Price and Quality Dispersion in an Offshoring Market: Evidence from Semiconductor Production Services*

David Byrne
Federal Reserve Board

Brian K. Kovak†
Carnegie Mellon University

Ryan Michaels
University of Rochester

December 2012

Abstract

As intermediate inputs are increasingly purchased from low-price markets such as China, it is essential to know whether low-price suppliers simply produce lower quality products or whether they offer real quality-adjusted discounts compared to their higher-price competitors. We study cross-country differences in price and quality in the market for semiconductor wafer manufacturing services. Using a proprietary transaction-level dataset, we document i) substantial constant-quality price differences across suppliers, and ii) shifts toward lower priced suppliers. Chinese producers on average charged 17% less than leading Taiwanese producers for otherwise identical products and increased their market share by 14.7 percentage points from 2000 to 2011. We introduce a model with costs of switching that can sustain realistic quality-adjusted price dispersion and predicts pricing dynamics that are present in the data but are inconsistent with alternate explanations. We conclude by discussing the implications of our findings for existing approaches to cross-country quality measurement and for productivity measurement when facing quality-adjusted price dispersion.

*An earlier version of this paper was prepared for the conference "Measurement Issues Arising from the Growth of Globalization," sponsored by the W.E. Upjohn Institute for Employment Research and the National Academy of Public Administration (NAPA). Thanks to Yong-Gyun Choi, Manuel Gomez, Steven Paschke, and Fabio Rueda for excellent research assistance. We would like to thank Bill Alterman, Ken Flamm, Erica Fuchs, Ross Goodman, Susan Houseman, Marshall Reinsdorf, Jodi Shelton, Sonya Wahi-Miller, Kim Zieschung, seminar participants at the Upjohn/NAPA conference, Carnegie Mellon, the U.S. Dept. of Commerce, NBER CRIW Summer Institute, and the Rocky Mountain Empirical Trade Conference for helpful discussions and Chelsea Boone at GSA, Peter Dziel at LSI, and Len Jelinek at iSuppli for help with data and background information on the industry. Remaining errors are our own. This paper reflects the views of the authors and should not be attributed to the Board of Governors of the Federal Reserve System or other members of its staff.

†Corresponding author: bkovak@cmu.edu

1 Introduction

In the last half century, reductions in transportation costs, communication costs, and barriers to trade have driven a fundamental change in the spatial and institutional organization of manufacturing production. It is now commonplace for individual design and fabrication steps in the manufacturing process to be implemented in different establishments, which are owned by different firms and located in different countries. This fragmentation of production means that a larger share of international transactions takes the form of processing trade: an input is fabricated, assembled, or processed abroad and then shipped back to the firm that designed it or elsewhere for additional processing (Hummels, Ishii and Yi 2001). As less developed economies such as China capture a greater share of the global market for tradable intermediates, the tasks performed overseas are increasingly done by relatively low-price suppliers. To determine the implications of this shift, it is essential to know whether low-price suppliers simply produce lower quality products or whether they offer real quality-adjusted discounts compared to their higher-price competitors.

We examine this question in the context of semiconductor wafer manufacturing services. Technological characteristics of production in this market make possible highly effective quality adjustment based on observable product characteristics. We utilize a novel proprietary database of semiconductor wafer transactions to document substantial shifts in production across geographic areas and show that suppliers with growing market shares typically charge much lower prices for observationally equivalent goods. Estimates from hedonic regressions reveal substantial constant-quality price differences across suppliers in different countries. For example, Chinese producers on average charged 17% less than firms in market leader Taiwan for otherwise identical products, and increased their market share from 7.1% in 2000 to 21.8% in 2011. By observing and directly controlling for quality differences across suppliers, we are able to circumvent the challenges faced by other papers in the literature. We thus follow Schott's (2008) suggestion to use "very detailed data about the hedonic attributes of goods produced and exported by China and developed economies" to reveal cross-country quality differences.

These sharp quality-adjusted price differences across suppliers raise the question of how price dispersion can persist without driving the high-priced suppliers out of the market. The contract semiconductor manufacturing industry is highly concentrated and faces very low transportation costs, so mechanisms generating quality-adjusted price variation based on search frictions or transportation costs are unlikely to apply in this context.[1] Mechanisms based on product differentiation and preference for variety are also difficult to justify when customers purchase manufacturing services for a product of their own design.[2]

[1] See, for example, Sorensen (2000) on search frictions in the prescription drugs market and Syverson (2004) on transport costs in the ready-mixed concrete market.

[2] See Hallak and Schott (2011) and Khandelwal (2010) for examples in which purchasers have, respectively, constant elasticity of substitution and nested logit preferences over differentiated products.

Instead, we develop a model reflecting the structure of the contract semiconductor manufacturing industry. In this model, sunk startup costs of producing a particular chip design with a particular supplier make it costly to switch suppliers (Klemperer 1995). The industry "leader" is the first producer of a leading-edge wafer. Lower-priced suppliers, "followers," enter the market later, but since it is costly for customers to switch suppliers, the higher-priced leader preserves a substantial market share. The high cost of switching effectively locks-in buyers, giving the industry leader the incentive to continue charging high prices compared to its competitors. The model allows for the possibility that the leader is more skilled at producing more complex wafers, but we show that the price difference across suppliers exceeds the premium implied by any quality differences across suppliers.

The model also predicts that the degree of price dispersion across suppliers diminishes over time. The buyers that initially purchase from the leader eventually complete their production and exit. At the same time, new buyers enter the market. The leader will compete more aggressively for these new buyers as their locked-in customers phase out, implying a declining price differential across suppliers over time. We stress that this pattern in the dynamics of price dispersion is clearly absent in models that include only fixed unobserved quality differences across suppliers.

We test for this declining price differential using data for producers located in Taiwan and China, which together accounted for 72% of the market in 2011. In practice, Chinese producers enter the market for a given semiconductor technology at least 2 years later than Taiwanese producers, so we treat Taiwan as the "leader" and China as the "follower." We find that on average the price gap between Taiwan and China closes substantially over the life of a given semiconductor technology, falling from 39% in the year of Chinese entry to 10% after 5 years.

Our focus on price differences across suppliers is in sharp contrast with prior theoretical and empirical studies of semiconductor prices. Although semiconductor pricing has been studied frequently, nearly all prior work examines the markets for commodity semiconductor products such as microprocessors and DRAM memory chips.[3] In these markets, there is little price variation across suppliers. Studies of processor and memory markets rule out quality-adjusted price variation and instead focus on learning-by-doing to explain stunning rates of average price decline.[4] We focus on the market for contract semiconductor wafer fabrication services provided by firms called "foundries." Due to a lack of detailed data, this market has been studied infrequently in the previous literature, and it differs in important ways from processor and memory markets.[5] Most foundry products are Application Specific Integrated Circuits (ASICs) requiring custom designs tai-

[3]For microprocessors, see Dulberger (1993), Grimm (1998), Doms, Aizcorbe and Corrado (2003), Holdway (2001), and Flamm (2007) and for memory see Flamm (1993), Grimm (1998),and Aizcorbe (2002)

[4]For models of learning by doing in semiconductor production, see Baldwin and Krugman (1988), Irwin and Klenow (1994), and Flamm (1996), all of which assume the law-of-one-price holds.

[5]Aizcorbe (2002) constructs price indexes with the limited available data for these products and argues for using a combination of prices of MPUs, memory chips, and average product prices as a proxy for the devices produced by foundries when data are not available.

lored to particular applications rather than being designed for general-purpose computing. As we show, this market exhibits substantial price variation across suppliers, which we argue is driven by product-specific fixed startup costs incurred because of the custom nature of each product. Thus, our modeling approach focuses on price variation between suppliers rather than on the sources of price declines across all suppliers, leaving for future work the task of modeling startup costs in the presence of declining marginal costs resulting from learning by doing.[6]

Our approach is more closely related to recent empirical work on firm dynamics by Foster, Haltiwanger and Syverson (2012). Using detailed plant-level data covering industries without much scope for quality differentiation, they find substantial variation across firms in quantity sold, even holding price fixed. Quantity sold increases as firms age, which the authors interpret as reflecting a dynamic process of "building a customer base," in which larger current sales increase future demand. Our framework based on costly switching across intermediate input suppliers generates a very similar demand structure; lowering the price to attract more customers in one period increases demand next period, as more customers are locked-in by the sunk startup cost. Thus, our model reflects a particular mechanism driving the more general phenomenon Foster et al. document. Their results suggest that various features of the market we discuss here may apply more broadly outside the semiconductor industry.

Our findings have important implications for at least two growing areas of research. First is the literature on trade and product quality, which seeks to understand how quality differences across suppliers affect the pattern of trade. Early papers in this literature interpreted all differences in price (unit value) across suppliers within narrow product classification as differences in quality.[7] Our results contradict this assumption in the semiconductor manufacturing industry, as all of the products in our data fall within one narrow product classification[8], and yet we observe substantial quality-adjusted price differences across suppliers. More recent work has developed tools for quality estimation in final goods markets based on product differentiation and consumers' love of variety.[9] These models allow for quality-adjusted price variation, with higher quality goods capturing larger market share conditional on price. We show that in the dynamic context of our model, both relative prices and relative market shares move in the opposite direction of relative quality across suppliers. Thus, the approaches used in the previous literature should be employed with care when studying intermediate input quality, particularly in markets with large fixed costs of producing a given product at a given supplier.

[6]Since our dynamic model relies on subgame perfect equilibria, integrating learning by doing would be substantially more complex than in the one-shot Cournot (precommitment) equilibria used in the previous work on semiconductor prices. See Fudenberg and Tirole (1983) for a discussion of precommitment vs. subgame perfect equilibria in this context and Besanko, Doraszelski, Kryukov and Satterthwaite (2010) and Besanko, Doraszelski and Kryukov (2011) for examples of subgame perfect equilibrium in models with learning by doing.

[7]See Hallak and Schott (2011) and Khandelwal (2010) for extensive lists of references.

[8]Code 8542.21.80.05 in the Harmonized Tariff Schedule of the U.S.

[9]Recent papers utilizing variations on this approach include Hallak and Schott (2011), Hummels and Klenow (2005), Khandelwal (2010), and Kugler and Verhoogen (2011).

Our results also closely relate to the emerging research on the implications of globalization for the accuracy of official aggregate price indexes and our understanding of the sources of economic growth. Feenstra, Mandel, Reinsdorf and Slaughter (forthcoming) examine three problems that bias standard import price measures upward: using a non-superlative index number formula, omitting the effect of tariffs on purchase prices, and omitting the effects of increased variety. Along with Houseman, Kurz, Lengemann and Mandel (2011), we investigate another source of bias. We argue that price measures produced by most government statistical agencies, such as the BLS International Price Program, implicitly assume that price differences across suppliers reflect quality differences rather than pure quality-adjusted price dispersion. This approach generates upward-biased measures of input price growth and productivity growth in the presence of quality-adjusted price differences and shifts to lower-priced suppliers.

We follow Reinsdorf (1993) to calculate price indexes allowing us to bound the effect of this bias. We find that conventional price measures miss shifts toward lower-priced suppliers accounting for an average annual price decline of up to 1.2 percentage points per year. More generally, as overseas production becomes increasingly sophisticated, researchers and practitioners must address quality adjustment for tradable intermediates, paralleling prior efforts relating to domestically produced products. The measurement challenges we document in the wafer fabrication market likely preview similar challenges that will emerge in other industries as further offshoring of intermediate input production occurs.

This paper proceeds by first discussing the technology of wafer fabrication. We outline the key physical attributes of wafers relevant for price setting and document the shifting geography of production away from market leading firms in Taiwan and toward more recent entrants in China. In Section 3, we investigate price differences between overseas manufacturers using a hedonic framework to control for quality differences across transactions. We find that, even after controlling for quality using our detailed transaction-level data, prices in this offshore market still differ substantially across suppliers. In Section 4 we present the model of startup costs and sequential entry explaining how large quality-adjusted price differences could persist in this market. In Section 5 we confirm the model's prediction that the price gap between leading and lagging suppliers should close over time, ruling out various alternative explanations for the price variation across suppliers. Section 6 discusses the implications of our findings for the prior literatures just discussed, and Section 7 concludes.

2 Industry Background

To understand the subsequent analysis, we need to briefly review three important features of the contract semiconductor wafer manufacturing industry. First, there are distinct, measurable technological attributes of wafers that significantly affect their price. Our data are remarkable in part

because they provide information on each of these technological characteristics, allowing us to control for quality differences in our analysis. Second, it has become more common for firms designing semiconductor chips to outsource the production of wafers to specialized manufacturing firms overseas. This business model results in the arm's-length purchases of semiconductor manufacturing services that will be the focus of our analysis and that have contributed to large shifts in the geographic distribution of wafer fabrication around the globe. Third, in contrast to general-purpose processors and memory chips that have been the focus of previous work, contract semiconductor manufacturing focuses primarily on custom designs produced in much smaller quantities. This custom aspect to production increases the importance of large fixed costs of producing a given chip at a particular foundry, which we argue drives quality-adjusted price variation across suppliers.

2.1 Semiconductor Wafer Technology

Semiconductor fabrication involves creating networks of transistors on the surface of a thin piece of semiconducting material.[10] The process begins with the design and layout of a new chip. Semiconductor designers use complex software suites to specify the functionality of the chip, convert that logic into the corresponding network of transistors, determine the physical layout of those transistors, and simulate the behavior of the proposed design for debugging purposes.

Semiconductors are manufactured in a facility called a "fab". Transistors are created on the surface of the wafer through a photolithography process, in which successive layers of conducting and insulating materials are deposited on the surface of the wafer and chemically etched away in the appropriate places to form the desired pattern of transistors and necessary interconnections. Design layout software determines the etching pattern for each layer. This is then projected onto the wafer through a mask containing the desired pattern, in a process similar to developing a photograph by projecting light through a negative. Each step of the etching process is repeated multiple times across the wafer, resulting in a grid pattern of many copies of the chip. Once all transistors and connection layers are complete, the chips are tested in a process called "wafer probe," and any faulty chips are marked to be discarded. The wafer is cut up, leaving individual chips, called "die," which are placed inside protective packages and connected to metal leads that allow the chip to be connected to other components.

Semiconductor fabrication technology has advanced over time in discrete steps, defined by wafer size and line width (also called feature size). Increases in wafer size allow larger numbers of chips to be produced on a wafer. Most fabs currently produce 150mm (roughly 6 inches), 200mm (8 inches), or 300mm (12 inches) diameter wafers. Although larger wafers cost more to produce, each wafer contains many more die, so the move to a larger wafer has generally reduced the cost per die by approximately 30 percent (Kumar 2007).

[10] Turley (2003) provides an accessible overview of semiconductor technology, manufacturing, and business.

Line width is the size of the smallest feature that can be reliably created on the wafer. Decreased line width means that individual transistors are smaller. A 30 percent decrease in line width approximately doubles the density of transistors on a chip. This makes chips of a given functionality smaller, lighter, faster, and more energy efficient, and also makes it feasible to include more functions on a single chip. The number of transistors that can be produced on a chip has grown exponentially over time, following Moore's Law.[11] Figure 1 shows the maximum number of transistors per chip and the minimum line width used to produce Intel processors over the last 40 years (both plotted on logarithmic scales).

Current line widths are measured in microns (μm) or nanometers (nm). The smallest line width currently being produced in volume is 20nm (Barak 2012). As a rule of thumb, Kumar (2007) estimates that moving a given chip design to a 30 percent smaller line width results in cost savings of approximately 40 percent, assuming the same number of defects in both processes. The primary drawback of smaller line widths is increased cost per wafer, particularly early in the technology's life span. Masks are much harder to produce when creating smaller features. In addition, new process technologies often result in higher defect rates and lower yields, the fraction of chips on a wafer that function correctly. In spite of these challenges, the benefits of increased die per wafer and better performance have outweighed the problems of decreased yields, which increase as the fabrication technology matures. Given the benefits of smaller line widths, semiconductor manufacturers have steadily moved toward newer technology. This is apparent in Figure 1 for Intel processors and can be seen even more clearly in Figure 2, which plots the technology composition of sales at Taiwan Semiconductor Manufacturing Company (TSMC), the largest contract semiconductor manufacturer.

There are a number of options regarding the chemical compounds used to create the transistors themselves and how the transistors are arranged to implement logical functions. The most common technology, called complementary metal-oxide semiconductor (CMOS), a silicon-based chemical process, accounted for 97 percent of worldwide semiconductor production in 2008.[12] We therefore restrict our analysis to CMOS and refer to each combination of wafer size and line width as a "process technology" (e.g., 200mm wafer, 180nm line width).

In order to examine constant-quality wafer price variation, we must define the set of technological characteristics that determine the quality of a given wafer. To guide this choice, we have consulted pricing models used by engineers to estimate production costs. Kumar (2008) presents a wafer cost model based on wafer size, line width, and logic family. The commercial cost estimation firm IC Knowledge distinguishes wafer cost estimates by wafer size, line width, logic family,

[11] Gordon Moore, Intel's cofounder, famously observed that the number of transistors on a chip doubled every eighteen months (Moore 1965). This regularity later slowed to doubling every two years.

[12] Share of wafer starts reported in SICAS Semiconductor International Capacity Statistics. Other transistor arrangements, such as bipolar logic, and other chemical processes, such as gallium arsenide (GaAs) or silicon germanium (SiGe), generally focus on niche markets for high-frequency, high power, or aerospace devices, rather than the storage and computational logic products comprising the majority of the CMOS market.

number of polysilicon layers, and number of metal layers.[13] All of these discrete characteristics are observable in our data, allowing us to construct credible constant-quality price measures in the analysis below.

2.2 Offshoring and the Foundry Business Model

In the early 1970s nearly all semiconductor producers were vertically integrated, with design, wafer fabrication, packaging, testing, and marketing performed within one company. Firms that perform both design and wafer fabrication are referred to as integrated device manufacturers (IDM). By the mid-1970s, IDMs began moving packaging and test operations to East Asia to take advantage of lower input costs (Scott and Angel 1988, Brown and Linden 2006). In spite of offshoring these relatively simple steps in the production process, firms maintained the more complex wafer fabrication operations in the home country.

As wafer fabrication technology advanced, however, it became prohibitively costly for younger and smaller semiconductor firms to stay at the frontier of process technology. The cost of building a fabrication facility has increased nearly 18 percent per year since 1970 and now stands at $4.2 billion (IC Knowledge 2001, Global Foundries 2009). Consequently, during the middle of the 1980s, younger and smaller firms began contracting with larger U.S. and Japanese IDMs to produce some of their more advanced designs in the latter's existing facilities (Hurtarte, Wolsheimer and Tafoya 2007). Around the same time, new contract manufacturing services firms sprang up overseas that were entirely dedicated to manufacturing wafers designed by other parties. These firms, operating principally in Asia, are known as wafer "foundries." Taking advantage of these new overseas facilities, a number of young U.S. semiconductor firms began outsourcing all of their wafer fabrication. These factoryless goods producers, which have little or no in-house wafer manufacturing capability, are called "fabless" firms. In general, fabless firms perform chip design and layout, and use foundries and other contractors for mask production, wafer fabrication, packaging, and testing.

The fabless business model has grown quickly over the last 30 years. It now accounts for about a quarter of total semiconductor industry revenue, as shown in Figure 3.[14] Some of the most prestigious U.S. chip makers, such as Fortune 500 firms Broadcom and AMD, are fabless firms.

Along with the shift from an integrated manufacturer to a foundry business model came a shift in production capacity toward Asia, where most large foundries are located. Table 1 shows how the share of worldwide foundry capacity has evolved in the last decade.[15] In 2000, the majority of foundry capacity was already in Asia, mainly Taiwan. Since then, the share of capacity in Asia as

[13] See http://www.icknowledge.com/. Polysilicon and metal layers are used in the construction of transistor "gates" and interconnections.

[14] Note that the shares in Figure 3 are likely to understate the extent of fabless production activity because it counts only companies that derive 75 percent or more of their semiconductor revenue from fabless production. Many companies not counted as fabless, such as Texas Instruments, nevertheless rely heavily on foundries.

[15] The sharp increase in European capacity from 2008 to 2010 marks the founding of Global Foundries, which was the fab division of integrated device manufacturer AMD.

a whole has only increased modestly, but there has been a notable shift in capacity within Asia. In particular, China has more than tripled its share of foundry capacity, largely at the expense of the industry leader, Taiwan. Thus, in Table 1, we get our first look at the emergence of China in the market for wafer fab services.

2.3 Foundry Production vs. Memory and Processor Markets

Economists have devoted substantial attention to studying semiconductor production in an effort to uncover the sources of rapid constant-quality price declines observed for high-tech products such as computers (Berndt and Rappaport 2001) and communications equipment (Doms 2005, Byrne and Corrado 2012). Attention has focused on the most important semiconductor components of computers, namely microprocessors (Dulberger 1993, Grimm 1998, Doms et al. 2003, Holdway 2001, Flamm 2007) and memory chips (Flamm 1993, Grimm 1998, Aizcorbe 2002).

However, general-purpose microprocessors and memory chips account for a minimal share of the market studied in this paper. Foundries instead specialize in custom chips for specific models of electronic devices such as cellular phones, hard drives, automobiles, and many others. These Application-Specific Integrated Circuits (ASICs) have been the subject of limited previous research and differ from memory and processors in important ways.[16] ASICs are produced in smaller batches, each model requires a substantial investment in design, and they are more likely to be produced using technology one generation or more behind the leading edge.[17] The most important characteristic for our analysis is the custom nature of each ASIC model. The uniqueness of each design generates substantial fixed costs of producing a given chip at a particular foundry, which we argue drives quality-adjusted price variation across suppliers (see Section 4).

3 The Distribution of Wafer Prices Across Countries

We use a new database of semiconductor wafer transactions to measure constant-quality price dispersion in semiconductor manufacturing services across supplying countries. The database provides information on all major technological characteristics that are relevant to product quality. We find that wafers produced in China sell at a 17% discount compared to otherwise identical wafers produced in Taiwan.

3.1 Wafer Price Data

Information on wafer prices comes from a proprietary database of semiconductor wafer purchases from foundries, collected by the Global Semiconductor Alliance (GSA), a nonprofit industry orga-

[16] Efforts to construct price indexes for the custom devices produced by foundries have been relatively limited, in large part due to data limitations (Aizcorbe 2002, Aizcorbe, Flamm and Khurshid 2002).

[17] Note the long persistence of technology nodes in Figure 2.

nization. The dataset consists of 6,916 individual responses to the Wafer Fabrication & Back-End Pricing Survey for 2004-2010. The survey has been conducted quarterly since 2001 and accounts for a representative sample of about 20 percent of the wafers processed by the foundry sector worldwide.[18]

The GSA dataset is unique in the amount of detail it provides for contract manufacturing of a high-technology product. For example, it includes information on the technological attributes that industry analysts and engineers report as being the key price-forming characteristics of wafers. These are logic family, line width, wafer size, and layers, as discussed in Section 2. In addition, GSA reports the location (country) of the foundry for each transaction and the price paid. This information allows us to examine how average prices vary by foundry location after controlling for all relevant technological characteristics determining the quality of the manufacturing services being purchased. An important limitation of the GSA data is the lack of firm identifiers. That is, we see information on individual transactions, including the country in which the producing foundry is located, but not the identity of the producing firm or buyer of wafer fabrication services.

As mentioned in Section 2, in an effort to focus our analysis on the ASIC market that dominates foundry production, we only analyze CMOS wafer transactions. Our sample thus omits niche markets for chips used in aerospace and high-power applications that require different production techniques. For the same reason, we also limit our sample to omit observations for products produced in countries that focus heavily on non-ASIC products: Europe, Israel, Japan, and Korea.[19] Compared to the countries in our sample, foundries producing primarily in the dropped countries derived a much larger share of their revenue from analog devices (41.5% vs. 11.0%), a larger share from discrete devices and memory products (19.0% vs. 7.7%), and a much smaller share from computational logic devices characterizing the ASIC market (35.3% vs. 70.3%).[20] The remaining countries in our sample accounted for 86.9% of foundry revenue in 2010.

Descriptive statistics for key variables in the GSA database are shown in Table 2.[21] We observe 223 transactions per quarter, on average. All figures are weighted using data on shipments by country and technology from the Pure Play Foundry Market Tracker database produced by market research firm iSuppli.[22] The changing technological characteristics of the fabrication process are evident in the statistics for wafer size and line width. Pilot production lines for 300 mm wafers

[18]The survey was substantially revised in 2004, so the sample we received from GSA begins in the first quarter of that year.

[19]We also estimated all analyses without omitting these countries, and the results were nearly identical. The only exception that the index results in Section 6.2 are a bit stronger, with a difference of 1.7 percentage points per year between the alternate indexes.

[20]Authors' calculations based on market data from iSuppli.

[21]See Appendix A for details on data sources and data cleaning.

[22]Appendix A describes how weights were constructed. We also implemented all analyses without these weights. The only substantive difference was that the indexes discussed in Section 6.2 were much more volatile, and the the difference between alternative indexes was smaller, at only 0.3 percentage points. Otherwise, all regression results presented below were substantively unchanged.

were first introduced in 2000 and the share for this emerging technology rises from 3 percent of contracts to 20 percent of contracts over the survey. Similarly, new generations of lithography increase in share over time: 65 nanometer technology reached volume production in the overall semiconductor industry in 2006 and slowly gained share in the foundry market, accounting for 8 percent by 2010; 45 nanometer contracts were just emerging in 2010. Meanwhile, older technologies, with line widths larger than 250 nanometers, dwindle in prominence from 40 percent in 2004 to 33 percent in 2010. The number of metal and mask layers per wafer also rose somewhat over the period studied, reflecting a trend toward foundries handling increasingly complex designs.

3.2 Cross-Country Wafer Price Dispersion

Given these detailed data, we move to investigating the cross-country variation in wafer prices in a simple hedonic regression framework, controlling for all of the relevant technological aspects of quality variation across transactions. Specifically, we regress the log of the wafer price on a vector of quarter indicators, indicators for each wafer size and each line width, layer indicators, and the quantity of wafers purchased in the transaction.[23] Table 3 presents the results.

The signs of all regression coefficients are intuitive. The omitted category is for 200mm wafers with 180nm line width, produced in Taiwan. More advanced production technologies, with larger wafers and smaller line widths, command higher prices. For example, the coefficient 0.671 implies that a 300mm wafer is 96% more expensive than an otherwise identical 200mm wafer produced in the same quarter.[24] Similarly, a wafer with 150nm line width was 18% more expensive than an otherwise identical wafer with 180nm line width in the same quarter. Wafers involving more layers are more expensive, as these require more raw materials and more steps in the production process. Larger orders also command a bulk discount. All of these coefficients are very precisely estimated and highly statistically significant.

The regression also includes indicators for the country where the producing foundry is located. The coefficient -0.186 implies that a wafer produced in China sells at a 17% discount compared to an otherwise identical wafer produced in Taiwan. Singaporean and Malaysian producers also exhibit discounts relative to Taiwan, while producers in the U.S. charge higher prices for otherwise identical wafers. These large price differences across supplying countries are very precisely estimated and robust to changes in specification.[25]

These large price differences could reflect unmeasured quality differences across suppliers that are not captured by our regressors. One measure of quality difference not included in our hedonic

[23] Note that hedonic regression estimates reflect a joint envelope of demand and supply functions (Rosen 1974). Including an order size regressor to control for non-linear pricing is reasonable in this context, unlike when trying to separately identify a demand or supply curve. That said, all results are similar when omitting this control.

[24] $\exp(0.671) - 1 = 95.6\%$.

[25] For example, controlling for technology more flexibly by including indicators for each combination of wafer size and line width has no meaningful effect on the country coefficients.

regression is the difference in yields across countries, though these differences are unlikely to account for our results. First, the vast majority of yield improvements occur during the engineering phase of product development, which is omitted from our data.[26] Second, managers at fabless firms report that they receive similar service from the major Taiwanese and Chinese foundries in terms of timeliness and yields.[27] Third, to the extent possible, we have analyzed yield data directly.

The GSA survey included a yield question from 2005 to 2008 that reported yield quartiles (0-25%, 26-50%, 51-75%, 76-100%), but this measure is quite rough.[28] Nearly all responses reported yields in the 76-100% range, and the question was dropped from the survey.[29] This does, however, rule out very large yield differences across suppliers. At our request, GSA added a continuous yield measure to the survey for the third quarter of 2011, allowing respondents to enter any number from 0 to 100%. Controlling for technology as in Table 3, there were no statistically significant differences in yields across countries for those observations where yields were reported, nor were there statistically significant differences in the probability of reporting the yield by country or technology that would suggest selection bias from non-response.[30] However, in both cases the estimates are very noisy, since the non-response rate for the continuous yield question was very high. This analysis is consistent with the anecdotal evidence against cross-country yield differences, but the large amount of non-response curtails the data's ability to generate strong statements regarding yield differences.

Another potential source of differences in quality across suppliers relates to the intellectual property, design tools, and engineering support each foundry provides to its customers to help facilitate the transition from chip design to the foundry's manufacturing process. It is possible that Taiwanese foundries provide better tools than their Chinese competitors, which could partly explain the observed differences in wafer prices between the two countries. In the absence of data on the quality of these tools and other services, we are unable to examine this hypothesis directly.

However, these alternative explanations cannot explain an important feature of the data. Constant yield differences or differences in the quality of service or design tools would lead to stable price differences across suppliers over time. Below we show that the price gap between Taiwanese and Chinese suppliers starts out very large and falls over the life of a process technology, behavior

[26] We consider the costs associated with the engineering phase to be part of the large startup costs modeled in the next section.

[27] We had extensive conversations on this issue with a design manager at LSI Corp, a major U.S. fabless firm, with the CEO and Managing Director of a large Korean fabless firm, and with the GSA Supply Chain Performance Working Group, an advisory panel consisting of executives and managers from various fabless firms.

[28] The only dataset on yields that we are aware of from the prior literature comes from the Competitive Semiconductor Manufacturing (CSM) program at the University of California at Berkeley (Leachman 2002). Thanks to Ken Flamm for bringing this survey to our attention. This data set covers the 1996-2000 time period, does not identify fab location, and relies heavily on foundries producing primarily memory products rather than ASICs, so it cannot be used to analyze cross-country yield differences in our context.

[29] Including indicators for each yield quartile in a version of Table 3 covering 2005-08 are nearly identical to those in 3, as expected given the small amount of variation in the yield quartile measures.

[30] Results available upon request.

that these alternate explanations cannot explain. In contrast, the model of quality-adjusted price dispersion we present in the following section predicts just this pattern of closing price gaps within process technology.

4 A Model of Price Dispersion Across Intermediate Input Suppliers

In this section, we study price dispersion in a duopoly pricing game that is designed to reflect the key features of the semiconductor wafer fabrication market. On the demand side, buyers pay a startup cost to initiate production with a supplier, and if a consumer were to switch suppliers at a later date, it would have to pay this startup cost again. In this sense, the startup cost acts like a switching cost. On the supply side, one supplier enters the market first and enjoys a temporary monopoly. In the following period, the lagging supplier enters.

This setup yields two intuitive results. First, the switching cost generates equilibrium price dispersion. The idea here is that the switching cost partly insulates the leading firm from competition from the lower-price lagging supplier. This allows the leader to retain its past customers even when charging a higher price than the lagging supplier. Second, the price difference between the leading and lagging supplier narrows over time for a particular process technology. When the lagging supplier enters, the leader has a large base of customers. It wishes to charge a relatively high price to these buyers, who are partially locked-in because of the startup cost. However, as the leader's original customer base exits the market, the leader will have a stronger incentive to compete aggressively for newly entering customers. This reduces the price gap between the leader and follower. This result is a simple but important property of the model, as it yields a testable prediction about the source of price dispersion that is not consistent with many forms of unobservable quality differences across suppliers. This gives us a way of distinguishing between competing explanations for the observed price differences in Section 5.

4.1 Model

4.1.1 Environment

The model has three periods. There are two types of agents in the market – buyers and manufacturers of an intermediate good. A cohort of buyers of mass one enters in each of the three periods. The period-1 cohort is present in periods 1 and 2, the period-2 cohort is present in periods 2 and 3, and the period-3 cohort is present in period 3.[31] We assume that buyers have inelastic demand

[31] If there were no entry in the final period and if the switching cost were reasonably high, then suppliers would simply "gouge" their period-2 customers, charging the (common) reservation price. As a result, the equilibrium price would jump up in the terminal period, which we do not observe in the data. The entry of a new cohort in period 3 introduces competition into the terminal period that avoids this counterfactual behavior.

such that each buyer will purchase the intermediate good from one of the suppliers.

Even though buyers purchase the same input (i.e. the same wafer size, line width combination) from both suppliers, the manufacturing process must be tailored to each buyer's unique design. In other words, within a given size and line width pair, there is heterogeneity in chip design, and some designs are more difficult to fabricate than others. Formally, we follow in the spirit of Klemperer (1995) and assume that design complexity, y, is distributed uniformly from 0 (lowest quality) to 1 (highest quality). This heterogeneity across designs would be unobservable to an econometrician who has data only on wafer size and line width. In this sense, the model nests as a special case the theory that price dispersion merely reflects a constant quality difference.

Turning to the manufacturers, Firm A (Taiwan) is the leader and is present in the market from period 1 onward. Firm B (China) is the follower; it joins the market in period 2. We assume Firm A is at the technology frontier. For example, in the wafer market, it is thought that Taiwan's fabrication plants have intellectual property that enables them to more efficiently produce a highly complex design. Accordingly, although Firm B can fabricate any chip, the consumer must pay a cost to monitor and consult with this supplier. To be precise, we assume that buyers who purchase from Firm B pay a per-period premium that is increasing in design complexity, τy (where $\tau > 0$).

The relative efficiency of Firm A renders it a clear advantage. However, for any given design, we assume that Firm B faces a lower unit cost of production, c^B. Specifically, we assume that both firms have constant unit costs, and that $c^A > c^B$. This is intended to reflect the difference in input costs across suppliers in Taiwan and China.

When a buyer initiates production with a supplier, it must pay a startup cost, s. This cost has to be paid again if the buyer switches suppliers. Thus, if a buyer purchased from Firm A in period 1 but switches to Firm B in period 2, it must pay s again (independent of its quality). Hence, it would pay a price, p_t^A, to remain with Firm A in period t, or it may switch to Firm B, in which case it pays $p_t^B + \tau y + s$, where p_t^B is Firm B's posted price this period, τy is the monitoring cost, and the startup cost s acts as a switching cost.

Lastly, we rule out hold-up problems by assuming perfect contracting and follow the related literature on switching costs in prohibiting price discrimination. These assumptions are consistent with wafer supplier contracts, which i) enumerate measurable requirements for buyers and suppliers, ii) specify sanctions if either party does not meet the specified requirements, and iii) explicitly limit the supplier's freedom in charging appreciably different prices across its customers. Hence, the price p_t^A $\left(p_t^B\right)$ applies to all Firm A (B) buyers in period t.

4.1.2 Solution

The problem is solved by backward induction. To analyze the terminal-period problem, we first conjecture that, in the prior period, Firm B attracts all period-2 entrants with $y \leq \hat{y} < 1$. In other words, we assume the least efficient producer will attract buyers with the least complex designs.

This conjecture will be confirmed in equilibrium. In what follows, since y is uniformly distributed, we refer to the mass of buyers $\hat{y}$ as Firm B's "customer base"; the mass of higher-quality buyers $1-\hat{y}$ makes up Firm A's customer base.

Firm A's terminal-period problem may now be stated as follows. There are three groups of buyers to whom Firm A may sell: members of its own customer base, members of Firm B's customer base, and buyers who enter in period 3 (period-3 entrants). The demand schedules for each of these cohorts is given below. Throughout, we let σ_t^{jj} represent the share of Firm j's customer base that it retains in period t; σ_t^{j0} the share of period-t entrants that it attracts; and σ_t^{ji} the share of Firm i's base acquired by Firm j. Hence, since y is drawn from a uniform distribution, we have

$$\begin{aligned}
\sigma_3^{AA}\left(p_3^A;\ p_3^B,\hat{y}\right) &\equiv \Pr\left[p_3^A \le p_3^B + \tau y + s \mid y > \hat{y}\right] = \tfrac{s+\tau-p_3^A+p_3^B}{(1-\hat{y})\tau} \\
\sigma_3^{AB}\left(p_3^A;\ p_3^B,\hat{y}\right) &\equiv \Pr\left[p_3^A + s \le p_3^B + \tau y \mid y \le \hat{y}\right] = \tfrac{-s+\hat{y}\tau-p_3^A+p_3^B}{\hat{y}\tau} \\
\sigma_3^{A0}\left(p_3^A;\ p_3^B,\hat{y}\right) &\equiv \Pr\left[p_3^A \le p_3^B + \tau y\right] = \tfrac{\tau-p_3^A+p_3^B}{\tau}
\end{aligned} \tag{1}$$

where p_t^j denotes the price of Firm j in period t. It follows that total sales by Firm A in period 3 are given by

$$Y_3^A \equiv \sigma_3^{AA}\left(p_3^A;\ p_3^B,\hat{y}\right)(1-\hat{y}) + \sigma_3^{AB}\left(p_3^A;\ p_3^B,\hat{y}\right)\hat{y} + \sigma_3^{A0}\left(p_3^A;\ p_3^B,\hat{y}\right).$$

Then Firm A selects p_3^A to maximize profits, $\left(p_3^A - c^A\right)Y_3^A$.

The solution to Firm A's problem is a best-response function, $P_3^A\left(p_3^B;\ \hat{y}\right)$. Figure 4 illustrates this solution for a moderate value of $\hat{y}$ (and conditional on the calibration discussed below). Here, we provide intuition for a number of features of the policy rule, with a formal analysis in Appendix B.

First, for given $\hat{y}$, Firm A's policy rule as a function of Firm B's price is partitioned into three regimes. For sufficiently small p_3^B, Firm A sells only to its customer base. To attract period-three entrants in the face of a low Firm B price, Firm A would have to lower its price, too. But if Firm B's price is sufficiently small, this will not be optimal. Since it is costly for its customer base to switch suppliers, Firm A can set a higher price and earn greater profits by selling exclusively to its partially "locked-in" buyers.

Second, at the point where Firm A elects to compete for new entrants, its optimal price changes *discretely*. To see why, suppose the opposite – imagine Firm A's strategy were continuous. In that case, its share of period-three entrants would vary continuously as a function of p_3^B. This would mean that there exists a Firm B price such that Firm A chooses to attract a share of new (period-three) entrants just greater than zero. But this cannot be optimal because $s > 0$. If Firm A were in this position, it could raise its price, discretely increasing profits from its old buyers because they face the discrete cost to switch and losing only an infinitesimal amount of profit by foregoing sales

to entrants. Therefore, profit from "gouging" its customer base must be higher.[32]

Firm A's price will also change discretely at the point where it starts poaching from Firm B's customer base. This reflects the fact that buyers – in this case, members of Firm B's customer base – face a discrete cost to switch suppliers. As a result, Firm A will have to cut its price discretely to compensate firms which switch from B to A. Such a discrete price cut is only optimal if Firm A attracts enough Firm B customers. This will be true if Firm B's price is sufficiently high.

Firm B faces a symmetric problem and so has a similar best response function. The intersection of the two yields the terminal-period equilibrium, given $\hat{y}$. We denote the equilibrium prices by $\mathcal{P}_3^A(\hat{y})$ and $\mathcal{P}_3^B(\hat{y})$.[33]

It remains to determine the threshold, $\hat{y}$, which represents Firm B's share of period-2 entrants. To this end, we now sketch the period-2 problem. Although it is not uncommon in the literature to assume myopic consumers, buyers in our model have perfect foresight, so we find subgame-perfect Nash equilibrium.[34] A period-2 entrant with design y will purchase from Firm A only if the discounted sum of period-1 and 2 prices is less than those the entrant would face if purchasing from Firm B. This implies that $\hat{y}$ satisfies the indifference relation,

$$p_2^A + \beta \min\left[\mathcal{P}_3^A(\hat{y}),\ \ \mathcal{P}_3^B(\hat{y}) + \tau\hat{y} + s\right] = p_2^B + \tau\hat{y} + \beta \min\left[\mathcal{P}_3^A(\hat{y}) + s,\ \ \mathcal{P}_3^B(\hat{y}) + \tau\hat{y}\right], \quad (2)$$

where $\beta < 1$ is the discount factor. This implicitly defines the threshold, $\hat{y}$, as a function of period-two prices, $\hat{y}\left(p_2^A, p_2^B\right)$.[35] Thus, Firm A's demand schedule among period-2 entrants is $1 - \hat{y} = \sigma_2^{A0}\left(p_2^A; p_2^B\right)$.

In addition, Firm A may have some its old buyers poached by Firm B. Specifically, Firm B attracts a buyer y in Firm A's base if $p_2^B + \tau y + s \leq p_2^A$. Since Firm A captures all (mass one) of period-1 entrants, this means that Firm B attracts a share of them equal to $\frac{p_2^A - p_2^B - s}{\tau}$. Thus, Firm A's demand schedule among members of its customer base is $\sigma_2^{AA}\left(p_2^A; p_2^B\right) = 1 - \frac{p_2^A - p_2^B - s}{\tau}$.

[32] This point is particularly easy to see if all old buyers are infra-marginal – the cost to switch is so large that all old buyers are retained if Firm A raises its price in this scenario. But the point is more general. Even if Firm A loses some of its lower-quality buyers when it raises its price (discretely), the price change is still optimal. Intuitively, if the firm is not going to capture a sufficiently large share of the new entrants, it still prefers to "gouge" the higher-quality buyers in its customer base.

[33] Given the discontinuities in the best response function, under certain conditions there may be no equilibrium. In that case if Firm A makes a marginal adjustment to its price, that may trigger a discrete shift in the optimal Firm B price. For instance, if Firm A raises its price slightly, it may be profitable now for Firm B to discretely lower its price, enter the market and capture a discrete share of new buyers. Firm A reacts to this by adjusting its price down. Firm B sees the lower Firm A price as a reason to give up on its attempt to attract new buyers, so it raises its price, again discretely. And the process repeats – at no point is either firm satisfied with its price. Below, we examine a calibration that does yield a unique Nash equilibrium in which both firms compete for entrants in periods 2 and 3.

[34] See discussion in the survey by Farrell and Klemperer (2007).

[35] It is not immediate that there is a unique solution for $\hat{y}$. But under certain conditions (honored in our calibrated model), this can be established. Intuitively, the discounted sum of Firm B prices is relatively low when $\hat{y}$ is low (i.e. when Firm B's customer base is small, it sets lower prices to attract new entrants). But the discounted sum of Firm B prices is also increasing at a relatively fast rate as $\hat{y}$ rises. This directly reflects the quality premium, as captured by $\tau\hat{y}$. Together, these features imply a single crossing, with the right side of (2) cutting the left side from below.

Putting this together, Firm A's problem is maximize

$$\left(p_2^A - c^A\right) Y_2^A(p_2^A; p_2^B) + (\mathcal{P}_3^A(\hat{y}) - c^A) Y_3^A(p_2^A; p_2^B),$$

where

$$Y_2^A = \sigma_2^{A0}\left(p_2^A; p_2^B\right) + \sigma_2^{AA}\left(p_2^A; p_2^B\right)$$

$$Y_3^A = \sigma_3^{A0}(\mathcal{P}_3^A(\hat{y}), \mathcal{P}_3^B(\hat{y})) + \sigma_3^{AA}(\mathcal{P}_3^A(\hat{y}), \mathcal{P}_3^B(\hat{y})) + \sigma_3^{AB}(\mathcal{P}_3^A(\hat{y}), \mathcal{P}_3^B(\hat{y})).$$

$$\hat{y} = \hat{y}\left(p_2^A, p_2^B\right)$$

The solution to this optimization problem yields Firm A's best response as a function of p_2^B. Firm B solves a symmetric problem, and combining the two firms' best responses yields the equilibrium period-2 prices. With these in hand, one recovers $\hat{y}$ and thus the period-3 prices, $\mathcal{P}_3^A(\hat{y})$ and $\mathcal{P}_3^B(\hat{y})$.

Because of the discontinuities in the pricing rules, it is difficult to solve the model in closed form. We instead choose an illustrative calibration and solve it numerically. There are five parameters, $\left(\beta, c^B, c^A, \tau, s\right)$, to select. First, we assume the period is one year and so set $\beta = 0.95$, which implies an annual real interest rate that is slightly higher than 5 percent. Second, as it is the relative price that affects the allocation of buyers across suppliers, we can fix $c^B = 1$. We then set $c^A = 1.20$ to imply a unit cost of production that is 20 percent higher at the leader. This choice results in a close match between the leader-follower price gap and the Taiwan-China price gap in the period of entry. Note that we do not target the *change* in the price gap; we let the model speak in this regard. Next, as equation (1) suggests, the extent of switching hinges on the value of s relative to τ. For example, a Firm B customer weighs the cost of switching to Firm A against the premium, τy, that it saves by ending its contract with Firm B. Accordingly, we normalize $\tau = 1$ and then select s.

Two factors guide our choice of s. First, since switching appears to be relatively rare in this market, s must be sufficiently large. On the other hand, if s is too large, there may be no equilibrium. Intuitively, as s increases, the "jumps" in the policy rule in Figure 4 grow larger: at the point where Firm A elects not to compete for new entrants, it will raise its price by a lot to gouge its locked-in buyers. But as these jumps grow larger, it becomes more difficult to find a point of intersection between the two firms' best responses. Our strategy, then, is to set s to be nearly the highest possible value consistent with the existence of equilibrium, in this case $s = 0.5\tau$.

Table 4 reports the results. There are a few we wish to highlight. First, the model implies that the degree of price dispersion declines over the product life cycle. The model implies a gap of roughly 40 log points in the period in which the lagging supplier enters, and a gap of around 20 log points in the next period. These results are very intuitive: the leader keeps its price elevated in period 2 in order to exploit its locked-in customers and then competes more aggressively for entrants once the original cohort of buyers exits the markets. These price gaps are quite close to

the empirical estimates reported in the next section. For this reason, we believe that our calibrated model, although quite simple, provides a useful perspective on the dynamics of price dispersion in this market.

Second, the average log gap is 16 percent higher than it would be in the frictionless model. To see this, note that the average of the logarithmic price gaps under $s > 0$ is 29. If we set $s = 0$ in our model, one may show that the optimal prices are given by $p^A = \frac{2}{3}\left(\tau + c^A\right) + \frac{1}{3}c^B$ and $p^B = \frac{1}{3}\left(\tau + c^A\right) + \frac{2}{3}c^A$. These results imply a constant log difference in prices of 25 points. Hence, the gap implied by $s > 0$ is $\frac{4}{25} = 16$ percent higher. Lastly, the calibration of s implies minimal switching – 14 percent of Firm A's customer base (carried over from period 1) switches to Firm B in period 2. No buyer switches away from either supplier in period 3.

4.2 Discussion

In this section, we wish to first discuss the mapping from the model to features of the wafer industry which it is intended to represent. Second, though the model is written with a particular market in mind, it is instructive to relate it to the wider literature on costly switching.

4.2.1 Relation to Wafer Industry

The two suppliers in the model are intended to represent the Taiwan Semiconductor Manufacturing Company (TSMC) and China's Semiconductor Manufacturing International Corporation (SMIC). Each firm is the largest foundry in its respective country. TSMC is the overall industry leader. Its foundries consistently enter the market for a given process technology (wafer size, line width combination) at least 8 quarters ahead of Chinese foundries.[36] This is why we assume in the model that the leader enters the market before its competitor.

Although the reason for China's later entry is not central to our argument, it likely occurs for a number of reasons. Chinese foundries may adopt new technology more slowly due to a relative lack of technical expertise in bringing new process technologies into the foundry industry. The larger and older Taiwanese firms have much more experience in this area. There are also restrictions on the export of the most advanced wafer fabrication equipment from manufacturers in the U.S., Europe, and Japan. These restrictions include the multilateral Wassenaar Arrangement and strong restrictions implemented by the Taiwanese Government that restrict the trade of dual-use technologies (EE Times 1998).[37]

[36] Authors' calculations using iSuppli data. Observable delays range between 8 and 12 quarters, though Taiwanese entry is censored for many technologies.

[37] Semiconductor manufacturing equipment is produced in the U.S. (45%), Japan (40%), and Europe (15%), based on 2006 Gartner reports. All of these producers are located in Wassenaar Arrangement participant countries, though Brown and Linden (2006) suggest that these restrictions have not substantially limited fabrication equipment exports to China. Industry insiders suggest that the Taiwanese restrictions may be more important, with the government willing to prosecute those illegally transferring technology to mainland China.

The cost of switching in the model is motivated by the large fixed costs of starting production of a particular design at any one supplier. Perhaps the most prominent example of the start-up cost s, is the cost of a new mask (see Section 2). Each foundry has its own proprietary processes and technologies that are generally incompatible across manufacturers, which means that a mask set cannot be reused at a foundry other than the one where the wafer is initially produced. A typical mask set designed for fabricating a 90nm wafer is priced at around $1.2 million. As a result, as one industry association noted, "The time and cost associated with [switching] tend to lock customers into a particular foundry."[38]

Other examples of startup costs include the many chemical and mechanical adjustments and calibrations to manufacturing equipment that are implemented during the engineering phase of production for a particular design. These adjustments are so delicate and specific that customers are reluctant to switch to another production line in the same facility, much less to another foundry. There are also other less tangible resource costs of beginning a new foundry relationship. Negotiating a new supplier agreement requires a vast amount of information to be exchanged about the specific details of the wafer design and the technological requirements needed to manufacture the chips. These talks can absorb much attention of senior management.[39]

In comparison to the framework used here, alternative theories of price dispersion such as costly search (Stigler 1961, Burdett and Judd 1983) are less well suited to the wafer market and other similar intermediate input markets. The wafer market is highly concentrated, with a few large suppliers, so the cost to search for a foundry does not seem first-order. Instead, the rents that bind parties in long-term relationships in this market more likely result from the sort of relationship-specific investments that we have discussed.[40]

4.2.2 Relation to Literature

Although the model of section 4.1 was written with a particular market in mind, it is related to the treatment of costly switching in the broader literature. For example, the model is similar to Klemperer (1987). On the supply side, both models allow for a form of product differentiation. On the demand side, both have overlapping generations of buyers. They differ insofar as Klemperer

[38]This quotation is taken from the Common Platform alliance. This industry group consists of a few large chip manufacturers, such as IBM and Samsung, that have begun to advocate for a "common platform" that would standardize production technology for certain high-volume wafers (such as 90nm chips). However, this alliance has not yet had a material impact on standardizing mask sets (McGregor 2007). We are grateful to Ross Goodman for insightful discussions on this topic.

[39]Allan (2002) gives a good example of this, discussing how a new fabless firm began a relationship with TSMC. The executives at the fabless firm stressed the importance of establishing clear channels of communication so that the foundry management knew precisely who to contact regarding details of the wafer design and production needs. Otherwise, if there were a single liaison in the fabless firm fielding all questions, that person "will be getting 250 communications per day" from the foundry.

[40]We suspect that this may be true of many other input markets as well, insofar as they tend to be relatively concentrated. Data from the 2007 U.S. Economic Census indicate that in a quarter of 6-digit manufacturing industries, the largest 20 producers account for at least 90 percent of sales.

assumes the cost of switching is so high as to prohibit any switching in equilibrium. In addition, since s is so substantial, Klemperer must assume that a sufficiently high share of period-1 buyers attrit after that period in order to induce both suppliers to compete for period-2 entrants.[41] Otherwise, a supplier may find it optimal to simply "gouge" its own customer base in period 2 rather than lower its price to compete for new buyers. A high attrition rate depletes the customer base and gives the supplier a reason to compete.[42]

With respect to more recent research, we wish to highlight three points. First, our simple three-period structure allows us to more easily focus on short-run equilibrium pricing dynamics with entry. Our objective differs from that of Dubé, Hitsch and Rossi (2009), who wish to estimate switching costs assuming that firms are very long-lived. Accordingly, they abstract from entry, assume infinitely lived sellers, and focus on steady-state equilibrium.

Second, it is common in the literature to allow for time-varying heterogeneity among buyers.[43] In our context, this would amount to assuming that buyers take a new draw from the complexity distribution in each period. This form of heterogeneity is technically convenient, as it can eliminate the discontinuity in sellers' demand schedules.[44] We have not introduced this feature simply because it has no obvious counterpart in the specific market we study.

Third, as in Klemperer (1987) and others, we prohibit price discrimination. Recent research by Cabral (2012) relaxes this restriction and allows suppliers to charge different prices to incumbent customers and new buyers. In our context, this approach would eliminate the fundamental tradeoff between gouging prior customers and competing for new entrants. Moreover, it is not clear this is a descriptively better model. As we noted, foundry contracts appear to restrict the extent of price discrimination. A more complete theory would, of course, generate this outcome endogenously, but as that model is beyond the scope of the present paper, we prohibit price discrimination directly.

5 Dynamics of Price Dispersion

The preceding model yields a clear testable prediction with respect to the dynamics of the price gap between leading and lagging suppliers of a given technology: the price gap for a given technology should fall over time following entry of the lagging supplier. In this section we take this prediction to the data.

[41] In his footnote 8, Klemperer reports that the attrition rate may have to exceed 40 percent to ensure an equilibrium in which each supplier competes for new (period-2) entrants.

[42] To (1996) presents an extension in which buyers have finite lives, but the suppliers live forever. He maintains the assumption that switching costs are so large as to rule out switching in equilibrium.

[43] See, for instance, Shin and Sudhir (2009). Their paper, like ours, studies a short-horizon duopoly game.

[44] Allowing buyers to redraw y in each period ensures that there are some existing Firm B customers who now prefer purchasing from Firm A at prices high enough to be rejected even by old Firm A buyers. This means the demand schedules facing Firm A – from its own buyers and from members of its competitor's customer base – link up in a continuous way. In contrast, in our setting there is a "gap" in the demand curve: Firm A would have to drop its price discretely to poach, since Firm B is still a relatively better match for its old buyers.

We focus our analysis on the two largest foundry markets, China and Taiwan. We restrict the sample to technologies that were produced by both Chinese and Taiwanese foundries and drop observations for other production locations. The left side of Table 5 re-estimates the hedonic specification from Table 3 on this restricted sample. The China discount coefficient is very similar in Tables 3 and 5, as are the technology coefficients, so the sample restriction does not affect the results substantially. Then we measure the quarter in which a Chinese foundry first began producing wafers of each line width and wafer size combination, and in each quarter calculate the elapsed time since Chinese entry for each technology.[45] We use an event-study style framework to show that Chinese foundries charge less than Taiwanese foundries upon entry, and that this gap closes as more time elapses since Chinese entry.

We let p_{ijkt} refer to the price in transaction i, produced in country j, using technology k, in quarter t. Define δ_{kt} as the number of years since Chinese entry for the relevant technology (wafer size and line width pair) in the relevant quarter, i.e. $\delta_{kt} = 0$ in quarters 0-3 following Chinese entry, $\delta_{kt} = 1$ in quarters 4-7, etc. The estimation equation takes the following form.

$$\ln p_{ijkt} = \beta_0 + \beta_c \mathbf{1}(China_j) + \sum_{s=1}^{T} \beta_s \mathbf{1}(\delta_{kt} = s) + \sum_{s=1}^{T} \beta_{sc} \mathbf{1}(China_j) \cdot \mathbf{1}(\delta_{kt} = s) + \Theta X_{ik} + \gamma_t + \epsilon_{ijkt} \quad (3)$$

This specification can be thought of as a series of difference-in-differences estimates in which β_c reports the price gap between China and Taiwan in the year of Chinese entry, and each β_{sc} reports the change in the price gap from the initial period to the sth year. The remaining terms capture common trends and observables. The β_s capture the average price decline across both countries as the technology ages. X_i is a vector of all of the technological controls associated with transaction i. These include wafer size, line width, number of layers, and order size. The γ_t's are quarter indicators.

This model allows us to estimate the changing price gaps across Chinese and Taiwanese producers for otherwise identical wafers. The model predicts $\beta_c < 0$ (Chinese products are cheaper upon entry), $\beta_{sc} > 0$ for all s (the price gap is smaller after entry), and that β_{sc} grows in magnitude as s increases (the gap closes as the elapsed time since Chinese entry increases).

The estimation results are presented in Table 5. The coefficient on the China dummy (β_c) is negative and statistically significant, indicating a 39% average discount in the year of Chinese entry into the market.[46] The time dummy coefficients (β_s) are also negative and increase in magnitude with elapsed time since Chinese entry, indicating decreasing average prices relative to the overall market as a technology ages. Of central interest are the coefficients on the interaction terms (β_{sc}).

[45] We measure Chinese entry by production technology using information from the iSuppli Foundry Market Tracker and our own analysis of foundry company reports. See Appendix A for the details of this process.

[46] $\exp(-0.494) - 1 = 39\%$.

As anticipated, these are all positive and significant, indicating that the price gap between Taiwan and China for a particular technology closes in the years following Chinese entry. Moreover, the magnitude increases with time, indicating that the price gap closes more as time since Chinese entry elapses for each technology. This can be seen more clearly by adding each interaction term coefficient to the China indicator coefficient to measure the gap itself in each year following Chinese entry. The resulting estimates of the gap for each year following Chinese entry are presented in Figure 5, along with a 95% confidence interval.[47] It is clear that the gap closes over time. One year after entry (4-7 quarters), the price gap has closed by 48%. With a small exception in the third year after entry (12-15 quarters), the price gap continues to close and eventually closes by 79% after 5 or more years following entry.

These results are quite striking in how closely they fit the model's predictions. China enters a particular process technology market with a substantial price discount, and that discount falls as more time since Chinese entry has elapsed. This finding provides evidence consistent with our argument that China's delayed entry into the market for each process technology and substantial fixed startup costs drive the quality-adjusted price differences that we documented in Section 3.

These findings are inconsistent with unobserved quality differences that are constant over time. Thus, explanations for the China-Taiwan price difference based on brand loyalty, differences in customer service, fixed yield differences, or other time-constant unobservable differences across suppliers cannot explain this pattern in the data. Only the small price gap that remains 5 years following Chinese entry is likely to reflect these fixed quality differences across Taiwanese and Chinese suppliers. The findings are also inconsistent with unobserved quality differences that vary steadily over time irrespective of timing relative to the introduction of the relevant process technology. Thus, the price dynamics also cannot be explained by changes in intellectual property rights, changing exchange rates or trade policies, or other time-varying factors that progress steadily across process technologies; variation based on such changes would be captured by the quarter fixed effects in (3).

A remaining alternative argument that could justify these findings is that Chinese foundries enter a process technology with extremely low yields compared to Taiwanese foundries and that these yield differences close over time, explaining the converging price gap. Given the very large price discount of 39% upon entry, if yields drove price differences, they would likely have been so large that they would have been picked up even by the yield quartile measures present in the 2005-08 GSA data. There is no evidence of such large yield differences - in fact, not a single transaction produced at a Chinese foundry in that period reports yields below the 76-100% range. Thus, the empirical evidence strongly suggests that the observed price differences in Section 3 reflect the kind of mechanisms based on sequential entry and startup costs highlighted in the model in Section 4.

[47] The associated standard errors in Table 5 and Figure 5 are clustered by quarter.

6 Additional Implications

In this section we discuss a few additional implications of quality-adjusted price differences across intermediate input suppliers, focusing on two areas of recent research interest. The first is the burgeoning literature studying how product quality differences across countries affect patterns of international trade. Our findings support recent work in that literature by casting doubt on the assumption that cross-country differences in unit values are fully explained by quality differences, even within narrow product classifications. However, our results reveal dynamic aspects of intermediate input markets that are not captured by static quality-measurement procedures. The second area focuses on the effects of offshoring of production on measures of input prices and productivity growth produced by government statistical agencies. We will show that the presence of quality-adjusted price differences across producers and substitution toward low price suppliers likely results in upward-biased measures of input price growth and productivity growth.

6.1 Cross-Country Product Quality Measurement

A large existing literature analyzes how to measure quality variation across suppliers of similar traded goods. Early papers in the literature assumed that the law-of-one-price holds within narrow product classification, and thus attributed all observed price (unit-value) differences across suppliers to unobserved differences in product quality.[48] Subsequent studies documented positive relationships between unit values and trading country characteristics thought to be associated with quality, such as capital intensity and income (Hallak 2006, Hummels and Klenow 2005, Schott 2004), suggesting that unit values do in fact contain information about quality variation across supplying countries.

In our data, price dispersion does reflect a degree of quality difference, but our results caution against assuming a one-to-one mapping between price and quality. Without technological controls, the average Chinese wafer price is 30% lower than Taiwan's, reflecting Chinese foundries' focus on trailing-edge products. After controlling for product characteristics, the price difference narrows; our results in Section 3 show that Chinese wafers are 17% cheaper than otherwise identical Taiwanese wafers. This finding rejects the price-as-quality assumption, which does not allow for the possibility of quality-adjusted price variation.

Reacting to various shortcomings of the price-as-quality assumption, more recent research accounts for quality-adjusted price variation across similar goods by assuming a love of variety, implying that goods are imperfect substitutes in consumption. In this context, goods from all sources will be consumed in at least some small amount, including goods that are expensive relative to their quality. However, such goods will capture a very small share of the market. This is the basis for identifying quality differences across countries; conditional on price, variation in market share

[48] See Hallak and Schott (2011) and Khandelwal (2010) for extensive lists of references.

reflects unobserved quality differences across suppliers. Recent papers utilizing variations on this approach include Hallak and Schott (2011), Hummels and Klenow (2005), Khandelwal (2010), and Kugler and Verhoogen (2011).

While this cross-sectional approach likely captures persistent differences in product quality across countries, our results highlight dynamic aspects of intermediate input markets that may confound quality inference based on market shares. In the model described in Section 4, differences in market share across suppliers do not necessarily reflect differences in product quality. Rather, the leading firm achieves a large market share simply by entering the market first and locking in a large number of customers early on. In the semiconductor foundry context, the average difference in Taiwanese and Chinese market shares would overstate the quality difference between the two.[49] Moreover, the leader's relative price and relative market share fall as its relative quality increases from period 2 to 3. In this case both price (unit value) and market share based approaches to inferring quality would suggest declining relative quality, when it was actually increasing. We conclude from this exercise that even if long-run market share is a faithful indicator of average quality, the theoretical case for identifying the *dynamics* of product quality using changes in market share is not necessarily strong. Identifying these dynamics is critical to characterizing the effects of entry by new offshore suppliers. We suspect that progress on this front will likely require a more explicit application of models with product market frictions.

6.2 Input Price and Productivity Measurement

Quality-adjusted price dispersion across suppliers poses a particularly difficult challenge for price index construction. As an example, consider an import price index. Imagine that the agency preparing the index observes a buyer purchasing a high-priced good from Taiwan in one period and a lower-priced good from China in the subsequent period. The agency must decide what portion of the observed price decline to include in the index, based on what portion of the price difference reflects a real quality-adjusted discount rather than lower quality. This decision requires extremely detailed data on product characteristics. In the absence of this information, the U.S. Bureau of Labor Statistics typically assumes that there are no quality-adjusted price differences across suppliers and that any observed discount reflects lower quality. In practice, this assumption involves starting a new price series for the good purchased from the lower-cost supplier, so the price drop incurred by shifting to a cheaper supplier is omitted from the index. This approach is correct when the law of one price holds across suppliers. But when there are quality-adjusted price differences across suppliers, the resulting index will miss the price declines that occur with substitution toward lower cost intermediate input suppliers. Missing these input price declines

[49] If one considers early entry into a given production technology as an aspect of "quality," then one could interpret the increased market share for early entrants as an aspect of quality that would be captured by the prior literature's approach. We are sympathetic to this interpretation, but believe that much is still lost in omitting the dynamic considerations we introduce.

leads to an upward-biased input price index, a downward-biased estimate of input quantity growth, and a corresponding upward-biased measure of productivity growth.

Houseman et al. (2010, 2011) examine the implications of this bias for productivity mismeasurement in the U.S. manufacturing sector using the micro data underlying the U.S. import price index. They attempt to identify quality-adjusted price dispersion by, in part, comparing prices across countries within the same 10-digit Harmonized System (HS) product category. Their results imply an upward bias in the official productivity growth estimate of 0.1 to 0.2 percentage points per year from 1997 to 2007.

We view our empirical analysis in this section as complementary to that in Houseman et al. While our data do not span the entire manufacturing sector, they do provide much more detail on product attributes. In our context, this detail is needed to implement credible quality adjustments, since there is substantial quality variation across wafer suppliers even within 10-digit HS categories.[50]

We now apply our GSA data to revisit the implications of shifts in product sourcing for aggregate price indexes. Specifically, we calculate two price indexes for semiconductor wafer fabrication services. One includes price declines that coincide with shifts toward lower-price suppliers. The other mimics BLS practice and excludes these changes (see below for more details). Both are matched-model indexes, which start with quarterly price relatives for each "model" and aggregate across models using a Fisher index.[51]

The two indexes differ in their definition of what constitutes a model. The first, which we call the "technology-country index," considers both technology (wafer size and line width) and country of production as characteristics that define a model. This index reflects the statistical agency approach just described - technologically identical goods produced in separate countries are treated as distinct models. The aggregate price index is an average over model-specific inflation rates. This means that shifts in market share toward countries with lower price *levels* have no effect on this index. The second index, which we call the "technology-only index," defines a model based on technology alone. This index does reflect increases in the market share of low price level suppliers, since such shifts lower the average price of a given technology.[52]

[50] Along with comparisons across countries within 10-digit HS category, Houseman et al. (2011) also use the quality measurement procedure developed in Mandel (2010), which is similar to those described in Section 6.1. We have already discussed the potential difficulties that this type of quality adjustment procedure is likely to encounter in our context.

[51] First we calculate Laspeyres and Paasche indexes, respectively, as $P_L^t = \sum_m s_m^{t-1} \frac{p_m^t}{p_m^{t-1}}$ and $P_P^t = \left[\sum_m s_m^t \frac{p_m^t}{p_m^{t-1}}^{-1}\right]^{-1}$, where m represents each model, t is time (quarter), p is the average price for a given model, and s_m is the share of total output in time t accounted for by model m. As the Laspeyres index overstates the welfare-theoretic "true" price change and the Paasche understates it, we construct the superlative Fischer index, which is a geometric mean of the Laspeyres and Paasche indexes: $P_F^t = \sqrt{P_L^t P_P^t}$.

[52] This approach follows Reinsdorf's (1993) analysis of outlet substitution bias in the U.S. CPI. We also implemented time-dummy hedonic and hedonic-imputation versions of the analysis. Both hedonic indexes fell somewhat more

In order to visualize the difference between the two price indexes, Figure 6 shows an example for a particular technology (300mm wafer, 90nm line width) from the third quarter of 2007 to the end of 2008.[53] In this example, Taiwan is the sole producer until China enters in the second quarter of 2008, at a substantially lower price. The technology-only index, which is based on the *average price* across producing countries, falls below the Taiwanese price series immediately upon China's entry and continues to do so as China's market share increases over time. In contrast, the technology-country index averages *the rates of inflation* across countries and omits variation in the price level due to shifting sourcing patterns. Since the average Chinese price remains constant over time, while the Taiwanese price falls, the technology-country index lies above the Taiwanese price in spite of China's lower price level.

This pattern holds more generally across the full sample of technologies and quarters. The results of the index calculations are presented in Table 6 and Figure 7. The technology-country index falls by 10.4% per year, while the technology-only index falls much more quickly, 11.6% per year. This difference between the two indexes reflects substitution toward lower price suppliers within each wafer size, line width combination. Given the evidence already presented that a portion of the observed price differences across suppliers reflects real quality-adjusted price differences, the 1.2 percentage points per year difference between the two indexes likely reflects some degree of upward bias in the technology-country index, which assumes away quality-adjusted price differences.

Ideally, we would compare our price indexes to the analog in official statistics, but the BLS does not publish a similarly disaggregate index.[54] Still, our analysis is informative regarding the *process* of index calculation utilized in official indexes such as the BLS International Price Program (IPP). Of central importance is the definition of a "product." In particular, we define products based on the technological characteristics of the manufacturing services being purchased from the foundry by the fabless firm. This approach supports the technology-only index calculation just described. In contrast, Nakamura and Steinsson's (forthcoming) analysis of the IPP micro data persuaded them that a product in the IPP is effectively defined as a "contract between a particular buyer and seller." Consistent with this interpretation, between 2004 and 2007, the IPP micro data

slowly than the matched-model indexes generated here, paralleling the results of Doms et al. (2003). In both cases, controlling for country of production results in an index that falls more slowly than versions without country controls, as with the matched-model indexes presented here, though the gaps between the indexes are somewhat smaller: 0.8 percentage points per year for time-dummy hedonic index and 0.6 percentage points per year for the hedonic imputation index.

[53] This technology and time period were chosen because they yield a particularly clean comparison between the two index approaches. In this sense it is a special case, but the overall index results presented in Table 5 demonstrate that the relationship between the indexes demonstrated in Figure 6 holds more generally.

[54] The closest comparison index is the Bureau of Labor Statistics' International Pricing Program (IPP) index for Harmonized System Code (HSC) 8542, electronic integrated circuits and microassemblies. This product category differs from our indexes in three respects: i) our data shows prices for wafer fabrication services purchased by chip makers around the world, while the BLS samples only U.S. importers, ii) Our data include only arms-length transactions while the IPP also includes intra-company transfers, and iii) the IPP index is much more aggregated than our indexes, as the IPP includes finished semiconductor chips and microassemblies in addition to processed wafers and die.

exhibit no examples of semiconductor products switching suppliers, presumably because the new supplier relationship was considered a new product.[55] This product definition closely mimics our technology-country index by omitting differences in price level across suppliers. In fact, it is even more subject to substitution bias because it not only omits price declines due to substitution across countries but also omits price declines due to substitution across suppliers within the same country.

7 Conclusion

In this paper, we have investigated pricing and product sourcing in the offshore market for semiconductor wafer fabrication services. Using detailed, transaction-level data, we detect notable variation in prices across suppliers for observationally equivalent goods, with China in particular offering substantial discounts. We also found significant shifts in production toward lower priced suppliers.

We interpret these findings within a model of sequential entry of suppliers and fixed costs of beginning production with a given supplier. The model shows that these features amplify price differences across suppliers beyond what would be seen in a frictionless model. The model also implies a falling price gap between leading and lagging suppliers of a particular technology, a pattern that is not present in a frictionless model and cannot be explained by constant unobservable quality differences across suppliers. We confirm this pattern in the case of Taiwanese and Chinese wafer manufacturers.

Our findings have important implications for prior work studying quality variation across suppliers of otherwise similar goods. We consider a mechanism for generating quality-adjusted price variation that differs from those in the previous literature studying cross-country quality and price differences. Our findings support recent work in that literature by rejecting the assumption that differences in unit values are fully explained by quality differences, but also highlight dynamic aspects of intermediate input markets that are not captured by static quality-measurement procedures. We also contribute to emerging research demonstrating the challenges posed by offshoring when calculating official aggregate price indexes. Our analysis shows that shifts toward lower-priced suppliers accelerated the average annual price decline of semiconductor wafers by 1.2 percentage points per year and argue that these declines are likely to be omitted from standard official index calculations.

Going forward, there are at least two important avenues for related research. First, we conjecture that frictions are not unique to the market for wafer fabrication services. Although the difficulties of switching suppliers are particularly large in this sector, the costs involved in initiating a new

[55] We thank Ben Mandel for this calculation. The IPP staff also kindly provided us with similar calculations for the May 2009 to may 2011 time period. They report that less than 0.25 percent of the items in the IPP's HSC 8542 sample reported a shift in source country.

overseas relationship are likely to be appreciable in most cases. To this end, analyses of price dispersion in other offshoring markets would be revealing.

Our findings also suggest that official index calculations could benefit from the types of quality adjustment procedures we implement here. Although this would likely be difficult in certain product markets, our results suggest that it could be quite beneficial to implement more detailed quality adjustment procedures for sectors with rapid technological change and fixed costs generating quality-adjusted price dispersion across producers. The U.S. Producer Price Index already implements such hedonic quality adjustments for selected high tech products (Holdway 2001). As offshore contract manufacturers move into more complex markets, other indexes such as the IPP would benefit from similar adjustments.

References

Aizcorbe, Ana, "Price Measures for Semiconductor Devices," *Board of Governors of the Federal Reserve System (U.S.), Finance and Economics Discussion Series*, March 2002, *2002-13*.

___ , Kenneth Flamm, and Anjum Khurshid, "The role of semiconductor inputs in IT hardware price decline: Computers vs. Communications," *Board of Governors of the Federal Reserve System (U.S.), Finance and Economics Discussion Series*, 2002, *2002-37*.

Allan, Gabriel, "A Marriage of Unequals," *Electronics Design Chain*, Summer 2002, *1*, 17–20.

Baldwin, Richard E. and Paul R. Krugman, "Market Access and International Competition: A Simulation Study of Random Access Memories," in Robert C Feenstra, ed., *Empirical Methods for International Trade*, MIT Press, 1988, pp. 171–202.

Barak, Sylvie, "Intel, Micron on sub 20-nm and insatiable thirst for memory," *Electronic Engineering Times*, April 2012.

Berndt, Ernst R. and Neal J. Rappaport, "Price and Quality of Desktop and Mobile Personal Computers: A Quarter-Century Historical Overview," *American Economic Review*, 2001, *91*, 268–273.

Besanko, David, Ulrich Doraszelski, and Yaroslav Kryukov, "The Economics of Predation: What Drives Pricing when there is Learning-by-Doing?," *mimeo*, 2011.

___ , ___ , ___ , and Mark Satterthwaite, "Learning-by-Doing, Organizational Forgetting, and Industry Dynamics," *Econometrica*, 2010, *78*, 453–508.

Brown, Claire and Greg Linden, "Offshoring in the Semiconductor Industry: Historical Perspectives," in "Brookings Trade Forum: 2005," Washington, DC: Brookings Institution Press, 2006.

Burdett, Kenneth and Kenneth Judd, "Equilibrium Price Dispersion," *Econometrica*, 1983, *51*, 955–69.

Byrne, David and Carol Corrado, "Prices for Communications Equipment: Rewriting a 46-Year Record," *mimeo*, 2012.

Cabral, Luis, "Switching Costs and Equilibrium Prices," *mimeo*, 2012.

Doms, Mark, "Communication Equipment: What has Happened to Prices?," in "Measuring Capital in the New Economy," Vol. 65 of *NBER Studies in Income and Wealth*, University of Chicago Press, 2005, pp. 323–362.

___ , Ana Aizcorbe, and Carol Corrado, "When Do Matched-Model and Hedonic Techniques Yield Similar Measures?," *Federal Reserve Bank of San Francisco, Working Papers in Applied Economic Theory*, June 2003, *2003-14*.

Dubé, Jean-Pierre, Günter Hitsch, and Peter Rossi, "Do Switching Costs Make Markets Less Competitive?," *Journal of Marketing Research*, 2009, *46*, 435–445.

Dulberger, Ellen R., "Sources of Price Decline in Computer Processors: Selected Electronic Components," in "Price Measurements and their Uses," Vol. 57 of *National Bureau of Economic Research Studies in Income and Wealth,*, University of Chicago Press, 1993, pp. 103–124.

EE Times, "U.S., China at Odds on Fab-Gear Export," *EE Times*, April 1998.

Feenstra, Robert C., Benjamin R. Mandel, Marshall Reinsdorf, and Matthew Slaughter, "Effects of Terms of Trade Gains and Tariff Changes on the Measurement of U.S. Productivity Growth," *American Economic Journal: Economic Policy*, forthcoming.

Flamm, Kenneth, "Measurement of DRAM Prices: Technology and Market Structure," in "Price Measurements and their Uses," Vol. 57 of *National Bureau of Economic Research Studies in Income and Wealth*, University of Chicago Press, 1993, pp. 157–197.

___ , *Mismanaged Trade: Strategic Policy and the Semiconductor Industry*, Brookings, 1996.

___ , "The Microeconomics of Microprocessor Innovation," *mimeo*, 2007.

Foster, Lucia, John Haltiwanger, and Chad Syverson, "The Slow Growth of New Plants: Learning about Demand?," *mimeo*, 2012.

Fudenberg, Drew and Jean Tirole, "Learning-by-Doing and Market Performance," *Bell Journal of Economics*, 1983, *14* (2), 522–530.

Global Foundries, "Fact sheet: Fab 2 module 1," March 2009.

Grimm, Bruce T., "Price Indexes for Selected Semiconductors, 1974-96," *Survey of Current Business*, 1998, *78 (2)*, 8–24.

Hallak, Juan Carlos, "Product Quality and the Direction of Trade," *Journal of International Economics*, 2006, *68(1)*, 238–265.

___ and Peter Schott, "Estimating Cross-Country Differences in Product Quality," *Quarterly Journal of Economics*, 2011, *126(1)*, 417–474.

Holdway, Michael, "Quality-Adjusting Computer Prices in the Producer Price Index: An Overview," 2001. Accessed January 26, 2011.

Houseman, Susan, Christopher Kurz, Paul Lengemann, and Benjamin Mandel, "Offshoring Bias in U.S. Manufacturing," *Journal of Economic Perspectives*, 2011, *25(2)*, 111–132.

___ , ___ , Paul Lengermann, and Mandel Benjamin, "Offshoring Bias in U.S. Manufacturing: Implications for Productivity and Value Added," *Board of Governors of the Federal Reserve System, International Finance Discussion Papers*, 2010, (1007).

Hummels, David and Peter J. Klenow, "The Variety and Quality of a Nation's Exports," *American Economic Review*, 2005, *95(3)*, 704–723.

___ , Jun Ishii, and Kei-Mu Yi, "The nature and growth of vertical specialization in world trade," *Journal of International Economics*, 2001, *54*, 75–96.

Hurtarte, Jorge S., Evert A. Wolsheimer, and Lisa M. Tafoya, *Understanding Fabless IC Technology* 2007.

IC Knowledge, "Can the semiconductor industry afford the cost of new fabs?," March 2001.

Irwin, Douglas A. and Peter J. Klenow, "Learning-by-Doing Spillovers in the Semiconductor Industry," *Journal of Political Economy*, 1994, *102* (6), 1200–1227.

Khandelwal, Amit, "The Long and Short (of) Quality Ladders," *Review of Economic Studies*, 2010, *77(4)*, 1450–1476.

Klemperer, Paul, "The Competitiveness of Markets with Switching Costs," *RAND Journal of Economics*, 1987, *18* (1), 138–150.

___ , "Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade," *Review of Economic Studies*, 1995, *62*, 515–539.

Kugler, Maurice and Eric Verhoogen, "Prices, Plant Size, and Product Quality," *Review of Economic Studies*, 2011, *79(1)*, 307–339.

Kumar, Rakesh, "The Business of Scaling," *IEEE Solid-State Circuit Society News*, 2007, *12(1)*, 22–27.

___ , *Fabless Semiconductor Implementation*, McGraw-Hill Professional, 2008.

Leachman, Robert C., "Competitive Semiconductor Manufacturing: Final Report on Findings from Benchmarking Eight-inch, sub-350nm Wafer Fabrication Lines," *mimeo. University of California at Berkeley*, 2002.

Mandel, Benjamin, "Heterogeneous Firms and Import Quality: Evidence from Transaction-Level Prices," *Board of Governors of the Federal Reserve System, International Finance Discussion Papers*, 2010, (991).

McGregor, Jim, "The common platform technology: A new model for semiconductor manufacturing," *In-Stat*, 2007.

Moore, Gordon E., "Cramming More Components onto Integrated Circuits," *Electronics*, 1965.

Nakamura, Emi and Jon Steinsson, "Lost in Transit: Product Replacement Bias and Pricing to Market," *American Economic Review*, forthcoming.

Reinsdorf, Marshall, "The Effect of Outlet Price Differentials on the U.S. Consumer Price Index," in "Price Measurements and their Uses," Vol. 57 of *National Bureau of Economic Research Studies in Income and Wealth*, University of Chicago Press, 1993, pp. 227–254.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 1974, *82* (1), 34–55.

Schott, Peter, "Across-Product Versus Within-Product Specialization in International Trade," *The Quarterly Journal of Economics*, 2004, *119(2)*, 647–678.

Schott, Peter K., "The relative sophistication of Chinese exports," *Economic Policy*, 2008, *23*, 5–49.

Scott, A.J. and D.P. Angel, "The Global Assembly-Operations of U.S. Semiconductor Firms: A Geographical Analysis," *Environment and Planning A*, 1988, *20(8)*, 1047–67.

Shin, Jiwoong and K. Sudhir, "Switching Costs and Market Competitiveness: Deconstructing the Relationship," *Journal of Marketing Research*, 2009, *46* (4), 446–449.

Sorensen, Alan T., "Equilibrium Price Dispersion in Retail Markets for Prescription Drugs," *Journal of Political Economy*, 2000, *108(4).*

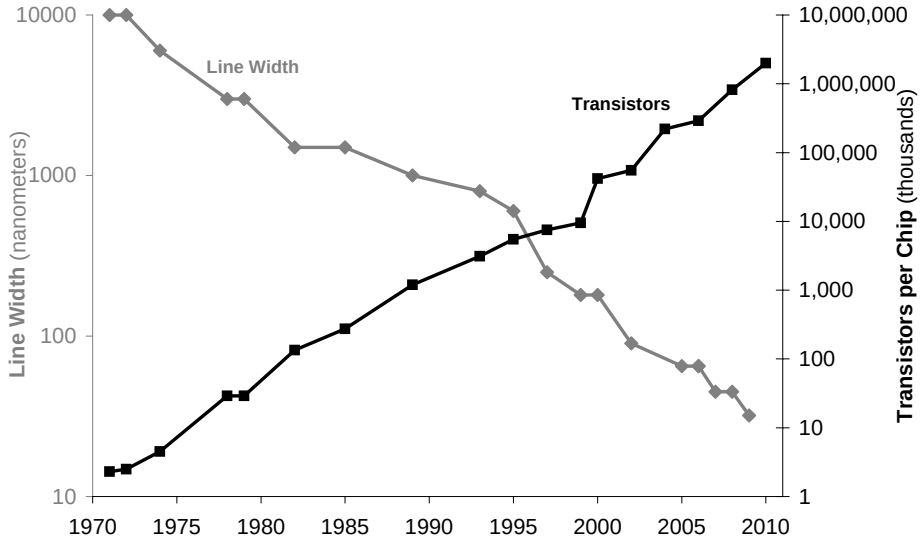
Stigler, George J., "The Economics of Information," *Journal of Political Economy*, 1961, *69*, 213–25.

Syverson, Chad, "Market Structure and Productivity: A Concrete Example," *Journal of Political Economy*, 2004, *112* (6), 1181–1222.

To, Theodore, "Multi-Period Competition with Switching Costs: An Overlapping Generations Formulation," *Journal of Industrial Economics*, 1996, *44*, 81–87.

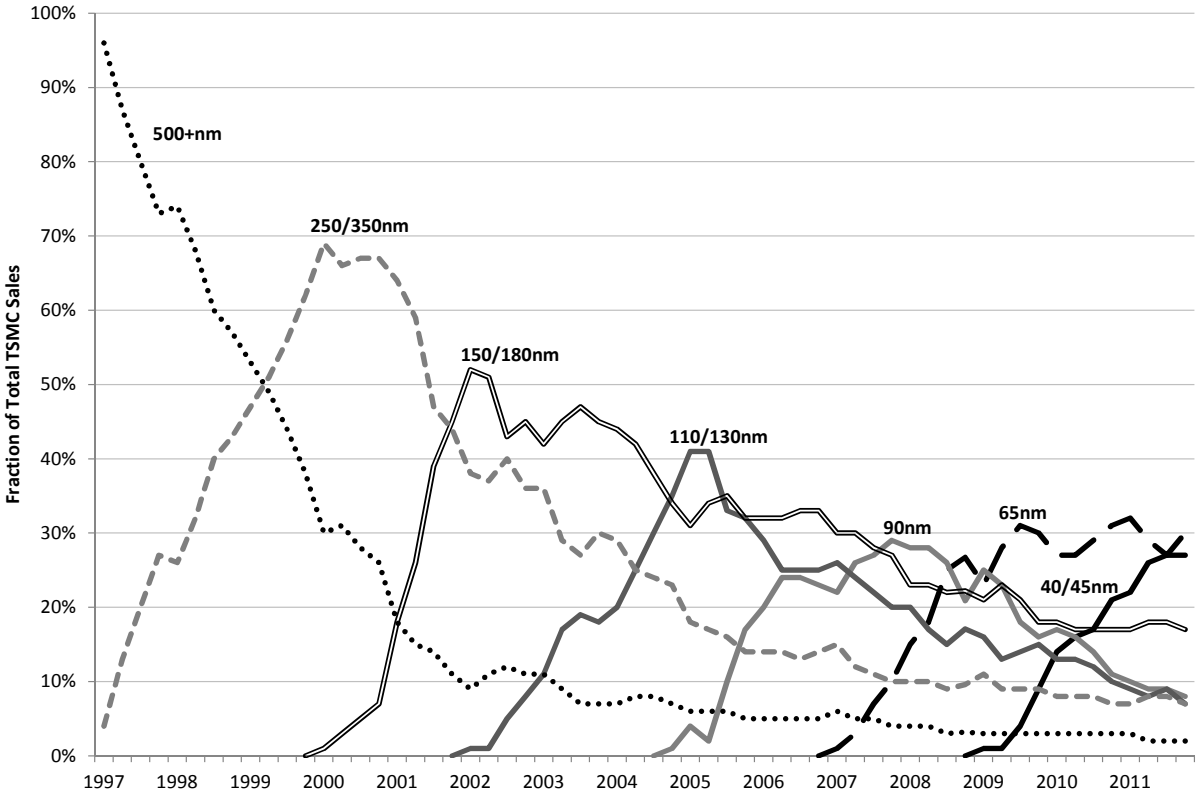
Turley, James L., *The Essential Guide to Semiconductors*, Prentice Hall PTR, 2003.

Figure 1: Moore's Law - Intel Processors



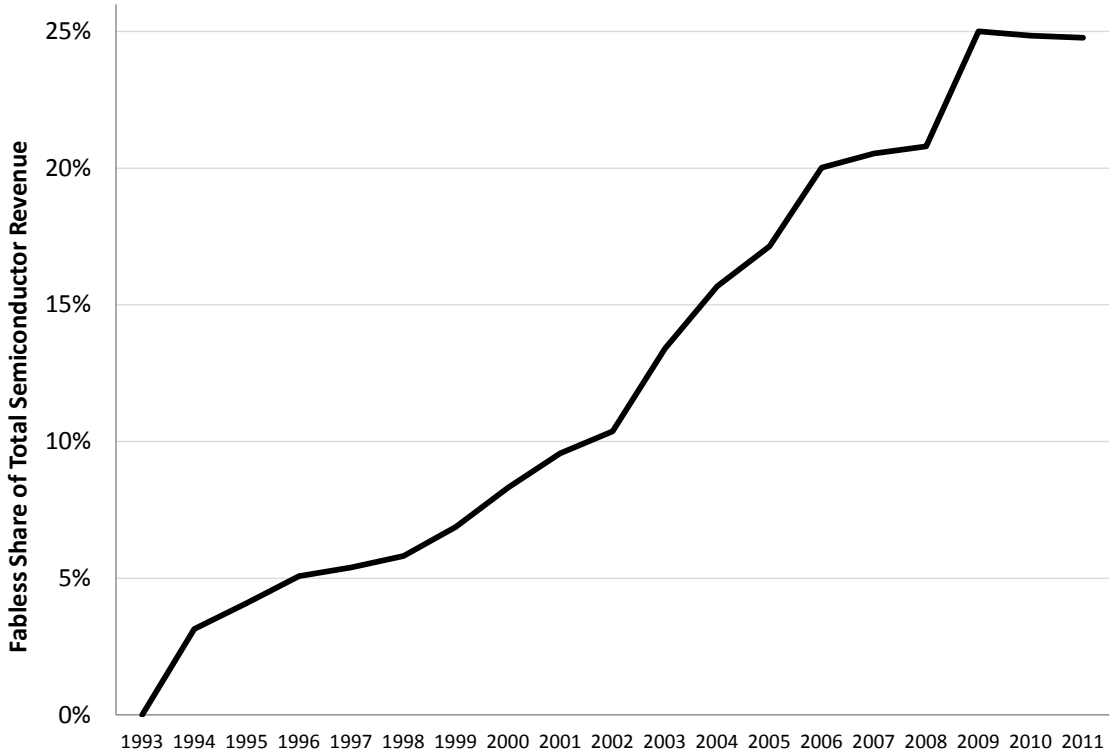
Sources: http://www.intel.com/technology/timeline.pdf
http://www.intel.com/pressroom/kits/quickreffam.htm

Figure 2: Technology Cycle - TSMC Sales by line width



Source: TSMC quarterly reports

Figure 3: Growth of the Fabless Business Model



Sources: Global Semiconductor Alliance (GSA) and Semiconductor Industry Association (SIA)

Figure 4: Firm A Best Response in Period 3

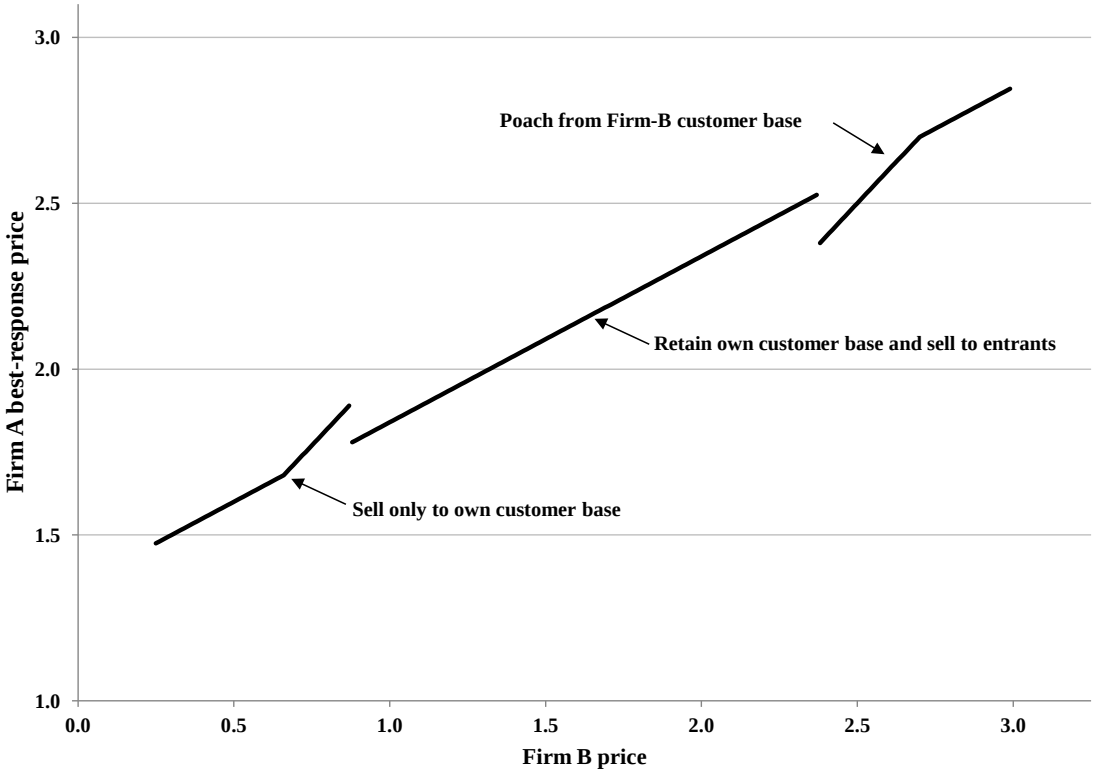


Figure 5: Closing Price Gaps Within Technology

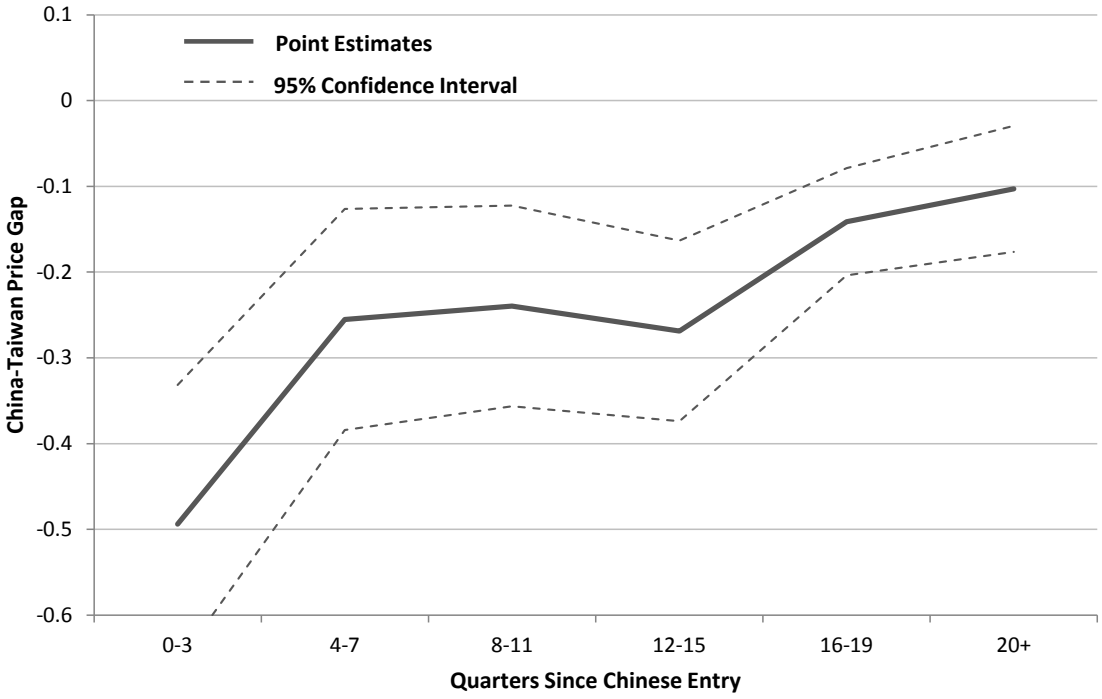


Figure 6: Index Calculation Example

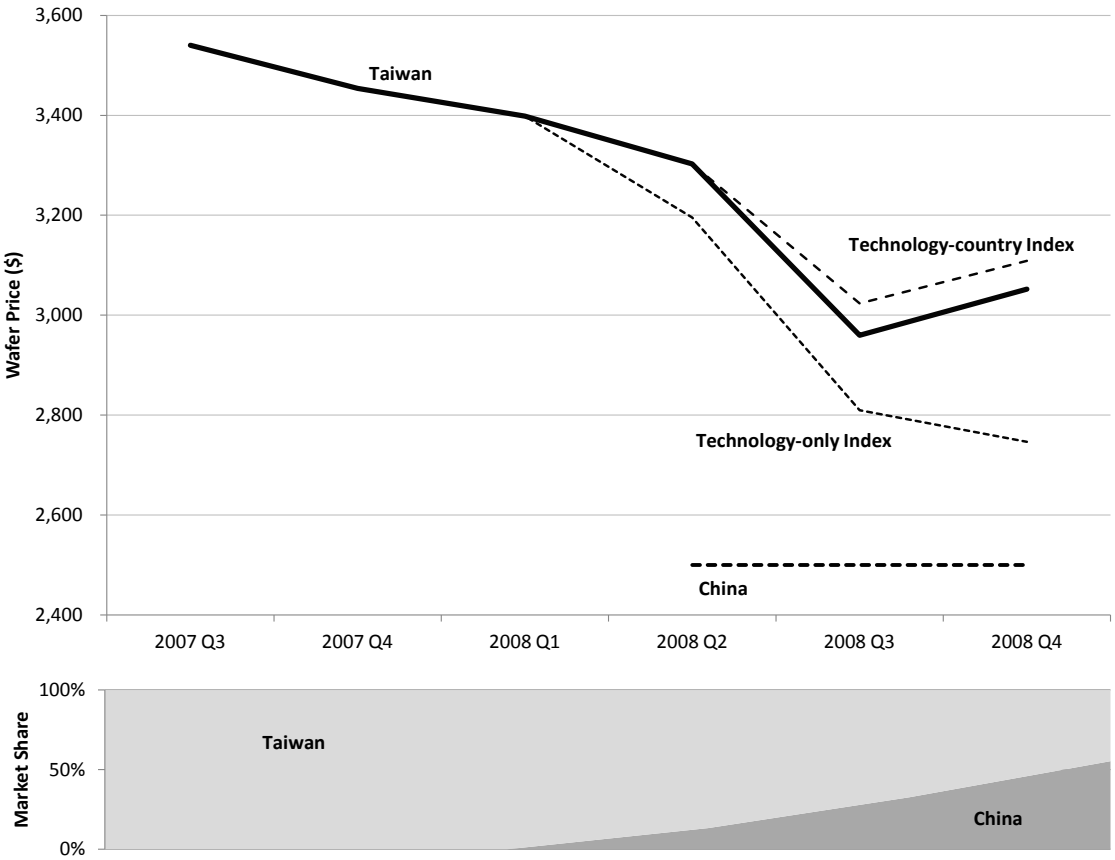


Figure 7: Matched-Model Price Index Results

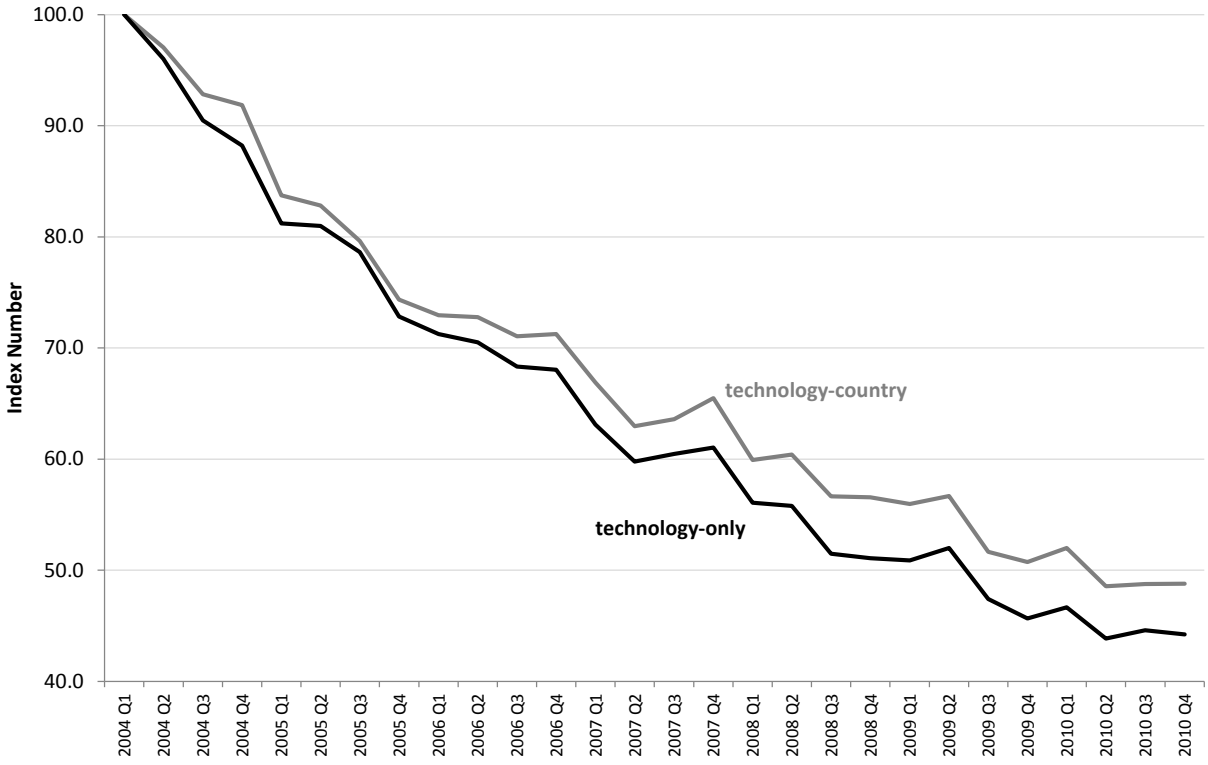


Table 1: Foundry Capacity by Country

Table 1. Foundry Capacity by Country

	Global Capacity									
	Foundry	Total Industry	Taiwan	China	Singapore	Europe	USA	Japan	S. Korea	Malaysia
2000	875	9,462	63.2%	7.1%	7.5%	6.8%	4.4%	7.8%	3.2%	0.0%
2001	972	8,286	60.5%	8.9%	8.4%	6.3%	4.2%	7.1%	2.9%	1.7%
2002	1,011	8,646	56.0%	10.7%	10.0%	6.2%	5.3%	7.0%	2.9%	1.8%
2003	1,150	9,018	51.1%	15.5%	10.1%	6.5%	5.1%	6.4%	3.2%	2.2%
2004	1,429	10,000	50.3%	19.0%	9.6%	5.6%	4.1%	5.3%	3.1%	3.0%
2005	1,739	11,073	48.5%	23.2%	9.2%	4.9%	3.6%	4.5%	3.0%	3.1%
2006	1,951	12,320	48.0%	24.1%	9.9%	4.4%	3.3%	4.1%	3.2%	3.0%
2007	2,157	13,588	48.9%	23.4%	10.0%	4.6%	3.1%	3.7%	3.5%	3.0%
2008	2,401	14,297	49.9%	22.1%	11.2%	5.1%	2.9%	2.9%	3.3%	2.7%
2009	2,546	14,058	49.2%	22.0%	10.9%	6.5%	2.8%	2.7%	3.5%	2.5%
2010	2,812	14,230	49.4%	21.5%	11.3%	7.4%	2.6%	2.5%	3.3%	2.1%
2011	3,177	14,923	50.2%	21.8%	10.7%	7.9%	2.4%	2.3%	3.0%	1.8%

Capacity measured in thousand 8-inch equivalent wafer starts per month.

Note: Includes pure-play foundries only.

Source: iSuppli

Table 2: Wafer Price Descriptive Statistics

			Yearly Means						
	Mean	Std. Dev	2004	2005	2006	2007	2008	2009	2010
Price Per Wafer ($)	1,204.22	949.28	1,321.01	1,293.51	1,298.54	1,189.68	1,120.34	1,080.39	1,126.09
Number of Wafers Contracted	11,865.00	22,302.74	8,344.61	14,912.17	10,544.97	16,767.34	9,260.24	9,692.42	13,533.24
Layers									
Metal Layers	4.60	1.77	4.22	4.56	4.85	4.76	4.62	4.56	4.64
Mask Layers	25.72	7.31	23.69	24.52	26.37	26.74	25.86	26.27	26.58
Polysilicon Layers	1.23	0.46	1.36	1.21	1.20	1.19	1.30	1.14	1.20
Wafer Size									
150 mm	0.16	0.37	0.18	0.17	0.15	0.10	0.17	0.16	0.17
200 mm	0.73	0.44	0.79	0.76	0.76	0.81	0.68	0.68	0.63
300 mm	0.11	0.32	0.03	0.07	0.09	0.09	0.15	0.16	0.20
Line Width									
45 nm	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.03
65 nm	0.03	0.16	0.00	0.00	0.00	0.01	0.04	0.07	0.08
90 nm	0.04	0.21	0.00	0.02	0.04	0.05	0.08	0.06	0.06
130 nm	0.15	0.36	0.12	0.16	0.18	0.17	0.14	0.14	0.15
150 nm	0.06	0.23	0.07	0.08	0.10	0.09	0.03	0.01	0.01
180 nm	0.26	0.44	0.25	0.23	0.24	0.28	0.26	0.31	0.26
250 nm	0.12	0.33	0.17	0.17	0.13	0.15	0.10	0.07	0.08
older vintage	0.33	0.47	0.40	0.35	0.30	0.25	0.34	0.35	0.33

6253 Observations

Source: Authors' calculations based on GSA Wafer Fabrication & Back-End Pricing Survey and iSuppli shipments data

Table 3: Hedonic Wafer Price Regression

dependent variable: log of price per wafer

Variable	Coefficient	Std. Err.	t-Stat
Foundry Location			
China	-0.186	(0.027)***	-6.87
Malaysia	-0.278	(0.042)***	-6.64
Singapore	-0.061	(0.016)***	-3.77
United States	0.068	(0.030)**	2.23
Wafer Size			
150 mm	-0.467	(0.032)***	-14.54
300 mm	0.671	(0.021)***	31.60
Line Width			
$\geq$ 500 nm	-0.245	(0.053)***	-4.61
350 nm	-0.167	(0.033)***	-4.99
250 nm	-0.061	(0.026)**	-2.34
150 nm	0.169	(0.027)***	6.21
130 nm	0.356	(0.018)***	19.80
90 nm	0.479	(0.032)***	14.77
65 nm	0.676	(0.030)***	22.49
45 nm	0.962	(0.062)***	15.62
Number of Metal Layers	0.076	(0.007)***	10.56
Number of Polysilicon Layers	0.027	(0.024)	1.10
Number of Mask Layers	0.005	(0.002)***	2.98
Epitaxial Layer Indicator	0.064	(0.037)*	1.72
log Number of Wafers Contracted	-0.056	(0.004)***	-13.18
R-squared	0.909		
Observations	6253		

Specification also includes quarterly indicator variables

non-CMOS production not included

Baseline case (omitted category) is Taiwan, 200mm, 180nm

Weighted using iSuppli shipment weights

Standard errors adjusted for 28 quarter clusters

* significant at 10% level, ** 5%, *** 1%

Table 4: Simulation Results

	$\ln(p^A / p^B)$		Pr(switching)	Firm A market share
	s > 0	s = 0		
Period 2	0.394	0.25	0.14	0.67
Period 3	0.186	0.25	0.00	0.55

Note: The column "s > 0" gives the log difference in prices in the model with a switching cost. The column, "s = 0", relates to the model with no switching cost, in which price dispersion purely reflects differences in product design. The column, "Pr(switching)" gives the share of Firm A's customer base that switches to Firm B. The column "Firm A market share" reports the Firm A's share of the total market (which consists of a continuum of buyers of mass two).

Table 5: Closing Price Gaps Within Technology

dependent variable: log of price per wafer

	Baseline		Closing price gap	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.
China	-0.198	(0.029)***	-0.494	(0.079)***
Time since Chinese entry				
4-7 quarters			-0.128	(0.026)***
8-11 quarters			-0.262	(0.028)***
12-15 quarters			-0.393	(0.043)***
16-19 quarters			-0.450	(0.040)***
$\geq$ 20 quarters			-0.478	(0.063)***
Interaction terms				
China * (4-7 quarters)			0.239	(0.102)**
China * (8-11 quarters)			0.255	(0.095)**
China * (12-15 quarters)			0.225	(0.089)**
China * (16-19 quarters)			0.353	(0.081)***
China * ($\geq$ 20 quarters)			0.391	(0.073)***
Technological controls from Table 3	X		X	
Quarter indicators	X		X	
R-squared	0.919		0.932	
Observations	5303		5303	

Specifications include quarter indicator variables, wafer size indicators, line width indicators, layer controls, and contract size.
Observations pertaining to Chinese and Taiwanese foundries only.
Technologies with no Chinese production omitted
non-CMOS production not included
Baseline case (omitted category) is Taiwan, 200mm, 180nm
Weighted using iSuppli shipment weights
Standard errors adjusted for 28 quarter clusters
* significant at 10% level, ** 5%, *** 1%

Table 6: Matched-Model Price Index Results

	technology-country index	technology-only index
2004 Q1	100.0	100.0
2004 Q2	97.0	96.0
2004 Q3	92.8	90.5
2004 Q4	91.9	88.2
2005 Q1	83.7	81.2
2005 Q2	82.8	81.0
2005 Q3	79.6	78.6
2005 Q4	74.3	72.8
2006 Q1	72.9	71.2
2006 Q2	72.8	70.5
2006 Q3	71.0	68.3
2006 Q4	71.3	68.0
2007 Q1	66.9	63.1
2007 Q2	63.0	59.8
2007 Q3	63.6	60.5
2007 Q4	65.5	61.0
2008 Q1	59.9	56.1
2008 Q2	60.4	55.8
2008 Q3	56.6	51.5
2008 Q4	56.6	51.1
2009 Q1	56.0	50.9
2009 Q2	56.7	52.0
2009 Q3	51.7	47.4
2009 Q4	50.7	45.7
2010 Q1	52.0	46.7
2010 Q2	48.6	43.9
2010 Q3	48.8	44.6
2010 Q4	48.8	44.2
2004	95.4	93.7
2005	80.1	78.4
2006	72.0	69.5
2007	64.7	61.1
2008	58.4	53.6
2009	53.8	49.0
2010	49.5	44.8
Avg. Yearly Change '04-'10	-10.4%	-11.6%

A Data

In this appendix, we describe the data sources and steps taken in the construction of the dataset used in our analysis.

A.1 Wafer Prices

As discussed in Section 3.1, our wafer price data come from a proprietary database of semiconductor wafer purchases from foundries, collected by the Global Semiconductor Alliance (GSA). We implement a number of data cleaning procedures before the main analysis.

First, we implement a variety of sample restrictions. We drop observations corresponding to engineering runs that occur during the design stage prior to volume production. We omit a few observations reporting the obsolete 100mm wafer diameter. As discussed in Section 2.1, we keep only observations corresponding to CMOS process technology, which dominates the foundry market, and omit other processes that are quite distinct and serve niche markets for high-power, defense, aerospace applications. We also keep only observations for wafers produced by so-called "pure-play" foundries, and omit transactions involving integrated device manufacturers (IDMs) that do both design and fabrication.

Table A.1 shows that 6,916 observations satisfy these sample restrictions. We drop an additional 663 observations for a variety of reasons. 576 observations do not report the location of production, 1 observation reports an implausibly large order (which distorts the price per wafer), and 86 observations report technologies (wafer-size, line-width combinations) for which other sources (see below) report no production in the reported country and quarter.

Finally, we also combine closely related line widths. Any line width greater than or equal to 500nm is combined into the 500nm code. 140 and 150nm widths are combined into the 150nm code. 80 and 90nm widths are combined into the 90nm code. 60 and 65nm widths are combined into the 65nm code. 40 and 45nm widths are combined into the 45nm code. In all of these cases, one of the combined widths is vastly dominant in the market, and it would be difficult to separately identify prices less prevalent line width with very few observations.

A.2 Semiconductor Wafer Shipments

Our price indexes use quantity information to weight observations across process technologies (though the qualitative nature of the results does not depend upon the choice of weighting scheme). The GSA data include quantity information for each shipment, but aggregating this information yields quite volatile aggregate quantity measures for each technology. Instead of using this quantity information from GSA, we employ data published by HIS iSuppli in their "Pure Play Foundry Market Tracker." This report is a global census of semiconductor foundries, including 91 fabs belonging to 20 companies. Annual and quarterly frequency data begin in 2000 and 2002, respectively. Characteristics provided for each fab include company of ownership, location, wafers shipped per month at full capacity, and diameter of wafers shipped. This allows us to construct capacity by region and wafer size (Table 1).

In addition, the report provides information on wafer shipments for specific technologies (line widths) by *company*, but not by *plant*. For 11 of the companies covered, this information is sufficient to construct the needed wafer size-by-line width-by-country weights for our analysis without further assumptions. In two cases, only one fab is active in the period we cover, and in nine cases, all the company's fabs employ the same wafer diameter and are located in the same country.

The remaining companies either have fabs in multiple regions, or fabricate wafers of multiple diameters, or both. In these cases, we first estimate wafer shipments by technology for each fab, which allows us to divide production between countries for each technology. To do this, we employ three additional resources: the partial information on the timing of technology introduction by plant from the IHS iSuppli report;

company information provided in public statments; and extensive discussions with iSuppli.[56] Specifically, the "Market Tracker" lists the technologies employed (without output shares) in each fab at the time of publication, which we have for previous vintages of the database beginning in Q1 2010. This information allows us to construct technology-by-country-by-quarter weights for an additional 2 companies. In these cases, each with two fabs in operation, we were able to identify one company fab exclusively employing a single category of legacy technology (500nm and above), leaving the remaining fab to account for the residual company production.

In the remaining 7 cases, to arrive at the weights, we need further information about the technology employed at specific plants. A search of company press releases and industry press reports yielded information on the timing of introduction of particular line widths at specific fabs in several cases, but information on the relative importance of each technology *by fab* is not available. To fill this gap we appeal to information on industry norms gathered from iSuppli and other reports and discussions with industry analysts. We assume that several rules hold in general: (1) fabs add new technologies progressively-adding a line width more advanced than all technologies used in previous periods; (2) fabs ramp up output of new technologies linearly over a four-quarter period; (3) companies introduced new technologies first at the company's most advanced fab; (4) when not ramping a new technology, fabs split production evenly among the 2-4 technologies in production. Individual companies often required deviations from these rules based on information regarding specific fabs to match the overall company technology mix.

Table A.2 shows the specific assumptions we made for each foundry in generating the weights, and we have made available the resulting set of weights by country, wafer size, line width, and quarter at:

`http://www.andrew.cmu.edu/user/bkovak/bkm_semicon_data/index.html`

We also use these data to calculate the time since Chinese entry for each production technology in Section 5.

[56] We thank Len Jelinek of iSuppli for invaluable assistance in understanding and supplementing the information in the report.

Table A.1: Dropped Observations

Total observations	6,916
Used in analysis	6,253
Dropped	663
Missing foundry location	576
Implausibly large order reported*	1
Inconsistent	86

There may be multiple reasons to drop an observation

* Confirmed with GSA

Table A.2: Technology Assumptions by Country

Company	# of Fabs	# of Countries	# of Wafer Diameters	Notes on Assumptions
Altis Semiconductor	1	1	1	
ASMC	3	1	2	Small diameter fab produces legacy technology. Large diameter fab produces remaining, more advanced technology.
CRMC	4	1	2	All fabs are same diameter until 2009. No assumptions needed. 2009-2011: small diameter fabs produce all reported legacy technology and amount of 350nm indicated by historical pattern. Remaining shipments from large diameter fabs.
Dongbu Hi Tek	2	1	1	
Episil	4	1	1	
Globalfoundries/ Chartered	11	2	3	2004-2008: Single country Small diameter fab capacity split evenly between 350nm and legacy technology. Large diameter fabs produce all shipments using 130nm and more advanced technology. Medium diameter fabs produce remaining, more advanced technology. 2009-2011: Begin with 2008 mix from existing locations. Ramp up 65nm and 45nm with timing indicated in reports. Residual is production in single remaining location.
Grace	2	1	1	
He Jian	2	1	1	
HHNEC	3	1	1	
HuaLi	1	1	1	
Landshunt Silicon Foundry GmbH	2	1	1	
Phenitec	3	1	1	
Silterra	3	1	1	
SMIC	10	1	2	All 150nm and less advanced technology produced at medium-diameter fabs. All 65nm and 90nm is produced at large-diameter fabs. Assign 130nm production based on guidance from industry analysts.
SSMC	1	1	1	
TowerJazz	5	3	2	Small diameter fab produces 350nm and legacy technology. Remainder split among other locations proportional to capacity.
TSMC	13	4	3	Production at one location known with certainty. Second set of fabs known to be divided between 250nm and 350nm, assumed to be split evenly. Third set of fabs assumed to be split evenly among 5 technologies (130nm, 150nm, 250nm, 350nm, 500nm) Small diameter fab accounts for nearly all legacy technology and small share of 350nm & 250nm. Small amount of legacy production in second location indicated by data. Large diameter fabs account for all production using 90nm and more advanced technologies. Residual capacity at these fabs split evenly between 130nm, 150nm, and 180nm until drawing down to minimal to offset ramp-up of 65nm technology. Residual goes to medium diameter production at remaining location.
UMC	12	3	3	Production at one location known with certainty. Small diameter fab split evenly between 350nm and legacy technology. Remaining 350nm and legacy technology and all 150nm, 180nm, and 250nm production from medium-diameter fabs. Large diameter fabs account for all production using 65nm and more advanced technologies. assumptions required to split 90nm-130nm between 200mm and 300mm. 2008-2011: Remaining medium-diameter capacity split evenly between 90nm and 130nm. Residual production using these technologies at large-diameter fabs.
Vanguard	2	1	1	
X-Fab Semiconductor Foundries AG	6	4	2	Split legacy technology production between small-diameter fabs in two locations proportional to capacity. 350nm at known location, medium diameter. Residual is implied at third location.

Notes: All companies require an estimate of the share of production using CMOS process.

B Details of the Terminal-Period Problem

This appendix provides analytical results for the model's terminal period problem. Throughout, we omit the period-3 subscript from the notation.

B.1 Four Cases

The analysis of Firm A's optimal policy in the terminal period is organized around two observations. First, Firm A can charge up to $p^B + s + \tau\hat{y}$ and still retain all of its customer base, whereas the firm must charge no more than $p^B + \tau$ to attract any new entrants. Therefore, if s is sufficiently large such that $s > \tau(1-\hat{y})$, then Firm A retains *all* of its customer base when it sets a price that is low enough to attract *any* new buyers. Otherwise, the firm may relinquish its relatively low-complexity old buyers and yet still compete for new higher-complexity entrants. Second, Firm A can charge up to p^B and still capture all new entrants, whereas it must charge no more than $p^B + \tau\hat{y} - s$ to capture any members of Firm B's customer base. Therefore, if s is sufficiently large such that $s > \tau\hat{y}$, then Firm A captures *all* new entrants when it sets a price that is low enough to poach *any* of Firm B's customer base. Otherwise, Firm A may be able to poach Firm B's highest-complexity customers even if it does not attract all new entrants.

This discussion indicates that the form of the optimal policy depends on how s compares to $\tau(1-\hat{y})$ and $\tau\hat{y}$. There are 4 separate cases (and one knife-edge case) to consider:

Case #1: The first is the case of a large startup cost in the sense that $s > \max[\tau(1-\hat{y}),\ \tau\hat{y}]$. This implies that i) if a firm participates in the new buyer market, it necessarily retains all old buyers and ii) if it poaches its competitor's customer base, it necessarily captures all period-3 entrants.

Case #2: In the second case, $\tau\hat{y} > s > \tau(1-\hat{y})$. This occurs if the startup cost is moderate and Firm A's customer base includes only the highest-quality buyers ($\hat{y}$ is above 1/2). Since these buyers are most reluctant to switch to Firm B, Firm A can charge them a relatively high price. It follows that, if Firm A lowers its price enough to attract any period-3 entrants, it will also retain its entire customer base. In addition, since $\hat{y}$ is high, Firm B's customer base also includes some relatively high-quality buyers, who pay the monitoring cost to purchase from Firm B. Therefore, Firm A may be able to poach some of its competitor's customers even if it does not sell to all new entrants.

Case #3: In the third case, $\tau(1-\hat{y}) > s > \tau\hat{y}$. This occurs if the startup cost is moderate and Firm A's customer base is relatively large ($\hat{y}$ is below 1/2). This means Firm A may capture high-quality new entrants and yet shed some of its lowest-quality old buyers. Thus, it does not necessarily retain its customer base if it attracts any new entrants. At the same time, since $s > \tau\hat{y}$, the price that Firm A must charge to poach is so low that it will also capture all new entrants if it poaches at all.

Case #4: Lastly, we consider the case $s < \min[\tau(1-\hat{y}),\ \tau\hat{y}]$. In this case the startup cost is small, such that Firm A may shed some of its old buyers and yet still compete for the higher-quality new entrants. Similarly, it also may poach the higher-quality members of Firm B's customer base even if it has not captured all new entrants.

Knife-edge Case: There is also a "knife-edge" case in which $s = \tau\hat{y} = (1-\hat{y})\tau$. If this holds, then the price at which Firm A just retains all old buyers ($p^A = p^B + s + \tau\hat{y}$) precisely coincides with the price at which Firm A begins to capture new entrants ($p^A = p^B + \tau$). This means that sales evolve smoothly from one regime to the next: as Firm A's price falls just below $p^B + s + \tau\hat{y}$, it does not acquire any more of its customer base (it has already captured all of them) but instead, its share of new entrants starts to rise. Moreover, the Firm A price at which it captures all new entrants ($p^A = p^B$) coincides with the price at which it begins to attract buyers in Firm B's customer base ($p^A = p^B - s + \tau\hat{y}$). Hence, in this case, Firm A's sales and profits will vary continuously as a function of its own price. This yields a smooth best-response function. As we will see, the same is not true of the best-response function outside of this knife-edge case.

B.2 Case #2: $\tau\hat{y} > s > \tau(1-\hat{y})$

The calibration we present in Section 4.1 is consistent with Case #2. We provide details of the terminal period problem for this case here. It is straightforward to apply similar logic to the other cases. In Case #2, Firm A selects a price within one of four mutually exclusive regimes:

Regime I: In this regime, Firm A charges $p^A \in \left(p^B + \tau\hat{y} + s,\ p^B + \tau + s\right]$, which means it sells exclusively to members of its customer base. Since y is uniformly distributed on the unit interval, sales to its old buyers are given by $1 - \frac{p^A - p^B - s}{\tau}$, and so its problem is to maximize $\frac{\tau - p^A + p^B + s}{\tau}\left(p^A - c^A\right)$. There is an interior solution to this maximization problem as long as $p^A \in \left[p^B + \tau\hat{y} + s,\ p^B + \tau + s\right]$, as we have assumed. It follows that the optimal price in the interior of this regime is $p^A = \frac{s + \tau + p^B + c^A}{2}$ and profits are $\tau^{-1}\left(\frac{s + \tau + p^B - c^A}{2}\right)^2$.

It is helpful to know the set of Firm B prices consistent with an interior solution in this regime. The Firm B price that induces Firm A to set $p^B + s + \tau\hat{y}$ satisfies $\frac{s + \tau + p^B + c^A}{2} = p^B + s + \tau\hat{y}$. This yields $p^B = \bar{p}_I^B \equiv -s - \tau(2\hat{y} - 1) + c^A$. The Firm B price such that Firm A wishes to set $p^B + \tau + s$ solves $\frac{s + \tau + p^B + c^A}{2} = p^B + \tau + s$, which yields $p^B = \underline{p}_I^B \equiv -s - \tau + c^A$. Hence, in Regime I, an interior solution obtains if $p^B \in \left[\underline{p}_I^B,\ \bar{p}_I^B\right)$.[57]

If Firm B's price is outside the range $\left[\underline{p}_I^B,\ \bar{p}_I^B\right)$, then Firm A's optimal strategy is to charge a price at the corner of the Regime I range. For example, if Firm A faces $p^B > \bar{p}_I^B$, its desired price is less than the minimum admitted under Regime I and would imply that Firm A captures a share of its customer base that is greater than 1. This is clearly impossible, so Firm A will instead charge the minimum price of $p^B + \tau\hat{y} + s$. A similar argument applies to $p^B < \underline{p}_I^B$. In that case Firm A charges the maximum price consistent with Regime I of $p^B + \tau + s$

Figure B1 shows Firm A's Regime I best response function for the calibration described in Section 4.1. Given the calibration, $\underline{p}_I^B = -0.3$ and $\bar{p}_I^B = 0.66$. Thus, for p^B between 0 and 0.66, Firm A's Regime I optimal price represents an interior solution, so the slope of that portion of the best response function is $\frac{1}{2}$. For p^B greater than 0.66, Firm A's Regime I optimal price is a corner solution such that $p^A = p^B + \tau\hat{y} + s$ and the slope is 1.

Regime II: In this region, Firm A selects the best price within the range, $\left(p^B + \tau\hat{y} - s,\ p^B + \tau\right]$. The price, $p^B + \tau$, makes the highest-quality new buyer indifferent between the two firms, whereas $p^B + \tau\hat{y} - s$ is just low enough to attract the highest-quality Firm B (old) buyer. Notice that the set of prices between $p^B + \tau$ and $p^B + \tau\hat{y} + s$ (the minimum feasible price under Regime I) is excluded from Regime II. This is for good reason: if Firm A set a price in this range, it would not attract any new entrants but would receive a lower price from each of its locked-in (old) buyers. Hence, it never chooses a price in this set. In other words, Firm A's demand schedule jumps downwards at $p^B + \tau\hat{y} + s$.

Since Firm A retains all of its customer base in Regime II, its problem is to maximize $\left((1-\hat{y}) + \left(1 - \frac{p^A - p^B}{\tau}\right)\right)\left(p^A - c^A\right)$. This has solution $p^A = \frac{(2-\hat{y})\tau + p^B + c^A}{2}$, which implies profits of $\tau^{-1}\left(\frac{(2-\hat{y})\tau + p^B - c^A}{2}\right)^2$.

We can again solve for the set of Firm B prices that is consistent with an interior solution in this regime. The Firm B price that induces Firm A to set $p^A = p^B + \tau$ is $p^B = \underline{p}_{II}^B \equiv -\hat{y}\tau + c^A$. The Firm B price such that Firm A sets $p^A = p^B + \tau\hat{y} - s$ is $p^B \equiv \bar{p}_{II}^B \equiv 2s + 2(1-\hat{y})\tau + \underline{p}_{II}^B$. Thus, in this regime, there is an interior solution if $p^B \in \left[\underline{p}_{II}^B,\ \bar{p}_{II}^B\right)$

As above, if Firm B's price is outside the range $\left[\underline{p}_I^B,\ \bar{p}_I^B\right)$, then Firm A's optimal strategy is to charge a price at the corner of the Regime II range. If $p^B > \bar{p}_{II}^B$, then Firm A charges the Regime II minimum price

[57] Of course, if both $\underline{p}_I^B$ and $\bar{p}_I^B$ are negative, then Regime I is immaterial. We assume that at least $\bar{p}_I^B > 0$ (and this is true under our calibration).

of $p^B + \tau\hat{y} - s$. If $p^B < \underline{p}_I^B$, then Firm A charges the Regime II maximum price of $p^B + \tau$. This Regime II best response function is shown in Figure B2 for the calibration described in Section 4.1.

Regime III: In this region, Firm A charges $p^A \in \left(p^B,\ p^B + \tau\hat{y} - s\right]$. Since it attracts a share of *both* new buyers and some of Firm B's higher-quality old buyers, Firm A's problem is to maximize $\left((1-\hat{y}) + \left(1 - \frac{p^A - p^B}{\tau}\right) + \left(1 - \frac{p^A - p^B + s}{\tau\hat{y}}\right)\hat{y}\right)\left(p^A - c^A\right)$. The optimal price is $p^A = \frac{\tau - (s/2) + p^B + c^A}{2}$ and profits are $\frac{2}{\tau}\left(\frac{\tau - (s/2) + p^B - c^A}{2}\right)^2$. It follows that the range of Firm B prices consistent with an interior solution is $\left[\underline{p}_{III}^B,\ \bar{p}_{III}^B\right)$, where $\bar{p}_{III}^B = \tau - (s/2) + c^A$ and $\underline{p}_{III}^B \equiv -2\left(\tau\hat{y} - s\right) + \bar{p}_{III}^B$. As in the other regimes, if p^B falls outside $\left[\underline{p}_{III}^B,\ \bar{p}_{III}^B\right)$, Firm A will charge corner solution prices $p^B + \tau\hat{y} - s$ or p^B. The Regime III best response function is shown in Figure B3 for the calibration described in Section 4.1. Note that for the calibration chosen, the Regime III interior region is quite small, indicating that there is only a small range of prices for which Firm A can poach from Firm B's customer base without having captured all new entrants.

Regime IV: In this regime, Firm A selects the best price in the range, $p^A \in \left(p^B - s,\ p^B\right]$. The lower end of this set, $p^B - s$, is the price at which even the lowest-quality buyer in Firm B's customer base is indifferent between the two suppliers. Since Firm A captures all new entrants in this regime, its problem is to maximize $\left((2-\hat{y}) + \left(1 - \frac{p^A - p^B + s}{\tau\hat{y}}\right)\hat{y}\right)\left(p^A - c^A\right)$. The solution is given by $p^A = \frac{2\tau - s + p^B + c^A}{2}$ and profits are $\tau^{-1}\left(\frac{2\tau - s + p^B - c^A}{2}\right)^2$. It follows that the range of Firm B prices consistent with an interior solution is $\left[\underline{p}_{IV}^B,\ \bar{p}_{IV}^B\right)$, where $\bar{p}_{IV}^B = 2\tau + s + c^A$ and $\underline{p}_{IV}^B \equiv 2\tau - s + c^A$. If p^B falls outside this range, Firm A will charge corner solution prices p^B or $p^B - s$. The Regime IV best response function is shown in Figure B4 for the calibration described in Section 4.1.

Global Best Response: In the best-response policy shown in Figure 4, note that the left segment corresponds to Regime I, the central segment corresponds to Regime II, and the right segment corresponds to Regime IV. For the calibration shown, regime III is dominated by the other regimes for all p^B. Finally, note that while the terminal-period best response functions shown here were derived based on the analytical results just discussed, they correspond exactly with the results of the numerical optimization procedure described in Section 4.1.

Figure B1: Firm A Period-3 Regime I Best Response

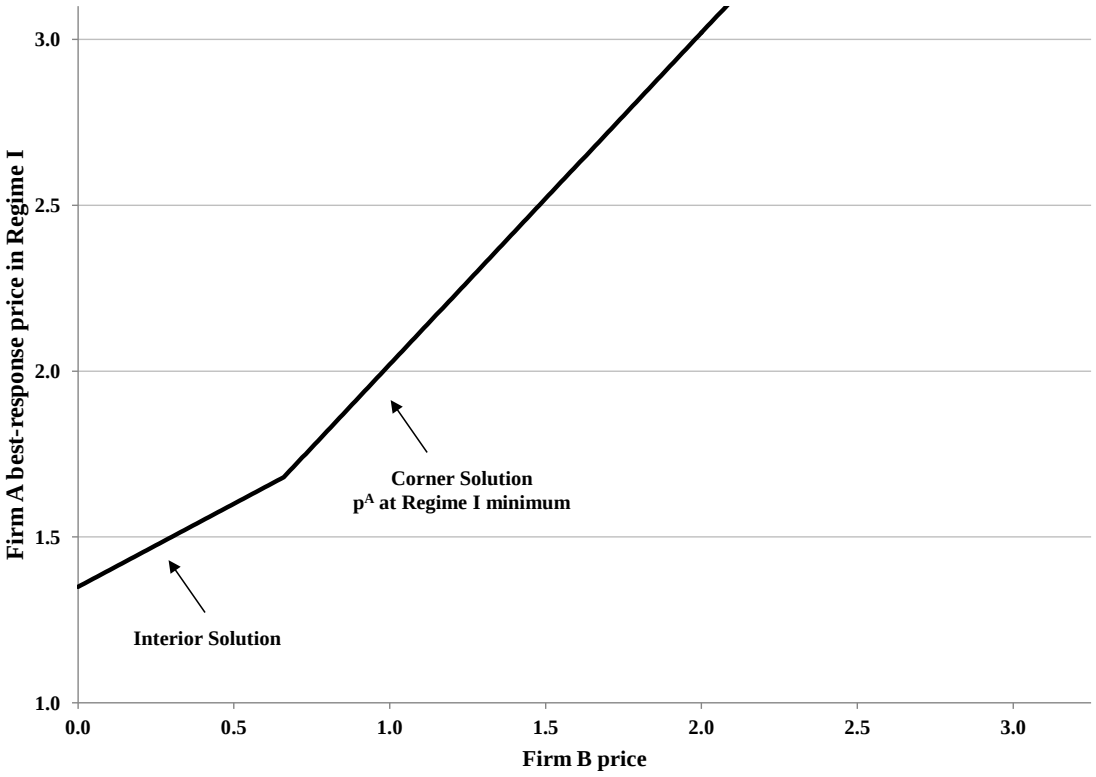


Figure B2: Firm A Period-3 Regime II Best Response

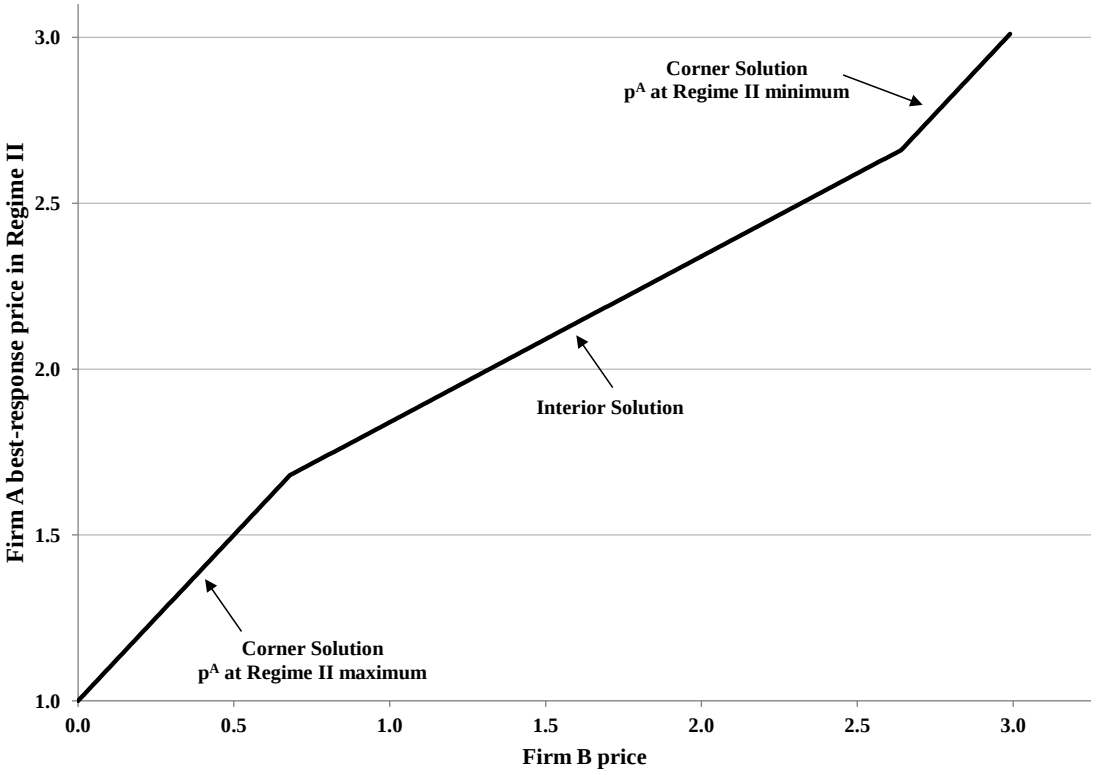


Figure B3: Firm A Period-3 Regime III Best Response

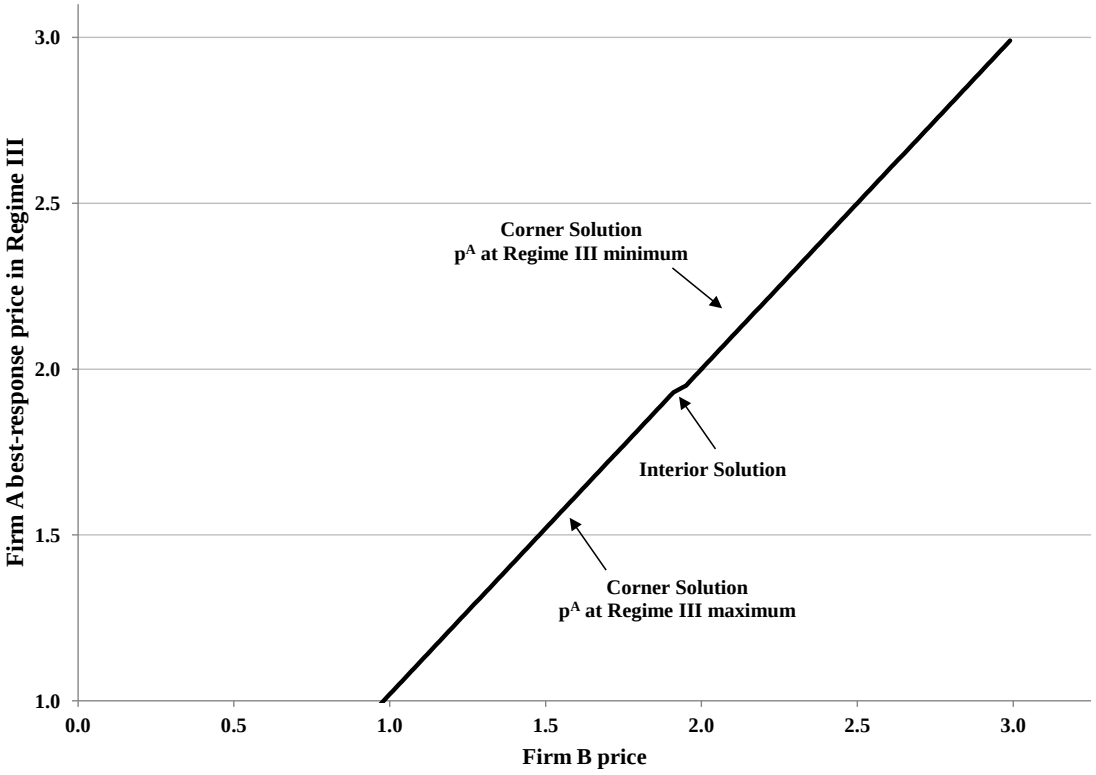


Figure B4: Firm A Period-3 Regime IV Best Response

