Tom Humphrey: An Appreciation

John A. Weinberg

This issue of the *Economic Quarterly* marks the end of Tom Humphrey's tenure as editor. Tom, who is retiring at the end of 2004, took on the role of editor of the *Monthly Review* in 1975 and continued in that post as the publication evolved, first into the bimonthly *Economic Review*, and eventually into its current form as the *Economic Quarterly*. Over that time, Tom has guided to publication hundreds of articles by Department economists and visiting scholars. He has also found the time to write more than 70 articles for this publication, its predecessors, and the Bank's *Annual Report*, not to mention numerous articles for external publications. His editorial gaze has seen many changes in our publication, our department, and the economics profession over the last 30 years. Despite these changes, Tom's editorial guidance has provided a constancy to the quality of our publications, just as his own work has stressed a certain constancy to the ideas and debates that have engaged economists throughout the history of our profession.

Tom came to this Bank in 1970, an opportune time for a young monetary economist with an interest in the history of thought on the subject. Much of the profession's thinking on aggregate fluctuations and inflation was still dominated by the Keynesian view that emphasized movements in aggregate demand as the force driving real output. The taxonomy of "demand-pull" and "cost-push" inflation allowed for a variety of causes, mostly unrelated to the central bank's monetary policy. Empirical evidence on the Phillips curve trade-off seemed to suggest that price stability could only be had at an unacceptable cost of long-lasting, high unemployment. In the Keynesian view, money was just one part of a broad spectrum of liquid assets, the evolution of which further limited the central bank's ability to control inflation.

The Keynesian view, however, faced a challenge from the monetarist school of thought that placed the quantity of money at the center of the determination of the price level and other nominal variables. This quantity theory view, which presumed the long-run neutrality of money, allowed for short run real effects when money growth and inflation deviated from their expected rates. Indeed, under this view erratic monetary policy was seen as a primary cause of economic fluctuations. Where the Keynesian school saw causation running from prices to money, the monetarist view was the opposite; the central bank, through its control of the monetary base, could achieve price stability without a long-run inflation-unemployment tradeoff.

Of course, the 1970s experience of high unemployment and high inflation presented a particular challenge to the traditional Keynesian view. It all appeared to underscore the importance of expectations, a view that gave rise ultimately to the rational expectations revolution. This evolution of thinking played itself out in the profession throughout Tom's career at the Richmond Fed and in the *Economic Quarterly* and its predecessors under Tom's editorial eye. Along the way, he established himself as a leading scholar on the history of monetary thought.

Tom's notable contribution has been to make clear that the debates that took place during his career as a Federal Reserve economist were in fact not new to the 1970s, or even to the 20th century. Applying his knowledge of the history of monetary thought, Tom traced the debate from the mercantilist writers John Law and James Steuart to their classical quantity theory critics David Hume and others in the 18th century, to the Bullionist-Antibullionist controversy in the early 19th century, to the Currency School-Banking School debate in the middle decades of that century, to the mid-19th / early 20th century disputes between cost-pushers Thomas Tooke and James Laurence Laughlin and their opponents Knut Wicksell and Irving Fisher, and finally to the German hyperinflation debate of the early-to-mid-1920s. He showed that the debate keeps recycling because people forget the lessons of the past and because, for better or worse, politicians and the public have tended to believe that central banks have the power to boost output, employment, and growth permanently. The result is that Keynesian ideas and their antecedents gained currency when unemployment was the main concern, just as monetarist ideas tended to reign when price stability was the dominant problem.

This historical perspective makes it clear that the central dimensions along which people have thought and debated about monetary policy and inflation have always remained essentially the same, even as economic thinking and methodology have evolved. It also shows that intellectual debate can be affected by political forces and the fluctuating degree of social concern for different problems. This last point provides an important cautionary tale for current and future economists and policymakers. Indeed, Tom's article in this issue shows that the recurrence of debates and the interaction of political forces with intellectual discourse are not unique to monetary concerns. Similar patterns can be seen in the history of thinking on such questions as the effects of technology on labor.

We at the Richmond Fed have gained much from our association with Tom Humphrey. We have benefited both from the red pen he has wielded to make our papers more readable and, more importantly, from the long-view perspective he has brought to our thinking about economics and monetary policy. For all this we thank him, and we offer him our best wishes. We won't say good-bye, though, as I'm sure we'll be hearing from him again.

Ricardo versus Wicksell on Job Losses and Technological Change

Thomas M. Humphrey

Are technological innovations net destroyers of jobs? Many think so and point to the information technology (IT) revolution and its progeny, the offshore outsourcing of service activities, as prime current examples. Here the simultaneous advent of (1) undersea installation of mega-bandwidth fiber optic cable allowing virtually costless transmission and storage of data, (2) global spread of personal computers, and (3) standardization of software applications allegedly have made it profitable to export abroad service functions once performed in the United States, thereby throwing Americans out of work (Friedman 2004).

Others, however, disagree and contend that new technology, including outsourcing, creates at least as many jobs as it destroys (Drezner 2004). It lowers costs, cheapens prices, stimulates demand, boosts output, and provides new employment opportunities. Historical experience, these observers contend, reveals such to be the case. Since the start of the industrial revolution, the number of jobs has grown as fast as the level of technology. Were the opposite true and innovation continually to displace workers, firms employing ever-advancing technology requiring ever-fewer hands to operate it eventually would produce the entire GDP with a labor force of one person. That outcome, the observers note, has not happened.

Concern with the jobs-versus-technology issue is hardly new—think of Karel Capek's famous 1920 play *R.U.R.* Its plot has factory automation permanently replacing human workers with robots, a possibility Paul Samuelson modeled mathematically in 1988. Samuelson and a few others aside,

For valuable comments and suggestions the author is grateful to his Richmond Fed colleagues Yongsung Chang, Huberto Ennis, John Hejkal, Roy Webb, and John Weinberg. The views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

however, commentators all too often have addressed the jobs-innovation question in an ad hoc, anecdotal manner conducive to selective reasoning, ambiguous conclusions, and emotional rather than rational responses. Too rarely has a coherent analytical framework capable of yielding dispassionate, clear-cut answers disciplined the discussion.[1] This article traces the first attempts to overcome this deficiency and to resolve the issue of technology's effect on jobs theoretically with the aid of a rigorous analytical model.

The model in question is David Ricardo's famous machinery example. It has capital-embodied innovation converting the wage-fund stock of consumable goods that sustains workers over the production period into fixed machinery that cannot sustain them. The result is to lower permanently the demand for labor, the number of jobs, and the level of output. Reversing his original position that innovation benefits all, Ricardo in 1821 constructed his model to demonstrate that workers have much to fear from technical change. "All I wish to prove," he said, "is, that the discovery and use of machinery may be attended with a diminution of gross produce: and whenever that is the case, it will be injurious to the labouring class, as some of their number will be thrown out of employment, and population will become redundant, compared with the funds which are to employ it" (Ricardo [1821] 1951, 390). Almost one hundred years later, Knut Wicksell deployed essentially the same model, albeit with a different assumed coefficient of elasticity of labor supply and a different theory of labor demand, to argue that Ricardo's predictions were flawed and that jobs and real output need not be lost to technological progress.

Wicksell's contribution was to refurbish Ricardo's model with new ideas emerging from the celebrated marginal revolution in economic theory that occurred in the 1870s, 80s, and 90s. He replaced Ricardo's classical wage-fund theory of labor demand with a neoclassical marginal productivity explanation. Likewise, he substituted a fixed-factor-endowment interpretation of labor supply for Ricardo's old-fashioned subsistence-wage approach. These improvements rendered the machinery model amenable to marginal analysis, thereby bringing it closer to modern theorizing on the jobs-innovation issue. They enabled Wicksell to challenge Ricardo's melancholy predictions within the framework of his own rehabilitated model. In short, in their respective readings of the model, Ricardo was the pessimist and Wicksell the optimist as far as innovation's impact on jobs and the well-being of labor were concerned.

Among the few who have commented extensively on these opposing outlooks is Paul Samuelson. In his 1989 *Scandinavian Journal of Economics* article "Ricardo Was Right!" Samuelson writes that "in the famous suit K.

---

[1] Exceptions include research on the jobs-innovations question recently initiated by Gali (1999), Basu, Fernald, and Kimball (1998), and Francis and Ramey (2002). These studies use formal modeling to conclude that technical progress reduces employment in the short run, but not the long. Job loss is transitory, not permanent.

Wicksell vs. D. Ricardo—in which Knut Wicksell denied that a viable invention could reduce aggregate output [and jobs]," a "modern judge must rule . . . against the plaintiff. My title therefore could have been . . . Wicksell was wrong!" (Samuelson 1989, 47–8).

What follows takes issue with Samuelson, arguing, contrary to him, that while both men were right in theory—that is, within the context of their particular variants of the hypothetical machinery model—only Wicksell was right in practice. Realizations match the predictions emerging from his reading of the model, but not from Ricardo's. True, with respect to theory, both economists employed impeccable logic and valid reasoning in constructing and manipulating their versions of the model to grind out the solutions they did. Their versions left nothing to be desired on internal consistency grounds. With respect to practice, however, only Wicksell's optimistic predictions have stood the test of time. He rightly foresaw that output and jobs would expand with labor-saving technological progress. He likewise predicted that labor-neutral and labor-using innovations would boost real wages as well. History has confirmed his predictions and falsified Ricardo's. It has revealed his version of the model to be the more realistic of the two.

Besides providing historical perspective on the outsourcing issue, the Ricardo-Wicksell controversy is of interest for at least six other reasons. First, it shows how the same analytical model can, with different assumptions about the values of its coefficients and the shapes of its functions, yield opposite results. In Ricardo's machinery model where labor demand is key, technology essentially enters the labor demand function as a variable bearing a negative sign. It thereby ensures that innovation harms, rather than helps, labor. A positively signed technology variable, Wicksell noted, would reverse that result. So too would a negatively signed variable if offset or negated by compensating profit-sharing schemes. Another key is the assumed slope of the labor supply curve. Depending on that slope, labor-saving innovation either shrinks or expands real output just as it destroys or preserves jobs.

Second, in spotlighting these polar results, the controversy shows how a single model under alternative parameter settings can, when used to organize discussion, encompass the entire range of opinion on the issue of jobs and innovations. Whether one believes innovations on balance are job destroyers, job creators, or merely job preservers, one's stance on this issue (albeit not necessarily one's acceptance of the model) falls somewhere between the extremes of Ricardo and Wicksell.

Third, the controversy shows that even the greatest economists' most cherished beliefs are not fixed and immutable. Ricardo recanted his long-held position that technical progress is Pareto-improving (that is, benefits all parties and harms none) only when he became convinced that he had been in error and that innovation could hurt labor even while it profited capital.

Fourth, the controversy shows that mainstream economists, notwithstanding their theoretical differences and social sympathies, tend to favor, on efficiency grounds, public policies conducive to technical progress. With respect to innovation, both Ricardo and Wicksell recommended that governments refrain either from suppressing or discouraging it regardless of whether it destroys jobs (Ricardo) or preserves them (Wicksell). Ricardo in particular argued that anti-innovation policy magnifies job destruction and intensifies harm done to labor. That is to say, he thought that while innovation hurts workers, attempts to prevent it only make matters worse. And Wicksell, though a redistributionist, welcomed pro-innovation policies. They would, he believed, help maximize the size of the pie—gross product—to be shared.

Fifth, the controversy shows how the study of a practical social issue such as technology's effect on jobs spurs new concepts and ideas that advance economic science. Here, in addition to the machinery model itself, the new concepts include Wicksell's distinction between labor-saving, labor-using, and labor-neutral innovations, namely those that lower, raise, or leave unchanged, respectively, labor's marginal productivity relative to capital's. Still another novel idea was the compensation principle according to which winners in an economic change compensate losers so as to make both groups better off. Wicksell devised this concept to argue that capitalists could profitably bribe workers to accept technological innovations that otherwise would hurt them.

Sixth and most of all, the controversy serves as a cautionary tale. Economists (not to mention general observers) have been discussing the effects of innovations on labor for a long time. The analysis has always been fraught with pitfalls, so one should be careful in jumping to conclusions, especially regarding policy responses. A common pitfall (albeit one largely avoided by Ricardo and Wicksell) is failure to distinguish between immediate and longer-run effects of innovation. Initially, technical progress is quite likely to hurt groups of workers possessing specific acquired skills and abilities. One must weigh this short-run cost of innovation against potential long-run benefits. In the long-run, workers will invest in acquiring a different set of skills that will enable them to operate the new technology. But this adjustment process may involve a painful transition period, and society may wish to ease the pain of those adversely affected during the transition. Yet because pain provides an incentive to undertake the necessary changes, too much assistance may delay the adjustment for an inefficiently long time.

1. THE MACHINERY QUESTION

Fears of job destruction through new technology antedate both today's outsourcing scare and David Ricardo. Think of the manuscript copyists whose skills Johannes Gutenberg's 1436 invention of the printing press rendered obsolete. Later, like their medieval counterparts, 18th- and 19th-century observers watching the mechanization of textile and other key manufactures also saw machinery as the source of technological unemployment (Rashid 1987; Berg 1980). Workers and their advocates then posed the celebrated *machinery question*: Could new machines embodying advanced technology permanently destroy jobs? Like modern economists, 18th-century economists generally answered in the negative, and with the same reasoning, too. New machines lower production costs. Lower costs mean cheaper prices. Cheaper prices extend the market. They stimulate demand for consumption goods and make it profitable for firms to expand output to satisfy the demand. Since extra output requires hands to produce it, increased production absorbs the initially laid-off workers and other workers as well. Technical advance, in addition to benefiting workers by giving them lower prices, begets more jobs than it destroys. Josiah Tucker said it all in his 1757 explanation of the effects of machinery:

> What is the Consequence of this Abridgment of Labour, both regarding the Price of the Goods, and the Number of Persons employed? The Answer is very short and full, *viz.* That the Price of Goods is thereby prodigiously lowered from what otherwise it must have been; and that a much greater Number of Hands are employed. . . .
>
> And the first Step is that Cheapness, *ceteris paribus* is an inducement to buy—and that many Buyers cause a great Demand—and that a great Demand brings on a great Consumption; which great Consumption must necessarily employ a vast Variety of Hands, whether the original Material is considered, or the Number and Repair of Machines, or the Materials out of which those Machines are made, or the Persons necessarily employed in tending upon and conducting them: Not to mention those Branches of the Manufacture, Package, Porterage, Stationary Articles, and Book-keeping, &c. &c. which must inevitably be performed by human Labour. . . .
>
> That System of Machines, which so greatly reduces the Price of Labour, as to enable the Generality of a People to become Purchasers of the Goods, will in the End, though not immediately, employ more Hands in the Manufacture, than could possibly have found Employment, had no such machines been invented (Tucker [1757] 1931, 241–2, quoted in Rashid 1987, 265).

Other classical economists, including Adam Smith, Jean Baptiste Say, and most notably David Ricardo, echoed this optimistic view. Ricardo, for example, wrote that mechanization ("a general good") benefits all social classes including workers, capitalists, and landlords alike. Mechanization conserves scarce resources, improves efficiency, increases output, and lowers production costs. The resulting fall in prices gives all consumers more purchasing power to spend on an augmented bundle of goods. In this way "the labouring class . . . equally with the other classes, participate[s] in the . . . general cheapness of commodities arising from the use of machinery" (Ricardo [1821] 1951, 388).

2. LABOR UNREST AND RICARDO'S ABOUT-FACE

Ricardo, in other words, initially believed that mechanization benefited workers by giving them more and cheaper goods. And it did so without destroying jobs or lowering money wages. "I thought that no reduction of wages would take place," he wrote, "because the capitalist would have the power of demanding and employing the same quantity of labour as before, although he might be under the necessity of employing it in the production of a new, or at any rate of a different commodity" (Ricardo [1821] 1951, 387). Cheaper prices at accustomed money wages together with availability of jobs in the innovating and non-innovating sectors of the economy—what more could workers want? They should welcome mechanization, not oppose it. That was certainly Ricardo's initial expectation.

Then came episodes of labor unrest—the violent strikes, riots, protests, and machine-breaking of 1811–21—that overlapped with periods of high unemployment in the post-Napoleonic War years of 1815–30. Famous among the rioters of this time were organized bands of English handicrafters know as Luddites. Taking their name from Ned Ludd, an apocryphal 18th-century Leicestershire handloom weaver who supposedly destroyed two stocking frames in a fit of rage, the Luddites conspired to smash the textile and cloth-finishing machines that they thought were threatening their jobs. Observing these uprisings, Ricardo changed his views radically in the famous Chapter 31 "On Machinery," which he added to the third (1821) edition of his *Principles of Political Economy and Taxation*.

In that chapter, which Samuelson (1988, 274) calls the best in the book, Ricardo took labor agitation seriously. He had always modeled agents as rational maximizers acting in their own self-interest (Maital and Haswell 1977, 365). Might not workers, as such agents, have a legitimate case against machines? Might not machines be inimical to their interests as they themselves maintained? Answering in the positive, Ricardo proceeded to construct a formal model (with a numerical example as its core) to demonstrate that "the opinion entertained by the labouring class, that the employment of machinery is frequently detrimental to their interests, is not founded on prejudice and error, but is conformable to the correct principles of political economy" (Ricardo, [1821] 1951, 392).

3. OVERVIEW OF THE MODEL

Ricardo's general equilibrium model says that when a capitalist installs new labor-saving technology in the form of a machine, that same capitalist permanently displaces labor and renders it superfluous. That is the initial effect. The intermediate, or transition, effects come when the redundant workers, in an effort to regain their lost jobs, bid down the wage rate. Since in Ricardo's model, as in the labor-market models of most classical economists, the initial wage rate already is at the equilibrium (or Malthusian minimum subsistence) level where the work force barely maintains its size with neither increase nor diminution, the fall in wages below that level means that fewer workers can survive and indeed must die off (see Samuelson 1994, 621). They continue to die off in sufficient numbers until the wage rate returns to its subsistence equilibrium. In that new, long-run equilibrium, the output-reducing effect of labor force diminution dominates the output-raising effect of the machine's greater efficiency so that gross output falls. Final steady-state equilibrium features these conditions: smaller output, fewer jobs, fewer workers to fill those jobs, subsistence wages, and raised profits (a necessary condition for the capitalist to install new machinery in the first place).[2]

It is hard to avoid noticing the model's current relevance. Replace the word "machinery" with "outsourcing" and downplay the Malthusian overtones. What you get is the typical current complaint that technical progress in the form of offshore outsourcing hurts labor at the same time it helps capital.

4. RICARDO'S EXAMPLE

The model itself has a group of laborers working for a single capitalist farmer who represents the entire productive sector of the economy. The capitalist initially has a total capital stock of £20,000, of which £7,000 is *fixed capital* (buildings, equipment, and the like), and £13,000 is *circulating capital* (stores of food and necessaries used to provision, or grubstake, labor over the period of production and thus the wherewithal to employ, or demand, workers). The importance to the model of circulating capital cannot be overstressed. It and it alone constitutes the capitalist's ability to employ workers. Nothing else, neither the lower prices and higher profits that innovation yields, nor the increased spending spurred by them, can affect employment in the model. To Ricardo, circulating capital, rather than demand for commodities, constitutes demand for labor. Anything that shrinks the stock of such capital automatically shrinks labor demand. No compensating mechanism such as the previously mentioned price and profit effects leading to increased demand for goods can offset, or negate, the resulting adverse employment effects of reductions in the stock of circulating capital.

Ricardo makes the foregoing point exceedingly clear in his example. He begins by assuming that year after year in stationary equilibrium the capitalist and his workers produce annual output worth £15,000. Of this sum, £13,000

[2] In his Chapter 31 Ricardo always speaks of the new machine as raising profit, or net income. Yet in his model and numerical example, profit remains constant. There is no inconsistency here. Ricardo recognized that profit must rise by some positive amount, however small, call it epsilon, to motivate the capitalist to invest in the risky new machine. To simplify his model, however, he let epsilon assume a limiting value of zero. Nothing would have changed if he had assigned it a positive value. See Barkai 1986, 599-600, footnote 2.

goes to replace the circulating capital stocks of food and necessaries workers have consumed over the year, and £2,000 goes to the capitalist as profit (a 10.0 percent profit rate) to reward him for the use of his capital. Ricardo assumes that the capitalist consumes, rather than invests, his profit such that no capital growth occurs.

Things change when the capitalist decides on profit grounds to divert half his labor force from the production of food and necessaries to the fabrication of a new machine. Since the workers reassigned to machine-building produce no food and fiber, farm output is halved to £7,500 while fixed capital rises from £7000 to £14,500 by the £7,500 value of the new machine. The machine, of course, is counted in final output during the time of its construction. But it is not so counted afterward when, its fabrication completed, it assumes its place in the economy's stock of fixed capital assets and production reverts to farm product only. When the capitalist extracts his £2,000 profit (still 10.0 percent of his capital stock) from the £7,500 value of farm output, barely £5,500 worth remains to provide for the maintenance of labor in the following year. In other words, circulating capital, or means of employing labor, falls from £13,000 to £5,500. Given that circulating capital constitutes demand for labor in Ricardo's model, the capitalist can now employ but 42.3 percent, or 5,500/13,000, of the labor he employed before to produce a gross output of half its former size. In short, switching labor from food production to machine installation permanently reduces the fund available to grubstake and therefore to hire workers. "There will," Ricardo gloomily concludes, "necessarily be a diminution in the demand for labour, population will become redundant, and the situation of the labouring class will be that of distress and poverty" (Ricardo [1821] 1951, 390).

Attempting to regain their lost jobs, the redundant workers put downward pressure on the real wage rate forcing it to drop below the minimum subsistence level, which the Malthusian iron law of wages—represented in Ricardo's model by a horizontal labor supply curve—dictates as the equilibrium wage. The resulting starvation of workers shrinks the population, the labor force, and with it the gross product until the real wage returns to its subsistence level.

Here then is the second crucial component of Ricardo's machinery model, namely the iron law of wages. Developed by Richard Cantillon, Adam Smith, and above all by Thomas Malthus, it says that population and labor force numbers respond to gaps between actual and subsistence wages. Their response together with diminishing returns to extra doses of labor applied to the fixed factor land keeps wages gravitating to subsistence. Thus below-subsistence wages lead to starvation, high death rates, low birth rates, and population and labor force decline. With fewer workers tilling the fixed amount of land, the land-to-man ratio rises, which means that each laborer has more land to work with and so experiences a rise in his productivity. Real wages rise with productivity until both return to the subsistence level where population shrinkage ceases and the labor force stabilizes in size. Conversely, above-subsistence wages encourage population growth and the crowding of more workers on the fixed land. Each worker has less land to work with and so experiences a fall in his productivity and real wage, both of which converge to the subsistence equilibrium where population growth ceases and the labor force stabilizes. In short, diminishing returns together with the feedback of wage deviations from subsistence on population growth operate to keep wages at subsistence. Operating through these channels, mechanization in Ricardo's model not only displaces workers but kills them off as well. Workers indeed have a legitimate case against machinery, or more precisely, against the ultra labor-saving bias of the technical progress embodied therein.

5. REACTION TO THE MODEL

Ricardo's demonstration appalled his classical contemporaries who found it incompatible with the rest of his work. Typical was the reaction of John Ramsay McCulloch who complained that Ricardo's machinery chapter ruined the book (see St. Clair [1957] 1965, 234, 237). How could Ricardo, creator of the comparative-advantage theory of gains from trade, contend that technological innovation, a key source of comparative advantage, hurts labor? How could he be so inconsistent? "[N]othing can be more injurious," wrote McCulloch to Ricardo on June 5, 1821, "than to see an Economist of the highest reputation strenuously defending one set of opinions one day, and unconditionally surrendering them the next" (McCulloch [1821] 1951, 382). "I will take my stand," declared McCulloch, "with the Mr. Ricardo of the first not of the third edition [of the *Principles*]" (385).

Ricardo's peers also feared his analysis might discredit the free-market precepts of classical economics, not to mention the aid and comfort it would provide to anti-market reformers. "[A]ll those who raise a yell against the extension of machinery," wrote McCulloch to Ricardo, "will fortify themselves by your authority" and claim that "the laws against the Luddites are a disgrace to the Statute book" (384–5).

6. RICARDO'S QUALIFICATIONS

Ricardo himself seemed sufficiently uncomfortable with his theoretical demonstration to express reservations about its practical relevance. At the end of his chapter he noted that capitalists often mechanize their operations gradually instead of suddenly, thus allowing time for smoother adjustment. He also noted that machine installation may be a manifestation of saving-financed *growth* in capital rather than of conversion of the circulating-into-fixed components of a capital stock of constant size. With no conversion, or shrinkage, of circulating capital there is no displacement of labor. Jobs are not destroyed.

Indeed, circulating capital (and jobs) conceivably might expand together with fixed capital. Tracing a causal chain from mechanization to falling production costs to cheaper product prices to rises in the real purchasing power of nominal profit incomes, Ricardo suggested that such increased real profit incomes could generate the saving from which investment in circulating, as well as fixed, capital would come. Alternatively, capitalists might spend their profit increases on the hiring of menial servants or on the purchase of luxury consumption goods. These expenditures would create new demands for labor. But such demands, Ricardo realized, could reabsorb but a fraction of the workers displaced by wage-fund contractions that exceeded profit expansions in size. He further pointed out that, in the context of an expanding population, mechanization, far from occurring autonomously, is often induced by rising money wage rates relative to the cost of machines. (The money wage hikes are, of course, necessary to maintain real wages at subsistence in the face of rising food prices caused by diminishing returns as the growing population resorts to more intensive cultivation of the fixed land). Capitalists then attempt to economize on costly labor by substituting relatively cheap machines for it. This point, however, refers to pure capital-labor substitution under given technology. It does not refer to technological change and so hardly qualifies as an exception to Ricardo's example.

Most of all, Ricardo warned of the futility and harmfulness of limiting or discouraging the introduction of new machines in a world where foreign competitors would introduce them anyway. By lowering the return on domestic relative to foreign capital, such restrictions would spur the export of capital, leading to even less demand for labor at home. In short, whereas conversion of circulating capital into machinery lowers domestic labor demand, capital exported abroad annihilates the demand altogether (Ricardo [1821] 1951, 397). Another point recognized by Ricardo is that the banning of machines makes a nation less efficient than its trade partners so that it obtains fewer labor hours' worth of imports per each labor hour's worth of exports given up. In other words, rejection of machinery turns the country's double factoral terms of trade against it (O'Brien 1975, 226).

The upshot of these considerations is that no restrictions should be placed on the introduction and use of machines. As Ricardo put it, "he would not tolerate any law to prevent the use of machinery. The question was,—if they gave up a system which enabled them to undersell in the foreign market, would other nations refrain from pursuing it? Certainly not. They were therefore bound, for their own interest, to continue it" ([1823] 1951 303). Ricardo's disapproval of anti-machinery policies aimed at preserving jobs indicates that were he alive today he would likewise oppose all restrictions on offshore outsourcing.

Nevertheless, Ricardo's reservations and doubts about his model evidently were not so serious as to invalidate his conclusion that capital-embodied innovations may harm labor. Thus when speaking before the House of Commons on May 30, 1823, he abandoned all mention of doubts and reservations and instead firmly reiterated “his proposition . . . that the use of machinery was prejudicial . . . to the working classes generally. It was the means of throwing additional labour into the market, and thus the demand for labour, generally, was diminished” (Ricardo [1823] 1951, 303).

7. MCCULLOCH ON THE MODEL

Ricardo’s model was a very special one with several curious features. Job destruction results solely from the conversion of circulating into fixed capital. The introduction of machinery leaves the total stock of capital (albeit not its composition) unchanged. Fixed capital bears no depreciation charges, implying that it has infinite life. Wages cannot fall permanently below their Malthusian minimum subsistence limit. Profits, too, cannot fall and indeed must rise by some amount, however small—call it epsilon—to induce innovation. (Here Ricardo created unnecessary confusion by having epsilon assume a limiting value of zero so that profits apparently remain unchanged.) Output falls.

Classical economist John Ramsay McCulloch, who as we have seen objected to Ricardo’s analysis, focused on some of these peculiarities (see O’Brien 1975, 227–28). He argued that displaced workers would find jobs in making machines, including new machines to replace worn-out ones. On this point he disagreed with Ricardo who, thinking that replacement was of little importance, modeled machines as lasting forever and so incurring no depreciation.

Regarding profits, McCulloch claimed that the capitalist would require a rise (rather than the apparent zero change) in them to compensate for the uncertainty of investing in untried new technology. Without additional profits, the capitalist would have no incentive to install the risky new machine. This criticism, too, missed its mark because, as previously mentioned, Ricardo agreed that profit rises were necessary. The zero rises in his model were but a proxy for and lower limit to the required positive rises.

McCulloch concentrated the bulk of his attention on the model’s output result. Machines, he said, raise, not lower, output. They do so through a causal chain running from lower production cost to lower product prices to increased consumer demand in response to cheaper prices, and thence to the profitability of producing extra output (and hiring extra hands) to satisfy that demand. Replacing Ricardo’s concept of circulating capital as demand for labor with the alternative notion of demand for goods as demand for labor, McCulloch argued as follows (see O’Brien 1975, 227–8): If product demand is unitary elastic such that price falls induce proportionate rises in quantity demanded, then labor re-absorption is complete. The machine-installing sector rehires the displaced workers. Similarly, if product demand is elastic such that price falls induce more-than-proportionate rises in quantity demanded, then labor re-absorption is more than complete. The sector rehires more workers than it laid off.

Conversely, if product demand is inelastic such that price falls induce a less-than-proportionate rise in quantity demanded, then labor re-absorption is incomplete. Even so, the price cuts in this last case still leave consumers with more purchasing power to spend on other goods, leading to increased hiring of workers to produce those other goods.

Of course, consumers may choose not to spend all the extra purchasing power that price cuts bring. If so, those consumers save. The saving, upon its deposit in banks, is loaned out to capitalists to finance investment in new capital goods. Demand for those goods and the labor to produce them rise.

Finally, if capitalists fail to pass cost reductions on into price reductions, the resulting extra profits they receive are used either to increase their own consumption or their purchases of investment goods. Either way, demand for goods and, in turn, for labor, rises, and displaced workers are reabsorbed. To be sure, re-absorption implies that workers must acquire new skills to enable them to adapt to the better technology. Likewise, it implies that they must learn new trades so that they can occupy new jobs to replace the old ones lost to mechanization. These adjustments may involve pain. But such distress is a reasonable cost to pay considering the gains to be made. Here in a nutshell was McCulloch's elaboration of Tucker's earlier analysis.

8. WICKSELL'S CRITIQUE

McCulloch's 19th-century critique of the machinery model was quite perceptive. But it remained for the Swedish neoclassical economist Knut Wicksell, writing a hundred years after Ricardo, to deliver the definitive critique. In his 1901 *Lectures on Political Economy* and his 1923 manuscript "Ricardo on Machinery and the Present Unemployment"—a manuscript that *Economic Journal* editor John Maynard Keynes rejected for publication in 1923 and that Lars Jonung shepherded into print in that same journal in March 1981—Wicksell argued that Ricardo had it all wrong. The latter's long-run steady state equilibrium had no room for lower wages and the resulting re-absorption of displaced labor. Nor did it have room for the increased output that the fully employed labor force equipped with improved technology could produce. Using the classical assumption that the long run equilibrium wage rate is fixed exogenously at minimum subsistence, Ricardo was denied the neoclassical insight that the equilibrium wage rate is instead determined endogenously by labor's marginal productivity at full employment. Deprived of that understanding, he failed to see that innovations do not reduce production, but rather augment it. In short, there is no floor to equilibrium wages. The labor supply curve is vertical rather than horizontal. The demand for labor determines the wage rate rather than the level of employment. Labor's marginal productivity, not the stock of circulating capital, constitutes labor demand. Labor-saving "machinery"—a word Wicksell uses to denote disembodied technical progress rather than fixed-capital-embodied technical progress in Ricardo's sense of the word—drives that demand through its influence on worker marginal productivity. If innovation is biased against labor, marginal product falls although gross product rises.

Incorporating these changes into Ricardo's model ensures that neither jobs nor output are lost to the machine, that is, to innovation. On the contrary, Wicksell ([1923] 1981, 200, 203) thought that with a sufficient drop in wages, all the workers displaced by the machine would be rehired and, with the aid of the new technology, would produce more output than they did before. The innovation-induced fall in wages, variously estimated by him ([1901] 1934, 138; [1923] 1981, 202) to be between 10.0 percent and 1.0 percent in size, was absolutely crucial.[3] It ensured continual equality between the wage rate and labor's lowered marginal productivity, this equality being a necessary condition for output to reach its maximum allowed by the innovation.

9. REDISTRIBUTION SCHEMES, OR PARETO OPTIMAL BRIBES

As for Ricardo's claim that the lower wages would invariably decimate labor through starvation, Wicksell ([1923] 1981, 204–5) denied it. True, Wicksell recognized that the post-innovation reduction in wages necessary to clear the labor market and to allow output to reach its maximum level makes workers worse off. Their jobs are preserved, but at dwindled pay. And he also realized that if the resulting reduced wage is below subsistence, the labor force would have to undergo Malthusian shrinkage just as in Ricardo's case. But Wicksell insisted that this outcome was not inevitable. Distinguishing between technical conditions necessary for maximum production on the one hand versus distributional requirements of maximum social welfare or satisfaction on the other, he noted that lower wages (equaling as they do labor's marginal product at full employment) satisfy the first set of conditions but not necessarily the second. Maximizing satisfaction requires that everyone's welfare, notably labor's, be improved. To obtain that maximum, the government, Wicksell said, must supplement wages with welfare relief payments sufficient to maintain workers at the subsistence standard of living or above.

[3] These wage falls are relatively small. In later writings, however, Wicksell entertained the notion that wages might have to fall to zero or close to it to clear the labor market following the introduction of new labor-saving technology (see Boianovsky and Hagemann 2003, 24–5). But he seems to have regarded such extreme wage falls as purely hypothetical. In his careful and detailed 1901 and 1923 critiques of Ricardo's model, he posits small, not large, reductions in wages.

Of course these relief payments ultimately would come from taxes on profits. Even so, profits net of tax would be higher with the machine than without it, thanks to the machine's capacity to raise the profit rate. Nor would the profit tax itself discourage production and so dry up the very proceeds that constitute the source of relief payments. No such disincentive effects could wreck the scheme; for Wicksell ([1896] 1958, 256–7) elsewhere had used a model of imperfect competition to prove that a lump sum profit tax, being independent of the level of output, is like a fixed cost. It does not affect producers' marginal cost and marginal revenue schedules and so leaves the profit-maximizing level of output unchanged. The tax, in other words, shifts the hump- or inverted U-shaped profit function downward by the amount of the levy. But it does not change the output level where the function reaches its peak or maximum value. Desiring to reach that peak, maximizing capitalists might complain about the tax. Still, they would be doing the best for themselves by maintaining the level of production rather than by curtailing it.[4]

The upshot was that society could devise a post-innovation tax-transfer scheme that would leave capitalists better off and workers at least no worse off than before. In this way, the fruits of technical progress could be shared by all. Via income transfers, capitalists could effectively bribe workers to accept those innovations that threatened to lower labor's marginal product and so real wages.

Wicksell, of course, realized that not all innovations would lower labor's marginal productivity and real wages. On the contrary, he thought that some, perhaps most, innovations would raise those productivity and real wage variables instead of lowering them. "[T]he great majority of inventions and technical improvements," he wrote, "tend to increase the marginal productivity of both labour and land, together with their share in the product" (Wicksell [1901] 1934, 143). For such labor-using innovations, transfers and bribes would be unnecessary since workers would benefit anyway.

And although he excluded capital accumulation from his disembodied-technical-change version of Ricardo's model, he elsewhere stressed the modern view that such accumulation creates jobs while raising labor's marginal productivity and real wages. "[T]he capitalist saver," he wrote, "is thus, fundamentally, the friend of labour, though the technical inventor is not infrequently its enemy" (Wicksell [1901] 1934, 164). It follows that innovation accompanied by, or embodied in, new capital requires no income transfers to benefit labor.

---

[4] Wicksell failed to note that, under certain circumstances, taxing the profit that innovation yields may dry up the future supply of that activity. If profit includes a cost payment, or normal rate of return, necessary to coax forth innovation, then removing that return would destroy the incentive to innovate. In other words, if the supply of innovation is elastic with respect to profit, taxing profit will reduce the quantity of innovation supplied.

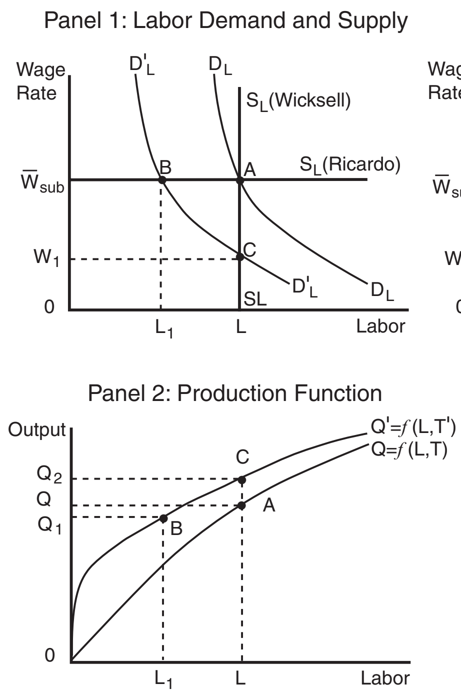
To summarize, Wicksell disputed Ricardo's ideas of (1) a lower bound, or floor, to wages, (2) a post-innovation decline in output, jobs, and the labor force, and (3) the absence of tax-transfer profit-sharing schemes.[5] Discarding these notions, Wicksell showed that the freedom of wages to fall to market-clearing levels where labor receives its marginal product promotes the re-hiring of displaced workers. Equipped with the improved technology, these workers together with their already employed counterparts produce additional output. Redistribution mechanisms then allow labor to share the extra output with capital so that both parties enjoy higher incomes after the innovation than before it. In Wicksell's own words, "the only completely rational way to achieve the largest possible production [is] to allow all production factors, including labour, to find their equilibrium positions unhindered, under free competition, however low they may be, but at the same time to discard resolutely the principle that the worker's only source of income is his wages. He, like every other citizen, ought rather to be entitled to a certain share of the earnings of the society's natural resources, capital, and (where they cannot be avoided) monopolies" ([1924] 1969, 257).

10. DIAGRAMMATIC ANALYSIS

Geometrical diagrams illustrate Ricardo's and Wicksell's cases (see Figure 1, suggested by Samuelson [1989], 53). Panel 1 shows how Ricardo's capitalist—when converting circulating (wage fund) capital into fixed capital via the installation of the machine—causes the labor demand curve to shift downward and to the left. The shifted demand curve intersects the horizontal labor supply curve, the height of which is fixed by the Malthusian minimum subsistence wage rate, at new equilibrium *B*. There the labor force is halved. Despite the machine's effect in enhancing efficiency, shown by the upward shift in Panel 2's aggregate production function, fewer workers spell less output so that gross product falls. At the same time, the innovation, by shifting outward Panel 3's factor price frontier, or menu of alternative maximum wage rate-profit rate combinations, reveals that the rate of profit rises from *A* to *B*. The end result is that output is down, jobs are down, the labor force is down, the wage rate is unchanged, labor income (wage rate times labor force) is down, and the profit rate and profit income (profit rate times total capital, a constant) are both up.

[5] Hansson (1983, 55) argues that Wicksell's criticism of the lack of a tax-transfer redistribution mechanism in Ricardo's model is misguided for two reasons. First, no such mechanism existed in Ricardo's time when welfare aid to unemployed workers, such as it was, consisted of poor relief and charity. Second, 19th-century English capitalists operated in a political system that catered to their interests. Given this state of affairs, they would have no incentive to depart from the Ricardian equilibrium and agree to income transfers. No pro-labor social and legal sanctions were in place to make them to do so.

Figure 1 Wicksell vs. Ricardo on Technological Innovation and Job Loss



Ricardo: Conversion of circulating into fixed capital via the installation of a machine shifts down the labor demand curve in Panel 1. At the same time, the advanced technology embodied in the machine shifts up Panel 2's production function and Panel 3's factor price frontier. The horizontal labor supply curve in Panel 1 dictates that equilibrium move from *A* to *B* in all panels. Output, jobs, and the labor force drop. Wages remain at subsistence. Profits rise.

Wicksell: Panel 1's vertical labor supply curve dictates that innovation moves equilibrium from *A* to *C* in all panels. Jobs and the labor force remain unchanged. Output and profits rise. But wages fall below subsistence. The remedy for reduced wages is a tax-financed subsidy that redistributes profit income from capital to labor. Move from *C* to *B* along Panel 3's factor price frontier to restore labor's subsistence standard. Move further from *B* toward *E* to make both parties better off than they were at initial point *A*.

Here is Ricardo's conclusion that machine-embodied technical progress hurts labor and helps capital.

Wicksell's case, by contrast, replaces Ricardo's horizontal labor supply curve with a vertical supply curve corresponding to the assumption of fixed factor endowments fully employed. As before, the labor-saving innovation shifts down Panel 1's labor demand curve at the same time it shifts up Panel

2's production function and Panel 3's factor price frontier. To ensure that the post-innovation production function is consistent with the downwardly shifted labor-demand curve, the former has been drawn in the relevant range with a flatter slope that its pre-innovation counterpart. Since the slope of the production function represents labor's marginal productivity—which, in turn, constitutes the demand-for-labor curve in Wicksell's analysis—it follows that a flatter post-innovation production function signifies a lower marginal product of labor and so corresponds to the lower labor demand curve.

Now, however, because the labor supply curve is vertical, labor demand determines the wage rate rather than the level of employment. Equilibrium moves from *A* to *C* rather than from *A* to *B*, as in Ricardo's analysis. The wage rate is allowed to fall to its new market-clearing level where all workers, including those temporarily displaced by the machine, are (re)hired. The wage fall is crucial. It keeps the wage rate equal to labor's lowered marginal productivity and allows output to rise to *C*, the maximum permitted by the unchanged labor force working with the new technology. Most of all, the wage fall permits the rise in the profit rate that spurs capitalists to expand production and re-hire labor.

Of course the new equilibrium wage rate is below subsistence. But workers need not starve. The government can compensate—indeed more than compensate—labor for below-subsistence wages by taxing profits and redistributing the proceeds to workers in the form of relief payments. The resulting move from *C* to *B* and thence toward *E* on the new factor price frontier is equivalent to restoring wages to and then raising them above their subsistence level. While helping labor, such redistribution hardly hurts capital. On the contrary, the transfer leaves both parties, capital and labor, better off than they were at initial point *A*. With extra output to share, everybody gains.

11. WICKSELL ON OUTSOURCING

Wicksell's analysis can be applied to the current offshore outsourcing problem. His advice to labor and the policymakers would go something like this: Don't discourage outsourcing. Like Ricardo's machine, it has the potential to benefit all parties through the extra output it permits. Instead, prevent domestic job losses by letting wages fall to market-clearing levels where it becomes profitable to re-hire laid-off workers. Offset the wage reductions if you must with compensatory profit-sharing or tax-transfer schemes. Such schemes, designed in cooperation with employers and/or the government, can spread the gains from outsourcing over all parties, labor as well as capital. In this way, outsourcing will prove to be unanimously beneficial despite being sharply labor saving.

12. CONCLUSION

Innovation destroys jobs in Ricardo's model. But that model, the first rigorous treatment of the machinery question, is too sparsely specified and idiosyncratic to support the generalizations he drew from it. His assumptions of a horizontal supply-of-labor curve, a minimum bound to wages, and a wage-fund-determined demand-for-labor curve—all essential to his contention that technological change decimates jobs, output, and the labor force—already were becoming anachronistic descriptions of the English labor markets of his day. Certainly his assumptions are unrealistic characterizations of labor markets in developed nations now. Drop the assumptions, and you get Wicksell's optimistic results.

Labor-saving innovations, Wicksell often noted, represent the worst-case scenario as far as job losses are concerned. And if such innovations cannot hurt labor under flexible wages and compensatory profit-sharing schemes, how much less do workers have to fear from labor-neutral and labor-using innovations? Indeed, Wicksell considered labor-saving innovations of the kind depicted in his rendition of the machinery model to be the outliers, and labor-neutral and labor-using innovations the norm. Counting on future technical progress to raise, not lower, labor's marginal productivity, he expected such advances to boost the demand for labor so much that the resulting wage increases would render profit-sharing schemes unnecessary. Historical evidence, showing that innovation, employment, and real wages have advanced together for centuries, supports his view and contradicts Ricardo's.

As an economic theorist, Ricardo was in a class by himself. Arguably the best pure theorist who ever lived, he was at least Wicksell's equal and head and shoulders above Tucker and McCulloch. But on the machinery question, their vision of the job-creating power of technical change seems far more convincing than his pessimistic view. President Richard Nixon in 1972 famously said, "We are all Keynesians now." Similarly, most economists today are Tucker/McCulloch/Wicksellians when it comes to technological progress. They would say with some assurance that innovation and its offspring, offshore outsourcing, are beneficial for the overall American economy and promise to create more jobs in the long run than they destroy in the short.

REFERENCES

Barkai, H. 1986. "Ricardo's Volte-Face on Machinery." *Journal of Political Economy* 94 (June): 595–613.

Basu, S., J. Fernald, and M. Kimball. 2004. "Are Technology Improvements

Contractionary?" NBER Working Paper 10592.

Berg, Maxine. 1980. *The Machinery Question and the Making of Political Economy, 1815–1848*. Cambridge: Cambridge University Press.

Boianovsky, M., and H. Hagemann. 2003. "Wicksell on Technical Change, Real Wages and Employment." Economics Department/University of Brasilia Working Paper 274 (January 17).

Drezner, Daniel W. 2004. "The Outsourcing Bogeyman." *Foreign Affairs* 83(3) (May/June): 22–34.

Francis, Neville, and Valerie Ramey. 2003. "Is the Technology-Driven Real Business Cycle Hypothesis Dead? Shocks and Aggregate Fluctuations Revisited." Unpublished manuscript.

Friedman, Thomas L. 2004. "Small and Smaller." *The New York Times*. Late Edition. 4 (March): 29.

Gali, Jordi. 1999. "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?" *American Economic Review* 89 (March): 249–71.

Hansson, Bjorn. 1983. "Wicksell's Critique of Ricardo's Chapter 'On Machinery.' " *Journal of Economic Studies* 10 (Winter): 49–55.

Jonung, Lars. 1981. "Ricardo on Machinery and the Present [1923] Unemployment: An Unpublished Manuscript by Knut Wicksell." *Economic Journal* 91 (March): 195–8.

Maital, S., and P. Haswell. 1977. "Why Did Ricardo (Not) Change His Mind? On Money and Machinery." *Economica* 44 (November): 359–68.

McCulloch, John R. [1821] 1951. Letter to Ricardo, 21 June. In *Letters, 1819–1821. The Works and Correspondence of David Ricardo*, vol. 8. Ed. Piero Sraffa and Maurice H. Dobb. Cambridge: Cambridge University Press.

O'Brien, D. P. 1975. *The Classical Economists*. New York: Oxford University Press.

Rashid, S. 1987. "Machinery Question." In *The New Palgrave: A Dictionary of Economics*, vol. 3. Ed. J. Eatwell, M. Milgate, and P. Newman. New York: Stockton Press: 264–67.

Ricardo, David. [1821] 1951. *On the Principles of Political Economy and Taxation*, 3rd edition. Vol. 1 of *The Works and Correspondence of David Ricardo*. Ed. Piero Sraffa and Maurice H. Dobb. Cambridge: Cambridge University Press.

__________. [1823] 1951. Speech of 30 May 1823 on Wages of Manufactures—Use of Machinery. In *Speeches and Evidence*; vol. 5 of

The Works and Correspondence of David Ricardo. Ed. P. Sraffa, Cambridge: Cambridge University Press.

__________. [1819–1821] 1951. *Letters*, 1819–1821. Vol. 8 of *The Works and Correspondence of David Ricardo*. Ed. Piero Sraffa and Maurice H. Dobb. Cambridge: Cambridge University Press.

Samuelson, Paul. 1988. "Mathematical Vindication of Ricardo on Machinery." *Journal of Political Economy* 96 (April): 274–82.

__________. 1989. "Ricardo Was Right!" *Scandinavian Journal of Economics* 91(1) (March): 47–62.

__________. 1994. "The Classical Classical Fallacy." *Journal of Economic Literature* 32 (4) (July): 620–39.

St. Clair, Oswald. [1957] 1965. *A Key to Ricardo*. London: Routledge & Kegan Paul. Reprinted New York: A. M. Kelley.

Tucker, J. [1757] 1931. *Instruction for Travellers*. Reprinted in R.L. Schuyler's *Josiah Tucker*. New York: Columbia University Press.

Wicksell, Knut. [1901] 1934. *Lectures on Political Economy*. Vol. 1. 1915 edition. Translated by E. Classen. New York: Macmillan.

__________. [1923] 1981. "Ricardo on Machinery and the Present Unemployment." *Economic Journal* 91 (March): 199–205.

__________. [1924] 1969. "Protection and Free Trade." In *Selected Papers on Economic Theory*. Edited with an introduction by E. Lindahl. New York: A. M. Kelley.

__________. [1896] 1958. "Taxation in the Monopoly Case." Appendix to Chapter 2 of his *Finanztheoretische Untersuchungen, nebst Darstellung und Kritik des Steuerwesens Schwedens*, pp. 19–21. Jena: Gustav Fischer. Reprinted in *Readings in the Economics of Taxation*, eds. R. A. Musgrave and C. S. Shoup, pp. 256–7. Homewood, Ill: Richard D. Irwin, Inc.

(Un)Balanced Growth

Andreas Hornstein

Since the late 1800s, real output in the United States has been growing at a steady rate of about 3.5 percent per year (see Figure 1).[1] With the exception of the 20 years between 1930 and 1950, the real aggregate capital stock of the United States has also been growing at that same steady rate. Thus, although output tripled and capital increased by a factor of 2.5 over this time period, the capital-output ratio remained roughly constant before 1930 and after 1950. Available data also indicate that the relative price of capital in terms of consumption goods has not changed much since the 1950s. In this article I review to what extent the stability of the aggregate capital accumulation pattern actually masks substantial changes in the composition of the aggregate capital stock—namely, changes in the relative importance of equipment and structures.

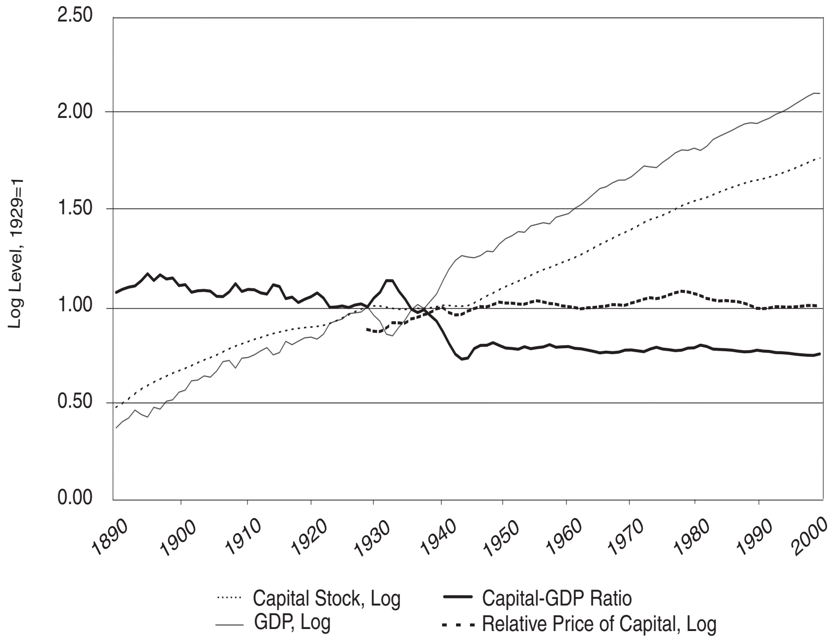
The observed stability of output and capital growth rates and the capital-output ratio are part of the "stylized facts" of growth (Kaldor 1957, 1961). The stylized facts also include the observations that the rate of return on capital and factor income shares have remained stable over long time periods in the United States and other industrialized countries.[2] These observed regularities suggest that a common theoretical framework might be able to account for the output and capital accumulation path of the U.S. economy and other industrialized economies over the last 100 years. Indeed, neoclassical growth theory was built around the stylized facts of growth.

I would like to thank Kartik Athreya, Yongsung Chang, Jason Cummings, and Marvin Goodfriend for helpful comments. I also would like to thank Jason Cummings for making his data available to me, and Jon Petersen for excellent research assistance. The views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Detailed descriptions of the data used in this paper are in the Appendix.

[2] The rate of return on capital as measured by the real return on equity in the United States has not changed much over time. Siegel (1998) calculates an average rate of return on equity of 7.0 percent for the time period 1802–1870, 6.6 percent for 1871–1925, and 7.2 percent for 1926–1997. Using BEA figures on nonfarm private business factor incomes, I calculate wage income shares for the time period 1929–2001. The average wage income share is 0.66 and varies between 0.63 and 0.72 with no discernible trend. If I exclude the housing sector, the average wage income share is 0.75 and varies between 0.71 and 0.84, again with no discernible trend.

Figure 1 Real Output and Capital in the Private Business Sector, 1889–2001



Notes: A detailed description of the data is in the Appendix.

Neoclassical growth theory assumes that there are two inputs to production: non-reproducible labor and reproducible capital. For a given level of technology, production is constant returns to scale in all inputs, and there are diminishing marginal returns for individual inputs. Technical change is taken as exogenous and is assumed to increase the marginal products of capital and labor for given amounts of inputs. Both inputs are assumed to be paid their marginal product, and the higher marginal product of capital induces more capital accumulation. In an equilibrium, capital accumulation proceeds at a rate such that the return on capital remains constant. Since the labor endowment is fixed, higher productivity and more capital increases payments to labor over time.

Within the framework of neoclassical growth theory, the stylized facts are interrelated. The rate of return on capital, r, is the gross rental rate, u, minus the value of depreciation, δp, plus any capital gains due to changes in the

Figure 2 The Relative Price of New Capital Goods, 1929–2001

Notes: A detailed description of the data is in the Appendix.

price of capital, Δp, divided by the price of capital:

$$r = \frac{u - \delta p + \Delta p}{p} = \frac{uk}{y}\frac{y}{pk} - \delta + \frac{\Delta p}{p}. \qquad (1)$$

Conditional on a constant depreciation rate, δ, and a constant price of capital, $\Delta p = 0$, the stability of either two of the three time series—capital-output ratio, capital income share, and rate of return on capital—implies the stability of the third time series.

In neoclassical growth theory, growth is driven by technological change, and capital accumulation responds to technical change, but the source of technical change is not explored. Greenwood, Hercowitz, and Krusell (1997) have argued that technical change in the sector that produces equipment capital is a major source of growth. Their argument for relatively faster technical change in this sector is based on the long-run decline of the relative price of equipment capital. Since the 1960s, the price of equipment capital relative to the price of consumption goods has been falling by about 40 percent, whereas the relative price of structures has been increasing by about 10 percent (see Figure 2).[3] If the relative price of equipment capital has been declining, then the producers of equipment capital must have become relatively more efficient.

Greenwood et al. (1997) evaluate the long-run contribution of different sources of technical change, including the response of capital accumulation to technical change. This is a reasonable procedure since long-run growth depends not only on exogenous technical change, but also on the endogenous capital accumulation response to technical change. But in order to determine the capital accumulation response to hypothetical time paths of technical change, one needs a theory of growth. Greenwood et al. (1997) use a straightforward extension of the aggregate neoclassical growth model to their multi-sector view of the economy, and they use a standard characterization of long-run growth. In particular, they assume that the long-run equilibrium growth path is balanced; that is, all variables grow at constant but possibly different rates. The stylized growth facts represent a balanced growth path (BGP) for the aggregate economy.

In order to obtain a balanced growth path, Greenwood et al. (1997) have to assume that the elasticity of substitution between inputs is unitary; that is, production is Cobb-Douglas (CD) in all sectors of the economy. I argue that the accumulation of equipment and structures in the second half of the 20th-century United States is not characterized by balanced growth. In particular, I argue that equipment capital has become relatively more important over time. This means that the stylized facts of growth do not apply to a more disaggregate view of the economy.[4] It also means that one cannot argue for a unitary elasticity of substitution between inputs based on growth path properties alone.

The changing composition of the aggregate capital stock does have implications for the implicit aggregate depreciation rate. Because the relative share of equipment capital has increased and equipment capital depreciates relatively faster than structures, the implicit aggregate depreciation rate has increased substantially since the 1980s. With a constant return on capital and a constant capital income share, neoclassical growth theory would then imply that the economy should move towards a new BGP with a lower aggregate capital-output ratio, which we have not observed. Thus, growth theory now appears to be at odds with the stylized facts of growth, at least for the last 20 years.

---

[3] Alternative measures of the relative price of equipment and software that try to account better for changes in product quality (Cummins and Violante 2002) indicate an even bigger decline, about 90 percent, over the same time period.

[4] This observation alone should not be too surprising since we have known for a long time about the extent of structural change in the U.S. economy and its limited impact on the growth path of the aggregate economy. See, for example, Kuznets' (1971) work on the secular decline of the agricultural sector and the corresponding increase of the service sector.

In Section 1, I describe a balanced growth path for a simple growth model where the relative price of capital is changing on the BGP. In the model are two types of capital: equipment and structures. I show that the capital-output value ratios are constant over time on the BGP but that the capital-output quantity ratios are not constant. In Section 2, I first review the long-run evidence on the ratio of the real capital stock to output. I find that from the late 1800s to the present, the ratio of structures to GDP has been steadily declining, and the ratio of equipment capital to GDP has been steadily increasing since 1950. I then review the evidence on the ratio of the capital value to output value for equipment and structures for the United States from 1950 on and find that these ratios also do not remain constant. It thus appears that we can find evidence for BGP properties of the U.S. economy at the level of aggregate capital only. I then return to the evidence of stable aggregate capital-output ratios and argue that this in fact appears to be evidence against, rather than for, balanced growth since the 1980s because the depreciation rate of the aggregate capital stock has increased substantially since the 1980s. In Section 3, I discuss some implications for modeling long-run growth and economic policy.

1. BALANCED GROWTH WITH MULTIPLE CAPITAL GOODS

I now describe a simple three-sector growth model with two types of capital and a fixed labor supply based on Greenwood, Hercowitz, and Krusell (1997). Technical change is labor-augmenting and sector specific. I will discuss under what restrictions balanced growth can occur—all variables grow at constant but not necessarily equal rates. In particular, I am interested in BGPs where the relative price of capital changes over time. I will focus on the production structure of the economy and disregard any restrictions on preferences necessary for a BGP. I will assume that, for the implied time paths of prices, a constant labor supply and the implied constant consumption growth rate are consistent with an equilibrium.

A Three-Sector Economy

Consider an economy that produces three goods: a consumption good, c, and two investment goods—equipment, x_e, and structures, x_s. Inputs to the production of any of the three goods are labor, n, and the stocks of the two capital goods—equipment, k_e, and structures, k_s. Assume that there is perfect mobility of labor and capital between the production sectors. Let $\rho_i \geq 0$ denote the fraction of labor allocated toward the production of the type $i = c, e, s$ good. Analogously, let $\phi_i \geq 0$ and $\mu_i \geq 0$ denote the fraction of the equipment capital stock and the structures stock allocated toward the production of type i goods. Capital stocks and labor allocated to each sector are limited by the total endowment of each input:

$$\phi_c + \phi_e + \phi_s \leq 1, \mu_c + \mu_e + \mu_s \leq 1, \text{ and } \rho_c + \rho_e + \rho_s \leq 1. \tag{2}$$

Production is constant returns to scale (CRS), and technical change is labor-augmenting:

$$c = C\left(\phi_c k_e, \mu_c k_s, A_c \rho_c n\right), \tag{3}$$

$$x_e = E\left(\phi_e k_e, \mu_e k_s, A_e \rho_e n\right), \text{ and} \tag{4}$$

$$x_s = S\left(\phi_s k_e, \mu_s k_s, A_s \rho_s n\right). \tag{5}$$

The effective labor input in a sector equals employment times labor-specific productivity, A_i. Labor-specific productivity may differ across sectors and may change at different but constant rates. Time is continuous, and the rate of change of labor-specific productivity is $\hat{A}_i = \gamma_i$.[5] Investment augments the existing capital stock after depreciation, $\delta_i > 0$:

$$\dot{k}_e = x_e - \delta_e k_e, \text{ and} \tag{6}$$

$$\dot{k}_s = x_s - \delta_s k_s. \tag{7}$$

Balanced Growth with Constant Returns to Scale

On a BGP, all variables change at constant rates.[6] Since total employment is fixed by assumption, this means that, on a BGP, employment in each sector is constant. Thus, the fraction of labor allocated to each sector, ρ_i, is constant. Given the resource constraint for equipment capital, the rate at which the use of equipment capital grows in each sector must be the same as the growth rate of total equipment capital. Therefore, the fraction of equipment capital, ϕ_i, allocated to each sector remains constant on a BGP. By the same argument, the fraction of structures, μ_i, allocated to each sector remains constant.

The equations for capital accumulation, (6) and (7), imply that on a BGP, the investment-capital stock ratio is constant for each capital type:

$$\hat{k}_e = \frac{x_e}{k_e} - \delta_e, \text{ and } \hat{k}_s = \frac{x_s}{k_s} - \delta_s. \tag{8}$$

Thus, investment in new capital goods grows at the same rate as does the stock of capital:

$$\hat{k}_e = \hat{x}_e, \text{ and } \hat{k}_s = \hat{x}_s. \tag{9}$$

---

[5] In the following, a dot represents the time derivative of the variable, $\dot{y} = dy/dt$, and a hat denotes the growth rate of the variable, $\hat{y} = \dot{y}/y$.

[6] Variables may remain constant on a BGP; that is, their rate of change is zero.

Since production is CRS, we can rewrite the production functions as

$$c/k_e = C\left(\phi_c, \mu_c k_s/k_e, n\rho_c A_c/k_e\right), \tag{10}$$

$$x_e/k_e = E\left(\phi_e, \mu_e k_s/k_e, n\rho_e A_e/k_e\right), \text{ and} \tag{11}$$

$$x_s/k_s = S\left(\phi_s k_e/k_s, \mu_s, n\rho_s A_s/k_s\right). \tag{12}$$

A sufficient condition for a BGP to exist is then that each argument in the rescaled production functions, (10) to (12), is constant; that is,

$$\hat{k}_s = \hat{k}_e = \hat{A}_c = \hat{A}_e = \hat{A}_s. \tag{13}$$

Note that, without imposing any additional restrictions on the form of the production functions, a BGP exists only if labor-augmenting technical change proceeds at the same rate in each sector.

Relative goods prices do not change on the BGP of a competitive equilibrium if technical change proceeds at the same rate in each sector. In a competitive equilibrium, profit maximizing firms take prices as given and hire inputs until an input's rental rate is equalized with the value of the marginal product of that input. Because inputs are perfectly mobile across sectors, they are paid the same rental rate no matter where they are employed. The conditions for optimal input use are then:[7]

$$\begin{aligned} u_e &= p_e E_e\left(\phi_e k_e, \mu_e k_s, A_e\rho_e n\right) \\ &= p_s S_e\left(\phi_s k_e, \mu_s k_s, A_s\rho_s n\right) = C_e\left(\phi_c k_e, \mu_c k_s, A_c\rho_c n\right), \end{aligned} \tag{14}$$

$$\begin{aligned} u_s &= p_s S_s\left(\phi_s k_e, \mu_s k_s, A_s\rho_s n\right) \\ &= p_e E_s\left(\phi_e k_e, \mu_e k_s, A_e\rho_e n\right) = C_s\left(\phi_c k_e, \mu_c k_s, A_c\rho_c n\right), \text{ and} \end{aligned} \tag{15}$$

$$\begin{aligned} w &= C_n\left(\phi_c k_e, \mu_c k_s, A_c\rho_c n\right) A_c \\ &= p_e E_n\left(\phi_e k_e, \mu_e k_s, A_e\rho_e n\right) A_e = p_s S_n\left(\phi_s k_e, \mu_s k_s, A_s\rho_s n\right) A_s. \end{aligned} \tag{16}$$

Normalize the price of the consumption good to one, and let p_e (p_s) denote the price of new equipment (structures) in terms of consumption goods. Let w denote the real wage, that is, the price of labor in terms of consumption goods, and let u_e (u_s) denote the rental rate of equipment (structures) capital. The rate of return on investment in either of the capital goods is also equalized:

$$r = \frac{u_e - \delta_e p_e + \dot{p}_e}{p_e} = \frac{u_s - \delta_s p_s + \dot{p}_s}{p_s}. \tag{17}$$

We now see that if labor-augmenting technical change proceeds at the same rate in each sector, the relative price of capital will be constant on a BGP. Because production is CRS—that is, homogeneous of degree one—the first derivatives of the production function (marginal products) are homogeneous

[7] The notation, I_i for $I = C, E, S$ and $i = e, s, n$, denotes the partial derivative of the function I with respect to the input i.

of degree zero. We can therefore rewrite equation (14) as

$$p_e = \frac{C_e\left(\phi_c, \mu_c k_s/k_e, n\rho_c A_c/k_e\right)}{E_e\left(\phi_e, \mu_e k_s/k_e, n\rho_e A_e/k_e\right)}, \tag{18}$$

and, on a BGP, all ratios on the right-hand side are constant. The same argument applies to the relative price of structures.

Balanced Growth With Unitary Elasticity of Substitution

We can construct a BGP where relative goods prices change at a constant rate if production in each sector is of the Cobb-Douglas variety. On a BGP, the total derivative of consumption (3) with respect to time is[8]

$$\dot{c} = C_e\phi_e\dot{k}_e + C_s\mu_s\dot{k}_s + C_n\dot{A}_c\rho_c n. \tag{19}$$

This expression gives us the growth rate of consumption goods in terms of the growth rates of inputs and labor-augmenting technical change:

$$\frac{\dot{c}}{c} = \frac{C_e\phi_e k_e}{C}\frac{\dot{k}_e}{k_e} + \frac{C_s\mu_s k_s}{C}\frac{\dot{k}_s}{k_s} + \frac{C_n A_c\rho_c n_c}{C}\frac{\dot{A}_c}{A_c}. \tag{20}$$

Let $\eta_{Ij} \equiv \left(\partial I/\partial k_j\right)\left(k_j/I\right)$ denote the input elasticity of type I production with respect to the type j capital good. Then we can write the BGP growth rates for outputs as[9]

$$\hat{c} = \eta_{C,e}\hat{k}_e + \eta_{C,s}\hat{k}_s + \left(1 - \eta_{C,e} - \eta_{C,s}\right)\hat{A}_c, \tag{21}$$

$$\hat{x}_e = \eta_{E,e}\hat{k}_e + \eta_{E,s}\hat{k}_s + \left(1 - \eta_{E,e} - \eta_{E,s}\right)\hat{A}_e, \text{ and} \tag{22}$$

$$\hat{x}_s = \eta_{S,e}\hat{k}_e + \eta_{S,s}\hat{k}_s + \left(1 - \eta_{S,e} - \eta_{S,s}\right)\hat{A}_s. \tag{23}$$

On a BGP, investment grows at the same rate as the capital stock (9), implying the following system of equations for the growth rates of consumption and capital goods:

$$\hat{c} = \eta_{C,e}\hat{k}_e + \eta_{C,s}\hat{k}_s + \left(1 - \eta_{C,e} - \eta_{C,s}\right)\hat{A}_c, \tag{24}$$

$$\left(1 - \eta_{E,e}\right)\hat{k}_e - \eta_{E,s}\hat{k}_s = \left(1 - \eta_{E,e} - \eta_{E,s}\right)\hat{A}_{e,} \text{ and} \tag{25}$$

$$-\eta_{S,e}\hat{k}_e + \left(1 - \eta_{S,s}\right)\hat{k}_s = \left(1 - \eta_{S,e} - \eta_{S,s}\right)\hat{A}_s. \tag{26}$$

Thus, a BGP with potentially different rates of labor-augmenting technical change exists if the input elasticities are constant. That is, the production functions are of the Cobb-Douglas (CD) variety. For example, the production

[8] I have used the fact that the total labor endowment is constant, and, on a BGP, the fraction of resources allocated to each sector remains constant.

[9] Because of constant returns to scale, the input elasticities sum to one.

function for consumption goods is

$$c = \left(\phi_c k_e\right)^{\eta_{C,e}} \left(\mu_c k_s\right)^{\eta_{C,s}} \left(A_c \rho_c n\right)^{1-\eta_{C,e}-\eta_{C,s}}, \tag{27}$$

with $\eta_{C,e}, \eta_{C,s}, 1 - \eta_{C,e} - \eta_{C,s} \geq 0$. Analogous expressions hold for the production of new equipment and structures. For CD production functions, the elasticity of substitution between inputs is unitary. That is, cost minimizing firms respond to a 1 percent increase in the relative price of an input with a corresponding 1 percent reduction in the relative usage of that input such that the cost share of the input remains constant at the input elasticity.[10]

We can now derive expressions for the rate of change of relative capital goods prices. Using the CD production structure, we can rewrite the condition for the profit maximizing use of equipment capital (14) as

$$u_e = \frac{\eta_{C,e} c}{\phi_c k_e} = p_e \frac{\eta_{E,e} x_e}{\phi_e k_e} = p_s \frac{\eta_{S,e} x_s}{\phi_s k_e}. \tag{28}$$

Given the constant input elasticities and the constant allocation shares of total equipment capital, this expression implies that the rates of change for relative prices are given by

$$\hat{p}_e = \hat{c} - \hat{k}_e \text{ and } \hat{p}_s = \hat{c} - \hat{k}_s \tag{29}$$

Thus, the relative price of capital in terms of consumption goods will change if capital accumulation proceeds at a different pace than does consumption growth. On the other hand, even if capital accumulation proceeds at a different pace than does consumption growth, the value of the capital stock relative to the value of consumption will remain constant. Finally, equation (28) for equipment and the corresponding expression for structures also imply that the rate of return on investment in either of the capital goods is constant.

The implications of the CD production structure for a BGP are most easily seen if we further simplify the production structure and assume that input elasticities (income shares) are equal in all industries, $\eta_{I,j} = \eta_j$ (see, for example, Greenwood, Hercowitz, and Krusell 1997). In this case, consumption and capital growth rates on the BGP are

$$\hat{c} = \eta_e \hat{A}_e + \eta_s \hat{A}_s + \left(1 - \eta_e - \eta_s\right) \hat{A}_c, \tag{30}$$

$$\hat{k}_e = \left(1 - \eta_s\right) \hat{A}_e + \eta_s \hat{A}_s, \text{ and} \tag{31}$$

$$\hat{k}_s = \eta_e \hat{A}_e + \left(1 - \eta_e\right) \hat{A}_s. \tag{32}$$

The rates of change for relative prices are

$$\hat{p}_e = \left(1 - \eta_e - \eta_s\right) \left(\hat{A}_c - \hat{A}_e\right), \text{ and } \hat{p}_s = \left(1 - \eta_e - \eta_s\right) \left(\hat{A}_c - \hat{A}_s\right). \tag{33}$$

[10] This fact is immediate from equations (14) to (16) and the constant input elasticities.

Note that the relative price changes directly reflect differences in the rates of labor-augmenting technical change.

Finally, we can define real output as total output in terms of consumption goods:

$$y = c + p_e x_e + p_s x_s.$$

Since capital investment grows at the same rate as capital stocks and since the value of capital grows at the same rate as consumption, real output grows at the same rate as consumption.

2. OBSERVATIONS ON CAPITAL ACCUMULATION IN THE UNITED STATES, 1869–2001

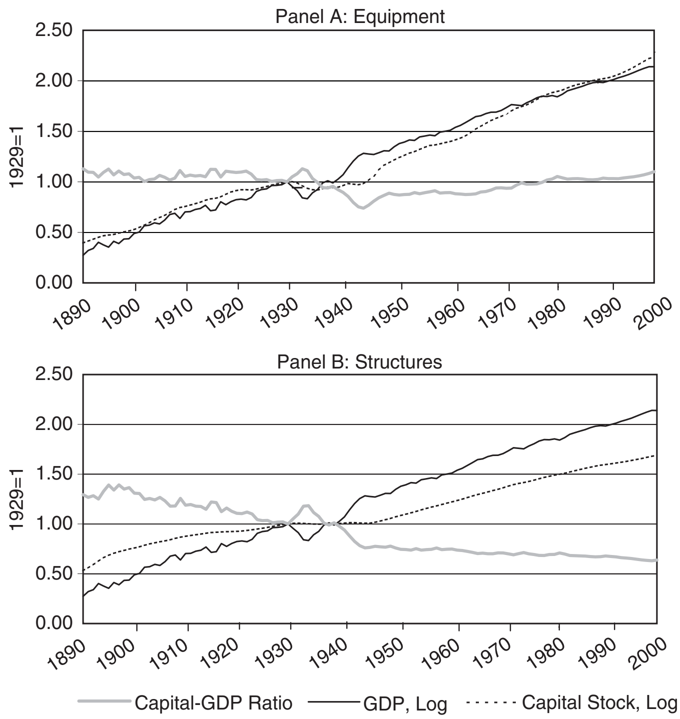
The fact that the relative price of equipment capital declined substantially, whereas the relative price of structures increased somewhat, suggests a differential productivity growth rate in the two broadly defined capital goods-producing sectors. An economy with productivity growth rates that are constant but different across capital goods-producing sectors can still achieve a BGP. Note that, however, on the BGP, the capital-output value ratios remain constant but that the capital-output quantity ratios do not. I now argue that there is no evidence for balanced growth in the United States economy at this more disaggregate level. Over the long run, capital-output quantity ratios do not appear to be stationary, and, for the post-WWII period, the equipment capital-output value ratio does not appear to be stationary either. Finally, I argue that one of the conditioning assumptions for balanced growth at the aggregate level—namely, constant depreciation rates—also does not hold for the late 20th century. Some of the changes in depreciation rates have to be attributed to the changing composition of the aggregate capital stock.

The structure of the U.S. economy changed drastically over the time period I consider. For example, government accounted for between 4 and 5 percent of GDP in the late 1800s, but since the 1930s, government's share in GDP has increased to about 10 percent. Within the private business sector, the share of value added that originated in agriculture declined from 30 percent in 1889 to 10 percent in 1930. It then stayed there until the mid-1940s and since then has declined to less than 1 percent.[11] These changes in GDP shares are also reflected in the changed employment shares of government and agriculture, but they are not the focus of this article.[12] In an attempt to limit the potential impact of this structural change on the BGP properties of capital-output ratios, I limit my analysis to output and capital in the nonfarm private business sector.

[11] The shares for the pre-1929 period are calculated from Kendrick's (1961) constant dollar estimates, and the shares for the post-1929 period are from BEA's current dollar estimates. In 1930, the shares based on constant and current dollar estimates are both roughly the same.

[12] See, for example, Kuznets (1971).

Figure 3 Real Capital and Output in the Nonfarm Private Business Sector, 1889-2001



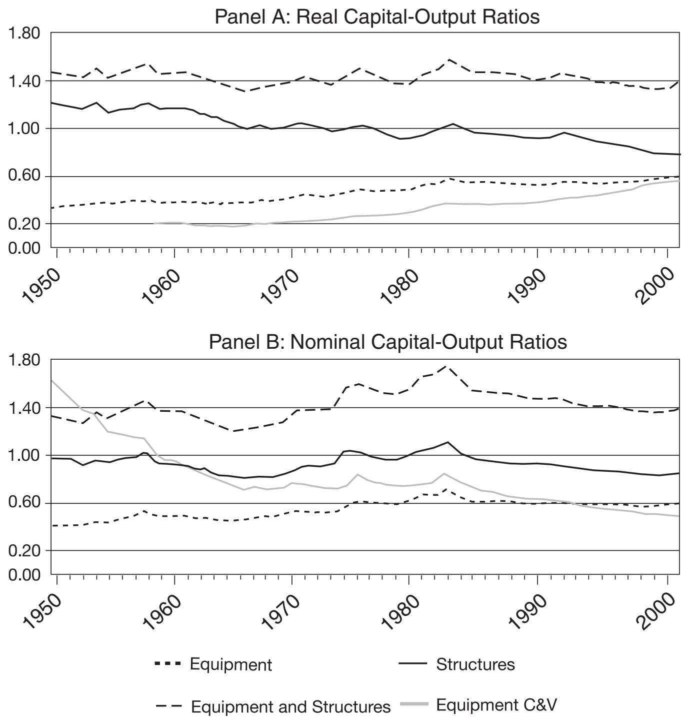
Notes: A detailed description of the data is in the Appendix.

Capital-Output Ratios for Equipment and Structures

Over the long run, capital-output quantity ratios for either equipment or structures do not appear to be stable. Figure 3 displays real nonfarm private GDP, nonfarm private equipment and structures, and the relevant capital-output ratios from 1889 to 2001. The stock of structures includes residential structures. There is a clear downward trend in the capital-output ratio for structures—the ratio has been falling steadily since the late 1800s, from 1.4 to 0.6 today. The behavior of the equipment capital-output ratio is more ambiguous. This ratio appears to be quite stable before 1929, then declines substantially until the 1950s and from then on shows a clear upward trend.

We now study the capital-output quantity and value ratios for the nonfarm private business sector, excluding housing. A big component of the housing sector output in National Income and Product Accounts (NIPA) consists of

Figure 4 Capital-Output Ratios in the Nonfarm Private Business Sector Excluding Housing, 1950–2001



Notes: A detailed description of the data is in the Appendix.

imputed rental income for owner-occupied housing. One could argue that NIPA data that includes residential housing is less reliable and that one should therefore focus on the non-residential business sector. Figure 4 graphs the capital-output quantity and value ratios for this sector. Even though residential structures are now excluded, the structures-output quantity ratio continues to decline for the period from 1.3 in 1950 to 0.8 in 2000. The structures-output value ratio, however, remains relatively stable over this time period. For equipment capital, we see that both the quantity and the value capital-output ratio increase from 0.4 in 1950 to 0.6 in 2000.

Figure 4 also includes real and nominal equipment-output ratios based on updated data from Cummins and Violante (2003). Cummins and Violante (2003) argue that the official NIPA figures overestimate the inflation rate for equipment capital because they do not appropriately account for quality change. Based on Cummins and Violante (2003), the real equipment-output ratio increased drastically from 0.2 in 1960 to 0.7 in 2000. On the other hand, according to their numbers, the nominal equipment-output ratio actually fell from 1.6 to 0.5.[13]

Other Related Work

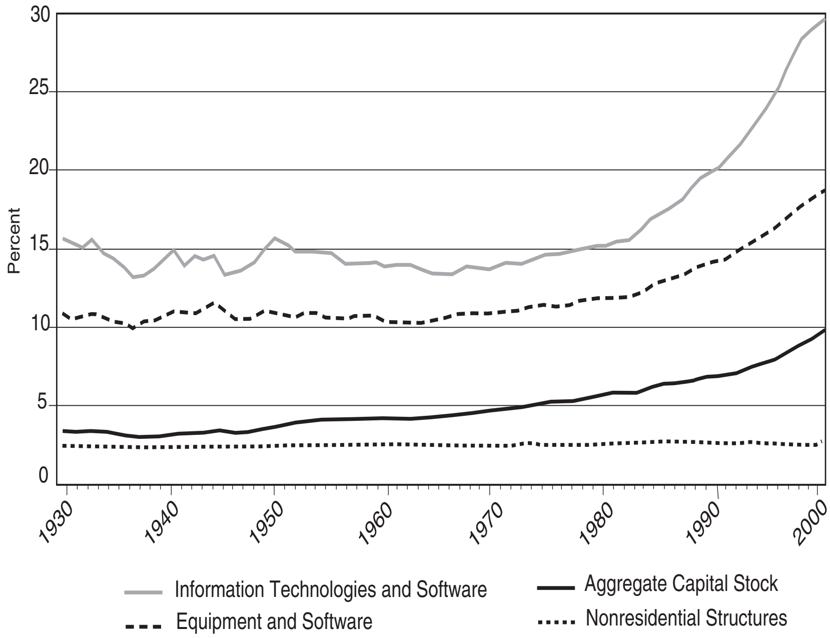
Maddison (1991) also argues that real capital-output ratios are not stationary, in particular, that the ratio of nonresidential capital to GDP changed substantially for some countries other than the United States. According to Maddison (1991, 67), from 1890 to 1987 the capital-GDP ratio doubled for the UK (from 0.95 to 2.02) and almost tripled for Japan (from 0.9 to 2.8). Maddison (1991) also calculates big increases for France and Germany from 1913 to 1987. Finally for the 19th-century United States, Gallman (1986, 192) argues that the real equipment capital-GDP ratios actually increased from 0.15 in 1840 to 0.91 in 1900 and that the corresponding nominal capital-GDP ratios increased from 0.23 in 1840 to 0.40 in 1900. Related to the stability of capital-output ratios on a BGP is the stability of expenditure shares in GDP. Along these lines King, Plosser, Stock, and Watson (1991) argue that the behavior of real private GDP, consumption, and investment in the United States satisfies the balanced growth conditions from 1949 to 1988. Their statistical tests indicate that output and consumption, as well as output and investment, are cointegrated and that the consumption-output and investment-output ratios are stationary. Whelan (2003) reviews the evidence for balanced growth of two expenditure components: consumption and investment. He also emphasizes that in the face of drastically changing relative prices of investment in producer-durable equipment and investment in structures, one should not expect that the ratios of real consumption to real investment in either of the two types remain constant, but rather that the ratios of nominal expenditures remain constant. Whelan shows that for an extended sample—the United States from 1949 to 2000—one can reject the null hypothesis of cointegration for real investment and real consumption, but one cannot reject the null hypothesis of cointegration for nominal consumption and nominal investment.

Depreciation

Depreciation rates for equipment capital have been increasing significantly since the 1980s (see Figure 5). This increase reflects mainly an aggregation

[13] At first, one might be surprised that the capital-output value ratios for Cummins and Violante (2003) are so different from the corresponding NIPA ratios. After all, we usually observe values, and the problem is one of obtaining a quantity index by deflating values with an appropriate price index. The problem with capital is that most capital is not traded, so we do not observe values. The NIPA procedure to obtain current values of the capital stock is to evaluate the estimated real capital stock at the current prices of capital stocks. I have used the same procedure to obtain capital stock values for Cummins and Violante (2003), and there is no reason to expect that the two estimates should agree on their values.

Figure 5 Depreciation Rates, 1926–2001



Notes: A detailed description of the data is in the Appendix.

effect since information technology (IT) equipment has higher depreciation rates than other equipment capital types, and the relative share of IT in total equipment has been increasing as part of the IT revolution. Assuming a stable real rate of return on capital, higher depreciation rates require a higher rental rate of capital to satisfy the optimal capital accumulation condition (1). This, in turn, implies that the aggregate capital-output ratio should have declined significantly, since the capital income share has remained stable for this time period. Thus, the stability of the aggregate capital-output ratio since the 1980s suggests that the behavior of the U.S. economy has not been well approximated by a BGP.

The aggregate depreciation rate on private nonfarm capital increased from about 4 percent before the 1970s to about 10 percent in the year 2000 (see Figure 5). Based on the neoclassical growth model, this increase in the depreciation rate should have resulted in a significant decline of the capital-output ratio. Before 1970, the capital-output ratio for the nonfarm private business sector was about 2.5 (1.5, excluding housing), and the capital income share was about 0.33 (0.25, excluding housing) (cf. footnote 2). Assuming a constant price index for the aggregate capital stock and use of the optimal capital accumulation condition (1) implies a 9.3 (12.5, excluding housing) percent rate of return on capital.[14] An increase of the depreciation rate to 10 percent would then imply that the capital-output ratio should decline to 1.7 (1.1, excluding housing), conditional on a constant return on capital and a constant capital income share. We have not observed such a decline in the capital-output ratios.[15]

I interpret the depreciation rates constructed from the BEA depreciation and capital stock series as physical depreciation of the capital stock. Whelan (2002) argues that this is not a valid interpretation of the BEA depreciation rates since the BEA frequently applies depreciation schedules in its procedures that include both physical and economic depreciation. While this is a valid concern, it does not necessarily affect our conclusions. First, to the extent that the BEA data overestimate the depreciation of the capital stock, they also underestimate the size of the capital stock. This means that the equipment capital-GDP ratio should have increased even faster over the last 20 years, implying that the aggregate capital-output ratio would have increased rather than remained stable. Second, it is not clear that the higher depreciation rate observed for the BEA data can only be attributed to faster economic depreciation. IT equipment depreciates at a rate of more than 20 percent, which is substantially higher than the 10 to 20 percent rates of other equipment capital.[16] Thus, the composition effect induced by an increase of the relative share of IT equipment alone increases the aggregate depreciation rate.

3. CONCLUSION

Does it matter whether the time path of the economy is characterized by balanced growth? Yes. Observed empirical regularities are important for the development of our understanding of how the economy works, and a breakdown in one of these empirical regularities might be viewed as a setback. The perceived stylized facts of growth certainly stimulated research on growth, and the neoclassical growth theory provided a simple interpretation of the facts that added to the appeal of that theory. On the other hand, since the behavior of the capital-output ratio over the last 20 years apparently no longer conforms

[14] Note that these estimates for the return on capital are significantly higher than other estimates, such as Siegel's (1998) estimate of 7 percent for the long-run rate of return on equity.

[15] One can argue that it takes some time for the economy to converge to the new BGP, given the new higher depreciation rates. Note, however, that for the observed capital income shares and depreciation rates, convergence—as predicted by the growth model—tends to be fast, and a higher depreciation rate speeds up the convergence process.

[16] In a recent reevaluation of the BEA depreciation scheme for PCs, Doms et al. (2003) have found annual depreciation rates of up to 35 percent, even accounting for quality change and economic depreciation. The work of Doms et al. (2003) has already been incorporated in the 2003 NIPA revisions.

to the stylized facts, it may just point to some structural break. After all, there was also a structural break in the aggregate capital-output ratio between 1930 and 1950, and the ratio was quite stable before 1930 and after 1950. We would then have to come up with some explanation for this structural break. The behavior of the aggregate capital-output ratio is also of some interest for the evaluation of monetary policy. Some observers have argued that the investment boom in the late 1990s was in part due to monetary policy that did not raise interest rates fast enough. The collapse of the investment boom and the ensuing 2001 recession are then attributed to having too much capital around, that is, overcapacities. In order to talk about whether the capital stock is too high, one needs an estimate of the "normal" capital stock. If the long-run capital-output ratio is quite stable, we might use this ratio to construct a good first indicator of the "normal" capital stock. Since the aggregate capital-output ratio at the end of the 1990s was not much out of line with its long-run average, we would therefore conclude that the investment boom of the 1990s did not result in any overcapacities. Thus, it did not contribute to the 2001 recession. On the other hand, taking into account the substantial increase of aggregate depreciation rates, we should conclude that the "normal" capital-output ratio was much lower than the observed capital-output ratio. Thus, the investment boom did result in "excess" capital.

APPENDIX

Economic time series are usually reported in current prices—nominal terms. For much of our analysis, we want to eliminate the effect of price level changes and use time series in "real" terms. These real time series are supposed to reflect quantity rather than price movements. Given the level of aggregation we usually deal with, the construction of aggregate quantity indexes is neither easy nor unambiguous. Most data on historical "real" series calculate constant-dollar base period quantity indexes. That is, aggregate quantities are constructed by weighting individual quantity series with a fixed set of base period prices. For the most part, aggregation procedures are a matter of convention, but based on economic theory, some indexes are better than others. One can make an argument that quantity indexes from constant base period prices are not very reliable indicators for quantity movements when large changes of relative prices occur during the period of interest. In response to these concerns, the Bureau of Economic Analysis (BEA), which provides the National Income Accounts for the United States, has shifted from quantity indexes that are based on constant base period prices to chain-weighted quantity indexes that deal better with relative price changes. The BEA constructs Fisher-Ideal quantity indexes that, for most applications, are well approximated by Divisia-quantity indexes. For a more detailed description of the BEA procedures, see Seskin and Parker (1998) or USBEA (1997).

Output

The GDP series used in this article are from Kendrick (1961) and the official NIPA series published by the BEA. Output in Figure 1 is real private business GDP; in Figure 3, output is real nonfarm private business GDP, and in Figure 4, output is real (nominal) nonfarm private business GDP, excluding the contributions of the housing sector. For the long-time series in Figures 1 and 3, I splice the constant 1929 dollar GDP series from Kendrick (1961) with the chained 1996 dollar GDP series from the BEA in 1929 at the level of 1929 BEA GDP. Obviously, there are potential problems since the two series use different methods to obtain estimates of real activity. Nevertheless the average growth rates of real private business GDP for the pre-1929 and post-1929 periods are remarkably similar—3.8 percent before 1929 and 3.7 percent after 1929. For the time series in Panel A of Figure 4, I construct real nonfarm private business GDP excluding housing as a Divisia index using the series on nonfarm private business GDP and housing GDP.

In this article I rely on the work of Kendrick (1961) for the time period before 1929. Balke and Gordon (1989) provide a useful survey on the different sources for GNP data in the United States before 1929, the most important source being Kuznets (1961). Kendrick (1961) essentially restates Kuznets' estimates of GNP for the definitions used by the Department of Commerce in its construction of the NIPAs. Balke and Gordon (1989) update this early work, but their concern is with the relative volatility of GNP in the period before and after WWII. Since they consider the work of Kuznets (1961) and Kendrick (1961) as providing acceptable estimates of trends, they construct their own GNP estimates around the Kuznets and Kendrick trend estimates. Thus Balke and Gordon's (1989) updates do not affect the information in Kendrick (1961) that is relevant for this article.

Cummins and Violante (2002), discussed below, have recently provided alternative estimates of the price deflator for producer-durable equipment investment. Using this alternative deflator affects the measure of real investment and, in turn, the measure of the real capital stock and real GDP. In Figure 4, Panel A, I also display an equipment capital-GDP ratio that is based on Cummins and Violante's (2002) measure of real nonfarm private business GDP, excluding housing, and their estimate of the real capital stock.

Capital Stock

The capital stock series in this article is based on Kendrick (1961) and on the official fixed durable asset series published by the BEA. Capital in Figure 1 is real total private capital; in Figure 3, it is real nonfarm private capital; and in Figure 4, it is real (nominal) nonfarm private capital excluding residential structures. From 1889 to 1953, Kendrick (1961) provides constant 1929 dollar estimates of farm structures and land, nonresidential structures and equipment, and residential structures.

Total capital is the sum of all these capital stocks. From 1929 to 2001, the BEA provides data on current and chained 1996 dollar estimates of capital stocks and depreciation for agricultural equipment (tractors and other farm machinery) and structures, nonresidential equipment (information technology and software, transportation equipment, etc.) and structures, and residential equipment and structures. The quantity index for individual asset classes is constructed based on the perpetual inventory method using deflated investment expenditures and estimates of the depreciation pattern for that class. The current dollar estimate is then calculated as the quantity index for the stock evaluated at current prices. To the extent that I have to construct quantity indexes from the BEA data, I construct them as Divisia indexes from the available current dollar and chained dollar estimates.

Real capital stock estimates by Kendrick (1961) and the BEA are quantity indexes of the real value of capital. That is, the components of the aggregate quantity index are weighted using their asset values. For the nominal capital-GDP ratios in Figure 4, Panel B, the nominal asset value series for capital are appropriate. From the point of view of production theory, one would prefer a quantity index that uses factor rental weights for the different assets in the aggregation procedure for the real capital-GDP ratios in Figures 1, 3, and 4 (Panel A).[17] A capital stock index that is based on factor rentals better reflects the role of capital as an input to production and is used in total factor productivity studies (see, for example, Jorgenson, Gollop, and Fraumeni 1987). A problem with this approach is that we do not have observations on the average factor rentals of individual asset categories. The usual procedure is then to impute factor rentals based on required returns on capital using a version of equation (17) (Hall and Jorgenson 1967).

For the long time series in Figures 1 and 3, I splice the constant 1929 dollar capital stock series from Kendrick (1961) with the chained 1996 dollar capital stock series from the BEA in 1929 at the level of 1929 capital stock. From 1929 to 1954, the Kendrick and BEA series overlap and do not behave very differently.

I also display in Figure 4 the equipment capital-GDP ratio based on Cummins and Violante's (2002) estimates of the real equipment capital stock. For the nominal capital-GDP ratio in Figure 4, Panel B, I evaluate Cummins and Violante's (2003) estimate of the real equipment capital stock using their price deflator for new equipment capital.

[17] For a discussion, see Whelan (2002).

Depreciation

The depreciation rates are based on official fixed durable asset series published by the BEA. The depreciation rates in Figure 5 are the ratio of nominal depreciation to nominal capital stock. The alternative calculation of depreciation rates as the ratio of constant-dollar depreciation to constant-dollar capital stocks yields slightly lower depreciation rates, but they are also increasing from the mid-1980s on.

Prices

For the period 1929 to 2001, the BEA provides current and chained 1996 estimates of personal consumption expenditures (PCE) for nondurable goods and services separately. I aggregate these to a nondurable goods and services index—consumption price index for short. For the same time period the BEA also provides price indexes for new investment in equipment and software and nonresidential structures. We can compare these price indexes with the capital price indexes implied by the current value and quantity indexes for the capital stock. Although the new investment price index and the implied capital price index are not the same, they follow each other closely, more so for structures than for equipment. The relative price of capital in Figure 1 is the implied capital price index for total private capital relative to the consumption price index. The relative capital prices in Figure 2 are the implied capital price indexes of nonfarm equipment and software and structures relative to the consumption price index. Figure 2 also displays an alternative relative price index for equipment and software constructed by Cummins and Violante (2002). Following the work of Greenwood, Hercowitz, and Krusell (1997), Cummins and Violante (2002) essentially extrapolate the quality adjustments of Gordon (1990) for producer-durable equipment prices from the time period 1947–1983 to the period 1983–2001.

REFERENCES

Balke, Nathan S., and Robert J. Gordon. 1989. "The Estimation of Prewar Gross National Product: Methodology and New Evidence." *The Journal of Political Economy* 97 (February): 38–92.

Cummins, Jason G., and Violante, Giovanni L. 2002. "Investment-Specific Technical Change in the United States (1947–2000): Measurement and Macroeconomic Consequences." *Review of Economic Dynamics* 5 (April): 243–84.

Doms, Mark E., Wendy E. Dunn, Stephen D. Oliner, and Daniel E. Sichel. 2003. "How Fast Do Personal Computers Depreciate? Concepts and New Estimates." Federal Reserve Bank of San Francisco Working Paper 2003–20.

Gallman, Robert E. 1986. "The United States Capital Stock in the Nineteenth Century." In Stanley L. Engerman and Robert E. Gallman, *Long-Term Factors in American Economic Growth*. (Studies in Income and Wealth 51): 65–206. Chicago: University of Chicago Press.

Gordon, Robert J. 1990. *The Measurement of Durable Goods Prices.* The University of Chicago Press: Chicago.

Greenwood, Jeremy, Hercowitz, Zvi, and Krusell, Per. 1997. "Long-Run Implications of Investment-Specific Technological Change." *The American Economic Review* 87(3): 342–62.

Hall, Robert E., and Dale W. Jorgenson. 1967. "Tax Policy and Investment Behavior." *The American Economic Review* 57 (June): 391–414.

Jorgenson, Dale W., Frank Gollop, and Barbara Fraumeni. 1987. *Productivity and U.S. Economic Growth.* Cambridge: Harvard University Press.

Kaldor, Nicholas. 1957. "A Model of Economic Growth." *The Economic Journal* 67 (December): 591–624.

____________. 1961. Ed. F.A. Lutz and D.C. Hague. "Capital Accumulation and Economic Growth." *The Theory of Capital.* London: Macmillan:177–222.

King, Robert G., Charles I. Plosser, James H. Stock, and Mark W. Watson. 1991. "Stochastic Trends and Economic Fluctuations." *The American Economic Review* 81 (September): 819–40.

Kendrick, John W. 1961. *Productivity Trends in the United States*. Princeton: Princeton University Press.

Kongsamut, Piyabha, Sergio Rebelo, and Danyang Xie. 2001. "Beyond Balanced Growth." *Review of Economic Studies* 60 (October): 869–82.

Kuznets, Simon S. 1961. *Capital in the American Economy: Its Formation and Financing.* Princeton: Princeton University Press.

____________. 1971. *Economic Growth of Nations—Total Output and Production Structure*. Cambridge: Harvard University Press.

Maddison, Angus. 1991. *Dynamic Forces in Capitalist Development—A Long-Run Comparative View.* Oxford: Oxford University Press.

Seskin, Eugene P. and Robert P. Parker. 1998. "A Guide to the NIPA's." *Survey of Current Business* (March): 26–48.

Siegel, Jeremy L. 1998. *Stocks for the Long Run.* 2nd ed. New York: McGraw-Hill.

U.S. Department of Commerce, Bureau of Economic Analysis. *National Income and Product Accounts of the United States, 1927–1997.* At http://www.bea.doc.gov/bea/an/nipaguid.pdf

U.S. Department of Commerce, Bureau of Economic Analysis. *Fixed Assets and Consumer Durables in the United States, 1925–97* (2003). At http://www.bea.doc.gov/bea/dn/Fixed_Assets_1925_97.pdf

Whelan, Karl. 2002. "Computers, Obsolescence, and Productivity." *Review of Economics and Statistics* 84 (August): 445–61.

———. 2003. "A Two-Sector Approach to Modeling U.S. NIPA Data." *Journal of Money, Credit and Banking* 35 (August): 627–56.

Auditing and Bank Capital Regulation

Edward Simpson Prescott

Capital regulations for banks are based on the idea that the riskier a bank's assets are, the more capital it should hold. The international 1988 Basel Accord among bank regulators set bank capital requirements to be a fixed percentage of the face value of assets. The only risk variation between assets was based on easily identifiable characteristics, such as whether it was a commercial loan or a government debt.

The proposed revision to the Accord, commonly called Basel II, is an attempt to improve upon the crude risk measures of the 1988 Accord. Under Basel II, banks use their internal information systems to determine the risk of an asset and report this number to regulators.[1] In an ideal sense, the proposal is eminently sensible. After all, who knows the risks of a bank's asset better than the bank itself? But a serious problem exists in implementation. What incentive does a bank have to report the true risks of its assets? Without adequate supervision and appropriate penalties, the answer is, "Not much."

Analysis of Basel II has been primarily focused on setting the capital requirements, commonly referred to as Pillar One of the proposal. But good capital requirements mean little if they cannot be enforced. For this reason, more attention needs to be focused on Pillar Two of the proposal, that is, supervisory review.[2] This pillar gives supervisors the authority to enforce compliance with the Pillar One capital requirements, and while not usually the focus of Basel II, it is fundamental to the success of the project.

I would like to thank Rafael Repullo, Pierre Sarte, Javier Suarez, John Walter, John Weinberg, and seminar participants at CEMFI for helpful comments. This article was prepared while I was visiting CEMFI. 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] Technically, in the proposed U.S. implementation, banks will use their internal systems to estimate several key numbers—like the probability of default and the loss given default. Banks then enter these numbers into a regulatory formula to determine capital requirements.

[2] The third and final pillar of Basel II is concerned with market supervision.

These issues are examined in models where regulatory audits affect the incentives for banks to send accurate reports. By the term "audit" we mean the process of determining if the reported number is accurate. In practice, our use of the term "audit" refers more to a supervisory exam than to an external audit, though our models are broad enough to incorporate this activity, too.

The models have strong implications for how supervisors should deploy their limited resources when examining banks. We find that stochastic auditing strategies are more effective than deterministic ones. Furthermore, the frequency of an audit should depend on the amount of capital held. The less capital a bank holds, the more frequent the audits need to be, even though the safest banks hold the least amount of capital. The reason for this counterintuitive result is that audits *prevent* risky banks from declaring that they are safe banks. Therefore, the safer a bank claims to be, the more prevention is needed and the more frequently it is audited.

1. THE MODEL

Verifying the risk of a bank's investment requires a model that illustrates the role of examinations and monitoring. The simplest model sufficient for the purposes of this study is the costly state verification model of Townsend (1979). In his model, a firm's cash flow is the information to be verified. Here, it will be the risk of a bank's investment. We study capital regulations in four variants of the basic model: an idealized one where the regulator observes the bank's risk characteristics; one where the regulator does not observe the risk characteristics; another where the regulator can audit deterministically to find out the risk characteristics; and a final model where the regulator may randomly audit, that is, conduct an audit with a probability anywhere between zero and one.

The Basic Model

In the model, there is one regulator and many small banks. Each bank has one investment of size one. Investments either succeed or fail. All successful investments return the same amount, and all failed investments produce zero. Banks' investment projects differ only in their probability of failure. The probability of a bank's investment failing is p, which lies in the range, $[\underline{p}, \bar{p}]$, with $0 \leq \underline{p} < \bar{p} \leq 1$. This probability is random to the bank and drawn from the density function, $h(p)$. The cumulative distribution function is $H(p)$. Shocks are independent across banks.

A bank's investment can be financed with either deposits or capital. Banks prefer less capital to more. For the moment, there is no need to be specific about the details of this preference. We only need banks to desire to hold less capital than the regulator wants them to. Such a desire by banks could come out of a model with a deposit insurance safety net or any model in which equity capital is costlier to raise than deposits. Let $K(p)$ be the amount of capital held by a bank with investment opportunity, p. Because each bank is of size one, $1 - K(p)$ is the amount in deposits each bank holds as well as being its *utility*.[3]

The regulator cares about losses from failure and the cost of capital. We assume that the failure losses depend on the amount of deposits that the regulator needs to cover in case there is failure. This function is $V(K(p))$ with V increasing and concave ($V' > 0$ and $V'' < 0$). Because V measures losses, we assume that $V(K(p)) \leq 0$ for all values of capital, with $V(1) = 0$ (see Figure 1). The regulator suffers no losses from a failed bank if it has 100 percent capital. The purpose of this function is to generate a desire on the part of the regulator for banks with riskier portfolios to hold more capital.

The regulator also cares about the cost of capital. Assuming that the per unit cost is q, this cost represents the foregone loss of liquidity services from a bank's use of capital rather than deposits.[4]

The problem for the regulator is to choose a *risk-based capital requirement*, $K(p)$, that balances the regulatory benefit of reducing losses of failure with the costs to banks of issuing capital. This problem is the maximization problem:

$$\max_{K(p)\in[0,1]} \int_{\underline{p}}^{\bar{p}} (pV(K(p)) - qK(p))dH(p).$$

The term $pV(K(p))$ is the expected failure loss to the regulator from a bank with risk p, while $qK(p)$ is the cost to the bank of raising capital.

It is straightforward to solve this problem. We assume that the solution is interior, so the first-order conditions are

$$\forall p, \quad pV'(K(p)) = q. \tag{1}$$

The expected marginal benefit of capital is set equal to the marginal cost of capital. Equation (1) implies that $K(p)$ increases with p. As the probability of failure grows, the regulator increases the capital requirement. For example, if $V(K) = -(1 - K)^{\alpha}$ with $\alpha > 1$, then (1) takes the simple form,

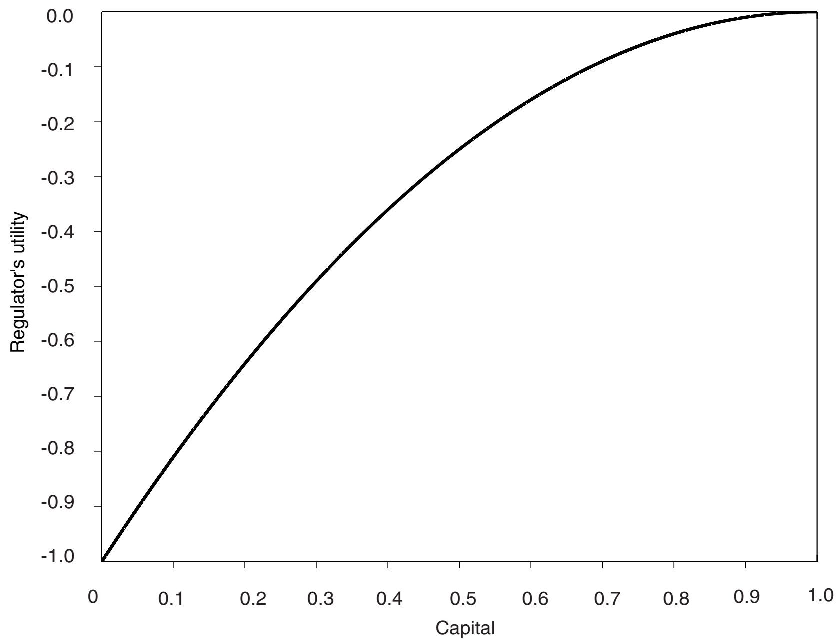
$$K(p) = 1 - (\frac{q}{\alpha p})^{1/(\alpha-1)},$$

assuming q and the range of p are such that $0 \leq K(p) \leq 1$ (see Figure 2). The positive relationship between default probability, p, and capital, K, is the goal of both the Basel I and II regulations.

[3] A bank's preferences over K are independent of its risk, p. Banks always prefer less capital to more. This assumption is strong, but it simplifies the analysis in several advantageous ways.

[4] We decided to model bank's preferences over capital by $1-K$ rather than formally including the cost of capital because it simplifies the algebra. This modeling decision has no impact on the article's results because the important feature is that the bank prefers less capital to more.

Figure 1 Example of $V(k)$ Function



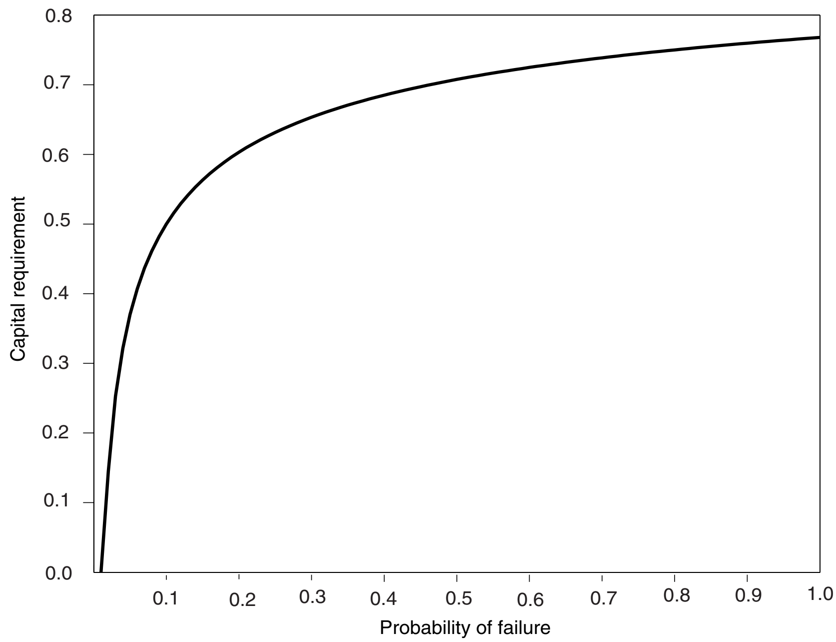
Notes: Figure 1 illustrates an example of the regulator's utility from failure losses as a function of a bank's capital given p, that is, $V(K)$. The more capital a bank holds, the less the loss to the regulator. The function is non-positive, increasing, and concave.

Private Information

The *fundamental* problem for Basel I and II is to determine the risk of a bank's assets. The premise of the Basel II reform is that a bank has the best information on its own assets so that by using its internal models and data, a regulator can get a better estimate of its risks than from the crude measures underlying Basel I. The problem for Basel II is that a bank has an incentive to understate the risk as long as it wants to save on capital costs.

For illustrative purposes, we start with the extreme assumption that the regulator knows almost nothing about the riskiness of a bank's investment opportunities except that the distribution of these risks is $H(p)$. Each bank, however, knows its own risk; that is, it has *private information*. Now, how should the regulator set capital requirements? The regulator would like to use the capital requirements illustrated in Figure 2, but that would be a disaster. Each bank would say that it was the safest bank; that is, report $\underline{p}$ to get the low capital of $K(\underline{p})$. All banks would do this, and there would be nothing

Figure 2 Full-Information Regulatory Capital Schedule



Notes: Figure 2 illustrates optimal regulatory capital as a function of bank risk when the regulation knows the bank's risk.

the regulator could do afterwards. The result for the regulator would be huge losses.

Instead, the regulator should design a capital schedule that takes into account each bank's private information. The effect of private information is modeled with an incentive constraint that says a capital schedule is only feasible if it is in the interest of a bank to report its risk truthfully.[5] Formally, the incentive constraint is

$$\forall p, \hat{p}, \quad 1 - K(p) \geq 1 - K(\hat{p}),$$

or, equivalently,

$$\forall p, \hat{p}, \quad K(p) \leq K(\hat{p}). \tag{2}$$

[5] The Revelation Principle is being used here.

This constraint says that the utility a bank with failure risk, p, receives from $K(p)$ is at least as much as it would receive if it claimed to have any other failure risk, $\hat{p}$.

The 1988 Basel Accord

The incentive constraint, (2), is very stringent, eliminating most capital schedules. The *only* schedules that satisfy it are those where $K(p)$ is a constant. If $K(p)$ varies with p at all, a bank assigned a higher $K(p)$ would simply claim that its assets are a risk that receives the lowest capital charge under the capital schedule. Consequently, all bank investments must face the same capital charge, regardless of how risky their portfolios are. Indeed, this lack of responsiveness of capital charges to risk looks exactly like the Basel Accord of 1988 as applied to assets within a particular risk class. For example, a commercial and industrial loan with a 10 percent chance of default is treated the same as one with a 2 percent chance of default.

It is precisely this equal treatment of different risks that has led to the development of Basel II. Basel II distinguishes between the riskiness of loans—the ps in the model—by allowing banks to report the risk characteristics of their loans. This is an admirable goal, as represented by (1), but in light of the incentive constraint (2), it is not attainable. That constraint says there can be no risk variation in capital requirements.

Something else is needed to make Basel II work. As will be discussed in the next section, that "something else" is audits and penalties. Unfortunately, these critical features are not usually discussed in the context of Basel II.

2. A ROLE FOR AUDITS

Risk-sensitive capital requirements could be implemented if the regulator could gather some information about the true risk of the investments. We assume that the regulator, devoting m units of resources, can observe a bank's risk characteristics. Other cost functions are possible. Indeed, some activities pose greater difficulty in gathering information than others do. Still, the fixed cost function is the simplest to study and illustrates the main points, so we will use it.

Audits are performed after the bank reports to the regulator on the risk characteristics of its investments. For the moment, we assume that auditing is deterministic; that is, in response to a particular report the regulator must either audit or not audit. Later, we will extend the model to allow the regulator to audit with some probability.

If an audit is performed and the bank is found to have misrepresented its asset risk, the regulator may impose a penalty. We model this penalty as a fixed utility amount, u. The utility of an audited bank found to have lied is $1 - K(p) - u$.

The addition of audits requires a slight modification to the regulator's decision problem and to the incentive constraints. Now the regulator must decide which reports of p to verify with an audit and which not to. Let A be the region of $[\underline{p}, \bar{p}]$ for which the regulator audits and N the region for which it does not. There are two sets of incentive constraints. The first set concerns misrepresentations in the no-audit region. These incentive constraints are

$$\forall p, \quad 1 - K(p) \geq 1 - K(\hat{p}), \quad \forall \hat{p} \in N$$

or, equivalently,

$$\forall p, \quad K(p) \leq K(\hat{p}), \quad \forall \hat{p} \in N. \tag{3}$$

Incentive constraints (3) state that a bank's capital must be less than it would receive if it claimed to have a p in the no-audit region, N. Like the earlier incentive constraints (2), these incentive constraints strongly restrict feasible allocations. However, the restriction only applies to p in the non-auditing region, N, so capital must be a constant *only* over this region. We refer to this amount of capital as K_N.

The second set of incentive constraints prevents misrepresentations in the audit region. These incentive constraints are

$$\forall p, \quad 1 - K(p) \geq 1 - K(\hat{p}) - u, \quad \forall \hat{p} \in A,$$

or, equivalently,

$$\forall p, \quad K(p) \leq K(\hat{p}) + u, \quad \forall \hat{p} \in A. \tag{4}$$

These incentive constraints are usually less important than (3). As long as u is high enough, they will be automatically satisfied.

To summarize, the main difference between the earlier model and the deterministic auditing model is the severity of the incentive constraints. In the earlier model, (2) forces the capital requirement to be the same for all risks while in the deterministic auditing model, (3) forces the capital requirement to be the same *only* for risks in the non-auditing region.

Even before writing out the program, two properties of optimal capital requirements can be derived. The first follows from (3). Because banks can always claim that their failure probability is some p in the non-auditing region, we know that

Proposition 1 $K(p) \leq K_N$.

The second proposition that we can prove is that the non-auditing region is convex and consists of the highest risk banks. This proposition will let us formalize the regulator's problem in a simple way.

Proposition 2 *The non-auditing region, N, is convex and consists of the highest risk banks.*

We do not provide a formal proof. Conceptually, the idea is simple. Assume that there is an audited bank that is riskier than some non-audited bank (and for simplicity both are equal fractions of the bank population). By Proposition 1, the non-audited bank holds more capital. Now, switching their regulatory requirements—switching the amount of capital each holds—and auditing the safe bank but not auditing the riskier bank satisfies the incentive constraints. It also increases the utility of the regulator since the capital is more effective when deployed against the risky bank rather than the safer bank.

These properties can be incorporated when formulating the regulator's problem. Let a be the cutoff between audited and non-audited banks. The regulator's program is:

Regulator's Program with Deterministic Auditing

$$\max_{a,K_N,K(p)} \int_{\underline{p}}^{a} (pV(K(p))-m-qK(p))dH(p)+\int_{a}^{\bar{p}} (pV(K_N)-qK_N)dH(p),$$

subject to the incentive constraints

$$\forall p < a, \quad K(p) \leq K_N \tag{5}$$

and (4).

For the purpose of our analysis, we are going to assume that the penalty u is high enough so that (4) does not bind. Furthermore, when we take the first-order conditions, we are going to ignore the incentive constraint (5) and show that the solution to the program without it still satisfies it. This property does *not* mean that the private information does not matter in this problem. Instead, it means that setting up the problem with a cutoff between the auditing and non-auditing regions and with constant capital in the non-auditing region is enough for incentive compatibility to hold.

The derivative with respect to K_N is

$$V'(K_N)\int_{a}^{\bar{p}} pdH(p) = q\int_{a}^{\bar{p}} dH(p). \tag{6}$$

The first-order conditions with respect to $K(p)$ are

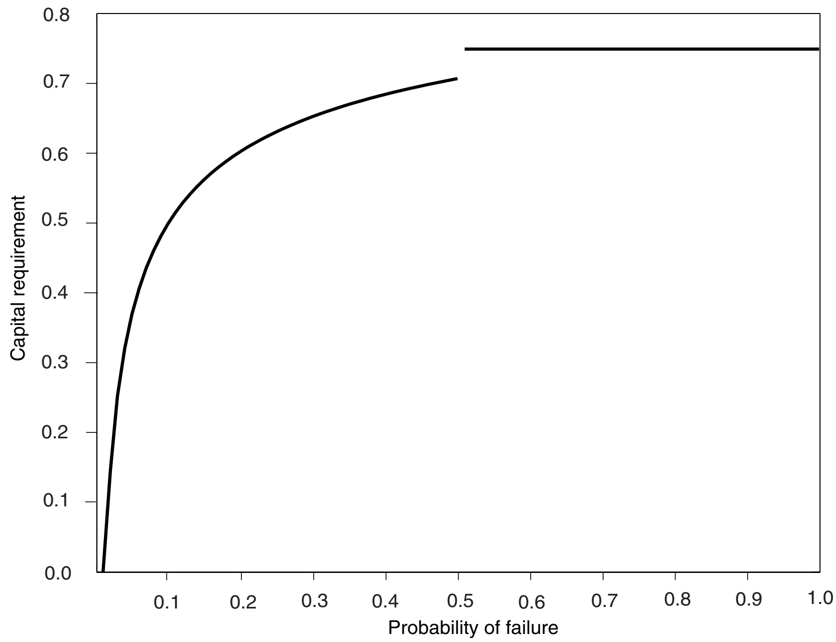
$$\forall p < a, \quad pV'(K(p)) = q. \tag{7}$$

Again, we assume that the solutions are interior.

Two properties of a solution follow from these two constraints. First, from (7), we know that $K(p)$ is increasing in p for $p \in A$. Second, there is a discontinuity in $K(p)$ at the cutoff a. Let $\tilde{K}(a) = \lim_{p\to a} K(p)$. Taking the limit of (7) at $p = a$ and substituting for q in (6) delivers

$$V'(K_N)E(p|p \geq a) = aV'(\tilde{K}(a)), \tag{8}$$

Figure 3 Optimal Regulatory Capital with Deterministic Auditing



Notes: Figure 3 illustrates optimal regulatory capital when banks have private information about their true risks and the regulator may undertake deterministic audits. The schedule is discontinuous at the point where the regulator stops auditing banks. The horizontal portion corresponds to the capital holdings of the risky banks, that is, K_N, none of which are audited.

where

$$E(p|p \geq a) = \frac{\int_a^{\bar{p}} p dH(p)}{\int_a^{\bar{p}} dH(p)}.$$

Because a is less than the average probability of failure in N, that is, over the range a to $\bar{p}$, (8) implies that $V'(\tilde{K}(a)) > V'(K_N)$, which, in turn, implies that $\tilde{K}(a) < K_N$. Thus, $K(p)$ is discontinuous at a. Furthermore, this result proves that constraint (5) is redundant.

The intuition for the discontinuity is that for $p \in A$, $K(p)$ is set as in the full-information problem, where (7) is satisfied when the marginal benefit of capital equals its marginal cost. But for $p \in N$, $K(p)$ is a constant, so K_N is set to equalize the *expected* marginal benefit of capital with its marginal cost. Figure 3 illustrates what a capital schedule might look like.

The final first-order condition is taken with respect to the cutoff point, a. It is

$$(aV(K_N) - (aV(K(a)) - m)) - q(aK_N - aK(a)) = 0.$$

Canceling terms and rearranging gives

$$aV(K_N) + qK(a) + m = aV(K(a)) + qK_N. \qquad (9)$$

The left-hand side of equation (9) is the marginal cost of increasing the cutoff point, and the right-hand side is the marginal benefit.

Back to Basel

The model's implications for capital regulation are very strong and, at first glance, counterintuitive. The highest risk banks do not need to be audited. Only banks that want to hold less capital than the maximal amount are audited. This result, however, should not be surprising since, for incentive reasons, there is no need to audit a bank willing to hold the maximal amount of capital. Indeed, if regulators have a maximum amount of risk, p, they are willing to allow banks to take, and assuming they have the power to shut down banks, they would have to audit every bank in operation.

The model demonstrates just how fundamental auditing and the penalties are to regulatory policy. Risk-sensitive regulation requires auditing of any bank holding less than the largest amount of capital. Presumably, this result would include most banks and likely would cause high auditing costs, which seems problematic. Fortunately, other regulatory policies may still implement risk-sensitive capital requirements at a lower cost. In the next section, we consider such policies in a model with *stochastic* auditing.

Still, the point remains that auditing and penalties cannot be avoided. Basel II contains many details on how a bank should justify its capital ratio, but these procedures can never be perfect. If they were, we could turn over investment decisions to regulators. Basel II is premised on the belief that banks know their risks better than regulators, and while regulators can gather some information on these risks, they can never know as much as the bank. For this reason, the incentive concerns detailed above are unavoidable.

3. STOCHASTIC AUDITING

In this section, we modify the model so that the decision to audit by the regulators can be stochastic. By stochastic we mean that in response to a bank's risk report, the regulator may audit with some probability. As we will see, this policy saves on supervisory resources. As before, we will assume that these audits fully reveal the information. Alternatives can be studied. For example, the regulator could observe only a signal correlated with the true risk, or the quality of the signal could depend on the intensity of the audit.

Stochastic auditing requires making a few changes to the model. First, we drop the distinction between the auditing and non-auditing regions. Let $\pi(p)$ be the probability of an audit, given that p is reported. As before, m is the cost of an audit, and u is the utility penalty that is imposed if a bank is found to have lied. The regulator's program is:

Regulator's Program with Stochastic Auditing

$$\max_{K(p)\in[0,1],\pi(p)\geq 0} \int_{\underline{p}}^{\bar{p}} (pV(K(p)) - \pi(p)m - qK(p))dH(p)$$

subject to the incentive constraint

$$1 - K(p) \geq 1 - K(\hat{p}) - \pi(\hat{p})u, \quad \forall p, \hat{p}. \tag{10}$$

Incentive constraint (10) differs from the deterministic case incentive constraints (3) and (4) in that $\pi(p)$ can take on any value from zero to one.

There are many incentive constraints in (10), but, fortunately, most of them are redundant. Notice that utility is decreasing in $K(p)$, and utility from reporting the wrong p does not depend on a bank's risk type. Therefore, if the incentive constraint holds for the type with the highest capital charge—for now, assume that it is the highest risk bank $\bar{p}$—then the incentive constraint holds for all other risk types. Formally, (10) can be replaced by

$$K(\bar{p}) \leq K(p) + \pi(p)u, \ \forall p. \tag{11}$$

Another simplification is possible. Audits are a deadweight cost, so it is best to minimize their probability. For a given capital schedule, the audit probabilities are minimized when (11) holds at equality. Therefore,

$$\pi(p) = \frac{K(\bar{p}) - K(p)}{u}. \tag{12}$$

We hinted above that the highest risk bank would be the type to hold the greatest amount of capital. This is intuitive, but it can be proven. Imagine that the bank assigned the highest amount of capital is not the highest risk one. For simplicity, assume that all types of banks occur with equal probability. Then, simply switch the capital requirement faced by the highest risk bank and the one holding the most capital. Incentive compatibility still holds and the regulator's objective function is higher since the highest risk bank holds more capital.

We could substitute (12) directly into the objective function, but for optimization purposes it is more convenient to consider

$$\pi(p) = \frac{\bar{K} - K(p)}{u} \tag{13}$$

and require that $K(p) \leq \bar{K}$, for all value of p. Equation (13) will be substituted into the objective function, and we will make $\bar{K}$ a choice variable.

As long as the solution has $K(p) \leq K(\bar{p})$, auditing probabilities will be non-negative. Furthermore, because auditing is a deadweight cost, any solution will necessarily set $\bar{K} = K(\bar{p})$. With these changes, the program is:

Simplified Regulator's Program with Stochastic Auditing

$$\max_{K(p)\in[0,1],\bar{K}} \int_{\underline{p}}^{\bar{p}} \left(pV(K(p)) - \frac{\bar{K} - K(p)}{u} m - qK(p) \right) dH(p)$$

subject to

$$\forall p, \quad K(p) \leq \bar{K}. \tag{14}$$

Even before studying the first-order condition, the solution has the following properties from (13) and the desire to lower $\bar{K}$. First, the probability of an audit is zero for any bank that holds the highest amount of capital. Second, the audit probability increases as capital declines.

The first set of first-order conditions for this problem is

$$\forall p, \quad (pV'(K(p)) + m/u - q) = \lambda(p), \tag{15}$$

where $\lambda(p)h(p) \geq 0$ is the Lagrangian multiplier on (14) for p. The remaining first-order condition is

$$m/u = \int_{\underline{p}}^{\bar{p}} \lambda(p) dH(p). \tag{16}$$

We already demonstrated that only the highest risk banks hold the greatest amount of capital. Therefore, $K(p) \leq K(\bar{p})$. For any bank with $K(p) < K(\bar{p})$, $\lambda(p) = 0$, so (15) implies that capital is increasing in risk in this range.

The first-order conditions can be used to derive two additional properties of a solution:

Proposition 3 *A range of banks at the upper tail of the distribution (more formally a range with positive measure) holds $K(\bar{p})$.*

This proposition is equivalent to showing that there is a range of p for which constraint (14) binds. A proof is contained in the Appendix.

The second result is differs from that of the deterministic auditing case.

Proposition 4 *The capital schedule $K(p)$ is continuous.*

This proof is also in the Appendix.

The properties of the stochastic auditing model are illustrated with an example. We also calculated the optimal deterministic auditing contract to compare the two. The example used the following parameter values: $h(p)$

is a uniform distribution over the range $\underline{p} = 0.1$, and $\bar{p} = 0.5$, $V(K) = -1.5(1-K)^2$, $m = 0.01$, $u = 1.0$, and $q = 0.5$.

Figure 4 illustrates optimal capital requirements under deterministic and stochastic auditing. The schedule for the deterministic case has a discrete jump at the non-audit point. The schedule for the stochastic case is continuous. In the deterministic case, there is a much bigger range of p for which capital is flat. Capital requirements are, necessarily, less finely tuned in this case. Also, for p in the audit range (roughly between 1.0 and 1.5), $K(p)$ is slightly smaller under deterministic auditing than under stochastic auditing. This difference comes from comparing the two problems' first-order conditions. Condition (15) has an additional term m/u that is not in (7). This term makes $K(p)$ higher in this range.

Figure 5 illustrates the audit probabilities for both models. Of course, the deterministic case probabilities are either zero or one. Probabilities for the stochastic case move smoothly and hit zero for the risk types that hold the highest amount of capital. As capital declines, audit probabilities increase. Finally, the stochastic auditing case saves on auditing resources. In the deterministic case, banks are audited 15.5 percent of the time and 13.7 percent of the time in the stochastic case.

The differences in the two types of arrangement are evident in the figures. Stochastic auditing is, of course, more efficient. Here it allows for more finely tuned capital requirements and uses less auditing resources.

4. CONCLUSION

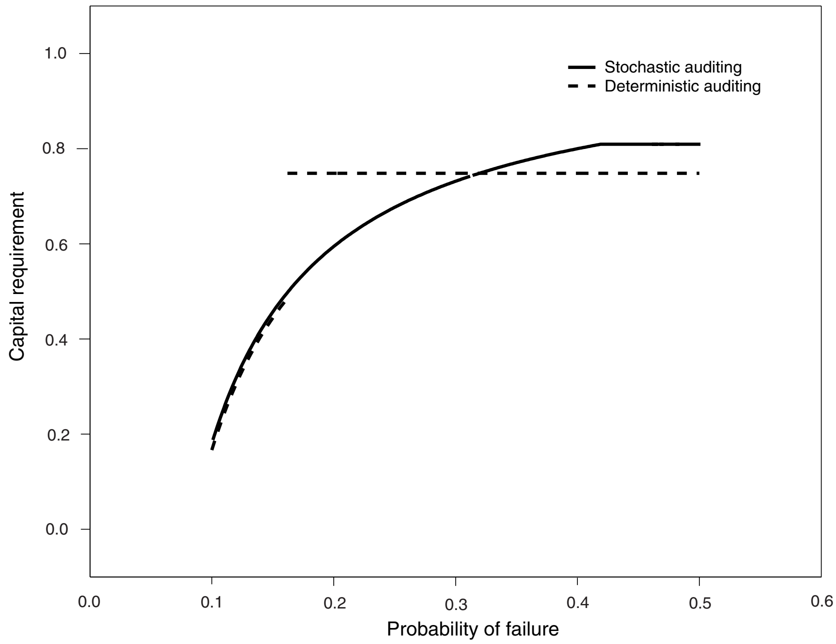
Banks know their own risks better than regulators. Basel II is based on the premise that these risks can be communicated by banks to regulators and then used to determine regulatory capital. But with this informational advantage, banks can control precisely what is communicated. For this reason, it is necessary to consider the incentives banks have for truthfully reporting their risks. This article argues that the penalties or sanctions imposed for noncompliance are critical for determining these incentives. Basel II is, unfortunately, relatively silent on this issue. As Basel II is adopted and implemented, these issues will have to be addressed.

The models developed in this article not only illustrate the role of penalties, but also illustrate various supervisory strategies for gathering information and imposing sanctions. Supervisory resources are scarce and costly. Therefore, finding the best way to deploy them is valuable. The stochastic auditing model demonstrates that randomized audits, or exams, could improve upon regularly planned audits.[6]

---

[6] Audits may be made to depend on other signals. Marshall and Prescott (2001) analyze a model where regulatory sanctions depend on the realization of bank returns.

Figure 4 Capital Requirements

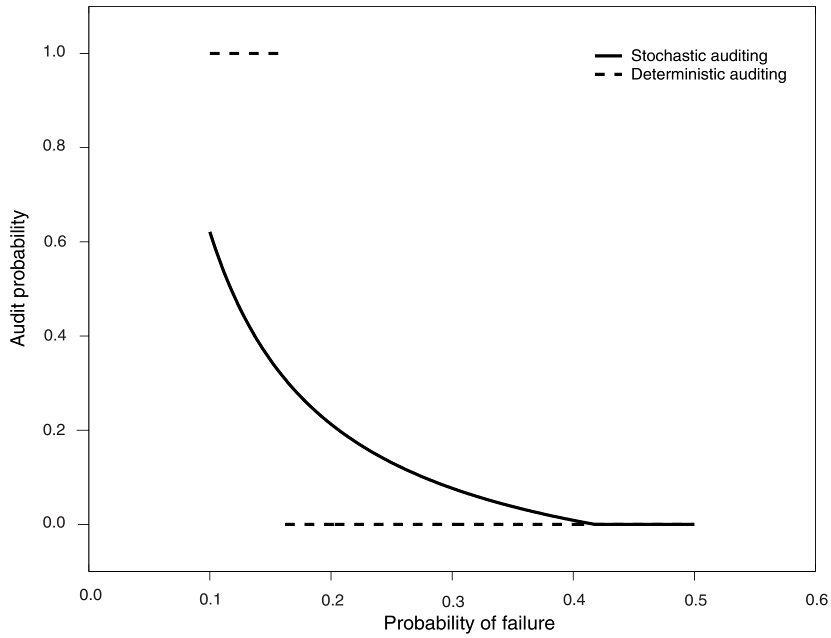


Notes: Figure 4 describes optimal capital requirements for both the deterministic and stochastic auditing cases. Where the discrete jump for the deterministic auditing case occurs is the point where the regulator stops auditing. Where the regulator audits in the deterministic case, the capital schedule is slightly lower for the deterministic case than in the stochastic case.

In the models, audit frequencies and capital requirements are inversely related. Less capital requires more frequent auditing for incentive reasons, implying, counterintuitively, that the safest banks are audited the most. The reason for this regulatory behavior is that the role of audits is to prevent risky banks from claiming to be safer than they really are. Because no one wants to claim to be riskier than they actually are, auditing a bank that claims it is the highest risk is unnecessary. This bank has agreed to hold more capital, and that is all the regulators desire.

The precise relationship between audit frequencies and capital requirements depends on parameters such as available penalties, auditing costs, the costs of capital, and the distribution of bank risk types. If these parameters differ between countries, then there should be different capital schedules in each country. Harmonization of regulations is not without its costs.

Figure 5 Audit Probabilities



Notes: Figure 5 describes optimal audit probabilities as a function of bank risk type for both the deterministic and stochastic cases. By necessity, the deterministic case probabilities are either zero or one. The probabilities vary smoothly for the stochastic case.

The models developed in this article omit other relevant dimensions to the problem. For example, audits are not perfect. Sometimes the information gathered is incorrect. One way to incorporate these important factors is to allow regulators to observe only a signal correlated with the true state. Other possibilities include making it costly for banks to hide information, e.g., Lacker and Weinberg (1989). Another important extension is to consider dynamic capital schedules. Supervisors interact over time with banks and may have latitude to generate the equivalent of penalties through their future treatment of the bank. The literature on dynamic costly state verification models should be relevant here and includes Chang (1990), Smith and Wang (1998), Monnet and Quintin (2003), and Wang (2003).

APPENDIX

Proposition 3 *There is a range of banks at the upper tail of the distribution (more formally a range with positive measure) that hold* $K(\bar{p})$.

If only the highest risk bank, $\bar{p}$, holds the greatest amount of capital, then $\lambda(p) = 0$ for all $p < \bar{p}$. But then $\int_{\underline{p}}^{\bar{p}} \lambda(p)h(p) = 0$, which contradicts (16). Therefore, $\lambda(p) > 0$ for a range of p with positive measure. These values of p have to be the highest risk values. If not, consider $p_1 < p_2$ with $K(p_1) = K(\bar{p})$ and $K(p_2) < K(\bar{p})$. We know that $\lambda(p_1) \geq 0$ and $\lambda(p_2) = 0$. Using (15), we have

$$p_1 V'(K(\bar{p})) + m/u - q = \lambda(p_1) \geq \lambda(p_2) = p_2 V'(K(p_2)) + m/u - q,$$

which implies that $p_1 V'(K(\bar{p})) \geq p_2 V'(K(p_2))$. But $V'(K(\bar{p})) < V'(K(p_2))$, so $p_1 > p_2$, which is a contradiction.

Proposition 4 *The capital schedule* $K(p)$ *is continuous.*

Let $\hat{p}$ be the lowest value of p at which $K(p) = K(\bar{p})$. The capital schedule is clearly continuous above and below this point. Take the limit of $K(p)$ as p approaches $\hat{p}$ from below. Call this limit $\tilde{K}(\hat{p})$. Evaluating (15) at the limit gives

$$(\hat{p} V'(\tilde{K}(\hat{p})) + m/u - q) = 0.$$

If $K(p)$ is not continuous at $\hat{p}$, then $K(\hat{p}) = K(\bar{p}) > \tilde{K}(\hat{p})$, which implies that

$$\lambda(\hat{p}) = (\hat{p} V'(K(\bar{p})) + m/u - q) < 0.$$

But $\lambda(\hat{p}) < 0$ is a contradiction, so $K(p)$ is continuous at $\hat{p}$ as well.

REFERENCES

Chang, Chun. 1990. "The Dynamic Structure of Optimal Debt Contracts." *Journal of Economic Theory* 52 (October): 68–86.

Lacker, Jeffrey M., and John A.Weinberg. 1989. "Optimal Contracts Under Costly State Falsification." *Journal of Political Economy* 97 (December): 1345–63.

Marshall, David A., and Edward Simpson Prescott. 2001. "Bank Capital Regulation with and without State-Contingent Penalties." Carnegie-Rochester Conference on Public Policy 54 (June): 139–84.

Monnet, Cyril, and Erwan Quintin. 2003. "Optimal Contracts in a Dynamic Costly State Verification Model." Manuscript, July.

Mookherjee, Dilip, and Ivan Png. 1989. "Optimal Auditing, Insurance, and Redistribution." *Quarterly Journal of Economics* 104 (May): 399–415.

Rochet, Jean-Charles. 1999. "Solvency Regulations and the Management of Banking Risks." *European Economic Review* 43 (April): 981–90.

Smith, Bruce D., and ChengWang. 1998. "Repeated Insurance Relationships in a Costly State Verification Model: With an Application to Deposit Insurance." *Journal of Monetary Economics* 42 (July): 207–40.

Townsend, Robert M. 1979. "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21 (October): 265–93.

Wang, Cheng. 2005. "Dynamic Costly State Verification." Economic Theory 25 (June): 887–916.

Using Manufacturing Surveys to Assess Economic Conditions

Matthew Harris, Raymond E. Owens, and Pierre-Daniel G. Sarte

Starting in the 1980s, the Richmond Fed began surveying District manufacturers as input into the Bank's Beige Book reports. The effort, which mimics the Institute of Supply Management's (ISM) national survey, was undertaken because little timely information on regional manufacturing activity was available. Surveys such as the ISM's are generally used because they are thought to provide a good balance between collection effort and the information obtained. While the earliest Richmond Fed Surveys appeared to be useful gauges of activity, they had an important shortcoming. They were conducted approximately every six to seven weeks—prior to the Fed's Beige Book reports, so that the results did not coincide with the regular monthly or quarterly findings from other surveys or economic reports. This irregular timing meant that Richmond Survey results could not be easily verified against other "benchmark" data, leaving unanswered the appropriate weight to assign the information. To overcome this shortcoming, the Richmond Survey was redesigned and conducted on a monthly basis starting in November 1993.

To address this question, we examine why surveys are conducted, and what information is collected. We also examine how the Richmond Fed Survey specifically compares to other benchmarks, including the ISM and the Philadelphia Fed Business Conditions Survey, how well it gauges regional economic activity, and what improvements may be made to the Survey going forward to increase its value.

We find that the ISM is a very good gauge of national economic activity as measured by GDP. Its accuracy is highly valued by analysts because it is

We thank Andreas Hornstein, Yash Mehra, and Roy Webb for their helpful comments. In addition, we thank Judy Cox for her assistance and help. The views expressed in this article do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System. Any errors are our own.

available up to three months before final GDP data. We also find that the Richmond Manufacturing Survey—alone and when used in conjunction with the Philadelphia Fed Survey of Business Conditions—is highly correlated with the ISM. In addition, we find the Richmond Survey to be a good predictor of several important measures of Fifth District Federal Reserve regional economic activity. It follows, therefore, that the value of the Richmond Survey would increase if it were released sooner and contained an overall measure of economic activity.

1. WHY SURVEY?

Prior to the Richmond Survey, information on Fifth District manufacturing activity was available primarily from the annual Gross State Product (GSP) reports of District states as well as manufacturing employment. But the GSP data are typically released one to two years after the period covered by the report. Other information, such as manufacturing employment, is received in a more timely manner, though still with a one- to two-month lag. Since manufacturing activity has historically shown cyclical behavior, the long lag in the GSP data is problematic. With lags, the cyclical nature of manufacturing activity raises the likelihood that current conditions in manufacturing activity differ from those described in the GSP report, rendering the data useful as a historical benchmark, but sharply reducing their value in assessing current conditions.

A second alternative was the monthly survey of manufacturing conditions provided by the National Association of Purchasing Management (NAPM), now called the ISM. Although timely, the ISM Survey gauges manufacturing activity at the national level rather than at the regional level. This broad geographic coverage raises questions about the NAPM's ability to represent accurately Fifth District manufacturing activity. The Richmond Fed's Survey was undertaken to fill this gap. The information gathered is timely, but has it been accurate? To address this question, an examination of the Richmond Survey and its results follows.

2. THE RICHMOND SURVEY

The Richmond Survey is distributed to approximately 200 manufacturers in the Fifth Federal Reserve District during the second week of each month, with approximately 40 of those manufacturers also receiving the Survey by e-mail. Responses are delivered to us by mail, fax, or via the Internet where respondents can directly input their data by the deadline. Responses must be received by the cutoff date—usually the first day of the following month—and typically number about 90 to 100. After compiling the results, the Richmond

Fed places them on the bank's Web site at 10:00 a.m. on the second Tuesday of the month following the survey month.

The survey sample is designed to approximate the distribution of manufacturing output by state, industry type, and firm size. Firms possessing the desired characteristics are typically identified through industry listings or other means. Once chosen, each manufacturer is invited by mail, e-mail, or by telephone to participate. Periodically, new names are added to the sample to improve the distribution's characteristics, to replace or to enlarge the sample, or to take advantage of a particular manufacturer's offer to participate.

The first portion of the Survey asks about business activity. Each survey includes questions on shipments, new orders, backlogs, finished goods inventories, employment, average workweek, vendor lead time, capacity utilization, and capital expenditures. Manufacturers are asked whether their firms experienced an increase, decrease, or no change in a variety of activity measures in each variable over the preceding month. They are also asked whether they expect an increase, decrease, or no change in the next six months. Raw data are combined to create diffusion indexes equal to the percentage of respondents reporting increases minus the percentage reporting decreases. Diffusion indexes are a standard survey tool and are used by many agencies, including the Philadelphia and Kansas City Feds.[1]

The diffusion index used for the Richmond Survey is centered on 0, meaning that 0 infers that the level of activity is unchanged from the prior month's level. A positive reading indicates a higher level, and a negative reading infers a lower level. Greater or lesser readings compared to the prior month are interpreted as faster or slower rates of change in activity, respectively. The diffusion index is computed according to the standard form,

$$\text{Index Value} = 100(I - D)/(I + N + D), \tag{1}$$

where I is the number of respondents reporting increases, N is the number of respondents reporting no change, and D is the number of respondents reporting decreases.

Once the raw diffusion indexes are derived, seasonal adjustment factors are applied. The factors are determined from the last five years of data using the Census X-12 program.[2]

The second portion of the Survey focuses on inventory levels. Manufacturers are asked how their current inventory levels compare to their desired

[1] For a recent detailed description of the Kansas City Fed Survey, see Keeton and Verba (2004).

[2] The Richmond Survey's results are bounded between -100 and 100 by construction. It has been suggested that the results could be transformed into an unbounded series using a logit transformation procedure before being seasonally adjusted. However, a comparison of this method with the simple add-on method reveals no substantial difference in the results.

levels. They may respond too low, too high, or correct. The manufacturers are also asked a similar question about their customers' inventories.

The third portion of the Survey covers price trends. We ask manufacturers to estimate recent annualized changes in raw materials and finished goods prices and price changes expected in the next six months. We report the simple means of their responses; no seasonal adjustment factors are applied.

The most recent survey form and the most recent press release are shown in Appendixes A and B. Unlike the ISM and the Federal Reserve Bank of Philadelphia, Richmond does not publish an overall or composite business index.[3] The construction is straightforward, however, and to allow for comparability, we construct a regional business index similar to that of the ISM. Our index differs from the ISM's in two respects. First the Richmond Survey asks only three questions similar to the five asked by the ISM. Given this, our weights on the questions differ from those of the ISM. The composite index, defined by the following components and weights, is used in the next section: shipments (0.33), new orders (0.40), and employment (0.27).

Before analyzing the usefulness of the Richmond Survey specifically, we first address the design and ability of the overall ISM to capture changes in economic activity at the national level.

3. THE ISM

The ISM Survey's indexes are highly regarded by business analysts because they have proven to be a reliable gauge of economic activity over a long period. The ISM's extensive history is a result of purchasing managers' long-standing desire to obtain industry-level information. The earliest purchasing manager survey was the local New York City's association poll of its members regarding the availability of specific commodities. The survey began in the 1920s and, by the 1930s, was broadened to capture a wider range of business activity measures. Following World War II, the report assumed a format similar to the current survey instrument, asking about production, new orders, inventories, employment, and commodity prices. Beginning in the 1970s, other series were added, including supplier deliveries and new export orders, and, in the 1980s, the Purchasing Manager's Index (PMI) was developed. The PMI is a weighted average of several of the seasonally adjusted series in the ISM survey and will be referred to as the ISM index in this article. The components and their weights are production (0.25), new orders (0.30), employment (0.20), supplier deliveries (0.15), and inventories (0.10).

At present, the Survey is sent to approximately 400 purchasing managers at industrial companies across the country each month. The sample is stratified

[3] The Federal Reserve Bank of Philadelphia does not construct an index from a weighted average of several questions. Rather, the survey directly asks about business conditions.

Figure 1 GDP Growth Rate

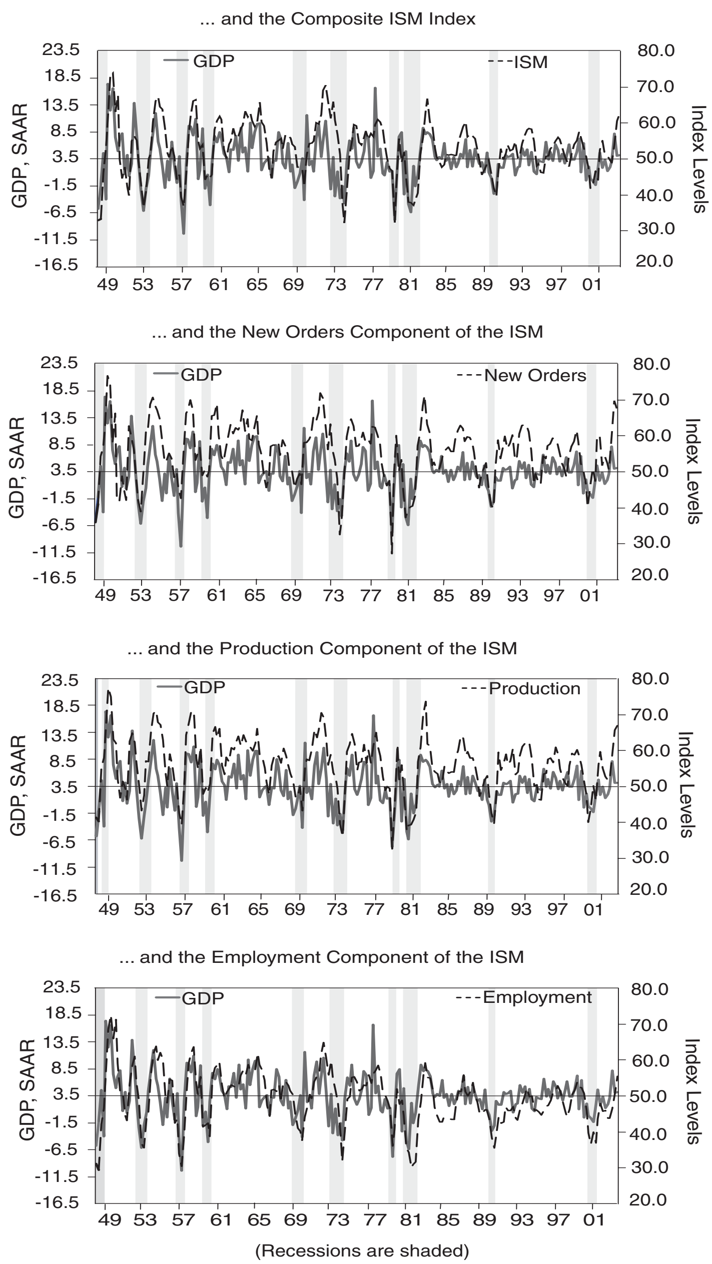
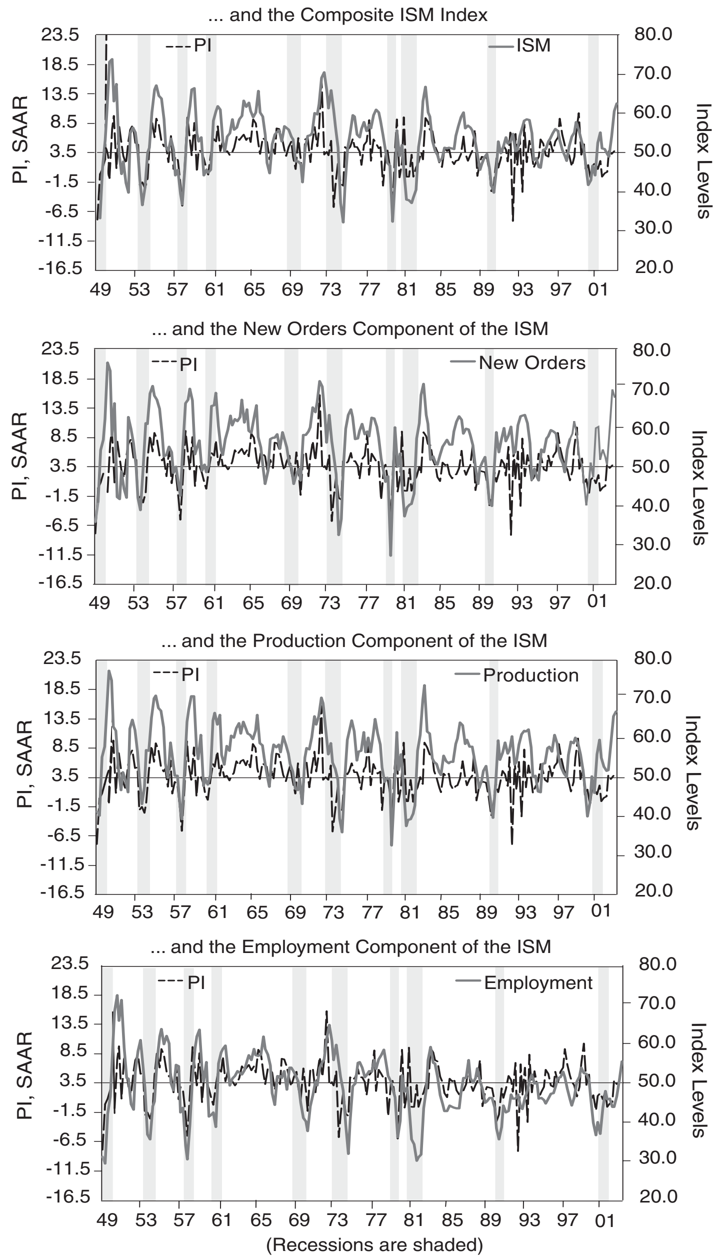


Figure 2 U.S. Personal Income Growth



to represent proportionally the 20 two-digit SIC manufacturing categories by their relative contribution to GDP. In addition, the Survey is structured to include a broad geographic distribution of companies (Kauffman 1999).

The ISM survey questions are not released by the organization, so we do not know precisely what questions respondents answer or whether the questions changed over time. In addition, the number of respondents is not revealed by the organization, making variations in response rates impossible to determine.

Despite a lack of detailed information on the survey instrument and response size, the purchasing manager's report has an enviable track record as an indicator of both national manufacturing and general economic conditions. A review of the ISM as an indicator of broader economic conditions follows.

4. THE ISM AND THE BUSINESS CYCLE

Figures 1 and 2 illustrate how various components of the ISM have moved with GDP and personal income, respectively, over the post-war period. The ISM appears to track movements in GDP closely. Note also that both the volatility of GDP growth and that of the ISM seem to have fallen together beginning in the early 1980s. Over the period from 1949 to 1984, the standard deviation of GDP growth was 5.0 percent, as compared to just 2.2 percent from 1984 to the present. This represents a decline of more than 50 percent between the two sample periods. Similarly, the standard deviation of the ISM fell from 8.8 percent over the 1949–1984 period to 4.6 percent since 1984. McConnell and Quiros (2000) argue that much of the reduction in output fluctuations over the last two decades can be attributed to a discrete fall in the volatility of durables output around 1984. Khan et al. (1999) then make the case that the fall in durables volatility itself reflects technological innovations in inventory management. To the degree that this explanation is an important factor driving the fall in output volatility starting in the early 1980s, one would expect the ISM to show precisely the kind of corresponding decrease in standard deviation it has experienced over the same period. In fact, all components of the ISM display a significant decrease in volatility after 1984.

Figures 3 and 4 show the cross-correlations between primary components of the ISM and GDP as well as personal income. Leads and lags in Figures 3 and 4 are measured in quarters. In both cases, the ISM correlates quite well with those measures, although the cross-correlations with personal income are generally smaller. Observe also that the cross-correlations are highest contemporaneously (i.e., $k = 0$) across components of the ISM, seemingly suggesting that the ISM offers no advance information on the state of the business cycle. However, the cross-correlations depicted in Figures 3 and 4 relate to revised GDP releases. Since GDP numbers for a given quarter are released in preliminary form with a one-month lag, and in revised form with up to a four-month lag, the ISM appears to provide surprisingly accurate real-time information on the business cycle, essentially one quarter or more ahead of the release of the final GDP report.

Interestingly, the cross-correlations with both GDP and personal income are highest not for the overall ISM but for its production component (as much as 70 percent contemporaneously in the case of GDP), which is not surprising. The production component of manufacturing most directly represents the sector's contribution to the value of real GDP in a contemporaneous setting. In contrast, new orders represent demand for some future period, and though they can offer insight about future production, they can also be canceled or altered.

The notion that the individual components of the ISM are not equally useful in terms of assessing current economic conditions is best reflected in its employment component. In the case of personal income, for instance, Figure 4 shows that the correlogram peaks at $k = 1$, indicating a one-quarter lag with respect to the business cycle. This lag is consistent with the idea that, once layoffs have taken place in a downturn and the economy subsequently begins to pick up, manufacturing firms at first are reluctant to hire new workers and would rather induce their current labor force to work longer hours. In other words, firms may adjust first along the intensive, rather than the extensive, margin.

While Figures 3 and 4 show that the ISM is highly correlated with GDP, the following rolling regressions show that it also generally improves the forecast performance of both GDP and personal income, as measured by the mean-squared forecast error. The regressions are run against two lags of the dependent variable and each of the ISM components, in turn, over the period 1949:Q1 to 1994:Q1, using a ten-year rolling window.

In Table 1, $\text{MSE}^{y,x}$ and MSE^{y} denote the mean-squared error of the y forecast with and without the ISM, or one of its components, respectively. Here, y refers to the cyclical component of GDP obtained from a Hodrick-Prescott (HP) filter decomposition.[4] Observe that the ratio of the MSEs is significantly less than one. This value demonstrates that including lags, either of the ISM or one of its components, always improves upon the current-quarter forecast of either GDP or personal income, relative to using their own lags alone.[5] Moreover, the ISM series performs better a quarter ahead for both GDP and personal income. The production series most improves the forecastibility

---

[4] GDP growth can be used in place of cyclical movements without substantial changes in the findings.

[5] Forecasting current-quarter GDP is a useful exercise because advance, preliminary, and final GDP data are released approximately one, two, and three months, respectively, after the quarter ends. In contrast, the ISM data are available one business day after the quarter ends.

Figure 3

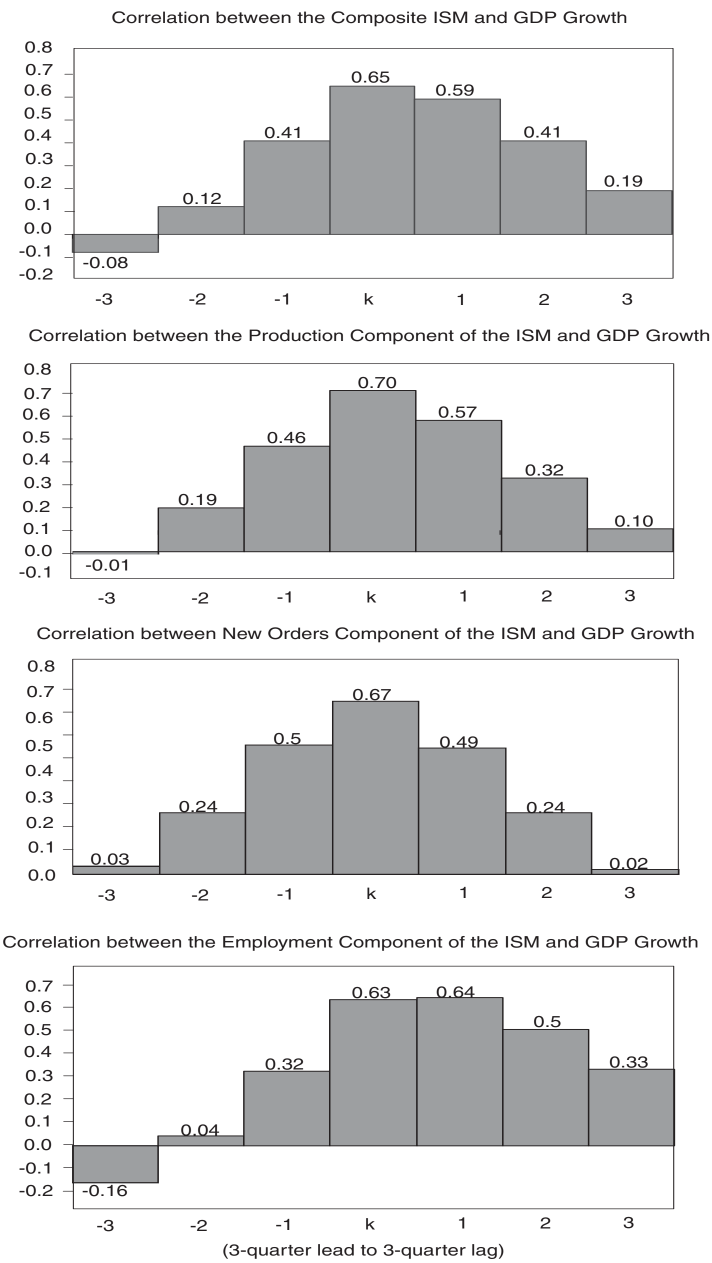

Correlation between the Composite ISM and GDP Growth
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.08
0.12
0.41
0.65
0.59
0.41
0.19
-3
-2
-1
k
1
2
3
Correlation between the Production Component of the ISM and GDP Growth
-0.01
0.19
0.46
0.70
0.57
0.32
0.10
Correlation between New Orders Component of the ISM and GDP Growth
0.03
0.24
0.5
0.67
0.49
0.24
0.02
Correlation between the Employment Component of the ISM and GDP Growth
-0.16
0.04
0.32
0.63
0.64
0.5
0.33
(3-quarter lead to 3-quarter lag)

Figure 4

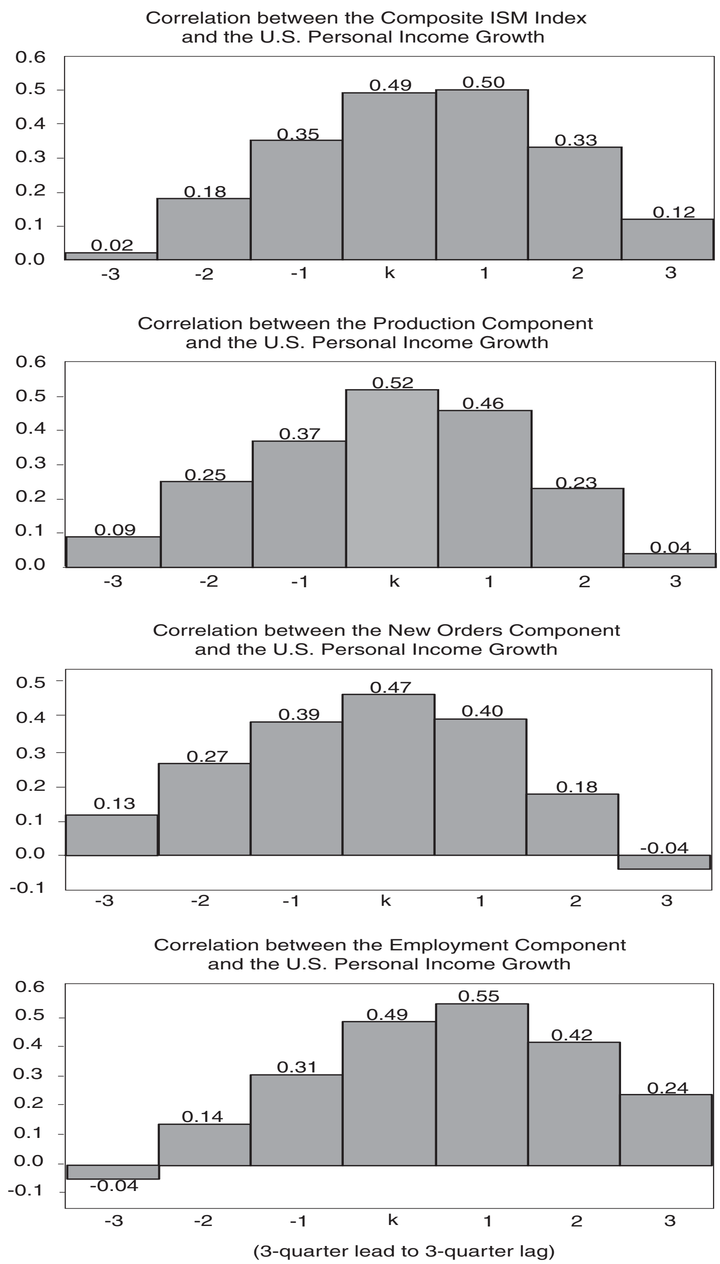

Correlation between the Composite ISM Index
and the U.S. Personal Income Growth
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.02
0.18
0.35
0.49
0.50
0.33
0.12
-3
-2
-1
k
1
2
3
Correlation between the Production Component
and the U.S. Personal Income Growth
0.09
0.25
0.37
0.52
0.46
0.23
0.04
Correlation between the New Orders Component
and the U.S. Personal Income Growth
-0.1
0.13
0.27
0.39
0.47
0.40
0.18
-0.04
Correlation between the Employment Component
and the U.S. Personal Income Growth
-0.04
0.14
0.31
0.49
0.55
0.42
0.24
(3-quarter lead to 3-quarter lag)

Table 1 Results from Rolling Regressions

$$y_t = \sum_{j=1}^{2} \alpha_j y_{t-j} + \sum_{j=0}^{1} \beta_j x_{t-j}$$

A: *y denotes detrended GDP*

Predictor, x:	Current Quarter $MSE^{y,x}/MSE^{y}$	1 Quarter Ahead $MSE^{y,x}/MSE^{y}$
ISM	0.78	0.69
ISM - Production	0.74	0.64
ISM - New Orders	0.78	0.71
ISM - Employment	0.75	0.64

B: *y denotes Personal Income*

Predictor, x:	Current Quarter $MSE^{y,x}/MSE^{y}$	1 Quarter Ahead $MSE^{y,x}/MSE^{y}$
ISM	0.86	0.84
ISM - Production	0.86	0.83
ISM - New Orders	0.87	0.85
ISM - Employment	0.87	0.85

of both GDP and personal income in the current quarter and one quarter ahead. This result is not surprising, as production most closely matches GDP conceptually and would be expected to perform well compared to personal income. In addition, the new orders component of the ISM generally improves both the current and one-quarter-ahead forecasts of GDP, although to a slightly lesser degree than the other components of the ISM one quarter ahead. This underscores the notion that new orders may not translate into shipments at a later date.

Although the ISM and its components improve the ability to forecast personal income in both the current quarter and one quarter ahead, Table 2 indicates that this improvement is somewhat reduced relative to GDP in Table 1. While personal income tends to track GDP over the long run, there are often substantial deviations between the two in the short run because of measurement error in personal income as well as differences in its definition. For example, personal income includes income from interest and rental sources which do not closely track movements in GDP.

While we have shown that the survey of purchasing managers is effective in tracking movements in GDP in real time (i.e., considerably ahead of the GDP release for the corresponding time period) and forecasting real growth, a more central question concerns its ability to alert us of impending recessions. Figure 1 shows that the ISM and its individual series tend to fall prior to recessions. As in Dotsey (1998), we can establish whether this behavior contains any predictive power most simply by assessing the signal value of

Table 2 Signal Value of the ISM and its Components One Quarter Ahead

Predictor, x:	$x < \mu - \frac{\sigma}{2}$	$x < \mu - \sigma$	$x < \mu - \frac{3\sigma}{2}$
ISM			
Total Signals	42	21	11
Frequency of True Signals (%)	61.90	71.43	90.91
Production			
Total Signals	29	14	5
Frequency of True Signals (%)	68.97	78.57	80.00
New Orders			
Total Signals	21	12	3
Frequency of True Signals (%)	66.67	83.33	66.67
Employment			
Total Signals	78	31	18
Frequency of True Signals (%)	38.46	67.74	72.22

the ISM series at different thresholds. Accordingly, let us define a signal as true if the ISM or one of its components falls below its mean (μ) by at least ϕ standard deviations (σ), where ϕ is alternatively $1/2$, 1, and $3/2$, and a recession occurs in the following quarter. We define a signal as false if no recession takes place in the quarter following one of the above signals. We can also carry out this exercise with respect to two-quarter-ahead predictions. In general, examining the relative frequency of true signals gives us a sense of how reliably the purchasing managers' survey anticipates recessions. Note, however, that this procedure says nothing about potential Type 2 errors—that is, situations in which a recession takes place without a signal occurring. As in Dotsey (1998), "this exercise lets us determine if" the ISM series "are like the boy who cried wolf or, in other words, if they correctly predict a weakening economy." The results from this non-parametric exercise are shown in Tables 2 and 3.

The results from Table 2 confirm the graphical intuition obtained from Figure 1 in that the ISM and its individual components generally represent a reliable, albeit imperfect, signal of future recessions. These results explain why both market participants and policymakers place so much emphasis on the monthly ISM release. For comparison, the unconditional likelihood of a recession over the period 1948:Q1 to 2004:Q1, as defined by the relative frequency of recession quarters, is just 20 percent. In contrast, conditioning on the ISM being one standard deviation below its mean, Table 1 indicates that the likelihood of being in a recession next quarter jumps to 71 percent. As expected, the weakest signal of an impending recession associated with the survey of purchasing managers stems from the employment series. For that series, the majority of false signals distinctly occurs towards the end of

Table 3 Signal Value of the ISM and its Components Two Quarters Ahead

Predictor, x:	$x < \mu - \frac{\sigma}{2}$	$x < \mu - \sigma$	$x < \mu - \frac{3\sigma}{2}$
ISM			
Total Signals	42	21	11
Frequency of True Signals (%)	42.86	42.86	54.55
Production			
Total Signals	29	14	5
Frequency of True Signals (%)	51.72	50.00	40.00
New Orders			
Total Signals	21	12	3
Frequency of True Signals (%)	42.87	50.00	33.33
Employment			
Total Signals	78	31	18
Frequency of True Signals (%)	26.94	38.72	44.44

recessions where the employment index remains low despite the end of the recession. As discussed earlier, this feature reflects firms' reluctance to hire new workers until they are convinced that the recession has come to an end. Table 3 indicates that the signal value of the ISM and its components in terms of foretelling recessions falls significantly two quarters ahead, although the frequency of true signals still hovers around 40 to 50 percent for most series. Again, the one exception is the employment series of which the signal value becomes barely more than the unconditional likelihood of a recession.

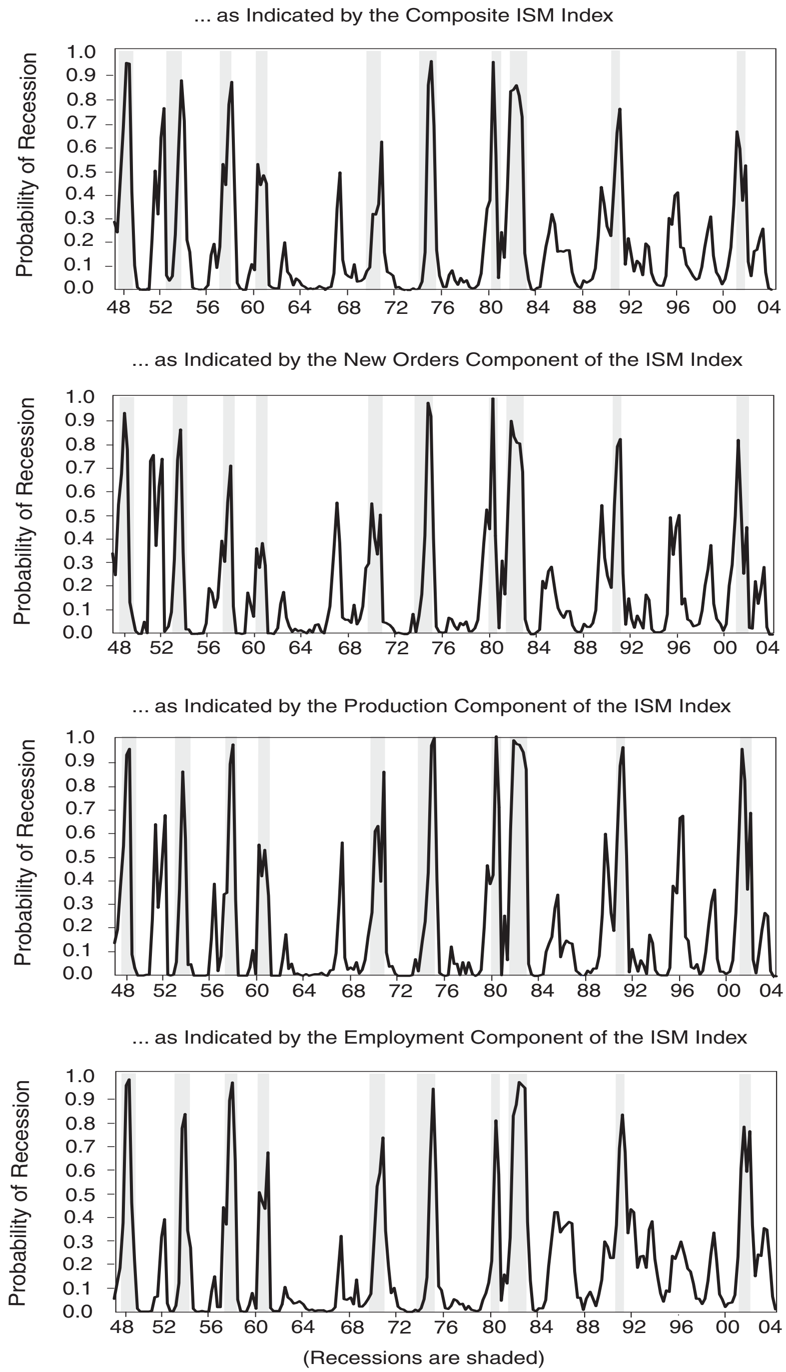
The above analysis can be refined by adding more structure to the way the likelihood of a recession is modeled conditional on observing the ISM or one of its components. In particular, one approach would be to model the probability of a recession as depending continuously on both the observed predictor, x (i.e., the ISM or one of its series), and some parameter, β, that translates the effect of the predictor on the likelihood of a recession. The probit model, for instance, expresses the likelihood of a recession as

$$\begin{aligned} \Pr(\text{recession}) &= \int_{-\infty}^{\beta x} \phi(\omega) d\omega \qquad (2) \\ &= \Phi(\beta x), \end{aligned}$$

where $\phi(\omega)$ is the normal density function that corresponds to the cumulative distribution, $0 \leq \Phi(\omega) \leq 1$. It follows that the likelihood of not being in a recession at a given date is simply $1 - \Phi(\beta x)$. Moreover, from (2), we can immediately see that the probability of a recession now increases continuously with the predictor, x.

Figure 5 shows the results from having estimated equation (2) using the ISM or one of its individual series as the conditioning variable. Observe that

Figure 5 Probability of Recession



actual recessions, shaded in gray, are generally associated with spikes in the estimated probability of a recession at those dates. This is especially true for the production series where many of the spikes are very near 1. Furthermore, consistent with the signal value analysis exercise carried out above, Figure 5 generally shows few cases of spikes taking place without a recession. In that sense, the ISM is typically not "a boy who cries wolf." Recall that our signal analysis had nothing to say about potential Type 2 errors—that is, situations where a recession took place without a signal from the survey of purchasing managers. In fact, Figure 5 suggests that these situations are seldom the case. One obvious exception concerns the 1960–1961 period where, despite a recession having taken place, the ISM, as well as all of its components, nevertheless implied a relatively low recession probability. This implication suggests that factors outside of manufacturing may have played an unusually large role in generating that specific downturn.

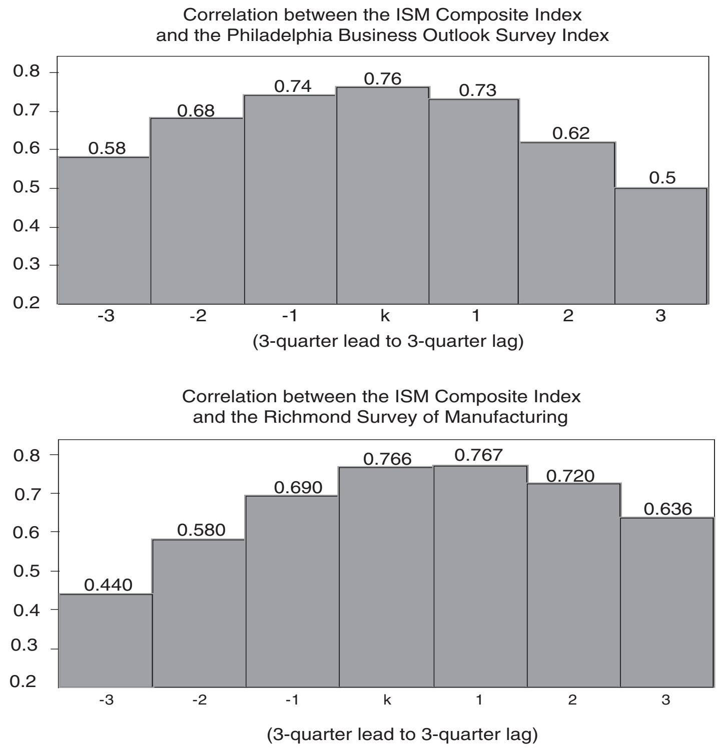
5. DO THE PHILADELPHIA AND RICHMOND SURVEYS HELP FORECAST THE ISM?

Among the regional Fed Surveys, Philadelphia has the longest running effort—stretching back to May 1968—and Richmond has the second oldest with monthly data beginning in November 1993. More recent surveys are those by Kansas City (quarterly, dating to late 1994) and New York (monthly, first released in 2002). In addition, Dallas is currently developing a manufacturing survey.

While the Philadelphia and Richmond Surveys are designed to gauge manufacturing conditions in their Districts, their results—seasonally adjusted and released monthly—also generally track the national ISM. It is noteworthy, however, that the regional Fed Banks collect and analyze their survey results prior to the release of the ISM data. The Philadelphia Survey, for example, is released on the third Thursday of the survey month compared to the first business day of the following month for the national ISM release. Similarly, while Richmond currently releases its index results to the public after the purchasing managers' index is made public, the Bank has preliminary results available internally well before the public release date. In any case, in the remainder of this analysis, the Richmond Survey information will be treated as if it is available to the public prior to the release of the ISM results.

A second issue related to the gathering of regional information has to do with the limits of the ISM. Ultimately, as with the Beige Book, dispersion matters. Although the current state of manufacturing nationally can be assessed with the ISM, information may also be gained by gauging manufacturing activity in regions. To see why, imagine a manufacturing sector composed of two industries, one stable and one volatile. If overall activity declines, but the source cannot be identified, the question of whether or not the decline is

Figure 6



a likely trend decline (if the stable industry declines) or a more temporary change (if the volatile sector declines) remains unanswered. But if the source of the decline can be identified, the question may be partially addressed. To the extent that more detailed information can be gathered by surveying regions with different manufacturing structures, insights may be gained by comparing their relative performances.

Figure 6 shows the cross-correlations of the ISM with the regional indexes constructed by the Federal Reserve Banks of Philadelphia and Richmond. Because these two Banks' Surveys are monthly and have long histories—like the ISM—they can be easily compared. From the figure, it is apparent that both regional indexes correlate very well with the ISM, over the period analyzed, although the Richmond index seems to lag the ISM slightly, relative to the Philadelphia regional index. Both Surveys display virtually identical contemporaneous correlations at 0.76. However, while these contemporaneous correlations with the ISM are very similar, they nevertheless stem from different regional information sets. Put another way, while the Philadelphia and

Richmond indexes correlate with the national survey to the same degree, we now show that they capture slightly different aspects of the ISM behavior.

In the following discussion, let P, R, and N denote, respectively, the survey indexes computed by Philadelphia, Richmond, and the national survey of purchasing managers. We assume that P, R, and N are random variables such that

$$E(N|P) = \alpha + \beta p \tag{3}$$

for all values p taken on by P. In other words, the expectation of the ISM number conditional on having observed the Philadelphia Survey index number is simply a linear function of that regional number. Under this assumption, one can show that

$$\alpha = \mu_N - \mu_P \frac{\varrho(N, P)\sigma_N}{\sigma_P}, \text{ and } \beta = \frac{\varrho(N, P)\sigma_N}{\sigma_p}, \tag{4}$$

where μ and σ denote means and standard deviations, respectively, while $\varrho(.)$ represents the correlation between two variables. In addition, we can interpret assumption (3) as deriving from the following equation,

$$N = \alpha + \beta P + \varepsilon, \; E(\varepsilon|P) = 0, \tag{5}$$

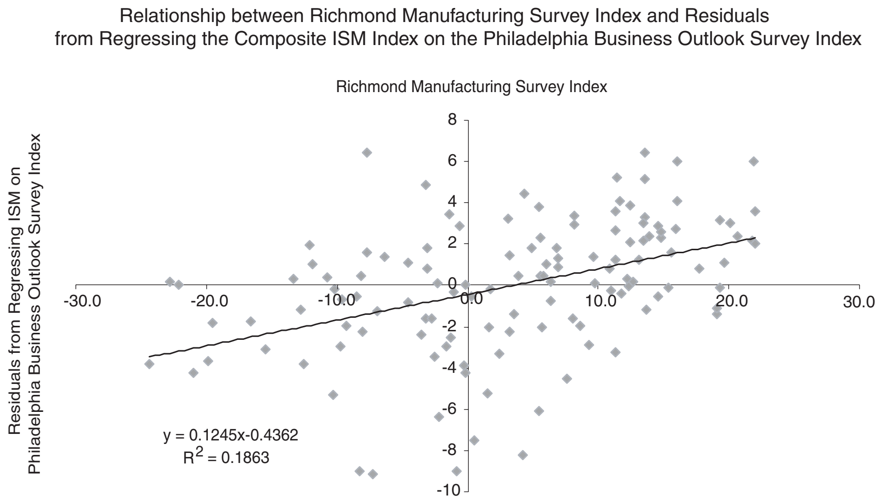
where ε denotes movements in the ISM that are not related to regional information captured by the Philadelphia Survey. Using equations (4) and (5), it is straightforward to show that

$$\varrho(N, R) = \varrho(N, P)\varrho(P, R) + \frac{\varrho(\varepsilon, R)\sigma_\varepsilon}{\sigma_N}. \tag{6}$$

Put simply, the degree to which regional information gathered in the Richmond Survey correlates with the ISM, $\varrho(N, R)$, can be split into two parts. The first term on the right-hand side of equation (6) tells us that the degree to which the Richmond Survey co-moves with the ISM is driven in part by the Richmond and Philadelphia Surveys sharing a common component, $\varrho(P, R)$, and the fact that the Philadelphia Survey itself moves with the ISM, $\varrho(N, P)$. Put another way, the correlation between the Richmond Survey and the ISM is explained by regional information common to both Philadelphia and Richmond. In contrast, the second term on the right-hand side of (6) depicts the co-movement between the Richmond Survey Index and variations in the ISM that are not captured by the Philadelphia Survey.

We know from Figure 6 that both $\varrho(N, R)$ and $\varrho(N, P)$ are around 0.77. Additional calculations yield that $\varrho(P, R) = 0.64$, so that approximately 64 percent of the correlation between the Richmond regional index and the ISM is accounted for by regional information common to Richmond and Philadelphia. This means that roughly 36 percent of the co-movement between the Richmond and purchasing managers indexes derives from the component of ISM movements, ε, orthogonal to the Philadelphia Survey index. The fact that the Richmond index is correlated with ε appears clearly in Figure 7.

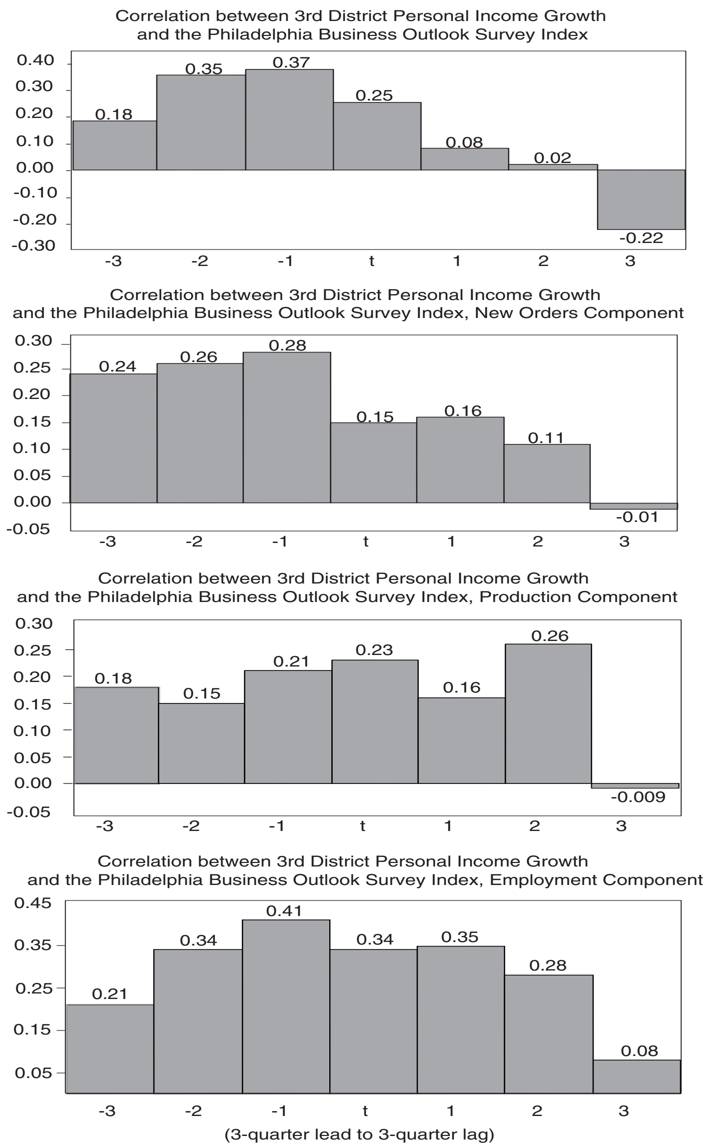
Figure 7



As mentioned earlier, the Philadelphia business outlook survey is typically released approximately ten or more days prior to the ISM. Therefore, given the ISM's ability to convey changes in business conditions outlined in the previous section, the exercise we have just carried out suggests that Philadelphia's regional index constitutes one of the earliest available gauges of the business cycle. Moreover, because the Richmond Manufacturing Survey captures variations in the ISM that are unexplained by Philadelphia's business outlook, we expect a simultaneous release of the two Surveys to convey most of the ISM's information in real time. Put another way, once regional information is gathered across the Third and Fifth Federal Reserve Districts, we already have a relatively accurate reading of what the national survey might indicate. But this reading cannot be fully exploited at present because the Richmond Survey results are released after the ISM results. As was mentioned earlier, though, we treat the Richmond results as if they were available in advance of the ISM. Tables 4 and 5 illustrate this point.

Analogous to the previous section, the first column of Table 4 tells us that when the Philadelphia business outlook index falls more than 0.5 standard deviations below its mean, the ISM behaves likewise almost 81 percent of the time within the same month. This number increases to 84 percent in the second column when both the Philadelphia and Richmond indexes fall below their respective means by at least 0.5 standard deviations. On the up side, the last column of Table 4 indicates that the ISM is above its mean by more than 0.5 standard deviations 88 percent of the time when both the Philadelphia and

Figure 8



Richmond Surveys behave likewise within the same month. Note that this finding represents an increase from 68 percent in the third column when the Philadelphia Regional Survey alone is considered.

Having established that the Richmond Survey—along with the Philadelphia Survey—is a good indicator of the ISM, the question of whether it also is

Table 4 Signal Value of the Philadelphia and Richmond Regional Surveys

ISM, z:	$z < \mu_z - \frac{\sigma_x}{2}$		$z > \mu_x + \frac{\sigma_x}{2}$	
Philadelphia, x:	$x < \mu_x - \frac{\sigma_x}{2}$	$x < \mu_x - \frac{\sigma_x}{2}$ and	$x > \mu_x + \frac{\sigma_x}{2}$	$x > \mu_x + \frac{\sigma_x}{2}$ and
Richmond, y:		$y < \mu_y - \frac{\sigma y}{2}$		$y > \mu_y + \frac{\sigma_y}{2}$
Total Signals	31	25	44	25
Freq. of True Signals	80.65%	84.00%	68.18%	88.00%

a good indicator of Fifth District economic conditions remains. We now turn our attention to that question.

6. THE RICHMOND SURVEY AND FIFTH DISTRICT ECONOMIC ACTIVITY

The Richmond Survey is useful in assessing some—though not all—aspects of regional economic activity. It is not, for example, a good gauge of gross state product (GSP) data. GSP data are only released on an annual basis, which, in terms of the Richmond Manufacturing Survey and the Fifth Federal Reserve District, represent only 13 data points. In contrast, personal income at the state level is available quarterly, and Figure 9 depicts the cross-correlations of the Richmond business surveys with Fifth District personal income. These cross-correlations are computed over the sample period for which the Richmond Manufacturing Survey numbers are available, 1994–2004.

Although the Richmond manufacturing index shows a generally high correlation with Fifth District personal income, it lags personal income by approximately one quarter. However, because state-level personal income data are released with a one-quarter lag, the Richmond results provide a more timely gauge of movements in Fifth District personal income.

More encouraging, as shown in Figure 11, the Richmond employment index distinctly leads changes in manufacturing employment by one quarter. This is noteworthy because changes in manufacturing employment are among the most timely and closely watched regional economic data.

7. CONCLUDING REMARKS

Given the strong interest in timely information on both national and regional economic conditions, the Richmond Survey of Manufacturing performs a useful role. In a national economic setting, the Survey appears capable of adding

Figure 9

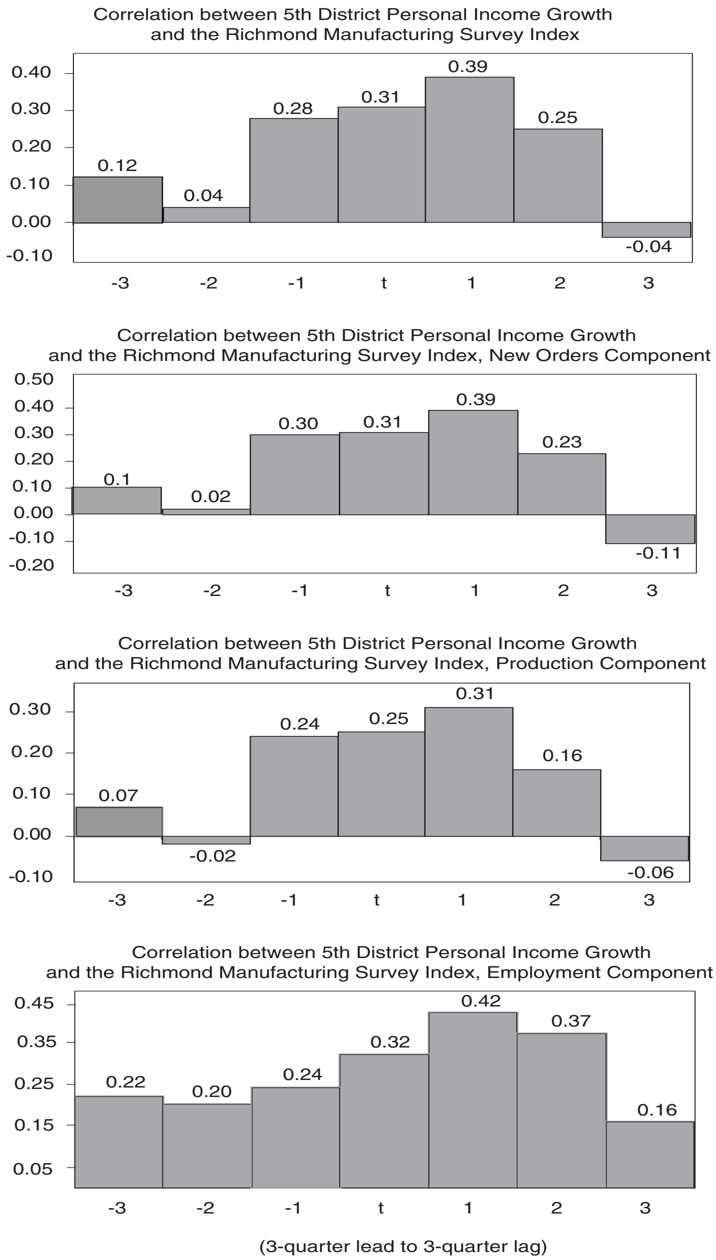

Correlation between 5th District Personal Income Growth
and the Richmond Manufacturing Survey Index
0.40
0.30
0.20
0.10
0.00
-0.10
0.12
0.04
0.28
0.31
0.39
0.25
-0.04
-3
-2
-1
t
1
2
3
Correlation between 5th District Personal Income Growth
and the Richmond Manufacturing Survey Index, New Orders Component
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
0.1
0.02
0.30
0.31
0.39
0.23
-0.11
-3
-2
-1
t
1
2
3
Correlation between 5th District Personal Income Growth
and the Richmond Manufacturing Survey Index, Production Component
0.30
0.20
0.10
0.00
-0.10
0.07
-0.02
0.24
0.25
0.31
0.16
-0.06
-3
-2
-1
t
1
2
3
Correlation between 5th District Personal Income Growth
and the Richmond Manufacturing Survey Index, Employment Component
0.45
0.35
0.25
0.15
0.05
0.22
0.20
0.24
0.32
0.42
0.37
0.16
-3
-2
-1
t
1
2
3
(3-quarter lead to 3-quarter lag)

Figure 10

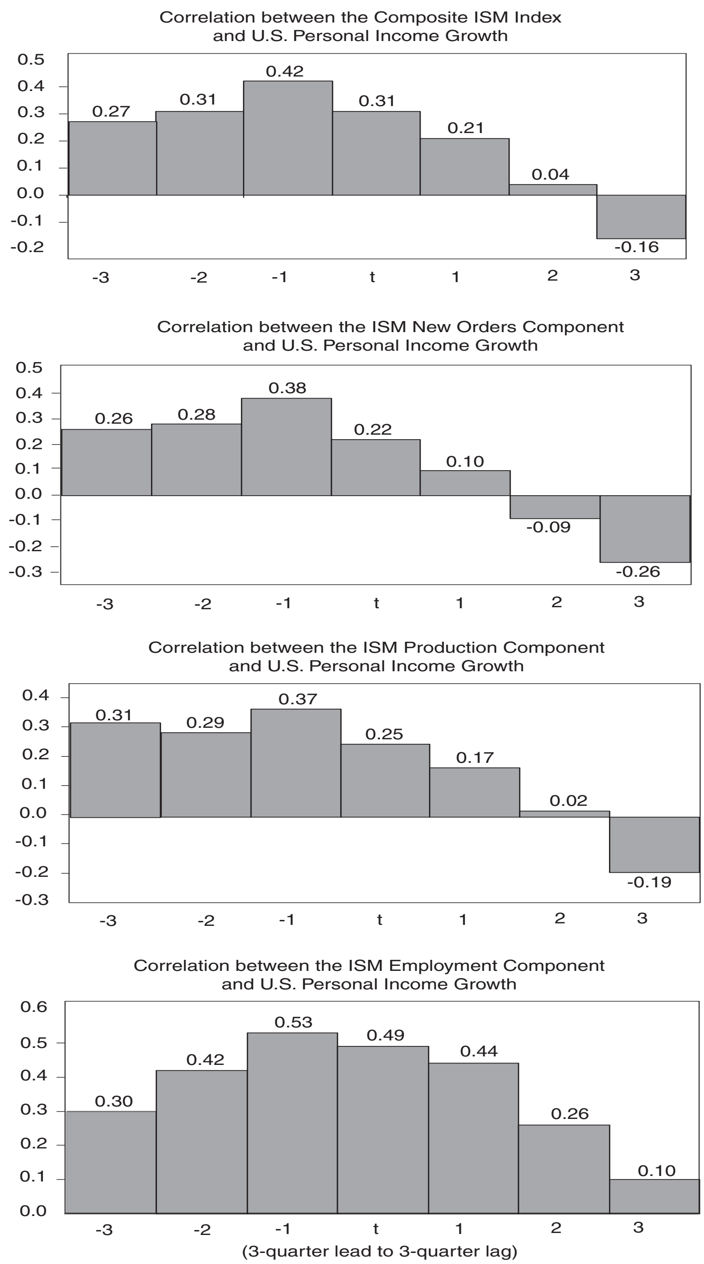

Correlation between the Composite ISM Index
and U.S. Personal Income Growth
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
0.27
0.31
0.42
0.31
0.21
0.04
-0.16
-3
-2
-1
t
1
2
3
Correlation between the ISM New Orders Component
and U.S. Personal Income Growth
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0.26
0.28
0.38
0.22
0.10
-0.09
-0.26
-3
-2
-1
t
1
2
3
Correlation between the ISM Production Component
and U.S. Personal Income Growth
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0.31
0.29
0.37
0.25
0.17
0.02
-0.19
-3
-2
-1
t
1
2
3
Correlation between the ISM Employment Component
and U.S. Personal Income Growth
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.30
0.42
0.53
0.49
0.44
0.26
0.10
-3
-2
-1
t
1
2
3
(3-quarter lead to 3-quarter lag)

Table 5 Signal Value of the Philadelphia and Richmond Regional Surveys

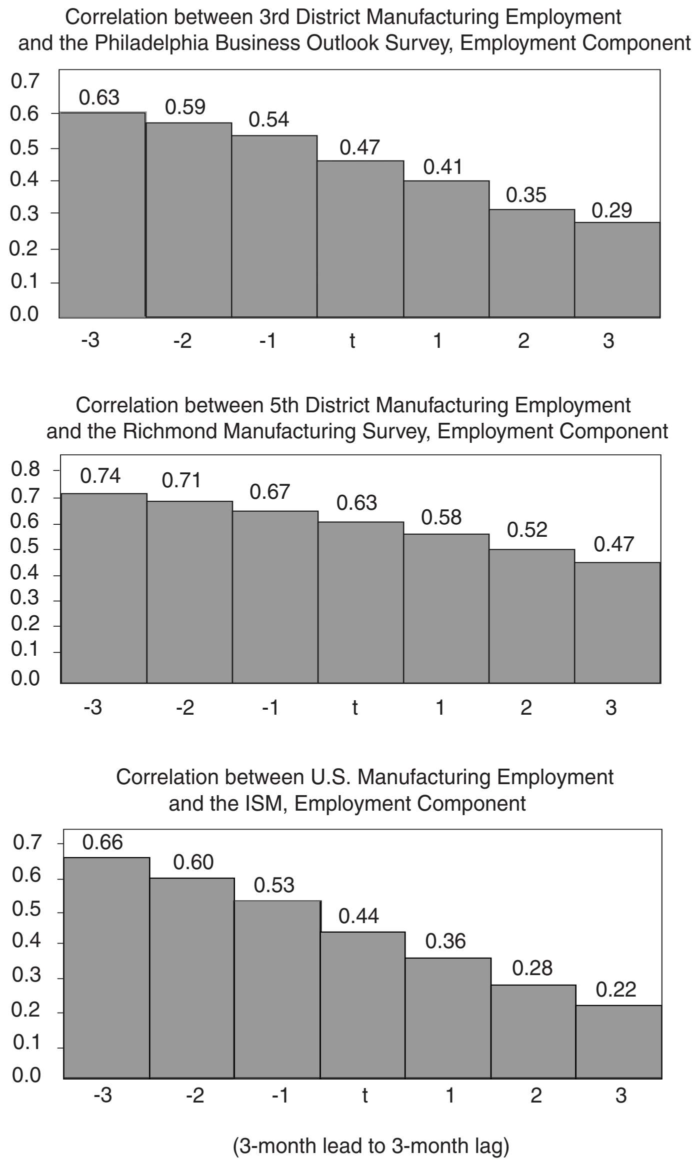
ISM, z :	$z < \mu_z - \sigma_z$		$z > \mu_z + \sigma_z$	
Philadelphia, x:	$x < \mu_x - \sigma_x$	$x < \mu_x - \sigma_x$ and	$x > \mu_x + \sigma_x$	$x > \mu_x + \sigma_x$ and
Richmond, y:		$y < \mu_y - \sigma_y$		$y > \mu_y + \sigma_y$
Total Signals	22	15	16	6
Freq. of True Signals	68.18%	86.67%	62.50%	66.67%

to the ability to forecast the PMI component of the ISM index, especially when combined with the results of the Philadelphia Fed's Survey results. This is important because the ISM has been a very good gauge historically. The ISM is released well ahead of GDP data, and it provides relatively accurate signals of both substantial changes in the growth rate of GDP and turning points in the economy.

Both the Philadelphia and Richmond Federal Reserve Banks produce monthly indexes that are highly correlated with the ISM. The Philadelphia Index is currently released well in advance of the ISM and serves as a valuable predictor of the ISM.

The Richmond Survey results are less useful at present. The results are reported as components only rather than in the format of an ISM-style weighted index. Moreover, the Richmond results are released after the ISM. But this memo suggests that some modification of the Richmond Manufacturing Survey could add substantial value to forecasters. First, as was done in this analysis, existing questions in the Survey could be combined and weighted in a manner similar to the construction of the ISM. One such construction, considered in the memo, is shown to correlate very well with the ISM. A second change would be to advance the release date of the Richmond Survey results. Because the information is currently available internally to the Richmond Fed well before it is released to the public, moving up the release date would provide the same advantage to the public. A second important finding is that the Richmond Survey is a good indicator of economic activity in the Fifth District. It provides a timely view of economic activity in the Fifth Federal Reserve District. While the Richmond Survey tends to lag its Federal Reserve District's personal income measure by around a quarter, the Survey's information is made available well in advance of the District personal income data and so effectively provides an advance look at Fifth District personal income. In addition, the Richmond Survey Index distinctly leads changes in Fifth District

Figure 11



employment, giving an advance indication of changes in the region's labor market.

APPENDIX A: SURVEY OF FIFTH DISTRICT MANUFACTURING ACTIVITY

Business Activity Indexes				
Compared to the previous month	September	August	July	3-month avg.
Shipments	22	18	6	5
Volume of new orders	8	13	13	11
Backlog of orders	-6	1	-3	-3
Capacity utilization	13	9	5	9
Vendor lead time	14	21	15	16
Number of employees	5	-2	6	3
Average workweek	1	-1	6	2
Wages	10	10	12	11
Six months from now				
Shipments	23	28	33	28
Volume of new orders	24	24	31	26
Backlog of orders	5	11	14	10
Capacity utilization	10	16	17	15
Vendor lead time	11	6	4	7
Number of employees	3	7	9	6
Average workweek	7	-4	0	1
Wages	34	42	46	41
Capital expenditures	9	19	17	15
Inventory levels				
Finished good inventories	16	16	19	17
Raw materials inventories	11	7	7	8

Price trends			
(percent change, annualized)	September	August	July
Current trends			
Prices paid	1.71	2.28	2.33
Prices received	1.25	2.17	3.20
Expected trends during next 6 months			
Prices paid	1.25	2.17	3.20
Prices received	0.08	1.37	2.59

Notes: Each index equals the percentage of responding firms reporting increase minus the percentage reporting decrease. Data are seasonally adjusted. Results are based on responses from 94 of 201 firms surveyed

All firms surveyed are located in the Fifth Federal Reserve District, which includes the District of Columbia, Maryland, North Carolina, South Carolina, Virginia, and most of West Virginia.

APPENDIX B: FIFTH DISTRICT MANUFACTURING ACTIVITY PRESS RELEASE

Manufacturing Output Strengthens in September; Employment Improves; Average Workweek Flat

On balance, manufacturing activity continued to generally strengthen in September, according to the latest survey by the Richmond Fed.[6] Factory shipments advanced at a quicker pace although the growth of new orders edged lower. Backlogs retreated into negative territory while capacity utilization inched slightly higher. Vendor lead-time grew more slowly than last month while raw materials inventories grew at a slightly faster rate. On the job front, manufacturers reported that worker numbers were higher at District plants; the average workweek was flat and wage growth stayed on pace of recent months.

Looking ahead, respondents' expectations were generally less optimistic than those of a month ago—producers looked for shipments and capital expenditures to grow at a somewhat slower pace during the next six months.

Price increases at District manufacturing firms continued to increase at a modest pace in September. Raw materials prices grew at a marginally slower rate, while finished goods prices grew at a slightly quicker rate. For the coming six months, respondents expected raw materials goods prices to increase only modestly and finished goods prices to be nearly flat.

Current Activity

In September, the seasonally adjusted shipments index inched up four points to 22, and the new orders index inched down five points to 8. In addition, the order backlogs index moved into negative territory, losing seven points to end at -6. The capacity utilization index advanced four points to 13 while the vendor lead-time index shed seven points to 14. The level of finished goods inventories was unchanged in September when compared to August, while the level of raw materials inventories increased. The finished goods inventory index held steady at 16, while the raw materials inventory index added four points to finish at 11.

Employment

Employment at District plants showed signs of improvement in September. The employment index posted a seven-point gain to 5 from -2; the average

[6] Released 12 October 2004.

workweek index picked up two points to 1 from -1. Wage growth remained modest, matching August's reading of 10.

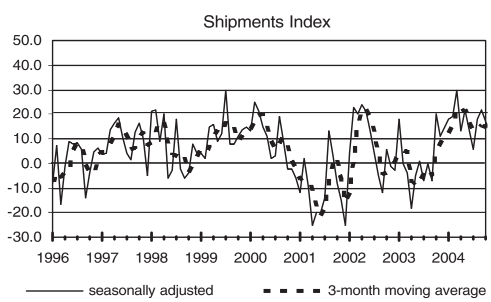
Expectations

In September, contacts were slightly less optimistic about demand for their products during the next six months. The index of expected shipments moved down five points to 23, while the expected orders index stayed at 24. The expected orders backlogs index dropped six points to end at 5 and the expected capacity utilization index shed six points to 10. The index for future vendor lead-time inched up five points to 11. In contrast, planned capital expenditures registered a ten-point loss to 9.

Manufacturers' plans to add labor in coming months were mixed. The index for expected manufacturing employment inched down four points to 3, while the expected average workweek index advanced eleven points to 7. The expected wage index posted a ten-point loss to 9.

Prices

Price changes remained modest in September. Manufacturers reported that the prices they paid increased at an average annual rate of 1.71 percent compared to August's reading of 2.28 percent. Finished goods prices rose at an average annual rate of 1.25 percent in September compared to 0.79 percent reported last month. Looking ahead to the next six months, respondents expected supplier prices to increase at a 1.25 percent annual rate compared to the previous month's 2.17 percent pace. In addition, they looked for finished goods prices to nearly match the pace of last month's expected 1.37 percent rate.



REFERENCES

Dotsey, Michael. 1998. "True Predictive Content of the Internet Rate Term Spread for Future Economic Growth." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (May): 31–51.

Keeton, William R., and Michael Verba. 2004. "What Can Regional Manufacturing Surveys Tell Us?—Lessons from the Tenth District." Federal Reserve Bank of Kansas City *Economic Review* 89 (Third Quarter): 39–69.

Lacy, Robert L. 1999. "Gauging Manufacturing Activity: The Federal Reserve Bank of Richmond's Survey of Manufacturers." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Winter): 79–98.

Kaufmann, Ralph G. 1999. "Indicator Qualities of the NAPM Report on Business." *Journal of Supply Chain Management* (Spring): 29–37.

Khan, James, Connell, Margaret, and Gabriel Quiros. 1999. "Inventories and the Information Revolution: Implications for Output Volatility." Mimeo, Federal Reserve Bank of New York.

McConnell, Margaret, and Gabriel Quiros. 2000. "Output Fluctuations in the United States: What Has Changed Since the Early 1980s?" *American Economic Review* 90 (December): 1464–76.