

Pricing and Product Sourcing 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

Initial Draft: October 30, 2009

This Draft: April 29, 2012

*Preliminary and incomplete
Please do not cite without permission.*

Abstract

Measuring the breadth of offshore outsourcing and its implications remains a focus of academics and practitioners. To this end, the present paper studies the offshoring of production and its effect on tradable intermediates prices in the market for semiconductor wafer manufacturing services. Specifically, the paper makes three contributions. First, we use a rich, proprietary dataset to document i) substantial constant-quality price differences across suppliers in different countries, and ii) shifts toward lower priced suppliers. Measuring prices for this exercise requires controlling for quality differences, and our database contains exhaustive product characteristic information necessary to implement direct quality adjustment techniques. Second, we introduce a model of price dispersion across suppliers that conforms to the features of intermediate input markets more closely than other models in the literature focusing on final goods markets. Third, we provide a formal accounting of the effect of offshoring on average semiconductor wafer prices, showing that shifting sourcing patterns accelerated the decline in the average wafer price, pushing it from 10.4 percent year to 11.6 percent year, with important implications for official index calculation procedures.

^{*} 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. Thanks to Yong-Gyun Choi, Manuel Gomez, and Steven Paschke for excellent research assistance. We would like to thank Bill Alterman, Ross Goodman, Susan Houseman, Marshall Reinsdorf, Jodi Shelton, Sonya Wahi-Miller, Kim Zieschung, and seminar participants at Carnegie Mellon and the U.S. Dept. of Commerce for helpful discussions and Chelsea Boone at GSA 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: Brian K. Kovak, Heinz College, Carnegie Mellon University. Email: bkovak@cmu.edu

1. INTRODUCTION

In the last half century, reductions in transportation and communication costs have driven a fundamental change in the spatial and institutional organization of manufacturing production. It is now commonplace for individual steps in the manufacturing process for a particular good 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: inputs are assembled or processed abroad and then, in certain cases, shipped back to the firm that designed the good. It is generally thought that the tasks performed overseas are increasingly done by relatively low-price suppliers, as less developed economies such as China capture a greater share of the global market for tradable intermediates.

Measuring the breadth of this trend in offshore outsourcing, and its implications, remains a focus of academics and practitioners. To this end, the present paper studies, for a particular market, the offshoring of production and its quantitative implications for tradable intermediates prices. Specifically, we focus on the market for semiconductor wafer manufacturing services. Drawing on a variety of rich, proprietary data, we document substantial shifts in production across geographic areas. We find that the suppliers that captured a greater share of the market typically charge much lower prices for observationally equivalent goods. Estimates from hedonic regressions reveal substantial constant-quality price differences across suppliers in different countries.¹ Although quality adjustment is always difficult, we stress that our data set includes information on those product characteristics that are considered by industry participants to be the most salient indicators of product quality.

We interpret these findings within a model of costly switching across suppliers (see Klemperer 1995). In this setting, the industry “leader” is the first to produce a particular good and preserves a majority of the market even after entry by other lower-priced suppliers, “followers,” because of the cost of relocating production to other facilities. Even allowing for unobserved heterogeneity across suppliers, the model shows that this friction amplifies price differences across suppliers. The high cost of switching locks in buyers, giving the industry leader an incentive to charge its customers high prices. Because products are finitely lived and

¹ 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,” in investigating the relationship between price and quality.

there is a staggering of customers entering the market over time, the leader must compete more aggressively for market share as its locked-in customers phase out. This implies a price differential across suppliers that is declining over time, a pattern that is consistent with our data but not present in a version of the model without frictions in switching suppliers.

Our findings have important implications for at least two growing areas of research. First is the growing literature on trade and product quality. We contribute new estimates of quality-adjusted price dispersion for a significant international market. While the prior literature has focused on final goods, this paper investigates tradable intermediates and demonstrates the relevance of product market frictions for price dispersion and quality measurement in these markets. Second, it informs the emerging research on the implications of offshoring for the accuracy of official aggregate price indexes. The shift of production overseas can pose challenges to measurement, and we provide direct evidence on this in the context of semiconductor production.

The semiconductor industry is frequently a target of academic study. This is, in part, because it has posted stunning rates of price *deflation* and productivity growth over the last 20 years.² It is a fascinating industry for those interested in fragmentation, with relatively low-skilled tasks, such as final assembly, being offshored as early as the late 1960s.³ Since the late 1980's, the industry has witnessed the offshoring of relatively more skill-intensive production tasks, such as the fabrication of computer chips on silicon wafers, and relatively low-price suppliers in China and Singapore have entered this market in recent years.

This recent trend in the wafer fabrication industry raises a more notable challenge to measuring prices in an offshore market. The rapid quality growth that has contributed to the stunning rates of price decline also complicates the accurate measurement of wafer prices. Yet as overseas production becomes increasingly sophisticated, researchers and practitioners must grapple with quality adjustment for tradable intermediates. Accordingly, measurement challenges in the wafer fabrication market likely preview similar challenges that will emerge in other industries as further offshoring of high-tech production occurs.

² For estimates of semiconductor price indexes, see Flamm (1993), Grimm (1998), and Aizcorbe (2002). Oliner and Sichel (2000) discuss the contribution of productivity growth in the semiconductor sector to the acceleration in aggregate productivity growth in the 1990s. Oliner, Sichel, and Stiroh (2007) update this work to include a look at the trends in the 2000s.

³ See Brown and Linden (2005) for a detailed discussion of the history of offshoring in this sector.

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. We focus on the market for contract semiconductor wafer fabrication services provided by firms called “foundries.” This market differs in important ways from the memory and processor markets studied in most prior work, with a strong emphasis on custom designs for particular applications rather than for general purpose computing. The custom nature of the products will be important for interpreting our results.

In Section 3, we investigate the differences in prices across different overseas manufacturers. This exercise requires “apples to apples” comparisons, that is, comparing prices for goods of the same quality. Our discussion of fabrication technology from Section 2 delineates the key indicators of product quality for which we must control. We are able to do this in a hedonic regression setting using survey data collected by an industry trade association, the Global Semiconductor Alliance (GSA). Accounting for quality receives substantial attention in domestic price measurement, but this paper is one of few studies to estimate the effect of offshoring using data that permits a direct accounting of cross-country dispersion in quality. We find that, even after controlling for quality using transaction-level micro data, prices in this offshore market still differ substantially across suppliers. During our sample period, 2004-2010, Chinese prices were 18.9% lower than Taiwanese prices for otherwise identical wafers.

In Section 4, we seek to understand how such large quality-adjusted price differences could persist in this market without being arbitrated away by customers quickly shifting toward suppliers with lower quality-adjusted prices. We introduce a model featuring sequential entry of leading and lagging suppliers and substantial fixed costs of starting production with a given supplier, both of which correspond to the foundry market. The model predicts quality-adjusted price dispersion across suppliers and closing price gaps between the leading and lagging suppliers. In Section 5, we confirm this prediction in the data.

In Section 6, we briefly discuss the implications of our results. As mentioned above, our work engages recent research on how product quality differences across countries affect patterns of international trade. We view our work as highly complementary to this line of research. While the literature has developed a number of tools appropriate to final goods markets, we focus on a market for intermediate inputs. In intermediate input markets, love of variety and

informational search frictions are less plausible as sources of price dispersion than they are in final goods markets.⁴ Instead, buyers and sellers tend to forge long-lived relationships that appear to be consistent with large costs of switching suppliers. In this context, recent approaches to quality measurement based on a supplier's market share will be confounded by the dynamic considerations raised in our model, stemming from sequential entry and fixed startup costs.

Our findings also complement the emerging literature focusing on the effects of offshoring of production on measures of input prices and productivity growth produced by government statistical agencies. Along with Houseman et al. (2011), we argue that price measures produced by most government statistical agencies, such as the BLS Import Pricing Program, implicitly assume that price differences across suppliers reflect quality differences rather than pure price dispersion. Such measures produce upward-biased measures of input price growth and productivity growth in the presence of quality-adjusted price differences when customers shift to lower-priced suppliers. We use the GSA data to show that shifts toward lower-priced suppliers accelerated the average annual price decline by 1.2 percentage points per year.

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

⁴ Recent papers utilizing variations on this approach include Hallak and Schott (2011), Hummels and Klenow (2005), Khandelwal (2010), and Kugler and Verhoogen (2011).

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.⁵ 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, which is 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 then cut up, leaving individual chips, called “die”. The die are then 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).

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. This makes chips of a given functionality smaller, lighter, faster, and more energy efficient, and also makes it feasible to

⁵ Turley (2003) provides an accessible overview of semiconductor technology, manufacturing, and business.

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, the Intel cofounder's famous observation that the number of transistors on a chip doubled every eighteen months (Moore, 1965).⁶ 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 will result 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, and 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 2008.⁷ 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. We 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).

⁶ This regularity later slowed to doubling every two years.

⁷ Share of wafer starts reported in SICAS Semiconductor International Capacity Statistics.

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, number of polysilicon layers, and number of metal layers.⁸ 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. By the mid-1970s, firms began moving packaging and test operations to East Asia to take advantage of lower input costs (Scott and Angel, 1988; Brown and Linden, 2005). In spite of outsourcing these relatively simple steps in the production process, firms maintained the more complex wafer fabrication operations in house. Firms that perform both design and wafer fabrication are referred to as integrated device manufacturers (IDM).

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, 2000; 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 et al., 2007). Around the same time, new firms sprung 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 companies, which have little or no in-house wafer manufacturing capability, are called "fables"

⁸ See <http://www.icknowledge.com/>. Polysilicon and metal layers are used in the construction of transistor "gates" and interconnections.

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. 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. In 2000, the majority of foundry capacity was already in Asia, mainly Taiwan. Since then, the share of capacity in Asia as 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 1989, Grimm 1998, Doms, Aizcorbe and Corrado 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. These Application-Specific Integrated Circuits (ASICs) have been the subject of limited previous research and differ from memory and processors in important ways.⁹ 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

⁹ 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 Kurshid 2002). Aizcorbe (2002) argues for using a combination of prices of MPUs, memory chips, and average product prices as a proxy for the devices produced by foundries.

the leading edge.¹⁰ 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 for a detailed discussion).

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 organization. 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 2004 and accounts for a representative sample of about 20 percent of the wafers processed by the foundry sector worldwide.

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 will allow 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.

¹⁰ Note the long persistence of technology nodes in Figure 2.

As mentioned in Section 2, in an effort to focus our analysis on the application-specific integrated circuit (ASIC) market that dominates foundry production, we only analyze CMOS wafer transactions, omitting 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. 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%).¹¹ 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 2.¹² 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.¹³ 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 were first introduced in 2000 and the share for this emerging technology rises from 3.6 percent of contracts to 20 percent of contracts over the survey. Similarly, new generations of lithography increase in share over time: 90 nanometer technology reached volume production in the overall semiconductor industry in 2004 and slowly gained share in the foundry market, ending at 7 percent in 2008; 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 incredibly 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

¹¹ Authors' calculations based on market data from iSuppli.

¹² See Appendix A for a more detailed description of the database and data cleaning.

¹³ See Appendix A for more detail on how weights were constructed.

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, the number of metal, polysilicon, and mask layers, an epitaxial layer indicator¹⁴, and the quantity of wafers purchased in the transaction. 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, a 300mm wafer is 67% more expensive than an otherwise identical 200mm wafer produced in the same quarter. Similarly, a wafer with 150nm line width was 17% 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. A wafer produced in China sells at a 19% 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.¹⁵

These large price differences could reflect unmeasured quality differences across suppliers that are not captured by our regressors. In what follows, we detail why this is unlikely to be the whole story. We first review the albeit limited data on one of the few sources of quality differences not included in our hedonic regression: differences in yields across countries. We find no evidence of systematic differences in yields. This corroborates our conversations with fabless firms, who have told us they receive similar service from the major Taiwanese and Chinese foundries in terms of both timeliness, and yields.¹⁶

¹⁴ An epitaxial layer is a thin ultra-pure layer of silicon crystal deposited on the surface of the raw silicon wafer that improves transistor performance.

¹⁵ 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.

¹⁶ 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.

Data on yields are very difficult to obtain, as producers and buyers are even reluctant to release yield information even for older production technologies.¹⁷ 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 proved too rough to be useful. Nearly all responses reported 76-100% yield, and the question was dropped from the survey. This does, however, rule out very large yield differences across suppliers. At our request, GSA added a continuous yield measure to the latest quarterly survey, allowing respondents to enter any number from 0 to 100%. Reflecting the reluctance to report detailed yield information, 57% of responses either did not report the yield or reported 0% yield, which is implausible for production wafer runs. 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.¹⁸ However, in both cases the estimates are very noisy, and large differences often cannot be ruled out. This analysis is consistent with the anecdotal evidence against cross-country yield differences, but the large amount of non-response curtails our ability to make strong statements regarding yield differences.

Another potential source of differences in quality across suppliers relates to the intellectual property and design tools 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, we are unable to examine this hypothesis directly.

That said, we remain skeptical that the price differences are entirely due to unobserved heterogeneity because there is a well defined and plausible alternative explanation. This alternative focuses on the existence of substantial costs of switching suppliers once production has begun at a particular foundry. In general, as shown by Klemperer (1995), this kind of friction can support "real" price dispersion by impeding arbitrage across suppliers. As we

¹⁷ 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.

¹⁸ Tables available upon request.

discuss in the following section, this notion of switching costs is consistent with important features of the wafer market, and may apply to other input markets as well.

4. A MODEL OF PRICE DISPERSION ACROSS INTERMEDIATE INPUT SUPPLIERS

In this section, we solve a simple price-setting game between two suppliers that allows us to address the relationship between unobserved heterogeneity across products and the presence of fixed costs of starting production with a particular supplier. The model considers the market for a given wafer technology, i.e. both suppliers produce the same wafer size and line width, though the model also allows for unobserved product quality differentiation within this market. We assume that one supplier enters the market before the other, and that there are large fixed costs of initiating production with a supplier.

The model yields two principal results. First, the observed difference in wafer prices across suppliers reflects a combination of unobserved heterogeneity and real price dispersion resulting from the startup cost. Second, the model shows that the startup costs within this market generate a pattern of price dynamics that is absent in a model with a constant unobserved quality difference. In Section 5, we show that this implication for the dynamics is consistent with the data.

4.1. Intuitive Argument and Supporting Features of the Industry

We begin with an intuitive description of the mechanism that we believe is a likely source of price dispersion in the semiconductor industry. One supplier (the “leader,” in this context Taiwan) enters the market first and has the market to itself for some time. When the lagging supplier (China) enters, even if it charges a lower price, it does not capture the leader’s customers due to the presence of large startup costs that would need to be repaid in order to switch to the lagging supplier. The leading supplier’s customers are partially “locked in” by having paid the startup cost with the leading firm. The lagging supplier is only able to capture newly entering customers who would have to pay the startup cost at either firm. These customers enter later in the cycle because their final products do not require the latest fabrication technology. Thus, the combination of sequential entry and large startup costs can maintain real price differences across suppliers.

However, each design has a finite lifespan. As the leading firm's customers' products phase out, the number of locked-in products declines and the leading firm must reduce its price relative to the lagging firm to successfully compete for newly entering customers, so the price gap between the leader and follower should decline as a given production technology ages.¹⁹ We formally model this mechanism below, but first discuss how the model's central assumptions relate to features of the contract semiconductor manufacturing (foundry) industry.

We restrict attention to the two largest foundry supplier countries, Taiwan and China (see Table 1). Taiwanese foundries consistently enter the market for a given process technology well ahead of Chinese foundries. In our data, Chinese foundries first begin volume production of a given process technology (wafer size, line width combination) at least 8 quarters after Taiwanese entry.²⁰ In the context of the model, Taiwanese foundries are the leading suppliers and Chinese foundries are the followers. Although the source of this consistent delay in Chinese entry is not central to our argument, it likely occurs for a number of reasons. Chinese foundries may adopt the 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 the U.S., European, and Japanese firms that dominate that market. These restrictions include the multilateral Wassenaar Arrangement and strong restrictions implemented by the Taiwanese Government (Electrical Engineering Times 1998) that restrict the trade of dual-use technologies.²¹

Even after Chinese entry at a substantially lower price, Taiwanese foundries still maintain a large market share. In order for the law-of-one-price to fail, there must be some friction impeding arbitrage across suppliers. In the market for semiconductor wafer manufacturing services, we believe that this friction comes primarily from very large fixed costs of starting production of a particular design at a particular supplier. For example, masks are an essential part of the semiconductor manufacturing process (see Section 2), and any change in the

¹⁹ We thank Ken Flamm for initially suggesting that our argument implies closing price gaps over time.

²⁰ Authors' calculations using iSuppli data. Observable delays range between 8 and 12 quarters, though Taiwanese entry is censored for many technologies.

²¹ 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.

semiconductor design requires the production of a new set of masks. Each foundry has its own proprietary processes and technologies that are generally incompatible across manufacturers, implying 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.”²²

Other examples of startup costs include the many chemical and mechanical adjustments and calibrations to the 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 production 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 by senior management.²³

Together, these fixed startup costs create a situation in which firms almost never switch foundries for a particular design.²⁴ Rather, a firm will begin working with a new supplier only for a completely new product or when producing a new version of an existing product in which design changes would require repayment of the various startup costs anyway. We integrate these observations into the model by explicitly modeling a large startup cost for producing a given design at a particular supplier and restricting attention to equilibria in which customers never choose to switch an existing design from one supplier to another. We will show that the lack of switching can be maintained in equilibrium as long as the startup cost is sufficiently large.

²² 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.

²³ A summer 2002 article in the magazine, *Electronics Design Chain*, gives a good example of this. It discusses 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.

²⁴ A design manager at LSI Corp, a major U.S. fabless firm, said he was aware of only one instance in his career when his firm shifted a product between foundries. Executives at a large foreign fabless firm also told us that they only switched foundries when their original partner has reached full capacity and was not able to fill their orders quickly enough.

4.2 Formal Model

We now describe the formal model, sketch the solution, and present main results. The detailed derivations are given in Appendix B.

4.2.1 Setup

The model has three periods. There are two types of agents in the market – buyers and manufacturers of an intermediate good. There are overlapping generations of buyers. A cohort of buyers of mass 1 enters in each of the three periods and remains in the market for two 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. We assume that buyers have inelastic demand over the relevant range, such that they will purchase the intermediate good from one of the suppliers.

Even though buyers are purchasing 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 design. To allow for this form of heterogeneity, we think of the *complexity* of the final good as varying across buyers (i.e. certain chip designs are more difficult to fabricate than others). Formally, we follow Klemperer (1995) and assume that final good quality, y , is distributed uniformly from 0 (lowest quality) to 1 (highest quality). We now detail the implications of this dispersion in final-good complexity for the price-setting problem between input suppliers.²⁵

Turning to the manufacturers, Firm A (Taiwan) is the leader – it is present in the market from period 1 onward. Firm B (China) is the follower and joins the market in period 2. We assume that Firm A is the more technologically sophisticated of the producers. For example, in the wafer market, it is typically thought that Taiwan's fabrication plants provide design tools and intellectual property that enable them to produce highly complex designs more easily. Buyers who purchase from Firm B pay a premium, τy (where $\tau > 0$), per period that represents the costs of monitoring and consulting with the manufacturer. The critical assumption is that these costs are increasing in the complexity of the design.

²⁵ Note that since the econometrician does not directly observe the degree of complexity, the dispersion of complexity represents a form of unobservable heterogeneity.

The heterogeneity across final goods raises the question of whether firms are permitted to price-discriminate according to y . Following Klemperer (1995) and the related literature on switching costs, we do not allow price discrimination. Our approach is also informed by the wafer supplier contracts that we have read, each of which explicitly limits the supplier’s freedom in charging appreciably different prices across its customers.²⁶ Without price discrimination, each firm charges a single price in each period. Let p_t^A denote Firm A’s price in period t and p_t^B denote Firm B’s price in period t .

The relative sophistication of Firm A renders it a clear advantage. At the same time, we assume that Firm B enjoys a lower marginal cost of production to balance that advantage. Specifically, we assume that both firms have constant marginal costs of production, and that $c^A > c^B$.

Lastly, and perhaps most importantly, there is a cost, s , of beginning production with a particular supplier. Thus, if a buyer purchased from Firm A in period 1 but switches to Firm B in period 2, it must repay the startup cost s (independent of its quality), plus monitoring cost τy and Firm B’s price. As long as s is sufficiently large, it will be optimal for customers to avoid switching the production of a particular design from one supplier to another (see Appendix B for the particular restriction on s).

We solve the model by backward induction beginning in period 3, focusing on equilibrium in which customers do not switch suppliers, reflecting the stylized facts in the industry.

4.2.2 Period-3 Problem

The critical feature of the model is that the demand facing each firm depends on its customer base, due to the presence of the startup cost, s . Define \hat{y} as Firm B’s share of the customer base as of the start of period 3, i.e. any entrant in period 2 with product $y \in [0, \hat{y}]$ chose Firm B and any entrant with a product $y \in [\hat{y}, 1]$ chose Firm A in period 2. We show below that such a partition exists in the period-2 problem.²⁷

²⁶ See, for instance, <http://contracts.corporate.findlaw.com/operations/sales/403.html>, which gives a TSMC contract. Of course, deriving this feature of pricing contracts as a profit-maximizing outcome would be preferable to simply imposing it. Providing a more substantive microfoundation for our assumption is an interesting extension for future work.

²⁷ The presence of a unique threshold, \hat{y} , implies that the two firms “split” the market, with Firm A selling to more complex final goods producers. Hence, this model is not designed to generate dispersion across prices in the same period *for given* y . One could instead consider a model which suppresses heterogeneity and trace through the implications of a startup cost (and we have begun to do this). Our purpose here was to give the benefit of the doubt

First, consider Firm A's problem in period 3. Firm A will attract a period-3 entrant if $p_3^A < p_3^B + \tau y$, since the new entrant must pay the startup cost at either firm. Using $y \sim U[0,1]$, the share of period-3 entrants choosing to purchase from Firm A is

$$\Pr[p_3^A < p_3^B + \tau y] = 1 - \frac{p_3^A - p_3^B}{\tau}. \quad (1)$$

Firm A retains a customer in its customer base ($y \in [\hat{y}, 1]$) if $p_3^A < p_3^B + \tau y + s$. Note that Firm A has an advantage with its returning customers, who have previously paid the startup cost with Firm A, but would need to repay it to switch to Firm B. The share of its customer base that it retains is

$$\Pr[p_3^A < p_3^B + \tau y + s \mid y \in [\hat{y}, 1]] = 1 - \frac{\frac{p_3^A - p_3^B - s}{\tau} - \hat{y}}{1 - \hat{y}}. \quad (2)$$

Finally, Firm A attracts a customer in Firm B's customer base ($y \in [0, \hat{y}]$) if it charges a price sufficiently low to offset the startup cost: $p_3^A + s < p_3^B + \tau y$. The share of Firm B's customers that Firm A is able to "poach" is then

$$\Pr[p_3^A + s < p_3^B + \tau y \mid y \in [0, \hat{y}]] = 1 - \frac{\frac{p_3^A + s - p_3^B}{\tau}}{\hat{y}}. \quad (3)$$

As shown in the appendix, (1)-(3) imply that Firm A faces the demand curve shown in Figure 4, where Y_3^A is period-3 sales by Firm A. The red portion of the demand curve corresponds to the case where Firm A charges a price so high that it only sells to its own customer base. The blue portion reflects intermediate prices in which Firm A retains all of its own customer base and sells to some of the entering cohort in period 3. Finally, the green portion reflects prices that are sufficiently low for Firm A to retain its customer base, capture all new entrants and poach some of Firm B's customer base. The kinks in the demand curve arise because of the discrete startup cost. For example, the price $p_3^B + s + \tau \hat{y}$ is just low enough for Firm A to defend its period-2 customer base, so its output is $1 - \hat{y}$. To compete for period-3 entrants, however, the firm must drop its price discretely. This is because the new entrants are not "locked in" by the discrete startup cost, which they will pay at either supplier in period 3.

to the unobserved heterogeneity hypothesis and then investigate what startup costs imply over and above dispersion in y . If we view the prices p^A and p^B as the counterparts of the prices we do observe in the data, then we uncover two results, as noted above: (i) startup costs amplify the dispersion in observed prices and (ii) imply a unique prediction with regard to the dynamics of the price differential. That the latter prediction (ii) is consistent with the data suggests to us that startup costs are present and hence (i) is likely to be true.

We focus on equilibria in which customers do not choose to switch suppliers, i.e. the blue portion of the demand curve in Figure 4. Firm A's profit maximizing price in period 3 follows from equating marginal revenue for this portion of the demand curve to marginal cost, c^A :

$$p_3^A = \frac{\tau + (1 - \hat{y})\tau + p_3^B + c^A}{2}. \quad (4)$$

Firm A's optimal price is increasing in Firm B's price and its own marginal cost, and Firm A charges a premium $\frac{\tau}{2}$ that reflects its advantage as the relatively more sophisticated producer. Firm A's price is also increasing in its period-2 market share, $1 - \hat{y}$. This reflects the "lock-in" effect: because of the startup cost, these customers are partially captive to Firm A, giving it greater market power.

Firm B's optimization problem yields a similar first-order condition:

$$p_3^B = \frac{\tau \hat{y} + p_3^A + c^B}{2}. \quad (5)$$

Combining these, we can see that Firm A charges a higher price than Firm B in period 3.

$$p_3^A - p_3^B = \frac{2}{3}(1 - \hat{y})\tau + \frac{c^A - c^B}{3} > 0. \quad (6)$$

4.2.3 Period-2 and Period-1 Problem

In period 2, the firms seek to maximize present discounted profits in periods 2 and 3. Firm A solves the following optimization problem,

$$\max_{p_2^A} (p_2^A - c^A)Y_2^A + \beta(p_3^A - c^A)Y_3^A, \quad (7)$$

where $\beta \in (0,1)$ is the discount factor. Using the solution to the period-3 problem, one can solve for p_3^A and Y_3^A as functions of parameters and Firm A's share of the period-2 cohort, $(1 - \hat{y})$.

$$p_3^A = \left(1 + \frac{1}{3}(1 - \hat{y})\right)\tau + \frac{2}{3}c^A + \frac{1}{3}c^B \quad (8)$$

$$Y_3^A = 1 + \frac{1}{3}(1 - \hat{y}) - \frac{1}{3} \frac{c^A - c^B}{\tau} \quad (9)$$

Both the price and sales in period-3 are increasing in the period-2 customer base, highlighting the dynamic nature of the problem. Firm A must set its period-2 price in a forward-looking manner, trading off the future benefit of setting a low price and growing a large customer base against the current benefit of setting a high price and "gouging" its period-1 customers who are partially locked-in because of the switching cost.

Given (8) and (9), it is straightforward to solve (7) once we have specified the demand curve faced by Firm A from period-2 buyers. To this end, note that firm A captures the entire period-1 cohort of measure 1 and $1 - \hat{y}$ from the period-2 cohort, so its period-2 sales are

$$Y_2^A = 1 + (1 - \hat{y}) \quad (10)$$

Then, to derive the demand schedule, we must uncover the relationship between \hat{y} and the period-2 price. This follows from considering the problem of a buyer entering the market in period 2. The buyer must take into account future prices when it decides on a supplier this period. Again assuming that buyers do not switch in the following period, an entrant compares the present discounted cost of purchasing from Firm A for two periods ($p_2^A + \beta p_3^A$) versus the cost of purchasing from Firm B for two periods ($p_2^B + \tau y + \beta(p_3^B + \tau y)$). Using (6), it follows that Firm A captures a share of entrants equal to

$$1 - \hat{y} = \frac{3+3\beta}{3+5\beta} - \frac{3}{3+5\beta} \frac{p_2^A - p_2^B}{\tau} - \frac{\beta}{3+5\beta} \frac{c^A - c^B}{\tau}. \quad (11)$$

We can now solve Firm A's optimization problem by plugging (8)-(11) into (7) and taking the first order condition with respect to p_2^A . This yields the optimal p_2^A as a function of p_2^B . Solving the parallel optimization problem for Firm B yields its optimal p_2^B as a function of p_2^A . Combining these first order conditions, we can show that

$$p_2^A - p_2^B = \frac{9+22\beta+13\beta^2}{27+41\beta} 2\tau + \frac{9+13\beta-2\beta^2}{27+41\beta} (c^A - c^B) > 0. \quad (12)$$

So Firm A charges a higher price than Firm B in period 2 as well. Moreover, the price gap is larger in period 2 than in period 3.

$$(p_3^A - p_3^B) - (p_2^A - p_2^B) = -\frac{\alpha_1 2\tau + \alpha_2 (c^A - c^B)}{27+41\beta} < 0. \quad (13)$$

$$\text{where } \alpha_1 \equiv \frac{18+87\beta+134\beta^2+65\beta^3}{3+5\beta} > 0 \quad \text{and} \quad \alpha_2 \equiv \frac{18+42\beta+14\beta^2-10\beta^3}{3+5\beta} > 0 \quad (14)$$

This prediction concerning the closing price gap is absent from a model in which there is no switching cost. Setting $s = 0$, one would, under certain conditions (see Appendix B), obtain dispersion in average prices within this wafer market. But unless τ is time-varying, this dispersion would be constant. Of course, one cannot rule out variation in τ , but as we discuss below, evidence for changes in τ on the order of what is needed to account for the empirical evidence on closing price differentials is scant. One can also show that the average price gap (averaged across periods 2 and 3) is higher in the model with a startup cost than in the corresponding frictionless model, which sets $s = 0$ but preserves heterogeneity in complexity.

This result is intuitive; the friction introduced by the switching raises the average price suppliers are capable of charging their customers.

The period-1 problem is trivial. Since Firm A is the only supplier, it will charge the highest price that the buyers are willing to pay, which we assume is above the prices charged in periods 2 and 3.

In order to maintain the equilibrium in which buyers choose not to switch suppliers and both firms compete for new customers, two conditions on the parameters must be satisfied. First, the startup cost must be sufficiently high that when a firm competes for new entrants, its customer base chooses not to switch to the other firm. Second, the difference in marginal costs must not be too large compared to the monitoring cost. This condition ensures that Firm A is able to compete for new entrants because the savings incurred by avoiding the monitoring cost are able to outweigh the price increase due to higher marginal cost. Appendix B describes these conditions in detail.

In summary, we have shown that sequential entry and fixed costs of switching support an equilibrium in which the leading firm charges a persistently higher price and customers choose not to switch suppliers. The model also suggests that the price gap between the leading and following supplier should fall as time elapses since the lagging supplier entered the market. We verify this pattern for Taiwanese and Chinese foundries in the following section.

5. CLOSING PRICE GAPS WITHN TECHNOLOGY OVER TIME

The preceding model shows how sequential entry into a given production technology and fixed costs of producing a given design at a given supplier can amplify price differences across producers. The model also yields an important prediction with respect to the dynamics of the price gap between leading and lagging suppliers of a given technology: the price gap should fall over time following entry of the lagging supplier.

We focus our analysis on the two largest foundry markets, China and Taiwan. We restrict the sample to technologies in which both Chinese and Taiwanese foundries produced wafers during the sample period and drop observations for other production locations. The left side of Table 4 re-estimates the hedonic specification from Table 3 on this restricted sample. The China discount coefficient is very similar in Tables 3 and 4, 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.²⁸ We then 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.

For each 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 in the relevant quarter. The estimation equation takes the following form.

$$\begin{aligned} \ln p_{ijkt} = & \beta_0 + \beta_C \mathbf{1}(\text{China}_j) + \sum_{s=1}^T \beta_s \mathbf{1}(\delta_{kt} = s) \\ & + \sum_{s=1}^T \beta_{sC} \mathbf{1}(\text{China}_j) \cdot \mathbf{1}(\delta_{kt} = s) + \Theta X_{ik} + \gamma_t + \varepsilon_i \end{aligned} \quad (15)$$

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 initial period following Chinese entry, and each β_{sC} reports the change in the price gap from the initial period to the s th period. As in Table 3, X_i is a vector of technological controls, and γ_t are quarter indicators. 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 4. The coefficient on the China dummy (β_C) is negative and statistically significant, indicating a 49% average discount in the year of Chinese entry into the market. 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}). 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

²⁸ 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.

Chinese entry are presented in Figure 5, along with a 95% confidence interval. 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 strong evidence in support of our argument that China's delayed entry into the market for each process technology and substantial fixed costs at least partially drive the quality-adjusted price differences that we documented in Section 3.

These findings are largely inconsistent with arguments that the observed price differences between China and Taiwan are actually driven by fixed differences in yields or by differences in the services and design tools provided by foundries in the two countries, as both of those hypotheses would predict stable price gaps following Chinese entry. 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.

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 49% upon entry, if yields drove price differences, the yield differences would be sufficiently 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.

6. IMPLICATIONS

In this section we briefly discuss the 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. The tools developed in that literature were designed to study quality variation in *final* goods, and are likely not as well suited to the study of intermediate input markets like the one we analyze here. We argue that the framework of section 5 may be more helpful to researchers as the study of offshoring focuses efforts increasingly on intermediate supplier markets. 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 can result in upward-biased measures of input price growth and productivity growth.

6.1 Cross-Country Product Quality Measurement

A growing literature studies the relationships between product quality, international trade, and economic development. Central to these studies are measures of quality variation across suppliers of otherwise similar 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.²⁹ 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. However, the unit value approach to measuring quality requires quite strong assumptions. For instance, it does not allow for the possibility of quality-adjusted price variation resulting from differences in manufacturing costs, consumer preferences, and market imperfections.

Reacting to this potential shortcoming of the price-as-quality assumption, more recent research accounts for quality-adjusted price variation across goods even within narrow product classifications by assuming that goods are imperfect substitutes in consumption and that consumers love variety. 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 basis for identifying quality differences across countries; conditional on price, variation in market share reflects unobserved

²⁹ See Hallak and Schott (2011) and Khandelwal (2010) for extensive lists of references.

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).

A complementary approach to studying price dispersion in final goods posits that customers are imperfectly informed as to the various alternatives and must expend real resources searching for the best good. This approach has a long tradition in economics (Stigler 1961, Burdett and Judd 1983). Sorensen's (2001) empirical study of price dispersion across equivalent prescription drugs is very much informed by this way of looking at law-of-one-price deviations. Though it has received somewhat less attention in the recent trade literature, this approach could presumably be applied to study purchasing patterns across imports and domestically produced items.³⁰

While these approaches seem very appropriate when analyzing final consumer goods, they are arguably less well suited to markets for many intermediate inputs. The notion of love of variety is difficult to apply when a customer designs the input or assembly process and purchases manufacturing or assembly services from a third party that implements the customer's design. In addition, there are only a few large suppliers in the input market we study, so a theory of search under imperfect information does not seem relevant. We suspect that this may be true of many other input markets as well. 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. It is somewhat hard to believe that imperfect information regarding producers is a first-order problem in these markets.

The model developed in Section 4 proposes an alternative mechanism generating quality-adjusted price dispersion. Price dispersion emerges because of the market friction generated by the fixed startup cost and sequential entry of suppliers. In that context, differences in market share across suppliers do not necessarily reflect differences in product quality. Rather, the leading firm (Firm A) 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 difference in Taiwanese and Chinese market shares would overstate the quality difference between the two. Thus, the approach used in the previous literature should be employed with

³⁰ Indeed, undergraduate texts on the *J*-curve typically invoke imperfect information as a reason for the delayed effect of changes in the real exchange rate on the trade balance (see Blanchard, 2000).

care when considering the prices of intermediate inputs, particularly in markets with large fixed costs of producing a given product at a given supplier.³¹

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 quality lower quality at the cheaper supplier. This decision requires a detailed understanding of the particular product market and extremely detailed data on product characteristics, making an informed decision infeasible in most cases. Instead, as we discuss in more detail below, the U.S. Bureau of Labor Statistics typically imposes 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 index for the good purchased from the lower-cost supplier, so the price drop incurred by shifting to a cheaper supplier is effectively ignored and omitted from the index. This approach is completely correct when the law of one price holds across suppliers, but when there are real 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 leads to an upward-biased input price index, and thus to an upward biased measure of productivity growth – one observes less expenditure on inputs and infers that this represents *fewer* units of input to produce the same output rather than the same quantity of *cheaper* inputs.

Houseman et al. (2011) discuss this implication of offshoring for productivity mismeasurement. They use the micro data underlying the U.S. import price index to estimate the impact of shifting sourcing patterns on the prices of inputs into the U.S. manufacturing sector. They find substantial upward bias in manufacturing productivity growth, 0.1 to 0.2 percentage

³¹ One caveat should be mentioned here. 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.

points per year from 1997 to 2007. To control for quality differences, they analyze the price changes that occur when a U.S. importer reports a different source country for the same product, defined as a 10-digit Harmonized system (HS) code.

Although the microdata afford considerable product-level information, the proprietary data we use in this paper provide much more detailed product information, which is necessary when analyzing highly differentiated goods. All the processed semiconductor wafers in our sample fall under the same 10-digit HS code, so Houseman et al. implicitly treat all wafers as the same product in their analysis. As shown in Table 3, when controlling for technological characteristics, Chinese foundries charge approximately 19% less than Taiwanese foundries for identical wafers. Without technological controls, the Chinese wafer price is 30% lower than Taiwan’s on average, reflecting Chinese foundries’ focus on trailing-edge products. Thus, even when using microdata at the 10-digit HS level, substantial quality differences remain across countries that may confound the analysis conducted by Houseman et al. This issue is likely to be less severe in sectors with more detailed HS codes and slower technological change. That notwithstanding, our findings demonstrate the potential value of using highly detailed transaction-level data.

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 that differ in whether they exclude or include the price declines incurred in shifting toward lower-price suppliers.³² Both are matched-model indexes, which start with quarterly price relatives for each “model” and aggregate across models using a Fisher index.³³

The two indexes differ in their definition of what constitutes a model. The first, which we will call the technology-country index, considers both technology (wafer size and line width) and

³² This approach follows Reinsdorf’s (1993) analysis of outlet substitution bias in the U.S. CPI.

³³ First we calculate Laspeyres and Paasche indexes, respectively, as

$$P_L^t = \sum_m s_m^{t-1} \cdot \frac{p_m^t}{p_m^{t-1}}$$

$$P_P^t = \left[\sum_m s_m^t \cdot \left(\frac{p_m^t}{p_m^{t-1}} \right)^{-1} \right]^{-1}.$$

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:

$$P_F^t = \sqrt{P_L^t P_P^t}.$$

country of production as price-forming characteristics that define a model. This index reflects the statistical agency approach just described – technologically identical goods produced in separate countries are included in different indexes, so price level differences across countries are omitted from the calculation. The second index, which we call the technology-only index, defines a model based on technology alone. As lower-cost suppliers of a given technology gain market share, the average price of the technology falls, and this is captured in the price relatives for that model and thus included in the technology-only index.

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 2008.³⁴ 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 corresponds to the Taiwan price through second quarter 2008, since one cannot calculate a price relative for the technology-country cell for China until two quarters of price data are available. 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 5 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% 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.

³⁴ 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 in general.

Ideally, we would compare our price indexes to the analog in official statistics, but the BLS does not publish a similarly disaggregate index.³⁵ 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 fables firm. This approach supports the technology-only index calculation just described. In contrast, Nakamura and Steinsson’s (2011) 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 exhibit no examples of semiconductor products switching suppliers, presumably because the new supplier relationship was considered a new product.³⁶ 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 sequential entry and the friction introduced by the fixed costs amplify price differences across suppliers beyond what

³⁵ 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 includes these and 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.

³⁶ 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.

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 model without frictions in switching 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 provide an alternative mechanism for generating quality-adjusted price variation that differs from those in the previous literature focusing on love of variety and imperfect substitutability across suppliers or informational search frictions. While these approaches are quite appropriate when studying final goods markets, we argue that they do not apply as readily to intermediate input markets. We also contribute to the 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 re-sourcing a product may be particularly large in this sector, the costs involved in initiating a new 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 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.

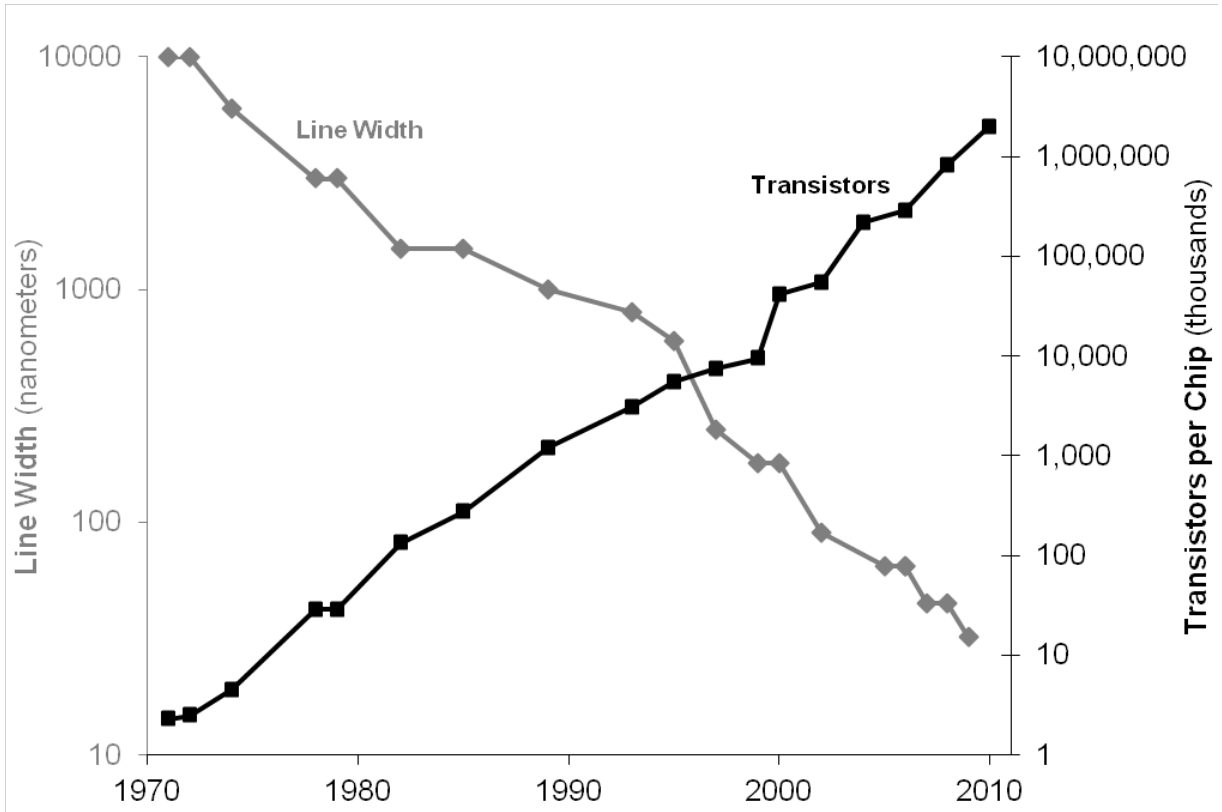
REFERENCES

- Aizcorbe, Ana (2002), "Price Measures for Semiconductor Devices," Finance and Economics Discussion Series 2002-13. Washington: Board of Governors of the Federal Reserve System (U.S.), March.
- Aizcorbe, Ana, Kenneth Flamm, and Anjum Khurshid (2002). "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*: 2002-37.
- Barak, Sylvie (2012), "Intel, Micron on sub 20-nm and insatiable thirst for memory." *Electronic Engineering Times*. April, 2012.
- Berndt, Ernst R., and Neal J. Rappaport (2001). "Price and Quality of Desktop and Mobile Personal Computers: A Quarter-Century Historical Overview," *American Economic Review*, vol. 91, no. 2, pp. 268-273.
- Brown, Clair and Linden, Greg (2006). "Offshoring in the Semiconductor Industry: Historical Perspectives," in Lael Brainard, Susan M. Collins, eds., *Brookings Trade Forum: 2005*. Washington, DC: Brookings Institution Press.
- Burdett, Kenneth, and Kenneth L Judd, (1983), "Equilibrium Price Dispersion." *Econometrica*. vol. 51 pp. 955–69.
- Byrne, David, and Carol Corrado (2012). "Prices for Communications Equipment: Rewriting a 46-Year Record," mimeo.
- Doms, Mark (2005). "Communication Equipment: What has Happened to Prices?" in Corrado, Carol, John Haltiwanger and Daniel Sichel eds., *Measuring Capital in the New Economy*. NBER Studies in Income and Wealth, vol. 65; Chicago and London: University of Chicago Press, pp. 323-362.
- Doms, Mark, Ana Aizcorbe, and Carol Corrado (2003), "When Do Matched-Model and Hedonic Techniques Yield Similar Measures?" Working Papers in Applied Economic Theory 2003-14. San Francisco: Federal Reserve Bank of San Francisco, June.
- Dulberger, Ellen R. (1993). "Sources of Price Decline in Computer Processors: Selected Electronic Components," in Foss, Murray F., Marilyn E. Manser and Allan H. Young eds., *Price Measurements and their Uses*. National Bureau of Economic Research Studies in Income and Wealth, vol. 57; Chicago and London: University of Chicago Press, pp. 103-124.
- Electronic Engineering Times (1998). "U.S., China at Odds on Fab-Gear Export," *Electronic Engineering Times*. April, 1998.
- Flamm, Kenneth (1993). "Measurement of DRAM Prices: Technology and Market Structure," in Murray F. Foss, Marilyn E. Manser and Allan H. Young, eds., *Price Measurements and their Uses*. National Bureau of Economic Research Studies in Income and Wealth, vol. 57. Chicago and London: University of Chicago Press, pp. 157-197.
- Flamm, Kenneth (2006). "The Microeconomics of Microprocessor Innovation" mimeo.
- Global Foundries Fact sheet: Fab 2 module 1. Global Foundries Available from http://globalfoundries.com/newsroom/2009/Fab_2_fact_sheet.pdf [cited March 2009].

- Grimm, Bruce T. (1998). "Price Indexes for Selected Semiconductors, 1974-96," *Survey of Current Business*, vol. 78 (2), pp. 8-24.
- Hallak, Juan Carlos (2006). "Product Quality and the Direction of Trade," *Journal of International Economics*. 68(1), pp.238-265.
- Hallak, Juan Carlos and Peter Schott (2011). "Estimating Cross-Country Differences in Product Quality," *Quarterly Journal of Economics*. 126(1), pp.417-474.
- Holdway, Michael (2001). "An Alternative Methodology: Valuing Quality Change for Microprocessors in the PPI," Washington: Bureau of Labor Statistics. Available via the Internet: www.bls.gov/ppi/ppicomqa.htm.
- Houseman, Susan, Christopher Kurz, Paul Lengermann, and Benjamin Mandell (2011) "Offshoring Bias in U.S. Manufacturing," *Journal of Economic Perspectives*. 25(2), pp.111-132.
- Hurtarte, Jorge S., Evert A. Wolsheimer, and Lisa M. Tafoya (2007). *Understanding Fabless IC Technology*. Oxford: Elsevier.
- IC Knowledge (2001). "Can the semiconductor industry afford the cost of new fabs?" IC Knowledge, www.icknowledge.com/economics/fab_costs.html (accessed March 2010).
- Khandelwal, Amit (2010). "The Long and Short (of) Quality Ladders," *Review of Economic Studies*. 77(4), pp.1450-1476.
- Klemperer, Paul (1995). "Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade," *Review of Economic Studies*. 62, pp.515-539.
- Kugler, Maurice, and Eric Verhoogen (2011). "Prices, Plant Size, and Product Quality," *Review of Economic Studies*. 79(1), pp.307-339
- Kumar, Rakesh. (2007). "The Business of Scaling," *IEEE Solid-State Circuit Society News*, vol. 12 (1), pp. 22-27.
- (2008). *Fabless Semiconductor Implementation*. New York: McGraw-Hill Professional.
- Leachman, Robert C. (2002). "Competitive Semiconductor Manufacturing: Final Report on Findings from Benchmarking Eight-inch, sub-350nm Wafer Fabrication Lines." mimeo. University of California at Berkeley.
- McGregor, Jim (2007). "The common platform technology: A new model for semiconductor manufacturing". Scottsdale, Ariz.: In-Stat, .
- Moore, Gordon E. (1965). "Cramming More Components onto Integrated Circuits," *Electronics*.
- Nakamura, Emi, and Jon Steinsson (2011), "Lost in Transit: Product Replacement Bias and Pricing to Market," Columbia University, June.
- Oliner, Stephen D., and Daniel E. Sichel (2000). "The Resurgence of Growth in the Late 1990s: Is Information Technology the Story?" *Journal of Economic Perspectives*, vol. 14 (4), pp. 3-22.
- Oliner, Stephen D., Daniel E. Sichel, and Kevin J. Stiroh (2007). "Explaining a Productive Decade," *Brookings Papers on Economic Activity*, (1), pp. 81-137.
- Reinsdorf, Marshall (1993). "The Effect of Outlet Price Differentials on the U.S. Consumer Price Index," in Murray F. Foss, Marilyn E. Manser and Allan H. Young, eds., *Price Measurements and their*

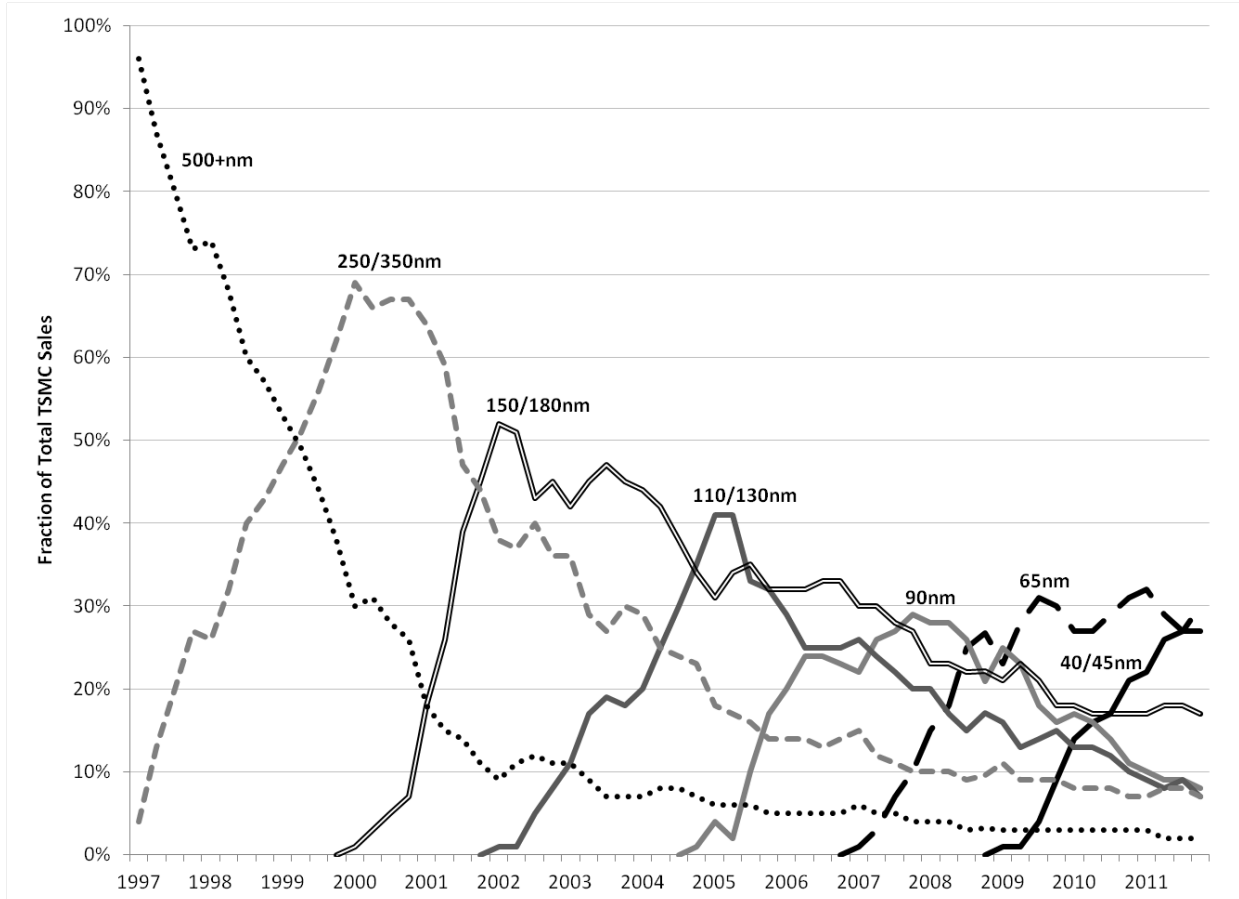
- Uses. National Bureau of Economic Research Studies in Income and Wealth, vol. 57. Chicago and London: University of Chicago Press, pp. 227-254.*
- Schott, Peter K. (2004) "Across-Product Versus Within-Product Specialization in International Trade," *The Quarterly Journal of Economics*. 119(2), pp.647-678.
- Scott, A. J., and D. P. Angel (1988). "The Global Assembly-Operations of U.S. Semiconductor Firms: A Geographical Analysis," *Environment and Planning A*, vol. 20 (8), pp. 1047-67.
- Sorensen, Alan T., (2000). "Equilibrium Price Dispersion in Retail Markets for Prescription Drugs," *Journal of Political Economy*. 108(4).
- Stigler, George J. (1961). "The Economics of Information." *Journal of Political Economy*. vol. 69 pp. 213-25.
- Turley, James L. (2003). *The Essential Guide to Semiconductors*. Upper Saddle River, New Jersey: Prentice Hall PTR.

Figure 1: Moore's Law – Intel Processors



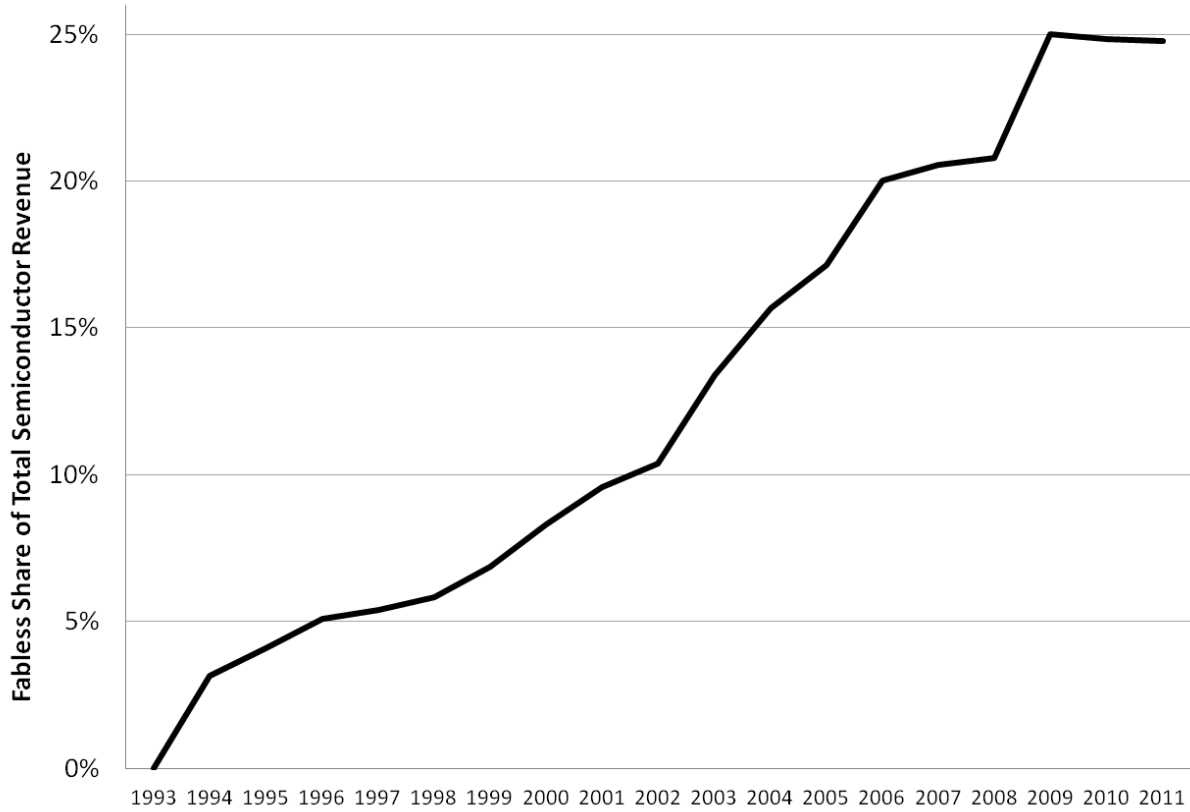
Sources: <http://www.intel.com/technology/timeline.pdf>
<http://www.intel.com/pressroom/kits/quickreffam.htm>

Figure 2: Technology Cycle – TSMC Sales by line width



Sources: TSMC quarterly reports

Figure 3: Growth of the Fabless Business Model



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

Figure 4: Demand Curve Facing Firm A in Period 3

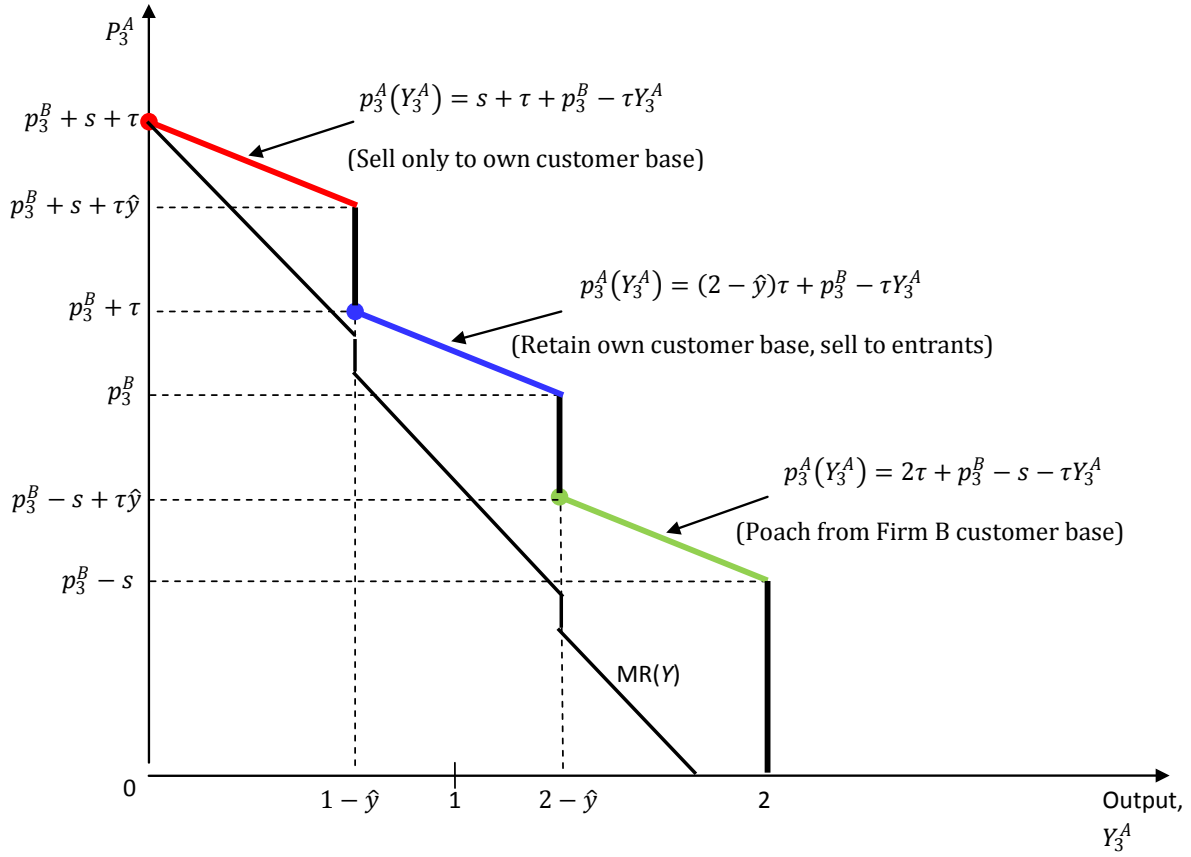


Figure 5: Closing Price Gaps Within Technology

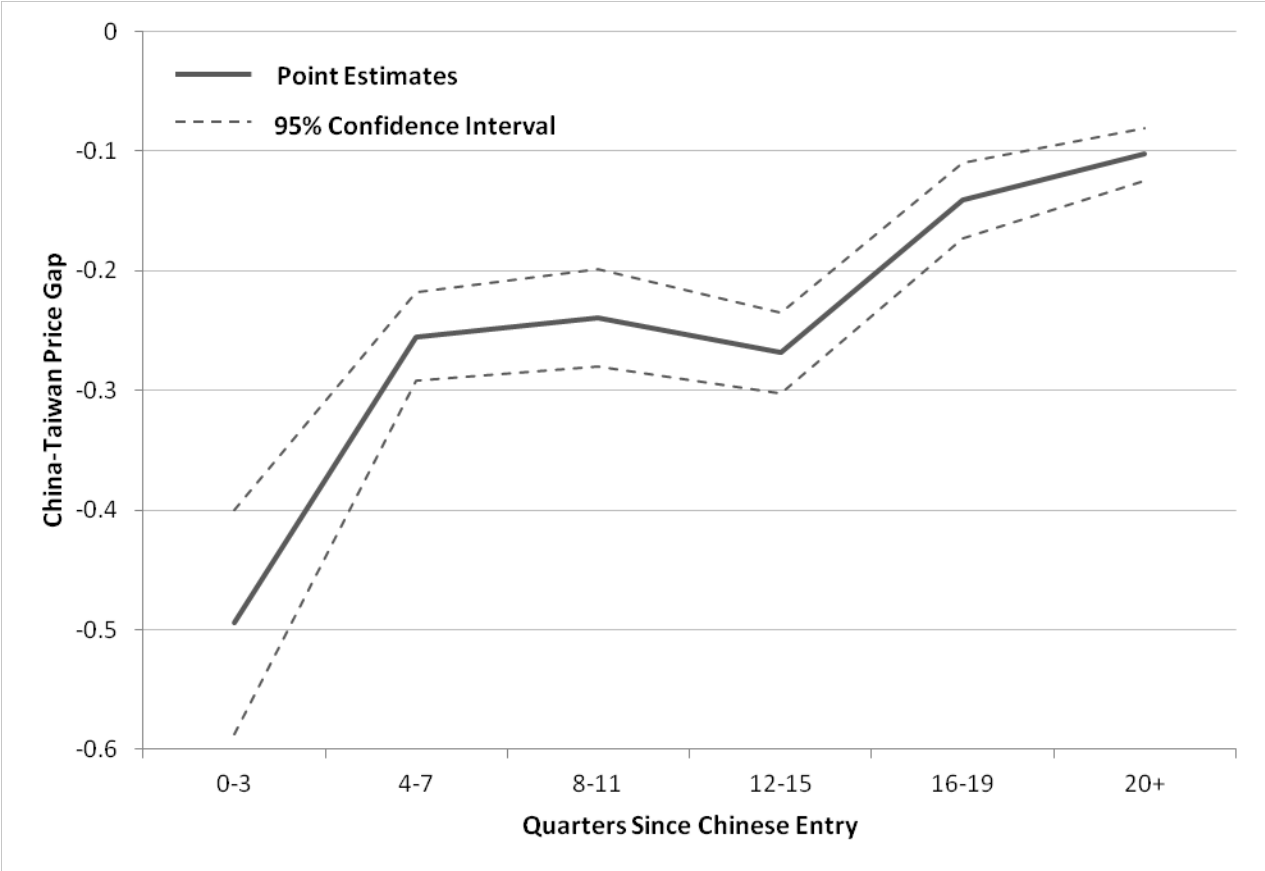


Figure 6: Index Calculation Example

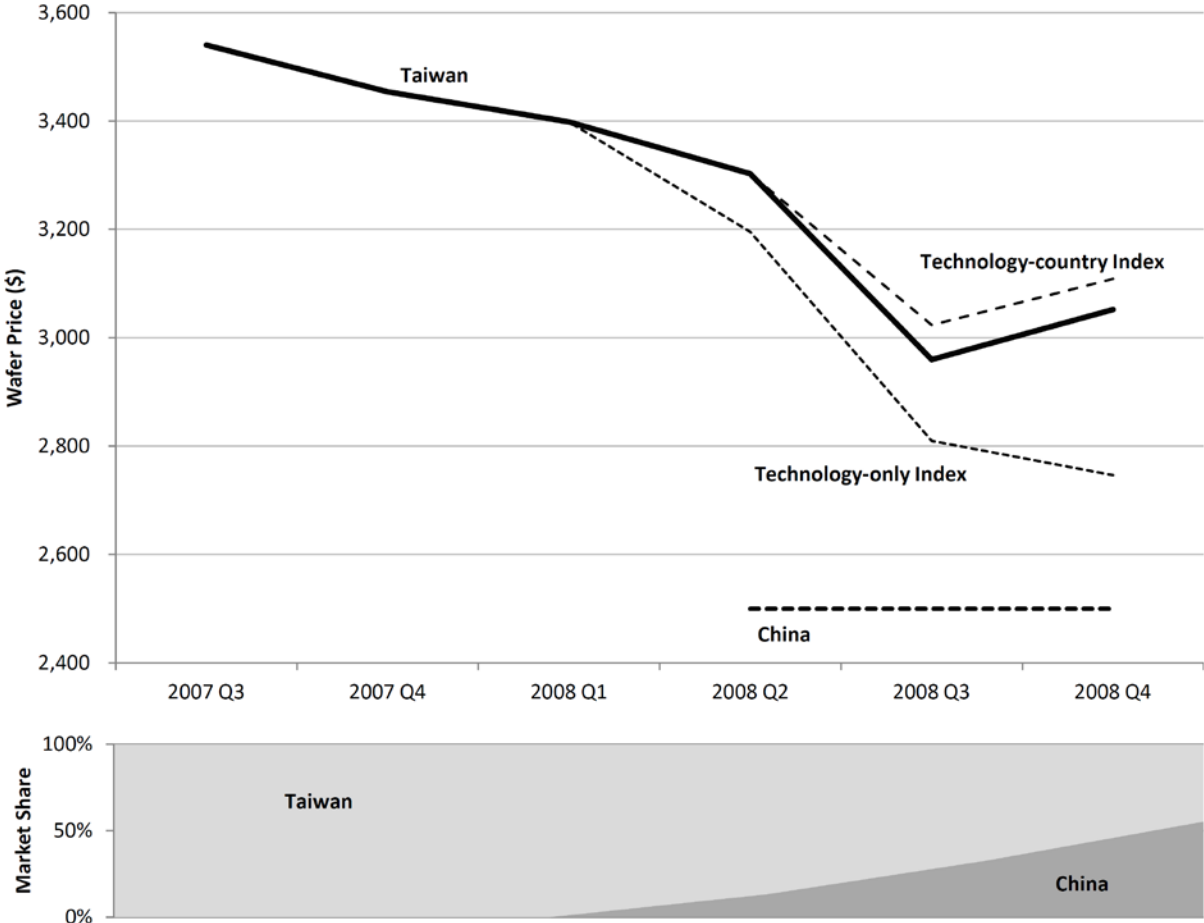


Figure 7: Matched-Model Price Index Results

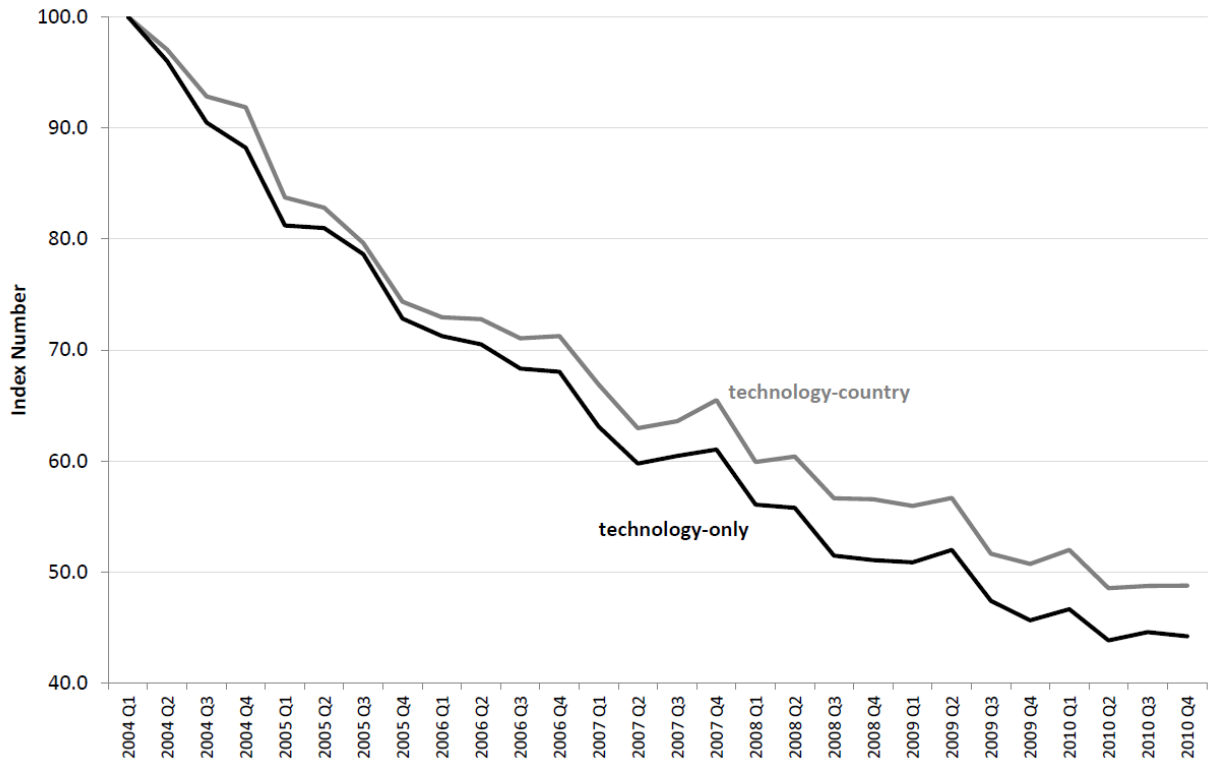


Table 1: Foundry Capacity by Country

	Global Capacity		Taiwan	China	Singapore	Europe	USA	Japan	S. Korea	Malaysia
	Foundry	Total Industry								
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

	Mean	Std. Dev	Yearly Means						
			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	(0.008)	-23.870
Malaysia	-0.278	(0.037)	-7.540
Singapore	-0.061	(0.008)	-7.320
United States	0.068	(0.015)	4.420
Wafer Size			
150 mm	-0.467	(0.013)	-37.380
300 mm	0.671	(0.013)	50.240
Line Width			
≥ 500 nm	-0.245	(0.016)	-15.690
350 nm	-0.167	(0.012)	-14.130
250 nm	-0.061	(0.010)	-6.090
150 nm	0.169	(0.012)	13.860
130 nm	0.356	(0.010)	36.470
90 nm	0.479	(0.019)	25.520
65 nm	0.676	(0.022)	31.340
45 nm	0.962	(0.044)	21.790
Number of Metal Layers	0.076	(0.003)	25.040
Number of Polysilicon Layers	0.027	(0.007)	3.880
Number of Mask Layers	0.005	(0.001)	6.720
Epitaxial Layer Indicator	0.064	(0.008)	7.540
log Number of Wafers Contracted	-0.056	(0.002)	-32.850
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

Table 4: Closing Price Gaps Within Technology

dependent variable: log of price per wafer

Variable	Baseline		Closing price gap	
	Coefficient	Std. Err.	Coefficient	Std. Err.
China	-0.198	(0.008)	-0.494	(0.048)
Time since Chinese entry				
4-7 quarters			-0.128	(0.013)
8-11 quarters			-0.262	(0.015)
12-15 quarters			-0.393	(0.016)
16-19 quarters			-0.450	(0.018)
≥ 20 quarters			-0.478	(0.022)
Interaction terms				
China * (4-7 quarters)			0.239	(0.051)
China * (8-11 quarters)			0.255	(0.052)
China * (12-15 quarters)			0.225	(0.050)
China * (16-19 quarters)			0.353	(0.050)
China * (≥ 20 quarters)			0.391	(0.049)
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

Table 5 – 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%

APPENDIX 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

To build the weights used in constructing our price indexes, and to examine trends in the geographic distribution of the industry, we employ data published by IHS 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 statements; and extensive discussions with iSuppli.¹ 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 for each production technology in Section 5.

¹ 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	
Phenittec	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 200nm and 300nm. 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.

APPENDIX B: MODEL DETAILED SOLUTION

This appendix contains the detailed solution and discussion of the model described in Section 4. Sections B.1 – B.4 derive and discuss the main results, while Sections B.5 and B.6 cover auxiliary results. The environment and notation are described in Section 4.2.1.

B.1 Static Model

It is helpful to first consider the frictionless model, in which $s = 0$. This is a static problem. In each period, a buyer of complexity y selects Firm A if $p^A < p^B + \tau y$. Since $y \sim U[0,1]$ and since the mass of buyers is normalized to 1, Firm A's total sales are

$$Y^A = 1 - \frac{p^A - p^B}{\tau}.$$

The firm sets p^A to maximize $(p^A - c^A)Y^A$. The first-order condition is

$$p^A = \frac{\tau + p^B + c^A}{2}.$$

Firm B solves the symmetric problem. It sets its price according to

$$p^B = \frac{p^A + c^B}{2}.$$

Thus, since $c^A > c^B$, Firm A charges the higher price,

$$p^A - p^B = \frac{\tau + c^A - c^B}{3}. \quad (\text{B1})$$

To guarantee Firm A positive sales, it must at least be attractive to the highest-quality buyer, that is, $p^A - p^B < \tau$. Substituting in from (B1), it follows that each firm earns positive sales only if

$$\frac{c^A - c^B}{2} < \tau, \quad (\text{B2})$$

This is an intuitive result. The left-hand side refers to the effect of Firm A's higher marginal cost on its price. The right-hand side is what the highest-quality buyer ($y = 1$) saves by purchasing from Firm A rather than B. To ensure that Firm A earns any sales, the savings to at least the highest-quality buyer must outweigh the cost differential. We will see that there is a result akin to (B2) in the model with a nonzero startup cost.

B.2 Period 3

Start with Firm A's problem in period 3. We assume the startup cost is "large" in the sense that it exceeds the premium for even the highest-quality product,¹

$$s > \tau. \quad (\text{B3})$$

We solve the model by backward induction, beginning in period 3. Define \hat{y} as Firm B's *customer base* as of the start of period 3, that is, any entrant in period 2 with product $y \in [0, \hat{y}]$ chose Firm B. Hence, any entrant with a product $y \in [\hat{y}, 1]$ chose Firm A in period. We show below that such a partition exists. Observe that \hat{y} may be 0 or 1.

Firm A attracts a period-3 entrant if $p_3^A < p_3^B + \tau y$. Since $y \sim U[0,1]$, it follows that the share of period-3 entrants who purchase from Firm A is²

$$\sigma_3^{A|0} \equiv \Pr[p_3^A < p_3^B + \tau y] = 1 - \frac{p_3^A - p_3^B}{\tau}.$$

Next, Firm A retains a customer, $y \in [\hat{y}, 1]$, in its customer base if $p_3^A < p_3^B + \tau y + s$. Note the role of the startup cost: it drives a wedge between the price that an entrant would be willing to pay and the price that a buyer in Firm A's customer base is willing to pay to remain with the firm. Again using $y \sim U[0,1]$, the share of its customer base that it retains is

$$\sigma_3^{A|A} \equiv \Pr[p_3^A < p_3^B + \tau y + s \mid y \in [\hat{y}, 1]] = 1 - \frac{\frac{p_3^A - p_3^B - s}{\tau} - \hat{y}}{1 - \hat{y}}.$$

Lastly, Firm A attracts a customer, $y \in [0, \hat{y})$, in Firm B's customer base if it charges a price sufficiently low to offset the startup cost, $p_3^A + s < p_3^B + \tau y$. The share of Firm B's customers that Firm A is able to "poach" is then given by

$$\sigma_3^{A|B} \equiv \Pr[p_3^A + s < p_3^B + \tau y \mid y \in [0, \hat{y}]] = 1 - \frac{\frac{p_3^A + s - p_3^B}{\tau}}{\hat{y}}.$$

Since we have normalized the mass of new buyers to one, total sales by Firm A, Y_3^A , are

$$Y_3^A = \hat{y}\sigma_3^{A|B} + \sigma_3^{A|0} + (1 - \hat{y})\sigma_3^{A|A}.$$

Figure 4 plots Firm A's demand schedule in bold and marginal revenue below.

To understand the picture, start with the red portion of the demand curve. This corresponds to the case where Firm A charges a relatively high price, so high in fact that it does not sell to any

¹ Note that a more stringent condition on the startup cost will be imposed below to ensure that the equilibrium reflects critical features of observed firm behavior.

² The shares are bounded below by zero and above by one., i.e., $\sigma_3^{A|0} = \min\left[1, \max\left[0, 1 - \frac{p_3^A - p_3^B}{\tau}\right]\right]$. To reduce clutter, we suppress the min and max operators. Section B.5 shows Firm B's period-3 problem with the min and max operators included.

period-3 entrants, nor does it poach from Firm B's customer base. As a result, $\sigma_3^{A|B} = \sigma_3^{A|0} = 0$. So its sales are given by

$$Y_3^A = (1 - \hat{y})\sigma_3^{A|A} = 1 - \frac{p_3^A - p_3^B - s}{\tau}.$$

Invert this to obtain the demand curve shown above. Similar calculations yield the blue and green portions of the demand curve. Along the blue portion, Firm A sells to all its period-2 customers and competes for at least a positive share of the period-3 entrants. Along the green portion, the firm captures all of the period-3 entrants and poaches a share of Firm B's period-2 customers.

The kinks in the demand curve arise because of the discrete startup cost. For instance, the price $p_3^B + s + \tau\hat{y}$ is just low enough for Firm A to defend its period-2 customer base, so its output is $1 - \hat{y}$. But to compete for period-3 entrants, the firm must drop its price discretely. This is because the period-3 entrants are not "locked in" by the discrete startup cost, which they must pay at either supplier. A similar argument may be made for the kink at $Y_3^A = 2 - \hat{y}$: to poach Firm B customers, Firm A must again drop its price discretely because Firm B buyers must repay the discrete startup cost if they switch.³

We focus on equilibria in which Firm A sells along the blue portion of the demand curve, as this is the empirically relevant case. In what follows, we will identify restrictions on the parameters that ensure the existence of this equilibrium. The marginal revenue schedule corresponding to this portion of the demand curve is given by

$$\frac{\partial}{\partial Y_3^A} p_3^A Y_3^A = (2 - \hat{y})\tau + p_3^B - 2\tau Y_3^A.$$

Setting this equal to marginal cost gives

$$p_3^A = \frac{\tau + (1 - \hat{y})\tau + p_3^B + c^A}{2}. \quad (\text{B4})$$

Firm A's price is increasing in Firm B's price and its own marginal cost. Firm A's price is also increasing in its period-2 market share, $1 - \hat{y}$. This reflects the "lock-in" effect: because of the startup cost, these customers are partially captive to Firm A, giving it greater market power. In addition, Firm A charges a premium $\frac{\tau}{2}$ that reflects its advantage as the relatively more sophisticated producer.

One can perform a similar analysis of Firm B's optimization problem (see Section B.5). If Firm B also retains its period-2 customers and competes for a positive share of the period-3 entrants, then

$$p_3^B = \frac{\tau\hat{y} + p_3^A + c^B}{2}. \quad (\text{B5})$$

³ Observe that any price between $p_3^B + s + \tau\hat{y}$ and $p_3^B + \tau$ cannot be optimal: the firm can raise the price, lose (and gain) no customers, and so raise its revenue. The same idea applies at the other kink.

Differencing (B4) and (B5), we see that Firm A's price is higher:

$$\Delta p_3 \equiv p_3^A - p_3^B = \frac{2}{3}(1 - \hat{y})\tau + \frac{c^A - c^B}{3} > 0. \quad (\text{B6})$$

Equations (B4) and (B5) are conditioned on the assumption that each firm retains its respective customer base and captures a positive share of period-3 entrants. We now must identify a restriction on the parameters that fulfills this assumption. To sell to any entrants, Firm A must appeal to at least the highest-quality buyer: $p_3^A - p_3^B < \tau$. In addition, to retain all of its period-1 customers, its price must be sufficiently low to discourage even the lowest-quality buyer from switching, $p_3^A - p_3^B < s$. Since $\tau < s$ by assumption, the former condition is the only binding one: if Firm A appeals to any entrants, its price will be low enough to retain its entire customer base and there will be no switching. Using (B6), we see that

$$p_3^A - p_3^B < \tau \Leftrightarrow \frac{c^A - c^B}{2} < \left(\hat{y} + \frac{1}{2}\right)\tau. \quad (\text{B7})$$

This is analogous to (B2) in the frictionless model. But here, the upper bound on the difference in marginal costs is determined by both the quality premium, τ , and Firm B's customer base, \hat{y} . The role of the former is familiar from (B2). What is the role of \hat{y} ? When this is low, Firm A's customer base is large. Since it wants to extract as much as possible from these buyers in the terminal period, its optimal price is relatively high. Thus, if \hat{y} is small and $c^A - c^B$ is large, the strategy of gouging the customer base will be more profitable than trying to compete for entrants with a much lower-cost supplier. Thus, \hat{y} must be sufficiently large relative to $c^A - c^B$ to maintain the desired equilibrium.

Equation (B7) is a restriction on the equilibrium outcome of the period-2 problem, since it contains \hat{y} . Below we will verify that the equilibrium choice of \hat{y} does in fact satisfy (B7).

B.3 Period 2

We now turn to the period-2 problem. Firm A seeks to maximize its present discounted profits by solving the following optimization problem,

$$\max_{p_2^A} (p_2^A - c^A)Y_2^A + \beta(p_3^A - c^A)Y_3^A, \quad (\text{B8})$$

where $\beta \in (0,1)$ is the discount factor.

Our solution to the period-3 problem allows us to express p_3^A and Y_3^A in terms of parameters and \hat{y} , the share of the period-2 cohort choosing Firm B.

$$p_3^A = \left(1 + \frac{1}{3}(1 - \hat{y})\right)\tau + \frac{2}{3}c^A + \frac{1}{3}c^B \quad (\text{B9})$$

$$Y_3^A = 1 + \frac{1}{3}(1 - \hat{y}) - \frac{1}{3}\frac{c^A - c^B}{\tau}. \quad (\text{B10})$$

The expressions (B9) and (B10) for the period-3 price and sales follow from solving (B4) and (B5) and using the demand schedule (along the blue portion of Figure 1). They are also intuitive.

The period-3 price is increasing in the customer base, $1 - \hat{y}$, since these buyers are partially locked in, and increasing in both firms' marginal costs. A high c^B implies a higher p_3^B , which enables Firm A to raise its price. Period-3 sales are also increasing in the size of the customer base and in the size of the quality premium, τ . Sales are declining in Firm A's relative marginal cost ($c^A - c^B$).

That period-3 revenue depends on the customer base cultivated in period 2 is the critical feature of the problem. Firm A must set its period-2 price in a forward-looking manner, trading off the future benefit of setting a low price and growing a large customer base against the current benefit of setting a high price and "gouging" its period-1 customers who are partially locked-in because of the switching cost.

Firm A's period-2 sales are given by $Y_2^A = 1 \cdot \sigma_2^{A|A} + \sigma_2^{A|0}$, where $\sigma_2^{A|A}$ is the share of the period-1 customers that it retains in period 2 and $\sigma_2^{A|0}$ is its share of the period-2 entrants. By definition, $\sigma_2^{A|0} = 1 - \hat{y}$. We will again focus on the case in which Firm A retains all of its period-1 customers, implying that $\sigma_2^{A|A} = 1$, so

$$Y_2^A = 1 + (1 - \hat{y}). \quad (\text{B11})$$

To solve the firm's optimization problem, we need to express (B8) in terms of parameters, p_2^A , and p_2^B by solving for \hat{y} . We can do so by developing the demand schedule that Firm A faces in period 2.

B.3.1 Demand from period-2 buyers

Period-1 buyers of quality y remain with Firm A iff $p_2^A \leq p_2^B + \tau y + s$. Hence, the share of period-1 buyers retained by Firm A is

$$\sigma_2^{A|A} \equiv 1 - \frac{p_2^A - p_2^B - s}{\tau}. \quad (\text{B12})$$

Period-2 entrants face a slightly different problem. They must take into account future prices when they decide on a supplier this period. Again, assuming buyers do not switch in the following period, which is ensured by assuming (B7), an entrant compares the present discounted cost of purchasing from Firm A for two periods ($p_2^A + \beta p_3^A$) versus the cost of purchasing from Firm B for two periods ($p_2^B + \tau y + \beta(p_3^B + \tau y)$). It follows that Firm A captures a share of entrants equal to

$$\sigma_2^{A|0} = 1 - \hat{y} = 1 - \frac{p_2^A - p_2^B + \beta(p_3^A - p_3^B)}{(1 + \beta)\tau}.$$

Using (B6), this implies

$$\sigma_2^{A|0} = 1 - \hat{y} = \frac{3+3\beta}{3+5\beta} - \frac{3}{3+5\beta} \frac{p_2^A - p_2^B}{\tau} - \frac{\beta}{3+5\beta} \frac{c^A - c^B}{\tau}, \quad (\text{B13})$$

This is Firm A's share of period-2 entrants expressed in terms of parameters and period-2 prices, which we needed to solve the optimization problem in (B8). Before returning to the optimization problem, for completeness we develop the full demand schedule for Firm A in period 2 and the parameter restrictions necessary to ensure that Firm A maintains its period-1 customers and competes for a share of period-2 entrants.

The demand schedule in Figure B1 is spliced together from (B10) and (B12). If the firm does not compete for period-2 entrants, then $\sigma_2^{A|0} = 0$ and $Y_2^A = \sigma_2^{A|A} = 1 - \frac{p_2^A - p_2^B - s}{\tau}$. Rearranging this, we have the inverse demand curve for this region, shown in red,

$$p_2^A = s + \tau + p_2^B - \tau Y_2^A. \quad (\text{B14})$$

If the firm instead retains all of its period-1 buyers and competes for at least a positive share of period-2 entrants, its period-2 sales equal $Y_2^A = 1 + \sigma_2^{A|0}$. Using (B13), the blue portion of the demand schedule is

$$p_2^A = \frac{6+8\beta}{3}\tau + p_2^B - \frac{\beta}{3}(c^A - c^B) - \frac{3+5\beta}{3}\tau Y_2^A. \quad (\text{B15})$$

To attract a positive measure of period-2 entrants, Firm A must set its price such that

$$p_2^A < \hat{p}_2^A \equiv (1 + \beta)\tau + p_2^B - \frac{\beta}{3}(c^A - c^B). \quad (\text{B16})$$

The threshold, \hat{p}_2^A , is the left endpoint of the blue demand schedule – it is the price that satisfies (B15) such that $\sigma_2^{A|0} = 0$ (and so $Y_2^A = 1$). Equation (B16) is a restriction on prices p_2^A and p_2^B in equilibrium – it is the period-2 analog to equation (B7). We will verify that it is satisfied once we solve for the period-2 prices.

We must also ensure that the startup cost is sufficiently high to ensure that if Firm A successfully competes for a positive share of entrants, its price must be low enough to ensure there is no switching among its period-1 buyers. This implies that \hat{p}_2^A must be no greater than $p_2^B + s$, which is the highest price consistent with Firm A retaining all of its period-one customers. This will hold if the startup cost is sufficiently high:

$$\hat{p}_2^A < p_2^B + s \Leftrightarrow \tau + \beta\tau \left(1 - \frac{1}{3}\frac{c^A - c^B}{\tau}\right) < s. \quad (\text{B17})$$

This gives a lower bound on s that ensures no switching in period 2. It is a stronger condition than equation (B3), $s > \tau$, which was sufficient for ruling out switching in period 3.

Firm B's problem can be handled symmetrically.

B.3.2 Profit maximization

Returning to Firm A's optimization problem in (B8), substituting in (B9)-(B13), and maximizing with respect to p_2^A gives the first-order condition that enforces this equilibrium, taking as given p_2^B .⁴

$$p_2^A = \frac{9+23\beta+14\beta^2}{18+28\beta} 2\tau + \frac{9+13\beta}{18+28\beta} p_2^B + \frac{9+14\beta-\beta^2}{18+28\beta} c^A + \beta \frac{1+\beta}{18+28\beta} c^B. \quad (\text{B18})$$

For this to be an equilibrium, Firm B must play the corresponding strategy, namely, it must set a price such it attracts a positive share of the entrants but does not poach any of Firm A's customer base. As shown in the appendix, this implies an optimal period-2 price,

$$p_2^B = \beta \frac{1+\beta}{18+28\beta} 2\tau + \frac{9+13\beta}{18+28\beta} p_2^A + \beta \frac{1+\beta}{18+28\beta} c^A + \frac{9+14\beta-\beta^2}{18+28\beta} c^B. \quad (\text{B19})$$

Taking the difference between (B18) and (B19) gives

$$\Delta p_2 \equiv p_2^A - p_2^B = \frac{9+22\beta+13\beta^2}{27+41\beta} 2\tau + \frac{9+13\beta-2\beta^2}{27+41\beta} (c^A - c^B). \quad (\text{B20})$$

There are two conditions that must be verified in order to support (B18) and (B19) as the equilibrium solution. First, to avoid corner outcomes where one firm captures the whole entrant cohort in period 2, (B16) requires that $\Delta p_2 < (1 + \beta)\tau - \frac{\beta}{3}(c^A - c^B)$. Substituting in from (B20), this implies

$$c^A - c^B < \frac{9(1+\beta)}{9+7\beta} \tau. \quad (\text{B21})$$

Second, we must verify (B7), which avoids corner outcomes in period 3. Combining (B7) and (B20), the following condition enforces an interior solution:

$$c^A - c^B < \frac{63 + 210\beta + 175\beta^2}{9 + 42\beta + 45\beta^2} \tau.$$

One may verify that that, for any $0 < \beta \leq 1$, this condition is non-binding in light of (B21).

B.4 Falling Price Gap

We now want to compare Δp_2 with Δp_3 . Using (B6), (B13), and (B20), we first compute

$$\Delta p_3 = \frac{1}{3 + 5\beta} \left\{ \frac{9 + 24\beta + 15\beta^2}{27 + 41\beta} 2\tau + \frac{9 + 42\beta + 45\beta^2}{27 + 41\beta} (c^A - c^B) \right\}.$$

Comparing this to (B20), we have

⁴ This is the same strategy we adopted in the analysis of the period-3 problem. But we cannot solve the dynamic problem graphically.

$$\Delta p_3 - \Delta p_2 = -\frac{\alpha_1 2\tau + \alpha_2 (c^A - c^B)}{27 + 41\beta} < 0.$$

where

$$\alpha_1 \equiv \frac{18 + 87\beta + 134\beta^2 + 65\beta^3}{3 + 5\beta} > 0$$

$$\alpha_2 \equiv \frac{18 + 42\beta + 14\beta^2 - 10\beta^3}{3 + 5\beta} > 0.$$

This says the price gap falls from period 2 to period 3.

B.5 Firm B's Period 3 Problem

There is a symmetric problem for Firm B. Its sales are given by

$$Y_3^B = (1 - \hat{y})\sigma_3^{B|A} + \sigma_3^{B|0} + \hat{y}\sigma_3^{B|B},$$

where $1 - \hat{y}$ is Firm A's customer base; $\sigma_3^{B|A}$ is the share of this base that is poached by Firm B,

$$\begin{aligned} \sigma_3^{B|A} &\equiv \min[1, \max[0, \Pr\{p^B + \tau y + s < p^A \mid y \in [\hat{y}, 1]\}]] \\ &= \min \left[1, \max \left[0, \frac{\frac{p^A - p^B - s}{\tau} - \hat{y}}{1 - \hat{y}} \right] \right]; \end{aligned}$$

$\sigma_3^{B|0}$ is the share of the period-3 entrants that is captured by Firm B,

$$\begin{aligned} \sigma_3^{B|0} &\equiv \min[1, \max[0, \Pr\{p^B + \tau y < p^A\}]] \\ &= \min \left[1, \max \left[0, \frac{p^A - p^B}{\tau} \right] \right]; \end{aligned}$$

\hat{y} is Firm B's customer base; and $\sigma_3^{B|B}$ is the share of this that it retains,

$$\begin{aligned} \sigma_3^{B|B} &\equiv \min[1, \max[0, \Pr\{p^B + \tau y < p^A + s \mid y \in [0, \hat{y}]\}]] \\ &= \min \left[1, \max \left[0, \frac{\frac{p^A - p^B + s}{\tau}}{\hat{y}} \right] \right]. \end{aligned}$$

The demand schedule in Figure B2 is again divided into three regions: (i) the firm does not compete for new entrants and retains only a fraction its period-2 customer base ($0 \leq Y_3^B \leq \hat{y}$); (ii) the firm retains its customer base and captures a fraction of the period-3 entrants ($\hat{y} < Y_3^B \leq$

$1 + \hat{y}$); and (iii) the firm captures all of the period-3 entrants and poaches a share of Firm A's customer base ($1 + \hat{y} < Y_3^B \leq 2$). The figure below illustrates the demand schedule in each region:

The marginal revenue schedule corresponding to the middle (blue) portion of the demand curve is

$$p^A + \tau\hat{y} - 2\tau Y^B. \quad (\text{B22})$$

If the firm finds it optimal to sell along the middle portion of the demand curve, then its price is obtained by equating (B22) to marginal cost, giving,

$$p^B = \frac{p^A + c^B + \tau\hat{y}}{2}. \quad (\text{B23})$$

To enforce (B23) as an optimum, it must be that marginal cost does in fact pass through the middle portion of the marginal revenue schedule. Evaluating marginal revenue at \hat{y} and $1 + \hat{y}$ yields $p^A - \tau\hat{y}$ and $p^A - \tau\hat{y} - 2\tau$. Hence, to induce the firm to sell within the region $[\hat{y}, 1 + \hat{y}]$, marginal cost, c^B , must satisfy

$$p^A - \tau\hat{y} - 2\tau < c^B < p^A - \tau\hat{y}.$$

Substituting in for p^A using (B23)-(B24) then yields

$$2\hat{y} > \frac{(c^A - c^B)}{\tau} - 1,$$

which matches the condition (B1) given in the main text.

Lastly, we calculate period-3 sales for Firm B. Firm B sales are given by $Y_3^B = \sigma_3^{B|0} + \hat{y} = \frac{p^A - p^B}{\tau} + \hat{y}$. Using (B5), this becomes

$$Y_3^B = \frac{2}{3} + \frac{1}{3} \left(\hat{y} + \frac{c^A - c^B}{\tau} \right). \quad (\text{B24})$$

One may confirm that $Y_3^B = 2 - Y_3^A$, as it should be (since the total mass of buyers in period 3 is 2).⁵ Thus, using (B24) and (2)-(B4), the demand curve implies that

$$p_3^B = \frac{2}{3}(\tau + c^B) + \frac{1}{3}(\tau\hat{y} + c^A). \quad (\text{B25})$$

B.6 Period-2 Optimization Problem

⁵ Since Firm A retains all its period-2 cohort ($\sigma_3^{A|A} = 1$), its total sales are $Y_3^A = \sigma_3^{A|0} + (1 - \hat{y})$. After solving (2)-(3) for the prices, p_3^A and p_3^B , in terms of the structural parameters, we can then solve for Y_3^A using the demand schedule: $Y_3^A = \frac{4}{3} - \frac{1}{3} \left(\hat{y} + \frac{c^A - c^B}{\tau} \right)$. It follows that $2 - Y_3^A = \frac{2}{3} + \frac{1}{3} \left(\hat{y} + \frac{c^A - c^B}{\tau} \right)$, as noted. Note that this also confirms $\sigma_3^{B|0} = 1 - \sigma_3^{A|0}$, since $\sigma_3^{B|0} + \hat{y} = Y_3^B = 2 - Y_3^A = 1 + \hat{y} - \sigma_3^{A|0}$. Canceling \hat{y} s, shows $\sigma_3^{B|0} = 1 - \sigma_3^{A|0}$.

Setting (conjecturing) $\sigma_2^{A|A} = 1$ and using (B9)-(B10) and (B15), we write the objective function (B8) as

$$(p_2^A - c^A) \left(\frac{6 + 8\beta}{3 + 5\beta} - \frac{3}{3 + 5\beta} \left(\frac{p_2^A - p_2^B}{\tau} + \frac{\beta c^A - c^B}{3\tau} \right) \right) \\ + \beta \tau \left(\frac{4 + 6\beta}{3 + 5\beta} - \frac{1}{3 + 5\beta} \frac{p_2^A - p_2^B}{\tau} - \frac{1 + 2\beta c^A - c^B}{3 + 5\beta} \frac{1}{\tau} \right)^2.$$

Differentiating with respect to p_2^A ,

$$p_2^A = \left(\frac{9 + 23\beta + 14\beta^2}{18 + 28\beta} \right) 2\tau + \frac{9 + 13\beta}{18 + 28\beta} p_2^B + \frac{9 + 14\beta - \beta^2}{18 + 28\beta} c^A + \beta \frac{1 + \beta}{18 + 28\beta} c^B.$$

We now solve Firm B's problem. Firm B maximizes present discounted profits,

$$(p_2^B - c^B) Y_2^B + \beta (p_3^B - c^B) Y_3^B.$$

Using (17) and (B23)-(B24), and noting that $Y_2^B = \sigma_2^{B|0}$, we can write this in terms of p_2^B ,

$$(p_2^B - c^B) \left(\frac{2\beta}{3 + 5\beta} + \frac{3}{3 + 5\beta} \left(\frac{p_2^A - p_2^B}{\tau} + \frac{\beta c^A - c^B}{3\tau} \right) \right) \\ + \frac{\beta}{3 + 5\beta} \tau \left(\frac{2 + 4\beta}{3 + 5\beta} + \frac{1}{3 + 5\beta} \frac{p_2^A - p_2^B}{\tau} + \frac{1 + 2\beta c^A - c^B}{3 + 5\beta} \frac{1}{\tau} \right)^2.$$

Differentiating with respect to the price,

$$p_2^B = \beta \frac{1 + \beta}{18 + 28\beta} 2\tau + \frac{9 + 13\beta}{18 + 28\beta} p_2^A + \beta \frac{1 + \beta}{18 + 28\beta} c^A + \frac{9 + 14\beta - \beta^2}{18 + 28\beta} c^B.$$

Figure B1: Firm A Period-2 Demand Curve

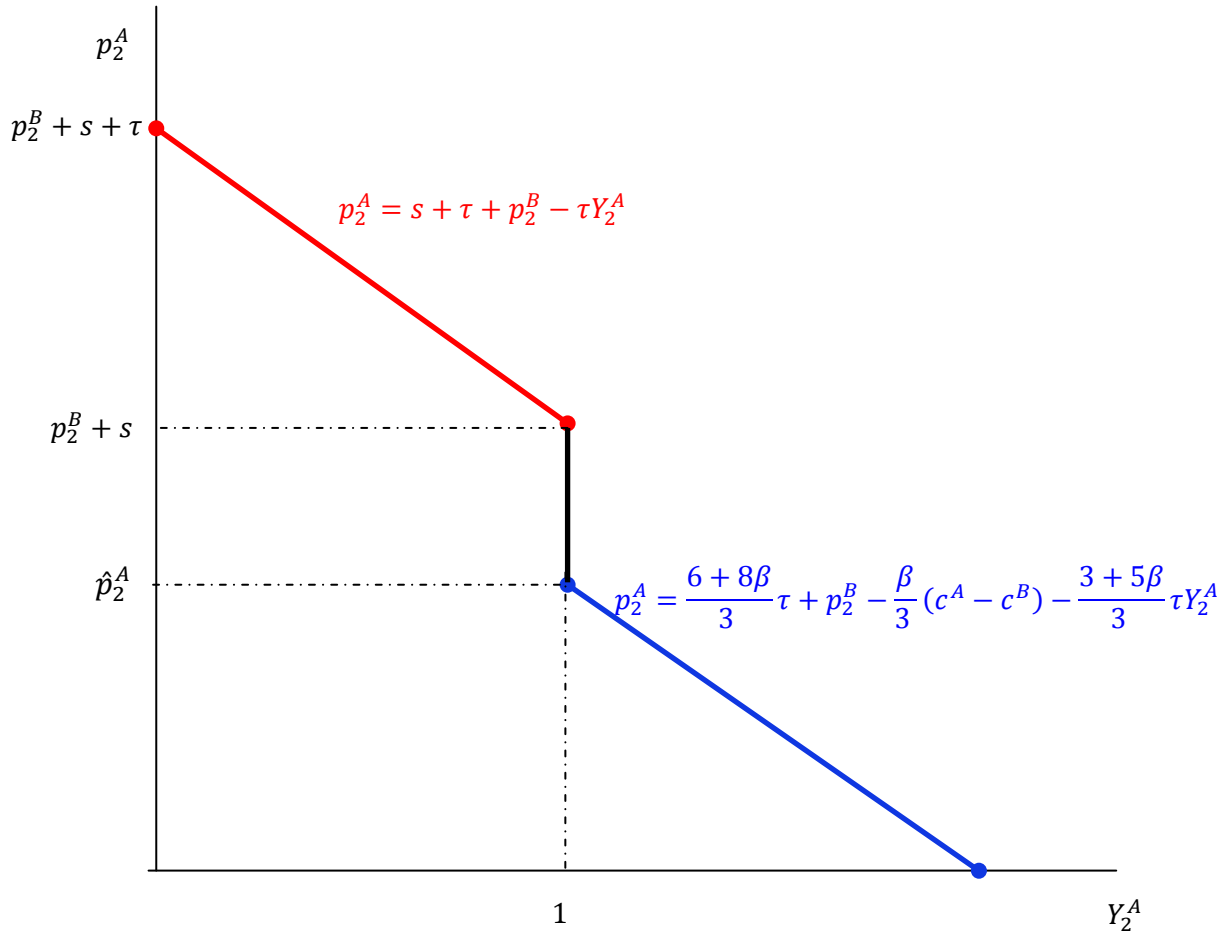


Figure B2: Firm B Period-3 Demand Curve

