Boom and Bust in Telecommunications

Elise A. Couper, John P. Hejkal, and Alexander L. Wolman

The telecommunications sector has experienced a spectacular decline from mid-2000 until the present, after experiencing a spectacular rise from early 1997. Equity valuations and capital spending soared and then plummeted, and a flood of initial public offerings turned into a flood of bankruptcy filings. The boom and bust in telecommunications coincided with the boom and bust in the U.S. equity market as a whole and with the "dot-com bubble" of Internet stocks. The dot-coms received most of the publicity initially, but the telecommunications industry accounts for a much larger share of market capitalization gained and lost than do the dot-coms.[1] This article documents the telecom boom and bust, and contends that it was caused by a combination of major changes in the regulatory landscape and rapid technological progress. Both factors made it difficult for telecommunications firms and outside investors to accurately forecast supply and demand conditions in the industry.[2]

The single most important telecommunications regulatory change in recent years was the Telecommunications Act of 1996. This Act was meant to bring competition to the local exchange carrier level, that is local telephone service. By 1996, long-distance telephone service had a significant amount of competition, whereas local service was largely monopolized by the regional Bell

Hejkal and Wolman are with the Federal Reserve Bank of Richmond. Couper is a Ph.D. student in economics at the University of California, Berkeley. We are grateful to Huberto Ennis, Andrew Foerster, Tom Humphrey, John Weinberg, and Roy Webb for comments on an earlier draft. The views here are the authors' and should not be attributed to the Federal Reserve Bank of Richmond, the Federal Reserve System, or the Board of Governors of the Federal Reserve System.

[1] As will become clear, the two industries are closely related.

[2] We concentrate on the U.S. telecommunications sector. A similar telecom boom and bust occurred in other countries; this does not seem to be at odds with our explanation for the U.S. experience, but further study is warranted.

operating companies, such as Bell Atlantic and Southwestern Bell. On the technological side, passage of the 1996 Act coincided with advances in fiber-optic technology that dramatically increased the capacity for data transmission and with more efficient use of the spectrum available for wireless communication. This was also a time of rapidly increasing Internet use. Growth of the Internet alone meant greater demand for telecommunications services. The combination of improving technology for data transmission and the possibility of a deregulated market for telecommunications services held out the potential that providers would be able to compete for *all* of a household's or firm's telecom needs. The confluence of these factors led to the tremendous investment surge and high stock valuations that were the hallmark of the telecom boom.[3]

Within four years of its passage, however, the Act's initial promise had faded. A series of legal battles had ushered in tremendous uncertainty about the industry's future. By early 2001, it became apparent that massive overinvestment had taken place in the sector, particularly in the area of long-distance fiber-optic cable. Stock prices plunged and investment collapsed. These problems were exacerbated by the U.S. economy's swing into recession early in 2001, and the telecommunications sector remains in a slump to this day.

We do not subscribe to the view held by many, that the boom and bust in the telecommunications industry represented a bubble that burst.[4] According to this view, telecom equity prices were high because people believed they would be high in the future, though there was no expectation of high future dividends. In turn, high equity prices drove the high levels of investment in the industry. Then, when the belief collapsed, equity prices and investment collapsed (the bubble burst). With the benefit of hindsight, it is clear that telecom equity prices and levels of capital spending were "too high" in the late 1990s. However, high equity prices and high investment seem to have been based on beliefs about future fundamentals, not simply on the expectation that prices would rise in the future. We are also skeptical about the view that WorldCom can be blamed for the industry's fluctuations.

Already much has been written about the fluctuations in the telecommunications industry around the turn of the 21st century. We look forward to thorough analyses of this episode in the years to come. Our purpose in this article is to document some basic facts about what happened in the telecommunications industry, and to propose an explanation for those facts. The facts alone make for an impressive tale. In addition, we hope that a tentative

[3] Firms seem to have viewed the prospect of offering a broad range of telecommunications services (being a "single provider") as carrying with it high profit margins. This raises interesting questions: Are consumers willing to pay higher prices to a single provider? Are there production efficiencies in being a single provider?

[4] A Google search on "telecom bubble" yields 1,860 hits. One might think that any two-word phrase would yield hundreds of hits when typed into Google. This is not true: "textile bubble" yielded only five hits.

understanding of what drove the telecommunications boom and bust can help inform policymaking in the immediate future.

1. THE TELECOMMUNICATIONS INDUSTRY IN THE UNITED STATES

For our purposes, telecommunications services will refer to two-way transmission of information (to include voice, text, audio, and video) "between parties that are not in physical contact with each other" (Cave, Majumdar, and Vogelsang, 2002, 3). Consumers purchase these services from telephone companies, which include local, long-distance, wireless, and cable, and from Internet providers. The divisions between these categories are increasingly blurred, with many companies providing more than one of the services. The blurring of divisions between different telecommunications services is, like the boom and bust, related to technological and regulatory changes. As the provision of telecommunications services has become less monopolized in the years since the breakup of AT&T, firms producing intermediate service inputs also have begun to play an important role in the industry.[5]

Telephone services include local and long distance, wireless, and related services such as voice mail, caller ID, and directory assistance. Local telephone service was originally provided by a single firm in each area, a regional Bell, or GTE. These firms are referred to as incumbent local exchange carriers, or ILECs. Since the 1996 Act, long-distance companies and local entrants known as competitive local exchange carriers (CLECs) have begun to compete with the incumbents for the local market.[6] The technology for both the incumbents and the entrants consists of the copper local loop (the portion of the lines connecting directly to the house or business), a fiber network for longer-distance transmission, and switching facilities that route calls along the network. The technology also includes facilities for providing other services, such as voice mail, alongside basic local service. Recently, cable companies have used their existing networks to provide phone service.

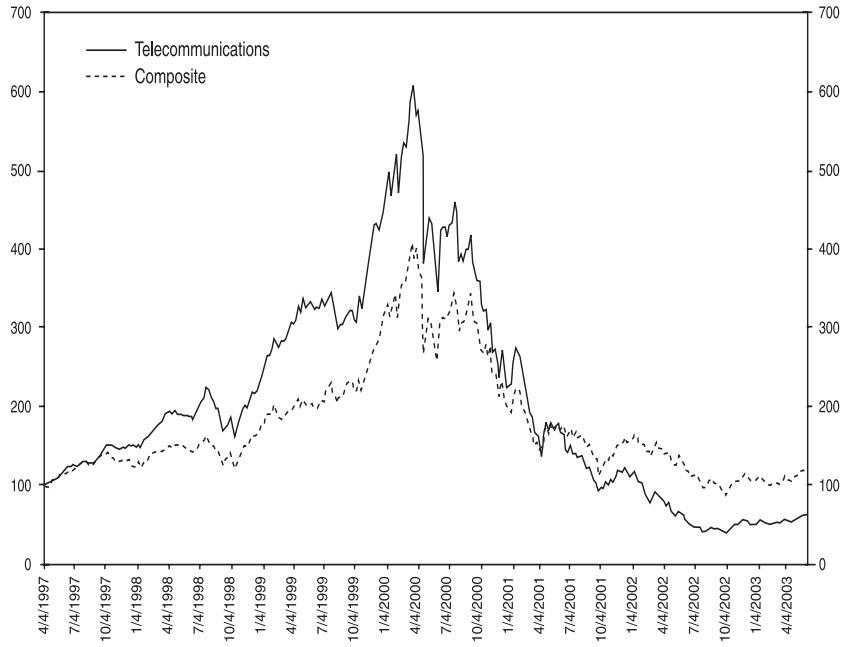
Since the breakup of AT&T in 1984, long-distance service has been provided primarily by a few large companies (such as AT&T, Sprint, or MCI) and many resellers. The 1996 Act conditionally opened the long-distance market to ILECs, and since then several of them have entered the market.

Wireless service was originally organized by the FCC as a duopoly. The FCC reserved one license for the incumbent local exchange carrier and auctioned the other. When the FCC auctioned rights to previously restricted parts

[5] On the industry's historical evolution in the United States, see Brock's chapter in Cave et al. (2002). Other chapters in that book also have been tremendously helpful to us in researching this article.

[6] Prior to 1996, four states had firms competing against the ILECs, but these accounted for only a small share of telecom revenues.

Figure 1 Nasdaq Telecommunications and Composite Indices



Note: End of week close. Normalized to April 4, 1997 = 100.

Source: Bloomberg

of the spectrum in 1995, many other firms entered the market; many areas now offer a choice of several wireless companies. Recently, wireless has become increasingly popular as a substitute for land lines (Noguchi 2002). Calls are transmitted from wireless phones to towers and then are connected to the local or long-distance networks.

Internet service is available from local phone companies, cable companies, and other providers such as AOL. Dial-up access, which still accounts for roughly 70 percent of the market (Noguchi 2003), allows users to connect to the Internet through the phone lines. Digital subscriber line (DSL) service also travels over the local loop, but is much faster than dial-up access. This service is most commonly offered by the ILEC, but any company can purchase capacity from the incumbents on a wholesale basis to resell to consumers. CLECs currently have a 20-percent market share in digital subscriber line service (Fitchard 2002). Cable companies also offer high-speed service over their own networks in some areas, and this has been more widely adopted than DSL. Both DSL and cable are commonly referred to as broadband connections.

Finally, wireless Internet services have recently gained popularity, offering access at home or at other locations with transmitters, such as coffee shops or airports.

A significant part of the telecommunications sector now consists of service wholesalers. These firms, such as Global Crossing and Level 3, constructed long-haul fiber networks in the 1990s in the hopes of selling capacity to telecom retailers and selling final services to large firms with high telecom demand.

2. QUANTIFYING THE BOOM AND BUST

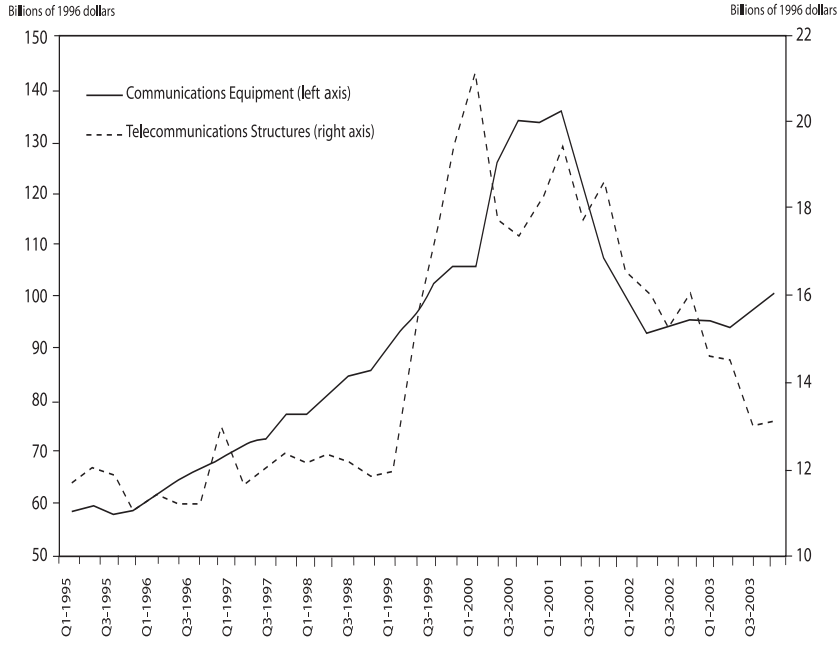
From April 1997 to March 2000, the Nasdaq index of telecommunications stocks rose spectacularly, from 198 to 1,230, an average annual increase of approximately 84 percent. As of May 16, 2003, the index stood at 136, an average annual decrease of approximately 50 percent since March 2000. To put these figures in perspective, the Nasdaq Composite Index rose and fell at respective annual rates of 61 percent and 32 percent over the same periods. Figure 1 displays a plot of the time series for the Nasdaq telecommunications and composite indices over this period, with both series normalized so that April 4, 1997, equals 100.

Equity price behavior illustrates the telecom boom and bust most vividly, but the evolution of the sector's investment spending, employment, and profitability is also dramatic. In contrast, increases in the consumption of telecommunications services and the price of local phone service, and decreases in the price of long-distance phone service have all been gradual.

From the first quarter of 1996 to the fourth quarter of 2000, investment in communications equipment grew from approximately $62 billion per year to over $135 billion per year in constant 1996 dollars (Figure 2). This represents average annual growth of nearly 18 percent. Since the final quarter of 2000, year-over-year communications investment growth was negative for seven straight quarters. In terms of investment levels, the low point came in quarter four of 2001, at under $93 billion—only 69 percent of the same figure one year earlier. As a percentage of total private investment, communications equipment fell from nearly 7 percent in 2000 to 4.8 percent at the end of 2002. Real investment in telecommunications structures was flat through most of the 1990s at approximately $12 billion. Enormous growth occurred in 1999 as investment in structures rose $9 billion in that year alone, to more than $21 billion in the fourth quarter. Such investment has fallen since then to about $13 billion at the end of 2002.

Telecommunications industry employment (services plus manufacturing) peaked at approximately 1.59 million workers in March 2001. Employment in telecom-related industries declined 22 percent—an average annual decrease of 8 percent—to about 1.30 million by July 2003 (Figure 3 shows services

Figure 2 Real Private Fixed Investment in Communications Equipment and Structures

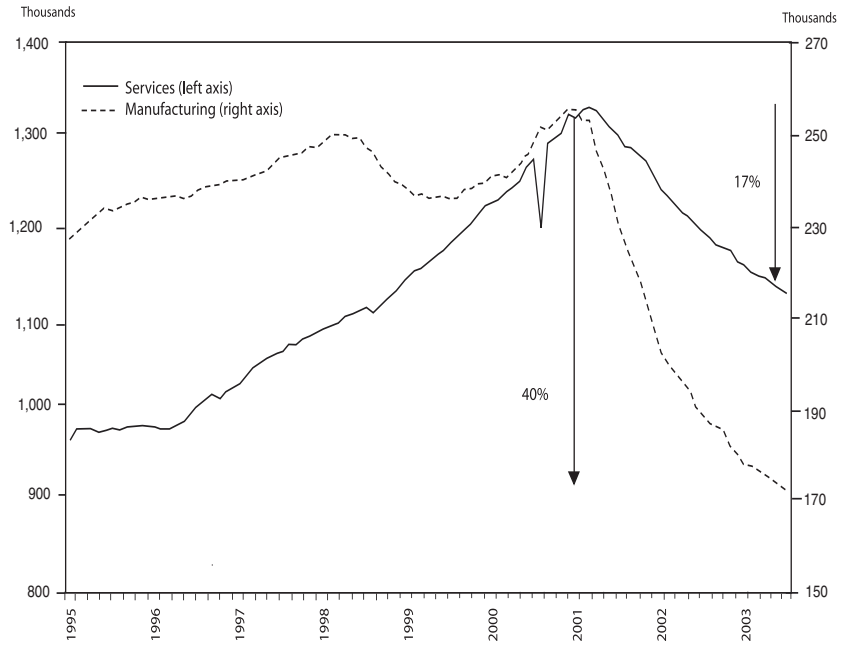


Source: BEA/Haver

and manufacturing employment separately). Announced figures for job cuts have been even more staggering, and media reports have cited numbers of over 500,000. That is nearly one-third of the sector's total employment at its peak. Observed declines in telecom employment have not been as large as the number of job cuts for two reasons. First, some new jobs were created even as others were being eliminated. Also, announced job-cut figures often include reductions in payroll through attrition, so there may be a significant lag between the announcement of cuts and observed employment declines. The boom and bust in employment is less dramatic than that in investment when measured relative to the U.S. economy. As a share of total employment, telecom employment fell only from 1.2 percent to 1.0 percent from March 2001 to July 2003.

Corporate profits for the communications industry started on a rapid downward trend after 1996. Current returns were negative for the year in which telecom stocks reached their highest market capitalization. Profits continued to be negative in 2001, the most recent year for which industry data is available:

Figure 3 Employment in Telecom Services and Manufacturing



Notes: Monthly observations. A strike at Verizon accounts for the downward spike in August 2000.

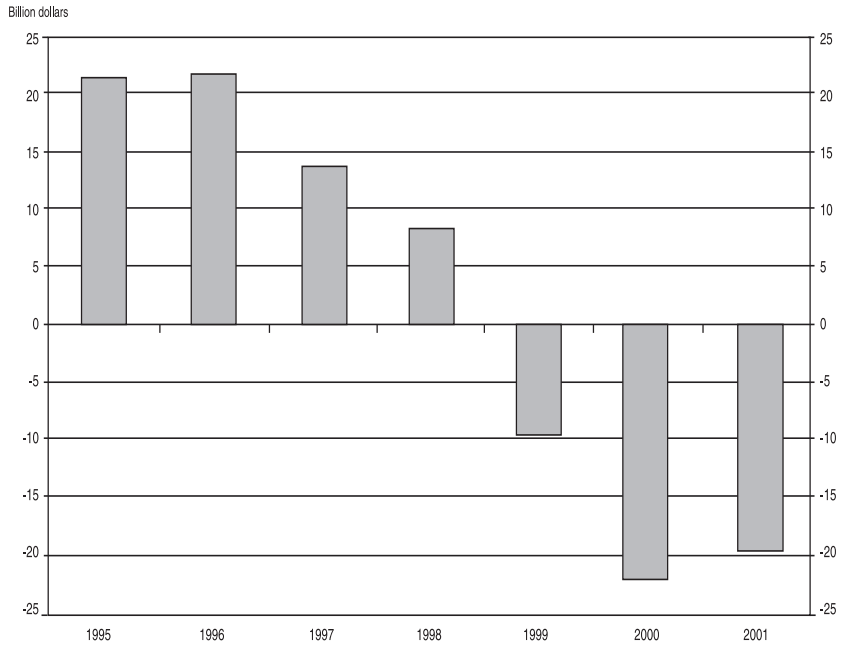
Source: BLS

the communications industry lost nearly $20 billion in 2001, as seen in Figure 4.

Consumption of telecommunications services grew steadily during the boom in investment and equity valuations, from approximately $88 billion in 1995 to $151 billion in 2001 in constant 1996 dollars. Telecom consumption's growth rate rose slightly during the boom—its average year-over-year growth was 6.7 percent from 1990 to 1995 and was 7.4 percent from 1996 to 2001. Consumption of telecom services grew faster than total consumption before, during, and after the boom. In 1995, consumption of telecom services amounted to approximately 1.7 percent of total personal consumption. By 2001, that number was 2.4 percent.

Figure 5 displays price indices for telephone service. Prices for long-distance telephone service fell 18.5 percent from December 1997 (the earliest date available) to March 2003, as measured by the consumer price index. Over the same period, prices for local service rose 21.7 percent. The rise in local

Figure 4 Corporate Profits after Tax in the Communications Industry



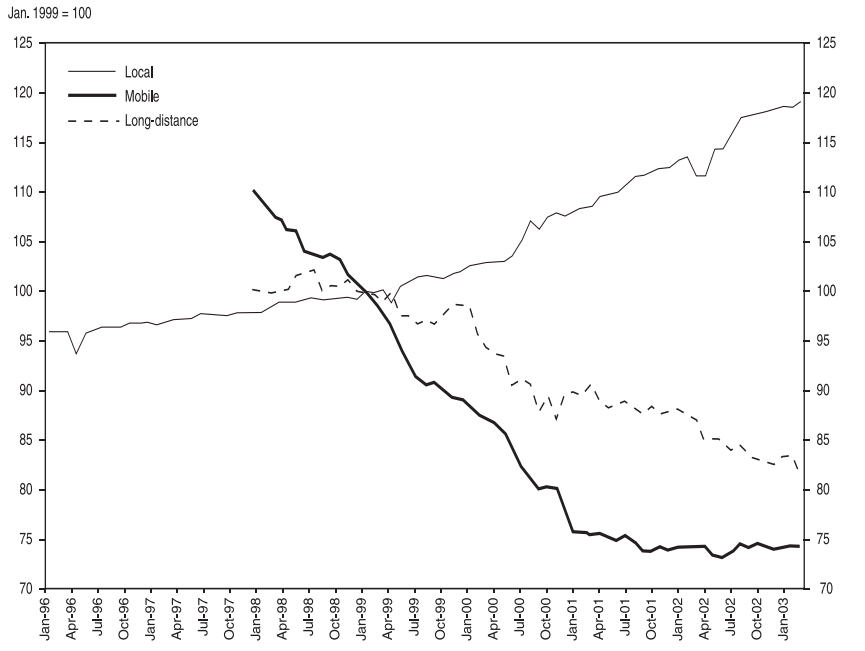
Source: BEA

service prices is particularly striking when compared with data from earlier in the 1990s. From January 1990 to January 1997, prices for local service increased only 8.9 percent. The price index for wireless service fell roughly 32 percent from July 1997 to early 2003, with most of the decline occurring before 2001.

3. UNDERSTANDING THE BOOM AND BUST

The interaction of technological and regulatory changes goes a long way toward explaining the behavior of the telecommunications industry at the turn of the 21st century. Technologies involved in producing telecommunications services advanced dramatically in the late 1990s, opening the door both to lower prices for existing services and to the introduction of a plethora of new services. At the same time, the regulatory environment appeared to be on the verge of transformation. The telecommunications boom was predicated on

Figure 5 Price Indices for Telephone Service



Source: BLS/DRI

technology and regulatory changes interacting propitiously.[7] In the event, the regulatory environment became clouded with uncertainty, undercutting the virtuous circle scenario on which the telecom boom was based.

With the benefit of hindsight, most people would say that telecommunications stocks were overvalued at their peak, and that too much investment took place in the telecommunications sector in the late 1990s. However, any time there is great uncertainty or rapid change in a market environment, one should not be surprised, *ex post*, to observe large forecast errors. Thus, our explanation for the telecommunications boom and bust does not involve fraud, irrationality, or a bubble. To be sure, as the bust became apparent, fraud did occur. But it is not clear that fraud played an important role in the boom and the early stages of the bust.

[7] Of course, technological progress was not entirely exogenous. Firms undertook research and development projects with the expectation of generating future profits.

Technology-Related Changes

While the period of the telecommunications boom saw significant improvements in technology, many of the basic elements forming the infrastructure remained the same. Switches and routers form a connection between the originator of the communication and its destination. Copper wire continues to connect most consumers to the nearest local switching center. For voice communication, an analog signal travels to a local switching center, where the signal is converted to a digital format.[8] Switches also direct the signal toward its destination. Fiber cables known as trunks carry the digital signal between switches. At some point sufficiently near the destination, the signal is converted back to analog format and directed to its destination in the local loop via copper wire.[9]

Fiber has proven to be far superior to copper in its ability to transmit data. However, the existing infrastructure running into homes and businesses primarily is made up of copper wire. Consequently, technology that increases the amount of data that can be carried over copper wire (in particular, digital subscriber line, or DSL) has been an important part of the development of telecommunications.

Technologies that increase the capacity of glass fiber also have been important. These arguably have been the most impressive advances in telecommunications in recent years. In 1996, a strand of fiber transmitted data at approximately 2.5 gigabits per second (Gbps). By 2000, the capacity of the same fiber could reach 100 Gbps.[10] This increase in capacity resulted from developments in "multiplexing," the transmission of more than one channel of information over a single medium (Freeman 1999). Instead of 2.5 Gbps over one wavelength, companies could replicate this flow over 40 wavelengths on the same fiber. Fiber capacity has since increased further, with equipment maker Cisco in July, 2002, claiming a maximum capacity of 320 Gbps over relatively short distances.[11]

A similar change took place in wireless communications. First-generation wireless was analog. Digital "second-generation" wireless networks, introduced in 1993, transmitted data at a much faster rate.[12] The shift from first- to

[8] Analog signalling uses variations in some physical property such as frequency or amplitude to transmit information. Digital signals are composed of discrete "on" or "off" units.

[9] For further explanation along these lines, see Sharkey's chapter in the *Handbook of Telecommunications* (2002).

[10] As a benchmark, a 56-kilobits-per-second (Kbps) dial-up connection is the same as a 0.000056 Gbps connection! To put the fiber capacity increase in perspective, compare the increase in capacity to the growth in the speed of integrated circuits, also considered quite rapid. Whereas the number of transistors per square inch on integrated circuits has doubled roughly every 18 months (Moore's law), fiber's capacity to transmit data doubled approximately every nine months between 1996 and 2000 (Doms Forthcoming).

[11] http://newsroom.cisco.com/dlls/prod_062402d.html

[12] Time division multiple access, the first second-generation technology, was introduced in 1993. The global system for mobile communication, based on time division multiple access tech-

second-generation technology increased the quality and security of the wireless network, and consequently increased the substitutability of wireless for wireline voice communication. Third generation digital wireless, defined by the International Telecommunications Union to be technologies with rates of 114 Kbps to 2 megabits per second (Mbps), is now widely available in South Korea. There, the maximum rate of transmission is 153 Kbps, nearly three times the top capacity of a typical dial-up connection. Tests have shown that rates as high as 1.8 Mbps are possible, but the technology has not been deployed to consumers.[13] Besides the improvements in data capacity, third-generation technology makes more efficient use of the spectrum, easing the constraints on areas with dense demand for mobile voice wireless service. However, third-generation technology is still unavailable in most areas in the United States, and its prospects for deployment are hampered by its incompatibility with earlier systems.

Another technological change affecting the telecommunications industry has been the shift from circuit to packet switching. Historically, voice calls have been circuit switched, meaning that an entire circuit—and therefore all the bandwidth on that circuit—is devoted to a single call end-to-end. Much of the capacity of the circuit goes unused. Over the past few years, as voice communication has moved to digital transmission and switching, telecommunications providers are gradually shifting to packet switching. With packet switching, the voice signal, which is analog by nature, is converted to digital packets of data. These packets can be transmitted separately to their destination, over whatever bandwidth is available. There, the data is reassembled and converted to sound again. This is the same basic process used for transmission of data over the Internet. Because bandwidth is distributed as needed, packet switching leads to more efficient use of available capacity. However, packets can be delayed or lost. Such losses are usually insignificant for data transmission, but they interfere with the quality of voice calls. Note that voice communication is transmitted and switched mainly in digital form even when circuit-based switching is used. Packet-based and circuit-based switching differ in how the network allocates bandwidth, but neither type handles information in an analog format—except at the level of the local loop.

If widely disseminated, these advances in basic technology for providing telecommunications services would have two implications. First, because the capacity of existing networks would increase dramatically, the price of existing services would be expected to fall. Second, the increase in capacity, and in speed, would lead to the development of new applications which

---

nology, is standard in Europe and most of the world. Some major U.S. carriers such as Cingular use it as well. Code division multiple access, which followed in 1995, is standard in South Korea and for U.S. carriers such as Verizon.

[13] QUALCOMM press release, Nov. 8, 1999. http://www.qualcomm.com/press/pr/releases1999/press378.html.

benefited from high-speed, high-capacity transmission. To cite one example that has already been observed, the World Wide Web is a telecommunications application which relied on relatively high-speed modems for its practicality. Looking ahead, high-quality streaming video is an application that relies on data transfer speeds greater than are currently available. The interaction between basic technology (speed and capacity) and new applications represents a virtuous circle in which new applications lead to demand for bandwidth, and demand for bandwidth provides the impetus for new supply of bandwidth, which in turn makes new, bandwidth-hungry applications feasible. To a large extent, belief in the relevance of this interaction fueled the telecommunications boom.

Changes in the Regulatory Environment

The Telecommunications Act of 1996 was designed to open up local phone service to competition. Similar liberalization of long distance in the previous decade had produced significant entry, and hopes were high that the 1996 Act would be equally successful. Prior to the 1996 Act, the telecommunications sector consisted of highly regulated monopolies in local service, competitive producers (and resellers) of long-distance services, and a large number of relatively small-scale Internet service providers. The distinctions between these sectors and between others such as cable were strictly preserved. As of September 1, 1995, a majority of states allowed competition in switched local service, but only four states (Illinois, Michigan, New York, and Washington) had any firms actively competing with the incumbent (Federal Communications Commission [Fall 1995]). And while the competitive access providers nearly doubled in size each year in the early 1990s, they accounted for less than 1 percent of revenues in 1993 (Federal Communications Commission [Spring 1995]). Meanwhile, the long-distance market had become increasingly competitive. AT&T's share of long-distance revenues had fallen to 55 percent in 1994; MCI, Sprint, and LDDS (WorldCom) together had 31 percent, and a fringe of smaller companies, 14 percent (Federal Communications Commission [Fall 1995]). By 1995, interstate toll call prices had fallen to roughly half their inflation-adjusted 1984 level (Federal Communications Commission [Spring 1995]).

The authors of the 1996 Act hoped to promote competition specifically in local phone services while maintaining universal service subsidies for residential users.[14] Economides (1999) identifies four crucial regulatory changes in the 1996 Act that were designed to encourage entry.

[14] Rural phone customers are more expensive to serve than their counterparts in more densely populated areas. In the interest of providing phone service to all at the same low prices, "universal service charges" average the cost over the two groups; the subsidy to rural customers comes at the expense of urban customers.

- All incumbents were required to sell unbundled network elements (such as rights to use the copper local loop or access to central office equipment) to entrants; the FCC and state utilities commissions would set the pricing methodology for unbundled network elements.

- Entrants were permitted to purchase at wholesale prices any ILEC service for resale.

- Incumbents and entrants were required to set reciprocal termination charges on their networks.

- Regional Bells that faced significant competition according to a list of criteria (the "competitive checklist") were permitted to enter the long-distance market.

Other rules pertained to cable, Internet, and long-distance service, but were not as sweeping (Economides 1999).

While the Act clearly aimed to bring competition to local telephone service, the specific means of implementation were ambiguous and difficult to interpret. The Act endowed the FCC with considerable discretion in implementing the Act's provisions; the telecoms used a variety of legal tactics to shape the FCC's interpretation of the Act. When the FCC's choices favored entrants, the incumbents challenged provisions in court, and vice versa when the FCC's choices favored incumbents. Of course, challenges were typically met with counter challenges (either by the FCC, state regulators or one segment of industry), further complicating implementation.

Incumbents challenged the FCC's rules concerning (1) whether the FCC had the authority to institute unbundled network element schemes, (2) which network elements must be unbundled, and (3) what conditions entrants must satisfy in order to gain access to those elements. A series of court cases ending with a January 1999 Supreme Court decision established the FCC's jurisdiction. In February 2003 the FCC completed revised rules for unbundling exempting upgraded systems from resale and allowing states to grant further exemptions, but leaving the unbundled network element platform largely intact.

Incumbents and state utilities commissions fought against entrants and the FCC over the FCC's choice of total element long-run incremental cost as the pricing methodology for unbundled network elements. This pricing scheme is based upon the forward-looking cost faced by a hypothetical efficient network, including "reasonable" profits for the incumbents. Believing that the pricing methodology would not allow them to recapture the costs of their network, the incumbents challenged the FCC's pricing order in court. Arguments that the methodology was contrary to the intent of the 1996 Act or was unconstitutional were rejected by the Supreme Court in May 2002.

The FCC's 1999 Collocation Order allowed entrants to place necessary equipment in incumbents' central offices and set a cost-recovery methodology for collocation. Incumbents who felt the entrants were given too much access challenged the order, and the Washington, D.C. Circuit Court of Appeals issued a mixed decision in March of 2000. The court agreed with the incumbents that the definitions of "necessary" and "physical collocation" were too broad; however, it approved other features of the Collocation Order, including the FCC's cost recovery methodology and a broad definition of the premises to which entrants had access (Ryan 2000).

Since 1978, the FCC has set rates for cable and telephone companies that were able to establish that electric utilities were charging monopoly rents for the right to string wires from utility poles; the 1996 Act gave the FCC authority to set pole attachment rates for *all* telecommunications providers. In 1998, the FCC added cable Internet and wireless attachments to the list of regulated attachments. The power companies challenged that policy in court, arguing that because "telecommunications services" did not include cable Internet, an "information service," the FCC could not set rates. The Supreme Court agreed with the FCC that the 1996 Act had in fact granted that authority, and the rules were upheld. Internet access charges, universal service subsidies, and the competitive checklist, among other things, have also been the cause of controversy.

The Industry Responds with Boom and Bust

With dramatic changes in basic technology, new products, and the regulatory environment, it is not surprising that during the period from 1996 to 2002 the telecommunications sector experienced significant volatility. The magnitude of the volatility, and the fact that it involved a sharp ascent followed immediately by a sharp descent, is nonetheless striking. Some observers have blamed fraud and irrationality for the boom and bust, and others have described the episode as a bubble. We see the boom and bust as—in large part—a rational response to the changing fundamentals of technology and regulatory environment.

Boom

In the wake of the 1996 Telecommunications Act, there was tremendous optimism about the eventual opening up of local telephony to competition. With the local exchange open to competition, all manner of firms would be free to compete to be the single provider of a household's or business's telecommunications services (that is, local, long distance, data, and wireless). It was expected that the 1996 Act would encourage the competition and innovation seen in the long-distance market after the breakup of AT&T in 1984. Services

would become cheaper for business users especially, and new services would become available. Writing in May 1996, Dennis R. Patrick, FCC chairman from 1987 to 1989, expounded the early optimistic view: "The Telecommunications Act of 1996 represents a significant milestone. It announces that the federal government is finally, largely, out of the way, or at least headed in that direction. It will usher in an era of radical transformation in the industry the scope and import of which will make divestiture [of AT&T] look like a footnote in history" (Patrick 1996).

Early optimism was mitigated somewhat by questions about exactly how the Act would be implemented, but these questions were expected to be resolved relatively quickly. Thus, the regulatory uncertainty that existed in the immediate aftermath of the Act's passage was a secondary factor; it may have affected where telecommunications investment was channeled, but did little to discourage investment in the industry as a whole.

Questions about the Act's implementation were most pressing in the short-run for new entrants, but early FCC rulings and court decisions seemed to bear out optimistic assessments of the entrants' prospects. The pricing methodology that the FCC had chosen for unbundled elements was favorable for entrants, making it appealing for those firms to compete by leasing at least some unbundled elements rather than by building entirely separate facilities. The Supreme Court's January 1999 decision in *AT&T vs. Iowa Utilities Board* supported the FCC's authority over pricing, and this was widely interpreted as a victory for entrants (CLECs). Robert Taylor, chief executive officer of Focal Communications, a Chicago competitive local exchange carrier, called the decision "great news for CLECs,"[15] and William Kennard, FCC chairman at the time, said that the ruling would create certainty in the industry.[16]

The competitive local exchange carriers—while relatively small—experienced a tremendous boom after the Act was passed. From 1996 to 2000 the number of CLECs rose from 30 to 711, and their revenue increased from less than $5 billion to $43 billion over the same period. From 1996 to 1999 CLECs' market capitalization rose from about $3 billion to $86 billion.[17] Over this same period, however, S&P 500 telecommunications services companies grew in market capitalization by about $500 billion. Thus, while the growth rate of the entrants was high by any measure, the increase in their market capitalization did not account for a large part of the telecom boom.

Investment from 1996 to 2000 was channeled primarily into long-haul fiber optic networks. There were few regulatory barriers to building such networks, and the value of these networks was expected to rise for two reasons. First, as mentioned above, eventual opening of local exchanges to competition

[15] Quoted in Schmelling (1999).

[16] Quoted in Mills (1999).

[17] Sources: FCC, Association for Local Telecommunications Services, and Progress and Freedom Foundation. Cited in Lenard (2002).

would allow owners of such networks to compete to be a single provider; this was viewed as a prize, particularly if a firm could attract a large number of customers.[18] Second, Internet use was growing rapidly, and with it the demand for bandwidth was increasing. From 1994 to 1996, traffic on Internet backbones in the United States is estimated to have grown from 16.3 to 1,500 terabits per month [19](Odlyzko 2002). Rapid growth in demand for bandwidth was widely forecast to continue as part of the virtuous circle, with new applications being developed to take advantage of bandwidth as it came online.[20] A May 1998 article about Qwest in *Wired* typified this view:

> Qwest is operating under an if-you-build-it-they-will-come vision. Bandwidth restrictions, the company believes, have held back development of all manner of innovation. Now the prospect of virtually endless throughput will free up the planet for a host of new applications in such areas as high-speed video and multimedia. (Diamond 1998)

Spurred by expected increases in demand for bandwidth from the Internet and by the promise of future access to local exchanges, construction of long-haul fiber networks exploded after 1996. Much of this investment was undertaken by new firms such as Qwest, Level 3, and IXC. In 1996, the "old guard" of AT&T, MCI, WorldCom, and Sprint together accounted for 72 percent of long-haul fiber in the United States, but by 1999 they accounted for only 30 percent of the total. Over this same period, annual fiber deployment increased more than four-fold (Dunay 2000). One of the major producers of fiber was Lucent Technologies. Early in 2000, Lucent was expanding its facilities to enable it to increase fiber output by 60 percent. A Lucent executive said, "We've seen fiber growth at 17 percent forever. Now we think the growth rate will be 30 percent this year. There's an enormous amount of fiber required to have the penetration needed by long-hauls, cable, and others."[21]

One of the mantras of the telecom boom was that Internet use doubles every three to four months. Many people attribute the origins of the statement to WorldCom (now called MCI) (Dreazen 2002). WorldCom carried the plurality of Internet traffic for a time, so their reports may have carried substantial weight (Sidak 2003).[22] Even so, the real effects of such a claim and the extent to which WorldCom should be faulted are hard to establish. According to research by Kerry Coffman and Andrew Odlyzko (2002), such growth did in fact occur for a time in 1995 and 1996. They estimate that the amount of data sent over the Internet has approximately doubled every year

---

[18] This reasoning relies on some form of increasing returns to scale.

[19] "A terabit is one trillion bits."

[20] The "virtuous circle" involves complementarity between applications and network capacity.

[21] Kuhl (2000). The executive quoted is Tim Cahall.

[22] Prior to 1995, the National Science Foundation administered the backbone for the Internet and kept accurate records of its growth. However, private backbones replaced the government's during 1995, hence public data was no longer available.

since then. However, throughout the boom major players outside WorldCom, such as Duane Ackerman, CEO of BellSouth, continued to assert that Internet traffic was doubling every 100 days (Calicchio 1999). In addition, although WorldCom was the biggest carrier, Sprint also carried a large portion of Internet traffic (16 percent to WorldCom's 37 percent, according to the U.S. Department of Justice's announcement that it was suing to block WorldCom from acquiring Sprint [2000]).

During the boom period, contrarian forecasts of Internet use and the resulting demand for fiber could be heard. Odlyzko has pointed out that growth rates of 100 percent every three months would have implied that between 1994 and 2000 Internet use grew by a factor of 17 million (Odlyzko 2002). Forecasts based upon those growth rates and 1994 Internet usage data have every Internet user in the year 2000 constantly downloading streaming video. Even admitting that in 1998 no one knew what applications would be available in 2000, it is difficult not to view this growth rate estimate as excessively optimistic. In the contrarian view, fiber deployment based on such optimistic forecasts would also be excessive: a May 7, 1999, opinion piece from the *Industry Standard* referred to "an unprecedented network overbuild and a looming glut of bandwidth and connectivity. Precious capital has been funneled into too much connectivity, and too few smart applications that could put all this bandwidth to use" (Aguirre and Bruneau 1999).

Pessimistic views regarding the progress in implementing the Act could also be heard. For example, the view that the Supreme Court's 1999 decision would create certainty was not held by all. Writing in the *Business Communications Review*, March 1999, Michael Weingarten argued that in the wake of the January 1999 decision, "matters may be as uncertain as ever" (Weingarten 1999). As Weingarten noted, the decision settled neither the precise set of unbundled elements which incumbents were required to provide, nor the precise pricing scheme to be used.

In the presence of rapidly changing technologies and market conditions it is not surprising that there was heterogeneity in forecasts. During the telecommunications boom, market outcomes evidently reflected the optimists more than the pessimists. Recent research in finance has suggested that when there are heterogeneous forecasts associated with new or rapidly changing technologies, pessimistic voices will have "too small" an effect on the market. These theories rely on restrictions on taking short positions in stocks. If the distribution of forecasts has a mean at the true expected value, it may nonetheless be the case that equity prices reflect a higher value.[23]

[23] See Ofek and Richardson (2003) and Scheinkman and Xiong (Forthcoming).

Bust

Even in 1996, industry observers did not believe that competition would arrive overnight in local telecommunications. By late 2000, however, four years had passed, meaningful competition had not arrived, and implementation of the 1996 Act was bogged down in the courts. In addition, the macroeconomy was weakening, and it was becoming clear that there was significant overcapacity in the long-haul fiber market. Together these factors spelled gloom for the telecommunications sector.

While the competitive local exchange carriers grew extremely fast from 1996 to 2000, their share of the local telephone market was still small, less than 8 percent in 2000.[24] Furthermore, only about 40 percent of that share comprised so-called facilities-based competition, that is, local service provided by competitors using their own lines rather than by reselling ILEC service or by purchasing some unbundled elements from incumbents. This strategy left them particularly exposed to the adverse ruling on the FCC's pricing methodology by the Eighth Circuit Court of Appeals in *Iowa Utilities Board vs. FCC*, which in July 2000 moved in the opposite direction from the 1999 decision. The market capitalization of CLECs fell 63 percent from $86.4 billion in 1999 to $32.1 billion in February of 2001 and then 88 percent to just $3.77 billion in February of 2002.[25] In contrast, the respective market values of two major ILECs, BellSouth and Qwest, each fell less than 15 percent from March to December 2000.[26] Relative equity valuations, together with the bankruptcy of many CLECs, suggests that the ILECs' market power increased after the Eighth Circuit's decision. This assessment is supported by the price indices displayed in Figure 5; the price of local telephone service relative to long-distance and wireless rose noticeably after July 2000.

On March 10, 2000, the Nasdaq telecom index peaked at 1,230.06; by the end of 2000 it had fallen by 62 percent. With hindsight, it is clear that 2000 was the year in which the telecommunications industry began its sharp decline. If anything, this decline was especially pronounced in the long-haul fiber segment. However, industry observers did not generally catch this development before late 2000. Early in 2000, we saw that Lucent was optimistic about demand for fiber, and even as share prices had begun to fall, in September of 2000 *Broadband Week* published an article with the headline, "Future Looks Bright for Fiber Optic Manufacturers." It soon became apparent, though, that there was massive overcapacity in long-haul fiber. Media reports of the glut in long-haul fiber became widespread early in 2001. In an article titled "The

[24] FCC, cited in Crandall (2002).

[25] Association for Local Telecommunications Services, and Progress and Freedom Foundation. Cited in Lenard (2002).

[26] Large mergers completed in 2000 greatly increased the market value of the other two ILECs, SBC and Verizon (formed from Bell Atlantic and GTE).

Coming Bandwidth Bubble Burst," Grahame Lynch wrote in *America's Network*, February 1, 2001, "It's the pain phase for America's fiber barons. Nearly 600,000 miles of new inter-city fiber is on the way. Capacity prices are dropping and major dot.com and CLEC customers are failing." And by June 2001, when Canadian equipment producer Nortel announced a $19 billion quarterly loss, the bust was clear to all. Compounding the problems that were specific to the telecommunications sector, the U.S. economy weakened over the course of 2000, with the National Bureau of Economic Research eventually declaring that a recession had begun in March 2001. This broad decline in economic activity coincided with the regulatory turmoil to send the industry into a sharp decline in 2000 and 2001, from which it still may not have emerged.

Overcapacity in long-haul fiber had three sources. First, the long-haul fiber industry was in its early stages, and it is typical in the evolution of an industry to see an initial overshooting of investment, followed by a shakeout period (Klepper 2002). Second, the dramatic increase in the capacity of a given strand of fiber may have been greater than anticipated when construction on various networks was begun (Sidak 2003, 216). Third, and perhaps most importantly, demand for long-haul fiber capacity had not grown as fast as many had forecast: the pessimists turned out to be right.

Above we explained the forecasts of high growth in demand for bandwidth as being based on the positive interaction between increases in bandwidth and the development of new applications to soak up that bandwidth. This interaction did occur; as average bandwidth to households has increased (mainly through digital subscriber line and cable broadband), it has become increasingly common for music to be disseminated over the Internet. However, the magnitude of increases in demand for bandwidth has been small compared to the forecasts embedded in equity valuations and investment numbers. The optimistic forecasts seem to have been based on a much wider adoption of fiber-to-the-home than actually occurred. Because the 1996 Act's implementation has been bogged down in the courts, neither ILECs or CLECs have undertaken large-scale investments in fiber-to-the-home, and thus bottlenecks at the level of the local loop remain (this is often referred to as the last-mile problem).

4. CONCLUSION

At any given time, some sectors of the U.S. economy are expanding and others are contracting. The behavior of the telecommunications sector since 1996 is particularly interesting because the magnitudes are so great. The decrease in market capitalization of S&P telecommunications firms alone from 2000 to 2002 was roughly $700 billion, more than 3.5 percent of the entire value of U.S. corporate equities at the stock market peak in 2000.

According to our analysis, the 1996 Telecommunications Act was an important factor in both the boom and the bust. High hopes for a new world of competition in telecommunications followed passage of the Act, and played a major role in the dramatic rise in equity valuations. Even as the boom was effectively over, in February 2000, then FCC Chairman William Kennard spoke of "the miracle of the American model for unleashing competition in telecommunications," competition that was "creating unprecedented investment and job growth in every sector of the communications industry."[27] Two years later, with the bust apparent to all, Kennard's successor Michael Powell described it in a speech as "an unbelievable disaster," and did not hesitate to assign some of the blame to "legal instability in the court system." Referring to the Telecommunications Act of 1996, Powell said

> I have rarely seen a 750,000-word document come out of the United States Congress with clarity, and I have rarely seen one that long and complex that isn't going to trigger years of uncertainty and litigation about the parameters of that statute. I was always sort of amazed by the degree to which people didn't have that expectation built into the way things would go.[28]

Of course, some people did have that expectation built into their forecasts, but market valuations were more optimistic. We do not have a definitive explanation for the market's valuations. However, theories of asset pricing in the presence of heterogeneous beliefs and restrictions on short sales imply that asset valuations will be driven by the market's optimists. Optimism about the fundamentals of the telecom sector was widespread during the boom years, leading us to be skeptical about claims that there was a bubble in telecom stocks.

In addition to the 1996 Act, technological advances in telecommunications also played important roles in both the boom and the bust. Investment in long-haul fiber was predicated on the idea that as-yet-unknown applications would be developed to take advantage of the new bandwidth. Failure of those applications to materialize at the rate that had been predicted translated into a capacity glut, and the glut was exacerbated by the dramatic advances in technology for increasing the capacity of each strand of fiber.

While our analysis of the telecommunications boom and bust has merely touched the surface of this issue, we do come away with two recommendations for policymakers. First, they should take seriously the idea that lack of clarity in the regulatory framework under which an industry operates can lead to substantial volatility in that industry. Our second recommendation is related

---

[27] Speech to National Press Club, February 8, 2000.

[28] Remarks of Michael K. Powell, Chairman, Federal Communications Commission, at the Thomas Weisel Partners Growth Forum 4.0, Santa Barbara, California, June 17, 2002.

to the mantra or myth of Internet traffic doubling every three months. While we are skeptical of the extent to which irrational belief in such growth rates drove the telecom boom, it is clear that good aggregate data on Internet use was difficult, if not impossible, to acquire after 1994. The federal government is involved in many data collection efforts, and the data it collects are viewed as a public good. With the benefit of hindsight, collection and dissemination of data on Internet use would have been a productive activity for the U.S. government to be involved in during this period, and will be in the future.[29]

There is much room for future work on the telecom boom and bust. Here we mention just two areas of interest. First, while telecommunications experienced particularly extreme fluctuations from 1996 to 2002, other sectors also rose and fell, as did the U.S. economy as a whole. Biotechnology, in particular, experienced fluctuations of nearly the same magnitude as telecom, though the spike in biotech was very brief and came toward the end of the telecom boom. A comparative study of biotech and telecom might be revealing about the causes of the fluctuations in both sectors. Second, there have been other episodes of sectoral booms and busts in the history of the United States, and one that immediately invites comparison with telecom is the railroad boom and bust of the 1870s. Like telecommunications, railroads consist of networks, and a comparative study of these episodes would shed light on the question of whether network industries are particularly prone to large fluctuations.[30]

REFERENCES

Aguirre, Pascal, and Mark R. Bruneau. 1999. "Too Much Bandwidth." *The Industry Standard* (7 May). http://www.thestandard.com/ (accessed June 4, 2003).

Calicchio, Dominick. 1999. "Front End." *InformationWeek* 721 (15 February): 14.

Cave, Martin E., Sumit K. Majumdar, and Ingo Vogelsang. 2002. *Handbook of Telecommunications Economics.* Boston: North-Holland/Elsevier.

Cisco Systems. 1995. "Cisco Roll Out Newest Metro DWDM Platform."

[29] Sidak (2003) cited FCC Commissioner Michael Copps as making this argument in testimony before a Senate Committee.

[30] The analogy between telecom and railroads has been made by many. The first reference we have found is the August 31, 2001, episode of PBS's *NewsHour with Jim Lehrer* (PBS, 2001).

News Release (24 June). http://newsroom.cisco.com/dlls/prod_062402d.html.

Coffman, K.G., and Andrew Odlyzko. 2002. "Growth of the Internet." In *Optical Fiber Telecommunications IV B: Systems and Impairments*, edited by I. P. Kaminow and T. Li. New York: Academic Press, 17–56.

Crandall, Robert. 2002. "A Somewhat Better Connection." *Regulation* 25 (Summer): 22–28.

Diamond, David. 1998. "Building the Future-Proof Telco." *Wired* (May). http://www.wired.com/wired/archive/6.05/qwest.html?topic=connectivity&topic_set=newtechnology (accessed April 30, 2003).

Doms, Mark. Forthcoming. "Communications Equipment: What Has Happened to Prices?" In *Measuring Capital in the New Economy.* Chicago: NBER/CRIW, University of Chicago Press. Cited with permission.

Dreazen, Yochi J. 2002. "Wildly Optimistic Data Drove Telecoms to Build Fiber Glut." *Wall Street Journal*, 26 September, B1.

Dunay, Neil G. 2000. "Miles to Go: Fiber Build Booms—But for How Long?" *Phone+* (May). http://www.phoneplusmag.com/ (accessed May 30, 2003).

Economides, Nicholas. 1999. "The Telecommunications Act of 1996 and Its Impact." *Japan and the World Economy* 11: 455–83.

Federal Communications Commission. 1995. "Common Carrier Competition Report." Staff report (Spring).

_____. 1995. "Common Carrier Competition Report." Staff report (Fall).

Fitchard, Kevin. 2002. "DSL Issues a Wake-Up Call to Cable." *Telephony* 243, no. 18 (December): 11. http://telephonyonline.com/ar/telecom_dsl_issues_wakeup/index.htm.

Freeman, Roger L. 1999. *Fundamentals of Telecommunications*. New York: Wiley.

Klepper, Steven. 2002. "Firm Survival and the Evolution of Oligopoly." *Rand Journal of Economics* 33 (Spring): 37–61.

Kuhl, Craig. 2000. "Worldwide Appetite for Fiber is Voracious." *Communications Engineering and Design* (May). http://www.cedmagazine.com (accessed February 18, 2003).

Lenard, Thomas M. 2002. "The Economics of the Telecom Meltdown." The Progress and Freedom Foundation *Progress on Point* 9.6 (February).

Lynch, Grahame. 2001. "The Coming Bandwidth Bubble Burst." *America's Network* 105 (February): 36–40.

Mills, Mike. 1999. "Supreme Court Backs FCC on Phone Rules." *The Washington Post*, 26 January, C01. http://www.washingtonpost.com /wp-srv/national/longterm/supcourt/stories/court101398.htm (accessed May 9, 2003).

Noguchi, Yuki. 2002. "Verizon to Cut 665 Jobs in the D.C. Area; Firms Losing Business to Cell Phones." *The Washington Post*, 16 April, E05. http://www.washigntonpost.com/ac2/wp-dyn?pagename=article &node=&contentId=A34191-2003Apr15&notFound=true.

_____. 2003. "Broadband's Spread Slowing." *The Washington Post,* 22 May, E05. http://washingtonpost.com/ac2/wp-dyn/A23117-2003May21?language=printer.

Odlyzko, Andrew. 2002. "Measurements and mismeasurements and the dynamics of data traffic growth." *Computer Measurement Group's 2002 International Conference*, Reno, Nev. (11 December).

Ofek, Eli, and Matthew Richardson. 2003. "DotCom Mania: The Rise and Fall of Internet Stock Prices." *Journal of Finance* 58 (June): 1113–38.

Patrick, Dennis R. 1996. "The Telecommunications Act of 1996: Intent, Impact and Implications." Progress and Freedom Foundation. http://www.pff.org/cad/patr051496.html (accessed May 14, 2003).

PBS. 2001. "Boom and Bust: The Telecommunications Industry is Suffering from Severe Financial Troubles." *NewsHour with Jim Lehrer* (31 August).

QUALCOMM. 1999. "QUALCOMM Unveils Wireless Internet Strategy." News Release (8 November). http://www.qualcomm.com/press /pr/releases1999/press378.html.

Ryan, Vincent. 2000. "Win-win for Incumbents, CLECs?" *Telephony* 238, no. 13 (February): 10. http://currentissue.telephonyonline.com /ar/telecom_winwin_incumbents_clecs/ (accessed May 15, 2003).

Scheinkman, Jose, and Wei Xiong. "Overconfidence and Speculative Bubbles." Forthcoming in *Journal of Political Economy*.

Schmelling, Sarah. 1999. "Let Them Decide." *Telephony* 236, no. 5 (February): 7. http://telephonyonline.com/ar/telecom _let_decide_supreme/index.htm (accessed February 5, 2003).

Sidak, J. Gregory. 2003. "The Failure of Good Intentions." *Yale Journal on Regulation* 20 (Summer): 207–67.

U.S. Department of Justice. 2000. "Justice Department Sues to Block WorldCom's Acquisition of Sprint." Press release, http://www.usdoj.gov/ (27 June).

Weingarten, Michael. 1999. "So You Thought the Supreme Court had Decided Things. . ." *Business Communications Review* 29 (March): 35.

Implications of the Capital-Embodiment Revolution for Directed R&D and Wage Inequality

Andreas Hornstein and Per Krusell

Wage inequality has increased dramatically in the United States since the late 1970s. In particular, we have witnessed growing wage differences between groups defined by observed skills such as education or experience. For example, the college premium—that is, the percentage difference between the average wages of college-educated and non-college-educated workers—increased by a factor of four. Since at the same time the relative supply of college-educated workers increased, we would have expected to see a fall of the college premium. The fact that a decrease did not occur suggests that something else changed too. A natural candidate is technical change that has been "biased" toward skilled labor over this time. If the nature of technical change makes skilled workers relatively more productive than unskilled workers, then the wage gap will widen, assuming that market wages reflect marginal productivities. But why should technical change be biased more toward skilled labor? In fact, technical change sometimes has been biased the other way. From a perspective of understanding the evolution of wage inequality, then, it is important to determine the possible bias of technical change.

In this article we investigate the long-term determinants of the bias of technical change using a dynamic model where R&D is endogenous and can be directed to specific inputs. One of the key determinants of the form of technical

Andreas Hornstein is with the Federal Reserve Bank of Richmond. Per Krusell is with the University of Rochester, the Institute for International Economic Studies, CAERP, NBER, and CEPR. We would like to thank Marvin Goodfriend, Kartik Athreya, and Bob Hetzel for helpful comments. The views expressed in this article are not necessarily those of the Federal Reserve Bank or the Federal Reserve System.

change, then, is wage inequality itself: with a high value of skilled workers—a high skill premium—the value of new technologies directed for use with skilled workers will rise. Thus, in our theory, wage inequality and technology are simultaneously determined through a two-way feedback. We first study the long-run determination of wages and technologies by considering long-run outcomes: steady states.

Our ultimate aim, however, is to understand what causes changes to the equilibrium wage inequality. In particular, we want to evaluate the role of the IT revolution in shaping the last thirty years of wages and productivity. We think of the IT revolution as having been initiated in the mid-1970s; the defining event was that the relative price of new capital, which is complementary to skilled labor, fell significantly. We then consider two quantitative experiments. First, we consider a one-time fall in the relative price of new capital, which allows us to trace out the short-run dynamics of this model: In response to this impulse, how do wage inequality and the induced directed R&D react? We then consider a gradual and persistent fall in the relative price of new capital aimed at matching the actual behavior of this price series as measured by U.S. data. Now the question is quantitative: What is the possible role of the IT revolution, viewed this way, in accounting for the observed increase in wage inequality and associated changes in productivity?

Why Is Wage Inequality Relevant to Macroeconomists?

Our quantitative theory has joint implications for wage inequality and technology. Thus, not only can such a theory tell us how technical change influences relative wages, but it allows us to use wages to understand the nature of technical change. In particular, not only do wages reflect current marginal productivities, but they are also relevant for understanding where current R&D efforts are directed—both its composition and its effect on aggregate productivity—and thus for predicting future productivity movements. We therefore believe that, on a general methodological level, the development of quantitative theories of the joint determination of wage inequality and technology is important for furthering our understanding of aggregate economic performance.

Because of the connection between wages and technology, wage data are an interesting testing ground for different theories about what is going on in the aggregate economy. Namely, there has been widespread interest in what has happened to aggregate productivity, especially in light of the "IT revolution": has IT technology, and all the changes in the workplace it seems to have led to, also delivered higher productivity? In conducting stabilization policy especially, monetary or otherwise, information on the behavior of productivity is useful. Relatedly, is there unmeasured quality improvement in the goods and services produced by the new economy? This information is particularly

important in understanding how inflation really has influenced the purchasing power of our money: with significant unmeasured quality improvements, we are better off than the inflation figures indicate. To the extent that wage inequality speaks indirectly about productivity advances of different sorts, it is therefore arguably an important variable to follow.

Aside from the role wage inequality has as an indicator of what is happening—and what will happen—to aggregate economic performance, it is also relevant in itself and for understanding the political debate. Most obviously, wage inequality is often part of the distributional goals of policymakers (and voters), and indications of widening wage inequality may be taken as cause for some kind of action by these groups. As economists, we perhaps have instinctive reactions to caution against policies aimed at reducing wage inequality, since we think they may reduce workers' efforts to work hard, accumulate human capital, and so on. The theory in this paper suggests that there are other reasons to react: reductions in wage inequality will certainly change the composition of R&D, and thus the nature of technology, and they are likely to change aggregate productivity growth as well.

To the extent that externalities in research and labor market frictions are not important, the market mechanism probably channels the R&D efforts to its different uses quite efficiently, and thus one should caution against policies leading to wage compression. However, with an imperfectly functioning market, the situation is more complicated. We do not characterize optimal policy in the environment we study, but one could. Do the market imperfections lead to too much or to too little wage inequality? The answer likely depends on details of the imperfections, including those in the labor markets (which we abstract from in this article). It is even a logical possibility that there is *too much* equilibrium wage inequality from the perspective of efficient R&D and that wage-compressing policies would be beneficial! However, it might also be the reverse: such policies might be even more harmful than indicated by our knee-jerk reactions. We hope to be able to address these important issues in future work.

Capital-Embodied Technical Change and Wage Inequality

A main purpose of our paper is the study of the short-, medium-, and long-run effects on the economy of an "IT revolution": of a burst in capital-embodied technical change. In particular, we focus on its role in wage inequality between skilled and unskilled labor and subsequent R&D efforts. In the postwar U.S. economy, capital-embodied technical change seems to have been an important source of growth. As argued in Greenwood, Hercowitz, and Krusell (1997), to a first approximation, capital-embodied technical change is reflected in the decline of the price of new capital goods (such as computers and other

equipment) relative to the price of consumption goods. Since in the United States the relative price of new capital has been falling at an annual rate of close to 3 percent, this channel has been responsible for a sizable fraction of overall growth.

The implications of capital-embodied technology for wages alone have been studied previously. In earlier work, Krusell, Ohanian, Rios-Rull, and Violante (referred to hereafter as KORV) (2000) estimate features of the aggregate production function and use these features to argue that a higher capital stock, induced by the fall in the price of new capital, must have increased the relative productivity, and thus wage, of skilled labor, that is, the skill premium. The argument in KORV (2000) is based on a partial equilibrium analysis and takes relative factor productivities and relative factor supplies as given. In this paper we also take the latter as given; we take the view that whereas the relative supply of skilled labor can be expected to change, it is unlikely to be very elastic. For example, if we identify skilled labor with college graduates, then we might expect that, because of inherent ability-based differences, the supply of college graduates has an upper limit or, alternatively, that the average quality of college graduates would tend to fall as more students choose to go to college.

The contribution of this paper is the analysis of the equilibrium response of relative factor productivities to changes in the relative price of capital. Unlike changes in the relative supply of labor, there does not seem to be a natural upper limit to technology improvements, in particular to the relative improvements of different applications. In a number of recent papers, Acemoglu (1998, 2002a, 2002b, 2003) has argued forcefully and repeatedly that technical change is endogenous and is purposefully directed to different uses, that is, specialized for different kinds of workers/machines. We apply Acemoglu's framework to the particular question of how changes in the relative price of capital affect the relative incentives for productivity improvements that are specific to capital, skilled labor, and unskilled labor.[1]

The argument in KORV (2000) that capital accumulation increases the skill premium is based on the different substitution possibilities between the inputs capital, skilled labor, and unskilled labor in the aggregate production function. For any pair of inputs, basic economic theory suggests that if firms minimize cost, then an input that becomes relatively more expensive is used relatively less, holding the output to be produced fixed. In other words, the relative input ratio falls as the relative price increases. The question is whether the relative

[1] Acemoglu (2002b) has studied how the interaction of directed R&D with a change in the relative supply of skilled labor affects wage inequality.

input ratio falls relatively more or less than the relative price increases. We say that two inputs are substitutes (complements) if following a 1 percent increase of the relative input price, the relative input use declines by more (less) than 1 percent.[2] Alternatively, we can ask by how much relative input prices have to change such that input markets clear if the relative supply of inputs changes. Thus, if two inputs are substitutes (complements) and the relative supply of one input increases by 1 percent, then the relative price of that input has to fall by less (more) than 1 percent such that the input markets clear.

Based on a wide range of empirical work and on independent estimation, KORV (2000) argue that skilled labor is more complementary to capital, whereas unskilled labor is more substitutable for capital. A higher capital stock reduces the supply of skilled labor and unskilled labor relative to capital. Holding the labor endowments and productivities fixed, the price of skilled and unskilled labor relative to the price of capital thus increases in an equilibrium. Since skilled labor is complementary to capital, whereas unskilled labor is a substitute for capital, the price of skilled labor relative to capital has to increase more than the price of unskilled labor relative to capital. Therefore, the wage of skilled labor increases relative to the wage of unskilled labor.

Directed Technical Change and Factor Productivity

A major technological event such as the IT revolution will affect not only the accumulation of capital but also the way R&D is conducted. In general, we expect that R&D is purposefully directed toward improving the productivity of activities where it will receive the highest rewards. From our perspective, the important distinction is whether R&D is directed toward improving the productivity of skilled labor or unskilled labor, or whether it is used to further increase the productivity of existing equipment capital. Many recent technology developments seem skill-biased; for example, the development of advanced software is performed by skilled labor. However, there are many examples of how IT technology might also help unskilled labor improve its productivity; cash registers, for example, have become very easy to use and have drastically improved efficiency. Finally, general software development can be viewed as improving the productivity of existing computers. Since all these developments are the result of intentional research activities, and since they have very different implications for the relative productivity of different factors, understanding how these research activities respond to a fall in the price of capital seems potentially quite important.

---

[2] With perfect complements the relative input use does not respond at all to a change in relative prices, and with perfect substitutes the relative input use may switch completely with a change in relative prices.

Acemoglu (1998, 2002b, 2003) describes a simple framework of endogenous technical change where R&D is purposefully directed toward the productivity improvement of different inputs. An important ingredient of this approach is that the returns to R&D that improve the productivity of an input are proportional to the total income of that input. This creates a "market size" effect of R&D: productivity-improving resources are allocated toward factor markets with large factor income. With endogenous technical change, it is quite possible that R&D resources are allocated to one factor at the expense of another factor if the market for the neglected factor is small. In the long run, the productivity of the neglected factor stagnates. Externalities in the R&D process—that is, productivity improvements to one factor that spill over to other factors—can overcome this effect such that in the long run productivity improvements proceed at the same rate for factors with small and big markets. We now describe how purposeful R&D affects the interaction of technical change and wage inequality.

We have already described how changes in the relative supply of capital together with different degrees of substitutability in production affect relative wages directly. More important, however, in an economy with directed R&D, relative supply changes also affect relative factor incomes, depending on the degree of substitutability. When factor productivities can change, the relevant factor supply is the product of factor endowment and factor productivity, that is, the number of available efficiency units. Now suppose that the effective supply of capital increases relative to the effective supply of skilled labor. Because capital and skilled labor are complements, in an equilibrium the wage of skilled labor relative to the price of capital has to increase by more than the supply of skilled labor relative to capital falls, and the total payments to skilled labor increase relative to payments to capital. Because of the market size effect, R&D is then redirected toward making skilled labor more productive relative to capital; that is, it increases the relative effective supply of skilled labor. This in turn lowers the relative income of skilled labor, and the R&D process is stable.

Now consider an increase of skilled labor productivity relative to unskilled labor productivity; that is, the effective relative supply of unskilled labor declines. Because unskilled labor is a substitute for skilled labor and capital, in an equilibrium the wage of unskilled labor relative to the wage of skilled labor has to increase by less than the relative supply of unskilled labor to skilled labor falls, and the total payments to unskilled labor decline relative to payments to skilled labor. Because of the market size effect, R&D is then redirected away from making unskilled labor more productive and the relative productivity of unskilled labor falls. This in turn again reduces the relative effective supply of unskilled labor, which in turn leads to even less R&D devoted to improve the productivity of unskilled labor, and so on. If this process is not stopped, the wage of unskilled labor will stagnate and over time will become negligible relative to the wage of skilled labor. This is a process that we have

not observed in the United States economic history.[3] While there have been changes in the skill premium, these changes have remained bounded.

In order to prevent unskilled wages from losing out relative to skilled wages, we assume that there are research spillovers between skilled and unskilled labor. This does not seem entirely unreasonable a priori, besides helping ensure that the long-run shares of skilled and unskilled labor remain balanced.[4] This mechanism is similar to Acemoglu (2002b), who studies the effects of directed R&D when the relative supply of skilled and unskilled labor is changing and skilled and unskilled labor are substitutes.[5]

Results

We find that capital-embodied technical change together with induced factor-specific technical change due to directed R&D significantly raises the skill premium, that is, increases wage inequality. We limit our analysis to the study of balanced growth paths where each variable grows at a constant rate. On these balanced growth paths factor income and expenditure shares are constant. We find that a one-time increase of productivity in the capital-goods-producing sector generates a small but very persistent increase of the skill premium. In the long run, however, wage inequality is not affected. As we have pointed out, capital-embodied technical change is not a one-time event, but a process that has been ongoing for a long time. We therefore consider a sequence of repeated productivity improvements in the capital-goods-producing sector, and this sequence generates a significant increase in the skill premium that persists for a very long time, even after there is no more capital-embodied technical change.

The remainder of our paper is outlined as follows. In Section 1 we describe the model—that is, the environment and the market structure—and then characterize balanced growth paths of the model economy. In Section 2 we parameterize the model to match the long-run growth characteristics of the U.S. economy. In Section 3 we study the short- and medium-term dynamics of the economy when there is capital-embodied technical change; in particular, we study how the skill premium and labor income share respond. Section 4 concludes.

---

[3] Goldin and Katz (1999) argue that in the United States the skill premium declined (increased) in the first half (second half) of the twentieth century.

[4] Research spillovers between labor of either sort and capital are not present in our model. Such spillovers would imply that there must be long-run technological change to augment the capital input, and this would make the capital-labor share unbalanced.

[5] A similar mechanism is used by Goodfriend and McDermott (1998) to explain the determinants of relative national per capita products in world balanced growth.

1. THE MODEL

Preferences and Technology

Preferences

The model has the simplest possible consumer preference structure: preferences are linear in consumption streams over time, with a constant rate of discount:

$$\sum_{t=0}^{\infty} \beta^t C_t, \tag{1}$$

where C_t is consumption at time t and β is the time discount factor. This preference specification implies that the goal of the consumer, or of any benevolent government planner, is simply to maximize present-value output using a constant interest rate that is equal to the consumer's rate of discount.

Production of Final Output: Capital-Skill Complementarity

A final output good Y (we omit time subscripts whenever there is no risk of confusion) is produced with three intermediate inputs, Y_k, Y_s, and Y_u, to an aggregate production function F. These intermediate inputs are in turn produced from the primary factors capital, skilled labor, and unskilled labor, respectively. We assume that F is of the nested constant-elasticity-of-substitution (CES) form, as in KORV (2000):

$$Y = F(Y_k, Y_s, Y_u) = \left\{ \lambda \left[\mu Y_k^{\rho} + (1-\mu) Y_s^{\rho} \right]^{\frac{\sigma}{\rho}} + (1-\lambda) Y_u^{\sigma} \right\}^{\frac{1}{\sigma}}, \tag{2}$$

with $\rho, \sigma \leq 1$. The elasticity of substitution between skilled labor and capital is $1/(1-\rho)$. This elasticity is less than one—that is, $\rho \leq 0$—since we assume that capital and skilled labor are complementary. On the other hand, the elasticity between unskilled labor and the aggregate of skilled labor and capital is $1/(1-\sigma)$. This elasticity is greater than one—that is, $\sigma \geq 0$—since we assume that capital and unskilled labor are substitutes.

Production of Intermediate Goods

The production of intermediate goods is central to our model: it is where the "directed technical change" appears. Following a large part of the recent literature on endogenous growth, we assume that productivity increases via an expansion in the variety of inputs with which each intermediate good is produced.[6] At any point in time, a type j intermediate good Y_j, $j = k, s, u$,

[6] See, for example, Romer (1990).

is produced with a continuum of specialized inputs, n_j:

$$Y_j = \left[\int_0^{n_j} Y_j(i)^{\frac{\nu-1}{\nu}} di\right]^{\frac{\nu}{\nu-1}}, \nu \geq 1. \tag{3}$$

Each specialized input $Y_j(i)$, $i \in [0, n_j]$, is produced from a primary factor with a distinct technology, which we discuss shortly. In a symmetric equilibrium, all specialized inputs to production of the same intermediate good are operated at the same level, $Y_j(i) = y_j$ for all i. This implies the following reduced form production function for an intermediate good:

$$Y_j = n_j^{\frac{\nu}{\nu-1}} y_j. \tag{4}$$

Production of Specialized Inputs

Finally, the production of specialized inputs is closely tied to the three primary factors. A unit of capital produces one unit of any type of specialized input used in the production of capital-based intermediate goods:

$$Y_k(i) = K(i), \tag{5}$$

where $K(i)$ is the amount of capital used for specialized input i. Analogously, we have for skilled and unskilled labor

$$Y_s(i) = S(i) \text{ and } Y_u(i) = U(i), \tag{6}$$

where $S(i)$ and $U(i)$ are the amounts of skilled and unskilled labor, respectively, used for specialized input i.

The resource availability for each primary factor is as follows. At each point in time t there is a fixed amount of capital K_t, and over time, K_t can be increased by foregoing consumption. The other primary factors, skilled and unskilled labor, S and U, are fixed. We can think of these as the amount of raw labor hours available in the two groups. We thus abstract from variations in the amount of hours supplied by each worker, in labor force participation, and in population growth. Perhaps more important, we abstract from education decisions; that is a topic worthy of further study.

In an equilibrium the demand and supply for primary factors is equalized, and in a symmetric equilibrium the total demand for a primary factor is equal to the product of the number and level of the specialized inputs using the factor

$$n_k y_k = K, n_s y_s = S, \text{ and } n_u y_u = U. \tag{7}$$

Given the reduced form production function for a symmetric equilibrium, we can relate intermediate goods production to the primary factors as follows:

$$Y_k = A_k K, Y_s = A_s S, \text{ and } Y_u = A_u U, \tag{8}$$

where $A_j \equiv n_j^{1/(\nu-1)}$. The variables A_j will play the role of the productivity specific to factor $j = k$, s, and u. Notice that the development of more specialized inputs, n_j, increases productivity, A_j, since $\nu > 1$. This development occurs through R&D and will be discussed below.

Investment-Specific Technical Change

The aggregate resource constraint in the economy is

$$C_t + I_t/q_t = Y_t, \tag{9}$$

where the price of new capital goods—that is, investment I_t—in terms of consumption is $1/q_t$. Investment increases the capital stock

$$K_{t+1} = (1 - \delta)K_t + I_t \tag{10}$$

after depreciation, $0 \leq \delta \leq 1$. An increase in q_t is a form of technical progress, because it makes investment cheaper, and we call this form of technical progress "capital-embodied" or "investment-specific." We will consider a gradual increase in q_t from an initial stable level to a new plateau, thus corresponding to a gradual fall in the price of new capital goods.

R&D

Finally, the development of new technologies occurs in a similar way for the three kinds of intermediate goods: there is a fixed amount of basic R&D input, R, that can be divided into producing new varieties of specialized inputs of type $j = k$, s, and u. One unit of research input produces $b_j \bar{n}_j$ new specialized j inputs, where $\bar{n}_j$ is a weighted average of existing research stocks (varieties). That is, researchers stand on the shoulders of past giants: with a larger available stock of past research in the form of many existing varieties, research productivity is higher. Besides R&D externalities from previously developed varieties to new varieties used in the production of the same intermediate input, there are also spillovers from R&D activities for one intermediate input type to other intermediate input types. In particular, we assume

$$\bar{n}_s = n_s^{\frac{1+\phi}{2}} n_u^{\frac{1-\phi}{2}}, \tag{11}$$

$$\bar{n}_u = n_u^{\frac{1+\phi}{2}} n_s^{\frac{1-\phi}{2}}, \tag{12}$$

$$\bar{n}_k = n_k. \tag{13}$$

We assume that R&D spillovers are limited to skilled and unskilled labor research. These research spillovers between skilled and unskilled labor are symmetric and captured by the parameter $\phi \in [-1, 1]$. Capital research does not lead to, nor does it receive, any spillovers.

Finally, the number of available specialized varieties depreciates at rates d_k, d_s, and d_u, respectively. Although we can interpret this assumption as exogenous obsolescence of ideas, it is essentially a technical requirement that is necessary to guarantee the local stability of balanced growth paths.

Summarizing the R&D sector, we have

$$n_{k,t+1} = (1 - d_k)n_{k,t} + b_k \bar{n}_{k,t} R_{k,t}, \tag{14}$$

$$n_{s,t+1} = (1 - d_s)n_{s,t} + b_s \bar{n}_{s,t} R_{s,t}, \tag{15}$$

$$n_{u,t+1} = (1 - d_u)n_{u,t} + b_u \bar{n}_{u,t} R_{u,t}, \tag{16}$$

where $R_{j,t}$ is the amount of R&D input devoted to type j product development. The market for R&D inputs clears

$$R_{k,t} + R_{s,t} + R_{u,t} = R. \tag{17}$$

Markets and Decentralized Equilibrium

The market structure we consider is quite standard for this kind of model setup. There is perfect competition in the final goods market. Intermediate goods are bought and sold by perfectly competitive firms, too, but their inputs—the specialized inputs—are provided by monopolistically competitive firms. Each such monopolistic firm thus owns a right (infinitely-lived patent) to produce its good that it once bought from an R&D firm, and it controls the quantity supplied in every period—with knowledge of the demand curve—in order to maximize profits. With free entry into the monopolistic industry, the stream of profits is enough to just cover the cost of the patent. Researchers, or R&D labs, are perfect competitors, as are the providers of the primary factors capital and labor. The output of research that has market value is the patent; the effect on research productivity of future research efforts is an externality.

We will now look at profit maximization conditions for the different kinds of firms, starting with the final output sector.

Final Output

We normalized the price of the final output at one. The profit of a competitive final goods producer is

$$F(Y_k, Y_s, Y_u) - P_k K - P_s S - P_u U. \tag{18}$$

A profit-maximizing final goods producer equates the marginal cost of a type j intermediate input—that is, its price, P_j—to the marginal value product of that input:

$$P_j = F_j(Y_k, Y_s, Y_u) = F_j\left(\frac{Y_k}{Y_s}, 1, \frac{Y_u}{Y_s}\right). \tag{19}$$

For the last equality we have used the fact that if F has constant returns to scale, then its derivatives are homogeneous of degree zero.

Intermediate Goods

A competitive intermediate goods producer takes prices and technology, in particular the number of available specialized inputs, as given. The profit of a producer of type $j = k, s, u$ intermediate goods is

$$P_j Y_j - \int_0^{n_j} p_j(i) Y_j(i) di, \tag{20}$$

where $p_j(i)$ is the price for specialized input i for intermediate good j. Again, a profit-maximizing choice equates the marginal value product of a specialized input with the marginal cost of the input, and we get

$$P_j \left[\frac{Y_j}{Y_j(i)} \right]^{\frac{1}{\nu}} = p_j(i). \tag{21}$$

Conditional on the price of the intermediate good and the level of intermediate goods production, this equation defines the quantity demanded $Y_j(i)$ for specialized input i as a function of its price, $p_j(i)$. This demand function has a constant price elasticity, ν.

In a symmetric equilibrium, $Y_j(i) \equiv y_j$ (it does not depend on i), $p_j(i) = p_j$, and $Y_j = n_j^{\nu/(\nu-1)} y_j = A_j^{\nu} y_j$, so that we have

$$p_j = P_j \left(\frac{y_j A_j^{\nu}}{y_j} \right)^{\frac{1}{\nu}} = A_j P_j. \tag{22}$$

Specialized Inputs

Specialized inputs are produced by monopolistically competitive firms; that is, they take into account the effect of the price they set on their sales while taking the level of demand and the price of the intermediate input for which they supply specialized inputs as given. For example, a firm that produces a specialized input for the capital-type intermediate good hires capital services $K(i)$ at the rental rate w_k and maximizes profits:

$$\max p_k(i) Y_k(i) - w_k K(i), \tag{23}$$

subject to the inverse demand function for $p_k(i)$ given in (21). Because of the demand function's constant price elasticity, the profit-maximizing choice sets the specialized input price as a constant markup $\mu \equiv \nu/(\nu - 1)$ over marginal cost:

$$p_k(i) = \mu w_k. \tag{24}$$

Period profits are then

$$\pi_k(i) = (\mu - 1) w_k K(i). \tag{25}$$

Similarly, we obtain for firms using skilled and unskilled labor

$$p_s(i) = \mu w_s \text{ and } \pi_s(i) = (\mu - 1) w_s S(i), \tag{26}$$

$$p_u(i) = \mu w_u \text{ and } \pi_u(i) = (\mu - 1) w_u U(i). \tag{27}$$

The capital value at time t of a specialized firm using factor j is

$$V_{j,t} = \pi_{j,t} + \beta(1 - d_j) V_{j,t+1}, \tag{28}$$

where we have used the fact that the firm dies randomly between one period and the next with probability d_j and firms discount future returns with the representative household's discount factor, β.

The Research Sector

Let w_R denote the price of the services provided by one unit of research. Each unit of research produces $b_j \bar{n}_j$ new varieties that use the primary factor j in the next period. Alternatively, in order to obtain one specialized input i, one needs to hire $1/b_j \bar{n}_j$ units of research services. Free entry in the research sector amounts to the requirement that the value of the patent to operate production of specialized input i using the basic input j from the next period on has to equal the cost of obtaining that patent. Thus,

$$w_{R,t} / b_j \bar{n}_{j,t} = \beta V_{j,t+1} \tag{29}$$

is the zero-profit condition for the research sector.

Consumer Savings

The intertemporal first-order condition for the consumer equates the marginal cost of a unit of investment good to the discounted value of its marginal value next period:

$$\frac{1}{q_t} = \beta \left[w_{k,t+1} + (1 - \delta) \frac{1}{q_{t+1}} \right]. \tag{30}$$

Balanced Growth

In this economy, there will be long-run productivity and output growth provided that the research activity is potent enough. We will assume that this is the case. What is of more interest, however, is the form that this growth will take. We will focus attention on balanced growth paths (BGPs)—that is, paths where each variable of interest grows at a constant percentage rate—and all factors are used in production and account for positive and constant shares of total income. This economy also allows for asymptotic growth paths where some factors become unimportant in the long run and their income shares become arbitrarily small. We do not study these asymptotic growth paths but restrict attention to locally stable balanced growth paths, and we assume that

initial conditions are such that the economy is in a locally stable neighborhood of the balanced growth path.

For the analysis of the BGP, we also assume that there is no capital-embodied technical change; that is, the relative price of capital is constant. When the relative price of capital is not constant but declines at a constant rate, a BGP exists only if the elasticity of substitution between all primary factors in the production function (2) is unitary (Greenwood, Hercowitz, and Krusell 1997). Equal and unitary elasticities of substitution are, however, inconsistent with the observed differences in factor-substitution elasticities.

A BGP with Labor-Augmenting Technical Change

We first establish that the BGP of our economy with endogenous directed technical change has the same properties as the BGP of the neoclassical growth model with exogenous labor-augmenting technical change.[7] Namely, output, capital, and the productivity of skilled and unskilled labor all grow at the same rate, and the productivity of capital is constant.

Prices for specialized inputs, p_k, p_s, and p_u (where we have removed the index i because of symmetry), are constant and equal markups on the prices of the associated primary factors K, S, and U (equations (24), (26), and (27)). From equations (19) and (22), it then follows that the relative incomes of the three factors satisfy

$$\frac{w_k K}{w_s S} = \frac{p_k K}{p_s S} = \frac{P_k A_k K}{P_s A_s S} = \frac{F_k A_k K}{F_s A_s S}, \tag{31}$$

$$\frac{w_u U}{w_s S} = \frac{F_u A_u U}{F_s A_s S}. \tag{32}$$

The marginal products of intermediate inputs depend only on the intermediate input ratios, $Y_k/Y_s = (A_k K)/(A_s S)$ and $Y_u/Y_s = (A_u U)/(A_s S)$, because the production function F is constant returns to scale. This in turn implies that the factor income ratios depend only on the intermediate input ratios. Since by assumption the two-factor income ratios are nontrivial constants on a BGP, the intermediate input ratios are then constant. Thus, on any BGP, (a) A_s and A_u grow at the same rate g, since U and S are constant; and (b) $A_k K$ grows at the same rate as A_s. This implies that intermediate inputs Y_k, Y_s, Y_u, and output Y all grow at rate g. Furthermore, because we assume that q is constant on a BGP, capital K has to grow at the rate of final output; otherwise, the investment share goes to zero or one. Because output grows at the same rate as A_s, so does K. Hence, A_k must be constant.

[7] For a similar environment, this was established by Acemoglu (2003).

The Equations that Characterize a BGP

The BGP is characterized by the constant (a) productivity growth rate g; (b) relative productivity of skilled and unskilled labor $\tilde{A}_u \equiv A_u/A_s$; (c) ratio of capital-skilled labor productivity $\tilde{K} \equiv K/A_s$; and (d) capital productivity A_k. We first turn to the R&D sector to derive two equations that determine the growth rate and relative productivity of unskilled and skilled labor. Capital productivity and the normalized level of capital then adjust to satisfy the optimal capital accumulation conditions.

Constant capital productivity A_k together with (14) imply that

$$d_k = b_k R_k. \tag{33}$$

Since productivity growth rates are constant on the BGP and the total amount of resources, R, available for R&D purposes is constant, the R&D resources directed to the different uses are also constant. The restriction on the total amount of R&D input resources then delivers one equation in the unknowns R_s and R_u: $R_s + R_u = R - d_k/b_k$. From equations (15) and (16) equal growth in A_s and A_u now implies that

$$1 - d_s + b_s \tilde{A}_u^{\frac{(1-\phi)(\nu-1)}{2}} R_s = 1 - d_u + b_u \tilde{A}_u^{\frac{-(1-\phi)(\nu-1)}{2}} (R - d_k/b_k - R_s). \tag{34}$$

This equation determines R&D resources devoted to the improvement of skilled labor productivity R_s as a function of the relative productivity of unskilled labor $\tilde{A}_u$. Together with the R&D equation for skilled labor, this determines aggregate growth:

$$g = -d_s + b_s \tilde{A}_u^{\frac{(1-\phi)(\nu-1)}{2}} R_s. \tag{35}$$

The economic incentives that determine the direction of technical change are described by the free-entry conditions for R&D (equation (29)). These conditions imply that the marginal payoffs from R&D in each of the three basic uses are equalized to the marginal cost of R&D:

$$w_R = b_j \bar{n}_j V_j \text{ for } j = s, u, k. \tag{36}$$

The capital value of a firm that produces a specialized input is equal to the expected present value of current and future profits from production. For example, from equation (26) a firm that produces specialized inputs from skilled labor has profits $(\mu - 1)\, w_s S/n_s = w_s S A_s^{1-\nu}$, and profits decline at the gross rate $(1+g)^{2-\nu}$ since more and more firms have to share the available stock of skilled labor. On a BGP the capital value of such a firm is

$$\begin{aligned} V_{s,t} &= \pi_{s,t} + \beta(1-d_s)\pi_{s,t+1} + \dots \\ &= (w_s S/n_s)\left[1 + \beta(1-d_s)(1+g)^{2-\nu} + \dots\right] \\ &= \frac{w_s S/n_s}{1 - \beta(1-d_s)(1+g)^{2-\nu}}. \end{aligned} \tag{37}$$

Notice that the capital value and therefore the return to R&D that improves the productivity of skilled labor is proportional to the total factor income of skilled labor. Similar expressions can be derived for the capital values of firms that use unskilled labor or capital.

Equalization of returns to R&D from productivity improvements for skilled and unskilled labor then implies the condition

$$\frac{b_s}{b_u} = \frac{w_u U}{w_s S} \tilde{A}_u^{(\nu-1)(\phi-1)} \frac{1-\beta(1-d_s)(1+g)^{2-\nu}}{1-\beta(1-d_u)(1+g)^{2-\nu}}, \tag{38}$$

which involves the growth rate g, relative productivity $\tilde{A}_u$, and the normalized capital stock $A_k\tilde{K}$ through the relative wages. Equalization of returns to R&D from productivity improvements for skilled labor and capital and manipulations similar to the ones above yield the condition

$$\frac{b_s}{b_k} = \frac{w_k K}{w_s S} \tilde{A}_u^{(\nu-1)(\phi-1)/2} \frac{1-\beta(1-d_s)(1+g)^{2-\nu}}{1-\beta(1-d_k)(1+g)}. \tag{39}$$

Note that the relative incentives to do R&D depend on the relative factor income shares.

Equations (34)–(39) involve four equations in four unknowns: R_s, g, $\tilde{A}_u$, and $A_k\tilde{K}$. We will briefly discuss the solution to this system below. Having solved for these four variables, we find the remaining endogenous variables by using the BGP version of our equations. First, we determine the constant productivity of capital A_k. Given the exogenous price of new capital q, we get a constant value for the rental rate of capital w_k from the optimal capital accumulation condition (30). Given markup pricing (24), the rental rate is equal to $p_k/\mu = P_k A_k/\mu = F_k A_k/\mu$, and since the marginal product of capital F_k depends on known factor input ratios, this delivers A_k.

To find levels of variables at a point in time, we need to initialize our state variables at time 0. The state variables of the system are K, A_k, A_s, and A_u, of which we already know A_k. Thus, let $K(0) = 1$. Then $A_s(0)$ is implied by $A_k\tilde{K} = A_k K/A_s$. Finally, $A_u(0)$ follows from knowing $\tilde{A}_u$. Given the growth rates of all variables, we can now solve for the levels of quantities and prices at all points in time. Perhaps the last variable to solve for is the factor rental of the research input, w_R; it equals a present value of profits, where each profit flow is a fixed fraction of labor costs per product.

Characteristics of Growth Paths

In our economy, capital-embodied technical change—that is, technical progress in the investment goods sector—temporarily increases the growth rate and the skill premium, but it does not affect growth or the skill premium in the long run. The temporary effects of a once-and-for-all productivity increase in the investment goods sector are, however, extremely persistent. In our economy, deviations from the BGP path are persistent because induced technical

progress can be self-fulfilling, which makes the economy potentially unstable and introduces the possibility of multiple BGPs. Counteracting this destabilizing force is a spillover between R&D activities devoted to productivity improvements of unskilled and skilled labor. In the next section we will show that for a calibrated version of the model economy, the research spillovers just overcome the self-fulfilling aspect of the growth process and the economy is just barely stable, which implies the high persistence of deviations from the BGP.

The Role of Investment Technology for Growth and Wage Inequality

The variable q represents the relative productivity of the investment goods sector. One unit of final output can be transformed into one unit of consumption or q units of new machines. Equivalently, $1/q$ is the relative price of new capital in terms of consumption goods. As we have just argued above, this technological parameter has no impact on long-run growth in this economy. Essentially, investment technology pins down the level of the marginal product of capital in production, but that is a level effect in this growing economy: it determines A_k, the productivity of installed capital that is constant over time. Growth is determined by R&D decisions, which respond to profits from innovation. Since profits are collected as a (constant) markup over costs, and costs are the expenditures on the primary factors, R&D decisions respond to factor income. The relative allocation of R&D resources toward factor-productivity improvements then depends on relative income shares. Finally, given the homogeneity of the production function, relative income shares depend on the relative input ratios (Y_k/Y_s and Y_u/Y_s), but not on the productivity of capital per se.

This result also applies to an economy where consumers desire to smooth consumption, that is, where utility is not linear. The optimal capital-accumulation condition (30) then includes the long-run growth rate g, but this variable has already been determined in the R&D sector. The optimal capital-accumulation condition is still limited to the determination of A_k.

The skill premium of this economy is

$$\frac{w_s}{w_u} = \frac{A_u}{A_s}\frac{F_s}{F_u}, \tag{40}$$

and it depends only on the relative input ratios. Since the relative input ratios are entirely determined in the R&D sector, the investment technology parameter q does not have a long-run impact on wage inequality either. Again, a permanent increase in q increases A_k, the productivity of installed capital, permanently; however, this variable does not influence F_s/F_u in the long run. In other words, the variables A_u/A_s and K/A_s will adjust over time until F_s/F_u returns to its initial value. Over the course of this adjustment, of course, there

are temporary effects on the skill premium, and the subject of the work below is to study these temporary effects.[8]

Can Technology Growth Be Self-Fulfilling?

In our economy R&D decisions depend on scale: if the productivity of a primary factor is large—that is, if there are many specialized inputs using this factor—then this factor gets paid a high rental rate and receives a high income, which in turn increases the incentive to do more R&D for this factor. This argument, however, applies to all factors, and given the finite resources that can be used for R&D, what matters is the relative allocation of these resources among competing uses. Thus the behavior of relative factor incomes determines the relative allocation of R&D resources. As was pointed out by Acemoglu (2002b), the impact on relative factor incomes is connected to the substitutability features of the intermediate goods in final output production.

Consider the case of capital and skilled labor first. Suppose the productivity of skilled labor increases, that is, the relative supply of skilled-labor-based intermediate inputs increases. Since capital and labor are gross complements, the relative income of skilled labor falls, and resources are redirected toward capital accumulation. This in turn increases the relative supply of capital-based intermediate inputs, and the process is stable.

Alternatively, consider the case of skilled and unskilled labor, which are substitutes. Now an increase of the relative supply of skilled-labor-based intermediate inputs increases the income of skilled labor relative to unskilled labor, which leads to even more R&D resources devoted to the creation of skilled-labor-using specialized inputs, which in turn increases the relative supply of skilled-labor-based intermediate inputs. This productivity growth process feeds on itself and the relative productivity of skilled labor increases more and more, such that in the end the economy is effectively specialized in skilled-labor-based intermediate inputs. In order for the economy to remain stable, we need another mechanism that counteracts the scale effects: technology spillovers between the two kinds of labor. With spillovers, productivity improvements for skilled labor lower the R&D cost for unskilled labor, and if these spillovers are strong enough, they can stabilize the R&D process and prevent a complete specialization. The strength of spillovers is reflected in the parameter ϕ: with $\phi = 1$, there are no spillovers and the strength of spillovers increases as ϕ declines.

---

[8] Notice that the basic supplies of skilled and unskilled workers, S and U, directly influence the long-run skill premium, even though they do not at all influence the relative total wage bills of the two groups. An interesting issue is how the endogenous accumulation of skills (e.g., education or on-the-job learning), which makes the relative supply of skilled labor endogenous, would interact with technological change to determine long-run wage inequality. We have argued before that there are limits to the extent that the relative skill endowment can be affected, and therefore we do not pursue this issue.

The possibility of self-fulfilling productivity growth paths in our economy suggests that there might be multiple BGPs. To simplify the study of multiple BGPs, assume that the number of specialized inputs depreciates at the same rate in all sectors, $d \equiv d_u = d_s = d_k$. We can then solve equations (34) and (35) easily for the growth rate:

$$g = b_s \frac{\tilde{A}_u^{\frac{(1-\phi)(\nu-1)}{2}}}{1 + \frac{b_s}{b_u}\tilde{A}_u^{(1-\phi)(\nu-1)}}(R - d_k/b_k) - d. \tag{41}$$

Note that with spillovers the growth rate is a non-monotone function of the relative productivity of unskilled labor $\tilde{A}_u$. Without spillovers ($\phi = 1$), the growth rate is a constant, independent of the relative productivity. Now use the nested CES aggregate production function (2) to derive explicit expressions for the factor income ratios:

$$\frac{w_u U}{w_s S} = \frac{\omega}{(1-\omega)(1-\lambda)}\left(\tilde{A}_u \frac{U}{S}\right)^{\sigma}\left(1 - \lambda + \lambda\left(\frac{A_k \tilde{K}}{S}\right)^{\rho}\right)^{\frac{\rho-\sigma}{\rho}}, \tag{42}$$

$$\frac{w_k K}{w_s S} = \frac{\lambda}{1-\lambda}\left(\frac{A_k \tilde{K}}{S}\right)^{\rho}. \tag{43}$$

Inserting these two expressions in equations (38) and (39), we obtain

$$1 = \#_1 \tilde{A}_u^{\sigma+(1-\phi)(\nu-1)}\left(1 - \lambda + \lambda\left(\frac{A_k \tilde{K}}{S}\right)^{\rho}\right)^{\frac{\rho-\sigma}{\rho}}, \tag{44}$$

$$1 = \#_2 \left(\frac{A_k \tilde{K}}{S}\right)^{\rho} \tilde{A}_u^{\frac{(1-\phi)(\nu-1)}{2}} \frac{1 - \beta(1-d)(1+g)^{2-\nu}}{1 - \beta(1-d)(1+g)}, \tag{45}$$

where $\#_1$ and $\#_2$ are constants and g depends on $\tilde{A}_u$. We now have two equations in two unknowns, $A_k \tilde{K}/S$ and $\tilde{A}_u$. They define two curves relating the two unknowns, and the balanced growth path is found as an intersection of the two curves. Is there a solution to this system, and if so, is there more than one? We will not go further here than to point out that both equations define upward-sloping curves so long as $\rho < 0 < \sigma$, which are the assumptions we use because of the data on cross elasticities between different inputs.[9] And with two upward-sloping curves, multiple solutions are not only possible but, as we have verified numerically, hard to avoid in this framework. This is in contrast to the setups in Acemoglu (2002b, 2003), which deliver unique steady states. Because of our three-factor setup here, multiplicity is hard to avoid.

[9] To simplify the exposition, we treat the growth rate in equation (45) as a constant; that is, we ignore the feedback from equation (41). The dependence of g on $\tilde{A}_u$ may cause non-monotonicities, but that is only a local property; globally, the equation defines an upward-sloping relation.

The interpretation is the one hinted at in several places above. On the right-hand side of equation (44) is the relative return on R&D with respect to productivity improvements of unskilled to skilled labor, and on the right-hand side of equation (45) is the relative R&D return with respect to capital and skilled labor. The two unknowns are the relative productivities of unskilled labor and capital (relative to that of skilled labor; in the case of capital, we measure the stock times the productivity). In equation (44), a higher productivity of unskilled labor raises the relative return on unskilled labor, because skilled and unskilled labor are substitutes ($\sigma > 0$) and because of the market size effect ($(1-\phi)(\nu-1) > 0$) if there are spillovers. To balance the increased relative return of unskilled labor, the productivity of capital has to increase. Because of capital-skill complementarity, $\sigma > 0 > \rho$, the higher capital productivity increases the return to skilled labor. In equation (45), an increase in the productivity of unskilled labor gives a reinforcing scale effect, because it can be viewed as a relative decrease in the productivity of capital, which is balanced in this case by an increase in the direct productivity of capital, since skilled labor and capital are complements ($\rho < 0$).

When there are multiple balanced growth paths, it is important to check "local stability" of each of these: do small deviations of the state variables from the balanced growth path lead back to the balanced path or do they lead away from it? In our numerical examples, we found one stable and one unstable path, the last of which is economically irrelevant (since no initial conditions would lead there). We also found cases where there is only one, unstable balanced growth path. In this case, the scale effects are simply too strong to admit convergence to a balanced outcome: any deviations from the balanced path would lead away from it. We tend to find at least one stable equilibrium when the spillovers are strong, i.e., when ϕ is low, and when knowledge depreciation is high, i.e., when d is close to one.

2. CALIBRATION

Our intention is to provide a quantitative statement on how a decline of the relative price of capital affects wage inequality. Furthermore, our model is sufficiently complicated such that we cannot analytically characterize the stability properties of its balanced growth path. We therefore solve the model numerically, and in order to do this we have to decide what are empirically relevant values of the model's parameters. In the following we parameterize the economy such that its balanced growth path is consistent with observations on the U.S. economy in the latter part of the twentieth century.

We assume that a time period represents one year, and we choose the time discount factor β such that the annual interest rate is 4 percent. The annual depreciation rate for equipment capital in the United States is $\delta = 0.125$. KORV (2000) estimate the elasticity parameters for the two-stage CES

production function (2) as $\rho = -0.5$ and $\sigma = 0.4$. We set the specialized input parameter $\nu = 11$ such that the equilibrium markup is 10 percent above marginal cost, $\mu = 1.1$. This choice is at the upper bound for estimates of profit rates in the U.S. economy. Acemoglu (2002a) provides various estimates of the factor income ratios of skilled to unskilled labor. We set the ratio $w_s S/w_u U = 0.5$, which corresponds to Acemoglu's estimate of this ratio in the 1990s for a broad definition of skilled labor. We set the capital income share in final output to one-third, which roughly corresponds to the capital income share in the United States.

Estimates by the National Science Foundation (NSF) suggest that in the United States R&D expenditures are less than 3 percent of GDP. The NSF estimates include public and private expenditures on R&D. In the following we interpret the R&D input as a type of labor and include the value of R&D inputs in the model economy's measure of GDP. Conditional on the factor income shares and assuming equal depreciation rates of knowledge, $d = d_j$, the R&D share in GDP determines the depreciation rate d. The R&D share in GDP is increasing in d, and with $d = 0$ the R&D share is 4.9 percent conditional on the other income shares. The BGP equilibrium is not stable for $d = 0$, but we obtain a stable BGP for $d = 0.01$, which implies a BGP R&D share of 5.9 percent. In the following we interpret the R&D input as another type of skilled labor.

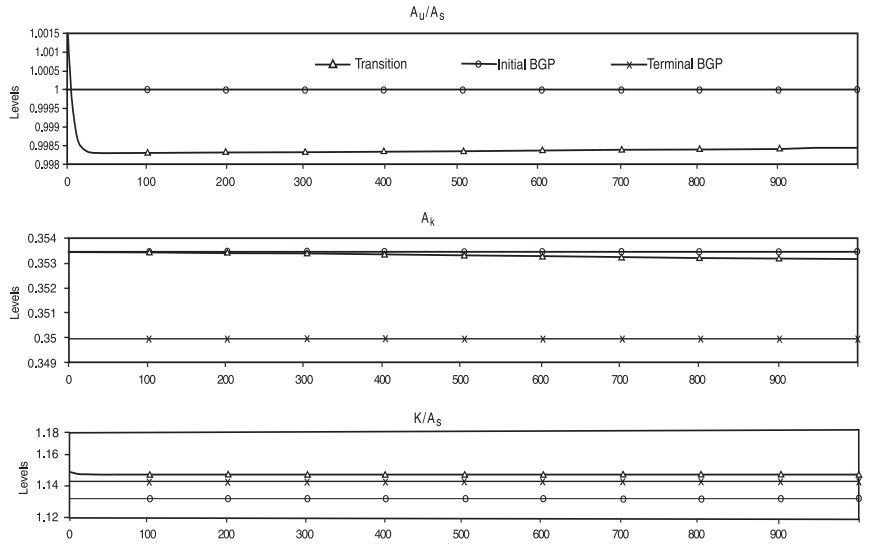
No quantitative evidence is available on the R&D externality. We set the R&D externality parameter for skilled and unskilled labor to $\phi = 0.5$. Larger externalities, smaller ϕ, have no appreciable impact on the medium-term to long-term dynamics. Smaller externalities, larger ϕ, make the effects of shocks more persistent, but for ϕ approaching 0.9 we can no longer find a stable BGP.

Direct observations on $\tilde{A}_u$, Y_u/Y_s, or Y_k/Y_s, are also not available. For the calibration exercise the values of these variables and of S/U, ω, and λ are not determined. This is not a problem since, conditional on the calibration so far, the local dynamics around the BGP are independent of the choice for these variables. In the following we normalize $\tilde{A}_u = Y_u/Y_s = Y_k/Y_s = 1$.

3. RESULTS

In the previous discussion of the BGP we argue that a permanent change of the relative price of capital does not affect the economy's long-run growth rate, factor income shares, or skill premium. We now want to argue that even though the effects of a permanent change in the relative price of capital are transitory, they are nevertheless very persistent. For this purpose we perform two experiments. First, we show that in response to a one-time permanent decline of the relative price of capital, the relative wage of skilled labor and the wage income share increases and these effects are extremely persistent. In

Figure 1 The Medium- to Long-Run Response of Endogenous State Variables to a Permanent Decline of the Price of Capital



the introduction we point out that capital-embodied technical change is not a one-time event, but an ongoing process. In a second experiment we therefore model ongoing embodied technical change through successive reductions of the relative price of capital and show that the skill premium and labor income share increase significantly over time and stay above their long-run values for a very long time.

We study a local approximation of the dynamic response of our economy to an exogenous shock. Since our economy is growing over time, we first have to transform the dynamic system such that all variables are stationary. This is possible since we study a BGP where all variables grow at constant rates. The state variables of the transformed system are $\left(A_{k,t}, \tilde{K}_t, \tilde{A}_{u,t}\right)$.

A preliminary observation is worth making before going into the details of the experiment. If one computes the relative wages of skilled and unskilled workers in this economy *treating productivity and investment levels as exogenous,* it is apparent that an increase in q, which automatically increases the capital stock, must increase the relative wage of skilled labor because $\sigma > \rho$. This can easily be seen by taking the ratio of F_s to F_u and using $\sigma > \rho$: this expression is increasing in K. This essentially is the argument in KORV (2000) about why the skill premium has been increasing. Our main question below is, how does capital accumulation and endogenous directed technical change respond to the fall in the price of new capital goods?

Experiment 1: A Permanent 1 Percent Decline of the Relative Price of Capital

Figure 1 shows the response of the state variables to a 1 percent permanent decline of the relative price of capital. On impact, as investment in capital becomes more attractive and the economy starts to accumulate more capital and more resources are devoted to the improvement of capital productivity, both $\tilde{K} \equiv K/A_s$ and A_k increase. After the initial impact, the economy devotes more resources to the improvement of skilled labor productivity since skilled labor and capital are complementary in production, and consequently the relative productivity of unskilled labor $\tilde{A}_u \equiv A_u/A_s$ declines. We have argued above that the BGP value of $\tilde{A}_u$ is independent of the relative price of capital and the relative productivity of unskilled labor returns to its long-run value over time. On the other hand, the BGP values of the normalized capital stock and the productivity of capital depend on the price of capital. In particular, the productivity of capital declines and the capital stock increases with the decline of the relative price of capital. From Figure 1 it is apparent that the shock has a very persistent impact on the state of the economy. Recall that one period represents a year. Even after 1,000 years the economy still has a long way to go to arrive at its new BGP.

The economy's GDP growth, the labor income share, the skill premium, and the relative wage of R&D labor all increase following a decline of the relative price of capital (see Figure 2).[10] As discussed above, the BGP growth rate, labor income share, and relative wages are independent of the relative price of capital. Whereas the impact on the growth rate dissipates very fast, the effect on relative wages and the labor income share is very persistent.[11] The quantitative effect of a one-time 1 percent reduction of the price of capital is small; for example, the skill premium increases by less than 1 percent.

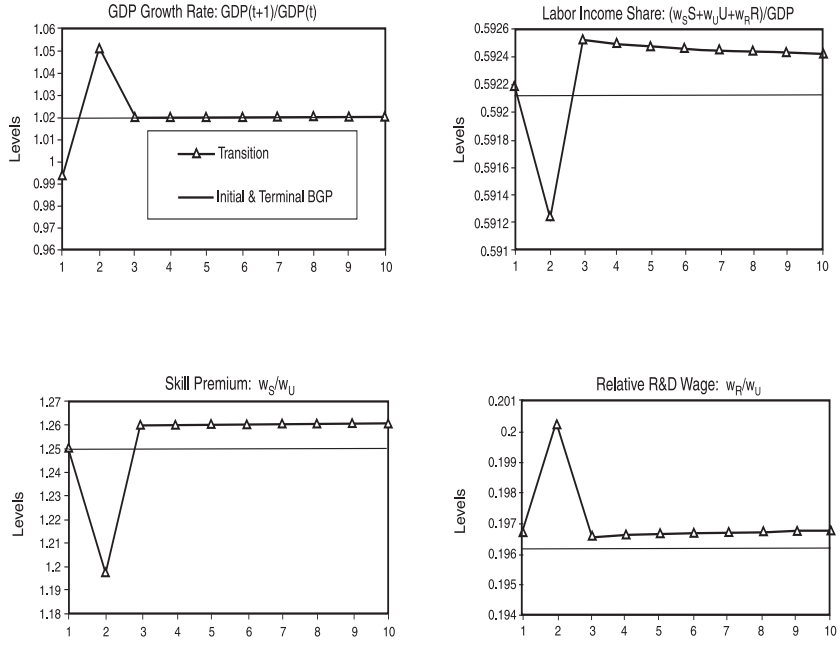
Experiment 2: A Sequence of Relative Price of Capital Reductions

In our economy a BGP does not exist if the relative price of capital declines at a constant rate. In order to model the effects of the observed secular decline of the relative price of capital, we therefore assume that this price declines at a constant rate for 100 years and then remains constant forever. We base our

[10] The substantial volatility for the GDP growth rate can be attributed to the fact that preferences are linear in consumption. With concave utility in consumption, there would be an incentive to smooth consumption and we would not see the wild swings in the GDP growth rate.

[11] It may appear odd that the relative wage of R&D labor is less than the wage of unskilled labor, but remember that we have said nothing about the units of R&D labor embodied in an R&D worker. Thus, the scale of the relative wage is arbitrary. The same can be said about the relative wage of skilled and unskilled workers.

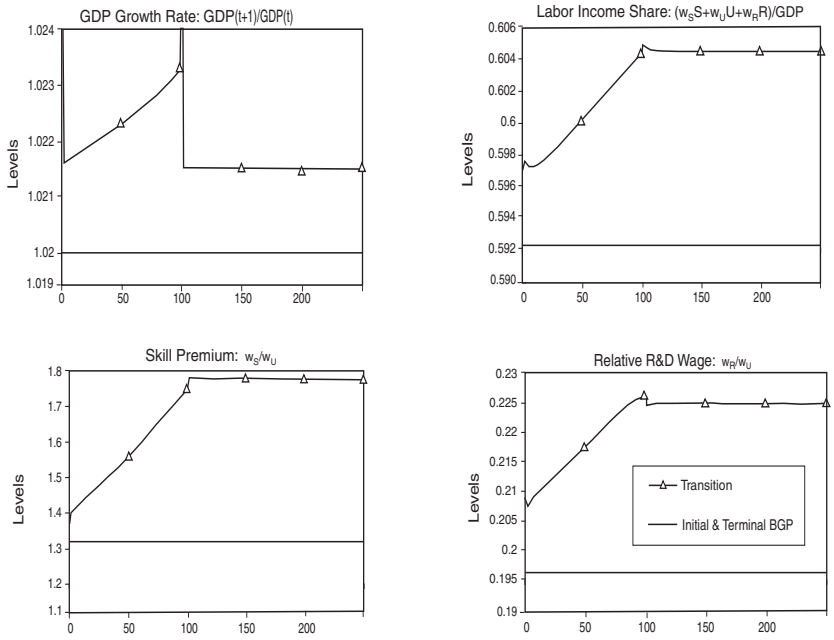
Figure 2 The Response of GDP Growth, Labor Income Share, and Relative Wages to a Permanent Decline of the Price of Capital



study of the medium- to long-run effects of the price decline on a local approximation of the economy's dynamics. We therefore want to avoid deviating too much from the BGP and limit ourselves to a 0.5 percent annual rate of price decline. This is substantially less than the 3 percent annual rate of decline for the relative price of equipment capital observed for the United States (Greenwood, Hercowitz, and Krusell 1997). Our example should therefore only be interpreted as a quantitative illustration of the effect of an ongoing decline of the relative price of capital.

Figure 3 shows that as long as the relative price of capital declines, the economy's growth rate, labor income share, and relative wage of skilled labor and R&D labor all increase. The impact of the capital price decline on relative wages is quantitatively important. Over the 100-year decline of the relative price of capital, the skill premium—that is, the price of skilled labor relative to unskilled labor—increases by about 40 percent, and the relative price of labor employed in the R&D sector increases by 15 percent. The change of the GDP growth rate and the labor income share would not be noticeable in the data. Changes of the magnitude implied by the model, one-tenth of a percentage point for the growth rate and half a percentage point for the labor

Figure 3 The Response of GDP Growth, Labor Income Share, and Relative Wages to an Ongoing Decline of the Price of Capital



income share, are dominated by other business-cycle-related fluctuations of these variables. Finally, all variables return to their initial BGP values once the relative price of capital no longer declines, but this process occurs at a very slow rate.

4. CONCLUSION

We find in this paper that a permanent decline in the relative price of capital has long-lasting, but not permanent, effects on wage inequality. In particular, we find that cheaper capital goods initially raise the relative wage of skilled workers due to capital-skill complementarity. In addition, cheaper capital goods also initially induce more technical change to augment the existing capital stock—a "scale effect" due to the incentives to do R&D—which works toward even larger wage inequality. However, in our model economy, the other factors of production eventually respond due to (a) complementarity in production with skilled labor and (b) spillovers from research into skilled-labor intensive industries to those mainly using unskilled labor. We do not have any way of knowing how strong such spillovers are; in the model we

assume that they are strong enough to counteract the initial impulse toward inequality. If they are in fact weaker than that, the long-run outcome likely would make the share of total income accruing to unskilled workers go to zero.

Our analysis focuses on how the initial impulse—the fall in the price of new capital—induces directed R&D. We have, however, abstracted from incentives to accumulate skill. In response to a higher wage premium to skill, one would expect more skill accumulation. How strong this effect is in reality is an open question. How it would interact with the other factors driving long-run inequality in our model is also an open question. We leave these interesting questions, as well as policy analysis, for future directed research.

REFERENCES

Acemoglu, Daron. 1998. "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality." *Quarterly Journal of Economics* 113 (November): 1055–90.

_____. 2002a. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature* 40 (March): 7–73.

_____. 2002b. "Directed Technical Change." *Review of Economic Studies* 69 (October): 781–810.

_____. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of European Economic Association* 1 (March):1–37.

Goldin, Claudia, and Lawrence F. Katz. 1999. "The Returns to Skill in the United States Across the Twentieth Century." NBER Working Paper 7126 (May).

Goodfriend, Marvin, and John McDermott. 1998. "Industrial Development and the Convergence Question." *American Economic Review* 88 (December): 1277–89.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell. 1997. "Long-Run Implications of Investment-Specific Technological Change." *American Economic Review* 87 (June): 342–62.

Krusell, Per, Lee Ohanian, Victor Rios-Rull, and Giovanni Violante. 2000. "Capital Skill Complementarity and Inequality." *Econometrica* 68 (September): 1029–53.

Romer, Paul. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (October, Part 2): S71–S102.

Why Does Consumer Sentiment Predict Household Spending?

Yash P. Mehra and Elliot W. Martin

The index of consumer sentiment is one of the most watched economic indicators. It is widely believed in both the financial press and academic circles that consumer sentiment has predictive content for household spending. This belief in the predictive content of consumer sentiment is in line with most previous research that indicates the sentiment contains information about future changes in household spending beyond that already contained in past values of other available indicators.

Why does consumer sentiment predict household spending? In an interesting paper, Carroll, Fuhrer, and Wilcox (1994)—denoted hereafter as CFW (1994)—have suggested two possible interpretations of the predictive content of sentiment for household spending. One is that sentiment predicts spending because it is an independent determinant of consumer spending; changes in consumer "attitudes" cause fluctuations in the economy.[1] An alternative interpretation is that sentiment simply foreshadows the overall outlook for the economy: when consumers are optimistic about the outlook for the economy, they give upbeat responses to interviewers. On average, those expectations are validated and spending eventually increases as foreshadowed by sentiment. Sentiment, according to this interpretation, is thus just a reflection of the overall state of the economy without being a causal economic force.

The empirical evidence that can discriminate between these two alternative interpretations of the predictive ability of sentiment for spending is rather limited. CFW (1994) report evidence that favors the first interpretation. In an

The authors would like to thank Robert Hetzel, Marvin Goodfriend, and Roy Webb for many helpful comments. The views expressed herein do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] We use the term *causal* to indicate the presence of Granger causality, meaning that sentiment has incremental predictive content for spending (Engle and Granger 1987).

economy where all consumers are forward-looking and behave according to the standard permanent income model as outlined in Hall (1978), consumption follows a random walk, and hence changes in spending are unforecastable from any past information known to consumers, including the lagged sentiment measures. However, following the suggestion in Campbell and Mankiw (1989, 1990) that some households follow a rule of thumb and set consumption equal to income, CFW (1994) have argued that in an economy containing both types of consumers, sentiment might predict spending without being an independent causal force. When the economic outlook is bright, forward-looking consumers will give optimistic readings on the economy. On average, their optimism will be vindicated and income will rise. When it does, the spending of rule-of-thumb consumers will increase. Thus, by this account, the survey responses of forward-looking households predict the spending of rule-of-thumb households. In order to test this hypothesis, CFW (1994) estimate consumption regressions in which spending depends on lagged sentiment as well as on expected change in current income. The response of consumption to current income is a proxy for the influence of current economic conditions on spending, reflecting the presence of rule-of-thumb consumers. They find that lagged sentiment remains significant in the consumption equation, suggesting that sentiment is a direct determinant of household spending.

In this article, we reexamine the evidence on why sentiment predicts household spending. In most previous research, including that of CFW (1994), the effect of sentiment on spending is investigated under a number of simplifying assumptions. One such key assumption is that there is no habit persistence in consumption. If this assumption is not correct, then current consumption might depend upon lagged consumption, income, and wealth variables. The sentiment measures might then spuriously determine spending, because they are correlated with these other determinants of spending that are omitted from the spending equation. Another key assumption made in previous work is that the real interest rate is constant, thereby ruling out the direct influence of the expected change in the real rate on household spending. Hall (1988) has argued that forward-looking consumers defer consumption in response to high real rates, and hence consumption may follow a random walk once we account for the response of consumption to the expected real rate. We examine whether the results in previous research are robust to changes in the underlying assumptions.

The empirical work presented here covers the sample period 1959Q1 to 2001Q2[2] and indicates that the result in CFW (1994)—showing that sentiment is a direct determinant of spending—is not robust to the consideration of

[2] The sample period covered here differs from the one used in CFW (1994), 1955Q1 to 1992Q3. We begin in 1959 motivated in part by the easy availability of consistent time series data on all the variables used here, including the series on household wealth.

influences of other economic variables on spending. In particular, the results indicate that current consumption is indeed correlated with lagged consumption, income, and wealth variables. Consumption is also sensitive to current changes in income and the level of the real rate. Sentiment has no direct role to play in predicting consumption once its indirect influences in predicting current changes in income and the real rate are accounted for in spending equations. The results indicate that lagged sentiment is significant in predicting current changes in income and the real rate. Together these results favor the second interpretation of why sentiment predicts household spending, which is that sentiment foreshadows current expectations about the economy and the interest rate but has no direct role in actually causing fluctuations in spending.

This article proceeds as follows: Section 1 presents the empirical methodology used for testing the influence of sentiment on spending, and Section 2 presents the empirical results. In Section 3 we discuss the results, and in Section 4 we offer concluding observations.

1. EMPIRICAL MODEL AND METHOD

Permanent Income Hypothesis, Consumption Growth Regression, and Consumer Sentiment

If all consumers in the economy are forward-looking and behave according to the permanent income hypothesis as outlined in Hall (1978), then consumption follows a random walk, changes in current consumption being unforecastable from any lagged information known to consumers, including sentiment. Intuitively, according to the permanent income hypothesis, households consume their permanent income and they form expectations of their permanent income rationally taking into account all available information. To the extent that information is available and relevant to consumption in period $t + 1 (C_{t+1})$, it is already imbedded in C_t. Hence, the difference $C_{t+1} - C_t$ reflects new information regarding permanent income available at time $t + 1$. Since households form their estimates of permanent income rationally, this change in consumption must be uncorrelated with any available information, including lagged sentiment measures.

In order to further explain the random walk implication of the permanent income hypothesis and highlight the underlying assumptions, let us consider an infinitely lived representative consumer who chooses current consumption based on the expected present discounted value of his future income, not just his current income. He maximizes expected discounted utility subject to an intertemporal budget constraint. Let us assume that the utility function maximized by the representative consumer is separable in time and depends only on contemporaneous consumption during each period, as shown in (1)

below:

$$E_t \sum_{t=0}^{\infty} (1+\beta)^{-t} U(C_t), \tag{1}$$

where C is consumption, β is the subjective rate of discount, and E is the expectation conditional on information available at time period t. Equation (1) is the expected discounted utility. Let us assume further that the representative consumer can borrow and lend at the constant real rate of interest (r) and that any amount borrowed—say, in period t—must be repaid in the future by setting consumption below labor income. The consumer is assumed to choose a pattern of consumption and asset holdings in order to maximize the expected discounted utility function (1) subject to an intertemporal budget constraint.[3] The first-order conditions for this problem include

$$E_t U'(C_{t+1})(1+r)/(1+\beta) = U'(C_t), \tag{2}$$

where U' is the marginal utility of consumption. Equation (2) is the Euler consumption equation, which says the expected present value of the marginal utility of consumption tomorrow equals the marginal utility of consumption today.

If we further assume that the real rate of interest equals the consumer's discount factor ($r = \beta$) and that the marginal utility function is linear in consumption, equation (2) reduces to $E_t C_{t+1} = C_t$, which says that consumption today is the optimal forecast of consumption tomorrow. Under the additional assumption that expectations are rational, we can express the above equation in the form of a consumption growth regression, as illustrated in (3):

$$C_{t+1} - C_t = \varepsilon_{t+1}, \tag{3}$$

where ε is a rational forecast error uncorrelated with any information known to the consumer at time t. Equation (3) is Hall's famous hypothesis that under the permanent income hypothesis, change in consumption is unforecastable. Hence, according to this version of the permanent income hypothesis, lagged sentiment should not help predict future consumption growth.[4]

---

[3] See, for example, Attanasio (1998) for a simple derivation of the Euler consumption equation.

[4] The random walk result can also be derived using the permanent income hypothesis (PIH) originally proposed in Friedman (1957). The Friedman PIH allows for the presence of a transitory component in measured consumption as well as in measured income. Permanent consumption follows permanent income. In the Friedman PIH, measured consumption is a random walk if permanent income follows a random walk and if there is no transitory component in consumption. In order to explain it further, consider the following time-series representation of the Friedman PIH, as in Falk and Lee (1990): $\dot{C}_t = \dot{C}_{pt} + \delta_t$, $\dot{Y}_t = \dot{Y}_{pt} + \eta_t$, and $\dot{C}_{pt} = \beta \dot{Y}_{pt}$, where $\dot{C}_t$ and $\dot{Y}_t$ are measured consumption and measured income, $\dot{C}_{pt}$ and $\dot{Y}_{pt}$ are permanent consumption and permanent income, and δ_t and η_t are transitory consumption and income. Transitory components are assumed to be white noise disturbances mutually uncorrelated and uncorrelated with the permanent components at all lags and leads. From this formulation, it is quite clear that measured

Consumer Sentiment in Consumption Growth Regressions, Including Expected Income and the Real Rate

The random walk hypothesis developed in Hall (1978) has not done well in empirical tests. Hall himself found that lagged changes in stock prices help predict changes in consumption, while Nelson (1987) showed that consumption growth is correlated with lagged growth in disposable income. In an extension of the basic model, Hall (1988) has argued that consumption is a random walk once any movements in the real interest rate are taken into account. Campbell and Mankiw (1989, 1990), on the other hand, have argued that consumption growth is a random walk once the response of consumption growth to the contemporaneous change in income is taken into account. Those who have empirically investigated the role of consumer sentiment in predicting consumption often find that lagged sentiment does have predictive content for future consumption growth in reduced form regressions, a result inconsistent with the random walk implication of the simple permanent income model.[5]

A possible explanation as to why the random walk implication of the permanent income model has not done well in empirical tests is that some of the underlying assumptions may not be consistent with the data. One key assumption pertaining to the random walk result is that the utility function is time-separable, so that the marginal utility of consumption today depends only upon today's consumption. This assumption rules out the presence of habit persistence in consumption behavior, which may be important in practice. If there is habit persistence in consumption, then current consumption might be correlated with lagged consumption and hence correlated with lagged income and wealth variables (Dynan 1993).

The other key assumptions underlying the random walk result are that the real rate is constant and that all consumers can borrow and lend at the constant real rate. These assumptions may not be valid. The real rate may vary over time, and some consumers may face borrowing constraints and hence may be unable to smooth consumption over time. If some consumers face borrowing constraints, then their consumption may be tied to current, not permanent, income. Campbell and Mankiw (1989, 1990) have argued that some consumers follow a rule of thumb and consume their current income.

---

consumption is a random walk if $\delta_t = 0$ for all t and if permanent income follows a random walk. However, consumption may not follow a random walk if there is a serially correlated transitory component in consumption, such as the one that may arise from the presence of serially correlated preference shocks. In that environment, permanent income may not be a random walk (Sargent 1987, 374).

[5] In reduced form regressions, spending is regressed on lagged values of the sentiment and other economic indicators including changes in income, the interest rate, stock prices, and the unemployment rate. See, for example, Leeper (1992), Carrol, Fuhrer, and Wilcox (1994), and Bram and Ludvigson (1998).

In the presence of rule-of-thumb consumers, aggregate consumption may appear sensitive to changes in current income. Other analysts have argued that consumption may also appear sensitive to changes in current income if the marginal utility of consumption depends upon factors other than consumption. For example, Baxter and Jermann (1999) have argued that consumers may substitute between home- and market-produced consumption goods, and hence the marginal utility of consumption may depend upon the labor-leisure choice, in addition to depending upon the level of consumption. Thus, consumption may appear sensitive to changes in current income.

Another interesting scenario in which the random walk result may not hold is outlined in Goodfriend (1992). The Hall model described above is the representative agent model in which the representative agent is assumed to fully know the income process. The aggregate income process is the individual income process, because all agents are assumed to be alike. Goodfriend, however, considers an economy with heterogeneous agents, where agents have individually specific income processes that may differ from the aggregate income process. If there is complete information about the aggregates, the random walk result holds at the aggregate level. However, if agents do not have contemporary information on the aggregate income, as is the case in practice since the aggregate income data are released with a lag, then aggregation yields a consumption equation that violates the random walk result. In particular, consumption is correlated with changes in lagged income. Intuitively, in the absence of contemporary information on the aggregate income, agents cannot distinguish between aggregate and relative shocks affecting their individual incomes. As a consequence, if there is an aggregate income shock, it may partially be interpreted as a shock to the individual-specific component of individual labor income. If the individual-specific component is less persistent than the aggregate component, then agents will fail to adjust their permanent incomes appropriately, and hence consumption observed will not move too much. However, in subsequent periods, as information on the aggregate income becomes available and the effect on actual income is observed to persist, consumption will adjust fully and will appear sensitive to lagged changes in actual income.[6]

In view of the considerations listed above, we examine the predictive content of sentiment for future changes in consumption using consumption growth regressions that allow for the lagged influences of other economic determinants of spending on current consumption. In particular, we consider

[6] Pischke (1995) extends Goodfriend's argument to the economy in which agents have no information on economy-wide variables.

consumption growth regressions of the form

$$\dot{C}_t = a + \lambda_y E_{t-1}\dot{Y}_t + \lambda_r E_{t-1} r_t + \sum_{s=1}^{k} b_s Z_{t-s} + \sum_{s=1}^{k} c_s S_{t-s} + \varepsilon_t, \quad (4)$$

where $E_{t-1}\dot{Y}_t$ is income growth expected for period t conditional on information at $t-1$; $E_{t-1}r_t$ is the real interest rate expected for period t conditional on information at $t-1$; Z is a set of control variables containing lagged values of consumption and other plausible economic determinants of spending; and S is an index of consumer sentiment. Equation (4) allows for the possibility that consumption is sensitive to current income growth as well as to the real rate. Furthermore, equation (4) also allows for the possibility that consumption is correlated with lagged values of economic factors (Z) other than consumer sentiment. For example, as indicated before, lagged consumption or other variables might enter directly into the consumption equation if there is habit persistence in consumption behavior or if the marginal utility of consumption depends upon factors other than the level of consumption.

In equation (4) consumer sentiment may help forecast consumption growth through two channels. The first channel is an indirect one: lagged sentiment helps predict consumption growth in period t because it is instrumental in predicting current income growth and the level of real interest rate for period t. The other channel is a direct one: lagged sentiment directly enters the consumption equation (4). It is possible that lagged sentiment may help predict consumption growth through both channels. CFW (1994) use the evidence on the presence of these two channels to distinguish between the two interpretations of why sentiment helps predict consumption growth. Sentiment may be considered an independent determinant of consumer spending if it directly enters the consumption equation (all $c_s \neq 0$ in (4)). In contrast, sentiment may be considered a passive predictor of spending because it just foreshadows current economic conditions. In this interpretation, lagged sentiment no longer directly enters the consumption equation (4) once its role as a predictor of current income and the real rate is allowed for in the consumption equation (all $c_s = 0$, but λ_y, $\lambda_r \neq 0$ in (4)). In this interpretation, sentiment is a predictor of household spending without being an independent causal force.

In previous research the predictive content of sentiment for household spending has been investigated using restricted versions of (4). For example, CFW (1994) investigate the role of sentiment using an aggregate consumption equation of the form

$$\dot{C}_t = a + \lambda_y E_{t-1}\dot{Y}_t + \sum_{s=1}^{k} c_s S_{t-s} \quad (5)$$

and find that sentiment enters the consumption equation directly. This empirical evidence is suspect. This specification of the consumption equation implicitly assumes that lagged values of consumption and other economic

variables do not enter the consumption equation directly. Moreover, consumption is assumed to be insensitive to the expected real rate. If other relevant variables are omitted from the consumption equation, then lagged sentiment may spuriously appear to predict consumption. Others have investigated the role of sentiment using reduced form consumption regressions of the form given below in (6) (Bram and Ludvigson 1998):

$$\dot{C}_t = a + \sum_{s=1}^{k} b_s Z_{t-s} + \sum_{s=1}^{k} c_s S_{t-s} + \varepsilon_t. \tag{6}$$

In this specification, even though there is a set of control variables including lagged values of consumption and other plausible economic determinants of spending, such as interest rates and income, consumption is still assumed to be insensitive to current income and the real rate. In view of these considerations, we reexamine the role of sentiment using instead the consumption equation (4).

Data, Estimation, and the Issue of Constancy of Second Moments

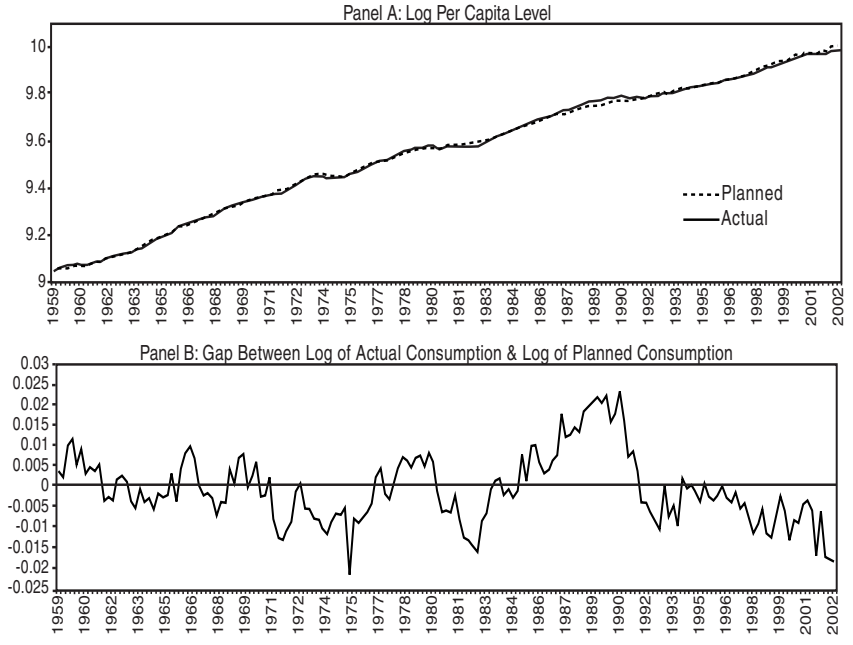
We investigate the role of sentiment in predicting spending using consumption equations of the form (4) and estimated using quarterly data over 1959Q1 to 2001Q2.[7] Consumption is measured as per capita consumption of nondurables and services, in 1996 dollars (C). Labor income is measured as disposable labor income per capita, in 1996 dollars (Y).[8] The real rate (r) is measured as the three-month Treasury bill rate minus the contemporaneous inflation rate; the latter is measured by the behavior of the consumption expenditure deflator. The index of consumer sentiment used here is the Expectations Component of the University of Michigan Sentiment Index.[9] The additional variables (Z) considered here include past values of consumption growth and the lagged

[7] The quarterly data used are of vintage 2002. We truncate the sample in 2001Q2 so that our results would not be affected by recent developments pertaining to terrorism or the war in Iraq.

[8] As in most previous research, we present results using disposable labor income rather than disposable personal income that also includes property income. The evidence in previous research is consistent with the presence of a different marginal propensity to consume out of labor and property incomes. Since the empirical work here includes the lagged residual from the cointegrating regression that includes labor income and wealth, the consumption regression indirectly captures the influence of property income. Labor income is defined as wages and salaries + transfer payments + other labor income – personal contributions for social insurance – taxes. Taxes are defined as [wages and salaries/(wages and salaries + proprietor's income + rental income + personal dividends + personal interest income)] personal tax and nontax payments.

[9] We use the Expectations Component because we are interested in examining the impact of beliefs about future economic conditions on current spending. For robustness, we do examine results using the Total Index. The results with the Total Index are qualitatively similar to those with the Expectations Component (see, for example, row 6 of Table 1). See the Appendix for the list of questions included in the sentiment surveys.

Figure 1 Cointegrating Regression: Actual and Planned Consumption



residual from the cointegrating regression estimated using levels of per capita consumption, labor income, and household net worth. The evidence in Mehra (2001) indicates that consumer spending is cointegrated with labor income and household wealth and that changes in current consumer spending depend in part upon lagged income and wealth variables through the error-correction term (Engle and Granger 1987). The lagged residual from the cointegrating regression, when included in the consumption equation of the form (4), captures in a parsimonious way the response of current consumption to lagged values of income and wealth variables. Wealth used in this cointegrating relationship is measured as per capita net worth of households, in 1996 dollars.

Equation (7) below reports the cointegrating regression estimated using real, per capita consumer spending, labor income, and household net worth over 1959Q1 to 2001Q2:

$$C_t = \underset{(21.4)}{3.7} + \underset{(46.1)}{.51}\, Y_t + \underset{(6.1)}{.07}\, W_t + \underset{(21.7)}{.002} T, \tag{7}$$

where all variables are in their natural log levels and where Y is per capita labor income; W is per capita household net worth; and T is a linear time trend. Parentheses below coefficients contain t-values corrected for the presence of

serial correlation and heteroscedasticity.[10] All variables appear with theoretically expected signs and are significant. Panel A in Figure 1 charts the (log) level of actual consumer spending and the level predicted by the cointegrating regression (7), and Panel B charts the gap between actual and predicted levels, which is the residual from the cointegrating regression (7). As can be seen in Figure 1, the actual and predicted consumption series move quite closely and the gap variable appears stationary over the sample period. In the consumption growth regression (4), the residual series is one of the variables that appear in the set Z.

The consumption growth regressions like (4) and (5) relate consumption to expected values of income growth and the level of the real rate and have been estimated using instrumental variables methods and assuming that expectations are rational (Hall 1988; Campbell and Mankiw 1989). Under the assumption of rational expectations, consumers take into account all known information in forming their expectations, and the forecast error is uncorrelated with any lagged information. Hence, period $t-1$ values of information variables are valid instruments. Hall (1988), however, notes that if the frequency with which consumption decisions are taken is higher than the frequency of observations (quarterly in our case), then under some assumptions the residuals of equations may have the first-order moving average structure. In that case, valid information for instruments will be any information dated $t-2$ or earlier. We follow Hall in using instruments lagged $t-2$ and before. The fact that aggregate data on income are available with a one-period lag also implies that period $t-2$ values will be in the information set of consumers (Goodfriend 1992). The instruments used are a constant, four lagged values of consumption growth, change in the unemployment rate, change in the real rate, and the level of the index of consumer sentiment. Following Campbell and Mankiw (1989), we also report the test of overidentifying restrictions, which is a test of the hypothesis that the instruments used are uncorrelated with the residual of the consumption equation.[11]

The consumption regression (4) relates consumption to income growth and the real rate among other factors. This regression assumes that second moments measuring volatility of economic variables are constant, implying that consumption is unaffected by second moments of expected income and the real rate. Mehra (2003) has recently argued that over the sample period (1959Q1 to 2001Q4) consumption is correlated negatively with the second moment of the real rate, which measures interest rate volatility. If the consumption equation

[10] The reported t-values have been correcting allowing for the presence of fourth-order serial correlation, as indicated by the underlying estimated autocorrelation coefficients.

[11] This test is performed by regressing the residual from the instrumental variables regression on the instruments, and then comparing T times the R-squared from this regression, where T is sample size, with the chi-squared distribution with $(K-1)$ degrees of freedom, K being the number of estimated parameters (Campbell and Mankiw 1989).

is estimated ignoring the presence of this negative correlation between consumption and interest rate volatility, then the estimated interest rate coefficient (λ_r) that measures the response of consumption to the expected real rate is biased downward. In view of such evidence, the consumption growth regression (4) is estimated including the interest rate volatility variable in a nonlinear fashion. In particular, the consumption regression is estimated including the interest rate volatility variable interacting with the real interest rate.[12]

2. EMPIRICAL RESULTS

Table 1 presents instrumental variables estimates of the consumption growth regressions like those in (4) and (5) for the full sample period, 1959Q1 to 2001Q2. Row 1 presents the consumption equation estimated including only current income growth as in Campbell and Mankiw (1989). The maintained hypothesis here is that consumption follows a random walk once we account for the sensitivity of consumption to current income, arising as a result of the presence of rule-of-thumb or liquidity-constrained consumers. χ_1^2 is a chi-square statistic that tests the hypothesis that the four lagged values of the sentiment measure are not jointly significant when included in the estimated consumption equation given in row 1. χ_2^2 is a chi-square statistic that tests the hypothesis that the four lagged values of the sentiment measure used in the prediction equation for current income growth are not jointly significant. χ_2^2 is large, suggesting that lagged sentiment contains information about current income growth. However, χ_1^2 is also large, implying that sentiment continues to have a predictive content for household spending, even after one accounts for its indirect role in predicting current consumption through the expected income channel. This result is qualitatively similar to the one in CFW (1994), interpreted to mean that sentiment is a direct determinant of consumer spending.

Row 2 in Table 1 estimates the consumption equation including expected income growth as well as the lagged residual from the cointegrating regression (7) that is estimated using levels of consumption, income, and wealth variables.

[12] The evidence in Mehra (2003) also indicates that the period from 1979 to the early 1980s accounts for the presence of negative correlation between consumption and interest rate volatility found in the full sample. This subperiod coincides with the Fed aggressively raising real rates in order to fight inflation. The increased volatility that accompanied the high level of real rates may have led to increased uncertainty about future real rates, deterring substitution of consumption in time. In view of this consideration, we further restrict the interactive interest rate volatility variable to take nonzero values only over the subperiod 1979Q3 to 1984Q4. However, results are qualitatively the same if the interactive variable is entered without the dummy as above (see Mehra 2003).

Table 1 Testing the Predictive Content of Sentiment

$$\dot{C}_t = a + \lambda_y E_{t-1}\dot{Y}_t + \lambda_r E_{t-1} r_t + b_o LRC_{t-1} + \sum_{s=1}^{k} b_s \dot{C}_{t-s} + \lambda_{rr}(r^* Vol)_t \qquad \textbf{(A)}$$

Row	λ_y	λ_r	b_o	Σb_s	λ_{rr}	χ^2_1	χ^2_2	χ^2_3	$\bar{R}^2$	p-value for overidentifying restrictions
1	0.53 (5.9)					14.7*	10.8*		0.03	0.14
2	0.57 (6.3)		−0.37 (2.0)			3.21	11.5*		0.01	0.61
3	0.49 (5.7)	0.20 (1.8)	−0.58 (3.5)		−0.37 (2.4)	3.2	12.8*	23.1*	0.20	0.84
4	0.32 (2.3)	0.19 (2.1)	−0.60 (3.7)	0.32 (1.6)	−0.27 (2.1)	0.71	12.7*	23.0*	0.46	0.91
5[a]	0.26 (3.3)	0.16 (1.9)	−0.71 (5.2)	0.39 (4.5)	−0.37 (3.3)	1.5	12.7*	60.5*	0.49	0.78
6[b]	0.33 (2.3)	0.22 (2.3)	−0.58 (3.5)	0.33 (1.7)	−0.28 (2.0)	1.8	8.3*	15.1*	0.44	0.95

Notes: The coefficients reported above are instrumental variables estimates of the consumption equation (A) over 1962Q1–2001Q2. $\dot{C}$ is consumption growth; $\dot{Y}$ is income growth; r is the real rate; $(r^* Vol)$ is the real rate interacting with the interest rate volatility variable; and LRC is the residual from the cointegrating regression (7) of the text. The instruments used are a constant, four lagged values of consumption growth, change in the unemployment rate, the real rate, consumer sentiment, and the lagged residual from the cointegrating regression. Instruments are dated period $t-2$ and earlier. χ^2_1 is the chi-square statistic that tests the hypothesis that four lags of consumer sentiment when included in the pertinent consumption equations are zero. χ^2_2 and χ^2_3 are chi-square statistics that test the joint significance of coefficients that appear on four lags of sentiment in the first-stage regressions for income and the real rate. The test for overidentifying restrictions tests whether the instruments used are correlated with the residual of the estimated consumption equation.

[a] Instruments are dated $t-1$ and earlier.

[b] Sentiment measure used is the Total Component of the University of Michigan Sentiment Index.

* Significant at the 0.05 level.

The lagged residual is significant in the estimated consumption equation, suggesting that current consumption is directly correlated with lagged income and wealth variables. Consumption is still sensitive to current income growth, and sentiment remains significant in predicting changes in current income (see the t-value on expected income and the chi-square statistic χ^2_2 in row 2, Table

1). However, sentiment no longer directly enters the estimated consumption equation (see the statistic χ_1^2 in row 2, Table 1). This result suggests that sentiment is not a direct determinant of household spending. Together these results suggest that since consumption is directly correlated with lagged income and wealth variables, their exclusion from the estimated consumption equation spuriously generates the result that sentiment is a direct determinant of household spending.

Row 3 in Table 1 estimates the consumption equation including expected income, the real rate, and the lagged residual from the cointegrating regression. As can be seen, consumption is sensitive to the expected real rate as well as to expected income (see t-values on these variables in row 3, Table 1). The lagged residual is also significant in the estimated consumption equation. However, the chi-square statistic χ_1^2 is small, implying that sentiment does not enter directly into the estimated consumption equation. χ_3^2 is the chi-square statistic that tests the hypothesis that lagged sentiment is not significant in predicting the real rate. This statistic is large, suggesting that sentiment does happen to contain information about current real rates.

In the consumption regressions discussed above, including the lagged residual from the cointegrating regression captures the dependence of current consumption on lagged income and wealth variables. The results do not change if the consumption equation is estimated including also lagged consumption growth. Row 4 of Table 1 reports the consumption regression estimated including three lagged values of consumption, in addition to the lagged residual of the cointegrating regression. As can be seen, the estimates are still consistent with the basic result: sentiment is not an independent determinant of consumer spending.

Row 5 in Table 1 presents the consumption equation estimated using instruments dated $t - 1$ and earlier. The estimated coefficients that appear on various variables change to a certain degree. However, the estimates still are consistent with the basic result that lagged sentiment is not a direct determinant of spending once we control for the influences of current income, the real rate, and other lagged income and wealth variables on spending. The results do not change if a consumption equation similar to the one in row 4 is estimated using instead the University of Michigan Total Sentiment Index (see row 6 in Table 1).

3. DISCUSSION OF RESULTS

The empirical work indicates that consumer sentiment has predictive content for future changes in income and the real rate.[13] However, sentiment has

[13] An additional table containing these first-stage regressions is available upon request from the authors.

no predictive content for consumption once we control for the influences of income and the real rate on consumption that work through the contemporaneous income and interest rate channels. Together these results suggest that sentiment is not a direct determinant of spending. One possible interpretation of these results based on Goodfriend's (1992) model discussed above is that sentiment surveys enable households to discriminate better between aggregate and relative shocks affecting their individual labor incomes, as sentiment surveys are available before data on the direct determinants of aggregate income are released. By sharpening the assessment of the current aggregate income and hence the aggregate shock, sentiment surveys enable more and more households to adjust their individual permanent incomes appropriately, thereby bringing consumption more in line with permanent income. If consumer sentiment surveys do help in this signal processing, then one would expect a diminished role of lagged income and hence lagged sentiment measures in predicting current consumption at the aggregate level. Hence, one may find that sentiment has no direct role in determining spending once one controls for the direct influence of current aggregate income on spending.

The fact that sentiment measures are so eagerly awaited and watched both in the financial press and by many serious economic analysts suggests they may be useful in sharpening the assessment of agents for the current state of the economy as measured by the behavior of aggregate income. The empirical result here indicating that sentiment measures lose their statistical significance in predicting current spending once one controls for the influences of the current state of the economy on spending suggests that these sentiment measures may have value as a summary statistic for the future course of consumption.

4. CONCLUDING OBSERVATIONS

Consumer sentiment might help predict household spending, either because sentiment is an independent determinant of spending or because it foreshadows current economic conditions. In order to distinguish empirically between these two interpretations of the predictive content of sentiment, we estimate the consumption equation that nests both these interpretations. In particular, consumer spending is assumed to be sensitive to current income and the real rate, in addition to depending upon lagged spending, income, wealth, and sentiment variables. The response of spending to current income and the real rate is a proxy for the influences of current economic conditions on spending, whereas the response of spending to lagged sentiment is a proxy for the direct influence of sentiment on spending. In previous research the predictive content of sentiment has generally been investigated using consumption equations without controlling for the sensitivity of current consumption to the expected

real rate and lagged income and wealth variables. The results here indicate that lagged sentiment has no direct role in predicting spending once we control for the direct influences of current income, the real rate, and other lagged determinants on spending.

Another interesting result is that consumer sentiment does have predictive content for future changes in income and the real rate, suggesting that sentiment measures are useful as a good barometer of the near-term course of the economy and hence consumption. Since in real time consumer sentiment measures are released before aggregate data on the current state of the economy are available, sentiment measures may be helpful in assessing the near-term direction of the economy. This may explain why sentiment measures are so eagerly awaited in the financial press and by many economic analysts.

APPENDIX: QUESTIONS IN THE MICHIGAN SURVEYS OF CONSUMERS

The University of Michigan publishes an overall index of consumer sentiment and two component indices measuring current economic conditions and consumer expectations. The overall index is based on answers to five survey questions, presented below. Two of the survey questions are used to calculate the current conditions component, and three questions underlie the expectations component.

Current Economic Conditions

Component Questions

Q_1 = "We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?"

Q_2 = "About the big things people buy for their homes—such as furniture, a refrigerator, stove, television, and things like that. Generally speaking, do you think now is a good or a bad time for people to buy major household items?"

Expectations

Component Questions

Q_3 = "Now looking ahead—do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?"

Q_4 = "Now turning to business conditions in the country as a whole—do you think that during the next 12 months we'll have good times financially, or bad times, or what?"

Q_5 = "Looking ahead, which would you say is more likely—that in the country as a whole we'll have continuous good times during the next 5 years or so, or that we will have periods of widespread unemployment or depression, or what?"

For details on the underlying methodology, see the papers, including the one by Richard T. Curtin, available at the public access Web site of the Institute for Social Research: http://www.sca.isr.umich.edu/.

REFERENCES

Attanasio, Orazio P. 1998. "Consumption Demand." NBER Working Paper 6466 (March).

Baxter, Mariane, and Urban J. Jermann. 1999. "Household Production and the Excess Sensitivity of Consumption to Current Income." *American Economic Review* 89 (September): 902–20.

Bram, Jason, and Sydney Ludvigson. 1998. "Does Consumer Confidence Forecast Household Expenditure? A Sentiment Index Horse Race." Federal Reserve Bank of NewYork *Economic Policy Review 4* (June): 59–78.

Campbell, John Y., and Gregory N. Mankiw. 1989. "Consumption, Income, and Interest Rates: Re-interpreting the Time Series Evidence." In *NBER Macroeconomics Annual*, edited by Olivier J. Blanchard and Stanley Fischer. Cambridge: MIT Press.

_____. 1990. "Permanent Income, Current Income and Consumption." *Journal of Business and Economic Statistics* 8 (July): 265–79.

Carroll, Christopher D., Jeffrey C. Fuhrer, and David W. Wilcox. 1994. "Does Consumer Sentiment Forecast Household Spending? If So, Why?" *American Economic Review* 84 (December): 1397–1408.

Dynan, Karen E. 1993. "Habit Formation in Consumer Preferences: Evidence from Panel Data." Working Paper 143, Economic Activity Section, Board of Governors of the Federal Reserve System.

Engle, Robert F., and C. W. Granger. 1987. "Co-integration and Error-Correction: Representation, Estimation, and Testing."

Econometrica 55 (March): 251–76.

Falk, Barry, and Bong-Soo Lee. 1990. "Time-Series Implications of Friedman's Permanent Income Hypothesis." *Journal of Monetary Economics* 26 (October): 267–83.

Friedman, Milton. 1957. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton University Press.

Goodfriend, Marvin. 1992. "Information-Aggregation Bias." *American Economic Review* 82 (June): 508–19.

Hall, Robert E. 1978. "The Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy* 86 (December): 971–87.

_____. 1988. "Intertemporal Substitution in Consumption." *Journal of Political Economy* 96 (April): 339–57.

Leeper, Eric M. 1992. "Consumer Attitudes: King for a Day." In Federal Reserve Bank of Atlanta *Economic Review* (July): 1–15.

Mehra, Yash P. 2001. "The Wealth Effect in Empirical Life-Cycle Aggregate Consumption Equations." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (Spring): 45–68.

_____. 2003. "Fed Policy and Estimation of Intertemporal Elasticity of Substitution in Consumption." Federal Reserve Bank of Richmond, mimeo.

Nelson, Charles R. 1987. "A Reappraisal of Recent Tests of the Permanent Income Hypothesis." *Journal of Political Economy* 95 (June): 641–46.

Pischke, Jorn-Steffen. 1995. "Individual Income, Incomplete Information and Aggregate Consumption." *Econometrica* 63 (July): 805–40.

Sargent, Thomas J. 1987. *Macroeconomic Theory.* Orlando, Fla.: Academic Press.

Firms, Assignments, and Earnings

Edward Simpson Prescott

The U.S. distribution of labor earnings is highly skewed to the right. Roughly, the lowest 50 percent of U.S. households, as measured by individual labor earnings, make 10 percent of total labor earnings. The next lowest 30 percent earn approximately 30 percent and highest 10 percent make 40 percent.[1]

Earnings are also related to a person's position within a firm and employment at a particular firm. Within a firm earnings tend to be associated with rank. The higher is an individual's authority and control, the higher is his compensation. The most extreme manifestation of this is the enormous pay of the top executives of large firms. In 1996 the median pay of chief executive officers of companies in the S&P 500 index was nearly 2.5 million dollars (Murphy 1999).

Across firms earnings tend to increase with firm size. This is particularly true for executives. The elasticity of executive pay with respect to firm size is in the range of 0.20 to 0.35 (Rosen 1992). Earnings for workers also increase with firm size. This is the well-documented wage-size premium (Brown and Medoff [1989] and Oi and Idson [1999]).

The standard neo-classical production function, where output equals a function of aggregate labor and aggregate capital, cannot simultaneously account for these facts. It can generate an unequal distribution of earnings, if some people's labor is more efficient than others. But it has only one economy-wide firm so it is necessarily silent on any relationship between earnings and firm assignments. And even with respect to the distribution of earnings, the

The author would like to thank Andreas Hornstein, Tom Humphrey, Pierre Sarte, and John Weinberg for helpful comments. The views expressed in this article do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] These are 1998 numbers taken from the Survey of Consumer Finances as reported by Rodriguez, Diaz-Gimenez, Quadrini, and Rios-Rull (2002). They define labor earnings as wages and salaries plus 85.7 percent of business and farm income.

inequality of labor efficiency that would imply such a distribution seems so unequal as to defy credulity.

For a theory to explain these facts, it needs to solve the problem of *jointly* assigning workers and managers to firms. This paper sketches such a theory that is based on the firm-size model of Lucas (1978) and on the hierarchy models of Rosen (1982, 1992). For simplicity, most of the analysis focuses on firms with only two types of jobs, executives and workers. This is enough to illustrate the connection between pay and rank within and across firms; it also has the advantage of allowing us to discuss the well-documented patterns in executive pay.[2]

In a firm the role of a manager is more important than that of any single worker, just as the role of the chief executive officer is more important than that of any subordinate, manager or worker. Firms are structured as hierarchies in which decisions made by a high-level manager affect the productivity of individuals in lower levels who report directly, or indirectly, to the manager. Decisions made at each successively higher level in a firm affect proportionately more people. Ultimately, the top executive's decisions affect the productivity of everyone within the firm. Figure 1 illustrates.

For this reason it matters a lot who is assigned to the top positions within a firm. For a firm, a small difference in managerial talent at the highest level leads to a big difference in output. As a consequence, within a firm it is best to place the most talented individual at the top, while across firms it is best to place the most talented individual at the largest firm. For both of these reasons, scarce managerial talent can be incredibly valuable. Within firms it leads to earnings inequality over rank. Across firms it lead to earnings inequality over firm size.

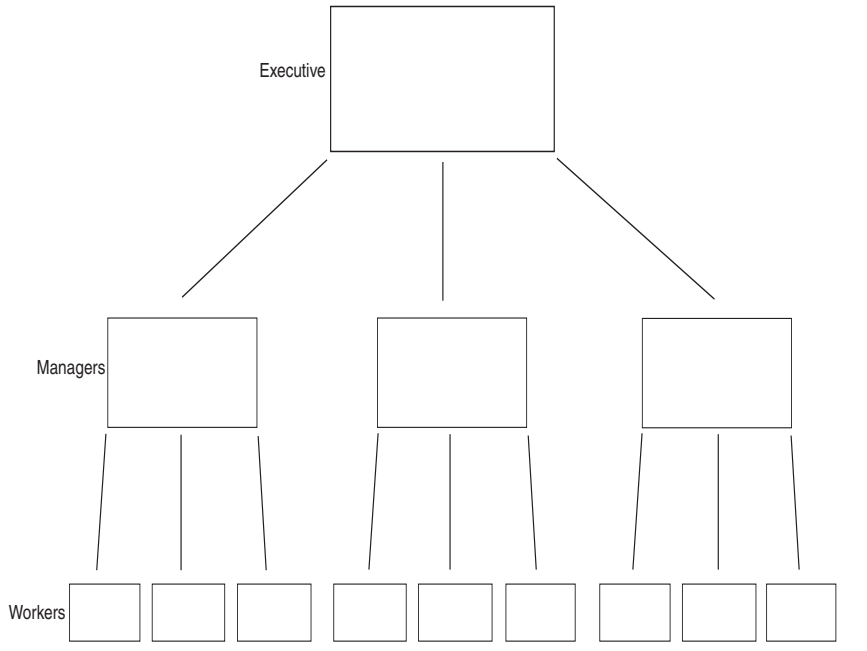
Section 1 studies a problem where people are assigned to be either workers or managers. All firms are a hierarchy with one level of management. This model is a simplification of Lucas (1978). Short discussions of executive pay, the wage-size premium, and marginal product pricing in assignment models are included. Section 2 studies a simple extension of the Lucas model to incorporate multi-level hierarchies as in Rosen (1982). Section 3 provides a concluding discussion.

1. TWO-LEVEL FIRMS

All production in this economy is done by firms. Each firm consists of a manager and a number of workers. A firm's production depends on the talent of

[2] Much of the recent literature on executive compensation has focused on the important question of *how* to pay executives in order to motivate them to act in the best interests of the corporation. That issue is not discussed in this paper. Surveys can be found in Rosen (1992) and Murphy (1999).

Figure 1 Organization of Production within a Firm



Notes: A three-level firm in which decisions made by a manager affect the productivity of all individuals who directly and indirectly report to him. The single individual at the top is the executive, the three individuals at the second level are lower-level managers, and the remaining nine individuals are workers.

the manager and the amount of labor supplied by the workers. The production function is $tf(l)$, where t is the talent of the manager and l is the number of workers working for him. The function $f(l)$ is concave so given a manager there is decreasing returns to scale in the number of workers who work for him. Decreasing returns at the firm level will lead to the existence of multiple firms rather than just one large firm with everyone working under the most talented manager. The number of workers working for a manager is often called a manager's *span of control*. The more talented a manager is the more workers who work for him, and the larger is his span of control.

People differ in their managerial talent. Talent is distributed by $h(t)$ across the population. $H(t)$ is the cumulative distribution function. There is an indivisibility in an individual's job. A person can either be a manager or a worker, but not both at the same time. The problem in this economy is to determine who will be a manager and then how many workers will be assigned to each manager.

Each person must decide whether to be a worker or a manager. If he chooses to be a worker then he receives the labor wage of w, which is independent of his talent. If he chooses to be a manager, he must decide how much labor to hire. He does this by solving

$$\max_{l \geq 0} tf(l) - wl.$$

The first-order condition is

$$tf'(l) - w = 0. \tag{1}$$

A manager's earnings π is equal to $tf(l) - wl$. Naturally, a manager must be paid at least as much as the wage or he would choose to be a worker. Since managerial earnings are increasing in talent there is a unique cutoff level of talent z for which all people with $t \geq z$ are managers and the rest are workers. Let $l(t)$ be the labor hired by a manager of talent t. Then,

$$zf(l(z)) - wl(z) = w. \tag{2}$$

This condition just states that a marginal manager's profit, $zf(l(z)) - wl(z)$, equals his opportunity cost of working, w.

There is also a resource constraint on the supply of labor. It is

$$\int_z^\infty l(t)dH(t) \leq H(z). \tag{3}$$

The left-hand side is labor hired while the right-hand side is labor supplied.

A *competitive equilibrium* is a cutoff level of talent z, an assignment of labor to managers $l(t)$ for $t \geq z$, and a wage w that satisfies the managers' first-order conditions, that is, (1) for all $t \geq z$, indifference for the marginal manager (2), and the resource constraint (3).

To illustrate the connection between firms and pay, we study the case where $f(l) = l^\alpha$ with $0 < \alpha < 1$. The number of employees assigned to a firm, $l(t)$, can be determined from (1). It satisfies

$$l(t) = (w/\alpha)^{1/(\alpha-1)} t^{1/(1-\alpha)}.$$

Because α is between zero and one, the number of employees grows more than proportionately with the manager's talent. Another measure of firm size is firm revenue or output $q(t)$. Its relationship with talent is nearly identical to that of $l(t)$. It is

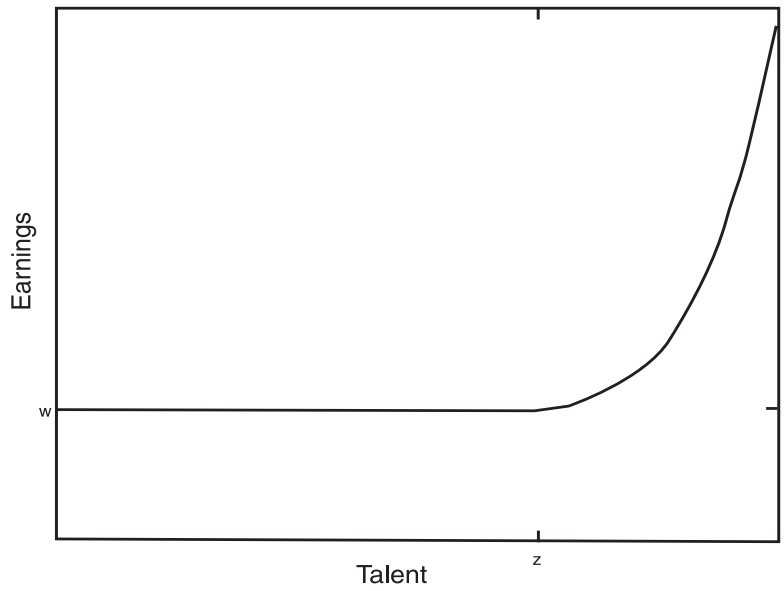
$$q(t) = tl(t)^\alpha = (w/\alpha)^{\alpha/(\alpha-1)} t^{1/(1-\alpha)}. \tag{4}$$

A similar relationship holds for managerial pay. Then,

$$\pi(t) = tl(t)^\alpha - wl(t) = \left((w/\alpha)^{\alpha/(\alpha-1)} - w(w/\alpha)^{1/(\alpha-1)}\right) t^{1/(1-\alpha)}. \tag{5}$$

Managerial pay grows more than proportionately with talent. Small differences in talent at the managerial level lead to large differences in pay (and firm size). The result is appealing because it implies that even with a symmetric

Figure 2 Earnings and Talent



Notes: Earnings as a function of talent. All individuals with talent $t < z$ are assigned to be workers and earn wages w. Individuals with more managerial talent are managers and their pay is an increasing convex function of their talent.

distribution of talent, which has some natural appeal, earnings and firm size will be skewed to the right, as is observed in the data.[3] Figure 2 illustrates the relationship between talent and earnings in this example.

While the relationship of firm size and executive pay to talent is of interest, the applicability of these theoretical results is limited. The talent distribution is not observed and there is little hope of directly observing it. However, the theory does predict a relationship between firm size and executive pay that, for this example, is independent of the talent distribution or talent level. The relationship follows directly from (1). Notice that

$$\frac{q(t)}{l(t)} = tl(t)^{\alpha-1} = \frac{w}{\alpha}. \tag{6}$$

[3] Like the earnings distribution, firm size is highly skewed to the right. This is true for a variety of firm size measures like assets, employment, sales, and others. See Simon and Bonini (1958).

Managerial pay with respect to firm size, as measured by $q(t)$, is

$$\pi(t) = q(t) - wl(t) = (1 - \alpha)q(t). \tag{7}$$

In this example, managerial pay is linear in firm size. (It is also linear in firm size if firm size is defined as number of employees.)

Implications for Executive Pay

Qualitatively, the theory seems on the mark. Executive pay grows with firm size. Quantitatively, however, some other functional form is needed. In the data the *log* of executive pay is linear in the *log* of firm size, which means that the level of executive pay takes the form

$$pay = b(size)^{\beta}. \tag{8}$$

Numerous studies find that the elasticity, β, is around 0.20-0.35. Elasticities in this range have been found in U.S. data during the 1940s and 1950s (Roberts 1956), U.S. data in the late 1930s (Kostiuk 1989), U.S. data from 1969–1981 (Kostiuk 1989), U.K. data during 1969–1971 (Cosh 1975), and U.S. banking data in the 1980s (Barro and Barro 1990). See Rosen (1992) and Murphy (1999) for more discussion.

One important feature of the data that the model is silent on is the large increase in the ratio of executive pay to worker pay observed over the last 30 years. In 1970 the average executive of an S&P 500 firm made 30 times the average worker wage. In 1996 this ratio was 90 for cash compensation and 210 for realized compensation, which includes the value of exercised stock options (Murphy 1999).

One strategy for addressing this question is to postulate that there was an exogenous change in the technology by changing the production technology to $tAf(l(t))$, where $A > 1$, and f is homogenous of degree α. Interestingly, this has no effect on the economy except to raise everyone's wealth by a factor of A. In particular, set the new wage to Aw and keep the z and the $l(t)$ unchanged from the above model. This allocation satisfies the first-order conditions. Worker pay grows by the factor A and so does managerial pay. Managers still supervise the same number of people and wages and managerial rents increase by the constant factor.

More promising strategies include postulating an exogenous change in the span of control technology, say, from advances in information technology, or by introducing capital. Lucas (1978) includes capital so that the production function is $tf(g(l, k))$, where g is a constant returns-to-scale technology. In his model, as an economy grows wealthier the capital-to-labor ratio in firms increases, there are less firms, and firm size increases. These forces increase executive pay, though the precise effect on the ratio of executive to worker pay is unclear because wages increase as well. Still, the growth in executive pay

in the last 30 years seems much greater than can be accounted for by changes in the capital stock so one would guess that other factors are also at work.

Talent as a Worker Input

In the simple model of the previous section there was a constant wage for all workers. A worker was paid the same no matter what firm employed him. In the data, however, there is a premium for working for a larger firm. This well-documented observation has been reported by Brown and Medoff (1989), Idson and Oi (1999), Oi and Idson (1999), Troske (1999), and others. Idson and Oi (1999) report an elasticity of wages with respect to plant size of 0.075 using 1992 data from the Census of Manufactures. This size elasticity implies that an employee who works for a plant that is twice the size of another plant earns 5 percent more than an employee at the smaller plant.

In this section, we modify the production function to allow talent to affect output at the worker level as well as at the managerial level. This will add more earnings inequality. Alone it does not necessarily generate a wage-size premium but it does give some insight into what might generate it.

Let $d(t)$ be the total *talent* of workers assigned to manager t and let the production function now be $tf(d(t))$. The wage w now refers to the payment per unit of hired talent so a worker of talent t is paid wt. A manager's maximization problem is

$$\max_{d\geq 0} tf(d) - wd.$$

The first-order condition is nearly identical to that of the previous problem. It is

$$tf'(d) - w = 0. \tag{9}$$

Marginal managers are indifferent to managing and working. This condition is

$$zf(d(z)) - d(z)w = zw. \tag{10}$$

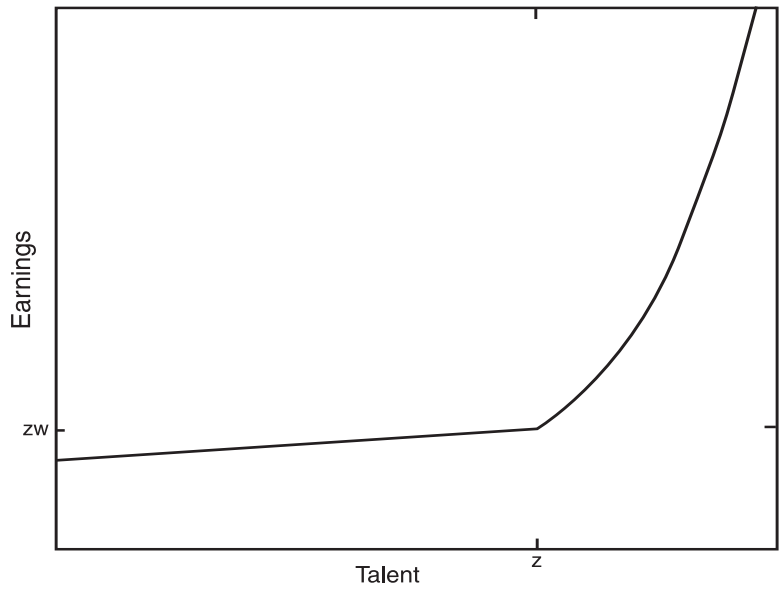
Notice that now the opportunity cost of managing is the marginal manager's talent times the wage.

Finally, the resource constraint on available talent is

$$\int_z^\infty d(t)dH(t) \leq \int_0^z tdH(t). \tag{11}$$

The primary advantage of this formulation is that worker pay varies with talent. Figure 3 describes the dependence of pay on talent for the production function $f(d) = d^\alpha$. Unlike the previous model, worker pay now varies and is linear with talent. However, the relationship between managerial talent and managerial pay is identical to that in the previous model. The connection between firm size and managerial pay is also the same.

Figure 3 Earnings and Talent



Notes: Earnings as a function of talent when talent is also an input into production. All individuals with talent $t < z$ are assigned to be workers and earn wages wt. Individuals with more managerial talent are managers and their pay is an increasing convex function in their talent.

The model is silent, however, on worker pay and firm size because for any given level of talent supplied to a firm, there are many combinations of differentially talented workers that can provide that total amount of talent. For example, a firm could have a small number of highly talented individuals or a large number of less talented individuals. Still, if there was a reason for the most talented workers to be assigned to the most talented managers and so on down the talent ladder until everyone was assigned, then there would be a wage-size premium. This kind of matching is referred to as *positive assortative matching*. One way to generate such a reason would be to make the production function highly complementary in the talent of the managers and workers. Kremer (1993) studies one such firm-level production function in which several tasks need to be performed simultaneously. If *any* of these tasks are performed unsuccessfully, then no output is produced. Talent improves the probability of success so this form of complementarity generates positive assortative matching and a positive wage-size premium.

Marginal Product

In this model all factors are paid their marginal product. This might not appear to be the case if one was to use the firm-level production function, $tf(l(t))$, to determine marginal product. However, that is not the right production function for determining the margin.

This model is an *assignment* model, and the right margin for determining marginal product is at the level of the production sector, which in this model is the entire economy. What this economy does is take as its inputs the numbers of people at each level of talent and then creates managers and workers, combines them into firms, and produces the output. Firms are really an intermediate good. The production function at the economy level is linear in these inputs so factors are paid their marginal products as in classical distribution theory.

More formally, the production sector solves

$$\max_{z,l(t)} \int_z^\infty tf(l(t))dH(t)$$

subject to

$$\int_z^\infty l(t)dH(t) \leq H(z). \tag{12}$$

The inputs into this production sector are the numbers of each type t. The supply of these inputs is, of course, the distribution $h(t)$. The economy is linear, or constant returns to scale, in the input. If w is the multiplier to (12) then the marginal product of $t \geq z$ is $tf(l(t)) - wl(t)$ and of $t < z$ is w. Managers are paid what is left over. Their pay is a residual, which is called a rent in the Ricardian tradition, but it is still a marginal product with respect to the production sector.

2. MULTIPLE-LEVEL HIERARCHIES

Extending the basic assignment model to hierarchies with more than two levels is conceptually straightforward, but it can be difficult analytically. In this section, a simple extension is provided.[4] The purpose is to generate a production hierarchy, like that illustrated in Figure 1, to introduce slightly more complicated managerial production functions, and to discuss relative pay levels between levels of management.

Production is limited to three-level hierarchies. As before, workers and managers jointly produce a good according to the production function $tf(l)$. However, this good is no longer final output but an intermediate good that is used by a second-level manager to produce the final output. Let the intermediate good be called m. Final output is $tg(m)$, where t is the talent of the second

[4] See Rosen (1992) for analysis of a problem where managers have a fixed span of control.

level manager. The intermediate goods are used to create a tractable example. The goal is to model firms organized like those illustrated in Figure 1.

The price of a unit of the intermediate good is λ and, as before, the price of labor is w. We need to solve for an assignment of labor to level-one managers, $l(t)$, an assignment of the intermediate good to level-two managers, $m(t)$, and cutoff values z_1 and z_2 that correspond to the cutoff talent levels between workers and level-one managers and between level-one managers and level-two managers, respectively.

The competitive equilibrium is set up so that the level-one manager hires the labor and creates the intermediate good, which he sells to the level-two managers. Despite this separation, we will interpret the level-one managers and workers who create the intermediate good for a level-two manager as being within the same firm. The problem can be formally set up in this way, but it is much more complicated to write down.

A level-two manager's problem is

$$\max_{m \geq 0} tg(m) - \lambda m.$$

The first-order condition is

$$tg'(m) - \lambda = 0. \tag{13}$$

The marginal level-two manager, z_1, must be indifferent to working as a level-one manager, that is,

$$z_2 g(m(z_2)) - \lambda m(z_2) = \lambda z_2 f(l(z_2)) - wl(z_2). \tag{14}$$

A level-one manager's problem is

$$\max_{l \geq 0} \lambda t f(l) - wl.$$

The first-order condition is

$$\lambda t f'(l) - w = 0. \tag{15}$$

The marginal level-one manager, z_2, must be indifferent to working as a worker,

$$\lambda z_1 f(l(z_1)) - wl(z_1) = w. \tag{16}$$

The final conditions for a competitive equilibrium are the resource constraints that the intermediate good used by level-two managers equals the intermediate good produced by combinations of workers and level-one managers

$$\int_{z_2}^{\infty} m(t) dH(t) \leq \int_{z_1}^{z_2} t f(l(t)) dH(t), \tag{17}$$

and that the labor used by level-one managers equals the labor supplied

$$\int_{z_1}^{z_2} l(t) dH(t) \leq H(z_1). \tag{18}$$

With these formulas, we can derive similar relationships to those derived earlier. For all managers at the same level in a hierarchy, the relationship of pay with respect to talent look similar to those studied earlier. The curvature of pay with respect to talent may differ between managers assigned to different levels of a hierarchy. This will depend on the properties of the production functions f and g. These functions need not be the same since managing at lower levels within a firm may be very different than managing at higher levels.

A commonly observed feature of managerial pay within a firm is that there is a much larger difference in pay between levels of a hierarchy than within given job classifications (Rosen 1992). In our simple model, such a feature could be obtained if level-one managers were appropriately assigned to level-two managers. But more generally, it would be desirable to formalize the model so that it mattered which level-one manager was assigned to which level-two manager. This brings up the issue of positive assortative matching raised earlier in our discussion of the wage-size premium. These effects would seem to matter for junior executives as well.

3. CONCLUSION

The hierarchy illustrated by Figure 1 captures some features of firms, but it really postulates that each branch within a firm operates separately from the others. Some parts of a firm operate in this way, but there are others, like personnel, maintenance, legal, and audit, that provide services to all parts of a firm. Their outputs are essentially intermediate inputs into the production of the final output by other parts of the firm. The literature rarely considers these features, yet if firms do anything special, it is that they do joint production of activities that are not as efficiently supplied on the market. It would seem desirable to introduce these features into some of the firm production functions that have been studied in the literature.

Finally, the model considered here is a static model. If taken to a dynamic environment, then the strategy would be to assume that managers and firms are reallocated each period to whichever type of firm the market sees fit to assign them. For some purposes, that abstraction is fine but it misses a very important part of the managerial assignment problem. Managers infrequently move between firms. Indeed, an important activity of a firm is to identify and develop managerial talent for future promotion through its ranks. The career ladder within a firm is a very important device for solving this assignment problem and to an extent it operates separately from the market. Understanding this mechanism might be quite important for understanding the internal distribution of pay between levels of a hierarchy within a firm. For some work along this

line see Lazear and Rosen (1981), Meyer (1994), or the papers surveyed in Rosen (1992).

REFERENCES

Barro, Jason R., and Robert J. Barro. 1990. "Pay, Performance and Turnover of Bank CEOs." *Journal of Labor Economics* 8: 448–81.

Brown, Charles, and James Medoff. 1989. "The Employer Size-Wage Effect." *Journal of Political Economy* 97 (October): 1027–59.

Cosh, Andrew. 1975. "The Renumeration of Chief Executives in the United Kingdom." *Economic Journal* 85 (June): 75–94.

Idson, Todd L., and Walter Y. Oi. 1999. "Workers are More Productive in Large Firms." *Papers and Proceedings of the American Economic Association* 89: 104–8.

Kostiuk, Peter F. 1989. "Firm Size and Executive Compensation." *Journal of Human Resources* 25: 90–105.

Kremer, Michael. 1993. "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics* 108: 551–75.

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89: 841–74.

Lucas, Robert E., Jr. 1978. "On the Size Distribution of Business Firms." *Bell Journal of Economics* 9 (Autumn): 508–23.

Meyer, Margaret A. 1994. "The Dynamics of Learning with Team Production: Implications for Task Assignment." *Quarterly Journal of Economics* 109 (November): 1157–84.

Murphy, Kevin J. 1999. "Executive Compensation." In *Handbook of Labor Economics* 3b, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier Science: 2485–563.

Oi, Walter Y. 1983. "Heterogeneous Firms and the Organization of Production." *Economic Inquiry* 21 (April): 147–71.

_____, and Todd L. Idson. 1999. "Firm Size and Wages." In *Handbook of Labor Economics* 3b, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier Science: 2165–214.

Roberts, David R. 1956. "A General Theory of Executive Compensation Based on Statistically Tested Propositions." *Quarterly Journal of Economics* 70 (May): 270–94.

Rodriguez, Santiago Budria, Javier Diaz-Gimenez, Vincenzo Quadrini, and Jose-Victor Rios-Rull. 2002. "Updated Facts on the U.S. Distributions of Earnings, Income, and Wealth." Federal Reserve Bank of Minneapolis *Quarterly Review* 26 (Summer): 2–35.

Rosen, Sherwin. 1982. "Authority, Control, and the Distribution of Earnings." *Bell Journal of Economics* 13: 311–23.

_____. 1992. "Contracts and the Market for Executives." In *Contract Economics*, edited by Lars Wein and Hans Wijkander. Cambridge, Mass. and Oxford: Blackwell: 181–211.

Simon, Herbert A., and Charles P. Bonini. 1958. "The Size Distribution of Business Firms." *American Economic Review* 48: 607–17.

Troske, Kenneth R. 1999. "Evidence on the Employer Size-Wage Premium from Worker-Establishment Matched Data." *Review of Economics and Statistics* 81 (February): 15–26.