.47	(0.40	0.54)	0.20	0.10	
2003Q3	0.50	(0.43	0.58)	0.17	0.08	
2003Q4	0.56	(0.47	0.64)	0.15	0.07	
2004Q1	0.66	(0.56	0.76)	0.78	0.39	3
2004Q2	0.40	(0.34	0.47)	0.50	0.25	3
2004Q3	0.52	(0.44	0.60)	0.37	0.18	
2004Q4	0.15	(0.96	1.34)	2.66	1.33	3
2005Q1	4.52	(3.77	5.28)	6.37	3.18	3
2005Q2	3.59	(3.00	4.19)	2.64	1.32	
2005Q3	4.72	(3.94	5.51)	4.21	2.11	
2005Q4	5.72	(4.80	6.64)	1.79	0.90	

Unit: $ (inflation adjusted - base: 2000)

We are also interested in what TAP would show with other PC firms not involved in the case. Since most PC firms not investigated for the case are small, we consider only Gateway, which has a significant market share but is not involved in the case. Table 14 presents the TAP results. We can see that marginal cost measures tend to either fall in the 95 percent confidence interval of marginal revenue or just below the interval. The result provides a nice contrast to the test result for Dell as we do not find excessive advertising for Gateway for both measures of marginal costs in any period. This result also supports the idea that the Intel's marketing program, if conducted as described, is driven by profit maximization rather than anticompetitive purposes for the case of Gateway. Overall, the findings with other PC firms seem to suggest that our model, despite its simplicity, represents demand reasonably well and captures the important features of the marketing program.

7. Robustness and Other Considerations

7.1. Motive, Recoupment, and Efficiency Motives

In this section, we discuss the industry background that speaks to the motives for predation, the prospect of successful predation and recoupment, and dynamic efficiency. Intel is a dominant firm in the CPU industry with about 80 percent of worldwide CPU sales. Its major (and only effective) rival is AMD, holding about 18 percent market share (Mercury Research, 2007). In 1999 and 2003, respectively, AMD introduced two new chips, the Athlon for personal computers and the Opteron for servers. These AMD chips were the high-end products intended to compete with Intel. By introducing these 64-bit processors, AMD enabled PC operating systems to handle large amounts of information more fastly and accurately (as compared to a 32-bit OS system). These attributes were welcomed by experts and consumers and it was generally agreed that these AMD chips were better-performing and cheaper than Intel counterparts. The case files denote that the Athlon "was almost universally recognized as being superior to Intel's then current top model for PCs, the Pentium III"(pp.14-15, Complaint,

US District Court of Columbia, SEC vs. Dell) and that "Opteron garnered virtually unanimous industry acclaim; AMD had succeeded with an innovative product design yielding performance advantages which effectively "leapfrogged" Intel."(pp.14-15, Complaint, US District Court of Delaware, State of New York vs. Intel). In addition, Table 5 of section 3.2 provides a supporting evidence from the data. The threat of new, high-performance processors from AMD may have induced Intel to engage in anticompetitive actions. These events provide the motive for Intel's predatory behavior. Indeed, many jurisdictions in the world accused Intel of using various anticompetitive tactics against AMD starting in 2002.

We are particularly interested in Intel's marketing subsidies. Predation involves short-run profit sacrifice and long-run recoupment. The TAP test is used to establish short-run profit sacrifice. However, we now turn to industry characteristics to examine the ease (or difficulty) with which Intel could successfully drive AMD out of the market and recoup lost profits by maintaining market power for a sufficiently long period after AMD's exit.

There are a number of factors that make long-run recoupment of profits likely to be successful in the CPU industry. To remain as a valid competitor in a rapidly changing, high-technology industry like the CPU industry, firms need to secure constant cash flows and keep investing in innovation. The CPU industry is capital-intensive, hence firms will incur substantial costs to construct and maintain manufacturing plants (called "fabs"). If a firm does not have sufficient internal funding, it must obtain external funding at market rates. According to industry experts, Intel is able to fund its fabs with revenue, while AMD must secure funding at market rates, which significantly raises AMD's cost of capital. Furthermore, obtaining external financing is complicated due to agency problems. Typically investors require firms to show a positive prospect of future profits, which is often based on current performance. Predation would make the future prospect of the prey look lower (and potentially negative) and ultimately induce it to exit the market. Thus, predation in the CPU market would be consistent with the long-purse (deep-pocket) theory of predation.

Second, since firms are continuously innovating, they may be

uncertain about how consumers will react to new products. New processors can have different characteristics possibly appealing to a different market segment from current customers. As mentioned before, the beginning of the anticompetitive use of the marketing program coincides with AMD's introduction of high-performance chips. By engaging in predatory behavior, Intel could send a (wrong) signal about the demand for new chips, which is consistent with the demand signaling theory (test-market theory) of predation.

Lastly, economies of scale exist in the CPU industry. The substantial investment in plants and technologies are sunk. Therefore, a firm needs to secure a certain amount of sales in order to recover the sunk costs and stay in business. It is easier for a dominant firm to exclude a rival and prevent new entrants in the presence of economies of scale. In this sense, predation is likely to be successful in driving AMD out of a market and Intel is likely to keep high profit margins for a sufficiently long time.

The CPU industry is inviting to predatory behavior for these reasons, and Intel is an incumbent with a dominant market share. Given that Intel's recoupment is very likely as a monopolist due to high entry barriers and that predation can successfully lead to exclude AMD in the CPU industry, showing sacrifice of short term profits would support that the marketing program is predatory.

The TAP test examines if the return on advertising (i.e., how it impacts demand) is high enough to justify marketing expenditures (as these are directed at increasing demand). Short-term profit sacrifice may be justified by dynamic efficiency reasons. Although the cost-based approach is widely used to show profit sacrifice in predatory pricing cases, pricing below cost does not necessarily mean the behavior is predatory. Short-term profit sacrifice can be rationalized by potential dynamic efficiency reasons such as learning-by-doing, promotional purposes (e.g., introductory prices), or network externalities. Our demand model includes only the current, short-term effect of advertising, hence the potential long-run benefit of the marketing program is not taken into account. However, just as the efficiency reasons for pricing below marginal cost are not usually applicable to an already dominant, incumbent firm with a large customer base, here too an unprofitable advertising subsidy by Intel is not easily justified by efficiency reasons.

Intel should already have achieved an efficient scale of operation, so learning-by-doing does not seem to justify short-term profit sacrifice. Given that Intel has been present for a long time and consumers already know about Intel and that the anticompetitive actions have been going on for four years, promotional motives are an unlikely explanation for short-term profit sacrifice. In addition, network externalities are not strong in the CPU market. For example, PC purchase guides, such as *Consumer Reports*, do not list the size of the customer base using Intel processors as an important factor for consumers to consider when purchasing a PC.

Our main concern is the brand-loyalty-building effect of the marketing program. Advertising is generally believed to build goodwill and this may be a reason for Intel to invest in marketing at the expense of short-term profits. Notice that, this incentive is constant across all periods, while the predatory motive is more pronounced during the period of AMD's new chip introductions. To consider this the TAP test includes two measures of marginal cost (MC_3 and MC_4) that serve as competitive benchmark as they are based on listed rebate rates that would have been applied prior to 2001/2002. In contrast, MC_1 and MC_2 are based on actual payment post AMD's introduction of the new chips. Hence, these measures would include brand loyalty building incentives plus anticompetitive motives, while MC_3 and MC_4 would be driven only by brand loyalty building incentive. We find that MC_1 and MC_2 (based on actual payment) are much larger than MC_3 and MC_4 for Dell. The results suggest that an anticompetitive motive induced Intel to sacrifice even more short-term profits as the difference between marginal marketing revenue and marginal costs are much larger when using MC_1 and MC_2. Also it is worth to note that, if advertising can establish strong brand loyalty, predatory marketing can be even more harmful as it would work as an endogenous entry barrier, deterring further entries and making recoupment even more likely. We provide further robustness checks regarding brand-loyalty in section 7.2.

7.2. Robustness Checks and Specification Tests

In this section, we conduct robustness checks of the model as it

relates to brand-loyalty building of advertising, cost pass-through, and model specification to insure the objective function did not converge to a local minimum (Knittle and Metaxoglou, 2008).

to be completed

8. Policy Implications

Price and quantity are not the only strategic variables that can be used for anticompetitive purposes. Advertising is another important strategic variable commonly employed by firms. However, antitrust authorities typically try to establish anticompetitiveness through pricing, but do not address the strategic use of advertising and, more generally, marketing campaigns. While the heart of the anticompetitive actions of Intel was their Intel-Inside marketing program, considerations of advertising/marketing predation where not at the forefront of the antitrust case. In this paper we focus on non-price anticompetitive behavior arising from marketing/advertising with a focus on the Intel case.

Our paper proposes a Test of Advertising Predation (TAP) that can be used to detect non-price predatory behavior. We provide a general Test of Advertising Predation (TAP) based on the presumption that, if a firm's marketing campaign is not predatory, marketing expenses should be profit maximizing and so should result in sufficient increased product demand to justify costs. To construct TAP, first we model consumer's demand for PCs from which we infer demand for CPU processors. Specifically, we estimate a random-coefficient model of demand for a PC-CPU, where the coefficients on PC and CPU characteristics and advertising vary with demographics.[36] Second, we compute Intel's

36 In previous literature that estimates CPU demand, it is generally assumed that final consumers directly purchase the CPU. We think it is more realistic to model consumers' choice of a CPU-PC combination. In addition, since we are interested in the e.ect of PC advertising on CPU demand, and PC advertising does not directly a.ect CPU demand, we model a consumer's discrete choice over CPU-PC combinations.

marginal revenue from the marketing subsidy using the demand side estimates. That is, we compute the marginal revenue of advertising dollars spent on Intel chips at the firm or product level.[37] The marginal revenue of the marketing program depends on the parameters of consumer utility (including advertising), CPU price and marginal manufacturing cost.

Test results suggest short-term profit sacrifice by Intel, supporting the predatory use of the Intel Inside campaign. To rationalize the short-term profit sacrifice, there should be something Intel can gain from the marketing program other than increasing CPU sales by boosting the willingness-to-pay for a PC. Antitrust authorities found evidence that the marketing subsidy is paid on anticompetitive condition that limits the use of AMD processors. This condition aimed at driving AMD out of a market and the prospect of future profit as a monopolist as a result would have rationalized the short-term profit sacrifice.

Our method can be used to guide antitrust authorities in future cases as it provides a general framework for testing for anticompetitive use of marketing campaigns. Computing the test requires little extra estimation over the typical demand estimation usually undertaken by antitrust authorities. Furthermore, the advertising data necessary to estimate the model parameters is usually not so difficult to obtain. Thus TAP is practical and easy to implement. Applying to the Intel case, this paper shows that our test can be used to show the predatory use of advertising/marketing. In addition, the benefit of looking at the advertising side is that, unlike predatory pricing which accompanies low price in the short term, predatory advertising/marketing does not have a clear benefit for consumers, even in the short term. In the long run, predatory marketing can be harmful if it has a long-lasting effect by establishing goodwill, which may become an endogenous entry barrier for potential competitors.

---

37 We do not model strategic decisions of PC firms and Intel. This makes the test we develop more stringent. Rather, PC firms' CPU choices are assumed to simply reffect consumers' demand, and not affected by the marketing campaign. Intel was accused of giving refunds to PC firms in the Intel Inside marketing program on the exclusionary condition that they limit the use of AMD chips. This implies the marketing program would affect PC firms' CPU choice and, hence, its anticompetitiveness would be even larger.

References

Angel, J (2002) “Retreat and Persist,” *Technology Marketing*.

Areeda, P. and Turner, D.F. (1975) “Predatory Pricing and Related Issues Under Section 2 of the Sherman Act, *Harvard Law Review*,”88, 697-733.

Baker, J.B. 1994. Predatory pricing after Brooke Group: an economic perspective. *Antitrust Law Journal* 64, 585-604.

Bagwell, Kyle (2007) “The Economic Analysis of Advertising,” Handbook of Industrial Organization, Elsevier, edition 1, volume 3, number 1. Mark Armstrong & Robert Porter (eds).

Baumol W. J. (1996) “Predation and the Logic of the Average Variable Cost Test,” *Journal of Law and Economics*, 39, 49-72.

Berry, S., J. Levinsohn, A. Pakes. (1995) “Automobile prices in market equilibrium,” *Econometrica* 63(4), 841-890.

Besanko, D., U. Doraszelskiz, and K.Yaroslav (2010) “The Economics of Predation: What Drives Pricing When There is Learning-by-Doing?,” Unpublished Manuscript .

Bolton, P., J. Broadley, and M. Riordan (2000) “Predatory Policy: Strategic Theory and Legal Policy,” *Georgetown Law Review*, 88, 2239-2330.

Bolton, P., J. Broadley, and M. Riordan (2001) “Predatory Pricing: Response to Critique and further elaboration,” *Georgetown Law Journal*, 89, 2496-2529.

Bolton, P., and D. Scharfstein (1990) “A theory of predation based on agency problems in financial contracting”, *American Economic Review*, 80, 93-107.

Bork, R. H.(1978) *The Antitrust Paradox*, New York: Free Press.

Bureau of Labor Statistics, All Urban Consumers CPI-U, U.S. city average (ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt).

Cabral, L. and M. H. Riordan (1997) “The Learning Curve, Predation, Antitrust, and Welfare,” *The Journal of Industrial Economics*, 2, 155-169.

Chen, Y. and W. Tan (2007) “Predatory Advertising: Theory and Evidence in the Pharmaceutical Industry,” Working Paper.

Dube, J. P., J. T. Fox, and C.-L. Su (forthcoming) “Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients

Demand Estimation," *Econometrica*.

Easterbrook, F.H. 1981. Predatory strategies and counterstrategies. *University of Chicago Law Review* 48, 263-337.

Easterbrook, F.H. 1984. The limits of antitrust. *Texas Law Review* 63, 1-40.

Edlin, E. (2012), Research *Handbook on the Economics of Antitrust Law*, chapter on predatory pricing, Einer Elhauge editor, Edward Elgar Publishing Massachuseets USA.

Eizenberg A. (2011) "Upstream Innovation and Product Variety in the United States Home PC Market," Working Paper.

Ellison, G. and S. F. Ellison (2011) "Strategic Entry Deterrence and the Behaviour of Pharmaceutical Incumbents Prior to Patent Expiration," *American Economic Journal: Microeconomics*, 3(1), 1-36.

Elzinga, K.G. and Mills, D.E. 1994. Trumping the Areeda-Turner test the recoupment standard in Brooke Group. *Antitrust Law Journal* 62, 559-84.

Elzinga, K.G and Mills, D.E. 2001. Predatory pricing and strategic theory. *Georgetown Law Journal* 89, 2475-93.

Genesove, D. and W. P. Mullin (2006) "Predation and Its Rate of Return: The Sugar Industry, 1887-1914," *Rand Journal of Economics*.

Goettler, R. and B. R. Gordon (2009) "Competition and innovation in the microprocessor industry: Does AMD spur Intel to innovate more?" Working paper, Columbia University, New York.

Gowrisankaran, G. and M. Rysman. (2007) "Dynamics of consumer demand for new durable consumer goods," Working paper, Boston University, Boston.

Gordon, B. (2009). "A Dynamic Model of Consumer Replacement Cycles in the PC Processor Industry,"Unpublished Working Paper, Columbia Business School.

Granitz, E. and Klein, B. 1996. Monopolization by 'raising rivals' costs': the standard oil case. *Journal of Law and Economics* 39, 1-47.

In-Stat (2005) "Intel Manufacturing Capacity and Die Cost".

In-Stat (2005) "Intel Rosetta Stone: Intel Processor Shipments, Forecasts, Technology, and Roadmaps".

Joskow, P. and A. K. Klevorick (1978) "A Framework for Analyzing Predatory Pricing Policy,"89 *Yale Law Journal*, 213, 219-220.

Moon, Y. E. and Darwall, C. (2005) "Inside Intel Inside," *Harvard Business School Case* 502-083.

Petrin, A. (2002) "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110(4), 705-729.

Prince, J. T. (2008), "Repeat Purchase Amid Rapid Quality Improvement: Structural Estimation of Demand for Personal Computers," *Journal of Economics and Management Strategy* 17(1), 1-33.

Reynaert M. and F. Verboven (2013) "Improving the Performance of Random Coefficients Demand Models-the Role of Optimal Instruments" forthcoming *Journal of Econometrics.*

Salgado, H. (2008a) "Dynamic Firms Conduct and Market Power: The Computer Processors Industry under Learning-by-Doing," Working Paper, University of California at Berkeley.

Salgado, H. (2008b) "Brand Loyalty, Advertising and Demand for Personal Computer Processors: The Intel Inside Effect," Working Paper, University of California at Berkeley.

Slade, M. E. (2005) "Product Rivalry with Multiple Strategic Weapons: An Analysis of Price and Advertising Competition," Working Paper.

Snider, C. (2009) "Predatory Incentives and Predation Policy: The American Airlines Case", Working Paper.

Song, M. (2007) "Measuring Consumer Welfare in the CPU Market: An Application of the Pure Characteristics Demand Model," *RAND Journal of Economics*, 38, 429-446.

Sovinsky Goeree, Michelle (2008) "Limited Information and Advertising in the US Personal Computer Industry," *Econometrica*, 1017-1074.

Thornhill, T. Lee C., and R. Shannon (2001) "Intel Corporation," *UBS Warburg Global Equity Research Report.*

Rojas, C. and E. Peterson (2008) "Demand for Differentiated Products: Price and Advertising Evidence from the U.S. Beer Market," *International Journal of Industrial Organization*, 26, 288-307.

Ying, F. (2010) "Ownership Consolidation and Product Quality: A Study of the U.S. Daily Newspaper Market," Working Paper.

Weiman, D. and Levin, R. 1994. Preying for monopoly? The case of Southern Bell Telephone Company, 1894-1912. *Journal of Political Economy* 102, 103-26.

Whinston, M. D. (2006) *Lectures on Antitrust Economics*, The MIT Press, Cambridge, Massachusetts.

Appendix

Appendix Table 1 Product Cross-Reference from Processor Core to Brand Name (i.e. Marketing Name) in Sample (Q1:2002 - Q4:2005)

Platform		Processor Core	Brand Name	Speed (Frequency: MHz)
Desktop	Mainstream	Willamette		1300-2000
		Northwood	Pentium 4	1600-3400
		Prescott		2260-3800
		Smithfield*	Pentium D	2667-3200
	Value	Tualatin	Pentium III	1000-1400
			Celeron	900-1400
		Willamette		
		Northwood	Celeron	1500-2000
				1600-2800
		Prescott	Celeron D	2133-3460
Mobile	Mainstream	Northwood	Mobile Pentium 4-M	1200-2600
		Prescott	Mobile Pentium 4	2300-3460
		Banias	Pentium M	1200-1800
		Dothan		1300-2267
	Value	Tualatin	Mobile Celeron	1000-1330
			Mobile Pentium III-M	866-1333
		Northwood	Mobile Celeron	1400-2500
		Banias	Celeron M	1200-1500
		Dothan		1200-1700

Appendix Table 1 (Continue)

Platform		Processor Core	Brand Name	Speed (Frequency: MHz)
	Low-Power	Tualatin LV	Mobile Pentium III-M	733-1000
		Tualatin ULV		700-933
		Tualatin LV	Mobile Celeron	650-1000
		Tualatin ULV		650-800
		Banias LV		1100-1300
		Banias ULV	Pentium 4	900-1100
		Dothan LV		1400-1600
		Dothan ULV		1000-1300
		Banias ULV	Celeron M	600-900
		Dothan ULV		900-1000

Notes: * Dual-core processor
Low-power mobile PCs are mini-notebook, tablet, and ultraportables.
(LV: low-voltage; ULV: ultra-low-voltage)

CHAPTER 4

Entrepreneurship, Small Businesses, and Economic Growth in Cities

by
Yong Suk Lee[*]
(Stanford University)

Abstract

This paper examines the impact of entrepreneurship on urban economic growth. I use the homestead exemption levels in state bankruptcy laws from 1975 to instrument for entrepreneurship and examine urban growth between 1993 and 2002. I find that a ten percent increase in the birth of small businesses increases MSA employment by 1.1 to 2.2%, annual payroll by 3.1 to 4.0%, and wages by 1.8 to 2.0% after ten years. I find no growth impact from entrepreneurship backed by the federal Small Business Loan program and further find that government-backed entrepreneurship crowds out privately financed entrepreneurship one for one.

* Lee: Freeman Spogli Institute of International Studies, Stanford University. I thank Nathaniel Baum-Snow, Kenneth Chay, Gilles Duranton, Leo Feler, Vernon Henderson, David Love, Chang-gyun Park, Junfu Zhang and seminar participants at the Korea Development Institute, Johns Hopkins University School of Advanced International Studies, Williams College, Brown University, the Urban Economics Association Annual Meeting, and the American Real Estate and Urban Economics Annual Meetings for comments.

1. Introduction

Entrepreneurship is widely believed to be a main source of economic growth. Entrepreneurs that succeed and contribute to the local economy become the spotlight of local media. Politicians and business advocates emphasize the role small businesses play in adding new jobs, and small businesses are a frequent topic in presidential debates. Governments both in the developing and developed world consider entrepreneurship as a way to jump start economic development and sustain economic growth. In the U.S., the Small Business Administration has actively promoted and supported small businesses since 1953. Employment statistics are often used to support the importance of entrepreneurship and small businesses in adding jobs to the economy.[1] However, while there are successful entrepreneurs, businesses also fail. According to the Bureau of Labor Statistics, only a third of all new establishments survive after 10 years. As important understanding entrepreneurship's contribution to economic growth may seem, we have surprisingly little empirical evidence on whether or not entrepreneurship promotes economic growth and if so by how much.

This paper's objective is threefold: (1) to estimate the impact of entrepreneurship measured by the birth of small establishments on urban employment and income growth; (2) to examine how entrepreneurship supported by government guaranteed loans perform relative to privately financed entrepreneurship regarding its impact on urban growth; and (3) to examine whether government backed entrepreneurship crowds out privately financed entrepreneurship. Overall, the paper will provide estimated magnitudes of the importance of entrepreneurship and shed light on policy's role in the promotion of entrepreneurship.

The extensiveness of the data required to examine business dynamics

1 Kleisen and Maues (2011) find that between 1992 and 2010 small firms with 1 to 19 employees provided about 30 percent of the gross new jobs in the economy, which is the largest percentage among the different firm size categories. However, they find that those small firms accounted for only 16 percent of the net new jobs, the smallest percentage among the different firm size categories. The cutoff for small businesses is relevant in this regard. The Small Business Administration uses the 500 employees cut off and report that small businesses account for 64% of net new jobs.

had been one of the main impediments in furthering our understanding of the relationship between individual business size and growth. However, recent research has made substantial improvements. Haltwinger et al.(2013), using the Census Longitudinal Business Dynamics data, examine the universe of all firms and establishments in the US and find that once firm age is controlled for smaller businesses grow no faster than larger businesses. They find that the main source of employment growth is attributed to small and young businesses. Neumark et al. (2011) also find similar results using the National Establishment Time Series data. Even though only a subset of news mall businesses survives, small businesses significantly contribute to the creation of jobs. These findings shed light on the importance of new small businesses. However, the implications of these studies are somewhat limited in its focus on average year to year growth. Given that many small firms die out and economic growth is assessed on intervals longer than one year, I focus on the impact of entrepreneurship after 5 or 10 years. Also, rather than focusing on individual businesses, I examine the impact of entrepreneurship on an aggregate economy, i.e., the metropolitan area.

Duranton and Puga (2013) review the determinants of urban growth and highlight entrepreneurship, along with human capital, as the main sources of dynamic aggregate growth in cities. The entrepreneurship literature has extensively examined the determinants of entrepreneurship including funding sources (Kerr et al. 2010, Samila and Sorenson 2011), housing collateral (Adelino et al. 2013, Brack et al. 2013), family (Bertrand and Schoar 2006), and peers (Lerner and Malmendier 2011). My focus on the aggregate economy is similar in spirit to Samila and Sorenson (2011), who examine the impact of venture capital on entrepreneurship and growth at the MSA level. The urban economics literature has examined the agglomeration benefits of production in cities (Greenstone et al. 2010, Henderson et al. 1995, Glaeser et al. 1992) and in particular, Rosenthal and Strange (2003) examine the agglomeration benefits to firm birth. Unlike many of the previous studies that have examined entrepreneurship as an outcome, my paper examines the impact of entrepreneurship on economic growth. The fundamental difficulty in examining this question is the joint determi-nation of the two, and finding a plausibly exogenous variation in entrepreneurship-

continues to be a challenge in the literature. One recent development has been Glaeser et al.(2012). They use proximity to mines in 1900 as instruments for average establishment size and find that cities with smaller average establishment size have higher employment growth.

I contribute to this nascent literature in three ways. First, in examining the impact of entrepreneurship on economic growth, I use a more direct measure of entrepreneurship, i.e. business births, in addition to the average establishment size proxies that have been used in the literature. Second, I introduce a new instrumental variable in generating a plausibly exogenous variation of entrepreneurship across cities. I use the homestead exemption levels set by state bankruptcy laws in 1975 as instrumental variables. States varied substantially in the degree to which debtors could avoid paying creditors back and such variation dates back to the nineteenth century. Posner et al. (2001) point out that the variation in the state's desire to promote migration in the 19th century and the legislative negotiation process, where negotiation starts based on initial exemption levels, caused state exemption levels to persist over a long period of time. Lastly, I separately examine privately financed and government backed entrepreneurship to assess policy's role in promoting entrepreneurship.

I find that cities with unlimited or higher exemption levels in 1975 see higher business births in 1993. Using this variation I find that a ten percent increase in entrepreneurship increases urban employment by 1.1 to 2.2%, annual payroll by 3.1 to 4.0%, and wage by 1.8 to 2.0% after ten years. These results are robust to additional controls of business environment, such as the minimum wage and the Right-to-work law, and past population. The instrumental variable regression estimates are smaller than the OLS estimates. This indicates that unobserved city level growth potentials impact entrepreneurial activity across cities and that OLS estimates are likely biased.

For every 100 businesses that are created in the private market there is one business created through government guaranteed loans. The Small Business Administration provides guaranteed loans to entrepreneurs who could not secure loans from the private market. I examine how these government-backed businesses impact urban employment and income growth. Using the universe of the Small Business Loan (SBL)

data I aggregate all loan approvals to the MSA year level and generate the number of new loan approvals and the total approved amount by MSAs. Examining the impact of government-backed entrepreneurship on urban growth in an OLS framework suffers from the endogeneity problem as before. Cities with higher growth potential may see more SBA loan applications and approval. On the contrary, cities that were declining with more people being laid off may see higher SBA loan applications and approval. In order to generate plausibly exogenous variation in SBA backed entrepreneurship, I use years since interstate banking deregulation and the number of SBA lender per capita in 1985 as instrumental variables. The banking sector was heavily deregulated during most of the 20th century. Gradually, each state allowed banks to operate across state borders. The new competition generated by multiple banks would provide more opportunities for personal and business finance. I find that metropolitan areas that deregulated earlier see more market entrepreneurial activity and less need to go through the SBA to finance a business in 1993. Cities with higher density of SBA lenders in 1985 would see more competition among SBA lenders which could facilitate capital constrained potential entrepreneurs. I indeed find that higher density of SBA lenders in 1985 increase SBA backed entrepreneurship in 1993. Whichever set of instruments I use, I find no impact of government-backed entrepreneurship on urban employment or income growth.

To further assess the role of government-backed entrepreneurship on urban growth, I examine whether government-backed entrepreneurship complements or substitutes market entrepreneurship. The cross-sectional variation initially indicates that the two are complements in entrepreneurial activity. However, when I examine within metropolitan areas over time I find statistically significant impact of crowd out. For one government-backed entrepreneurship, there is one less market entrepreneurship. The one for one crowd out and the fact that market entrepreneurship contributes to urban economic growth but that government-backed entrepreneurship does not, suggests that there is no efficiency gain from the government's involvement in promoting entrepreneurship.

The paper proceeds as follows. Section 2 discusses the theory that guides the empirical work. Section 3 discusses the data and variables

used in the analysis. Section 4 examines the impact of entrepreneurship on urban growth. Section 5 compares the impact of government-backed entrepreneurship and market entrepreneurship. Section 6 concludes.

2. A Simple Theory of Entrepreneurship and Urban Growth

I introduce entrepreneurship to a standard model of urban growth (Glaeser et al. 1992, Henderson et al. 1995) to guide the empirical work. Consider a representative firm in a city at time *t* where production is specified as $f(L_t) = A_t L_t^{\alpha},\ 0 < \alpha < 1$. A_t represents the level of technology and L_t the level of labor input at time *t*. The model abstracts away from other factors of production such as, capital and land, and hence will not be able to capture change in wage or employment due to labor substituting technological advances. I note that city subscripts are dropped in the description of the model for expositional brevity. Within this stylized framework, labor is paid the value of marginal product where output price is normalized to one, returning the labor demand function $w_t = f'(L_t) = \alpha A_t L_t^{\alpha-1}$. Putting this in a dynamic framework the growth of employment in a city can be represented as

$$(1-\alpha)\Delta \ln L_t = \Delta \ln A_t - \Delta \ln w_t \quad (1)$$

where $\Delta \ln L_t = \ln L_{t+1} - \ln L_t$, and similarly for the other variables. I specify the growth of the technology as:

$$\Delta \ln A_t = \ln A_{t+1} - \ln A_t = g(e_t, N_t, ini_t, \rho) \quad (2)$$

Where I define e_t as the aggregate entrepreneurship in the city at time *t*. Note that e_t is the *aggregate* entrepreneurial level and hence is impacted by the *number* of entrepreneurial activity as well as the *average* entrepreneurial ability of entrepreneurs in the city. N_t is the size of the city measured by population capturing traditional agglomeration externalities, and ini_t represents initial

economic condition that might explain growth of technology in the city, such as, initial employment, income, cost of living, and education level. ρ is the national growth rate of technology that is constant across cities.

I assume an upward sloping labor supply curve $w(L) = w_0 L^{\sigma}, \sigma > 0$. The upward sloping labor supply relaxes the perfect labor mobility and the cross-city wage equalization assumptions often used in the literature and allows workers to have preferences for cities. Hence, wage growth is no longer constant at the national level but can vary across cities. Incorporating labor supply into (1) and (2) returns the reduced form equations:

$$\Delta \ln L_t = L(e_t, N_t, L_t, w_t, ini_t)$$
$$\Delta \ln w_t = w(e_t, N_t, L_t, w_t, ini_t) \quad (3)$$

The main empirical test will be to examine whether entrepreneurship indeed promotes the growth of city employment and wages, i.e., whether

$$\partial \Delta \ln L_t / \partial e_t > 0 \text{ and } \partial \Delta \ln w_t / \partial e_t > 0.$$

A discussion of what I empirically refer to entrepreneurship in an MSA is warranted at this point. First, the terms *firm*, *establishment,* and *business* need clarification. As Neumark et al. (2011) point out, a *firm* is identified by a common owner and can own multiple *establishments*, and a *business* generally refers to either a *firm* or an *establishment*. A large firm opening a branch, e.g., Walmart opening a new branch in town, would show up as a new establishment in the data but we would not considered such expansion as entrepreneurship. An entrepreneur that starts a new business would appear as a new firm as well as a new establishment in the data. Hence, firm birth would be an ideal proxy. However, for firms, especially multi-establishment firms, the relation between geography and economic measures (employment, payroll) is more obscure, whereas for establishments, there is always a one to one

matching between location and employment (or payroll). Hence, a common proxy used to measure entrepreneurship over a fixed geography (MSA or county) is average establishment size over that geography (Glaeser et al. 2010, 2012). Since most entrepreneurship is associated with small businesses, average establishment size serves as a reasonable proxy for entrepreneurship and the establishment level data links economic activity of businesses to a location in a straightforward way. One concern could be that average establishment size could contain other information, i.e., the degree of competition in an area. A more direct measure of entrepreneurship, the birth of businesses, has also been used in the literature but as an outcome variable rather than a right hand side variable (Rosenthal and Strange 2003, Samila and Sorenson 2011). This paper will use birth of businesses in the metropolitan area, in additional to average establishment size, to proxy for entrepreneurship.

In practice, I run regressions following the model:

$$\Delta \ln Y_{i,1993-2002} = \beta \ln e_{i,1993} + \ln X_{i,1993} \cdot \gamma + \delta_d + \varepsilon_i \qquad (4)$$

for Metropolitan Statistical Areas (MSAs) in the United States for the years 1993 to 2002. I examine this ten year period primarily because the census definition of MSAs often change after each census cycle. By limiting my analysis to these years I am able to maintain a consistent geography for MSAs and examine the growth dynamics of cities in a consistent manner. *Y* denotes the dependent variable (employment, annual payroll, or wage) so that $\Delta \ln Y_{i,1993-2002}$ is the change in log employment or income between 1993 and 2002 for city *i*. Annual payroll includes all wages, salary, bonuses, and benefits paid to employees in the MSA. Wage is calculated as annual payroll divided by employment. $\ln e_{i,1993}$ is the log of entrepreneurship measured by business births or average establishment size in 1993. $\ln X_{i,1993}$ is the vector of log control variables, which include employment in 1993, median family income in 1990, population in 1990, percent college educated and above in 1990, and the housing price index in 1993. δ_d is the set of census division dummy variables. $\ln e_{i,1993}$ is the log entrepreneurship measured by the birth of new businesses for city *i* in 1993.

A fundamental difficulty in retrieving an unbiased estimate of β in equation (4) is the joint determination of urban entrepreneurial activity and urban economic growth. Cities with more growth potentials will likely see higher levels of entrepreneurial activity, which would render the estimate of β upward biased in equation (4). The challenge of generating a plausibly exogenous variation of entrepreneurship has hampered the development of the causal investigation of the impact of entrepreneurship on economic growth. I am not aware of any other paper that has attempted to examine this causal relationship other than Glaeser et al (2012).[2] This paper adds to this literature by using a different source of exogenous variation for urban entrepreneurial activity. I defer the discussion of my instrumental variable to the next section.

Entrepreneurial ability would encompass various facets ranging from one's knowledge of the business and legal environment, communication skills, personnel and time management, to leadership. Empirically, I will be examining entrepreneurial ability that is *relevant for economic growth*. Another question I am interested in is how entrepreneurial ability differs between privately financed and government-backed entrepreneurship. The rationale for government intervention in promoting entrepreneurship is market imperfection, that because of imperfect information concerning the ability of entrepreneurs and risk aversion in part from the lenders, the market is inefficiently allocating resources to entrepreneurs of differing abilities. Potential discrimination in the lending market is another argument for government intervention in small business lending. I do not examine the sources of market imperfection in this paper, but rather examine aggregate economic outcomes and based on such results infer the average entrepreneurial ability of government-backed entrepreneurs. In order to examine this margin, I differentially examine entrepreneurship financed by government-backed loans and entrepreneurship financed in the private market. In practice, I run regressions following the model:

---

2 Gilles Duranton has also presented on going work on a similar topic during the Journal of Regional Science lectures in the 2013 North Americal Regional Science Council Meetings.

$$\Delta \ln Y_{i,1993-2002} = \beta_1 \ln mrktent_{i,1993} + \beta_2 \ln govtent_{i,1993} + \ln X_{i,1993} \cdot \gamma + \delta_d + \varepsilon_i. \quad (5)$$

Whether government backed entrepreneurship will be on average lower or higher ability is not ex ante evident. There could be negative selection if the market correctly screens entrepreneurs, so that those who can start business only through government support are on average low ability and contribute less to growth. On the other hand, there could be positive selection, given that the application to get federally guaranteed loans is an arduous process. A potential entrepreneur has to navigate through the bureaucracy of the SBA and banks to secure a loan and may hence be an individual of high ability and contribute more to growth. Finally, in assessing government-backed entrepreneurship and privately financed entrepreneurship one would need to know whether government-backed entrepreneurship complements or crowds out privately financed entrepreneurship.

3. Data and Variables

To examine these questions, I construct a city level panel of MSAs in the United States from 1993 to 2002. The information on the births of establishments comes from the publicly available Statistics of U.S. Businesses (SUSB) Employment Change Data. Birth of establishments is stratified into three categories based on the *firm's size*, i.e., firms with 19 or less employees, 20-499 employees, and 500 employees or above. Any establishment births that appear in the 20-499 or 500 or above category are expansions by existing firms. For instance, an opening of a small establishment that is part of a large firm (e.g., a new Starbucks store) will appear in the 500 or above category. This paper does not consider expansion by large firms as entrepreneurship. Since a new firm starts with zero employee, all new firm creation appears only in the 19 or less category. New establishments created as an expansion by small firms (19 or less employees) are also included in this category. I denote this category *small business birth*. This birth measure will be my main proxy for entrepreneurship. The SUSB Employment Change Data also

provides the number of initial establishments for each MSA. The SUSB Annual Data provides static accounts of each MSA, including employment, number of establishments by the three size categories, and annual payroll which includes all forms of compensations, such as salaries, wages, benefits, and bonuses. The population data comes from information collected from the Census Bureau. I use the Federal Housing Finance Agency's House Price Index (HPI) to control for MSA level housing price. HPI is a measure of single-family house prices based on the average price change in repeat sales or refinancing of the same properties. Among the 329 MSAs in the 1993 to 2002 census data, I drop Anchorage, Honolulu, and MSAs that have missing information.[3] I eventually end up with a balanced panel of 316 MSAs. All analysis is performed on this set of metropolitan areas.

In order to generate MSA level government-backed entrepreneurship variables, I collect data on the universe of Small Business Administration loans approved between 1985 and 2012.[4] The data set contains a rich set of information including the loan amount, loan date, business location, lender, number of employees, and whether the loan was to a new business or existing business. I use this information to create MSA level aggregate variables. I identify each loan approval for a new business as an incidence of government-backed entrepreneurship. I then aggregate the count and approval amount of each incidence to generate MSA level entrepreneurship variables. Though the information provided in the data is quite comprehensive it does have some miscodes and missing information, particularly pertaining to the business location. I match the loan data to the MSA level census data based on the place name and zip code if available. The loans were first matched to a county and then linked to an MSA.[5]

---

3 MSAs not included in the sample are Anchorage, AK, Honolulu, HI, Cumberland, MD-WV, Enid, OK, Flagstaff, UT-AZ, Grand Junction, CO, Hattiesburg, MS, Jamestown, NY, Johnstown, PA, Jonesboro, AR, Missoula, MT, Pocatello, ID, Steubenville-Weirton, OH-WV.

4 I purchased this data from Coleman Publishing.

5 Some of the loan data had missing reports and miscodes. In the end I was able to match 93% of the data to a county, which were in turn matched to MSAs.

Table 1 Summary Statistics

Variable	Mean	Std.Dev.	Mix	Max
Change in log employment, 1993-2002	0.163	0.100	-0.261	0.550
Change in log annual payroll, 1993-2002	0.258	0.139	-0.169	0.752
Employment, 1993	252130	439654	20957	3495130
Annual payroll ($1,000), 1993	6553740	13300000	335607	123000000
Average establishment size, 1993	14.85	2.53	8.29	24.13
Employment of establishments with less than 20 employees, 1993.3	48003	79320.2	5317	644273
Employment of establishments with 20 to 499 employees, 1993.3	82163	144312.4	6868	1203297
Employment of establishments with more than 499 employees, 1993.3	121963	217507.3	6870	1666884
Number establishments with less than 20 employees, 1993.3	11856	20298.56	1234	180540
Number of establishments with 20 to 499 employees, 1993.3	2357	3774.05	245	31251
Number of establishments with more than 499 employees, 1993.3	1999	3107.24	213	22605
Birth of establishments by new firms or firms with less than 20 employees, 1992.3-1993.3	1387	2390.85	105	20602
Birth of establishments by new firms of firms with 20 to 499 employees, 1992.3-1993.3	119	201.29	6	1771
Birth of establishments by firms with more than 499 employees, 1992.3-1993.3	153	254.10	8	1866
Amount of SBA loans approved($1,000), FY1993	18400	30600	86	307000
Amount of SBA loans approved($1,000) for new businesses, FY1993	2809	4424	0	45700
Amount of SBA loans approved($1,000) for new businesses with less than 20 employees, FY1993	1823	2981	0	30500
Number of SBA loans approved, FY1993	68.6	101.97	1	879
Number of SBA loans approved for new businesses, FY1993	13.3	18.24	0	140
Number of SBA loans approved for new businesses with less than 20 employees, FY1993	11.7	16.44	0	130
Number of SBA lenders in 1985	4.7	6.55	0	54

Notes: Unit of analysis is the Metropolitan Statistical Area (MSA) and the number of MSAs in the data is 316.

The timing of birth variables warrants further explanation. The static variables in the SUSB data are for March or first quarter of each year. The birth variables count establishment births that occurred between

March of the previous year and March of the reference year. Initial establishment level is the number of establishments in March of the previous year. For example, birth of establishment number for 1993 is the number of establishment births that occurred between March 1992 and March 1993. The initial establishment number for 1993 is the number of establishments that existed as of March 1992. The SBA loan data follows a fiscal year. Hence, the number of SBA loans and the approved amount for 1993 are the aggregate values for all loans approved in FY1993, i.e., July 1992- June 1993.

Table 1 presents the summary statistics of the main variables used in the analysis. Employment growth during the ten year period is about 16 percent, which translates to an annualized growth rate of about 1.5 percent. The descriptive statistics indicate that small businesses are responsible for 73% of urban establishments but only 19% of urban employment. On average each metropolitan area saw a birth of 1387 small establishments where 13 of these were government-backed entrepreneurship. Small businesses accounted for 83.6% of all establishment births. Average establishment size in 1993 was about 15 employees.

4. The Impact of Entrepreneurship on Urban Growth

4.1. OLS Results

I begin the analysis by visually examining the relationship between entrepreneurship and urban employment growth. Figure 1 presents a scatterplot between the change in log MSA employment between 1993 and 2002 and the log small business birth in 1993. Figure 2 and 3 present a similar plot for MSA payroll and wage growth. A general upward sloping trend is observed. A higher share of small establishment birth is positively correlated with urban growth. I examine this relationship more formally in an econometric framework. Table 2 presents the OLS results as specified in equation (4), where the dependent variables are employment, payroll, or wage growth in 1993-2002. Table 3 presents corresponding results for the 5 years windows of

Table 2 Impact of Entrepreneurship on Urban Growth (10 year growth): OLS Estimates

	(1)	(2)	(3)	(4)	(5)
Panel A: Dependent variable:	*Change in log employment, 1993-2002*				
Log average establishment size in 1993	-0.183***				
	(0.0447)				
Log small business births in 1992-93		0.126***	0.157***	0.241***	0.267***
		(0.0262)	(0.0238)	(0.0311)	(0.0311)
Log establishment births by existing medium firms in 1992-93		0.0328		0.0264	
		(0.0212)		(0.0228)	
Log establishment births by existing large firms in 1992-93		0.0436**		0.0292	
		(0.0196)		(0.0223)	
Log employment in 1993	0.0453	-0.155***	-0.109***	-0.224***	-0.215***
	(0.0365)	(0.0311)	(0.0312)	(0.0587)	(0.0600)
Log median family income in 1990	0.0288	0.0252	0.0171	0.0907*	0.0890*
	(0.0517)	(0.0485)	(0.0486)	(0.0486)	(0.0492)
Log population in 1990	-0.0278	-0.0379	-0.0372	0.0130	0.0111
	(0.0368)	(0.0345)	(0.0351)	(0.0327)	(0.0332)
Percent college and above in 1990	0.00195*	0.00108	0.00127	0.000122	0.000257
	(0.00110)	(0.00110)	(0.00111)	(0.00103)	(0.00102)
Log housing price index 1993	0.0193	0.0238	0.0699	0.0438	0.0642
	(0.139)	(0.125)	(0.129)	(0.120)	(0.126)
Log establishments with less than 20 employees in 1992				-0.275***	-0.286***
				(0.0497)	(0.0512)
Log establishments with 20 to 499 employees in 1992				0.197***	0.202***
				(0.0604)	(0.0578)
Log establishments with 500 or above employees in 1992				-0.000338	0.0295
				(0.0434)	(0.0354)
Census division fixed effects	Y	Y	Y	Y	Y
R squared	0.379	0.446	0.43	0.504	0.498

Table 2 (continue)

	(1)	(2)	(3)	(4)	(5)
Panel B: Dependent variable:	*Change in log payroll, 1993-1998*				
Log average establishment size in 1993	-0.225***				
	(0.0581)				
Log small business births in 1992-93		0.176***	0.215***	0.317***	0.362***
		(0.0355)	(0.0316)	(0.0419)	(0.0391)
Log establishment births by existing medium firms in 1992-93		0.0560**		0.0546**	
		(0.0265)		(0.0269)	
Log establishment births by existing large firms in 1992-93		0.0378		0.0436	
		(0.0262)		(0.0282)	
Panel C: Dependent variable:	*Change in log wages 1997-2002*				
Log establishment per employee in 1993	-0.0420				
	(0.0271)				
Log small business births in 1992-93		0.0501***	0.0580***	0.0755***	0.0951***
		(0.0160)	(0.0149)	(0.0204)	(0.0185)
Log establishment births by existing medium firms in 1992-93		0.0232		0.0282**	
		(0.0149)		(0.0139)	
Log establishment births by existing large firms in 1992-93		-0.00584		0.0144	
		(0.0133)		(0.0151)	
Base controls	Y	Y	Y	Y	Y
Initial establishment controls				Y	Y
Census division fixed effects	Y	Y	Y	Y	Y

Notes: The unit of analysis is the MSA and the number of observations is 316. Establishment births for 1993 are counted between March 1992 and March 1993. Establishment per employee is the number of all establishments divided by the number of all employees in the MSA. The "small business births" variable includes all new firm creation and expansions by firms with less than 20 employees. The "establishment birth by existing medium firms" variable refers to expansion by firms with 20-499 employees. The "establishment birth by existing large firms" variable refers expansion by firms with over 500 employees. Base controls refer to the five control variables in Panel A Column (1). Initial establishment controls are the three number of establishment controls in Panel A. The nine census division dummies are included as controls. The dependent variable is the change in log total MSA employment between 1993 and 2002 in Panel A, the change in log total annual payroll, which includes all wages, salary, bonuses, and benefits between 1993 and 2002 in Panel B, and the change in wage, which is payroll divided by employment, in Panel C. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

Figure 1 Scatterplot of MSA employment growth (1993-2002) and small business births (1993)

Figure 2 Scatterplot of MSA payroll growth (1993-2002) and small business births (1993)

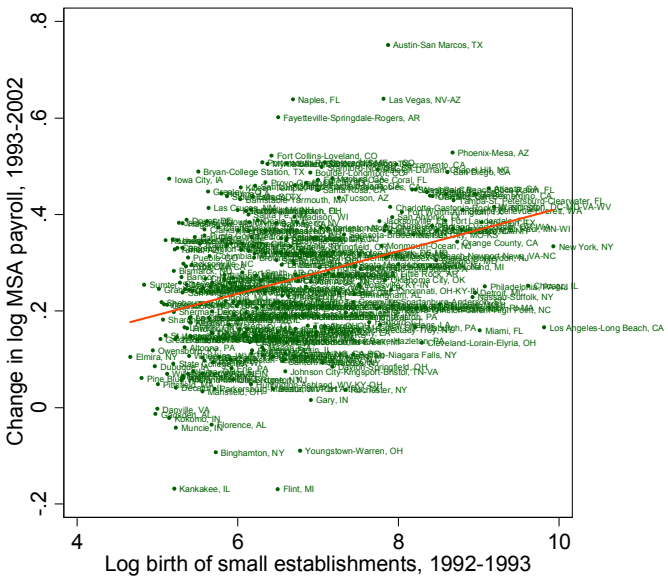
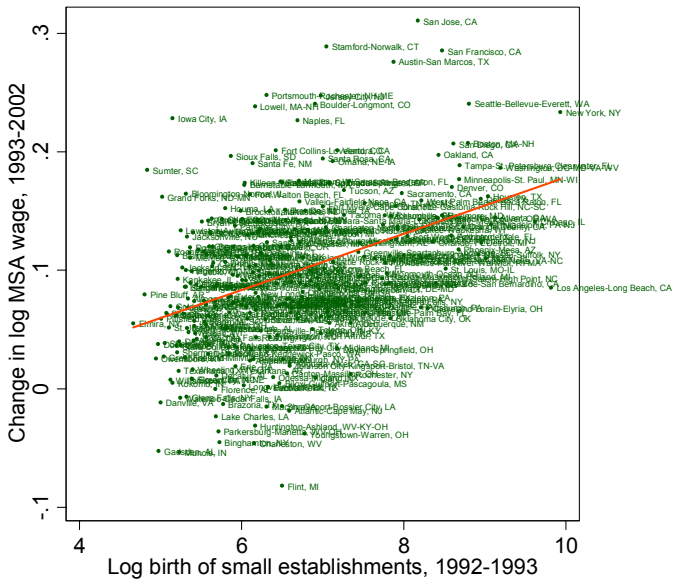


Figure 3 Scatterplot of MSA wage growth (1993-2002) and small business births (1993)



1993-1998 and 1997-2002. The main variable of interest is log small business birth, my main proxy for entrepreneurship. I also examine the impact of average establishment size, which has been used to proxy for entrepreneurship in the literature. I include the log establishment births by medium (20 to 499 employees) and large (500 or more employees) firms as controls. The birth of establishments by the larger firms represents an expansion of existing firms and does not capture any new firm birth. Panel A examines the growth of MSA employment, Panel B growth of total annual payroll, and Panel C growth of wage, which is payroll divided by employment. All specifications in Table 2 include initial employment, median family income, population, percent college educated and above, the house price index, and the nine census division dummies as base controls.

Column (1) first examines the average establishment size effect. A 10 percent decrease in average establishment size in 1993 is associated with a 1.8 percent higher employment, 2.3 percent higher payroll, and

0.4 percent higher wages after 10 years. The employment and payroll effects are statistically significant at one percent level. Cities with smaller establishments on average have higher economic growth. However, given that average establishment size can imply various aspects of an urban economy, I next examine my main proxy for entrepreneurship, small business births. Column (2) indicates that a 10 percent increase in small business birth is associated with a 1.3 percent higher employment, 1.8 percent higher payroll, and 0.58 percent higher wages after 10 years. All coefficient estimates are statistically significant at the 1 percent level. Not only is there employment growth, there is also productivity growth from entrepreneurship. The contribution of establishment births by the expansion of larger firms on employment growth is considerably smaller. The birth of small businesses contributes to urban growth at considerably higher degrees than establishment expansions by larger firms. When I focusing on the small business birth results in column (3), the coefficient estimates increase slightly for employment, payroll, and wages. In examining the impact of new firm births, the number of births relative to the total number of initial establishments could matter for growth. Also, there could be mean reversion in the number of establishments within MSAs. Hence, in columns (4) and (5) I control for the log number of establishments in 1992 for the three employee size categories. The coefficient estimates on small business births increase almost twofold. Focusing on column (5), a 10 percent increase in small business births is associated with a 2.7 percent higher employment and 3.6 percent higher payroll after 10 years. The larger coefficient estimates on entrepreneurship for payroll growth than that for employment growth imply that wage would increase with entrepreneurship. Panel C documents this pattern. A 10 percent increase in small business birth results in about 1 percent higher wages after 10 years. The coefficient estimates on the log initial number of small establishments and log employment in Panel A are negative and statistically significant. These estimates are consistent with mean reversion in employment and small establishments. However, cities that initially have a higher number of medium sized establishments see higher employment growth.

Table 3 Impact of Entrepreneurship on Urban Growth (5 year growth): OLS Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Dependent variable:	*Change in log employment, 1993-1998*			*Change in log payroll, 1993-1998*			*Change in log wage, 1993-1998*		
Log average establishment size in 1993	-0.0729**			-0.0612			0.0118		
	(0.0330)			(0.0436)			(0.0213)		
Log small business births in 1992-93		0.137***	0.164***		0.180***	0.218***		0.0429**	0.0539***
		(0.0219)	(0.0220)		(0.0297)	(0.0285)		(0.0185)	(0.0176)
Log establishment births by existing medium firms in 1992-93		0.0390***			0.0472***			0.00823	
		(0.0125)			(0.0156)			(0.00976)	
Log establishment births by existing large firms in 1992-93		0.0182			0.0337*			0.0155	
		(0.0154)			(0.0192)			(0.0113)	
Panel B: Dependent variable:	*Change in log employment, 1997-2002*			*Change in log payroll, 1997-2002*			*Change in log wage, 1997-2002*		
Log average establishment size in 1997	-0.142***			-0.194***			-0.0519***		
	(0.0309)			(0.0357)			(0.0194)		
Log small business births in 1996-97		0.216***	0.234***		0.285***	0.291***		0.0690***	0.0566***
		(0.0318)	(0.0292)		(0.0415)	(0.0378)		(0.0228)	(0.0213)
Log establishment births by existing medium firms in 1996-97		0.0373			0.0168			-0.0206	
		(0.0252)			(0.0323)			(0.0143)	
Log establishment births by existing large firms in 1996-97		-0.00205			-0.0102			-0.00817	
		(0.0153)			(0.0193)			(0.0101)	
Base controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Initial establishment controls		Y	Y		Y	Y		Y	Y
Census division fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y

Notes: The unit of analysis is the MSA and the number of observations is 316. Establishment births for year t are counted between March of year t-1 and March of year t. Establishment per employee is the number of all establishments divided by the number of all employees in the MSA. The "small business births" variable includes all new firm creation and expansions by firms with less than 20 employees. The "establishment birth by existing medium firms" variable refers to expansion by firms with 20-499 employees. The "establishment birth by existing large firms" variable refers expansion by firms with over 500 employees. Base controls refer to the five control variables in Table 2 Panel A Column (1). Initial establishment controls are the three number of establishment controls in Table 2 Panel A. The nine census division dummies are included as controls. Panel A examines the five year growth between 1993 and 1998. Panel B examines the five year growth between 1997 and 2002. The dependent variables are the change in log total MSA employment, the change in log total annual payroll, which includes all wages, salary, bonuses, and benefits, and the change in wage, which is payroll divided by employment. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

Table 3 examines five year economic growth. The coefficient estimates show similar pattern to Table 2, with the impact of entrepreneurship on 1997 to 2002 growth being larger than on 1993 to 1998 growth. Focusing on the small business birth results in column (3), a 10 percent increase in entrepreneurship is associated with an annualized employment growth rate of about 0.32 percent in Panel A, and 0.46 percent in Panel B. The 10 year growth results in Table 2 column (5) returns an annualized growth rate of about 0.26 percent. The fact that the annualized growth rates for the five year periods are higher than for the ten year period is consistent with faster growth of businesses when they are young as documented by Haltwinger et al (2013). The larger coefficient estimates on entrepreneurship for payroll growth than that for employment growth imply that wage would increase with entrepreneurship. Columns (7) to (9) document this pattern. A 10 percent increase in entrepreneurship is associated with 0.54 to 0.56 percent higher wages after five years.

Tables 2 and 3 depict an equilibrium relation rather than a causal impact of entrepreneurship on urban growth. Unobserved factors that increase a city's growth potential would increase urban entrepreneurial activity as well as actual growth. Such omitted variable would render the OLS coefficient estimates on entrepreneurship biased. To alleviate some of the concerns that arise in the cross-sectional analysis, I present first difference estimates in Table 4 based on the following model:

$$\Delta \ln Y_{i,1997-2002} - \Delta \ln Y_{i,1993-1998} = \beta \Delta \ln e_{i,1993-1997} + \Delta \ln X_{i,1993-1997} \cdot \gamma + \varepsilon_{i,1993-1997}. \quad (6)$$

This specification essentially takes the difference between the specifications in Table 3 Panels A and B and runs an OLS estimation. The first differencing would deal with unobserved constant MSA fixed effects, such as static metropolitan area growth potentials. However, first differencing a dynamic framework mechanically introduces endogeneity if the error terms are correlated over time, a very likely scenario. Hence, one should examine the Table 4 results with such caveat in mind.

The coefficient estimates are considerably smaller than those observed in Table 3. For instance, the coefficient estimate on employment growth in Table 4 column (3) is 0.12 compared to 0.16 and 0.23 in Table 3

Table 4 Impact of Entrepreneurship on Urban Economic Growth: First-difference Estimates

	(1)	(2)	(3)
Panel A: Dependent variable:	*Change in 5 year employment growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		
ΔLog average establishment size between 1993 and 1997	-0.406***		
	(0.117)		
ΔLog small business births between 1993 and 1997		0.116***	0.122***
		(0.0310)	(0.0325)
ΔLog establishment births by medium firms between 1993 and 1997		0.0359	
		(0.0221)	
ΔLog establishment births by large firms between 1993 and 1997		-0.00115	
		(0.0115)	
R-squared	0.576	0.598	0.585
Panel B: Dependent variable:	*Change in 5 year payroll growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		
ΔLog average establishment size between 1993 and 1997	-0.347**		
	(0.143)		
ΔLog small business births between 1993 and 1997		0.114***	0.115***
		(0.0421)	(0.0420)
ΔLog establishment births by medium firms between 1993 and 1997		0.0146	
		(0.0236)	
ΔLog establishment births by large firms between 1993 and 1997		-0.00761	
		(0.0163)	
R-squared	0.576	0.582	0.581
Panel C: Dependent variable:	*Change in 5 year wage growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		
ΔLog average establishment size between 1993 and 1997	0.0593		
	(0.0737)		
ΔLog small business births between 1993 and 1997		-0.00230	-0.00723
		(0.0232)	(0.0230)
ΔLog establishment births by medium firms between 1993 and 1997		-0.0214***	
		(0.00785)	
ΔLog establishment births by large firms between 1993 and 1997		-0.00647	
		(0.00838)	
Base controls	Y	Y	Y
Initial establishment controls	Y	Y	Y
R-squared	0.552	0.571	0.559

Notes: The unit of analysis is the MSA and the number of observations is 316. Establishment births for year t are counted between March of year t-1 and March of year t. Establishment per employee is the number of all establishments divided by the number of all employees in the MSA. The "small business births" variable includes all new firm creation and expansions by firms with less than 20 employees. The "establishment birth by existing medium firms" variable refers to expansion by firms with 20-499 employees. The "establishment birth by existing large firms" variable refers expansion by firms with over 500 employees. Base controls include the change in log employment, payroll, population, and house price index, and the 1990 percent college educated and log median family income. Initial establishment controls are the change in the three establishment number variables. The dependent variable is the change in five year employment growth in Panel A, the change in five year annual payroll growth in Panel B, and the change in wage growth in Panel C. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

column (3). Similarly the coefficient estimates are smaller for payroll growth and wage growth. Dealing with unobserved MSA level static growth potential by first differencing seems to have mitigated the omitted variable bias in the cross sectional analyses of Tables 2 and 3.

I also separately examine the impact of firm expansion by existing medium and large firms on urban economic growth in Appendix Table 1. I run regressions where log establishment births by medium or large firms in 1993 are the main covariate of interest. As in previous tables, I run both OLS regressions and first-differenced regressions. The coefficient estimates on firm expansion are nearly three folds smaller than the estimates on small business births. Furthermore, the coefficient estimates in the first-differenced regression are no longer statistically different from zero for employment and payroll, regardless of firm size. Though new business birth and existing firm expansion are correlated within cities, new business birth is driving the five years and ten years urban economic growth results.

4.2. Homestead Exemption Levels as Instrumental Variables and 2SLS Results

If there are unobserved time varying MSA level growth potentials that are correlated with entrepreneurship, then dealing with MSA fixed effects will not be sufficient for obtaining unbiased estimates. For example, if potential entrepreneurs perceive that in 1993 that a city will be increasingly favorable for growth and start businesses then the endogeneity concern remains. To deal with these potential problems, I also estimate the impact of entrepreneurship on urban growth using the homestead exemption levels in 1975 as instrumental variables. When a non-incorporated business is no longer financially viable, the debt of the business becomes personal liability of the business owner and he or she can file for personal bankruptcy.[6] However, in these unfortunate

6 Over 70% of small businesses are sole proprietors. Partnerships are also unincorporated and hence are eligible for personal bankruptcy procedures. Limited liability companies and corporations limit the financial liability of the owner or shareholder. http://www.sba.gov/community/blogs/top-10-questions-about-small-business-incorporation-answered

instances property exemption laws in the US have protected a part of the debtor's assets. Such property exemption has existed in the US since 1845 when Texas became a US state, and by 1898 people could file for bankruptcy under a federal bankruptcy law and receive protection

Table 5 Homestead Exemption in 1975 and Year Interstate Banking was Permitted by State

State	Homestead exemption level in 1975	Year of interstate banking deregulation	State	Homestead exemption level in 1975	Year of interstate banking deregulation
AK	19,000	1987	MT	40,000	1993
AL	4,000	1982	NC	2,000	1990
AR	U	1986	ND	80,000	1985
AZ	15,000	1989	NE	8,000	1987
CA	20,000	1987	NH	5,000	1986
CO	15,000	1988	NJ	0	1989
CT	0	1983	NM	20,000	1982
DE	0	1988	NV	25,000	1985
DC	N/A	1985	NY	4,000	1991
FL	U	1985	OH	0	1985
GA	1,000	1985	OK	U	1987
HI	50,000	1995	OR	12,000	1986
IA	U	1985	PA	0	1986
ID	14,000	1986	RI	0	1984
IL	10,000	1986	SC	2,000	1986
IN	1,400	1991	SD	U	1988
KS	U	1992	TN	7,500	1985
KY	2,000	1984	TX	U	1987
LA	15,000	1987	UT	11,000	1984
MA	24,000	1978	VA	10,000	1988
MD	0	1985	VT	10,000	1985
ME	6,000	1983	WA	20,000	1987
MI	7,000	1986	WI	25,000	1988
MN	U	1986	WV	0	1987
MO	2,000	1988	WY	20,000	1987
MS	30,000	1986			

Notes: Exemption amounts are nominal and were collected from Posner et al.(2001). U dentotes unlimited exemption. Exemption amount was not available for DC. Year of interstate branching collected from the St. Louis Fed publication at www.stlouisfed.org/publications/re/2007/b/pdf/dereg.pdf.

according to each state's homestead exemption level (Posner et al. 2001). Homestead exemption protects ownership on real property, such as house or land, up to the specified level. If an entrepreneur owns $50,000 equity in a house and files for bankruptcy in a state where the homestead exemption level is $20,000, the entrepreneur would keep $20,000 and the rest would go to the (unsecured) creditors.

As Table 5 indicates the homestead exemption levels in 1975 were set by each state and vary significantly across states. The exemption levels ranged from zero in Connecticut, Delaware, Maryland, New jersey, Ohio, Pennsylvania, Rhode Island, and West Virginia to unlimited in Arkansas, Florida, Iowa, Kansas, Minnesota, Oklahoma, South Dakota and Texas. An entrepreneur filing for bankruptcy in Iowa could keep his or her home and land in entirety, where as one in Ohio would have lost his house if debt was greater than equity in his house. Given that there are unlimited exemption levels, I cannot simply use the continuous exemption level as the instrumental variable. Hence, I first construct two state exemption level variables: UN_s, a dummy equal to one if the state has unlimited exemption and equal to zero if the state has limited or no exemption, and EX_s, the state exemption level. EX_s is set to zero for states with unlimited exemption. For MSAs not contained entirely within one state, I average each variable across the states each MSA overlaps with. Hence, the final set of MSA level instrumental variables are:

$$UN_i = \frac{1}{N[s \in i\}} \sum_{s \in i} UN_s,\ lnEX_i = \log(\frac{1}{N[s \in i\}} \sum_{s \in i} EX_s + 1). \qquad (7)$$

Where i indexes for MSAs and s for states. Two conditions are needed for the above set of homestead exemption level variables to serve as a valid instrument for entrepreneurship in equation (4). The first is that exemption levels need to impact entrepreneurship. The literature provides direct evidence on this relationship. Fan and White (2003) discuss how higher exemption levels serve as a wealth insurance and induce risk averse potential entrepreneurs to start a business. They empirically confirm this using household level data. I will find strong evidence of this correlation at the aggregate level in my data as well.

The second condition, that conditional on city economic conditions in 1993, the 1975 homestead exemption level impacts 1993-2002 urban growth only through its impact on entrepreneurship warrants further understanding of the variance in exemption levels across states. What explains the astonishingly wide variance in exemption levels? As Posner et al. (2001) points out, hypotheses relating the difference in the demand for insurance, or in altruism are unlikely to explain such wide variance. They examine the cross sectional variation in homestead exemption level in a regression framework by including multiple variables, such as income, charitable giving, population density, farm proprietors share, and find that only the historical exemption levels in 1920 predict current exemption levels. Their argument that (1) initially sparsely populated states in the 1800s set high homestead exemption levels to compete for migrants and that (2) whenever state lawmakers would negotiate the exemption level the bargaining point would be the then current levels provides a convincing explanation of the persistent variation of exemption level across states. The assumption for instrument exogeneity holds if unobserved MSA level static and dynamic growth potential between 1993-2002, controlling for 1993 economic conditions and entrepreneurship, is not correlated with the homestead exemption levels in 1975.

Table 6 presents the instrumental variable regression results. The estimation in practice is identical to equation (4) where the entrepreneurship variable is instrumented with the homestead exemption variables in equation (7). Regressions that examine the impact of average establishment size include the base controls and the Census division dummies. Specifications that use small business births as the main proxy for entrepreneurship additionally control for the number of small, medium, and large establishments in 1992. Panel A presents the first stage of the 2SLS estimation. Column (1) examines the impact of the unlimited exemption variable on the average establishment size variable. Average establishment sizes is about 2 percent smaller in metropolitan areas with unlimited exemption versus not. Column (2) examines the impact of the unlimited exemption variable on small business births. Small business births are eleven percent higher in metropolitan areas with unlimited exemption versus not. Columns (3)

Table 6 | Impact of Entrepreneurship on Urban Economic Growth: 2SLS Estimates

	(1)	(2)	(3)	(4)
Panel A - 1st Stage Dependent variable:	*Logaverage establishment size in 1993*	*Logaverage establishment size in 1993*	*Log small business births in 1992-1993*	*Log small business births in 1992-1993*
Log homestead exemption level in 1975		-0.00346*		0.00834**
		(0.00179)		(0.00336)
Unlimited exemption in 1975	-0.0199***	-0.00776	0.111***	0.0819***
	(0.00612)	(0.00658)	(0.0264)	(0.0281)
Base controls	Y	Y	Y	Y
Initial establishment size controls			Y	Y
Census division fixed effects	Y	Y	Y	Y
R squared	0.932	0.934	0.985	0.986
Panel B - 2SLS : Dependent variable:	*Change in log employment, 1993-2002*			
Log average establishment size in 1993	-0.267*		-0.263*	
	(0.146)		(0.147)	
Log small business births in 1992-93		0.218*		0.106
		(0.119)		(0.106)
Hansen J-statistic p-value			0.4351	0.1385
Panel C - 2SLS : Dependent variable:	*Change in log payroll, 1993-2002*			
Log average establishment size in 1993	-0.489**		-0.493**	
	(0.210)		(0.210)	
Log small business births in 1992-93		0.399**		0.309**
		(0.170)		(0.146)

Table 6 (Continue)

	(1)	(2)	(3)	(4)
Hansen J-statistic p-value			0.7141	0.3754
Panel D - 2SLS : Dependent variable:	*Change in log wage, 1993-2002*			
Log average establishment size in 1993				
Log small business births in 1992-93				
Hansen J-statistic p-value			0.0629	0.6617
Instrumental variables:	*Unlimited exemption in 1975*		*Unlimited exemption in 1975, Log homestead exemption level,*	
1st stage F-statistic	15.44	17.75	7.76	11.77
Base controls	Y	Y	Y	Y
Initial establishment controls		Y		Y
	Y	Y	Y	Y

Notes: Panel A presents the first stage of the 2SLS regression and Panel B present the 2SLS estimates. The unit of analysis is the MSA and the number of observations is 316. Small business births for 1993 are counted between March 1992 and March 1993. Establishment per employee is the number of all establishments divided by the number of all employees in the MSA. The "small business births" variable includes all new firm creation and expansions by firms with less than 20 employees. Base controls are initial employment, median family income, population, percent college degree and above, and the house price index. Initial establishment controls are the three log number of establishment variables. The nine census division dummies are included as controls. The Kleibergen-Paap rk Wald F statistics are reported as the 1st stage F-statistics.* p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

and (4) add the continuous log exemption level variable. The coefficient estimates on both instrumental variables are negative but the statistical significance weakens quite a bit in column (3). In column (4) the coefficient estimates on both instruments are positive and statistically significant at the 5 percent level. A doubling of the exemption level increases small establishment birth by 0.6%. Overall, Panel A results indicate that (1) higher exemption levels increase entrepreneurship, (2) the unlimited exemption variable is a stronger instrument, and (3) that the instruments work better for small business births than average establishment size. The F-statistics at the bottom of Table 6 reflect this. The F-statistics are strong and above 10 in columns (1) and (2), are smaller in columns (3) and (4), and is less than 10 in column (3).[7]

Table 6 Panels B through D present the 2SLS results on employment, payroll, and wage growth using the homestead exemption variables as instruments. Columns (1) and (2) use only the unlimited exemption variable as the instrument and Columns (3) and (4) use both variables in the instrument set. First focusing on specifications that use the small business births variable, 10% more small business birth in 1993 leads to 1.1~2.2% more employment, 3.1~4% higher total annual payroll, and 1.8~2% higher wages after 10 years. The 2SLS estimates for employment growth are smaller in magnitude relative to the OLS estimate in Table 2 indicating that the instrumental variable estimates substantially corrected for potential omitted variables in employment growth. The 2SLS estimate on payroll growth decreases relative to the OLS estimate when both instruments are used but is actually larger when only one instrument is used.

However, the 2SLS estimates on average establishment size are larger in magnitude compared to the OLS estimates. A 10 percent decrease in average establishment size increases employment by about 2.6 percent, annual payroll by 4.9%, and wage by 2.2~2.3% after ten

7 Appendix Table 2 examines the impact of the homestead exemption variables on expansions by medium and large firms. Given that small business births and existing firm expansion are correlated within cities, I do find positive correlation between the instrumental variables and firm expansion. However, once I control form small business birth the impact of the homestead exemption variables on firm expansion goes away.

years. The coefficient estimates are quite robust regardless of whether I use one or both instruments. The finding that the 2SLS estimates change relative to the OLS estimates in opposite directions depending on which entrepreneurship variable I use is actually intuitive. The main omitted variable, unobserved MSA growth potential, will likely be positively correlated with small business births and thus be negatively correlated with average establishment size.[8]

Note that the 2SLS estimates implicitly assume that the variation in the homestead exemption levels impacts the number of births but not the average entrepreneurial ability in each MSA. However, it is unlikely to be the case. Consider a distribution of entrepreneurial ability in a city. If homestead exemption serves as a wealth insurance as in Fan and White (2003), cities with higher exemption will see more new businesses. Depending on whether the marginal entrepreneur's entrepreneurial ability is greater or lower than the existing average entrepreneurial ability in the city, the 2SLS estimate on the number of entrepreneurship may over or understate the true impact. If higher homestead exemption renders the marginal entrepreneur to be of lower ability than the average, the 2SLS estimates we get in Table 6 is likely a lower bound. On the other hand, if higher homestead exemption renders the marginal entrepreneur to be of higher ability than the average, the 2SLS estimates we get in Table 6 are likely to be larger than the true impact.[9] I do not have data to test which situation is likely to be the case. However, if we assume a model where the decision to become an entrepreneur is non-decreasing in wealth and entrepreneurial ability, and that the additional wealth insurance from higher homestead exemption levels mostly impacts the contribution of wealth on start-up decision, then the

8 Glaeser et al. (2013) examine the impact of average establishment size on 1982-2002 employment growth using distance to mines and the quantity of mineral deposits as instrumental variables. They find coefficient estimates that range from -0.87 to -0.96.

9 Note that this argument assumes a closed city or that all cities are identical. If entrepreneurs of different ability sort across cities to take advantage of higher homestead exemption, one would need to consider whether there is positive or negative selection across cities as well. I abstract away from this discussion. However, there is evidence that entrepreneurs disproportionately start their businesses in their hometowns (Michelacci and Silva, 2007).

marginal entrepreneur's ability would be lower than the average.[10] This would imply that the 2SLS estimates in Table 6 are lower bounds.

4.3 Sensitivity Analysis

A geographic pattern is noticeable from the homestead exemption levels in Table 4. The southern and western states tend to have higher exemption levels than the northeastern states. A natural concern is whether this variation is related to state business environment that could impact both state bankruptcy laws and entrepreneurship levels. In Table 7, I test whether such concern might be valid by adding variables that proxy for state business environment to the 2SLS regressions. I use the specification that uses the share of MSA with unlimited exemption level in 1975 as the instrument, since including both instruments weakens the first stage F-statistic. Appendix Table 3 presents the results when I use both instrumental variables. Panel A adds the minimum wage for each MSA. For MSAs that cross state borders, I use a population weighted average. The minimum wage level could potentially impact an entrepreneur's decision to start a business. However, the first-stage F-

10 Suppose a potential entrepreneur's decision to start a business depends on the individual's wealth *w* and entrepreneurial ability *a*. Further assume that wealth *w* and entrepreneurial ability *a* are uniformly distributed across a two-dimensional space. I assume that the decision to become an entrepreneur is non-decreasing in wealth *w* and entrepreneurial ability *a*. Wealth captures both collateral used to start a business, as well as risk preference, so that higher *w* will imply a higher propensity to start a business. Higher entrepreneurial ability will also imply a higher propensity to start a business. Given *w* and *a* there will be an expected payoff for entrepreneurship and working for others. If the expected payoff of entrepreneurship is greater than the wage earnings, one will start a business. In other words, one can think of a simple decision rule that can be expressed as below:

$$D_{entrepreneur} = 1, \quad if\ \tau w + \varphi a \geq c$$
$$D_{entrepreneur} = 0, \quad if\ \tau w + \varphi a < c$$

for some parameters τ and φ and cutoff c. $D_{entrepreneur}$ equals one for an entrepreneur and zero if one works for another. Depending on how higher exemption level might impact the relative importance of the two factors, i.e., the ratio τ/φ, the average ability of observed entrepreneurs in the metropolitan area will differ. If higher exemption serves as a wealth insurance and increases the relative importance of wealth, i.e., τ/φ increases, then average ability *E(a)* in the city will decrease.

statistics and the coefficient estimates on small business births are virtually unchanged. Panel B controls for state Right-to-work laws. I add a variable that captures the population share of MSA subject to Right-to-work laws. Similarly, the results do not change.

Cities that had were growing in the late 1980s or early 1990s might have had more entrepreneurial activity and had higher homestead exemption levels. In Panel C, I control for population patterns by additional including log population in 1985 as a control. The coefficient estimates on entrepreneurship barely changes, though the standard errors become slightly larger in columns (1) and (2). Panel D controls for industry composition by including the employment share in manufacturing, retail, and services. The coefficient estimates show similar patterns as before but standard errors are larger.

Lastly, before examining the agglomeration benefits of entrepreneurship, I examine other mechanisms that might explain the growth impact of entrepreneurship. One is the mechanical channel. If there is a constant subset of entrepreneurs that survive and grow, then more entrepreneurship across cities would imply higher number of surviving entrepreneurs and establishments down the road. Panel E columns (1) and (2) examines how small business births in 1993 impact total number of establishments in 2002. Ideally, I would directly examine surviving businesses but I do not have that data. The impact is strong and significant. A 10 percent increase in small business birth results in 3.6% more establishments 10 years later. Given that there were 16,212 establishments and 1,387 small business births in 1993, this result potentially suggests a large externality benefit of entrepreneurship to other firm creation. Another channel relates to the idea of creative destruction. Creative destruction suggests that new entrepreneurship generates growth by promoting the obsolete firms to exit. In Panel E columns (3) and (4), I examine how small business birth impacts establishment death the next period. I find a statistically significant birth to death elasticity of 0.63. A 10 percent increase in entrepreneurship generates a 6.3% increase in establishment death.

4.4. The Agglomeration Benefits of Entrepreneurship

The OLS, first difference, and instrumental variable estimates all indicate that entrepreneurship contributes to urban growth. In this section, I examine whether the growth impact of entrepreneurship is simply due to the growth in the newly created businesses or whether there is agglomeration benefit, i.e., growth associated with other firms in the economy. A 10 percent increase in small establishment birth in 1993 translates to about 139 more births at the mean. Using the preferred 2SLS estimates this will generate about 1.1 to 2.2% more employment ten years later, which amounts to 2,773 to 5,546 more jobs. The Bureau of Labor Statistics reports that about a third of new establishments survive after 10 years.[11] If I assume all of the employment increase came from the new businesses created in 1993 it would imply that on average each surviving business increased employment by 60 to 120. Unfortunately, I could not find information on the average growth of new businesses that survive after 10 years and hence cannot make a direct comparison. However, in the 1992-1993 period, there were 564,504 firm births in the less than 20 employee category, which in aggregate created 3,438,106 employment in the U.S. This returns on average 6.1 employees per new small business created in 1993. If the average new business that survives after ten years is unlikely to grow from 6.1 employees to 54 to 78 employees, the results here imply substantial agglomeration benefits from entrepreneurship.

Examining payroll growth provides a clearer picture of the agglomeration benefits of entrepreneurship. A 10 percent increase in entrepreneurship causes 3.1 to 4.0% higher annual payroll after 10 years, which translates to $203,166,000 to $262,149,600 in 1993 dollars. If this increase were distributed solely to the newly created employment (using the average of 4,160) each employee would get an annual pay of $48,838 to $63,016 in 1993 dollars. If we use the lower bound estimates for both employment and income growth each employee would get an annual pay of $73,265 in 1993 dollars. Given that the average pay for employees working in small establishments in 2002 was $30,004

11 http://www.sba.gov/advocacy/7495/29581

Table 7 | Robustness Tests

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Control for minimum wage	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log average establishment size in 1993	-0.265*		-0.486**		-0.221**	
	(0.146)		(0.210)		(0.112)	
Log small business births in 1992-93		0.217*		0.396**		0.179*
		(0.119)		(0.169)		(0.0955)
1st stage F-statistic	15.7	17.86	15.7	17.86	15.7	17.86
Panel B: Control for Right-to-work	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log average establishment size in 1993	-0.227		-0.442**		-0.215**	
	(0.142)		(0.201)		(0.107)	
Log small business births in 1992-93		0.198		0.391**		0.193*
		(0.131)		(0.185)		(0.103)
1st stage F-statistic	16.14	15.65	16.14	15.65	16.14	15.65
Panel C: Control for past population	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log average establishment size in 1993	-0.248		-0.463**		-0.216*	
	(0.152)		(0.220)		(0.126)	
Log small business births in 1992-93		0.207		0.403**		0.196*
		(0.140)		(0.197)		(0.115)
1st stage F-statistic	12.13	11.99	12.13	11.99	12.13	11.99

Table 7 (Continue)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel D: Control for industry composition	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log average establishment size in 1993	-0.158		-0.459		-0.301	
	(0.288)		(0.409)		(0.225)	
Log small business births in 1992-93		0.179		0.364		0.184
		(0.171)		(0.233)		(0.131)
1st stage F-statistic	6.71	9.42	6.71	9.42	6.71	9.42
Panel E: Other outcome variables:	*Log total establishment in 2002*		*Log establishment death in 1994*			
Log average establishment size in 1993	-1.222***		-1.866***			
	(0.131)		(0.198)			
Log small business births in 1992-93		0.369***		0.557***		
		(0.115)		(0.133)		
1st stage F-statistic	15.44	17.75	15.44	17.75		
Base controls	Y	Y	Y	Y	Y	Y
Initial establishment controls		Y		Y		Y
Census division fixed effects	Y	Y	Y	Y	Y	Y

Notes: All results are 2SLS estimates using the unlimited exemption in 1975 variable as the instrument. The unit of analysis is the MSA. Each panel adds additional controls to the specifications in Table 6 Panel B Columns (1) and (3). The MSA average minimum wage is added in Panel A, the Right-to-work status in Panel B, log population in 1985 in Panel C, and the employment shares of manufacturing, service, and retail in Panel D. The number of observations is 316 except for Panel D which is 306. Base controls are initial employment, median family income, population, percent college degree and above, and the house price index. Initial establishment controls are the three log number of establishment variables. The nine census division dummies are included as controls. The Kleibergen-Paap rk Wald F statistics are reported as the 1st stage F-statistics.* p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

($617,583,597,000/20,583,371 employees) in 2002 dollars or $24,100 in 1993 dollars, there seems to be substantial spill over effects of entrepreneurship to other firms in the economy. This simple accounting exercise suggests that there is agglomeration benefit of entrepreneurship, in addition to the creative destruction and mechanical growth channels discussed in Table 7.

5. The Impact of Government-backed Entrepreneurship on Urban Growth

5.1. Background on Small Business Loans

Given the finding that entrepreneurship contributes significantly to urban economic growth, I next ask how the federal government's effort to promote entrepreneurship performs in regards to economic growth. The US government established the Small Business Association (SBA) in 1953 to promote the creation and expansion of small businesses and has since served as the advocacy agency, provided guidance, and financially supported small businesses. The fact that there is government intervention implicitly implies that there is market failure in the small business loan market, i.e., capable potential entrepreneurs are unable to start or expand a business because of imperfect information, missing insurance markets, or discrimination. Commercial lenders are unwilling to lend to potential entrepreneurs without sufficient collateral, may not be able to properly assess the feasibility of businesses, or may discriminate against female or minority entrepreneurs. Because of such likely market imperfections, the SBA promotes entrepreneurship by guaranteeing loans provided through commercial lenders and taking over the debt in case the debtor defaults.

The SBA's main form of guaranteed lending is the Small Business Loan, also known as the 7(a) loan program.[12] The Small Business Loan

12 There also is the Certified Development Company Loan, also known as the 504 loan program. The Certified Development Company (CDC) loan provides financing for fixed assets, such as, land, buildings, or machines, through a certified development company. A certified development company is a non-profit corporation set up to

Table 8 Impact of Government-backed Entrepreneurship on Urban Economic Growth: OLS and First-difference Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Dependent variable:	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log number of SBA loans approved for new businesses, FY1993	0.0210**	0.0117	0.0235*	0.00867	0.00250	-0.00300
	(0.0104)	(0.00740)	(0.0141)	(0.00966)	(0.00613)	(0.00436)
Log amount of SBA loans approved for new businesses, FY1993	-0.00281		-0.00448*		-0.00167	
	(0.00189)		(0.00250)		(0.00119)	
Base controls	Y	Y	Y	Y	Y	Y
Initial establishment controls	Y	Y	Y	Y	Y	Y
Census division fixed effects	Y	Y	Y	Y	Y	Y
R squared	0.374	0.37	0.416	0.411	0.409	0.406

Table 8 (Continue)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Dependent variable:	*Change in 5 year employment growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		*Change in 5 year payroll growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		*Change in 5 year wage growth, (1997 to 2002 growth) - (1993 to 1998 growth)*	
ΔLog number of SBA loans approved for new businesses, 1993-97	0.00252	0.00123	0.00532	0.00552	0.00280	0.00429
	(0.00518)	(0.00447)	(0.00747)	(0.00632)	(0.00385)	(0.00323)
ΔLog amount of SBA loans approved for new businesses, 1993-97	-0.000527		8.23e-05		0.000609	
	(0.00102)		(0.00134)		(0.000813)	
Base controls	Y	Y	Y	Y	Y	Y
Initial establishment controls	Y	Y	Y	Y	Y	Y
R squared	0.568	0.568	0.572	0.572	0.562	0.562

Notes: The unit of analysis is the MSA and the number of observations is 316. The number of new SBA loans approved and the total amount approved between July 1992 and June 1993 in each MSA are proxies for government-backed entrepreneurship. Panel A reports the OLS estimates for the employment, payroll, and wage growth regressions. Panel B reports the First-difference estimates. In Panel A, base controls are initial employment, median family income, population, percent college degree and above, and the house price index. Initial establishment controls are the three log number of establishment variables. The nine census division dummies are included as controls. In Panel B, base controls include the change in log employment, payroll, population, and house price index, and the 1990 percent college educated and log median family income. Initial establishment controls are the change in the three establishment number variables.* p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

(SBL) is based on Section 7(a) of the Small Business Act and is provided by commercial lenders that structure loans according to SBA's guidelines and receive a guarantee from the SBA. The SBA usually guarantees up to 85% of the loan. The commercial lender is in charge of the process and the loan applicant must meet the commercial lender's criteria. The applicant and the commercial lender negotiate the loan term subject to the SBA requirements and the applicant must meet the SBA's firm size requirements and be for-profit. The purpose of this study is not to assess whether there is market failure in the small business lending market but to examine whether entrepreneurship supported by the SBA differ from market entrepreneurship in its contribution to urban economic growth. Ex ante, it is difficult to assess whether there is positive selection or negative selection in SBA supported entrepreneurship. If the SBA guarantee draws in entrepreneurs that were not only credit constrained but also of lower entrepreneurial ability, there could be negative selection into government-backed entrepreneurship. If high ability entrepreneurs were shun from the commercial lending, SBA guaranteed lending could create positive selection. Also, the complexity and the bureaucracy associated with the application process itself could generate positive selection. Hence, this is a question that needs to be assessed empirically. The variables used to measure SBA guaranteed entrepreneurship in an MSA are (1) the number of SBA loans approved to new businesses, and (2) the total dollar amount of SBA loans approved to new businesses. Descriptive statistics of these variables appear in Table 1.

promote local economic development with several hundred locations nationwide. An important difference is that the CDC is only available to existing small businesses that plan to expand its business and cannot be used to start a new business and hence is not subject of interest in this study. The loan portfolio is such that typically the applicant contributes 10% of the total cost, the commercial lender 50%, and the CDC 40% which is fully guaranteed by the SBA.

5.2. The Impact of Government Backed Entrepreneurship on Urban Economic Growth

Table 8 Panel A reports the OLS results. Estimation is based on equation (4) where the entrepreneurship variables are replaced by the SBA loan variables. All specifications include the initial year controls and the census division dummies. The cross-sectional analysis on employment in columns (1) and (2) indicates that more government backed entrepreneurship measured by the number of loans approved to new businesses results in higher employment growth. However, the approved dollar amount has no significant impact on employment growth with coefficient estimates that are negative.[13] Getting more entrepreneurs started seems to be more important for growth than giving out larger loans. When loan amount is not controlled for in column (2) the coefficient estimate on the number of loans is smaller and no longer statistically significant at the 5% level. The annual payroll results in columns (3) and (4) are statistically weaker in general and the negative impact of total loan amount is more pronounced in column (3). Columns (5) and (6)indicate that more SBA loans are not associated with any wage growth, which is in contrast with previous results showing that small business births increase wage growth.[13]

The cross-sectional analysis likely suffers from endogenous SBA loan application and approval that relates to unobserved city characteristics. Table 8 Panel B presents first difference estimates, which controls for the MSA fixed effect at the cost of introducing the potential for endogeneity through correlated error terms. All estimates are no longer statistically significantly different from zero at standard levels. The OLS and first-difference results suggest that a larger number of SBA loans were approved in cities that were growing, but a larger amount of SBA loans were approved in cities that were declining.

Table 9 further examines the impact of government guaranteed entrepreneurship using instrumental variables. I focus on the impact of

13 Samila and Sorenson (2011) also find that the number of firms receiving loans matter for growth but not the total amount when examining the impact of venture capital. The number of entrepreneurship seems to be driving force of growth and getting entrepreneurs off the ground is more important than giving out big loans.

Table 9 Impact Impact of Market versus Government-backed Entrepreneurship on Urban Economic Growth: OLS and 2SLS Estimates

	OLS					2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable:	*Change in log (1993-2002) employment*	*payroll*	*wage*	*Log number of SBA loans to new businesses*	*Log market entrepreneur-ship*	*Change in log (1993-2002) employment*	*payroll*	*wage*
Log number of SBA loans to new businesses	0.00563	0.000405	-0.00523			-0.0179	-0.0142	0.00376
	(0.00626)	(0.00830)	(0.00429)			(0.0227)	(0.0295)	(0.0135)
Log market entrepreneurship	0.262***	0.358***	0.0964***			0.185*	0.390***	0.205***
	(0.0306)	(0.0394)	(0.0190)			(0.0985)	(0.138)	(0.0744)
Log number of SBA lender per capita in 1985				0.275***	-0.0131			
				(0.0622)	(0.0145)			
Log years since interstate banking deregulation				-0.291**	0.0611*			
				(0.126)	(0.0345)			
Log homestead exemption level in 1975				0.0445***	0.00669*			
				(0.0137)	(0.00360)			
Unlimited exemption in 1975				-0.309**	0.0945***			
				(0.156)	(0.0273)			
Base controls	Y	Y	Y	Y	Y	Y	Y	Y
Initial establishment controls	Y	Y	Y	Y	Y	Y	Y	Y
Census division fixed effects	Y	Y	Y	Y	Y	Y	Y	Y
1st stage F-statistic						6.0	6.0	6.0
Hansen J-statistic p-value						0.13	0.42	0.77
R squared	0.498	0.535	0.451	0.699	0.986			

Notes: The unit of analysis is the MSA and the number of observations is 316. Columns (1) through (3) are OLS estimates, columns (4) and (5) are first stage estimates of the 2SLS estimation in columns (6) through (8). The number of new SBA loans approved between July 1992 and June 1993 in each MSA proxy for government-backed entrepreneurship. Market entrepreneurship is defined as total small business birth minus the number of new SBA loans. Base controls are initial employment, median family income, population, percent college degree and above, and the house price index. Initial establishment controls are the three log number of establishment variables. The nine census division dummies are included as controls. The Kleibergen-Paap rk Wald F statistics are reported as the 1st stage F-statistics. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

the number of SBA loans approved to new businesses in 1993 on MSA growth. I introduce a couple more instruments to generate plausibly exogenous variation in SBA guaranteed loans: the number of SBA lender per capita in the metropolitan area in 1985, and years since interstate banking was deregulated in each metropolitan area. Table 9 Panel A presents the first stage results of the 2SLS estimation, i.e., the impact of the instrumental variables on the number of SBA loans approved for new small businesses in 1993. All specifications in Table 9 control for the initial economic conditions and census division dummies. Column (1) indicates that the number of SBA lender per capita in 1985 strongly predicts the number of SBA guaranteed loans to new business in 1993. The idea behind this is that cities that have higher competition among lenders will likely give out more loans. The validity of the instrument relies on the assumption that the number of loans given out in 1993, conditional on MSA employment, income, population, education, and housing price in 1993, is related to the density of SBA lenders in 1985 but not to unobserved demand factors determining urban growth between 1993-2002.

Column (2) uses years since interstate banking deregulation as an instrument. Banks in the U.S. were severely restricted in their ability to branch within and across state borders during most of the 20th century. Such restrictions were based on the concern that large concentrated banks would help the wealthy at the cost of the poor (Beck et al. 2010). Only in recent decades did states start to permit banks to open new branch within state (intrastate branching) and out of state (interstate branching), and by 1994 all restrictions were lifted with the passage of the Riegle-Neal Interstate Banking and Branching Efficiency Act. Table 5 lists the years each state deregulated interstate banking. I use years since interstate branching deregulation in 1993 (1993- deregulation year) as an instrumental variable. For MSAs that overlap with multiple states, I use the average years across the overlapping states. The main intuition behind the instrument is that MSAs that deregulated interstate branching earlier would see more competition for commercial lending in 1993. This in turn would reduce the need for marginal entrepreneurs to go through the bureaucracy of the SBA to get loans. Column (2) confirms this relationship. The longer it has been since deregulation the lower is

SBA backed entrepreneurship in 1993. The validity of this instrument hinges on the assumption that the timing of deregulation was more or less idiosyncratic and unrelated to the growth potential of cities between 1993 and 2002. Previous studies have found the timing of deregulation to be unrelated to state economic conditions (Beck et al. 2010). Column (3) illustrates the first stage when both instruments are used.

Table 9 Panels B through D report the 2SLS results on employment, payroll, and wage. For each column the instrumental variables are the variables reported in Panel A. Whichever instrumental variables I use, the estimated impact of government guaranteed entrepreneurship on either urban employment or income growth is statistically indistinguishable from zero at standard levels. The first stage F-statistic is generally quite strong, and when multiple instruments are used the over-identification test results pass the first cut for instrument exogeneity.

Table 10 Crowd-out of Market Entrepreneurship by Government-backed Entrepreneurship

	(1)	(2)
Dependent variable:	*Log market entrepreneurship*	*Log market entrepreneurship*
Log number of SBA loans to new businesses	0.0174***	-0.0144***
	(0.00582)	(0.00364)
Log market entrepreneurship		
Control variables	Y	Y
Year fixed effects	Y	Y
MSA fixed effects		Y
Observations	3,223	3,223
R-squared	0.97	0.99

Notes: The unit of analysis is the MSA-year for 316 MSAs between 1993 and 2002. The number of new SBA loans approved between July 1992 and June 1993 in each MSA proxy for government-backed entrepreneurship. Market entrepreneurship is defined as total small business birth minus the number of new SBA loans. All models include employment, payroll, population, establishment, and the house price index as control variables. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

Table 10 directly compares the impact of market entrepreneurship versus government-backed entrepreneurship on urban economic growth. Since the establishment birth variables used in the previous section is the universe of births, I subtract the number of SBA guaranteed loans to new businesses from the number of small business birth to get the number of market entrepreneurship. All specifications control for initial economic conditions and the census division dummies. Columns (1) through (3) report the OLS results on employment, payroll, and wage growth. The coefficient estimates on market entrepreneurship is nearly identical to the estimates in Table 2. However, the coefficient estimates on government-backed entrepreneurship decreases relative to the estimates of Table 7. Once I control for market entrepreneurship, the impact of government-backed entrepreneurship weakens and is no longer statistically significant. I estimate the same specification using 2SLS using all four instruments. Columns (4) and (5) report the first stage and list the instruments used. Note that the instruments generally impact market entrepreneurship versus government-backed entrepreneurship in opposite directions. As I discussed with the deregulation instrument, a lending environment helpful for market entrepreneurship decreases the potential entrepreneur's need to seek government help and in turn suppresses government-backed entrepreneurship. Columns (6) through (8) report the 2SLS results using all four instrumental variables. The first stage F-statistics is 6 and the over-identification test reports relatively large p-values. Similar to the OLS results in columns (1) through (3), there is no impact of government-backed entrepreneurship on urban economic growth. The coefficient estimates on market entrepreneurship is 0.185 for employment growth, 0.39 for payroll growth and 0.205 for wage growth, which are similar to the 2SLS estimates reported in Table 5.

5.3. Does Government Backed Entrepreneurship Crowd out Market Entrepreneurship?

Given that market entrepreneurship promotes urban employment and income growth and that government backed entrepreneurship has no impact, I further examine whether government backed entrepreneurship

simply supplements market entrepreneurship or whether there is crowd out of market entrepreneurship because of government-backed entrepreneurship. Table 11 examines this relationship. In practice I run the following panel regression:

$$\ln mrktent_{i,t} = \beta \ln govtent_{i,t} + \ln X_{i,t} \cdot \gamma + \mu_i + \eta_t + \varepsilon_{i,t} \qquad (8)$$

where $\ln govtent_{i,t}$ is the log number of SBA guaranteed loans to new businesses and $\ln mrktent_{i,t}$ is the log number of market entrepreneurship, i.e., the number of small business births minus the number of SBA loans to new businesses. $X_{i,t}$ is the set of the employment, establishment, payroll, and housing price index variables, η_t is the vector of year fixed effects, and μ_i is the vector of MSA fixed effects. Column (1) estimates the above equation excluding the MSA fixed effects. I find a positive relationship between government and market entrepreneurship. However, once I control for MSA fixed effects and look within MSAs over time the relation becomes negative and statistically significant. Government-backed entrepreneurship crowds out market entrepreneurship. A doubling of SBA loans to new small businesses decreases market entrepreneurship by 1 percent. Using the averages in 1993, this implies that increasing the number of SBA loans to new businesses by 13 will decrease market entrepreneurship by 13. There is a one for one crowd out. The results imply that government-backed entrepreneurship replaces market entrepreneurship one for one but in itself has no positive impact on economic growth. Based on the crowd out result and the *average* impact of entrepreneurship, one could conclude that government-backed entrepreneurship actually interferes with urban economic growth.

Entrepreneurial ability and hence the contribution of each entrepreneur to urban economic growth is likely heterogeneous. An important question to ask is whether the SBA loans were crowding out the high ability or low ability market entrepreneurs. One way to assess this is to compare the 2SLS estimates that include all entrepreneurs in Table 6 and when we separate out the type of entrepreneurs in Table 10. The estimates on market entrepreneurship are not only statistically indistinguishable but also very similar. The SBA loans may have

crowded out certain entrepreneurs but it seems like it did not replace entrepreneurs that contributed to economic growth. There may not be growth benefits of government-backed entrepreneurship but there also seems to be no harm. Hence, a more complete assessment of government-backed entrepreneurship would require careful examination relating to equity concerns, a future area of research, in addition to the efficiency results found in this paper.

6. Conclusion

Entrepreneurship is widely believed to be a main source of economic growth. This paper estimated the impact of entrepreneurship measured by the birth of businesses on urban employment and income growth, and examined how entrepreneurship supported by government guaranteed loans compare with market entrepreneurship in relation to its impact on urban growth. I also examine whether government-backed entrepreneurship complements or crowd outs market entrepreneurship. The study of entrepreneurship and urban growth has been hampered by the joint determination of the two. I use the variation in entrepreneurship generated by the homestead exemption levels in state bankruptcy laws to examine urban growth between 1993 and 2002. I find that a ten percent increase in the birth of small businesses increases MSA employment by 1.1 to 2.2%, annual payroll by 3.1 to 4.0%, and wage by 1.8 to 2.0% after ten years. I next examine whether the Small Business Loan programs that guarantee loans to entrepreneurs unable to finance through the market generate urban growth. I find no growth impact from government-backed entrepreneurship and further find that government-backed entrepreneurship crowds out market entrepreneurship one to one. In sum, market entrepreneurship promotes urban employment and income but government-backed entrepreneurship does not.

While the results of this paper indicate that there are no efficiency gains from government-backed entrepreneurship, a complete assessment of government-backed entrepreneurship requires further examination regarding how equitable entrepreneurial activity is. The main rationale for government intervention is market failure in the small business

lending market, and particularly of discrimination. Blanchflower et al. (2003) find that black entrepreneurs are twice as likely to be denied credit compared to white entrepreneurs. A substantial literature has documented discrimination in the home mortgage lending market (Ladd 1998) and the employment market (Bertrand and Mullainathan 2004, Oreopoulos 2009). The SBA reports that the share of female and minority entrepreneurs are smaller relative to the overall economy and many state economic development agencies and the federal Minority Business Development Agency provide assistance to female and minority entrepreneurs. Commercial lenders are unwilling to lend to potential entrepreneurs without sufficient collateral. This may imply that on average we will see less entrepreneurship in demographics with lower wealth. However, the literature has also found preference-based discrimination in the lending market. Further examination on the extent and impact of such discrimination is needed for a complete assessment of government-backed entrepreneurship.

References

Acs, Zoltan et al. 2008. “Entrepreneurship and Urban Success: Toward a Policy Consensus”, Kauffman Foundation.

Adelino, Manuel, Antoinette Schoar, and Felipe Severino. 2013. “House Prices, Collateral, and Self-Employment.”NBER Working Paper 18868.

Beck, Thorsten, Ross Levine, and Alexey Levkov. 2010. “Big Bad Banks? The Winners and Losers from Bank Deregulation in the United States.” *Journal of Finance*, 45(5), 1637-1667.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991-1013.

Bertrand, Marianne, and Antoinette Schoar. 2006. "The Role of Family in Family Firms." *Journal of Economic Perspectives*, 20(2): 73-96.

Black, Sandra E. and Philip E. Strahan. 2002. “Entrepreneurship and Bank Credit Availability.” *Journal of Finance*, 57(6): 2807-2833.

Blanchflower, David G., Phillip B Levine, and David J. Zimmerman. 2003. “Discrimination in the Small-Business Credit Market.” *Review of Economics and Statistics*, 85(1): 930-943.

Bracke, Phillip, Christian Herber, and Olmo Silva. 2013. “Homeownership and Entrepreneurship: The Role of Commitment and Mortgage.” IZA Discussion Paper No. 7417.

Craig, Ben R., William E. Jackson III, and James B. Thomson. 2007. “Small Firm Finance, Credit Rationing, and the Impact of SBA-Guaranteed Lending on Local Economic Growth”, *Journal of Small Business Management*, 45(1), pp. 116-132.

Duranton, Gilles, and Matthew A. Turner. 2012. “Urban Growth and Transportation.” *Review of Economic Studies*, 79, 1407-1440.

Duranton, Gilles and Diego Puga, 2013. "The Growth Of Cities," Working Papers wp2013_1308, CEMFI.

Edmiston, Kelly, 2007. “The Role of Small and Large Businesses in Economic Development”, *Economic Review*, Federal Reserve Bank of Kansas City, pp. 73-97.

Fan, Wei and Michelle J. White. 2003. “Personal Bankruptcy and the Level of Entrepreneurial Activity.” *Journal of Law and Economics*, 46, 543-567.

Glaeser, Edward L., Kallal, Hedi D., Scheinkman, Jose A., and Andrei Shleifer. 1992. “Growth in Cities.” *Journal of Political Economy*, 100(6), 1126-1152.

Glaeser, Edward L., Sari Pekkala Kerr, and William R. Kerr. 2012. “Entrepreneurship and Urban Growth: An Empirical Assessment with Historical Mines.” NBER Working Paper 18333.

Glaeser, Edward L., William R. Kerr, and Giacomo A.M. Ponzetto. 2010. “Clusters of Entrepreneurship”, *Journal of Urban Economics*, 67 150-168.

Glaeser, Edward L., Stuart S. Rosenthal, and William C. Strange. 2010. “Urban Economics and Entrepreneurship”, *Journal of Urban Economics*, 67.

Haltwinger, John, Ron S Jarmin, and Javier Miranda. 2011. “Who Creates Jobs? Small vs. Large vs. Young”, *The Review of Economics and Statistics*, May 2013, Vol. 95, No. 2, pp. 347-361.

Henderson, Vernon, Kuncoro, Ari, and Matt Turner. 1995. “Industrial Development in Cities.” *Journal of Political Economy*, 103(5), 1067-1090.

Hamilton, Barton H. 2000. “Does Entrepreneurship Pay? An Empirical Analysis of the Returns of Self-Employment”, *Journal of Political Economy*, 108(3) 604-631.

Kerr, William, Josh Lerner, and Antoinette Schoar. 2010. “The Consequences of Entrepreneurial Finance: A Regression Discontinuity Analysis.” NBER Working Paper 15831.

Kliesn, Kevin L. and Julia S. Maues. 2011. “Are Small Businesses the Biggest Producers of Jobs?” *The Regional Economist*, Federal Reserve Bank of St. Louis, pp. 8-9.

Ladd, Helen F. 1998. "Evidence on Discrimination in Mortgage Lending." *Journal of Economic Perspectives*, 12(2): 41-62.

Lee, Seung-Hyun, Yasuhiro Yamakawa, Mike W. Peng, and Jay B. Barney. 2011. “How Do Bankruptcy Laws Affect Entrepreneurship Development Around the World?” *Journal of Business Venturing*, 26: 505-520.

Lerner, Josh, and Ulrike Malmendier. 2011. “With a Help From My (Random) Friends: Success and Failure in Post-Business School Entrepreneurship.” NBER Working Paper 16918.

Michelacci, Claudio and Olmo Silva. 2007. “Why So Many Local Entrepreneurs?” *The Review of Economics and Statistics*, Vol. 89, No. 4, pp. 615-633.

Neumark, David, Wall, Brandon, and Zhang, Junfu. “Do Small Businesses Create More Jobs? New Evidence for the United States from the National Establishment Time Series.” *The Review of Economics and Statistics*, February 2011, Vol. 93, No.

1, pp. 16-29.

Oreopoulos, Philip. 2009. “Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Six Thousand Resumes.” NBER Working Paper No. 15036.

Posner, Eric A., Hynes, Richard, and Anup Malani. 2001. “The Political Economic of Property Exemption Laws.” John M. Olin Law & Economics Working Paper No. 136.

Rosenthal, Stuart S. and William C. Strange. 2003. “Geography, Industrial Organization, and Agglomeration.” *The Review of Economics and Statistics*, Vol. 85, No. 2, pp. 377-393.

Samila, Sampsa and Olav Sorenson. 2011. "Venture Capital, Entrepreneurship, and Economic Growth." *The Review of Economics and Statistics*, Vol. 93, No. 1, pp. 338-349.

Appendix Table 1 Impact of Firm Expansion on Urban Growth

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Dependent variable:	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log establishment births by existing medium firms in 1992-93	0.0721***		0.115***		0.0431***	
	(0.0238)		(0.0288)		(0.0141)	
Log establishment births by existing large firms in 1992-93		0.0814***		0.115***		0.0332**
		(0.0249)		(0.0312)		(0.0146)
Panel B: Dependent variable:	*Change in log employment, 1993-1998*		*Change in log payroll, 1993-1998*		*Change in log wage, 1993-1998*	
Log establishment births by existing medium firms in 1992-93	0.0653***		0.0828***		0.0175*	
	(0.0139)		(0.0171)		(0.00940)	
Log establishment births by existing large firms in 1992-93		0.0509***		0.0761***		0.0252**
		(0.0157)		(0.0197)		(0.0113)
Panel C: Dependent variable:	*Change in log employment, 1997-2002*		*Change in log payroll, 1997-2002*		*Change in log wage, 1997-2002*	
Log establishment births by existing medium firms in 1996-97	0.0849***		0.0781**		-0.00676	
	(0.0254)		(0.0320)		(0.0132)	
Log establishment births by existing large firms in 1996-97		0.0215		0.0170		-0.00451
		(0.0184)		(0.0222)		(0.00999)

Appendix Table 1 (Continue)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel D: Dependent variable:	*Change in 5 year employment growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		*Change in 5 year payroll growth, (1997 to 2002 growth) - (1993 to 1998 growth)*		*Change in 5 year wage growth, (1997 to 2002 growth) - (1993 to 1998 growth)*	
ΔLog establishment births by medium firms between 1993 and 1997	0.0382*		0.0166		-0.0216***	
	(0.0227)		(0.0239)		(0.00775)	
ΔLog establishment births by large firms between 1993 and 1997		0.00232		-0.00462		-0.00694
		(0.0118)		(0.0164)		(0.00836)

Notes: The unit of analysis is the MSA and the number of observations is 316. Establishment births for 1993 are counted between March 1992 and March 1993. The "establishment birth by existing medium firms" variable refers to expansion by firms with 20-499 employees. The "establishment birth by existing large firms" variable refers expansion by firms with over 500 employees.The initial employment, median family income, population, percent college degree and above,the house price index, the three log number of establishment variables and the nine census division dummies are included as controls in Panels A through C. The change inlog employment, payroll, population, and house price index, the 1990 percent college educated and log median family income, and the change in the three establishment number variables are included as controls in Panel D. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

Appendix Table 2 Homestead Exemption and Firm Expansion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable:	*Log establishment births by medium firms in 1992-1993*	*Log establishment births by large firms in 1992-1993*	*Log establishment births by medium firms in 1992-1993*	*Log establishment births by large firms in 1992-1993*	*Log establishment births by medium firms in 1992-1993*	*Log establishment births by large firms in 1992-1993*	*Log establishment births by medium firms in 1992-1993*	*Log establishment births by large firms in 1992-1993*
Log homestead exemption level in 1975					0.0106**	0.0147***	0.00718	0.0113**
					(0.00537)	(0.00496)	(0.00505)	(0.00516)
Unlimited exemption in 1975	0.106**	0.107**	0.0584	0.0578	0.0691	0.0555	0.0353	0.0215
	(0.0449)	(0.0414)	(0.0437)	(0.0407)	(0.0473)	(0.0454)	(0.0457)	(0.0441)
Log small business births in 1992-1993			0.432***	0.445***			0.412***	0.415***
			(0.101)	(0.0878)			(0.101)	(0.0898)
Base controls	Y	Y	Y	Y	Y	Y	Y	Y
Initial establishment controls	Y	Y	Y	Y	Y	Y	Y	Y
Census division fixed effects	Y	Y	Y	Y	Y	Y	Y	Y

Notes: The unit of analysis is the MSA and the number of observations is 316. Small business births for 1993 are counted between March 1992 and March 1993. The "small business births" variable includes all new firm creation and expansions by firms with less than 20 employees. The "establishment birth by existing medium firms" variable refers to expansion by firms with 20-499 employees. The "establishment birth by existing large firms" variable refers expansion by firms with over 500 employees. Base controls are initial employment, median family income, population, percent college degree and above, and the house price index. Initial establishment controls are the three log number of establishment variables. The nine census division dummies are included as controls.* p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

Appendix Table 3 2SLS Estimates Using Both Homestead Exemption Variables as Instruments

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Control for minimum wage	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log establishment per employee in 1993	0.260*		0.493**		0.232**	
	(0.147)		(0.210)		(0.112)	
Log small business births in 1992-93		0.115		0.324**		0.208***
		(0.102)		(0.141)		(0.0767)
1st stage F-statistic	7.91	11.99	7.91	11.99	7.91	11.99
Panel B: Control for Right-to-work	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log establishment per employee in 1993	0.210		0.445**		0.234**	
	(0.142)		(0.201)		(0.107)	
Log small business births in 1992-93		0.0711		0.291*		0.220***
		(0.119)		(0.158)		(0.0849)
1st stage F-statistic	8.2	10.11	8.2	10.11	8.2	10.11
Panel C: Control for past population	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log establishment per employee in 1993	0.245		0.465**		0.220*	
	(0.152)		(0.220)		(0.127)	
Log small business births in 1992-93		0.0689		0.289*		0.220**
		(0.123)		(0.165)		(0.0896)
1st stage F-statistic	6.07	8.79	6.07	8.79	6.07	8.79

Appendix Table 3 (Continue)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel D: Control for industry composition	*Change in log employment, 1993-2002*		*Change in log payroll, 1993-2002*		*Change in log wage, 1993-2002*	
Log establishment per employee in 1993	0.0701		0.445		0.375	
	(0.281)		(0.402)		(0.233)	
Log small business births in 1992-93		0.00144		0.198		0.197*
		(0.144)		(0.190)		(0.101)
1st stage F-statistic	3.76	8.32	3.76	8.32	3.76	8.32
Panel E: Other outcome variables:	*Log total establishment in 2002*		*Log establishment death in 1994*			
Log establishment per employee in 1993	1.227***		1.880***			
	(0.131)		(0.198)			
Log small business births in 1992-93		0.432***		0.741***		
		(0.0960)		(0.111)		
1st stage F-statistic	7.76	11.77	7.76	11.77		
Base controls	Y	Y	Y	Y	Y	Y
Initial establishment controls		Y		Y		Y
Census division fixed effects	Y	Y	Y	Y	Y	Y

Notes: All results are 2SLS estimates using both the unlimited exemption status in 1975 and the log exemption level in 1975 as instruments. The unit of analysis is the MSA. Each panel adds additional controls to the specifications in Table 6 Panel B Columns (1) and (3). The MSA average minimum wage is added in Panel A, the Right-to-work status in Panel B, log population in 1985 in Panel C, and the employment shares of manufacturing, service, and retail in Panel D. The number of observations is 316 except for Panel D which is 306. Base controls are initial employment, median family income, population, percent college degree and above, and the house price index. Initial establishment controls are the three log number of establishment variables. The nine census division dummies are included as controls. The Kleibergen-Paap rk Wald F statistics are reported as the 1st stage F-statistics.* p<0.1, ** p<0.05, *** p<0.01. Robust standard errors are in parentheses.

CHAPTER 5

Changes in Competition of Small vs. Large Firms resulting from International Trade*

by
Changwoo Nam
(Korea Development Institute)
*Jiyoon Oh***
(Korea Development Institute)

Abstract

Using Korean plant-level manufacturing data, this paper examines the effect of lowering trade barriers on changes in markups of small and large firms, exporters and non-exporters. We find that large firms decide on higher markups in each sector as they have greater market power in integrated markets. Also, exporters set higher markups through higher observable productivity than non-exporters. Even after controlling for productivity and other firm characteristics, markups are proportional to market share, which can be interpreted to mean that market power only influences firm price strategy. Interestingly, the markup distribution, which is more closely related to the competition from globalisation, has been decreasing over time while the performance gap measured as sales has been stable over time. We caution that even if the performance gap measured in absolute terms may be widening, this does not imply that the level of competition between large and small firms is weakening.

* This working paper is funded by Economic Research Institute for ASEAN and East Asia (ERIA).
** Correspondent Author's Contact: Dr. Oh, Jiyoon: jiyoon.oh@kdi.re.kr

1. Introduction

Globalisation has been regarded as one of the main driving forces behind changes in market environments, such as degree of competitiveness between firms. Intuitively, more integrated markets confer a benefit on more productive firms as they can sell more products in a bigger market. The firm selling its product in the domestic market can grow as a global company. This leads to the exit of less productive firms in the market, thereby polarising firm performance. On the other hand, the influx of foreign products from world markets makes the market environment more competitive, so firms with monopolistic power due to market frictions can lose their market power in the domestic market. It creates a level playing field for all firms, so "second movers" with small market share can enjoy more equal benefits. It alleviates inequality between firms, especially in terms of firm size.

Since globalisation has two opposite effects on inequality between firms, it is a natural question whether more integrated markets benefit all firms equally. There is a wide range of academic literature on globalisation and its effect on aggregate output or firm performance, but relatively little work has been done on the different impacts of globalisation on firm performance.

For policy administration, firm size is a convenient measure. In many countries, firm policy has been implemented discriminately according to firm size. Tax benefits accrue to a greater extent to small firms, as regulations for large firms are stricter. Even if firm size contains many characteristics suggesting productivity, firm age, market power, size itself is actually obscure property. For example, firm size is not directly linked to productivity. There are ongoing debates about why large firms are large. Are they big because of their advanced technology or did they just benefit from early market entry, giving them a "first mover" advantage. Economists who had thought that small firms are the engine of growth and the entity of creative destruction have come to realize that many small firms are actually at the low level of innovation. (Hurst and Pugsley, 2011) However, there is a general consensus that market share is the obvious determinant of market power. It seems appropriate, therefore, to study the effect of globalisation on changes in the relative

market power of firms. Markups are commonly investigated as a proxy variable to capture market power.

In this paper, we investigate whether more integrated markets through globalisation expand or shrink the market power gap between large and small firms. Using plant-level data of the Korean annual survey of manufacturing, we rigorously estimate plant markups and assess markup trends over time. We then empirically examine the effect of lowering trade barriers on changes in markups of small and large plants. Through this exercise, we can test our educated guesses of markup variations in small and large firms in international models. Furthermore, we can directly observe the market power gap between small and large firms measured by markups, and investigate whether this gap converges when markets become more open through trade liberalization.

In terms of the theoretical literature, our paper is closely associated with the recent development of the heterogeneous model of international trade. Markups have received a lot of attention from economists and policymakers as they measure the effect of various competition and trade policies on market power. Recently, the theoretical study of firm heterogeneity in terms of productivity or size has been combined with heterogeneity in markups. Melitz and Ottaviano (2008) suggest a monopolistically competitive model of trade with firm heterogeneity. In their model, market size and trade affect the strength of competition. Larger and more integrated markets through trade tend to have lower markups. However, this paper does not point out the difference between small and large firms. Decreases in markups when market size is bigger through trade, is linear in terms of productivity.

Other types of theoretical models of endogenous markups, such as Atkeson and Burstein (2008), Oh (2013), and Edmund, Midrigan and Xu (2013) emphasize the increasing schedule of optimal markup with respect to a firm's market share. These types of models rely on a similar setting of monopolistically competitive markets, except that the number of competitors in a particular industry or for a particular product is small. In such settings, firms take into account the effect of their pricing decisions on the equilibrium of the prices of industry goods. Price elasticity of demand decreases with a firm's market share. Thus an

optimal markup, which is the inverse of price elasticity, increases with firm size. Large firms assign higher markups than small firms. A reduction in trade barriers reduces the industry share of domestic producers, thus reducing their markups. Interestingly, the optimal markup is convex-increasing in a firm size. Therefore, the adjustment in markup of large firms is larger than that of small firms with the same reduction in market share due to international trade.

This notion of differences in markups between small and large firms has hardly been investigated empirically. Roberts and Supina (1996) show that plant-specific markups of price over marginal cost vary according to the size of producers. For three products, markups decline in size and in two cases they increase. Edmund, Midrigan and Xu (2013) accurately calibrate their model with Taiwanese manufacturing plant-level data and argue that endogenous markup setting shows much larger gains from trade than Ricardian models. Bigger welfare gains in their model are driven by a significant reduction in large firms' markups. They imply that import competition reduces the gap between large and small in terms of firm markup. However, they never show any empirical evidence that plant-specific markups decrease after trade barriers are lower.

Our main empirical findings confirm many theoretical predictions. First, the level of markup is obviously higher in more productive firms, as predicted by Melitz and Ottaviano (2008). Second, markup increases as market share rises. This reinforces the increasing relationship between markup and market share. As Atkeson and Burstein (2008) and Oh (2013) predicted, larger market share leads to higher markup because large firms can enjoy more market power due to lower demand elasticity. Third, markups of exporters are higher on average than markups of non-exporters. This makes sense, that exporters are mostly more productive and larger firms, which can afford to pay fixed costs for exporting, as Melitz (2003) predicts. Fourth, we create distributions of firm markups at a given point every year, and compare them. Interestingly, the mean of markups has decreased over time, and so has the dispersion. Even though we cannot identify the main factors behind for convergence of markups, a competition effect from globalisation is definitely one of the plausible factors. In order to identify and quantify

the effect of import competition on markup dispersions, we regress industry markup dispersions on industry import penetrations. Generally speaking, import competition reduces markup dispersion. Lastly, although the overall picture of markup distribution has been increasingly condensed over time, we find that when individual firms expand their market share in overseas markets, their markups increase.

The remainder of this paper is organized as follows. We introduce our theoretical model and briefly provide theoretical predictions about empirical results in Section 2. Section 3 introduces our empirical framework and our estimation routine. Section 4 provides our main empirical results and a discussion of them. The final section concludes.

2. Theoretical Background

In this section, we illustrate how variations in firm size are theoretically related to differing levels of firm markup. We first lay out the market structure in the model to examine the mechanisms underlying market share and markup. This model is based on the monopolistic competition suggested by Dixit and Stiglitz (1977), except that it has a few competetors rather than a continuum of firms. The goods market features differentiated oligopoly competition within a quantity-setting game.

We construct a model of imperfect competition in which final goods consist of a continuum of industry goods and each industry goods market consists of N_j firms. The final good is produced using a constant returns to scale production function, which aggregates a continuum of industry goods.

$$Y = \left(\int_0^1 y_j^{1-\frac{1}{\eta}} dj \right)^{\frac{\eta}{\eta-1}} \quad (1)$$

where y_j denotes the output of industry y_j. The elasticity of substitution between any two different industry goods is constant and equals η. Final goods producers behave competitively.

In each industry, there are N_j firms producing differentiated goods that are aggregated into industry goods through an aggregate constant

elasticity of substitution (CES) production function. The output of goods in industry j[1] is given by

$$y_j = N^{\frac{1}{1-\theta}} \left(\sum_{i=1}^{N_j} y_{ij}^{1-1/\theta} \right)^{\frac{\theta}{\theta-1}} \tag{2}$$

where y_{ij} is the output of firm i in industry j. Within each industry of N_j firms, a firm sets its quantity. The elasticity of substitution between any two intra-industry goods is constant and equals θ. It is assumed that the elasticity of substitution between any two goods within an industry is higher than the elasticity of substitution across industries, $1 < \eta < \theta$.

The final good producer solves a static optimization problem that results in the usual conditional demand for each industry good,

$$y_j = \left(\frac{P_j}{P} \right)^{-\eta} Y,$$

where P_j is the industry j price and P is the price of final goods,

$$P = \left(\int_0^1 P_j^{1-\eta} dj \right)^{\frac{1}{1-\eta}}, \tag{3}$$

Denoting the price of good i in industry j by P_{ij},

$$P_j = N_j^{\frac{1}{\theta-1}} \left(\sum_{i=1}^{N_j} p_{ij}^{1-\theta} \right)^{\frac{1}{1-\theta}}, \tag{4}$$

the inverse demand functions for goods within an industry are given by:

1 The term N1-11P implies that there is no variety effect in the model.

$$\left(\frac{p_{ij}}{P}\right) = \left(\frac{y_{ij}}{y_j/N_j}\right)^{-1/\theta} \left(\frac{y_j}{Y}\right)^{-1/\eta}$$

Dixit and Stiglitz (1977) assume that each firm is small relative to the economy, and therefore does not influence the equilibrium price and quantity. In this model, the assumption of a small number, N_j, of firms in each industry implies that a firm's quantity choice affects the industry price. Within a given industry, each firm takes into account the effect that the pricing and production decisions of other firms has on the demand for its own goods. Therefore, the price elasticity of demand ϵ (S_{ij}) of firm (i) is a decreasing function of the firm's when the substitutability of within-industry goods is higher than that of between-industry goods ($\eta < \theta$). In equation (6), the demand elasticity is a market share weighted average of two values [η, θ]: when yij is close to zero, the perceived demand elasticity of firm i in industry j is equal to θ, which is the same as in Dixit and Stiglitz (1977). On the other hand, if y_{ij} is close to one, the demand elasticity of firm i is the same as that of the monopoly firm in industry j.[2]

$$\epsilon(S_{ij}) = \left[\frac{1}{\theta}(1 - S_{ij}) + \frac{1}{\eta}S_{ij}\right]^{-1} \tag{5}$$

From eq (4) and eq (5), these market shares can be written as a function of prices in equation (7)

$$S_{ij} = \frac{p_{ij}y_{ij}}{p_jY_j} = \frac{p_{ij}^{1-\theta}}{\sum_{i=1}^{N_j} p_{ij}^{1-\theta}} \tag{6}$$

Derived directly from the demand elasticity in equation (6), firm markup is an increasing function of its market share from (8).

2 If firms compete in a price-setting game (Bertrand competition) within an industry, the demand elasticity would be $\epsilon_{y,p} = \theta(1 - s_i) + \chi s_i$.

$$\mu_{ij}(S_{ij}) = \frac{\epsilon(S_{ij})}{\epsilon(S_{ij}) - 1} = \frac{1}{1 - \frac{1}{\theta}(1 - S_{ij}) - \frac{1}{\eta} S_{ij}} \tag{7}$$

Firm markups are combined into aggregate industry markup ($\bar{\mu}_j$). Aggregate markup can be expressed in two ways: the input-share weighted average of firm markup, which is equal to the revenue-share weighted harmonic average of firm markup.

$$\bar{\mu}_j = \sum_{i=1}^{N_j} x_{ij}\mu_{ij} = \left(\sum_{i=1}^{N_j} \frac{S_{ij}}{\mu_{ij}} \right)^{-1} \tag{8}$$

where $x_{ij} = \frac{Input_{ij}}{Input_j}$ is the input share[3] of firm *i* in industry *j*.

In a symmetric industry equilibrium, aggregate industry markup $\bar{\mu}_j$ is equal to aggregate markup $\bar{\mu}$. Going forward, we will restrict our attention to symmetric industry equilibrium.

The assumption of ($\theta > \eta$) implies that each firm's markup of its price over marginal costs is an increasing function of that firm's market share within an industry. At one extreme, if the firm has a market share S_i approaching zero, it faces only the industry elasticity of demand 0 and chooses a markup equal to θ / (θ-1). At the other extreme, if the firm has a market share approaching one, it has lower elasticity of demand across industries η and sets a higher markup equal to η / (η-1). The difference θ-η actually determines how much the demand elasticity changes in response to shifts in market share. As θ-η gets bigger, the effect of market share on demand elasticity and markup becomes increasingly significant.

$\Gamma(s)$ refers to the elasticity of the markup with respect to market

---

3 In the case that input prices are common to all firms, input shares of any input are equal within firms. For instance, the labour input share $\frac{h_{ij}}{H_j}$ of firm *i* in industry *j* is the same as the capital input share $\frac{h_{ij}}{K_j}$, if firms face the same wage rates and capital rental prices.

share. Note that $\Gamma(s)$ is an increasing and convex function of s. In the constant markup model, $\Gamma(s) = 0$,

$$\Gamma(s) = \frac{s}{1 - \frac{1}{\theta}(1-s) - \frac{1}{\eta}s}\left(\frac{1}{\eta} - \frac{1}{\theta}\right)$$

This convexity plays an important role in the dynamics of aggregate markup. Due to this convexity, aggregate markup increases as market shares across firms become more dispersed or unequal.

In addition to convexity, the level of aggregate markup is influenced by a composition effect. Since aggregate markup is the input-share weighted average of firm markups, a large firm's high markup weighted by its high input-share contributes significantly to raising aggregate markup, and vice versa. This composition effect implies that the pricing behaviours of large firms play a dominant role in the dynamics of aggregate markup.

It is worth mentioning that a firm's markup does not change unless its market share changes. When there are uniform changes such as cost reductions for all firms, relative prices do not change between firms; therefore, market share stays constant. This is an important departure from a generic sticky price model in which an exogenous price-setting friction causes variations in markup for the representative firm.[4] In our model, aggregate fluctuations cannot change aggregate markup. Only changes in relative productivity between firms matter in determining aggregate markup.

The model described above can apply to how globalisation can influence firm decisions in terms of markups. In terms of increases in importing, trade liberalization and the resulting greater influx of imported goods make domestic markets more competitive. Rises in import penetration in an industry naturally reduces the market share of domestic firms. Based on the theoretical framework above, this effect lowers the level of domestic markups. Furthermore, the speed of lowering markups is accelerated in large firms rather than small firms

4 It follows that our model can explain why large firms are reluctant to cut prices in recessions—due to the low demand elasticity they face.

due to the convex schedule of optimal markup. In this sense, we can say that globalisation generates more competition and reduces market power inequality between large and small firms.

When it comes to globalisation through exporting, applying this modified imperfect competition framework is ambiguous. It is obvious domestic firms lose domestic market share to foreign competitors. For exporting producers, domestic market share may not change at all after participating in exporting, but entry to exporting may change the firm distribution through the selection process.

In terms of the economic literature, we can lean on the endogenous markup model suggested by Melitz and Ottaviano (2008). Even if the details are different, the basic mechanism is closely related to our model above. Competition from entry lowers the level of markups. Melitz and Ottaviano (2008) prove theoretically that the mean of markups decreases and the average level of productivity of firms increases as markets become more integrated through trade liberalization. This makes sense, that the selection effect pushes the least productive firms out of the market. A more competitive environment makes firms reduce prices and markups as well. Interestingly, Melitz and Ottaviano also predict the dispersion of firm performance measures such as price, markup, and firm size: the variance of cost, prices, and markups are lower in bigger markets because the selection effect decreases support for these distributions. On the other hand, the variance of firm size (in terms of either output or revenue) is larger in bigger markets due to the direct magnifying effect of market size on these variables.

Regarding the dispersion of firm performances, the two different directions of price and quantity are very interesting. Even if the degree of competition increases, firm size distribution will be more unequal. To determine if globalisation actually increases the level of competition or is more beneficial for large firms, markup distribution is a better measure than firm size distribution. In Section 4 we will present our empirical results regarding the dispersion of markups.

However, the effect of increases in firms' exports on firms' markups is ambiguous in the sense that variations in markups across firms in cross-sectional analysis may not show the same pattern in time series analysis. Thus, the real effect of international trade on differing markups

according to size should be measured empirically. De Loecker and Warzynski (2012) showed that exporting makes firms increase markups in time series analysis as well as in cross-sectional analysis.

3. Estimation

3.1. Production Function Estimation

The problem of estimating the production function has been an important issue since the beginning of economics as production functions are a fundamental component of all economics. In fact, the econometric subject is the possibility that the major determinants of firm's production decision might be unobservable to econometricians. Thus, this measurement error induces the endogeneity problem due to the relation between observed inputs and unobserved productivity shocks. Olley and Pakes (1996, hereafter referred to as the OP model), Levinsohn and Petrin (2003, hereafter the LP model), Ackerberg, Caves and Frazer (2006, hereafter the ACF model), and De Loecker and Warzynski (2012, hereafter the DLW model) are seminal papers leading to the introduction of new techniques for identification of production functions. OP model and LP model cannot avoid the multicollinearity issue when they estimate the labour coefficient of production function in the estimation scheme. DLW model owes ACF model in terms of the full identification in the second stage of structural estimation. In addition, these papers are somewhat more structural in nature-using observed input decisions to control for unobserved productivity shocks (De Loecker and Warzynski (2012)). These techniques have been used in a large number of recent empirical papers including Pavcnik (2002), Fernandes (2007), Criscuola and Martin (2009), Topalova and Khandelwal(2011), Blalock and Gertler (2004), and Alvarez and Lopez (2005).

3.2. Markup Estimation

Estimating markups has a long tradition in industrial organization

and international trade. Re-searchers in industrial organization are interested in measuring the effect of various competition and trade policies on market power through estimating unobservable markups. In this paper, we use a simple empirical framework in the DLW model to estimate markups. Our approach of following the DLW model nests the price-setting model used in applied industrial organization and international trade and relies on optimal input demand conditions obtained from standard cost minimisation and the ability to identify the output elasticity of a variable input. This framework eliminates issues related to input adjustment costs. Also, this methodology suggests that the output elasticity of a variable factor of production is exactly equal to its expenditure share in total revenue as price equals marginal cost of production, solving the cost minimisation problem.

Markup estimates are obtained using production data where we observe output, total expenditure on variable inputs, and plant-level revenue datasets. The DLW model in particular requires a measure of output that disregards price differences across firms. Therefore, we use real output value from the Korean dataset. Recent literature makes empirical approach very suitable to those types of datasets from several countries. (Foster, Haltiwanger, and Syverson [2008]; Goldberg et al. [2010]; Kugler and Verhoogen [2008]; De Loecker and Warzynski [2012]).

Some assumptions are introduced following the DLW model. First, constant returns to scale is not imposed, and second, the user costs of capital do not need to be observed or measured in our model. This relaxation results in a flexible methodology and reliable estimates such as the DLW model. We then use our empirical model to verify whether exporters, on average, charge higher markups than their domestic counterparts in the same industry, and how markups change according to firm size, i.e., how the market share changes. This framework is well suited to relating markups to any observed plant-level activity potentially correlated with plant-level productivity.

3.3. Local Constant Kernel Model

In recent decades, the literature on nonparametric econometric methods

has offered solutions to the problems related to parametric misspecification of econometric regression models. This misspecification problem can be generically generated in production or markup estimations because the functional form of production is wholly determined by the researcher's arbitrary decision. However, nonparametric regression techniques basically do not oblige the researcher to assume and specify a functional form of production for the relationship between the firm's decision variables and the production variable (output production or value-added production). Fully nonparametric models are most often applied to cross-sectional data, while they are seldom applied to panel data sets (Czekaj and Henningsen [2013][5]).

There still exists a possibility that the DLW model has a multicollinearity problem because it uses nth order nonparametric series regression with inter-variable components in the first stage of structural estimation even though it fully estimates coefficients necessary to compute the markups in the second stage formed by the generalized method of moments (GMM) structure. Therefore, we use a local constant kernel model (hereafter, the LCK model) with unordered discrete data at the first stage of structural estimation. The LCK model is a fully nonparametric model that uses a time variable and an individual identifier as additional (categorical) explanatory variables (Racine and Li [2004]). In this formation we do not need to consider separately the production part of labour and capital, and the productivity shock observable to firm managers before their input decisions (on labour, investment, materials so on), but unobservable to econometricians. The fully nonparametric regression, that is, LCK, model only focuses on how to accurately estimate data. At the same time, the LCK model captures non-linear individual and time effects that do not need to be assumed to be additive and separable.

In our analysis, we use a fully nonparametric and nonseparable panel data model (LCK model) that has been suggested by Henderson and Simar (2005), Racine (2008), and Gyimah-Brempong and Racine (2010).

---

5 Czekaj and Henningsen (2013) only compare the fittability of ordinary least squares (OLS), semiparametric and fully nonparametric regressions. It is not their purpose to solve unbiased estimators for unobserved productivity shocks in firm decisions.

They estimate an undefined function as a fully nonparametric two-ways effects panel data model with individual effect and time effect as categorical explanatory variables using the nonparametric regression method proposed by Li and Racine (2004) and Racine and Li (2004). Those papers use both continuous and categorical explanatory variables for fully nonparametric specification. This estimator does not require any data transformation with a loss of observations. In addition, the intercept of the dependent variable and the slopes of the explanatory variables on the dependent variable are not fixed according to the interaction between time periods and individuals in the fully nonparametric model. Hence, this estimator does not imply any restrictions on the most general specification of panel data models. Furthermore, the bandwidths of the explanatory variables can be selected using data driven cross-validation methods. The overall shape of the relationship between the dependent variable and the covariates, the individual, and time is entirely determined by the data.

Finally, we compare the empirical results with LCK, DLW and conventional ordinary least squares (OLS) models. It is found that the LCK model fits better with the production data and is more consistent with the economic theory compared with the DLW and OLS models. This means that the LCK model captures the non-linear individual and time effects by using a discrete smoothing parameter, and the fitted value-added is determined by the local weighted average rather than by labour, capital, and material variables.

3.4. Structure to Estimate Markups

We explain the structural model to obtain plant-level markups relying on standard cost minimisation conditions for variable inputs following the DLW model. These conditions imply that the markup is the output elasticity of an input in relation to the share of that input's expenditure in total sales and the firm's markup (DLW model). To obtain output elasticities, we need estimates of the production function, for which we rely on proxy methods developed by the DLW model. We follow the restrictions that DLW imposes, and we discuss our model in detail in below given DLW model.

3.4.1. Deriving Markups

A firm i produces output at time t with the implicit production technology:

$$Q_{it} = Q_{it}(X_{it}^1, \dots, X_{it}^N, K_{it}, z_{it}),$$

in which it relies on N variable inputs such as labour, intermediate inputs, and electricity. In addition, the firm relies on a capital stock, K_{it}, which is treated as a dynamic input in production, which means the amount of investment at *t* is determined given the information at *t*-1. The productivity shock z_{it} evolves exogenously following a first order Markov process, and the labour in production is a non-dynamic input, which means the amount of labour at t is related to the observed productivity shock z_{it}. However, the only restriction we impose on Q_{it} to derive an expression of the markup is that Q_{it} is continuous and twice differentiable with respect to its arguments.

Producers have a cost-minimisation problem, such as the associated Lagrangian function:

$$\mathcal{L}(X_{it}^1, \dots, X_{it}^N, K_{it}, z_{it}) = \sum_{j=1}^{N} P_{it}^{X^j} X_{it}^j + r_{it} K_{it} + x_{it}\big(Q_{it} - Q_{it}(\cdot)\big),$$

in which $P_{it}^{X^j}$ and r_{it} show a firm's input price for a variable input *j* and capital, respectively. The first-order condition for any variable input is

$$\frac{\partial \mathcal{L}_{it}}{\partial X_{it}^j} P_{it}^{X^j} - x_{it} \frac{\partial Q_{it}(\cdot)}{\partial X_{it}^j} = 0,$$

in which x_{it} is the marginal cost of production at a given level of output as $\partial \mathcal{L}_{it} / \partial Q_{it} = x_{it}$. Then we can generate the following expression after some calculus:

$$\frac{\partial Q_{it}(\cdot)}{\partial X_{it}^j} \frac{X_{it}^j}{Q_{it}} = \frac{1}{x_{it}} \frac{P_{it}^{X^j} X_{it}^j}{Q_{it}}. \tag{9}$$

The equation (9) can be rewritten following DLW (2012) such that

$$\mu_{it} \frac{P_{it}^{X} X_{it}}{P_{it} Q_{it}} = \epsilon_{it}^{X},$$

in which the output elasticity on an input X is denoted by ϵ. This expression shows that the markup is the measure for the output elasticity on an input divided by the share of an input's expenditure in total sales such that

$$\mu_{it} = \frac{\epsilon_{it}^{X}}{\sigma_{it}^{X}}, \tag{10}$$

where σ_{it}^{X} is the share of expenditure on input X_{it} in total sales $P_{it} Q_{it}$. This means that an estimate of the output elasticity of one variable input in production and data on the expenditure share are enough to obtain a measure of plant-level markups using production data. The expenditure share can be directly obtained from observed micro data.

This derivation is standard and has been used throughout the literature, especially the DLW model (2012). DLW model's contribution is to provide consistent estimates of the output elasticities while allowing some inputs to face adjustment costs and recover firm-specific estimates of the markup related to various economic variables.

3.4.2. Output Elasticities and Markups

For estimates of the output elasticities E_{it}, production functions are implicitly assumed to contain a scalar Hicks-neutral productivity term and common technology parameters across the set of producers. But, when taking the log of production, the overall function can be estimated by using a fully nonparametric regression, LCK model. The latter does not imply that output elasticities of inputs across firms are constant, except for the special case of Cobb-Douglas.

The production function is

$$Q_{it} = G\left(X_{it}^1, \dots, X_{it}^N, K_{it}, z_{it}; \beta\right) = F\left(X_{it}^1, \dots, X_{it}^N, K_{it}, \beta\right) \exp(z_{it}),$$

in which a set of common technology parameters β govern the transformation of inputs to units of output, combined with the firm's productivity z_{it}.

This expression contains most specifications used in empirical work such as the translog production function. The main advantage of restricting production technologies of this form is proxy methods suggested by LCK, DLW, and OLS to obtain consistent estimates of the technology parameters β at the second stage. At the first stage, the total function of production G will be estimated. We consider the log version of equation (10) given that the output elasticity of a variable input j, $\epsilon_{it}^{X^j}$ is given by $\partial \ln \mathrm{G}(\cdot)/\partial \ln X_{it}^j$ and is by definition independent of a firm's productivity level.

We implicitly assume that measurement errors occur in the output data and we assume unanticipated shocks to production, which we combine into v_{it}. It is assumed that the log output is given by $q_{it} = \ln Q_{it} + v_{it}$, where v_{it} are unanticipated shocks to production and i.i.d. (independent, identically distributed) shocks including measurement error. Also, the first stage of our estimation separates the overall production part and the measurement error from the data. It is important to emphasize that we explicitly count on the fact that firms do not observe v_{it} before optimal input decisions.

Therefore, the production function we estimate for each industry separately, is defined as

$$q_{it} = f(x_{it}; \beta) + z_{it} + v_{it},$$

in which we collect all variable inputs in x_{it}, and β contains all relevant =coefficients. We consider flexible approximations to f (·), therefore we can use the LCK model, and explicitly write the production function we estimate on the data in general terms. For instance, our main empirical specification relies on any functional form that implies that f (-) is approximated by a fully non-parametric specification (LCK model), or a second order nonparametric series where all (logged) inputs, (logged) inputs squared, and interaction terms between all (logged) inputs are

included (DLW model). We recover the translog production function when we drop higher-order and interaction terms. The departure from the translog production function (DLW model) is important for our purpose, to compare the empirical results.

Our fully nonparametric approach can nest various specifications of the production function, and only need the proper order of approximation of production functions at the second stage of the structural estimation framework. However, in order to obtain consistent estimates of the production function at the second stage, we need to control for unobserved productivity shocks, which are potentially correlated with input choices such as the insight of OP and LP models, and we use the DLW model approach while relying on materials to proxy for productivity. In this case, we do not need to reconsider the underlying dynamic model when considering modifications to OP setup when dealing with additional state variables. We describe the estimation framework while relying on a dynamic control for capital and discuss the additional assumptions.

We follow the DLW model (2012) and use material demand,

$$m_{it} = m_t(k_{it}, z_{it}, x_{it}),$$

to proxy for productivity by inverting m(.), where we collect additional variables potentially affecting optimal material demand choice in the vector x_{it}. The inclusion of these additional control variables shows the only restriction we impose on the underlying model of competition (DLW model). Once those variables are appropriately accounted for in the estimation routine to obtain output elasticities, we can analyse how markups are different across firms and time, and how they relate to firm-level characteristics such as the globalisation or export status.

$z_{it} = m_t^{-1}(k_{it}, m_{it}, x_{it})$ is used to proxy for productivity in the production function estimation. The use of a material demand equation to proxy for productivity is important for researchers considering multicollinearity and estimating output elasticities and markups. Especially, as long as $\partial m/\partial z > 0$ conditional on the firm's capital stock and variables captured by $x_{it}, m_t^{-1}(k_{it}, m_{it}, x_{it})$ can be used to proxy for z_{it} being used to index a firm's productivity. In this setting, the DLW

model (2012) finds it useful to refer to Melitz and Levinsohn (2006) who also rely on intermediate inputs to proxy for unobserved productivity while allowing for imperfect competition. Melitz and Levinsohn (2006) show that this monotonicity condition holds as long as more productive firms do not set lower markups than less productive firms. This is the main part of the DLW model's idea.

3.4.3. Steps for Estimating Markups

Basically, our analysis departs from De Loecker and Warzynski (2012) and gives up on identifying any parameter at the first stage since conditional on a nonparametric function in capital, materials, and other variables affecting input demand, identification of the labour coefficient is not plausible. Even though they use nonparametric series regression with inter-variable, high-order components, we use the fully nonparametric regression with continuous and discrete data. We are concerned with more flexible production functions and allow for an undefined functional form between the various inputs, identification of the labour coefficients at the first stage.

Our procedure consists of two steps and follows the DLW model. However, let us consider a value-added production function with the general form, which is given by

$$q_{it} = \Phi(k_{it}, l_{it}, m_{it}, \iota_{i,}\iota_{t}, x_{it}) + v_{it}, \tag{11}$$

also for the comparison with the DLW model, given by

$$q_{it} = \beta_l l_{it} + \beta_k k_{it} + \beta_u l_{it}^2 + \beta_{kk} k_{ik}^2 + \beta_{lk} l_{it} k_{it} + z_{it} + v_{it}, \tag{12}$$

in which lower case means the natural logarithms. k_{it} and l_{it} are log labour and log capital in firm i in period t and q_{it} denotes log value-added, and l_i and l_t in (11) are the individual and time identifiers as categorical explanatory variables.

At the first stage we run a fully nonparametic kernel regression (LCK model) of (11), then we obtain estimates of expected output ($\widehat{\Phi}_{it}$) and an estimate for v_{it}. Expected output is given by

$$\widehat{\Phi}_{it} = \frac{\sum_{i=1}^{n} q_{it} K_{\delta}(k_{it}, l_{it}, m_{it}, \iota_{i,} \iota_{t}, x_{it})}{\sum_{i=1}^{n} K_{\delta}(k_{it}, l_{it}, m_{it}, \iota_{i,} \iota_{t}, x_{it})}, \tag{13}$$

in which *K* is the kernel function for the vector of mixed variables[6]. For the DLW model,

$$\widehat{\Phi}_{it} = \hat{\beta}_{l} l_{it} + \hat{\beta}_{k} k_{it} + \hat{\beta}_{u} l_{it}^{2} + \hat{\beta}_{kk} k_{ik}^{2} + \hat{\beta}_{lk} l_{it} k_{it} + \hat{f}_{t}(k_{it}, l_{it}, m_{it}) + v_{it},$$

in which $\hat{f}_t$ is estimated by high-order polynomial series of k_{it}, l_{it}, and m_{it}.

The second stage estimates coefficients for the production function through the law of motion for productivity such that

$$z_{it} = g(z_{it} - 1) + \eta_{it}.$$

Following the DLW model, we allow for the potential of additional (lagged and observable) decision variables to affect current productivity outcomes (in expectation), in addition to the standard inclusion of past productivity. By allowing plant-level decisions such as export participation and investment, which directly affect a firm's future profit, the DLW model tackles concerns of De Loecker (2010), who discusses potential problems of restricting the productivity process to be completely exogenous.

After the first stage, we can compute productivity for any value of β, where $\beta = (\beta_l, \beta_k, \beta_u, \beta_{kk}, \beta_{lk})$, using $z_{it} = \widehat{\Phi}_{it} - \beta_l l_{it} - \beta_k k_{it} - \beta_u l_i^2 - \beta_{kk} k_{ik}^2 - \beta_{lk} l_{it} k_{it}$. The innovation to productivity given β, $\eta_{it}(\beta)$is recovered by regressing $z_{it}(\beta)$ on its lag $z_{it-1}(\beta)$. Then, we use generalised moment conditions to estimate parameters of the production function such that

$$\mathbb{E}\left[\eta_t(\beta) \times (l_{it-1}, l_{t-1}^2, l_{t-1} k_t, k_{it} k_{ik}^2)'\right] = 0.$$

6 Please refer to Racine (2008) and Racine and Li (2004) for details of fully nonparametric estimation with continuous and discrete data, and how to find optimal smoothing parameters for discrete data. Also, see Appendix A for the basics of nonparametrics.

The moments above are from the DLW model and exploit the fact that the investment is assumed to be decided at a period ago and therefore should not be correlated with the innovation in productivity. We use lagged labour to identify the coefficients on labour since current labour is expected to react to shocks to productivity, and hence $\mathbb{E}[l_{it}\eta_{it}]$ is expected to be nonzero. In fact, DLW (2012) require input prices to be correlated over time while using lagged labour as a valid instrument for current labour, and they already find very strong evidence for that requirement by running various specifications that essentially relate current wages to past wages.

The estimated output elasticities are computed using the estimated coefficients of the production function. Under a translog value-added production function, the output elasticity for labour (1) is given by

$$\hat{\epsilon}_{it}^{l} = \hat{\beta}_l + 2\hat{\beta}_u l_{it} + \hat{\beta}_{lk} k_{it}.$$

In addition, a Cobb-Douglasproduction implies that the output elasticity of labour is simply given by $\hat{\beta}_l$. Finally, using expression (10) and our estimate of the output elasticity, we compute markups directly. However, we only observe $\tilde{Q}_{it}$, which is given by $Q_{it}\exp(v_{it})$. The first stage of our procedure gives us an estimate for v_{it} and we use it to compute the expenditure share such that

$$\hat{\sigma}_{it}^{X} = \frac{P_{it}^{X} X_{it}}{P_{it} \frac{\hat{Q}_{it}}{\exp(\hat{v}_{it})}}.$$

This correction like the DLW model is important as it removes any variation in expenditure shares resulting from variation in output not correlated with $\Phi\left(k_{it}, l_{it}, m_{it}, \iota_{i,}\iota_{t}, x_{it}\right)$, or output variation not related to variables impacting input demand including input prices, productivity, technology parameters, and market characteristics, such as the elasticity of demand and income levels. These estimates for the markup as given by equation (10) for plant i at time t are computed while allowing for considerable flexibility in the production function, consumer demand, and competition (DLW [2012]).

4. Empirical Results

In this section, we use our empirical model to estimate markups for Korean manufacturing firms, and test whether exporters and non-exporters, also large and small plants have, on average, different markups. In addition, we substantially investigate how markups change with correlation with market share and export status, additionally, industry import penetration, and as such we are the first, to our knowledge, to provide robust econometric evidence of this relationship with unbalanced fixed effect regression and dynamic unbalanced panel regression.

After estimating the output elasticity of labour and materials, we can compute the implied markups from the first order conditions as described above. We use our markup estimates to discuss several major findings. First, we compare our markup estimates to the DLW model and the OLS model. Second, we look at the relationship between markups and plant-level export status and market size, and industry import penetration effect in both the cross-section and the time series. Third, we briefly discuss the relationship between markups and other economic variables.

4.1. Background and Data

We use a plant-level dataset covering firms selected in Korean manufacturing during the period 1980-2001. The data are provided by the Korean Statistical Office and contains plant-level accounts for an unbalanced panel of 91,522. We have the information about market entry and exit, as well as detailed information on plant-level export status and export sales. At every point in time t, we know whether the firm is a domestic producer, an export entrant, an export quitter, or a continuing exporter. Table 1 provides some summary statistics about numbers of observations, observation period, manufacturing industries, and plants in data. In addition, Table 2 presents basic statistics of input variables related to production, value-added, export, material cost, labour and capital. The unit of variables except monthly average employees is Mil. KRW.

Table 1 Data Statistics

This table lists numbers of observations, observation period, manufacturing industries, and plants in data.

	Value
Number of Observations	576,690
Observation Period	> 5 year
Number of Industries	69
Number of Plants	91,522

Table 2 Statistics of Input Variables

This table lists basic statistics of input variables related to production, value-added, export, material cost, labour and capital. The unit of variables except monthly average employees is Mil. KRW.

Variable	Min	Median	Max	Mean	Std. Dev.
Nominal Production	2	400	17,100,000	4,150	81,154
Nominal Export	0	0	8,466,105	1,230	44,315
Nominal Material Cost	0.2	161	9,288,284	2,271	46,490
Real Material Cost	0.0	3.1	140,137	42	819
Monthly Average Employees	2	13	33,553	45	315
Property, Plant and Equipment	0.5	141	9,041,855	2,010	43,344
Real Production	2.1	495	16,500,000	4,676	82,190
Real Value-Added	0.0	195	5,107,007	1,461	26,035

4.2. Estimated Markups

We obtain an estimate of each plant's markup and unobservable productivity shock (or total factor productivity, TFP) and compare the average or median with the DLW and the OLS approach (simple regression of the first stage without the second stage of structural estimation) in Table 3. Although our focus is not so much on the exact level of the markup and TFP, we want to highlight that the estimated markups and TFP are comparable to those obtained with different methodologies, but are different in an important way.

Our procedure generates industry-specific production function coefficients,

which in turn deliver firm-specific output elasticity of variable inputs. The latter are plugged in the FOC of input demand together with data on input expenditure to compute markups. We list the median markup using a set of specifications to highlight our results in Table 3. We first present results using our standard methods using the LCK model. We present our results using value-added functions (for value-added production functions, we rely on the output elasticity of labour to compute markups), allowing for nonparametric series regression (DLW model) and a conventional OLS model (CD production).

Table 3 | Statistics of TFPs and Markups

This table lists the statistics of TFPs and markups estimated by a local constant kernel (LCK) model, a De Loecker and Warzynski (DLW, 2012) model, and an OLS model. The root mean squared error (RMSE) shows the deviation of fitted value-added (VA) from real value-added. The lower panel shows correlations of LCK, DLW, and OLS markups.

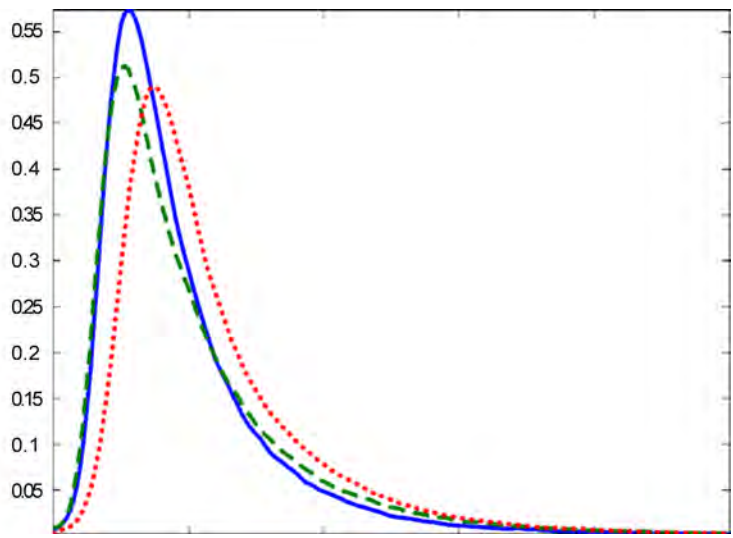
Model	RMSE	1%	Median	99%	Mean	Std. Dev.
LCK						
q	0.39					
TFP		1.08	3.45	4.55	3.32	0.65
Markup		0.41	1.61	8.93	2.09	2.51
DLW						
q	0.65					
TFP		0.72	3.45	6.32	3.33	0.98
Markup		-0.57	1.68	9.75	2.21	2.27
OLS						
Markup		0.62	2.05	9.35	2.57	2.39
Correlation	LCK	DLW		OLS		
LCK	1					
DLW	0.54	1				
OLS	0.87	0.48		1		

As can be seen in Table 3, the RMSE of the LCK model is much lower than that of the DLW model. This means the measurement error from the LCK model is estimated to be small as long as suitable to data compared to the DLW model. In addition, the median of the LCK model

is slightly lower than that of the DLW model, but much lower than that of the OLS model. The literature argues that the simple OLS model (based on CD function) has biased estimates for coefficients so that markup estimates from the OLS model might have relatively upward-bias compared to other structural estimation. However, the interesting thing is that OLS markups are higher correlated to LCK markups than DLW markups. The correlation between LCK and DLW markups is only 0.54, which is much lower than we expect because LCK and DLW markups basically share the estimation framework except the first stage for Φ_{it}. Figure 1 shows distributions of markups estimated by the LCK, DLW, and OLS models, respectively, which are left-skewed sequentially by list. In addition, Figure 2 presents

Figure 1 Distributions of Markups According to Estimation Models

This figure shows distributions of markups estimated by a local constant kernel (LCK) model, a De Loecker and Warzynski (DLW: 2012) model, and an OLS model. The vertical line shows the frequency of distributions, and the horizontal line shows markups from 0 to 10 in the figure. The solid line represents the distribution of LCK markups, the dashed line is for DLW markups, and the dotted line is for the distribution of OLS markups.



distributions of LCK markups over time from 1980 to 2001. As time goes by, the distributions of markups are getting denser and lower, which can be interpreted as the effect on firm's markups from changes in the competitive environment and the globalisation of the Korean economy.

Figure 2 Distributions of Markups over Time

This figure shows distributions of LCK markups and standard deviations of LCK markups and log of sales over time from 1980 to 2001. The vertical line in the upper panel shows the frequency of distributions over time, and the horizontal line shows markups from 0 to 5 in the figure. The arrow shows the direction of medians of markups over time. The solid line in the lower panel represents standard deviations of LCK markups and the dashed line depicts standard deviations of log sales in real terms.

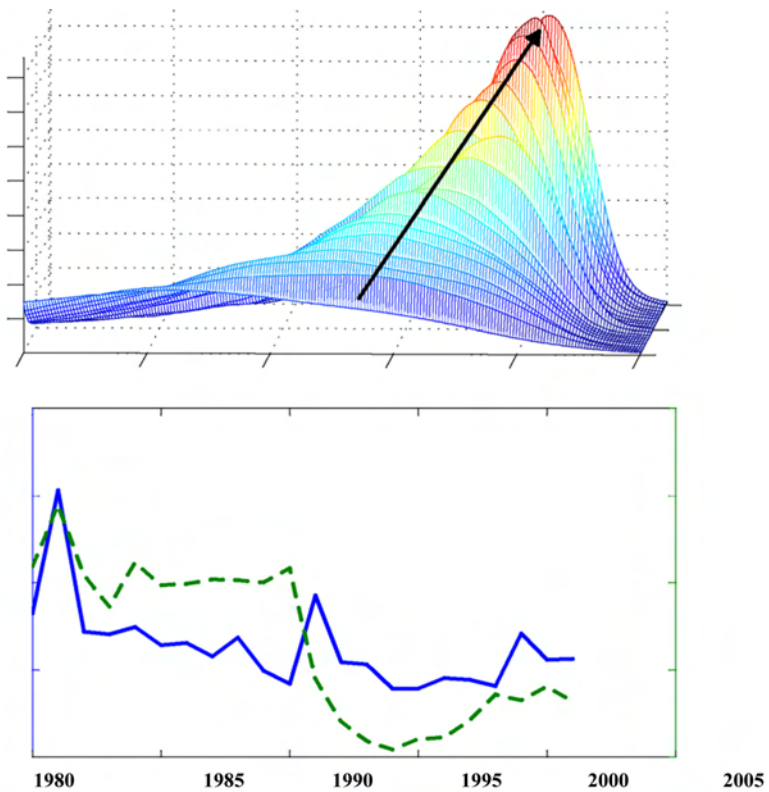


Table 4 presents means of four groups' markups (LCK model) as independent sorts of size and globalisation (export status). The small plants are in the lower 30 percent of sales in each industry at each time period, and the large plants are in upper 30 percent of sales in each industry at each time period, and the other sort is exporter or non-exporter.
As can be seen in Table 4, mean differences between large firms and small firms given export status is relatively larger than mean differences between exporters and non-exporters given firm size. We can interpret this to mean that firm markups are affected by firm size rather than by firm globalisation strategy. Figures 3-5 show distributions of exporters and non-exporters, large and small plants, and four groups' markups as independent sorts of size and globalisation. As can be seen from the lower panel in Figure 5, mean and median differences of large and small firms' markups given the export status is bigger than those of exporters and non-exporters given firm size over time. However, we can see that mean and median differences decrease over time, which is contrary to the notion that the polarisation between large and small or exporters and non-exporters would be deteriorating over time. This phenomenon might occur as a result of stronger competition in the industry, in other words, the markup gap decreases in the degree of competition intensified over time even though the innovation polarisation deteriorates over time.

Table 4 Means and Differences of Markups

This table lists means of four groups' markups (LCK model) as independent sorts of size and globalisation. The small plants are in the lower 30 percent of sales in each industry, the large plants are in the upper 30 percent of sales in each industry, and the other sort is exporter or non-exporter. Numbers of plants are annual averages from1980 to 2001. t-statistics in parentheses are for mean differences, defined as mean difference divided by the standard error (the standard deviation of mean difference divided by the square root of number of years).

	Means of Markups		
	Exporter(A)	Non-Exporter(B)	(A)-(B)
Large Plant(C)	3.63	2.84	0.79 (12.7)
Small Plant(D)	2.05	1.97	0.08 (2.96)
(C)-(D)	1.57	0.86	
	(16.4)	(16.2)	

	Numbers of Plants	
	Exporter	
	Small Plant	Large Plant
	2,569	7,551

Figure 3 | Distributions of Exporters and Non-Exporters' Markups

This figure shows distributions of exporters' and non-exporters' markups. The solid line in the upper panel represents the distribution of exporters' markups, and the dashed line is for the distribution of non-exporters' markups. The solid line in the lower panel shows the difference between medians of exporters' and non-exporters' markups over time, and the dashed line depicts the difference between means of exporters' and non-exporters' markups over time.

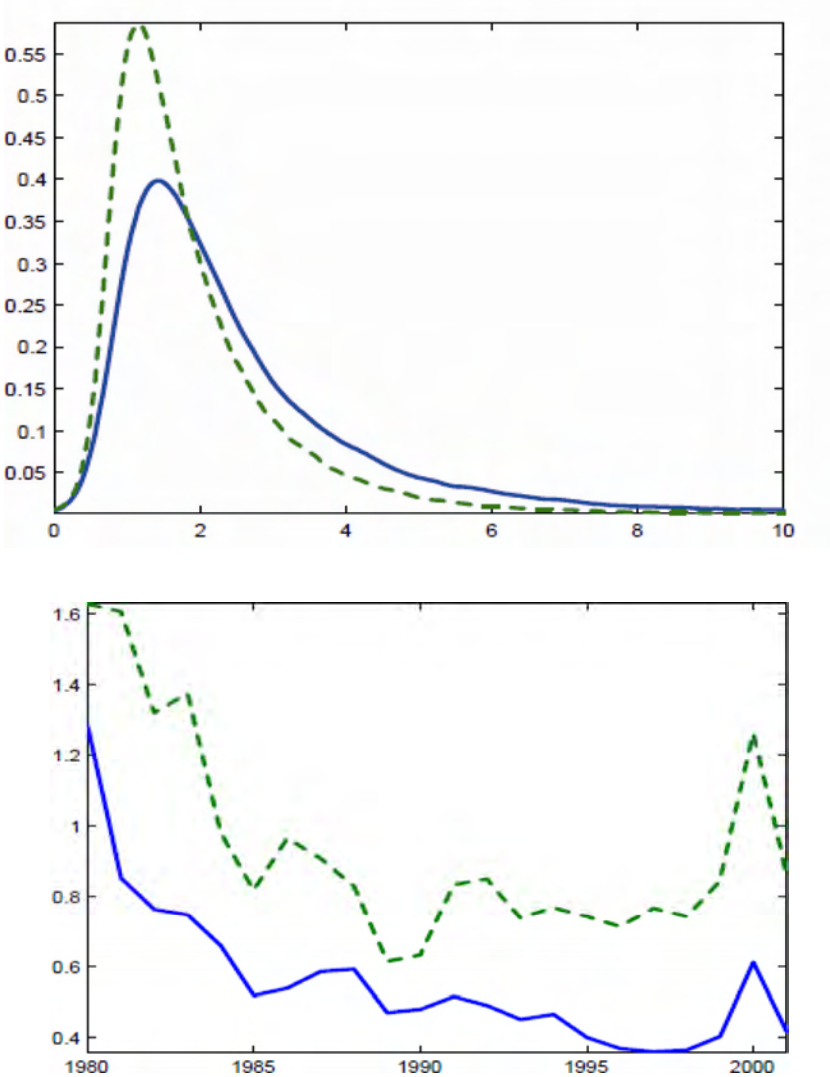


Figure 4 | Distributions of Small and Large Plants' Markups

This figure shows distributions of small plants' and large plants' markups. The small plants are in the lower 30 percent of sales, and the large plants are in the upper 30 percent of sales. The solid line in the upper panel represents the distribution of large plants' markups, and the dashed line is for the distribution of small plants' markups. The solid line in the lower panel shows the difference between medians of large and small plants' markups over time, and the dashed line depicts the difference between means of large and small plants' markups over time.

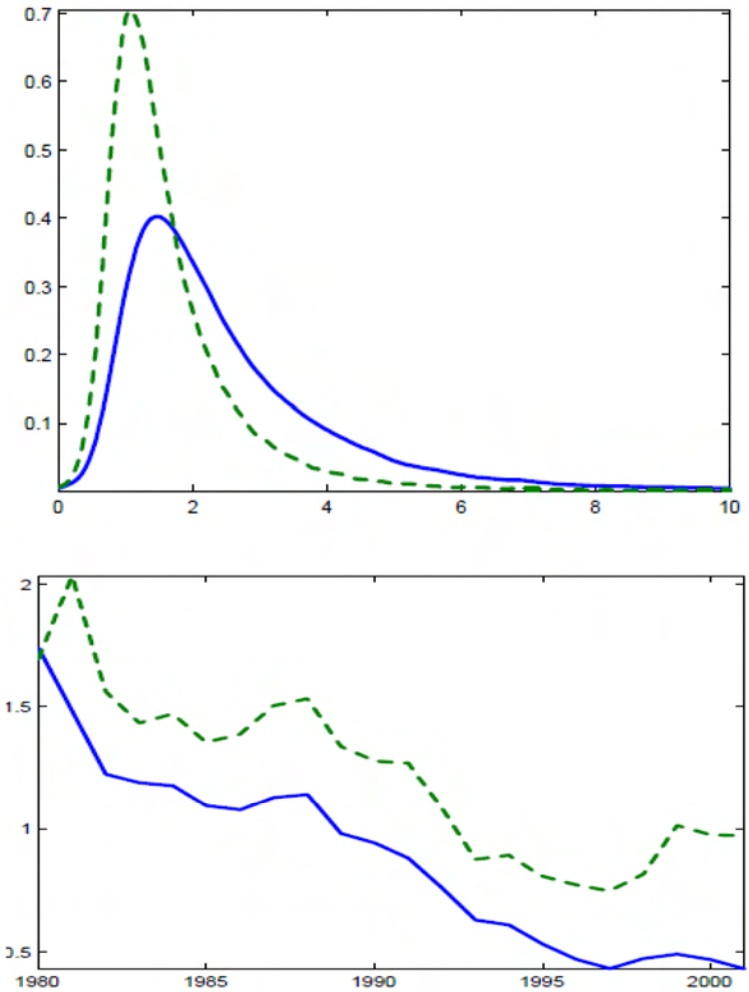
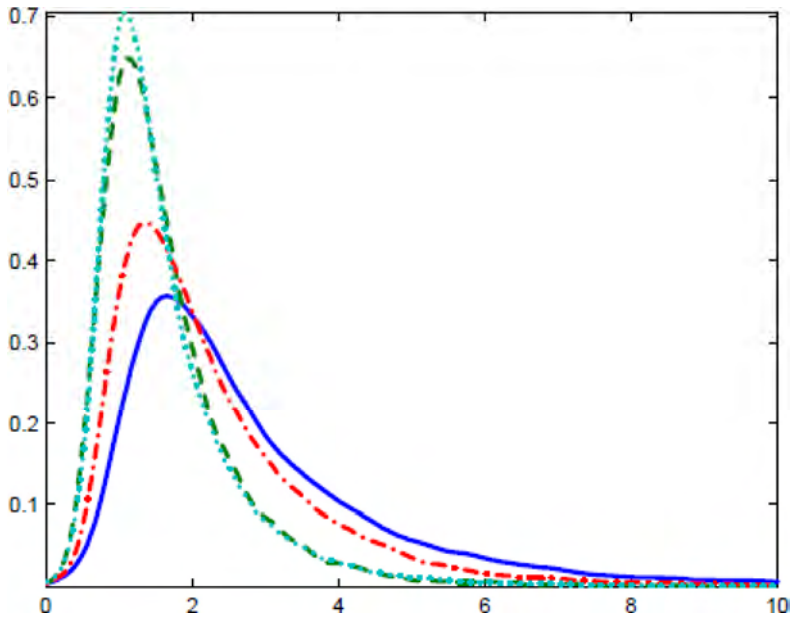
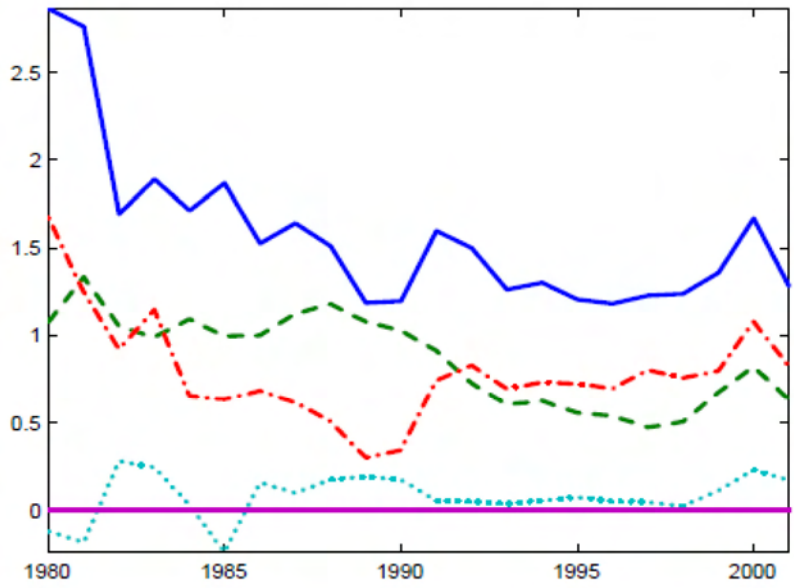


Figure 5 | Distributions of Markups of Plants sorted by Size and Globalisation

This figure shows distributions of four groups' markups as independent sorts of size and globalisation. The small plants are in the lower 30 percent of sales in each industry, the large plants are in the upper 30 percent of sales in each industry, and the other sort is exporter or non-exporter. The solid line in the upper panel represents the distribution of large and exporting plants' markups, the dashed line depicts the distribution of small and exporting plants' markups, the dashed-dotted line represents the distribution of large and non-exporting plants's markups, and the dotted line shows the distribution of small and non-exporting plants' markups. The solid line in the lower panel shows the difference between medians of large and small exporters' markups over time, the dashed line depicts the difference between medians of large and small non-exporters' markups, the dashed-dotted line shows the difference between medians of large exporters and non-exporters over time, and the dotted line represents the difference between medians of small exporters and non-exporters over time.





4.3. Unbalanced Panel Data Analysis for Markups

We can now turn to the main focus of our application—how markups on average are affected by size, globalisation, and productivity shock and whether markups change when the import penetration in industry increases. We discuss unbalanced panel data analysis for markups in fixed effects regression and dynamic panel regression.

The estimation framework introduced above was not explicit about firms selling in multiple markets. In light of our application we want to stress that our measure of markups for globalisation is a share-weighted average markup across the multi-markets, where the weight by market is the share of an input's expenditure used in production sold in that market. We can correctly compare markups across producers and time without requiring additional information on input allocation across production destined for different markets. To compare markups across markets within a plant, we do require either more data or more theoretical structure to pin down the input allocation by final market.

Given plant-specific markups, we can simply relate a plant's markup

to its size and globalisation (export status) in a regression framework. As noted above, we are not interested in the level of the markup and instead we estimate the percentage difference in markups depending on the plant's size (market share in industry and export status).

The unbalanced panel specification we take to the data is given by

$$\ln \hat{\mu}_{ijt} = X_{ijt}\gamma + \iota_i + \iota_j + \iota_t + \xi_{ijt},$$

in which ι_i, ι_j and ι_t are individual, industry, and time effects, respectively. X_{ijt} is an independent vector with industrial import penetration ratio, log (z_{it}), log (X_{ijt}/L_{ijt}), log(market share$_{jjt}$), export dummy, and export dummy x log(market shareijt). We control for labour and capital use, log(K_{ijt}/L_{ijt}), in order to capture differences in factor intensity, as well as full year-industry inter-actions to take out industry specific aggregate trends in markups (ι_j). We collect all the controls in a vector X_{ijt} with γ the corresponding coefficients.

We rely on our approach to test whether, on average, exporters have different markups as well as a different slope for exporters' market share. The latter, to our knowledge, has not been documented and we see this as a first important set of results. We are interested in the coefficients on the various control variables, so further below we will discuss the separate coefficients of other economic variables such as total factor productivity and industry import penetration. We estimate this fixed effect regression at the manufacturing level and include a full interaction of year and industry dummies. Once we have estimated coefficients of export dummy and export dummy x log(market share$_{ijt}$), we can compute the level of markup difference by applying the percentage difference to the constant term, which captures the domestic markup average. We denote this markup ratio between exporters' markup μ_{ijt}^{E} and non-exporters' markup μ_{ijt}^{N}, and we compute it by applying

$$\mathbb{E}\left[\frac{\mu_{ijt}^{E}}{\mu_{ijt}^{N}} \mid X_{ijt} \text{ except export status}\right] = \exp\left[\gamma^{E} + \gamma^{E\times\log(\text{market share})}\log(\text{market share}_{ijt})\right]$$

after estimating the relevant parameters. Table 5 presents our results.

Table 5 | Market Share and Export Effects on LCK Markups in Unbalanced Panel

This table shows results of fixed effect regressions in unbalanced panel data for markups estimated by local constant kernel (LCK) model such as

$$\ln \hat{\mu}_{ijt} = X_{ijt}\gamma + \iota_i + \iota_j + \iota_t + \xi_{ijt},$$

in which l_i, l_j, and l_t are individual, industry and time effects, respectively. X_{ijt} is an independent vector with industrial import penetration ratio, log(z_{it}), log(K_{ijt}/L_{ijt}), log(market share$_{jjt}$), export dummy, and export dummy x log(market share$_{ijt}$) - ** and * refer to the statistical significance levels at 1 percent and 5 percent, respectively. Robust standard errors in brackets are clustered within plants.

	(1)	(2)	(3)	(4)	(5)
Import Penetration	0.000		0.000		
	[0.000]		[0.000]		
Log(z_{ijt})	1.333**	1.065**			
	[0.027]	[0.018]			
Log(K_{ijt}/L_{ijt})	0.071**	0.072**	0.065**	0.062**	0.062**
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Log(market share$_{jjt}$)	-0.077**	-0.039**	-0.016**	0.012**	0.016**
	[0.002]	[0.001]	[0.002]	[0.001]	[0.001]
Dummy(exporter)	0.344**	0.329**	0. 176**	0.131**	0.042**
	[0.015]	[0.013]	[0.018]	[0.013]	[0.002]
Dummy(exporter)	0.047**	0.040**	0.022**	0.011**	
xLog(market share$_{ijt}$)	[0.002]	[0.001]	[0.002]	[0.001]	
Industry dummy (L~)	yes	yes	yes	yes	yes
R-sq: within	0.31	0.40	0.14	0.27	0.27
Num. of Plants	61,549	78,803	61,557	78,812	78,812
Num. of Obs	320,385	565,899	320,679	566,756	566,756

We run the fixed effect regression for the various estimates of the markups as described above. The parameter γ^{E} is estimated very significantly in all specifications (1)-(5) and values are between 0.042 and 0.344, which means that the exporters' markup is, on average, about 4.2 percent to 34.4 percent greater than non-exporters' markup, and values for coefficient - $\gamma^{E\times\log(\text{market share})}$ range from 0.011 to 0.047.

The parameter for the log market share ranges from -0.077 to 0.016. As expected, all the results except base of market share level relying on translog technology are very similar because the variation in markups is almost identical across the various specifications. One important message from this table is that the parameter of market share does not have consistent signs, suggesting that this unbalanced fixed effects regression might have omitted variables.

Under assumptions of dynamic unbalanced panel data analysis of Arellano and Bond [1991], we take to the data is given by

$$\Delta \ln \hat{\mu}_{it} = \alpha \Delta \ln \hat{\mu}_{it-1} + \Delta X_{it}\gamma + \Delta \xi_{it},$$

in which X_{it} is an independent vector with industrial import penetration ratio, $\log(z_{it})$, $\log(K_{it}/L_{it})$, $\log(\text{market share}_{it})$ export dummy, export dummy x $\log(\text{market share}_{it})$, and time dummy. The second lags of $\log(z_{it})$, $\log(K_{it}/L_{it})$, $\log(\text{market share}_{it})$, export dummy x $\log(\text{market share}_{it})$, and the first differences of industrial penetration ratio, export dummy, and time dummy are used as instrument variables in difference GMM system. Table 6 presents our results. The parameter γ^{E} is estimated very significantly in all specifications (1) - (5) and values are between 0.054 and 0.396, which are slightly higher than values in Table 5, and values for coefficient $\gamma^{E\times\log(\text{market share})}$ range from 0.010 to 0.043, which are similar to results of fixed effects regressions. The parameters for the log market share in all specifications have robust positive signs. The significances for the import penetration ratio are weak, thus we need to consider other variables to capture the industrial characteristics. In addition, similarly to DLW (2012), TFP increases the markup on average by 16.3 percent to 24.0 percent.

Table 6 Market Share and Export Effects on LCK Markups in Dynamic Unbalanced Panel

This table shows results of difference GMMs in dynamic unbalanced panel data (Arellano and Bond [1991]) for markups estimated by using a local constant kernel (LCK) model, such as

$$\Delta \ln \hat{\mu}_{it} = \alpha \Delta \ln \hat{\mu}_{it-1} + \Delta X_{it}\gamma + \Delta \xi_{it},$$

in which X_{it} is an independent vector with industrial import penetration ratio, log(z_{it}), log(K_{it}/L_{it}), log(market share$_{it}$), export dummy, export dummy x log(market share$_{it}$), and time dummy. The second lags of log(z_{it}), log(K_{it}/L_{it}), log(market share$_{it}$), export dummy x log(market share$_{it}$), and the first differences of industrial penetration ratio, export dummy, and time dummy are used as instrument variables in difference GMM system. ** and * refer to the statistical significance levels at 1 percent and 5 percent, respectively. Robust standard errors in brackets are estimated by the finite-sample corrected two-step covariance matrix.

	(1)	(2)	(3)	(4)	(5)
Log(markeup$_{it,1}$)	0.169**	0.175**	0.179**	0.185**	0.184**
	[0.005]	[0.004]	[0.006]	[0.004]	[0.004]
Import Penetration	0.000		0.001*		
	[0.000]		[0.000]		
Log(z_{it})	0.240**	0.163**			
	[0.075]	[0.038]			
Log(K_{it}/L_{it})	0.079**	0.045**	0.074**	0.036**	0.034**
	[0.008]	[0.005]	[0.009]	[0.005]	[0.006]
Log(market share$_{it}$)	0.080**	0.115**	0.079**	0.109**	0.118**
	[0.015]	[0.010]	[0.016]	[0.011]	[0.012]
Dummy(exporter)	0.121	0.396**	0.054	0.383**	0.063**
	[0.134]	[0.102]	[0.152]	[0.107]	[0.006]
Dummy(exporter)	0.010	0.043**	0.002	0.042**	
xLog(market share$_{ijt}$)	[0.017]	[0.013]	[0.020]	[0.014]	
Num. of Plants	48,674	76,472	48,686	76,502	76,502
Num. of Obs	199,926	370,917	200,203	371,701	371,701

For comparison of the DLW and OLS models, Table B.1-B.4 shows the results of unbalanced fixed effects and dynamic panel regressions. Tables for the DLW model show that the DLW model still has the negative signs for market share, and weak consistent signs for the parameter of productivity shock. The OLS model has negative signs for variables related to export dummy. Therefore, the results of the LCK model are robustly consistent with industrial organization and international economic theories compared with those of the DLW and OLS models.

For the last exercise, we directly quantify how import competition can influence the dispersion of markups. Since large firms set higher markups than small firms, the dispersion of markups is closely related to the gap in markups between small and large firms. Table 7 reports industry panel fixed-effect regressions. It shows that import competition measured as import penetration reduces the differential in firm markups. In the first column, the standard deviation of industry markups decreases by about 0.07 percent with a 1 percentage point increase in import penetration. The inequality of markups between firms decreases with intensified international competition, as the theory predicts.

Table 7 | Import Penetration Effect on Dispersion of Markups

This table shows results of industry panel fixed-effect regressions

$$\ln SD_{jt} = \alpha + \beta_1 \ln IMPR_{jt} + \beta_2 \ln \mathrm{K}/L_{jt} + \beta_3 \ln \mu_{jt}$$

in which $5D_{jt}$ is a standard deviation of markup in an industry j, IMPR is an industry import penetration, $\log(K_{jt}/L_{jt})$ is an industry capital-labour ratio, and$_{\mu it}$ is log industry average markup. Overall import penetration can be categorized into two types. One is only imports from China, and the other is imports from the rest of the world. This classification comes from Bernard, Jensen, and Schott (2006). They emphasize that the response of industry employment to import competition from low-wage countries such as China can be different from typical import competition from the rest of the world. The dependent variable of columns (1) and (2) is the standard deviation of industry markups, and columns (3) and (4) use an inter-quartile range of industry markups as a response variable. ** and * refer to the statistical significance levels at 1 percent and 5 percent, respectively.

	(1)	(2)	(3)	(4)
Import Penetration	-0.067**		0.068	
	[0.019]		[0.039]	
Import Penetration (other)		-0.099**		0.057
		[0.020]		[0.043]
Import Penetration (China)		1.162**		0.504
		[0.315]		[0.675]
Log(K_{jt}/L_{jt})	0.240**	0.211**	0.378**	0.368**
	[0.035]	[0.033]	[0.069]	[0.072]
Log(μ_{i_t})	0.040**	0.056**	0.113**	0.119**
	[0.016]	[0.015]	[0.031]	[0.033]
Num. of Industry	16	16	16	16
Num. of Obs	130	130	130	130

We further estimate whether import competition from low-wage countries such as China has a stronger effect on domestic firm behaviour. Interestingly, the second column reveals positive signs for the effect of import penetration from China on markup dispersion, while import penetration from the rest of the world retains its negative effect on markup dispersion. In some sense, it is embarrassed, but it can be

possible if forces of import competition are concentrated on only very small firms. Products from China are usually low quality and low price. These types of goods are commonly made by domestic small firms. It would be plausible to assume that if the good markets are segmented by high and low quality goods, and the substitution between high and low quality products is very low, import competition from low-wage countries would affect only low price goods. If domestic small firms face stronger competition from low wage countries, they have to cut prices to stay in the market, whereas large firms with high quality goods are generally able to maintain their prices. In this case, the standard deviation of markups rises with increased import penetration.

For our robustness check we use an inter-quartile range of markups as markup dispersion in each industry as a dependent variable. However, import penetration has an insignificant effect on markup dispersion in this case. It implies that changes arising from import penetration are concentrated on the lower tail or upper tail of the support of markups. If the changes in markup dispersion occur uniformly or in overall support, the result should be the same if we use the standard deviation as a dispersion measure. We can conjecture that very small firms or very large firms are more influenced by import penetration, considering the different results of the standard deviation and inter-quartile range.

5. Conclusions

In this paper, we show that large firms decide on higher markups in each industry as they have greater market powers in integrated markets. Also, exporters set higher markups through higher observable productivity than non-exporters. This is empirically consistent with the theory that firms conditional on higher observable productivity decide on higher markups. Interestingly, even after controlling for productivity and other firm characteristics, the level of markup is proportional to the market share. A one percent increase in market share leads to a 0.080.12 percent increase in markup. It draws attention since it is the evidence that the firm strategy of price is reflected by pure market power.

To answer the question whether globalisation confers unequal

benefits on small and large firms, we generate markup distribution and find that the mean and the dispersion of markups have been decreasing over time. However, the average firm size and firm size distribution have been increasing. These patterns are predicted exactly by the theoretical model of trade. The main hurdle is to identify the effect of globalisation. In order to investigate the effect of globalisation, we use industry panel fixed-effect regressions. For the proxy of globalisation, the import penetration is used. It is a disadvantage that import penetration only captures the effect of importing as globalisation includes both imports and exports. It turns out that import competition reduces the markup gap between small and large firms, as predicted by the theory.

Methodologically, we develop De Loecker and Warzynski (2012) and control the endogeneity problem by using the difference GMM in dynamic unbalanced panel data suggested by Arellano and Bond (1991). Compared to De Loecker and Warzynski (2012), our estimate of markups has smaller errors and reasonable levels of average and median of markups.

This paper has an important message regarding enterprise policies. A widening of the performance gap, measured as output or sales, between large and small firms, cannot be interpreted as meaning that globalisation negatively affects the welfare of consumers. It is likely that globalisation strengthens competition between all firms, so the gap in prices or markups shrinks as a result of the selection effect, which benefits consumers. Therefore, protective policies to shield SMEs from the effects of globalisation may interfere with the selection process and harm productivity growth.

References

Ackerberg, D., K. Caves and G. Frazer (2006), *Structural Identification of Production Functions*. Unpublished.

Atkeson, A. and A. Burstein (2008), 'Pricing-to-Market, Trade Costs and International Relative Prices', *American Economic Review*, 98(5), pp.1998–2031.

Alvarez, R. and R. Lopez (2005), 'Exporting and Performance: Evidence from Chilean Plants', *Canadian Journal of Economics*, 38(4), pp.1384–400.

Arellano, M., and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies* 58, pp.277–97.

Blalock, G. and P. Gertler (2004), 'Learning from Exporting: Revisited in a Less Developed Setting', *Journal of Development Economics*, 75(2), pp.397–416.

Choi, Y.-S. and C.H. Hahn (2010). The Effects of Imported Intermediate Varieties on Plant Total Factor Productivity and Product Switching: Evidence from Korean Manufacturing. ERIA Research Project Report 2010.

Criscuolo, C. and R. Martin (2009), 'Multinationals and U.S. Productivity Leadership: Evi-dence from Great Britain', *Review of Economics and Statistics*, 91(2), pp.263–81.

Czekaj, T. and A. Henningsen (2013), 'Panel Data Specifications in Nonparametric Kernel Regression: An Application to Production Functions', *IFRO Working Paper*, Department of Food and Resource Economics, University of Copenhagen.

De Loecker, J. (2010), 'A Note on Detecting Learning by Exporting', *NBER Working Paper Series*, No. 16548, Cambridge, MA: NBER.

De Loecker, J. and F. Warzynski (2012), 'Markups and Firm-level Export Status', *American Economic Review*, 102(6), pp.2437–71.

Dixit, A.K., and J.E. Stiglitz (1977), 'Monopolistic competition and Optimum Product Diversity', *American Economic Review* 67 (June): 297–308.

Edmund, C., V. Midrigan and D. Yi Xu 2013. Competition, Markups, and the Gains from International Trade. Mimeo.

Fernandes, A. (2007), 'Trade Policy, Trade Volumes and Plant-Level Productivity in

Colombian Manufacturing Industries', *Journal of International Economics* 71, pp.52–71.

Foster, L.S., J. Haltiwanger and C. Syverson (2008), 'Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?', *American Economic Review* 98 (1), pp.394–425.

Goldberg, P.K., A.K. Khandelwal, N. Pavcnik and P. Topalova (2010), Imported Intermediate Inputs and Domestic Product Growth: Evidence from India', *Quarterly Journal of Economics* 125 (4), pp.1727–67.

Gyimah-Brempong, K. and J.S. Racine (2010), 'Aid and Investment in LDCs: A Robust Approach', *The Journal of International Trade & Economic Development: An International and Comparative Review* 19, pp.319–49.

Henderson, D.J., and L. Simar (2005)' 'A Fully Nonparametric Stochastic Frontier Model for Panel Data', *Working Paper 0519*, Department of Economics, State University of New York at Binghamton.

Hurst, E., and B. Pugsley (2011). What Do Small Businesses Do? Brookings Papers on Economic Activity, pp. 73–142.

Kugler, M., and E. Verhoogen (2008). The Quality-Complementarity Hypothesis: Theory and Evidence from Colombia. *NBER Working Paper Series*, No. 14418, Cambridge, MA: NBER..

Levinsohn, J., and A. Petrin (2003), 'Estimating Production Functions Using Inputs to Control for Unobservables', *Review of Economic Studies*, 70(2), pp.317–41.

Li, Q. and J.S. Racine (2004), 'Cross-validated Local Linear Nonparametric Regression', *Statistica Sinica* 14, pp.485–512.

Melitz, M., and J.A. Levinsohn (2006). Productivity in a Differentiated Products Market Equilibrium. Unpublished

Melitz, M.J., and G.I.P. Ottaviano 2008, 'Market Size, Trade, and Productivity', *Review of Economic Studies* 75, pp.295–316.

Oh, J. (2013). The Cyclicality of Firm Size Distribution and its Effect on Aggregate Fluctuations. Mimeo.

Olley, S.G., and A. Pakes (1996), 'The dynamics of productivity in the telecommunications equipment industry', *Econometrica*, 64(6), pp.1263–97.

Pavcnik, N. (2002), 'Trade Liberalization Exit and Productivity Improvements: Evidence from Chilean Plants', *Review of Economic Studies* 69, pp.245–76.

Racine, J.S. (2008), 'Nonparametric Econometrics: A primer', *Foundations and Trends in Econometrics* 3, pp.1–88.

Racine, J.S., and Q. Li (2004), 'Nonparametric Estimation of Regression Functions with

both Categorical and Continuous Data', *Journal of Econometrics* 119, pp.99–130.

Roberts, M.J., and D. Supina (1996), 'Output Price, Markups, and Producer Size', *European Economic Review* 40, pp.909–21.

Topalova, P., and A. Khandelwal (2011), 'Trade Liberalization and Firm Productivity: The Case of India', *The Review of Economics and Statistics*, 93(3), pp.995–1009

Wooldridge, J.M. (2009), 'On Estimating Firm-Level Production Functions using Proxy Variables to Control for Unobservables', *Economic Letters* 104(3), pp.112–14.

Appendix

A. Local Constant Kernel Regression

A.1. Basics of LCK Regression

The nonparametric model is as follows:

$$y_i = g(x_i) + u_i, i = 1, \dots, n,$$

in which the functional form g (·) is unknown. If g (·) is a smooth function, we can estimate g (·) nonparametrically using kernel methods so that we consider g (·) as the conditional mean of y given x such that

$$g(x) = \mathbb{E}[y_i | x_i = x],$$

due to the general result of nonparametric theory. We note that $\mathbb{E}[y_i|x_i = x] = \int \mathrm{y}\, f_{y,x}(x,y)dy$ can be replaced with $\int \mathrm{y}\, \hat{f}_{y,x}(x,y)dy$ with the unknown probability density function $f_{y,x}(x,y)$ estimated by the kernel method such that

$$\hat{f}_{y,x}(x,y) = \frac{1}{nh_0 \dots h_q} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \kappa\left(\frac{y - y_i}{h_0}\right), \tag{14}$$

in which $K\left(\frac{x_i - x}{h}\right) = \kappa\left(\frac{x_{i1} - x_1}{y_1}\right) \times \dots \times \kappa\left(\frac{x_{iq} - x_q}{h_q}\right)$ and where κ is a kernel function satisfying basic conditions of nonparametrics, h_0 is the smoothing parameter associated with y, and $h_0 \ \dots \ h_q$ are bandwidths for x_i. From equation (14), we obtain the estimate

$$\hat{g}(x)\frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}, \tag{15}$$

which is simply a weighted average of yi, because we can rewrite (15) as

$$\hat{g}(x)\sum_{i=1}^{n} y_i w_i,$$

in which $w_i = K\left(\frac{x_i - x}{h}\right) / \sum_{j=1}^{n} K\left(\frac{x_i - x}{h}\right)$ is the weight attached to y_i

A.2. Cross-Validation Method for Bandwidth

Once we have the continuous explanatory variables xi, the optimal bandwidth h is determined by the cross-validation method, minimizing

$$\mathrm{CV}\left(h_1, \ldots, h_q\right) = \min_h \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_{-1}(x_i)\right)^2 m(x_i), \tag{16}$$

in which $\hat{f}_{-i}(x_i) = \frac{\sum_{i\neq 1}^{n} y_i K\left(\frac{x_i - x_1}{h}\right)}{\sum_{i\neq 1}^{n} K\left(\frac{x_i - x_1}{h}\right)}$ is the leave-one-out kernel estimator of $f(x_1)$ and $m(x_i)$ is a weight function that rules out boundary observations and $\leq m(\cdot) \leq 1$. Then, the asymptotic result of optimal bandwidth is

$$n^{1/(q+4)}\hat{h} = \hat{a} \rightarrow_p \ a,$$

in which a is uniquely defined, positive, and finite to asymptotically minimize the first leading term of *CV* (*h*).

A.3. LCK Regression with Mixed Data

We now turn to a nonparametric approach with continuous and discrete variables. From a statistical point of view, smoothing discrete

variables may introduce some bias, but it is also known that it reduces the finite-sample variance resulting in a reduction in the finite-sample mean squared error of the nonparametric estimator.

Coming back to a nonparametric regression model given by

$$y_1 = g\left(x_i^c, x_i^d\right) + u_i, i = 1, \dots, n,$$

in which x^c and x^d are continuous and discrete variables, respectively. Then we define the estimate of unknown PDF as

$$\hat{f}_{y,x}(x, y) = \frac{1}{nh_0 \dots h_q} \sum_{i=1}^{n} K_\delta \left(\frac{x_i - x}{h}\right) \kappa \left(\frac{y - \mathrm{y}_i}{y_0}\right)$$
$$= \frac{1}{nh_0 \dots h_q} \sum_{i=1}^{n} W_h \left(\frac{x_i^c - x^c}{h}\right) L\left(x^d, x_i^d, \lambda\right) \kappa \left(\frac{y - \mathrm{y}_i}{h_0}\right),$$

in which $\delta = (h, \lambda), W$ is a asymmetric, nonnegative univariate kernel function, and $L\left(x^d, x_i^d, \lambda\right) = \prod_{s=1}^{r} \lambda_s^{1\left(x_{is}^d \neq x_s^d\right)}$ where $1\left(x_{is}^d \neq x_s^d\right)$ is an indicator function which equals one when $x_{is}^d \neq x_s^d$ and zero otherwise. The smoothing parameter for x^d is assumed to be $0 \leq \lambda \leq 1$. Then,

$$\hat{g}\left(x_i^c, x_i^d\right) = \frac{\sum_{i=1}^{n} y_i K_\delta\left(x, x_i^c, x_i^d\right)}{\sum_{i=1}^{n} K_\delta\left(x, x_i^c, x_i^d\right)}$$

which is analogous to equation (13).

Least squares cross-validation selects 9 = (h, A) to minimize the following function:

$$CV(h, \lambda) = \min_{h,\lambda} \sum_{\mathrm{i}=1}^{\mathrm{n}} (y_i - \hat{g}_{-i}\left(x_i^c, x_i^d\right))^2$$

in which $\hat{g}_{-i}\left(x_i^c, x_i^d\right)$ and $m\left(x_i^c, x_i^d\right)$ are the same as (16). Note that when $\lambda = 1, L(x^d, x_i^d, \lambda)$ becomes unrelated to (x^d, x_i^d). Finally, the asymptotic results of smoothing parameter δ is

$$n^{1/(q+4)}\hat{h} = \hat{a} \rightarrow_p \ a,$$
$$n^{2/(q+4)}\hat{\lambda} = \hat{b} \rightarrow_p \ b,$$

in which a and b are uniquely defined, positive, and finite to asymptotically minimize the first leading term of CV (h, λ).

B. Additional Tables

Table Appendix 1 Market Share and Export Effects on DLW Markups in Unbalanced Panel

This table shows results of fixed effect regressions in unbalanced panel data for markups estimated by a De Loecker and Warzynski (DLW, 2012) model such as

$$\ln \hat{\mu}_{ijt} = X_{ijt}\gamma + \iota_i + \iota_j + \iota_t + \xi_{ijt}.$$

Other descriptions remain the same as in Table 5.

	(1)	(2)	(3)	(4)	(5)
Import Penetration	0.000		0.000		
	[0.000]		[0.000]		
Log(zzjt)	0.656**	0.238**			
	[0.043]	[0.013]			
Log(Kzjt/Lzjt)	0.030**	0.016**	0.027**	0.016**	0.015**
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Log(market sharezjt)	-0.086**	-0.042**	-0.075**	-0.036**	-0.027**
	[0.002]	[0.001]	[0.002]	[0.001]	[0.001]
Dummy(exporter)	0.316**	0.302**	0.263**	0.262**	0.029**
	[0.016]	[0.013]	[0.015]	[0.012]	[0.002]
Dummy(exporter)	0.043**	0.035**	0.036**	0.030**	
xLog(market sharezjt)	[0.002]	[0.001]	[0.002]	[0.001]	
R-sq: within	0.14	0.27	0.12	0.27	0.27
Num. of Plants	60,843	78,231	60,872	78,289	78,289
Num. of Obs	313,937	557,260	314,374	559,753	559,753

Table Appendix 2 | Market Share and Export Effects on DLW Markups in Dynamic Unbalanced Panel

This table shows results of difference GMMs in dynamic unbalanced panel data (Arellano and Bond (1991)) for markups estimated by a De Loecker and Warzynski (DLW: 2012) model such as

$$\Delta \ln \hat{\mu}_{it} = \alpha \Delta \ln \hat{\mu}_{it-1} + \Delta X_{it}\gamma + \Delta \xi_{it}.$$

Other descriptions remain the same as Table 6

	(1)	(2)	(3)	(4)	(5)
Log(markupit-1)	0. 199**	0.201**	0.197**	0.205**	0.205**
	[0.006]	[0.004]	[0.006]	[0.004]	[0.004]
Import Penetration	-0.000		-0.000		
	[0.000]		[0.000]		
Log(zit)	0.147**	-0.083**			
	[0.051]	[0.017]			
Log(Kit/Lit)	0.068**	0.018**	0.067**	0.026**	0.027**
	[0.009]	[0.006]	[0.009]	[0.006]	[0.006]
Log(market shareit)	0.082**	0.051**	0.073**	0.068**	0.083**
	[0.017]	[0.011]	[0.016]	[0.011]	[0.012]
Dummy(exporter)	0.286*	0.627**	0.292*	0.594**	0.042**
	[0.130]	[0.097]	[0.128]	[0.099]	[0.006]
Dummy(exporter)	0.032	0.077**	0.033	0.072**	
xLog(market shareijt)	[0.017]	[0.012]	[0.017]	[0.012]	
Num. of Plants	47,608	75,658	47,648	75,752	75,752
Num. of Obs	194,394	363,654	194,777	365,723	365,723

Table Appendix 3 | Market Share and Export Effects on OLS Markups in Unbalanced Panel

This table shows results of fixed effect regressions in unbalanced panel data for markups estimated by an OLS model such as

$$\ln \hat{\mu}_{ijt} = X_{ijt}\gamma + \iota_i + \iota_j + \iota_t + \xi_{ijt}.$$

Other descriptions remain the same as Table 5.

	(1)	(2)	(3)
Import Penetration	0.001		
	[0.000]		
Log(Kijt/Lijt)	0.077**	0.089**	0.088**
	[0.001]	[0.001]	[0.001]
Log(market shareijt)	-0.084**	-0.068**	-0.056**
	[0.002]	[0.001]	[0.001]
Dummy(exporter)	0.289**	0.278**	-0.024**
	[0.018]	[0.013]	[0.002]
Dummy(exporter)	0.043**	0.039**	
xLog(market shareijt)	[0.002]	[0.001]	
Industry dummy (tj)	yes	yes	yes
R-sq: within	0.13	0.29	0.29
Num. of Plants	61,579	78,834	78,834
Num. of Obs	321,010	567,279	567,279

Table Appendix 4 | Market Share and Export Effects on OLS Markups in Dynamic Unbalanced Panel

This table shows results of difference GMMs in dynamic unbalanced panel data (Arellano and Bond (1991)) for markups estimated by an OLS model such as

$$\Delta \ln \hat{\mu}_{it} = \alpha \Delta \ln \hat{\mu}_{it-1} + \Delta X_{ijt}\gamma + \Delta \xi_{ijt}.$$

Other descriptions remain the same as Table 6

	(1)	(2)	(3)
Log(markupit-1)	0.176**	0.181**	0.181**
	[0.006]	[0.004]	[0.004]
Import Penetration	0.001**		
	[0.000]		
Log(Kit/Lit)	0.095**	0.054**	0.052**
	[0.009]	[0.005]	[0.005]
Log(market shareit)	0.038**	0.079**	0.091**
	[0.016]	[0.010]	[0.011]
Dummy(exporter)	-0.009	0.292**	0.040**
	[0.155]	[0.104]	[0.005]
Dummy(exporter)	-0.002	0.033*	
xLog(market shareijt)	[0.020]	[0.013]	
Num. of Plants	48,744	76,573	76,573
Num. of Obs	200,519	372,262	372,262

CHAPTER 6

Endogenous Product Characteristics in Merger Simulation: A Study of the U.S. Airline Industry

by
*Jinkook Lee**
(Korea Development Institute)

Abstract

Standard merger simulations focus solely on price changes while constraining the set of product characteristics to be identical pre- and post-merger. Recent papers have begun to address this issue (see, e.g. Fan, August 2013 AER). To overcome the limitations of traditional simulations, I endogenize both prices and product characteristics by specifying a two-stage oligopoly game. After estimating demand and supply system, I simulate the effect of the Delta and Northwest Airlines merger on prices, product characteristics, and welfare. The simulation results show that (i) the merged firm tends to increase product differentiation post-merger; (ii) the higher product differentiation reduces the firm's incentive to raise prices; (iii) the changes in characteristics and prices increase not only the merged firm's profit but also consumer welfare. I also compare the predicted to actual post-merger outcome and find that endogenizing product characteristics is essential to better predict the actual outcome.

* Associate Research Fellow, Department of Industry and Service Economy, KDI, Email: ljk@kdi.re.kr

1. Introduction

Until recently, standard merger simulations have focused solely on price changes while implicitly constraining the set of product characteristics to be identical pre- and post-merger.[1] However, when an industry experiences a change in market structure such as entry, exit, or a merger, firms are likely to adjust product characteristics. As examples of the airline industry, Peters (2006) shows that a merged airline tends to reduce flight frequency on segments where the merging carriers were competing with each other, and Mazzeo (2003) finds that carriers are likely to deteriorate on-time performance when markets become less competitive.[2]

Ignoring this aspect can lead to a significant bias in predicted prices in several aspects. On the demand side, the set of characteristics is an important part from which consumers derive utility. Then, it is very natural that the post-merger changes in characteristics affect consumers' choices and the resulting market shares of products. On the supply side, merging firms consolidate their production facilities and change the way of conducting operations. This induces the combined firm to search a new set of optimal characteristics based on changes in marginal and fixed cost. Further, the product repositioning influences the extent of cross-price effect merged firm internalizes. Suppose that a merged firm's products become more differentiated than before, then cross-price elasticity between them becomes weaker so that the firm has less ability to increase prices than standard simulations predict. However, traditional simulations do not consider these three channels through which optimal prices are affected. Besides the predicted prices, subsequent welfare assessments can be biased in this sense.[3]

---

1 Throughout this paper, standard merger analysis refers to the simulation method based on differentiated product demand and firm conduct in oligopolistic markets. This empirical model is widely used since Berry and Pakes (1993), Berry (1994), Werden and Froeb (1994), and Berry, Levinsohn and Pakes (1995).

2 Merger effect is not the primary focus of Mazzeo (2003), but the finding on the link between market competition and product quality is closely related to this study.

3 Besides the intuitive understanding of limits, Crawford (2012) discusses potential econometric problems associated with exogenous product characteristics.

To overcome the limitations, I endogenize not only prices but also product characteristics to analyze merger effects in the U.S. airline industry. To be specific, I aim to answer the following four questions: (*i) How does merged firm adjust product characteristics?* A few studies has addressed this issue by comparing 'actual' pre- and post-merger data. Unlike the literature, I 'simulate' post-merger characteristics based on pre-merger data and structural model, assuming that actual outcomes are not available. *(ii) How and to what extent does the product repositioning affect post-merger prices?* After simulating price changes, the paper analyzes how much of the changes is caused through each of three channels (described above). Especially by separating the magnitude of cross-price elasticity, I attempt to see a change in the firm's ability of raising prices. *(iii) How does post-merger equilibrium affect welfare?* I introduce consumer heterogeneity in preferences for the characteristics in the demand specification. Given heterogeneous consumers, merged firm can reposition various subsets of products differently. I assess welfare changes of each type of consumers as well as profit changes by each group of products. *(iv) Does endogenizing product characteristics contribute to better predicting post-merger outcome?* Although a large body of literature has displayed interests in merger simulation, there has been very little studies testing this matter. With a focus on flight frequency, I evaluate the predictive performance of my simulation.

The U.S. air travel market in the late 2000's offers an ideal environment to this research. Above all, the industry has experienced at least nine completed or on-going mergers between 2008 and 2013, including the recently approved the American and US Airways merger.[4] Second, airline mergers involve very complicated integration procedures on various levels. In terms of overall operations, they reform engineering, maintenance, crew training, network design, flight schedule, and allocation of fleets. Also production facilities such as aircraft, gates, and ticket offices are consolidated. Regarding customer service, they create single reservation system and harmonize frequent flier program. All these

4 Since the U.S. airline market was deregulated in 1979, there have been more than thirty merger cases. They exhibit a variation in merging entity types such as legacy, regional, low cost carriers (LCCs). The comprehensive list of U.S. airline mergers is available at http://www.airlines.org/Pages/U.S.-Airline-Mergers-and-Acquisitions.aspx.

consolidations can impact operational characteristics of the airline's products. Third, a comprehensive and latest dataset is publicly available from the U.S. Department of Transportation (DOT). The data used include Origin and Destination survey, Air Travel Consumer Report, On-Time Performance, T-100 Domestic Segment, and other sources from the U.S. Bureau of Transportation Statistics.

To simulate merger effects, I set up a structural model of demand and supply in differentiated product markets. The demand model uses discrete choice setting (McFadden, 1981; Berry, Levinsohn and Pakes, 1995) and particularly adopts random coefficient logit model with finite consumer types (Berry, Carnall and Spiller, 2006; Berry and Jia, 2010) to see whether the tourists and the business passengers exhibit heterogeneous preferences for price and characteristics. In the supply model, I set up a two-stage oligopoly game where firms decide optimal product characteristics - flight frequency, on-time performance, mishandled baggage rate, and denied boarding rate - at the first stage, and then choose optimal prices at the second stage. Even though the sequential choice model involves technical difficulties, it is more realistic for industries where adjusting product characteristics requires a long time.[5]

After estimating the model parameters, I predict post-merger equilibrium by using three different games: traditional model with endogenous price (Price model, G^P), a new model with endogenous characteristics and endogenous price (Full model, G^{FL}), and a hypothetical model where firms can choose only prices under pre-merger situation, and product characteristics are given by post-merger characteristics of the full model. (Hypothetical model, G^H). Since the hypothetical model does not consider the ownership consolidation, price changes in the game arise from the adjustment of characteristics rather than from the cross-price effect. I compare price changes from the three simulations, and then identify two different cross-price effects, respectively, from the price model and the full model.

This study examines the Delta and Northwest Airlines merger, which

5 For example, Fan (2013) studied the U.S. Daily Newspaper Market by using a sequential choice model. Endogenous newspaper characteristics include non-advertising space, the number of staff for opinion sections, the number of reporters, and other measures. All these are not quite changeable in a short period of time.

created the largest commercial airline in the world as of 2008. Importantly, they competed in more than 450 markets with each other. Based on its greater scale of overlapped markets than other recent merger cases, we can expect the merger effect to be considerable. Further, the integration process had been completed early enough (December 2009) for the actual post-merger data to be available. This enables an evaluation of the simulation performance.

From the simulation results, I find that (*i*) the merged firm tends to increase product differentiation post-merger. I measure a product quality by taking an inner product of the set of endogenous characteristics and their respective parameters, and then compute a change in quality of each product. The result shows that the merged firm raises the quality of primary products, but largely decreases that of secondary products.[6] (*ii*) The higher product differentiation reduces the firm's incentive to raise prices especially of the primary products. On the contrary, the firm increases prices of the secondary products substantially with intent to move passengers from secondary to more profitable primary goods. (*iii*) Consumer and producer welfare changes substantially differ from those of standard merger analyses. While the price model predicts decrease in consumer surplus for both types of passengers, the full model predicts that the business travelers get welfare gains due to the quality improvement of the primary goods, and this leads to an overall increase in consumer surplus. Regarding producer surplus, both models show that merged firm earns higher profit and competitors have less profits, but the additional gain to the merged firm is much bigger in full model. (iv) Finally, endogenizing characteristics is essential to better predict the actual outcome. Based on the comparison between the pre-merger, the simulated, and actual post-merger frequency, I find that the simulated frequency becomes closer to actual post-merger frequency. In summary, the results highlight that the analysts need to endogenize product characteristics as well, when simulating the effects of a proposed merger.

This paper contributes to the existing literature in three ways. First, it extends merger literature. Focusing on airline merger studies, one group

6 A primary product refers to a major route where passenger traffic is large, and a secondary product indicates a route where small number of passengers travel.

of papers adopts comparative analysis or reduced form model to examine changes in price, output, or welfare. Borenstein (1990), Werden, Joskow and Johnson (1991), Kim and Singal (1993), Morrison (1996), Kwoka and Shumilkina (2010), and Luo (2011) belong to this group. Another group takes a more structural approach to simulate post-merger outcomes. Peters (2006) applies the discrete-choice demand and oligopolistic pricing game and suggests that merger simulation can better perform when it considers the changes in product characteristics. But he updates the characteristics by using actual post-merger data rather than endogenizes them in the model. Richard (2003) endogenizes flight frequency decision as well as quantity decision. However, the model is restricted to a single-firm optimization so that merged firm's decision is not affected by competitors, and the choice variables are decided simultaneously (a one-stage monopoly game). My research belongs to the latter group, but endogenizes both the set of product characteristics and prices in a sequential fashion in oligopolistic markets (a two-stage oligopoly game).

Second, this study contributes to the on-going literature on endogenous product choice (or quality). Starting from Mazzeo (2002), the issue has been continuously addressed by Crawford and Shum (2006), Gandhi et al. (2008), Draganska, Mazzeo and Seim (2009), Chu (2010), Byrne (2012), and Fan (2013).[7] Endogenizing product charac-teristics involves serious computational burden, especially when supply model adopts a two-stage oligopoly game with continuous character-istics. This is because one needs to compute derivatives of prices with respect to product characteristics for all products in a market. The literature avoids the complicated matrix by assuming reduced-form profit function without demand-driven market share (Mazzeo, 2002) or by adopting a one-stage game (Gandhi et al., 2008; Chu, 2010) or by analyzing monopoly market (Crawford and Shum, 2006; Byrne, 2012). The closest paper to mine is Fan (2013) in terms of specifying a two-stage oligopoly game. She derives the matrix by taking the total derivative of the second-stage

7 Crawford (2012) well summarizes this on-going literature. Also, Cho (2012) provides a great review by categorizing the literature according by types of product differentiation and consumer heterogeneity.

optimality condition as an application of the implicit function theorem. I empirically solve it in a more explicit way in which the optimal price function is derived from the second-stage optimality condition, and then I differentiate it with respect to product characteristics to solve the first-stage optimality condition. Since my approach directly computes the derivatives, it can be applied to more complicated optimization problems where multiple choice variables are correlated and decided sequentially (e.g. a three-stage oligopoly game).

Finally, looking at the overlap between above two subjects, this paper adds an empirical evidence on how a merger influences 'product positioning' or 'product variety'. This issue still remains controversial. A series of papers including Berry and Waldfogel (2001), Gandhi et al. (2008), and Sweeting (2010) shows that merged firm tends to increase product differentiation to avoid market cannibalization. On the other hand, Gotz and Gugler (2006) finds that higher concentration in retail gasoline market reduces product variety. This matter is critical because the consumer welfare is largely depending on how products are repositioned post-merger (Mazzeo, Seim and Varela, 2013). To provide a new evidence from the airline industry, I introduce two quality-distance measures: *within-firm distance* and *within-market distance* and analyze post-merger changes in the extent of product differentiation.

The remainder of the paper is organized as follows: Section 2 provides the structural model of air travel market and derives necessary optimality conditions. Section 3 describes the dataset. Section 4 presents an estimation procedure and reports model parameters. Section 5 simulates post-merger product characteristics and price, and analyzes welfare changes. The comparison analysis between the simulated and actual post-merger outcome is also addressed here. Section 6 concludes with a brief summary.

2. The Model

This section presents demand and supply model in the air travel market. In each market, carriers provide the set of differentiated products, and each consumer either purchases one product or takes the outside

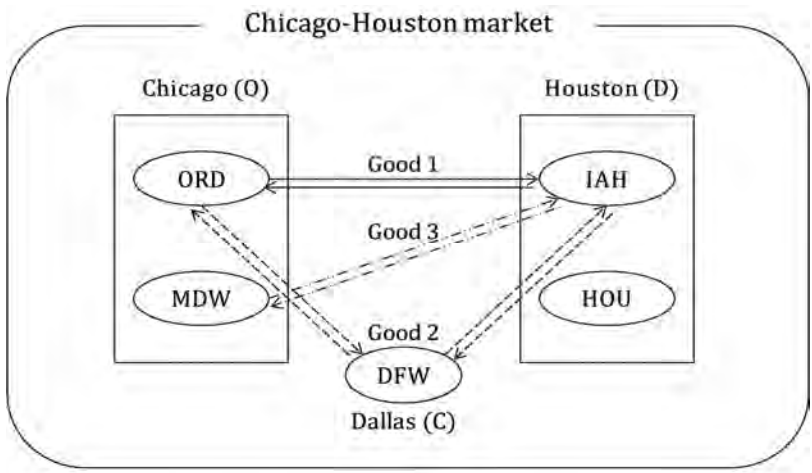
option of not flying. Importantly, the endogenous product characteristics are assumed to affect both consumers' utility and firms' cost.

2.1. Demand

The demand model follows discrete choice framework with heterogeneous consumer preferences (Berry, Levinsohn and Pakes, 1995, henceforth, BLP). In particular, I allow the consumer hetero-geneity to be represented by discrete distribution with only two types of consumers (Berry and Jia, 2010). As Borenstein and Rose (1994), Gerardi and Shapiro (2009), and several airline studies suggest, we can regard them as the business passengers and the tourists (r = 1 or 2).

A 'market' is a directional round trip between origin and destination city (see figure 1).[8] A 'product' is a unique combination of carrier-

Figure 1 An Illustration of Market and Product



Notes: O, D, C indicate origin, destination, and connecting city, respectively. Given carrier, roundtrip ORD-IAH, ORD-DFW-IAH, and MDW-IAH are considered as different products.

8 A city is a Metropolitan Statistical Area. In general, one city has one airport, but a few big cities have multiple airports. For example, Chicago has ORD (O'Hare) and MDW (Midway) airport as described in figure 1.

itinerary.[9] In other words, given an itinerary, all tickets sold by a carrier are aggregated to form a representative product.[10] This market and product definitions allow us to distinguish direct and connecting flights and to use information on characteristics for each airport.

Each consumer derives utility from price, observed product characteristics, and unobserved components. The conditional indirect utility of a passenger i who is of type r from choosing product j in market t is assumed to be

$$\begin{aligned} u_{ijt} &= p_{jt}\alpha_r + y_{jt}\psi_r + z_{jt}\varphi + \xi_{jt} + v_{it}(\lambda) + \lambda\varepsilon_{ijt} \\ &= x_{jt}\beta_r + \xi_{jt} + v_{it}(\lambda) + \lambda\varepsilon_{ijt} \end{aligned} \quad (1)$$

where p_{jt} is a passenger-weighted average ticket price of product j. y_{jt} is a two-dimensional vector including *Ontime*15 and *Layovers*. Ontime15 represents on-time performance of a product. A flight is counted as 'on-time' if it arrives at a gate less than 15 minutes after the scheduled arrival time. Since the original data contain a flight's scheduled arrival and actual arrival time on each non-stop segment, I measure *Ontime*15 as the geometric mean of percentage of flights that arrive on-time on each segment. *Layovers* is the number of connections per round-trip: 0 for direct flights and 2 for connecting flights.[11] I allow p_{jt} and y_{jt} to have random coefficients α_r and ψ_r, respectively, to see whether the business passengers and the tourists exhibit heterogeneous tastes for price, on-time arrival, and direct flight.

The vector z_{jt} includes several other characteristics for which both

9 An itinerary is an ordered sequence of airports for a round-trip.

10 The product definition is based on two considerations. First, the data on mishandled baggage rate and denied boarding rate are available only at carrier-level. Second, if each ticket is considered as a product, the estimation time tends to seriously increase, mainly because product shares need to be inverted at each iteration to derive unobserved product quality.

11 In the sample, passengers on direct flights have two coupons with no connection, and passengers on connecting flights have four coupons with two connections. Technically, 'direct' means that passengers do not change a plane between origin and destination, whereas 'non-stop' means that the flight does not stop between origin and destination. In this paper, I use both terms to refer to flights that do not stop between origin and destination.

types of passengers are assumed to have same level of marginal utility (φ). It contains *Frequency*, the number of average daily departures, to capture the benefits from convenient flight schedule with multiple departure times. *Frequency* is computed as the geometric mean of a flight frequency on each segment for a similar reason to *Ontime*15. *Mishandled baggage* is the number of mishandled baggages per 1,000 passengers. If a passenger's baggage is lost, damaged, or delayed, it is considered mishandled. z_{jt} also includes *Denied boarding* measured by the number of involuntary denied boardings per 10,000 passengers. Even though consumers hold confirmed reservations, they may be denied boarding from a flight due to airline overbooking.

In this research, the four characteristics - *Ontime*15, *Frequency*, *Mishandled baggage*, *and Denied boarding* - are modeled as endogenous variables (along with endogenous price). In previous studies, the characteristics are assumed to be exogenous based on the notion that firms cannot adjust them at least in the short run. However, this paper aims to simulate the merger effect. Since airline integration process takes a long time to be completed and the consolidation influences the overall operational characteristics of airline products, I reasonably set the characteristics to be a firm's choice variables.[12]

As additional controls, z_{jt} includes *HubDM*, the number of a carrier's hub airports on itinerary. This variable controls consumer valuation for frequent flier program and convenient gate access generated by a carrier's hub operation. I also expect that passengers' utility depend on *Constant, Distance* (the total round-trip distance), *Slot - control* (the number of slot-controlled airports on itinerary), and *Tour* (1 if a destination airport is located in either California, Florida, or Nevada).[13] Finally, I include several carrier dummies to control the brand-specific effect. Major airlines are Continental (*CO*), Delta (*DL*), Northwest (*NW*), United (*UA*), US Airways (*US*), and American (*AA*, the base carrier). Two low cost carriers are Southwest (*WN*) and AirTran (*FL*). The remaining carriers are defined as Other Carriers (*OT*).

---

12 The integration process for the Delta and Northwest Airlines merger continued for twenty one months after the initial announcement in April 2008.

13 In this study, slot-controlled airports include Chicago O'Hare (ORD), John F. Kennedy (JFK), LaGuardia (LGA), Reagan National (DCA) airport.

ξ_{jt} is an unobserved (to the econometrician) product quality which is not captured by the dataset. It represents ticket-or flight-level characteristics such as Saturday night stay-over, advance purchase, non-refundability, minimum or maximum stay restriction, and in-flight meal service quality.[14] v_{it} is the nested logit disturbance. It is constant across all airline products (inside goods) in market *t*, but differentiates air travels from the outside option of not flying. λ is the nested logit parameter which represents the degree of product differentiation between inside goods. It varies between 0 and 1. If λ = 1 (then v_{it} = 0), each airline product is perfectly differentiated. In this case, there will be no need to set outside option, and demand specification becomes a multinomial logit model. If λ = 0, all airline services are perfectly substitutable. ε_{ijt} is an *i.i.d.* (across consumers, products, and markets) logit error. The error structure v_{it} (λ)+ λε_{ijt} follows the Type I extreme value distribution to derive closed-form market share equation.

The indirect utility from the outside good (e.g. driving a car or taking a train) is given by

$$u_{i0t} = \xi_{0t} + \varepsilon_{i0t}. \quad (2)$$

A simple way to identify mean utility of the outside good is setting it as one of the inside goods. However, it is not desirable strategy because airline products are quite different services from those by other ground transportation modes. Alternatively, I normalize both ξ_{0t} and ε_{i0t} to be zero for all consumers. In this case, the coefficient of *Constant* will measure marginal utility from choosing any airline products.

Note that in the standard BLP model, consumer tastes vary with demographics and unobserved individual characteristics, following multivariate normal (or other continuous) distributions. Differently, the random coefficients here vary with finite passenger types, following discrete *r*-type distribution. Thus, I derive a market share function by computing the weighted sum of the market share for each type, rather

14 This restricted information is available only in transaction-level data from Computer Reservation System. An analysis which uses the specific information can be found in Puller, Sengupta and Wiggins (2012).

than by integrating purchase probability over continuous distributions. The weight is the percentage of type *r* consumers in the population (γ_r).

Assuming each type *r* consumer purchases one airline ticket which gives the highest mean utility $(x_{jt}\beta_r + \xi_{jt})$, the market share of *j*th product is given by

$$s_{jt}(x_t, \xi_t, \theta_d) = \sum_{r=1}^{2} \gamma_r \cdot \frac{e^{(x_{jt}\beta_r + \xi_{jt})/\lambda}}{\sum_{k \in J_t} e^{(x_{kt}\beta_r + \xi_{kt})/\lambda}} \cdot \frac{(\sum_{k \in J_t} e^{(x_{kt}\beta_r + \xi_{kt})/\lambda})^{\lambda}}{1 + (\sum_{k \in J_t} e^{(x_{kt}\beta_r + \xi_{kt})/\lambda})^{\lambda}}, \quad (3)$$

where x_t=$(x_{1t}, \ldots, x_{Jt})$, $\xi_t = (\xi_{1t}, \ldots, \xi_{Jt})$,and J_t is the set of all airline products in market *t*. θ_d is the set of all demand parameters (α_r, ψ_r, φ, λ, γ_r). Each market provides two groups of products: all the airline services and outside option, thus the first term indicates within-group share of airline product *j*, and the second term denotes to group share of all the airline products.

2.2. Supply

In this section, I describe a two-stage oligopoly game where each carrier chooses optimal product characteristics first and then decides optimal prices to maximize the expected profit under Bertrand-Nash competition. Airline network structures such as markets, routes, airports served, and location of hub airports are assumed to be exogenous.[15]

At the first stage, firm *f* decides the set of product characteristics, $\overline{x}_j = (\overline{x}_j^0, \overline{x}_j^F, \overline{x}_j^M, \overline{x}_j^D)$ to maximize the profit function[16]

---

15 This assumption is justified by the fact that most airlines sign 'long-term use-and-lease agreements' with airports to occupy the airport facilities. The detailed information on the contractual practices between airports and airlines can be found in Ciliberto and Williams (2010) and Lee (2013). Also, considering that an airline product is a carrier-route combination, the assumption is analogous to a typical setting where the number of products offered is exogenously given.

16 The superscript *O, F, M*, and *D* denote the first letter of the endogenous product characteristics, respectively.

$$\Pi_f^I = \sum_{j \in J_f} (p_j(\overline{x}) - mc_j(\overline{x}_j^F)) \cdot M \cdot s_j(p(\overline{x}), \overline{x}, \xi; \theta_d) - F(\overline{x}_f, \zeta_f; \tau) \quad (4)$$

where $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_J), \overline{x}_f = (\overline{x}_{1f}, \ldots, \overline{x}_{Jf})$, and $\zeta_f = (\zeta_{1f}, \ldots, \zeta_{Jf})$. J is the set of all products in a market, and J_f is the set of all products offered by firm f in a market. Throughout the supply model, a market subscript t is omitted for simplicity.[17] mc_j is the marginal cost of product j, and M is a market size which is the geometric mean of the MSA population of two end-point cities. $s_j(\cdot)$ is the demand-driven market share function of product j coming from equation (3), and $F(\cdot)$ is the fixed cost function.

In equation (4), a carrier's decision on the characteristics ($\overline{x}$) affects prices, marginal, and fixed cost. It also affects market share directly and indirectly.[18] To be specific, in each market, prices of all products are influenced by the characteristics of all products through $p_j(\overline{x})$ and $p(\overline{x})$. These interactions (arising from a two-stage oligopoly game) make the necessary equilibrium conditions difficult to be computed. The way of solving it is described in section 2.3.

Marginal cost of serving an additional passenger is given by the following linear function

$$mc_j = h_j\delta + \omega_j \quad (5)$$

where h_j denotes the set of cost characteristics. h_j includes *Frequency* ($\overline{x}_j^F$) to capture marginal cost effect of the aircraft utilization. Among the four endogenous characteristics, only *Frequency* is modeled to affect marginal cost because it is a quantity-related variable. h_j also controls *HubMC*, 1 if a flight departs from, connects at, or arrives at its hub airport. A carrier's hub operation can cause two countervailing effects on marginal cost. In hub-and-spoke system, a majority of passengers come from different origins and connect at a carrier's hub airport to reach their

17 Following Berry and Jia (2010), markets are assumed to be independent. Thus, all equations in this section are applied to each market without loss of generality.

18 Even though price, marginal cost, and market share function also depend on other control variables, equation (4) is expressed with a focus on the endogenous characteristics.

final destinations. This allows the carrier to generate high load factor on major routes, which contributes to decreasing the per-passenger cost. On the other hand, a carrier's hub operation causes massive air- and ground-side congestion at an airport. This can increase marginal cost. The coefficient reflects the net effect of the two factors. I control two distance measures, $Distance_{short}$ and $Distance_{long}$, consi-dering that fuel efficiency can differ depending on aircraft size, and different sizes of fleets are allocated on short-haul and long-haul routes.[19] Similarly, I control $Layovers_{short}$ and $Layovers_{long}$. Conne-cting flights involve an additional landing/takeoff during which airplanes burn a large fraction of fuel, and the amount of fuel consumed is known to vary with aircraft size.[20] Finally, I set carrier dummies to control carrier-specific cost effect.

δ indicates a vector of cost parameters, and ω_j represents unobservable (to the econometrician) marginal cost shocks. It includes fluctuations in oil prices, quality of on-board meals, charges levied for landing, and other unobserved factors.

Following Fan (2013), I adopt a quadratic function to approximate the fixed cost function. Specifically, the slope of the fixed cost with respect to an endogenous characteristic ($\overline{x}_j^k, k = O, F, M, D$) is given by

$$\frac{\partial F(\overline{x}_f, \zeta_f; \tau)}{\partial \overline{x}_j^k} = \tau_0^k + \tau_1^k \overline{x}_j^k + \zeta_j^k, \tag{6}$$

where τ is a vector of parameters, and ζ_j^k represents unobservable fixed-cost shock. Adjustments of the operational characteristics accompany consolidation of facilities (aircraft, gate, and ticket counter) and workforce (pilots, flight crew, gate/ticket takers, baggage handlers, and ticket booking agent). Using more or less of these resources influences

---

19 I create an indicator variable $I_{long} = 1$ if a market distance is longer than 3,500 miles and $I_{short} = 1$ if a market distance is shorter than 3,500 miles. Then the distance measures are computed as $Distance_{long} = Distance * I_{long}$ and $Distance_{short} = Distance * I_{short}$.

20 Similar to the distance measures, I compute the two $Layovers$ as $Layovers_{long} = Layovers * I_{long}$ and $Layovers_{short} = Layovers_{short} * I_{short}$.

the fixed cost. Other cost shocks such as advertising costs are captured by ζ_j^k.

Given a vector $\overline{x}_j$ chosen at the first stage, firm f decides price p_j at the second stage to maximize the following profit function,

$$\Pi_f^{II} \sum_{j \in J_f} \Pi_j^{II} = \sum_{j \in J_f} (p_j - mc_j) \cdot M \cdot s_j(p, \overline{x}, \xi; \theta_d) \tag{7}$$

While the first stage profit function is specified as the difference between the variable profit and the fixed cost, carriers now maximize the variable profit under the Bertrand-Nash competition.

In airline industry, prices are easily changeable, but the product characteristics are not. For example, when a carrier increases flight frequency, it needs to adjust aircraft size and to hire more employees who manage flight schedule. Further, it may reallocate gates based on contract with airport authority. However, price decisions can be made relatively quickly and flexibly at the final stage. Hence, this sequential choice model better reflects airlines' decision-making process.

2.3. Necessary Equilibrium Conditions

I solve carriers' optimization problems by deriving necessary equilibrium conditions for the product characteristics and prices. From the conditions, I will recover the structural errors in marginal cost function (ω_j) and fixed cost function ($\zeta_j^O, \zeta_j^F, \zeta_j^M, \zeta_j^D$) in section 4.

Starting with the second-stage game based on backward induction, I take the derivative of the second-stage profit function $\prod_f^{II}$ with respect to prices ($p_j, j = 1, \cdots, J_f$) to generates the first-order condition $\partial \prod_f^{II} / \partial p_j$),

$$s_j(p, \overline{x}, \xi; \theta_d) + \sum_{h \in J_f} (p_h - mc_h) \cdot \frac{\partial s_h(p, \overline{x}, \xi; \theta_d)}{\partial p_j} = 0. \tag{8}$$

Stacking all J_f products together yields

$$s_f(p, \overline{x}, \xi; \theta_d) + \Omega_{s_f, P_f} \cdot (p_f - mc_f) = 0, \tag{9}$$

where $s_f = [s_1, \ldots, s_{J_f}]'$, $p_f = [p_1, \ldots, p_{J_f}]'$, $mc_f = [mc_1, \ldots, mc_{J_f}]'$ and Ω_{s_f, P_f}, is a $J_f \times J_f$ matrix given by

$$\Omega_{s_f, P_f} = \begin{bmatrix} \frac{\partial_{s_1}}{\partial_{p_1}} & \cdots & \frac{\partial_{s_{J_f}}}{\partial_{p_1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial_{s_1}}{\partial_{p_{J_f}}} & \cdots & \frac{\partial_{s_{J_f}}}{\partial_{p_{J_f}}} \end{bmatrix}. \tag{10}$$

Rearranging terms in equation (9) derives a carrier's optimal price function,

$$p_f = h_j \delta + \omega_j - \Omega_{s_f, P_f}^{-1} \cdot s_f(p, \overline{x}, \xi; \theta_d). \tag{11}$$

The right hand side be composed of two parts. The first two terms indicate the marginal cost and the remaining term (including negative sign) constitutes markup. Through the two components, the optimal price is affected by product characteristics. This dependency provides a link between the first stage and the second stage game.

Moving on to the first-stage game, I differentiate the profit function $\prod_f^I$ with respect to the product characteristics ($\overline{x}_j^k, j = 1, \ldots, J_f, k = O, F, M, D$) to yield the first-order condition $\partial \prod_f^I / \partial \overline{x}_j^k$,

$$\sum_{h \in J_f} \frac{\partial \prod_h^{II}}{\partial \overline{x}_j^k} + \sum_{h \in J_f} \sum_{h' \in J} \frac{\partial \prod_h^{II}}{\partial p_{h'}} \frac{\partial p_{h'}}{\partial \overline{x}_j^k} - \tau_0^k - \tau_1^k \overline{x}_j^k - \zeta_j^k = 0. \tag{12}$$

While the adjustment of $\overline{x}_j^k$ has a direct effect on variable profit of product h ($\prod_h^{II}$), it also has an indirect impact on $\prod_h^{II}$ h by affecting

prices of all products in a market.[21] Main computational difficulty arises from $\frac{\partial p_{h'}}{\partial \overline{x}_j^k}$ in the second term. This requires the derivative of all equilibrium prices with respect to all products' characteristics.[22] As a great way of computing it, Fan (2013) applies the implicit function theorem by taking the total derivative of the second-stage optimality condition (9) with respect to prices and product characteristics.[23] Since this approach relies on the observed product characteristics, one needs to rule out corner solutions where the equation (9) does not hold. I empirically solve it in a more explicit way. I plug the optimal price function (11) into the first-stage profit function (4) and differentiate the profit function with respect to each product characteristic. While both methods need an assumption that the optimal price function is smooth and differentiable with respect to the characteristics, they produce the same computational result for a two-stage oligopoly game.[24] However, since my approach directly computes the derivatives, it can be applied to more complicated optimization problems where multiple choice variables are correlated and decided sequentially (e.g. a three-stage oligopoly game).

3. Data

3.1. Sources

I collected the data from a variety of sources (see table 1). The primary data set is the Airline Origin and Destination Survey (DB1B)

---

21 For *Frequency* ($\overline{x}_j^F$), the exact expression for the first-order condition is equation (13) below, because the frequency affects marginal cost function. However, the optimal price function includes marginal cost in it, thus equation (12) and (13) are essentially same for *Frequency*.

$$\sum_{h\in J_f} \frac{\partial \prod_f^{II}}{\partial \overline{x}_j^F} + \sum_{h\in J_f}\sum_{h'\in J} \frac{\partial \prod_h^{II}}{\partial p_{h'}} \frac{\partial p_{h'}}{\partial \overline{x}_j^F} - \sum_{h\in J_f} \frac{\partial \prod_h^{II}}{\partial mc_h} \frac{\partial mc_h}{\partial \overline{x}_j^F} - \tau_0^F - \tau_1^F \overline{x}_j^F - \zeta_j^F = 0 \quad (13)$$

22 Technically, the derivative requires us to compute $\frac{\partial(\Omega_{s_f^-,P_f}^{-1})}{\partial \overline{x}_f^k}$ and $\frac{\partial(\Omega_{s_j^-,P_j}^{-1})}{\partial \overline{x}_j^k}$.

23 This approach was initially introduced by Villas-Boas (2007).

24 The four endogenous product characteristics in this research are reasonably continuous.

produced by the U.S. Department of Transportation (DOT).[25] Based on the DB1B, I defined the market and product, and created the variables varying by product (*Fare* and *Layovers*), carrier (brand dummies), airport (*Slot-control*), carrier/airport (*HubDM, HubMC*), and market (*Distance* and *Tour*).

Table 1 Data Sources

Database	Variables	Level of observation	Sample periods
O&D Survey (DB1B)	Market/Product, Fare, Controls	ticket	'07. 2Q.~'08. 1Q.
On-Time Performance	Ontime15	carrier-route	'07. April~'08. March
T-100 Domestic Segment	Frequency	carrier-route	'07. April~'08. March
Air Travel Consumer Report	Mishandled baggage, Denied boarding	carrier	'07. April~'08. March
Air Carrier Financial Report	Employee statistics	carrier	'07
Weather Underground	Wind, Rain, Snow	airport	'07. April~'08. March
MSA Population	Market size	MSA city	'07 estimates

Notes: Controls include Layovers, Distance, HubDM, hubMC, Slot-control, Tour, Carrier dummies.

The endogenous product characteristics come from three different sources, also produced by the DOT. I calculated Ontime15 based on the Airline On-Time Performance Data. The data contain monthly information on scheduled and actual departure/arrival times for a flight, covering all U.S. carriers that account for at least one percent of domestic scheduled passenger revenues.[26] *Frequency* was constructed by using T-100 Domestic Segment Data. Among several departure-related terms, I used 'departures performed' which counts takeoffs by each carrier at an airport. Finally, I used Air Travel Consumer Report to create *Mishandled baggage* and *Denied boarding*. While the report is filed on a monthly basis, the statistics on *Mishandled baggage* and *Denied boarding* are updated by monthly and quarterly, respectively. Hence, I computed *Mishandled baggage* as the average value of the mishandled baggage rate

25 The DB1B database is a 10% sample of airline tickets from reporting carriers, produced on a quarterly basis. There are three subcomponents to the DB1B: market, coupon, and ticket dataset. This study combines the last two dataset.

26 With this data, one can create other discrete variables (e.g. Ontime30, Ontime60) or continuous variables (e.g. *Average minutes late*). Besides departure and arrival times, the data also provide information on the causes of delay and cancelations.

of each month during a quarter.

Further, I used airline employment data and weather data to construct instrumental variables for the endogenous characteristics. The employment data come from Air Carrier Financial Reports (Schedule P-10). It contains annual employee statistics by labor category such as pilots/ copilots, maintenance employees, and passenger handling employees. The weather data was collected from Weather Underground. This is a commercial weather service which gathers its most information from the National Weather Service (NWS). Typically, the weather reporting location for a particular city is its airport, which is appropriate for this research. The instruments will be explained in more detail in Section 4.

3.2. Sample Selection and Description

The Delta and Northwest Airlines merger was announced the second quarter of 2008. I define pre-merger period as the four quarters predating the announcement. Hence, the sample period for estimating pre-merger demand and supply is from the second quarter of 2007 to the first quarter of 2008.

The criteria for sample selection is as follows. In ticket level, I focus on round-trip itineraries within U.S. continent with at most four coupons. Also, I drop tickets whose prices are lower than $50 or higher than $1,800. The lower bound is to eliminate tickets purchased using frequent flyer miles, and the higher bound is to restrict the sample to coach-class travel. In product level, I drop observations with fewer than five passengers because they are likely to be non-regular services.[27] I exclude products associated with open-jaw.[28] An open-jaw trip does not fit for applying the typical definitions of origin and destination city. Further, they are known to be subject to different pricing scheme relative to the ordinary round-trip tickets. In market level, I focus on medium to large metropolitan areas whose populations are more than 850,000. This is for reducing heterogeneity of demand and supply. As Berry and Jia (2010)

27 Since the DB1B is a 10% random sample, those airline products are likely to carry less than fifty passengers during a quarter.

28 An open-jaw trip is essentially a round trip in which the outward point of departure and the inward point of arrival are not the same.

Table 2 Variable Definitions and Summary Statistics for the Estimation Sample

Variable	Description	Mean	Std.dev.	Min	Max
Endogenous variables					
Fare	Average ticket fare ($100)	3.74	1.13	0.55	13.17
Ontime15	Percentage of flights that arrive less than 15 minutes late	0.75	0.08	0	1
Frequency	No. of average daily departures per quarter	4.32	2.42	0.01	26.18
Mishandled baggage	No. of mishandled baggages per 1,000 passengers	6.44	1.73	2.61	13.52
Denied boarding	No. of involuntary denied boardings per 10,000 passengers	1.19	0.72	0.01	4.48
Control variables					
Layovers	No. of connections per round trip	1.67	0.75	0	2
Distance	Market distance round trip (1,000 miles)	3.18	1.43	0.22	6.94
HubDM	No. of hub airports given carrier and itinerary	0.72	0.60	0	3
HubMC	1 if a flight departs from, connects at, or arrives at hub airport	0.64	0.48	0	1
Slot-control	No. of slot-controlled airports on itinerary	0.28	0.59	0	3
Tour	1 if destination airport is in either CA, FL, or NV	0.32	0.47	0	1
Carrier dummies					
AA	1 if a carrier is American Airlines	0.13	0.34	0	1
CO	1 if a carrier is Continental Airlines	0.07	0.25	0	1
DL	1 if a carrier is Delta Airlines	0.14	0.35	0	1
NW	1 if a carrier is Northwest Airlines	0.10	0.30	0	1
UA	1 if a carrier is United Airlines	0.10	0.30	0	1
US	1 if a carrier is US Airways	0.11	0.32	0	1
FL	1 if a carrier is AirTran Airways	0.05	0.22	0	1
WN	1 if a carrier is Southwest Airlines	0.17	0.38	0	1
OT	1 if a carrier is other carrier	0.13	0.33	0	1

Notes: The sample contains 87,906 unique products in 9,117 markets. Sample period is from '07. 2Q. through '08. 1Q.

states, the demand pattern and the operation cost among small-sized markets tend to be different from those among medium to large-sized markets.

The final sample contains 87,906 unique products in 9,117 markets.[29] Table 2 provides summary statistics for the estimation sample. Focusing on the endogenous characteristics, the mean value of *Ontime*15 indicates 75% of flights arrived on-time during the sample period. As extreme cases, 24 products have 100% on-time performance record. All of them are direct flights, and more than half of them are produced by the Southwest Airlines. As the worst cases, 160 products have 0% on-time performance. When using a rougher measure *Ontime*30, the on-time performance increases to 86%. Also, a continuous measure *Average minutes late* shows that flights arrived 12.6 minutes late on average. The statistics for *Frequency* indicate that flights departed 4.3 times a day on average. It varies significantly across markets and products. To be specific, frequency is higher in tourism markets (4.52) than in others (4.23), and higher in short-haul markets (4.41) than in long-haul markets (4.20). Further, flights originating from a carrier's hub airports show high frequency (4.91) than others (4.27). Lastly, the number of mishandled baggages are 6.4 per 1,000 passengers, and the number of denied boardings are 1.2 per 10,000 passengers. They also exhibit large variations across carriers and across quarters for each carrier.

4. Estimation

To estimate the model parameters, I recover the structural errors in the demand and supply specification as a function of model parameters and data. The errors include unobserved quality (ξ_t), marginal cost shock (ω_t), and fixed cost shocks $\zeta_t^O, \zeta_t^F, \zeta_t^M, \zeta_t^D$). ξ_t is derived by inverting the market share function: $\xi_t = s^{-1}(x_t, \dot{s}_t, \theta_d)$. Given demand parameters $\theta_d = [\alpha_r, \psi_r, \varphi, \lambda, \gamma_r]$ and data x_t, I solve for ξ_{jt} that equates the predicted market share to observed market share by using a contraction mapping (Berry, Levinsohn and Pakes, 1995; Berry and Jia, 2010),

29 Travels on same itinerary but in different quarters are considered as different products in different markets.

$$\zeta_{jt}^{H} = \zeta_{jt}^{H-1} + \lambda\left[\ln \dot{s}_{jt} - \ln s_{jt}(x_t, \xi_t, \theta_d)\right], \tag{14}$$

where H denotes the H^{th} iteration, $\dot{s}_{jt}$ is the observed market share, and $s_{jt}(x_t, \xi_t, \theta_d)$ is the predicted market share defined by equation (3). This convergence process is carried out market by market because market share of product *j* depends on the characteristics of all products in market *t*.[30]

The marginal cost shock is recovered by necessary optimality conditions at the second stage. From the optimal price function (11), I derive ω_{jt} as a function of marginal cost characteristics h_{jt} and parameters δ,

$$\omega_{jt} = p_{jt} - h_{jt}\delta + \Omega_{s_{ft}, P_{ft}}^{-1} \cdot s_{jt}(p_t, \overline{x}_t, \xi_t; \theta_d). \tag{15}$$

Finally, the fixed cost shock for each endogenous characteristic is obtained by the optimality condition at the first stage. The first-order condition (12) yields $\zeta_{jt}^{k}(k = O, F, M, D)$ as,

$$\zeta_{jt}^{k} = \left(\sum_{h \in J_{f_t}} \frac{\partial \prod_h^{II}}{\partial \overline{x}_{jt}^{k}} + \sum_{h \in J_{f_t}} \sum_{h' \in J_t} \frac{\partial \prod_h^{II}}{\partial p_{h'}} \frac{\partial p_{h'}}{\partial \overline{x}_{jt}^{k}} \right) - \tau_0^k - \tau_1^k \overline{x}_{jt}^{k}. \tag{16}$$

The marginal and fixed cost shocks are computed carrier by carrier within a market, considering that each firm maximizes profit from its own products. Notice that demand parameters θ_d enters the specifications of all structural errors. While θ_d enters the unobservable quality ξ_{jt} on the demand side, it becomes a factor of marginal cost shock ω_{jt} through the market share function, and of $\zeta_{jt}^{k}(k = O, F, M, D)$ through the profit function. Moreover, marginal cost parameters δ included in ω_{jt} enters ξ_{kjt} through the profit function. This interrelation motivates us to jointly estimate the demand and supply parameters for enhancing efficiency.

I estimate the parameters by using the two-stage nonlinear Gener-

---

30 I iterate the contraction mapping until the maximum difference between each iteration is smaller than 10^{-12} : $||\xi^M - \xi^{M-1}||_\infty = \max\{|\xi_1^M - \xi_1^{M-1}|, \cdots, |\xi_{J_t}^M - \xi_{J_t}^{M-1}|\} < 10^{-12}$.

alized Method of Moments. For product j in market t, let $W_{jt} = [W_{jt}^d \; W_{jt}^c \; W_{jt}^k]$ be a set of instruments for endogenous variables in demand, marginal cost, and fixed cost specification, respectively. As an identification assumption, I set the moment conditions by taking expectations of each structural error interacted with the exogenous instruments

$$\begin{aligned} &\forall j, t: \\ &E[W_{jt}^{d'} \xi_{jt}(\theta_d) = 0, \\ &E[W_{jt}^{c'} \omega_{jt}(\theta_d, \delta) = 0, \\ &E\left[W_{jt}^{k'} \zeta_{jt}^k(\theta_d, \delta, \tau^k)\right] = 0, \quad k = O, F, M, D. \end{aligned} \tag{17}$$

Let $g(\Theta)$ be the stacked vector of sample analogues to the moments (17), where $\Theta = [\theta_d, \delta, \tau^k]$. I minimize the first-stage objective function $Q = g(\Theta)' V g(\Theta)$ with a weighting matrix $V = (W'W)^{-1}$, assuming all error terms are homoscedastic. After obtaining parameter estimates $\widehat{\Theta}^1$, I compute the structural errors $\hat{\eta} = [\hat{\xi} \; \hat{\omega} \; \hat{\zeta}^k]$ to obtain the optimal weighting matrix $V = (W'\hat{\eta}\hat{\eta}'W)^{-1}$ for second stage. The objective function is minimized once again to produce the final parameter estimates $\widehat{\Theta}^2$.

4.1. Instruments

Carriers observe the product quality ξ_{jt} and the cost shocks (ω_{jt}, $\zeta_{jt}^k, k = O, F, M, D$) before they decide optimal product characteristics and prices. Therefore, the carriers' decisions are correlated with the structural errors. As an example of price, airline tickets restricted to Saturday night stay-over, advance purchase, or non-refundability requirement tend to be cheaper than unrestricted tickets (Puller, Sengupta and Wiggins, 2012). Further, when carrier face significant marginal cost shocks (e.g. fuel cost, landing fee) and fixed cost shocks (e.g. insurance, FAA registration fee, advertising cost), they may reorganize flight operations and production facilities which can affect the product characteristics. In this sense, the price and the characteristics are endogenous.

Table 3 Instrumental Variables for Endogenous Product Characteristics

Instruments	Ontime15	Frequency	Mishandled Baggage	Denied Boarding
Weather (wind, rain, snow)	Yes	Yes	Yes	–
Carrier's Hub Status				
Hub Origin	Yes	Yes	–	Yes
Hub Connection	Yes	Yes	Yes	–
Hub Destination	–	Yes	–	–
Labor Category (%)				
General Managers	Yes	Yes	Yes	Yes
Pilots & Copilots	–	Yes	–	–
Passenger Svc. & Admin.	–	–	Yes	Yes
Maintenance	Yes	–	–	–
Aircraft Traffic Handling	–	Yes	–	–
Aircraft Control	Yes	Yes	–	–
Passenger Handling	–	–	–	Yes
Cargo Handling	–	–	Yes	–
Statistical	–	–	–	Yes
Traffic Soliciters	–	Yes	–	Yes
Validity of instruments				
F-statistics	724.5	806.6	1,600.5	2,903.6
R-squared	0.083	0.114	0.154	0.284

Notes: The weather data come from Weather Underground which gathers its most information from the National Weather Service (NWS). I used information on wind, rain, and snow condition at origin and destination airport. The employment data come from Air Carrier Financial Reports (Schedule P-10).

The exogenous instruments for prices include the information on market and airport-carrier level. The number of routes within a market can represents the degree of the market competition, which is correlated with overall price level. Next, the number of cities directly connected from an origin airport by a carrier measures the carrier's network size from each airport. This airport-carrier specific variable is related to the attractiveness of frequent flier program and thus can capture a substantial portion of price premium. Finally, exogenous variables in the demand and supply specification are included.

The identification strategy for the product characteristics is to find exogenous factors influencing airline operations. I apply weather conditions at an airport, a carrier's hub status at an airport, and a carrier's employ-

ment statistics (see table 3). First, weather conditions such as wind, rain, and snowfall are beyond carriers' controls, but affect several product characteristics either directly or indirectly. *Ontime*15 is affected most. Adverse weather conditions are the direct cause of most flight delays because it requires extra preparations for takeoff and landing. *Mishandled baggage* also falls under the direct effect since a delayed baggage is counted as a mishandled one.[31] The bad weather indirectly influences *Frequency* through flight cancelations. Since I measured *Frequency* based on departures performed (not on departures scheduled), the cancelation due to bad weather is correlated with *Frequency*. However, it is not easy to find a close relationship between the weather conditions and *Denied boarding*. Notably, when the U.S. DOT measure *Denied boarding*, it does not consider passengers affected by canceled, delayed, or diverted flights.

Second, a carrier's hub status at an airport, which can be treated as exogenous, significantly affects the product characteristics.[32] As Rupp, Owens and Plumly (2006) states, flights originating from hub airports tend to have lower on-time performance, because some of aircraft services such as cleaning, refueling, or catering occur only at hub airports, requiring a longer preparation time for the next same-day departure. Differently, flights connecting to hub airports tend to have better on-time performance in order to reduce inconvenience to connec-ting passengers. About *Frequency*, flights to and out of hub airports tend to exhibit high frequency to accommodate the dense traffic flows (Brueckner and Zhang, 2001). Since my sample is supporting the pattern, I include all hub-related variables in instruments. Baggages are mostly mishandled when transferred through hub airports during congested peak periods

31 Wyld, Jones and Totten (2005) provides a good example. During Christmas holiday season in 2004 when severe weather created disruptions, US Airways misplaced thousands of baggages across the Midwest, accumulating them at airports along the East Coast.

32 The exogeneity assumption on a carrier's hub status at an airport is supported by DOT (1999). In most medium and large airports in the US, major airlines have entered into long term use-and-lease agreements, including residual, compensatory, and hybrid agreements to attain the status of hub (or signatory) carrier. The average length of the agreement was 28 years for a residual agreement, 17 years for a compensatory agreement, and 20 years for a hybrid agreement.

(Jayaraman and O'Connell, 2011). Hence, I consider only connection at a hub as an instrument for *Mishandled baggage*. Since *Denied boarding* is positively correlated with high load factor mostly observed from flights out of hubs, the origination from a hub is included.

The final group of instruments contains a carrier's employment statistics. Using Air Carrier Financial Report, I calculated the percentage of workers in each labor category over total number of employees in an airline company and identified how their works were related to each product characteristic. Suppose that significant malfunctions of aircraft systems are detected just before departure time, then many skilled maintenance workers would be necessary for the flight to be on-time. Similarly, carriers need a large number of pilots, copilots, and aircraft controllers to keep high *Frequency*. *Mishandled baggage* and *Denied boarding* can be affected by the number of cargo handling employees and the number of staff in statistical posts, respectively.

I conduct F-test by running reduced-form regressions. The test statistics (in bottom panel of table 3) indicate that the instruments are valid at 99% significance level.

4.2. Estimation Results

4.2.1. Demand Parameters

The first column in table 4 reports the estimated demand parameters. First, price parameters are identified by sensitiveness of product shares in response to changes in prices. The coefficients of $Fare_1$ and $Fare_2$ are -0.098 and -0.999, respectively. While both groups receive disutility from price increase, type 2 passengers exhibit about ten times as much price sensitivity as type 1 passengers. Based on industry knowledge, we can regard type 1 as the business travelers and type 2 as the tourists.[33]

33 In Berry and Jia (2010), estimates of price coefficients are -0.07 and -0.78 for the business passengers and the tourists, respectively (using 1999 data). Berry, Carnall and Spiller (2006) reported 0.068 and 0.696 for the business passengers and the tourists, respectively (using 1985 data). My estimates are close to them in terms of coefficient of each type and difference between the two coefficients.

Table 4 Estimation Results on Model Parameters

Mean Utility		Marginal Cost ($100)	
Endogenous Characteristics		Frequency	-0.021** (0.002)
$Fare_1$	-0.098** (0.004)	HubMC	-0.184** (0.017)
$Fare_2$	-0.999** (0.019)	$Layovers_{short}$	0.175** (0.010)
$Ontime15_1$	1.641** (0.164)	$Layovers_{long}$	0.300** (0.019)
$Ontime15_2$	2.114** (0.155)	$Distance_{short}$	0.226** (0.005)
Frequency	0.084** (0.003)	$Distance_{long}$	0.130** (0.004)
Mishandled baggage	-0.054** (0.005)	US	-0.036** (0.008)
Denied boardings	-0.253** (0.018)	DL	-0.168** (0.011)
		NW	-0.289** (0.012)
Controls		UA	-0.010 (0.008)
$Layovers_1$	-1.255** (0.013)	CO	-0.015 (0.010)
$Layovers_2$	-1.085** (0.009)	FL	-0.559** (0.019)
Distance	0.105** (0.005)	WN	-0.223** (0.018)
HubDM	0.056** (0.010)	OT	-0.190** (0.016)
Slot-control	-0.071** (0.006)	Constant	1.721** (0.029)
Tour	0.238** (0.006)		
US	0.143** (0.014)	**Slope of Fixed Cost ($100)**	
DL	0.007 (0.027)	$Ontime15_{constant}$	2416.8** (0.742)
NW	-0.511** (0.019)	Ontime15	-2970.2** (0.518)
UA	0.124** (0.017)	$Frequency_{constant}$	6.605** (0.064)
CO	-0.119** (0.024)	Frequency	0.224** (0.005)
FL	-0.875** (0.020)	$Mishandled\ baggage_{constant}$	-8.126** (0.059)
WN	-0.329** (0.021)	Mishandled baggage	0.491** (0.005)
OT	-0.023 (0.017)	$Denied\ boarding_{constant}$	-31.634** (0.148)
Constant	-7.184** (0.130)	Denied boarding	7.005** (0.020)
λ	0.618** (0.003)	Number of observations: 87,906	
γ_1	0.052** (0.004)		

Note: Standard errors are in parentheses. ** indicates 99% level of significance. Subscript 1 and 2 attached to *Fare*, *Ontime*15, and *Layovers* indicate consumer types.

Positive coefficients of $Ontime15_1$ and $Ontime15_2$ suggest that better on-time performance increases passengers utility who do not want flight delay during their travels. It should be noted that consumers do not know whether they would experience flight delays or not at the time of ticket purchase. However, as Suzuki (2000) and Mazzeo (2003) state, passengers can form expectations of flight delays based on the carrier's

past on-time performances on a specific route. In that sense, the parameters can be interpreted as marginal utility from the expected on-time arrival.[34] To calculate willingness-to-pay (WTP) for on-time performance, I divide the coefficients of $Ontime15_1$ and $Ontime15_2$ by those of $Fare_1$ and $Fare_2$, respectively. The result implies that business travelers show nearly eight times higher WTP than the tourists do: $\frac{\psi_{11}}{\alpha_1} / \frac{\psi_{12}}{\alpha_2} = 7.9$.

Next, an increase in *Frequency* has a positive effect on consumers' utility. The parameter estimate is 0.084. Consumers value a flight schedule with multiple departures because they are more able to depart at their preferred time. Increases in *Mishandled baggage* and *Denied boarding* will decrease the quality of airline products hurting passengers satisfaction. Reasonably, both characteristics have negative coefficients: -0.054 for *Mishandled baggage* and -0.253 for *Denied boarding*.[35]

All other demand parameters have the expected signs. The coefficients of $Layovers_1$ and $Layovers_2$ are -1.255 and -1.085, respectively, indicating that connecting flights generate disutility to both groups. Going through an additional stopover at the connecting airport makes their travels not as smooth as flying on direct flights. In terms of WTP, the business group exhibits about twelve times higher WTP than the tourists do: $\frac{\psi_{21}}{\alpha_1} / \frac{\psi_{22}}{\alpha_2} = 11.8$. *Distance* has a significantly positive coefficient, 0.105. In short-haul markets, airline products are competing with other transportation modes such as cars, buses, or trains. As a traveling distance increases, however, the substitutability to the outside goods becomes worse so that demand for air travel can grow.[36] *HubDM* also has a positive coefficient, 0.056. It indicates that carriers attract more passengers at their hub airports.

---

34 More specifically, one can set up a dynamic model where a consumer's decision at time t depends on past experiences of flight delays at time t-1,...,t-N. One good reference is Suzuki (2000) who developed an aggregate-level Markovian type model.

35 Similar to on-time performance, I posit that consumers form their expectations on whether the baggages will be damaged, lost, or delayed, and wether they will be denied boarding from flights based on past experiences.

36 Many studies controlled distance squared to capture the curvature of demand. For example, Berry and Jia (2010) found negative sign of distance squared, implying that further increase in distance makes the travel less pleasant.

Borenstein (1989) called this phenomenon *airport dominance* by major carriers. The positive parameter is consistent with the finding. [37] The coefficient of *Slot-control* is -0.071, indicating that passengers get disutility from traveling through slot-controlled airports. An obvious source is flight delays frequently observed at these airports. However, since this study controls the delays by *Ontime*15, I interpret the disutility to mean fatigue and discomfort passengers endure at the congested airports. It can include a longer waiting time at ticket check-in counter and security check gates. The positive coefficient of *Tour* supports the well-known fact that tourist places attract more passengers.

The nested logit parameter λ measures the degree of product differentiation between all airline products. If λ is equal to 1, air transportation services are perfectly differentiated. The estimate 0.618 implies that there exists a mild substitution possibility among airline services. Finally, γ_1 measures the percentage of type 1 passengers in the population. The parameter 0.052 indicates that the business group accounts for only 5.2% of the potential travelers. However, the business passengers are much more likely to actually buy ticket compared to the price-sensitive tourists. Based on the consideration, I calculate the percentage of each type of consumers in the sample and find that the business group makes up 40.5% of the actual travelers.[38]

---

37 Borenstein (1989) pointed out the airport dominance as the main cause of hub premium. A body of related studies suggest that the airport dominance is possible because of more convenient gate access and higher expected value from frequent flier program at hub airport. Recently, Lee (2013) suggests that the airport dominance is based on the gate contract between airport and major carriers. The estimates of a structural model reveal that a major carrier's gate dominance at its hub airport has a positive effect on consumers' utility.

38 The estimates are close to those in Berry, Carnall and Spiller (2006). Based on various specifications, they reported that the business travelers make up 2.5%-7.7% in the population, and 26.8%-39.9% in the sample. I calculate the percentage of the business group in the sample as:

$$\sum_{t=1}^{T} M_t \cdot \hat{\gamma}_1 \cdot \frac{D_{1t}^{\lambda}}{1+D_{1t}^{\lambda}} / \sum_{t=1}^{T} \sum_{r=1}^{2} M_t \cdot \hat{\gamma}_r \cdot \frac{D_{rt}^{\lambda}}{1+D_{rt}^{\lambda}}, \quad \text{where}$$

$$D_{rt} = \sum_{k \in J_t} e^{x_{kt}\hat{\beta}_r + \hat{\xi}_{kt}/\hat{\lambda}}.$$

4.2.2. Marginal and Fixed Cost Parameters

The second column in table 4 presents the cost parameters. Marginal cost parameters are estimated by regressing the difference between price and estimated markup on the marginal cost characteristics. Starting with *Frequency*, the parameter -0.021 indicates that when a carrier adds one more departure per day for a specific route, the cost of serving an additional passenger tends to decrease by $2.1. Greater *Frequency* contributes to increasing aircraft utilization (block hours per day) and to reducing turnaround times at airports. This makes per-flight and per-passenger cost decrease. The parameter of *HubMC* (-0.184) indicates that the existence of hub airport on an itinerary tends to decrease marginal cost by $18.4. Among two countervailing effects (described in section 2.2), the negative sign supports that cost reduction from high load factor is greater.

As expected, $Layovers_{short}$ and $Layovers_{long}$ have positive coefficients, 0.175 and 0.300, respectively. The additional fuel that a connecting flight spends during extra landing/takeoff increases marginal cost substantially. $Distance_{short}$ and $Distance_{long}$ also have positive coefficients, 0.226 and 0.130, respectively. As a market distance increases, the cost of carrying one more passenger rises. Interestingly, given the tendency of larger airplanes to serve long-haul markets, the *Layovers* and *Distance* coefficients imply that larger aircrafts tend to consume relatively more fuel during landing and takeoff phases, but tend to exhibit high fuel efficiency in the air.

The coefficients of carrier dummy variables show that American Airlines (omitted as a base carrier) appears to have the highest marginal cost, followed by US Airways, Delta, and Northwest in order of high cost. In order to check the validity of the carrier-specific cost effect, I looked into each carrier's operating cost per available seat mile (CASM) during the sample periods, using Air Carrier Financial Statistics (Schedule P-12). Table 5 and figure 2 indicate that the order of US Airways-Delta-Northwest still stands in CASM data. However, American Airlines reports the lowest CASM, which seems curious. This implies that there

can exist other factors which are not captured by the model.[39] The low cost carriers Southwest and AirTran have reasonably low level of marginal costs than the legacy carriers.

The fixed cost parameters are estimated by regressing the derivative of the variable profit function on the fixed cost characteristics. Notice that the dependent variable is equivalent to the slope of fixed cost by equation (6) and (12), and thus the constant terms measure the marginal effect of the characteristics on the fixed cost. The coefficient of $Ontime15_{constant}$ indicates that as on-time performance improves from 0% to 100%, the fixed cost increases by $0.24 million. Although the on-time performance is largely affected by exogenous factors such as weather, carriers can still take steps to improve it. They can make an investment to

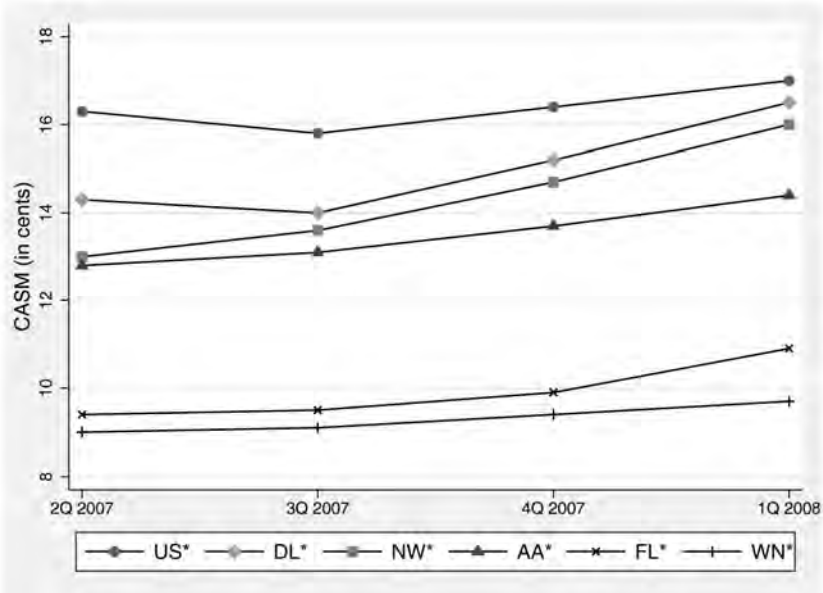
Table 5 Operating Cost per Available Seat Mile (CASM, in cents)

	2Q 2007	3Q 2007	4Q 2007	1Q 2008	Average
(**By carrier**)					
US*	16.3	15.8	16.4	17	16.4
DL*	14.3	14	15.2	16.5	15.0
NW*	13	13.6	14.7	16	14.3
UA	13	13.3	14.6	14.9	14.0
CO	13.5	13.5	14	14.9	14.0
AA*	12.8	13.1	13.7	14.4	13.5
FL*	9.4	9.5	9.9	10.9	9.9
WN*	9	9.1	9.4	9.7	9.3
(**By carrier type**)					
Legacy carriers	13.4	13.6	14.5	15.3	14.2
LCC carriers	9.6	9.6	9.5	10	9.7

Notes: Data sources of CASM are Air Carrier Financial Statistics (Schedule P-12) and T-100 Domestic Segment from U.S. DOT. The asterisk indicates that the brand-specific effects of the carriers are statistically significant.

39 One interesting point in figure 2 is that all carriers experienced substantial increases in CASM from the fourth quarter of 2007. This is mainly due to high fuel prices during the U.S. economic recession (beginning December 2007). Specifically, Schedule P-12 data reveal that legacy and LCC carriers, respectively, spent 23.0% and 27.3% of their operating costs for fuel in 2006, however, the proportions increased up to 30.2% and 34.7% by 2008.

Figure 2 | Operating Cost per Available Seat Mile (CASM)



Notes: Data sources of CASM are Air Carrier Financial Statistics (Schedule P- 12) and T-100 Domestic Segment from U.S. DOT.

adopt newer aircraft with fewer maintenance problems, more efficient fuel/food delivery system, advanced crew scheduling, and better boarding procedures. All these steps increase the fixed cost significantly.

$Frequency_{constant}$ also has positive coefficient 6.605. Although raising *Frequency* reduces marginal cost, it increases the fixed cost. Considering that increased frequency on a certain route requires more economic resources such as fleets, pilots and crew-member, ground-side services, the result makes sense.

Coefficients of $Mishandled\ baggage_{constant}$ and $Denied\ boarding_{constant}$ are -8.126 and -31.634, respectively. They suggest that as each of them decreases, the fixed cost increases. Since decreases in the characteristics make airline products better, the negative signs make intuitive sense. In order for baggages to be in the right place at the right time, efficient equipment and well-trained agents (e.g. check-in agents, ramp agents, and baggage handlers) are necessary at each baggage-handling point.

Similarly, reducing the number of involuntarily bumping passengers (without hurting the load factor) needs to apply sophisticated forecasting system and to increase the aircraft capacity to some extent. All these improvements lead to higher fixed cost.

5. Merger Simulations

The primary purpose of this paper is to simulate how a merged carrier adjusts the product characteristics and prices, and how the post-merger equilibrium affects welfare. In this section, I simulate the Delta and Northwest Airlines merger based on pre-merger data, the structural model, and the parameter estimates. Section 5.1 describes simulation methodology, and section 5.2 provides the detailed simulation results. In section 5.3, I report changes in consumer and producer welfare. Finally, section 5.4. evaluates the simulation result by comparing it with actual post-merger data.

5.1. Simulation Methodology

I perform the simulation based on the last two quarters in the pre-merger sample,[40] and focus on the markets where the Delta and Northwest Airlines competed with each other.[41] Table 6 describes several statistics for the simulation sample. The merging airlines competed in 1,129

40 At an early stage of this study, I simulated based on the last quarter in the pre-merger sample as most merger studies did. The results from that sample are largely consistent with what will be reported in section 5.2. However, since the last quarter has only three monopoly markets after ownership consolidation, I expand the simulation sample to provide more robust results not only for oligopoly markets but for monopoly markets.

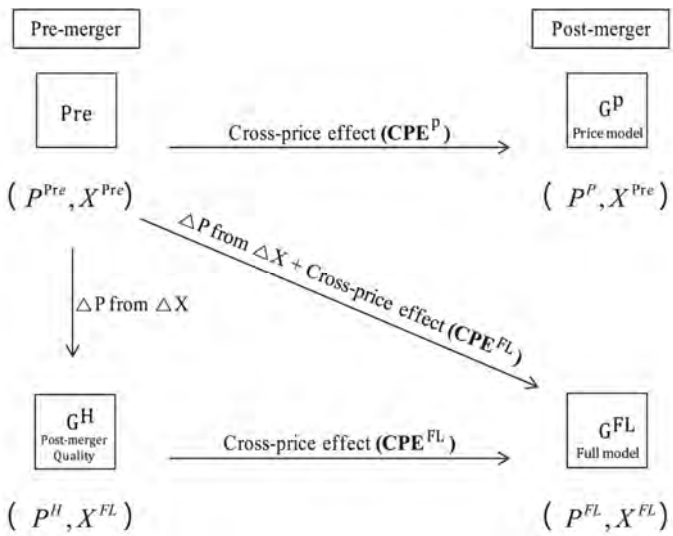
41 The focus on overlapped markets does not necessarily mean that merger effect in other markets is negligible. A series of papers studied the spill-over effect over non-overlapped markets. However, standard merger simulation has focused on the overlapped markets where merger effect arises from a loss of competition. To compare it with my simulation, I also concentrate those markets.

Table 6 Description of Simulation Sample

Integration steps		
Announcement	2Q. 2008	
Completion	4Q. 2009	
Resulting entity	Delta Airlines	
Simulation sample (pre-merger)		
Sample period	4Q. 2007~1Q. 2008	
Number of markets overlapped	1,129	
- Duopoly / Oligopoly markets	9 / 1,120	
Statistics by carrier	Delta Airlines	Northwest Airlines
Passenger share (per market) (%)	0.17	0.17
Number of products (per market)	2.62	2.13
Fare ($100)	3.81	3.35
On-time performance (%)	0.80	0.70
Flight frequency (per day)	4.51	3.35
Mishandled boarding rate (per 1,000 passengers)	7.51	4.67
Denied boarding rate (per 1,0000 passengers)	1.45	0.70

Notes: Simulation sample consists of the last two quarters of the estimation sample. The simulation is conducted for 1,129 overlapped market pre-merger.

Figure 3 Simulation Design for Decomposing Sources of Price Change



Notes: *Pre* is the actual pre-merger data. G^P and G^{FL} indicate games based on the price model and the full model. P is a price and X is a vector of the endogenous product characteristics. G^H is a hypothetical game.

overlapped markets, including nine duopoly markets prior to the merger. They had very similar passenger share per market, but their products were significantly different in terms of the price and the characteristics.

Figure 3 illustrates three separate games: price model G^P, full model G^{FL}, and hypothetical model G^H. Pre is actual pre-merger data where P^{Pre} is a price, and X^{Pre} is a vector of the endogenous characteristics. In the price model, carriers can change only prices post-merger, holding the characteristics fixed at pre-merger level. This game corresponds to the standard merger simulation where change in price $P^P - P^{Pre}$ measures the cross-price effect CPE^P from the merger. On the other hand, the full model allows carriers to adjust both prices and the characteristics. In this case, $P^{FL} - P^{Pre}$ represents not only the cross-price effect CPE^{FL} but also demand and cost-driven effects $\Delta P \; from \; \Delta X$ (explained in section 1). I decompose the price change in the full model into two separate effects by simulating the hypothetical model. This game assumes pre-merger situation as if the Delta and Northwest Airlines are separate carriers, but the characteristics are hypothetically equated to the post-merger characteristics in full model X^{FL}. Since this game does not consider the ownership consolidation, the price change comes from the adjustment of characteristics, that is, $P^H - P^{Pre}$ measures $\Delta P \; from \; \Delta X$. Consequently, $P^{FL} - P^H$ identifies CPE^{FL}.[42] Notice that magnitudes of two cross-price effects will be different because the product repositioning in the full model can cause higher differentiation or higher substitutability between the merged firm's products. In the case of higher differentiation, we expect $CPE^P > CPE^{FL}$, otherwise $CPE^{FL} > CPE^P$.

The full model derives post-merger characteristics and prices sequentially. Based on the post-merger ownership of the products, the simulation searches $X_f^{FL} = (\overline{x}_f^{k*}, k = O, F, M, D)$ for firm f by solving

42 This simulation design is an application of price-location game in Gandhi et al. (2008).

$$X_f^{FL} = \text{argmin} \sum_{\text{k}} \frac{\partial \prod_f^{I\prime}}{\partial \overline{x}_f^k} \frac{\partial \prod_f^I}{\partial \overline{x}_f^k}, \quad k = O, F, M, D \tag{18}$$

where $\frac{\partial \prod_f^I}{\partial \overline{x}_f^k}$ is the necessary optimality condition (12) at the first-stage game. After deriving X^{FL} for all carriers in all markets, it continues to derive P_f^{FL} by solving the optimal price function

$$P_f^{FL} = \widehat{mc}_f^* - \Omega_{s_f, P_f}^{\text{post}}\left(P^{FL}, X^{FL}, \hat{\xi}; \hat{\theta}_d\right)^{-1} \cdot s_f(P^{FL}, X^{FL}, \hat{\xi}; \hat{\theta}_d) \tag{19}$$

where $\widehat{mc}_f^*$ is marginal cost estimates calculated by using post-merger characteristics X_f^{FL} and Ω_{s_f, P_f}^{post} is an analogous matrix to (10) based on post-merger ownership structure. I iterate this sequential process once more in spirit of best-response iteration.[43] The price model and the hypothetical model skip the derivation of a new vector of product characteristics and search only new optimal prices.

After simulating the post-merger equilibrium, I compute quality index Q for each product by taking an inner product of the set of endogenous characteristics and their respective parameters,

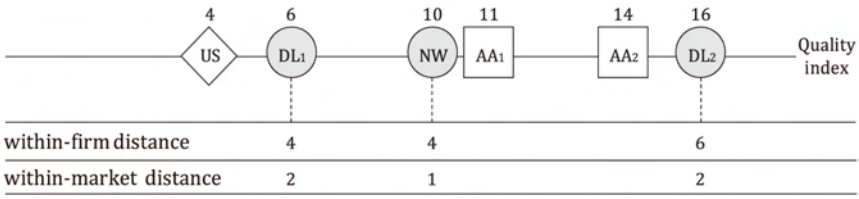
$$Q_{jt} = \sum_{r=1}^{2} \sum_{k} \gamma_r \cdot \overline{x}_{jt}^k \beta_r^k, \quad k = O, F, M, D \tag{20}$$

where γ_r is the percentage of type r passengers in the population.[44]

43 I used a very tight tolerance to compute the equilibrium. The tolerance levels of product characteristics and prices are 1e-12 and 1e-15, respectively.

44 A primary reason of introducing the scalar-valued quality index is to provide more intuitive interpretations for changes in product characteristics. It does not affect simulation results.

Figure 4 Measures for the Extent of Product Differentiation: Within-firm distance and Within-market distance



Notes: Within-firm distance of a product is the closest quality-distance from itself to other goods produced by the same firm. Within-market distance of a product the closest quality-distance from itself to other goods produced by competitors.

Further, I quantify the magnitude of product differentiation with two quality-distance measures: *within-firm distance and within-market distance* (see figure 4). I define a within-firm distance of product *j* as the closest quality-distance to other goods produced by the same firm. For the merged firm's products, if the distance increases post-merger, it implies that the product becomes more differentiated so that the cross-price effect can be weaker. On the other hand, a within-market distance of product *j* is measured by the closest quality-distance to other goods produced by competitors in the same market. A longer within-market distance post-merger implies that the merged carrier can raise price easily based on less substitutability to its competitors' products.

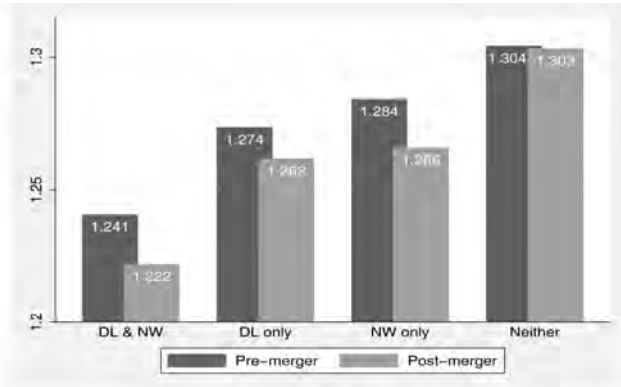
5.2. Simulation Results

5.2.1. Changes in Product Characteristics

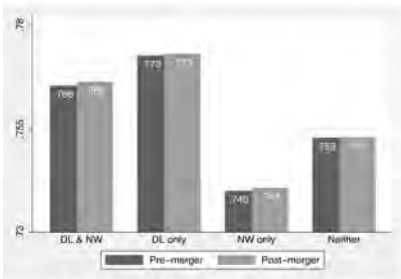
The histograms in figure 5 through 7 describe how the merged firm changes the product characteristics.[45] They show us two important findings: *overall quality degradation and higher product differentiation* post-merger. The first implication is shown by figure 5 (a). While the average quality index decreases for all groups of markets after the merger, the quality degradation is severe in markets where the merging

45 Each number mounted on each bar in figure 5 through 7 is an average value of product quality or characteristics over the corresponding products.

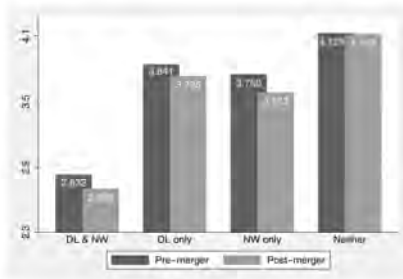
Figure 5 Quality Changes of Merged Firm's Products in All Markets By market power



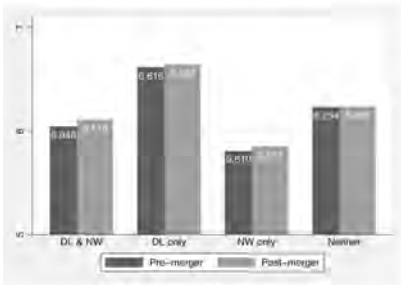
(a) Quality index



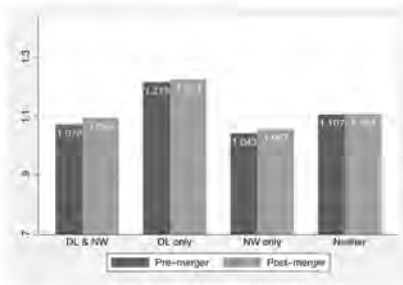
(b) Ontime15



(c) Frequency



(d) Mishandled baggage rate



(e) Denied boarding rate

	DL & NW	DL only	NW Only	Neither
Number of markets	29	219	215	666
Number of products	122	1,011	816	3,410

carriers had market power before the merger.[46] Specifically, markets where both carriers had market power (indicated by DL & NW) show that the quality decreases by 1.5% from 1.241 to 1.222. Sub-figures from 5 (b) to (e) suggest that flight frequency changes the most, decreasing by 4.7% from 2.832 to 2.698. It amounts to 12 less flights for each product during a quarter. The mishandled baggage rate and denied boarding rate increase, respectively, by 1.1% from 6.048 to 6.113 and 2.1% from 1.072 to 1.095, also supporting the quality degradation. When either carrier had market power (DL only, NW only), the quality decreases as well. However, when neither carrier had market power (Neither), the quality rarely changes. In short, the combined firm lowers the product quality especially when it has a strong market power. However, the incentive of quality degradation becomes weaker when

Table 7 Changes in Price and Characteristics of Merged Firm's Products in Oligopoly Markets

A. Large share goods

	Pre-merger	Post-merger	Change
Quality index	1.319	1.330	0.011
Ontime15	0.756	0.756	0.000
Frequency	4.369	4.438	0.069
Mishandled Baggage rate	6.287	6.249	-0.038
Denied Boarding rate	1.141	1.131	-0.010
Quality Distance			
Within-firm distance	0.179	0.200	0.021
Within-market distance	0.111	0.107	-0.004

	Pre	$G^P - Pre$	$G^{FL} - Pre$	$G^H - Pre$	$G^{FL} - G^H$
Fare($)	368.6	4.9	3.7	1.5	2.2
Marginal cost ($)	209.2	0.0	-0.1	-0.1	0.0
Passengers	327.4	-2.6	0.3	4.1	-3.8
Profits ($100)	591.2	1.9	6.4	5.4	1.0

Notes: Large share goods include 692 products in 680 markets. Each number indicates average values over the large goods.

46 I define an airline has market power if it carries more than 25% of total market enplanements.

Table 7 (Continue)

B. Medium share goods

	Pre-merger	Post-merger	Change
Quality index	1.306	1.315	0.008
Ontime15	0.759	0.759	0.000
Frequency	4.097	4.166	0.068
Mishandled Baggage rate	6.241	6.222	-0.019
Denied Boarding rate	1.134	1.128	-0.006
Quality Distance			
Within-firm distance	0.096	0.106	0.011
Within-market distance	0.080	0.081	0.001

	Pre	$G^P - Pre$	$G^{FL} - Pre$	$G^H - Pre$	$G^{FL} - G^H$
Fare($)	352.8	9.5	7.2	0.3	6.9
Marginal cost ($)	229.3	0.0	-0.1	-0.1	0.0
Passengers	47.7	-2.0	-0.8	1.2	-2.1
Profits ($100)	69.4	0.3	1.7	1.3	0.4

Notes: Medium share goods include 1,041 products in 676 markets. Each number indicates average values over the medium goods.

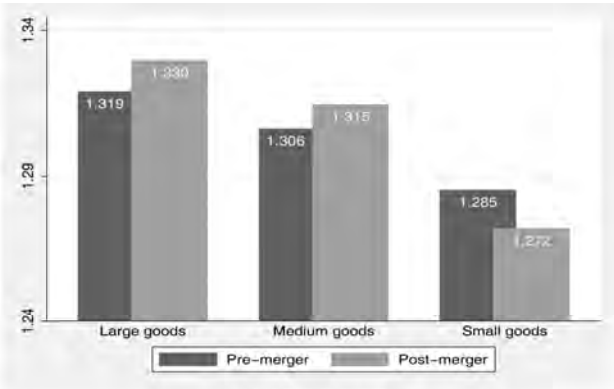
C. Small share goods

	Pre-merger	Post-merger	Change
Quality index	1.285	1.272	-0.013
Ontime15	0.754	0.754	0.000
Frequency	3.889	3.776	-0.113
Mishandled Baggage rate	6.230	6.263	0.033
Denied Boarding rate	1.108	1.119	0.011
Quality Distance			
Within-firm distance	0.064	0.075	0.012
Within-market distance	0.058	0.064	0.006

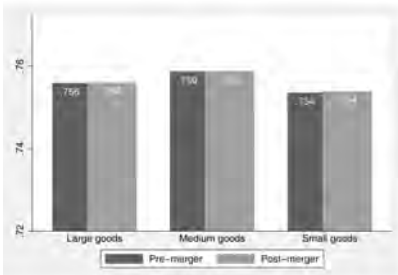
	Pre	$G^P - Pre$	$G^{FL} - Pre$	$G^H - Pre$	$G^{FL} - G^H$
Fare($)	361.1	24.1		1.9	34.2
Marginal cost ($)	233.9	0.0	0.2	0.2	0.0
Passengers	12.4	-1.0	-1.0	0.2	-1.2
Profits ($100)	19.5	0.0	0.1	0.2	-0.1

Notes: Small share goods include 3,598 products in 865 markets. Each number indicates average values over the small goods.

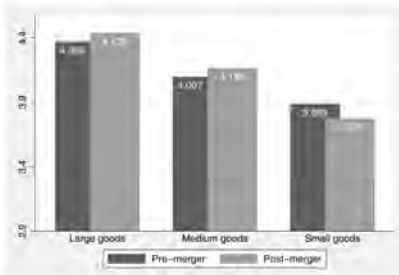
Figure 6 Quality Changes of Merged Firm's Products in Oligopoly Markets By product group



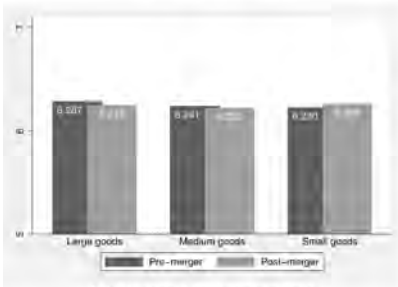
(a) Quality index



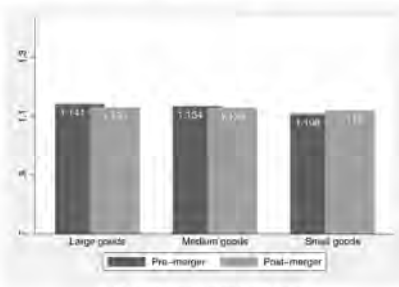
(b) Ontime15



(c) Frequency



(d) Mishandled baggage rate



(e) Denied boarding rate

	Large goods	Medium goods	Small goods
Number of markets	680	676	865
Number of products	692	1,041	3,598

strong competitors exist because the potential risk of losing passengers increases.

To look at the second implication, I divide the merged firm's products into large-, medium-, and small-share goods (henceforth large, medium, and small goods, respectively).[47] Since a product is a unique combination of carrier and route, we can regard a large good as a major route (or a primary good), and a small good as a minor route (or a secondary good) in a market. A medium good refers to a route serving medium-sized enplanements. Figure 6 (a) illustrates the changes in average quality of three product groups in oligopoly markets. For the large goods including 692 products in 680 markets, the merged firm improves the quality by 0.8% from 1.319 to 1.330. Specifically, frequency increases by 1.6%, corresponding to adding 6 more flights per product during a quarter. Mishandled baggage rate and denied boarding rate decrease by 0.6% and 0.9%, respectively. The medium goods show very similar patterns. However, the small goods including 865 products in 3,598 markets deteriorate substantially. The quality index decreases by 1.0% from 1.285 to 1.272 and the underlying characteristics become worse. The frequency reduces by 2.9%, indicating 10 less flights per product during a quarter, and mishandled baggage rate and denied boarding rate increase by 0.5% and 1.0%, respectively.[48] To sum up, the merged firm increases the product differentiation post-merger by upgrading large and medium goods and downgrading small goods.

The higher product differentiation post-merger can be understood by the firm's profit-maximizing behavior. As I will show in subsection 5.2.2, the average profit from large goods is much bigger than that from small goods. This motivates the merged firm to move passengers from small goods to large or medium goods by adjusting the product quality. To verify this argument, I compute the number of passengers of each type

---

47 If a product serves more than 50% of the carrier's enplanements in a market, it is defined as a large good. If less than 50% but at least 20%, it belongs to medium goods. The remaining goods with less than 20% constitute small goods.

48 I tested different definitions of large, medium, and small goods. The quality index changed slightly depending on the definitions, but directions of quality changes were highly robust. The test results are available upon request.

who purchase the large, medium, and small goods pre- and post-merger.[49] It reveals that while total number of tickets sold in oligopoly markets decreases post-merger, the proportion of large goods increases from 70.6% to 71.6% and that of small goods decreases from 14.0% to 13.0%.[50] Specifically, table 14 shows that the business group purchases more large goods and less small goods, and the tourists' consumptions decrease the most for small goods after the merger. Intuitively, the large goods are associated with major routes where a carrier's hub airports exist and they generate considerable profits. Therefore, the merged firm takes better care of those routes to attract more passengers to them.

Figure 7 (a) provides the quality adjustment in monopoly markets. The simulation sample contains only 28 products in 9 monopoly markets, but the pattern of higher product differentiation post-merger still stands. A notable thing in monopoly markets is that the quality changes are greater in absolute value, relative to oligopoly markets. The quality of large goods increases by 6.4% from 1.342 to 1.428, and that of small goods decreases by 7.4% from 1.328 to 1.230. We can understand the greater quality changes by monopolist with table 14 again. After the merger, the business group takes higher proportion of large goods in monopoly markets (66.7%) than in oligopoly markets (53.9%). Also, the tourists buy bigger proportion of small goods in monopoly markets (61.6%) compared to oligopoly markets (55.1%). Without competition, monopolist can adjust product quality more flexibly toward extracting more profits from each group. Even though the provided qualities can be higher or lower than most preferred level by each type, consumers are more forced to choose a particular product as the monopolist leads.

---

49 For example, I computed a percentage of the business passengers who actually bought large goods in the sample as:

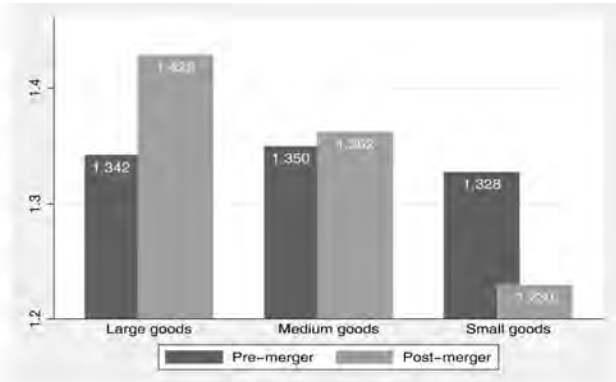
$$\sum_{t=1}^{T} M_t \cdot \hat{\gamma}_1 \cdot \frac{L_{1t}}{D_{1t}} \frac{D_{1t}^{\lambda}}{1 + D_{1t}^{\lambda}} / \sum_{t=1}^{T} \sum_{r=1}^{2} M_t \cdot \hat{\gamma}_r \cdot \frac{L_{rt}}{D_{rt}} \frac{D_{rt}^{\lambda}}{1 + D_{rt}^{\lambda}}, \quad \text{where}$$

$D_{rt} = \sum_{k \in J_t} e^{x_{kt}\hat{\beta}_r + \hat{\xi}_{kt}/\hat{\lambda}}$, $L_{rt} = \sum_{k \in J_t} e^{(x_{kt}\hat{\beta}_r + \hat{\xi}_{kt})/\hat{\lambda}}$, and L_t is the set of large goods produced by Delta or Northwest Airlines in overlapped market *t*.

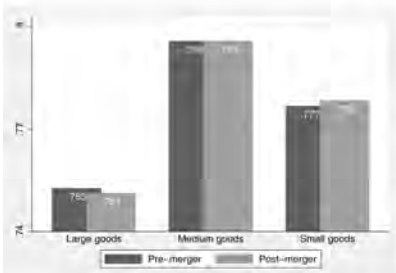
50 We can check this based on table 14.

Figure 7 Quality Changes of Merged Firm's Products in Monopoly Markets

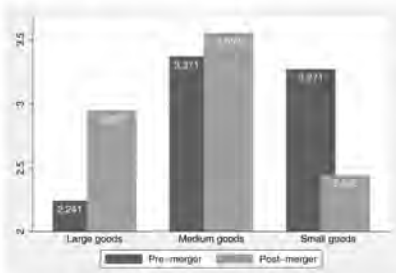
By product group



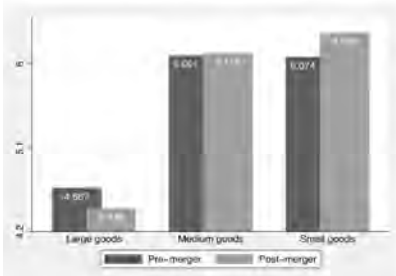
(a) Changes in Quality



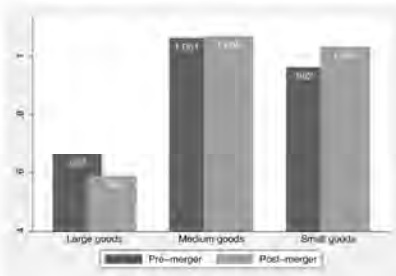
(b) Ontime15



(c) Frequency



(d) Mishandled baggage rate



(e) Denied boarding rate

	Large goods	Medium goods	Small goods
Number of markets	6	5	6
Number of products	6	8	14

5.2.2. Changes in Prices

How and to what extent does the product repositioning affect post-merger prices? I present the results in table 7 and 8. Each table reports the quality index, the endogenous characteristics (whose values are identical to those in figure 6 and 7), and the quality-distance measures pre- and post-merger. Importantly, the bottom panel describes the simulated price changes from three separate games. [51]

Table 7A is about the large goods in oligopoly markets. Notably, the within-firm distance increases by 0.021 post-merger. This implies that the large goods become more differentiated from medium and small goods due to the quality improvement. Meanwhile, the within-market distance decreases very little, indicating that they become slightly more substitutable to competitors' products. Based on the repositioning of large goods, we can expect the merged firm to have a difficulty in internalizing cross-price effect.

I examine this hypothesis by comparing the cross-price effects CPE^P and CPE^{FL} in the bottom panel. The second column indicates the cross price effect in the price model: CPE^P =\$4.9. The third and fourth column present price changes under the full model and the hypothetical model, respectively, and finally the cross-price effect in the full model is calculated in the last column: CPE^{FL} =\$2.2. Consistent with the hypothesis, the merged carrier internalizes a lower cross-price effect given the quality adjustment: $CPE^P > CPE^{FL}$. Even though the quality improvement causes price increase: $\Delta P \; from \; \Delta X = \$1.5$, it does not surpass the reduction of cross-price effect. Therefore, when the product characteristics are endogenized, the simulation predicts a lower price increase than the typical merger analysis due to the higher product differentiation: $P^P > P^{FL}$.

If so, why does the merged firm raise the product differentiation? This issue leads us to see profit changes. In table 7A again, the full model predicts lower marginal cost and more passengers relative to the price model. The cost reduction is possible due to the increased frequency, and

---

51 Each number in the table 7 and 8 indicates the average value over the corresponding products.

the attraction of more consumers is based on the enhanced quality. All these changes allow the merged firm to increase profit per product by \$540 in the full model, which is much bigger than \$190 in the price model. The additional profits correspond to \$0.37 million and \$0.13 million, respectively, when multiplied by the number of large goods. To sum up, although the higher product differentiation reduces the ability of raising price, it can generate more profit.[52]

The medium goods show very similar patterns to the case of large goods (see table 7B). One difference is that the within-market distance slightly increases post-merger. It can positively affect the cross-price effect in the full model. However, it still predicts smaller price increase than the price model by showing $CPE^{P}(\$9.5) > CPE^{FL}(\$6.9)$ and $P^{P} > P^{FL}$.

Interestingly, small goods show very different aspects compared to large and medium goods. Even though the small goods become more differentiated from other product groups due to their quality degradation (the within-firm distance increases by 0.012 post-merger), the full model predicts a greater cross-price effect than the price model: $CPE^{FL}(\$34.2) > CPE^{P}(\$24.1)$. It seems implausible, but is still consistent with profit-maximizing decision in two aspects. First, unlike other product groups, the within-market distance considerably increases, implying that the small goods become less substitutable to competitors' products. This encourages the merged firm to increase price. Second, small goods generate very small profit per product pre-merger. The profits are \$59.1k, \$6.9k, and \$2.0k for large, medium, and small goods, respectively. Thus, the significant price increase (together with the quality degradation) of small goods can contribute to transferring consumers to other profitable goods.

52 Section 5.3 will address the profit analysis more thoroughly.

Table 8 Changes in Price and Characteristics of Merged Firm's Products in Monopoly Markets

A. Large share goods

	Pre-merger	Post-merger	Change
Quality index	1.342	1.428	0.086
Ontime15	0.753	0.751	-0.001
Frequency	2.241	2.950	0.709
Mishandled Baggage rate	4.667	4.449	-0.217
Denied Boarding rate	0.663	0.590	-0.073
Quality Distance			
Within-firm distance	0.097	0.295	0.199
Within-market distance	–	–	–

	Pre	$G^P - Pre$	$G^{FL} - Pre$	$G^H - Pre$	$G^{FL} - G^H$
Fare($)	428.0	60.9	26.6	1.6	25.0
Marginal cost ($)	115.1	0.0	-1.5	-1.5	0.0
Passengers	53.3	-3.6	3.8	5.5	-1.7
Profits ($100)	201.7	1.5	14.0	12.2	1.8

Notes: Large share goods include 6 products in 6 markets. Each number indicates average values over the large goods.

B. Medium share goods

	Pre-merger	Post-merger	Change
Quality index	1.350	1.362	0.012
Ontime15	0.796	0.796	0.000
Frequency	3.371	3.553	0.182
Mishandled Baggage rate	6.091	6.115	0.024
Denied Boarding rate	1.061	1.069	0.008
Quality Distance			
Within-firm distance	0.059	0.209	0.150
Within-market distance	–	–	–

	Pre	$G^P - Pre$	$G^{FL} - Pre$	$G^H - Pre$	$G^{FL} - G^H$
Fare($)	370.3	65.0	48.0	2.3	45.7
Marginal cost ($)	175.1	0.0	-0.4	-0.4	0.0
Passengers	16.3	-2.5	0.8	2.9	-2.1
Profits ($100)	28.5	1.0	4.9	4.3	0.5

Notes: Medium share goods include 8 products in 5 markets. Each number indicates average values over the medium goods.

Table 8 (Continue)

C. Small share goods

	Pre-merger	Post-merger	Change
Quality index	1.328	1.230	-0.098
Ontime15	0.777	0.779	0.002
Frequency	3.271	2.445	-0.826
Mishandled Baggage rate	6.074	6.335	0.261
Denied Boarding rate	0.962	1.032	0.070
Quality Distance			
Within-firm distance	0.102	0.218	0.116
Within-market distance	–	–	–

	Pre	$G^P - Pre$	$G^{FL} - Pre$	$G^H - Pre$	$G^{FL} - G^H$
Fare($)	336.8	113.4	147.0	1.8	145.2
Marginal cost ($)	191.3	0.0	1.7	1.7	0.0
Passengers	7.8	-1.8	-2.6	-1.1	-1.4
Profits ($100)	11.0	0.4	-0.4	-1.0	0.5

Notes: Small share goods include 14 products in 6 markets. Each number indicates average values over the small goods.

Table 8 reports the simulation results for monopoly markets. While they exhibit similar patterns, one difference is that prices of all product groups increase to a greater extent in the monopoly markets than in the oligopoly markets. One possible explanation is that the monopolist does not consider the within-market substitutability.

5.3. Welfare Analysis

This section assesses how the post-merger equilibrium affects the welfare. The demand parameters indicate that two types of consumers have heterogeneous tastes for the characteristics and price. On the supply side, the simulation results reveal that the merged firm repositions the large, medium, and small goods differently. This encourage us to examine the consumer surplus of each type and the producer surplus from each product group.

5.3.1. Consumer Welfare

I measure changes in consumer welfare by the compensating variation.

Table 9 Change in Consumer Surplus (CS) after the Delta and Northwest Airlines Merger

Markets	Price model (G^P)		Full model (G^{FL})		Quality change of DL/NW	Number of products (markets)
	Change in CS ($100K)	% Change in CS (%)	Change in CS ($100K)	% Change in CS (%)		
All markets						
Total	-15.59	-0.18	18.53	0.21	-0.006	18,430 (1,129)
Business	-9.87	-0.12	22.16	0.28		
Tourists	-5.72	-0.72	-3.63	-0.45		
By market competitiveness						
Monopoly	-0.14	-4.18	0.06	1.73	-0.027	28 (9)
Business	-0.08	-2.63	0.07	2.30		
Tourists	-0.06	-22.81	-0.01	-5.13		
Oligopoly	-15.45	-0.18	18.47	0.21	-0.006	8,402 (1,120)
Business	-9.79	-0.12	22.09	0.28		
Tourists	-5.66	-0.71	-3.62	-0.45		
By quality change						
QI markets	-12.75	-0.19	19.86	0.30	0.012	12,782 (712)
Business	-8.22	-0.14	22.05	0.36		
Tourists	-4.53	-0.69	-2.19	-0.33		
QD markets	-2.84	-0.14	-1.34	-0.07	-0.042	5,648 (417)
Business	-1.65	-0.09	0.11	0.01		
Tourists	-1.18	-0.83	-1.44	-1.01		

Notes: QI and QD markets indicate quality-increase and quality-decrease markets, respectively. If a passenger-weighted average quality of merged firm's products increases after the merger, it belongs to QI markets, otherwise it belongs to QD markets. Number of products counts not only merged firm's products but also competitors' products.

Following Small and Rosen (1981), the compensating variation for a type *r* passenger in market *t* is given by

$$CV_{rt}\frac{V_{rt}^{pre}-V_{rt}^{post}}{\alpha_r}, \tag{21}$$

where $\alpha_r < 0$ is the marginal disutility from price increase. Pre-merger term is defined as $V_{rt}^{pre} = \ln\left[1 + \left(\sum_{j\in J_t} e^{(x_{jt}^{pre}\beta_r+\xi_{jt})/\lambda}\right)^{\lambda}\right]$, and V_{rt}^{post} is analogously defined to V_{rt}^{pre} replacing x_{jt}^{pre} by x_{jt}^{post}. Then, the change in the average per-passenger surplus in market t is measured by $CS_t = \sum_{r=1}^{2}\gamma_r \cdot CV_{rt}$, and the change in total consumer surplus is the sum of CS_t in all markets: $CS = \sum_t M_t \cdot CS_t$ where M_t is the market size.

Table 9 reports the welfare effect based on the price model G^P and the full model G^{FL}. In the first panel covering all markets, G^P predicts a decrease in consumer welfare for both types. Since this game predicts substantial price increase, holding the product characteristics fixed at pre-merger level, the welfare loss is a natural result.

However, G^{FL}predicts substantially different outcomes in two aspects. First, the overall consumer welfare increases. To be specific, the tourists still experience the welfare losses (-$0.36 million), but the business passengers benefit significantly from the merger ($2.22 million). For the price-insensitive business group, utility gains from large and medium goods (associated with the quality improvement and the lower price increase) are greater than their losses from small goods (associated with the quality degradation and the greater price increase). However, for the price-sensitive tourists, the welfare losses from the price increase surpass the potential gains from quality improvement of large and medium goods. In overall, the amount of benefits to the business group largely surpasses the losses to the tourists.[53] The result reveals that if the set of repositioned

53 As section 4.2 showed, the tourists account for 59.5% of actual travelers and for 94.8% of potential travelers. Considering the significant proportions, it seems unreasonable that the tourists' welfare losses are much smaller than the business travelers' gains. However, the tourists are shown to exhibit about ten times as much price sensitivity as the business group ($\alpha_2/\alpha_1 = 10.2$). When computing the compensating variation for each type, a change in the indirect utility is divided by the respective price coefficient

products exhibits more differentiation post-merger, it mitigates the welfare loss from the price increase and even leads to increases in consumer welfare. Recent studies on endogenous product choice have found that merger can have a positive effect on consumer welfare if the merged firm changes its product offerings which consumers value more (see, e.g. Mazzeo, Seim and Varela, 2013). My finding is consistent with the literature and provides new evidence from the airline industry.

Second, the tourist group experiences smaller loss in G^{FL} than in G^{P}. As table 14 shows, the number of tourists who purchase large or medium goods is not much different pre- and post-merger, but for small goods, it largely decreases. That is, a substantial portion of the tourists moves to large, medium, or outside goods, facing the dramatic increase in price of small goods.[54] Since G^{FL} predicts lower prices for large and medium goods than G^{P}, the welfare losses by the tourists become smaller in G^{FL}.

In the second panel table 9, we can observe the same patterns of welfare changes in both monopoly and oligopoly markets. The bottom panel presents another aspect. I divide markets into Quality-increase (QI) and Quality-decrease (QD) markets. If a weighted average quality of the merged firm's products in a market increases post-merger, I define it as QI market, otherwise it belongs to QD market.[55] While the welfare changes in the QI markets follow the overall trend well, the consumer surplus in QD markets become worse off. This is because the overall quality degradation prevents the business group from getting significant welfare gains. I compare the features of two groups of markets (see table 12) and observe that QI markets consist of more competitive routes where the merging carriers had relatively small market presence pre-merger. This confirms that the lack of market competition results in worse product quality, and thus negatively affect the consumer welfare.

---

for converting it into dollar value. This makes a scale of the tourists' losses drop to a tenth so that it becomes largely surpassed by business travelers' gains.

54 In table 7 and 8, G^{FL} predicts that small goods become more expensive than large and medium goods on average post-merger.

55 I use the number of passengers of each product as the weight.

5.3.2. Profits and Social Welfare

Endogenizing product characteristics crucially affects firms' profits as well. In the top panel of table 10, the price model G^P predicts that merged carrier increases profits by \$0.17 million, but the merger lowers the competitors' profits by \$0.30 million, causing the overall producer surplus to decrease. On the other hand, the full model G^{FL} forecasts further increases in the merged firm's profit by \$0.68 million and smaller decreases in the competitors' profits by \$0.12 million, leading to increase in the producer surplus. A closer look at the computed outcome reveals that the higher gain to the merged firms is due to higher markup, and the lower losses to the competitors are based on increased number of consumers who switch from the merged firm due to the overall quality degradation and the price increase.[56] Once two carriers are combined, the pre-merger characteristics may no longer be at the profit-maximizing level. G^P ignores this, but G^{FL} finds new equilibrium characteristics and prices allowing higher profits for the merged carrier. The pattern of profit changes is observable in monopoly, oligopoly, and QI markets.

In QD markets, however, not only competitors but also the merged firm loses profit in G^{FL}, even though the amount of loss by the merged firm is very small. The main reason is a large decrease in the passenger enplanements. The merged firm carries less 0.9% of passengers in QI markets, but it loses 2.6% of passengers in QD markets, which is large enough to completely offset the gains from higher markups. Finally, the bottom panel shows that the merged firm increases profit from all groups of products, but mostly from large and medium goods.

Table 11 describes change in the social welfare. Expectedly, two simulations produce completely different outcomes. While G^P leads the social welfare to decrease by \$1.70 million, G^{FL} predicts it to increase by \$2.41 million based on the increase not only in consumer surplus but

56 The average markups of the merged firm are \$150.1 and \$157.4 in G^P and G^{FL}, respectively, and the average passenger enplanements by competitors are 172,304 and 172,733 in G^P and G^{FL}, respectively.

Table 10 Change in Producer Surplus (PS) after the Delta and Northwest Airlines Merger

Markets	Pre-merger profit ($100K)	Price model (G^P)		Full model (G^{FL})		Number of products	Number of markets
		Change in PS ($100K)	% Change in PS (%)	Change in PS ($100K)	% Change in PS (%)		
All markets							
Total	2632.3	-1.38	-0.05	5.58	0.21	18,430	
DL/NW	553.3	1.66	0.30	6.82	1.23	5,359	1,129
Competitors	2079.0	-3.03	-0.15	-1.24	-0.06	13,071	
By market competitiveness							
Monopoly	1.6	0.02	1.40	0.12	7.35	28	9
Oligopoly	2630.7	-1.40	-0.05	5.46	0.21	18,402	
DL/NW	551.7	1.63	0.30	6.70	1.21	5,331	1,120
Competitors	2079.0	-3.03	-0.15	-1.24	-0.06	13,071	
By quality change							
QI market	2026.5	-0.92	-0.05	5.72	0.28	12,782	
DL/NW	384.3	1.18	0.31	6.83	1.78	3,549	712
Competitors	1642.2	-2.10	-0.13	-1.12	-0.07	9,233	
QD market	605.8	-0.46	-0.08	-0.14	-0.02	5,648	
DL/NW	169.0	0.47	0.28	-0.01	-0.01	1,810	417
Competitors	436.8	-0.93	-0.21	-0.12	-0.03	3,838	
By product quantity (for DL/NW only)							
Large goods	410.3	1.32	0.32	4.51	1.10	698	686
Medium goods	72.5	0.30	0.41	1.80	2.48	1,049	681
Small goods	70.5	0.03	0.05	0.51	0.72	3,612	871

Notes: QI and QD markets indicate quality-increase and quality-decrease markets, respectively. If a passenger-weighted average quality of merged firm's products increases after the merger, it belongs to QI markets, otherwise it belongs to QD markets. Markets for large, medium, and small goods are not mutually exclusive.

Table 11 Change in Total Surplus (TS) after the Delta and Northwest Airlines Merger (unit: $100K)

Markets	Price model (G^P)			Full model (G^{FL})		
	Change in CS	Change in PS	Change in TS	Change in CS	Change in PS	Change in TS
All markets	-15.59	-1.38	-16.96	18.53	5.58	24.11
Monopoly	-0.14	0.02	-0.12	0.06	0.12	0.17
Oligopoly	-15.45	-1.40	-16.85	18.47	5.46	23.93
QI markets	-12.75	-0.92	-13.67	19.86	5.72	25.58
QD markets	-2.84	-0.46	-3.29	-1.34	-0.14	-1.47

Notes: QI and QD markets indicate quality-increase and quality-decrease markets, respectively. If a passenger-weighted average quality of merged firm's products increases after the merger, it belongs to QI markets, otherwise it belongs to QD markets.

also in producer surplus. The quality improvement of large and medium goods contributes to increasing utility gains especially of business travelers, on the other hand, the merged firm extracts more profits from the consumers who switch to more profitable goods. In overall, when a merger simulation endogenizes product characteristics, it produces quite different results from traditional simulation in terms of the post-merger equilibrium and the welfare effects.

5.4. Comparison between Simulation Result and Actual Post-merger Outcome

In this section, I evaluate the predictive performance of my simulation by comparing the simulated result with actual post-merger data. Notice that both the price model G^P and the full model G^{FL} rely on the same set of assumptions regarding demand, cost, and firms' conduct. One difference is that only G^{FL} allows the changes in product characteristics post-merger. In this sense, the comparison can be a test of the endogeneity assumption.

To make such a comparison feasible, I exclude several markets from the comparison sample. Specifically, I drop a market if the merged carrier does not serve it any longer, or if the number of carriers, LCCs,

Table 12 Market Competitiveness of QI and QD markets Pre-merger

	Quality-increase markets	Quality-decrease markets
Number of rival firms (within a market)	4.56	3.93
Number of LCCs (within a market)	1.73	1.45
Number of rival routes (within a market)	12.06	8.60
Delta/Northwest only		
Passenger share	0.29	0.37
Percentage of flights originating from hub	0.08	0.09

Notes: Each number indicates average values. QI and QD markets indicate quality-increase and qualitydecrease markets, respectively. If a passenger-weighted average quality of merged firm's products increases after the merger, it belongs to QI markets, otherwise it belongs to QD markets.

and routes within a market substantially change after the merger.[57] This is for controlling the exogenous changes such as entry and exit occurrence in routes or markets that the model does not take account of. The final comparison sample consists of 244 markets. The bottom panel of table 13 shows that the characteristics of the selected markets do not change much over the integration period.

Importantly, I restrict this analysis to comparing flight frequency. I consider that the set of product characteristics is the first to be derived in the sequential choice model and the frequency changes the most among the endogenous characteristics. Hence, if the simulated frequency is substantially different from the actual post-merger frequency, further comparison analyses on prices and welfare effects would not be a very meaningful tasks.[58] I compare the frequencies by market level rather than by product level because the set of merged firm's products in a

57 Among 1,129 markets in the simulation sample, I exclude 154 markets the merged carrier exited as of first quarter of 2010. Further, I drop 731 markets where a change in the number of carriers is greater than three, or a change in the number of LCCs is greater than two, or a change in the number of routes is greater than five.

58 Other product characteristics are not appropriate for the comparison analysis. The data on mishandled baggage rate and denied boarding rate are available only at carrier level (see table 1), whereas the simulation outcomes are carrier-route specific. Also, on-time performance rarely changes according to the simulation.

Table 13 Comparison of Average Market Frequency (AMF): Pre-merger vs. Post-merger (Simulated) vs. Post-merger (Actual)

	Pre-merger	Post-merger **(Simulated)**	Post-merger **(Actual)**	Number of markets
Average market frequency				
All markets	14.10	13.77	13.64	244
Monopoly	8.17	6.15	4.43	3
Oligopoly	14.17	13.86	13.75	241
QI markets	15.27	16.02	17.79	117
QD markets	13.02	11.69	9.82	127
Measures of market similarity				
Number of carriers	6.00	5.00	5.03	244
Number of LCCs	1.35	1.35	1.38	244
Number of routes	10.81	10.81	9.59	244

Notes: Market frequency (MF) is defined as sum of frequency of each product provided by merging/merged carriers in a market: $MF_t = \sum_{j=1}^{J_{f_t}} Frequency_{jt}$, where j is a product and t is a market. Then, average market frequency (AMF) is mean value of market frequency across markets: $AMF = \frac{1}{T}\sum_{t=1}^{T} MF_t$.

market has changed post-merger.[59]

The top panel of table 13 presents the result. The first column reports average market frequency (henceforth, AMF) pre-merger which G^P relies on.[60] The second and the third column report the simulated AMF from G^{FL} and actual AMF post-merger, respectively. The table shows two clear trends. First, the simulated and actual AMF decrease from pre-merger AMF by 0.33 and 0.46, respectively. The reductions correspond to 7,247 (=0.33 × 90 × 244) and 10,102 (=0.46 × 90 × 244) less flights

59 For example, the merging carriers served the Chicago to New Orleans market with two connecting flights: (in order of origin-outward connecting-destination-inward connecting airport) ORD-ATL-MSY-ATL and ORD-MEM-MSY-MEM pre-merger, but MDW-ATL-MSY-ATL and ORD-MEM-MSY-ATL post-merger. Even though the number of routes are same, airports in the itineraries are slightly different. This prevents the comparison by product level.

60 Market frequency (MF) is defined as sum of frequency of each product provided by merging/merged carriers in a $MF_t = \sum_{j=1}^{J_{f_t}} Frequency_{jt}$, where j is a product and t is a market. Then, average market frequency (AMF) is mean value of market frequency across markets: $AMF = \frac{1}{T}\sum_{t=1}^{T} MF_t$.

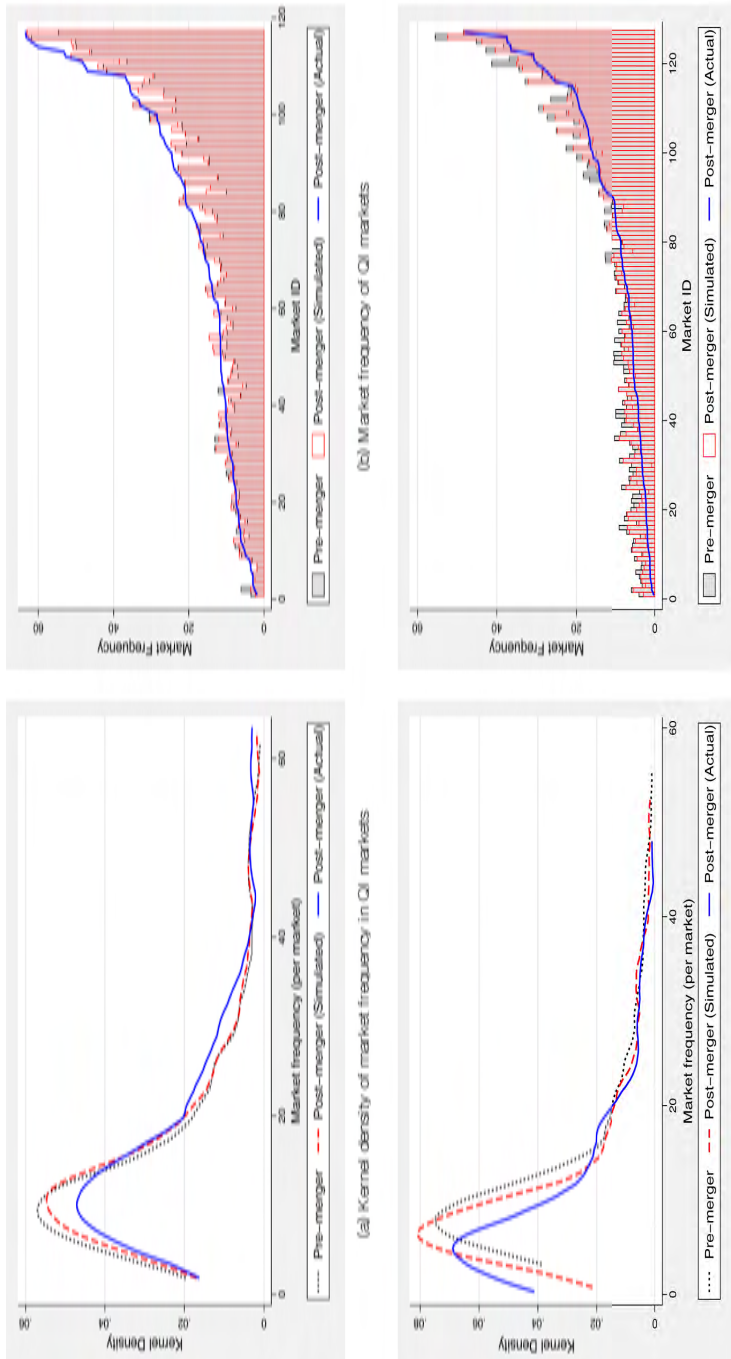
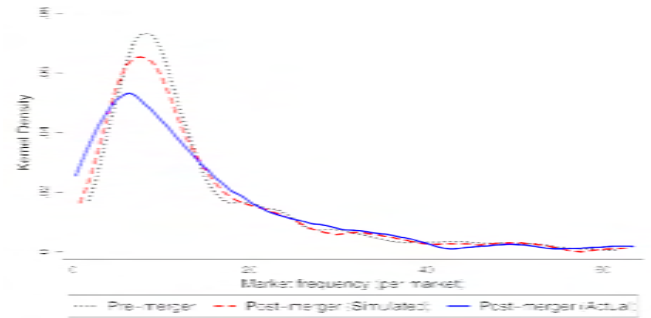
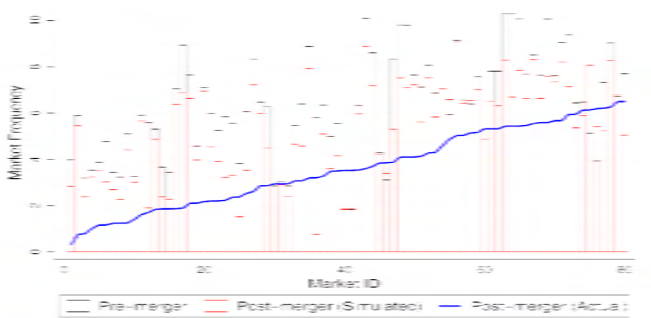


Figure 8 Distribution of Market Frequency in QI and QD Markets: Pre-merger vs. Post-merger (Simulated) vs. Post-merger (Actual)

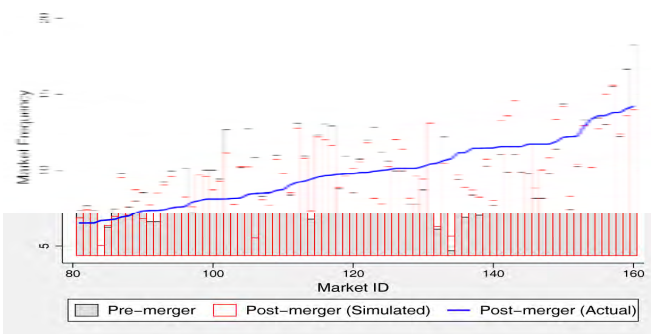
Figure 9 Distribution of Market Frequency in All Markets: Pre-merger vs. Post-merger (Simulated) vs. Post-merger (Actual)



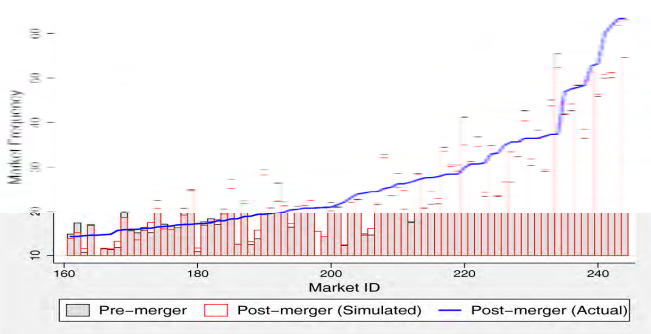
(a) Kernel density of market frequency in all markets



(b) Market frequency of markets (ID: 1-80)



(c) Market frequency of markets (ID: 81-160)



(d) Market frequency of markets (ID: 161-244)

during a quarter in the selected markets. This pattern is observable in both monopoly and oligopoly markets. Second, the simulated and actual AMF move in the same direction. They increase in QI markets and decrease in QD markets, while actual AMF changes more. The results implies that even though G^{FL} under- or overestimates actual AMF, it better predicts post-merger outcomes than G^{P}.

To be more specific, I illustrate market frequency (henceforth, MF) of each market in figure 8. Figure 8 (a) shows that the probability density function of actual MF (solid line) shifts to the right from that of pre-merger MF (dotted line) in QI markets. Figure 8 (b) shows that actual MF generally lies above pre-merger MF. Both figures indicate the increase of actual MF in QI markets. A notable thing is that the simulated MF is located closer to actual one than pre-merger MF is. The density function of the simulated MF shifts to the right, and its bar plot in each market fits actual MF line better. Figure 8 (c) and (d) describe the result in QD markets. They show quite the opposite situation where the density function of the simulated MF shifts to the left following actual MF and its bar plot also fits in well with actual MF line mostly lying below pre-merger MF.

Figure 9 describes an overall pattern by getting QI and QD markets together. In figure 9 (a), the density function of actual MF significantly deviates from that of pre-merger MF, and the simulated MF is located between them in general. Finally, the bar plots in figure 9 (b) to (d) confirm again that the simulated MF better follows actual MF line than pre-merger MF does. In short, the comparison analysis suggests that endogenizing product characteristics is essential to better predict the actual post-merger outcome.

6. Conclusion

When a merger simulation ignores changes in product characteristics post-merger, it can lead to a significant bias in predicted prices and welfare effects. This paper overcomes the limitations by endogenizing both price and product characteristics in a two-stage oligopoly game. Using data from the U.S. Department of Transportation, I estimate the

model parameters and then simulate the effect of the Delta and Northwest Airlines merger on product characteristics, fares, and welfare. To evaluate the predictive performance of the simulation, I compare the simulated outcome with actual post-merger data.

The main findings are as follows. First, the merged firm tends to increase product differentiation post-merger. The firm increases the quality of large and medium goods, but decrease that of small goods. The magnitude of the changes are stronger in monopoly markets. Since the large and medium goods are more profitable, the merged firm takes better care of their qualities to attract more passengers to them.

Second, the higher product differentiation affects the merged firm's incentive to raise prices. For large and medium goods, the full model predicts smaller cross-price effects than the price model. But for small goods, the cross-price effect is greater in the full model. The decreased quality and the increased price of small goods contributes to moving the consumers to large or medium goods.

Third, endogenizing product characteristics leads to quite different welfare effects. While the price model predicts decrease in consumer welfare for both types of passengers, the full model predicts that the business passengers benefit from the merger ($2.22 million) and the tourists experience smaller losses (-$0.36 million). This leads to an overall increase in consumer welfare. The finding highlights that a merger can increase consumer welfare if the merged firm brings the repositioned products that consumers can value more. About producer surplus, both models predicts higher profit for the merged firm and less profits for the competitors, but the additional profit gain for the merged firm is much bigger in the full model ($0.68 million) than in the price model ($0.17 million).

Finally, endogenizing product characteristics contributes to better predicting actual post-merger outcome. In QI markets, the simulated and actual post-merger frequency increase from pre-merger frequency. In QD markets, on the other hand, they decrease from pre-merger one. For both cases, the probability density function of the simulated is located between those of pre-merger and actual post-merger frequency.

This paper concludes with two notes for future research. First, the comparison analysis suggests that simulated result still under-or overes-

timates the actual post-merger outcome. The insufficient performance possibly comes from ignoring a change in unobserved product quality post-merger. If the link between unobserved and observed characteristics can be modeled through either structural or reduced form method, it can further improve the predictive performance of merger simulation. Second, this paper addresses the consumer heterogeneity with two types of passengers and estimates the percentage of each type in the population (γ_r). An interesting point is that the distribution of the consumers can vary from market to market. For example, the Houston to Las Vegas market may have a higher proportion of the tourists to the business passengers than the Las Vegas to Houston market. Responding to these distributions, a carrier can choose different product offerings for each market. The relationship between the distribution of consumer types and firm' decisions on product offerings can be an interesting topic for future research.

References

Berry, Steven. 1994. “Estimating discrete-choice models of product differentiation.” *RAND Journal of Economics*, 25(2): 242-262.

Berry, Steven, and Ariel Pakes. 1993. “Some applications and limitations of recent advances in empirical industrial organization: Merger analysis.” *American Economic Review,* 83(2): 247-252.

Berry, Steven, and Joel Waldfogel. 2001. “Do mergers increase product variety?: Evidence from radio broadcasting.” *Quarterly Journal of Economics*, 116(3): 1009-1025.

Berry, Steven, and Panle Jia. 2010. “Tracing the woes: An empirical analysis of the airline industry.” *American Economic Journal: Microeconomics*, 2: 1-43.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. “Automobile prices in market equilibrium.” *Econometrica*, 60(4): 889-917.

Berry, Steven, Michael Carnall, and Pablo T. Spiller. 2006. “Airline hubbing, costs and demand .” *In D. Lee, ed., Advances in Airline Economics*, 1(1).

Borenstein, Severin. 1989. “Hubs and high fares: Dominance and market power in the U.S. airline industry .” *The RAND Journal of Economics*, 20(3): 344-365.

Borenstein, Severin. 1990. “Airline mergers, airport dominance, and market power .” *American Economic Review*, 80(2): 400-404.

Borenstein, Severin, and Nancy L. Rose. 1994. “Competition and price dispersion in the U.S. airline industry .” *Journal of Political Economy*, 102(4): 653-683.

Brueckner, Jan K., and Yimin Zhang. 2001. “A model of scheduling in airline networks: How a hub-and-spoke system affects flight frequency, fare and welfare.” *Journal of Transport Economics and Policy*, 2(35): 195-222.

Byrne, David P. 2012. “The impact of consolidation on cable TV prices and product quality.” SSRN Working Paper No.1802347.

Cho, Sungick. 2012. “Three essays on firms’ quality choice.” Ph.D. dissertation, Texas A&M University.

Chu, Chenghuan Sean. 2010. “The effect of satellite entry on cable television prices and product quality.” *RAND Journal of Economics*, 41(4): 730-764.

Ciliberto, Federico, and Jonathan W. Williams. 2010. “Limited access to airport facilities and market power in the airline industry .” *Journal of Law and Economics*, 53(3): 467-495.

Crawford, Gregory S. 2012. “Endogenous product choice: A progress report.” *International Journal of Industrial Organization*, 30(3): 315-320.

Crawford, Gregory S., and Matthew Shum. 2006. “The welfare effects of endogenous quality choice: The case of cable television.” mimeo, University of Arizona.

DOT. 1999. “Airport business practices and their impact on airline competition: FAA/OST task force study.” U.S. Department of Transportation.

Draganska, Michaela, Michael Mazzeo, and Katja Seim. 2009. “Beyond plain vanilla: Modeling joint product assortment and pricing decisions.” *Quantitative Marketing and Economics*, 7: 105-146.

Fan, Ying. 2013. “Ownership consolidation and product characteristics: A study of the U.S. daily newspaper market.” *American Economic Review*, 103(5): 1598-1628.

Gandhi, Amit, Luke Froeb, Steven Tschantz, and Gregory J. Werden. 2008. “Post-merger product repositioning.” *Journal of Industrial Economics*, 56(1): 49-67.

Gerardi, Kristopher S., and Adam H. Shapiro. 2009. “Does competition reduce price dispersion? New evidence from the airline industry.” *Journal of Political Economy*, 117(1): 1-37.

Gotz, Georg, and Klaus Gugler. 2006. “Market concentration and product variety under spatial competition: Evidence from retail gasoline.” *Journal of Industry, Competition and Trade*, 6(3-4): 225-234.

Jayaraman, Sugandhi, and John F. O’Connell. 2011. “An investigative study into the annual US$2.5 bn mishandled baggage problem.” *Journal of Airport Management*, 5(4): 325-334.

Kim, E. Han, and Vijay Singal. 1993. “Mergers and market power: evidence from the airline industry.” *American Economic Review*, 83(3): 549-569.

Kwoka, John, and Evgenia Shumilkina. 2010. “The price effect of eliminating potential competition: Evidence from an airline merger.” *Journal of Industrial Economics*, 58(4): 767-793.

Lee, Jinkook. 2013. “Airport-airline vertical contract and market power in the U.S. airline industry.” Working Paper.

Luo, Dan. 2011. “The price effects of the Delta-Northwest Airline merger.” *Review of Industrial Organization*, 1-22.

Mazzeo, Michael. 2002. “Product choice and oligopoly market structure.” *Rand Journal of Economics*, 33(2): 221-242.

Mazzeo, Michael. 2003. “Competition and service quality in the U.S. airline industry.” *Review of Industrial Organization*, 22: 275-296.

Mazzeo, Michael, Katja Seim, and Mauricio Varela. 2013. “The welfare consequences of mergers with product repositioning.” Working Paper.

McFadden, Daniel. 1981. “Econometric models of probabilistic choice, in Charles F. Manski and Daniel McFadden, eds., Structural analysis of discrete data with econometric applications.” Mas- sachusetts: MIT Press.

Morrison, Steven A. 1996. “Airline mergers: A longer view.” *Journal of Transport Economics and Policy*, 30(3): 237-250.

Peters, Craig. 2006. “Evaluating the performance of merger simulation: evidence from the U.S. airline industry.” *Journal of Law and Economics*, 49(2): 627-649.

Puller, Steven L., Anirban Sengupta, and Steven N. Wiggins. 2012. “Does scarcity drive intra-route price dispersion in airlines?” Working paper: Texas A&M University.

Richard, Oliver. 2003. “Flight frequency and mergers in airline markets.” *International Journal of Industrial Organization*, 21(6): 907-922.

Rupp, Nicholas, Doug Owens, and L. Wayne Plumly. 2006. “Does competition influence airline on-time performance?” In D. Lee, ed., Advances in airline economics, vol.1: Competition policy and antitrust.

Small, K.A., and H.S. Rosen. 1981. “Applied welfare Economics with discrete choice models.” *Econometrica*, 49: 105-130.

Suzuki, Yoshinori. 2000. “The relationship between on-time performance and airline market share: a new approach.” *Transportation Research Part E*, 36: 139-154.

Sweeting, Andrew. 2010. “The effects of mergers on product positioning: Evidence from the music radio industry.” *RAND Journal of Economics*, 41(2): 372-397.

Villas-Boas, Sofia Berto. 2007. “Vertical relationships between manufacturers and retailers: Inference with limited data.” *Review of Economic Studies*, 74: 625-652.

Werden, Gregory, Andrew Joskow, and Richard Johnson. 1991. “The effects of mergers on price and output: Two case studies from the airline industry.” *Managerial and Decision Economics*, 12(5): 341-352.

Werden, Gregory J., and Kuke M. Froeb. 1994. “The effects of mergers in differentiated products industries: Logit demand and merger policy.” *Journal of Law, Economics, and Organization*, 10(2): 407-426.

Wyld, David C., Michael A. Jones, and Jeffrey W. Totten. 2005. “Where is my suitcase? RFID and airline customer service.” *Marketing Intelligence and Planning*, 23(4): 382-394.

CHAPTER 7

Knowledge, Entrepreneurship and Creation of New Competence: Foundations of the Creative Economy*

by
Hong Y. Park
(Saginaw Valley State University)

Abstract

Entrepreneurship studies the birth of a new product or service. How a new product or service is conceived and who is involved in the process is the subject of this study. Discovery and exploitation of entrepreneurial opportunity are commonly referred as to Kirznerian entrepreneurship (Kirzner, 1973). Perhaps the best known concept of entrepreneurship in economics is Schumpeter's entrepreneurship (Schumpeter, 1934). Schumpeter regards the entrepreneur as the innovator who creates new products, processes, markets, sources of supply, or industrial combinations. This study examines the roles of knowledge in Kirznerian discovery of entrepreneurial opportunity and exploitation, and in Schumpeterian innovation. The entrepreneur is a risk taker for profit (Knight, 1921), and the paper will study how knowledge is used to discover entrepreneurial opportunities and reduce the Knightian risk and uncertainty. The paper builds a philosophical foundation for the creative economy.

* This paper is prepared for the 2014 KDI-KAEA Joint Conference. This research was supported by the Braun research fellowship at Saginaw Valley State University. I am grateful for the support. I would also like to express my deep appreciation to Mr. Mark Whiteman for his sharing of the knowledge and competence creation practices of Dow Global Technologies, Inc.

1. Introduction

Recent emphasis on the firm and endogenous economic growth has catapulted knowledge and entrepreneurship to the forefront of economic policy and corporate strategy debates. Policy makers have realized that entrepreneurship helps create jobs and foster economic growth, primary factors in their political successes. Economists and management scholars began to pay attention to the origination and innovation of the firm because the global economy is going through a period of profound change and transformation. Entrepreneurs are leading the way, and newly emerging corporations such as Facebook, Twitter, Tesla, Google and Apple, frequently in the news as success stories, are major job creators. Their strategies also contribute to the firms' successes and economic growth. Entrepreneurs originate new products, processes, and services in response to changes in the economic environment which create disturbance in the equilibrium (Kirznerian entrepreneurship, 1973). Kirzner's (1973) theory of entrepreneurship emphasizes the equilibrating role of entrepreneurship, conceived as alertness to profit opportunities, according to Holcombe (2003). Holcombe illustrates how entrepreneurship discovers the stock of entrepreneurial opportunities and points to entrepreneurship as the engine of economic growth.

The Schumpeterian entrepreneur (Schumpeter, 1934, 1947), on the other hand, creates disequilibrium by innovating products, processes, services and markets. Foss and Klein (2010) argue that Schumpeterian entrepreneurship is exercised within the firm when it introduces new products, processes, or strategies, but routine operations of the firm need not involve entrepreneurship at all. Therefore, Schumpeterian entrepreneurship can be studied from the competence perspective of the firm, which consists of three subgroups in the theories of the firm (Foss and Mahnke, 2000): the resource-based, knowledge-based and evolutionary theories of the firm. This competence perspective of the firm has become a leading topic in recent industrial organization and strategy studies. Creating new capabilities (innovation) has become the most important source of competitive advantage. The outcomes of new

capabilities are sources of above normal rewards for entrepreneurship; these provide incentives for entrepreneurship, though the creation of new capabilities is accompanied by uncertainty and risk.

However, entrepreneurs make judgments on the outcomes of their innovation (Cantillon, 1775; Mises, 1949; Schumpeter; 1934; Knight, 1921). Their decisions to innovate occur in the present, but they bear fruits in the future. Not all innovations necessarily lead to successful outcomes. Since innovation involves uncertainty and risk, reducing uncertainty has drawn significant attention (Arrow, 1971). Knowledge plays a crucial role in entrepreneurship and risk management.

Knowledge, entrepreneurship and competence building are crucial sources for competitive advantages of the firm (Foss, 1999; Schiuma, 2009). Langstrom et al. (2012) ague that "to successfully develop entrepreneurship research in the future, we need to relate new research opportunities to earlier knowledge within the field, which calls for a stronger 'knowledge-based' focus" (p. 1154). Casson(2014) urges researchers to involve greater use of formal models and give greater attention to cognition and information processing. However, there is a paucity of studies on knowledge and entrepreneurship (Foss and Mahnke, 2000). Therefore, in an attempt to fill the gap this paper investigates the roles of knowledge in entrepreneurship and risk management. The study also attempts to bridge the gap between abstract theory and descriptive empiricism by applying knowledge and entrepreneurship theories to a case study.

The paper is organized as follows: Section two reviews the nature of knowledge; Section three reviews the entrepreneurship theories of Kirzner, Schumpter and Knight; Section four analyzes the links between knowledge and entrepreneurship, and how knowledge can reduce risk and uncertainty in entrepreneurship; Section five evaluates knowledge, entrepreneurship and risk and uncertainty in the context of corporate entrepreneurship based on the case of the Dow Chemical Company; Section six discusses knowledge and entrepreneurship in the case of Dow; Section seven draws policy implications; finally, Section eight concludes the paper.

2. The Nature of Knowledge

We need to have deeper understanding of the nature of knowledge and theories of entrepreneurship before integrating them. Therefore, we begin to discuss the nature of knowledge here, the next section reviews theories of entrepreneurship, and the following section integrates knowledge and entrepreneurship. Knowledge plays a key role in the discovery of entrepreneurial opportunity and innovation in new products, processes and services. Scholars in Western epistemology follow Plato's concept of knowledge as "justified true belief" (Nonaka, 1994; Kimball, 1999). Knowledge scholars have also developed several dimensions of knowledge in the history of epistemology and in economics. In our discussion, we employ four dimensions of knowledge (Ryle, 1946, 1949; Polanyi, 1962, 1966; Hayek, 1945 Lam, 2000): Ryle's "know-how" and 'know that', Polanyi's "tacit and explicit knowledge", Hayek's information as knowledge, and Lam's epistemological and ontological dimensions.

2.1. Know How and Know That

First, Ryle's distinction between knowing how and knowing that (Ryle, 1946, 1949) links knowledge and entrepreneurship. Ryle maintains that know-how and know that are different kinds of knowledge. Ryle's distinction is accepted in philosophy, although there are two other propositions: knowing that is a species of know-how (Hartland-Swan, 1956, 1957), and knowing how is a species of know that (Fantle, 2008; Stanley and Wilson; 2001). For Ryle, knowing how is practical knowledge, an ability, competency or skill, whereas know that is a relationship between a thinker and a true position (Ryle, 1946, 1949). Despite the debates on know-how and know that among scholars in philosophy, Ryle's perspective on the ability and dispositional account of know-how offers an interesting insight into entrepreneurship. Ryle's view is that know-how is ability. Knowing how to do something entails having the ability to do it and being able to do it. For example, the individual knows how to ride a bicycle and is also able to ride a bicycle. According to Roland (1958), knowing how to ride a bicycle is a capacity which implies having learned how to ride through practice. Fantl (2008),

however, argues that there are various grades of know-how, and one's know-how to ride a bicycle requires the addition of relevant pieces of know that.

Ryle's dispositional account of know-how as "know" in the case of moral judgments and rules of conduct acts as a tendency word, rather than a capacity word (Roland, 1958). For Ryle, if exercises of know-how to ride a bicycle can involve disparate activities, then we can construe know-how to ride a bicycle as a disposition to engage in these activities in proper ways, as he claims in his book, *The Concept of Mind:* Knowing how, then, is a disposition, but not a single-track disposition like a reflex or a habit. Its exercises are observances of rules or canons or the application of criteria, but they are not tandem operations of theoretically avowing maxims and then putting them into practice. Further, its exercises can be overt or covert, deeds performed or deeds imagined, words spoken aloud or words heard in one's head, pictures painted on canvas or pictures in the mind's eye. Or they can be amalgamations of the two. (Ryle, 1949, p. 46)

Ryle's know-how to do something means to be disposed to behave in certain ways (tendency), but it is very hard to specify in advance what those ways are. However, know-how may be disposed to behave in accord with various rules (i.e., know that). Fantl (2008) points out an interpretation of Ryle that makes know-how a matter of ability; it will have to be a certain kind of ability – an ability to act out of the relevant rules (or canons or criteria) governing the intelligent execution of how to do the activity.

This requires understanding of know that, a second dimension, and knowing how demonstrates observance of these rules or canons or application criteria.

2.2. Tacit and Explicit Knowledge

Second dimensions, Polanyi's tacit and explicit knowledge, are crucially important in knowledge and competence creation. Polanyi (1962) recognizes the indispensable role belief plays in all knowing:

We must recognize belief once more as the source of all knowledge. Tacit assent and intellectual passions, the sharing of an idiom and of

cultural heritage, affiliation to a like-minded community: such are the impulses which shape our vision of the nature of things on which we rely for our mastery of things. No intelligence, however critical or original, can operate outside such a fiduciary framework. (Polanyi, 1962, 266)

Polanyi's statement includes many important aspects of current knowledge creation debates, such as belief, sharing and vision. How does a knower form belief? Polanyi (1966) argues that "our body is the ultimate instrument of all our external knowledge" (p. 15). We may probe things outside ourselves by a sentient extension of our body. Our awareness of our body for attending to external things comprises focal and subsidiary awareness. Focal awareness, the object of our attention here, depends on subsidiary awareness. Polanyi (1969) pointed out that "focal and subsidiary awareness are not two degrees of attention but two kinds of attention given to the same particulars" (*Knowledge and Being,* 128). He uses recognition of a countenance as an example. Mitchell (2006) summarizes Polanyi's example: the particular features of the physiognomy are subsidiarily known, and the integration of the particulars such as the nose and eyes produces the recognizable face. According to Mitchell (2006), "Polanyi's most significant insight concerns the basic operation of mind: all knowing consists of the integration of subsidiary and tacitly sensed particulars into a focal and articulated whole" (p. 70). Polanyi identifies the object of our attention, the subsidiaries of our attention, and the knower as the triad (three components) of this tacit knowing. The knower integrates the subsidiary and the focal awareness into the active process that constitutes tacit knowing. However, the triad is not permanent, and the knower can dissolve the triad by looking differently at the subsidiaries.

Polanyi (1969) pointed out that tacit knowing is indispensable in the discovery of new knowledge, and all knowledge "is either tacit or rooted in tacit knowledge" (*Knowledge and Being*, 1969, p. 195). If all knowing is either tacit or rooted in the tacit, the knower is indispensable in knowing. The knower integrates subsidiary elements within a focal whole; therefore, explicit knowledge is rooted in tacit knowledge, and tacit knowledge is the origin of knowledge. The knower acquires tacit knowledge by experience (indwelling) and engages in discovery of new

knowledge by uncovering hidden reality. Polanyi (1966) argues that indwelling is the proper means of knowing (p. 16). An individual indwelling in her job is gaining experiences, acquiring intuition for a discovery of new knowledge through her experiences. For Polanyi, intuition is making a guess and guessing correctly. He uses the notion of intimation of hidden reality. Individuals who are indwelling in their jobs develop a good intuition and easily detect the intimation from the hidden reality. As the knower (employee) accumulates tacit and explicit knowledge, the knower improves the discovery of knowledge. In the same vein, Cook and Brown's (1999) generative dance between knowledge and knowing explain how the interplay between knowledge and knowing can generate new knowledge and new ways of knowing. Polanyi (1966) indicates that there are many particulars of hidden reality; they are inexhaustible. Therefore the discovery of knowledge from the hidden reality can be continuous. The discoveries of Nonaka and Takeuchi's (1995) kneading for bread making and Orr's (1996) knowledge sharing among service workers for Xerox copy machines provide paradigmatic examples of this discovery of knowledge. Every organization needs a process that can wring out tacit knowledge and encourage employees to engage in the discovery of knowledge (Park et al., 2014a).

2.3. Knowledge and Information

A third dimension is that, in economics, knowledge is information (Hayek, 1945; Arrow, 1962). According to Hayek, knowledge is dispersed and is not in a useable form, the so called Hayekian knowledge problem in the literature (Hayek, 1945; Holcombe, 2003; Sautet, 2000). Knowledge needs to be discovered through managerial and entrepreneurial actions. Spontaneous decisions should be made by those who have superior information. Consequently, limited knowledge (Arrow, 1962) and asymmetry of information (Stigler, 1961; Akerlof, 1970; Alchian and Demsetz, 1972; Spence, 1973; Riley, 1975; Stiglitz, 1975) have become huge issues in financial markets, human resource management and entrepreneurship. In the context of entrepreneurship, Arrow (1962) pointed out that there is no market for future goods.

Therefore, future prices are not known and cannot reveal information about the market. In the absence of markets for future goods, expectations for the future are related to both quantities as well as prices. As Arrow stated, the future state of quantity and price of the good or service that the entrepreneur is producing is based on his/her anticipation of future quantity and price. He also argues that future-oriented decisions are made under conditions of dispersed and costly information (Arrow, 1974). Arrow (1962) brings expectations and uncertainties to the forefront of economic discussions; foresight, expectation and uncertainty are key issues because entrepreneurs hire inputs now to produce goods and services for future markets. Knight (1921) addresses this issue in his book, *Risk, Uncertainty and Profit* (1921). Therefore, it is essential for entrepreneurs to develop foresight on quantities and prices of their new products and services.

2.4. Ontology of Knowledge

Finally, Lam's ontological dimension of knowledge (Lam, 2000) explains where knowledge within the firm can reside: individually or collectively. Her epistemological dimension of knowledge follows Polanyi's classification of knowledge: tacit and explicit. Table 1 represents Lam's dimensions of knowledge. Understanding the ontological dimension of knowledge is crucial in discussion of enactment of knowledge for entrepreneurship. Various enactment structures have emerged to wring out knowledge from individuals and groups (Weick, 1979; Park et al.,

Table 1 Lam's Dimensions of Knowledge

		Ontological dimension	
		Individual	Collective
Epistemological dimension	explicit	Embrained knowledge	Encoded knowledge
	tacit	Embodied knowledge	Embedded knowledge

2014a). Nonaka's (1994) SECI model (socialization, externalization, combination and internalization), for example, is a pioneering model for knowledge creation.

3. Entrepreneurship Theories of Kirzner, Schumpeter and Knight

The formal description of entrepreneurship goes as far back as Cantillon, 1775. According to Schumpeter (1934), Cantillon introduced the term "entrepreneur" and defined this as the agent who buys means of production at certain prices in order to combine them into a product that will sell at prices that are uncertain at the moment at which he commits himself to his costs. Today's Austrian entrepreneurship scholars (Mises, 1966; Schumpeter, 1934; Kirzner, 1973) follow closely Cantillon's explanation of entrepreneurship. Entrepreneurs both originate new firms and lead innovation in existing firms. Langlois (2005) rightly points out that the 'firm exists because of entrepreneurship'. Langlois (2005) argues that the cluster of ideas central to the entrepreneurship literature flows out of the tradition of Knight (1921), Schumpeter (1934), and Kirzner (1973). We follow this tradition in our paper. Before reviewing the entrepreneurial ideas of Cantillon, Knight, Schumpeter and Kirzner, however, we need to define the field of entrepreneurship to link current entrepreneurship to them. Shane and Venkataraman (2000) define the field of entrepreneurship thus:

The scholarly examination of how, by whom, and with what effects opportunities to create future goods and services are discovered, evaluated, and exploited (Venkataraman, 1997).

Consequently, the field involves the study of sources of opportunities; the process of discovery, evaluation, and exploitation of opportunities; and the set of individuals who discover, evaluate, and exploit them (Shane & Venkataraman, 2000, 218).

Shane and Venkataraman(2000) invoke three aspects of entrepreneurship, and Langlois (2005) points out that Knight, Schumpeter and Kirzner represent these three aspects of entrepreneurship which Shane and Venkataraman point to. According to Langlois (2005), Kirzner is about

discovery, the alertness to new opportunities; Knight is about evaluation, the faculty of judgment in economic organization; and Schumpeter is about exploitation, the carrying out of new combinations and the creative destruction that often results therefrom. We briefly discuss these three aspects of entrepreneurship.

3.1. Kirznerian Entrepreneurship

First, Kirzner(1973) continues the tradition of the Austrian entrepreneurship scholars. He has developed the concept of the discovery of entrepreneurial opportunities which has drawn attention from many scholars (Alvarez and Barney, 2007; Sautet, 2000; Shane, 2000; Yu, 2001; Murphy and Marvel, 2007). Alvarez and Barney (2007) propose two internally consistent theories of how entrepreneurial opportunities are formed: An opportunity discovery theory and an opportunity creation theory. Kirzner's (1973, 1997, 2000) theory of alert entrepreneurial discovery explains the market phenomena from the perspective of entrepreneurial decisions. He developed his entrepreneurial discovery based on Mises (1949) and Hayek (1937, 1945), who represent the Austrian school. According to Mises (1949), the driving force of the market process is provided by the promoting and speculating entrepreneurs, and profit-seeking speculation is the driving force of the market, as it is the driving force of production (p. 328-329). For Mises, entrepreneurs buy where and when they deem prices are too low, and they sell where and when they deem prices too high. Profit is the difference between the buying and selling prices. Entrepreneurial profit then is the driving force of entrepreneurial actions which equilibrate prices. Mises (1949) explains the role of entrepreneurs as follows:

The entrepreneurs take into account anticipated future prices, not the final prices or equilibrium prices. They discover discrepancies between the height of the prices of the complementary factors of production and the anticipated future prices of the products, and they are intent upon taking advantage of such discrepancies. These endeavors of the entrepreneurs would finally result in the emergence of the evenly rotating economy if no further changes in the data were to appear (Mises, 1949, p. 329).

Actions of profit seeking entrepreneurs bring about a tendency toward an equalization of prices for the same goods in all subdivisions of the market; thus entrepreneurs make the market system work efficiently and contribute to economic growth. Boettke and Sautet (2011) point out that for Mises "the market is not only a space where people may haggle over prices; it is also a process, by which knowledge is generated, information comes to be known, and prices are determined throughout society" (p. 32). In the market process, entrepreneurs eliminate price differentials and improve the overall efficiency of the economic system and economic growth. This function of the entrepreneur is similar to the role of Cantillon's entrepreneur, as stated before. Cantillon and Mises were keenly aware of the speculative nature of entrepreneurial actions. Therefore, knowledge is important in dealing with the uncertainty in speculative entrepreneurial actions.

Kirzner (1973) expands an Austrian (i.e., a Misesian) perspective on markets, which are envisaged as the dynamically competitive **market process**. He argues that in that process, markets tend to continuously move towards equilibrium. Markets are continuously in disequilibriaum due to continual changes in the exogenous environment, such as technologies, consumer preferences and regulations. Kirzner also introduces earlier errors made in the course of market exchanges as a source of entrepreneurial opportunities. Alert agents discover entrepreneurial opportunities generated by external shocks and earlier errors. Yu (2001) points out that the basic concept in Kirzner's theory of entrepreneurship is alertness, which helps entrepreneurs discover opportunities hitherto unnoticed. Kirzner (1984) argues that disequilibrium prices offer pure profit opportunities that attract the notice of alert, profit-seeking entrepreneurs. Yu (2001) elaborates on the subjective perspective of Kirzner's concepts of entrepreneurial alertness and discovery. He argues that the entrepreneurial discovery process is associated with the actor's interpretation framework, or the stock of knowledge, which is derived from everyday life experiences. Entrepreneurial alertness leads to discovery of the stock of knowledge generated in the market process and the tacit knowledge gained from everyday life experiences. The driving force of the discovery and exploitation of entrepreneurial opportunities is profit. For Kirzner, disequilibrium generates entrepre-

neurial opportunities, and the entrepreneur engages in arbitrage that helps the market move towards equilibrium. Alert entrepreneurs discover opportunities stemming from external shocks, such as markets, technologies, regulations, and prior errors which are caused by entrepreneurial actions. Kirzner (1997) argues that opportunities created by earlier entrepreneurial errors have resulted in shortages, surpluses, and misallocated resources. Entrepreneurial errors occur due to overly optimistic/pessimistic entrepreneurial decisions. Ground for errors stemming from entrepreneurial decisions can be found in Popper's epistemological philosophy (Popper, 1982).

Popper (1982) points out that all living things face problems and solutions to those problems always come with errors. Popper's evolutionary epistemology (1982) offers a theoretical framework for error in entrepreneurial decisions; he argues that "all organisms are constantly, day and night, engaged in solving-problems" (p. 110). He states that problem-solving is always preceded by methods of trial and error; his fundamental evolutionary sequence of events is [P1→TS→EE→P2], where P1=initial problem, TS=tentative solutions, EE=error elimination, P2=new problems. For Kirzner correcting this error is an entrepreneurial action.

3.2. Schumpeterian Entrepreneurship

Secondly, the Schumpeterian entrepreneur (Schumpeter 1934, 1942, and 2008) generates surplus profits by breaking circular flow. Schumpeter defines the concept of circular flow in his book, *The Theory of Economic Development* (1934):

Under the assumption of constant conditions, consumers' and producers' goods of the same kind and quantity would be produced and consumed in every successive period because of the fact that in practice people act in accordance with well-tried experience, and that in theory we regard them as acting in accordance with a knowledge of the best combination of present means under the given conditions. But there is also another connection between how the successive period operates with goods which an earlier period prepared for it, and in every period goods are produced for use in the next (p. 41, 42).

Goods would be bought and sold at the same prices year after year in this circular flow, which is a description of the concept of Walrasian general equilibrium. Every business firm finds that its selling price exactly equals its cost of production; there is no surplus profit and no economic growth in the circular flow. Breaking this circular flow is the function of the entrepreneur.

For Schumpeter (1934), economic development is spontaneous, and discontinuous change in the channels of the flow, disturbance of equilibrium, forever alters and displaces the equilibrium state previously existing. The entrepreneur carries out new combinations and contributes to economic development. These entrepreneurial actions then are the main mechanism in the process of economic development:

This concept covers the following five cases: (1) The introduction of a new good – that is one with which consumers are not yet familiar – or of a new quality of a good. (2) The introduction of a new method of production, that is, one not yet tested by experience in the branch of manufacture concerned, which need by no means be founded upon a discovery scientifically new, and can also exist in a new way of handling a commodity commercially. (3) The opening of a new market, which is a market into which the particular branch of manufacture of the country in question has not previously entered before, whether or not this market has existed before. (4) The conquest of a new source of supply of raw materials or half-manufactured goods, again irrespective of whether this source already exists or whether it first has to be created. (5) The carrying out of new organization of any industry, like the creation of a monopoly position (for example through trustification) or breaking up a monopoly position (Schumpeter, 1934, p. 66).

Schumpeter's (1942) later work links these new combinations of entrepreneur with creative destruction. He argues that "the opening up new markets, foreign or domestic, and organizational development from craft shop to new organization illustrate the process of mutation that incessantly revolutionizes the economic structure from within, incessantly destroys the old one, incessantly creates a new one" (p. 83). He refers to this process as the process of creative destruction. This fact bears upon our problem in two ways: (1) In the process every element takes considerable time in revealing its true feature and ultimate effects,

so we must judge its performance over time; (2) Every piece of business strategy acquires its true significance only against the background of that process and within the situation created by it. It must be in its role in the perennial gale of creative destruction; it cannot be understood irrespective of it or, in fact, on the hypothesis that there is a perennial lull (pp. 83, 84).

Thus Schumpeter (1934) explains economic development from the creative destruction perspective:

Development in our sense is a distinct phenomenon, entirely foreign to what may be observed in circular flow or in the tendency towards equilibrium. It is spontaneous and discontinuous state previously existing. Our theory of development is nothing but a treatment of this phenomenon and the process incident to it (Schumpeter, 1934, p. 64).

For Schumpeter (1947), gradual or routine adaptive response to changes in data is not entrepreneurial, but creative response to changes in data is entrepreneurial. Creative response is something that is outside of existing practice. According to him, a study of creative response in business becomes coterminous with a study of entrepreneurship. Therefore, the Schumpeterian entrepreneur is a creative entrepreneur, compared to a Kirznerian alert entrepreneur. Carrying out new combinations is referred to as *enterprise.* Schumpeterian entrepreneurs are individuals whose function is to carry out the new combination. In contrast, managers merely operate an established business or direct routine daily tasks in circular flow and do not receive profit, though they receive wages.

Schumpeter (1934) characterizes three corresponding pairs of opposites in characterizing the entrepreneur: (1) opposition of two real processes: the circular flow or the tendency towards equilibrium on the one hand, vs. a change in the channels of economic routine or a spontaneous change in economic data arising within the system, on the other; (2) opposition of two theoretical apparatuses: statics and dynamics; (3) opposition of two types of conduct: mere managers vs. entrepreneurs. Entrepreneurs respond creatively to a spontaneous change in economic and environmental data and work under the dynamic theoretical framework. Schumpeter argues that entrepreneurs are motivated by the psychology of a non-hedonic character. First, there is the dream and the

will to found a private kingdom, usually, though not necessarily, also a dynasty. Second, there is the will to conquer: the impulse to fight, to prove oneself superior to others, to succeed for the sake of, not the fruits of success, but success itself. Third, there is the joy of creating, of getting things done, or simply of exercising one's energy and ingenuity (p.93).The Schumpeterian entrepreneur is bold, self-confident, creative and innovative.

For Schumpeter, entrepreneurs do not bear risk, because new combinations are financed by capitalists, although entrepreneurs may own capital in some cases. New combinations are financed by banks that bear the financial risk. He points out that interest is paid out of their profit. Today venture capital firms also finance new combinations.

3.3. Knightian Entrepreneurship

Thirdly, Knight (1921) provides extensive reviews on profit, the reward that the entrepreneur receives from taking uninsurable risk (uncertainty). The entrepreneur makes changes, and with changes come risk. He distinguishes insurable risk from the uncertainty that is not measurable or insurable. He argues that profit cannot be the reward of management, for this can be performed by hired labor. If the manager takes no risk, this individual is no longer an entrepreneur. For Knight (1921) there is a fundamental distinction between the reward for taking a known risk and the reward for assuming a risk whose value itself is not known (p. 44).

We have discussed Kirzner's alert entrepreneurship, Schumpeter's creative entrepreneurship and Kight's uncertainty. Scholars have compared these entrepreneurships and tried to reconcile them (Kirzneer, 2008; Alvarez and Barney, 2007, 2010) by presenting three assumptions based discovery and creation theories of entrepreneurial action (see Table 2).

Table 2 Alvarez & Barney's Central Assumptions of Discovery and Creation Theories of Entrepreneurial Action

	Discovery Theory	Creation Theory
Nature of Opportunities	Opportunities exist, independent of entrepreneurs. Applies a realist philosophy.	Opportunities do not exist independent of entrepreneurs. Applies an evolutionary realist philosophy.
Nature of Entrepreneurs	Differ in some important ways from non entrepreneurs, ex ante.	May or may not differ from non entrepreneurs, ex ante. Differences may emerge, ex post.
Nature of Decision Making Context	Risky	Uncertain

Table 2 summarizes three key issues in entrepreneurship: discovery of entrepreneurial opportunity, creation of entrepreneurial opportunity, and uncertainty. Venkataraman (2003) indicates that Shane (2003) builds a framework of the individual-opportunity nexus. According to Venkataraman, Shane takes two positions: 1) Entrepreneurial opportunities exist independent of the actors in a system; 2) A human being is required to provide this agency, so that when *a market can come to be, it will come to be (italics in the original)*. This agency is entrepreneur. Entrepreneurial decision-making is risky since entrepreneurs exploit existing opportunities within the critical realist philosophy (Bhaskar, 1975). However, in creative view, opportunities do not exist independent of entrepreneurs, and in an evolutionary realist or process philosophy, opportunities are social construction (Campbell, 1960; Whitehead, 1929).

3.4. Entrepreneurship and Foresight

As we discussed in Mises, Kirzner and Schumpeter, the entrepreneurs take actions for change. Changes are speculative because there is a time gap between taking actions and the outcomes of these actions. These changes result in temporal monopoly profit. The great source of monopoly profit is to be found in the fact that the actuarial risk of any given undertaking is not the same for different entrepreneurs, owing to

their differences in ability and environment. Profit results from risks wisely selected, and some entrepreneurs have better foresight and prescient knowledge of the future. Knight (1921) deals with the uncertainty issue which Arrow (1974) defines as follows:

Uncertainty means that we do not have a complete description of the world which we fully believe to be true. Instead, we consider the world to be in one or another of a range of states.

Each state of the world is a description that is complete for all relevant purposes. Our uncertainty consists in not knowing which state is the true one (Arrow, 1974, 33).

Entrepreneurs do not know which state will result when they make their entrepreneurial decisions ex ante. For Arrow, uncertainty of entrepreneurial actions is due to the absence of markets for future goods. Because entrepreneurs take actions under uncertainty, profits are their reward.

Since profit is reward to the entrepreneur, it is requisite to define the entrepreneur's function. Baumol (1968) defines this function as follows:

The entrepreneur (whether or not he in fact also doubles as a manager) has a different function. It is his job to locate a new idea and to put them into effect. He must lead, perhaps even inspire; he cannot allow things to get into a rut and for him, today's practice is never good for tomorrow. In short, he is the Schumpeterian innovator and some more.

He is the individual who exercises what in the business literature is called "leadership." And it is he who is virtually absent from the received theory of the firm (Baumol, 1968, p. 65).

Since knowledge has become an important factor in locating new ideas and putting them into effect, today's entrepreneurs seek knowledge from employees, consumers, suppliers, competitors and universities. Knowledge provides entrepreneurial ideas, and foresight helps entrepreneurs deal with uncertainty.

The exercise and development of entrepreneurial foresight is requisite for successful entrepreneurial outcomes. Whitehead (1933) discusses the conditions for these. He argues that the habit of foreseeing is elicited by the habit of understanding; understanding can be acquired by a conscious effort and can be taught. He also points out that the training of foresight is by medium of understanding, and foresight is the product of

insight (p. 89). Understanding commerce stems from personal experience of commerce and first-hand practice. His insight on business routine and the change of routine illustrates how changes in business routines are made and how those changes are directed toward progressiveness. Whitehead (1933) argues that routine is the god of every social system, the essential component in the success of every factory; novel ideas emerge from routine activities and these novel ideas end up founding novel methods and novel institutions. Changes in civilization are made toward progressiveness in human life. Therefore, when entrepreneurs understand Whitehead's process, they are able to capture novel ideas from exper-iences of routine business activities and draw foresight on emerging entrepreneurial ideas. Entrepreneurs feel the quality of changes in the air and quantify the changes for new products and services, processes and new markets. This is akin to the "awareness" of Kirznerian entrepreneurs.

4. Knowledge and Entrepreneurship and Uncertainty

4.1. Knowledge and Entrepreneurship

We have discussed the nature of knowledge and presented three perspectives on entrepreneurship in sections 2 and 3. Having better understanding of knowledge and entrepreneurship, we can build knowledge as a foundation for entrepreneurship and risk and uncertainty management.

According to Holcombe (2003), the assumption of perfect knowledge in the classical model of perfect competition rules out the possibility that any unrecognized profit opportunities could exist in the economy. Therefore, the classical model ignores entrepreneurship and surplus profit. The Austrian model deals with the market processes, which are dynamic; economic data or economic environments continue to change due to external economic shocks, and changes in environment generate entrepreneurial opportunities. Kirznerian entrepreneurs engage in arbitrage and help the market move toward equilibrium, whereas Schumpeterian entrepreneurs innovate products, services, production processes, organi-

zations, markets and inputs.

Reconciliation or integration of the discovery of entrepreneurial opportunities and creation of entrepreneurial opportunities has become a hot issue. Kirzner (1999, 2008) himself recognizes the critical contribution of Schumpeter to entrepreneurship and justifies his position on alertness: creation opportunity needs alertness. Alvarez and Barney (2007, 2010) argue that the theories are difficult to reconcile because they are grounded in two separate philosophical realisms. Discovery opportunities are studied with theories embedded in critical realism, while creation opportunities are studied with theories embedded in evolutionary realism. Therefore, the theories may be reconciled or integrated only in an epistemological perspective. Mathews (2006) also attempted to reconcile Schumpeter with Kirtzner.

Entrepreneurial ideas stem from individuals who possess knowledge. Ryle's know that and know-how can be utilized both to discover opportunities and to create opportunities. Know that is useful to discover existing entrepreneurial opportunities. If an actor or agent knows that there are price differentials, he/she can engage in arbitrage and make a profit. If the actor has knowledge that the future price will rise, he/she can buy now and sell in the future. Crises continue to recur in the world. The event of crisis may be a critical reality, but there are a set of opportunities in a crisis. The subprime mortgage crisis, for example, was an external shock which presented an entrepreneurial opportunity because the housing market was below the market equilibrium. Entrepreneurs' arbitrages made the current housing prices rise and improved economic efficiency and growth. As we increase experiences with bubbles, we are able to identify their probability distribution. Therefore, entrepreneurs take Knightian risk. Entrepreneurs in the U. S. housing bubble realized large profits by arbitraging price differentials in and out of the crisis. We observed the same pattern in the Asian financial crisis, which included Korea. There were also ample entrepreneurial opportunities in Greece, Spain and Ireland under the European economic crisis.

Casson (2014) urges researchers to give greater attention to cognition and information processing in entrepreneurship research. Entrepreneurs need to acquire knowledge and understanding through thought, exper-

ience and the senses on changes in data or economic environment. It is widely known that the IBM company had technology for data storage, but did not recognize the entrepreneurial opportunity for it and the Oracle company's CEO unsderstood the potential for entrepreneurship and the Oracle company was born based on the data storing technology. The Kodak company owned the patent for digital technology, but did not capitalized the technology and lost the ensuing huge entrepreneurial opportunities from the digital technology. Entrepreneurs can process information and discover entrepreneurial opportunities which help solve the Hayek's knowledge problem. Current environmental awareness of consumers may lead to changes in consumer preferences and offers entrepreneurial opportunities. Entrepreneurial opportunities in treatment of carbon, solar energy and alternatives to carbon-based energy will emerge. These are examples of Kirzner's awareness (1973) and Casson's cognition and information processing.

Know-how offers creation of entrepreneurial opportunities. Know how is about how to do things, how to create new combinations via innovation in products, services, processes and markets (Schumpeter 1934). Know-how is tacit knowledge which individuals gain by doing things, by performing routine tasks and using products. Knowledge is also dispersed and does not exist as useful form, as Hayek (1945) draws to our attention. Because the reality in this perspective is hidden and continues to emerge (Campbell, 1960, 1974; Whitehead, 1929), entrepreneurs need to enact this knowledge to create entrepreneurial opportunities (Weick, 1977, 1979).

Weick (1979) developed the enactment theory, seen as a process for people to achieve continuity and coordination by bringing out tacit knowledge from individuals. Knowledge in modern corporations is often created by a team; individuals in a team bring out their tacit knowledge as they interact with team members. Weick's enactment process is seen as a structure, and structuration (Giddens, 1979, 1984) has emerged as a leading field in social sciences. Giddens' structuration theory (1979, 1984) can be employed as a framework for explaining the relationship between structure and agent in knowledge creation, as it is concerned with understanding the activities of knowledgeable human actors and the structuring of social systems.

4.2. Knowledge and Uncertainty in Entrepreneurship

Does an entrepreneur have better foresight than a non-entrepreneur? How does he form prescient foresight on future goods and services? These questions are frequently raised in the entrepreneurship research because according to Knight (1921), profit stems from entrepreneurs' bearing risk and uncertainty. Hayek (1937) argues that an entrepreneur's foresight can be formed from two sources. First, entrepreneurs' understanding and interpretation of data can affect foresight. Second, general tendency or routine can also be a source of foresight. Whitehead (1933) illuminates foresight from the understanding and routine perspective; he points out that "if we knew enough of the laws, then we should understand that the development of the future from the past is completely determined by the details of the past by these scientific laws which condition all generation" (p. 87). He indicates that unfortunately our knowledge of scientific laws is woefully defective; thus our knowledge of the relevant facts of the present and past is scanty in the extreme. Whitehead (1933) further argues that foresight depends on understanding, which can be acquired by conscious effort and can be taught:

Thus the training of Foresight is by the medium of understanding. Foresight is the product of Insight (Whitehead, 1933, p. 89). Hosinski (1993) observes Whitehead's distinction between understanding and knowledge. Understanding is achieved in insight and expressed in hypotheses, and knowledge is tested understanding. Hosinski (1993) further points out that Whitehead (1929) illuminates the knowledge that we acquire through this method is not final, absolutely certain knowledge. It is always partial, liable to errors in judgment, and open to future improvement and correction. Alert entrepreneurs consciously search entrepreneurial opportunities in the field in which they have experience and knowledge, since previous knowledge is a source of entrepreneurial opportunities (Shane, 2003).

Whitehead (1933) is cognizant of the importance of routine as the bedrock of efficiency, but at the same time recognizes that foresight is crucial for success in innovation (change). He points out that the modern commercial mentality requires many elements of discipline, scientific

and sociological. However, he indicates that the great fact remains: details of relevant knowledge cannot be foreseen. He then argues that the quantitative aspect of social change is the essence of business relations, and the habit of transforming observation of qualitative change into quantitative estimates should be characteristic of a business mentality. He points out that the business of the future must be controlled by a somewhat different type of person compared to those of previous centuries and illuminates requisite abilities for business mentality:

Thus even for mere success, and apart from any intrinsic quality of life, an unspecified aptitude for eliciting generalizations from particulars and for seeing the divergent Illustration of generality in diverse circumstances is required. Such a reflective power is essentially a philosophic habit: it is the survey of society from the standpoint of generality. This habit of general thought, undaunted by novelty, is the gift of philosophy, in the widest sense of the term (Whitehead, 1933, pp. 97-98).

Those entrepreneurs who have learned this habit of general thought may form foresight on the emergence of novelty in business. Entrepreneurial actions based on this foresight may reduce uncertainties in entrepreneurial actions because knowledge and foresight reduce entrepreneurial errors. However, entrepreneurial decisions are liable to errors in judgments, as stated above. Popper's (1969) fallibility also supports this uncertainty. An entrepreneur may know that her discovered entrepreneurial opportunity can be profitable and take entrepreneurial action, but this knowing is fallible, and entrepreneurial action can result in an uncertain outcome (Fantle and McGrath, 2009).

Fantle's (2008) interpretation of Ryle (1946, 1949) can be relevant in entrepreneurship and uncertainty. According to Fantle, Ryle's know-how is a matter of ability and there are various grades of know-how. Entrepreneurs themselves have various grades of ability to discover, evaluate, and exploit opportunities. Successful entrepreneurs should have a higher grade of abilities in discovering, evaluating and exploiting opportunities; these higher grades of require knowledge of know that and practice. As stated by Roland (1958), knowing how to ride a bicycle is a capacity which implies having learned how to ride through practice.

A capacity of entrepreneurship implies that entrepreneurs learned know-how to become successful through practice. Very few entrepreneurs succeed in their first entrepreneurial action.

Various structures for knowledge creation and new combination have been developed in the market. The Dow Chemical Company, a leading global innovating firm, provides an interesting case study for knowledge creation and corporate entrepreneurship. The Dow Chemical Company has customers in 180 countries and had annual sales over $57 billion in 2013. The company employed approximately 53,000 people worldwide and has more than 6000 products manufactured at 201 sites in 36 countries across the globe.

5. Corporate Entrepreneurship: The Dow Chemical Company

Companies such as Dow Chemical depend on knowledge and competence creation for their survival. To encourage knowledge creation, The Dow Chemical Company has established Idea Central, a database. The variations in experiences by Dow employees in all functions, regions and ages are key sources of knowledge creation, so when an employee or a group of employees has an idea, he/she can post it to Idea Central. All employees are encouraged to propose ideas. Dow believes that no ideas are bad ideas; this creates the culture that all ideas are appreciated and valued. Employees who work for Dow learn from their daily experiences in dealing with customers and suppliers or in their labs. Dow obtains tacit knowledge from them and from all functions, ages and regions in the world, since it is a global company. Variations and diversity of ideas are crucial sources of knowledge creation because they can be combined to create new knowledge. Therefore, Dow management encourages employees to express diverse ideas and create an environment for diversity.

Knowledge is necessary for the entrepreneur to recognize an entrepreneurial opportunity when one appears (Holcombe, 2003). The process of Idea Central helps employees capture tacit knowledge they experience at a specific time and place. Since tacit knowledge is a prime

source of new knowledge creation, as Polanyi (1966) argued, it is acquired through lived experiences (i.e., doing, using, and indwelling), and employees receive intimation from the realities of their daily job. But tacit knowledge of their experiences is only momentarily realized, so it will be lost if employees do not capture it at that moment, because knowledge is context-specific in time and place (Hayek, 1945; Polanyi, 1966). When Idea Central operates in the organization, employees develop innovation dispositions, as Bourdieu (1977, 1990) advocated. The Idea Central of Dow Chemical is also a device to utilize dispersed knowledge (Hayek, 1945), since Dow's employees are dispersed around the world. This process of knowledge creation has multiple aspects.

5.1. Structure in External Knowledge Mobilization (Environmental Context)

In addition to Idea Central, Dow Chemical also has scouting departments whose jobs are obtaining knowledge from outside sources, such as academic journals, conferences, universities and government agencies (e.g., NIH). Scouting, designed to obtain knowledge from outside of Dow Chemical, is divided into many different specialties, such as engineering and science, and acquires external knowledge from various sources. Employees in a scouting department also search out information and knowledge from suppliers, customers and competitors. Dow Chemical arranges a license agreement with those who have patents, if the patent is deemed necessary for Dow. This knowledge creation practice of Dow provides evidence of Nonaka and Toyoma's (2003) boundary crossing. Chesbrough (2003) refers to this as the open innovation system, the use of purposive inflows and outflows of knowledge to accelerate innovation. Chesbrough (2003) argues that such open innovation saves time and costs compared to closed innovation.

External sources of knowledge have been gaining importance in the knowledge economy. Apple Company, for example, is known for offering strong incentives to small companies and individuals with new knowledge. If knowledge from external sources is adopted in Apple products, a significant portion of the additional profit generated is offered to the firm or the individual who supplied the knowledge. This

is a strong incentive to the supplier of knowledge, because Apple Company has a large market, and this creates value for both Apple and knowledge suppliers. Through this process Apple Company acquired multi-touch sensing capabilities from FingerWorkks (Isaacson, 2011), and Google acquired Android.

5.2. Screening

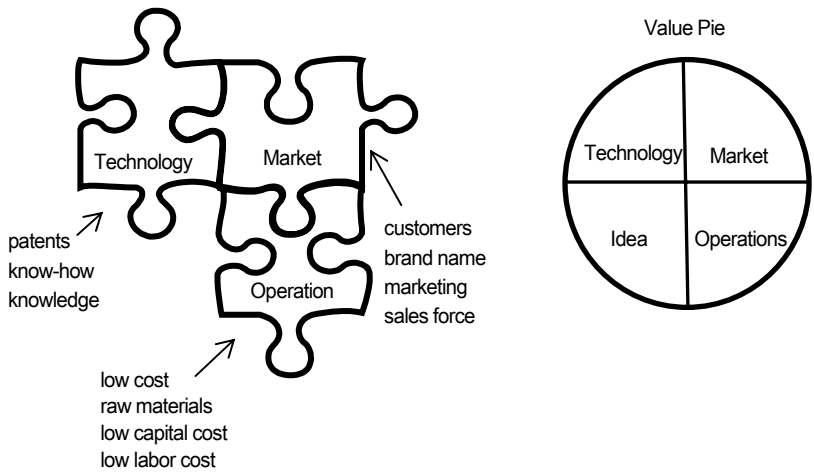
The next phase, screening of Idea Central, is done on a monthly basis. A cross-functional team consisting of four or five members from various functions goes through all posited proposals. This team makes judgments on the ideas, based on deeper analysis of the diverse proposals submitted by employees from all over the world. The team initially selects the ten best proposals from Idea Central and then narrows the list down to a few ideas for funding. Dow saves all proposals for future use. Some proposals may not be viable at the time of screening, but could become value-generating ventures in the future, as the economic environment changes (e.g., new technologies and market conditions). Since all proposals are stored in Idea Central and are accessible by Dow employees, Idea Central serves as an effective organizational memory. Therefore, costly knowledge assets that Dow created are not lost, and Idea Central mediates the problem raised by Nonaka et al. (2006) and stemming from lack of coherence in decision making. Dow can also make use of knowledge assets stored in Idea Central by combining it with future proposals for creating new products, processes or services. Screening can be regarded as a validation process of new knowledge, because knowledge is "verified true belief."

The way that Dow Chemical screens proposals is interesting, like putting pieces of a puzzle together. Dow maintains that sustained profitability of any business is possible only when three advantages all exist. The first piece of the puzzle is technology advantage. Does Dow have a patent or patents for the proposed idea? Patents that need to be evaluated are composition, process, and application patents. If Dow does not have these, they need to solve that puzzle by contacting patent holders for license agreements or propose a joint venture. The second piece of the puzzle is operations advantage, which includes low cost raw

materials, low capital options, process experience, synergistic cite, and supply chain and regulatory constraints such as environmental health and safety regulations. Dow assesses raw material requirements as well as accessibility of required raw materials, facilities and capital. Costs of acquiring them need to be competitive. The third piece of the puzzle is market channels, which includes customers, brand names, reputation, synergistic sales, channel partners, market sales forces and requirements of regulatory agents, such as the Environmental Protection Agency (EPA) and Food and Drug Administration (FDA).

Within the three large pieces there are smaller puzzle pieces. In assessing each piece of the puzzle, limitations or constraints are analyzed. If Dow is missing technology pieces, they may develop them within the organization or acquire them from outside. Dow needs to assess the availability of raw materials, facility and capital as well as missing pieces of the operations puzzle. When they find missing pieces (constraints), they have to solve them. The market channel piece of the puzzle includes figuring out customers, brand names and sales forces.

Figure 1 | Screening process



They also examine market penetrability; some markets are very difficult to penetrate because new comers have insurmountable barriers to overcome. When all the missing pieces of the puzzles are identified, they put them together to create value for Dow, which will benefit from the entire value pie. Figure 1 is a graphical presentation of this screening process.

Once Dow has established the new competency at competitive levels, competitive advantage depends on recognizing new business opportunities and recombining these capabilities and unexploited proposals at data central, as emphasized by Helfat et al. (2007) and Argyres (2011). Argyres argues that "organizational economics can offer insights into features and processes that can promote or hinder opportunity recognition within organizations" (p.1141). According to Argyres, "governance choices are endemic to any capability development process, because such processes involve structuring incentives, allocating authority, and stimulating information flow" (p.1142), which influence knowledge creation and sharing.

It is hard to make judgments about what pieces of the puzzle can or cannot be found, and even if the pieces are found, the cost structure may not be competitive enough to embark on the project. Dow has learned the hard way not to waste time and resources for a project that has missing pieces, or is too costly or impossible to find. Appropriate structures are established based on experiences of past failures to prevent future failures. This establishment of structures is an example of the realization of Giddens' structures and agents theories (1984) and Bourdieu's reflexive feedback and self-monitoring (1977, 1990). Once a knowledge creation structure is established agents will become reflexive and self-monitor in response to the structure and the structure should be changed based on the feedback of agents. Therefore, the knowledge creation process is both emerging and ongoing, as Whitehead advocated (1929).

5.3. Collaboration Options

What can Dow Chemical do with ideas when they cannot overcome constraints or find missing pieces of the puzzle? Dow figures out factors that the company is missing in the puzzle, then decides to create them

internally or to fill them from outside. They license out (sell) ideas to those who have the capabilities to use them or form a joint venture. Dow Chemical may search for a firm that needs the knowledge that Dow has and negotiate a license agreement. A joint venture agreement is another solution for utilizing their ideas. They form a joint venture by connecting a network of companies which have pieces of the puzzle (technologies, ideas, and operation). They can often find small companies who have good ideas or Dow's missing puzzle pieces to establish these joint ventures. This process shows that innovation capabilities can be distributed across firm boundaries, and managers can combine distributed innovation capabilities to create a new capability through structures which coordinate the contributions of the various participating firms (Coombs and Metcalfe, 2002).

Coombs and Metcalfe (2002) argue that we need to consider how the capabilities perspective can be extended to embrace a multi-firm perspective on innovation. They further argue that capabilities themselves become an important unit of analysis which is not coterminous with the firm. Madhok (2002) points out that "the reasons for collaborations between firms is to combine synergistically two sets of complementary but dissimilar resources and capabilities in a manner which will generate returns that will either create a market transaction or complete internalization" (p.277). Dow Chemical recognizes that advantages frequently occur in different organizations; collaboration is therefore critical because each organization has a unique pile of "advantaged" puzzle pieces.

There are many possibilities to find matches with other collaboration partners who have technology, operations and/or market advantage that Dow Chemical does not have. Dow can buy, sell, create, and trade pieces with collaboration partners such as universities, large companies, small and startup businesses, market specialists and incumbents. Finding optimum partners is crucial for the success of collaboration. Dow has developed good strategic mixes for acquiring advantage in collaboration, a necessary factor (Whiteman, 2013). The practices of Dow's scouting, collaboration and joint venture can be regarded as an open innovation system (Chesbrough, 2003). The openness of technology sharing economizes costs in new product development by accomplishing economy of scale and scope, as Chesbrough (2003) argues. According to Chesbrough

(2003), open innovation is the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and to expand the markets for external use of innovation, respectively. Dow's practice is an example of such open innovation.

However, the cross-firm structures for new capabilities innovation need to address governance issues on opportunism. Firms need capabilities to write contracts that efficiently deal with opportunistic behavior. Joint ventures among many firms require trust among the participating firms. Past experiences help build trust and reputation within the industry. If a firm becomes known for taking advantage of participating firms and not sharing ideas, other firms are likely to avoid future joint ventures. Thus building a good reputation can be a strategic asset in the long run, because business dealings can be ongoing and repeating.

Once the value pie is created, the firms split it based on contributions or strengths of each participant (see Figure 1 for a value pie). The initiating or proactive company can secure a better share of the value pie. This example shows that how the governance choices of the firm (Williamson, 1991, 1999; Argyres, 2011) are involved in any capability development.

Dow forms over 100 joint ventures in a year, and having good intuition for what consumers would like to have is a key factor for success in joint ventures, as well as in Dow's new product development. Firms can search to find an addressable market and size and explore market needs based on voices of customers (VOC). As von Hippel (1988) indicates, customers and suppliers are the most important sources of innovation, as the voices of customers provide information on market needs. However, managers' intuition on what consumers want plays a key role in the success of a new product. The cofounder of Apple Company, Steve Jobs, was known for his good intuition on consumers' preferences, and Steve Jobs and his designer, Ive, worked to simplify new design. Apple's first brochure proclaimed "Simplicity is the ultimate sophistication," as Steve Jobs had aimed for the simplicity that comes from conquering complexities, not ignoring them (Isaacson, 2011, p. 343). Samsung (a large Korean chaebol) is taking a different track but also values consumers' knowledge and experiences (Chen, February 11,

2013). Knowledge about what consumers want is a hidden reality and managers/knowers can capture intimation from that hidden reality. Their intuition is a key element in the success of a new product. Dow's success rates are known to be 20-30 percent. This process shows how the network form of organization emerges as an adaptation to changes in the environment for knowledge creation (Hannan and Freeman, 1984; Levinthal, 1997).

Dow Chemical's practice for new product development also illustrates an integration of macro-micro approaches. The Idea Central (macro level) structure is designed to solicit tacit knowledge from individuals (micro level). The variations in knowledge creation structures among organizations may have a different impact on individuals, and individuals' responses to the same structure may differ based on personal traits. If an Individual's ideas are verified by a team and the project proposed by an individual becomes a new competency by team efforts (meso level), interactions among individuals result in new qualities on the macroscopic level of the system (Fuchs, 2003). Interactions among individuals from cross-functions lead to new knowledge and competency creation in Dow's process. Variations of ideas offer opportunities for the organization to recombine these for new competency (Schumpeter, 1934; Kogut and Zander, 1992; Nonaka, 1994; Nonaka and Takeuchi, 1995). This example of Dow's knowledge creation illustrates how to engage in knowledge creation activity and how this knowledge evolves over time (Levinthal, 2006; Winter, 2006).

Dow Chemical's Idea Central can thus be seen as a "ba" (a knowledge creating place: Nonaka and Kono, 1998, Nonaka et al., 2000a). Bas may vary depending on the contexts of knowledge creating firms. The market selects a good ba and the selected ba retains good characteristics of the ba. Variations in knowledge creating bas will continue in the market; the knowledge-creating firm is, therefore, dynamic. According to Nonaka and Toyama (2003), "ba" can transcend time, space, and organization boundaries to create knowledge. A knowledge-based view of the firm regards the firm as a knowledge creating entity, and in this view a firm is changed from an entity of being to an entity of becoming (Nonaka and Toyama, 2002, 2003), as seen in Prigogine's scientific

view of the world (2003) or Whitehead's process philosophy (1929). In the same vein Schiuma (2009) argues that the ability to harness knowledge assets dynamics lies at the core of organizational value creation capacity. Schiuma (2009) introduces the managerial foundations of knowledge assets dynamics and discuss three processes: knowledge assets identification, knowledge assets mapping and knowledge assets flow. He points out that understanding the evolutionary stages of knowledge assets flow helps organizations produce timely products for the markets.

Collaboration saves precious time in developing new products, processes and services. Collaboration may be a preferred practice in a high-velocity economic environment because firms can bring new products to market quickly. Nonaka et al. (2000b) point out that building one's own knowledge comes with cost, i.e., time. Since building up knowledge assets through a firm's own knowledge creating process takes time, it is costly (Nonaka et al., 2000b). Teece (2007) argues that the opportunity cost is especially high when the industry that the firm is in is a high velocity economic environment. If the firm develops the missing puzzle pieces by themselves, it may fall behind the market competition.

5.4. The Role of Incentives

The intervention of managers to create knowledge can occur at both institutional and individual levels (Abell et al., 2008). First, managers need to provide opportunities for agents to interact. Managers and individuals together design an institution to be conducive for dialogue and interactions. Second, for diversity of ideas, managers need to hire individuals from many different universities and institutions and provide incentive systems for individuals and groups to acquire and share knowledge. When agents acquire useful knowledge (genotype) and skills, firms develop technology (phenotype) based on this useful knowledge and knowledge sharing (intensity of knowledge), as Mokyr (2000) argues. When agents accumulate more knowledge and develop dispositions of imagination and creativity, the firm will have easy access to both knowledge and creativity (Mokyr, 2000), because knowledge is

stored in agents who supply knowledge and skills to the firm. Therefore, developing an incentive system for agents to constantly acquire knowledge and skills is crucial in building the dynamic capabilities of the firm. Individuals' choice to work in an organization, and their acquiring and sharing knowledge, depends on firm characteristics. As discussed before, structuring incentives and allocating authority are thus important in knowledge creation and sharing.

Because managers at Dow appreciate diversity and createa culture for employees to supply many different ideas, it rewards those who produce the most ideas as well as the best idea.

6. Discussion

In process philosophy (Whitehead, 1929) and modern sciences (Prigogine, 2003) the universe is emerging. For Whitehead (1929) reality is becoming and ongoing. Oranizations emerge and continue to make changes in the universal context of Whitehead's reality. Entrepreneurs are actors carrying on changes and they make changes based on ideas from knowledge. Entrepreneurship theories explain these changes.

Leading entrepreneurship theories are the discovery theory and creation theory. Kirznerian entrepreneurs engage in arbitrage and Schumpeterian entrepreneurs create new combinations. Knowledge, an important source of entrepreneurship, is utilized to deal with entrepreneurial risk and uncertainty. Sources of entrepreneurial opportunities in the discovery theory (Kirzner) are disturbance in equilibrium caused by technologies, markets, regulations, errors made in previous entrepreneurial actions and Schumpeterian innovation. There is no entrepreneurial opportunity when the industry or market is in equilibrium or circular flow. Alert Kirznerian entrepreneurs discover entrepreneurial opportunities and exploit them, engaging in arbitrage. Schumpeterian entrepreneurs create new entrepreneurial opportunities by new combinations, which Schumpeter refers to as innovation: 1) product innovation, 2) process innovation, 3) organization innovation, 4) market innovation and 5) input innovation. For Schumpeter, entrepreneurs respond creatively, not routinely, to changes in data such as technologies, markets

and regulations. Schumpeterian entrepreneurs carry out new combinations and engage in innovation in product, process, organization, market and input to create opportunities. For Schumpeter innovation never stops and is disruptive; he referred to it as the perennial gale of creative destruction (1942).

Errors in entrepreneurship are bound to recur as seen in the case of Dow. The Dow Chemical Company's success rate is between 20 and 30 percent of new ventures. Errors are stemming from many sources such as lack of foresight and knowledge. Lack of market information leads to errors in judgement because entrepreneurs produce goods for the future and there are no markets for future goods (Arrow, 1962, 1974).

Entrepreneurial actions are the mechanism for the process of economic development. Entrepreneurs foster firm growth, which causes the economy to grow. As discussed before, knowledge, an important source of innovation is dispersed and does not exist in a useful form, known as the knowledge problem (Hayek, 1945). Knowledge is tacit and embedded in individuals (Polanyi, 1966). When a firm wrings out knowledge from individuals, knowledge becomes a public resource and knowledge is abundant. However, making use of knowledge is scarce (Nonaka, 1994; Hislop, 2009).

How do firms enact knowledge for innovation? The case of The Dow Chemical Company described above illuminates the enactment process (Weick, 1977) or structure (Giddens, 1984). Giddens' structuration theory (179, 1984) provides a framework for knowledge creation structures and agency. If we assume that there are two employees in a knowledge creation team, both have tacit knowledge and lived experiences from their respective jobs. They have acquired knowledge from customers, investors, partners, competitors and the scientific community. Therefore, individuals engaged in knowledge creation are knowledgeable, conscious and reflexive. Individuals can also anticipate possible future states, based on their abilities to detect intimation from the hidden reality in their fields (Polanyi, 1969). They can anticipate change in technologies, markets and regulations. A knowledge creation team consists of individuals with these traits who participate in knowledge creation. The structure for enacting knowledge has two charateristics: 1) Eliciting tacit knowledge and 2) constraining creativity. The Idea Central at the Dow

Chemical Company brings out tacit knowledge from stakeholders, but this structure can inhibit creativity by using criteria for choosing a project (Shaviro, 2009). Therefore, the structure of knowledge creation can cause the antinomy of eliciting tacit knowledge and hampering creativity. Organizations need to address the antinomy of the Idea Central by making changes in structure based on the feeback of stakeholders and storing proposed ideas as accessible data for any potential future uses.

Since knowledge is a public resource, everyone can have access to the knowledge resource and it does not deplete. Today, firms compete for access to this abundant knowledge resource. Open innovation systems (Chesbrough, 2003) have emerged as a mechanism to exploit this knowledge resource. Firms obtain knowledge from employees, suppliers, customers, competitors, regulators and academic communities. User knowledge can be inexpensive to acquire because of information technology (Park et al., 2014b), and open innovation has become the newly emerging practice which captures dispersed knowledge. More corporations are forming networks to share their patents and knowledge, and collaboration among corporations is increasing. Google, for example, has formed collaborative relations with several firms such as Samsung, and Samsung is collaborating with Sony, Facebook and BMW. Korean firms such as Samsung, LG, Hyundai and Posko are also known for their continuous knowledge creation and innovation.

Three dimensions of knowledge: know how and know that (Ryle, 1946, 1949); tacit knowledge and explicit knowledge (Polanyi, 1962, 1966) and information as knowledge (Hayek, 1937, 1945) can be integrated in discovering and exploiting entrepreneurial opportunities. Tacit knowledge is the source of explicit knowledge and it can lead to know how by organizations as seen in the case of the Dow Chemical company. Entrepreneurs identify entrepreneurial opportunities with information in price differentials among various regions. Therefore, deeper understanding of dimensions of knowledge and the relationship among them helps identify sources of entrepreneurial opportunites.

Current changes in technologies, markets and regulations present abundant entrepreneurial opportunities. New technologies such as 3D printing, digitization and carbon fiber open up entrepreneurial oppor-

tunities similar to IT and BT in the '90s. These technologies may provide an impetus for new waves of entrepreneurship and economic growth. These three new technologies have been referred to as the third industrial revolution by *The Economist* (2012). Emerging global markets are continuing sources of entrepreneurial opportunities; the Affordable Care Act in the U.S. also offers tremendous entrepreneurial opportunities. Changes in global mandates stemming from the degradation of environment offer entrepreneurial opportunities in new energy sources and fuel efficient automobiles, recycling carbon dioxide, biofuels and electric cars. External shocks or disturbances in equilibrium, such as the recent financial crisis, also offer entrepreneurial opportunities. Recent real estate prices have risen from the price levels at the outset of the crisis and are likely to rise in the future. European countries such as Spain, Greece, Italy, Portugal and Ireland, which have faced severe crises, also offer entrepreneurial opportunities. Alert entrepreneurs can realize large profits when they exploit the opportunities created by price differentials. Such entrepreneurs engage in arbitrage and generate a tendency towards equilibrium. However, the new state of the economy is always emerging and continuously changing. Therefore, entrepre-neurial opportunities are ubiquitous and alert entrepreneurs will continue to discover and exploit them.

Entrepreneurs create new products and services for profits. The creativity is the important souce of entrepreneurship. We thus need to elucidate the origin of the creativity. According to Whitehead (1929) the creativity is the actualization of potentiality, and the process of actualization is an occasion of experiencing. A new actuality always arises out of a given world that conditions the creativity which transcend it (p.43). Whitehead (1929) considers two meanings of potentiality: (a) the "general" potentiality, which is the bundle of possibilities, mutually consistent or alternative…. and (b) the "real" potentiality, which is conditioned by the data provided by the actual world (1929, p. 65). Whitehead's real potentiality is akin to Schumpeter's (1934, 1942) innovation in products and services responding to changes in data. Therefore, changes in data by the actual world offer entrepreneurial opportunity. **The creative economy is actualization of the potentiality.** For Whitehead, reality is emerging and becoming. Whitehead (1929)

points out that the creativity drives the world. We observed that entrepreneurial actions at Dow creates new products for profits and innovation in products drives the Dow company. Entrepreneurship is not just about creating new products for new firms, but it also applies to **new as well as existing firms**. It is crucial for policy makers to balance their emphasis on both new and existing firms. If they overly emphasize the newly creating firms or technologies, they may neglect the existing firms, which are crucially important for the creative economy. Entrepreneurship is essential for the creative economy and it drives the economy, facilitates economic growth and creates jobs. **The creative econmy may be defined as actualizing the potentiality of the nation's economy.** For Whitehead (1929) creativity is a transition of many into a novel one. Reality has many facets and these many make a transition into a novel one. Employees, consumers and other stakeholders experience many aspects of a firm's product. Experiences of these stakeholders on a product (Park et al., 2014b) help create a new product or service. As seen from the case of the Dow Chemical Company, not all potentials are actualized. Potentials that are not utilized in a new product are stored and they remain as potentials for future products.

The creative economy is about designing an effective process of a transition of many stakeholders' experiences into a new product. Entrepreneurs carry out the task of the new way of production. The potentiality of the economy is to continue to emerge and becoming, and entrepreneurship is therefore ongoing. According to Whitehead (1929) reality is ongoing and becoming, and there is next and future because of creativity. Whitehead's philosophy is a metaphysics of becoming and philosophy of creativity. According to Shaviro (2009, viii), Whitehead asks, "How is it that there is always something new?" Shaviro (2009) argues that Whitehead's creativity is precisely the manner in which something radically new can emerge out of the prehension of already existing elements. Creation of new competence is all a matter of " 'the subjective form' which is how that subject prehends that datum" (Whitehead, 1929, p. 23). Entrepreneurs and policy makers need to properly prehend changes in economic environment.

The creative economy in Korea is about creating something new in

products, processes, services markets and organizations by prehending current economic environment (data). Therefore, Whitehead's philosophy can lay a ground work for trajectories of knowledge assets flow (Schiuma, 2009) and provide philosophical grounds and foundations for the creative economy in Korea. We have to keep in mind that the creative destruction (Schumpeter, 1942) is disruptive and entrepreneurs and policy makers are required to prepare for ensuing disruptions of the creative destruction.

What does cause creativity? Shaviro (2009) points out that a final cause (or teleological) is always at work, alongside the efficient (mechanistic) cause in Whitehead's potential for change or novelty. Whitehead insists that every entity is essentially dipolar, with its physical and mental poles (Whitehead, 1929, p. 239). For Whitehead (1929) the final cause is the "decision" (p. 43) by means of which an actual entity becomes what it is. The future Korean economy will, therefore, be created by the immanent mechanistic (physical) elements such as smart phones for now and the entrepreneurs and policy makers' decisions (mental) in investment for the chosen field (i.e. R & D in information technologies in the 90s) and entrepreneurship.

Knowledge assets dynamics in the literature are emerging as the managerial foundations of organizational value creation (Schiuma, 2009). According to Schiuma (2009) through knowledge assets dynamics, organizations are able to continuously develop, upgrade and extend their capabilities by improving their capacity of exploiting internal resources and their abilities to identify and shape new business opportunities (p. 292). He argues that knowledge assets dynamics reside at the core of organizational value creation capacity and analyzes the knowledge assets dynamics through knowledge assets identification, mapping and flow. We can see that Dow identifies knowledge assets domains through screening, maps knowledge assets by focusing on value creation driver and evaluates knowledge assets flow by analyzing the future direction of knowledge assets.

Based on theories of entrepreneurship and the case study of Dow, we can draw policy implications, elaborated in the following section.

7. Policy Implications

Recently, policies fostering entrepreneurship have become prominent in policy debates on economic growth and job creation because entrepreneurship is seen as the engine of economic growth. The classical economic growth model had limited success in explaining the growth of mature economies. Naturally, the endogenous growth theory has emerged as a leading alternative to the classical growth theory. Entrepreneurship is the core of the endogenous growth theory. As discussed before, entrepreneurial ideas stem from knowledge. Knowledge is abundant and nondepletable. In today's knowledge economy, economists are required to learn the optimal uses of the abundant knowledge resource. The emerging knowledge economy causes a paradigm shift (Kuhn, 1962) from the economics of scarcity to economics of abundance because economics traditionally deals with making choices under the scarcity of resources. Knowledge resource is abundant, but making use of knowledge requires traditional scarce resources. It is interesting to discover how firms exploit abundant knowledge. The Dow Chemical Company case illustrates the value of eliciting knowledge from employees and external stakeholders. Thus we can draw a few policy recommendations from this analysis.

First, policy makers and corporate leaders need to have a deeper understanding of the nature of knowledge: Ryle's know that and know-how; Polanyi's tacit and explicit knowledge; and Hayek's knowledge as information. Policy makers need to provide forums for integration of epistemology in philosophy, physical and social sciences and arts. Deeper interactions among scholars, practitioners and policy makers in these fields are essential for the creative economy because interactions and combinations are sources of creativity. The creative economy should be built on sound philosophical ground because sound theoretical and philosophical ground provides an impetus for continuous economic growth and prosperity. Entrepreneurship is about the process of creating a new competence and Whitehead's process or creative philosophy is exactly about the process of creativity.

Second, policy makers and corporate leaders need to design a structure to elicit tacit knowledge from employees, all stakeholders and

academic communities (Park et al., 2014a). Interactions among knowledge creating workers and openness and trust among them are essential in the structuring of knowledge creation. Firms also need to make changes in structure based on conscious feedback of stakeholders (Park et al., 2014a). Idea Central at the Dow Chemical Company is a structure enacting tacit knowledge for Dow's new competence creation. This structure makes it possible for Dow to introduce new products to the market. The structure enacts tacit knowledge, but at the same time it may inhibit creative ideas because good ideas may not be selected by applying a set criteria for selecting a project (Shaviro, 2009). Therefore, a structure of enactment needs to be flexible and adaptable to a changing environment (Park et al., 2014a).

According to knowledge scholars knowledge is abundant, but making use of knowledge is scarce (Hislop, 2009). Knowledge is a public good, it is not depleted by using it, and everybody can access it. Designing good structures is essential, because public goods will be under-produced if incentives are not structured properly. Knowledge is abundant in Korea. However, the retirement age in Korea is 55 or 60. Those who retired at the age of 55 or 60 have an untapped wealth of experience and knowledge. Knowledge does not deplete by using, but gets obsolete. Knowledge continues to be renewed and updated that can be mobilized for entrepreneurship if Korea creates an appropriate structure. Policy makers and managers also need to continuously evaluate knowledge assets flow in exploring their future entrepreneurial opportunities. It is high time for scholars to study determining factors of knowledge assets flow because the Korean government is currently pursuing an economic policy of "the creative economy." Korean policy makers and entrepreneurs were relatively successful in identifying the direction of knowledge assets flow in the past and created new competencies in a timely fashion which led to temporal monopoly profits and economic growth. Therefore, we need to have abetter understanding of the evolutionary stages of knowledge assets flow dynamics in terms of their trajectories (Schiuma, 2009).

Third, entrepreneurship should be practiced in all fields: corporate entrepreneurship, public entrepreneurship and social entrepreneurship. Policy makers need to encourage entrpreneurship and discourage rent-

seeking behaviors by enforcing rules and regulations and eliminating unnecessary rules, regulations and inefficient social customes. There are large public sectors in the Korean economy and rentseeking behaviors in public enterprises cause huge inefficiency.

Fourth, education to encourage creativity is crucial. To take one example, the Korean education system is known for being very costly for the same outcome compared to advanced economies such as the U.S and Sweden. Creativity is the most importance source of entrepreneurship.

Fifth, research and development is crucial for small and medium size firms world-wide. Korean entrepreneurial firms, for example, need to pool their resources and public resources. Taiwan already has such a research center for pooling participating firms and public resources, and more such models are needed with research outcomes made available to all participating firms. Since firms are competiting in the global market competition among small firms within a nation has become less severe or non-isssue.

Sixth, policy makers need to foster entrepreneurship and innovation. Because entrepreneurs are alert, they can capture the quality of changes in the air and translate them into new products and services. Today's emerging new technologies such as 3D printing, digitization and carbon fiber can offer huge entrepreneurial opportunities. Consumers' changes in preferences for carbon reduction can offer entrepreneurial opportunities. Policy makers need to institute support systems for these entrepreneurial opportunities. Continuous entrepreneurial actions need to take place at all levels of organizations: new firms, governments, established small and large firms, and society as a whole. Governmental and corporate leaders should become entrepreneurial. Organizations are required to become flat and more tolerent for employees' risk taking and mistakes as long as they learn from them and do not repeat.

Seventh, policies managing risk and uncertainty need to be refined. Private venture capital is relatively well developed in the U.S; state funding for new ventures is often available, and cooperative funding for new ventures is growing. Angel funding is another source for a new venture. Securing funds from many small investors, called crowd financing, is also gaining popularity. Optimal risk taking should be

defined and operationalized, because too much risk taking can also be a waste of financial resources. Financing new ventures involves principal and agent issues and asymmetry of information, which cause adverse selection and moral hazard problems. Adverse selection and moral hazard have been huge problems in financial markets, causing recurring financial crises in the world. There are new strict regulations for these financing methods in the U.S. because they are risky and uncertain. Other countries may need similar regulations.

Eighth, policy makers and corporate leaders need to manage both routine and innovation effectively for a seamless transition from one routine to a new routine. Innovation is always disruptive, as Schumpeter (1942) refers to as the creative destruction. Entrepreneurs carry on disruptive innovation and persuade organizational members who resist changes (Langlois, 1998). Whitehead (1933) argues that routine is the god of factory manufacturing, but the firm also needs to innovate products, services and organization to thrive. Making innovation in products, services and organization seamless is the most important task in today's high velocity economy. Rapid changes in economic environment such as markets, technologies and domestic and global regulations are entrepreneurial opportunities as well as challenges. Policy makers and corporate leaders need to identify barriers to accomplishing potentials and eliminate them. Langlois (1998) points out that innovative entrepreneurs engender rapid and sustained economic growth. We have witnessed entrepreneurs' contributions to economic growth and job creation in Korea as well as in other countries. Founders of Samsung, LG, Hyundai, Ford, GE, Apple and Google are examples of entrepreneurs' contributions to economic growth and job creations.

Finally, knowledge is fallible and there is no market for future goods and services. Therefore, entrepreneurs face uncertainty and failures. Knowledge and deeper understanding can reduce them, and entrepreneurs' foresights and prescient knowledge are crucially important for managing uncertainty. Entrepreneurial actions need to be persistent and accumulate knowledge from their mistakes and experiences.

8. Conclusion

Entrepreneurship is essential in the creative economy. Entrepreneurship studies the discoveries and the creation of entrepreneurial opportunities as entrepreneurs generate new ideas and put them into effect. Ideas stem from knowledge; because knowledge is tacit, it needs to be enacted to elicit knowledge from individuals who hold that knowledge. Entrepreneurs can wring out knowledge from individuals because knowledge exists in individuals, and they put knowledge into use for innovation in products, services, processes, organizations, markets and inputs. Tacit knowledge becomes know how for individuals and organizations. The know how in the organization is competency of the organization. In the traditional economy inputs such as labor, land and capital are scarce, but in today's knowledge economy knowledge is the most important input and knowledge is abundant. Knowledge is the most important source of the entrepreneurship. Tacit knowledge requires enactment (structure) and entrepreneurs are enactors. Therefore, we conclude that knowledge and entrepreneurship are foundations of the creative economy and entrepreneurship is the key source of rapid and sustainable economic growth. The continuous success of entrepreneurship is essential to secure the economic growth and prosperity of a nation and Whitehead's creativity lays a philosophical foundation on the creative economy.

References

Abell, P., Felin, T., & Foss, N. (2008). "Building microfoundations for the routines, capabilities, and performance links," *Managerial and Decision Economics,* 29 (6), pp. 489-502.

Akerlof, G. A. (1970). The market for "lemon": Quality uncertainty and the market mechanism, *The Quarterly Journal of Economics,* 84 (3), 488-500.

Alchian, A. A. and Demsetz, H. (1972). "Production, information cost and economic organization," *American Economic Review,* 62 (4), 777-795.

Alvarez, S. A. and Barney, J. B. (2007). Discovery and creation: Alternative theories of entrepreneurial action, *Strategic Entrepreneurship Journal*, 1(1-2), 11-26.

Alvarez, S. A. and Barney, J. B. (2010).Entrepreneurship and epistemology, *The Academy of Management Annals,* 4(1), 557-583.

Argyres, N. (2011). "Using organizational economics to study organizational capability development and strategy,"*Organization Science*, 22(5), pp. 1138-1143.

Arrow, K. (1962). "Limited knowledge and economic analysis," *The American Economic Review*, 64 (1), 1-10.

Arrow, K. (1971). *Essays in the Theory of Risk-Bearing*, Amsterdam, Holland: North-Holland Publisher.

Arrow, K. (1974). *The Limits of Organism*, New York: W.W. Norton.

Baumol, W. J. (1968). Entrepreneurship in economic theory, *The American Economic Review*, 58 (2)64-71.

Bhaskar, R. (1975). *A Realist Theory of Science,* London: Leeds Books Ltd.

Boettke, P.J. and Sautet, F. (2011).The genius of Mises and the brilliance of Kirzner, GMU Working Paper in Economics No. 11-05.

Bourdieu, P. (1977). *Outline of a theory of practice* (translated by Richard Nice), Cambridge, UK: Cambridge University Press.

Bourdieu, P. (1990). *The logic of practice* (translated by Richard Nice), Stanford, CA: Stanford University Press.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge Processes, *Psychological Review*, 67(6), 380-400.

Campbell, D. T. (1974). Evolutionary epistemology, in P.A. Schilipp (ed.), *The Philosophy of Karl Popper*, 14, 413-463. La Salle, IL.: Open Court.

Cantillon, R. (1775). *An Essay on Economic Theory*, English translation by C. Saucier and edited by M. Thornton in 2010, Auburn, Alabama: Ludvig von Mises Institute.

Casson, M. (2014).*Entrepreneurship: A personal view*. International Journal of the Economics of Business, 21 (1), 7-12.

Chen, B. S. (2013). "Challenging Apple's Cool: As a phone rival Samsung takes a different track, *" New York Times,* February 11, 2013, p. B1.

Chesbrough, H. W. (2003). *Open Innovation*. Boston: Harvard University Press.

Cook, S. D., & Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science*, *10*(4), 381-400.

Coombs, R., & Metcalfe, J.S. (2002)."Organizing for innovation: Co-coordinating distributed innovation capabilities," Foss, N. and Mahnke, V. (ed). *Competence, Governance, and Entrepreneurship:* Advances in Economic Strategy Research, Oxford: Oxford University Press, pp. 209-231.

The Economist (1912). A third industrial revolution, *The Economist*, April 21, 2012, p. 15.

Fantle, J. (2008). Knowing-how and knowing that, *Philosophy Compass*, 3(3), 451-470.

Fantl, J. and McGrath, M. (2009). *Knowledge in an Uncertain World*, Oxford, England: Oxford University Press.

Foss, N. (1999). Networks, capabilities, and competitive advantage, *Scandinavian Journal of Management,* 15 (1), 1-15.

Foss, N.J. and Mahnke, V.(2000). *Competence, Governance, and Entrepreneurship*. Oxford University Press, Oxford, U.K.

Foss, K., N.J. Foss and P.G. Klein.(2007). Original and derived judgment: An entrepreneurial theory of economic organization, *Organization Studies*, 28(12), 1893-1912.

Foss, N. J. and Klein, P. G. (2010). Alertness, action and the antecedents of entrepreneurship, The *The Journal of Private Enterprise,* 25 (2), 145-164.

Fuchs, C. (2003). "Structuration theory and self-organization," *Systemic Practice and Action Research*, 16 (2), pp. 133-167.

Giddens, A., (1979). *Central problems in social theory*, London: Macmillan/Berkeley: University of California Press.

Giddens, A. (1984). *The Constitution of society,* Berkeley and Los Angeles: University of California Press.

Hannan, M. T. and Freeman J.,(1984).“Structural inertia and organizational change,” *American Sociological Review*, 49 (2), pp. 149-164.

Hartland-Swan, J. (1956) The Logical status of knowing that. *Analysis* 16(5), 111-115.

Hartland-Swan, J. (1957) Knowing involves deciding. *Philosophy* 32 (120) 39-57.

Hayek, F. A. (1937). Economics and knowledge, *Economica,* 4, 33-54.

Hayek, F. A. (1945). The use of knowledge in society, *The American Economic Review*, 35(4), 519-530.

Helfat, C. E., Finkelstein, S., Mitchell, W., Peteraf, M. A., Singh, H., Teece, D. J., & Winter, S. G. (2007). Dynamic capabilities: Understanding Strategic Change in Organizations. Malden, MA: Blackwell Publishing.

Hislop D (2009) *Knowledge Management*. Oxford University Press, Oxford.

Holcombe, R. G. (2003). The origin of entrepreneurial opportunities, *The Review of Austrian Economics,* 16(1), 25-43.

Hosinski, T.E. (1993) Stubborn Fact and Creative Advance: An Introduction to the Metaphysics of Alfred North Whitehead.Landham: Rowman& Littlefield publisher.

Isaacson, W. (2011).*Steve Jobs*, New York: Simon & Schuster.

Kirzner, I. M. (1973).*Competition and Entrepreneurship,* Chicago: University of Chicago Press.

Kirzner, I. M. (1984).Economic planning and the knowledge problem*, Cato Journal*, 4(2), 407-418.

Kirzner, I. M. (1997). Entrepreneurial discovery and the competitive market process: An Austrian Approach, *Journal of Economic Literature*, 35(1), 60-85.

Kirzner, I. M. (1999). Creativity and /or alertness: A reconsideration of the Schumpeterian entrepreneur, The Review of Austrian Economics, 11 (12), 5-17.

Kirzner, I, M. (2000).Creativity and /or alertness: A reconsideration of the Schumpeterian entrepreneur, The Review of Austrian Economics, 11 (12), 5-17.

Kirzner, I. M. (2008). The alert and creative entrepreneur: A clarification, *IFN Working Paper* No. 760.

Kimball, R. H. (1999). “Error in causal efficacy,”*Process Studies*, 38 (1-2), pp. 56-67.

Knight, F. H. (1921).*Risk, Uncertainty and Profit,* Boston, MA: Houghton Mifflin Co.

Kogut, B., & Zander, U. (1992).Knowledge of the firm, capabilities, and the replication of technology. Organization Science, 3(3), 383-397.

Kuhn, T. (1962) *The Structure of Scientific Revolutions.* Chicago: The University of Chicago Press.

Landstrom, H., Harirchi, G. &Astrom, F. (2012). Entrepreneurship: Exploring the

knowledge base. *Research Policy*, 41(2012), 1154-1181.

Lam, A. (2000). Tacit knowledge, organizational learning and societal institutions: An integrated framework. *Organization Studies, 21*(3), pp.487-513.

Lnanglois, R. N. (1998). Personal capitalism as charismatic authority: the organizational economics of a Weberian concept, *Industrial and Corporate Change* 7(1), 195-213.

Langlois, R. N. (2005). The entrepreneurial theory of the firm and the theory of the entrepreneurial firm, *Economics Working Papers*. Paper 200527.

Levinthal, D.A. (1997). "Adaptation on rugged landscapes,"*Management Science,* 43(7), pp.934-950.

Levinthal, D.A. (2006). "The Neo-Schumpeterian theory of the firm and the strategy field," *Industrial and Corporate Change,* 15(2), 391-394.

Madhok, A., (2002). "Inter-firm collaboration: Contractual and capabilities-based perspectives," Foss, N. and Mahnke, V. (ed), *Competence, Governance and Entrepreneurship: Advances in Economic Strategy Research,* Oxford: Oxford University Press, pp. 276-303.

Mathews, J. A. (2006). A strategic and evolutionary perspective on entrepreneurial dynamics Reconciling Schumpeter with Kirzner, *MGSM working paper*.

Mises, von L. (1949). *Human Action*: *A Treatise on Economics*, New Haven: Yale University Press.

Mitchell, M.T. (2006). *Michael Polanyi: The art of knowing*, Wilmington, Delaware: ISI Books.

Mokyr, J. (2000). Knowledge, technology, Economic growth during the industrial revolution.In B. van Arts, S. Kuipers, & G. Kuper (Eds.), Productivity, Technology and Economic Growth (pp. 253-292). Boston, MA: Kluwer Academic Publishers.

Murphy, P. J. and Marvel, M. R. (2007). The opportunity-based approach to entrepreneurial discovery research, http://works.bpress.com/profpjm/11

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation.*Organization Science*,5 (1), 14-37.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company*. Oxford, UK: Oxford University Press.

Nonaka, I., &Kono, N. (1998). The concept of "Ba": Building foundation for knowledge creation, *California Management Review*, 40(3), 40-54.

Nonaka, I., Toyama, R., &Kono, N. (2000a). SECI. Ba and leadership: A unified model of dynamic knowledge creation. *Long Range Planning*, 33 (2000), 5-34.

Nonaka, I., Toyama, R., & Nagata, A. (2000b). A firm as a knowledge-creating entity: A new perspective on the theory of the firm. *Industrial and Corporate Change*, 9(1), 1-20.

Nonaka, I., & Toyama, R. (2002). A firm as a dialectical being: towards a dynamic theory of a firm. *Industrial and Corporate Change*, 11(5), 995-1009.

Nonaka, I.&Toyoma, R. (2003).The knowledge-creating theory revisited: Knowledge creation as a synthesizing process.Knowledge Management Research & Practice, 1(1), 2-10.

Nonaka, I., von Krogh, G. &Voelpel, S. (2006). Organizational knowledge creation theory: Evolutionary paths and future advances. *Organization Studies*, 27(8), 1179-1298.

Orr, J. E. (1996). *Talking about Machines: Ethnography of a Modern Job*. Ithaca, NY: Cornell University Press.

Park, Hong Y., Chang, Hyejung& Park, Yong-Seung, (2014a).Firm's knowledge creation structure and new product development, *IFKAD Conference on Knowledge and Management Model for Sustainable Growth,* Matera, Italy, June 11-13, 2014.

Park, Hong Y., Cho, I., Jung, S.,& Main, D. (2014b). Information technology and user knowledge-driven innovation in services, presented at *the 14th International Conference on Knowledge, Culture and Change in Organizations,* August 4-5, 2014, Said Business School, Oxford University, London, U.K.

Polanyi, M. (1962). *Personal Knowledge*. Chicago: The University of Chicago Press

Polanyi, M. (1966). *The Tacit Dimension*. New York, NY: Anchor Day Books.

Polanyi, M. (1969). *Knowing and Being*. (ed) by Grene M. Chicago: The University of Chicago Press.

Popper, K. (1969). Conjectures and Refutation: The Growth of Scientific Knowledge (3rd ed.), London, England: Routlege& Kegan Paul.

Popper, K. (1982). Of clouds and clocks in *Learning, Development, and Culture* (ed) by loykin, H. C., New York: John Wiley & Sons, 109-119.

Prigogine, I. (2003), *Is Future Given?* London: World Scientific. P. 39.

Riley, J. G. (1975). "Competitive signaling," *Journal of Economic Theory*, 10 (2), 174-186.

Roland, J. G. (1958). On "knowing how" and "knowing that," *The Philosophical Review*, 67(3), 379-388.

Ryle, G. (1946). Knowing how and knowing that, *Proceedings of the Aristotelian Society,* XLVI, vol. 2. 212-225.

Ryle, G. (1949). *The Concept of Mind,* London, U.K.: Hutchinson & Co. LTD.

Sautet, F. E. (2000). *An Entrepreneurial Theory of the Firm*, London: Rutledge.

Shane, S. (2000). Prior knowledge and discovery of entrepreneurial opportunities, *Organization Science*, 11(4), 448-469.

Shane, S. and Venkataraman, S. (2000).The promise of entrepreneurship as a field of research*, Academy of Management Review*, 25 (1), 217-226.

Shane, S. (2003). A General Theory of Entrepreneurship: The Individual-Opportunity Nexus, Cheltenham, U.K., Edward Elgar Publishing, Inc.

Shaviro, S. (2009). Without Criteria: Kant, Whitehead, Deleuze, and Aesthetics, Cambridge, MA: The MIT Press.

Schiuma, G. (2009). The managerial foundations of knowledge assets dynamics, *Knowledge Management Research &Practicee*, 7(4), 290-299.

Schumpeter, J. A. (1934). *The Theory of Economic Development*, New York: Oxford University Press.

Schumpeter, J. A. (1942). *Capitalism, Socialism and Democracy*, New York: Harper & Brothers.

Schumpeter, J. A. (1947). The creative response in economic history, *Journal of Economic History*, 7 (2), 149-159.

Schumpeter, J. A. (2008). *Essays on Entrepreneurs, Innovations, Business Cycles, and the Evolution of Capitalism,* (ed.) Richard V. Clemence, New Brunswick: Transaction Publishers.

Spence, A. M. (1973). "Job market signaling," *The Quarterly Journal of Economics*, 87(3), 355-374.

Stanley, J. and Wilson, T. (2001). Knowing how, *Journal of Philosophy*, 98(8), 441-444.

Stigler, G. J. (1961). "The economics of information," *The Journal of Political Economy*, 69 (3), 213-225.

Stiglitz, J. E. (1975). "Incentive, risk and information: Notes towards a theory of hierarchy," *The Bell Journal of Economics,* 58(2), 531-537.

Teece, DJ (2007) Explicating dynamic capabilities: the nature and micro foundations of (sustainable) enterprise performance. *Strategic Management Journal* 28 (13), 1319-1350.

Venkataraman, S. (1997). The distinctive domain of entrepreneurship research: An editor's perspective. In *Advances in Entrepreneurship, Firm Emergence, and Growth*, Katz, J. and Brockhaus R. (ed.). Greenwich, CT: JAI Press.

Venkataraman, S. (2003). Foreword in Shane, S. (2003). *A General Theory of Entrepreneurship: The Individual-Opportunity Nexus,* Cheltenham, U.K., Edward Elgar Publishing, Inc.

von Hippel E (1988). *The Sources of Innovation*. MIT Press, Cambridge.

Yu, T. F. (2001). Entrepreneurial alertness and discovery, *The Review of Austrian Economics*, 14 (1), 47-63.

Weick, K. E. (ed.). (1977). *Enactment Process in Organizations*, Chicago: St. Clair Press.

Weick, K. E. (1979). *The Social Psychology in Organizing*, Reading, MA: Addison-Wesley.

Whitehead, A. N. (1929). Process and reality. New York: Free Press.

Whitehead, A. N. (1933). Foresight, Chapter VI in the *Adventures of Ideas,* New York: The Free Press, 87-99.

Whiteman, M. (2013). Collaboration to develop new business ideas, Dow Chemical Company, Midland, Michigan.

Williamson O. E. (1991).“Strategizing, economizing, and economic organization,” *Strategic Management Journal,* 12 (s2),75-94

Williamson, O.E. (1999). “Strategy research: governance and competence perspectives*,” Strategic Management Journal,* 20(12), 1087-1108.

Winter, S. G. (2006). “Toward a neo-Schumpeterian theory of the firm,”*Industrial and Corporate Change*,” 15(1), 125-141.

CHAPTER 8

Canary in a Coal Mine – Analysis of Systemic Risk

by

Gabjin Oh[*]

(Chosun University)

Hyeongsop Shim[**]

(Ulsan National Institute of Science and Technology)

Yong-Cheol Kim[***]

(University of Wisconsin-Milwaukee)

Abstract

We develop a measure of systemic risk based on symbolic transfer entropy (STE) in a network of 48 industries, both financial and real sectors, by incorporating the strength and asymmetry of information flows. Time variation of systemic risk in the United States shows that from 2001 systemic risk start to rise, peaking in 2004, and accelerating until the financial crisis of 2007/8. Year 2004 coincides with several financial events that potentially contribute to the crisis: sudden acceleration of subprime mortgage, financialzation of commodity, CDS activities, and others. Furthermore, we document that not only the economy-wide systemic risk of US economy has started to build up from 2004, but also the asymmetry of information flows between

* Division of Business Administration, College of Business, Chosun University, Gwangju 501-759, Republic of Korea.

** School of Technology Management, Ulsan National Institute of Science and Technology, Ulsan 689-798, Republic of Korea. Author's Contact Information: Kim: ykim@uwm.edu, (414) 229-2483; Oh: phecogjoh@chosun.ac.kr, 82-62-230-6811; Shim: hshim@unist.ac.kr, 82-52-217-3132;

*** Corresponding author: Lubar School of Business, University of Wisconsin-Milwaukee, PO Box 742, Milwaukee, Wisconsin 53201-0742, USA.

financial and real sector accelerated from the year 2004 and grow continuously until it reaches a peak in 2007. In addition, we find that systemic risks are closely linked to a battery of macro-economic variables, and show our systemic risk measure is robust with unemployment, treasury rate, return and volatility of stock index among other macro-economic variables. Systemic risk has predictive power for future stock returns, and it can serve as a warning signal for the future. Interestingly, the contemporary systemic risk remains at high plateau similar to the financial crisis period of 2007/2008.

1. Introduction

The financial crisis of 2007/2008 caught virtually everyone by surprise, and prompted a flurry of academic research on why and how it happened, and evoked policy and regulatory discussions what to do to minimize the impact on the economy as well as how to prevent it to happen again. U.S. Federal reserve adopted unusual quantitative easing monetary policy of bond purchase. Our motivation and contribution of this paper is to analyze the whole process of financial crisis, from slow and consistent build-up of systemic risk culminating to the crisis, and develop a warning signal for a potential future crisis that might occur similar to the 2007/2008 crisis. Useful measure and warning signal should be broad enough as concrete and detailed events and activities are inherently unpredictable as financial institutions pursue highly profitable but at the same highly risky projects and business activities. At the same time, a measure of systemic risk is useful when it captures as much as concrete signals using available data and information. In this paper we develop a numerical systemic risk measure of U.S. economy by analyzing the network of stock returns of 48 industries to capture economy-wide systemic risk.

Academic research has taken several different approaches to post-diagnose the financial crisis.[1] Among them are the analysis of sub-prime

1 There are vast and growing literatures related to the financial crisis of 2007/2008 and the aftermath. Here we sample only few articles to show the nature of articles that it is a partial analysis of the whole elephant of financial crisis.

mortgages (Demyank and Hemert, 2011), downward spiral effect on the balance sheet of financial intermediaries triggered by sudden asset price decline (Brunnermeier et al., 2009), and credit, loan and liquidity channel from monetary policies to banks and firms. More broad approach to explain the crisis is the idea of collective moral hazard (Farhi, and Tirole, 2012). As the crisis is characterized by too-big-to-fail and too-interconnected-to-fail, researches have taken various statistical approaches, such as conditional value at risk (CoVaR) (Adrian and Brunnermeier, 2010), and others. Billio and et al. (2012) conduct detailed analysis of interconnections of four financial institutions. We take a view that the systemic risk is most meaningful when it is considered as a risk of the whole economy. Our major contribution is to develop a measure of systemic risk based on a network of industries in the whole economy as well as systemic risk and contribution of each industry and sector to the economy-wide systemic risk. We show that the systemic risk based on parsimonious analysis of network of industries can explain the whole process of financial crisis, identify the timing of the beginning of the crisis and the building up processes, and, as such, serve as a warning signal for the future.

Financial crisis in 2007/2008 highlights the vulnerability of real economy to financial sector meltdown, ultimately leading to the biggest and painful recession after the Great Depression in 1929. and creates keen interests of why, how and what happened to the financial sector and the whole economy, As a result of sudden unexpected financial crisis, the issue of how to prevent another potential crisis and crash in the future caught immediate interests for academics, practitioners, regulators and policymakers. Policy makers in the U.S. adopted one of the most stringent regulations in 2010, adopting Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010. Furthermore, regulators try to mend several weaknesses that are considered to be responsible for the crisis. Requiring CDS to be traded through the Clearing House is one example. For academics, plethora of research, both theoretical and empirical, on different aspects of financial crisis added to the better understanding of the causes, processes of the current and past crises. Those researches have implications on several preventive and preemptive measures to forestall potential future crisis.

One of key issues in the prevention of future crisis is the measurement problem.[2] The relevant questions are as follows: what is the symptom that potentially leads to the crisis? Are there any measures that signals potential crisis? How to measure systemic risk?

While the idea of systemic risk is considered to be important and relevant in understanding the current crisis and the prevention of potential future crises, there is no clear consensus on the concept and measurement of systemic risk. For example, Billio et al. (2012) takes a viewpoint that more formal definition (of systemic risk) is any set of circumstances that threatens the stability of public confidence in the financial system, and analyze monthly returns of four types of financial institutions: hedge funds, and publicly traded banks, broker/dealers, and insurance companies. Lehar (2005) proposes to measure systemic risk, defined as the probability of a given number of simultaneous bank defaults, from equity return data. Works of Avesani et al. (2006) and Basurto and Padilla (2006), among others, are examples of stress testing exercises on the financial sector using market –based information.

In this paper, we take much broader view of systemic risk by including whole economic system that includes financial sectors and real sectors in one network framework. We define systemic risk as the risk of collapse of an entire economy and financial system.[3] The challenge to implement this broader concept of systemic risk operationally is how to incorporate wide range of economic sectors that are inter-linked and interdependent, and how to identify and measure complex web of connections among firms, industrial and financial firms. Key components in the measurement of systemic risk are: connectedness, strength and direction of causality and information flows from one node to another in a network, degree of concentration of information

2 Bisias et al. (2012), in their survey article, present taxonomies of systemic risk measures by various criteria: by data requirements, by supervisory scope, by event/decision time horizon, and by research method.

3 Crisis and the collapse of an entire economy and financial system can happen for various routes. One possibility is that troubles and difficulties on one part of the system transmit and cascade to other sectors rapidly (domino effect or cascading effect). In this case, the initial cause of problems might occur internally or external to the system. another possibility is like a "Tsunami" where outside forces destabilize the system.

originations in a particular node or nodes, where nodes represent, in this paper, 48 industry group classified by Fama and French (1997). In addition, we capture contributions of the industry and sectors to total systemic risk as wells as the economy-wide systemic risk. Billio et al. (2012) empirically estimate the network structure of financial institutions generated by stock-return interconnections, by simply measuring correlation directly and unconditionally using principal components analysis and by pairwise Granger-causality tests. They use these metrics to gauge the degree of connectedness of the financial system. We propose symbolic transfer entropy (STE) to capture not only connectedness, but also the direction and strength of causality and information flows. The advantage of our systemic risk is that the STE captures the strength as well as asymmetry of information flows. Furthermore, our systemic risk measure are versatile, i.e. we can measure systemic risk for one industry, for example banking industry, or we can measure systemic risk for financial sector comprising of three industry group of banking, insurance, and trading, or the systemic risk of the whole economy. Most importantly, as we use current market information such as daily stock return, we can measure up-to-date magnitude of systemic risk in almost real time at a reasonable computer time and costs.[4]

One of most striking results from our empirical analysis is that our measure not only identifies the current crisis and other past crises in 1987, and 1997/1998 Asian and Russian crisis, but also we find that systemic risk started to increase continuously from 2001, reaching a local peak in 2004. After 2004, systemic risk remains at high level and continues to increase to a global peak in our sample when the financial crisis in 2007/2008 erupts. In addition, we find that systemic risk obtained from the analysis of network of industries is closely linked to macroeconomic measures. We use a battery of macro-economic variables, and show our systemic risk measure is robust with unemployment, treasury rate, stock return and volatility of stock among other macro-economic variables.

4 In this paper, we focus on three financial industry groups. However, with further calibration, we can apply symbolic transfer entropy (STE), and measure systemic risk by any level of aggregation. For example, we can calibrate SR measure, at a micro level, to a particular institution.

Network analysis and literature review on systemic risk and measurement is summarized in section 2. The measurement of systemic risk based on symbolic transfer entropy (STE) is described and discussed in section 3. Section 4 describes data and present empirical results in both graphs and tables. Conclusion follows in section 5.

2. Network Analysis and Literature Review

Financial crisis of 2007/2008 highlighted the fragility of financial institutions and the systemic risk originating from financial institutions in the global economy. Academics produced plethora of literature on financial crisis and systemic risk after the financial crisis of 2007/8. In this section, we review the findings and various approaches of economic and financial research to lay economic foundations to our measure of systemic risk. Crisis, by its nature, affects broad spectrum of economic agents and activities, and, as such, the causes and processes how the crisis evolves can be better captured by analyzing the network and inter-linkage of economic agents and activities.

One strand of research is to analyze physical and financial linkage and relationship of individual firms and financial institutions, and show that inter-linkage drives chain and domino effects to other counter-parties by empirically analyzing the counterparty exposure in a network of banks with banks, banks with firms and firms with firms. For example, De Bandt and Hartmann (2000) delineate the systemic events into horizontal and vertical systematic events. In horizontal events, the bad news or even failure of a financial institution bring about adverse effects on one or several other financial institutions in a sequential way like domino.[5] In a vertical systemic event, the failure of financial sector should adversely affect real sectors. The transmission of shocks could take a form of credit crunch or debt deflation. Acharya and Naqvi (2012) show that banking sector ignites asset price bubbles, and create potential seeds of crisis. Gan (2007) investigates Japanese land price and loan- and firm-level of lending channel. She discusses the linkages and

5 See Brunnermeier (2009) to recall the chronological order of financial contagion.

transmission mechanism from 1) asset price decline (land price in Japan) - constrain bank with mortgage assets - less lending to firms, and 2) show that firms rely on bank credit (no other substitutes) exercise less investment and lower valuation.

Second strand of research is to investigate the effect of monetary policies on the financial institutions and real economy. Bernanke and Gertler (1995) look at the credit channel to see how changes in monetary policy affect bank lending activities. Ashcraft (2006) find evidence consistent with a bank lending channel of monetary policy, where the effect of monetary policy on bank lending is amplified by the inability of some banks to replace an outflow of insured deposits with large certificates of deposit (CDs) and federal funds. In other words, there is a mismatch of loan growth on asset side (liquid assets), and the liability side of balance sheet (deposits, commercial papers and other external funds, large CDs and federal funds).[6]

Third strand is to look at the liquidity channel at macro and aggregate level. One of stylized fact of the market liquidity is the commonality in securities' liquidity. Among others, Chordia et al. (2000) and Hasbrouck and Seppi (2001) find evidence of commonality in liquidity of traded securities. Extant research provides empirical evidence of the source of the commonality. One strand of literature (e.g. Fujimoto, 2004) investigates macroeconomic factors such as the Federal fund rates to explain the commonality. Fujimoto finds liquidity declines with higher Federal fund rates, tighter monetary policies, and in recessions. Other strand of literature investigates both macroeconomic and microeconomic factors to explain commonality. Chordia et al. (2001) analyze macroeconomic impact on market liquidity and they find that the changes in liquidity are affected by volatility, market return, and macroeconomic announcements. Chordia et al. (2003) study the link between 'macro' liquidity and 'micro' liquidity by exploring the co-movements in the stock and the Treasury market's liquidity. In their

6 Ashcraft (2006) states, by focusing on loan supply and demand, that "It follows that it may be hard to distinguish a differential shift in loan supply across bank leverage (the lending channel) from a differential shift in loan demand across firm creditworthiness(the balances sheet channel) when low capital banks concentrate their lending with less creditworthy firms.". (page 756)

studies, the 'macro' liquidity or 'money flows' is proxied by the bank reserves, the Fed fund rates, and mutual fund investments.

Adrian et al. (2009) and Adrian and Shin (2010) document that balance sheet structure of market-based financial intermediaries is more critical in explaining the liquidity channel and spiral than other institutions. Capital market liquidity is increasingly dominated by market-based intermediaries, and market-based assets are substantially larger than bank assets. Furthermore, they document pro-cyclical growth of leverage, i.e. the growth of assets and leverage of financial intermediaries move in the same direction and the same rate, and as a result, equity ratio also grows at a constant rate over time. The co-movement of assets and liabilities imply the unique linkage and importance of financial intermediaries in the liquidity channel of the financial market.

This research on liquidity channel suggests the network structure of financial intermediaries and the monetary policy variables. In other words, the layouts of liquidity channel are related to the commonality and contagion channel of economic transmission of a shock. For example, monetary policy variable is an external factor for a subsystem of financial intermediaries, while it is an internal factor in the whole economy when interest rate policy and the liquidity of financial intermediaries are Granger-causal relationship in both ways.

In a similar context, Brunnermeier et al. (2009) discuss economic interpretations of the process where assets and liability of financial institutions show feedback and downward spiral effect when asset market crash or sudden dry-up of funding. One unanswered question in this liquidity spiral is that we still don't know which one, asset market crash or funding liquidity shortage is a primary source of events. Our analysis with network approach can give a clue in distinguishing commonality and contagion, and better understanding of the liquidity spiral and asset market decline.

Fourth strand is to discuss risk taking and moral hazard of top management to explain the crisis. Farhi and Tirole (2012) and Kim et al. (2013) show the collective moral hazard as one of prime reasons for the financial crisis of 2007/2008. Research on risk-taking behavior of bank holding companies show CDS and other financial derivatives are a

medium that accentuates and prorogates risks across a broad range of financial institutions.

While multiple aspects of network of economic agents are considered in explaining the bubble and crisis, another important question is the measurement of systemic risk and its predictability. In regard to the complexity of network of agents and events, we can broadly classify into two approaches to understand the issue of systemic risk and macro-prudential regulation: empirical macroeconomics based and financial market based approach. These two approaches have different methodologies, emphases, and purposes. Macroeconomics base approach focuses on the impact of systemic events on real economy while financial sector is considered as amplification mechanism.[7] Financial market based approach focuses on the financial market structure with background information of macro economy environments. This approach puts more emphasis on the interaction among financial institutions, the nonlinear feedback effect, and the identification of individual institutions that are systemically important. Better understanding of systemic risk for the whole economy should incorporate both approaches.

Three popular cross sectional measures of systemic risk are conditional value at risk (CoVaR) by Adrian and Brunnermeier (2010), distressed insurance premium (DIP) by Huang, Zhou, and Zhu (2011), and systemic expected shortfall (SES) by Acharya et al. (2012). These measures aim at estimating the magnitude of losses when many financial institutions simultaneously fall into difficulty. CoVaR computes the value-at-risk (VaR) of financial institutions under the condition that other institution is in financial distress. DIP computes required insurance premium to cover the losses arising from distressed banking system. Acharya et al. (2012) discuss each financial institution's contribution to systemic risk can be measured as its systemic expected shortfall (SES), i.e., its propensity to be undercapitalized when the system as a whole is undercapitalized. SES measures the expected loss to each financial institution under the poor performance of entire set of financial

7 See De Nicolo and Lucchetta (2010). Macroeconomics base approach helps us to better understand the fundamental linkage between the real economy and the financial sector, especially in the long run.

institutions. Acharya et al. (2012) show the ability of financial firm's marginal expected shortfall, MES (i.e., its losses in the tail of the aggregate sector's loss distribution) to forecast systemic risk. Bisias et al. (2012) show, in their survey article, taxonomy of methodologies applying to systemic risk.

While these measures have some merits in the context of its own problem set, we are interested in the comprehensive and economy-wide systemic risk that incorporates all sectors and industry, financial and non-financial. Furthermore, the network of agents that includes financial institutions, firms, and investors, monetary policy makers, regulators, and others are linked in a much more complex way than can be identified by their actual financial relationships. Furthermore, we are interested in the building-up process leading to the financial crisis as well as the aftermath of the crisis, and more detailed analysis of the contribution of each industry to the systemic risk.

Crisis by definition is a surprise and an event that occurs totally unexpected with wide impact for the society. However, crisis does not happen in isolation, and there are series of events leading up to the crisis, and like bubble, could have long dormant pre-crisis period. While the exact timing when the current crisis has started to brew is not clear, events leading up to current sub-prime mortgage crisis is likely to have started around 2003/4 when the volume of sub-prime loans jumped dramatically. The circumstances and the sequence of events in the mortgage borrowings and risk transformation process by financial institutions can be summarized as follows: in a macro-economic environment where market interest rate is kept at low level through central bank monetary policy and rising real estate prices, mortgage borrowers have easy access to loans as banks are lenient in lending, and encourage sub-prime borrowers to borrow. Banks in turn create mortgage-backed securities (MBS) and sell risky loan portfolio to other financial institutions (CDOs) through special purpose vehicle (SPV). SPV is a subsidiary outside of regulatory supervision. CDOs slice up mortgage portfolio and create reformulated mortgage portfolio in tranches. Each tranche represent a collection of mortgage portfolio, and the risk of each tranche is reduced partly through diversification. Furthermore, CDS and rating agency make sliced-up mortgage portfolio

more credit worthy, and it is sold at a higher price. Final portfolio of mortgages of risk-transformed mortgage portfolio attached with CDSs is held by various financial institutions like hedge fund. The credit risk of risky borrower is not borne by originating banks, and the more risky the initial loans are, there are more gains for banks from risk transformation. Banks pursue more risky loans through originators as the risk transformation and credit enhancements boost bank profitability. In the sequence of risk transformation, each party participating in the process seems to act rationally following their own incentives. When we understand the process of events leading up to the crisis, it is clear that it is important to see the big picture and the analysis of crisis have to consider the context and the activities of each party as a part of whole.

The contribution of our paper is to adopt network approach and analysis that covers whole economy and industries, and to produce economically meaningful interpretation and implications, and to provide timely warning signal to economic agents, a canary in a coal mine. Closely related work to ours is Billio et al. (2012). Billio et al. (2012) propose connection based measures of systemic risk in a network of 4 distinct types of financial institutions: bank, insurance, trading companies, and hedge funds. They propose measures using well established econometric approaches: principal component analysis and pairwise Granger-causality tests. Their main measure of systemic risk is the interconnectedness of nodes based on the idea that linkages among financial institutions are the key element of systemic risk. Their work is based on the theoretical development which argues that financial crisis is more likely when the degree of correlation among the holdings of financial institutions is higher. They measure the connection between two financial institutions with Granger-causality of monthly stock returns. Their connection measure is of binary value (0 or 1) by applying the confidence level cut-off to Granger-causality tests.

Our motivations of network study is 1) to measure systemic risk for the whole economy as well as individual industry, 2) to identify economically meaning changes in the systemic risk over time, 3) to separate industries that play a significant role in the economy, and 4) to compare a network including external variables such as interest rates to investigate the role and influence of monetary policy on the economy-

wide systemic risk. Reflecting the complexity of inter-linkage of economic agents and business relations, we adopt symbolic transfer entropy (STE) methods. Billio et al. (2012) considers network structure of four financial institutions, while our network covers 48 industries. In addition, we can identify source or sink, by measuring the strength and direction of information flows between and among each industry, together with the measurement of aggregate systemic risk.

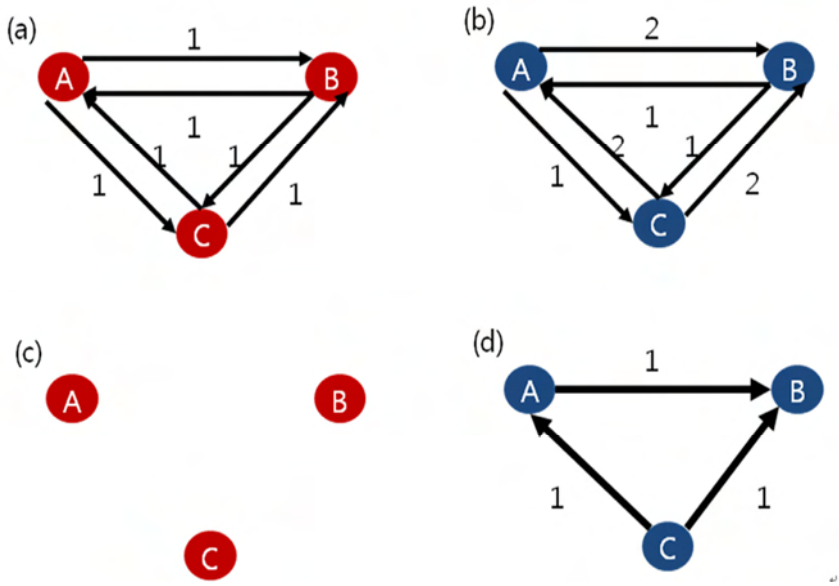
We take a network analysis that considers not only the interconnectedness, but also the strength of directed connectedness (SDC) and the asymmetry of SDC (ASDC) between all nodes (industries) in a system, and develop a measure of systemic risk utilizing all industrial sectors which include financial industries as well as industries in real sector. In contrast, Billio et al. (2012) use correlation measure of principal components analysis (PCA) and a binary connection from Granger causality test. Our measure of symbolic transfer entropy (STE) can capture the strength of directed information flows from one industry to all other industries, and direction of information flows.

3. Measurement of Systemic Risk

In this section we develop and elaborate the idea how to measure systemic risk, and the components that are critical to the systemic risk measure, and economic relevance of those concepts. We use examples of networks of three nodes to illustrate the essential nature of SDC and ASDC, as well as the comparison of our methodology to the measures discussed in Billio et al. (2012) and others.

Panel (a) of figure 1 depicts a neutral network, and shows that information flows among three nodes (industry sectors in our study) have equal strength. It delineates a network of symmetric information flow which could be observed in the binary Granger-causality. The number in the diagram is binary value (0, 1) from linear Granger causality tests, where 1 indicates the existence of Granger causality. It shows that there are symmetric information flows in both directions for all nodes, and the net flow of information is all zero in all three connections.

Figure 1 Network structure of symmetric information flows and asymmetric information flows.



In contrast, Billio et al. (2012) treats this network as most unstable based on the number of connectedness which is 6. The number of connections including the directions of information flow which captures the systemic importance of single node is 4 (2 inflows and 2 outflows). Now, let's reinterpret the number as a measure of strength of directed connectedness (SDC) and the network structure changes to panel (b). Under the measure that considers only interconnectedness based on the Granger causality test, panel (a) and panel (b) are not different, as the number of connectedness remains the same.

In contrast, our measure of systemic risk uses both the strength of directed connectedness (SDC) and the asymmetry of SDC, net strength of inward and outward information flows. Now consider the network in panel (a). This type of interconnection among nodes is equivalent to the diagram (c), as if there are absolutely no interconnections, making the system neutral. We call this network "symmetric network", and the

symmetric network structure is still an unstable network, as the total SDC is 6. Note that in this case there is no source or sink that disturbs the network structure, where source is the node from which information originates, and the sink is the node that receives information from other nodes. Our measure of systemic risk considers both SDC and the ASDC in multiplicative fashion.[8] Suppose, in a network of panel (a), that the strength of all information flow increases to 2 from 1. In this case the net flow is still zero, i.e. ASDC=0. However, the interconnectedness in our measure increases from 6 to 12, as SDC is the sum of strength of connectedness. Even though the ASDC does not change from 0, our systemic risk increases as our measure incorporates both SDC and ASDC. Real world relevancy of network structure could be the case of banking network. Popular view of the instability of financial industry is summarized as "too-interconnected to fail" and "too-big to fail". In out context, SDC is a proxy for the degree of interconnectedness, and ASDC is a proxy for the degree of concentration (too-big) of financial institution or economic sector. We posit that the combination of the interconnectedness (sum of SDC) and the ASDC contributes to the systemic risk.

Another economic example is as follows. Suppose banks use derivatives to manage their primary loan portfolio using derivatives with other banks. In this case, SDC increases even if there are no new connections, as the strength of information flows with existing connections increases. On the other hand, if the number of connectedness increases and at the same time information flows become more asymmetric, our measure of systemic risk definitely increases. When the connectedness among nodes (industry sectors) increases, i.e. higher SDC, and at the same time ASDC decreases (i.e. strength of information flows becomes similar), systemic risk can decrease. Oppositely, when

8 Billio et al. (2012) measure systemic risk based on the number of connections and the degree of Granger causality. Billio et al. (2012), in their Granger causality test, use the sum of in- and out-flow, and interpret the sum as a partial measure of systemic risk. On the other hand, net flows, net of inflows and out flows of information can better capture the origin or source of information flow and the asymmetry of information flow.

both the strength of directed connectedness (SDC) and the degree of asymmetry (ASDC) increase, systemic risk is going to increase.

The SDC and ASDC can be used to interpret the studies on collateral channel between banks and firms (Gan, 2007). When asset price declines sharply, firm's financial status when firms use real estate as collateral worsens, creating more information (higher strength) emanating from firms to banks, and the asymmetry of SDC. On the other hand, when banks hold mortgage loans, banks' holdings of mortgage become problem loans, and in turn, firms that rely on banks reduce investments. The results of Gan (2007), in the context of systemic risk, can be interpreted as the increasing strength of information (SDC) from asset market crash, and the increase in the asymmetry of information flows between banks and firms, higher ASDC, that creates more instability in a network between banks and firms. Panel (b) network structure corresponds to this case.

Panel (b) shows an example where the number of connections is the same as panel (a), but there exists asymmetry of information flows. When we capture net information flows, then panel (b) can be collapsed to panel (d) network. In addition, the network (d) shows the origin and the source of information flows. In this example, node "C" is indisputable "source" of information flows, while node "B" is the recipient of information from both nodes, and we call a node that has net negative information flow as "sink". In this context, Node "A" is a sink with node C, but a source with respect to node B. In addition to the local source and sink based on net information flows between two nodes, we are also interested in the global source and sink, where global source is defined as the sum of net information flows originating from one node to all the remaining nodes, and global sink is defined as the sum of net information flows from all other nodes to one particular node. In our example, node "C" is global source and the sum of ASDC is positive 2, while node "B" is global sink and the sum of ASDC is negative 2. The sum of ASDC is considered as a measure of the importance of the node in the network structure. Node "C" has a node that destabilizes network structure with a force 2, i.e. node "C" is an important node in this network. On the other hand, node "B" is global sink with negative

global ASDC. Node "B" is, in the global context, a sink, but it still contributes to the instability of network by transmitting information to node "A" and "C". We design our systemic risk measure to capture global information flows as well as this local contagion effect by adjusting the weights of local factor. One extreme case is when a node is a perfect sink, i.e. it receives information from other nodes without giving any information to any nodes, and in this case the node (industry) is perfectly neutral. Opposite case is a perfect source that transmits information to all other nodes without getting any information from any nodes. However, it is very unlikely to observe either perfect source or sink scenarios in the real world.

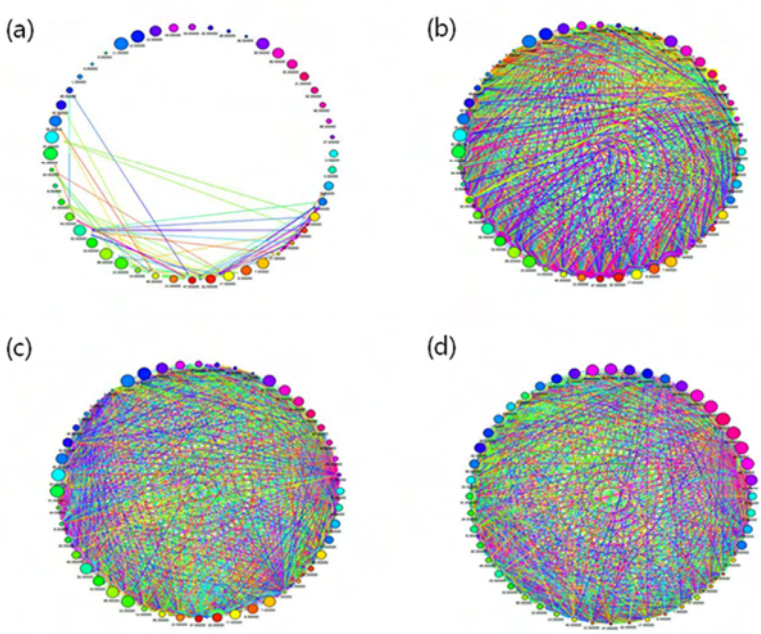
The discussion of these networks suggests, by analyzing the strength of directed connectedness (SDC) and the asymmetry of the strength of information flows (ASDC), that we can identity important industries in the stability of network structure, and the stability of the network, i.e. systemic risk of the whole network.[9] Furthermore, by analyzing network structure in terms of SDC and ASDC, we can show more subtle and nuanced separation of systemic risk whether the systemic risk of an industry or the whole economy are internally generated and contagious to other sectors or common factors from outside of the network.[10]

Figure 2 shows the network diagram of SDC and ASDC in both normal and crisis periods. We find that the number of connections among 48 industry sectors in terms of SDC (ref. to diagram (a) and (c)) and ASDC (ref. to diagram (b) and (d)) significantly increases from 2000-2001 to 2007-2008.This result indicates that the network

9 Billio et al. (2012) use PCA that is determined by correlation of nodes without showing the direction, while we are interested the in direction of information flows. They also use closeness and the Eigen vector centrality for the similar purpose, and report scores for each financial institution in their measure of importance of financial institutions. Lo/s closeness and Eigen vector centrality try to capture the importance of a node in the network frameworks. They show that banks and insurance sectors are in the center of networks of 4 financial institutions. In our case we capture the importance of an industry sector by separating source industry and sink industry, i.e. source (sink) industry has a positive (negative) net flow.

10 Billio et al. (2012), in page 549, separate the contagion and common factor exposure, based on the pattern of Granger causality.

Figure 2 Network structure of SDC that are significant at the certain threshold value among the daily return of the 48 industry sectors over 2000 to 2001 (a) and over 2007 to 2008 (b). (c) and (d) displays the ASDC network for normal period (2000-2001) and financial crisis (2009-2010), respectively.



structure estimated by STE reflects various economic conditions over time. Based on the network properties of both SDC and ASDC, we propose a novel measurement of systemic risk in the next section.

3.1 Entropy

The seminal work of Blackwell (1953) stipulates in formativeness ordering. An information structure is more informative than another whenever the latter is a garbling of the former, i.e., the less informative signal can be interpreted as observing the more informative signal with noise. Entropy is a measure based on informativeness ordering, and more information according to Blackwell's ordering corresponds to a

decrease in entropy. In decision-theoretic literature, information ordering and ranking is a basis of decision making in the face of uncertainty and incomplete information.[11] In our paper, we use entropy of stock returns of 48 industry sectors as a measure of information uncertainty of each industry. We use entropy from daily stock return as a measure of information for each industry sector. Several research use entropy as an adequate measure of the value of information in economics (Marschak (1959); Sims (2003); Blackwell (1953); Backus et al. (2014)).[12] Entropy is a measure of dispersion, a generalization of variance of sample data after daily stock return data is transformed into informativeness ordering. Another feature of entropy is that entropy extends easily to multiple periods, and we use transfer entropy at each point of time, and derive a measure of systemic risk.[13]

Entropy introduced by Shannon (1984) can measure the degree of uncertainty from various data sets, and it is defined as $H(q) = -\Sigma\, k \in K\, q\,(k)\, \ln\, q(k)$, where q(k) is probability measure of certain pattern. The entropy calculated from a specific probability density function(pdf) is a measure of degree of uncertainty and directly related to the state of system. i.e., when pdf is delta function, the entropy shows the minimum level of uncertainty, whereas it indicates the maximal uncertainty when the entropy shows uniform distribution. The uncertainty measure based

---

11 Cabrales et al. (2013) investigate the question "When can one say that a new piece of information is more valuable to such an agent than another?", and show that " in formativeness ordering is represented exclusively by the decrease in entropy of the agent's beliefs.". They conclude that "entropy as the unique objective" way to speak of the in formativeness of information structures when dealing with preferences, wealth levels and decision problems in the classes we consider." Blackwell (1953) and lit afterwards, decision makers having preferences in a particular class use complete order of all information structures.

12 Sims (2003) uses entropy to model limits in human information processing capabilities, which he called "rational inattention". (see Sims (2007) for a summary of other contributions.)

13 Cabrales et al. (2013) use the change of entropy with information priors by analyzing information ordering with posterior entropy, and argue that "in formativeness ordering is represented exclusively by the decrease in entropy of the agent's beliefs (over time between priors and posteriors)." In our paper we don't look at the change of entropy, rather the trend of entropy at each day.

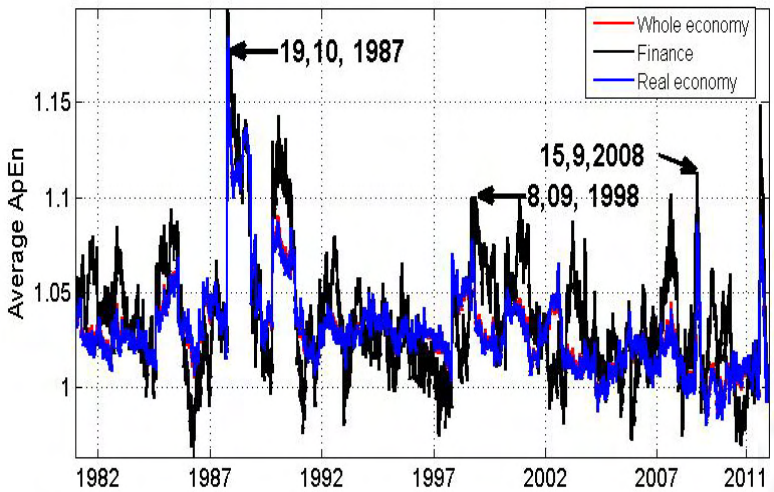
on the Shannon entropy is calculated by statistical feature in time series. However, Shannon entropy is hard to apply to the financial time series with both statistical and temporal correlation properties. To overcome the weakness of Shannon entropy, we employ the approximate entropy (ApEn) proposed by Pincus (1991). ApEn is a measure of irregularity or randomness in time series data. The ApEn is defined as follows: $ApEn(m,r) = \varphi^m(r) - \varphi^{m+1}(r)$, where m is the length of pattern, embedding dimension and r is similarity between patterns. $\varphi^m(r)$ is given by

$$\varphi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln\left[\frac{B_i(r)}{N-m+1}\right]$$

where $B_i(r)$ is the number of pattern within a similarity r and defined as $B_i(r) \equiv d[x(i), x(j)] \leq r$. The distance d between two patterns in R^m is given as $d[x(i), x(j)] = \max_{k=1,2,\ldots,m}(|r(i+k-1) - r(j+k-1)|)$, where r(t) is a return time series at day t[14]. We can measure the ApEn value by comparing the difference of the relative magnitude between repeated pattern occurrences for the pattern length, m and m+1. The ApEn that is closer to the maximum value corresponds to a random walk process, whereas minimum value of ApEn in time series indicates regular pattern. ApEn, as a class of entropy, is closely related to the market uncertainty or market volatility. In economic and financial literature, standard deviation is used as a measure of volatility and risk. Similarly, ApEn is a measure to quantify the degree of irregularity or uncertainty in the financial time series with nonlinear properties. In our measure of systemic risk, ApEn will be used as an indicator of the market stability.

14 See Steven Pincus and Rudolf E. Kalman (2004). ApEn measure shows to be a useful indicator of system stability in both real market data and artificial time series generated based on the efficiency market hypothesis.

Figure 3 Time evolution of average approximate entropy (ApEn) measure for whole economy, financial sectors, and real economy sectors



The Figure 3 shows the systemic risk of the economy based on the average ApEn value of 48 industries over the sample period of 1980 to 2012. Approximate entropy (ApEn) is a measure of randomness and uncertainty of financial time series of industry returns. There are several peaks in the figure 3 that correspond to extremely risky financial events in the real world. The peak of highest entropy occurs on 9/19/1987 (Black Monday). Second local peak happens on 9/8/1998 when LTCM crisis erupts. Third peak occurs on 9/15/2008 when the bankruptcy of Lehman Brothers is announced. Even as this single series of entropy suggests the timing and magnitude of systemic risk, it does not show the building-up of randomness or uncertainty, as the single series ignore the interconnectedness of industries. Since we are interested in the contribution of each industry to the total entropy as well as the economy wide randomness or uncertainty, we separated financial and nonfinancial sector. Figure 3 shows the systemic risk of financial sector is higher than real sector around all three events of Black Monday, LTCM and 2007/8 financial crisis. Interestingly, the relative systemic risk of financial sector is larger than real sector over the whole period. The results are indicative of the importance of incorporating transfer entropy

to identify the source of entropy, and to gauge the relative contribution of each sector to the economy-wide uncertainty.

3.2 Transfer Entropy

Next, as we are interested in the information flows in a network structure, we use transfer entropy as a measure of transmission of information from one node (industry) to other nodes (industries). In this fashion we can capture dynamic interaction between industries and the systemic risk arising from interconnection of industries. In addition to the connectedness of each industry with others, we want to measure the strength, direction, and the combination of strength and direction of connectedness based on information flows. We use two measures to fully characterize the network structure of 48 industries. One measure is the strength of directed connectedness (SDC), i.e. strength of information flow, summing up the total information flows regardless of the direction of information flows between two industries. Second, as the SDC does not capture the direction of net information flows, we separately measure the asymmetry of SDC (ASDC) to identify whether the particular industry is a source of information transmission to all other industries, or sink. We call the industry to be a source (sink), when net strength of information flows to (from) all other industries is positive (negative).

In this section, we develop a formal measure of information flow which enables us 1) to measure systemic risk, and 2) to identify the source and sink nodes in a total economic system comprised of industries in real and financial sectors.[15] As we are interested in the dynamic shock propagation in the total economic system especially during the financial crisis period, we adopt the transfer entropy (TE) measure from nonlinear dynamics literature. (Schreiber (2000); Staniek and Lehnertz (2008)) TE is measured first, by quantifying the information flow between two nodes, and second, by aggregating the pairwise information flow for a node from all other nodes in the system. The resulting measure of TE can tell whether the node plays a role of

15 The information flow could be considered as the direction of shock transmission.

either source or sink.[16] A source node has a positive aggregate net information flow while a sink node has a negative aggregate net information flow with respect to other industries. We assume that internally or externally generated systemic shocks will move from source nodes to sink nodes. In other words, a source node can either generate a shock or transmit shocks to another source or sink node while a sink node does absorb shocks or transfer shocks to other sink nodes according to the hierarchy of the network.

Introduction of transfer entropy gives us at least two pieces of information about network structure. One is the total SDC for a node defined as the sum of all information flows originating from one node to other remaining nodes, regardless of the direction of flows. Second measure is the degree of asymmetry of information flows from one industry to other industries, netting out inflows and outflows. Second, we include both global and local ASDC in SR global and local ASDC in our systemic risk measure. Global ASDC is calculated by adding all asymmetric SDC (ASDC) originating from (to) one industry to (from) all other industries. Local ASDC is a pairwise ASDC between two industries, where those industries is directly adjacent with each other. Ideally, we can measure ASDC between two industries with intermediate industries removed, and so on to get full picture of network structure. However, we use only local ASDC from two industries directly linked with each other, and use this local ASDC as a representative measure of indirect contribution of the industry to the total systemic risk.[17] The parsimonious measure using only directly linked pairwise local ASDC is used as a proxy of network structure and included in our SR measure

16 In principle, we could have a neutral node whose aggregate information flow is exactly zero. But, the probability of the occurrence of neutral node is almost surely zero.

17 The reason we consider only directly connected nodes is partly due to computational complexity. Bur more important reason is that more detailed reflection of network can create more noises according to the "Occam's razor". Furthermore, We can still capture heterogeneous network structure by the distribution of pairwise ASDC. Uniform distribution of pairwise ASDC is likely to make the network more unstable than other distributions, and we design SR measure to reflect the nature of the distribution of pairwise ASDC.

to capture the (in)stability of network structure.[18] Remaining question is how much global and local ASDC contribute the systemic risk. It can be reasonably assumed that global ASDC is likely to contribute more to systemic risk than local ASDC. In our measure of SR we apply different weights to global and local ASDC, and check the sensitivity of SR to the various combination of global and local ASDC.

Our measure of SR critically depends on the total SDC, global and local ASDC. Figure 5 shows the average SDC and average global ASDC in each year and the correlation of two measures for the sample period 1984 and 2010. Our original sample data starts from 1980, and we use 4-year rolling windows to calculate SDC and ASDC. Figure 5 shows the time variation of the measures, and the correlation between SDC and ASDC of the network of 48 industries. Focusing on the peak in both measures which is higher than average for the whole sample period we can identify one period around 1989, and the period after 2004. SDC peaks at 1989, and the ASDC peaks at 1990, at the same time correlation of two measures remain positive for 5 years. Relating this period to LTCM crisis, we can infer that crisis worsens and systemic risk increases when both SDC and ASDC increase at the same time. In 2004 both SDC and ASDC peaks in 2004, and while SDC and ASDC continues to remain at high level above the norm, and, at the same time, correlation remains positive and continues to be positive until recent period in the sample. The crude observation supports the validity of network-based measurement of SR. the following table illustrates the contribution of SDC and ASDC to the SR.

Table 1 Interactions of SDC and ASDC

		SDC (Strength of directed Interconnectedness)	
		High	Low
Asymmetric SDC	High	1	2
	Low	3	4

18 Full consideration of network structure is not only complicated but also it can introduce noises to SR measure.

In this table we argue that the SR reaches the highest (lowest) level when both SDC and ASDC are high (low) in the cell 1 (cell 4). In other cases (cell 2 and 3) where SDC is high (low) and ASDC is low (high), SR level is determined by the total SDC and the relative magnitude of global and local ASDC.

In figure 5, we find year 2004 is when both measures, SDC and ASDC, reach local peak. This period corresponds to the financial events when subprime loans started to increase rapidly, and the financialization of commodity started to be conspicuous (Cheng and Xiong, 2013). Mayer and Pence (2008) reports that subprime mortgage sharply increased around 2004 based on three sources of data on subprime mortgages: mortgages in securitized pools marketed as subprime by the securitize (Loan Performance); mortgages with high interest rates (HMDA higher-priced); and mortgages originated by lenders specializing in subprime mortgages (HMDA HUD). CFTC staff report (2008) and Masters (2008) reports that the total value of various commodity index-related instruments purchased by institutional investors has increased from an estimated $15 billion in 2003 to at least $200 billion in mid-2008. In addition to sub-prime mortgage and commodity financialization, private equity investment, CDS, and other financial instruments began to grow exponentially from 2004.[19]

3.3. Symbolic Transfer Entropy (STE)

Information flows among the economic entities including companies in both financial and real sector is measured by symbolic transfer entropy. Staniek and Lehnertz (2008) extend the measure of information transfer of Schreiber (2000) by utilizing the technique of symbolization. The symbolic transfer entropy (STE) measure has some merits over the original transfer entropy measure in a couple of ways: symbolic transfer entropy is a robust and computationally implementable method to quantify synchronization and nonlinear interactions between dynamical systems than original transfer entropy method which is related to the

19 CDS data from BIS: http://www.bis.org/statistics/derdetailed.htm

causality measure of Granger (1969).[20]

In order to better understand symbolic transfer entropy, we first discuss transfer entropy. Suppose we are interested in the interaction between X and Y systems and observe the sequences of variable *x* and *y* from the systems at t =1, 2, ...T. The transfer entropy or information flow from Y to X system is defined in the following way

$$TE_{Y\to X} = \sum p(x_{t+1}, x_t, y_t)\, log\, \frac{p(x_{t+1}|x_t,y_t)}{p(x_{t+1}|x_t)}, \quad (1)$$

where $x_t = x(t)$ and $y_t = y(t)$, for t =1, 2, 3, ..., T, and they represent the sequences of observations from systems X and Y at time t. The p indicates the transition probability density function. The joint probability density function $p(x_{t+1}, x_t, y_t)$ is the probability of three events, x_{t+1} , x_t , and y_t occurring in conjunction. The conditional probabilities $p(x_{t+1}|x_t,y_t)$ and $p(x_{t+1}|x_t)$ are the probabilities of some event x_{t+1} at time t + 1, given the occurrence of the return x_t, y_t and x_t, respectively, at time t. Transfer entropy is the weighted sum of joint probability of three states where the weights are the log ratio of two conditional probabilities. If there is no information flow from Y to X, then X and Y must be independent or $p(x_{t+1}|x_t,y_t) = p(x_{t+1}|x_t)$ so that transfer entropy becomes zero by definition. Otherwise, Y is informative to predict the transition probability of state *i* from time t to time t + 1. This measure is essentially asymmetric in that $TE_{X\to Y}$ could be different from $TE_{Y\to X}$.

Calculating symbolic transfer entropy (STE) involves the following two steps. First, we generate the symbolic time series from the original return series data using the symbolization technique.[21]We define symbols by reordering the amplitude values of return time series x_t and y_t. Given time delay τ and embedding dimension *m*, we rearrange the set of return sequences $X_t = \{x_t, x_{t+\tau}, \cdots, x_{t+(m-1)\tau}\}$ in an ascending order $\{x_{t+(k_{t1}-1)\tau} \leq x_{t+(k_{t2}-1)\tau} \leq \cdots \leq x_{t+(k_{tm}-1)\tau}\}$.

---

20 See Barnett et al. (2009) for the relationship between transfer entropy and Granger causality. They show that transfer entropy is equivalent to Granger causality when variables are Gaussian.

21 Refer to the concept of permutation entropy of Bandt and Pompe (2002) to concretely understand the symbolization technique.

This sequence of indexes is utilized to define a symbol $\widehat{x_t} \equiv (k_{t1}, k_{t2}, \cdots, k_{tm})$. For example, a set of return sequences with embedding dimension 3 and time delay 1, {0.1, 0.4, 1.1}, is symbolized into (1,2,3) according to the position of ascending amplitude. Another set of returns sequences with dissimilar amplitudes but same ranks {-0.2, 0.3, 1.2} is identically symbolized into (1,2,3). When we have equal amplitude in the set, we ensure that every x_t has unique mapping onto one of the m! permutations by figuring out the relevant indexes.[22]We can estimate the joint and conditional probabilities of the sequence of permutation indices with the relative frequency of symbols.

Second, we estimate transfer entropy using the symbolic time series data and $\hat{R}_t^i$ and $\hat{R}_t^j$

$$STE_{\hat{R}_t^i \to \hat{R}_t^j} = \sum p\left(\hat{R}_{t+\delta}^j, \hat{R}_t^j, \hat{R}_t^i\right) log \frac{p(\hat{R}_{t+\delta}^j | \hat{R}_t^j, \hat{R}_t^i)}{p(\hat{R}_{t+\delta}^j | \hat{R}_t^i)}, \quad (2)$$

where $\hat{R}_t^i$ and $\hat{R}_t^j$ denote the symbolic time series at time t reconstructed with the embedding dimension m and time delay τ from the return time series $r_t^i$and r_t^j, respectively. The symbolic transfer entropy (STE), $STE_{R_t^i \to R_t^j}$, can quantify the amount of information flow from R_t^i to R_t^j.

3.4. Information Flow Asymmetry: Pairwise and Aggregate

The measurement of systemic risk based on STE uses directionality and strength of information flows, while Billio et al. (2012) use binary Granger causality. First, our directionality measure provides the strength of directionality while binary Granger causality network has three discrete levels of directionality of information flow from *i* to *j*, that is +1($i \to j$), 0(bidirectional feedback), and -1($j \to i$). Even though there is feedback effect between two institutions, it is possible that one institution weakly dominates the other in terms of influence. Our measure can detect this asymmetric information flow while the binary

22 Suppose $x_{t+(k_{t1}-1)\tau}$ is equal to $x_{t+(k_{t2}-1)\tau}$. We write that $x_{t+(k_{t1}-1)\tau}$ is less than or equal to $x_{t+(k_{t2}-1)\tau}$ if k_{t1} is less than k_{t2}

network cannot. Second, we can construct a network topology of the asymmetry of SDC based on the strength of information flows.

To measure the asymmetry of information flows we consider the pairwise information flow using the STE, defined as

$$D^S_{i\to j} = STE_{R^i_t \to R^j_t} - STE_{R^j_t - R^i_t} \tag{3}$$

If the pairwise information flow $D^S_{i\to j}$ has a positive value, R^i_t is a driving force for another process R^j_t, while if it has a negative values, R^j_t drives R^i_t. We call *i* and *j* pairwise source and sink each if the pairwise information flow, $D^S_{i\to j}$ takes a positive value. In the case that R^i_t and R^j_t follow independent and identically distributed (IID) processes, that is, $\frac{p(\hat{R}^j_{t+\delta}|\hat{R}^j_t,\hat{R}^i_t)}{p(\hat{R}^j_{t+\delta}|\hat{R}^i_t)} = 1$, then the STE, $T_{R^i_t \to R^j_t}$ is also zero. That is, there is no information flow when two processes are random.

We define another information flow which accounts for the aggregate information flow from one node (industry sector) to all other nodes (industry sectors) in a system. Suppose that there are N nodes in a system S. First, we define outward information flow from node *i* to all other nodes in S. For the brevity of expression, we define alternative representation of symbolic transfer entropy: $STE_{R^i_t \to R^j_t} \equiv STE_{i\to j}$.

$$IF^{out}_i = \frac{1}{N-1}\sum_{i\neq j} STE_{i\to j} \tag{4}$$

The inward information flow to node *i* from all other nodes is defined by the following equation:

$$IF^{in}_i = \frac{1}{N-1}\sum_{i\neq j} STE_{j\to i} \tag{5}$$

To estimate the total strength of directed connectedness (TSDC) we consider all possible connections in a given system and SDC is defined by the following equation:

$$\mathrm{T}SDC = \frac{1}{N}\sum_{i} IF_i^{out} + IF_i^{in} \tag{6}$$

SDC is high when the interrelation among the industry sectors is very close based on the concept of too-interconnected-to-fail. To consider the contagion effect among the industry sectors we define the asymmetry of the strength of directed connectedness of an individual node i by taking the value of difference between outward information flow from node i and inward information flow to nodei.

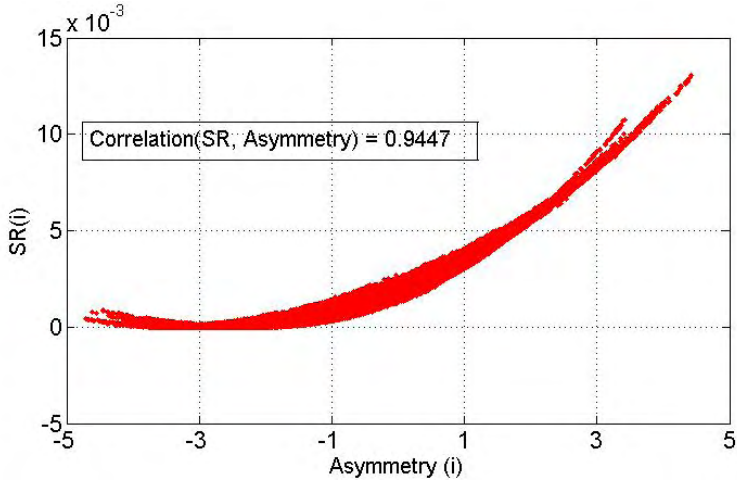
$$ASDC_i = \frac{1}{N-1}\sum_{j \neq i} \frac{STE_{i \to j} - STE_{j \to i}}{STE_{i \to j} + STE_{j \to i}} \tag{7}$$

If the $ASDC_i$ of node *i* is positive (negative), then the node plays the role of information source (sink) of the system. The ASDC measure spans from -1 to +1. Two boundary cases occur when all nodes in a system S form a star network where only one node *i* is connected with all other nodes *j(≠i)* and all directions of information flow from *i* to *j* are the same. The ASDC of node *i* takes the value of +1 (-1) in case that the direction of information flow is outward from (inward to) *i*.

3.5. Measurement of Systemic Risk

The (in) stability of network structure or system can be characterized by total information flows which is defined as the total strength of information in the system as a whole, i.e. sum of SDC, and the net information flows for each node with the rest of nodes (ASDC). We also assume that, for tractability, only global source and sink nodes are potentially affect the (in)stability of the network structure, where global source (sink) node is defined as a positive (negative) sum of ASDC. Figure 4 shows the positive relationship between the systemic risk and an asymmetry of information flows among industry sectors.

Figure 4 Relationship between the measure of systemic risk and asymmetric of information flow



In nature, the movement of particles is produced when difference in potential energy between any two places in the same space is great. In a stream of river water flows and does not stand still, when there is an asymmetry between inflows and outflows in any place. Likewise, when the asymmetry of information inflows and outflows in the economic system increases, information will flow more strongly from one to another node. However, when there is no difference between inflows and outflows in all sectors of the economy, the information flow is stagnant, i.e. the financial system will be stable and in equilibrium. As such, the asymmetry of information flows between and among industrial sectors and the total amount of information in the system are two major factors in determining the characteristics of network and market stability.

The principal goal of developing a measure of systemic risk is to capture the impact of the combination of strength and asymmetry of information flow on the systemic risk and stability of network. By Considering just two dimensions of systemic risk, namely strength of directed interconnectedness (SDC) and the asymmetry of SDC (ASDC), we can still get rich and economically sensible measure of systemic risk; 1) systemic risk of each industry which is defined as the contribution of each industry to the total systemic risk, 2) the nature of node (real sector

and financial sector and each industry separately) whether it is a source or sink.

To further clarify the economic meaning of SR measurement, we can think of two extreme cases. First, suppose an industry is a source to all remaining industries, then the source industry has high systemic risk. Note that our economy-wide systemic risk is the sum of systemic risk of all industries. When an industry is independent of all other institutions, then both pairwise and aggregate information flow of the industry is zero. This independent industry does not contribute to the systemic risk because it cannot transmit the shock generated by one industry to other industries. The most stable system with the least likelihood of systemic events is a collection of independent industries without any source and sink in the system.

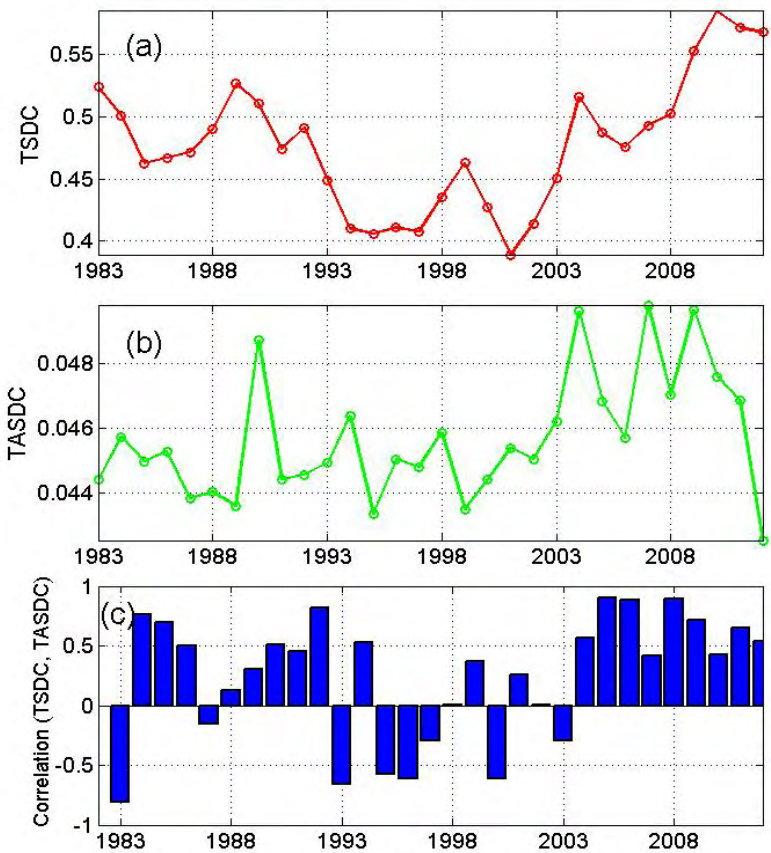
When ASDC of an industry is -1, the industrycontributes nothing to the systemic risk of system as it just play the role of black hole which absorb all possible systemic shocks created from other industries. However, an industry i in a system S start increasing the system wide risk when it plays the role of aggregate source. Systemic risk will increase as we have more aggregate sources in the system. Let's first consider the pair of a global source industry. This can serve as a path through which shocks can be transmitted. As we have more this kind of paths, it is more likely that the shocks transmit further through the chain of relays like domino effect. If this chain finally ends up with a ring, then we observe a feedback which will have a larger impact on the systemic risk.

The diagrams in Figure 5 illustrates the importance of SDC, and ASDC that captures source-sink information flow. Finally we postulate SR as a function of strength and asymmetry of information flows.

$$SR(t) = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \left(SDC_{i \to j}\right) \times (ASDC(i) + \alpha) \times (LASDC_{i \to j} + \alpha) \quad (8)$$

where local ASDC (LASDC)is the asymmetry of SDC for pairs of industry sectors, $LASDC = \frac{(STE_{i \to j} - STE_{j \to i})}{(STE_{i \to j} + STE_{j \to i})}$. We add α to make a

Figure 5 Evolution of the strength of directed connectedness and its asymmetry: (a) 1 year rolling estimates of the total summation of strength of directed connectedness; (b) Total asymmetry of SDC; (c) linear correlation between SDC and ASDC



positive value for $(ASDC(i) + \alpha)$, as $ASDC(i)$ takes negative value when the node i is a sink node. By choosing minimum α, all institutions, no matter whether the institution is a source or a sink except when the institution is a perfect sink, positively contribute to the systemic risk of the whole network system.[23] In this paper, we use $\alpha = 0.1$. To check the

23 The maximum meaningful value of α is 0.1, since SR collapses to SDC when α is higher than the maximum. At α = 0.1, the correlation of SR with both SDC and ASDC is the same at 0.52. See Figure 8 in the appendix.

robustness of α value, we calculate the systemic risk according to various α values and find that observed results is statistically independent of α value. Figure 1 in the Appendix shows the correlation of SR with ASDC and SDC as α changes.

Table 2 The 48 industry group classification by Fama French 1997

Industry	Industry name	Industry	Industry name
1	Agriculture	25	Shipbuilding, Railroad Equipment
2	Food Products	26	Defense
3	Candy & Soda	27	Precious Metals
4	Beer & Liquor	28	Non-Metallic and Industrial Metal Mining
5	Tobacco Products	29	Coal
6	Recreation	30	Petroleum and Natural Gas
7	Entertainment	31	Utilities
8	Printing and Publishing	32	Communication
9	Consumer Goods	33	Personal Services
10	Apparel	34	Business Services
11	Healthcare	35	Computers
12	Medical Equipment	36	Electronic Equipment
13	Pharmaceutical Products	37	Measuring and Control Equipment
14	Chemicals	38	Business Supplies
15	Rubber and Plastic Products	39	Shipping Containers
16	Textiles	40	Transportation
17	Construction Materials	41	Wholesale
18	Construction	42	Retail
19	Steel Works Etc	43	Restaurants, Hotels, Motels
20	Fabricated Products	44	Banking
21	Machinery	45	Insurance
22	Electrical Equipment	46	Real Estate
23	Automobiles and Trucks	47	Trading
24	Aircraft	48	Almost Nothing

Table 3 Descriptive Statistics

We provide the descriptive statistics of systemic risk measure and macroeconomic variables we link to the systemic risk in the following analyses. Macroeconomic variables include three month treasury rate, London interbank offered rate (LIBOR), the return and volatility of S&P 500 index of daily frequency, and unemployment rate of monthly frequency.

Mean		Median	Std. Dev.	Max	Min	Mean	Median	Std. Dev.	Max	Min
Panel A: Systemic risk						**Panel B: Unemployment rate**				
1982-1987	103.03	102.14	5.97	122.61	91.50	7.97	7.40	1.49	10.80	5.70
1988-1993	102.51	100.57	10.28	131.11	81.18	6.25	6.50	0.88	7.80	5.00
1994-1999	89.65	88.95	5.26	105.59	78.50	5.10	5.15	0.69	6.60	4.00
2000-2005	100.75	99.77	13.45	132.58	79.24	5.15	5.40	0.72	6.20	3.90
2006-2011	122.20	122.19	11.53	147.88	97.87	7.31	8.70	2.29	10.00	4.40
Panel C: Three month T-bill rate						**Panel D: London Interbank Offered Rate**				
1982-1987	8.30	8.10	2.20	15.49	5.18	7.01	7.00	0.73	9.31	5.63
1988-1993	5.86	6.00	2.08	9.45	2.67	6.44	6.88	2.32	10.63	3.13
1994-1999	5.01	5.13	0.56	6.07	2.98	5.50	5.63	0.58	6.50	3.25
2000-2005	2.79	1.99	1.79	6.42	0.81	3.08	2.19	1.92	6.87	1.00
2006-2011	1.84	0.22	2.10	5.19	0.00	2.47	1.45	2.18	5.73	0.25
Panel C: Return of S&P500 index						**Panel F: Volatility of S&P500 index**				
1982-1987	0.13	0.16	0.15	0.46	-0.26	9.08E-05	7.72E-05	6.80E-05	4.56E-04	4.00E-05
1988-1993	0.08	0.09	0.11	0.30	-0.26	1.26E-04	7.62E-05	1.42E-04	4.92E-04	2.94E-05
1994-1999	0.18	0.19	0.10	0.40	-0.04	8.27E-05	5.36E-05	5.80E-05	2.01E-04	2.24E-05
2000-2005	-0.02	0.05	0.17	0.36	-0.42	1.49E-04	1.68E-04	7.76E-05	3.01E-04	4.11E-05
2006-2011	0.01	0.08	0.22	0.52	-0.67	2.22E-04	1.31E-04	2.45E-04	8.29E-04	3.46E-05

4. Empirical Results

4.1. Data

Basic data for the systemic risk is daily stock returns from CRSP between Jan 1, 1980 and Dec 31, 2012, value weighted index returns for 48 industrial sectors is used adopting Fama and French (1997) categorization. Financial sector consists of banking(industry group #44), insurance(# 45), and trading industries(# 47).The detailed description is reported in Table 2. We classify all other manufacturing and service industries except for financial sectors as real economic sectors. To investigate the temporal evolution of network structure and systemic risk, we construct a sequence of networks from one-year moving window of daily returns. The components of SR, i.e. SDC, ASDC and LASDC is calculated using STE method from a network structure of information flows for each trading day. The created network for each sub-period is a set of significant network connections in terms of the path of information flow, which is the level of systemic risk induced by network topology over time.

Macroeconomic variables come from the Federal Reserve Economic Data (FRED) at Federal Reserve Bank of Saint Louis: three month treasury bill rate, ten year treasury bond rate, London interbank offered rate (LIBOR), S&P 500 index returns of daily frequency and unemployment rate of monthly frequency. Table 3 shows descriptive statistics of systemic risk measure and macroeconomic variables related to the systemic risk.

4.2. Evolution of Systemic Risk

Figure 6 shows time evolution of systemic risk and the decomposition of systemic risk by sectors. The measure of SR using STE methods is robust and is not driven by chance, we compared our SR measure to the Monte-Carlo simulation of randomly generated industry index returns, and find there is a significant difference in probability density function between the two samples. The bottom part of figure 6 (a) shows mean and two standard deviation of systemic risk from simulations. This

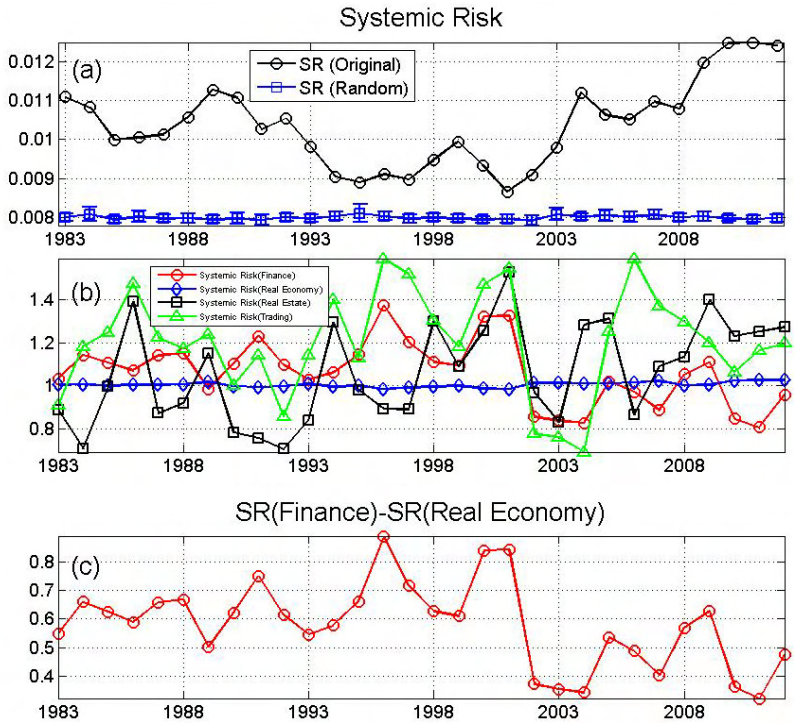
comparison indicates that our network structure is generated by specific connections among industry sectors in the sample.

It is clearly visible that the magnitude of systemic risk waxes and wanes over time in diagram (a). The level of systemic risk moderately oscillates in 1980s until it climbs up to a peak in period of Black Monday in 1987. Systemic risk shows downward trend in early 1990s, and it reaches the lowest level in 1994 in the sample period, and subsequently maintains low level in late 1990s until it bumps up to a peak. The peak coincides with Russian and Asian crisis in late 1990s. Most interesting time variation of systemic risk is the period in 2000s relative to the previous Asian crisis. There is a turning point in early 2000s - systemic risk continues to grow after 2001, reach interim peak in 2004 until it reached a peak in 2009/2010, while it subsides after 1997 Asian currency crisis for a relatively long period until 2001. In contrast, systemic risk remains at high level relative to the average level of systemic risk for the sample after the 2008 until the end of sample period.[24] Focusing only on peaks of systemic risk, Figure 6 (a) shows that the sharp increase in the systemic risk coincides with several identified crisis periods. Systemic risk reaches local peaks around 1987 LTCM crisis, 1997 Russian crisis, and 2008 financial crisis. This match between actual crisis periods and high peak of systemic risk levels supports the validity and usefulness of systemic risk in detecting real crises in the US economy.

Figure 6 depicts the unfolding of systemic risk over time. The peaks of systemic risk are particularly interesting and worth close investigations, as we are interested in the systemic risk as a symptom of instability of economy. There are three notable peaks of systemic risk in 1990, 2004 and 2010 in our sample period. The peak in 1990 might be related to the LTCM crisis in 1989, and systemic risk has been abated in the following years, possible by higher interest rates. Interesting peak is in 2004, about three years before the actual financial crisis of 2007/2008. Considering the start-ups of several financial surge in 2004 – sudden increase in

24 The changing pattern of systemic risk in 20008/2009, i.e. systemic risk remain at high plateau, is worth noting. The systemic risk in the market are not abated in spite of or because of the quantitative easing (QE) of U.S. Federal reserve liquidity policy.

Figure 6 Time evolution of systemic risk measure: (a) 1 year rolling estimates of the systemic risk; (b) systemic risk of industry sectors such as finance, real economy, real estate, and trading sectors, and it is normalized by the average of systemic risk of 48 industries; (c) ratio of difference of systemic risk between financial and real economy sectors.



subprime mortgage, financialization of commodity, CDS activities, and others, we decompose the economy-wide systemic risk into financial and real sectors in Figure 6 (c). Figure 6 (c) shows the systemic risk of financial sector is higher than real sector around all three events of black Monday, LTCM and 2007/8 financial crisis. However, the systemic risk of financial sector relative to real sector, i.e. contribution to the economy-wide systemic risk starts to decrease approximately one year before the actual events. The results show the importance of considering transfer entropy, and identify sources of systemic risk, and the relative contribution of each sector to the economy-wide systemic risk. While

Table 4 The rank order of systemic risk in 48 industry sectors in U.S. market

We examine the rank order of industry sectors according to the degree of portion in systemic risk over different periods of time.

Periods /Rank	1983-1987	1988-1993	1994-1999	2000-2005	2006-2011
1	14	34	47	20	19
2	18	45	30	11	46
3	47	43	21	32	24
4	38	21	42	15	34
5	22	17	35	23	47
6	16	14	11	40	8
7	11	28	41	1	20
8	34	8	40	37	40
9	44	18	38	35	30
10	8	13	44	46	39
11	36	26	45	17	38
12	21	42	33	10	43
13	35	11	22	19	15
14	13	41	34	43	29
15	31	2	46	24	23
16	17	47	14	6	1
17	39	31	23	25	48
18	5	19	18	47	12
19	32	22	19	14	35
20	24	36	48	13	10
21	28	32	9	28	36
22	26	7	17	21	27
23	43	10	6	38	28
24	42	24	20	16	37
25	41	9	43	34	16
26	45	12	10	9	17
27	19	6	28	41	18
28	10	16	2	42	2
29	48	40	32	7	22
30	15	38	24	22	41
31	9	15	26	18	26

Table 4 (Continue)

Periods /Rank	1983-1987	1988-1993	1994-1999	2000-2005	2006-2011
32	46	46	13	48	5
33	3	3	36	39	25
34	6	35	5	44	32
35	25	44	29	36	7
36	40	20	16	33	21
37	7	29	37	31	11
38	12	4	3	29	14
39	37	33	12	45	31
40	20	25	25	27	9
41	2	48	8	8	4
42	23	30	31	12	3
43	4	39	15	3	33
44	1	27	39	2	13
45	30	5	7	5	45
46	29	23	1	4	44
47	33	1	4	30	6
48	27	37	27	26	42

the systemic risk of finance sector dominated the systemic risk of real sector before 2004. The trend changes in 2004 - the economy-wide systemic risk is evenly attributed to both sectors. In other words, real sector instability started to drive economy systemic risk as much as financial sector. To gain insights on which specific industry contribute to the economy-wide systemic risk, we dissected further into real estate and financial trading in figure 6 (b). Real estate industry in 2004 and financial trading industry in 2006 contribute more than industry average to the economy-wide systemic risk.

Table 4 shows rank order of industries sorted by the systemic risk of each industry. The industry number corresponds to the list of 48 industries in table 2 in appendix. Focusing on the last three periods, 994-1999, 2000-2005 and 20006-2011, there is a noticeable change in the contribution of a particular industry to economy-wide systemic risk. First, in terms of contribution to systemic risk, three financial industries

(44, 46, and 47) belong to the top one-third of industries in 1994-1999. However, systemic risk of financial sector becomes lower in the following period, and in 2006-2011, only trading industry (47) is one of top 5 industries in the order of systemic risk among 48 industries, and other two financial industries belong to the bottom. Second, the changing order of the systemic risk of real estate industry (46) in the three periods is prominent. The contribution of real estate to the economy-wide systemic risk, i.e. rank order of systemic risk, is 15, 10, and 2 in 1994-1999, 2000-2005, and 20006-2011 respectively. Third, other industries in real sector shows changing pattern as well. In the two periods, 2000-2005 and 2006-2011, that includes financial crisis, top ten industries with highest systemic risk are real sectors, with an exception of financial industry 47 (trading) in 2006-2011. In the period 2000-2005 that is considered as a crisis build-up period, real sector industries are systemically important industries at the top ranks up to the rank of $18^{th,}$ and financial trading industry (47) takes the rank of 19^{th}.

In summary, we find that the major drivers of financial crisis of 2008/2009 is not simply banks and financial intermediaries, but the trading activities of financial institutions, more specifically, trading activities of financial institutions in real assets, such as real estate and commodity. In addition, real sectors play a major role during crisis period. In the context of components of systemic risk, we can interpret the results as follows: financial sector was a source of instability before the crisis and immediately prior to the crisis, it is real sector that exacerbated the crisis. This sequence of events corresponds to the findings of numerous financial researches on financial crisis in our literature review. For example, banking sector creates asset bubble, asset market crash create crisis in financial sector speeding to all economic sectors. (Acharya and Naqvi, 2012).

4.3 Systemic Risk and Macroeconomic Variables

Table 5 reports the relationship between systemic risk and selected macroeconomic variables, which are considered to be relevant in the literature. Financial macroeconomic variables include three month

Table 5 The relationship between the measure of systemic risk (alpha=0.1) and macroeconomic variables

We examine the relationship between our systemic risk measure and macroeconomic variables over different periods of time. The estimation model is followed:

$$SR(t) = \alpha + \beta \times \text{macroeconomic variables} + \varepsilon(t)$$

Periods / macroeconomic variables	1982 - 1987		1988 - 1993		1994 - 1999		2000 - 2005		2006 - 2011	
	beta	t-value	beta	t-value	beta	t-value	beta	t-value	beta	t-value
3 Month T-bill rate	-0.07	-2.42**	0.79	50.84***	0.10	3.82***	-0.79	-50.37***	-0.87	-69.20***
10 year T-bill rate	-0.09	-3.31***	0.90	78.48***	-0.57	-26.98***	-0.84	-59.02***	-0.75	-43.97***
Libor rate	0.36	8.64***	0.83	57.24***	0.25	9.93***	-0.78	-48.40***	-0.86	-64.53***
S&P500 index return	0.51	21.24***	-0.12	-4.80***	0.36	14.77***	0.30	12.23***	-0.24	-9.55***
index volatility	0.14	4.85***	0.14	5.59***	0.23	8.99***	-0.22	-8.74***	0.08	3.24***
Unemployment rate	0.42	2.69***	-0.66	-5.97***	-0.68	-6.23***	0.78	8.32***	0.81	9.20***

Treasury bill rate, ten year Treasury bond rate, and LIBOR rate, and the return and volatility of S&P 500 index. Monthly unemployment rate is selected to real macroeconomic condition. Table 5 reports the coefficient and statistical significance (t-value) from the simple regression of systemic risk on each macroeconomic variable. The estimation model is:

$$SR(t) = \alpha + \beta \times \text{macroeconomic variables} + \varepsilon(t)$$

Regression results are separately reported for 5-year windows reflecting the time evolving nature of systemic risk. While all macro variables are significantly correlated with systemic risk in all periods, the signs of correlation are opposite across certain sub-periods.

Three month Treasury bill rate, LIBOR rate, and Federal Funds rate are positively associated with systemic risk measure in 1980s and 1990s while they are negatively associated in 2000s. Especially, all three interest rate variables, 3-month T-bill, 10-year T-bond and Libor are negatively correlated to systemic risk, i.e. lower interest rates increases systemic risk.

This implies that the systemic risk might be elevated due to the continuous fall of short term interest rate or cheap credit. It is interesting that the volatility of S&P 500 index is positively related in all sub periods except for 2000-2005 sub periods. In other words, systemic risk and stock return volatility diverge in opposite directions, and the S&P 500 index returns are increasing in the same direction as systemic risk. During this sub period systemic risk kept rising, while short term interest rate continues to fall. One plausible scenario is that low interest rates create a surge in liquidity, and in turn those funds invest in risky assets including stock market. As fund flows into stock market increases, volatility of stock market is reduced.

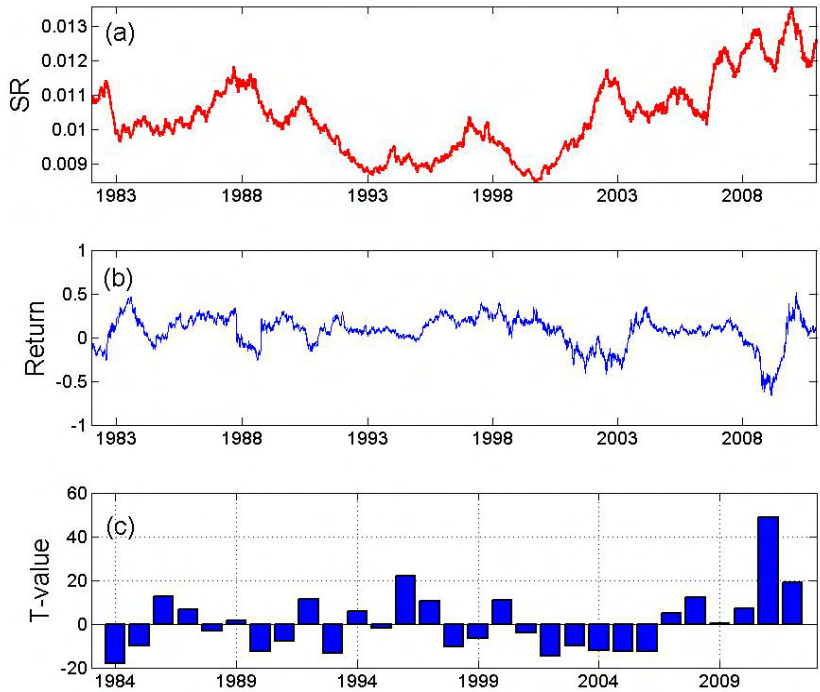
Unemployment rate is negatively related to systemic risk in 1980s and 1990s whereas unemployment rate is positively related to systemic risk in 2000s. The reversal of correlation between unemployment rate and systemic risk from positive to negative correlation during these sub periods is noticeable. Positive correlation in 2000s indicates that heightened systemic risk of real economic sector hampers real sectors to

Table 6 Predictive power of our systemic risk measures. Regression coefficients and t-statistics for regressions of aggregation return of S&P500 index on systemic risk measure calculated by proposed method.

The systemic risk at present has a predictive power of return time series in future. Statistics that significant at the 1%, 5%, and 10% level are showed in ***, **, and *.

Year	Beta coefficient	t-value	Year	Beta coefficient	t-value
1984	-5.668797	**-3.7330112*** **	1998	-4.4217589	**-4.044795*** **
1985	10.308282	**14.294051*** **	1999	-6.0253508	**-2.3794897*** **
1986	3.9830118	**4.2881295*** **	2000	-0.6023354	-1.3818786*
1987	-4.381641	**-8.238737*** **	2001	-8.1405656	**-10.341469*** **
1988	-7.4946731	**-4.4153538*** **	2002	1.230101	0.72206367
1989	-9.9642619	**-11.812977*** **	2003	-9.0892414	**-9.8145291*** **
1990	-8.3203572	**-10.204909*** **	2004	-15.866158	**-9.5725*** **
1991	13.123054	**19.466266*** **	2005	-10.518364	**-6.8002758*** **
1992	10.059414	**14.019158*** **	2006	2.2694231	**6.732826*** **
1993	2.0172971	**4.016253*** **	2007	1.4866821	**8.0788017*** **
1994	-0.97584333	-1.8695624**	2008	1.3420514	**5.6571508*** **
1995	-4.4357146	**-10.966582*** **	2009	-11.192896	**-6.5888183*** **
1996	-9.0512814	**-6.8673625*** **	2010	40.413091	**14.023274*** **
1997	-7.2782949	**-9.8092009*** **	2011	12.123111	**18.543751*** **

Figure 7 | Time evolution of t-value of beta coefficient



hire workers. This is in contrast to the opposite relationship between interest rates and systemic risk in the same period.

4.4. Predictability and Warning Signal of Systemic Risk

Ultimately we are interested in the ability of our measure of systemic risk in predicting future instability of network of industries and serve as a warning signal for the future. Figure 7 and table 6 show out-of sample predictability by checking the sensitivity of out-of sample systemic risk to in-sample measure of systemic risk. Table 6 reports the beta of the regression of S&P 500 index returns (out-of sample) to the systemic risk (in-sample) to gauge the predictability of systemic risk on future stock returns. It shows that coefficient of the beta in the regression is statically significant at 1% for most of the sample years. Most striking results is the change of signs, and especially negatively significant signs of the

coefficient in the period of 1987-1990, 1995-1999, and 2001-2005. Those are the periods of unusually climbing and irrationally exuberant stock market. As this is an out-of sample prediction, the negatively significant signs of the coefficient in prolonged periods seem to be canaries in a coal mine. Note that the mean value and significance of beta shows the forecasting power of in-sample systemic risk in predicting out-of sample stock returns. Notice that stock returns are considered as one of important measures for the effectiveness of monetary policies.

The systemic risk stays at high level in the post-crisis period from 2009 and after, but it is fluctuating. The fluctuation of systemic risk at high level indicates the heightened probability of double dip. Table 1 in section 3.2 illustrates the interaction of SDC and ASDC, and we discussed that the economy in the cell 2 or 3 can switch back to either cell 1 with a potential crisis or cell 4 with stabilizing network. While the measure of systemic risk alone is an imperfect predictor of crisis, the heightened system risk warrants close scrutiny in terms of SDC and ASDC. Furthermore, we show that the build-up of systemic risk is not confined in the financial sectors, but rather real asset activities of financial institutions contribute to the build-up of the financial crisis. Acharya and Naqvi (2012) develop a theory where banks flush with liquidity sow seeds of a crisis, by creating asset price bubble, and at the end, suffer, like a boomerang, as asset price bubble bursts.

5. Conclusions

We develop a measure of systemic risk in a network of 48 industries based on the strength and asymmetry of information flow measured by symbolic transfer entropy (STE). Our measure is versatile, in that not only we measure economy-wide systemic risk, but also we can decompose total into the systemic risk of each industry and sector, i.e. contribution of each industry and both financial and real sectors to total systemic risk. The evolution of systemic risk in the United States using size-weighted index returns of Fama and French 48 industries shows that there is a turning point in 2001 in that systemic risk has stated to

build up from 2001, reaches a peak in 2004 and accelerates until it reaches a peak in 2007. Year 2004 when systemic risk increases rapidly are documented in financial literature as a beginning of risky financial activities in terms of financialization of commodities, acceleration of sub-prime mortgage, and other derivative transactions.

We identify that financial sectors contribute to total systemic risk significantly before the crisis. However, immediately before the crisis, real sectors are major source and contributor to economy-wide systemic risk except financial trading industry. While systemic risk of trading industry remains as one of major contributors, systemic risk of real estate industry keeps rising over the sample period and it is the second highest risky industry in 2006-2011. The validity of systemic risk is supported by strong correlations between systemic risk and macroeconomic variables. Three month Treasury bill rate, LIBOR rate, and Federal Funds rate are positively associated with systemic risk in 1980s and 1990s, while they are negatively associated in 2000s. This implies that the systemic risk might be elevated due to the continuous fall of short term interest rate or cheap credit.

Current crisis is considered to be a result of activities of financial intermediaries in sub-prime mortgage, commodity trading, and others. As such, the regulatory and policy measures are focused on the control of those activities. Central banks take measures in capital ratio stipulated in Basle II, requirement that all CDS trading has to be traded through Chicago Mercantile Exchange, and several measures to limit risk activities of financial institutions. Dodd-Frank act is a comprehensive measure to regulate activities of financial institutions. However, financial institutions find ways to boost risk activities. Already we see some signs of financial institution engaging in real asset activities that are not directly related to the main functions of financial institutions, which are approved by financial regulators.[25]

Policy prescription on asset price overvaluation, securitization,

25 Some examples are: Senate Democrats ratcheted up their criticism of big banks' activities in the physical commodities markets Wednesday, pressuring the Federal Reserve to act quickly in an ongoing review of the banks' ability to trade in materials such as oil and aluminum. (WSJ, Jan. 15, 2014). Big Banks Find Cash Storing Commodities - (WSJ July 5, 2011)

derivatives, capital ratio, etc. are all measures after the fact, and considering the "genius" of wall street and financial institutions, the future crisis, if any, is likely to take different forms and shape. Admitting that economic agents pursue self-interests, the feasible actions to preempt or mitigate potential bubble or crisis should rely on the scenario that is broad enough to reflect any potential activities that might build up to the crisis, and the signal that can be obtained with given data and information. Our network approach to systemic risk considers wide range of financial activities, and suggests a measure of systemic risk based on industry stock returns. The time variations of systemic risk that capture the information flows in two dimensions of strength and asymmetry can be explained by the stock return induced entropy. Further, this measure of SR has significant predictive power, and enough confidence to frame policy measures, e.g. interest rate policy or quantitative measures.

The financial crisis of 2007/2008 created economic problems comparable to the Great Depression of 1929/1930, and every part of economy, domestic and global, and financial and real sectors, experienced great adversities, pose great challenges in figuring out what happened, why it happened, how it happened, what can be done to alleviate the difficulties, what can be done not to repeat the same calamities. Furthermore, as the crisis, by its nature, is unpredictable, and we are unprepared when it happens, it is a keen interest to economic agents and utmost urgency to policy makers to find out implementable early warning signals to better prepare or preempt potential future crises.

Figure 8 The correlation between systemic risk and essential quantity, such as SDC, ASDC according to various α value from 0 to 1.

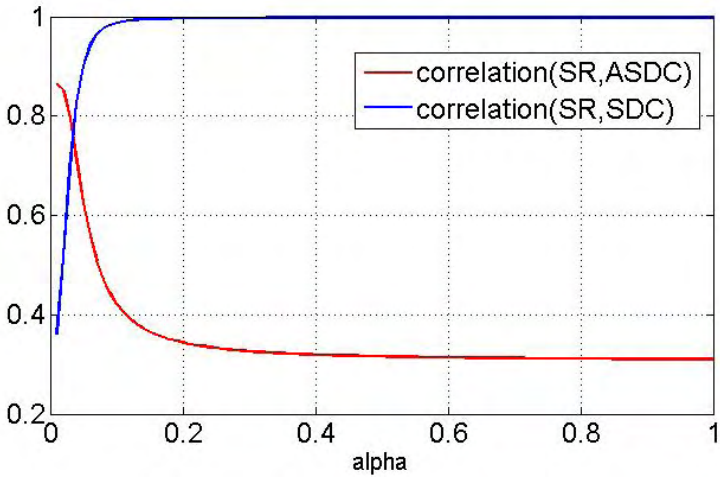


Figure 9 Time evolution of both systemic risk and 3 month T-bill rate.

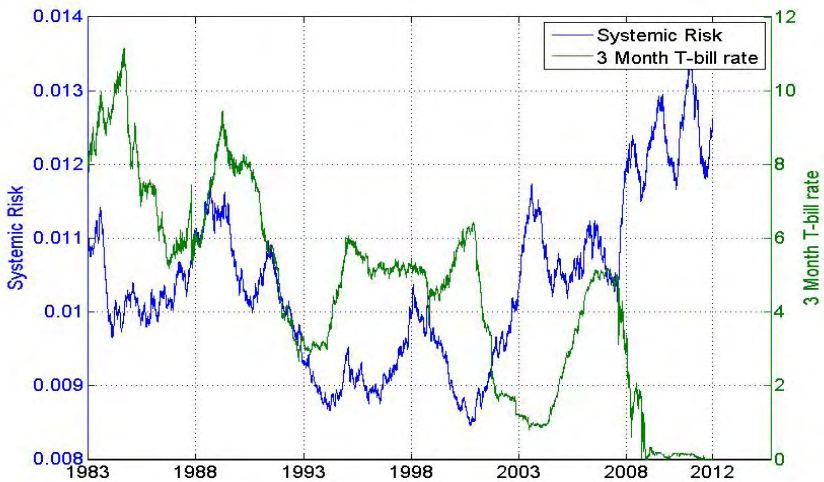


Figure 10 Correlation between SR and 3 month T-bill rate over different periods of time.

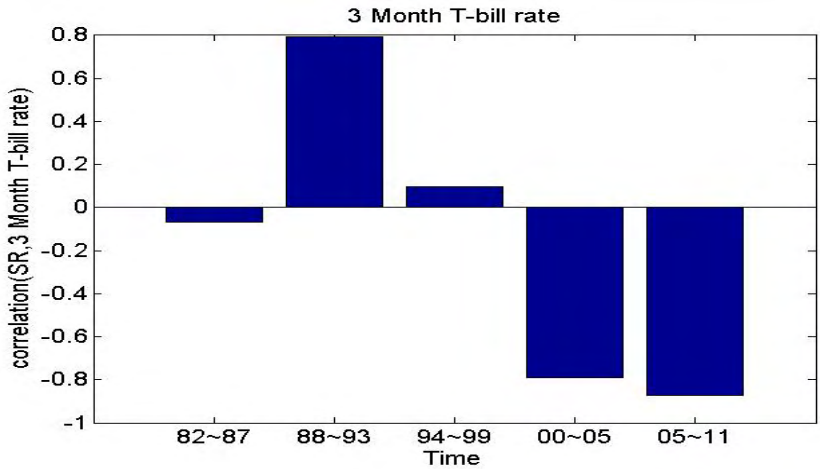


Figure 11 Time evolution of both systemic risk and 10 year T-bill rate.

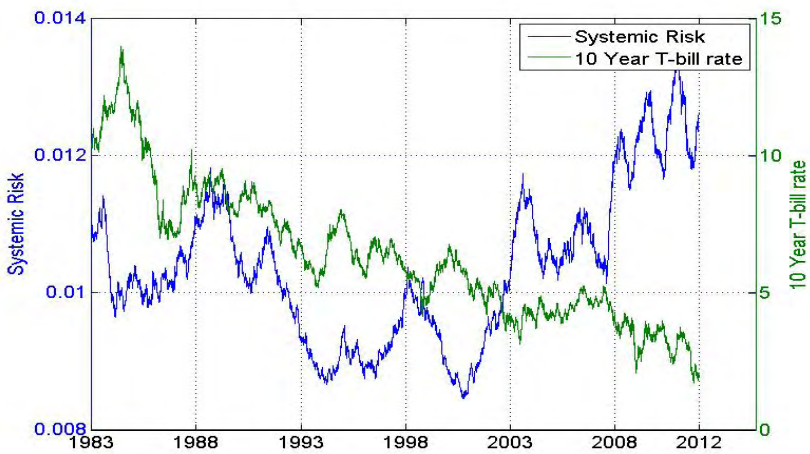


Figure 12 Correlation between SR and 10 year T-bill rate over different periods of time.

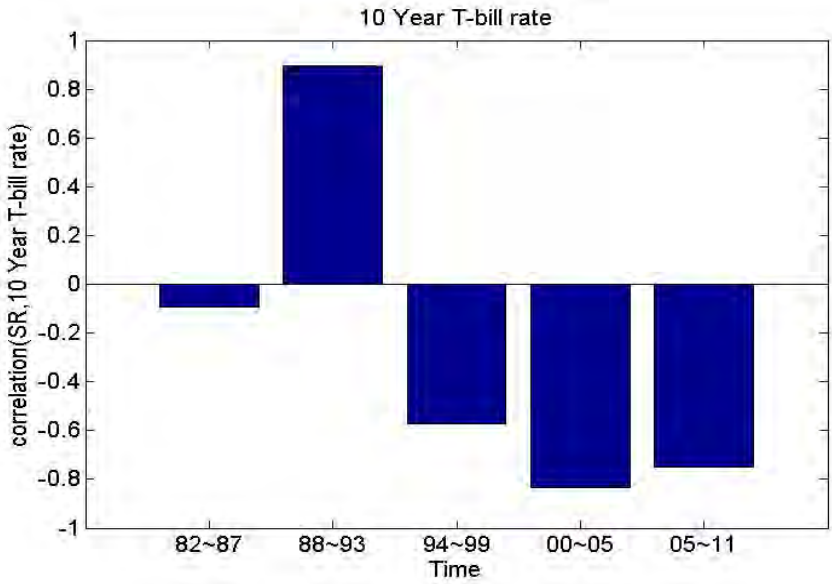


Figure 13 Time evolution of both systemic risk and Libor rate

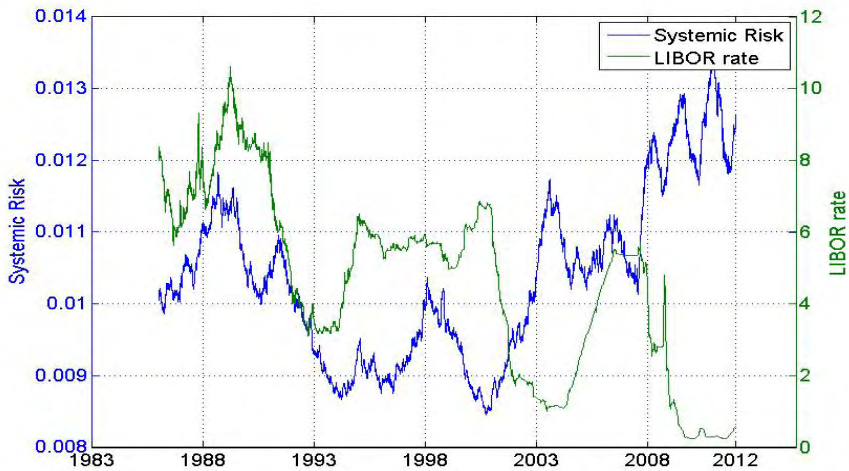


Figure 14 | Correlation between SR and Libor rate over different periods of time.

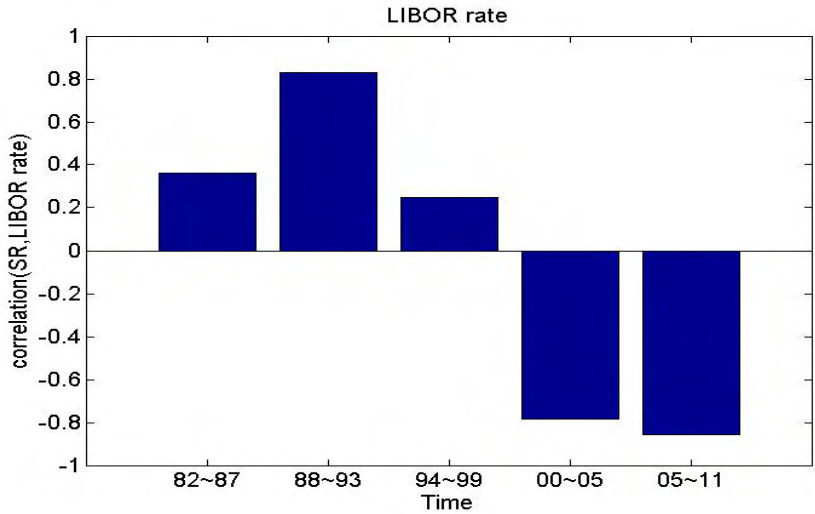


Figure 15 | Time evolution of both systemic risk and S&P500 index return

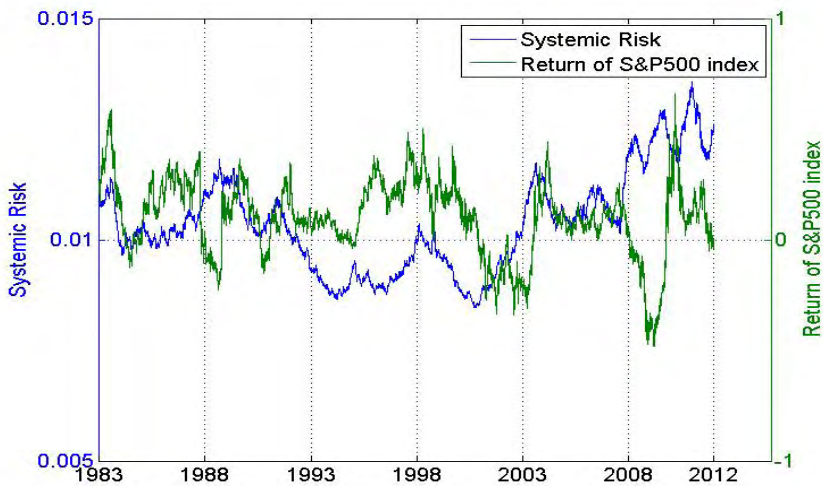


Figure 16 Correlation between SR and S&P500 index return over different periods of time.

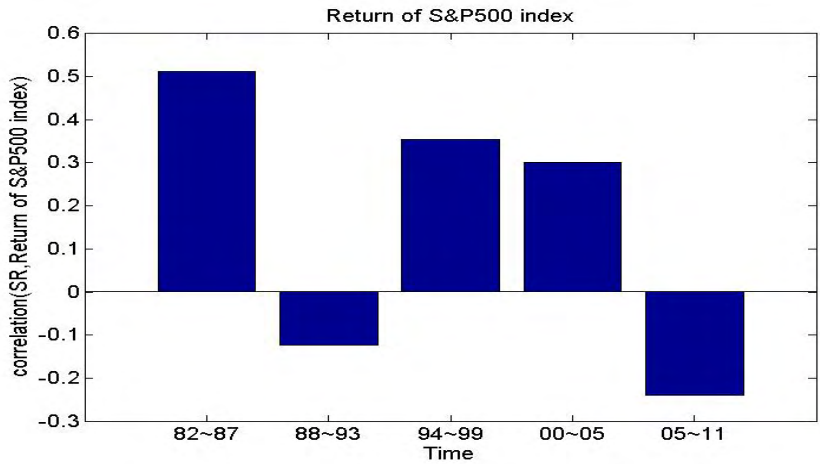


Figure 17 Time evolution of both systemic risk and volatility of S&P500 index.

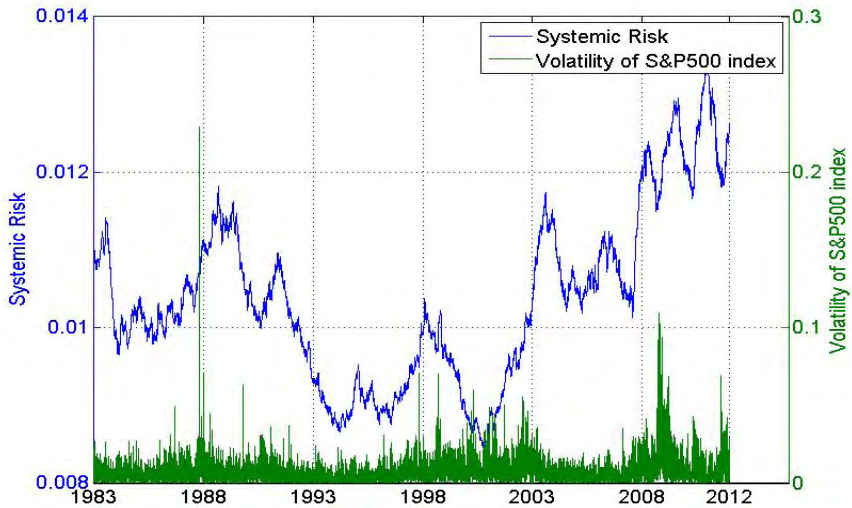


Figure 18 Correlation between SR and volatility of S&P500 index over different periods of time.

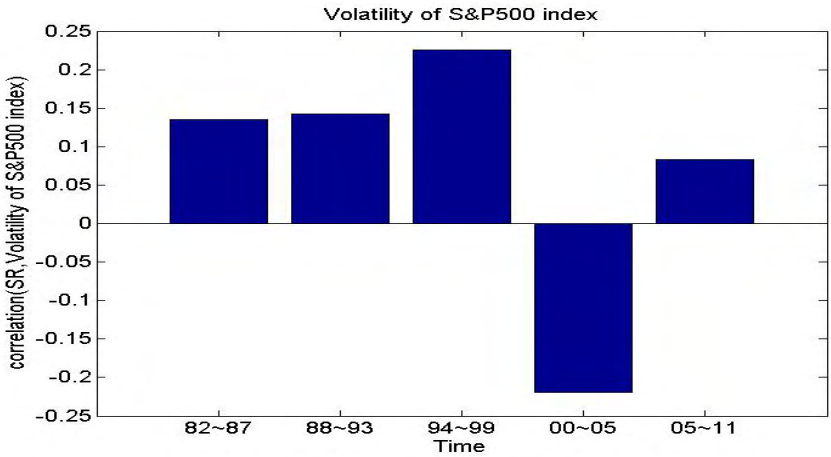


Figure 19 Time evolution of both systemic risk and unemployment rate

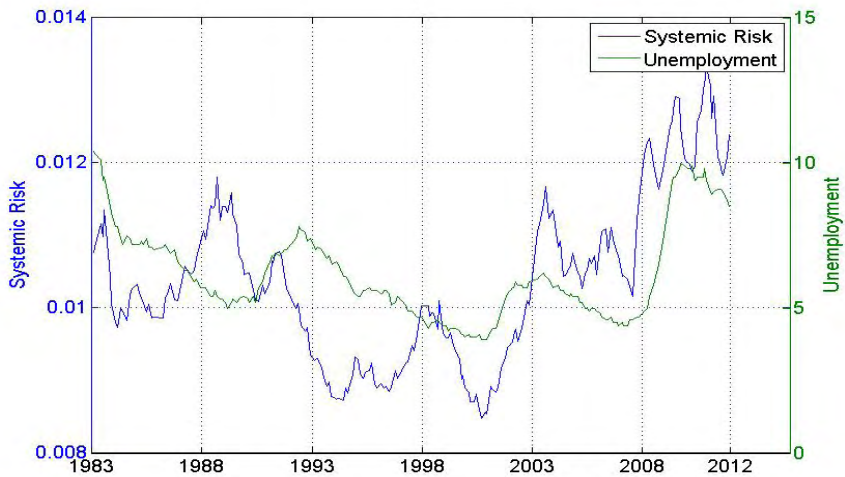
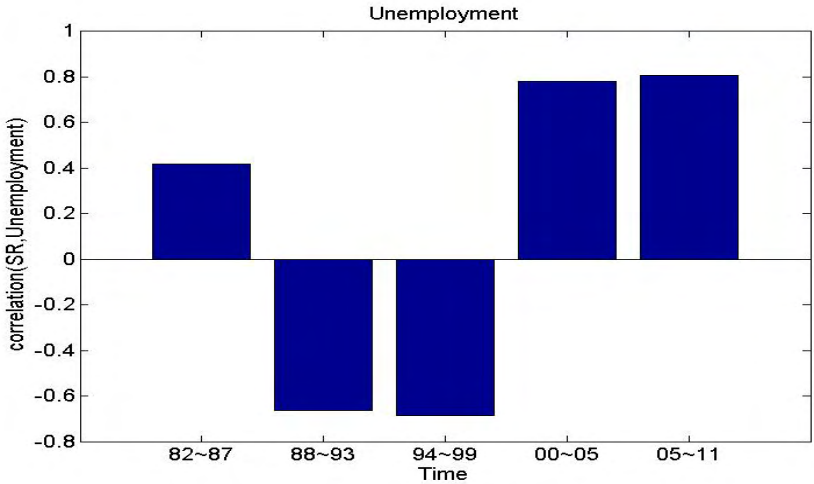


Figure 20 Correlation between SR and unemployment rate over different periods of time.



References

Acharya, Viral and Hassan Naqvi, 2012, The Seeds of a Crisis: A Theory of Bank Liquidity and Risk Taking over the Business Cycle, Journal of Financial Economics 106, 349-366.

Acharya, Viral V., Lasse H. Pedersen, Thomas Philippon, and Matthew Richadson, 2010, Measuring systemic risk, Unpublished working paper, New York University.

Adrian, Tobias, and Marcus Brunnermeier, 2010, CoVaR, Unpublished working paper, Princeton University and Federal Reserve Bank of New York.

Adrian, Tobias, and Shin, Song Hyun, 2008, Liquidity, Monetary policy, and Financial Cycles, Current Issues in Economics and Finance, Federal Reserve Bank of New York, 14(1): 1-7.

Adrian, Tobias, Moench, Emanuel, and Shin, Song Hyun, 2009, Financial Intermediation, Asset Prices, and Macroeconomics Dynamics, Federal Reserve Bank of New York, Working Paper Series.

Allen, F., Babus, A., Carletti, E. ,2012, Asset commonality, debt maturity and systemic risk. Journal of Financial Economics 104, 519–534.

Ashcraft, Adam Blair, 2006, New Evidence on the Lending Channel, 2006, the Journal of Money, Credit, and Banking, Volume 38, Issue 3, pp. 751-776, April.

Backus, D., Chernov, M. and ZIN, S. (2014), Sources of Entropy in Representative Agent Models, The Journal of Finance 69 (1), 51–99.

Bandt, C. and G. Keller and B. Pompe, Entropy of interval maps via permutations, Nonlinearity 15 (2002), 1595-1602.

Barnett, Lionell, Adam B. Barret, and Anil K. Seth, 2009, Granger causality and transfer entropy are equivalent for Gaussian variables, Physical Review Letters 103, .

Bernanke, Ben S. and Mark Gertler, 1995, Inside the Black Box: The Credit Channel of Monetary Policy Transmission, Journal of Economic Perspectives 9(4), 27-48.

Billio, Monica, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon, 2012, Econometric measures of connectedness and systemic risk in the finance and insurance sectors, Journal of Financial Economics 104, 535-559.

Bisias, Dimitrios Mark Flood, Andrew W. Lo, Stavros Valavanis, 2012, A Survey of

Systemic Risk Analytics, Office of Financial Research Working Paper #0001.

Blackwell, D. ,1953, Equivalent comparison of experiments," Annals of Mathematical Statistics, 24, 265-272.

Brunnermeier, Markus B., 2009, Deciphering the liquidity and credit crunch 2007-2008, Journal of Economic Perspectives 23, 77-100.

Brunnermeier, Markus K., and Pedersen, Lasse Heje, Market Liquidity and Funding Liquidity, Review of Financial Studies, 2009, 22(6): 2201-2238.

Cabrales, Antonio, Olivier Gossner and Roberto Serrano, 2013, Entropy and the Value of Information for Investors, AER Vol. 103, No. 1, FEBRUARY, 360-377.

Chaney, Thomas, David Sraer and David Thesmar, 2012, The Collateral Channel: How Real Estate Shocks Affect Corporate Investment(pp. 2381-2409), AER Vol. 102, No. 6, OCTOBER.

Cheng, Ing-Haw and Wei Xiong, 2013, The Financialization of Commodity Markets, NBER Working Paper No. 19642.

Chordia, Tarun, Roll, Richard, and Subrahmanyam, Avanidhar, 2001,Market Liquidity and Trading Activity, Journal of Finance, 56(2): 501-530.

Chordia, Tarun, Sarkar, Asani, and Subrahmanyam, Avanidhar, 2005,An Empirical Analysis of Stock and Bond Market Liquidity, Review of Financial Studies, 18(1): 85-129.

De Bandt, Olivier, and Philipp Hartmann, 2000, Systemic risk: a survey, European Central Bank working paper no.35.

De Nicolo, Gianni, and Marcella Lucchetta, 2010, Systemic Risks and the Macroeconomy, International Monetary Fund working paper no. 29.

Demyank, Y., and O. Van Hemert. 2011. Understanding the Subprime Mortgage Crisis. Review of Financial Studies 24:1848–81.

Drehmann, Mathias and Nikola Tarashev, 2011, Systemic importance: some simple indicators, Bank of International Settlement working paper.

Fama, Eugene, and Kenneth French 1997, Industry cost of equity, Journal of Financial Economics 43, 153-193.

Farhi, Emmanuel, and Jean Tirole. 2012. "Collective Moral Hazard, Maturity Mismatch, and Systemic Bailouts." American Economic Review, 102(1): 60-93.

Gan, Jie, 2007, The Real Effects of Asset Market Bubbles: Loan- and Firm-Level Evidence of a Lending Channel, Review of Financial Studies 20, 1941–1973

Granger, C. W., 1969, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37, 424-438.

He, Zhiguo, and Arvind Krishnamurthy, 2012, A macroeconomic framework for

quantifying systemic risk, working paper.

Jimenez, Gabriel, Steven Ongena, Jose-Luis Peydro and Jesus Saurina, 2012, Credit Supply and Monetary Policy: Identifying the Bank Balance-Sheet Channel with Loan Applications(pp. 2301-2326) AER Vol. 102, No. 5, AUGUST.

Kahle, Kathleen M., and Rene M. Stulz, 2011, Financial Policies, Investment, and the Financial Crisis: Impaired Credit Channel or Diminished Demand for Capital? Fisher College of Business Working Paper No. 2011-3.

Lehar, Alfred, 2005, Measuring systemic risk: A risk management approach, Journal of Banking and Finance 29, 2577-2603.

Longstaff,, Francis A., 2010, The Subprime Credit Crisis and Contagion in Financial Markets, Journal of Financial Economics 97, 436-450.

Marschak, J , 1959, Binary-Choice Constraints and Random Utility Indicators", in Arrow, Karlin and Suppes, editors, Mathematical Methods in the Social Sciences, Stanford University Press.

Saretto, Alessio and Heather E. Tooke, 2013, Corporate Leverage, Debt Maturity, and Credit Supply: The Role of Credit Default Swaps, Rev. Finance. Stud. 26 (5): 1190-1247.

Schreiber, Thomas, 2000, Measuring information transfer, Physical Review Letters 85, .

Schularick, Moritz and Alan M. Taylor, 2008, Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, AER Vol. 102, No. 2, APRIL, 1029-1061.

Sims, C. A. , 2003, Implications of rational inattention," Journal of Monetary Economics, 50(3), 665-690.

Staniek, Matthaus, and Klaus Lehnertz, 2008, Symbolic transfer entropy, Physical Review Letters 100.

Williamson, Stephen D., 2012, Liquidity, Monetary Policy, and the Financial Crisis: A New Monetarist Approach(pp. 2570-2605), AER Vol. 102, No. 6, OCTOBER.

CHAPTER 9

Monetary Policy Transmission Via Risk-taking Channel in the Mortgage Market*

by
*Min-Ho Nam***
(Bank of Korea)

Abstract

This paper aims to analyze the transmission of accommodative monetary policy to the overall economy through the risk-taking channel operating in the mortgage market. To achieve this aim, the analysis procedure undergoes two steps. Firstly, the empirical relationship between short-term interest rates and LTV ratio is estimated using the U.S. data to verify the existence of the risk-taking channel. Secondly, the estimated relationship is incorporated into a DSGE model featuring a borrowing constraint and housing to construct the virtual economy in which the channel takes effect to analyze the impacts of a monetary policy shock on it. The results of the analysis suggest that under a low interest rate environment, the effects of the risk-taking channel should be taken into account in monetary policy analysis as it amplifies the impacts of a monetary policy shock. Furthermore, there is a scope for policy authorities to smooth both real and financial volatilities by lowering a ceiling on LTV ratio to discourage excessive risk-taking.

* The views expressed in this paper are solely my own and should not be interpreted as reflecting the views of the Bank of Korea. I thank seminar participants at the 2014 KDI Journal of Economic Policy Conference for useful discussions.

** Research Department, Bank of Korea (E-mail: minho@bok.or.kr, Tel: +82-2-759-4241)

1. Introduction

It has been maintained that the deregulation and liberalization of housing finance since the 1980s have broadened credit availability which in turn has led to more pronounced fluctuations in housing prices. Specifically, banks and governmental mortgage agencies were allowed under the liberalization process to produce a wide range of mortgage loan products, set lending interest rates at their own discretion, determine the level of the loan-to-value (LTV) ratio based on their own judgement rather than regulatory prescriptions, and so forth. Moreover, non-bank financial corporations were given permission to enter the mortgage market, thereby heightening the degree of competition in this market.[1] Consequently, easier credit supply to the housing market has increased the volatility of the demand for housing and house prices, and in turn amplified further repercussions of the housing sector on consumption and residential investment. These developments in housing finance retain substantial implications for the analysis of monetary policy transmission. Recent findings support the view that the financial liberalization process has rendered the housing sector and the rest of the economy more responsive to a monetary policy shock as interest rates affect credit availability more significantly in a deregulated environment (Iacoviello and Minetti, 2003; IMF, 2008; Calza et al., 2009).[2] [3]

The aforementioned findings are obtained under an assumption that

---

1 The characteristics of housing finance in each country is distinct. IMF (2008) and Calza *et al*. (2009) provide indicators of the differing developments in mortgage financing in industrialized countries. ECB (2009) surveys the recent circumstances of housing finance in the Euro area since 1999.

2 Regardless of the recognized importance of credit availability in housing finance, there exist a limited number of findings about the relationship between credit availability and house prices. The reason for this is that there are few trustworthy measures of *credit availability* itself. Even though the amount of mortgage debt seems a plausible proxy for the variable, the amount registered in banks' books is not a proper measure since it is the realized value of *credit availability*. Furthermore, changes in mortgage depend on other factors besides it.

3 Another relevant finding is in Almeida *et al.* (2006) which confirms that the response of house prices to income shocks is more rapid in those countries having a higher ceiling on the LTV ratio.

mortgage market characteristics are exogenously determined by the financial deregulation process. Recently, a line of research has raised the issue of a possible causal link between an accommodative monetary policy stance and bank lending behavior. Researchers pursuing this line of reasoning highlight the observed facts regarding lending markets in the run-up to the recent housing boom which numerous developed countries underwent. During that period, lending criteria were loosened appreciably, the minimum down-payment decreased considerably while policy rates were deemed relatively lower than a specific judgement criterion, for example, the Taylor rule or estimated neutral interest rates. The crux of the findings of these researchers is that low interest rates for such a protracted period increased banks' appetite for higher risk in lending and other investments. This transmission channel of monetary policy is labeled the *risk-taking channel* by Borio and Zhu (2008). However, the underlying rationale for the *risk-taking channel* has been highlighted by central bankers. Greenspan (2010), for example, ascribes the failure of the banking system during the recent financial crisis to the possibility that the prolonged period of a relaxed policy stance might have driven banks to neglect the negative tail of investment risk (Greenspan, 2010); this comment implies that the overall perception of risk was positively biased. Voices from the European Central Bank (ECB) have expressed apprehension from a similar viewpoint; low interest rates for a prolonged period abet moral hazard in banks' investments imbuing them with a myth that the cental bank may not be able to reverse interest rates rapidly because of worries about asset market collapse (ECB, 2005; Trichet, 2005; Papademos, 2006).[4]

These ref1ections provide a motivation to consider the relationship between low interest rates and banks' risk taking attitude. I submit the hypothesis that a positive monetary policy shock causes banks to raise the LTV ratio and supply ample liquidity to the housing market, thereby rendering the path of house prices and consumption more volatile. In a nutshell, this paper has two aims: firstly, to verify the existence of the risk-taking channel in the mortgage market, and secondly, to estimate

4 The government's intervention to relieve troubled banks through bailout programs also has been referred to as a cause of the 'too big to fail' myth.

the repercussions of an expansionary monetary policy shock on the economy as a whole via this channel. These aims are attained through a two-stage analysis. To examine the existence of the *risk-taking channel*, namely, a negative relationship between monetary policy rates and the LTV ratio, two kinds of empirical analyses, i.e. regression and VAR, are conducted using U.S. time series data. Although there are various indicators of the degree of banks' risk-taking, the LTV ratio is chosen as an effective one as mortgage lenders tend to set the ratio depending on their own perception of the risk latent in housing-collateralized lending. Subsequently, a DSGE model is developed incorporating an estimated regression equation for the *risk-taking channel* to analyze its role in a broader economy. In the DSGE model, the LTV ratio is defined as a function of policy rates and house price inflation and is set less than unity. This variable plays a key role in amplifying and propagating an initial shock to the economy.[5] In addition, the model follows the lead of Iacoviello (2005) and Iacoviello and Neri (2010). These papers adapted the financial accelerator mechanism of Kiyotaki and Moore (1997) to investigate the dynamics of the housing sector and its spillover into the rest of the economy.

There are two notable contributions in this paper which set it apart from the rest of the literature on bank risk-taking and financial friction in lending. Firstly, this analysis is the first attempt, to the best of my knowledge, to delve into the effects of the *risk-taking channel* employing a general equilibrium framework. Secondly, the model in this paper *endogenizes* the LTV ratio for the first time.

To elaborate on the second point, the LTV ratio in existing models is assumed to be a fixed constant (Monacelli, 2008; Calza *et al.*, 2009; Iacoviello and Neri, 2010). An improvement over the constant LTV ratio is the assumption of a time-varying exogenous stochastic process as in Pariés & Notarpietro (2008) and Gerali *et al.* (2010), but still this stochastic ratio is not affected by other variables in the model. However, in practice, since banks adjust the LTV ratio on the basis of an

5 The ceiling on the LTV ratio in practice can exceed one, as in the U.S. which raised the maximum ratio up to 125%. However, only a small portion of borrowers can take advantage of this ceiling as other income requirements and lending criteria should be satisfied.

evaluation of default risk and the redeemable value of collateral in case of foreclosure, the existing way of treating the LTV ratio in economic models is clearly unsatisfactory. Endogenizing this ratio bases the model more firmly on realistic aspects of housing finance. An additional advantage of introducing the LTV ratio in this manner is the resultant parsimony of the model. As opposed to the models in, for instance, Goodfriend and McCallum (2007), Cúrdia and Woodford (2008), and Gerali *et al*. (2010) which introduce a separate block for financial intermediation, the supply side of credit can be reflected in my model by allowing the LTV ratio to vary depending on the banks' decision.

The main findings of the analysis are twofold. First, the results from the regression and VAR analysis lend firm support to the assertion that there is a negative relationship between short-term interest rates and the LTV ratio. It implies mortgage suppliers have tended to be more aggressive in housing-collateralized lending in the period of low interest rates. Secondly, a positive monetary policy shock in the model with the *risk-taking channel* included produces enhanced deviations of consumption and financial debt from the steady state than the model without this channel. These findings can shed light on the conundrum why central bankers, before the sub-prime crisis, failed to forecast the catastrophic results stemming from low interest rates for a prolonged period; presumably they dismissed the *risk-taking channel* when analyzing the transmission effects of their monetary policy decisions. Furthermore, the results justify the need for central banks to pay more attention to the possible existence of more as yet undiscovered transmission channels of monetary policy and for financial supervisory authorities to regulate banks' risk-taking behavior.

The remainder of this paper is organized as follows. Section 3.2 provides a brief review of the rationale for, and summarises existing findings on, the *risk-taking channel*. It also presents an explanation for the working of the *risk-taking channel* in bank lending and its repercussions on the housing market and broader economy. In section 3.3 empirical evidence for the *risk-taking channel* in the U.S. mortgage banking sector is presented. In section 3.4, a baseline DSGE model is developed and, in the following section, the monetary policy transmission is analyzed in the absence of the *risk-taking channel*. In

section 3.6, the *risk-taking channel* is accommodated in the baseline model to examine how the transmission effects change in the presence of the channel. Section 3.7 sets out the conclusion.

2. Risk-taking Channel and Mortgage Lending

This section provides a short summary of the theoretical considerations underlying the *risk-taking channel* and summarises relevant empirical findings. I will then demonstrate the implications of the channel for the mortgage market and its impact on the housing market and the economy as a whole.

2.1 Rationale and Empirical Findings

The *risk-taking channel* was introduced explicitly by Borio and Zhu (2008) as an additional monetary policy transmission channel. It is defined as follows:

> the *risk-taking* channel in the transmission mechanism (is) defined as the impact of changes in policy rates on either risk perceptions or risk-tolerance and hence on the degree of risk in the portfolios, on the pricing of assets, and on the price and non-price terms of the extension of funding. (*Ibid*, pp. 9)

This issue has received growing interest in academia due to the failure of the global banking system caused partly by the excessive risk-taking behavior of mortgage suppliers observed in the first half of the last decade. The Fed maintained that the overheated housing market prior to the crisis was not associated with past monetary policy decisions (Greenspan, 2010; Bernanke, 2010). However, an alternative possibility was proposed that an accommodative monetary policy stance for the extended period might affect the risk-taking behaviour of economic agents, especially financial intermediaries. Thereafter, the main focus has been on the causal chain or correlation between an easy policy stance and banks' risk-taking. A preliminary consensus has been established that two necessary conditions must be fulfilled: 'too low'

interest rates as the first condition and for 'too long' a period as the second one.

The question naturally follows: in what ways does a loose monetary policy stance encourage banks to take more risk? There are three possibilities. The foremost and fundamental driver, from my viewpoint, would be the tendency of 'search for yield' in the period of low interest rates as noted by Rajan (2006) and others. To take an example from the banking sector[6], the yields of bonds are more likely to be lower than these of other investments such as stock and collateralized lending to households. While the exposure to stock is circumscribed within a ceratin level since it is classified as a highly risky asset and hence harmful for satisfying minimum capital requirements set by the BIS (Bank for International Settlements), collateralized lending ensures higher profitability and also limited possible loss in the case of default. These two attractions drive banks to expand lending against housing as collateral by loosening lending criteria and increasing lending to borrowers with low creditworthiness.

Secondly, during a period of low monetary policy rates coupled with moderate economic growth as during the Great Moderation, banks are more likely to neglect the possibility that assets held by them can turn sour or non-performing and borrowers' real income growth can become negative in the future. To apply this reasoning to the mortgage market, if the value of collateral and the income level of borrowers were increasing, then mortgage suppliers would perceive the risk in lending as lower than they would otherwise. The underestimation of risk results mainly from the expectation that robust growth in income and collateral value will persist into the distant future. Finally, as pointed out by the ECB sources, banks are more likely to undertake riskier and more profitable investments as long as interest rates remain low in the belief that the lender of last resort will come to the rescue in order to prevent the overall economy from collapsing. If it were not for the 'too big to fail' myth, preference for a riskier position could be subdued to some extent and the

6 Investment banking or shadow banking is not considered in this example.

degree of moral hazard could be lessened.[7]

Empirical findings on the *risk-taking channel* has been increasing recently. Drawing on expansive and detailed data on individual bank loans from the Spanish Credit Register in the period from 1984 to 2006, Jiménez *et al.* (2008) find that lower overnight lending rates cause banks to loosen lending criteria and expand credit line to borrowers with mediocre credit records despite higher default risk. These findings gain support from Ioannidou *et al.* (2009) who provide the evidence that the default probability of bank loans rises and lending to riskier borrowers tends to increase in Bolivia as the U.S. federal funds rates (FFR) decrease.[8] In another study, Altunbas *et al.* (2010) investigate whether low interest rates affect the risk position of 643 banks in 15 industrialized countries using balance sheet data for the period from 1998 to 2008. Banks' risk is measured by the expected default frequency (EDF), an indicator for the probability that a company will default in a certain time horizon. They find that the low short-term interest rates for a sustained period caused an increase in banks' default risk. Gambacorta (2009) provides evidence on the negative relationship between interest rates and banks' default risk by using an econometric approach similar to the one employed in Altunbas *et al*. (2010).[9] Delis and Kouretas (2011), by consulting the balance sheet information of banks in the Euro area during the period 2001-2008, find that the ratio of risky asset value to total asset value, as well as the ratio of non-performing loans to total loans, increased. Maddaloni and Peydró (2011) focus explicitly on the influence of monetary policy rates on relaxed lending standards. By using the responses from bank lending surveys carried out in the Euro area and U.S., they identify the positive influence of an accommodative policy stance on the loosening of lending

---

7 Essentially the first and second points are in line with Borio and Zhu (2008) who explain how the *risk-taking channel* works in general instead of focusing on bank lending.

8 The authors maintain that the U.S. FFR is a proper measure of Bolivian monetary policy stance since over 90 percent of Bolivian deposits and credits are transacted in the U.S. dollar.

9 The data used is obtained from the balance sheets of 600 banks in Europe and U.S. during the period 2007-2008.

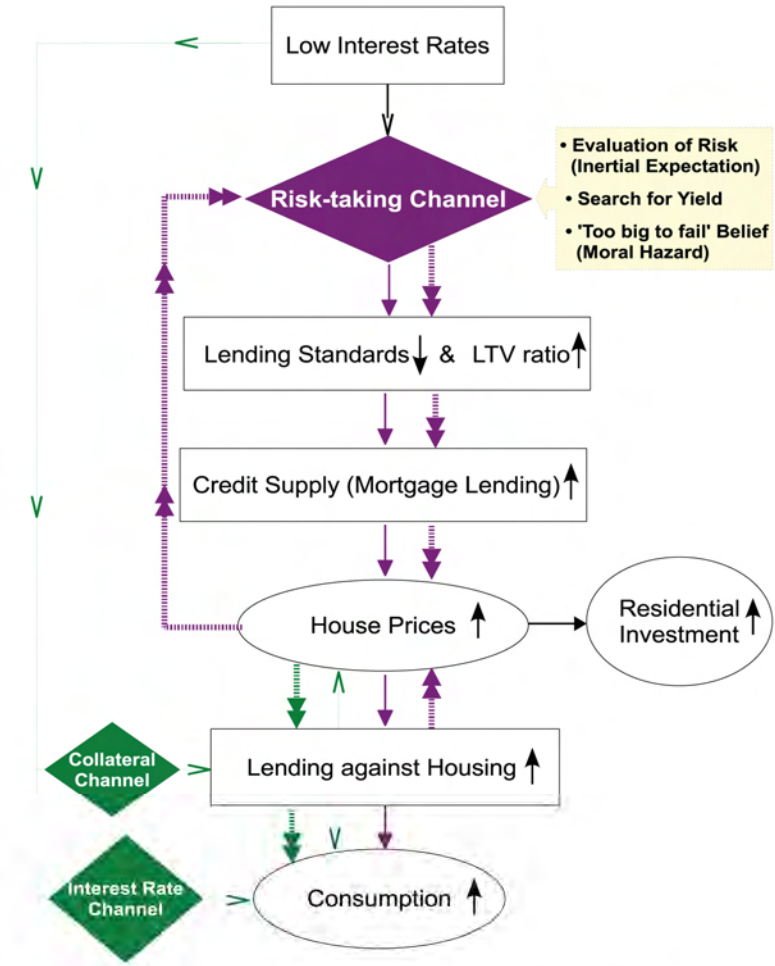
standards during the period from 2002 Q4 to 2008 Q3.

A similar strand of research in the U.S. has also recently been in the spotlight. Adrian and Shin (2009, 2010b) stress the importance of the role of short-term interest rates in generating business cycle by causing dramatic changes in the banking sector's credit supply. In a recent paper, Adrian and Shin (2010a) introduce the concept of *risk-taking channel* and maintain that banks are liable to estimate risk as lower and hence take a riskier investment position as lower short-term interest rates widen the margin between the interest rate on deposits and return on the assets in the balance sheet. However, their research is theoretical and hence needs sound support from empirical research.

2.2. Implications of Risk-taking Channel for Mortgage Market and Economy

The empirical studies reviewed in the previous sub-section suggest that lower interest rates lead banks to soften lending criteria and supply more credit than they would otherwise. The specific dependent variables in these analyses include the probability of banks' default in the future, the ratio of risky asset value to the total asset value and the percentage of banks tightening their lending standards. Given the importance of the effect of leverage on general consumption and housing purchases, the LTV ratio needs to be added to the list of measures of banks' risk appetite. The rationale for considering the LTV ratio as a measure of risk-taking attitude is consistent with the rationale for the *risk-taking channel*. During a prolonged period of an accommodative monetary policy stance, collateralized lending to households satisfies the dual targets of profitability and safety. This leads banks to expand lending against housing as collateral by raising the LTV ratio even though the collateral value stays constant. There are additional factors inducing banks to lower the LTV ratio. As long as house price inflation triggered by low interest rates continues, lenders would take the default risk of borrowers less seriously compared with the period of a bearish housing market. Furthermore, if low interest rates are maintained for an extended period, expectation of robust house prices in the future would encourage complacency in evaluating the risk of mortgage lending. Lenders can

Figure 1 | Risk-taking Channel in Mortgage Market



Note: 1. Rectangles and ellipses represent changes in financial and real variables respectively.
2. The thick dashed line represents the feedback effects between the risk-taking channel, housing market and mortgage.

also decrease the price of lending as long as households' net worth is increasing given the low interest rate environment. As such, the realized appreciation of housing prices and expectation about further increases induce banks to perceive the overall risk of mortgage lending as lower

and to increase the LTV ratio. Enhanced credit supply and higher value of housing will persist until interest rates reverse their direction.

The *risk-taking channel* operating in the mortgage market has unambiguous implications for the wider economy. More funds would be available to households than in the absence of the channel. The funds borrowed against housing are spent not only on purchasing houses but also on consuming other durable and non-durable goods. Residential investment increases as the demand for housing expands fueled by ampler liquidity with low borrowing costs. Notable is that once the channel begins to operate, a self-reinforcing feedback loop would come into play between risk-taking, mortgage lending supply, house prices and real economic activity. Figure 1 illustrates the causal chain running from low interest rates to the housing market and macroeconomic activities via the *risk-taking channel*.

3. Empirical Estimation of Risk-taking Channel

In this section, we examine the presence of the *risk-taking channel* in the U.S. mortgage market. Two empirical methodologies are employed: simple regression and VAR approach. However, before conducting the analysis, we first discuss some relevant aspects of the data on the LTV ratio.

3.1. Data

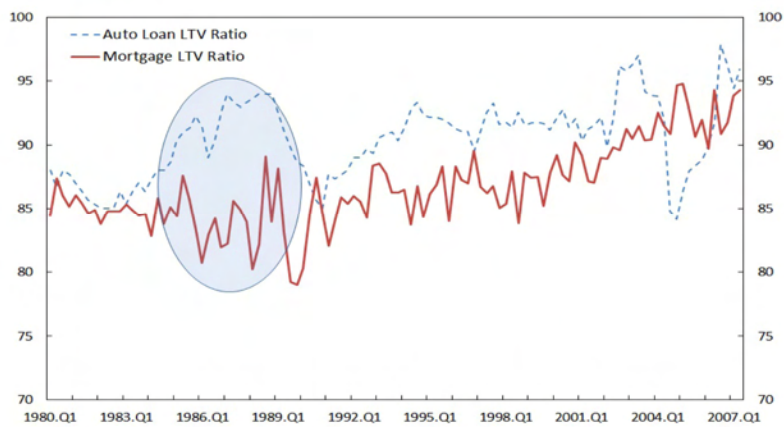
There does not exist an officially compiled historical series on the LTV ratio. Hence, one has to depend on work done by other researchers to obtain this database. In this regard, a special comment should be given about the LTV time series estimated by Duca *et al*. (2011) for first-time home buyers in the U.S. The series is very useful for researchers in the field of housing since the frequency is very high (quarterly) and the time span extends as far back as 1979.[10] The

10 Using the estimated data, the paper evaluates the influence of credit condition changes on house prices.

estimated data can be broadly classified into two types depending on the type of mortgage issuers; private mortgages and all types of mortgages including those from government-sponsored agencies such as the Federal Housing Finance Agency (FHFA). The series for private mortgages is considered more pertinent for an analysis of the *risk-taking channel* since the mortgages from the FHFA omit non-standard loans which convey substantial information on the risk-taking behavior of private mortgage lenders.

The data series has proven to be highly reliable judging by its close co-movement with the data on the auto loan LTV ratio published officially by the Fed.[11] The property of an automobile as a durable good justifies the comparison. Figure 2 shows the long-term series of the two kinds of LTV ratios regarding home mortgages and auto loans, respectively, from 1980 Q1 to 2007 Q4. These two time series appear to have an upward trend, although there is pronounced decoupling between them during the period from 1985 to 1989. Especially after 1990, the co-movement of the two trends is pronounced in the same figure.

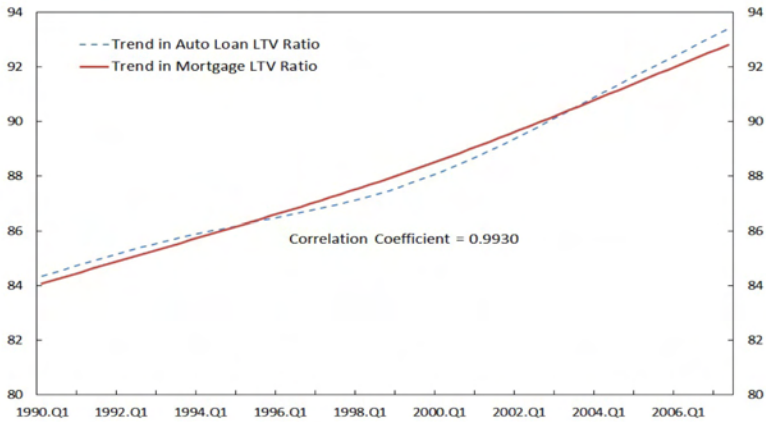
Figure 2 Auto Loan and Home Mortgage LTV Ratio in U.S.



Data Source: Federal Reserve, Duca *et al.* (2011)

11 The data can be downloaded from the 'Terms of credit' menu on the webpage of http://www.federalreserve.gov/releases/g19/hist

Figure 3 | Trend Components of Auto Loan and LTV Ratio



Data Source: Federal Reserve, Duca *et al*. (2011)

To evaluate the correlation between the trends of both time series, the trend components of each time series are extracted using the Hodrick-Prescott (HP) filter with the smoothing parameter 100,000. The trend components are graphed in Figure 3. Not surprisingly, and as expected from the original data series, the correlation coefficient turns out to be 0.9930 which implies a near perfect co-movement. This lends great plausibility to the data series estimated by Duca *et al*. (2011) and we will use this in our subsequent estimations.

3.2. Estimation

3.2.1. Regression Analysis

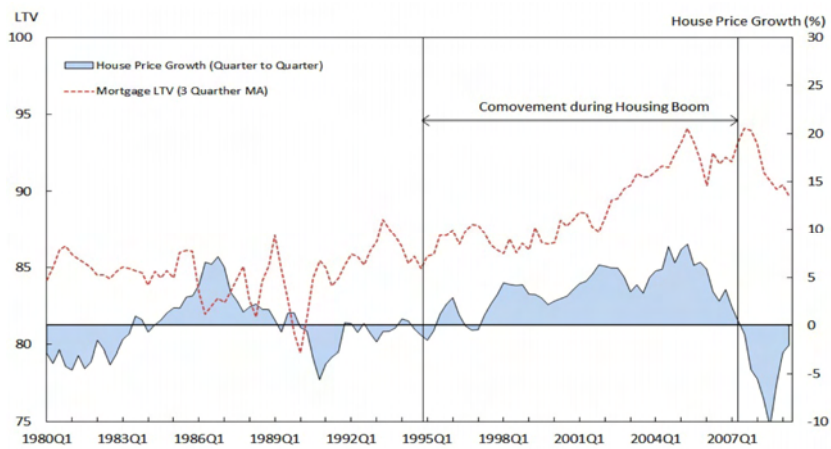
To examine whether the *risk-taking channel* exists or not in the U.S. mortgage market, that is, if lenders take on more risk by increasing the LTV ratio as interest rates decrease, the following regression equation is estimated with the LTV ratio as the dependent variable.

$$LTV_t = \alpha + \beta_1 LTV_{t-1} + \beta_2 FFR_{t-1} + \beta_3 HPG_{t-1} + \epsilon_t \qquad (1)$$

The determinants of the LTV ratio in the current period, LTV_t,

includes the one-quarter lagged LTV ratio, LTV_{t-1}, the lagged Federal Fund Rates (FFR_{t-1}), which are the overnight interest rates fluctuating closely around policy rates, the lagged growth rates of real house prices, HPG_{t-1}which is computed using the National House Price Index (NHPI) published by FHFA. The equation above hypothesizes that mortgage lenders set the LTV ratio of the current period based on the level of short-term interest rates and house price inflation in the previous period while avoiding overly rapid changes by adjusting the LTV ratio in the previous period to a small extent. The inclusion of house price inflation is justified on the basis of the reasoning about the *risk-taking channel*. As house prices continue to rise for a prolonged period, lenders are more likely to focus on the positive side in the distribution of the housing price risk and hence increase the LTV ratio. The apparent positive correlation in Figure 4 between the house prices and the LTV ratio in the U.S. from 1995 to 2007 seems to support this speculation.

Figure 4 Mortgage LTV Ratio and House Price Growth in U.S.



Data Source: Federal Reserve Board, Duca *et al.* (2011)

The regression results for two different time periods are shown in Table 1. The first time period for estimation is from 1980 Q1 to 2009 Q2 to utilize all the data on the regression variables. The coefficient of FFR_{t-1}, which is of main interest, is significant at the 10% significance

level as well as that of HPG_{t-1}. The signs of the coefficients are consistent with the *risk-taking channel* hypothesis above; low interest rates and the robust housing market induce lenders to assume more risk by raising the LTV ratios and become more willing to supply credit.

Table 1 Estimation of LTV Equation using Level Data

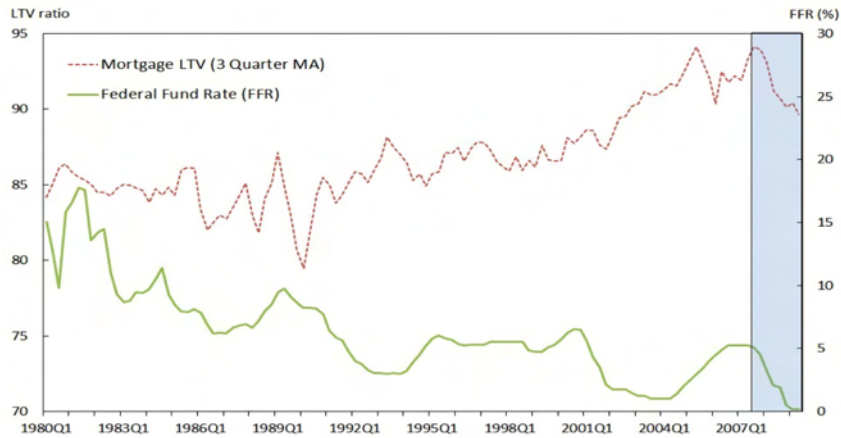
	1980 Q1 - 2009 Q2			1985 Q1 - 2007 Q2		
	value	*t* - statistic	*p*-value	value	*t* - statistic	*p*-value
α (constant)	28.855***	4.701	0.000	40.392***	5.124	0.000
$\beta_1(LTV_{t-1})$	0.676***	9.917	0.000	0.558***	6.466	0.000
$\beta_2(FFR_{t-1})$	-0.135*	-1.974	0.051	-0.425***	-3.195	0.002
$\beta_1(HPG_{t-1})$	0.104***	1.729	0.087	0.146*	1.692	0.094
	$R^2 = 0.643, DW = 2.283$			$R^2 = 0.649, DW = 2.153$		

In the regression for the second time period from 1985 Q1 to 2007 Q2, the coefficients of these two variables prove to be significant again.[12] Notably, the coefficient of FFR_{t-1} is significant even at the 1% significance level. The statistical significance improves since the new time period excludes the data from 2007 Q2 to 2009 Q2. Over theose two years, the housing market crash subdued the LTV ratio despite appreciable decreases in policy rates, as the shaded area in Figure 5 illustrates. Overall, the two sets of regression results corroborate the existence of the *risk-taking channel* in the U.S. and support inclusion of the channel when we inspect the effects of monetary policy decisions on the mortgage market. In terms of macroeconomic modeling, this implies that the LTV ratio should be allowed to vary based on changes in interest rates and house prices to analyze the full aspects of monetary policy transmission.

Since the equations estimated using the level data are not suitable for the log-linearized DSGE model to be presented in the following section,

12 The motivation for starting the sample period from 1985 lies in the possibility that the financial liberalization might begin exerting its real influence only after the mid-1880s. Moreover, the samples after the breakout of the sub-prime crisis are excluded because the relationship between interest rates and house prices in the post-crisis period is positive which is abnormal from the viewpoint of the established empirical findings.

Figure 5 Mortgage LTV Ratio and Federal Fund Rates in U.S.



Data Source: Federal Reserve Board, Duca *et al.* (2011)

a separate equation needs to be estimated using detrended or demeaned data. As the overall fitness of regression using the detrended data by HP filter is not satisfactory, demeaned data are used for all variables in the regression equation. Demeaning implies that the long-term average is assumed to be the steady state of each variable. The regression equation to be estimated is given below. [13]

$$\widehat{LTV}_t = \gamma_1 \widehat{LTV}_{t-1} + \gamma_2 \widehat{FFR}_{t-1} + \gamma_3 \widehat{HPL}_{t-1} + \epsilon_t \qquad (2)$$

where the hatted variables represent percentage deviation from the steady state and *HPI* represents the level of the house price index from the FHFA (instead of the growth rate denoted by *HPG* in the preceding regression analysis).[14]

---

13 Not only for the purpose of obtaining a dynamic solution to the model but also by the assumption of naive type of backward-looking expectation, the dependent variables take only one-period lagged terms.

14 A separate regression equation is estimated which includes *realized* federal funds rates instead of *nominal* ones. The estimation results reveal that the magnitude of the coefficient of γ_2 is slightly below the level obtained by the estimation of the equation including nominal *FFR*. This implies that the regression result is robust.

Table 2 | Estimation of LTV Equation using Demeaned Data

	1980 Q1 - 2007 Q2			1995 Q1 - 2007 Q2		
	value	t - statistic	p-value	value	t - statistic	p-value
$\gamma_1(\widehat{LTV}_{t-1})$	0.427***	4.885	0.000	0.329*	2.012	0.036
$\gamma_2(\widehat{FFR}_{t-1})$	-0.577	-1.335	0.185	-3.207***	-4.370	0.000
$\gamma_3(\widehat{HPI}_{t-1})$	0.128***	4.568	0.000	0.181***	6.235	0.000
	$R^2 = 0.676, DW = 2.045$			$R^2 = 0.792, DW = 2.214$		

Table 2 shows the two sets of regression results for the different time periods. For the first time period from 1980 Q1 to 2007 Q2, the coefficient of $\widehat{FFR}_{t-1}$ turns out to be insignificant even at the 10% significance level, whereas it is highly significant for the second time period from 1995 Q1 to 2007 Q2. Based on the statistical significance, the regression equation for the latter time period is set as the benchmark LTV equation to be incorporated into the DSGE model. To express it in an equation form, the estimated equation of (2) is given by

$$\widehat{LTV}_t = 0.211\widehat{LTV}_{t-1} - 3.207\widehat{FFR}_{t-1} + 0.181\widehat{HPI}_{t-1} + \epsilon_t \qquad (3)$$

The estimated coefficients imply that lenders respond more aggressively to the deviation of short-term rates from the steady state than that in house prices. A basic intuition is provided by the fact that the *risk-taking channel* induces low interest rates to influence lenders' behavior in more ways than house prices do. In a low interest rate environment, lenders expect higher house prices, search for higher yield and estimate downside risk to collateral as lower.

The error term, ϵ_t, which is a shock to the LTV decision process, retains an important implication for the housing market. The shock encompasses, for instance, the changes in regulation relating to the discretion of mortgage lenders to decide on their LTV ratios, the invention of new lending products such as exotic mortgages, and changes in the degree of information asymmetry between lenders and borrowers. For example, if the financial authorities grant more latitude to mortgage lenders in determining the LTV ratio or allow them to sell mortgage products with a smaller downpayment, then the ratio will increase given a specific level of house prices and interest rates.

3.2.2. VAR Analysis

A VAR analysis is now employed to examine the existence of the risk-taking channel. A specific representation of the relationship to be identified by VAR will differ from one obtained through regression analysis between the three variables of interest, i.e. short-term interest rates, house prices and the LTV ratios. However, as each variable in a VAR model is also expressed in a similar type of equation to a regression equation with the LTV ratio as a dependent variable, the same qualitative aspect of the *risk-taking channel*, if it exists, can be ascertained through an impulse response analysis. The specific VAR model is defined as below.

$$\Gamma_{yt} = c + A(L)_{yt} + \Sigma_{e_t} \quad (4)$$

where y is a vector of endogenous variables, $A(L)$ is the parameter matrix in the lag operator L, and Σ is the variance-covariance matrix of the structural shocks. The vector y includes three endogenous variables, (i) the federal funds rates denoted by *INT*, (ii) the growth rates of the house price index, *HPG*, and (iii) the LTV ratio. The model is estimated using the same data as in the regression analysis from 1980 Q1 to 2007 Q2. In order to let the impulse response of the LTV ratio be more sensitive to a monetary policy shock, the data after 2007 Q2 is excluded.

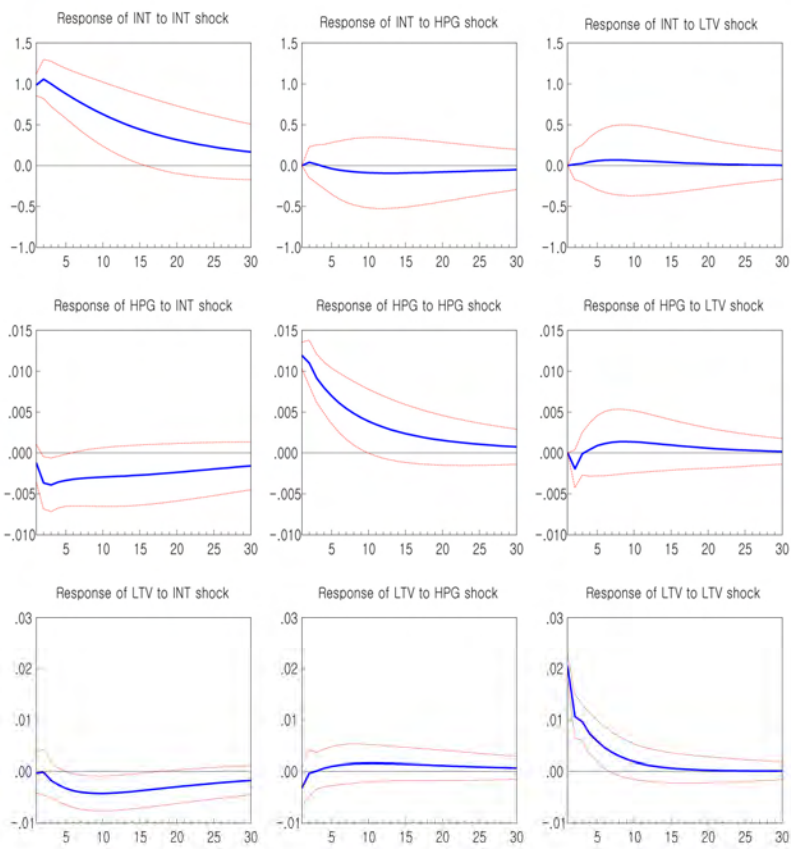
The three structural shocks from the model are identified through Cholesky decomposition which include a monetary policy shock and shocks to house prices and the LTV ratio. The endogenous variables are ordered as follows:

$$y_t = [INT_t, HPG_t, LTV_t] \quad (5)$$

This recursive identification scheme restricts interest rates, *INT*, from responding contemporaneously to the house price growth (*HPG*) and the LTV ratio. In a similar vein, the LTV ratio bears no influence on *HPG* in the same quarter. These restrictions are harmonious with the fundamental purpose of this VAR analysis for diagnosing the existence of the *risk-taking channel*, despite the common observations that these

three variables exert influences on each other simultaneously. The lag order of the model is set as 2 based on Schwarz information criterion and the F -statistics for model reduction.

Figure 6 Impulse Response from VAR Analysis



Note: 1. *INT* and *HPG* denote Federal Fund Rates and the growth rate of realized NHPI.
2. Confidence bands are based on the 95% significance level.
Data Source: Federal Reserve, Federal Housing Finance Agency, Duca *et al.* (2011)

Figure 6 includes the impulse responses of each endogenous variable to the three structural shocks. Our main interest lies in whether the LTV ratio responds negatively to a monetary policy shock. In addition, we can see whether the LTV ratio reacts positively to *HPG* as indicated by

the results of the regression analysis presented above. The left panel of the bottom row in the same figure is of primary interest to us. It shows that the response path of the LTV ratio to a monetary policy shock and it is consistent with the hypothesis presented above on the *risk-taking channel*. A positive increase of 100bp in short-term interest rates leads the LTV ratio to decrease by a maximum of 0.5%. This implies that an expansionary monetary policy shock will result in higher the LTV ratio through the *risk-taking channel*. The impulse response function is consistent with the regression results presented above.

The second panel in the bottom row of that figure shows that a positive shock to house prices increases the LTV ratio. The result also confirms the legitimacy of the regression result despite the relatively subdued magnitude of the response of the LTV ratio. As stated previously, lenders are likely to underestimate the risk latent in housing-collateralized lending as housing prices appreciate. Another notable feature of this impulse response analysis is the response of *HPG* to a contractionary monetary policy shock. The first panel in the middle row is consistent with recent findings using a VAR approach (Bjørnland and Jacobsen, 2010; Musso *et al*., 2010) in that an increase in short-term interest rates induces housing prices to deflate.

4. Developing a DSGE Model

In this section, a DSGE model will be developed to analyze the influences of the *risk-taking channel* on the overall economy. The model features two types of households which are patient households (savers) and impatient households (borrowers). Households supply firms with labor as an input to production and spend their labor income to accumulate residential housing and consume other goods. Savers who are more patient than borrowers save a fraction of their labor income and lend the funds to borrowers facing a borrowing constraint. In return for the funds lent, savers earn interest from borrowers. Firms produce wholesale consumption goods using only labor. Monopolistically competitive retailers buy the intermediate goods from firms and price these goods for sale. However, as in Calvo (1983), only a certain proportion

of retailers can adjust retail prices subject to the predetermined probability that a random signal arrives. The restriction on price re-optimization introduces nominal rigidities into the model.

A simplifying assumption is introduced regarding the use of housing. Housing in this economy is solely for residential purposes unlike that in Iacoviello (2005) and Iacoviello and Neri (2010) where it is used for production purposes as well. In these two papers, the production technology comprises housing as an input. This simplification makes the model consistent with the estimation of the LTV ratio equation in Section 3 since data comprising only home mortgages was used for the estimation.[15]

4.1. Patient Households

There is a continuum of identical patient households (savers) denoted by P. A representative patient household maximizes a lifetime utility function given as below.

$$E_0 \sum_{t=0}^{\infty} \beta_P^t \left[\ln c_t^P + j \ln h_t^P - \frac{(L_t^P)^{\varphi}}{\varphi} + \chi \ln \left(\frac{M_t^P}{P_t} \right) \right] \tag{6}$$

Consumption c_t^P, holding of housing h_t^P and real money balances $\frac{M_t^P}{P_t}$ affect the level of utility positively whereas hours worked L_t^P brings disutility to households. β_P refers to the discount factor with $0 < \beta_P < 1$, j and χ denote preference for housing and real money balances respectively, and φ is related to the elasticity of labor supply. The budget constraint faced by patient households when maximizing expected utility is given as follows.

15 If data on the LTV ratio of business properties were available, this baseline model can be expanded to include an entrepreneurial sector. We leave this for future research.

$$c_t^P + q_t(h_t^P - h_{t-1}^P) + s_t^P = w_t^P L_t^P + \frac{R_{t-1}s_{t-1}^P}{\pi_t} + F_t + T_t^P - \Delta\left(\frac{M_t^P}{P_t}\right) \quad (7)$$

where q_t denotes real house prices ($\frac{Q_t}{P_t}$), s_t is real savings, w_t^P is real wages.

Patient households consume goods and accumulate housing while saving a certain fraction of the total income which comprises labor income, real interest income $\frac{R_{t-1}s_{t-1}^P}{\pi_t}$, dividends from the retailers (F_t) and transfer from the central bank (T_t).[16] Increments in real money balances are funded by the various sources of total income.

4.2. Impatient Households

The group of impatient households (borrowers) denoted by *B*, also has unit mass and maximizes the same type of utility function as savers.

$$E_0 \sum_{t=0}^{\infty} \beta_B^t \left[\ln c_t^B + j \ln h_t^B - \frac{(L_t^B)^{\varphi}}{\varphi} + \chi \ln\left(\frac{M_t^B}{P_t}\right)\right] \quad (8)$$

However, the discount factor of the impatient households is less than that of the patient ones, i.e. $\beta_B < \beta_P$. This condition ensures that the borrowing constraint for the impatient households binds near the steady state with reasonably small shocks.[17] The budget constraint is different from that of savers only in that impatient households are the borrowing entities and pay interest to savers.

---

16 Nominal interest income from lending s_t to borrowers at the previous period is $R_{t-1}s_{t-1}$. where S_{t-1} is nominal savings equal to $P_{t-1}s_{t-1}$. Hence nominal interest income from lending the savings can be rewritten as $(R_{t-1}P_{t-1}s_{t-1})$. Dividing it with overall price level P_t renders real interest income at the current period $(R_{t-1}s_{t-1}P_{t-1})/P_t$. Since P_{t-1}/P_t is the reciprocal of the gross inflation rate $\pi_t = P_t/P_{t-1}$, the real interest income at the current period is expressed finally as $(R_{t-1}s_{t-1}^P)/\pi_t$.

17 Appendix B-1 proves that the borrowing constraint binds at the steady state.

$$c_t^B + q_t(h_t^B - h_{t-1}^B) + \frac{R_{t-1}b_{t-1}^B}{\pi_t} = b_t^B + w_t^B L_t^B + T_t^B - \Delta\left(\frac{M_t^B}{P_t}\right) \quad (9)$$

where b_t refers to the debt owed to patient households.

Additionally and importantly, the impatient households are subject to a borrowing constraint the role of which lies at the heart of propagation and amplification of a monetary policy shock in this model. The impatient households provide the current housing stock as collateral and borrow funds against the expected value of the collateral in the next time period. However, mainly because of the uncertainty latent in future house prices and borrowers' ability to repay the debt, the impatient households are entitled to borrow only a fraction of the total collateral value. To express the constraint,

$$b_t \leq m_t E_t\left(\frac{q_{t+1}h_t^B\pi_{t+1}}{R_t}\right) \quad (10)$$

This borrowing constraint implies the total amount of real debt should be less than a fraction of the discounted expected value of the housing provided as collateral.[18] m_t is the the LTV ratio with $0 < m_t < 1$ and the multiplicative term $m_t E_t(q_{t+1}h_t^B\pi_{t+1}/R_t)$ can be considered as the upper bound of the collateral value which lenders can secure in redeeming a possible default in the following period. Put differently, $(1 - m_t)$ fraction of the collateral value is considered by the lenders as the minimum sum of various costs to be incurred by a default such as the cost for legal proceedings, foreclosing and reselling collateral.

Even though m_t is time-varying in practice and determined by the patient households in the model with the *risk-taking channel* to be presented later, we assume for the moment it is fixed as $\overline{m}$ to provide a benchmark for measuring the effect of the *risk-taking channel*.

18 In nominal terms, $B_t \leq m_t E_t\left(\frac{Q_{t+1}h_t}{R_t}\right)$, and if both sides are divided by P_t, $\frac{B_t}{P_t} \leq$ *mtEt htRtQt+*1*Pt+*1*Pt+*1*Pt*

Henceforth, I will designate the version of the model with the fixed LTV ratio as the *baseline model* and the LTV-endogenized version as the *risk-taking model*.

4.3. Wholesale Goods Firms

The firms produce wholesale goods Y_t by hiring labor from households using the following technology in which labor is a unique input.

$$Y_t = A(L_t^P)^{\alpha}(L_t^B)^{(1-\alpha)} \tag{11}$$

where A represents total factor productivity and α is the labor income share of patient households. Since the main focus of the analysis is put on the transmission effects of the shocks to monetary policy and the LTV ratio, we ignore technolo-gical shocks and set A=1 for the purpose of simplicity. The wholesale firms maximize profit, i.e. revenue of Y_t/X_t less cost of $w_t^P L_t^P + w_t^B L_t^B$, as below.

$$\max_{L_t^P, L_t^B} \frac{Y_t}{X_t} - w_t^P L_t^P - w_t^B L_t^B \tag{12}$$

where X_t is the markup of final goods over wholesale goods defined by a ratio of retail prices to wholesale ones, P_t/P_t^W.

4.4. Retailers

To introduce price rigidities into the model, monopolistic competition and Calvo-type price optimization are assumed at the retail level as in the standard New Keynesian model. A continuum of retailers of mass unity buy wholesale goods from the firms at P_t^W and sell them to consumers at P_t. Aggregate final goods index (Y_t^F) is the integration of demand of each retailer, indexed by i, for intermediate goods as follows.

$$Y_t^F = \left(\int_0^1 Y_t(i)^{\frac{\varepsilon-1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}} \tag{13}$$

where ε represents the elasticity of substitution among differentiated intermediate goods and is over unity ($\varepsilon > 1$). The aggregate price index also derives from integration of the individual price index which the retailers are facing.

$$P_t = \left(\int_0^1 P_t(i)^{1-\varepsilon} di \right)^{\frac{1}{1-\varepsilon}} \tag{14}$$

Given these two aggregate indices, retailers maximize the expected lifetime utility function under a standard type of budget constraint. The maximization yields the following individual demand function for final goods which each retailer faces.

$$Y_t(i) = \left(\frac{P_t(i)}{P_t} \right)^{-\varepsilon} Y_t^F \tag{15}$$

Taking the demand function and the wholesale price, P_t^w, as given, each retailer chooses the optimal price $P(i)_t^*$ to maximize the current value of the profit made under the condition that the chosen price remains effective. However, only a fraction, 1-θ, of retailers receive random signals during each period and reset the prices while the remaining fraction θ maintains the same price as in the previous period. The optimal price can be obtained by solving the following problem.

$$\max_{P(i)_t^*} \sum_{k=0}^{\infty} \theta^k E_t \left\{ A_{t,k} \left(\frac{P_t^*(i)}{P_{t+k}} - \frac{X}{X_{t+k}} \right) Y_{t+k}^*(i) \right\} = 0 \tag{16}$$

where $Y_{t+k}^*(i) = P_t^*(i)/P_{t+k})^{-\varepsilon} Y_{t+k}$ is the demand for each retailer's differentiated goods and $A_{t,k} = \beta_P (c_t^P / c_{t+k}^P)$ is the usual stochastic discount factor of the patient household. Without price rigidities, $\theta = 0$, the first order condition of this maximization problem boils down to the condition that the optimal price $P(i)_t^*$ needs to be equalized to the real marginal cost times the desired markup $X = \frac{\varepsilon}{\varepsilon - 1}$. Retailers rebate profits

$F_t = (1 - 1/X_t)Y_t$ to patient households. The first order condition of the maximization problem is given as below.

$$P_t^* = X \sum_{k=0}^{\infty} \left[\frac{(\theta\beta)^k E_t(\Lambda_{t,k} Y_{t+k}^{*f} P_{t+k}^{-1})}{\sum_{k=0}^{\infty} (\theta\beta)^k E_t(\Lambda_{t,k} Y_{t+k}^{*f} P_{t+k}^{-1})} \right] E_t \left(\frac{1}{X_{t+k}^n} \right) \tag{17}$$

Under the Calvo pricing environment, the aggregate price dynamics of the economy is as follows.

$$\pi_t^{1-\varepsilon} = \theta + (1-\theta) \left(\frac{P_t^*}{P_{t-1}} \right)^{1-\varepsilon} \tag{18}$$

where π_t refers to gross inflation $\frac{P_t}{P_{t-1}}$. Linearizing the equation (17) around the steady state and combining it with (18) above yields the standard New Keyensian Phillips Curve (NKPC).[19]

4.5. Monetary Policy

In order to close the model, the central bank is assumed to determine nominal policy rates R_t in response to the deviations of inflation and output from the desired level. The specific type of the Taylor rule is given by

$$R_t = R_{t-1}^{r_R} \left(\pi_{t-1}^{1+r_\pi} \left(\frac{Y_{t-1}}{Y} \right)^{r_Y} \overline{rr} \right)^{1-r_X} e_t^R \tag{19}$$

where $\overline{rr}$ and Y denote the steady-state real interest rate and output respectively, and e_t^R is an independently and identically distributed monetary policy shock with zero mean and variance σ_R^2. The exponent r_R represents the degree of inertia in adjusting policy rates in practice.

19 A detailed description of the derivation of NKPC is given on pp. 43-49 of Gali (2008).

4.6. Equilibrium

If the necessary conditions for optimization and a set of market clearing conditions are satisfied, the model reaches a unique stationary equilibrium in the absence of shocks to the system.[20] The market clearing conditions are for the housing market, $H = h_t^P + h_t^B$, the total oupout, $Y_t = c_t^P + c_t^B$, and the lending market, $s_t = b_t^B$. The housing stock is assumed to be fixed for simplicity. As stated above, impatient households borrow up to the maximum amount savers are willing to lend. By linearizing the set of first-order conditions and market-clearing ones around the steady state, the baseline model boils down to a system of 14 log-linearized equations as presented in Appendix B-3.

5. Analysis of Monetary Policy Transmission in Baseline Model

5.1. Parameter Values for Calibration

To conduct a qualitative analysis of the monetary policy transmission using the baseline model, I choose specific values for parameters based mainly on Iacoviello and Neri (2010) and related papers with a similar motivation and model structure, for example, Iacoviello (2005), Calza *et al*. (2009) and Gerali *et al.* (2010). The number of parameters to calibrate is 12 and the chosen values of the parameters are listed in Table 3.

The values of the parameters in Table 3 are somewhat different from those in other studies. For examples, Gerali *et al*. (2010) sets the impatient households' discount factor β_B to be 0.98 and Gali (2008) sets the price rigidity parameter θ to be 0.66. However, the differences in these parameter values do not substantially affect the quantitative aspects of the baseline model analysis. Furthermore, the choice of the parameter values specified on Table 3.3 ensures a proper solution to the model. Given the prime role of the borrowing constraint, it is worth

20 Appendix B-2 includes the necessary conditions for each sector.

elaborating on the level of the long-term average LTV ratio. There have been few statistical sources for calculating the long-term average LTV ratio which can be considered a steady-state value for the parameter $\overline{m}$. As there is no consensus on the level of average $\overline{m}$, researchers have used different values for it. Referring to Table 4 in respect of the U.S., Monacelli (2008) sets the annual average of the LTV ratio as 0.75 for the period 1952-2005, Iacoviello (2005) chooses 0.55, while Iacoviello and Neri (2010) use 0.85. For the Euro area, Calza *et al*. (2009) use 0.7.

In contrast to the literature mentioned above, I used the quarterly LTV data estimated by Duca *et al*. (2011) for the U.S. in the previous section. The average of the quarterly LTV ratio from 1980 Q1 -2008 Q4 is 0.87 which is close to that in Iacoviello and Neri (2010).

Table 3 Calibrated Parameters

Parameter	Value	Description
Households		
β_P	0.99	Patient households' discount factor
β_B	0.95	Borrowers' (Impatient households') discount factor
φ	1.01	Labor supply aversion
j	0.12	Weight on housing in households' utility function
α	0.64	Labor income share of patient households
Price Rigidities		
X	1.05	Steady-state gross markup
θ	0.75	Probability of maintaining prices
TFP		
A	1.00	Total Factor Productivity
Monetary Policy		
r_R	0.73	Smoothing parameter of the Taylor rule
r_π	0.27	Inflation coefficient of the Taylor rule
r_Y	0.13	Output gap coefficient of the Taylor rule
LTV ratio		
$\overline{m}$	0.87	LTV ratio

Table 4 | Average LTV Ratio for Home Mortgage

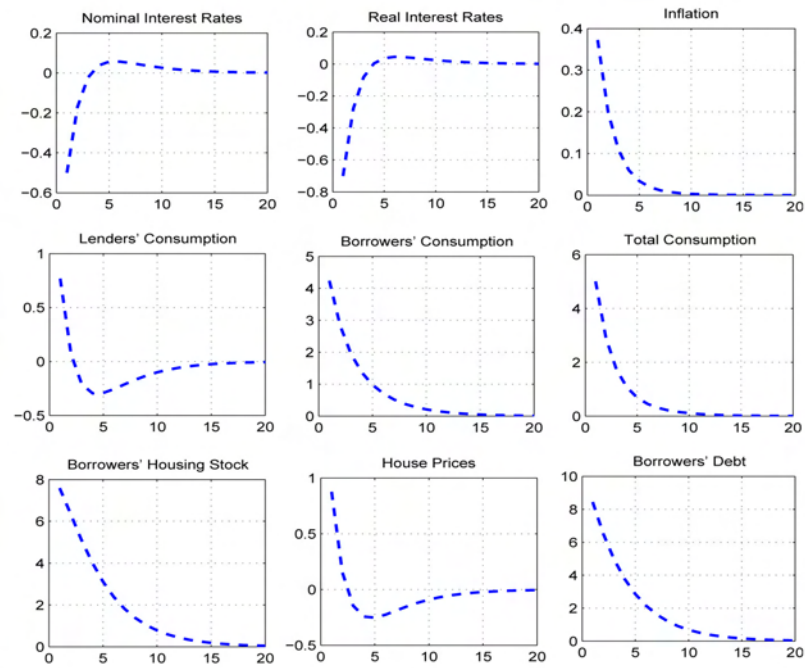
Literature	LTV ratio	Period	Country	Data Source
Calza *et al.*(2009)	0.70[1)]	-	Euro Area	various sources[2)]
Iacoviello (2005)	0.55	-	U.S.	-
Iacoviello and Neri (2010)	0.85	1973-2006	U.S.	Finance Board's Monthly Survey
Monacelli (2007)	0.75	1952-2005	U.S.	Federal Housing Finance Board

Note: 1) Gerali *et al.* (2010) follows the same value of Calza *et al.*(2009).
2) For more details, see the Table 1. on pp. 38 in the paper.

5.2. Impulse Response to Monetary Policy Shock

I will focus on the impulse responses of the main variables of interest to a monetary policy shock: consumption, house prices and debt. The main task of this analysis is in examining if the *risk-taking transmission channel* of monetary policy generates enhanced effects on the paths of

Figure 7 | Impulse Responses to a Monetary Policy Shock



Note: The y-axis measures percent deviation from the steady state.

these variables. Figure 7, in which the time period is one quarter, plots impulse responses to an expansionary monetary policy shock, i.e. a negative shock to nominal interest rates (R_t). The transmission process starts with a decrease in R_t which reduces the real interest rate by the Taylor principle. The sufficient condition for satisfying the Taylor principle, as clarified by Bullard and Mitra (2002), is $r_\pi > 1$ in the Taylor rule specified above which implies that real interest rates rise with an increase in the nominal interest rates.

Lower level of real interest rates induce households to expand consumption. In particular, the interest rate channel exerts a stronger influence on the consumption of impatient households than patient ones. The assumption that the discount factor of impatient households is lower than that of patient households implies the former has an incentive to increase current consumption by expanding their borrowing. Another transmission mechanism operates through the changes in house prices caused by an upward pressure on demand in the housing market. As the impatient households spend the additional funds borrowed not only in consuming final goods but also in buying houses, housing prices increase. In turn, the appreciation in the collateral value increases the maximum amount the impatient households can borrow. Owing to this so-called equity withdrawal effect, households again can consume more than previously. This second channel is an application of the *credit cycle* in Kiyotaki and Moore (1997) to the housing market and analogous to the *financial accelerator mechanism* in Bernanke *et al*. (1999). These two channels, the *interest rate channel* and *house price channel*, compose the transmission mechanism of monetary policy in Iacoviello (2005) and explain why the impatient households' consumption deviates further from the steady state than that of patient households.[21] [22] These two transmission channels can be illustrated by the following causal flows in which hatted variables denote percent deviations from the steady state.

---

21 The term *house price channel* is sometimes termed the *collateral channel*.

22 There is one more channel titled the *debt deflation channel* in Iacoviello(2005).

■ Interest Rate Channel

• Patient Households : $\hat{R}_t \downarrow \rightarrow \widehat{rr}_t \downarrow \rightarrow \hat{c}_t^P \uparrow$ and $\hat{s}_t^P \downarrow$

• Impatient Households : $\hat{R}_t \downarrow \rightarrow \widehat{rr}_t \downarrow \rightarrow \hat{b}_t^B \uparrow \rightarrow \hat{c}_t^B \uparrow$

■ House Price Channel: Equity Withdrawal Effect

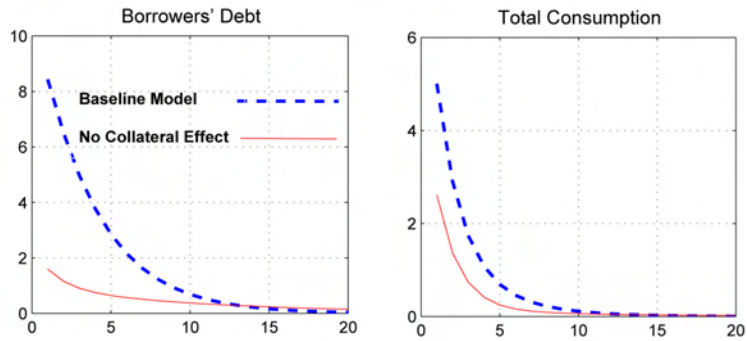
• Impatient Households $\hat{R}_t \downarrow \rightarrow \widehat{rr}_t \downarrow \rightarrow \hat{b}_t^B \uparrow \rightarrow \hat{q}_t \uparrow \rightarrow b_t^B \uparrow\uparrow \rightarrow \hat{c}_t^B \uparrow$

To get an idea of the quantitative influence of the equity withdrawal effect, I compare the impulse response of consumption in the baseline model with that from the model in which impatient households cannot borrow with housing as collateral. Figure 8 juxtaposes the impulse responses of consumption to a positive monetary policy shock from the baseline model with that of the model without a borrowing constraint. As shown in the same figure, the impatient households in the economy without equity withdrawal borrow and consume less than they would otherwise. This means that a monetary policy shock is amplified through borrowing against collateral in the baseline model.

6. Effects of Risk-taking Channel in Mortgage Market

In this section, I will examine whether the *risk-taking channel* in the presence of a positive monetary policy shock generates more volatile paths of the main variables relative to the baseline model. With this in mind, the regression equation for the *risk-taking channel* estimated in Section 3 is incorporated into the baseline model. Endogeneizing the LTV ratio also allows us to examine how a shock to this ratio affects the whole economy.

Figure 8 | Comparison of Impulse Responses of Consumption



Note: The y-axis measures percent deviation from the steady state.

6.1. Monetary Policy Shock

6.1.1. Backward-looking LTV Ratio Decision Rule

The main hypothesis of this analysis is that the *risk-taking channel* intensifies the effects of a monetary policy shock in the baseline model since lenders raise their LTV ratio in reaction to the shock. Accordingly, impatient households can borrow more than they would otherwise and increase their consumption and holding of housing stock. By incorporating the benchmark LTV equation into the baseline model, we will examine whether the hypothesis can be supported by the *risk-taking model*. In this model, banks reset the level of the LTV ratio in every period on the basis of the rule expressed in equation (20), that is, based on short-term interest rates and house prices in the previous period.

$$\hat{m}_t = \gamma_1 \hat{m}_{t-1} + \gamma_2 \hat{R}_{t-1} + \gamma_3 \hat{q}_{t-1} + \hat{e}_t \tag{20}$$

where $\hat{m}$, $\hat{R}$ and $\hat{q}$ refer to the deviation of the LTV ratio, interest rates and house prices, respectively, from their steady states. ϵ_t refers to an exogenous shock to the decision process.[23] Low interest rates in the

23 In the model, banks are assumed to deposit and lend funds at the same interest rate and not to impose any transaction cost. For the purpose of simplification, savers are assumed to act as lenders and banks at once.

previous period drive them to take more risk for the reasons state previoiusly. To reiterate, these reasons include higher yields from collateralized lending than bonds, underestimation of downside risk to house prices, expectations of robustness in future house prices, and a belief in the 'too big to fail' myth.

Table 6 lists the parameter values of the LTV decision rule above. The values are based on the results from the estimation of the regression equation (3).

Figure 9 shows the impulse responses to an unexpected decrease in policy rates by 0.5%p.

Table 5 | Parameter Values for Backward-looking LTV Decision Rule

	γ_1	γ_2	γ_3
Value for Calibration	0.7	-3.2	0.3

The solid line depicts the results from the *risk-taking model* while the dashed line illustrates the same impulse responses from the baseline model as in Figure 7. It is evident that the traditional *interest rate* and *house price* channels in the *risk-taking model* generate similar positive responses of the variables of interest as the baseline model does. However, the *risk-taking channel* pushes up consumption to a higher level during the first year. Specifically, in the *risk-taking model*, the expansionary monetary policy shock induces consumption to deviate positively by 6.5% from the steady-state whereas it generates an increase of 5.0% in the baseline model. The additional increment in the deviation of consumption results from the increase in borrowers' debt. As suggested by our previous discussion about the *risk-taking channel*, lower interest rates lead lenders to forecast higher house prices in the future and consequently to under-estimate the risk of collateralized lending. Lenders are now willing to provide more funds even if there is no change in the collateral value or income level of borrowers.

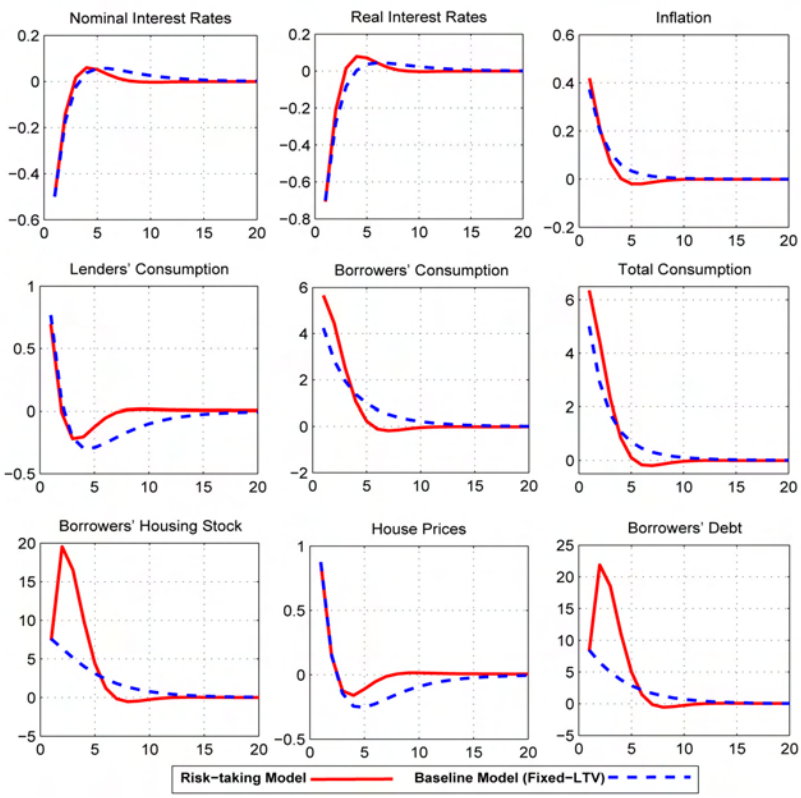
The difference in the paths of the economy generated by the baseline model and the *risk-taking model* sheds some light on why central bankers failed to forecast the full effects of the long-lasting accommodative policy on the economy. The two traditional channels, i.e., interest

rate and collateral channels, might be taken into account in the estimation of the future path of the economy. However, missing the causal chain between low policy rates and bankers' lending behavior might lead policy makers to underestimate the influence of their decision to maintain low interest rates for a prolonged period in the first half of the last decade.

To continue our discussion, two peculiar features of the impulse responses of house prices and borrowers' debt are noted below.

Firstly, comparing the impulse responses, there exists no difference in the two paths of house prices over the first four quarters. This indicates that the impact of the *risk-taking channel* on consumption results mainly from the decline in policy rates and a subsequent increase

Figure 9 Impulse Responses to a Monetary Policy Shock



Note: The y-axis measures percent deviation from the steady state.

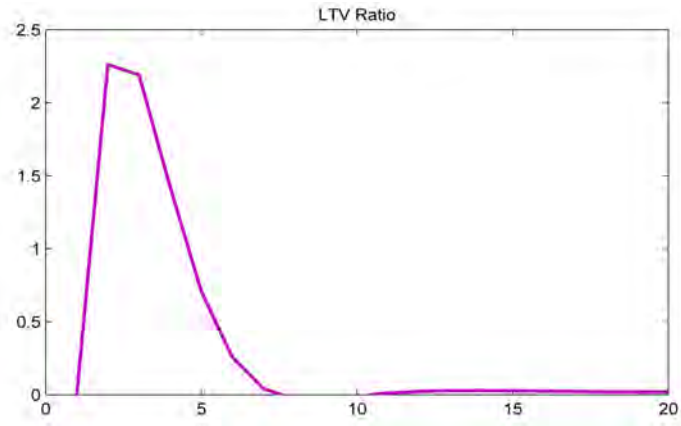
in debt rather than the increase in house prices in the first four quarters. In a sense, this is at odds with the hypothesis of the *risk-taking channel*. In the hypothetical economy, lenders adjust their risk-taking attitude by responding to changes not only in policy rates but also in house prices. Fluctuations in house prices send strong signals to mortgage lenders about how risky collateral will be in the future. The *risk-taking model* presented here fails to reflect the causal chain running from realized house prices to lenders' risk-taking attitude in the first four quarters. However, house prices do increase sharply after the first year in the risk taking model. This shows that it takes a while for the risk taking channel to have its full impacts; nevertheless eventually house prices do increase much more with the risk taking channel. This is not inconsistent with observations during the recent housing boom.

Secondly, the impulse response of borrowers' debt shows no response in the first period and then a rapid upturn in the second period. This result arises from the assumption that lenders decide their LTV ratios on the basis of interest rates and house prices in the previous quarter. As seen from Figure 10, the impulse response of the ratio reveals an unnatural kink possibly because of the backward-looking behavior of lenders. Under this backward-looking decision rule, the response of borrowers' debt to policy change appears to have a more volatile path than the other economic variables. To overcome these shortcomings, an alternative rule is introduced below.

6.1.2. Forward-looking LTV Ratio Decision Rule

To make the assumption on the behavior of lenders more consistent with reality, a forward-looking decision rule of the LTV ratio is introduced below. Lenders adjust the LTV ratio on the basis of the observed level of policy rates and house prices in the current period and also their own expectations of the evolution of these two variables in the next period. Another distinction from the backward-looking rule is the absence of the lagged LTV ratio itself. The omission of the term implies that gradualness in adjusting the LTV ratio is not in the lenders' interest *per se*.

Figure 10 Impulse Response of LTV Ratio



Note: The y-axis measures percent deviation from the steady state.

$$\hat{m}_t = \xi_1 \hat{R}_t + \xi_2 \hat{q}_t + \zeta_1 E_t(\hat{R}_{t+1}) + \zeta_2 E_t(\hat{q}_{t+1}) + \hat{e}_t \quad (21)$$

The parameter values in the forward-looking decision rule are obtained by a regression using the same data used in estimating the backward-looking equation. The responsiveness to the contemporary policy rate is lower than that of the backward-looking rule. The forward-looking coefficient, i.e., ζ_1 is still lower. This result makes sense in that lenders put less weight on their own expectation of policy rates because of the uncertainty inherent in forecasting.

Table 6 Parameter Values for Forward-looking LTV Decision Rule

	ξ_1	ξ_2	ζ_1	ζ_2
Value for Calibration	-1.95	0.21	-1.77	0.16

The impulse responses to a shock to monetary policy with a 50bp decrease are illustrated in Figure 11. As anticipated, the two *risk-taking* models with a backward-looking rule and forward-looking rule generate more volatile paths of the economic variables than the baseline model. However, even though the two *risk-taking* models share the common characteristics of the *risk-taking channel*, the overall economy displays

enhanced deviations in the case of the forward-looking rule. In other words, a monetary policy shock is amplified more when lenders adjust their LTV ratios depending on their forecasts for policy rates and house prices rather than on past information on these variables.

House prices, in particular, show a further deviation from the steady state. As lenders expect future policy rates and housing prices will move and they use these expectations when deciding the LTV ratio for every period, they supply more credit to borrowers. If lenders employed only the information in the previous period, their willingness to make loans might be less than it would otherwise. As a result of more mortgage supply, borrowers can consume more housing and non-residential goods. House prices and consumption display more fluctuations. As such, the model employing the forward-looking LTV decision rule overcomes the shortcomings of the model with the backward-looking rule.

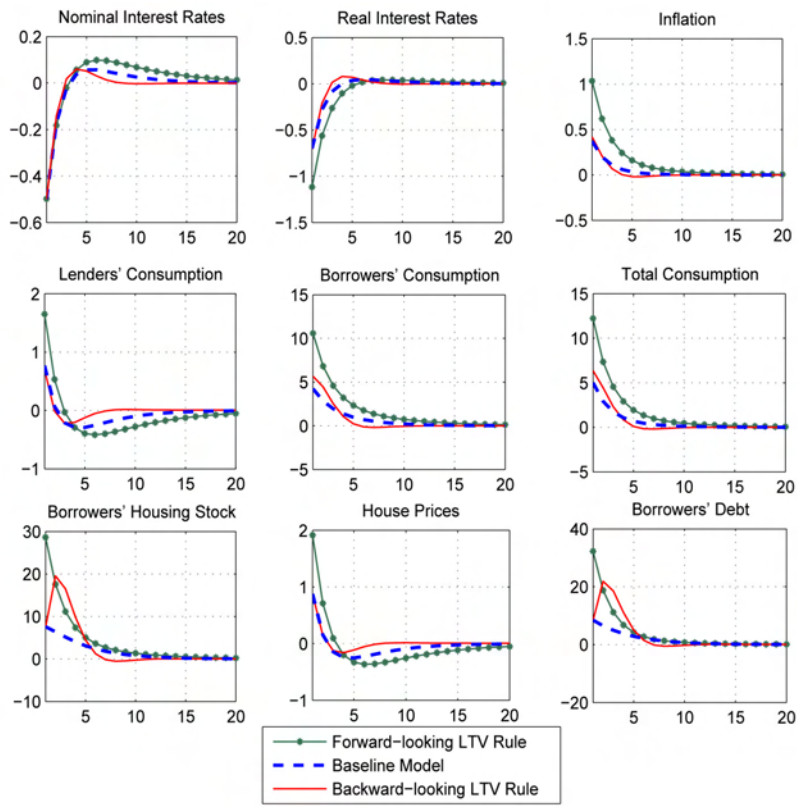
This result provides a forceful insight into the role of the expectations of credit suppliers in the housing market. In retrospect, it is acknowledged that the housing bubble in the run-up to the sub-prime crisis could not have developed only due to the irrational and myopic expectations of housing buyers. One of the main driving forces in the housing market at that time was the infinitely elastic credit supply in response to the demand for loans. Regarding the behavior of bankers, some researchers have raised the possibility that bankers at that time were as irrational as home buyers in forming their expectations of future housing market situation. The impulse responses generated by the model with the forward-looking rule supports this hypothesis. If lenders draw on their own expectations in deciding the LTV ratio instead of using the realized value of policy rates and housing prices, the economy shows a more volatile path than it would otherwise.

6.2. LTV shock

We now turn to what happens to the whole economy if a shock to the LTV ratio decision process arrives using the forward-looking decision rule introduced above. The shock to the decision process can be a wider latitude to adjust downpayment or change the ceiling on the LTV ratio which are caused by changes in banking regulations. In the context of

risk-taking attitude of lenders, the shock can be interpreted as changes in the preference for the risk related to housing-collateralized lending which can be caused, for example, by lenders' optimistic expectation about future house prices and economic activities.

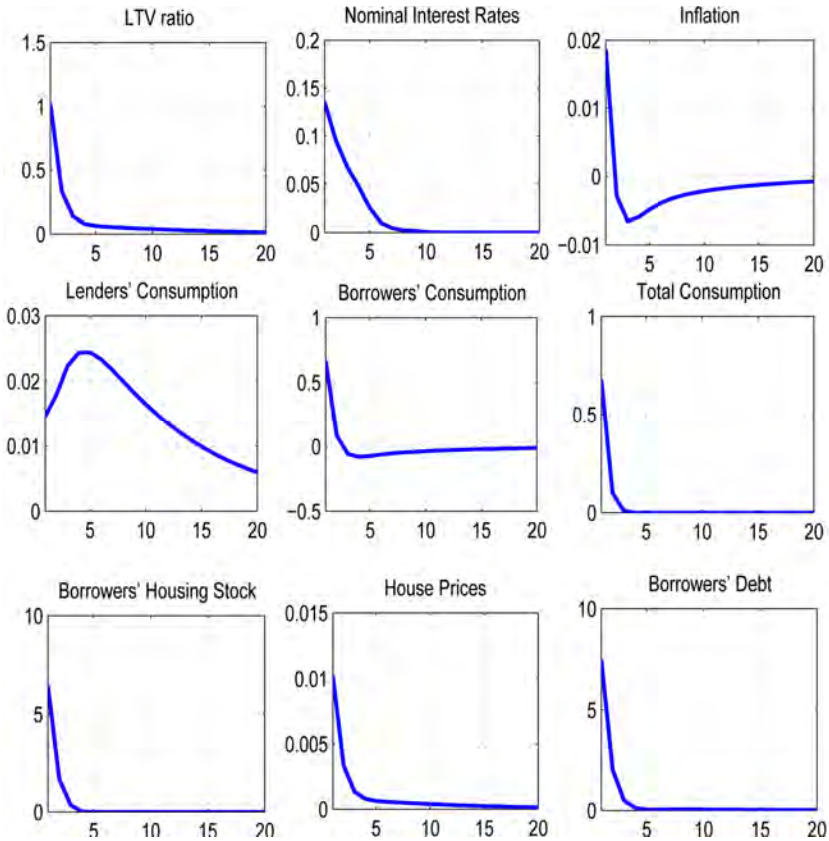
Figure 11 Impulse Responses by Forward-looking LTV Decision Rule



Note: The y-axis measures percent deviation from the steady state.

Figure 12 shows the impulse responses to a positive shock to the LTV decision process which has the magnitude of one percent deviation from the steady-state the LTV ratio. The positive shock implies that lenders become more aggressive in expanding housing-collateralized loans. After the shock hits the economy, lenders supply more credit

Figure 12 | Impulse Responses to LTV Shock



Note: The y-axis measures percent deviation from the steady state.

given a specific level of housing value and borrowers can increase debt not only for consumption but also for house purchases. As the first panel in the same figure reveals, the shock increases the LTV ratio instantaneously by one percent from the baseline, which implies heightened credit availability and leads to a corresponding increase in borrowers' debt. The impulse responses of consumption and house prices are in line with expectation: total consumption increases by 0.8% and house prices rise further 0.05%. The central bank reacts to these output and inflation gaps by increasing policy rates. However, since the gaps are narrow enough to be bridged by a soft response, the magnitude of rate hike is not substantial.

As noted above, an LTV shock exerts only limited influence on the variables compared with a monetary policy shock. The extent to which the variables deviate from the steady state in the presence of an LTV shock is quite smaller than when a monetary policy shock hits the economy. For example, borrowers' consumption deviates by only 0.6% in response to an LTV shock whereas a monetary policy shock causes it to deviate by almost 6% as shown by Figure 11. Similarly, the response of housing prices to an LTV shock is less than with a monetary policy shock; house prices change by only 0.01% with the LTV shock whereas they deviate by 2% in reaction to a monetary shock. These differences can be explained by the difference in the channels through which these two shocks are transmitted. As elaborated above, monetary policy is transmitted through three channels: the interest rate channel, house price channel and risk-taking channel. These channels have a long-lasting effect on the economy. On the other hand, an LTV shock directly affects only the amount of lending by patient households to borrowers. Only a 1% deviation of the LTV ratio falls short of exerting strong impact on the behavior of economic variables, which is consistent with the reality.

However, the qualitative property of the simulated paths of consumption and house prices are consistent with what we observed in certain developed countries during the period before the sub-prime crisis. Mortgage lenders enhanced the LTV ratio to 100 percent, in some extreme cases, even up to 120 percent, and consequently existing home owners could withdraw more equity from their houses. The funds borrowed against housing as collateral were spent in purchasing houses and consuming other goods. The increased demand for housing pushed house prices to an unsustainable level. The loop reiterated itself until mortgage related assets, such as mortgage-backed securities (MBS), turned non-performing as a result of the housing market crash and the functioning of the whole banking sector was then crippled.

7. Conclusion

The motivation for the analysis originates from the following queries: (i) Why did central banks and most macroeconomists fail to forecast the

rapid economic downturn after the sub-prime crisis? (ii) How did the prolonged period of low interest rates affect banks' mortgage supply? (iii) Does there exist any unidentified relationship between the accommodative monetary policy maintained for a protracted period and the devastating aftermath of the financial crisis?

To answer these questions, a micro-founded model is developed which incorporates the hypothesized *risk-taking channel* of monetary policy into a workhorse DSGE model featuring housing-collateralized lending and a borrowing constraint. In the model, lenders become increasingly aggressive towards risk by increasing the LTV ratio as a reaction to a decrease in policy rates. There are two prominent reasons for lenders assuming more risk. These include *search for yield* and the tendency to under-estimate risk in housing-collateralized lending in the presence of robust growth in collateral value. The specific procedure for setting up a DSGE model which mobilizes the *risk-taking channel* underwent two steps. First, two kinds of empirical analysis were conducted using U.S. data during the period from 1980 to 2007: (i) a set of simple regressions with the LTV ratio as the dependent variable and (ii) a VAR model with short-term interest rates, the LTV ratio, and house prices as endogenous variables. In turn, the estimated regression equation chosen as a benchmark equation merges into the baseline model to make the *risk-taking channel* operative. The overarching aim of the *risk-taking model* is investigating whether the *risk-taking channel* amplifies an expansionary monetary policy shock. Additionally, the model enables us to see how the economy behaves in response to a shock to the LTV decision process.

The analysis of impulse responses generated by the *risk-taking model* confirms the hypothesis that with the *risk-taking channel*, the trajectories of consumption and mortgage debt in the model become more volatile. If the *risk-taking channel* operates, an initial monetary policy shock produces more significant deviations of consumption and borrowers' debt from the steady state relative to the baseline model. In particular, if lenders decide the LTV ratio using their own expectations about the future paths of policy rates and house prices, the *risk-taking channel* turns out to have a greater effect on the economy as a whole.

From the analysis, we can derive several implications for monetary

policy implementation and financial regulation. First and foremost, to evaluate accurately the influence of monetary policy decisions on economic and financial activities, we need to take into account the impact generated through the *risk-taking channel* in addition to the impact from the traditional transmission channels. If the effects are not given proper consideration, accommodative monetary policy decisions can instead destabilise the whole economy since the response of banks and households to the easy stance will be underestimated. Secondly, regulations on the LTV ratio can contribute to the stability of the economy by curbing the aggressive risk-taking behavior at the credit supply side. If the *risk-taking channel* is operating, counter-cyclical regulatory interventions in the lending market (by imposing a lower ceiling on the LTV ratio) can smooth the paths of financial and real economic variables alike.

For future research, two points are worth mentioning. To capture the whole picture of the influence on house prices and the whole economy, we need to consider the *risk-taking channel* on the credit demand side in the mortgage market. Secondly, the process of expectation formation still remain a "black box" in evaluating, both empirically and theoretically, the effects of monetary policy decisions on the housing sector. Hence we need to invest more of our resources in identifying and estimating the effects produced via the expectation channel.

References

Adrian, T., and Shin, H. S. (2009). Money, liquidity, and monetary policy. *American Economic Review Papers and Proceedings*, 99, 600-605.

Adrian, T., and Shin, H. S. (2010a). Financial intermediaries and monetary economics. In Friedman, B. M., and Woodford, M. (eds.), *Handbook of Monetary Economics Vol. 3A*, pp. 601-650, Amsterdam: North-Holland.

Adrian, T., and Shin, H. S. (2010b). Liquidity and leverage. *Journal of Financial Intermediation,* 19, 418-437.

Almeida, H., Campello, M., and Liu, C. (2006). The financial accelerator: evidence from international housing markets. *Review of Finance*, 10, 321-352.

Altunbas, Y., Gambacorta, L., and Marqués-Ibánẽz, D. (2010). Does monetary policy affect bank risk-taking? *ECB Working Paper 1166*. European Central Bank.

Bernanke, B. (2010). Monetary policy and the housing bubble. *Speech delivered at the Annual Meeting of the American Economic Association*. Atlanta, Georgia, January 3.

Bernanke, B., Gertler, M., and Gilchrist, S. (1999). The financial accelerator in a quantitative business cycle framework. In Taylor, J. B., and Woodford, M. (eds.), *Handbook of Macroeconomics Vol. 1C*., pp. 1341-1393, Amsterdam: North-Holland.

Bjørnland, H. C., and Jacobsen, D. H. (2010). The role of house prices in the monetary policy transmission mechanism in small open economies. *Journal of Financial Stability*, 6, 218-229.

Borio, C., and Zhu, H. (2008). Capital regulation, risk-taking and monetary policy: a missing link in the transmission mechanism? *BIS Working Paper 268*. Bank for International Settlements.

Bullard, J., and Mitra, K. (2002). Learning about monetary policy rules. *Journal of Monetary Economics*, 49, 1105-1129.

Calza, A., Monacelli, T., and Stracca, L. (2009). Housing finance and monetary policy. *ECB Working Paper 1069*. European Central Bank.

Clavo, A. G. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, 12, 383-398.

Cúrdia, V., and Woodford, M. (2009). Credit frictions and optimal monetary policy. *BIS*

Working Paper 278. Bank for International Settlements.

Delis, M. D., and Kouretas, G. P. (2011). Interest rates and bank risk-taking. *Journal of Banking and Finance*, 35, 840-855.

Duca, J. V., Kurt, J., Muellbauer, J., and Murphy, A. (2011). U.S. house prices, home-ownership rates, and downpayment constraints for first-time home-buyers. *Manuscript*. Federal Reserve Bank of Dallas.

ECB (2005). Asset price bubbles and monetary policy. *ECB Monthly Bulletin*, April, 47-60.

ECB (2009). Housing finance in the Euro area. *ECB Monthly Bulletin*, August, 77-91.

Gali, J. (2008). *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton: Princeton University Press.

Gambacorta, L. (2009). Monetary policy and the risk-taking channel. *BIS Quarterly Review*, December, 43-53. Bank for International Settlements.

Gerali, A., Neri, S., Sessa, L., and Signoretti, F. M. (2010). Credit and banking in a DSGE model of the Euro area. *Journal of Money, Credit and Banking*, 29, 107-141.

Goodfriend, M., and McCallum, B. T. (2007). Banking and interest rates in monetary policy analysis: A quantitative exploration. *Journal of Monetary Economics*, 54, 1480-1507.

Greenspan, A. (2010). The crisis. *Brookings Papers on Economic Activity*, Spring, 201-246.

Iacoviello, M. (2005). House prices, borrowing constraints, and monetary policy in the business cycle. *American Economic Review*, 95, 739-764.

Iacoviello, M., and Minetti, R. (2003). Financial liberalization and the sensitivity of house prices to monetary policy: theory and evidence. *The Manchester School*, 71, 20-34.

Iacoviello, M., and Neri, S. (2010). Housing market spillovers: evidence from an estimated DSGE model. *American Economic Journal: Macroeconomics*, 2, 125-164.

IMF (2008). *World Economic Outlook*. April edn., Washington D. C.: International Monetary Fund.

Ioannidou, V., Ongena, S., and Peydró, J.-L. (2009). Monetary policy, risk-taking andpricing: Evidence from a quasi-natural experiment. *CentER Discussion Paper No. 2009-31S*.

Jiménez, G., Ongena, S., Peydró, J.-L., and Saurina, J. (2008). Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking? *Bank of Spain WP 0833*. Bank of Spain.

Kiyotaki, N., and Moore, J. (1997). Credit cycles. *Journal of Political Economy*, 105,

211-248.

Maddaloni, A., and Peydró, J.-L. (2011). Bank risk-taking, securitization, supervision, and low interest rates: Evidence from the Euro-area and the U.S. lending standards. *Review of Financial Studies*, 24, 2121-2165.

Monacelli, T. (2008). Optimal monetary policy with collateralized household debt and borrowing constraints. In Campbell, J. Y. (ed.), *Asset Prices and Monetary Policy*, pp. 103-146, Chicago: University Of Chicago Press.

Musso, A., Neri, S., and Stracca, L. (2010). Housing, consumption and monetary policy: How different are the US and the Euro area? *ECB Working Paper 1166*. European Central Bank.

Papademos, L. (2006). Price stability, financial stability and efficiency, and monetary policy. *Speech at the Third Conference of the Monetary Stability Foundation on Challenges to the Financial SystemAgeing and Low Growth*. Frankfurt, July 7.

Pariés, M. D., and Notarpietro, A. (2008). Monetary policy and housing prices in an estimated DSGE model for the US and the Euro area. *ECB Working Paper 972*. European Central Bank.

Rajan, R. (2006). Has finance made the world riskier? *European Financial Management*, 12, 499-533.

Trichet, J.-C. (2005). Asset price bubbles and monetary policy. *MAS Lecture*. June 8.

Appendix

Appendix 1: Proof of the Borrowing Constraint Binding at Steady State

The Euler equation derived from the patient households' optimization is given by

$$\frac{1}{c_t^P} = \beta_P E_t \left(\frac{R_t}{c_{t+1}^P \pi_{t+1}} \right)$$

and the corresponding steady state is

$$\frac{1}{c^P} = \beta_P \frac{R}{c^P \pi}$$

and as inflation is assumed to be zero at the steady state, namely, $\pi = 1$,

$$R = \frac{1}{\beta_P} \tag{S1}$$

Turning to the Euler equation for the impatient households,

$$\frac{1}{c_t^B} = \beta_P E_t \left(\frac{R_t}{c_{t+1}^B \pi_{t+1}} \right) + \lambda_t R_t$$

and the corresponding steady state is

$$\frac{1}{c^B} = \beta_B \frac{R}{c^B \pi} + \lambda R \tag{S2}$$

Substituting (S1) into (S2) and arranging the terms about λ yields

$$\lambda = \frac{\beta_P - \beta_B}{c} \tag{S3}$$

Since $\beta_P - \beta_B > 0$ by assumption, λ is over zero at the steady state. As the original λ_t measures the increment in the lifetime utility of impatient households accrued by increases in borrowing, there is always room for increasing utility as long as λ is positive. Hence impatient households borrow to the upper limit of the borrowing constraint.

Appendix 2: Necessary Equilibrium Conditions

1. Patient Households as Lenders

$$\frac{1}{c_t^P} = \beta_P E_t \left(\frac{R_t}{c_{t+1}^P \pi_{t+1}} \right) \tag{N1}$$

$$w_t^P = c_t^P (L_t^P)^{\eta - 1} \tag{N2}$$

$$\frac{q_t}{c_t^P} = \frac{j}{h_t^P} + \beta_P E_t \left(\frac{q_{t+1}}{c_{t+1}^P} \right) \tag{N3}$$

$$m_t = m_{t-1}^{\rho m} (R_{t-1}^{\rho_R} R_{t-1}^{\rho_q})^{1-\rho m} e_t^m \tag{N4}$$

2. Impatient Households

$$\frac{1}{c_t^B} = \beta_B E_t \left(\frac{R_t}{c_{t+1}^B \pi_{t+1}} \right) + \lambda_t R_t \tag{N5}$$

$$w_t^B = c_t^B (L_t^B)^{\eta - 1} \tag{N6}$$

$$\frac{q_t}{c_t^B} = \frac{j}{h_t^B} + E_t \left(\beta_B \frac{q_t}{c_t^B} + m_t \lambda_t q_{t+1} \pi_{t+1} \right) \tag{N7}$$

$$c_t^B + q_t(h_t^B - h_{t-1}^B) + \frac{R_{t-1}b_{t-1}^B}{\pi_t} = b_t^B + W_t^B L_t^B + T_t$$

$$-\Delta\left(\frac{M_t^B}{P_t}\right) \quad \text{(N8)}$$

$$b_t = m_t E_t\left(\frac{q_{t+1}h_t^B\pi_{t+1}}{R_t}\right) \quad \text{(N9)}$$

3. Firms

$$w_t^P = \alpha\frac{Y_t}{X_t L_t^P} \quad \text{(N10)}$$

$$w_t^B = (1-\alpha)\frac{Y_t}{X_t L_t^B} \quad \text{(N11)}$$

$$\frac{Y_t}{X_t} = w_t^P L_t^P + w_t^B L_t^B \quad \text{(N12)}$$

4. Retailers

$$P_t^* = X\sum_{k=0}^{\infty}\left[\frac{(\theta\beta)^k E_t(\Lambda_{tk}Y_{t+k}^{*f}P_{t+k}^{-1})}{\sum_{\mathrm{k}=0}^{\infty}(\theta\beta)^{\mathrm{k}}\,E_t(\Lambda_{tk}Y_{t+k}^{*f}P_{t+k}^{-1})}\right]E_t\left(\frac{1}{X_{t+k}^n}\right)$$

$$P_t = [\theta P_{t-1}^{\varepsilon} + (1-\theta)(P_t^*)^{1-\varepsilon}]^{\frac{1}{1-\varepsilon}} \quad \text{(N13)}$$

5. Central Bank

$$R_t = [R_{t-1}]^{r_R}\left[\pi_{t-1}^{1+r_\pi}\left(\frac{Y_{t-1}}{Y}\right)^{r_Y}\overline{rr}\right]^{1-r_R}e_t^R \quad \text{(N14)}$$

6. Market Clearance

$$c_t^P + c_t^B = Y_t \quad \text{(N17)}$$

$$h_t^P + h_t^B = \overline{H} \quad \text{(N18)}$$

$$s_t^P = b_t^B \quad \text{(N19)}$$

Appendix 3: Log-linearized Conditions

1. Patient Households as Lenders

$$\hat{c}_t^P = E_t\hat{c}_{t+1}^P - \widehat{rr}_t \quad \text{(L1)}$$

$$\hat{q}_t = \beta_P E_t(\hat{q}_{t+1}) + \iota\hat{h}_t^B + \hat{c}_t^P - \beta_P E_t(\hat{c}_{t+1}^P) \text{ where}$$

$$\iota = (1-\beta_P)\frac{h^B}{h^P} \quad \text{(L2)}$$

$$\widehat{m}_t = \rho_m\widehat{m}_{t-1} + (1-\rho_m)\left(\rho_R\hat{R}_{t-1} + \rho_q\hat{q}_{t-1}\right)\hat{e}_t^m \quad \text{(L3)}$$

2. Impatient Households

$$\hat{q}_t = (1-m\beta_P)\hat{c}_t^B - \beta_B(1-m)E_t\hat{c}_{t+1}^B + [\beta_B + m(\beta_P - \beta_B)]$$
$$E_t\hat{q}_{t+1} + m(\beta_P - \beta_B)\widehat{m}_t - m\beta_P\widehat{rr}_t - \frac{q}{c^B}\frac{j}{h^B}\hat{h}_t^B \quad \text{(L4)}$$

$$c^B\hat{c}_t^B = -qh^B\Delta\hat{h}_t^B - Rb^B\left(\hat{R}_{t-1} + \hat{b}_{t-1}^B - \pi_t\right) + b^B\hat{b}_t^B +$$
$$(1-\alpha)\frac{Y}{X}\left(\hat{Y}_t - \hat{X}_t\right) \quad \text{(L5)}$$

$$\hat{b}_t^B = \widehat{m}_t + E_t\hat{q}_{t+1} + \hat{h}_t^B - \widehat{rr}_t \quad \text{(L6)}$$

3. Aggregate Supply

$$\hat{Y}_t = \hat{X}_t + \eta\hat{L}_t^P + \hat{c}_t^P \quad \text{(L7)}$$

$$\hat{Y}_t = \hat{X}_t + \eta\hat{L}_t^B + \hat{c}_t^B \quad \text{(L8)}$$

$$\hat{Y}_t = \alpha\hat{L}_t^P + (1-\alpha)\hat{L}_t^B \quad \text{(L9)}$$

4. Inflation Dynamics: New Keyensian Phillips Curve

$$\hat{\pi}_t = \beta_P E_t\tilde{\pi}^{t+1} - \kappa\hat{X}_t$$
$$\text{Where } \kappa = \frac{(1-\theta)(1-\beta_P\theta)}{\theta} \quad \text{(L10)}$$

5. Central Bank

$$\hat{R}_t = r_R \hat{R}_{t-1} + (1 - r_R)\left[(1 + r_\pi)\hat{\pi}_{t-1} + r_Y \hat{Y}_{t-1}\right] + r_R \hat{R}_{t-1} + \hat{e}_t^R \quad \text{(L11)}$$

6. Equilibrium in Goods, Housing and Lending Markets

$$\hat{Y}_t = \frac{c^P}{c^P + c^B}\hat{c}_t^P + \frac{c^B}{c^P + c^B}\hat{c}_t^B \quad \text{(L12)}$$

$$0 = h^P \hat{h}_t^P + h^B \hat{h}_t^B \quad \text{(L13)}$$

$$\hat{s}_t^P = \hat{b}_t^B \quad \text{(L14)}$$

CHAPTER 10

Enhancing the Link between Higher Education and Employment

by
Kye Woo Lee
(KDI School)[*]
Miyeon Chung
(Hyupseong University)

Abstract

This study aims to improve the efficiency of fiscal assistance programs for higher education by investigating variables that influence college graduates' employment rates. An empirical analysis of 2010-2011 higher education statistics shows that two variables — educational expenditure per student and the number of students per full-time faculty member — consistently and significantly affect college graduates' employment rates, even after location and type of school are controlled. Although scholarship rates also affect employment rates positively, the number of students per industry-academe liaison officer do not have a statistically significant effect. Moreover, as educational expenditure per student or the student/faculty ratio increases beyond a certain level, graduate employment improves at an increasing rate. The two variables also affect the employment rate interactively. At a relatively higher level of per-student expenditure, employment rates increase even as the student/faculty ratio rises. However, at a relatively lower level of per-student expenditure, employment rates decline as the student/faculty ratio rises.

* Corresponding author

The policy implication is that fiscal assistance programs for higher education should accord a much greater weight to these key variables when selecting and assessing institutional recipients.

1. Introduction

This study aims to improve the efficiency of fiscal assistance programs for higher education with the goal of raising college graduates' employment rates. The low employment rate of college graduates is a world-wide phenomenon, and many governments have responded by trying to enhance macroeconomic growth rates, labor market flexibility, and higher-education quality.

The Korean government has also tackled the problem with a series of fiscal assistance programs for higher education institutions. However, despite a steady increase in fiscal support, graduate employment rates have remained stagnant, barely exceeding 50%. Since the enactment of Korea's Higher Education Law of 1997, government fiscal assistance has reached almost 10% of total education expenditures or ₩5 trillion (Ahn 2010). Furthermore, due to fiscal constraints, the level of fiscal support for higher education as a percentage of total educational expenditures stands only at 73% of the OECD member countries' average. Under the circumstances, Korea should improve the efficiency of the fiscal assistance programs for higher education.

Understanding what determines graduates' employment rates is essential to this goal, but literature on these determining factors is sparse, and a rigorous analysis of the links between institutional fiscal assistance and employment has not been carried out.

Existing fiscal assistance programs have used sets of educational indicators to select and assess the colleges involved in the fiscal assistance programs (Table 1); this practice incentivizes schools to prioritize those criteria. As a result, higher education quality is in fact managed around these indicators without knowing their effectiveness. Ideally, college education should be assessed based on output indicators, such as the employment rate. However, the current assessment practice

Table 1 Fiscal Assistance Programs: Management Indicators and Their Weights Used

Eligibility and Assessment Indicators/ Fiscal Assistance Programs	Emplo-yment Rate	Rate of Enrollment to Authorized Capacity	Full-Time Faculty Rate	Rate of Education Expenses to Tuition	Management of Academic Affairs	Rate of Scholarship to Tuition	Student Loan Repayment Rate	Moderation of Tuition Burden	Corporate Indicators	Total
Educational Capacity Strengthening Program	20	20	10	10	20	10	-	10	-	100
Students Loan Programs	20	30	7.5	7.5	5	5	10	10	5	100
Other Fiscal Assistance Programs	20	30	7.5	7.5	10	10	-	10	5	100

Source: Kim, Soo Kyung (2013)

includes both educational output and input indicators with some intuitive or arbitrary weights assigned to them and without a clear understanding of the theoretical and practical relationship between the input and output indicators.

Therefore, it is critical to define the causal relationship between the educational management indicators as inputs and the employment rate as an output. However, no critical analyses have yet been made on whether such a relationship exists between these management indicators and graduate employment rates. In the absence of such studies, political proposals to help students, such as halving college fees and tuition, expanding the number of university-industry liaison professors to balance the demand for and supply of college graduates, or establishing a system of assessing college students' capacity, launching entrepreneurship education, lack an empirical basis (Joo 2014, Oh 2014).

In order to ensure the effectiveness of the fiscal assistance programs for higher education, the selection and assessment of participating colleges should be made on the basis of the essential educational management indicators that would efficiently raise education quality and college graduates' employment rates.

Accordingly, this study will analyze the determinants of college graduates' employment rates, not at an individual student level, but at the higher educational institution level, utilizing the educational management indicators actually used in Korean fiscal assistance programs.

This study consists of the following sections. Section 2 of this study will review earlier studies on educational productivity and examine Korea's fiscal assistance programs for higher education and related studies. Section 3 will explain how the methodology and data of this study can overcome the limitations of earlier studies. Section 4 will discuss the results of the empirical analysis, and the last section will summarize the analysis and explain the policy implications for enhancing college graduates' employment rates.

2. Review of Literature and Fiscal Assistance Progams

2.1. Literature on Educational Productivity

The conceptual approach to the relationship between higher education inputs and outputs began with the theory of educational production functions. This economic approach regards schools as sites where various educational resources interact to produce educational outputs. Under this model, if educational investment is efficient, productivity will be high, resulting in a high level of educational quality (Ban 1994).

Investigating scholars differ on the nature of educational inputs. Hanushek (1989, 1995) divides educational inputs into 7 groups: student-teacher ratio, teacher educational level, teacher experience level, teacher compensation level, per-student educational expense, school management, and school facilities. U.S. educational sociologist Coleman (1996) classifies educational inputs as school-related factors, peer groups, individual cognitive ability, and family background factors. Hadderman (1998) uses six educational inputs: per-student educational expenses, student-teacher ratio, teacher educational level, teacher compensation level, school facilities, and school management.

Scholars also differ on educational outputs. In general, educational achievement level is used in many educational production function analyses. However, studies that use the employment rate as the educational output are few. Many U.S. studies have used the graduation rate as the educational output at different levels of education.

In Korea, there have not been many empirical studies that analyze the relationship between educational input and output indicators at the college level. Moreover, very few studies have analyzed college graduates' employment rates as an educational output, as the amount of educational input and output data has been insufficient for meaningful analyses. For example, Ban (1998) estimates the rate of return to higher education as an educational output by using such financial inputs as scholarship recipients and per-student educational expenses. Jun and Min (2009) also analyze the rate of return to higher education by using such financial inputs as instructors and school facilities. They find that

as the student-teacher ratio declines, the school facility size increases, and the return to educational investment increases.

The launch of the College Information Opening System (2008) enabled scholars to evaluate educational outputs more rigorously, especially the employment rate. Choi (2008), for example, analyzes variations in college employment rates by using such financial input indicators as scholarship recipient rates and tuition dependency rates, and such personnel input indicators as instructional hours at both individual and college levels. The study finds that the scholarship recipient rate consistently influences the employment rate positively, and private universities show higher employment rates than public universities. However, this study uses only one-year data and does not cover such important educational management inputs as per-student expense, student-teacher ratio, and student to industry-academe cooperation professor ratio, and their influence on the employment rate.

Jang (2004) also studies educational inputs in relation to college graduates' employment rate. This study shows that increases in the student-teacher ratio have negative effects on the employment rate, and recommends more investment in the teaching force and imposing an upper limit on the student-teacher ratio. The study also argues that per-student educational expenses should be divided into personnel and non-personnel expenses, and that personnel expenses have statistically positive effects on the employment rate, while non-personnel expenses, especially school facility expenses, tend to reduce personnel expenses and affect the employment rate negatively in the short run. In the long run, however, investment in school facilities helps improve the quality of educational services and in turn enhances the employment rate. Therefore, this study fails to establish a clear-cut priority for educational investment.

2.2. Fiscal Assistance Programs

From the mid-1990s to the beginning of the 2000s, the Korean government provided fiscal assistance to improve the quality of all higher educational institutions. Since 2004, however, the government ceased to provide general fiscal assistance and instead has offered

Table 2 Selective Fiscal Assistance Programs

Programs	Objective	Period	Budget	Client Colleges/ Univ	Features	Core Assessment Indicators
Brain Korea 21(BK21)	Fostering research capacity	2006~ 2012	₩1.5T	Research Univ.		Student-teacher ratio Full-time teacher ratio 70% or higher faculty participation
Educational Capacity Strengthening (PEUEC)	Strengthen colleges' educational capacity	2008~ Present	₩296B	All higher education institutes	Selective focus; formula funding	Employment rate Enrollment rate Full-time faculty rate Scholarship rate Per-student educational expense
World-Class University (WCU)	Fostering world-class colleges	2008~ 13	₩625B	Research-oriented colleges		Research records Panel review Overseas peer review Overall assessment
Restructuring Higher Educational Institutions	Restructure and improve quality	2005~ present	₩321B	All higher educational institutions	Formula funding	Employment rate Enrollment rate Full-time faculty ratio Educational expenses and tuition ratio
Balanced Regional Development Program	Fostering provincial colleges' competitiveness	2009~ present	₩621B	Provincial colleges	Formula funding	Employment rate Enrollment rate Full-time faculty ratio Scholarship rate Per-student educational expense
Leaders in Industry-University Cooperation (LINC)	Fostering LINC colleges	2012~ present	₩170 B (₩552B, including on-going program)	Industry-university centered institutions	Integrate existing programs of similar nature	Employment rate Patent applications per faculty member Ratio of faculty members with practice experience Industry-university cooperation officers Per-faculty research projects and funds Per-faculty technology transfer contracts

Table 2 (Continue)

Programs	Objective	Period	Budget	Client Colleges/ Univ	Features	Core Assessment Indicators
New University for Regional Innovation (NURI)	Regional development through growth of provincial universities	2004~ 08	₩1.2T	Industry-University Cooperation centered colleges	Predecessor of the LINC Program	Full-time faculty rate Enrollment rate Employment rate Curriculum development records Industry-University cooperation records

Source: Ministry of Education

selective and concentrated fiscal assistance through monitoring and evaluation of higher educational institutions. In particular, the government established a college/university evaluation system covering both achievement and accreditation evaluation aspects. For example, the colleges/universities participating in the Educational Capacity Strengthening Program are selected and provided additional fiscal assistance on the basis of the evaluation results. The incentive system based on the evaluation formula naturally promotes competition among higher educational institutions. The selective fiscal assistance programs that have been carried out to date are summarized in Table 2.

Despite the varied name, these fiscal assistance programs have all pursued the improvement in the quality and competiveness of higher educational institutions, which can be ultimately assessed and measured by graduate employment rates. However, many of them do not include graduate employment rates as an eligibility, appraisal, monitoring or evaluation indicator. Even when the graduate employment rate is used explicitly as an indicator for selection and monitoring indicator, other indicators are also used simultaneously, and some arbitrary weights are assigned to them without any theoretical and empirical basis. Among the fiscal assistance programs listed above, the NURI and PEUEC are selected for a detailed review of their characteristics and evaluation studies since these programs include college graduates' employment rates as an indicator of monitoring and evaluation.

2.2.1. The New University for Regional Innovation (NURI)

The NURI fiscal assistance program aims to achieve regional development by promoting enterprise-college cooperation and professional human resources development. During 2004-2008, a total of ₩1.24 trillion was invested outside the Seoul Metropolitan area. The program fund was distributed to colleges in 13 regions selected on the basis of enrollment rate, which must be at least 60% of the authorized capacity for a college alone, and above 90% for a group of participating colleges. In addition, participating colleges must satisfy required and elective achievement indicators. The required achievement indicators are: faculty recruitment rate, student enrollment rate, and graduate employment rate. The elective achievement indicators are: human development, exchange among participating colleges, industry-academe cooperation, employment promotion, employment rate, scholarship, etc. It is questionable whether these indicators represent achievement (outputs) or management (inputs) by nature.

Among the NURI evaluation reports, Jung (2008) regards the graduate employment rate as the core achievement indicator and finds NURI had a positive impact on that rate. Ryu et al. (2010) compare the participating colleges and non-participating colleges, using the difference-in-difference approach, and find that colleges selected for the NURI program on average show a considerably higher employment rate than the control group. The study also reveals relationships between the employment rate and the educational management indicators. For example, while the per-student NURI assistance amount exerts a positive effect on the employment rate, the per-instructor assistance amount shows insignificant effects, indicating it would be desirable to increase NURI assistance for individual students rather than for instructors. The study also indicates that the type of college also affects the employment rate, i.e., private colleges achieved higher employment rates.

2.2.2. The Program for Enhancing Universities' Educational Competency (PEUEC)

The PEUEC program was launched in 2008 to strengthen the educational capacity and competiveness of colleges/universities. While earlier fiscal assistance programs were based on college/university applications, this program selected recipients based on an educational capacity assessment formula and provided fiscal assistance in a discriminatory manner, thus inducing schools to improve competitively. The capacity assessment formula is a set of quantitatively measurable educational achievement and management indicators that have been refined over time. A special feature of the formula is that almost all educational management indicators are weighted equally. This means that each indicator affects the achievement indicator equally, an assumption or policy that needs to be tested with empirical analysis. Another feature is that the achievement indicators comprise two quantitative and qualitative indicators: the ratio of enrollment to authorized capacity and the graduate employment rate. However, it is questionable whether the enrollment ratio is an achievement indicator.

Table 3 | PEUEC Indicators (2011)

Achievement Indicators		Management Indicators							Total
Employment Rate	Enrollment Ratio	Internationalization Index	Full-time Faculty Ratio	Management of Curriculum and Academic Affairs	Per-student Educational Expenditure	Scholarship Rate	Educational Expense Ratio to Total Tuition	Admission Test Index	
20%	20%	5%	10%	10%	10%	10%	10%	5%	100%

Source: Hong (2012)

Park (2010) analyzes the impact of the PEUEC on the educational achievement and management indicators. The average graduate employment rates of participating universities improved more than those of non-participating universities when compared before (2007) and after (2008) the program's implementation. The study also shows that the

participating universities experienced improvement in such indicators as the enrollment ratio, full-time faculty ratio, and scholarship ratio when compared with the year before the program. Participating universities also spent twice as much on per-student educational expenses as non-participating schools. However, the impact evaluation requires a more rigorous and scientific comparison between participating and non-participating universities over the time before and after the program's implementation.

In sum, the previous studies have weaknesses. First, although existing studies have confirmed positive impacts of fiscal assistance programs on the employment rate, they do not shed light on the causal relationship between educational management indicators that would affect the employment rate and the employment rate itself. Consequently, the studies do not confirm whether improving the educational management indicators is necessary. Moreover, the studies do not clarify whether educational management indicators have equal effects on the employment rate.

Second, although previous studies cover the impact of fiscal assistance programs on some educational management indicators (e.g. fiscal support expenses per student or instructor), their coverage is limited. It is necessary to include all the educational management indicators involved in the selection, monitoring, and assessment of colleges involved in the fiscal assistance program, and whether and to what degree these indicators have been influenced by the fiscal assistance programs. Moreover, it is too simplistic to assume that individual educational indicators would affect the employment rate at the same rate and independently. It is possible, for example, that they would affect the employment rate at an increasing or decreasing rate (non-linear effects) and in interaction with other indicators (interactive effects). Therefore, impact evaluation studies should include a squared term and interactive terms of various educational indicators. A sound design of the fiscal assistance programs should be based on the findings of these empirics.

Third, existing studies focus on the evaluation of an individual fiscal assistance program. However, there have been several fiscal assistance programs for higher education at one time. It is more reasonable to assume that the several fiscal assistance programs together would have

influenced the educational quality and employment rate with a scale effect and in interaction with each other. Consequently it is desirable to include all fiscal assistance programs in an impact analysis.

Fourth, previous studies focus on the colleges and universities participating in various fiscal assistance programs. However, the determinants of employment rates could be different between participating and non-participating schools, depending on the institutions' characteristics. Therefore, it is desirable to include all higher educational institutions in impact evaluation studies.

This study therefore aims to overcome the weaknesses of existing studies and complement their findings. Accordingly, before carrying out any impact evaluation study of fiscal assistance programs, this study, first, seeks to identify the educational indicators that determine the employment rate, which is the critical achievement indicator of the fiscal assistance programs. Second, this study includes all educational management indicators used in the selection, monitoring, and assessment of higher educational institutions involved in fiscal assistance programs aimed at enhancing the employment rate. Third, this study covers all higher educational institutions and tries to determine whether institutions' characteristics influence the employment rate. Finally, this study analyzes the impact of the educational indicators on the employment rate, allowing for non-linear and interactive effects of educational indicators. The findings of such analyses would help construct sound design, implementation and assessment frameworks to improve graduate employment rates.

3. Method and Data

The estimation model used in this study for the college graduate employment rate is specified in the following functional form:

$$Y = a_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + e \quad (1)$$

where

Y: college graduate employment rate
a_1: constant term
x_1: the university type (public or private)
x_2: the university location (Seoul metropolitan area or other areas)
x_3: the scholarship rate
x_4: per student educational expense
x_5: the number of students per full-time faculty member
x_6: the number of students per industry-academe liaison officer
x_7: year (2010 or 2011)
e: error term

The dependent variable is the college graduate employment rate and is defined in the Higher Education Statistics on the basis of health insurance system statistics, as follows:

$$\text{College Graduates' Employment Rate} = \frac{\text{Health Insurance System Members based on Employment with Employers}}{\text{Graduates} - (\text{further education} + \text{military service} + \text{foreign students} + \text{disabled})} \text{X } 100$$

The independent variables used in this study are selected from the list of educational management indicators considered to affect the employment rate in the design and study of the fiscal assistance programs. The independent variables are grouped into four categories: university characteristics, financial indicators, faculty indicators, and industry-academe liaison indicators. Each of the independent variables is explained in the following table.

The university characteristics variables are not an indicator included in fiscal assistance programs' assessment formulas. However, many fiscal assistance programs in fact consider these characteristics, and earlier studies also indicate that the employment rate varies in response to the characteristics. Therefore, these dummy variables are included. The type dummy variable assigns 0 for public universities and 1 for private universities. The location dummy variable assigns 0 for the Seoul metropolitan area and 1 for outside areas.

Table 4 Explanations of the Independent Variables

Group	Variable	Explanation	Programs using Indicators
Characteristics Variables	Type of college (Dummy)	Public=0, Private=1	Considered in all fiscal assistance programs
	Location of college (Dummy)	Seoul Metropolitan=0, Others =1	
Financial Variables	Scholarship rate	(Total scholarship / Tuition and fees) * 100	PEUEC, Structural Adjustment
	Per-student educational expenditure	(Net operating expenses + depreciation) / total number of students	PEUEC, Structural Adjustment, OECD educational statistics
Faculty-related Variables	Number of students per full-time faculty member	Total number of students/Total number of full-time faculty members	PEUEC, OECD educational statistics
Industry-University Cooperation Variables	Number of students per Industry-University Cooperation Professors	Total number of Students/Total number of Industry-University Cooperation Professors	LINC
Year	Dummy	2010=0, 2011=1	

The scholarship rate variable is the ratio between scholarship amounts and total tuition and fees. Students consider this variable to indicate financial soundness and quality of a university, since most scholarships are based on students' academic achievements. Therefore, a higher scholarship rate means a greater proportion of academically superior students. Since the employment rate is regarded as a proxy of the quality of a university, a positive sign is expected in our estimation.

Per-student educational expense is a core indicator of the quality of a university. Per-student expense is the total educational expenses that a university spends on a student in a year and therefore includes expenses financed by tuition and fees, as well as contributions made by the university, government and other entities. In general, it is assumed that higher per-student expense leads to greater educational services for students; hence, a higher level of educational quality. Therefore, a

positive sign is expected in the estimation of the variable.

The number of students per full-time faculty member is most often used in OECD analyses, together with per-student expenses. The fiscal assistance programs adopt the rate of full-time faculty members, which means the ratio between the number of full-time faculty members recruited and the total number of full-time faculty positions authorized. However, since this rate does not indicate how the full-time faculty members are actually utilized for education, the number of students per full-time faculty member is a better indicator of educational quality. Since it is generally understood that reducing the number of students per instructor will improve educational quality, a negative sign is expected in the estimation.

The number of students per industry-academe cooperation professors has not been used in earlier studies as a determinant of the employment rate. However, this indicator is included in the estimation model since it is used in the LINC program, which succeeded NURI, and aims to enhance the employment rate by increasing relevance of higher education to the real world of work. It is expected to show a positive sign in the estimation.

In addition, a year dummy variable is used to see if there is any difference in the employment rate between the two years, 2010 and 2011, when the two-year data are pooled for estimation.

The data used in this study come from the Higher Education Statistics: 2010 and 2011, compiled by the Korea Education Development Institute (KEDI). Earlier annual data are not used mainly because the employment rate definition changed in 2010. The Higher Education Statistics are used since they include educational management indicators available from higher educational institutions.

The total number of Korean higher educational institutions used in this study is 259 colleges and universities. Excluded from the Higher Education Statistics are micro higher educational institutions, which have total enrollments of fewer than 1,000 students, annual intake of fewer than 300 students, or only one college. Also excluded are special purpose institutions, such as educational, religious, pharmaceutical, and medical colleges, and new colleges/universities, which have not yet produced any graduates. These institutions are mostly professional

colleges and may have extreme values in the employment rate that could distort the overall estimation and are in general free from graduate employment problems.

The total number of higher educational institutions classified by type and location is as follows:

Table 5 | Higher Educational Institutions Classified by Type and Location

Type of Colleges	Location of Colleges		Total
	Seoul Metro. Area	Other Areas	
Private	85	132	217
Public	6	36	42
Total	91	168	259

4. Analysis Results

4.1. Descriptive Statistics

The summary of statics is as follows:

Table 6 | Descriptive Statistics of the Regression Analysis

Variable	Obs.	Mean	Std. Dev.	Min	Max
Employment Rate (%)	259	53.69	8.9	22	100
Type Dummy	259	0.16	0.37	0	1
Location Dummy	259	0.65	0.48	0	1
Scholarship Ratio (%)	259	19.17	6.61	7.11	58.82
Per Student Expense (Won '000)	259	9,231.54	366.99	5,100.4	33,439.5
Students per Full- Time Faculty	259	32.87	7.84	10.5	56.6
Students per Industry Cooperation Professor	259	428.16	880.462	32.8	9,067
Year_Dummy	259	0.5	0.5	0	1

Table 6 (Continue)

Variable	Obs.	Mean	Std. Dev.	Min	Max
Per Student Expense Squared(Won '000)	259	9.86e+07	1.13e+08	2.60e+07	1.12e+09
Students per Faculty Squared	259	1,141.941	508.49	110.07	3,200.33
Per Student Expense* Students per Faculty	259	287,524	78,631.63	102,783.4	564,125.4

The average employment rate of all higher educational institutions is 53.7%, the scholarship rate 19.2%, per-student annual educational expense ₩9,231,540, the number of students per full-time faculty member 32.9, and the number of students per industry-academe cooperation professor 428 students.

The employment rate by the type and location of higher educational institutions is as follows:

Table 7 Employment Rates by Type and Location of Higher Educational Institutions

Variable		2010		2011	
		No. of Observation	Employment Rate	No. of Observation	Employment Rate
Type of Colleges	Public	21	53.0	21	55.3
	Private	108	51.9	109	55.2
Location of Colleges	Seoul Metropolitan	45	51.1	46	53.8
	Other Areas	84	52.6	84	56.0
Total/Average		258	52.1	260	55.1

The employment rate was higher in the non-capital city (provincial) areas than in the Seoul metropolitan area, which is consistent with earlier studies. However, the employment rate was higher among public higher educational institutions than private ones, which contradicts the findings of earlier studies (Ryu et al. 2010, Choi 2008).

The OLS estimation of equation (1) shows that the set of the independent variables explains the variations in the employment rate by 22%.

Table 8 The Results of the OLS Estimation of Equation (1)

Dependent Variable	College Graduates' Employment
Independent Variables	Model [1] (2010~2011)
Type	-4.454*** (1.517)
Location	2.644** (1.200)
Scholarship Rate	0.214** (0.088)
Per Student Expense	0.001*** (0.000)
Number of Students per Faculty	-0.222*** (0.081)
Number of Students per Industry-University Cooperation Professor	-0.001 (0.001)
Year	2.829*** (0.984)
Constant	49.214*** (4.358)
Adjusted R-square	0.22
Number of observations	259

Note: 1) ( ): standard error
2) * p<0.1; ** p<0.05; *** p<0.01

As expected, per-student educational expense and the number of students per full-time faculty member variables affect the employment rate at the one percent level of statistical significance. The employment rate rises as per-student educational expense increases and as the number of students per full-time faculty member decreases. The scholarship rate, as well as the location and type of institutions, and the year variables also affect the employment rate at the five percent significance level. However, the number of students per industry-university cooperation professor variable does not produce any statistically significant effects. The cause of insignificance is unclear. Either the policy of appointing a separate group of industry-university

cooperation professors for conducting collaborative training programs and internships on site in industry may be unworthy or the number of students per industrial cooperation professor may be too many to be efficient.

The independent variables that affects the employment rate at the highest level are the scholarship rate and the number of student per full-time faculty member, followed by per student expense. As the scholarship rate rises by 1%, the employment rate increases by 0.21 percentage points. Also, as the number of students per faculty member decreases by 1%, the employment rate increases by 0.22 percentage points. Therefore, to increase the employment rate, priority should be given to increasing scholarships to students in relation to tuition and fees, decreasing the number of students per faculty member, and increasing per-student educational expenses.

4.2. Increasing Return and Interactive Effects on the Employment Rate

The estimation model adopted in equation (1) assumes that the relationship between various inputs and the output (the employment rate) are sustained over the whole range of the changes in inputs. However, the production function theory in economics shows that as input increases, output can increase either at an increasing or decreasing rate. Likewise, per-student educational expense and the number of students per faculty member may not be linearly related to the employment rate over the whole range of the changes in these educational input variables. Therefore, to test whether the employment rate changes at a steady, increasing, or decreasing rate over the whole range of these input variables, a squared value of these two input variables can be added to the estimation equation (1).

The two input variables may also affect the employment rate by interacting with each other. For this purpose, an interactive term of the two input variables can be added to equation (1). Paek (2000) argues that although both human resources and physical resources contribute to educational output, they have both complementary and substitute relationships in producing educational outputs. The number of students

per faculty member is a human resource indicator, while per-student educational expense is an indicator that combines both human resources and physical resources.

As discussed earlier, to enhance the employment rate, per-student expenses can be increased and/or the number of students per faculty member can be decreased. To raise the level of per-student expenses, a policy maker should decide whether physical resources (educational facilities, equipment or materials, etc.) or human resources (the number of instructors or their compensation level, etc.) should be increased. If per-student educational expense is raised through increases in the number of instructors, the number of students per instructor will decline, which will improve the employment rate at an increasing rate. If per-student educational expense is raised by increasing compensation, the number of students per instructor will not change, which will not increase the employment rate at an accelerated rate. Therefore, depending on the nature of interaction of the two educational input variables (per-student expense and the number of students per instructor), the effects on the employment rate will be different, and it is desirable to add an interactive term of two educational input variables to equation (1).

Equation (2) shows the revised estimation model.

$$Y = a_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_4^2 + b_9x_5^2 + b_{10}x_4x_5 + e \quad (2)$$

where

x_4^2: (per student expense)^2
x_5^2: (the number of students per faculty member)^2
x_4x_5: (per student educational expense) x (the number of students per faculty member)

The rest of the symbols are defined in the same way as in equation (1).

The results of the OLS estimation of equation (2) are presented in comparison with equation (1), as follows.

Table 9 Results of the OLS Estimation of Equation (2) with Squared and Interactive Terms

Dependent Variable	College Graduates' Employment Rate	
Independent Variables	Model [1] (2010~2011)	Model [2] (2010~2011)
Type	-4.454*** (1.517)	-4.472*** (1.422)
Location	2.644** (1.200)	3.649*** (1.134)
Scholarship Rate	0.214** (0.088)	0.147* (0.082)
Per Student Expense	0.001*** (0.000)	-0.006*** (0.001)
Number of Students per Faculty	-0.222*** (0.081)	-2.747*** (0.663)
Number of Students per Industry-University Cooperation Professor	-0.001 (0.001)	-0.000 (0.001)
Year	2.829*** (0.984)	2.289** (0.916)
Per Student Expense Squared		6.90e-08*** (0.000)
Number of Students per Faculty Squared		0.016** (0.007)
Per Student Expense*Number of Students per Faculty		1.75e-04*** (0.000)
Constant	49.214*** (4.358)	113.929*** (16.324)
Adjusted R-square	0.22	0.34
Number of observations	259	259

Note: 1) ( ): standard error
2) * p<0.1; ** p<0.05; *** p<0.01

The results of the analysis show that R^2 increased ($\Delta R^2 = 0.12$), and the squared and interactive terms, which were added to equation (1) are significant at the usual significance level, affecting the employment rate positively. As in the estimation of equation (1), the university characteristics (type and location) and the year dummy variables all have a statistically significant coefficient. Also, as in equation (1), the number of students per industry-university cooperative professor is statistically insignificant. Further, as in equation (1), the number of students per instructor also has a statistically significant negative sign.

The only difference is that although per-student educational expense has a statistically significant positive sign in equation (1), it has a statistically significant negative sign in equation (2). However, it cannot be interpreted that per-student expense has a negative relationship with the employment rate. This is because the interactive term between per-student expense and the number of students per instructor and the squared per student expense variable has a statistically significant positive sign. This means that although per-student expense alone cannot affect the employment rate, it affects the employment rate in interaction with the number of students per instructor. Therefore, the impact of per-student expense cannot be judged by the per-student expense variable alone, but by its marginal product (E_a) of the employment rate, which can be derived from equation (2) as follows.

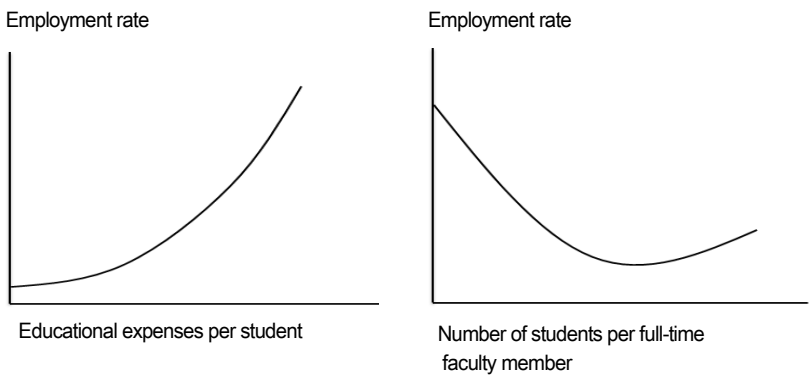
$$E_a = b_4 + 2b_8x_4 + b_{10}x_5 \tag{3}$$

The statistically significant coefficients b_8 (squared per student expense) and b_{10} (interactive term), and the average value of x_4 and x_5 can be taken from Table 6 to estimate equation (3). The result is a positive value, and therefore the per-student educational expense variable has a positive relationship with the employment rate in equation (2) as well.

Among the three added variables, the two squared variables (square of per-student expense and square of the number of students per faculty member) have a statistically significant positive coefficient. This means that per-student expense and the number of students per faculty member each brings an increasing return on the employment rate. On the one

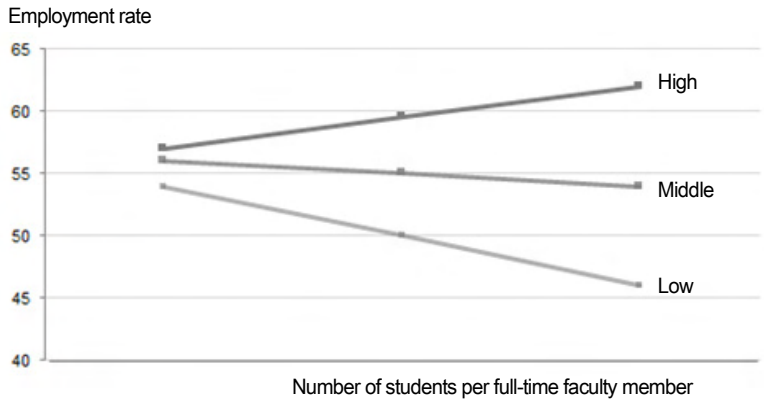
hand, although per-student expense has a statistically significant negative effect on the employment rate, as the level of expense rises, its effect on the employment rate increases at a statistically significant increasing rate (i.e. it increases non-linearly). On the other hand, as the number of students per faculty member increases, the employment rate declines. However, if the number of students per faculty member increases above a certain level, the employment rate starts rising. The negative effects of the high number of students per instructor must have been offset by the increases in per-student expense, resulting in the positive net effect on the employment rate. Figure 1 shows the relationships that per-student education expense and the number of students per faculty member have with the employment rate.

Figure 1 Relationship between per-student expense and the number of students per faculty with the employment rate



The estimation of equation (2) also confirms the existence of the interactive effect of two educational input indicators on the employment rate. The interactive term of per-student expense and the number of students per faculty member has a statistically significant positive coefficient. Figure 2 depicts the interactive effects visually.

Figure 2 Interactive effects of per-student expense and the number of students per instructor



When per-student educational expense is relatively high (₩12,800,000 and higher), even as the number of students per faculty member increases, the employment rate keeps increasing. Per-student expense must have offset the negative effect of the higher number of students per instructor on the employment rate, and the net effect of the interactive term on the employment rate is still positive. This indicates that the relatively higher per-student expense was achieved not by increasing the number of faculty members, but by recruiting better-quality instructors, raising faculty compensation, or improving educational facilities and equipment.

When per-student educational expense is relatively low (₩5,500,000 and lower), as the number of students per faculty member increases, the employment rate declines. This result can be explained as follows: When educational expense is relatively low, it cannot offset the negative effects of the increasing student-faculty ratio on the employment rate, and consequently the net effect of the interactive term on the employment rate is negative. This also indicates that the relatively lower per-student expense was achieved by a low investment in quality instructors, faculty compensation, educational facility and equipment, or mainly by a high number of students per instructor.

5. Conclusion

As higher educational institutions are responsible for fostering the high-level human resources needed for national development, almost all countries provide fiscal assistance to such institutions to enhance their educational quality. The Korean government has focused on the graduate employment rate at higher educational institutions as an important indicator reflecting educational quality. The need for enhancing the higher education quality and therefore the graduate employment rate is accentuated by the relatively low employment rates in recent years.

However, the government has carried out various fiscal assistance programs without clearly understanding the determinants of the graduate employment rate. Such an understanding, and a creative strategy for enhancing the employment rate based on that understanding, are prerequisites for sound program design and efficient operation. As Korea's level of assistance for higher education is relatively lower than in other OECD countries, efficient design and operation of the fiscal assistance programs are essential to ensure good educational outcomes.

This study empirically analyzes the determinants of the employment rate at higher educational institutions based on 2010-2011 data. The study results show that the three factors consistently contributing to employment rate determination are, besides the scholarship rate, the educational expenditure per student and the number of students per full-time faculty member.

These variables affect the employment rate significantly even after the type and location of higher educational institutions are controlled. To check the consistency of their influence on the employment rate, the squared variables of per student expenditure and the number of students per faculty member are added to the estimation model. The regression results show that as the values of these two variables increase, the employment rate improves at an increasing rate.

Also, to check whether these two variables interact and increase their influence on the employment rate, an interactive term of these two variables is added to the estimation model. The regression results show that when the educational expenditure per student is at the relatively high level, the employment rate keeps increasing, even as the number of

students per faculty member rises. Such phenomena can be explained by the theory that the relatively high educational expenditure per student is driven not by increases in the number of faculty members, but by recruitment of high quality faculty, greater compensation to existing faculty, or improvements in educational facilities and equipment. Accordingly, the positive effects of the rise in per-student educational expenditures on employment rates more than offset the negative effects of the increase in the student/faculty ratio on the employment rate, resulting in net positive effects. In contrast, when per-student educational expenditure is at the relatively low level, the employment rate keeps declining, as the number of students per faculty member increases. The relatively low educational expenditure per student must have been driven initially, not by measures to improve the quality of education, but by increases in the number of students per instructor.

The amount of scholarship vis-à-vis tuition and fees also makes positive effects on employment rates. Students on the scholarship must have tried hard to maintain the good academic standing, which must have led to a relatively higher employment rate. Therefore, universities with a higher scholarship rate could achieve better employment rates.

However, the number of students per industry-academe cooperation professor does not affect the employment rate in any statistically significant way. The fiscal assistance programs have supported various forms of industry-university cooperation (e.g. appointment of cooperation professors, conducting on-site training programs in plants, internship, and operation of special courses contracted with industry and entrepreneurship education, etc.). Either the programs are ineffective or the number of cooperation professors may be insufficient.

On the basis of this study, the following policy recommendations can be made for enhancing college graduates' employment rates through the improvement of the quality of higher educational institutions. The government has two options. One is to reward those higher educational institutions which achieve an absolutely high level of employment rates and/or which make good progress in improving employment rates on the basis of periodic assessments. An alternative is to play a more pro-active role by supporting higher educational institutions to manage the educational management indicators that are proven to be effective on the

basis of periodic empirical studies, as this study has shown, and monitor and evaluate their performance continuously. The essence is to select and concentrate on critical indicators. Instead of giving fiscal support for a large number of educational management indicators without any empirical basis, as has been done to date, priority of fiscal supports should be given to the scholarship rate, educational expenditure per student and the number of students per full-time faculty member when selecting and monitoring educational institutions in the fiscal assistance programs. In particular, the government should give greater weights to the scholarship rate and educational expenditure per student in selecting institutions and monitoring and evaluating their performance. These variables raise the employment rate irrespective of the number of students per faculty member through the effect of interaction with the number of students per instructor, resulting in net positive effects on the employment rate. When the educational expense per student is relatively high, even the increases in the number of students per faculty member will result in rises in educational budgets per student, which will keep increasing the employment rate. The current policy of the industry-university cooperation professors should be reviewed in-depth and be improved for a pilot program before launching a massive program. Instead of appointing a large group of separate industry-academe cooperation professors, the fiscal assistance funds can be used for the reorientation of the curriculum toward the demand by industry and the improvement in the quality of all existing faculty members and facilities.

Further investigations should follow up this study. The unit of observation in this study is the individual higher educational institution, and the available employment data for these institutions distinguish neither between the majors of studies, nor between regular and temporary workers. Future studies of employment rates should distinguish between the two. Also, this study does not utilize a time-series data set since the definition of the employment rate was modified in 2010. Future studies should be conducted on the basis of a panel data, which may shed better light on this study's topic.

References

<Korean Literature>

Ahn, Min Suk (2010) The Situation and Directions for Improvement of Lee Myung Bak Government's Fiscal Assistance for Higher Education, National Assembly Audit Policy Papers No.2, Seoul.

Ban, Sang Jin (1994) Educational Investment and Effective Education: Review of New Research Subjects of Educational Economics, *Educational Research* 9, Dongkuk University Education Research Institute, Seoul.

Ban, Sang Jin (1997) The Micro Approach to Educational Investment: Productivity of School Education, *Educational Finance and Economics Research* 6(2): 276-97.

Choi, Jung Yun (2008) *Analysis of the Quality of Higher Education in Korea* (II), Korea Education Development Institute, Seoul.

Higher Education Statistics (2011) Korea Education Development Institute, Seoul.

Hong, Min Sik (2012), Directions for Improvement of Evaluation Indicators for Higher Education Evaluation: with emphasis on the PEUEC, *Higher Education* 175: 68-75.

Jang, Soo Myung (2004) Analysis of Achievements and Directions for Improving Efficiency of Fiscal Assistance Programs for Higher Education, Research Papers, Ministry of Education and Human Resources, Seoul.

Joo, Hwi Jung (2014) The Current Situation and Challenges of High Level Human Resources Development through Industry-University Cooperation, *The HRD Review* (July).

Jun, Jae Sik and Min, Joo Hong (2009) Linking Education and Labor Markets and Achievement (1): Returns to Investment in Higher Education Quality, Korea Research Institute of Vocational Education and Training 137, Seoul.

Jung, Jin Hwa (2008) Comprehensive Evaluation of the 3-Year NURI Programs, Policy Research Paper, Korea Academic Research Promotion Foundation, Seoul.

Kim, Soo Kyung (2013) Directions for Improving the Competitiveness of Higher Education, *Higher Education* 179(23).

Oh, Ho Young (2014) Strengthening Universities' Employment Support Capacity and Promoting Entrepreneurship Education, *The HRD Review* (July).

Paek, Il Woo (2000) *Education Economics*, HakJi Sa : 318-320, Seoul.

Park, Kyung Ho (2010) Impact of the PEUEC on the Educational Conditions and Achievements: The First-Year Experience, *Educational Administration Research* 28(4):63-82.

Ryu, Jang Soo et al. (2010) Evaluation of the New University Regional Innovation, Policy Research Paper, Korea Research Foundation, Seoul.

<Overseas Literature>

Coleman, J.S., et al. (1996) *Equality of educational opportunity,* Washington, D.C.: Government Printing Office.

Hadderman, Margaret (1998) School productivity (ED420092). ERIC Clearinghouse on Educational Management, Eugene.

Hanushek, E.A. (1989) The Impact of Differential Expenditures on School Performance. *Educational Researcher,* 18(4), 45-51.

Hanushek, E.A., & Pace. R.R (1995) Who Chooses To Teach (and Why)? *Economics of Education Review*, 14(2), 101-117.

OECD (2011) Education at a glance 2011.

Ryan, J.F. (2004) The relationship between institutional expenditures and degree attainment at baccalaureate colleges, *Research in Higher Education*, 47(5).

CHAPTER 11

Effect of College Major on Earnings and Gender Gap in Labor Markets: Evidence from Young Adults in South Korea

by
Sungjin Cho
(Department of Economics Seoul National University)
Jihye Kam
(School of Education University of Wisconsin-Madison)
Soohyung Lee[*]
(Department of Economics and MPRC University of Maryland)

Abstract

This paper measures the impact of college major on employment and earnings. We use a nationally representative dataset from Korea, where there is little concern for endogenous major choice due to the nature of the college admission system and high school curriculum. We find that "Engineering" and "Medicine and Public Health" majors dominate in all labor market outcomes, but, different from existing studies of the U.S., in South Korea, graduates with an "Education" degree perform better than those who majored in "Natural Science or Mathematics." Finally, the large gender disparity in choice of major may account for approximately 50 percent of the gender gap in labor market outcomes.

* Corresponding author

1. Introduction

The impact of college major on earnings has attracted substantial attention from economists, sociologists and educational researchers (see Hamermesh and Donald, 2008, and Altonji et al, 2012, for overview of the literature). College graduates majoring in engineering are consistently found to have the highest earnings, usually followed by business and science in the United States, and graduates majoring in humanities and education earn the least.[1] The impact of college major on earnings has important implications for the gender gap in labor market outcomes because of the large gender disparity in choice of major and the disproportionate fraction of women who choose majors with low earnings potential for both sexes (e.g., Turner & Bowen, 1999; Joy, 2003; Gemici & Wiswall, 2013; Goldin, 2014). Despite the extensive examination of the relationship between college major and earnings, there is a need for additional research on this topic particularly because little research in this area accounts for endogenous selection into college major (see more in Hamermesh & Donald, 2008; Altonji et al., 2012). A few exceptional studies that address endogeneity include Kinsler and Pavan (2013), who use a structural model to account for college major choice, and Hastings et al. (2013), who apply a regression discontinuity approach to Chilean administrative datasets.

This paper aims to contribute to this literature by examining a setting where there is little concern for endogenous major choice. In South Korea, high school seniors have to select their college majors when they apply to a college, and switching their college major is extremely difficult, although they can select a secondary major to study a subject of interest. Moreover, in high school, students have virtually no opportunity to learn what they could expect from a specific college major in terms of course work and career prospects. For example, the Korean Ministry of Education tightly regulates the high school curriculum, which leaves no discretion to high schools in terms of subjects they would like to provide.

1 For example, Kinsler and Pavan (2013) report that college graduates with a science or business related degree earn up to 25% higher wages than other college graduates in the United States.

The current high school curriculum does not include career sessions or advanced studies that can influence students' choices of college majors. While, in theory, students can gather such information outside the formal school system, this is unlikely because the college admission system in Korea is extremely competitive and students devote as much time and effort as possible into achieving the highest possible score on the nationwide academic test.

Due to these institutional features, we can measure the effect of college major on labor market outcomes by regressing the outcomes on college major, controlling for students' characteristics at high school and local labor market conditions. We use the dataset from the "Graduates Occupational Mobility Survey (GOMS)," a nationally representative survey of young adults in Korea who graduated from either a 2-year or 4-year college program in a given survey year. Our sample consists of individuals who graduated from a college in August 2004 to February 2008 and includes their outcomes 20 months since the year of college graduation and 2 years afterward. We classify college majors into 7 groups: Engineering, Humanities, Social Science, Education, Natural Science and Mathematics, Medicine and Public Health, Arts and Athletics. The labor market outcomes we examine include labor market participation, employment status, job quality,[2] wages, and wage growth.

We find that, for young adults in our sample, college major makes fairly little impact on labor market participation and employment rate, but its impact is significant in terms of earnings and whether a college graduate works in a long-term fulltime position or works for a large company. For example, upon graduation, an engineering major is 13.4 percentage points more likely to be a long-term contract worker than a humanities major, conditional on nationwide college entrance test score, gender, age, survey year, and other characteristics. Conditional on having a long-term contract, an engineering major is 17.9 percentage points more likely to work for a large firm with 100 or more employees than a humanities major. As for earnings, an engineering major earns 20

2 Specifically, we measure whether a person has a job at a firm whose number of employees is greater than 100 individuals and works for a full-time job with long-term labor contracts (i.e., regular jobs instead of "irregular jobs").

percent more than his/her counterparts who majored in humanities. Such gaps between college majors persist 2 years after graduation. In our sample, engineering majors are comparable in labor market outcomes to those who majored in health-related fields to be trained as medical doctors, nurses, pharmacists, physiologists, chiropractors, dental hygienists, nutritionists, therapists, and other healthcare providers. They are followed by individuals who majored in either social sciences or education, then by those who majored in natural sciences or mathematics, in terms of having a regular job and earnings. Arts and athletics majors are the lowest and humanities majors are the second lowest in terms of labor market outcomes. The finding that engineering majors have better labor market outcomes than the rest of the majors in our sample is consistent with the findings in the literature explained earlier. However, contrary to the findings in the U.S., we find that those who majored in natural science or mathematics have worse outcomes than those who majored in social sciences or education.

Finally, we examine the effect of college major in explaining the gender gap in labor market outcomes. We conduct this analysis because South Korea is known for its large gender disparity in labor market outcomes; such a gap persists even after gender equality is achieved in educational attainment (OECD, 2010)[3]; and the South Korean government has been developing various incentives for firms to hire and maintain female employees, which appear to have had limited impact so far. Although our sample consists of young adults, almost all of them are single and have no children. Women exhibit weaker labor market performance than their male counterparts in terms of employment rate, having a long-term job, and earnings. For example, 20 months after the year of college graduation, women are 2 percentage points less likely to

3 The gender pay gap in South Korea was about 39 percent, which was 2.6 times larger than 15 percent, the average among the 28 OECD countries, followed by Japan, which was 21 percent, according to 2010 OECD statistics. It was reported as 29 percent in Germany and Israel, [this doesn't seem to make sense: if Germany and Israel had 29 percent, wouldn't they be second-highest, not Japan with 21 percent?] and 19 percent in the U.S., Finland, Switzerland, and Austria. In the U.K. and Czech Republic, women were paid an average of 18 percent less than men. In contrast, the lowest gender pay gap, 6 percent, was in Hungary, Poland, and Spain.

be employed, 10 percentage points less likely to work for a large firm with a long-term contract, and have 12 percent less earnings than their male counterparts. We find that controlling for college major reduces these gender gaps by 50 percent. However, college majors account very little for the gender gap in wage growth. These findings suggest that gender equality in college major choices can significantly reduce the gender gap in employment and having a quality job, but not the growth rate of earnings.

Our work proceeds as follows: section 2 presents the institutional background and data. Section 3 and Section 4 present the empirical framework and the main results, respectively. Section 5 examines the sensitivity of our results. Finally, section 6 concludes.

2. Institutional Background and Data

2.1. Institutional Background

This section conducts brief overview of the Korean education system, focusing on when and how a person chooses his/her college major. At a senior year at high school, a student needs to apply for a college as well as a major in Korea. Admission outcomes depend on an applicant's test score from the nationwide college entrance exam which is called the College Scholastic Ability Test (CSAT), relative academic performance at his/her high school, and academic and non-academic performances outside his/her school (e.g., award at the Mathematical Olympiad, and volunteer work). The format of CSAT managed by the Korea Institute of Curriculum and Evaluation is similar to the scholastic assessment test (SAT) in the U.S. The test is offered one time per year, typically in November. In general, high school graduates take CSAT in the beginning of November and receive the result in the end of November. They begin to apply for colleges with a specific major in the mid December. Those who pass the initial screening process based on certain criteria for each college receive an in-person interview or/and an essay exam in January. The admission results are announced in the beginning of February and admitted students should confirm their

intention to enroll in the program by late February.

The institutional autonomy is constrained in Korea (OECD, 2003). The Ministry of Education predetermines the number of seats in colleges and universities located within the National Capital Region (NCR)[4] as well as public universities instead of meeting the demand by students. The Ministry of Education also predetermines the number of seats in majors in health professions or teaching. Although private colleges and universities in local area have autonomy of the number of seats in other majors, they still need to submit a proposal for admission seats to the Ministry of Education. They can proceed in accordance with their plan upon the Ministry of Education's acceptance (Ministry of Education, 2013).

In Korea, a college has to announce the total number of seats across different majors before the college application process begins and the policy on how to evaluate applicants in terms of relative weight on each of the three categories. The collective evaluation is heavily weighted by CAST scores in general although it varies across colleges. A college can offer admissions only up to the seats allocated to each major. If an offer is turned down, then college can offer admissions to the next best applicant. Upon admission, students are technically allowed to change their majors or select additional major as their "double" major or "minor". Although some universities allow for change in major, it is practically impossible for a student to move to any popular major because of high competition and requirements for transfer application. To be eligible to change major, students must be in good academic standing and have completed one or two continuous academic years in a full-time degree program. A written exam or/and a personal interview are usually required and the number of permitted students is limited to less than 20 percent of each program. Students cannot change their major to any majors in teaching or health professions such as nursing, medicine, and pharmacy. Besides, the original major mostly matters in practice regardless of the double major or minor. Students are supposed to spend most of their time in taking classes of the original major while

4 The NCR is referred to Seoul and surroundings area where the colleges tend to be ranked higher than other area colleges.

declaring double major or minor requires taking much smaller number of courses and moreover, the student's academic rank for merit-based scholarships or fellowships informing how well he/she performed at college is determined based on the original major.

These institutional features imply that choice of college majors in our setting is relatively more exogenous in labor outcomes compared to the U.S. setting for two reasons. First, students need to make their decisions on college major without knowing much of what coursework and job opportunity a major has. In Korea, the opportunity of taking advanced classes such as honors classes, advanced placement (AP) courses, and the international baccalaureate (IB) program courses is not provided; high school curriculum is designed to put heavy weight on reading and literature, English, and mathematics, and other subjects unrelated to CSAT are so briefly taught in class that students do not get much of information regarding college majors and careers. Compulsory education time for high school averages 29.17 hours per week[5] that does not account for additional time spent in non-compulsory instruction time like self-study time after regular school hours and extra classes with private tutors and in private educational institutions, so-called *hagwon*. The proportion of the compulsory curriculum is devoted to reading and literature, English, and mathematics (38.29%).

Furthermore, career prospect is in some sense not a first order concern when a student chooses a major. For instance, consider a student who is good but not great to the point that she can get into some unpopular majors at the Seoul National University (SNU), considered the top university in Korea but cannot get into the college of medicine[6] at the SNU. Then, her choice will be made between an unpopular major at the SNU and the college of medicine at a less prestigious university regardless her interests or aptitude. These two factors make differences in the choice of college majors that reduce the issue of endogeneity.

---

5 Authors' calculations: 50 minutes per class × 7 classes per days × 5 days per week

6 In Korea, the colleges of medicine were comprised of 2-year premedical courses and 4-year medical program as an undergraduate program unlike the U.S. The new medical education system, so-called "4+4" system, 4-year medical program after bachelor's degree awarded, was introduced since 2006 but decided to abolish from 2015.

Thus, we do not address the causes of the ability sorting across majors.

2.2. Data and Summary Statistics

This study relies on two data sources. One is the Graduates Occupational Mobility Survey (GOMS), a nationally representative survey of young adults in Korea who graduated from either a 2-year or 4-year college program. The GOMS is a short-term panel that contacts survey participants after 2 years for follow-up. The data includes demographic information of individuals and their labor market outcomes 20 months after the college graduation and 2 years from the initial survey. Our sample consists of three waves of GOMS from 2005, 2007, and 2008. The 2005 GOMS includes individuals who graduated from a college in August, 2004 or February, 2005 and it surveys their initial labor market outcomes in 2005 and then conducts the follow-up survey after 2 years in 2007.[7] Note that the 2006 GOMS does not exist because of reconstructing the survey design in 2007 and the 2008 GOMS is the latest wave available to the public. 85.1 percent in the 2005 GOMS, 81.6 percent in the 2007 GOMS, and 85.0 percent in the 2008 GOMS survey participants re-participated in the follow-up survey. We narrow our sample only to 4-year college graduates (66.81 percent of the survey participants) for two reasons. First, for 4-year colleges, we have reliable information of their quality of students measured at the admission. Therefore, we will use that information to control for students' underlying cognitive ability that can affect students' major choice. Second, 2-year colleges in Korea are vocational schools typically tied to certain firms where they send their graduates to work, and vocational and 4-year colleges are not comparable to each other even if they provide the same major. By combining the initial and follow-up data, we conduct analysis of the dynamic effect of gender gap in the labor market outcomes in terms of employment status, wage level and growth as well as job mobility.

---

7 In Korea, the academic year begins in March and continues through December. An academic year had two regular semesters of spring and fall so that students graduate in February.

The second data source is the reports of college rankings from 2006 to 2013, which informs the minimum academic qualification for students to get an offer from a specific major in a college. There is no formal document informing college rankings based on CSAT scores but informal reports analyzed by large-scale *hagwons* with reputable brand name are distributed to help students applying to college. There reports suggest estimated cutoff points for admission based on CSAT score distribution in prior year. Three variables are available to identify CSAT proxies in GOMS: the regions where college is located, school types informing whether college is private or national/public, and school characteristics informing whether college is regular or a college of education[8]. Using 2006 data, we construct the index for mean rankings[9] over dummies for regions, school types, and school characteristics and

Table 1 Summary Statistics

Region (%)		Found (%)		School (%)	
- Seoul	20.62	- Private	80.93	- Regular	94.33
- Busan	6.70	- National/Public	19.07	- Colleges of Education	5.67
- Daegu	1.55				
- Incheon	3.61				
- Gwangju	3.61				
- Daejeon	4.12				
- Ulsan	0.52				
- Gyeonggi Privince	13.92				
- Gangwon Privince	5.15				
- North Chungcheong Privince	5.15				
- South Chungcheong Province	9.79				
- North Jeolla Province	4.64				
- South Jeolla Province	5.15				
- North Gyeongsang Province	9.79				
- South Gyeongsang Province	4.12				
- Jeju Province	1.55				

8 There are 11 colleges of education where provide specifically public elementary school teacher training programs. These colleges are government-run institutions.

9 A sorting algorithm is implemented by ranking them from the highest number to the lowest for a given year. The higher number means the higher ranked-institution.

Table 2 CSAT Proxies

Rank 2006	
Private	-42.784***
	(9.169)
University of Education	63.212***
	(14.899)
Region (Reference to Seoul)	
- Busan	-65.979***
	(13.164)
- Daegu	-50.921*
	(24.763)
- Incheon	-17.058
	(16.679)
- Gwangju	-90.058***
	(16.679)
- Daejeon	-50.634**
	(15.680)
- Ulsan	-52.661
	(40.950)
- Gyeonggi Privince	-45.513***
	(10.092)
- Gangwon Privince	-85.317***
	(14.415)
- North Chungcheong Privince	-67.817***
	(14.415)
- South Chungcheong Province	-69.702***
	(11.274)
- North Jeolla Province	-87.724***
	(15.067)
- South Jeolla Province	-101.974***
	(14.646)
- North Gyeongsang Province	-95.638***
	(11.282)
- South Gyeongsang Province	-82.731***
	(15.852)
- Jeju-do	-84.921***
	(24.763)
R-square	0.527
No. of observations	194

Notes: OLS regression model. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3 | Correlations

	rank 2006	rank 2009	rank 2010	rank 2013	rank 2014
rank 2006	1.000				
rank 2009	0.850	1.000			
rank 2010	0.853	0.986	1.000		
rank 2013	0.903	0.894	0.893	1.000	
rank 2014	0.917	0.883	0.880	0.980	1.000

Table 4 | Spearman's Rank Correlation Coefficients

	rank 2006	rank 2009	rank 2010	rank 2013	rank 2014
rank 2006	1.000				
rank 2009	0.850***	1.000			
rank 2010	0.843***	0.978***	1.000		
rank 2013	0.900***	0.903***	0.905***	1.000	
rank 2014	0.913***	0.904***	0.900***	0.971***	1.000

Notes: The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

test its validity using the ordinary least squares (OLS) regression and a Spearman's rank correlation with the CSAT proxies in other years. Table 1 shows summary statistics of the input data. Nearly 35 percent of 4-year colleges are located in Seoul or Gyeonggi province. 81 percent are private colleges and 6 percent are special colleges of education. When we regress rankings of average minimum scores for each college over dummies for regions, school types, and school characteristics, R-square is 0.527 as seen in Table 2. The R-square is high enough to believe that our CSAT proxies are in fact predicted by regions, school types, and school characteristics. The correlation and Spearman's rank correlation of our CSAT proxies with ones in other years show that there is a statistically significant correlation between measurements at 1 percent level of significance (see Table 3 and 4).

Table 5 reports summary statistics from the initial and follow-up surveys, respectively. Before we explain details of those tables, several variables require explanations. We consider a person is employed if the person works at least 1 hour in the previous week from when the survey was conducted or has a job but not working due to family and childcare,

Table 5 Summary Statistics

	Initial Survey		Follow-up Survey	
	Male (1)	Female (2)	Male (3)	Female (4)
No. of observations	22,953	18,305	19,382	14,568
Age	27.96	25.62	29.91	27.56
Married (%)	10.75	7.25	28.79	19.16
College major (%)				
• Humanities	9.21	18.96	9.13	18.16
• Social Science	22.85	22.81	22.84	22.80
• Education	4.86	14.02	4.54	15.16
• Engineering	39.91	10.12	40.51	10.19
• Natural Science/Mathematics	13.28	16.32	13.46	16.44
• Medicine/Public Health	3.18	5.12	3.10	5.14
• Arts/Athletics	6.73	12.65	6.43	12.12
In the labor force (%):	78.47	75.71	88.64	81.97
Employed in the labor force (%)	96.07	95.40	86.10	84.79
Among those employed:				
- Monthly Earnings (10,000 2010 won)	244.00	192.65	280.26	217.64
- Not regular position (%)	16.86	26.10	9.73	16.95
- Regular position (%)	83.14	73.90	90.27	84.79
- Among regular position (%):				
• Working at a large-scale firm	47.24	30.99	48.00	31.66
Imputed SAT score (standardized)	-0.32	-0.19	-0.31	-0.18

Notes: All gender differences are statistically significant at 1 percent level (average is based on t-test and the distribution of college majors is based on Kolomogrow and Smirnov test).

strike, or stoppage of operation. Among employed, a person holds a "irregular position" if the person has a fixed-term contract, part-time job, or freelancer. "A large-scale firm" refers to a firm hiring 100 or more employees. A person's earning is reported in a yearly, monthly, weekly or hourly basis in GOMS. We convert the reported earnings to monthly basis using reported hours of work. Finally, for college majors, we classify college majors into 8 groups: humanities, social science, engineering, education, natural science/mathematics, medicine/public health[10], and arts/athletics.

Initial survey is composed of 22,953 male and 18,305 female

10 Medicine/public health includes veterinary medicine, nursing, and pharmacy.

respondents, while follow-up survey contains smaller number of respondents due to attrition. On average, male respondents are two years older than female respondent in both surveys. This is because Korean men over the age of 18 have to provide about 2-year compulsory military service unless they are disabled or under special condition. Thus, men enter the labor market 2 years later than women in common. A notable gender-gap was observed in labor market outcomes between men and women. For example, the average monthly earnings are 2,440,000 won[11] for men, over 30 percent higher than that of female employees. Labor force participation rate was higher for male respondents and employment rate was slightly higher for male respondents in both surveys while female respondents have higher imputed CSAT score compared to male respondents. Male respondents are shown to be more likely to earn more, to hold regular position, and to be employed at large-scale firm. In other dimensions, male college graduates outperform female graduates in the sense that the former are more likely to have a regular position and working for a large-scale firm that tend to provide higher earnings and more job security. All of these differences are statistically significant at 1 percent level, based on two-sided t-tests. Overall, the labor market participation rate, the proportion of regular workers, the likelihood of working at large firm, and average monthly wage are increased in both men and women but the gender-gap in the labor market outcomes remains similar in the follow-up surveys.

For the rest of this paper, we decompose the gender gap in the labor market outcomes so that we can examine the extent to which the gender-gap in college majors may account for. Among male respondents, more than 75 percent majored in engineering (40%), natural science/mathematics (13%), or social science (23%). On the other hand, female respondents' college majors are distributed across humanities (19%), education (23%), education (14%) or natural science/mathematics (16%). The distribution of college majors among men is different from that of women at 1 percent significance level, based on the Kolmogorv-Smirnov test. The difference in college majors could account for gender gap if a college major may attribute to different types of human capital

11 Approximately, 2,440 dollars.

and their values in the labor markets are different. To discuss gender differences in the labor market outcomes more in detail, we perform probit analyses on the data. We establish the link between the difference in college majors and labor market outcomes between men and women. This paper discuss what extent college majors explain the gender gap in labor market outcome in terms of employment status, wage level and growth, and job mobility across firms and industries.

3. Empirical Framework

We estimate probit models to examine the role of college major on labor market participation, employment, and likelihood of having a quality job. We denote by $Y_{i,j,c,l,t}$, a labor market outcome of person i who majored in j, graduated from a college in year c, lives in location l and was surveyed in round $r \in \{0,1\}$:

$$Y_{i,j,c,l,r} = 1(Y^*_{i,j,c,l,r} > 0)$$

$$Y^*_{i,j,c,l,r} = \alpha_r female_i + \beta_r CSAT_i + \gamma_r age_{i,r} + \delta_r age^2_{i,r} + \theta_{j,r} + \rho_{c,r} + \mu_{l,r} + \varepsilon_{i,j,c,l,r}, \quad (1)$$

where $female_i$ is 1 if person i is female and 0 if male; $CSAT_i$ is person i's test score for college admission tests, and $age_{i,r}$ is the persons' age at survey round r. Round r is 0 if the survey is conducted 20 months since the year of college graduation, and 1 if it is conducted 2 years after the initial survey. Variables $\rho_{c,r}$ and $\mu_{l,r}$ capture cohort and location specific fixed effects, respectively. The parameters are allowed to vary by survey rounds. The parameter of interest is $\theta_{j,r}$ that measures the relationship between a person's college major and his/her labor market outcomes. The identification assumption is that conditional on a person's CSAT score and other observable characteristics, a person's choice of college major is exogenous. Such assumption is likely to hold because students have very limited information of college major at the time of their choice and thus it is unlikely that students will sort themselves according to their comparative advantage as explained in

Section 2. When we analyze earnings, we estimate Mincerian equations that regress the logarithm of income on college major and other observables (Mincer, 1976).

4. Findings

4.1. Employment

Table 6 and Table 7 show the effect of college major on employment status in initial and follow-up surveys, respectively. We calculate the distribution of labor market status depending on a person's college major and regress employment status on college majors using a probit model. The results in the initial survey and the follow-up survey are similar to each other. Test score shows no significant effect on labor force participation or employment but has significantly positive effect on the probability to be a regular worker and to work at big firm. College major shows significant effects, especially on employment related probabilities.

Surprisingly, the probability to be in the labor force is higher for female in initial survey while it is lower for female in follow-up survey. Medicine/public health major has the highest likelihood to be in the labor force in both surveys. On the other hand, natural science/mathematics and humanities majors have the lowest likelihood to be in the labor force in initial and follow-up survey, respectively. Among those in the labor force, employment probability is lower for female. Engineering major are the most likely to be employed and arts/athletics are the least likely to be employed. Among those employed, the probability to be a regular worker is the highest for engineering major in both surveys while it is the lowest for arts/athletics in initial survey and medicine/public health in follow-up survey. Among regular workers, medicine/public health major has the highest likelihood and education major has the lowest likelihood to work at big firm in both surveys. On average, females have lower probability to be a regular worker and to work at large-scale firm.

Table 6 Gender Gap in Employment: Initial Survey

Dependent variable	1: labor force 0: not in the labor force	1: employed 0: not employed but in the labor force	1: regular workers 0: not regular workers but employed	1: regular at big firm 0: not regular workers at big firm but regular workers
	(1)	(2)	(3)	(4)
No. of observations	41,258	31,796	30,454	24,100
Panel A: No Major Controls	0.018**	-0.020***	-0.031***	-0.105***
Female (a)	(0.006)	(0.003)	(0.007)	(0.010)
	0.002	0.006***	0.032***	0.047***
Test score	(0.003)	(0.001)	(0.003)	(0.004)
	0.016	0.062	0.054	0.046
Pseudo R-square				
Panel B: Major controls	0.019**	-0.011***	-0.006	-0.054***
Female (b)	(0.006)	(0.003)	(0.007)	(0.010)
	0.004	0.002	0.028***	0.094***
Test score	(0.003)	(0.001)	(0.003)	(0.005)
College major (Reference=Engineering)				
Humanities	-0.023**	-0.026***	-0.134***	-0.179***
	(0.007)	(0.005)	(0.010)	(0.010)
Social Science	0.005	-0.019***	-0.032***	-0.131***
	(0.006)	(0.004)	(0.007)	(0.008)
Education	0.007	-0.004	-0.042***	-0.394***
	(0.008)	(0.005)	(0.011)	(0.006)
Natural Science/Mathematics	-0.070***	-0.035***	-0.095***	-0.130***
	(0.007)	(0.006)	(0.009)	(0.010)
Medicine/Public Health	0.099***	-0.029***	-0.114***	0.067***
	(0.009)	(0.008)	(0.014)	(0.018)
Arts/Athletics	0.029***	-0.140***	-0.183***	-0.258***
	(0.008)	(0.010)	(0.012)	(0.009)
Pseudo R-square	0.022	0.112	0.069	0.102
Panel C: Gap explained by Majors	-0.058	0.381***	0.813***	0.480***
1 - (b)/(a)	(0.096)	(0.074)	(0.195)	(0.058)

Notes: Probit model, marginal effects reported. Sample of the 2005-2008 GOMS1. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 7 Gender Gap in Employment: Follow-up Survey

Dependent variable	1: labor force 0: not in the labor force	1: employed 0: not employed but in the labor force	1: regular workers 0: not regular workers but employed	1: regular at big firm 0: not regular workers at big firm but regular workers
	(1)	(2)	(3)	(4)
No. of observations	33,950	29,112	24,909	21,754
Panel A: No Major Controls				
Female (a)	-0.052***	-0.024***	-0.049***	-0.158***
	(0.005)	(0.005)	(0.006)	(0.009)
Test score	0.008**	0.010***	0.026***	0.043***
	(0.003)	(0.003)	(0.003)	(0.005)
Pseudo R-square	0.017	0.071	0.039	0.041
Panel B: Major controls				
Female (b)	-0.045***	-0.013*	-0.028***	-0.076***
	(0.005)	(0.005)	(0.006)	(0.010)
Test score	0.007*	0.007*	0.022***	0.091***
	(0.003)	(0.003)	(0.003)	(0.005)
College major (Reference=Engineering)				
Humanities	-0.057***	-0.024**	-0.094***	-0.194***
	(0.008)	(0.008)	(0.010)	(0.010)
Social Science	-0.023***	-0.013*	-0.021**	-0.121***
	(0.006)	(0.006)	(0.007)	(0.009)
Education	0.015*	-0.007	-0.032***	-0.405***
	(0.007)	(0.008)	(0.010)	(0.006)
Natural Science/Mathematics	-0.053***	-0.055***	-0.073***	-0.137***
	(0.007)	(0.008)	(0.009)	(0.010)
Medicine/Public Health	0.046***	-0.001	-0.137***	0.071***
	(0.009)	(0.011)	(0.015)	(0.019)
Arts/Athletics	-0.034***	-0.129***	-0.125***	-0.249***
	(0.008)	(0.010)	(0.012)	(0.011)
Pseudo R-square	0.024	0.083	0.055	0.099
Panel C: Gap explained by Majors	0.123***	0.444***	0.422***	0.522***
1 - (b)/(a)	(0.034)	(0.118)	(0.061)	(0.040)

Notes: Probit model, marginal effects reported. Sample of the 2005-2008 GOMS3. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Then, we explore to what extent college majors explain labor market outcomes using delta method with estimates from probit regression models. As seen in Panel B of Table 6, the size of f coefficients of female dummy significantly decreases as compared to ones in Panel A. In Panel C of Table 6, the large gender disparity in choices of major account for approximately 38 percent and 81 percent of the gender gap in employed status and regular worker status among the employed at 1 percent significance level. Approximately, 48 percent of the gender gap in regular worker at large-scale firm status can be explained by college majors. After 2 year, nearly 12 percent of the gender gap for labor market participation can be revealed by college major choice. Around 40 percent of the gender gap for employed status and regular worker status can be captured by the gender gap in college major choice as reported in Panel C of Table 7. The gender disparity in choices of major can explain approximately 50 percent of the gender gap for regular worker status at large-scale firm.

4.2. Wages

Table 8 shows the returns to postsecondary education with respect to college majors (columns 1 & 2) and the wage growth between the initial placement and 2-year later. The sample is restricted to employed workers. We regress wage level and growth on college majors using an OLS regression model. On average, female workers earn less and experience slower wage growth compared to male workers. Test score has significantly positive effect on the level of earnings but has significantly negative effect on the growth of earnings. Age as a proxy for work experience has significantly positive effect on wage level but negative effect on wage growth. Married workers earn higher income but experience slower wage growth, which can be attributed to longer work experience of married workers. With respect to college major, medicine/public health majors earn the highest level of income while arts/athletics majors earn the lowest level of income in both surveys. Natural science/mathematics majors experience the fastest wage growth while education majors experience the slowest wage growth. The gap also widens for education major over two years. This is because teacher's

Table 8 Gender Gap: Earnings

Dependent var.	Log monthly earnings	Log monthly earnings	Change in the log monthly earnings
Sample	Employed, Initial survey (1)	Employed, Follow-up survey (2)	Employed, Both surveys (3)
No obs.	30,242	24,767	22,717
Panel A: No Major controls			
Female	-0.121***	-0.191***	-0.052***
	(0.008)	(0.007)	(0.008)
Test score	0.104***	0.086***	-0.018***
	(0.004)	(0.003)	(0.003)
Age	0.089***	0.076***	-0.018**
	(0.006)	(0.006)	(0.006)
Age-squared	-0.001***	-0.001***	0.000
	(0.000)	(0.000)	(0.000)
Married	0.105***	0.090***	-0.035***
	(0.010)	(0.006)	(0.009)
R-squared	0.138	0.161	0.015
Panel B: Major controls			
Female	-0.078***	-0.147***	-0.050***
	(0.008)	(0.007)	(0.008)
Test score	0.100***	0.088***	-0.011**
	(0.004)	(0.003)	(0.004)
Age	0.084***	0.076***	-0.017**
	(0.006)	(0.006)	(0.006)
Age-squared	-0.001***	-0.001***	0.000
	(0.000)	(0.000)	(0.000)
Married	0.097***	0.081***	-0.033***
	(0.010)	(0.006)	(0.009)
College major (Reference=Engineering)			
- Humanities	-0.207***	-0.196***	0.010
	(0.009)	(0.009)	(0.009)
- Social Science	-0.048***	-0.058***	-0.012
	(0.007)	(0.007)	(0.007)
- Education	-0.049***	-0.102***	-0.060***
	(0.011)	(0.010)	(0.010)
- Natural Science/Mathematics	-0.151***	-0.115***	0.036***
	(0.009)	(0.008)	(0.008)
- Medicine/Public Health	0.134***	0.102***	-0.003
	(0.013)	(0.013)	(0.013)
- Arts/Athletics	-0.329***	-0.303***	0.011
	(0.010)	(0.010)	(0.010)
R-squared	0.186	0.208	0.019
Panel C: Gap explained by Majors	0.350***	0.233***	0.030
1 - (b)/(a)	(0.031)	(0.016)	(0.036)

Notes: OLS regression model. Sample of the 2005-2008 GOMS1 & 3. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

wage has not experienced a dramatic increase at the early period of work experience. The statistically insignificant coefficients in column 3 reveal that there is a hierarchy of wage structure for each occupation. That means there is a similar wage growth within same firm regardless of workers' college majors based on their seniority.

Except female dummy, most point estimates do not show notable change when college major dummies are included. That is, the downward bias of estimate for female dummy, which can be attributed to different college major choice by gender, is relieved by the inclusion of college majors. Nearly 35 percent and 23 percent of the gender gap between wages in initial placement and 2-year later can be explained by differences in college major choice by gender as presented in Panel C of Table 8. College majors, however, do not have a statistically significant impact on wage growth.

4.3. Dynamics

Table 9 show the job mobility between the initial placement and 2-year later. The sample is restricted to who are employed in both initial year and the follow-up years to see the chance to change their firm or industry. We regress job mobility across firms (column 1) and industries (column 2) using a probit model. When a worker is employed in both periods, not including college major dummies, the probability to change a firm is higher for female workers but becomes insignificant once we control for college major. Higher test score is associated with lower probability to change a firm regardless of inclusion of major dummies. Arts/athletics majors show the highest likelihood while education majors show the lowest likelihood to change a firm. This is because public school teachers have to teach for 4 years in one school and then can move to another school while private school teachers are unlikely to move to another school. Surprisingly, approximately 77 percent of the gender gap for changes in firm can be explained by gender differences in college major choice.

On the other hand, female workers face lower probability to change industry regardless of the inclusion of college major dummies. Higher test score is associated with lower likelihood to change industry. With

Table 9 Gender Gap in Job Mobility

Dependent variable	1: firm change 0: not change	1: industry change 2: not change
Sample	Employed, Both surveys (1)	Employed, Both surveys (2)
No. of observations	22,985	22,985
Panel A: No Major Controls		
Female (a)	0.026**	-0.019*
	(0.009)	(0.008)
Test score	-0.074***	-0.041***
	(0.004)	(0.003)
Pseudo R-square	0.047	0.027
Panel B: Major controls		
Female (b)	0.006	-0.024**
	(0.009)	(0.008)
Test score	-0.065***	-0.030***
	(0.004)	(0.004)
College major (Reference=Engineering)		
Humanities	0.107***	0.038***
	(0.012)	(0.009)
Social Science	0.030***	0.020**
	(0.009)	(0.007)
Education	-0.040***	-0.101***
	(0.012)	(0.008)
Natural Science/Mathematics	0.073***	0.038***
	(0.011)	(0.009)
Medicine/Public Health	0.042**	-0.054***
	(0.016)	(0.011)
Arts/Athletics	0.186***	0.091***
	(0.014)	(0.012)
Pseudo R-square	0.059	0.042
Panel C: Gap explained by Majors		
1 - (b)/(a)	0.771***	-0.288*
	(0.292)	(0.166)

Notes: Probit model, marginal effects reported. Sample of the 2005-2008 GOMS1 & 3. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 10 Gender Gap in Employment by High School Track

Dependent variable	1: labor force 0: not in the labor force	1: employed 0: not employed but in the labor force	1: regular workers 0: not regular workers but employed	1: regular at big firm 0: not regular workers at big firm but regular workers
	(1)	(2)	(3)	(4)
Panel A: Humanities and Social Science Track – Initial				
Female with no majors (a)	-0.005	-0.016***	-0.029**	-0.062***
	(0.009)	(0.005)	(0.011)	(0.015)
Female with majors (b)	-0.001	-0.012**	-0.017	-0.060***
	(0.009)	(0.004)	(0.011)	(0.015)
Gap explained by Majors: 1 -(b)/(a)	0.760	0.143*	0.418**	0.017
	(1.338)	(0.082)	(0.176)	(0.075)
No. of observations	18,159	14,078	13,419	10,317
Panel B: Humanities and Social Science Track – Follow-up				
Female with no majors (a)	-0.040***	-0.026***	-0.055***	-0.147***
	(0.007)	(0.008)	(0.008)	(0.014)
Female with majors (b)	-0.035***	-0.015	-0.037***	-0.082***
	(0.008)	(0.008)	(0.009)	(0.015)
Gap explained by Majors: 1 -(b)/(a)	0.117*	0.407***	0.322***	0.441***
	(0.062)	(0.156)	(0.068)	(0.057)
No. of observations	14,305	12,170	10,354	9,023
Panel C: Math and Science Track – Initial				
Female with no majors (a)	0.043***	-0.010*	-0.003	-0.096***
	(0.009)	(0.004)	(0.010)	(0.015)
Female with majors (b)	0.048***	-0.004	0.019	-0.046**
	(0.009)	(0.004)	(0.009)	(0.016)
Gap explained by Majors: 1 -(b)/(a)	-0.112*	0.498**	6.490	0.516***
	(0.059)	(0.250)	(19.152)	(0.105)
No. of observations	19,274	14,667	14,204	11,644
Panel D: Math and Science Track – Follow-up				
Female with no majors (a)	-0.051***	-0.020*	-0.044***	-0.158***
	(0.008)	(0.008)	(0.009)	(0.015)
Female with majors (b)	-0.047***	-0.013	-0.021*	-0.073***
Panel A: Humanities and Social Science Track – Initial	(0.008)	(0.008)	(0.009)	(0.016)
Gap explained by Majors: 1 -(b)/(a)	0.077	0.352*	0.494***	0.545***
	(0.051)	(0.190)	(0.118)	(0.066)
No. of observations	14,755	12,784	11,025	9,677

Notes: Probit model, marginal effects reported. Sample of the 2005-2008 GOMS1 & 3. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include test scores, age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 11 Gender Gap: Earnings by High School Track

Dependent variable	Log monthly earnings: Initial Survey	Log monthly earnings: Follow-up survey	Change in the log monthly earnings
Sample	Employed, Initial survey (1)	Employed, Follow-up survey (2)	Employed, Both surveys (3)
Panel A: Humanities and Social Science Track			
Female with no majors (a)	-0.120***	-0.194***	-0.033**
	(0.012)	(0.010)	(0.012)
Female with majors (b)	-0.098***	-0.155***	-0.034**
	(0.012)	(0.010)	(0.012)
Gap explained by Majors: 1 - (b)/(a)	0.181***	0.200***	-0.045
	(0.031)	(0.022)	(0.055)
No. of observations	13,328	10,300	9,730
Panel B: Math and Science Track			
Female with no majors (a)	-0.057***	-0.168***	-0.067***
	(0.012)	(0.010)	(0.012)
Female with majors (b)	-0.030**	-0.128***	-0.064***
	(0.012)	(0.010)	(0.012)
Gap explained by Majors: 1 -(b)/(a)	0.471***	0.240***	0.042
	(0.117)	(0.026)	(0.041)
No. of observations	14,109	10,974	10,968

Notes: OLS regression model. Sample of the 2005-2008 GOMS1 & 3. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include test scores, age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

respect to college major, arts/athletics majors are the most likely to change industry while education majors are the least likely because of the college major specific human capital. It is an interesting feature that humanities, social science, and natural science/mathematics majors are less likely to change industry rather than engineering majors. When dummies for college major are included, 28 percent for the gender gap of change in industry can be explained by gender-specific college major choice.

Table 12 Gender Gap in Job Mobility by High School Track

Dependent variable	1: employer change 0: not change	1: industry change 2: not change
Sample	Employed, Both surveys (3)	Employed, Both surveys (4)
Panel A: Humanities and Social Science Track		
Female with no majors (a)	0.012	-0.008
	(0.014)	(0.012)
Female with majors (b)	0.002	-0.010
	(0.015)	(0.012)
Gap explained by Majors: 1 - (b)/(a)	0.851	-0.373
	(1.051)	(0.706)
No. of observations	9,846	9,846
Panel C: Math and Science Track		
Female with no majors (a)	0.026	-0.037***
	(0.014)	(0.011)
Female with majors (b)	0.014	-0.035**
	(0.014)	(0.011)
Gap explained by Majors: 1 - (b)/(a)	0.436	0.031
	(0.271)	(0.080)
No. of observations	11,091	11,091

Notes: Probit model, marginal effects reported. Sample of the 2005-2008 GOMS1 & 3. Dummies for entrance years, survey years, and residence fixed effects are included. Other controls include test scores, age, age-squared and dummy for being married. The standard errors are in parentheses. The asterisks *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

5. Robustness Checks

Two robustness tests for our results are studied in order to address validity in this section. First, we include individuals who are not in the labor force to simply test two concerns. One is that in Korea, young adults try to stay at a graduate school or other form while they look for a job. They are considered effective unemployed but shown as not in the labor force. The other is that for the policy makers point of view, not in

the labor force could be a cost in that not supplying for qualified labor forces to the labor market especially considering the fact that Korea has been experiencing drastic aging of workers. The coefficients are not significantly different across these two distinct definitions of employment status.

Second, we conduct a simple robustness check to allow for differences in high school tracks. In Korea, students choose a curriculum track between humanities/social science track and mathematics/science when they become sophomores in high school. Reading, literature, and English are more emphasized in the humanities/social science track while mathematics and natural science education is more highlighted in the mathematics/science track. The CSAT subject tests that students take are based on their high school track. A large proportion of male students choose the mathematics/science track while many female students are inclined to choose the humanities/social science track. Throughout, we report regression results with two different high school tracks in Tables 10, 11, and 12.

Regardless of both tracks, we observe that the probability to be in the labor force, to be employed, to be a regular worker, and to be a regular worker at large-scale firm is lower for female workers as reported in Table 10. According to Table 11, female workers earn less and experience slower wage growth compared to male workers in both high school tracks. In Table 12, no statistically significant gender gap is observed in humanities/social science track, while female workers face lower likelihood to change employer or change industry in mathematics/science track. The magnitude is larger in the follow-up survey compared to the initial survey. The overall tendency for estimates to shrink toward zero is common across surveys and tracks when college major dummies are included. The proportions of gender gap for labor market outcomes in terms of employment status, wage level and growth, and job mobility that can be explained by gender differences in college major choices are also similar to our results.

6. Conclusion

Using a nationally representative datasets of young Korean adults, we have examined the impact of college major on labor market outcomes and its role in accounting for gender gap. We find sizable returns from majoring in engineering and medicine/public health, followed by social science and education, then by national science/ mathematics. Majors of humanities, arts/athletics pay the least. These variations across college major account for approximately 50 percent of the gender gap in employment, working for a long-term contract job, and earnings.

A limitation of our current study is that we focus on the labor market outcomes up to three years after the college graduation, because of data availability. Within this threeyear window, we find that the gap across college major and gender persisted and often widened. Therefore, it will be an important future research of examining a longer-term effect of college major on labor outcomes and gender disparity. Another related issue for further research is to examine if students optimally choose college major and if not, what would be policy tools to improve the efficiency. A few recent studies report that a person's subjective expectation on earnings plays an important role in selecting college major in the U.S. (e.g., Betts, 1996; Arcidiacono et al, 2012; Zafar, 2012; Zafar and Wiswall, 2013). Therefore, it will be important to examine the accuracy of such expectation and the gender gap in forming accurate expectation associated with college major choice.

References

Arcidiacono, P., Hotz, V.J., & Kang, S. (2012). "Modeling College Major Choices using Elicited Measures of Expectations and Counterfactuals." *Journal of Econometrics*, 166(1), 3-16.

Altonji, J. G., Blom, E., & Meghir, C. (2012). "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers." *Annual Review of Economics, 4*(1), 185-223.

Betts, J. R. (1996). "What Do Students Know About Wages? Evidence from a Survey of Undergraduates." *Journal of Human Resources, 31*(1), 27-56.

Black, D., Taylor, L., & Sanders, S. (2003). "The Economic Reward to Studying Economics." *Economic Inquiry, 41*(3), 365–377.

Gemici, A., & Wiswall, M. (2013). "Evolution of Gender Differences in Post-Secondary Human Capital Investments: College Majors." *International Economic Review, 55*(1), 23-56.

Goldin, C. (2014). "A Grand Gender Convergence: Its Last Chapter." *American Economic Review, 104*(4), 1-30.

Grogger, J., & Eide, E. (1995). "Changes in College Skills and the Rise in the College Wage Premium." *Journal of Human Resources, 30*(2), 280–310.

Hastings, J. S., Neilson, C., & Zimmerman, S. D. (2013). "Are Some Degrees Worth More than Others? Evidence from College Admission Cutoffs in Chile," (Working paper No. 19241). Retrieved from National Bureau of Economic website: http://www.nber.org/papers/w19241.

Joy, L. (2003). "Salaries of Recent Male and Female College Graduates: Educational and Labor Market Effects," *Industrial and Labor Relations Review, 56*(4), 606-621.

Kinsler, J., & Pavan, R. (2013). "The Specificity of General Human Capital: Evidence from College Major Choice," Working paper, University of Rochester and University of London. Retrieved from http://www.econ.rochester.edu/people/Kinsler/major_HC_specificity.pdf.

Mincer, Jacob (1974) "*Education and Earnings*." New York, NY: Columbia University.

Organisation for Economic Cooperation and Development (2003). "Changing Patterns

of Governance in Higher Education." Education Policy Analysis. Retrieved from OCED website: http://www.oecd.org/education/skills-beyond-school/35747684.pdf.

Organisation for Economic Cooperation and Development (2010). "OECD Factbook 2009." Retrieved from OECD website: http://www.oecd-ilibrary.org/content/book/factbook-2009-en.

Turner, S., & Bowen, W. (1999). "Choice of Major: The Changing (Unchanging) Gender Gap." *Industrial and Labor Relations Review*, 52(2), 289-313.

Wise, D., (1975). "Academic Achievement and Job Performance." *American Economic Review, 65*(3), 350-366.

Zafar, B. (2013). "College Major Choice and the Gender Gap." *Journal of Human Resources, 48*(3), 545-595.

Zafar, B., & Wiswall, M. (2013). "Determinants of College Major Choice: Identification using an Information Experiment." (Staff Report No. 500). Retrieved from Federal Reserve Bank of New York website: http://www.newyorkfed.org/research/staff_reports/sr500.pdf.

CHAPTER 12

Economic Growth and Labor Market Institutions in East Asian Structural Transformation*

by
Seungjoon Oh
(University of Michigan)
*Seung-Gyu Sim***
(University of Tokyo)

Abstract

This paper develops a new growth model by incorporating labor market friction and human capital accumulation into the multi-sector growth framework to analyze the underlying link between economic growth and labor market institutions in a transitional economy. The model, calibrated based on the Japanese and South Korean structural transformation episodes, demonstrates that lifetime employment (and the implied lengthy job tenure) has contributed to endogenous formation of a Ricardian comparative advantage in non-agricultural sector, by enhancing specific human capital accumulation and facilitating investment. It has enabled Japan and South Korea to achieve unprecedentedly rapid economic growth. The counterfactual experiment finds that had the job

* We thank Seung Mo Choi, Shin-ichi Fukuda, Junichi Fujimoto, Hidehiko Ichimura, Soohyun Oh, Dan Sasaki, Michio Suzuki and all seminar participants at various conferences and workshops. Seung-Gyu Sim also appreciates the hospitality of E. Han Kim and the Ross School of Business, University of Michigan, a substantial part of this study having been completed during his stay in Ann Arbor. All errors are our own.

** Corresponding author (Address: University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-0033.Email: sgsim@e.u-tokyo.ac.jp)

durations of a typical worker been 1 year (roughly one tenth of the actual average job duration) for 1960-1990 in the Japanese labor market and 1970-2000 in the Korean labor market, the non-agricultural GDP per capita in 1990 would have accounted for 71 and 76 percent of the actual values, respectively.

1. Introduction

There is extensive and still growing interest in non-stationary multi-sector economic growth, or shortly structural transformation. However, since most of the previous studies on structural transformation presume that it is the evolution of Walrasian equilibria guided by the exogenous shock of total factor productivity (hereafter TFP), they are incapable of answering the question, "What enables some countries to perform rapid structural transformation and others not?" Motivated by Mortensen and Pissarides (1994, 1998), this paper develops a new growth model that incorporates labor market friction and human capital accumulation into the multi-sector growth theory to analyze how different labor market institutions contribute differently to economic growth. In particular, our quantitative assessment based on East Asian structural transformation episodes reveals that their lifetime employment system characterized by lengthy job tenures has enabled postwar Japan and South Korea, by contributing to formation of Ricardian comparative advantage in sectors in which it had not previously been experienced, to achieve unprecedentedly rapid structural transformation.

In their seminal work, Aghion and Howitt (1994) show that fast economic growth may reduce unemployment (capitalization effect) or add fuel to unemployment (creative destruction effect). Mortensen and Pissarides (1998) subsequently document the rapid progress of disembodied technology to decrease, and of embodied technology to increase, the steady state unemployment rate. By extending the aforementioned model, Pissarides and Vallanti (2007) and Miyamoto and Takahashi (2011) also evaluate the effect of technological progress on the labor market. All those pioneering models, focusing only on the one-sided effect of technological progress on steady state unemployment, are somewhat far

from the neoclassical growth models in which the feedback effect of human capital accumulation in the labor market on economic growth is a central issue. Chen, Chen, and Wang (2011) recently constructed a search model with endogenous human capital accumulation and labor market participation to evaluate the effectiveness of human capital policies. But their framework is not widely applicable to various environments of interest in the growth literature given that they solve only for the one-sector[1] steady state model in which all households, regardless of sector and employment status, share the same rate of human capital accumulation.

Following Matsuyama (1992), we construct an endogenous growth model of a small open economy with agricultural and non-agricultural sectors populated by a continuum of entrepreneurs and workers who consume both agricultural and non-agricultural products. Entrepreneurs employ capital and labor input to produce products in their respective sectors. Each sectoral labor market is subject to search and matching friction, as in Mortensen and Pissarides (1994). Employed workers provide labor and earn wages in the respective sectors, according to the Stole and Zwiebel (1996) bargaining rule, and accumulate job-specific human capital through learning-by-doing on the job. Unemployed workers collect unemployment benefits and search for jobs. They switch to the other sector if the opportunity cost of staying dominates the value of unemployment in their own sector. We assume the population and labor-augmenting technology (general human capital) of workers to grow at a constant rate, technology and capital stock endogenously through investment by forward-looking entrepreneurs. Analyzing the transition path from initial states to balanced growth paths under rational expectation enables us to investigate rich counterfactual scenarios. To solve for the entire transition path, we iterate the forward- and backward-shooting algorithms following Lipton, Poterba, Sachs, and Summers (1982) and Ishimaru, Oh, and Sim (2013).

We calibrate the model using the structural transformation episode of Japan and South Korea. In addition to Japan, so called Newly Industrializing Countries (NICs) of East Asia, such as Hong Kong, Singapore, South

1 More precisely, it means "one productive sector." Their model consists of two sectors, a productive sector and non-productive sector.

Korea, and Taiwan, have experienced rapid economic growth by transforming from agricultural to non-agricultural economies through specialization in heavy chemical industries. It is an interesting question why only those NICs and Japan were able to form Ricardian comparative advantages on those heavy chemical industries. At least part of the answer likely lies in interaction between specific human capital and the "lifetime employment system" that resulted from the Confucian tradition which ethically discouraged job turnover by employees and dismissal by employers.[2] Although general human capital contributes growth of the economy, it hardly creates or reinforces structural transformation toward a particular sector. In contrast, specific human capital acquired and utilized in a particular sector or firm can be a key driving force of endogenous formation of Ricardian comparative advantage. More precisely, the interaction between specific human capital and lifetime employment system has little effect on the agricultural economy, which does not rely on significant (physical or human) capital accumulation. Rather it accelerates and intensifies specialization toward the non-agricultural sectors that require skilled workers and physical investment on complicated facilities, such as the electronics, automobiles, and heavy-chemical industries.[3] In light of this, the main goal of our calibration is to quantify the social returns to specific human capital and highlight the channel through which the lifetime employment system enhances economic growth.

The counterfactual experiment finds that had the job durations of a typical worker been 1 year (roughly one tenth of the actual average job

---

2 According to (OECD) Employment Outlook 1993 (http://www.oecd.org/els/oecdemploymentoutlookdownloadableeditions1989-2011.htm), average (median) job tenure for male workers in Japan, the United Kingdom, and the United States was 12.5 (10.1), 9.2 (5.3), and 7.5 (3.5) years, respectively. The figures imply even shorter job durations for the majority of U.K. and U.S. workers. Esteban-Pretel and Fujimoto (2012) report the quarterly job separation rate in Japan at around 0.02. Chang, Nam, and Rhee (2004) report that of South Korea at between 0.02 and 0.03, roughly one quarter of the U.S. separation rate of 0.1 reported in Shimer (2005).

3 Davidson, Martin, and Matusz (1999) argue that the labor market institution can determine comparative advantage as well, showing by means of the two-country, two-sector, two-factor trade model that a country with better matching efficiency has comparative advantage over a sector with a higher rate of separation.

duration) for 1960-1990 in the Japanese labor market and 1970-2000 in the Korean labor market, the non-agricultural employment share in 1990 would have accounted for 80-85 percent of their actual values and the non-agricultural GDP per capita in 1990 for 71 and 76 percent of their actual values, respectively, suggesting sluggish structural transformation. Also, if the Japanese and South Korean labor markets had transplanted from the flexible U.S. labor market[4] the matching technology and high separation rate, non-agricultural GDP per capita would have been lower by 5-8 percent in both countries, whereas agricultural GDP per capita would have not been affected significantly. It suggests that at least during the period of structural transformation from agricultural to non-agricultural, the stable labor markets can perform better than flexible ones.

This paper adds to the extant literature on structural transformation several distinctive features. First, it introduces imperfect mobility of productive resources, that is, labor. Previous papers on structural transformation including Hansen and Prescott (2002), Ngai and Pissarides (2007), Duarte and Restuccia (2010), and Buera and Kaboski (2012) assume perfectly competitive labor markets absent any rigidity in factor mobility. Departing from the "immediate full employment assumption," we borrow the concepts of both search friction from Mortensen and Pissarides (1994, 1998) and an inter-sectoral labor barrier from Hayashi and Prescott (2008).[5] The current paper analyzes the dynamic path of non-stationary economic growth impeded by intra- and inter-sectoral labor rigidity as in Ishimaru, Oh, and Sim (2013). Intra-sectoral rigidity (search friction) is introduced to incorporate the effect of job security and job tenure on human capital accumulation and inter-sectoral rigidity (inter-sectoral barrier) to control the direction and duration of the non-stationary structural transformation.

Second, it incorporates both general and specific human capital. General human capital represents the accumulated knowledge or general equipment of the economy. Although not many workers in 1960 knew how to use automobiles, computers, or even telephones, most workers in

4 Precisely speaking, the matching technology used and calibrated in other studies on the U.S. labor market

5 Hayashi and Prescott (2008) show that the labor barrier in their neoclassical two-sector growth model depressed economic growth in the prewar Japanese economy.

developed countries now take advantage of cars, computers and cell-phones. These are captured by labor augmenting technology (general human capital) in our model. It works as a permanent TFP shock in other papers analyzing a balanced growth path with a permanent TFP shock. In addition to general human capital, workers individually acquire skills through learning-by-doing in many workplaces. The acquired skills through repetition of similar tasks may be specific to the occupation, job, or working environment. We capture it as job specific human capital of individual workers. Our numerical exercises quantify the effect of general human capital on the overall growth of the economy and that of specific human capital on the unbalanced growth of the sectors, structural transformation.

Third, rather than assuming an exogenously embedded TFP shock, the present paper demonstrates the endogenous formation of a Ricardian comparative advantage. Hayashi and Prescott (2008), Esteban-Pretel and Sawada (2009), and Uy, Yi, and Zhang (2013) analyze the structural transformation episode of Japan and South Korea.[6] In their models, however, the main driving force of structural transformation is the exogenously embedded sectoral TFP shock. Unless their sources are clearly identified, there is a risk of overstating the impact of exogenous sectoral TFP shocks on economic growth. In sharp contrast, this paper, by replacing "perfect foresight" with "rational expectation," develops a non-stationary endogenous growth model in which the dynamic path of the transitional economy is endogenously determined by forward-looking decisions of contemporary economic agents. Apparently, invoking rational expectation enables us to conduct rigorous counterfactual experiments.

The remainder of the paper is organized as follows. We develop the model in section 2 and, present numerical analysis using the Japanese and South Korean episodes in section 3. Based on the calibration in section 3, we conduct counterfactual experiments in section 4, and conclude in section 5.

6 The former allows time-varying sectoral TFP shocks, while the latter constant sectoral TFP shocks.

2. The Baseline Model

2.1. Environment

Consider a small open economy populated by continuum of entrepreneurs and workers who consume both agricultural and non-agricultural products. Entrepreneurs in both sectors, denoted by subscript $i \in \{a, m\}$, manage their own firms and take profit flow π_t at every instant $t \in \{0, \infty\}$. Workers are either employed or unemployed, the latter in both sectors receiving the same unemployment insurance, b_t, at every instant, the employed workers in sector i earning wage w_{ijt} at time t, depending on skill $j \in \{h, l\}$. The labor market is assumed to be subject to search and matching friction following the Diamond-Mortensen-Pissarides model. Time is continuous and all economic agents discount the future at rate r.

Workers At every instant, each individual worker having income flow $\overline{w}_t$ chooses the consumption bundle (c_{at}, c_{mt}) to maximize

$$\left(c_{at}^{\frac{\sigma-1}{\sigma}} + c_{mt}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \tag{1}$$

subject to the budget constraint $p_a c_{at} + p_m c_{mt} = \overline{w}_t$, where p_a and p_m represent the world prices of agricultural and non-agricultural products, respectively. It is assumed that workers are allowed neither to save nor to borrow. Since we are looking at a small open economy, we assume that (p_a, p_m) are exogenously given and constant over time.[7] In particular, we take the non-agricultural products as *numéraire* and normalize p_m to be unity. Note that the elasticity of substitution should be strictly positive, $\sigma > 0$.[8] Solving the static utility maximization problem and plugging the individual demand for each product into the direct utility flow yields the

---

7 Matsuyama (1992) shows that agricultural productivity is positively (negatively) correlated with economic growth in a closed (small open) economy.

8 When $\sigma \to 0, 1$, and ∞, the utility function in (1) converges to a Leontief function, Cobb-Douglas function, and von-Neumann function, respectively. Refer to Arrow, Chenery, Minhas, and Solow (1961).

indirect utility flow, as follows.

$$v(p_a, p_m, \overline{w}_t) = \overline{w}_t P^{-1} = v(\overline{w}_t), \quad \text{where} \quad P = (p_a^{1-\sigma} + p_m^{1-\sigma})^{\frac{\sigma}{1-\sigma}} \quad (2)$$

Denote by L_t the total population of workers at time $t \in \{0, \infty\}$. At every instant, ρL_t measure of workers retire and χL_t measure of newly born workers enter the labor market as unemployed, which implies that the total population grows at rate $\chi - \rho (\geq 0)$ at every instant. Hence,

$$L_t = e^{(\chi-\rho)t} L_0 \text{ at each } t \in \{0, \infty\} . \quad (3)$$

The sectoral ratio of newly-born workers is *ex ante* assumed to be the same as that of existing workers at every instant. But the new workers in one sector can immediately switch to the other sector at switching cost $s \sim Logistic(0, \varepsilon_t/\omega)$. For simplicity, the location parameter of the logistic distribution is normalized to be zero. The scale parameter of the logistic distribution is composed of parameter ω and the labor-augmenting technology ε_t (The role of the labor-augmenting technology will be explained later).

Following Mcfadden (1974), Rust (1987), and Kennan and Walker (2011), we obtain the probability that the newly-born unemployed worker in sector *i* switches to sector $-i$ as follows.

$$w_{-it} = \frac{1}{1 + \exp\left[-(U_{-it} - U_{it})\mathrm{w}/\varepsilon_t\right]} 1 - w_{it} , \quad (4)$$

where U_{it} represents the lifetime value of the unemployed worker in sector *i* at time *t*.[9] Unemployed workers retire at rate ρ and receive job offers at rate $f(\theta_{it})$[10]. If employed, the worker enjoys the lifetime value of W_{ilt}. Unemployed workers in sector *i* also get revision shocks at rate ξ and

9 Given random switching cost *s*, the worker in sector *i* switches to sector -*i* if and only if $U_{it} < U_{-it} - s$.

$Pr\{U_{it} < U_{-it} - \mathrm{s}\} = \frac{1}{1 + \exp\left[-(U_{-it} - U_{it})\omega/\varepsilon_t\right]}$.

10 It will be discussed later.

switch to the other sector if the net gains from doing so is positive. The Hamilton-Jacobi-Bellman (HJB hereafter) equation for the worker is given by

$$rU_{it} = v(b_t) - \rho U_{it} + f(\theta_{it})(W_{ilt} - U_{it}) + \xi\mathbb{E}[\max\{U_{-it} - U_{it} - \mathrm{s}, 0\}] + \dot{U}_{it}\,, \tag{5}$$

where

$$\mathbb{E}[\max\{U_{-it} - U_{it} - \mathrm{s}, 0\}] = \omega^{-1}\varepsilon_t \log(1 + \exp[(U_{-it} - U_{it})]\,\omega/\varepsilon_t) =: \Delta_{it} \tag{6}$$

The detailed derivation of (6) is presented as well in Ishimaru, Oh, and Sim (2013). The left-hand side of (5) can be interpreted as the opportunity cost of holding the asset, unemployment at time *t*. The terms on the right-hand side represent the dividend flow from holding the asset U_{it}, the potential loss from retirement, the potential gains from job finding, the potential gains from inter-sectoral migration, and the gains from changes in valuation of the asset, respectively. To ensure the existence of the balanced growth path, we assume, following Mortensen and Pissarides (1998), that $b_t = b\varepsilon_t$. An unskilled worker employed in sector *i* receives wage flow w_{ilt} per instant. The worker is separated from the job and becomes unemployed at rate δ and acquires skills at rate ζ. The HJB equations for the skilled and unskilled worker in sector *i* are, respectively,

$$rW_{iht} = v(w_{iht}) - \rho W_{iht} + \delta(U_{it} - W_{iht}) + \dot{W}_{iht}, \text{and} \tag{7}$$

$$rW_{ilt} = v(w_{ilt}) - \rho W_{ilt} + \delta(U_{it} - W_{ilt}) + \zeta(W_{iht} - W_{ilt})\dot{W}_{ilt}. \tag{8}$$

The left-hand sides of (7) and (8) represent the opportunity cost of holding asset W_{iht} and w_{ilt}, respectively. The right hand side in equation (7) consists of the dividend flow from the asset, the potential loss from retirement, the loss from job separation, and the gains from changes in valuation of the asset, respectively. The right hand side of (8) has one additional term, gains from skill acquisition. It is assumed that "skill" is job-specific so that skilled workers as well as unskilled workers lose their

skill and get U_{it} when they are separated from their job. Alternatively, one can think of a variant with "general skills", which requires to extend the dimension of the model. The alternative may cause quantitative changes in our argument, but the qualitative implication of this paper will be maintained. Furthermore, in Japan, South Korea, and other East Asian countries having the Confucianism tradition, it is not common for an ordinary worker in one firm to switch to another rival firm and utilize the skills that she/he acquired in the former job. Once she/he is separated from a job, the worker is more likely to get the next job in a different occupation through long unemployment periods. Hence, this paper keeps the specific skills.

Entrepreneurs There are n_i -measure of entrepreneurs in sector $i \in \{a, m\}$, who share the same preferences with workers. Entrepreneurs in each sector produce homogeneous products using identical production technology, given by

$$y_{it} = a_{it}^{\beta_{ai}} k_{it}^{\beta_{ki}} (\varepsilon_t \tilde{l}_{it})^{\beta_{li}}, \tag{9}$$

where $(a_{it}, k_{it}, \hat{l}_{it})$ represent the TFP component, capital stock, and labor input at time *t*, respectively. It is assumed that $\beta_{ai} + \beta_{ki} + \beta_{li} = 1$. The labor-augmenting technology ε_t is introduced to capture the growth of the accumulated (general) knowledge of the economy. It can be interpreted as general human capital. Sustained economic growth on the balanced growth path is obtained by assuming ε_t to grow permanently, such that

$$\dot{\varepsilon}_t = \psi \varepsilon_t \text{ for each } \quad t \in \{0, \infty). \tag{10}$$

The total factor productivity component, a_{it}, can be formed gradually through R&D investment and interaction with other productive resources (i.e. human capital). Throughout the paper we call it 'technology'. That many R&D outcomes including new technologies and equipment can be utilized together with a certain level of skills accounts for gradual and persistent economic growth in a transitional economy. We posit

$$\dot{a}_{it} = -\eta_a a_{it} + \lambda_i z_{it}^{\kappa_i} (\varepsilon_t L_{iht})^{1-\kappa_i}, \tag{11}$$

where λ_i is the efficiency parameter of technology investment, κ_i is the elasticity of technology investment, and η_a is the technology deterioration rate. L_{iht}, L_{ilt}, and L_{it} are the number of the skilled employed, unskilled employed, and all workers, respectively, in sector *i* at time *t*. Each entrepreneur in sector *i* makes R&D investment of z_{it} at every *t*. The law of motion in (11) implies that that R&D investment becomes more efficient as the skilled population in sector *i* increases (human capital externality). [11] This, together with deterioration, embodies the gradual formation of the Ricardian comparative advantage and disadvantage. The capital stock, which can be immediately accumulated by purchasing from the international capital market, is depreciated at rate η_k.

$$\dot{k}_{it} = -\eta_k k_{it} + x_{it}, \tag{12}$$

where x_{it} is the capital investment in sector *i* at time *t*. Regarding the labor input, it is assumed that

$$\tilde{l}_{it} = l_{ilt} + \alpha_i l_{iht}, \tag{13}$$

where (l_{ilt}, l_{iht}) represent the masses of unskilled and skilled workers, respectively, employed by the entrepreneur at time *t*. The coefficient α_i reflects the fact that a skilled worker produces $\alpha_i (\geq 1)$ times more than an unskilled worker. Again, "skill" is assumed to be job-specific. Let v_{it} be the number of vacancies waiting for unemployed workers. The entrepreneur finds at rate $q(\theta_{it})$ a worker who starts producing as an unskilled worker. Unskilled workers get learning shocks at rate ζ on the job and become skilled workers. All workers leave the entrepreneur due to separation shock at rate δ and retirement shock at rate ρ. The law of motion of the employed workers is described by

11 This positive externality is motivated by Lucas (1988) and Choi (2011). By calibrating the model and conducting a counterfactual analysis, Choi (2011) argues that human capital externality has a significant effect on economic growth. Romer (1986, 1987) proposed the growth models with physical capital externalities in advance.

$$\dot{l}_{iht} = -(\delta + \rho) l_{iht} + \zeta l_{ilt}, \text{and} \tag{14}$$

$$\dot{l}_{ilt} = -(\delta + \rho + \zeta) l_{ilt} + q(\theta_{it}) v_{it}. \tag{15}$$

The profit flow of the entrepreneur in sector *i* at time *t* is given by

$$\pi_{it} = p_i y_{it} - \sum_{j=h,l} w_{ijt} l_{ijt} - p_k x_{it} - p_z z_{it} - \gamma \varepsilon_t v_{it}, \tag{16}$$

where (p_k, p_z) represent the cost of capital investment and R&D investment, respectively, and $\gamma \varepsilon_t$ represents the cost of creating a vacancy. Following Mortensen and Pissarides (1998), we assume the vacancy creation cost to grow together with ε_t, which is necessary to ensure the existence of the balanced growth path. The entrepreneur in sector *i* having $(\overline{\varepsilon}, \overline{a}_i, \overline{k}_i, \overline{l}_{hi}, \overline{l}_{li})$ at time *t* chooses the schedule of $(x_{i\tau}, v_{i\tau}, z_{i\tau})$ for each $\tau \in \{t, \infty)$to maximize

$$E_{it}\left(\overline{\varepsilon}, \overline{a}_i, \overline{k}_i, \overline{l}_{hi}, \overline{l}_{li}\right) = \max_{z_{i\tau}, x_{i\tau}, v_{i\tau} \geq 0} \int_t^{\infty} e^{-r(\tau - t)} \pi_{i\tau} P^{-1} d\tau \tag{17}$$

subject to (10), (11), (12), (14), (15) and the initial condition (ε_t, a_{it}, k_{it}, l_{iht}, l_{ilt}) =$\overline{\varepsilon}, \overline{a}_i, \overline{k}_i, \overline{l}_{hi}, \overline{l}_{li}$. Lemma 1 solves the optimal control problem by the entrepreneur.

Lemma 1 *The entrepreneur in sector* $i \in \{a, m\}$ *optimally chooses* (x_{it}, v_{it}, z_{it}) *for each* $t \in \{0, \infty)$ *such that*

$$p_z = \kappa_i \lambda_i z_{it}^{\kappa_i - 1} (\varepsilon_t L_{iht})^{1 - \kappa_i} \int_t^{\infty} e^{-(r + \eta_a)(\tau - t)} \frac{\partial \pi_{i\tau}}{\partial a_{i\tau}} d\tau \tag{18}$$

$$p_k = \int_t^{\infty} e^{-(r + \eta_k)(\tau - t)} \frac{\partial \pi_{i\tau}}{\partial k_{i\tau}} d\tau, \qquad \text{and} \tag{19}$$

$$\gamma \varepsilon_t = q(\theta_{it}) \int_t^{\infty} e^{-(r + \delta + \rho + \zeta)(\tau - t)} \left[\frac{\partial \pi_{i\tau}}{\partial l_{il\tau}} + \zeta \frac{\partial \pi_{i\tau}}{\partial l_{iht}} \right] d\tau. \tag{20}$$

In equations (18), (19), and (20), the left-hand side represents the

marginal cost of investment, while the right-hand sides represent the marginal benefit. The proof of Lemma 1 is delayed in Appendix A.

Labor Market The degree of the labor market tightness in sector *i* is defined as the ratio of the measure of total vacancy to that of job seekers in the sector. For each $i \in \{a, m\}$,

$$\theta_{it} \coloneqq \frac{v_{it} n_i}{u_{it}}. \tag{21}$$

Let $m(v_{it} n_i, u_{it})$ be the number of matches successfully formed at time *t* for each $i \in \{a, m\}$. Given constant returns to scale matching technology, we obtain

$$\begin{aligned} f(\theta_{it}) &= \frac{m(v_{it} n_i, u_{it})}{u_{it}} = m(\theta_{it}, 1) = \theta_{it} m\left(1, \theta_{it}^{-1}\right) \\ &= \frac{\theta_{it} m(v_{it} n_i, u_{it})}{v_{it} n_i} = \theta_{it} q(\theta_{it}). \end{aligned} \tag{22}$$

We denote by $(L_{iht}, L_{ilt}, u_{it})$ the measure of skilled employees, unskilled employees, and unemployed workers, respectively, in sector *i* at time *t*. The total population, L_t, at time *t* is given by

$$L_t = \sum_{j=h,l} [L_{ajt} + L_{mjt}] + u_{at} + u_{mt} \tag{23}$$

The dynamic worker flows are summarized as follows.

$$\dot{L}_{iht} = -(\rho + \delta) L_{iht} + \zeta L_{ilt}, \tag{24}$$

$$\dot{L}_{ilt} = -(\rho + \delta + \zeta) L_{ilt} + \theta_{itq}(\theta_{it}) u_{it}, \text{and} \tag{25}$$

$$\begin{aligned} \dot{u}_{it} = &-(\rho + \theta_{it} q(\theta_{it}) + \xi \omega_{-it}) u_{it} + \xi \omega_{it} u_{-it} + \\ &\delta(L_{ilt} + L_{iht}) + \chi \omega_{it} L_t, \end{aligned} \tag{26}$$

where L_t is presented in (3). The last term of the right-hand side of (26) is

derived by the following argument: At time t, there are $\chi(L_{aht} + L_{alt} + u_{at})$ -measure of newly born workers in the agricultural sector and $\chi(L_{mht} + L_{mlt} + u_{mt})$ -measure of newly born workers in the manufacturing sector; because they can immediately switch, the total measure of newly born workers in sector i after the immediate switching is given by $\chi\omega_{it}L_t$.

Wage Bargaining In accordance with Stole and Zwiebel (1996)[12], successfully matched workers and entrepreneurs individually negotiate wages by splitting the marginal surplus. This implies that for each $j \in \{h, l\}$,

$$(1-\phi)\left(W_{ijt} - U_{it}\right) = \phi \frac{\partial E_{it}}{\partial l_{ijt}} \quad \text{at every} \quad t \in \{0, \infty), \tag{27}$$

where ϕ is the worker's bargaining power. In wage bargaining, the firm takes $1-\phi$ portion of the joint (marginal) surplus and gives ϕ portion to the worker in the form of wage payment. Note that (27) is implicitly based on an equilibrium restriction such that in any equilibrium $W_{ijt} - U_{it} \geq 0$ and $\frac{\partial E_{it}}{\partial l_{ijt}} \geq 0$ for each $i \in \{a, m\}$, $j \in \{h, l\}$, and $t \in [0, \infty)$. Since the bargaining rule in (27) is true at every t, continuity and differentiability implies that

$$(1-\phi)\left(\dot{W}_{ijt} - \dot{U}_{it}\right) = \phi \frac{\partial}{\partial t}\left(\frac{\partial E_{it}}{\partial l_{ijt}}\right) \text{ for almost everywhere}$$

$$t \in [0, \infty). \tag{28}$$

Lemma 2 *The implied wage in sector* $i \in \{a, m\}$ *is given by*

$$w_{iht} = A_i^0 \alpha_i p_i \beta_{li} \alpha_{it}^{\beta_{ai}} k_{it}^{\beta_{ki}} \varepsilon_t^{\beta_{li}} (l_{ilt} + \alpha_i l_{iht})^{\beta_{li}-1} + A_{it}^1, \text{and} \tag{29}$$

$$w_{ilt} = A_i^0 p_i \beta_{li} \alpha_{it}^{\beta_{ai}} k_{it}^{\beta_{ki}} \varepsilon_t^{\beta_{li}} (l_{ilt} + \alpha_i l_{iht})^{\beta_{li}-1} + A_{it}^1, \tag{30}$$

12 Helpman and Itskhoki (2010), Helpman, Itskhoki, and Redding (2010), Felbermayr, Prat, and Schmerer (2011), and Ishimaru, Oh, and Sim (2013) adopt the bargaining rule proposed by Stole and Zwiebel (1996) in their studies on international trade.

where

$$A_i^0 = \frac{\phi}{1-\phi+\phi\beta_{li}} \text{ and}$$

$$A_{it}^1 = (1-\phi)b\varepsilon_t + \gamma\varepsilon_t\phi\theta_{it} + (1-\phi)\xi\Delta_{it}P.$$

Note that in both (29) and (30) the first terms are proportional to the marginal product of labor. The second terms reflect the labor market condition. For later use, we remark that A_{it}^1 grows at rate ψ if θ_{it} is constant over time and Δ_{it} grows at rate ψ. The detailed derivation is provided in Appendix A.

Equilibrium We finish this section by defining the equilibrium of our interest. The following definition summarizes the overall shape of our model.

Definition An *equilibrium* consists of bounded time series of choice rules $\{\omega_{it}, x_{it}, u_{it}, z_{it}\}_{i\in\{a,m\}}$, labor market tightness parameters $\{\theta_{it}\}_{i\in\{a,m\}}$, wages $\{w_{it}\}_{i\in\{a,m\}}$, value equations $\{E_{it}, W_{iht}, W_{ilt}, U_{it}\}_{i\in\{a,m\}}$, and laws of motions $\{\varepsilon_t, a_{it}, k_{it}, l_{iht}, l_{ilt}, L_t, L_{iht}, L_{ilt}, u_{it}\}_{i\in\{a,m\}}$ at every $t \in [0,\infty)$ such that:

(*i*) unemployed workers as well as newly born workers optimally choose where they work when they are hit by an exogenous shock,

(*ii*) each entrepreneur in sector *i* optimally chooses $\{z_{it}, x_{it}, v_{it}\}$ at every *t*,

(*iii*) aggregate consistency requires that

- the market tightness $\{\theta_{it}\}_{i\in\{a,m\}}$ should be consistent with its definition,
- wages $\{w_{it}\}_{i\in\{a,m\}}$ should be consistent with the imposed bargaining rule,
- $L_{iht} = n_i l_{iht}$, $L_{ilt} = n_i l_{ilt}$, and $L_t = \sum_i (L_{iht} + L_{ilt} + u_{it})$ for each $i \in \{a, m\}$.

(*iv*) the evolution of the entire system is recursively governed by the law of motions of (3), (5), (7), (8), (10), (11), (12), (14), (15),(17),

(24), (25), and (26), given $\{E_{i0}, W_{ih0}, U_{il0}, U_{i0}\}_{i\in\{a,m\}}$ and $\{\varepsilon_0, a_{i0}, k_{i0}, l_{ih0}, l_{il0}, L_0, L_{ih0}, L_{il0}, u_{i0}\}_{i\in\{a,m\}}$.

2.2. On the Balanced Growth Path

In this subsection, we characterize the balanced growth paths by applying the previous definition with a small modification on the law of motions (v). Stationarity on the balanced growth path requires that

$$\frac{\dot{L}_t}{L_t} = \frac{\dot{L}_{iht}}{L_{iht}} = \frac{\dot{L}_{ilt}}{L_{ilt}} = \frac{\dot{u}_{it}}{u_{it}} = \chi - \rho, \frac{\dot{a}_{it}}{a_{it}} = \frac{\dot{k}_{it}}{k_{it}} = \chi - \rho + \psi, \text{and}$$
$$\frac{\dot{U}_{it}}{U_{it}} = \frac{\dot{W}_{iht}}{W_{iht}} = \frac{\dot{W}_{ilt}}{W_{ilt}} = \frac{\dot{\varepsilon}_t}{\varepsilon_t} = \psi. \tag{31}$$

Let L_{ih}, L_{il} and u_i denote the proportion of skilled, unskilled and unemployed workers, respectively, in sector i to the total population. That is,

$$L_{ih} = \frac{L_{iht}}{L_t}, L_{il} = \frac{L_{ilt}}{L_t}, \text{and } u_i = \frac{u_{it}}{L_t}. \tag{32}$$

By the stationarity condition of the balanced growth path (31), these ratios are constant over time. The stationarity condition dictates that $\Delta_{it}/\varepsilon_t$ is constant over time on the balanced growth path. Also, from (15), (21), and (31), we get

$$v_{it} = \frac{(\chi + \delta + \zeta) l_{ilt}}{q(\theta_{it})} = \frac{(\chi + \delta + \zeta) L_{ilt}}{q(\theta_{it}) n_i} = \frac{\theta_{it} u_{it}}{n_i} \tag{33}$$

on the balanced growth path. Since L_{ilt} and u_{it} grow together at the rate $(\chi - \rho)$, θ_{it} should be constant on the balanced growth path to make the last equality hold over time. We drop the time subscript t from the variables that are constant on the balanced growth path.

Using (31), we rewrite (5) as follows.

$$(r+\rho-\psi)U_{it} = b\varepsilon_t P^{-1} + f(\theta_i)(W_{ilt} - U_{it}) + \xi\Delta_{it}. \tag{34}$$

Combining (20) and (27) yields

$$\gamma\varepsilon_t = Pq(\theta_{it})\frac{\partial E_{it}}{\partial l_{ilt}} = Pq(\theta_{it})\frac{(1-\phi)(W_{ilt}-U_{it})}{\phi}. \tag{35}$$

Subtracting U_{at} from U_{mt}, dividing by ε_t, and combining with (35) yields

$$(U_{mt}-U_{at})/\varepsilon_t = \frac{\gamma\phi(\theta_m-\theta_a)}{(r+\rho-\psi+\xi)P(1-\phi)} \text{ and} \tag{36}$$

$$\omega_a = \frac{1}{1+\exp\left[(U_{mt}-U_{at})\omega/\varepsilon_t\right]} = 1-\omega_m. \tag{37}$$

These imply that (ω_a, ω_m) and $(U_{mt}-U_{at})/\varepsilon_t$ are constant over time on the balanced growth path and uniquely determined by (θ_a, θ_m). From (25), (26), and (31), the triplet of (L_{ih}, L_{il}, u_i) is characterized as follows. Given (θ_a, θ_m), for each $i \in \{a, m\}$,

$$0 = (\chi+\delta)L_{ih} - \zeta L_{il}, \text{and} \tag{38}$$

$$0 = (\chi+\delta+\zeta)L_{il} - \theta_i q(\theta_i)u_i, \text{and} \tag{39}$$

$$0 = (\chi+\theta_i q(\theta_i)+\xi\omega_{-i})u_i - \xi\omega_i u_{-i} - \delta(L_{il}+L_{ih}) - \chi\omega_i, \tag{40}$$

Solving (37), (39) and (40), and multiplying by L_t results in Lemma 3.

Lemma 3 *Given* (θ_a, θ_m), *on the balanced growth path*,

$$L_{iht} = \frac{(\chi+\delta)f(\theta_i)\chi\omega_i(\chi+\xi\omega_i+f(\theta_{-i})\chi/(\chi+\delta)+\xi\omega_{-i})L_t}{\zeta(\chi+\delta+\zeta)\left\{\left(\chi+\xi\omega_{-i}+\frac{f(\theta_i)\chi}{\chi+\delta}\right)\left(\chi+\xi\omega_i+\frac{f(\theta_{-i})\chi}{\chi+\delta}\right)-\xi^2\omega_i\omega_{-i}\right\}}, \tag{41}$$

$$L_{ilt} = \frac{f(\theta_i)\chi\omega_i(\chi+\xi\omega_i+f(\theta_{-i})\chi/(\chi+\delta)+\xi\omega_{-i})L_t}{(\chi+\delta+\zeta)\left\{\left(\chi+\xi\omega_{-i}+\frac{f(\theta_i)\chi}{\chi+\delta}\right)\left(\chi+\xi\omega_i+\frac{f(\theta_{-i})\chi}{\chi+\delta}\right)-\xi^2\omega_i\omega_{-i}\right\}}, \text{and} \tag{42}$$

$$u_{it} = \frac{\chi\omega_i(\chi+\xi\omega_i + f(\theta_{-i})\chi/(\chi+\delta) + \xi\omega_{-i})L_t}{\left\{\left(\chi + \xi\omega_{-i} + \frac{f(\theta_i)\chi}{\chi+\delta}\right)\left(\chi + \xi\omega_i + \frac{f(\theta_{-i})\chi}{\chi+\delta}\right) - \xi^2\omega_i\omega_{-i}\right\}}. \quad (43)$$

Notice that $\frac{\partial y_{it}}{\partial l_{ilt}}$ grows at rate ψ on the balanced growth path due to the growth of the labor-augmenting technology. Lemma 4 summarizes the entrepreneur's behavior on the balanced growth path.

Lemma 4 *Given* (θ_a, θ_m) *on the balanced growth path, the entrepreneur in each sector optimally chooses*

$$z_{it} = \varepsilon_t L_{iht}\left[\frac{\lambda_i\kappa_i\beta_{ai}p_i(1-\beta_{li}A_i^0)}{p_z(r+\eta_a)\beta_{li}a_{it}}\frac{\partial y_{it}}{\partial l_{ilt}}(l_{ilt}+a_i l_{iht})\right]^{\frac{1}{1-\kappa_i}}, \quad (44)$$

$$x_{it} = (\chi-\rho+\psi+\eta_k)\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i(l_{ilt}+a_i l_{iht})}{\beta_{li}p_k(r+\eta_k)}\frac{\partial y_{it}}{\partial l_{ilt}}, \text{and} \quad (45)$$

$$v_i = \frac{\theta_i\chi\omega_i(\chi+\xi\omega_i + f(\theta_{-i})\chi/(\chi+\delta) + \xi\omega_{-i})L_t}{\left(\chi + \xi\omega_{-i} + \frac{f(\theta_i)\chi}{\chi+\delta}\right)\left(\chi + \xi\omega_i + \frac{f(\theta_{-i})\chi}{\chi+\delta}\right) - \xi^2\omega_i\omega_{-i}}, \quad (46)$$

where

$$\frac{\partial y_{it}}{\partial l_{ilt}} = \left[\beta_{li}\left(\frac{\lambda_i L_{iht}}{\chi-\rho+\psi+\eta_a}\right)^{\beta_{ai}(1-\kappa_i)}\right.$$
$$\left(\frac{\lambda_i\kappa_i\beta_{ai}p_i(1-\beta_{li}A_i^0)}{p_z\beta_{li}(r+\eta_a)}(l_{ilt}+\alpha_i l_{iht})\right)^{\beta_{ai}\kappa_i}$$
$$\left.\left(\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i(l_{ilt}+\alpha_i l_{iht})}{\beta_{li}p_k(r+\eta_k)}\right)^{\beta_{ki}}\varepsilon_t^{\beta_{li}}(l_{ilt}+\alpha_i l_{iht})^{\beta_{li}-1}\right]^{\frac{1}{1-\beta_{ki}-\beta_{ai}\kappa_i}}. \quad (47)$$

In addition, these imply that

$$a_{it} = \left[\frac{\lambda_i\varepsilon_t L_{iht}}{\chi-\rho+\psi+\eta_a}\right]^{1-\kappa_i}\left[\frac{\lambda_i\kappa_i\beta_{ai}p_i(1-\beta_{li}A_i^0)}{p_{zi}\beta_{li}(r+\eta_a)}\frac{\partial y_{it}}{\partial l_{ilt}}(l_{ilt}+\alpha_i l_{iht})\right]^{\kappa_i}, \quad (48)$$

$$k_{it} = \frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i(l_{ilt}+\alpha_i l_{iht})}{\beta_{li}p_k(r+\eta_k)}\frac{\partial y_{it}}{\partial l_{ilt}}, \quad (49)$$

$$l_{iht} = \frac{(\chi+\delta)f(\theta_i)\chi\omega_i(\chi+\xi\omega_i + f(\theta_{-i})\chi/(\chi+\delta)+\xi\omega_{-i})L_t}{n_i\zeta(\chi+\delta+\zeta)\left\{\left(\chi+\xi\omega_{-i}+\frac{f(\theta_i)\chi}{\chi+\delta}\right)\left(\chi+\xi\omega_i+\frac{f(\theta_{-i})\chi}{\chi+\delta}\right)-\xi^2\omega_i\omega_{-i}\right\}}, \text{and} \quad (50)$$

$$l_{ilt} = \frac{f(\theta_i)\chi\omega_i(\chi+\xi\omega_i + f(\theta_{-i})\chi/(\chi+\delta)+\xi\omega_{-i})L_t}{n_i(\chi+\delta+\zeta)\left\{\left(\chi+\xi\omega_{-i}+\frac{f(\theta_i)\chi}{\chi+\delta}\right)\left(\chi+\xi\omega_i+\frac{f(\theta_{-i})\chi}{\chi+\delta}\right)-\xi^2\omega_i\omega_{-i}\right\}}. \quad (51)$$

Plugging (48), (49), and (51) into (20) and reordering yields

$$\frac{\gamma}{\beta_{li}q(\theta_{it})}+\frac{A_{it}^1/\varepsilon_t}{(r+\rho+\delta-\psi)\beta_{li}} \quad (52)$$
$$= \left[\frac{1-\alpha_i}{r+\delta+\rho-\psi+\zeta}+\frac{\alpha_i}{r+\delta+\rho-\psi}\right]\left[\left(\frac{\lambda_i L_{iht}}{\chi-\rho+\psi+\eta_a}\right)^{\beta_{ai}(1-\kappa_i)}\right.$$
$$\left(\frac{\lambda_i\kappa_i\beta_{ai}}{p_z(r+\eta_a)}\right)^{\beta_{ai}\kappa_i}\left(\frac{\beta_{ki}}{p_k(r+\eta_k)}\right)^{\beta_{ki}}(1-\beta_{li}A_i^0)p_i$$
$$\left.(l_{ilt}+\alpha_i l_{iht})^{-\beta_{ai}(1-\kappa_i)}\right]^{\frac{1}{1-\beta_{ki}-\beta_{ai}\kappa_i}},$$

for each $i \in \{a, m\}$. As mentioned before, (35) is based on the implicit restriction such that $W_{it} - U_{it} > 0$ and $\frac{\partial E_{it}}{\partial l_{it}} > 0$ for each $i \in \{a, m\}$ and $t \in \{0, \infty)$. If it is violated, the right-hand side of (52) can be negative so that there is no solution. If the implicit restriction is satisfied, we get two equations described in (52) to solve for two unknowns (θ_a, θ_m).

Proposition 1 *There exists a balanced growth path if and only if the system of equations described in (52) has a solution of* (θ_a, θ_m).

Given the complexity of the overall system, it is difficult to analytically determine under what parametric condition the balanced growth path exists and whether it is unique. Instead, we acknowledge that in our numerical experiment with reasonable parameter values, we obtain the same result regardless of the random starting points (we repeat the same experiment with more than 20 different random initial guesses).

2.3. On the Postwar Transition

Here, we characterize the transition path from a particular initial state to the balanced growth path, with the evolution of the economy governed by the system of differential equations. One advantage of our model is that the transition path is fully described by an autonomous control.

The lifetime value of unemployment evolves as follows. For each $i \in \{a, m\}$,

$$\dot{U}_{it} = (r+\rho)U_{it} - \frac{\theta_{it}\gamma\varepsilon_t\phi}{P(1-\phi)} - \xi\Delta_{it} - b\varepsilon_t P^{-1} \text{ with } \lim_{t\to\infty} U_{it} = U_i \quad (53)$$

Given $(\theta_{at}, \theta_{mt})$, equation (53) together with (6) determines $(U_{at}, U_{mt}, \Delta_{at}, \Delta_{mt})$ at every $t \in \{0, \infty)$. Plugging (53) into (4) also yields $(\omega_{at}, \omega_{mt})$. Starting from $(L_{ah0}, L_{al0}, L_{mh0}, L_{ml0}, u_{a0}, u_{m0})$, $(L_{aht}, L_{alt}, L_{mht}, L_{mlt}, u_{at}, u_{mt})$ evolve following (24)-(26) toward

$$\lim_{t\to\infty}(L_{iht}, L_{ilt}, u_{it}) = (L_{ih}, L_{il}, u_i). \quad (54)$$

Note that $l_{iht} = L_{iht}/n_i$, and $l_{ilt} = L_{ilt}/n_i$.

Lemma 5 *Given $(\theta_{at}, \theta_{mt}, L_{aht}, L_{alt}, L_{mht}, L_{mlt})$ at every $t \in \{0, \infty)$, we obtain*

$$k_{it} = \left[\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i a_{it}^{\beta_{ai}}\varepsilon_t^{\beta_{li}}(l_{ilt}+a_i l_{iht})^{\beta_{li}}}{(r+\eta_k)p_k}\right]^{\frac{1}{1-\beta_{ki}}} \quad (55)$$

$$a_{it} = \int_0^t e^{-\eta_a(t-\tau)}\lambda_i z_{i\tau}^{\kappa_i}(\varepsilon_\tau L_{ih\tau})^{1-\kappa_i}\, d\tau + e^{-\eta_a t}a_{i0}, \quad (56)$$

l_{iht}=$L_{iht}/n_{i,}$ and $l_{ilt} = L_{ilt}/n_i$, (57)

Where

$$z_{it} = \varepsilon_t L_{iht}\left[\frac{\lambda_i\kappa_i}{p_z}\int_t^\infty e^{-(r+\eta_a)(\tau-t)} \left(1-\beta_{li}A_i^0\right) d\tau\ \beta_{ai}p_i a_{i\tau}^{\beta_{ai-1}} a_{i\tau}^{\beta_{ki}} k_{i\tau}^{\beta_{ki}}\varepsilon_\tau^{\beta_{li}}(l_{il\tau}+a_i l_{ih\tau})^{\beta_{li}}\right]^{\frac{1}{1-\kappa_i}}$$

Lemma 5 summarizes the dynamic path followed by entrepreneurs. Given $(\theta_{at}, \theta_{mt})$ at every $t \in \{0, \infty)$, equations (24)-(26) jointly determine $(L_{aht}, L_{alt}, L_{mht}, L_{mlt}, u_{at}, u_{mt})$ and, together with Lemma 5, the unique path of the economy. Combining (20) and (27) yields

$$\gamma \varepsilon_t P^{-1} = q(\theta_{it}) \frac{(1-\phi)(W_{ilt} - U_{it})}{\phi}, \tag{58}$$

which restores $(\theta_{at}, \theta_{mt})$ at every $t \in \{0, \infty)$. Because $(W_{ilt}, U_{it}) \to (W_{il}, U_i)$, we can get the convergence point (θ_a, θ_m). As in the previous subsection, (58) raises the equilibrium restriction that $W_{ilt} - U_{it} > 0$ for all *i* and *t*. Equations (24)-(26) and (53)-(58) provide the full description of the transition path of the model.

Again, the complexity of the system prevents us from providing an analytical proof on the existence and uniqueness of the solution. Instead, we acknowledge that we obtain the same result when we repeat the experiment with 20 different sets of "initial guesses."

3. Numerical Analysis

This section presents a quantitative assessment of the underlying link between labor market institutions and economic growth, with a focus on two East Asian countries, Japan and South Korea. The Japanese episode is a comprehensive example of economic growth through structural transformation up to the starting point of 'lost decades'. Lack of data for the 1950s and 1960s precludes analysis of the initial periods of structural transformation independently of the 'postwar effect' of World War II (1939-45) and the Korean War (1950-53). Initially less confounded by the postwar effect, the structural transformation of South Korea accelerated in the early 1970s and continued until the Korean foreign currency crisis in 1997. This complementary analysis of the Japanese and South Korean episodes supports concrete conclusions about labor market institutions and economic growth in transitional economies.

In subsection 3.1, we calibrate the model by combining three Japanese data sources: the World Bank database; the Japanese national account; and

the dataset employed by Hayashi and Prescott (2008). Lacking earlier data, we focus on the period beginning in 1960 and, to exclude the lost decades of Japan in the 1990s, ending in 1990. In subsection 3.2, we calibrate the model using three South Korean sources: the World Bank database; the Korean national account; and the estimation results presented by Chang, Nam, and Rhee (2004). We focus on the period starting in 1970, the implied starting point of structural transformation in South Korea, and, to avoid its foreign currency crisis, ending in 1995. [Table 1] and [Table 2] summarize our choices of parameter values associated with Japan's structural transformation, [Table 3] and [Table 4] our choices of parameter values associated with South Korea's structural transformation.

3.1. Japanese Structural Transformation

Exogenously Fed Parameters We set $r = 0.0095$ during the sample period to fix the annual interest rate at 3.8%, which is consistent with the estimate of the annual discount rate in Esteban-Pretel and Sawada (2009). But, the discount rate does not have a pronounced effect on the macro variables in our model. We normalize the price of non-agricultural products to be 1, and fix the relative price of agricultural products at 1.2, which is consistent with the average of the Engel coefficient from 1960-1990. The Engel coefficient, the ratio of food expenditures to total expenditures by households, is obtained by

$$\frac{p_a^{1-\sigma}}{p_a^{1-\sigma} + p_m^{1-\sigma}} \approx 0.37.$$

Japan's Engel coefficient was around 0.49 in 1960 and declined continuously to around 0.25 in 1990.[13] Because it assumes homothetic preferences, our model predicts a constant Engel coefficient, unless it additionally assumes the subsistence level, as in Matsuyama (1992). For the sake of simplicity, we alternatively take a simple average of the Engel coefficient during the

13 The Family Income and Expenditure Survey of the Ministry of Internal Affairs and Communications Statistics Bureau of Japan provides total consumption expenditure and food expenditure.

target period.

The retirement rate is set to be 0.011, which results in roughly 90 percent of workers, who enter the labor market at age 25, retiring before age 70. The implied average 'market duration', the average elapsed time between labor market entry and exit, is thus less than than 25 years. Not being a lifecycle model with aging, our model considers a worker who moves into the non-labor force to be a retiree and a worker who moves from the non-labor force into the labor force to be a newly born worker. Average market duration is thus much shorter than average lifespan. Japan's population grew from 58.63 million in 1965 to 81.0 million in 1990.[14] The implied quarterly growth rate is $\ln(81.0/58.63)/\,100 \approx 0.003$, which determines the birth rate $\chi = \rho + 0.003 = 0.014$ during the sample period. The separation rate δ is set to 0.02 to obtain the average job duration of 12.5 years among non-retirees in the OECD data set. This choice is consistent with Esteban-Pretel and Fujimoto (2012). We posit

$$q(\theta_i) = 0.6 \times \theta_i^{-0.6} \text{ for each } i \in \{a, m\}, \tag{59}$$

borrowing from Kano and Ohta (2002), who estimate the matching function of the Japanese labor market consistent with the plausible range of empirical elasticity of 0.5 to 0.7 proposed by Petrongolo and Pissarides (2001). We equalize the bargaining parameter to the elasticity of the matching technology by invoking Hosios's condition, that is, $\phi = 0.6$. These choices on bargaining parameter and matching function parameters are consistent with Esteban-Pretel and Fujimoto (2012). The growth rate of labor augmenting technology ψ is set at 0.007. The latter choice, which implies a balanced growth path of 0.028 annually, is based on the average annual growth rate (a weighted average of country growth rates) of GDP in all OECD countries for the period 1980-1990.[15]

The capital depreciation rate η_k, calculated as the ratio of depreciation

---

14 These values are obtained from the Japanese Population Census, which can be downloaded from the Statistics Bureau of Japan.

15 Their lost decades starting from 1990's preclude calibration based on the balanced growth path. Instead, we use the average growth rate of other OECD countries for the period 1980-1990. Table V.1. Growth Performance in OECD countries is from OECD Outlook 2000 V. Recent Growth Trends in OECD Countries.

to real capital stock each year, results in $\eta_k = 0.033$, following Jorgenson (1996). [16] Calculating the technology depreciation rate η_a based on the depreciation rate of service industry equipment results in $\eta_a = 0.041$, again following Jorgenson (1996). Due to the lack of sectoral data, we employ the aggregate economy level of the depreciation rate. The elasticity of substitution σ is set to 3.8, following Bernard, Eaton, Jensen, and Kortum (2003), Felbermayr, Prat, and Schmerer (2011), and Ishimaru, Oh, and Sim (2013). Values exogenously assigned are summarized in [Table 1].

Calibrated Parameters For the remaining parameters, we choose a vector that matches the aggregate of the time series data for sectoral GDP growth per capita, sectoral labor share, sectoral capital share, sectoral wage growth, the net EXP/GDP, replacement ratio and average unemployment rate.

In [Figure 1], the dots represent actual data values, the smooth lines the time series predicted by the model. Overall, [Figure 1] suggests that our calibration strategy captures fairly well the trend in transition. Panels (a) and (b) present the time-series data and the model's prediction of sectoral GDP growth per capita from 1970 to 1990 (lacking data for earlier periods). Panels (c) and (d) report sectoral labor shares, panels (e) and (f) sectoral capital shares, from 1960-1990. Both labor and capital shares declined sharply in the agricultural, and rose rapidly in the non-agricultural, sector. In Panels (g)-(h), we exploit the evolution of sectoral wages from 1960-1990 to calibrate the labor productivity parameters $(\alpha_a, \alpha_m, \zeta)$. Based on historical wage data reported by Japan's Ministry of Health, Labor and Welfare,[17] from 1960-1990 workers' real wages grew by 29 and 152 percent in the agricultural and non-agricultural sectors, respectively. Panel (i) reports the unemployment rate in the actual data and the model. Without business cycle fluctuation, we set the target of the average unemployment rate at 2.0 percent. Panel (j) shows the Net Export/GDP ratio predicted by the model to be reconciled with the average Net Export/GDP ratio of 10.5 percent and its linear trend from 1960 to 1990.

16 Calculating the average annual depreciation rate of agricultural (0.0971), construction (0.1722), mining and oil field (0.1650), metalworking (0.1225), and special industry machinery (0.1031) results in 0.132 annually, or 0.033 quarterly.

17 URL: www.mhlw.go.jp.

Panel (k) presents the population growth from the data and the model. Panel (l) reflects adjustment of the model's parameters to keep unemployment benefits at around 40 percent of the average wage over time. These choices are summarized in [Table 2].

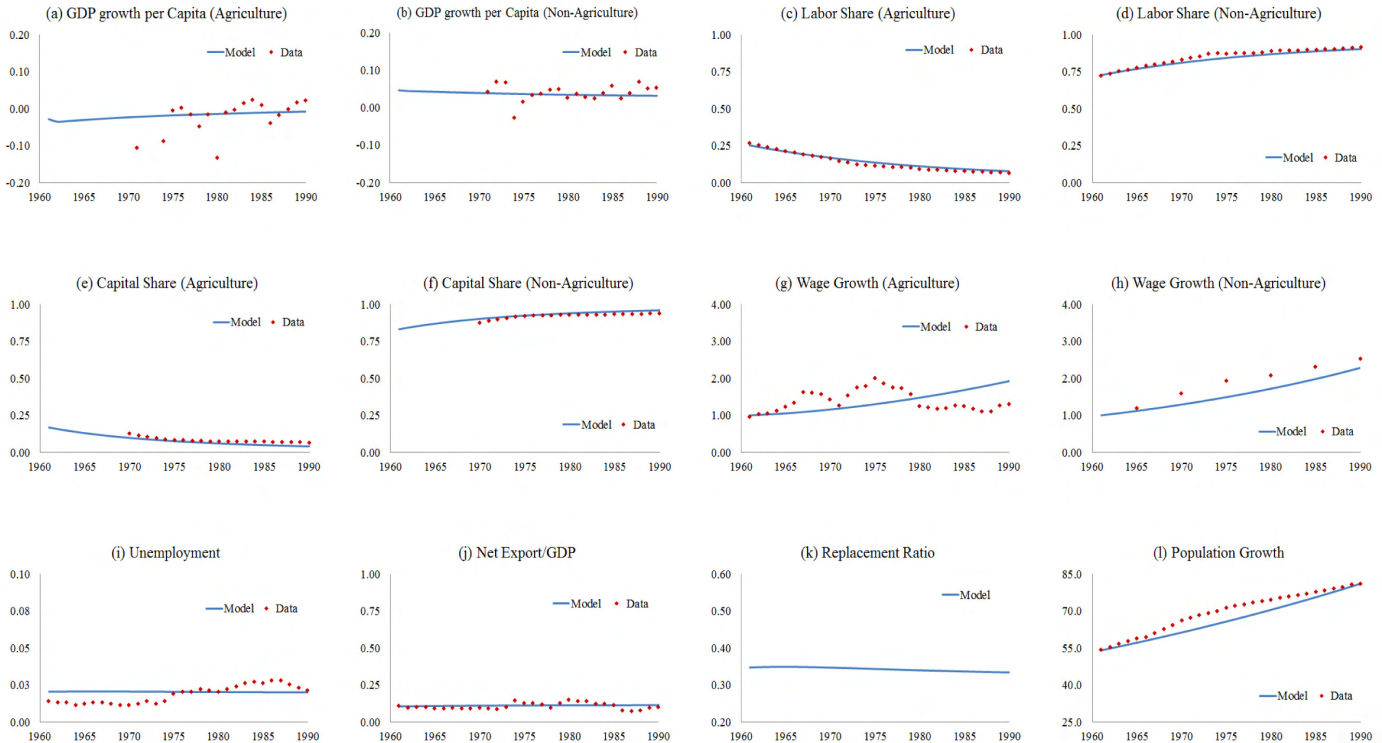
Table 1 Parameter Values: Exogenously Assigned (Japan)

Parameters	Description (Source/ Target)
$r = 0.0095$	discount rate (Esteban-Pretel and Sawada (2009))
$p_{a/}p_m = 1.2$	price of agriculture good (Engel coefficient)
$\rho = 0.011$	retirement rate (retirement age)
$\chi = 0.014$	birth rate (population growth rate)
$\delta = 0.02$	separation rate (Esteban-Pretel and Fujimoto (2012))
$q(\theta) = 0.6\theta^{-0.6}$	matching technology (Kano and Ohta (2002))
$\phi = 0.6$	bargaining parameter (Hosios (1990))
$\psi = 0.007$	growth rate of the labor augmenting technology (average growth rate of OECD countries)
$\eta_a = 0.041$	technology depreciation rate (Jorgenson (1996))
$\eta_k = 0.033$	capital depreciation rate (Jorgenson (1996))
$\sigma = 3.8$	elasticity of substitution (Ishimaru, Oh, and Sim (2013))

Table 2 Parameter Values: Endogenously Targeted (Japan)

Parameters	Description (Source/Target)	
$\omega = 0.03$	sensitivity of inter-sectoral migration	
$\xi = 0.007$	arrival rate of revision shock	
$\alpha_a = 1.05$	agr labor Productivity	
$\alpha_m = 1.5$	non-Agr labor Productivity	
$\zeta = 0.6$	human capital accumulation	(the time series of sectoral GDP growth, sectoral labor share, sectoral capital share, sectoral wage growth, net EXP/GDP ratio, replacement ratio, and unemployment rate)
$\gamma = 0.12$	cost of vacancy	
$p_k = 0.3$	capital cost	
$p_z = 0.4$	technology investment cost	
$\beta_{ka} = 0.193$	agr capital share in production	
$\beta_{km} = 0.23$	non-Agr capital share in production	
$\beta_{la} = 0.58$	agr labor share in production	
$\beta_{lm} = 0.5$	non-Agr labor share in production	
$\lambda_a = 0.3$	efficiency of technology investment	
$\lambda_m = 0.5$	efficiency of technology investment	
$k_a = 0.1$	elasticity of technology investment	
$k_m = 0.4$	elasticity of technology investment	
$b = 1.3$	unemployment benefit	

Figure 1 Calibration Results (Japan)



3.2. South Korean Structural Transformation

Exogenously Fed Parameters We set the real interest rate at 0.01 to match South Korea's annual average real interest rate of around 4.0 percent during 1970-1995 period.[18] As in the Japanese episode, we normalize the price of non-agricultural products to be 1 and fix the relative price of agricultural products at 1.2, consistent with the historical average Engel coefficient of 0.37 for South Korea for the 1970-1995 period. The share of food expenditure to total expenditure declined from 0.44 in 1970 to 0.26 in 1995. The retirement rate is set at 0.011 for consistency, and South Korea's quarterly growth rate was also 0.003, which implies the birth rate $\chi = \rho + 0.003 = 0.014$ during the sample period. According to World Bank data, South Korea's population grew from 31.92 million in 1970 to 45.09 million in 1995, which implies a quarterly growth rate of 0.003.

The separation rate δ is set to 0.027 based on the estimation results in Chang, Nam, and Rhee (2004), who show that the monthly separation rate of the aggregate economy declined from 0.012 in 1981 to 0.006 in 1994. Note that taking the average of these two values results in a quarterly separation rate of 0.027 and corresponding average job duration of 9.25 years. We also borrow from Chang, Nam, and Rhee (2004) the estimates of the matching technology in the South Korean labor market.

$$q(\theta_i) = 0.45 \times \theta_i^{-0.5} \text{ for each } i \in \{a, m\} \tag{60}$$

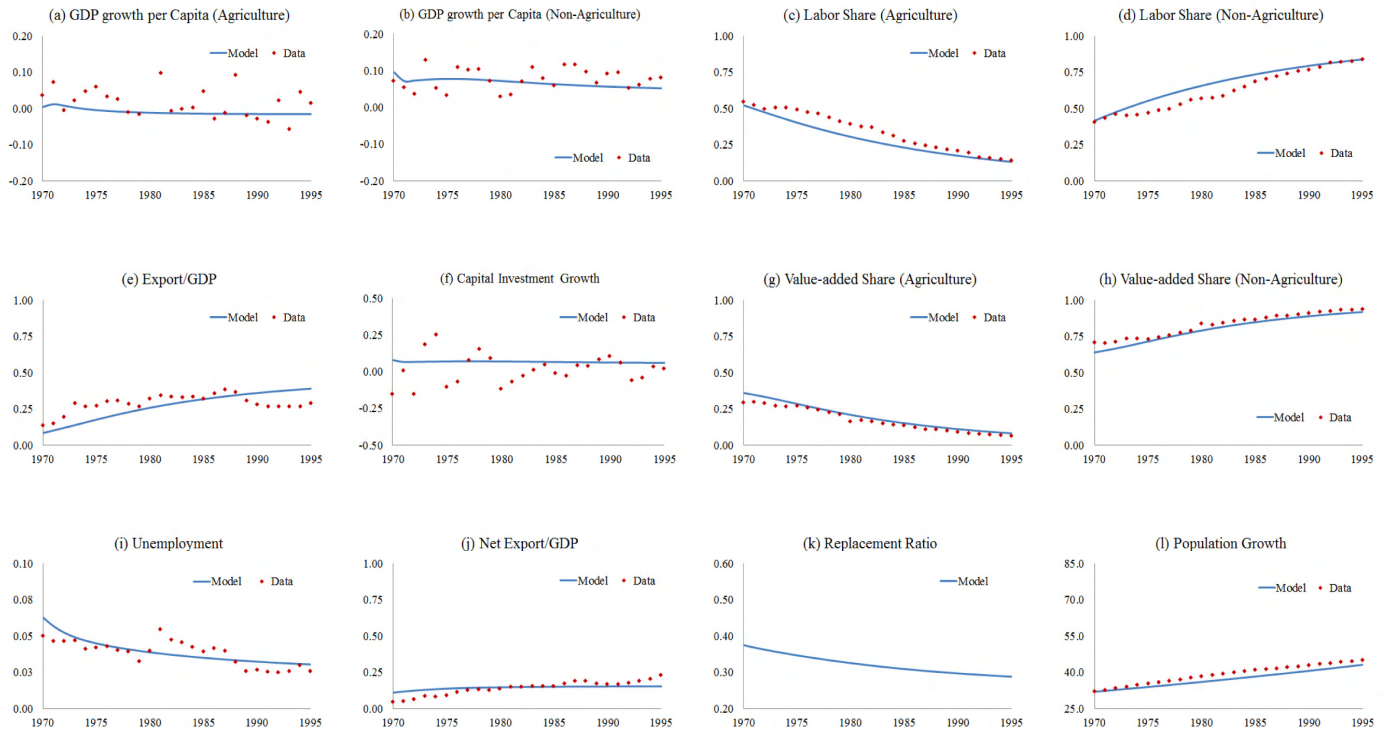
Invoking Hosios's condition, we set $\phi = 0.5$. We also set the growth rate of the labor augmenting technology ψ at 0.01, which is equivalent to an annual GDP growth rate of 4.0 percent on the balanced growth path based on South Korea's average annual GDP per capita growth rate from 2000 to 2010.[19] We use the same values as in the Japanese episode for the remaining parameters $((\eta_a, \eta_k, \sigma)$.

---

18 Chang, Nam, and Rhee (2004) use the annual real interest rate of 4.17 percent during the 1975-1994 period.

19 Because the Japanese economy is suffering still from the ongoing 'lost decades', whereas the South Korean economy was considered recovered from the foreign currency crisis in the early 2000s, we adopt different calibration strategies for their potential growth rates on the balanced growth path.

Figure 2 Calibration Results (South Korea)



Calibrated Parameters We choose a vector of the remaining parameters to match the aggregate of the time series data for sectoral GDP growth per capita, sectoral labor share, the EXP/GDP ratio, capital investment growth, sectoral value-added share, and the net EXP/GDP and replacement ratios and average unemployment rate. Given the similarity between the Japanese and Korean economies and lack of sectoral capital share data, we keep the same parameter values for $(p_k, p_z, \beta_{ka}, \beta_{km}, \beta_{la}, \beta_{lm})$ as in the Japanese episode. Most of the panels in [Figure 2] report the Korean equivalents of the Japanese data. The exceptions are sectoral capital share and wage growth. The latter time series for the sample period being unavailable in the South Korean dataset, we substitute the time-series data for the EXP/ GDP ratio in Panel (e), the aggregate capital investment growth rate in Panel (f), and the sectoral value-added shares in Panels (e) and (f). The actual data show the labor and value-added shares to have declined substantially in the agricultural, and risen sharply in the non-agricultural, sector. Overall, [Figure 2] suggests that our calibration strategy captures fairly well the behavior of the target time series data for South Korea for 1970-1995 period.

Table 3 Parameter Values: Exogenously Assigned (South Korea)

Parameters	Description (Source/ Target)
$r = 0.01$	discount rate (Chang, Nam, and Rhee (2004))
$p_a/p_m = 1.2$	price of agriculture good (Engel coefficient)
$\rho = 0.011$	retirement rate (retirement age)
$\chi = 0.014$	birth rate (population growth rate)
$\delta = 0.027$	separation rate (Chang, Nam, and Rhee (2004))
$q(\theta) = 0.45\theta^{-0.5}$	matching technology (Chang, Nam, and Rhee (2004))
$\phi = 0.5$	bargaining parameter (Hosios (1990))
$\psi = 0.01$	growth rate of the labor augmenting technology
	(the average growth rate of South Korea from 2000-2010)
$\eta_a = 0.041$	technology depreciation rate (Jorgenson (1996))
$\eta_k = 0.033$	capital depreciation rate (Jorgenson (1996))
$\sigma = 3.8$	elasticity of substitution (Ishimaru, Oh, and Sim (2013))

Table 4 | Parameter Values: Endogenously Targeted (South Korea)

Parameters	Description (Source/Target)	
$\omega = 0.035$	sensitivity of inter-sectoral migration	
$\xi = 0.006$	arrival rate of revision shock	
$\alpha_a = 1.1$	Agr labor Productivity	
$\alpha_m = 1.55$	Non-Agr labor Productivity	
$\zeta = 0.7$	human capital accumulation	(the time series of
$\gamma = 0.213$	cost of vacancy	sectoral GDP growth,
$p_k = 0.3$	capital cost	sectoral labor share,
$p_z = 0.4$	technology investment cost	sectoral value-added share,
$\beta_{ka} = 0.193$	Agr capital share in production	aggregate capital growth,
$\beta_{km} = 0.23$	Non-Agr capital share in production	EXP/GDP ratio,
$\beta_{la} = 0.58$	Agr labor share in production	net EXP/GDP ratio,
$\beta_{lm} = 0.5$	Non-Agr labor share in production	replacement ratio, and
$\lambda_a = 0.2$	efficiency of technology investment	unemployment rate)
$\lambda_m = 0.5$	efficiency of technology investment	
$\kappa_a = 0.1$	elasticity of tech. investment	
$\kappa_m = 0.45$	elasticity of tech. investment	
$b = 1.15$	unemployment benefit	

4. Counterfactual Experiments

Using the calibrated model, we conduct counterfactual experiments by substituting different arrival rates of separation shock and matching technologies. The baseline simulations, represented by the solid lines, are the results with the calibrated parameters. In the first counterfactual experiment, represented by the dotted lines (Counterfactual Experiment 1, or CE-1), setting the separation rate at 0.239 and retirement rate at 0.011 results in a one-year job tenure. The second experiment, represented by the dashed lines (Counterfactual Experiment 2, or CE-2), posits the counterfactual question: "If the Japanese and South Korean labor markets had transplanted from the flexible U.S. labor market the efficient matching technology and high separation rate, what would have happened in 1990?"

[Figure 3] shows how varying job duration affects human capital accumulation and labor productivity, as calculated in each counterfactual

experiment. Long job duration enables each entrepreneur to maintain high employment at a lower cost, and improves the average productivity of employed workers through learning-by-doing on the job. Panels (a)-(d) show the fraction of skilled workers to trend downward in the agricultural and upward in the non-agricultural sectors, and to be lower in counterfactual experiments involving a high separation rate. Panels (e)-(h) show the fraction of unskilled workers to exhibit moderate changes over time, Panels (i)-(l) labor productivity to be substantially higher in the baseline simulation.

[Figure 4] shows the transitional paths of technology stock, capital stock, and total production, as determined in each simulation. Panels (a)-(h) show agricultural investments in technology and capital to be slightly lower, and non-agricultural investments to be substantially lower, in the counterfactual experiments than in the baseline simulation. That technology and capital accumulation are more pronounced in the non-agricultural than in the agricultural sector is important. Taken together with the sectoral human capital accumulation patterns depicted in [Figure 3], [Figure 4] shows long job tenure and enhanced labor productivity to stimulate technology and capital investment more greatly in the non-agricultural sector. Consistent with these results, Panels (g)-(h) show aggregate production to be higher in the baseline simulation than in the counterfactual experiments. GDP growth in the non-agricultural sector, in which human capital accumulation significantly improves labor productivity and encourages investment, is especially accelerated by long job duration.

[Table 5] summarizes the main results of our numerical experiments in terms of key ratios of counterfactual simulation results to baseline simulation results in 1990. Because structural transformation was still ongoing in South Korea in 1990, these ratios are relatively underestimated in the South Korean experiments. Panel (a) presents the results from the first counterfactual experiment (CE-1), in which we set the separation rate at 0.239 and retirement rate at 0.011 to make job tenure one year. This result shows that if the job duration of Japanese and South Korean workers had been one year, the non-agricultural employment share in 1990 would have accounted for about 80 percent, and non-agricultural GDP per capita in 1990 for about 70 percent, of their actual values. This suggests that long job duration is a crucial determinant of structural transformation that

enables the economy to achieve rapid growth by driving resource reallocation from the agricultural to the non-agricultural sector.

Panel (b) depicts the second counterfactual simulation model (CE-2), which shows that, whereas agricultural GDP per capita would have not been affected, non-agricultural GDP per capita would have been lower by 5-8 percent in both countries. At the level of the aggregate economy, the difference between the baseline simulation and CE-2 implies that lower job duration would have reduced total GDP per capita by 7 and 5 percent in Japan and South Korea, respectively. The counterfactual analysis of technology stock, capital stock, and human capital accumulation provides an explanation for the gap in GDP per capita across different labor market institutions. We show technology and capital stocks in the non-agricultural sector to be significantly lower in CE-2 than in the baseline simulations. In CE-2, the sectoral accumulations of technology and capital in the non-agricultural sector are 90-93 percent and 92-95 percent of the baseline simulation result in Japan and South Korea, respectively, and skilled labor in the non-agricultural sector declines by 11 percent in Japan and 9 percent in Korea, whereas unskilled labor in both sectors increases significantly in both countries. This indicates that high separation rates drive lower human capital accumulation.

Panel (c) illustrates the effect of only making job duration shorter, as in the US labor markets, without transplanting the matching technology. These counterfactual simulation results show GDP per capita to be lower than in the baseline simulation or CE-2, which indicates that a high exogenous separation rate would have had a greater impact, some part of which, however, would have been absorbed by the efficient matching technology in CE-2.

Figure 3 Counterfactual Results: Labor Market Institutions and Labor Productivity

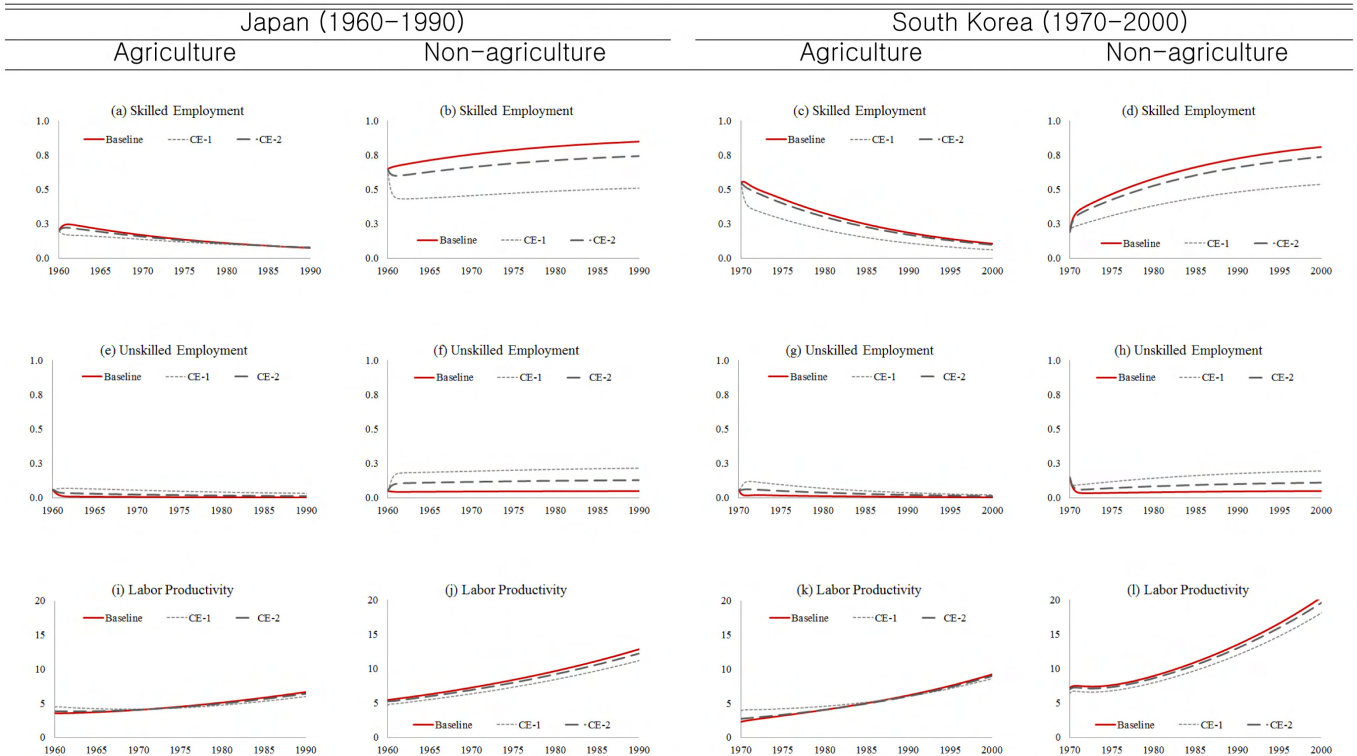


Figure 4 Counterfactual Results: Labor Market Institutions and Economic Growth

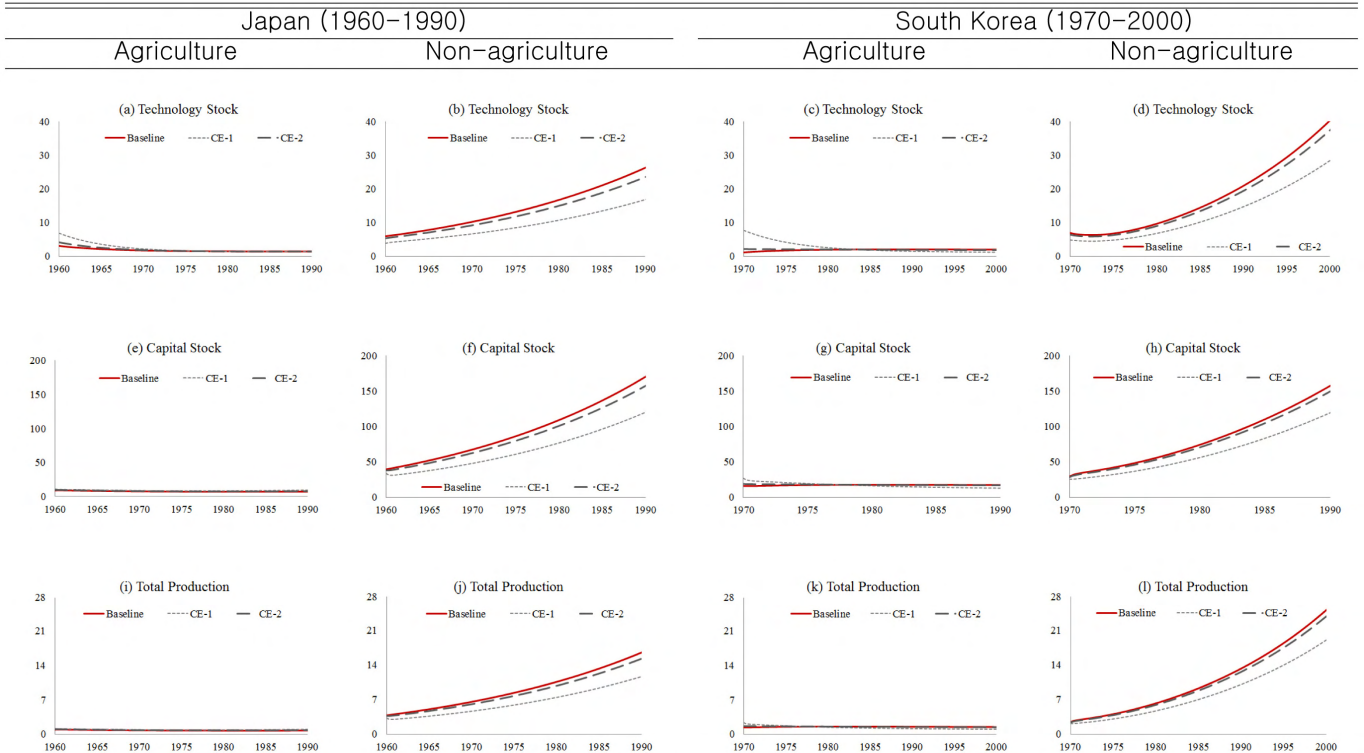


Table 5 Counterfactual Experiments

	Japan (1990)			South Korea (1990)		
	Agriculture	Non-Agriculture	Aggregate	Agriculture	Non-Agriculture	Aggregate
Panel A. Experiment with δ=0.239 to fix job tenure at 1 year						
GDP per capita	1.29	0.71	0.74	0.75	0.76	0.76
Technology Stock	1.02	0.64	0.66	0.74	0.70	0.71
Capital Stock	1.29	0.71	0.73	0.75	0.70	0.76
Employment	1.43	0.81	0.86	0.76	0.85	0.84
Skilled Employment	1.06	0.60	0.64	0.59	0.66	0.65
Unskilled Employment	10.34	4.39	4.75	5.23	3.81	4.00
Panel B. Experiment with δ=0.089 an $dq(\theta) = 1.35\theta^{-0.72}$						
GDP per capita	1.09	0.92	0.93	0.98	0.95	0.95
Technology Stock	1.01	0.90	0.90	0.94	0.93	0.93
Capital Stock	1.09	0.92	0.93	0.98	0.95	0.95
Employment	1.13	0.97	0.98	1.00	0.99	0.99
Skilled Employment	1.02	0.88	0.89	0.92	0.91	0.91
Unskilled Employment	3.84	2.61	2.69	3.04	2.17	2.29
Panel C. Experiment with δ=0.089						
GDP per capita	1.08	0.90	0.91	0.88	0.93	0.92
Technology Stock	0.99	0.87	0.88	0.86	0.91	0.90
Capital Stock	1.08	0.90	0.91	0.88	0.93	0.92
Employment	1.12	0.94	0.96	0.89	0.97	0.95
Skilled Employment	1.01	0.85	0.86	0.82	0.89	0.88
Unskiled Employment	3.80	2.54	2.62	2.70	2.12	2.20

5. Conclusion

We incorporate labor market friction and learning-by-doing into a transitional two sector growth framework to create a new endogenous growth model for exploring the underlying link between labor market institutions and economic growth. As workers stay in their jobs longer, they become more productive through learning-by-doing, which concomitantly lowers the labor cost to entrepreneurs. Enhanced labor productivity stimulates physical capital investment by forward-looking entrepreneurs. When the economy specializes in a sector where learning-by-doing does not improve labor productivity significantly, such as the agricultural sector, economic growth is more moderate. When the economy shifts towards the industrialized sector, on the other hand, in which the effect of learning-by-doing is significant, a stable labor market with a long job duration stimulates and accelerates economic growth by encouraging investment as well as improving labor productivity.

In the numerical analysis, we apply our model to the East Asian episode of structural transformation. Those countries' lifetime employment systems, borne of the Confucian tradition that ethically discourages job turnover by employees and imposes on employers the duty of continuing employment, contributes to the formation of Ricardian comparative advantage in the non-agricultural industrialized sector. The counterfactual experiment finds that had the average job duration of Japanese and South Korean workers been one year, the non-agricultural employment share of each country would have accounted for about 80 percent, and non-agricultural GDP per capita for about 70 percent, of their actual values in 1990, which suggests sluggish structural transformation. Had the Japanese and Korean labor markets transplanted and maintained the efficient matching technology and high separation rate of the flexible U.S. labor market, non-agricultural GDP per capita in 1990 would have declined by 8 percent in Japan and 5 percent in South Korea, and aggregate GDP per capita by 7 percent and 5 percent, respectively. Lengthy job tenure was apparently one of the key ingredients in the structural transformation towards the non-agricultural sector.

Overall, this paper studies the underlying link between labor market institutions and the economic growth in transitional economies by

highlighting matches between particular labor market institutions and sectoral characteristics. We hope that it delivers useful implications for many developing countries that are currently experiencing or expecting structural transformation towards the industrial sector. One may ask if our model can account for the recent growth slowdown or stagnation in NICs and Japan. This paper focuses only on structural transformation from the agricultural to non-agricultural (especially manufacturing) sectors, with an emphasis on the 'capitalization effect'. In contrast, the emergence after the 1990s of new service sectors, such as information technology and the health and financial industries, may stimulate 'creative destruction'. Potential requirements for different processes of human capital accumulation and/or different types of human capital, being beyond the scope of this paper, are left to future research.

References

Aghion, P., and P. Howitt (1994): "Growth and Unemployment," *Review of Economic Studies*, 61(3), pp. 477–494.

Arrow, K. J., H. B. Chenery, B. S. Minhas, and R. M. Solow (1961): "Capital Labor Substitution and Economic Efficiency," *The Review of Economics and Statistics*, 43(3), pp. 225–250.

Bernard, A. B., J. Eaton, J. B. Jensen, and S. Kortum (2003): "Plants and Productivity in International Trade," *American Economic Review*, 93(4), 1268–1290.

Buera, F. J., and J. P. Kaboski (2012): "Scale and the origins of structural change," *Journal of Economic Theory*, 147(2), 684–712.

Chang, Y., J. Nam, and C. Rhee (2004): "Trends in unemployment rates in Korea: A search-matching model interpretation," *Journal of the Japanese and International Economies*, 18(2), 241–263.

Chen, B.-L., H.-J. Chen, and P. Wang (2011): "Labor-Market Frictions, Human Capital Accumulation, and Long-Run Growth: Positive Analysis and Policy Evaluation," *International Economic Review*, 52(1), 131–160.

Choi, S. M. (2011): "How Large Are Learning Externalities?," *International Economic Review*, 52(4), 1077–1103.

Davidson, C., L. Martin, and S. Matusz (1999): "Trade and search generated unemployment," *Journal of International Economics*, 48(2), 271–299.

Duarte, M., and D. Restuccia (2010): "The Role of the Structural Transformation in Aggregate Productivity," *The Quarterly Journal of Economics*, 125(1), 129–173.

Esteban-Pretel, J., and J. Fujimoto (2012): "Life-cycle search, match quality and Japans labor market," *Journal of the Japanese and International Economies*, 26(3), 326–350.

Esteban-Pretel, J., and Y. Sawada (2009): "On the Role of Policy Interventions in Structural Change and Economic Development: The Case of Postwar Japan," Discussion papers 09001, Research Institute of Economy, Trade and Industry (RIETI).

Felbermayr, G., J. Prat, and H.-J. Schmerer (2011): "Globalization and labor market

outcomes: Wage bargaining, search frictions, and firm heterogeneity," *Journal of Economic Theory*, 146(1), 39–73.

Hansen, G. D., and E. C. Prescott (2002): "Malthus to Solow," *American Economic Review*, 92(4), 1205–1217.

Hayashi, F., and E. C. Prescott (2008): "The Depressing Effect of Agricultural Institutions on the Prewar Japanese Economy," *Journal of Political Economy*, 116(4), 573–632.

Helpman, E., and O. Itskhoki (2010): "Labour Market Rigidities, Trade and Unemployment," *Review of Economic Studies*, 77(3), 1100–1137.

Helpman, E., O. Itskhoki, and S. Redding (2010): "Inequality and Unemployment in a Global Economy," *Econometrica*, 78(4), 1239–1283.

Hosios, A. J. (1990): "On the Efficiency of Matching and Related Models of Search and Unemployment," *Review of Economic Studies*, 57(2), 279–98.

Ishimaru, S., S. Oh, and S.-G. Sim (2013): "Unemployment and Inequality after Trade Liberalization," Discussion papers.

Jorgenson, D. (1996): "Empirical Studies of Depreciation," *Economic Inquiry*, 34, 24–42, Econometrics 1, pp. 73-96.

Kano, S., and M. Ohta (2002): "An Empirical Matching Function with Regime Switching: The Japanese Case," Discussion paper, Institute of Policy and Planning Sciences, No. 967.

Kennan, J., and J. R. Walker (2011): "The Effect of Expected Income on Individual Migration Decisions," *Econometrica*, 79(1), 211–251.

Lipton, D., J. Poterba, J. Sachs, and L. Summers (1982): "Multiple Shooting in Rational Expectations Models," *Econometrica*, 50(5), pp. 1329–1333.

Lucas, R. J. (1988): "On the mechanics of economic development," *Journal of Monetary Economics*, 22(1), 3–42.

Matsuyama, K. (1992): "Agricultural productivity, comparative advantage, and economic growth," *Journal of Economic Theory*, 58(2), 317–334.

Mcfadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in econometrics*, ed. by P. Zarembka, pp. 105–142. Academic Press, New York.

Miyamoto, H., and Y. Takahashi (2011): "Productivity growth, on-the-job search, and unemployment," *Journal of Monetary Economics*, 58(6-8), 666 – 680.

Mortensen, D. T., and C. A. Pissarides (1994): "Job Creation and Job Destruction in the Theory of Unemployment," *Review of Economic Studies*, 61(3), 397–415.

Mortensen, D. T., and C. A. Pissarides (1998): "Technological Progress, Job Creation

and Job Destruction," *Review of Economic Dynamics*, 1(4), 733–753.

Ngai, L. R., and C. A. Pissarides (2007): "Structural Change in a Multisector Model of Growth," *American Economic Review*, 97(1), 429–443.

Petrongolo, B., and C. A. Pissarides (2001): "Looking into the Black Box: A Survey of the Matching Function," *Journal of Economic Literature*, 39(2), 390–431.

Pissarides, C. A., and G. Vallanti (2007): "The Impact Of TFP Growth On Steady-State Unemployment," *International Economic Review*, 48(2), 607–640.

Romer, P. M. (1986): "Increasing Returns and Long-run Growth," *Journal of Political Economy*, 94(5), 1002–37.

_____(1987): "Growth Based on Increasing Returns Due to Specialization," *American Economic Review*, 77(2), 56–62.

Rust, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55(5), 999–1033.

Shimer, R. (2005): "The Cyclical Behavior of Equilibrium Unemployment and Vacancies," *American Economic Review*, 95(1), 25–49.

Stole, L. A., and J. Zwiebel (1996): "Intra-firm Bargaining under Non-binding Contracts," *Review of Economic Studies*, 63(3), 375–410.

Uy, T., K.-M. Yi, and J. Zhang (2013): "Structural change in an open economy," *Journal of Monetary Economics*, 60(6), 667–682.

Appendix

A. Numerical Algorithm

A.1. On the Balanced Growth Path

In this appendix, we describe the computational algorithm for solving the balanced growth path. To start, guess (θ_a, θ_b).

(a) Given (θ_a, θ_m), we obtain $(U_{mt} - U_{at})/\varepsilon_t$ and (ω_a, ω_m) using (4) and (36).

(b) Let

$$LHS_i = [\gamma(r + \delta + \rho - \psi)/(\beta_{li} q(\theta_i)) + (A_{it}^1/\beta_{li}\varepsilon_t)]^{1-\beta_{ki}-\beta_{ai}\kappa_i}, \text{ and}$$

$$RHS_i = \left[\frac{\lambda_i m_i}{\chi - \rho + \psi + \eta_a}\right]^{\beta_{ai}(1-\kappa_i)} \left[\frac{\lambda_i \kappa_i \beta_{ai}}{p_z(r + \eta_a)}\right]^{\beta_{ai}\kappa_i} \left[\frac{\beta_{ki}}{p_k(r + \eta_k)}\right]^{\beta_{ki}} (1 - \beta_{li} A_i^0) p_i.$$

If $\sum_i RHS_i - LHS_i)^2$ is less than the preassigned tolerance level, go to step (*c*). Otherwise, using the Nelder and Meade method, update (θ_a, θ_m) and go back to step (a).

(c) Given (θ_a, θ_m), solve for variables $\{L_i, u_i w_i, k_i, x_i, a_i, z_i, v_i, Y_i\}_{i\in\{a,m\}}$.

A.2. On the Transition Path

In this subsection, we present the solution algorithm for the transitional path to the new balanced growth path. First, we analyze the transition path from an initial point to the autarky balanced growth path. Assume that the economy converges to the new balanced growth path

within T. We already know both endpoints.

(a) Pick up a sufficiently large amount of time for T and construct the set of evenly spaced grid points $\{t_0, t_1, \cdots, t_n\} \subset [0, T]$.
(b) Guess the entire transition path of $(L_{at}, L_{mt}, u_{at}, u_{mt}, W_{at}, W_{mt}, U_{at}, U_{mt}, p_{at}, p_{mt})$.
(c) Take all the other series as given and calculate the new series of $(\hat{L}_{at}, \hat{L}_{mt}, \hat{u}_{at}, \hat{u}_{mt})$ by forward shooting. Then, update $L_{at} = (1-a)\, L_{at} + a\hat{L}_{at}$ for a sufficiently small but positive a. Using the same weight a, update $(\hat{L}_{mt}, \hat{u}_{at}, \hat{u}_{mt})$.
(d) Take all the other series as given and iterate the new series of $(\widehat{W}_{at},\ \widehat{W}_{mt}, \widehat{U}_{at}, \widehat{U}_{mt})$ by the backward shooting and weighted updating procedure as in step (c).
(e) Take all the other series as given and iterate the prices $(\hat{p}_{at}, \hat{p}_{mt})$ using the market clearing condition and weighted updating procedure. In case of the open economy, skip this step.
(f) Iterate step (c), (d), and (e), until all the series converge to the below of a certain tolerance level. If the differentials between the initial value and the updated values are small enough, move onto step (g).
(g) Prolong the time interval into $[0, \tilde{T}]$, where $\tilde{T} > T$. Repeat step (b)-(f). If the (point-wise) maximum difference between the two sets of the converged series in $[0, T]$ is below a certain tolerance level, stop here. Otherwise, update $T = \tilde{T}$, enlarge $\tilde{T}$ and repeat step (a)-(g).

B. Mathematical Appendix

Proof of Lemma 1 The entrepreneur chooses $(x_{i\tau}, z_{i\tau}, v_{i\tau})$, at every $\tau \in [t, \infty)$ to maximize

$$E_{it}\left(\overline{a}_i, \overline{k}_i, \overline{l}_{ih}, \overline{l}_{il}\right) = \max_{z_{i\tau}, x_{i\tau}, v_{i\tau} \geq 0} \int_{\mathrm{t}}^{\infty} e^{-r(\tau - t)}\, \pi_{i\tau} P^{-1} d\tau \tag{B1}$$

subject to

$$\dot{a}_{i\tau} = -\eta_a a_{i\tau} + \lambda_i z_{i\tau}^{\kappa_i} (\varepsilon_\tau L_{ih\tau})^{1-\kappa_i} \tag{B2}$$

$$\dot{k}_{i\tau} = -\eta_k k_{i\tau} + x_{i\tau} \tag{B3}$$

$$\dot{l}_{ih\tau} = -(\delta + \rho) l_{ih\tau} + \zeta l_{il\tau} \tag{B4}$$

$$\dot{l}_{il\tau} = -(\delta + \rho + \zeta) l_{il\tau} + q(\theta_{i\tau})\, v_{i\tau} \tag{B5}$$

$$a_{it} = \overline{a}_i, k_{it} = \overline{k}_i, l_{iht} = \overline{l}_{ih}\text{ , and } l_{ilt} = \overline{l}_{il} \tag{B6}$$

First, we ignore the non-negative restriction in the domain and solve for the optimal control problem. Then, we check if the optimal decision is binding. The Hamiltonian for the above problem is

$$\begin{aligned}\mathcal{H} = e^{-r(\tau-t)}[&p_i \alpha_{i\tau}^{\beta_{ai}} k_{i\tau}^{\beta_{ki}} \varepsilon_\tau^{\beta_{li}} (l_{il\tau} + \alpha_i l_{ih\tau})^{\beta_{li}} - \\ &\sum_{j=h,l} w_{ij\tau}\, l_{ij\tau} - p_k x_{i\tau} - \gamma \varepsilon_\tau v_{i\tau} - p_z z_{i\tau}] P^{-1} \\ &-\mu_a[\eta_a a_{i\tau} - \lambda_i z_{i\tau}^{\kappa_i} (\varepsilon_\tau L_{ih\tau})^{1-\kappa_i}] - \mu_k[\eta_k k_{i\tau} - x_{i\tau}] \\ &-\mu_h[(\delta + \rho) l_{ih\tau} - \zeta l_{il\tau}] - \mu_l[(\delta + \rho + \zeta) l_{il\tau} - q(\theta_{i\tau}) v_{i\tau}].\end{aligned}$$

The maximum principle implies that

$$z_{i\tau}:\ e^{-r(\tau-t)} P^{-1} p_z z_{i\tau}^{1-\kappa_i} = \mu_a \kappa_i \lambda_i (\varepsilon_\tau L_{ih\tau})^{1-\kappa_i} \tag{B7}$$

$$x_{i\tau}:\ e^{-r(\tau-t)} P^{-1} p_k = \mu_k \tag{B8}$$

$$v_{i\tau}:\ e^{-r(\tau-t)} P^{-1} \varepsilon_\tau \gamma = ulq(\theta_{i\tau}) \tag{B9}$$

$$a_{i\tau}:\ \dot{\mu}_a = -e^{-r(\tau-t)} P^{-1} \frac{\partial \pi_{i\tau}}{\partial a_{i\tau}} + \mu_a \eta_a \tag{B10}$$

$$k_{i\tau}:\ \dot{\mu}_k = -e^{-r(\tau-t)} P^{-1} \frac{\partial \pi_{i\tau}}{\partial k_{i\tau}} + \mu_k \eta_k \tag{B11}$$

$$l_{ih\tau}:\ \dot{\mu}_h = -e^{-r(\tau-t)} P^{-1} \frac{\partial \pi_{i\tau}}{\partial l_{ih\tau}} + \mu_h(\delta + \rho) \tag{B12}$$

$$l_{il\tau}: \dot{\mu}_i = -e^{-r(\tau-t)}P^{-1}\frac{\partial \pi_{i\tau}}{\partial l_{il\tau}} + \mu_h\zeta + \mu_l(\delta+\rho+\zeta) \quad \text{(B13)}$$

From (B10),

$$e^{-\eta_a(\tau-t)}\dot{\mu}_a - \eta_a e^{-\eta_a(\tau-t)}\mu_a = e^{-(r+\eta_a)(\tau-t)}P^{-1}\frac{\partial \pi_{i\tau}}{\partial a_{i\tau}}$$
$$\Leftrightarrow \mu_a = e^{\eta_a(\tau-t)}\int_\tau^\infty e^{-(r+\eta_a)(\tau'-t)}P^{-1}\frac{\partial \pi_{i\tau'}}{\partial a_{i\tau'}}d\tau' + C_a e^{\eta_a(\tau-t)}$$

Since the shadow price μ_a cannot diverge as $\tau \to \infty$, $C_a = 0$. Thus, we get

$$\mu_a = e^{\eta_a(\tau-t)}\int_\tau^\infty e^{-(r+\eta_a)(\tau'-t)}P^{-1}\frac{\partial \pi_{i\tau'}}{\partial a_{i\tau'}}d\tau' \quad \text{(B14)}$$

Then, plugging (B14) into (B7) yields

$$z_{it} = \varepsilon_t L_{iht}\left[\frac{\lambda_i\kappa_i}{p_z}\int_t^\infty e^{-(r+\eta_a)(\tau-t)}\frac{\partial \pi_{i\tau}}{\partial a_{i\tau}}\,d\tau\right]^{\frac{1}{1-\kappa_i}} \quad \text{(B15)}$$

Along the similar reasoning, by combining (B8) and (B11), we get

$$p_k = \int_t^\infty e^{-(r+\eta_k)(\tau-t)}\frac{\partial \pi_{i\tau}}{\partial k_{i\tau}}\,d\tau. \quad \text{(B16)}$$

From (B12) and (B13), we get

$$\mu_h = e^{(\delta+\rho)(\tau-t)}\int_\tau^\infty e^{-(r+\delta+\rho)(\tau'-t)}P^{-1}\frac{\partial \pi_{i\tau'}}{\partial l_{ih\tau'}}d\tau', \text{and} \quad \text{(B17)}$$

$$\mu_l = e^{(\delta+\rho+\zeta)(\tau-t)}\int_\tau^\infty e^{-(r+\delta+\rho+\zeta)(\tau'-t)}P^{-1}\left[\frac{\partial \pi_{i\tau'}}{\partial l_{il\tau'}} - \frac{\partial \pi_{i\tau'}}{\partial l_{ih\tau'}}\right]d\tau'$$
$$+e^{(\delta+\rho)(\tau-t)}\int_\tau^\infty e^{-(r+\delta+\rho)(\tau'-t)}P^{-1}\frac{\partial \pi_{i\tau'}}{\partial l_{ih\tau'}}d\tau' \quad \text{(B18)}$$

Plugging (B18) into (B9) yields

$$\frac{\varepsilon_t \gamma}{P} = q(\theta_{it}) \int_t^{\infty} e^{-(r+\delta+\rho)(\tau-t)} P^{-1} \left[e^{-\zeta(\tau-t)} \frac{\partial \pi_{i\tau}}{\partial l_{il\tau}} + (1 - e^{-\zeta(\tau-t)}) \frac{\partial \pi_{i\tau}}{\partial l_{ih\tau}} \right] d\tau \quad \text{(B19)}$$

Finally, since

$$\frac{\partial E_{it}}{\partial l_{ilt}} = \int_t^{\infty} e^{-(r+\delta+\rho)(\tau-t)} P^{-1} \left[e^{-\zeta(\tau-t)} \frac{\partial \pi_{i\tau}}{\partial l_{il\tau}} + (1 - e^{-\zeta(\tau-t)}) \frac{\partial \pi_{i\tau}}{\partial l_{ih\tau}} \right] d\tau, \quad \text{(B20)}$$

connecting (B19) and (B20) results in

$$P^{-1} \varepsilon_t \gamma = q(\theta_{it}) \frac{\partial E_{it}}{\partial l_{ilt}} \quad \text{(B21)}$$

Proof of Lemma 2 Let $E^h_{iht} := \frac{\partial E_{iht}}{\partial l_{iht}}$ and $E^l_{ilt} := \frac{\partial E_{ilt}}{\partial l_{ilt}}$. From the bargaining rule proposed by Stole and Zwiebel (1996),

$$(1-\phi)\left(W_{ijt} - U_{it}\right) = \phi E^j_{ijt} \text{ and } (1-\phi)\left(\dot{W}_{ijt} - \dot{U}_{it}\right) = \phi \dot{E}^j_{ijt} \quad \text{(B22)}$$

By equation (B21) and (B22), we get as follows.

$$(1-\phi)(w_{iht} - b\varepsilon_t - \xi \Delta_{it} P) = \phi\left(\frac{\partial \pi_{it}}{\partial l_{iht}} + \varepsilon_t \gamma \theta_{it}\right) \quad \text{(B23)}$$

$$(1-\phi)(w_{ilt} - b\varepsilon_t - \xi \Delta_{it} P) = \phi\left(\frac{\partial \pi_{it}}{\partial l_{ilt}} + \varepsilon_t \gamma \theta_{it}\right) \quad \text{(B24)}$$

Rewriting this yields

$$w_{iht} + \phi \sum_{j=l,h} \frac{\partial w_{ijt}}{\partial l_{iht}} l_{ijt} = \phi p_i \frac{\partial y_{it}}{\partial l_{iht}} + (1-\phi)(b\varepsilon_t + \xi \Delta_{-it} P^{-1}) + \phi \varepsilon_t \gamma \theta_{it} \quad \text{(B25)}$$

$$w_{ilt} + \phi \sum_{j=l,h} \frac{\partial w_{ijt}}{\partial l_{ilt}} l_{ijt} = \phi p_i \frac{\partial y_{it}}{\partial l_{ilt}} + (1-\phi)(b\varepsilon_t + \xi \Delta_{-it} P^{-1}) + \phi \varepsilon_t \gamma \theta_{it} \quad \text{(B26)}$$

The solution of the above differential equation is given by

$$w_{iht} = \alpha_i A_i^0 p_i \frac{\partial y_{it}}{\partial l_{ilt}} + A_{it}^1, \text{and } w_{ilt} = A_i^0 p_i \frac{\partial y_{it}}{\partial l_{ilt}} + A_{it}^1 \quad \text{(B27)}$$

where

$$A_i^0 = \frac{\phi}{1-\phi(1-\beta_{li})} \text{ and } A_{it}^1 = (1-\phi)(b\varepsilon_t - \xi \Delta_{-it} P) + \varepsilon_t \gamma \phi \theta_{it} \quad \text{(B28)}$$

Proof of Lemma 3 Rewriting (39) and (40) in matrix notation yields

$$\begin{pmatrix} \chi+\delta & 0 & -f(\theta_a) & 0 \\ 0 & \chi+\delta & 0 & -f(\theta_m) \\ -\delta & 0 & \chi + f(\theta_a) + \xi\omega_m & -\xi\omega_a \\ 0 & -\delta & -\xi\omega_m & \chi + f(\theta_m) + \xi\omega_a \end{pmatrix} \begin{pmatrix} L_{ah} + L_{al} \\ L_{mh} + L_{ml} \\ u_a \\ u_m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \chi\omega_a \\ \chi\omega_m \end{pmatrix}$$

The matrix on the left-hand side is non-singular. Multiplying the inverse matrix of it on both side yields $(L_{ah} + L_{al}, L_{mh} + L_{ml}, u_a, u_m)$. Since $(\chi+\delta)L_{ih} = \zeta L_{il}$ for each $i \in \{a, m\}$ on steady states, we get $(L_{ah}, L_{al}, L_{mh}, L_{ml}, u_a, u_m)$.

Proof of Lemma 4 Since

$$\frac{\partial \pi_{it}}{\partial a_{it}} = \beta_{ai} p_i (1 - \beta_{li} A_i^0) a_{it}^{\beta_{ai}} k_{it}^{\beta_{ki}} [\varepsilon_{it} (l_{ilt} + \alpha_i l_{iht})]^{\beta_{li}} \frac{1}{a_{it}}, \tag{B29}$$

We get

$$z_{it} = \varepsilon_t L_{iht} \left[\frac{\lambda_i \kappa_i P}{p_z} \int_t^{\infty} e^{-(r+\eta_a)(\tau - t)} P^{-1} \frac{\partial \pi_{i\tau}}{\partial a_{i\tau}} \, d\tau \right]^{\frac{1}{1-\kappa_i}}$$

$$= \varepsilon_t L_{iht} \left[\frac{\lambda_i \kappa_i \beta_{ai} p_i (1 - \beta_{li} A_i^0)}{p_z (r + \eta_a) \beta_{li} a_{it}} \frac{\partial y_{it}}{\partial l_{ilt}} (l_{ilt} + \alpha_i l_{iht}) \right]^{\frac{1}{1-\kappa_i}},$$

and

$$a_{it} = \frac{\lambda_i \varepsilon_t L_{iht}}{\chi - \rho + \psi + \eta_a} \left[\frac{\lambda_i \kappa_i \beta_{ai} p_i (1 - \beta_{li} A_i^0)}{p_z (r + \eta_a) \beta_{li} a_{it}} \frac{\partial y_{it}}{\partial l_{ilt}} (l_{ilt} + \alpha_i l_{iht}) \right]^{\frac{\kappa_i}{1-\kappa_i}}$$

$$= \left[\frac{\lambda_i \varepsilon_t L_{iht}}{\chi - \rho + \psi + \eta_a} \right]^{1-\kappa_i} \left[\frac{\lambda_i \kappa_i \beta_{ai} p_i (1 - \beta_{li} A_i^0)}{\beta_{li} p_z (r + \eta_a)} \frac{\partial y_{it}}{\partial l_{ilt}} (l_{ilt} + \alpha_i l_{iht}) \right]^{\kappa_i}. \tag{B30}$$

From (19),

$$p_k = \int_t^{\infty} e^{-(r+\eta_k)(\tau - t)} \frac{\partial \pi_{i\tau}}{\partial k_{i\tau}} d\tau = \int_t^{t+dt} e^{-(r+\eta_k)(\tau - t)} \frac{\partial \pi_{i\tau}}{\partial k_{i\tau}} d\tau +$$

$$e^{-(r+\eta_k) dt} p_k$$

Then, reordering and dividing by *dt* yields

$$\frac{1 - e^{-(r+\eta_k) dt}}{dt} p_k = \frac{1}{dt} \int_t^{t+dt} e^{-(r+\eta_k)(\tau - t)} \frac{\partial \pi_{i\tau}}{\partial k_{i\tau}} d\tau$$

Finally, sending $dt \to 0$ and reordering results in

$$k_{it} = \left[\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i a_{it}^{\beta_{ai}}\varepsilon_t^{\beta_{li}}(l_{ilt}+\alpha_i l_{iht})^{\beta_{li}}}{(r+\eta_a)p_k}\right]^{\frac{1}{1-\beta_{ki}}} \text{ and} \tag{B31}$$

$$x_{it} = \chi - \rho + \psi + \eta_k \left[\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i a_{it}^{\beta_{ai}}\varepsilon_t^{\beta_{li}}(l_{ilt}+\alpha_i l_{iht})^{\beta_{li}}}{(r+\eta_k)p_k}\right]^{\frac{1}{1-\beta_{ki}}} \tag{B32}$$

Plugging (B30) and (B31) into the expression of $\frac{\partial y_{it}}{\partial l_{ilt}}$ and reordering yields

$$\frac{\partial y_{it}}{\partial l_{ilt}} = \left[\beta_{li}\left(\frac{\lambda_i \varepsilon_t L_{iht}}{\chi-\rho+\psi+\eta_a}\right)^{\beta_{ai}(1-k_i)}\left(\frac{\lambda_i k_i \beta_{ai} p_i\left(1-\beta_{li}A_i^0\right)}{\beta_{li}p_z\left(r+\eta_a\right)}\right.\right.$$
$$\left.(l_{ilt}+\alpha_i l_{iht})\right)^{\beta_{ai}\kappa_i}\left(\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_i(l_{ilt}+\alpha_i l_{iht})}{\beta_{li}p_k(r+\eta_k)}\right)^{\beta_{ki}}$$
$$\left.\left.\varepsilon_t^{\beta_{li}}(l_{ilt}+\alpha_i l_{iht})^{\beta_{li}-1}\right]^{\frac{1}{1-\beta_{ki}-\beta_{ai}k_i}}\right.$$

Since $L_{ih} = l_{ih}n_i, L_{il} = l_{il}n_i$, and $q(\theta_i)v_i = (\chi+\delta+\zeta)$, we get (46).

Proof of Proposition 1 First, suppose that there exists a balanced growth path. The stationarity conditions dictate that θ_{at}, and θ_{mt} should be constant on the balanced growth path. Then, by construction, equation (52) should be satisfied on the balanced growth path. Second, suppose that the system of equations described in (52) has a solution of (θ_a, θ_m). Then, by invoking Lemma 1 through 4, we know that any pair of (θ_a, θ_m) can generate the solution of the model satisfying the equilibrium configuration (*i*) through (*iv*). Therefore, the solution of (52) solves for the balanced growth path.

Proof of Lemma 5 By the same reasoning as in the proof of Lemma 4, we get

$$k_{it} = \left[\frac{(1-\beta_{li}A_i^0)\beta_{ki}p_{it}a_{it}^{\beta_{ai}}\varepsilon_t^{\beta_{li}}(l_{ilt}+\alpha_i l_{iht})^{\beta_{li}}}{(r+\eta_k)p_k}\right]^{\frac{1}{1-\beta_{ki}}}. \tag{B33}$$

Since

$$\frac{\partial \pi_{it}}{\partial a_{it}} = \beta_{ai}p_i\left(1-\beta_{li}A_i^0\right)a_{it}^{\beta_{ai}}k_{it}^{\beta_{ki}}[\varepsilon_{it}(l_{ilt}+\alpha_i l_{iht})]^{\beta_{li}}\frac{1}{a_{it}}, \tag{B34}$$

we get

$$z_{it} = \varepsilon_t L_{iht}\left[\frac{\lambda_i\kappa_i}{p_z}\int_t^{\infty} e^{-(r+\eta_a)(\tau-t)}\beta_{ai}p_i\left(1-\beta_{li}A_i^0\right)a_{i\tau}^{\beta_{ai}-1}k_{i\tau}^{\beta_{ki}}[\varepsilon_{i\tau}(l_{il\tau} + \alpha_i l_{ih\tau})]^{\beta_{li}}d\tau\right]^{\frac{1}{1-\kappa_i}}$$

It completes the proof.

CHAPTER 13

Wage Dynamics with Private Learning-by-doing and On-the-job Search

by
Seung-Gyu Sim[*]
(University of Tokyo)

Abstract

This paper develops an equilibrium job search model in which the employed worker privately accumulates human capital and continually searches for a better paying job. Firms encourage production and discourage turnover by rewarding with bonus payments and long service allowances, respectively, workers with better performance and longer job tenure. Wage growth attends human capital accumulation (productive promotion) and job tenure (non-productive promotion) as well as job-to-job transition. The model is estimated using indirect inference to investigate the effect of human capital accumulation on

* This is the revised version of the second chapter in my Ph.D thesis. I deeply thank Shin-ichi Fukuda, Hugo Hopenhayn, Hidehiko Ichimura, John Kennan, Francis Kramarz, Rasmus Lentz, Shouyong Shi, Yongseok Shin, and Chris Taber for their helpful guidance and comments. I am also grateful to all participants at various seminars and conferences including 2014 AEA Meeting, Dallas Fed, Penn State University, 2013 SED Annual Meeting, Seoul National University, 2014 Society of Labor Economists Meeting, St. Louis Fed, University of Iowa, University of Maryland, University of Tokyo, University of Washington, University of Wisconsin, and University of Virginia. All remaining errors are mine. Please address correspondence to: Seung-Gyu Sim, Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-0033. Tel: +81-3-5841-5620. E-mail: sgsim@e.u-tokyo.ac.jp.

individual wage growth. In NLSY79 data, the average wage of white male high school graduates after 20 years of market experience is 1.88 times higher than the average of the first full-time wages. A counterfactual experiment using the structural parameter estimates shows that the wage of a typical worker unable to accumulate human capital would grow by 41.8%.

1. Introduction

In the U.S. labor market, the wage of a typical male[1] worker doubles over the course of a 40-year career.[2] Wage growth is generally understood to be the outcome of human capital accumulation subsequent to entering the labor market. But human capital accumulation being neither a sufficient nor necessary condition for wage growth in the frictional labor market, it is necessary to develop a structural model to identify and quantify the relevant sources of wage growth. This paper builds and estimates an equilibrium job search model by focusing on firms' strategic responses to workers' private learning-by-doing and on-the-job search behavior.

Burdett and Mortensen (1998) develop a wage-posting model with on-the-job search in which some firms post high wages to attract workers from other firms and workers ascend the wage ladder only through job-to-job transition. Burdett and Coles (2003) depart from the aforementioned model by building an equilibrium in which firms optimally back-load some portion of wage payments to discourage job turnover and extract more surplus from early leavers. All firms post, and choose different starting points on, a common wage-tenure schedule. In equilibrium, wages grow with the back-loading schedule and job

1 The masculine pronoun is used throughout the paper because a male sample of workers is used in the empirical analysis.

2 Topel and Ward (1992) report this fact in cross sectional data. The average wage in the focal sample of white male high school graduates in the 1979 National Longitudinal Survey of Youth after 5, 10, and 20 years of market experience is, respectively, 1.43, 1.61, and 1.88 times higher than the average of the first full-time wages.

turnover. Burdett, Carrillo-Tudela, and Coles (2009) extend the Burdett and Mortensen (1998) framework by adding human capital accumulation. They assume that each worker accumulates human capital through the deterministic, on the job, learning-by-doing process, and is paid following a single wage-output ratio (the piece rate sharing rule), and that wages grow consequent to job turnover and learning-by-doing. Why firms stick to the piece rate sharing rule is unclear, however.

In fact, each worker privately accumulates human capital on the job through learning-by-doing, while continually searching for a better paying job. Incentives offered by firms unable to monitor the learning-by-doing process or job search out-comes include bonus payments and long-service allowances for superior performance and longer tenure, respectively. This motivates the present model's incorporation of private learning-by-doing into the equilibrium wage-tenure contract framework of Burdett and Coles (2003) and Stevens (2004). The resulting wage in a job rises with worker performance and job tenure. Market equilibrium implies multiple wage-ladders, each associated with workers identified with a particular level of human capital. Workers gradually climb their respective wage ladders along the back-loading wage schedule (non-productive promotion), jump to a higher rung through job-to-job transition, and switch ladders through human capital accumulation or depreciation (productive promotion).

The private learning-by-doing process originates two types of upward pressure in workers' compensation, (i) to induce truthful revelation, a commitment to higher value for better output (internal pressure), and (ii) productive workers being more attractive to poaching firms, a commitment to pay more to retain workers (external pressure). The latter, if it dominates the former, pushes the wage payment beyond the incentive compatible wage level. The incentive compatibility constraint should then be slack and information asymmetry no longer matter. Because, given the convexity of the offer and earning distributions, the optimal productive promotion schedule post human capital accumulation may not be well defined, another wage determination mechanism, such as the piece rate sharing rule, is required.[3] If, on the other hand, internal

3 Also required is anti-discrimination legislation, as in Burdett, Carrillo-Tudela, and Coles (2009), which, in assuming payment of exactly the same piece rate to all workers

dominates external pressure, the optimal productive promotion schedule post human capital accumulation is determined by the (least cost) incentive compatibility constraints, which should be binding. The learning-by-doing process, practically speaking, being unobservable by firms, the present paper focuses on the case in which internal dominates external pressure.

The empirical analysis estimates the model and performs some counterfactual experiments to investigate returns to both human capital and tenure. A sample of white male high school graduates from the 1979 National Longitudinal Survey of Youth is constructed and tracked with respect to employment and wage history. The structural parameters of the model are estimated using indirect inference. The model implies human capital accumulation to have a permanent effect on wage growth and the effect of job tenure to be reset when a worker experiences unemployment. The difference between usual wage growth and re-employment wage[4] growth is exploited to capture the effect of human capital accumulation (productive promotion) independently of that of job tenure.

In the sample, the average wage after 20 years of market experience is 88% higher than the average of first full-time wages. The counter-factual analysis, which finds that wages would grow by 41.8% without human capital accumulation, implies that the return to human capital is, at best, the other 46% wage growth in the first 20 years. Considering the interactions, this might be construed to be an upper bound on returns to human capital. The limited effect of human capital accumulation is consistent with the sample's estimated slope coefficient in the re-employment wage-experience regression, being almost half the coefficient in the usual wage-experience regression.

Altonji and Shakotko (1987) attribute the lion's share of individual wage growth to returns to experience (general human capital), identified by 'within job wage growth transferred to the next job', and only limited

---

regardless of job tenure or performance, forestalls back-loading subsequent to, and extraction of all surplus prior to, human capital accumulation. The present paper assumes that two workers with the same job tenure and responsible for the same output cannot be paid differentially.

4 By re-employment wage is meant the first wage after unemployment.

effects to returns to tenure (job specific human capital), identified by 'within job wage growth non-transferable to the subsequent job'. Although not at odds with the empirical findings in Altonji and Shakotko (1987) and Altonji and Williams (2005) in the sense that 'within job wage growth transferred to the next job' is significant in wage growth, the present paper argues that care be exercised in interpreting 'within job wage growth transferred to the next job' as returns to human capital; because non-productive promotion is also transferred to subsequent jobs in the model used here, wage growth through productive promotion is redefined as returns to human capital, and wage growth through non-productive promotion as returns to tenure.

Bagger, Fontaine, Postel-Vinay, and Robin (2006), who study wage dynamics by combining learning-by-doing on the job with the ex post offer matching framework proposed by Postel-Vinay and Robin (2002), assume firm and worker to sign on to a particular piece rate. An employed worker who finds a recruiting firm is bid new piece rates by the existing and recruiting firms, and accepts the offer with the higher lifetime value. Wages grow through human capital accumulation, that is, 'search and stay' and 'search and switch'. Estimating the model finds wage growth through 'search and stay' to have, consistent with the present paper, a substantial effect. The key difference between the two papers lies in the way returns to tenure are modeled; whereas the earlier paper assumes an ex post offer matching process,[5] the present paper relies on an ex ante, preemptive back-loading scheme.

The paper proceeds as follows. The theoretical model is built and the equilibrium of interest characterized in section 2. The sample is constructed and relevant variables are defined in section 3. The estimation protocol and results are presented in section 4. Section 5 concludes. All proofs and data construction are in the Appendix.

5 Ex post offer matching is vulnerable to the criticism that it provides the wrong incentives for a worker to search for an outside offer. See Shimer (2006).

2. The Model

2.1. Environment

Consider a labor market populated by a unit measure of infinitely-lived and homogeneous risk-neutral firms, and mortal and heterogeneous, in terms of the level of human capital, risk-averse workers $y_i \in \mathcal{Y} := \{y_1, y_2, \cdots, y_n\}$. For expositional convenience, a worker having y_i units of human capital is designated a 'y_i-type worker'. Human capital is y_1 units upon a newly-born worker's entry to the labor market, and accumulates throughout his career. A worker who stochastically retires (or dies) is replaced immediately by another newly-born worker. The model is set in continuous time, and all firms and workers discount the future at rate r.

Workers A worker is either unemployed or employed. A y_{i-1}-type unemployed worker collects unemployment benefits b per instant, finds a job offer at rate λ_u, retires at rate ρ, and becomes a y_{i-1}-type by losing human capital at rate η. Denote as U_i the equilibrium asset value of the y_i-type unemployed worker, and let $F_i(\cdot)$ be the cumulative distribution function of lifetime values offered by recruiting firms to y_i-type workers. Equilibrium support and the cumulative distribution function are endo-genously determined later. The HJB equation for the y_i-type unemplo-yed worker is given by

$$rU_i = u(b) + \lambda_u \int \max\{x - U_i, 0\}\, dF_i(x) - \rho U_i + \eta(U_{i-1} - U_i). \quad (1)$$

There being no worker with human capital strictly less than y_1 units, ignoring U_0 or setting $U_0 = U_1$ in equation (1) implies that a y_1-type worker is not subject to depreciation shock. In the asset value equation (1), the left-hand side is interpreted as the opportunity cost of holding the asset y_i-type unemployment. The terms on the right-hand side are interpreted as the benefit flow from holding the asset U_i, which consists of the dividend flow from the asset, potential gains from job finding, potential loss from retirement, and potential loss from human capital depreciation, respectively.

A y_i-type employed worker can produce y_i units of output at every instant. Being time-varying private information, firms cannot pay different wages by worker type. It is assumed, instead, that anti-discrimination legislation dictates that a firm offer the same wage to workers with the same job tenure and performance.[6] Let $\phi : [0, \infty) \rightarrow \{y_1, y_2, \cdots, y_n\}$ be a mapping from the interval of job tenure to the set of types to which a worker potentially pretends to be. Operating firms determine wages as a function of job tenure and performance, that is, $w(t, \phi(t))$. It is further assumed that the flow disutility of a y_i-type employed worker who mimics a $\phi(t)$-type and produces $\phi(t)$-units of output at time *t* is given by

$$c_i(\phi(t)) = \begin{cases} \alpha_0 - \alpha_1(y_i - \phi(t)) & \text{if } \phi(t) \leq y_i \\ \infty & \text{otherwise} \end{cases}.$$

This implies that the disutility from working is proportional to hours worked. A y_i-type worker who produces y_i units of output incurs the disutility of α_0. A y_i-type worker who elects to produce $\phi(t)(< y_i)$ units can finish the job earlier and expend his disutility as leisure. The private benefit from misreporting is captured by $\alpha_1(y_i - \phi(t))$. Note that in the market equilibrium, no efficiency loss accrues to the information asymmetry in each match.

A y_i-type employed worker finds another job offer at rate $\lambda \in (0, \lambda_u)$[7], privately accumulates human capital at rate μ, loses it at rate η, separates from his job at rate δ, and retires at rate ρ. The expected lifetime value of the y_i-type employed worker at tenure *t* who chooses production schedule $\phi(\cdot)$ under contract m, $E_i(t; \phi, m)$is given by

---

6 The anti-discrimination legislation concept is borrowed, with modifications, from Burdett, Carrillo-Tudela, and Coles (2009), whose assumption that a firm should pay exactly the same piece rate to all workers implies that workers who produce the same output should be paid the same wage. The present paper assumes that wages can vary with employee job tenure and performance.

7 Recruiting firms are assumed to be contacted more frequently by unemployed than by employed workers.

$$E_i(t;\phi,m)=\int_{\mathrm{t}}^{\infty} e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u\big(w(s,\phi_i(s);m)-c_i\big(\phi(s)\big)+z(s;m)\big]ds,$$

where

$$z_i(s;m)=\delta U_i+\lambda\int_{\underline{E_i}}^{\overline{E}_i}\max\{x,\max_{\phi(\cdot)}\{E_i(s;\phi,m)\}\}dF_i(x)$$
$$+\mu\max_{\phi(\cdot)}\{E_{i+1}(s;\phi,m)\}+\eta\max_{\phi(\cdot)}\{E_{i-1}(s;\phi,m)\}.$$

As above, $E_0(s;\phi,m)$ and $E_{n+1}(s;\phi,m)$ are ignored throughout the paper. A y_i-type employed worker under contract *m* can choose and update his own production schedule to maximize his lifetime value, which is is characterized at time *t* by

$\max_{\phi(\cdot)}\{E_i(t;\phi,m)\}$.

When a y_i-type employed worker truthfully chooses $\phi(\cdot)=y_i$, $E_i(t;m)$, is used instead of $E_i(t;\phi,m)$. The HJB equation for a y_i-type employed worker with a truthful production schedule is given by

$$\begin{aligned}rE_i(t;m)=&\,u\big(w(t,y_i(s);m)\big)-c_i(y_i)+\dot{E}_i(t;m)\\&+\lambda\int\max\{x-E_i(t;m),0\}\,dF_i(x)-\rho E_i(t;m)\\&+\delta\big(U_i-E_i(t;m)\big)+\mu(E_{i+1}(t;m)-E_i(t;m))\\&+\eta(E_{i-1}(t;m)-E_i(t;m)).\end{aligned}\qquad(2)$$

Firms Each firm maintains one vacancy at every instant. A firm recruits a worker by posting a labor contract, *m*, that specifies the action profile (or 'terms of trade') that stipulates the worker's output schedule and the lifetime value delivered by the firm under the truthful revelation assumption. That is, *m* is characterized by $\{(y_i(\cdot),E_i(\cdot\,;m))\}_{i=1}^{n}$.

Definition *Contract m is* incentive compatible for y_i-type *if*

$$E_i(t;m) \geq \max_{\phi}\{E_i(t;\phi,m)\} \quad \text{at each } t \in (0,\infty). \tag{3}$$

In particular, when the contract is incentive compatible for all types, it is called incentive compatible.

As a tie-breaking rule, a y_i -type worker who is indifferent is assumed to truthfully produce y_i. The least cost incentive compatibility is defined separately, as follows.

Definition *Contract, m, is* least cost incentive compatible for y_i-type *if the following statements hold*;

(i) Contract m is incentive compatible for y_i-type workers.

(*ii*) *There exists at least one $\phi : [0,\infty] \to \mathcal{Y} \cap \{y_i\}^c$ such that $E_i(t;m) = E_i(t;\phi,m)$ at any $t \in [0,\infty)$.*

In particular, when the contract is least cost incentive compatible for all types, it is called least cost incentive compatible.

When a menu of contracts is accepted by a worker, firm and worker together begin producing immediately. If y_i units are produced by an employee with job tenure *t*, the operating firm earns revenue y_i and makes wage payment $w(t, y_i; m)$. The match is destroyed when the worker leaves the job, whether voluntarily or involuntarily.[8] Denote as $J_i(t;m)$ the expected value of an operating job with a y_i-type worker under contract *m*. For expositional convenience, let

$$\hat{z}_i(s;m) = \mu J_{i+1}(s;m) + \eta J_{i-1}(s;m).$$

Given the promised value $\{E_i(0;m)\}_{i=1}^n$, the operating firm with a y_i-type worker chooses the schedule of $\{w(\cdot,\phi_i(\cdot);m\}_{i=1}^n$ to maximize the expected value

8 Because, in equilibrium, all jobs yield positive expected profit to firms, there is no endogenous firing.

$$\int_{t}^{\infty} e^{-\int_{t}^{s}\left[r+\rho+\delta+\lambda\left(1-F_i\left(E_i(\tau;m)\right)\right)+\mu+\eta\right]d\tau}\left[y_i(s)-w(s,y_i(s);m)+\hat{z}_i(s;m)\right]ds \tag{4}$$

subject to the sets of least cost incentive compatibility and promise-keeping constraints. The least cost incentive compatibility constraints presume conditions (*i*) and (*ii*) above. The promise-keeping constraints, described in (2), imply that a firm that commits $\{E_i(\cdot\,;m)\}_{i=1}^{n}$ in terms of the truthful revelation value should deliver same through wage schedules.

Let u_i and $G_i(x)$ be the proportion of y_i-type unemployed and employed workers, respectively, receiving the value less than *x*. Denote as $\overline{m}$ and $\underline{m}$ the contract offered by the most generous and least generous recruiting firm, respectively.[9] Then, for each type,

$$G_i\left(E_i\left(0;\underline{m}\right)\right)=0, \frac{\partial G_i(\cdot)}{\partial x}\geq 0, \text{and } \sum_{i=1}^{n}\left[u_i+G_i\left(E_i(0;\overline{m})\right)\right]1.$$

Denote as $\mathcal{M}$ the set of equilibrium contracts. The equal profit condition implies that

$$\sum_{i=1}^{n}\left(\lambda G_i\left(E_i(0;m)\right)+\lambda_u u_i\right)J_i(0;m) \begin{cases}=\pi, & \text{if } m\in\mathcal{M},\\ <\pi & otherwise.\end{cases} \tag{5}$$

Aggregating all recruiting firms' strategies yields the distribution of lifetime values offered to each type, $\{F_i\}_{i=1}^{n}$. Given $\{F_i\}_{i=1}^{n}$, both employed and unemployed workers behave optimally. Given $\{F_i\}_{i=1}^{n}$ and workers' optimal behaviors, operating firms choose the productive and non-productive promotion schedules that determine the steady state distribution $\{u_i, G_i(\cdot)\}_{i=1}^{n}$. Then, $\{u_i, G_i(\cdot), J_i, E_i\}_{i=1}^{n}$ should be consis-

9 The argument that some firms make the most generous offer to some types and other firms to other types is rendered moot by least cost incentive compatibility.

tent with the equal profit condition (5). The equilibrium is consequently defined as follows.

Definition *A market equilibrium requires that the following two conditions be met.*

(*i*) *Given* $\{F_i\}_{i=1}^n$*, a* y_i*-type unemployed worker accepts the contract* $\{(y_i, E_i(0;\ m))\ \}_{i=1}^n$ *if and only if*

$$\max_{\phi}\{E_i(0;\phi,m)\} \geq U_i.$$

(*ii*) *Given* $\{F_i\}_{i=1}^n$*, a* y_i*-type employed worker optimally chooses the production schedule and accepts a new contract m' if and only if*

$$\max_{\phi}\{E_i(0;\phi,m')\} \geq \max_{\phi}\{E_i(t;\phi,m)\}.$$

(*iii*) *Given* $\{F_i\}_{i=1}^n$*, an operating firm with contract m optimally chooses the wage schedules to deliver* $\{E_i(0;m)\}_{i=1}^n$*. The contracts described by* $\{y_i, E_i(t;m)\}_{i=1}^n$ *are least cost incentive compatible at any* $t \in \{0,\infty)$.

(iv) *Given* $\{F_i\}_{i=1}^n$*, the optimal behavior of each economic agent determines* $\{u_i, G_i(\cdot)\}_{i=1}^n$.

(v) *Given* $\{u_i, G_i(\cdot)\}_{i=1}^n$*, a recruiting firm optimally posts contract m given the equal profit condition described in* (5).

(vi) *The equilibrium distributions* $\{F_i, G_i(\cdot)\}_{i=1}^n$ *are stationary.*

2.2. Equilibrium Characterization

The equilibrium in which all firms offer least cost incentive compatible contracts is characterized below. Technically, all least cost incentive compatible constraints are assumed to be binding, and whether firms have incentives to deviate from the least cost incentive compatible

contracts is numerically checked. To make the model tractable, the following equilibrium restriction is imposed.

Eq'm Restriction Let $\mathcal{M}$ be the set of contracts offered on equilibrium. For all $E_i(t;m) \in (\underline{E}_i, \overline{E}_i)$,

(*i*) F_i is continuously differentiable and satisfies $F_i'(E_i)$ is bounded away from zero.

(*ii*) $F_1\big(E_1(t;m)\big) = F_2\big(E_2(t;m)\big) = \cdots = F_n\big(E_n(t;m)\big)$ for any $m \in \mathcal{M}$ and $t \in [0,\infty)$.

This restriction, which implies that the acceptance (and retention) probability under contract *m* is same across all types, is based on pre-imposed equilibrium condition that firms have no incentive to screen out any type. Note, too, that the second condition (*ii*) renders the least generous contract $\underline{m}$ and the most generous contact $\overline{m}$ well-defined.

Lemma 1 *Contract m is least cost incentive compatible if and only if*

$$u\big(w(t,y_i;m)\big) = u\big(w(t,y_1;m)\big) + \alpha_1(y_i - y_1), \text{at any } t \in [0,\infty). \tag{6}$$

Lemma 1 characterizes the least cost incentive compatible contracts. The set of least cost incentive compatible contracts, given that no firm has incentives to screen out any particular type, is characterized by

$$\min_i\{E_i(0;\underline{m}) - U_i\} = 0. \tag{7}$$

This implies that the lifetime value of employment delivered by the least generous firm should be not less than the values of unemployment for each type *i*, and at least one of them should be same to it. The equal profit condition dictates that, given $\underline{m}$,

$$\sum_{i=1}^{n} \lambda_u u_i J_i\big(0;\underline{m}\big) \sum_{i=1}^{n} \lambda G_i(E_i(0;\overline{m}) + \lambda_u u_i) J_i(0;\overline{m}). \tag{8}$$

There being no contract that dominates contract $\overline{m}$, the operating firm with contract $\overline{m}$ has no incentive to increase a worker's value over $\{E_i(0;\overline{m})\}_{i=1}^{n}$. That is, for any $i = 1,2,\cdots,n$,

$$E_i(t;\overline{m}) = E_i(0;\overline{m}), \text{at every } t \in [0,\infty). \tag{9}$$

This implies that, given *i*,

$$E_i(t;\overline{m}) = \frac{u\big(w(t,y_i;\overline{m})\big) - c_i(y_i) + \delta U_i + \mu E_{i+1}(t;\overline{m}) + \eta E_{i-1}(t;\overline{m})}{r+\rho+\delta+\mu+\eta}$$

$$= \overline{E}_i, \tag{10}$$

$$J_i(t;\overline{m}) = \frac{y_i - w\big(t,y_i;\overline{m}\big)}{r+\rho+\delta+\mu+\eta} = \overline{J}_i, \text{and} \tag{11}$$

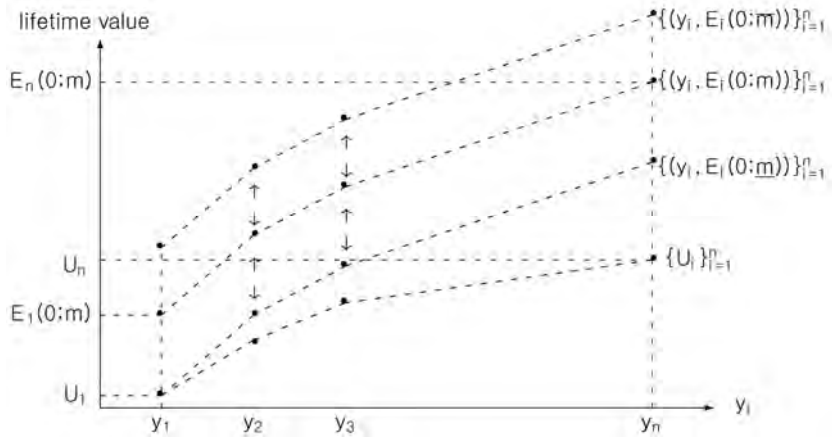
$$w(t,y_i;\overline{m}) = w(0,y_i;\overline{m}) = \overline{w}_i \text{ for any } t \in [0,\infty). \tag{12}$$

The second statement implies that, given $E_i(0;\overline{m})$ by the most generous recruiting firm, no firm has incentives to pay more than the value, which renders the optimal wage schedule for each type by the most generous firm constant. In equilibrium, the wage schedules of all firms are bounded.

The foregoing is summarized in [Figure 1]. If both *m* and m' are least cost incentive compatible, then $\{(y_i, E_i(0\,;m))\}_{i=1}^{n}$ and $\{(y_i, E_i(0\,;m'))\}_{i=1}^{n}$ cannot intersect. Lemma 1 determines the size of information rent, which is represented by the slope of the contract curve in [Figure 1]. The gap between $E_i\big(0\,;\underline{m}\big)$ and U_i should be zero at at least one y_i.[10] Given $\{F_i(\cdot), G(i,\cdot)\}_{i=1}^{n}$ and $\underline{m}$, the equal profit condition determines $\overline{m}$.

10 In most numerical experiments, $E_1\big(0;\underline{m}\big) = U_1$ and $E_{-1}\big(0;\underline{m}\big) > U_{-1}$.

Figure 1



It remains to examine the strategy of the least generous firm. Given $\underline{m}$, the firm chooses an optimal wage schedule to deliver the committed value $\{E_i(0;\underline{m})\}_{i=1}^n$. For expositional convenience, denote as w_i the wage schedule of the y_i-type worker who produces truthfully under the least cost incentive compatible contract $\underline{m}$. Let

$$\hat{\psi}(t) = \exp\left[\int_0^t (r+\rho+\delta+\lambda(1-F_1(E_1(s;m)))\,ds\right].$$

Lemma 2 *Given* $\{F_i\}_{i=1}^n$ *and* $E_1(0;\underline{m})$*, the optimal wage-tenure schedules solve for*

$$\dot{J}_i = -(y_i - w_i + \eta J_{i-1} + \mu J_{i+1}) + [r+\rho+\delta+\lambda(1-F_i(E_i)) + \mu + \eta]J_i,$$

$$\dot{E}_i = -u(w_i) + \alpha_0 + (r+\rho+\delta+\lambda(1-F_i(E_i)) + \mu + \eta]E_i - \delta U_i - \lambda\int_{E_i}^{\overline{E}} x dF_i(x) - \mu E_{i+1} - \eta E_{i-1},$$

$$\dot{w}_i = \left[\frac{\psi_1\hat{\psi}_1 u''(w_1)}{[u'(w_1)]^2} - \sum_{i=2}^{n}\frac{x_{ji}\hat{\psi}_1 u''(w_i)}{[u'(w_i)]^2}\cdot\frac{u'(w_1)}{u'(w_i)}\right]^{-1}$$

$$\left[\widehat{\psi}(t)\sum_{i=1}^{n}x_{ji}\lambda F_i'(E_i)J_i+\frac{\psi_1\dot{\widehat{\psi}}}{u'(w_1)}-\sum_{i=2}^{n}\frac{x_{ji}\dot{\widehat{\psi}}}{u'(w_i)}+\frac{\dot{\psi}_1\widehat{\psi}_1}{u'(w_1)}-\sum_{i=2}^{\mathrm{n}}\frac{x_{ji}\widehat{\psi}_1}{u'(w_i)}\right]$$

$$w_i=u^{-1}(u(w_{i-1})+\alpha_1\Delta),\qquad \text{and}$$

$$\dot{x}_i=x_{i-1}\mu+x_{i+1}\eta-x_i[r+\rho+\delta+\lambda\big(1-F_i(E_i)\big)+\mu+\eta]$$

subject to the boundary conditions:

$$\lim_{t\to\infty}\{J_i,E_i,w_i,x_i\}_{i=1}^{n}=\{\bar{J}_i,\overline{E}_i,\overline{w}_i,0\}_{i=1}^{n}.$$

The general optimal wage-tenure contract $m(\neq \underline{m})$ of other recruiting firms is now considered. The optimal paths represented by the system of differential equations and their boundary conditions in Lemma 2 are uniquely determined by the baseline salary scale borrowed from Burdett and Coles (2003), which is important because it can be extended to prescribe the wage schedules offered by all firms in a steady state. That the initial value $E_1\big(0;\ \underline{m}\big)$ establishes only a starting point means that firms that post $E_1(0;\ m)>E_1\big(0;\ \underline{m}\big)$ move along the same path but from different starting points. That is, recruiting firms choose different starting points on the baseline salary scale,

$$E_i(0;\ m)=E_i\big(t;\ \underline{m}\big)=E_i(t),J_i(0,m)=J_i\big(t;\ \underline{m}\big)=J_i(t),\text{and}$$

$$w_i(0;\ m)=w_i\big(t;\ \underline{m}\big)=w_i(t).$$

Given the baseline property, $F\colon[0,\infty]\to[0,1]$ is defined as the distribution of starting points on the baseline. Let $G_i(t)$ be the proportion of y_i-type employed workers that

receives less than $E_i\big(t;\underline{m}\big)$. Then,

$$F_i(E_i(0;\ m))=F_i(E_i\big(t;\ \underline{m}\big)=F(t)\ \text{ and } G(i,E_i(0;\ m))$$
$$=G(i,E_i\big(t;\ \underline{m}\big))=G_i(t).$$

The y_i-type employed workers retire, are laid off, and accumulate or lose human capital at rate $\rho + \delta + \mu + \eta$ such that the outflow from $\overline{G_i}$ is $(\rho + \delta + \mu + \eta)\overline{G_i}$, where $\overline{G_i} = G_i(\infty)$, y_i-type unemployed workers get jobs at rate λ_u, and y_{i-1}-type and y_{i+1}-type employed workers enter into $\overline{G_i}$ at rate μ and η, respectively. The outflow and inflow at steady state are equated to solve for $\overline{G_i}$ and u_i, and, by the same reasoning, for $\dot{G}_i(t)$. The baseline property presents steady state $\{(u_i, G_i\}_{i=1}^n$ in the following.

Lemma 3 *In the steady state equilibrium*,

$$\dot{G}_i(t) = \lambda_u F(t) u_i + \mu G_{i-1}(\mathrm{t}) + \eta G_{i+1}(\mathrm{t}) - (\rho + \delta + \lambda\big(1 - \mathrm{F(t)}\big) + \mu + \eta) G_i(\mathrm{t}),$$

where

$$u_1 = \frac{\delta \overline{G}_1 + \rho + \eta_2 u_2}{\lambda_u + \rho},$$

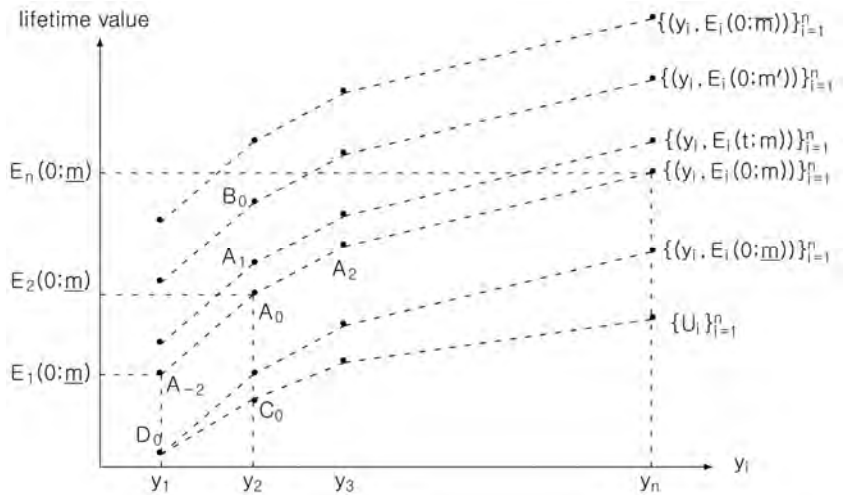
$$u_i = \frac{\delta \overline{G}_i + \eta u_{i+1}}{\lambda_u + \rho + \eta}, i = 2,3, \cdots, n \text{ and}$$

$$\overline{G_i} = \frac{u_i \lambda_u + \mu \overline{G}_{i-1} + \eta G_{i+1}}{\delta + \rho + \mu + \eta}, i = 1,2, \cdots, n.$$

[Figure 2] summarizes Lemma 1 through Lemma 3. A y_2-type employed worker under contract m, denoted by A_0 in the figure, who switches to a better paying job with contract m' jumps to point B_0 on his value ladder. If he is able to remain there without any shocks and is gradually promoted, he will be found, after some time, at A_1. Accumulating (or losing) Δ units of human capital will result in his moving to point $A_2(A_{-2})$ on a neighboring ladder. Becoming unemployed will find him moved down to C_0. Upon retiring, he is replaced by a newly-born

worker whose career starts at point D_0. Lemma 1 and 2 determine the gains from human capital accumulation and speed of non-productive promotion, respectively. Lemma 3 presents the earnings distribution over the type-value space.

Figure 2



Proposition 1 *The sufficient conditions for a market equilibrium are as follow.*

(*i*) *Given F, Lemma 1-3 are satisfied.*

(*ii*) *F is stationary and satisfies the equilibrium restriction*.

Because providing a fixed point algorithm to find a market equilibrium does not guarantee the existence of an equilibrium, and it is unclear, moreover, when the least cost incentive compatibility constraints are binding, in lieu of a theoretical proof, the model is solved numerically and whether the implied equilibrium outcome satisfies the sufficient condition is checked. For a broad range of model parameters, a unique fixed point is obtained with all constraints binding. In particular, when μ is small such that a relatively large mass is of the y_i-type, the constraints are binding.[11]

11 The estimates yield $\mu = 0.023$.

3. Data

The source of the data used in the study is the 1979 National Longitudinal Survey of Youth (NLSY79), which contains weekly work records from 1978 through 2010. The model implies that workers' wages vary with their (unobserved) productivity and (observed) job tenure. NLSY79 is well suited to the analysis of careers (and human capital accumulation) because it reports weekly labor force status from the high school period, which enables individual workers' entire work histories, from their first jobs, to be examined. The manner of constructing the sample such that five jobs are tracked in each survey round is detailed in Appendix B. The focus here is mainly on defining the adopted variables.

Workers The survey begins with individuals 14-22 years old in 1979. The sample extracts from these individuals white male high school graduates, the largest demographic group in NLSY79, who completed 12th grade or received the equivalent degree (GED) between the ages of 17 and 20 after 1978, and reported no more than 12 years of education until the most recent survey. The age and year restrictions were imposed to exclude individuals with unusual or hidden experience. Receiving a high school diploma earlier than age 17, later than age 20, or before 1978 is presumed to reflect such experience, which is not captured in the survey. By similar reasoning, individuals who entered military service were excluded. Following this selection rule yields a sample of 776 individuals.

Full-time Employment and Non-employment Full-time employment is understood to mean working, on average, more than 30 hours per week. Calculating average hours worked as the weighted (by the number of weeks) mean of hours worked at the match (with the same employer) throughout a worker's career precludes a transition between part-time and full-time work with the same employer, but includes the case of working less than 30 hours per week for a short period as a trainee or intern and more than 30 hours for a long period as a regular worker. Periods reported as out of the labor force, no information,

unemployment, and employment with average hours worked less than 30 hours per week are recoded as non-employment, the counterpart of unemployment in the model. When hours worked per week is unavailable, hours worked is calculated as 'hours worked per day' times five working days per week.

Labor Market Entry Following Farber and Gibbons (1996) and Yamaguchi (2009), a worker employed full-time for more than half of three consecutive years for the first time is assumed to have made the transition from school to work. Each individual's work history is tracked from his first transition. By construction, all workers in the sample begin their careers as employed workers. To mitigate any risk of potential bias, the first unemployment period before the first job is ignored.

Tenures Job tenure is defined as the length of a continuous period of work with one employer, employment tenure as the duration of consecutive job spells without non-employment. The source of the difference is job-to-job transition-switching to a new from an old job resets job, but not employment, tenure-but how to determine job-to-job transition in the sample is not clear. To accommodate instances like brief vacations interposed between jobs, non-employment spells of less than three weeks duration, being assumed to more likely be an outcome of on-the-job search than an employment-unemployment-employment transition, are discarded. This selection integrates into subsequent job tenure 1,702 instances of short term non-employment.

Although the model implies that there are neither recalled jobs nor returned workers, NLSY79 includes numerous instances of workers who returned to jobs they left for varying periods. In instances like unpaid vacations or hospitalization, in which the absence is planned in advance by both parties and the worker's return considered in the previous labor contract, the former and recalled job are considered one job, and the new contract, post return, is affected by the previous contract. Instances of unplanned absence are considered two different jobs because the fact of the worker's return affects neither the previous nor the new labor contract. As in Pavan (2008), an intermediate period

of sufficiently short duration is naturally presumed to be more likely to have been planned. In the sample, if a worker returned to a previous job within one quarter, the intermediate work history is dropped and the two jobs are considered one continued job. Otherwise, they are considered two different jobs. This provision results in 923 short-term (less than one quarter) non-employment periods, and 84 temporary jobs being classified 'planned return' and dropped. In 555 cases, workers returned to an old job after one quarter. The final sample contains 4,325 employer-employee matches[12] and 4,880 jobs.

Experiences Because the model assumes human capital to be accumulated only on the job, it is necessary to distinguish worker from market experience. Hence, worker experience is defined as the sum of all employment spells, market experience calculated by subtracting age at entry from a worker's current age. In the NLSY79 data set, wages are reported on the interview date and, if it ended, on the job's end date. Beginning with the 1985 survey, the first wage on the job was solicited. Because first wage data might be biased by the reporting of the first wage for jobs started before 1985 and kept until the 1985 survey, the simulation uses the first wage only for jobs started after or continued until the 1985 survey. For the first wages reported, the re-employment wage, a key variable for estimating the effect of human capital accumulation, is defined as the first wage after non-employment. The sample of 13,735 wage observations includes some with potential coding errors.[13] Because of the difficulty of fitting all data points (especially at the ends) using a simple model, the top and bottom 2.5% of wage observations were discarded and the remaining 95% made the focus of analysis.

The final sample contains 665 individuals, 4,325 employer-employee matches, 4,880 jobs, and 14,298 observations. Construction of the data set is detailed in the Appendix.

---

12 Because NLSY79 does not distinguish 'job' from 'employer-employee match', all returning cases are considered one job.

13 For example, $0.03 per hour was the lowest, $862.69 per hour the highest, wage reported (after adjustment by monthly CPI).

4. Estimation

4.1. Estimation Procedure

Maximum likelihood inference not being numerically feasible, indirect inference was used, as in Bagger, Fontaine, Postel-Vinay, and Robin (2006). Indirect inference requires[14] that the structural model replicate the true data generating process in terms of specific target moments given a true value of the structural parameter vector θ_0. Denote as $g(\theta)$ the vector of the target moments simulated by the parameter vector θ, which is estimated by minimizing the distance between the set of sample moments from NLSY79 and set of moments from the simulations. The moment vector is simulated and calculated k times and the average taken. The simulated moments estimator of θ_0 is defined as

$$\hat{\theta} = \arg\min_{\theta}(\overline{g}_k(\theta) - g(\theta_0))^T \hat{w}_n (\overline{g}_k(\theta) - g(\theta_0)),$$

where $\hat{w}_n$ is a positive definite matrix that converges in probability to a deterministic positive definite matrix W. The covariance matrix of the auxiliary statistics is estimated by re-sampling 665 individuals with replacement 1,000 times (n=1000) and taking the inverse. The entire wage and employment history of a selected individual i is included in the sample. For each set of simulated moments, the simulation is repeated 200 times and the average of the moments from each simulation (k=200) taken.

In the absence of any theoretical evidence of the uniqueness of the minimum value, the objective function is minimized using both the Nelder-Mead and simulated annealing algorithms. The Nelder-Mead method is used repeatedly. When the distance reaches a local minimum, the size of the simplex is reset and the algorithm restarted from the local minimum. If the program stops at a point sufficiently close to the local minimum, the simulated annealing method is invoked. Although the

14 For details on 'indirect inference', see Gourieroux, Monfort, and Renault (1993) and Gourieroux and Monfort (1997).

latter involves heavy computation, the probability of reaching a global minimum can be increased by applying the simulated annealing method repeatedly. In the present paper, the process is repeated with four different starting points. Obtaining the same estimates for the structural parameters is taken to be a global minimizer.

4.2. Estimation Specification

CARA (exponential) utility with risk aversion parameter γ is assumed for the empirical implementation.

$$u(w) = -\exp(-\gamma w)$$

Normalizing $y_{nj} = 1.0$, the most productive worker produces one unit of output; $y_1 = 0.4$ and $\Delta = 0.1$ are then set so as to have seven types of workers ($n_j = 7$).[15] The latter choice is arbitrary, but absent output data it is difficult to obtain inference from these parameters. The number of equilibrium contracts is fixed at 20 levels[16] and $s = 0.01$. The highest being nearly eight times the lowest wage in the sample, the highest wage is eight or nine times greater than the lowest wage depending on parameter values. The interest rate *r* is fixed at 0.012.

The short history of the NLSY79 data set makes it is difficult to estimate the arrival rate of the retirement shock ρ. Assuming the average worker to remain in the labor market for 40 years fixes ρ at 1/160. To match the actual survival probability, an 'attrition probability' of 2.5%, introduced in each survey round, accounts for the survey's loss, over time, of some workers who would otherwise remain in the labor market.

4.3. Estimation

Seven structural parameters are estimated: four Poisson arrival rates $(\delta, \lambda_u, \lambda, \mu)$, the risk aversion parameter γ, the unemployment benefit b (or w_{max}), and cost function parameter c_1.

---

15 Level of human capital is discretized into seven levels.

16 The cases with 19 and 21 equilibrium contracts are reported later.

The average nonemployment spell, average job spell, and average length of unemployment over the first five years are used to capture the dynamic flow of workers. The model implies that the rate at which workers are promoted increases with the accumulation of human capital, and that job turnover is more likely among young workers with less human capital. Examining the sample's total non-employment (or employment) period over the first five years reveals an average unemployment duration of 0.471 years, job spell of 2.175 years, and that the average worker keeps a full time job during 88.3% of the first five years.

One of the main objectives of the present empirical study being to estimate the effect of human capital accumulation independently of the effect of strategic promotion and job turnover, the log re-employment wage($\widehat{w}$), re-employment wage being defined as the first wage after unemployment, is regressed on work experience,

$$\widehat{w}_k = \beta_0 + \beta_1 \times work\ experience_k + \varepsilon_k,$$

where ε_k is a statistical residual. Adopting β_1 as the auxiliary moment captures wage growth attributable to the accumulation of work experience. The regression coefficient being insufficient to identify the frequency and magnitude of the human capital accumulation shock, human capital accumulation in the model is determined to occur at rate μ and increase workers' wages by a certain amount affected by c_1. Information on the re-employment wage distribution is exploited to capture the frequency and magnitude of individual shocks. The ratios of the 3rd to the 1st and 2nd to the 1st quartile of the distribution are calculated. The auxiliary regression indicates a coefficient β_1 of 0.109, and the quartile ratios are 1.775 and 1.281, respectively.

The slope of the wage-tenure profile is captured by regressing wages reported in the first five years ($\widetilde{w}$) on market experience,

$$\widehat{w}_k = \alpha_0 + \alpha_1 \times market\ experience_k + u_k,$$

where α_1 is adopted as one auxiliary moment. The focus on wages reported in the first five years makes it more likely that promotion rates

are dependent on the level of human capital, and the focus on a narrow and identical group serves to more accurately capture the slope. In the sample, $\alpha = 0.052$.

Finally, to capture overall wage growth (or wage-age profile), I add some additional auxiliary moments. Denote by w_1 the first wage reported within the first 6 months after the transition to work. Also denote by w_5, w_{10}, and w_{20} the average of wages reported first after 5 years, 10 years, and 20 years of market experience, respectively. I take the ratios $w_5/w_1, w_{10}/w_1$, and w_{20}/w_1, which are 1.430, 1.616, and 1.881, respectively. The auxiliary moments from the sample and the bootstrapping standard errors are summarized in the second column of [Table 1]. It also reports the estimates of corresponding moments from the simulation based on estimates of structural parameters.[17]

Overall wage growth (or wage-age profile) is captured by adding some auxiliary moments. Denote as w_1 the first wage reported within the first six months after transitioning to work. Denote as w_5, w_{10}, and w_{20}, respectively, the average of wages reported first after 5, 10, and 20 years of market experience. The ratios $w_5/w_1, w_{10}/w_1$, and w_{20}/w_1 are 1.430, 1.616, and 1.881, respectively. The auxiliary moments from the sample and bootstrapping standard errors are summarized in the second

Table 1 Auxiliary Moments

	sample moment	simulated moment
average unemployment duration (yr)	0.471 (0.013)	0.467 (0.052)
average job duration (yr)	2.175 (0.044)	2.185 (0.458)
average unemployment periods in the first 5years	0.117 (0.004)	0.123 (0.025)
$\Delta log(\tilde{w}) / \Delta work\ experience$	0.023 (0.002)	0.025 (0.020)
3rd/1st quartile ratio of reemployment wage dist.	1.775 (0.031)	1.774 (0.467)
2nd/1st quartile ratio of reemployment wage dist.	1.281 (0.019)	1.282 (0.091)
$\Delta log(w) / \Delta market\ experience$	0.052 (0.004)	0.054 (0.034)
w_{20}/w_1	1.881 (0.042)	1.913 (0.361)
w_{10}/w_1	1.616 (0.033)	1.645 (1.001)
w_5/w_1	1.430 (0.026)	1.478 (0.070)

Notes: Standard errors in the second column are estimated using bootstrap. The asymptotic standard errors of the estimated moments are reported in parentheses in the third column.

17 The asymptotic standard errors will be reported soon.

column of [Table 1], which also reports the estimates of corresponding moments from the simulation based on estimates of the structural parameters.

The second column of Table 1 shows auxiliary moments calculated from NLSY79. The numbers in parentheses are the bootstrapping standard errors used to estimate the weight matrix. The third column provides the estimates of the moments from simulation. [Table 2] reports the estimates of the structural parameters.

4.4. Counterfactual Analysis

A counterfactual experiment is performed to explore how human capital accumulation contributes to wage growth. The experiment is designed to show how much would be earned by a representative worker unable to accumulate any human capital. It is necessary to keep all players' strategies unchanged. As before, firms optimally choose

Table 2 Parameter Estimation

Parameter	Estimates
δ (separation shock)	0.091
λ_u (offer finding rate by unemployed workers)	0.580
λ (offer finding rate by employed workers)	0.446
μ (human capital accumulation shock)	0.023
b (unemployment benefit)	0.413
c_1 (cost parameter)	0.302
γ (risk aversion parameter)	0.450

their strategies assuming that workers stochastically accumulate human capital. But here it is further assumed that a worker remains in the same state when hit by the human capital accumulation shock. The experiment is repeated with 665 workers and an artificial data set is constructed.

[Table 3], which compares the average wage growth of the two groups, shows the average wage to grow by 43%, 61.4%, and 88.1% in the first 5, 10, and 20 years, respectively, and in the present estimation,

Table 3 Counter Factual Analysis

	w_5/w_1	w_{10}/w_1	w_{20}/w_1
data	1.430	1.614	1.881
estimation with human capital accumulation	1.478	1.645	1.913
without human capital accumulation	1.354	1.416	1.418

by 47.8%, 64.5%, and 91.3%, respectively. In the absence of human capital accumulation, the average wage grows partly due to non-productive promotion and partly due to job turnover; the growth rates without productive promotion are 35%, 41.6%, and 41.8% in the first 5, 10, and 20 years, respectively.

The finding that wages grow by 41.8% without human capital accumulation seems surprising, and may be construed to contradict the conclusions of Altonji and Shakotko (1987) and Altonji and Williams (2005), who show 'returns to job tenure' to account for, at most, 11%, and 'returns to experience' for the preponderance, of wage growth, leading them to conclude that human capital accumulation accounts for most wage growth. The source of this seemingly different result is different definitions. The aforementioned authors define the job tenure effect as the wage loss that would be incurred were a worker to move to a new job, with the same values for the error components; they interpret all partial effects of market experience as general human capital accumulation effects. In the model developed here, wage growth with one employer, through both productive and non-productive promotion, is transferred to the next job through the reservation value. Strictly applying the earlier definition to the present model, 'returns to job tenure' is zero because workers lose nothing in job-to-job transition. Instead, the effect of human capital accumulation is overestimated because 'returns to experience' also includes wage growth through non-productive promotion.

That wage growth through non-productive promotion and job-to-job transition exhibits a concave pattern is also of interest. Wages grow by 35% without human capital accumulation in the first five years, after which the growth rate moderates. When back-loading of wage payments

is allowed, a recruiting firm has an incentive to post a contract with low contingent values and promote a recruited worker to a higher valued contract later. The longer a worker stays, the higher the wage increases and the more the job turnover and promotion rates decline. This shows faster wage growth in early periods to be explained by the strategic back-loading scheme as well as by a concave learning curve.

5. Conclusion

This paper develops and estimates an equilibrium job search model with unobserved human capital. An equilibrium is built with multiple wage-ladders that a worker can climb or navigate among. A worker climbs gradually through non-productive promotion, jumps to a higher rung through job-to-job transition, or, upon accumulating human capital, switches to a higher-valued ladder through productive promotion. In the empirical study, the model is estimated using indirect inference and the effect of human capital accumulation captured using the re-employment wage after unemployment. A counterfactual experiment, conducted after estimating the model, finds that the wage of a typical worker unable to accumulate human capital would grow by 41.8%.

The sample being composed entirely of white male high school graduates, who can hardly be expected to be homogeneous, an attempt is made to add ex ante heterogeneity on the worker side to the framework developed in the paper. The model proposes that lifetime value is a function not of worker type, but of what a worker actually produces. Adding ex ante heterogeneity at a worker's level of human capital thus does not require any additional state variables. But more careful attention to the sufficient condition is warranted.

References

Altonji, J., and R. Shakotko (1987): "DoWages Rise with Job Seniority?," *Review of Economic Studies*, 54(3), 437-459. 3, 17.

Altonji, J., and N. Williams (2005): "Do Wages Rise with Job Seniority? A Reassessment," *Industrial and Labor Relations Review*, 58(3), 370-397. 3, 17.

Bagger, J., F. Fontaine, F. Postel-Vinay, and J. Robin (2006): "A Feasible Equilibrium Search Model of Individual Wage Dynamics with Experience Accumulation," Manuscript, December 2006. 3, 14.

Burdett, K., C. Carrillo-Tudela, and M. G. Coles (2009): "Human Capital Accumulation and Labor Market Equilibrium," Manuscript, June 2009. 2, 3, 5.

Burdett, K., and M. Coles (2003): "Equilibrium Wage-Tenure Contracts," *Econometrica*, 71(5), 1377-1404. 2, 10.

Burdett, K., and D. T. Mortensen (1998): "Wage Differentials, Equilibrium Size, and Unemployment," *International Economic Review*, 39(2), 257-273. 2.

Farber, H., and R. Gibbons (1996): "Learning and Wage Dynamics," *Quarterly Journal of Economics*, 111(4), 1007-1047. 12, 24.

Gourieroux, C., and A. Monfort (1997): *Simulation-Based Econometric Methods*. Oxford University Press, New York, NY. 14.

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85-S118. 14.

Pavan, R. (2008): "A flexible Model of Individual Wage Dynamics and Job Mobility Outcomes," August 2008, University of Rochester. 13.

Postel-Vinay, F., and J.-M. Robin (2002): "The Distribution of Earnings in an Equilibrium Search Model with State-Dependent Offers and Counter-Offers," *International Economic Review*. 3.

Shimer, R. (2006): "On-the-job Search and Strategic Bargaining," *European Economics Review*, 50(4), 811-830. 4.

Stevens, M. (2004): "Wage-Tenure Contracts in a Frictional Labor Market: Firms' Strategies for Recruitment and Retention," *Review of Economic Studies*, 71(2), 535-551. 2.

Topel, R., and M. Ward (1992): “Job Mobility and the Careers of Young Men,” *Quarterly Journal of Economics*, 107(2), 439-479. 2.

Yamaguchi, S. (2009): “Job Search, Bargaining, and Wage Dynamics,” (2007-03). 12, 24.

Appendix

A. Mathematical Appendix

Proof of Lemma 1 i) First, suppose that contract m is least cost incentive compatible. We want to show that condition (6) should be true for all types. Since contract m is least cost incentive compatible, it should be least cost incentive compatible for each type y_i. Consider the case of $i = 2$ first. There should $\hat{y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)$ with $\hat{y}_2 : [0, \infty] \rightarrow \mathcal{Y} \cap \{y_2\}^c$ and $\hat{y}_i : [0, \infty] \rightarrow \mathcal{Y}$ for each $i \neq 2$, such that $E_2(\cdot; \hat{y}, m) = E_2(\cdot; m)$. At every $t \in \{0, \infty)$,

$$
\begin{aligned}
&E_2(t; m) - E_2(t; \hat{y}_1, m) \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)} [u(w(s, y_2; m)) - c_2(y_2) + z_2(s; m)] ds \\
&- \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)} [u(w(s, y_1; m)) - c_2(y_1) \\
&\qquad + z_2(s; m)] ds \qquad \text{(A1)} \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)} [u\big(w(s, y_2; m)\big) - c_2(y_2) + u\big(w(s, y_1; m)\big) \\
&\qquad + c_2(y_1)] ds \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)} [u\big(w(s, y_2; m)\big) - u\big(w(s, y_1; m)\big) \\
&\qquad - \alpha_1(y_2 - y_1)] ds = 0,
\end{aligned}
$$

which implies that

$$
u\big(w(\cdot, y_2; m)\big) = u(w(\cdot, y_1; m) + \alpha_1(y_2 - y_1). \qquad \text{(A2)}
$$

Now, we want to show that once condition (6) is true for $i \leq k$, it is also true for $i = k + 1$. Assume that condition (6) is true for all $i = 1,2, \cdots, k$.

That is,

$$u\big(w(\cdot, y_1; m)\big) = u(w(\cdot, y_1; m) + \alpha_1(y_i - y_1) \text{ for any } i = 1,2,\cdots,k. \tag{A3}$$

Let $i = k+1$. Since contract m is least cost incentive compatible for the y_i-type, there exists at least one $\hat{y}' : [0,\infty] \to \mathcal{Y}^n$ with $\hat{y}'_{k+1} : [0,\infty] \to \mathcal{Y} \cap \{y_{k+1}\}^c$ and $\hat{y}'_i : [0,\infty] \to \mathcal{Y}$ for each $i \neq k+1$, such that $E_i(\cdot\,; m) = E_i(\cdot; \hat{y}', m)$. Difine set $T_j \subset \{0,\infty)$ such that if $\hat{y}'_{k+1}(t) = y_j$, then $t \in T_j$, where $j = 1,2,\cdots,k$. Then, for every $t \in \{0,\infty)$,

$$\begin{aligned}
&E_i(t; m) = E_i(t; \hat{y}', m) \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s, y_i; m)) - c_i(y_i) + z_i(s; m)]ds \\
&- \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s, \hat{y}'_i(s); m)) - c_i(\hat{y}'_i(s)) + z_i(s; m)]\,ds \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s, y_i; m)) - c_i(y_i) - u(w(s, \hat{y}'_i(s); m)) \\
&\qquad + c_i(\hat{y}'_i(s))]ds \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s, y_i; m)) - u\big(w(s, \hat{y}'_i(s); m)\big) \\
&\qquad - \alpha_1(y_i - \hat{y}'_i(s))]ds \\
&= \sum_{j=1}^k \int_{T_j \cap [t,\infty]} e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}\,[u(w(s, y_i; m)) - u(w(s, y_j; m)) \\
&\qquad -\alpha_1(y_i - y_j)]ds \\
&= \sum_{j=1}^k \int_{T_j \cap [t,\infty]} e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}\,[u(w(s, y_i; m)) - u(w(s, y_1; m)) \\
&\qquad -\alpha_1(y_i - y_1)]ds \\
&= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}\,[u(w(s, y_i; m)) - u(w(s, y_1; m)) \\
&\qquad -\alpha_1(y_i - y_1)]ds = 0.
\end{aligned}$$

Thus, we obtain that when $i = k+1$,

$$u\big(w(t,y_i;m)\big) = u(w(t,y_1;m)) + \alpha_1(y_i - y_1) \text{ at any } t \in [0,\infty). \text{ (A4)}$$

The mathematical induction yields condition (6).

ii) Conversely, I want to show that if condition (6) holds under contract m, it should be least cost incentive compatible. It is sufficient to show that the y_i-type worker has no incentive to deviate. Consider an arbitrary downward deviation with

$$\phi''(t) = \begin{cases} y_1 & \text{if} \quad t \in T_{i1} \\ \vdots & \vdots \quad\quad \vdots \\ y_{i-1} & \text{if} \quad t \in T_{ii-1} \end{cases} \tag{A5}$$

and $\bigcup_{j=1}^{i-1} T_{ij} = [0,\infty)$. For any $t \in [0,\infty)$ and $i \in \{1,2,\cdots,n\}$,

$$\begin{aligned} & E_i(t;m) - E_i(t;\phi'',m) \\ &= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s,y_i;m)) - c_i(y_i) + z_i(s;m)]ds \qquad \text{(A6)} \\ &\quad - \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s,\phi_i''(s);m)) - c_i(\phi_i''(s)y) + z_i(s;m)]ds \\ &= \int_t^\infty e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}[u(w(s,y_i;m)) - c_i(y_i) - u(w(s,\phi_i''(s);m)) \\ &\qquad\qquad + c_i(\phi_i''(s))]ds \\ &= \sum_{j=1}^{i} \int_{T_{ij}} e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}\,[u(w(s,y_i;m)) - c_i(y_i) - u(w(s,y_j;m)) \\ &\qquad\qquad +c_i(y_j)]ds \\ &= \sum_{j=1}^{i-1} \int_{T_{ij}} e^{-(r+\rho+\delta+\lambda+\mu+\eta)(s-t)}\,[u(w(s,y_i;m)) - u(w(s,y_1;m)) \\ &\qquad\qquad -\alpha_1(y_i - y_1)]ds = 0. \end{aligned}$$

There is no profitable deviation and there exists $\phi'': [0,\infty] \to \mathcal{Y} \cap \{y_i\}^c$ such that $E_i(\cdot;m) = E_i(\cdot;\phi'',m)$. Since this is true for all $i = 1,2,\cdots,n$, contract m is least cost incentive compatible. *Q.E.D*

Proof of Lemma 2 Rewrite the operating firm's problem:

$$\max_{w_1(\cdot)} \int_0^{\infty} \psi_1[y_1 - w_1 + \mu J_2]\, ds$$

$$s.t.\ \ \dot{\psi}_1 = -[r + \rho + \delta + \lambda(1 - F_1(E_1)) + \mu]\psi_1$$

$$\dot{E}_i = -u(w_i) + \alpha_0 + (r + \rho + \delta + \lambda\big(1 - F_i(E_i)\big) + \mu + \eta)$$

$$E_i - z_i^*(E_i, m)$$

$$\dot{J}_i = -y_i + w_i + \big(r + \rho + \delta + \lambda\big(1 - F_i(E_i)\big) + \mu + \eta\big]J_i - \overline{\varphi}_i(m)$$

$$u(w_i) = u(w_{i-1}) + \Delta = u(w_1) + (i-1)\Delta$$

The Hamiltonian of the problem is

$$\mathcal{H} = \psi_1[y_1 - w_1 + \mu J_2] - x_\psi\big(r + \rho + \delta + \lambda\big(1 - F_i(E_1)\big) + \mu\big]\psi_1$$
$$+ \sum_{i=1}^{n} x_{ei}\,[-u(w_1) + \alpha_0 + \alpha_1(y_i - y_1) + (r + \rho + \delta + \lambda$$
$$(1 - F_i(E_i)) + \mu + \eta)E_i - \varphi_i(E_i; m)]$$
$$+ \sum_{i=2}^{n} x_{ji}\,[-y_i + u^{-1}(u(w_1) + \alpha_1(y_i - y_1)) + (r + \rho + \delta + \lambda$$
$$(1 - F_i(E_i)) + \mu + \eta)J_i - \tilde{\varphi}_i(m)]$$

Applying the maximum principle yields the following first order condition and differential equations.

$$0 = -\psi_1 - \sum_{i=1}^{n} x_{ei}\, u'(w_1) + \sum_{i=2}^{n} \frac{x_{ji} u'(w_1)}{u'(u^{-1}(u(w_1) + (i-1)\Delta))} \quad \text{(A7)}$$

$$\dot{x}_\psi = -[y_1 - w_1 + \mu J_2] + x_\psi\big[r + \rho + \delta + \lambda\big(1 - F_1(E_1)\big) + \mu\big] \quad \text{(A8)}$$

$$\dot{x}_{ji'} = x_{ji'} - 1\mu + x_{ji'} + 1\eta$$
$$- x_{ji'}\big[r + \rho + \delta + \lambda\big(1 - F_{i'}(E_{i'})\big) + \mu + \eta\big] \quad \text{(A9)}$$

$$\dot{x}_{ei} = x_{ei-1}\mu + x_{ei+1}\eta - x_{ei}\big[r + \rho + \delta + \lambda\big(1 - F_i(E_i)\big) + \mu + \eta\big]$$
$$+ x_{ji}\lambda F_i'(Fi)J_i \quad \text{(A10)}$$

where $x_{j1}(t) = -\psi(t), i' = 2,3,\cdots,n$ and $i = 1,2,\cdots,n$. From (A8), I obtain

$$\dot{x}_{\psi}\psi_1(t) + x_{\psi}\dot{\psi}_1(t) = -[y_1 - w_1 + \mu J_2]\psi_1(t)$$

$$\Leftrightarrow x_{\psi} = \int_t^{\infty} [y_1 - w_1 + \mu J_2] \frac{\psi_1(0,\tau)}{\psi_1(0,t)} d\tau + A_{\psi}\psi_1^{-1}(t)$$

$$= \int_t^{\infty} \psi_1(t,\tau)[y_1 - w_1 + \mu J_2]\, d\tau = J_1.$$

Plugging (A11) into (A8) yields

$$\dot{J}_1 = -[y_1 - w_1 + \mu J_2] + J_1\big[r + \rho + \delta + \lambda\big(1 - F_1(E_1)\big) + \mu\big]. \qquad \text{(A11)}$$

Summing up all equations in (A10) and reordering yields

$$\sum_{i=1}^{n} \dot{x}_{ei} = \sum_{i=1}^{n} x_{ji}\lambda F_i'(E_i)J_i - \Big(r + \rho + \delta + \lambda\big(1 - F_1(E_1)\big)\Big)\sum_{i=1}^{n} x_{ei}.$$

Let

$$\hat{\psi}(t) = \exp\,[\int_0^{\mathrm{t}} (r + \rho + \delta + \lambda(1 - F_1\big(E_1(s;m)\big)))ds]. \qquad \text{(A12)}$$

Multiplying by integrating factor on both sides of equation (A12) yields

$$\hat{\psi}(t) = \sum_{i=1}^{n} \dot{x}_{ei} + \Big(r + \rho + \delta + \lambda\big(1 - F_1(E_1)\big)\Big)\hat{\psi}(t)$$

$$\sum_{i=1}^{n} x_{ei} = \hat{\psi}(t)\sum_{i=1}^{n} x_{ji}\,\lambda F_i'(E_i)J_i. \qquad \text{(A13)}$$

Then, multiplying by on both sides of equation (A7), dividing by $u'(w_1)$, taking derivative with respect to *t*, and combining with equation (A13) yields

$$\hat{\psi}(t) = \sum_{i=1}^{n} x_{ji}\lambda F_i'(E_i)J_i$$
$$= \left[-\frac{\psi_1}{u'(w_1)} + \sum_{i=2}^{n} \frac{x_{ji}}{u'\left(u^{-1}(u(w_1) + (i-1)\Delta)\right)}\right]\dot{\hat{\psi}}$$
$$-\frac{\dot{\psi}_1}{u'(w_1)} + \sum_{i=2}^{n} \frac{\dot{x}_{ji}}{u'(w_i)}$$
$$+\left[\frac{\psi_1\hat{\psi}_1 u''(w_1)}{[u'(w_1)]^2} - \sum_{i=2}^{n} \frac{x_{ji}\hat{\psi}_1 u''(w_i)}{[u'(w_i)]^2} \cdot \frac{u'(w_1)}{u'(w_i)}\right]\dot{w}_1$$

Rewriting this, I obtain

$$\dot{w}_1 = \left[\frac{\psi_1\hat{\psi}_1 u''(w_1)}{[u'(w_1)]^2} - \sum_{i=2}^{n} \frac{x_{ji}\hat{\psi}_1 u''(w_i)}{[u'(w_i)]^2} \cdot \frac{u'(w_1)}{u'(w_i)}\right]^{-1}$$
$$\left[\hat{\psi}(t) = \sum_{i=1}^{n} x_{ji}\lambda F_i'(E_i)J_i + \frac{\dot{\psi}_1\hat{\psi}_1}{u'(w_i)}\right.$$
$$-\sum_{i=2}^{n} \frac{x_{ji}\dot{\hat{\psi}}}{u'\left(u^{-1}(u(w_1) + (i-1)\Delta)\right)} + \frac{\dot{\psi}_1\hat{\psi}_1}{u'(w_1)}$$
$$\left. -\sum_{i=2}^{n} \frac{\dot{x}_{ji}\hat{\psi}_1}{u'(w_i)}\right]$$

It derives Lemma 2. *Q.E.D*

Proof of Lemma 3 Consider the outflow from and inflow into y_i-type unemployment for any arbitrarily small time interval $dt > 0$. By equating them, I obtain that when $i = 1$.

$$\delta dt\overline{G}_1 + \eta dt u_2 + \rho dt(1 - u_1) = (\lambda + \mu)dt u_1 \qquad \text{(A14)}$$

and when $i > 1$,

$$\delta dt\overline{G}_i + \mu dtu_{i-1} + \eta dtu_{i+1} = (\rho + \lambda + \mu + \eta)dtu_i. \quad \text{(A15)}$$

Also, consider the proportion of y_i-type employed workers. Equating the inflow and outflow yields

$$\lambda dtu_i + \mu dt\overline{G}_{i-1} + \eta dt\overline{G}_{i+1} = (\rho + \delta + \mu + \eta)dt\overline{G}_i, \quad i = 1,2,\cdots,n \quad \text{(A16)}$$

Sending $dt \to 0$ and combining (A14), (A15) and (A16) all together yields the initial values in Lemma 3. By the similar reasoning, the differential equations in Lemma 3 are obtains. *Q.E.D*

B. Construction of the Sample

Appendix B describes the steps used to construct the sample.

1. The analysis is restricted to white male high school graduates, the largest demographic group in the NLSY79 dataset. In the first survey round, 8,736 of 12,686 respondents are identified as 'white'[18] and 6,403 as 'male'. Combining these responses yields an initial sample of 4,393 white males. Of these, 1,990 individuals have completed the 12th grade or received the equivalent degree (GED) without reporting further education until the most recent (2010) survey.

2. Individuals with hidden or unusual experience are excluded by dropping 994 individuals who graduated before January 1, 1978 or before age 17 (the 204th month) or after age 20 (the 240th month). Eliminating individuals who entered military service at least once before the last survey results in another 220 individuals being dropped. These selection rules leave 776 workers and

18 The sample of 'white' is obtained using question (R01727.00) rather than question (R02147.00).

34,010 work-records in the sample.[19]

3. Full-time employment is defined as the match with weekly hours worked, on average, in excess of 30. Average hours worked is calculated as the weighted (by the number of weeks) mean of hours worked at all matches with the same employer throughout a worker's career. For weekly hours worked less than 10, the maximum number between weekly and daily hours worked is multiplied by five working days. When only daily hours worked are reported, they are multiplied by five working days. Before calculating the average, the hours worked per week are top-coded so that they cannot exceed 96.[20]

4. Following Farber and Gibbons (1996) and Yamaguchi (2009), a worker who works full-time for more than 78 weeks in three consecutive years for the first time is assumed to have made the transition from school to work. Work records before this transition are dropped. This leaves 752 workers, 5,955 full-time employers, 573 part-time employers, and 27,606 work records.

5. All non-full-time employment, such as 'out of labor force', 'no information', 'unemployment', and employment with average weekly hours worked less than 30 are recoded as non-employment, and all consecutive non-employment spells merged. This leaves 752 workers, 5,955 full-time employers, 5,918 non-employment spells, and 24,808 work records in the sample.[21]

---

19 In the non-military sample, 20 individuals reported more than one graduation date; assuming these to be coding errors, the graduation dates closest to age 18 (the 222nd month) were selected.

20 Ninety-eight cases are top-coded.

21 Although the NLSY79 does not distinguish employers from job, each is defined separately. In particular, a worker who returns to an old employer can be considered a planned return or a random re-match. Returns not planned should be considered a different job with the same employer.

6. Workers after long term non-employment being assumed no longer to be attached to the regular labor market, observations subsequent to three years of non-employment spells are dropped. This leaves 752 workers, 5,801 (full-time) employers, 5,413 non-employment spells, and 24,019 work records.

7. Quitting an old and starting a new job within a span of three weeks is considered a job-to-job transition, in which case the intermediate period in the next job spell is included. A return to an old job within 13 weeks is assumed to be a planned return and recoded as a single continuous job. This leaves the final sample of 665 workers, 4,796 jobs, and 14,298 observations.