Monetary Policy and Inflation Targeting in the United States

Based on a speech given by President Anthony M. Santomero, at the NABE/GIC Conference, October 4, 2004

BY ANTHONY M. SANTOMERO

Should monetary policymakers in the U.S. adopt explicit inflation targeting? After all, the Fed has steadily reduced inflation over the past 25 years without resorting to an explicit inflation target. But having achieved price stability, we must now deal with the matter of maintaining it. In this quarter's message, President Anthony Santomero returns to the topic of inflation targeting, which he first discussed in the spring of 2003. This time, he expands that discussion by proposing a specific inflation targeting program.

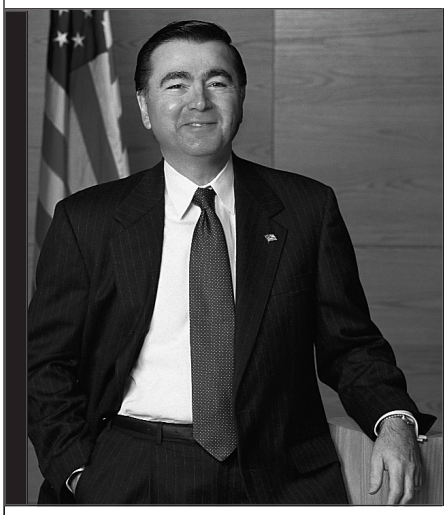
I would like to address a topic that I first discussed in the spring of 2003. Back then, I said the time had come for the Fed to adopt an explicit inflation targeting program. I noted that quite a few countries around the globe had already done so successfully. I acknowledged that for an inflation targeting program to be successful in the U.S., it would have to address our unique circumstances here, as well as resolve some practical challenges. Nonetheless, I expressed some optimism that these issues could be resolved.*

Since that time, there has been a good bit more discussion and research into the concept of inflation targeting for the U.S. I would like to extend that discussion by proposing a specific inflation targeting program for consideration. My hope is that, by laying out a specific inflation targeting proposal, I can help move the public discussion to the level of detail necessary to develop and implement inflation targeting in the U.S.

People might ask why the U.S. should move forward on inflation targeting now. For 25 years, the Fed has steadily reduced inflation and inflation expectations, without resorting to an explicit inflation target. Indeed, the case can be made that the Fed has been remarkably successful over this period, and we have now achieved essential price stability. Hence, it might seem that there is no reason to adopt an explicit inflation target at this point.

However, having achieved price stability, we must still deal with the matter of maintaining it. I believe an explicit inflation target can help us do that. By defining what it means by price stability, the Fed would not only better inform market participants of its intentions but would also strengthen their capacity to monitor the Fed's performance. This makes it more difficult for the central bank to compromise its objective some time down the line and hence heightens the commitment. Thus, a well-formulated inflation targeting program would give the public good reason to be more confident in the Fed's commitment to a stable price environment.

I do not foresee undue inflationary pressures building any time soon. I believe price pressures are, and will remain, well contained. So my point is not that higher inflation is an imminent threat and we need an inflation target to avert it. My point is just the opposite. I think that setting an inflation target at a time of price stability makes eminently good sense.



Anthony M. Santomero, President, Federal Reserve Bank of Philadelphia

* See "Flexible Commitment or Inflation Targeting for the U.S.?" *Business Review*, Third Quarter 2003.

It would clarify the Fed's definition of its price stability goal – and seal its commitment to price stability – without requiring the Fed to shift its current policy strategy. Plus it would not disrupt the economy.

Suppose the Fed were to announce an explicit inflation target at this point. I daresay that the Fed's credibility right now is such that the market would expect the Fed to adhere to it. So the public's inflation expectations would match the Fed's intentions. Having achieved that match, the Fed would be loath to break it and would risk doing so only in the most extraordinary circumstances. That, in short, is why this is a good time for the Fed to adopt an inflation target.

So, what exactly would such an inflation targeting program look like, and what specifically would my desired goal be?

WHY INFLATION TARGETING?

Let me begin by briefly reviewing the potential benefits of inflation targeting. Both economic theory and experience demonstrate that securing a low and stable rate of inflation is the most important contribution a central bank can make to sustaining maximum economic growth. Price stability helps promote economic efficiency by allowing economic agents to discern relative price changes clearly and allocate current resources to their best use. It reduces uncertainty about the future and encourages both higher levels of investment and investment in the most productive projects. Equally important, it reduces the harmful distributional effects of both anticipated and unanticipated inflation.

Note that the stable price environment I have just described includes not only low inflation in the present but also market participants' expectations of low inflation in the future. Thus, it is important for the central bank not only to achieve low inflation but also to make a commitment to maintaining low inflation.

Finally, the public must find the commitment credible. This is a fundamental element of true price stability. In simple terms, a central bank's commitment can be deemed fully credible only if the public's expectations about the future course of inflation exactly match the central bank's stated intentions with respect to the future course of inflation.

So the question is: Why would the public expect inflation to behave in accordance with the central bank's stated intentions? Economists would say that the only reason to believe that individuals or institutions will act in accordance with their stated intentions is that it would be in their interest to do so when the time for action comes. So, it is reasonable to believe that the central bank will take the actions necessary to maintain low and stable inflation only insofar as the central bank will always see that the benefits of doing so outweigh the costs.

The problem is that the central bank may occasionally be tempted to pursue overly stimulative monetary policies in order to boost economic activity today at the expense of containing inflation tomorrow. This is all the more likely to happen after inflation has been under control for some time and the action thus seems relatively "harmless." Recognizing this, the public has reason to be skeptical that the central bank will carry out its stated intentions to maintain low and stable inflation over the long term without some form of commitment to the goal. Establishing an explicit inflation target overcomes this problem by increasing the benefits of sticking with the low inflation policy and increasing the costs of deviating from it.

An inflation targeting program is like a contract between the public and the Fed. It states, in explicit terms, what the Fed means by its goal of price stability. As long as the Fed abides by the contract — that is, achieves its inflation target as specified — the public, too, will abide by the contract — that is, will expect the Fed to continue achieving its inflation target as specified. Conversely, if the Fed allows inflation to stray outside the target range, essentially failing to live up to its part of the contract, the public will alter its expectations about future inflation accordingly.

Such a shift in expectations would be costly to the economy because it is only when the central bank's inflation intentions, the public's inflation expectations, and actual inflation all match up that we have a stable price environment. In such circumstances, investors are making optimal decisions, and the economy is delivering optimal outcomes — both in terms of price stability and real growth.

An explicit inflation target would give the Fed a stronger incentive to act as it says it intends to. Recognizing this, the public would consider the Fed's commitment to low and stable inflation more credible, and thus, the full benefits of a stable price environ-

> It is important for the central bank not only to achieve low inflation but also to make a commitment to maintaining low inflation.

ment are more likely to be realized. Put another way, an explicit inflation target would make the socially optimal monetary policy what economists call a time-consistent one.

THE DUAL MANDATE

Before I go any further, let me address one other point. As a member of the FOMC, I recognize that by law the Federal Reserve has a mandate that goes beyond just price stability. In advocating inflation targeting, I am sensitive to our dual mandate and the need to pursue our economic stabilization goals.

Undoubtedly, a commitment to maintain a stable price environment limits the latitude the Fed has in pursuing these stabilization goals. But this is not the result of adopting an inflation target. It is simply a reality, given the way monetary policy affects the economy. All else constant, adding monetary stimulus helps generate output and employment and also puts upward pressure on the price level.

Therefore, recognizing our dual mandate does not speak against the wisdom of inflation targeting as much as it recognizes the difficulty of conducting appropriate monetary policy where our objectives are intrinsically intertwined. In fact, a case can be made that the clarity and confidence afforded by an explicit inflation target may actually enhance the Fed's capacity to achieve the dual goals of price stability and strong economic performance. Once the public's inflation expectations are well anchored, changes in short-term nominal interest rates, and in current prices, send a clearer signal about changes in real interest rates and intertemporal shifts in relative prices. These improved signals should evoke stronger responses from market participants and hence heighten the responsiveness of the economy to monetary policy actions.

In short, I do not think of inflation targeting as restricting the Fed to the pursuit of just one goal, but rather as empowering the Fed to pursue its two goals with alacrity.

Having said that, I recognize that the inflation targeting program we establish must afford the Fed enough flexibility to respond appropriately to various economic disturbances. We should remember that the U.S. achieved its current price stability with a policy framework I have previously referred to as one of "flexible commitment." In my view, an inflation targeting framework that precluded an appropriate policy response to economic disturbances would be sub-optimal from the social point of view and would not be credible in the eyes of the public.

> In setting the [target] band, the Fed would also communicate to the public its assessment of the short-run volatility in inflation that is consistent with the Fed's maintaining price stability in the long run.

THE PROPOSAL

With this as background, let me move to my inflation targeting proposal. Here the devil is in the details. In my view, any inflation proposal must address at least three distinct but interrelated questions: Should the target be a single number or a range? Over what time interval should the target be set? And which measure of inflation should be targeted? To advance the discussion on an inflation target for the U.S., let me offer a concrete proposal: *The Fed should establish a target band of 1 to 3 percent for annual inflation, as measured by the 12-month moving average rate of change in the core PCE deflator.*

Why these specific features?

Setting a Range. Let's address the first question: Why should we set an annual target band rather than a single long-run value or central tendency, as advocated by my colleague Ben Bernanke? My answer is based on two features of the practical world in which we live.

First, FOMC members may have different opinions about the optimal long-run inflation rate. Different models of the economy and different assessments of society's preferences can create legitimate differences of opinion. So it is unlikely, and even unreasonable, to expect that the FOMC members could agree to a single number for an inflation target. Nonetheless, the FOMC members could presumably be comfortable with a target band for inflation.

Once the FOMC members established a target band, it would help them coordinate their decision-making. FOMC members would know that as long as inflation is well within the target band, they have the latitude to consider a variety of stabilization strategies. But as inflation approaches the boundaries of the target band, FOMC members would realize that they need to focus on strategies for keeping inflation within the band, properly taking into account the lags and uncertainties surrounding the impact of monetary policy.

Second, establishing a target band for inflation would also help the FOMC communicate its intentions to the public more clearly. People would know that inflation outside the target band is clearly unacceptable to the

Fed, and they could expect the Fed to take action to bring it back "in bounds."

In setting the band, the Fed would also communicate to the public its assessment of the short-run volatility in inflation that is consistent with the Fed's maintaining price stability in the long run. The reality is that monetary policy cannot deliver the same degree of price stability over intervals as short as a month or a quarter as it can over the course of a year. More precisely, it can do so only at the expense of creating an unacceptable degree of short-run instability in economic activity. So the width of the Fed's target band would indicate how much short-run inflation volatility the Fed has decided to accept in order to limit short-run output volatility, given the economy's underlying structure. This information should help prevent unnecessary inflation or deflation "scares" when the inflation rate accelerates or decelerates.

So in my view an annual band for inflation serves as a practical device for coordinating decisions among FOMC members and communicating clearly with the public.

Nonetheless, one might object that a single long-run value would provide a more precise anchor for long-run inflation expectations than would an annual band. Mathematically speaking, that may be true, of course. Yet, from a practical point of view, I think an annual target band establishes the stronger anchor. An annual target band gives the public a clear criterion by which it can monitor and assess the Fed's performance against its stated inflation objective on a continuous basis. Recognizing this, the Fed would always have the incentive to keep inflation in the target band and would weigh seriously any policy actions that risked pushing inflation outside the band. Public monitoring would not be brought to bear in the same way if the Fed had a single-value long-run inflation target.

To use a simple metaphor, let me suggest that we are building a road, or highway, to continued price stability. On my highway, I would paint white lines along the shoulders, making it clear when the car is veering too far from the center of the road. If the car drifts toward the white line, the driver will likely react. If the driver does not, the passenger probably will. With the single long-run target, the highway has no white lines to encourage the driver to stay on course. So the car is more likely to stray from the center of the highway, and the driver and passenger are less likely to react in time to keep from running off the road. For this reason, both driver and passenger are likely to feel more confident about reaching their destination if they take a well-marked highway.

Similarly, the Fed is more likely to follow policy consistent with long-run price stability, and the public is likely to be more confident that it will, under an inflation targeting program that specifies an annual inflation target band.

To recap, I think setting an annual inflation target band has significant practical benefits. It would help coordinate decision-making among FOMC members; it would improve the FOMC's communication with the public; and it would build public confidence that prices will remain stable over the long run.

Focusing on an Annual Target Range. In regard to a target band for annual inflation, I believe the 12-month moving average of an inflation measure provides a relatively clear signal to the public – and to the Fed – of the Fed's inflation performance. Taking the 12-month moving average eliminates the "noise" of highly transitory movements in prices.

At the same time, it strengthens the "signal" of a change in trend inflation that a longer term moving average might obscure. For example, suppose the Fed focused on the three-year moving average of inflation rather than the one-year moving average. Further suppose that inflation had averaged 1 percent for two-and-a-half years, then ratcheted up to 4 percent for six months. A three-year moving average of inflation would still be only 1.5 percent, not an alarming number. But I would argue that six months of 4 percent inflation should certainly trigger a change in policy.

The Fed is more likely to follow policy consistent with long-run price stability, and the public is likely to be more confident that it will, under an inflation targeting program that specifies an annual inflation target band.

In short, targeting the inflation rate over an annual interval seems to bring the right focus both to the Fed's decision-making and to the public's monitoring of the Fed's performance.

A Band of 1 Percent to 3 Percent to Start. Having explained why I think a year is the right interval for the inflation target, let me turn to the companion issue: why I think 1 percent to 3 percent is the right band for the inflation target.

Setting the right band pres-

ents an interesting problem. I think there is consensus that an annual inflation rate of more than 3 percent does not represent price stability and that annual inflation below 1 percent provides too little cushion against the risk of deflation in times of economic weakness. So a band of 1 percent to 3 percent seems to be a reasonable starting point.

But should the band be tighter? Perhaps. As I will explain in a moment, our recent history suggests that a tighter band is feasible. Ultimately, perhaps, we will want to move in that direction.

However, I think starting with a band wide enough to command broad agreement is crucial. The reason is that it must be beyond dispute that any inflation targeting program respects the Fed's dual mandate: to maintain stable prices and a stable economy. While an inflation targeting program must preclude the Fed from compromising on its delivery of low and stable inflation, it must also allow the Fed the flexibility to respond appropriately to economic disturbances. As Chairman Greenspan indicated some time ago, monetary policy is, in the end, an exercise in risk management. Any policy regime must permit an appropriate, and immediate, response by the central bank to short-term disturbances without concern about signaling a regime change.

Admittedly, a plan that gives the Fed too much flexibility will do little to increase public confidence in the Fed's commitment to price stability. But I would point out that a plan that gives the Fed too little flexibility would turn out to be equally unconvincing and potentially dangerous.

Suppose an overly restrictive inflation targeting program was in place. Now, further suppose that a significant shock hit the economy. The Fed would face a difficult choice: pursue the appropriate stabilization policy or follow its very restrictive inflation targeting program. If the Fed sticks to its inflation plan, and that program is, in fact, overly restrictive, the Fed will have needlessly compromised the economy's performance. If the Fed deviates from the inflation plan and pursues the appropriate stabilization policy, people will not know whether the action represents a lack of commitment to price stability or a temporary deviation from the inflation target path that does not compromise the Fed's commitment to price stability.

At the end of the day, a 1 percent to 3 percent inflation target is a reasonable way to implement a policy aimed at preserving price stability.

The bottom line is that an overly restrictive inflation targeting program either needlessly compromises the Fed's performance on the stabilization front or needlessly undermines public confidence in the Fed's commitment to price stability. My sense is that the wise strategy is to start with a relatively wide band for our inflation target and perhaps consider narrowing it in the future, as we gain experience.

Some might argue that the 1 percent to 3 percent band is itself too narrow and could constrain the Fed from effectively pursuing its economic stabilization objective. However, I do not think that should be a serious problem. Consider our recent economic history.

Since the mid-1990s, we have experienced several international financial crises, a stock market boom and bust, and a direct attack on the nation's capital and its financial center. We have been through an entire business cycle from strong expansion, through recession, through recovery, and into expansion again. In the course of responding to these events, the Fed has moved its target fed funds rate over a range of 500 basis points. Over that period, the 12-month average core PCE inflation rate has moved within a band of 1 percent to 2 percent. That suggests that a two-percentage-point band on this measure of inflation should provide the Fed with sufficient latitude to conduct stabilization policy.

Of course, we cannot know for sure what lies ahead. We may yet encounter some very unusual situation in which responding effectively to a disturbance would push inflation outside the inflation target band. Yet, the very rarity of the situation may allow the Fed to respond without any loss of credibility. Presumably, the unusual nature of the situation would make it easily recognizable, much as the events I just catalogued were. As long as the Fed clearly communicated how it was dealing with the situation and, once it passed, began moving inflation back within its target band, its credibility could be preserved.

At the end of the day, a 1 percent to 3 percent inflation target is a reasonable way to implement a policy aimed at preserving price stability. It would serve to keep inflation low and stable, without overly constraining the Fed from reacting to economic or financial disturbances.

Why the Core PCE Deflator? Now let me turn to the last question of the proposal: the inflation measure itself. The choice of the PCE deflator is relatively straightforward.

The Fed has been focusing on this measure in recent years, and I see no reason to change that. The PCE deflator is a broader measure than the consumer price index. Also, it is a chain-weighted index and so takes account of consumers' shifting among goods and services as relative prices change. Consequently, it reflects recent changes in the overall price level more accurately than the CPI, which is based on a fixed basket of goods and services.

However, I prefer targeting the core PCE deflator, that is, the deflator less its food and energy price components. Like using a 12-month moving average, using the core PCE helps reduce the "noise" in the inflation signal, enhancing its value as a monitoring device. In light of the recent run-up in oil prices, it is worth emphasizing that the choice of the core PCE deflator essentially insulates the Fed from having to respond to such shocks in order to achieve its inflation target. Large as it was, the recent run-up in oil prices has had relatively little impact on core PCE. Thus, the inflation targeting program will not induce the Fed to tighten aggressively when oil prices rise or ease aggressively when they fall.

In short, my proposal is not an elaborate one by any means. It does not codify any new Fed procedures. It does not specify a particular reaction function for monetary policy. It does not set a timetable for returning inflation to target when deviations occur. It simply defines what the Fed means by price stability and thereby reinforces the Fed's commitment to, and the public's confidence in, its preservation.

SUMMARY

I believe a program of explicit inflation targeting is a logical next step for the Fed to take in its commitment to preserve the stable price environment it has worked so long and so hard to achieve. A specific inflation target such as the one I propose here – 1 percent to 3 percent inflation in the 12-month moving average of the core PCE – could be that step.

Recent theoretical work on optimal monetary policy offers three lessons. First, it is optimal for the central bank to establish a low and stable rate of inflation. Second, it is optimal for the central bank to respond to disturbances that affect economic activity. Third, optimal monetary policy does the most good when people understand what that optimal policy is and expect the central bank to execute it. To put it another way, optimal monetary policy is most effective when it is both transparent and credible.

By those standards, I believe my proposed inflation targeting program would move the Fed a step closer to conducting optimal monetary policy. It would help the Fed establish a low and stable rate of inflation. It would not unduly restrict the Fed's ability to respond to economic disturbances. And by increasing the transparency and credibility of Fed policy, it would enhance the Fed's overall effectiveness.

True Confessions:
Should Banks Be Required to Disclose More?

BY MITCHELL BERLIN

Can market participants, such as bondholders and depositors, play a significant role in ensuring that banks limit their risk-taking? Although regulators find this idea increasingly attractive, to evaluate banks' risk-taking, investors need good information about a bank's activities and balance sheet. In light of this, would stiffer mandatory disclosure requirements for banks — as in the recent Basel II proposal — be a good thing? While there are no definitive answers to this question, Mitchell Berlin reviews some recent economic literature that can offer useful insights to policymakers.

The largest banks are now very complex organizations — complex enough that regulators place full-time examiners on-site, rather than conduct periodic regulatory examinations, as they did until the 1990s. This is just one symptom of the greater difficulty of keeping a close watch on the activities of giant financial companies engaged in a continually changing mix of activities. Even with examiners working full time at individual banks, regulators face a complex job keeping pace with new financial products and activities with uncertain implications for bank risk.

Mitchell Berlin is an assistant vice president and economist in the Research Department of the Philadelphia Fed.

Regulators have increasingly been attracted to the idea that market participants — bondholders, depositors, and, perhaps, stockholders — can play a significant role in disciplining banks, that is, pressuring banks to limit their risk-taking. Banking economists argue that market participants have strong incentives to evaluate the creditworthiness of banks in which they invest.[1] Some argue that market discipline can substitute for regulatory discipline to a significant extent, while others view the two as potentially complementary.

But to evaluate a bank's creditworthiness, investors must have good information about the bank's activities and balance sheet. One part of the new Basel Accord (Basel II) proposes to improve market discipline through enhanced disclosure requirements for banks. (See *Market Discipline: The Third Pillar.*)

Putting aside the costs of producing all this information, it might seem that more required disclosures must be a good thing.[2] Economists have long noted that firms may produce and disclose too little information because they can't capture all the gains from producing it; others can't be excluded from using the information themselves. But both economic theory and empirical evidence suggest a more circumspect approach and raise questions about the benefits of mandatory disclosure: (1) Empirical evidence and theoretical work indicate that firms disclose information voluntarily. Under what conditions will they disclose too little? (2) Is more information necessarily better? After all, banks are specialists in producing information. Might more disclosure undermine bank profitability? (3) Bank regulators already examine banks. Would more information for investors merely be redundant?

Current economic knowledge offers no definitive answers to these questions, but the recent economic literature can offer some useful insights to policymakers.

[1] Investors are not the only market participants who might impose discipline. Customers — for example, borrowers with loan commitments from the bank — will be concerned about the bank's creditworthiness.

[2] Of course, a complete evaluation of the net benefits of disclosure should not ignore these costs, as discussed at length in Sherrill Shaffer's article.

Market Discipline: The Third Pillar of Basel II

The Basel Accord, an agreement reached in 1988 by the banking regulators of the G-10 countries, sets common standards for capital adequacy and risk management for banks. Although the document has no legal status, most countries have adopted the accord's guidelines.

The New Basel Accord, or Basel II, will create new guidelines for capital adequacy and risk management. Implementation of Basel II is expected to occur in 2007.

The proposed third pillar of Basel II, which covers market discipline, has three sections:*

(1) Bank holding companies (BHCs) would provide detailed information about their corporate structure, that is, a full description of the BHC's subsidiaries and ownership positions in other firms. While information about ownership positions is already reported routinely in the U.S., the reporting requirements for banks in other countries are not uniform, and bank regulators around the world don't all have equal powers to compel banks to provide such information. It is difficult to evaluate the risk of the firm embedded in a labyrinthine organizational structure. Without detailed information about the organizational structure, it is impossible. (2) BHCs must also provide a complete account of how they calculated their capital level, for example, providing details about the required capital for complicated or innovative financial instruments. The capital requirements are the first pillar of Basel II. (3) Basel II will also require both qualitative and quantitative disclosures concerning the BHC's risk position, including (a) credit risk, mainly, the risk of default on the bank's loans; (b) market risk, the risk of loss on the bank's trading portfolio; (c) the risk of the BHC's equity positions in firms; and (d) operational risk, the risk of system breakdowns. Consider the disclosures concerning credit risk. The qualitative disclosures would include: (a) a detailed discussion of the BHC's risk management policies; (b) an account of the relationship between external ratings systems and the bank's internal method for assigning loans to risk classes; (c) definitions of past due and impaired loans. The quantitative disclosures would require information about the BHC's credit exposures broken down by industry concentration, geographic concentration, and counterparty concentration. (A counterparty is any customer whose default would affect the bank's profits.)

* See Jose Lopez's article for a more detailed summary of the third pillar.

MANDATORY REQUIREMENTS MAY BE UNNECESSARY

Firms May Disclose Voluntarily to Signal Quality. While most business observers would probably agree that firms are often reluctant to disclose bad news voluntarily, several classic articles explain that firms may have powerful incentives to disclose all information, both good and bad, voluntarily.[3]

Sanford Grossman, Oliver Hart, and Paul Milgrom consider markets in which: (1) firms can disclose information at very low cost; (2) no firm can be forced to disclose information involuntarily; and (3) a firm suffers heavy penalties for disclosing false information; thus, it can remain silent, but it can't lie. In the literature, information is called *verifiable* when the firm can't lie. The third assumption is not as unrealistic as it may at first seem. Even when the law doesn't impose penalties for lying, firms may opt to increase their penalties for misrepresentation by offering warranties. (Later, I'll discuss how Grossman, Hart, and Milgrom's conclusions change when these assumptions are relaxed.)

To take a concrete example, consider the market for washing machines. Washers are familiar consumer goods that vary widely in quality. Think about durability as the main issue for buyers of washing machines. Why would a firm whose machine breaks down frequently disclose this information to customers?

The idea is simple. As long as customers remain skeptical and interpret a firm's silence as an admission of very low quality, the following will happen. Firms with the most

[3] The articles are those by Sanford Grossman, by Grossman and Oliver Hart, and by Paul Milgrom.

durable machines will certainly disclose this information to distinguish themselves from all others. In turn, firms with slightly less durable machines will disclose this to distinguish themselves from all others with even lower quality. (Since all claims are verifiable, these firms can't falsely claim that their machines are more durable than they actually are. At best, a firm can remain silent.) Continuing this logic of *unraveling*, all firms will truthfully and voluntarily disclose information about the durability of their washing machines except perhaps the lowest quality producer.

Voluntary Disclosure Reduces the Wasteful Production of Information. Douglas Diamond's paper provides a second explanation for why firms might disclose information voluntarily. By disclosing information, the firm dissuades its investors from wasting the time and effort of collecting information, and it also increases the accuracy of the information that investors use to price the firm's securities.

One part of Diamond's conclusion is obvious. If producing information is costly — and anyone who has prepared her own tax forms knows that collecting, organizing, and reporting information takes time, effort, and money — having a single firm produce information can save expenses for lots of investors who can then avoid duplicative research. Investors are willing to pay more for a firm's stock when they don't have to perform as much costly research.

The second part of Diamond's argument is more subtle. The price of a firm's stock is the result of trading decisions by lots of investors who sell the stock when they think its current price is too high and buy more stock when they think its current price is too low. That is, the stock price incorporates the judgments of many investors. Suppose each investor makes his or her judgment based on research into the firm, for example, by reading a report from an investment research firm. Of course, any two people looking at the same report will interpret it slightly differently, and part of this difference will simply be *noise* — in this case, differences in individual judgments that are essentially random and unrelated to the firm's true profitability. Noise is just another source of uncertainty that reduces the firm's stock price because investors don't like risk.

If producing information is costly, having a single firm produce information can save expenses for lots of investors who can then avoid duplicative research. Investors are willing to pay more for a firm's stock when they don't have to perform as much costly research.

Compared with a situation in which individual investors must produce the information on their own, the amount of noise can be reduced substantially if the firm releases information on its own. Thus, the information about the firm's profitability is more accurate, and the firm's stock price will be higher. So it will pay for the firm to release information, which raises the value of its own stock by preventing its investors from engaging in duplicative, noisy research. Note, as long as investors are capable of uncovering bad news about the firm's profitability, this argument holds for disclosures of both bad news and good news.[4]

More Information Is Not Necessarily Better. Our folk wisdom is filled with homilies celebrating the virtues of ignorance: Curiosity killed the cat. What you don't know can't hurt you. Jack Hirshleifer was probably the first to articulate and explore the idea that beneficial insurance arrangements may be impossible when individuals have too much knowledge.

Consider the health-insurance market. Suppose some fraction of the population has a strong genetic predisposition to contract a disease for which no cure exists. In this situation, the availability of insurance makes everyone better off. Everyone pays a premium, although only those who actually contract the disease receive insurance payments. Those who contract the disease can use the insurance payments to cover their spouse's home payments and their children's college expenses, while those who never contract the disease have more peace of mind because they know their families are protected if they do. (Note: When we think about the value of insurance, it makes sense to take the perspective of someone *before* he or she contracts the disease.)

Now, if scientists discover a low-cost way to uncover a genetic marker that routinely predicts the disease — but without offering a cure — everyone will be worse off because insurance markets thrive on uncertainty. Individuals who learn they won't contract the disease will refuse to pay premiums, and the insurance market will break down for lack of funding. Since there is no

[4] The recent corporate governance scandals show that crooked accountants can make it very hard for investors to collect accurate information about a firm. Later, I will discuss models in which firms may lie as well as refuse to disclose information.

cure, the information has no value to those who will contract the disease. In this example, more information clearly makes everyone worse off.

Although few real world examples are quite as straightforward as this, many financial institutions and contracts have a risk-sharing function. One function of banks is to provide individuals or firms with protection against liquidity shocks, that is, a sudden need for funds. Thus, a bank is (partly) a web of insurance contracts. Individuals are willing to deposit their funds and firms are willing to pay upfront commitment fees for credit lines so that they can borrow — at attractive terms — should they suddenly need funds.

Such risk-sharing contracts may not be feasible if depositors are fully informed about the loans in the bank's portfolio and if the interest the bank must pay depositors is highly sensitive to available information.[5] Depositors would demand a high interest rate whenever the bank was providing funds to many firms with sudden liquidity needs, and the bank could no longer profitably provide insurance against liquidity shocks. Of course, nobody would ever propose that banks disclose detailed information about each loan in their portfolios. But this example shows there are limits to the gains from disclosure in light of financial intermediaries' insurance functions.[6]

In a related vein, Oved Yosha, among others, has argued that banks are specialists in maintaining proprietary information about their loan customers. In his article, Yosha argues that some firms avoid public securities markets and borrow from banks to avoid revealing proprietary information to their competitors. This places limits on the information that banks can reveal about their customers while remaining profitable. For example, there are limits to the disclosures that banks, when acting as swap dealers, can make about their customers without revealing and undermining their customers' hedging strategies.[7]

Like many other classic results in economics, the Grossman, Hart, and Milgrom unraveling result — that all firms will be forced to reveal the truth voluntarily — can be enlightening even for those who disagree with its conclusion.

FIRMS MAY NOT DISCLOSE VOLUNTARILY

Like many other classic results in economics, the Grossman, Hart, and Milgrom unraveling result — that all firms will be forced to reveal the truth voluntarily — can be enlightening even for those who disagree with its conclusion. A disciplined way to think about the issue is to ask: How do the Grossman, Hart, and Milgrom results change when each of their strong assumptions is relaxed? When we approach the question this way, we can gain insight into the conditions under which voluntary disclosure may not occur and when mandatory disclosure may help.[8]

Disclosures Are Biased When Misrepresentation Is Possible. When penalties for misrepresentation are moderate and if firms can sometimes lie without getting caught, the Grossman, Hart, and Milgrom results must be qualified (although not overturned). In their working paper, Evelyn Korn and Ulf Schiller find that instead of each firm revealing its true quality, firms divide into two groups when the penalty for lying is relatively low. High-quality firms voluntarily disclose, but they all make the same report. Thus, reports are biased upward for all but the highest quality firms. In contrast, low-quality firms don't disclose at all, so they are indistinguishable from each other.[9]

Two points about Korn and Schiller's findings should be kept in mind. Although firms don't disclose

[5] See my article with Loretta Mester for empirical evidence that U.S. banks offer firms intertemporal insurance against liquidity shocks and that the insurance offered declines when banks have fewer core deposits, that is, when banks' funding costs are more sensitive to changing market conditions.

[6] Charles Jacklin first made the argument that too much information might undermine banks' risk-sharing functions. However, we should be careful not to take this argument too far. Too little information about bank risks can undermine depositors' willingness to place their funds in banks.

[7] A simple interest rate swap may involve two firms that exchange interest payments on their debt. For example, a firm whose interest rate fluctuates may agree to swap with a firm whose interest rate is fixed.

[8] Note, mandatory disclosure doesn't automatically help when voluntary disclosure is inadequate. For example, assume that a firm can be forced to disclose what it knows but that the firm must undertake costly investigations or testing to learn the quality of its own product. In this case, Steven Matthews and Andrew Postlewaite argue that mandatory disclosure rules may *reduce* actual disclosure by leading a firm to choose to remain ignorant about its product's quality.

[9] Korn and Schiller find that when penalties are higher, a range of middle-quality firms that disclose truthfully can arise, thus moving even closer to Grossman, Hart, and Milgrom's results.

all information, as in Grossman, Hart, and Milgrom's model, the basic logic of the unraveling argument still remains: High-quality firms distinguish themselves from low-quality firms by making an informative (albeit biased) disclosure. Also, Korn and Schiller's results don't provide any clear rationale for mandatory disclosure requirements.

Firms May Not Disclose If Too Few Customers Are Sophisticated. An implicit assumption of the Grossman, Hart, and Milgrom model is that all customers are sophisticated enough to understand the disclosure. While this is certainly plausible in many cases — markets for familiar consumer goods — it is less convincing for other cases. For example, the implications of a firm's quarterly report for its future profitability (and therefore for its stock price) may be hard for many investors to interpret.

Michael Fishman and Kathleen Hagerty's 2003 article shows that voluntary disclosure depends on sophisticated customers who act as policemen for the market. The authors relax only one of Grossman, Hart, and Milgrom's assumptions: Firms still can't misrepresent their quality, but only a fraction of customers can evaluate the firm's disclosure. Fishman and Hagerty examine the case of a monopolist, say, a producer of a revolutionary new washing machine.[10] The performance of the washing machines is partly a matter of chance; managerial troubles at the plant might lead to a run of low-quality products.

What happens if the monopolist makes no disclosure? Since all customers are equally ignorant about the quality of the firm's washing machines, the firm will charge a single price to all customers and both sophisticated and unsophisticated customers will buy both high- and low-quality washers.[11] Customers know there is some likelihood of buying a high-quality machine and some likelihood that their washer will be a pile of nuts and bolts in a matter of months; therefore, the price is the average value customers place on a washing machine of unknown quality.

Does the Grossman, Hart, and Milgrom unraveling logic hold here? That depends on the reasoning of a firm with high-quality goods, which, in turn, depends on the number of sophisticated customers. The firm will reason: "If I disclose my quality and raise my price, I will attract customers sophisticated enough to evaluate my claims, but I will lose all of my unsophisticated customers because of the higher price." If the fraction of sophisticated customers is low, the high-quality firm would prefer not to disclose; there is no unraveling and no voluntary disclosure. However, if the fraction of sophisticated customers is high, the unraveling logic leads to full disclosure.

Interestingly, Fishman and Hagerty show that mandatory disclosure may be valuable when the shortage of sophisticated customers leads to too little disclosure. With mandatory disclosure, sophisticated customers are better off because they avoid purchasing the low-quality washers. Unsophisticated customers are no worse off, as long as they pay a price no higher than they would in the case without disclosure. Of course, the firm opposes mandatory disclosure because it sells fewer (low-quality) goods and its profits are reduced.[12]

Standardization Can Make Disclosure More Informative. While most of the literature treats disclosure as a simple, verifiable, one-dimensional statement — for example, "My washing machine will last three years without needing repairs" — real-world disclosures are often complicated mixtures of verifiable and unverifiable information: think of a corporation's quarterly profit report as an example. In their 1990 article, Fishman and Hagerty show that this complication may actually provide a rationale for mandatory disclosure rules. In particular, they demonstrate that disclosures can be made more informative if firms are given less discretion over what they may disclose.

In their model, firms can make verifiable disclosures, but only for a subset of the types of information they might choose to disclose. The problem with full discretion is that with enough flexibility, firms can always find something positive to report, so skeptical customers discount all positive information. By limiting the types of information the firm may disclose, a mandatory disclosure rule can increase firms' credibility by increasing the difficulty of reporting positive information. For example, a rule might require firms to report a single standardized quarterly profits figure and not allow them to report self-serving, pro forma information about profits.

When Firms Have Correlated Returns, They May Disclose Too Little. Consider a market with multiple banks with significant risk exposures to the telecom sector. If any

[10] For Fishman and Hagerty's results, it is not essential that the firm be a monopolist, only that the market structure permit firms to make positive profits in equilibrium; that is, they have some market power.

[11] In fact, Fishman and Hagerty show that different outcomes are also possible for the same underlying conditions, and they discuss the various reasons for choosing one outcome as the most reasonable. I simplify things by leaving these complications aside.

[12] Fishman and Hagerty present conditions in which the gains to customers outweigh the firm's losses.

one bank discloses information about the performance of its telecom loans, rational investors will also re-evaluate the prospects for other banks, even if the other banks don't disclose any information about their own portfolios. In cases where firms have correlated returns and disclosure is costly — think of the time and effort of producing and communicating the information to investors — Anat Admati and Paul Pfleiderer's article shows that each firm may have inadequate incentives to disclose voluntarily because it doesn't take into account the benefits of its disclosure for other firms and their customers.[13]

Actually, Admati and Pfleiderer make a somewhat stronger point. In their model, a firm can raise the value of its stock by committing to disclose information that increases the accuracy of the information available to investors. But in many financial situations, small increases in accuracy will have no value for the firm; the firm will find it unprofitable to bear any costs unless there is a relatively *large* increase in accuracy.[14] While this would not keep an isolated firm from increasing its expenditures to increase the accuracy of its investors' information, firms with correlated returns may get stuck in a situation in which no firm discloses at all. Each firm, as well as each firm's investors, would be better off if the firms could agree to produce more information.

While Admati and Pfleiderer's analysis suggests that mandatory disclosure rules may be beneficial, their main conclusion is that while firms may have inadequate incentives to disclose, it is very difficult to draw practical conclusions about when mandatory disclosure rules would actually improve matters.[15]

> The weight of the evidence indicates that neither regulators nor market participants are superfluous.

DISCLOSURE AND BANK REGULATION

One factor that differentiates banks from many other firms is that banks are heavily regulated. Most relevant, regulators routinely examine banks and put pressure on those banks found to be excessively risky. This raises an obvious question: With regulators already on the job, would more disclosure by banks to the public be redundant? Do we need to worry about providing more information to bank investors — bondholders, stockholders, and depositors — as long as regulators are watching over banks for them?

Of course, a similar question might be posed from the opposite direction. Financial economists have traditionally viewed financial markets as places where investors, driven by the profit motive, have very strong incentives to produce and process information about firms. Some economists have wondered whether market discipline — aided by extensive disclosures — would be an effective *substitute* for regulatory discipline of banks.[16]

It is important to note here that when we discuss banks and disclosure, there is a shift in emphasis from much of the literature on disclosure. Historically, bank regulators have been particularly concerned about the *safety and soundness* of banks, that is, the likelihood that banks will experience financial problems or failure. Furthermore, regulators are concerned that individual bank failures might have wider economic repercussions. Thus, bank regulators would not view market discipline as a successful substitute for regulatory discipline if a bank with insured depositors could choose a high-risk investment strategy, make full disclosure of the risks to all market participants, and sell its securities to investors with an appetite for high-risk, high-return investments.[17]

Do Market Participants Have Useful Information That Regulators Do Not (and Vice Versa)?

To sort out these issues, we must first determine whether bank regulators and market participants actually know different things and if they learn them at different times. The weight of the evidence indicates that neither regula-

[13] This is an application of the argument that information may be under-produced since it is a public good.

[14] Think, for example, of a firm considering selling stock to uninformed investors. If stockholders suspect that insiders have adverse information about the firm's prospects, they may be unwilling to buy the stock at any price. To increase the price investors are willing to pay for the stock, the firm can make disclosures that increase the accuracy of outside investors' information about the firm. But it may take a large increase in accuracy before the stock price rises high enough to make the sale worthwhile for the firm.

[15] Admati and Pfleiderer also discuss the possibility of subsidizing information production as an alternative to mandatory disclosure rules.

[16] The recent literature on the potential role for market discipline of banks has addressed a number of questions that I don't consider in this article. See Flannery and Nikolova's review and the introduction to Krainer and Lopez's working paper for more complete reviews of the literature.

[17] I won't discuss whether bank regulators' perspective is the correct one. The main argument for this perspective is that instability at one bank can generate instability for other banks and for the rest of the economy. An individual bank's investors and customers won't take these external effects into account.

tors nor market participants are superfluous. For example, Allen Berger, Sally Davies, and Mark Flannery find evidence that the information generated by bank regulators and by market participants is complementary; that is, each has information about banks' performance that the other doesn't.

They show that a change in a bank's credit rating from Moody's — a measure of the information available to bond market participants — and a change in the bank's regulatory rating *both* help predict changes in the bank's financial condition. Neither regulatory rating changes nor rating agency changes are superfluous. Interestingly, the bank's regulatory rating has predictive power only if the rating is of recent vintage, that is, only if the bank was examined on-site no earlier than the previous quarter.[18]

How Does Increased Disclosure Affect Market Discipline? It is important to keep in mind that most information is not like manna from heaven; it must be produced, and to be useful, it must be interpreted. We can't assess the effects of more mandatory disclosure without asking how it would affect market participants' willingness to produce information and to pay for the services of specialists in interpreting information, such as industry analysts.

Industry analysts are a major channel through which information disclosed by firms is interpreted and disseminated in a useful form to investors. It is certainly possible that more disclosure, especially more standardized disclosure, might make banks' performance easier to decipher, thus reducing the profitability of interpreting banking industry data and leading analysts to concentrate their attention on other industries. Thus, mandatory disclosure could, in principle, *reduce* investors' ability to interpret the information by reducing analyst coverage of the banking industry.

However, evidence from the empirical accounting literature indicates that more disclosure *increases* information production.[19] In particular, firms that disclose more (according to a number of different measures of the quality of disclosure) are covered by more analysts. To be sure, we should interpret these empirical results with some care. The precise measures of the quality of disclosure are controversial, and the studies don't completely rule out the possibility that analyst coverage and disclosure are related through some factor that has nothing to do with the quality of the information disclosed.[20]

These caveats aside, the positive relationship between disclosure and analyst coverage suggests that the information available to market participants won't reduce their incentive to produce information. Instead, disclosure and information production may be complements.

Can Market Discipline Substitute for Regulatory Discipline? The production of information about a firm and the ability to affect the firm's behavior are not the same thing. There is now ample empirical evidence that bank bondholders and depositors respond to information about banks' creditworthiness in a timely and sensible way.[21] When a bank's riskiness rises — for example, if a bank announces that expected losses on its loan portfolio have increased — its bond prices fall, the rates it pays for uninsured deposits rise, and many uninsured depositors find another bank. However, the need to pay bondholders and depositors a higher rate may not have much of an effect on the behavior of banks' managers.

The article by Matthew Billet, Jon Garfinkel, and Edward O'Neal provides evidence that it may be unusually difficult for a bank's investors to effectively discipline the bank's managers because of the availability of insured deposits. In their article, they showed that when rating agencies downgrade a bank's bonds, the bank typically substitutes toward insured deposits; that is, the bank increases the total quantity of insured deposits, as well as the share of insured deposits among various funding sources.

The authors' argue that this shift toward insured deposits indicates that banks view regulatory discipline as *less burdensome* than market discipline. In response to the rating agency's downgrade, the bank must

[18] This finding is consistent with much of the recent literature, which finds that regulatory ratings get stale within half-a-year. See the references in Flannery's 1998 article.

[19] See Paul Healy and Krishna Palepu's review of the empirical literature on disclosure for the relevant references.

[20] For example, high-tech firms might have strong incentives to make disclosures and may independently have a relatively high following by analysts. We might conclude (incorrectly) that better disclosure leads to high analyst coverage.

[21] The literature prior to the 1990s suggested that bank creditors were not responsive to changes in bank risk. (But see Daniel Covitz, Diana Hancock, and Myron Kwast's recent article for evidence that bondholders were more sensitive than previously believed.) It is widely believed that investors have become more responsive because they no longer believe that the government will bail out failing banks' investors (except for insured depositors). However, there is a line of thought suggesting that one ground for the special treatment of banks (the safety net and extensive regulatory monitoring) is that they are unusually *opaque* — difficult for investors to analyze— compared with other types of firms. This issue clearly relates to the potential gains from mandatory disclosure. To date, this literature is inconclusive. For opposing views, see the articles by Donald Morgan and by Mark Flannery, Simon Kwan, and M. Nimalendran.

pay a higher rate to retain uninsured deposits. This higher rate is the primary way in which market participants discipline the bank for an increase in risk. Regulatory discipline could take a number of forms. Although the bank doesn't suffer an explicit regulatory penalty for heavier reliance on insured deposits, an increase in bank risk could lead regulators to more closely monitor the bank's activities. The bank's substitution of insured deposits for uninsured deposits is evidence that the higher cost of retaining insured deposits is *lower* than the cost of closer monitoring by regulators.[22]

Can Market Discipline Supplement Regulatory Discipline?

If market participants have information about banks that regulators don't — especially if the information is more current — but if we have doubts about bank investors' ability to discipline bank management, we might nonetheless hope that regulators could make increased use of market information. In this middle view, banks would be required to make more detailed information available to market participants, and bank regulators would pay closer attention to market information, for example, stock and bond prices, to help identify potential problem banks.

John Krainer and Jose Lopez's article provides evidence that by paying attention to market signals, regulators might enhance their ability to detect developing problems in a timely way. They show that a bank's excess stock returns — the part of the bank's stock return that doesn't result from the general movements in all stocks — predict the likelihood of a regulatory downgrade up to one year ahead. That is, the bank's stock price will fall up to a year ahead of a regulatory downgrade. This finding is especially interesting because some banking scholars have argued that stock prices may not be helpful to regulators in light of stockholders' and regulators' potentially opposing views about risk.

If we have doubts about bank investors' ability to discipline bank management, we might nonetheless hope that regulators could make increased use of market information.

Higher stock returns could mean that investors believe that a bank's risky lending strategy is likely to be highly profitable, even though it increases the probability of default.

On the other hand, stocks trade much more widely and frequently than do bonds, so we expect stock prices to respond to new information more rapidly than do bond prices.[23] Lopez and Krainer's evidence suggests that, in practice, despite the potentially opposing interests of stockholders and regulators, stock prices may have useful information for regulators.[24]

CONCLUSION

The literature on information disclosure provides some useful perspectives on the potential benefits of enhanced disclosure requirements for banks as proposed in Basel II.

While firms may have incentives to disclose both good and bad information voluntarily, there are plausible instances in which mandatory disclosure rules would increase the information disclosed by firms and make the firm's customers and investors better off. If most investors find it difficult to interpret the firm's disclosures, mandatory disclosure will increase disclosure and make customers better off. When firms are disclosing large amounts of (partially) unverifiable information, standardized disclosure may help.

From the accounting literature, the empirical evidence suggests that disclosure can lead to a virtuous circle, in which more disclosure by firms can increase information production by market participants.

Finally, the banking literature provides reasons to believe that providing better information to market participants may provide useful information to regulators; thus, market discipline and regulatory discipline may be complements.

[22] Robert Bliss and Mark Flannery's article also provides evidence that market discipline may not be very effective.

[23] Diana Hancock and Myron Kwast's article discusses some of the difficulties of using subordinated bond price spreads as indicators of bank risk.

[24] Krainer and Lopez find somewhat weaker evidence that excess returns predict regulatory downgrades over and above bank balance-sheet information that is already available to regulators.

REFERENCES

Admati, Anat, and Paul Pfleiderer. "Forcing Firms to Talk: Disclosure Regulation and Externalities," *Review of Financial Studies*, 13, 2000, pp. 479-519.

Basel Committee on Banking Supervision. Consultative Document, Pillar 3 (Market Discipline), May 2003.

Berger, Allen, Sally Davies, and Mark Flannery. "Comparing Market and Supervisory Assessments of Bank Performance: Who Knows What When?" *Journal of Money, Credit and Banking*, 32, August 2000 Part 2, pp. 641-66.

Berlin, Mitchell, and Loretta Mester. "Deposits and Relationship Lending," *Review of Financial Studies*, 12, Fall 1999, pp. 579-607.

Billet, Matthew, Jon Garfinkel, and Edward O'Neal. "The Cost of Market Versus Regulatory Discipline in Banking," *Journal of Financial Economics*, 48, 1998, pp. 333-58.

Bliss, Robert, and Mark Flannery. "Market Discipline in Governance of U.S. Bank Holding Companies: Monitoring versus Influencing," in F. Mishkin (ed.): *Prudential Supervision: What Works and What Doesn't?* NBER, University of Chicago Press, 2001.

Covitz, Daniel, Diana Hancock, and Myron Kwast. "A Reconsideration of the Risk Sensitivity of U.S. Banking Organization Subordinated Debt Spreads: A Sample Selection Approach," *Economic Policy Review*, Federal Reserve Bank of New York, September 2004.

Diamond, Douglas. "Optimal Release of Information by Firms," *Journal of Finance*, 60, September 1985, pp. 1071-94.

Fishman, Michael, and Kathleen Hagerty. "The Optimal Amount of Discretion to Allow in Disclosure," *Quarterly Journal of Economics*, May 1990, pp. 427-44.

Fishman, Michael, and Kathleen Hagerty. "Mandatory Versus Voluntary Disclosure in Markets with Informed and Uninformed Customers," *Journal of Law, Economics, and Organization*, 19, 2003, pp. 45-63.

Flannery, Mark. "Using Market Information in Prudential Bank Supervision: A Review of the U.S. Empirical Evidence," *Journal of Money, Credit, and Banking*, 30, August 1998, Part 1, pp. 273-305.

Flannery, Mark, Simon Kwan, and M. Nimalendran. "Market Evidence on the Opaqueness of Banking Firms' Assets," *Journal of Financial Economics*, March 2004.

Flannery, Mark, and Nikova Stanislava. "Market Discipline of U.S. Financial Firms: Recent Evidence and Research Issues," Working Paper, University of Florida, November 2003.

Grossman, Sanford. "The Informational Role of Warranties and Private Disclosure About Product Quality," *Journal of Law and Economics*, 24, 1981, pp. 461-83.

Grossman, Sanford, and Oliver Hart. "Disclosure Laws and Takeover Bids," *Journal of Finance*, 35, pp. 323-34.

Hancock, Diana, and Myron Kwast. "Using Subordinated Debt to Monitor Bank Holding Companies: Is It Feasible?" *Journal of Financial Services Research*, 20, October/December, 2001, pp. 147-87.

Healy, Paul, and Krishna Palepu. "Information Asymmetry, Corporate Disclosure, the Capital Markets: A Review of the Empirical Disclosure Literature," *Journal of Accounting and Economics*, 31, 2001, pp. 405-40.

Hirshleifer, Jack. "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, 64, June 1974, pp. 373-74.

Jacklin, Charles. "Demand Deposits, Trading Restrictions, and Risk-Sharing," in Edward Prescott and Neil Wallace (eds.): *Contractual Arrangements for Intertemporal Trade*, 1987, pp. 26-47.

Korn, Evelyn, and Ulf Schiller. "How to Create a Hype. Voluntary Disclosure of Partially Verifiable Information," Working Paper, University of Tubingen, February 2002.

Krainer, John, and Jose Lopez. "Incorporating Equity Market Information Into Supervisory Monitoring Models," Working Paper, Federal Reserve Bank of San Francisco, 2002.

Lopez, Jose. "Disclosure as a Supervisory Tool: Pillar 3 of Basel II," *Economic Letter*, Federal Reserve Bank of San Francisco, July 2003.

Matthews, Steven, and Andrew Postlewaite, "Quality Testing and Disclosure," *Rand Journal of Economics*, 16, Autumn 1985, pp. 328-40.

Milgrom, Paul. "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 12, 1981, pp. 380-91.

Morgan, Donald. "Rating Banks: Risk and Uncertainty in an Opaque Industry," *American Economic Review*, 2003.

Shaffer, Sherrill, "Rethinking Disclosure Requirements," Federal Reserve Bank of Philadelphia *Business Review*, May/June 1995, pp. 15-29.

Yosha, Oved. "Information Disclosure and the Costs of Financing Source," *Journal of Financial Intermediation*, 4, January 1995, pp. 3-20.

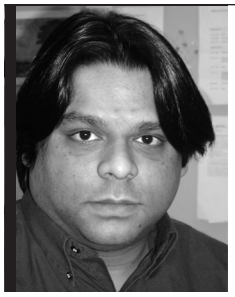
Why Are Married Women Working More?
Some Macroeconomic Explanations

BY AUBHIK KHAN

For the past 60 years, the number of hours worked per person in the U.S. has changed very little. Nonetheless, the labor force has undergone some pronounced shifts over that same period. One prominent change is the sharp increase in the number of hours worked by married women. In this article, Aubhik Khan discusses how the composition of the labor force has changed since 1945 and how macroeconomists explain these changes.

Since the Second World War, there has been little overall change in the number of hours worked per person in the United States. Hiding under this apparent constancy lie some pronounced shifts in the composition of the labor force. The share of employment attributable to men and women, to older workers, and to married households has changed, in some instances rather dramatically. How have these shares changed, and what does recent economic research have to say about these compositional changes?[1]

[1] My discussion of changes in hours worked uses the work of Ellen R. McGrattan and Richard Rogerson closely. The advanced reader should consult their 1998 and 2004 papers for a far more thorough analysis.



Aubhik Khan is a senior economist in the Research Department of the Philadelphia Fed.

Perhaps the most prominent change in the composition of the labor force has been the sharp rise in the hours worked by married women. Motivated by this rather striking phenomenon, macroeconomists have developed models to explain the asymmetric rise over the past 40 years in weekly hours worked by married women. Three basic changes in the economy likely have contributed to a rise in hours worked by married women (in no particular order of importance): (1) technological progress that has made durable consumer goods more productive; (2) a reduction in the gender wage gap associated with lower pay for women than for men; and (3) a change in social attitudes toward married women working outside the home.

CHANGES IN COMPOSITION OF THE LABOR FORCE

If we ignore differences in the sex, age, or marital status of workers and look at aggregate average hours worked, the number of weekly hours of market work per person has remained roughly constant over the postwar period from 1950 to 2000 (Table 1).[2] This is not to suggest that there have not been short-term fluctuations. For example, we know that hours worked per person fall during recessions (as firms' demand for employment decreases) and rise during expansions. However, aside from such cyclical fluctuations, the long-run value changed little between 1950 and 2000: The data indicate that average weekly hours worked per person were 22.34 in 1950 and rose slightly to 23.90 in 2000.

Average weekly hours are, of course, considerably less than the familiar 40-hour workweek, since not all persons are employed. However, the first indication that the aggregate measure hides changes across different groups of workers comes from an examination of the employment to population ratio. Over the same 50 years, this ratio has risen from 0.53 to 0.59. Thus, while five out of 10 people were working in 1950, 50 years later nearly six out of 10 people were employed in the economy. This substantial rise in the employment-to-population ratio and the smaller increase in average weekly hours per person together imply that the hours worked by the typical employed individual have fallen. Indeed, on average, workers worked two fewer hours per week in 2000 then they did in 1950.

Of course, the constancy of average weekly hours per person does

[2] I survey the postwar data using the decennial U.S. census, which is taken in the final year of every decade. All tables are based on data taken from McGrattan and Rogerson's 2004 article. I thank Ellen McGrattan for making these data available to me.

TABLE 1

Average Weekly Hours

Year	Average Weekly Hours Worked Per Person	Average Weekly Hours Worked Per Worker	Employment-to-Population Ratio %
1950	22.34	42.40	52.69
1960	21.55	40.24	53.55
1970	21.15	38.83	54.47
1980	22.07	39.01	56.59
1990	23.86	39.74	60.04
2000	23.90	40.39	59.17
% Change 1950-2000	6.98	-4.74	12.30

Source: Table based on data presented in McGrattan and Rogerson (2004); original source for the data is the U.S. census.

not reflect a constancy of earnings. Average real compensation per hour is a common measure of real labor earnings, which includes workers' benefits and controls for the effects of inflation on nominal earnings. Between 1950 and 2000, average real compensation per hour rose more than 150 percent (Figure 1). Thus, while workers are earning much more, the population as a whole is not working more.[3]

A SIMPLE ECONOMIC MODEL OF THE LABOR-LEISURE TRADEOFF

The constancy of hours worked in light of changes in wages is interesting to economists, as it offers some insight into workers' preferences. To see this, consider the following very primitive model of labor supply sometimes used by macroeconomists.

Assume for simplicity that each worker values two goods. We call the first *consumption*, a single commodity that represents all the different goods and services we use. The second good is *leisure*, which is not produced but is granted to the worker as time. The worker may devote his time to either leisure, which he values, or to labor, which earns him a wage.[4] Since wages pay for consumption goods, any worker faces this fundamental tradeoff: The time spent enjoying leisure could have been spent working for wages. Given any real wage — the amount of consumption goods that can be purchased with a given money wage — the worker must choose how much of his time to allocate to labor, and the remainder is leisure.

A rise in his wage will induce a *wealth effect*: With no change in hours worked, the worker is now wealthier. The wealth effect tends to make the worker consume more of most goods – economists call these *normal goods*.[5] Economists think of leisure as one such commodity. Thus, the wealth effect tends to reduce the quantity of labor supplied in response to a rise in wages as people wish to have more leisure. Nonetheless, since the worker may earn more than before — if he does not reduce his hours worked too sharply — both leisure and consumption will rise.

The rise in wages also implies a *substitution* effect: The cost of leisure has now risen, since each hour of leisure means an hour less of work, which is now worth more. As the cost of leisure rises, demand falls, and this by itself should increase the worker's hours of work. The wealth and substitution effects conflict, and, in general, there is no way to tell which will dominate. However, the observation that average hours worked per person have not changed in response to a 150 percent rise in earnings has led many macroeconomists to suggest that at least for the average or representative household in the economy, the wealth and substitution effects have offset each other. Thus, this offsetting is one explanation for the observed lack of trend in hours worked.

WOMEN'S WORK PATTERNS CHANGED DRASTICALLY

While this net cancellation may summarize the behavior of the representative household, it does not accurately reflect the changes in labor supplied by men and women, nor the young and old. There have been large movements in hours worked across all these groups. However, as we have already noted, the sum of these

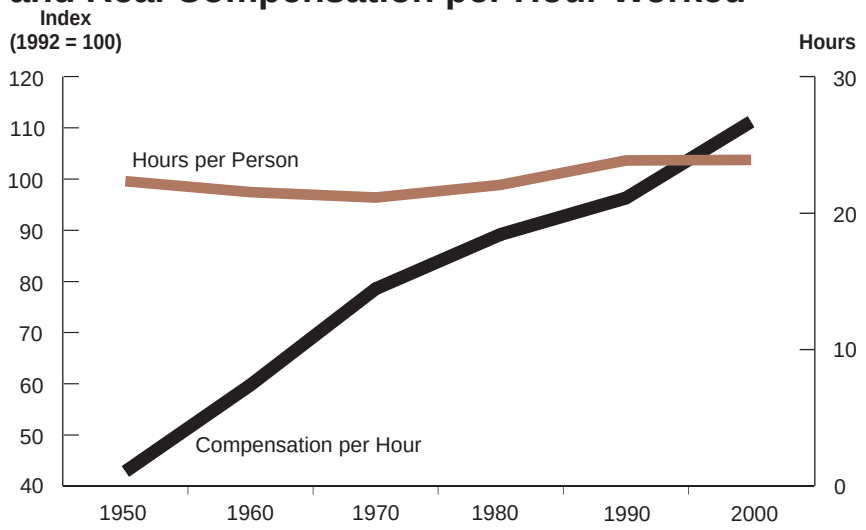
[3] Average real compensation per hour is a better measure of earnings than wages. For example, it would make little sense to focus on a measure that ignored health insurance provided by an employer. However, I will use the term wage in what follows as shorthand for compensation per hour.

[4] Thus, this simple economic model assumes we do not value jobs directly, but rather indirectly through the goods available to us as a result of earning a salary.

[5] For example, generic paper towels may not be a normal good. As your wage rises, you might switch to a name brand; hence, your expenditure on generic paper towels would fall.

FIGURE 1

Average Weekly Hours Worked per Person and Real Compensation per Hour Worked*



* The compensation series is an index of hourly compensation in the business sector, deflated by the consumer price index for all urban consumers. The index is constructed to equal 100 in 1992.

TABLE 2

The Distribution of Hours Worked by Gender

	Average Weekly Hours Worked per Person by Gender		
Year	Total Population	Males	Females
1950	22.34	34.18	10.87
1960	21.55	31.93	11.84
1970	21.15	29.72	13.32
1980	22.07	28.70	16.02
1990	23.86	29.11	19.03
2000	23.90	28.34	19.78
% Change 1950-2000	6.98	-17.09	81.97

Source: Table based on data presented in McGrattan and Rogerson (2004); original source for the data is the U.S. census.

movements has had little overall effect on average hours worked per person. Separating weekly hours per worker, we find that hours worked by males fell 17 percent, while hours worked by females rose an astounding 82 percent between 1950 and 2000 (Table 2).

Almost all of the rise in female hours worked is explained by an increase in the average weekly hours spent in employment by married women (Table 3). The weekly hours worked by married women ages 25 to 54 rose, on average, more than 200 percent! The corresponding figure for single women is actually -1.3 percent (second panel of Table 4).[6] This does *not* mean that single women are working less than married women. Rather, in 1950, the U.S. census shows married women working far fewer hours than single women. However, 50 years later, these differences had largely evaporated as the hours worked by married women rose to match the initially longer workweek of single women. Moreover, most of the change in married women's hours of market work happened between 1950 and 1990.

To see this clearly, take, for example, the weekly hours of married and single women, between 35 and 44 years of age, in 1950. The census conducted that year shows that, on average, married women in this age group worked about 9.5 hours a week. Single women in this age group worked far more: 30.5 hours a week. Now re-examine the weekly hours of women in the same age group, but 50 years later. In 2000, married women ages 35 to 44 were working 26 hours, on average. Single women in this age group worked an average of about 29.5 hours a week, actually slightly less than their predecessors 50 years ago.

Across all age groups, the length of the average workweek of married women (with spouses present) rose about 200 percent, from about 7 hours to over 20 hours a week, between 1950 and 2000 (Figure 2).[7]

[6] Generally, the changes in hours worked by single men and single women, of any age, are rather similar.

[7] The census refers to married men (or women) with spouses present as a separate group from married men (or women) with spouses absent. A married person with a spouse present is living in a household with the wife or husband. Married people with spouses absent include those with spouses in the military or living in institutions. The 2000 census indicates there are 2 million households composed of a married person with spouse absent. Hereafter, when I describe a married person whose spouse is present, I will simply describe them as married.

TABLE 3

Changes in Hours Worked by Married People

Status		Gender	Year	Weekly Hours Worked Per Person by Age (in Years) 15-24	25-34	35-44	45-54	55-64	65-74	75-99
Spouse Present*	*Total*	Males	1950	39.68	42.20	43.46	42.06	37.62	23.39	9.74
			1960	39.50	42.33	42.77	41.21	35.77	14.74	6.23
			1970	37.14	41.86	42.60	40.90	34.73	11.59	4.31
			1980	37.80	40.99	42.01	39.79	30.53	8.15	3.08
			1990	38.75	42.52	42.88	40.88	28.68	7.69	2.56
			2000	37.17	40.99	41.88	40.39	29.17	8.15	3.06
			% Change 1950-2000	-6.33	-2.87	-3.64	-3.97	-22.46	-65.16	-68.58
		Females	1950	9.03	7.93	9.43	8.43	4.42	1.67	0.52
			1960	10.00	9.10	12.35	13.56	8.66	2.27	0.98
			1970	14.67	12.23	15.04	16.26	11.77	2.53	1.22
			1980	18.95	19.25	20.13	18.61	12.18	2.44	0.79
			1990	21.75	24.36	25.95	24.51	14.00	2.84	0.70
			2000	20.49	24.49	26.03	27.27	16.99	3.38	1.02
			% Change 1950-2000	126.91	208.83	176.03	223.49	284.39	102.40	96.15

Source: Table based on data presented in McGrattan and Rogerson (2004); original source for the data is the U.S. census.

*A married person with a spouse present is living in a household with the wife or husband. Married people with spouses absent include those with spouses in the military or living in institutions.

Finally, it is important to emphasize that these figures are hours per person, not hours per worker. To a significant extent, the increase in hours worked by married women is due to their greater participation in the labor force. In sharp contrast to the behavior of married women, the average hours worked per single woman remained relatively unchanged, rising 11 percent (Figure 3). Over the same period, hours worked by married men with spouses present fell eight hours, or 20 percent. Finally, hours worked by single men fell 7 percent.

In their 2003 paper, Larry Jones, Rodolfo Manuelli, and Ellen McGrattan adopted an interesting perspective on these changes in the labor force. They noted that in 1950, a married couple's total hours worked in the market were much fewer than those we would obtain by summing the hours worked by the average single man and woman. Their census data indicate that this *artificial* couple (formed by combining a single man and a single woman) would have worked, on average, 60.5 hours a week in 1950; their hours worked would have changed little over the next 40 years, falling slightly to a little over 59 hours by 1990. As I mentioned, the total hours worked in the market by the married couple together was initially far less, 49.5, in 1950. However, by 1990, differ-

TABLE 4

Changes in Hours Worked by Single People

			Weekly Hours Worked Per Person by Age (in Years)						
Status	Gender	Year	15-24	25-34	35-44	45-54	55-64	65-74	75-99
Single	Males	1950	20.02	32.81	34.14	32.07	27.19	15.44	6.00
		1960	14.92	31.99	30.78	29.19	24.35	9.74	5.07
		1970	13.20	30.88	30.30	28.14	22.31	8.76	4.27
		1980	16.58	31.80	30.25	27.21	19.90	6.11	2.30
		1990	16.75	32.86	30.88	27.25	18.19	5.99	2.21
		2000	16.25	33.29	30.68	27.81	19.12	6.42	3.63
		% Change 1950-2000	-18.83	1.46	-10.13	-13.28	-29.68	-58.42	-39.50
	Females	1950	14.25	30.64	30.53	28.61	22.72	10.31	3.14
		1960	10.76	29.46	29.49	29.07	24.40	10.62	3.40
		1970	10.44	28.72	27.70	27.69	24.38	8.34	3.08
		1980	13.53	30.28	28.89	26.73	20.60	5.20	1.36
		1990	14.17	30.75	30.99	28.30	19.11	5.23	1.17
		2000	13.86	30.52	29.56	28.53	19.67	5.43	1.51
		% Change 1950-2000	-2.74	-0.39	-3.18	-0.28	-13.42	-47.33	-51.91

Source: Table based on data presented in McGrattan and Rogerson (2004); original source for the data is the U.S. Census.

ences in hours worked between these two pairs — the married couple and the artificial couple — had largely disappeared. The average married couple was working 61 hours by then. Thus, we see that in terms of total hours spent in the market, married couples are now behaving much more like single people. Why has the behavior of married households changed?

THE HOUSEHOLD EMPLOYMENT DECISION

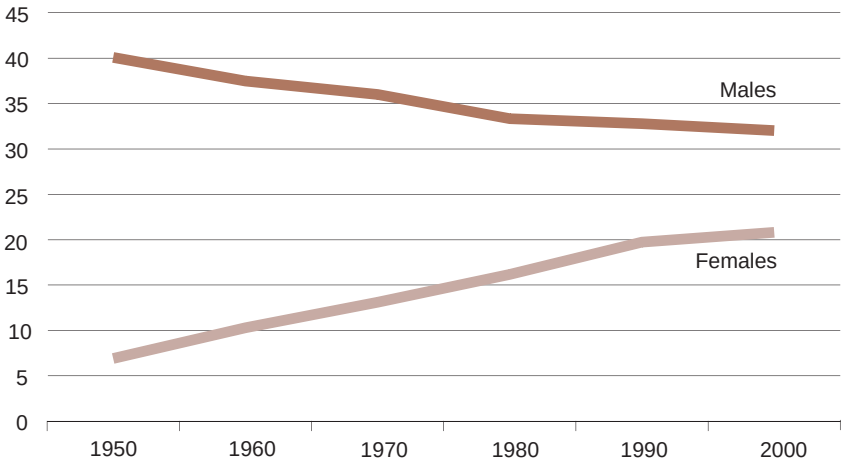
To answer the question about changes in hours worked, we must consider the determinants of hours worked. Many economic factors determine an individual's employment decisions. The number of dependents and earning ability are just two characteristics that come to mind. In turn, these characteristics are themselves affected by an individual's decisions about the number of children to have and the years of schooling to invest in. I won't attempt to discuss the general economic theory of labor supply, a rich theory that has been developed over several decades by many economists. Instead, I will focus on more recent macroeconomic models developed to understand why married women's hours of work in the market have risen so sharply.[8]

Three basic changes in the economy have likely contributed to a rise in hours worked by married women: (1) technological progress that has made durable consumer goods, such

[8] Given this focus, the models I will discuss will abstract from many issues that affect individuals' employment choices but do so more or less uniformly across individuals of different gender and marital status.

FIGURE 2

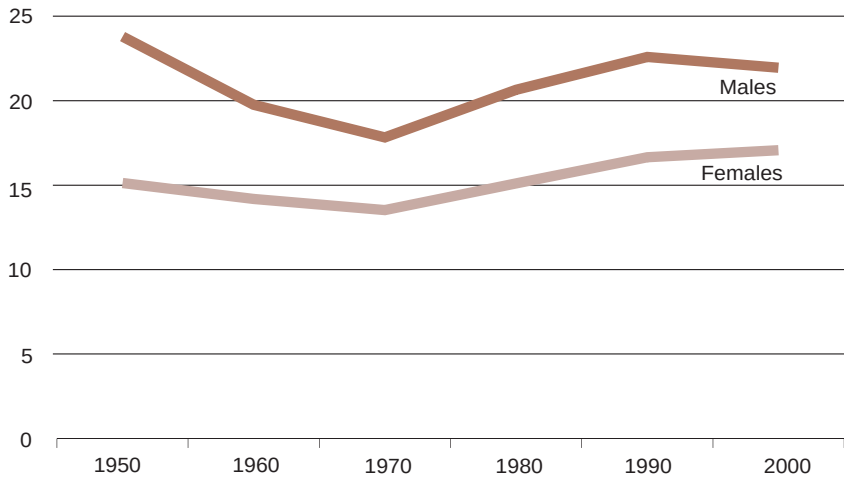
Average Weekly Hours Worked per Person Married, Spouse Present*



* A married person with a spouse present is living in a household with the wife or husband. Married people with spouses absent include those with spouses in the military or living in institutions.

FIGURE 3

Average Weekly Hours Worked per Person Not Married (Single, Widowed, Divorced, Married-Spouse Absent*)



* A married person with a spouse present is living in a household with the wife or husband. Married people with spouses absent include those with spouses in the military or living in institutions.

as home appliances, more productive, (2) a reduction in the gender wage gap that yields lower pay for women than for men for the same work, and (3) a change in social norms.

Producing at Home or Working in the Market. The data do not imply that married women are working more hours but that they are working more hours in formal employment, that is, in the market. This has come at the expense of fewer hours worked at home. A necessary starting point for understanding changes in married women's market hours is a discussion of how a household allocates the time of its adult members between the home and the market. This involves understanding the economic model of home production.

All households, whether composed of a married couple or a single adult, value goods bought in the market – for example, restaurant meals, wine, and the inevitable dose of aspirin – and goods produced at home – for example, breakfast.

To purchase market goods, the typical household must work in the market. The earnings from this employment allow the household to consume market goods. Other market goods a household buys are not directly used but are themselves inputs in the production of a different set of goods produced at home. These inputs are durable consumer goods, such as refrigerators and washing machines, that are used at home. Finally, some of the household's earnings may be saved toward future consumption.

Economists find it useful to think of three broad uses of time. Of the hours in a day not spent sleeping, the time may be spent engaged in home production, market work, or at leisure. An example of time spent in home production — labor that does not earn a wage or salary but contributes to the production of goods at home — is time spent washing one's

automobile or cooking a meal. A clean car and a cooked meal are both goods most of us enjoy. An example of time spent enjoying leisure may include driving that car across the countryside or eating that meal. Alas, there are only so many hours in the day, and some of those hours must be spent sleeping or engaged in basic activities such as bathing. It then follows that, given an individual's consumption of leisure, time spent in home production tends to reduce the amount of time an individual can work in the market. A household must then decide how much time its adult members should devote to market work and how much to home work.

Members of married households have historically specialized between work in the market and work done at home. As such, they provided an illustration of Adam Smith's famous theory of the division of labor. Once married, the majority of women spent all or most of their time in home production. Men specialized in market work. According to Adam Smith's theory, two individuals, each specialized in one task, are likely to be far more productive as a team than when each individual spends part of his or her time on each task. This division of labor is a large part of the reason that most people do not sew their own clothes and build their own houses. This specialization may well hold true for time a household divides between market and home production.

Technology may have had more subtle effects in leading households to specialize between market and home work. Because technological progress has enabled the production of more and more commodities, goods that once required substantial levels of home production are now widely and cheaply available in the marketplace. A prominent example is prepared food. Restaurant meals are far more common now than they were 50 or 100 years ago. Home delivery of meals was almost unknown until the postwar period. If, some time in the past, many of the goods consumed by a typical household involved a substantial level of home production, the household would have had to allocate far more time to home production. As more substitutes prepared in the market became available, the household was able to devote less time to home production and increase the time it devoted to market work. If the male adult was already working in the market, the female adult could then increase her hours of market work.

Members of married households have historically specialized between work in the market and work done at home.

MACROECONOMIC THEORIES OF THE RISE IN MARRIED WOMEN'S MARKET HOURS

Several macroeconomic theories seek to explain the pronounced increase in the hours worked in the market by married women. Each of these theories involves a completely specified model of the economy in which households' and firms' decisions interact to determine total quantities, such as total hours worked, and prices, such as real wages. These macroeconomic models make predictions about how hours worked by women will change in response to a change in some outside factor, such as a change in preferences or technology.[9]

The Gender Wage Gap. In the first macroeconomic theory we examine, the outside factor is a discrimination tax that leads to a gender wage gap. Researchers Ellen McGrattan, Rodolfo Manuelli, and Larry Jones studied the extent to which changes in married women's market hours may be explained by the gender wage gap. Their analysis does not attempt to explain the gap between men's and women's wages. Rather, taking this gap as given, they wanted to evaluate how much of the difference in hours worked in the market may be attributed to the gap. While the ratio of female to male wages has been less than one over the entire postwar period, the gap has been narrowing. In her paper, economist Francine Blau reported that in 1969, women who worked full time earned about 56 percent of what men earned. By 1994, they earned 72 percent of what men earned. Part of this difference in pay may be attributable to differences in occupation and skills. The remainder is effectively a tax on women's hours of work, which we call a discrimination tax.

Think of the problem faced by a household that must allocate the time of its two working-age adults to home production, market work, and leisure. If the woman's market wages are lower than the man's, the woman will spend more time at home. She may spend some time in market work, but it must be less than the time the man spends if his wage is higher.

[9] My survey of the macroeconomic literature does not include a large body of microeconomic empirical research into women's labor supply decisions. Such research has offered important insight into why married women's hours of work may have changed that is complementary to the macroeconomic theories we discuss here. For example, Lawrence Katz and Claudia Goldin discuss the role of birth control; Mark Rosenzweig and Paul Schulz study the effects of changes in fertility; and Lawrence Katz and Kevin Murphy examine the changes in the difference between male and female wages.

Lower market wages for women, at least those with the same skills as men, amount to a tax on women's wages relative to those of men. By devoting less of the woman's time to market work, the household avoids the tax.

As the gap between the man's and the woman's market wages narrows, the time spent by each in home and market work should become more nearly equal. This is exactly what we have seen. Indeed, as the wage gap associated with women's market work falls, the total amount of household members' time spent in the market may rise.

An interesting aspect of the gender wage gap theory is that it does not require a very large gap to explain the observed disparity between married men's and married women's hours. This is because even a small difference in the wages paid to women leads women to acquire less human capital, that is, invest less in education.

In such economic models, people invest in education because firms pay higher wages for more educated workers. Since the gender wage gap reduces women's wages, women find it less worthwhile to invest in education. A gap in women's wages relative to those of men reduces the returns to schooling for women relative to returns earned by men. Lower investment in education by women further reduces their potential market earnings, above and beyond that implied by the gender wage gap. Women's lower investment in human capital reinforces the extent of their specialization in the home.

The theory implies that as the gender wage gap narrows, women's investment in human capital should rise relative to that of men. There is evidence for this prediction. For example, in 1960, women, when compared with men, were 60 percent as likely to be college graduates. This fraction rose steadily until, in 2002, women were 88 percent as likely to be college graduates.[10]

Another interesting feature of the gender wage gap argument is that it can explain the large change in hours worked by married women without incorrectly predicting — contrary to the evidence — concomitant large movements in hours worked by single women. Single women, not having a partner, cannot specialize in home production. As single women value both home and market goods, and the latter cannot be bought without earning market wages, they will specialize far less than married women, despite the wage gap. The theory leaves unexplained the origin of the gender wage gap and how it might affect incentives to marry.

Another interesting feature of the gender wage gap argument is that it can explain the large change in hours worked by married women without incorrectly predicting concomitant large movements in hours worked by single women.

Technological Progress at Home. Economists Jeremy Greenwood, Ananth Seshadri, and Mehmet Yorukoglu (2003) argue that technological progress is largely responsible for the rise in married women's market employment. They suggest that improvements in labor-saving equipment used in the home has freed up women's time for market work. Thus, like the gender wage gap theory discussed above, the technological progress explanation also uses as its basic framework a model of home production.

Home production, just like production in the market, requires capital. We all understand that firms combine workers with capital, in the form of both equipment and structures, along with materials and energy in order to produce output. Similarly, in home production, consumer durable goods serve as capital in the production of goods and services made at home. Stoves, dishwashers, washing machines, and refrigerators are all examples of capital used in home production.

Over the past 100 years, as electricity has reached more and more households in the United States, the technology of home production has undergone a dramatic change. Households have begun to invest in capital goods — consumer durables — that have increased labor productivity in the home. Investment in household appliances, as a percentage of gross domestic product, has nearly doubled over the past 100 years. As a series of new household appliances has increased productivity in the home, the amount of time a worker must devote to home production, to produce any desired level of goods and services, has fallen sharply.

A refrigerator is one example of a labor-saving consumer durable that has become common in households. The availability of refrigeration allows meals to be prepared far in advance of when they are consumed. Households with refrigerators are able

[10] The data are taken from Table 228 of the *Statistical Abstract of the United States: 2003.*

to prepare several meals at one time, thus reducing the labor required to cook meals at home. In the 1920s, almost no households in the United States had refrigerators. Twenty years later, about half of all households had such equipment. By 1960, almost all households had refrigeration. Richer families bought refrigerators first; as the real cost of the technology fell, additional families adopted it.

Electric washing machines and irons have also sharply reduced the amount of time people must work in the home. Consider the following example, taken from Greenwood, Seshadri, and Yorokoglu. In 1900, 98 percent of households used a scrub board to wash their clothes. The process of washing clothes required water to be transported to the stove, then heated by burning wood or coal, which itself had to be brought into the house. Next, the clothes were cleaned and rinsed. Afterward, the water used had to be removed and the clothes had to be hung on a clothesline to dry. Finally, the clean clothes had to be ironed, using flatirons heated on the stovetop. A study by the Rural Electrification Authority in 1945 found that washing and ironing 35 pounds of clothes required 8.5 hours when done by hand but only two hours and 16 minutes when done using a washing machine and an electric iron.

Such examples confirm that the introduction of household appliances, combined with electricity, central heating, and indoor plumbing, has dramatically increased the productivity of home work. Thus, Greenwood and his co-authors argued that these capital-specific productivity improvements have allowed the substitution of capital for labor at home. As a result, the amount of time required in the home has fallen, and this has freed up time for market work, especially women's time. (This argument assumes that, historically, the majority of home work was done by women.) Greenwood and co-authors found that more productive capital in the home leads to reduced time spent working in the home. This is in contrast to standard macroeconomic models in which more capital would tend to increase employment because capital accumulation raises the value of labor.

Changing Social Norms. A final explanation for the rise in hours worked by married women shifts the focus away from home production and toward changing social norms. In their paper, Raquel Fernandez, Alessandra Fogli, and Claudia Olivetti suggest that changes in men's attitudes about their wives' working have been important in bringing about the rise in the fraction of married women who work.

They discussed evidence indicating that in the early part of the century, men strongly disapproved of married women working, a disapproval that has lessened over time. In 1938, a Gallup poll asked, "Do you approve or disapprove of a married woman earning money in business or industry if she has a husband capable of supporting her?" Of the men surveyed, 81 percent did not approve. However, by 1972, the percentage of negative responses had fallen to 38; 10 years later this percentage had decreased further to 25. By 1998, only 17 percent of men surveyed gave negative responses. Clearly, married men at the end of the century were more willing to have their wives work than were men of 60 years before.

Fernandez and her co-authors suggest that the change in married men's preferences has come about as a result of being raised in households in which their own mothers worked. They showed that the sons of mothers who are skilled and who work are more likely to marry wives who are also skilled and who work. They concluded that a few mothers set an example that led their sons to become more accepting of women working. As these sons themselves married, their wives found it easier to work in the market. Thus, women, who could now work without hurting their marriage possibilities, undertook more education in order to earn higher wages. Each generation of households increased the acceptance of married women working, and over time, the fraction of working wives increased.

The introduction of household appliances, combined with electricity, central heating, and indoor plumbing, has dramatically increased the productivity of home work.

The authors provide evidence of the growing acceptability of marrying educated women, women who are far more likely to work in the market. Between 1890 and 1950, the likelihood that a college-educated woman would eventually marry, originally much lower than that for men, had risen to about the same as that for men. When studying women born in 1890, Fernandez and co-authors found that 31 percent of those with a college degree did not marry, while only 7.8 percent of those without college educations did not marry. In contrast, during this same period, there was no comparable difference in the marriage rate for men; they had a 10 percent chance of not marrying, whether or not they had a college degree.

If we look at women born 60 years later in 1950, who were attending college in 1970, the percentage of those with college degrees who did not marry fell to 7.9 percent. For comparison, 5.5 percent of those without college degrees did not marry. The probabilities for men were similar.

Fernandez, Fogli, and Olivetti developed a model in which the sons of educated mothers, who are more likely to work in the market, are less unhappy about marrying educated women. In other words, there is a direct transmission of preferences from mother to son, and sons of educated mothers find educated women less unsuitable as partners.

These co-authors' model predicts that, over time, more and more women will choose to obtain an education, marry, and work in the market. Their model matches several facts. First, both women's market work and educational level have risen. Second, men's attitudes toward working women have improved over time as more and more of them are themselves the sons of working mothers. Finally, the marriage rate of working women has risen relative to that of women who do not work in the market.

CONCLUSION

Over the postwar period there has been a large change in the composition of the labor force that is hidden when one examines the overall change in total hours worked per person. The labor force participation of married women has risen sharply, while the hours worked by others has fallen a little. Economists are using the home production model to better understand the determinants of these changes. Other explanations center on changing social norms.

The working behavior of the old and the very young has undergone significant changes over this period. Economic theory now must integrate the changes in working behavior across all ages and marital status.

REFERENCES

Blau, Francine D. "Trends in the Well-Being of American Women, 1970-85," *Journal of Economic Literature*, 36, March 1998, pp. 112-65.

Fernandez, Raquel, Alessandra Fogli, and Claudia Olivetti. "Marrying Your Mom: Preference Transmission and Women's Labor and Education Choices," NBER Working Paper 9234 (September 2002).

Greenwood, Jeremy, Ananth Seshadri and Mehmet Yorukoglu. "Engines of Liberation," unpublished manuscript, 2002.

Jones, Larry E., Rodolfo E. Manuelli, and Ellen R. McGrattan. "Why Are Married Women Working So Much?" Federal Reserve Bank of Minneapolis *Staff Report* 317 (2003) (available at: http://research.mpls.frb.fed.us/research/sr/sr317.html).

Katz , Lawrence F., and Claudia Goldin. "The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions," *Journal of Political Economy*, 100, 2002, pp. 730–70.

Katz, Lawrence F., and Kevin M. Murphy. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 57, 1992, pp. 35-78.

McGrattan, Ellen R., and Richard Rogerson. "Changes in Hours Worked Since 1950," Federal Reserve Bank of Minneapolis *Quarterly Review*, 22, 1, Winter 1998, pp. 2-19 (available at: http://research.mpls.frb.fed.us/research/qr/qr2211.html).

McGrattan, Ellen R., and Richard Rogerson. "Changes in Hours Worked, 1950-2000," Federal Reserve Bank of Minneapolis *Quarterly Review*, 28,1 July 2004 (available at http://minneapolisfed.org/research/qr/qr2812.html).

Rosenzweig, Mark, and Paul Schulz. "The Demand for and Supply of Births," *American Economic Review*, 75, 1985, pp. 992-1015.

U.S. Census Bureau, *Statistical Abstract of the United States: 2003* (123rd Edition) Washington, DC, 2003.

Sprawl: What's in a Name?

BY TIMOTHY SCHILLER

What lies behind concerns about the way metropolitan areas have been spreading out over the past several decades? This spreading out, commonly known as sprawl, is reflected in lower density and centralization in metropolitan areas. In this article, Tim Schiller looks at some recent trends toward lower population and employment density in metro areas and discusses some of the underlying forces propelling these trends.

In the elections of 2003, voters in 16 states passed measures to use public funds to preserve undeveloped land. One such measure was passed in Montgomery County, Pennsylvania, a suburban county neighboring the city of Philadelphia. These political proposals, often referred to as anti-sprawl initiatives, reflect a public desire to slow or halt the extension of urban land uses for housing, stores, factories, and office buildings.

The use of the word sprawl to describe the growth of metropolitan areas first became common in the later half of the 20th century.[1] By then, there was a public perception that growth in metropolitan areas was not only more extensive and less orderly than in the past but also economically inefficient.

Sprawl has been described in a number of ways: the lack of continuity in development, sometimes called leapfrogging (Marion Clawson); awkward or poorly planned spreading out (Charles Abrams); fragmented, incomplete, ad hoc, and uncentered development (Robert Geddes); consuming land for urbanization at a faster rate than the growth of population (William Fulton et al.); and low-density, automobile-dependent urban growth (Gregory Squires).[2] The common elements in all of these definitions are low density and less centralization.

Critics of sprawl argue that lower density and decentralization result in less efficient land use and other negative consequences.[3] Issues such as whether there are inefficiencies resulting from lower density, and how large they might be, are not settled (see *What Is the Efficient Level of Density?*). That debate is beyond the scope of this article. Here we describe the trends in metropolitan development with respect to density and centralization.

Is it really the case that metropolitan areas began to grow in a less orderly way about the middle of the 20th century? Has density taken a sudden turn down? Measures of density and concentration reveal both continuity and change in the way metropolitan areas have grown. The trend toward lower population density is a long-standing one, and it continues today. A more recent trend is an acceleration in the decline of employment density, as employers and the jobs they provide have spread out. Some of the forces leading to less dense development, such as improvements in transportation technology, are old and familiar; others, such as changes in business activity that require less concentration of firms, are new and challenging.

TRENDS IN DISTRIBUTION OF POPULATION IN METRO AREAS

Land development — converting land to urban uses (built-up

Tim Schiller is a senior economic analyst in the Research Department of the Philadelphia Fed.

[1] See, for example, the article by Marion Clawson, the article by David Mills, and the publication from the Real Estate Research Corporation.

[2] These definitions are representative, not a complete list, and are confined to those with measurable attributes.

[3] Readers interested in the arguments and counter-arguments on the effects of recent development will find many of them summarized in the article by Anthony Downs and in the 2000 article by Peter Gordon and Harry Richardson.

What Is the Efficient Level of Density?

When economists analyze metropolitan structure, they focus on the efficiency of the spatial distribution of employment and population. Is the density of an area greater or less than it would be if employers and residents considered the social costs of concentrating or spreading out when deciding where to set up their businesses or buy their homes? Researchers have cited several factors as evidence that residential density would be higher if homeowners took into account all the costs of spreading out.[a] These factors are failure to account for the amenity value of open space, the social cost of road congestion, and the infrastructure cost of new development. Counter-arguments point out that density can be inefficiently high, imposing social costs on the population and raising the cost of living.[b]

Open Space. It can be argued that people who live near open space benefit from it, but the owner of the space is not compensated for providing this benefit. Its value as open space is not explicitly recognized, so the loss of this value is not considered when the space is developed. This leads to more development than is socially desirable. One solution to this problem is to tax the development of open land. However, calculating the appropriate amount of such a tax is difficult.[c] Furthermore, such a tax imposes all of the cost on the property's owner, and property-rights issues also limit the use of this approach. As an alternative, outright purchase of open space for preservation by local, state, and federal governments and private groups has become increasingly popular. But any policy that preserves open space raises the price of other land and therefore the cost of any economic activity that takes place on it.

Road Congestion. The social cost of road congestion is the cost that each driver imposes on other drivers by increasing their drive time. Drivers do not take this cost into account when they plan their commute. One solution to this problem is to impose peak-time tolls that provide incentives for drivers to reduce their use of roads at busy times by changing their schedules, car pooling, or using public transportation. Kenneth Small argues that peak-time tolls can be computed fairly accurately. Although road congestion can cause inefficiency, studies indicate that commuting time has not been increasing rapidly, and survey data indicate that in a significant number of large metropolitan areas, commuting times are lower in the suburbs than in the city.[d]

Infrastructure Cost. The infrastructure cost of new development usually involves a one-time expense for such things as extending roads and utility systems to the new areas. Typically, these costs are shared among current residents and residents of the newly settled areas. Under this arrangement, the new residents are not paying the full costs of the new infrastructure they require, and demand for new development is higher than it would be if the new residents paid the full cost. One solution is to impose "impact fees" on new homeowners to cover the one-time cost of extending infrastructure. A number of municipalities have implemented these fees. Note that infrastructure cost is not always higher in low-density areas. There appears to be a density level above which infrastructure costs stop falling and begin to rise.[e]

Employment Concentration. Besides the costs of spreading out that residents do not take into account, there are benefits to greater concentration and density that business firms do not take into account. When there are spillover benefits from firms' locating close to one another but the firms are not compensated for providing these benefits, they are less likely to locate as close to each other as would be mutually beneficial. Zoning laws are one way of inducing firms to locate closer together; subsidies or tax abatements are other ways. However, a practical way to determine exactly the level of concentration that maximizes spillover benefits net of congestion costs in a given area has yet to be achieved; so policymakers should approach this issue cautiously.[f]

Clearly, there are factors that might lead to less density and concentration, at a given point in time and in a given area, than is most economically efficient. Public policies should address these factors, but policymakers should proceed carefully in order to avoid imposing costs in excess of benefits. Regardless of the efficiency of metropolitan structure at any given time and place, history suggests that the economically efficient level of density and concentration has been declining over time because of changes in transportation and communications technology and the rising demand for more residential space.

---

[a] See Jan Brueckner's 2001 article.

[b] See the 2000 article by Gordon and Richardson and the one by Pietro Nivola.

[c] See the article by Glenn Blomquist and John Whitehead.

[d] See the article by Gordon and Richardson and the publication by the U.S. Census Bureau.

[e] For more on demand for new infrastructure and its costs, see Brueckner's 1997 article. The article by Alan Altshuler and Jose Gomez-Ibanez talks about impact fees. Helen Ladd's article discusses the relationship between infrastructure cost and density.

[f] For more on concentration, density, and business firms, see the article by Antonio Ciccone and Robert Hall. See the articles by Esteban Rossi-Hansberg and Robert Lucas and Rossi-Hansberg for more about spillovers and compensation to firms. For more about optimal concentration and spillover benefits, see Glaeser and Kahn's 2003 article.

residential, commercial, industrial, and public uses) from nonurban ones (agriculture, forest, wetlands, or simply vacant land) — has taken place throughout history.

Cities Spread Out. Urban areas have been spreading out for more than two centuries, and economic activity has been dispersing geographically.[4] Large-scale development of suburban land became more common in the post-World War II years, and many people point to the creation of Levittown, New York, as the beginning of large-project development on a scale hitherto unknown. Indeed, Levittown was the largest single housing project undertaken to that time.[5] Nevertheless, metropolitan areas have been expanding and their density declining for many years.

The primary reason for this spreading out has been changes in transportation technology that enabled people to travel and goods to be shipped more quickly, less expensively, over longer distances, and with less regard to the physical features of the landscape.[6] When waterborne commerce was the most efficient means of transportation, cities developed close to oceans, lakes, and rivers. As railroads replaced rivers, cities developed at rail hubs. Within metropolitan areas, the railroad allowed people to live farther from their place of employment — to move to the suburbs and work in the central city. With the invention of the automobile, the road system became the network on which cities depend to bring people to work and to ship the products they use or produce. Besides extending the distance over which daily commuting was feasible — by bus or car — the road network allowed lower residential density because it was more extensive than the railroad network and did not require people to live near a limited number of railroad stations. The road network permitted metropolitan areas to increase in population and area with less need for concentration of population and employment, thus enabling a decline in density.[7]

The most recent U.S. census data indicate that this trend toward lower density and less centralization of population has continued as metropolitan areas have grown. As more land area has been converted to urban uses, the density of development on that land has been less than that in older urbanized areas. [8] From 1982 to 1997, the U.S. population increased 17 percent, while urbanized land increased 47 percent, or about 2.75 times as much.[9] Consequently, population per acre of urbanized land declined, hence, the term sprawl.

From 1982 to 1997, this decline was smaller in the West than in the rest of the nation (Table 1). That's because, in the West, geographic barriers, such as deserts and mountains, make it more difficult for development to spread out. Urbanized land increased only one and a half times faster than population in the West, but in other regions, it grew two to six times faster than population. For the metropolitan areas in the three states of the Third Federal Reserve District (Pennsylvania, New Jersey, and Delaware), population per acre of urbanized land decreased most in areas that had slow population growth or population losses and decreased least in those areas that had relatively rapid population growth. Thus, while sprawl issues tend to arise in growing areas, the decline in population per acre of urbanized land is not confined to areas with rapid population growth.

When waterborne commerce was the most efficient means of transportation, cities developed close to oceans, lakes, and rivers. As railroads replaced rivers, cities developed at rail hubs.

A Model of City Growth. Cities typically grow at their edges, and population density is typically lower at the edges than in the center of the city. This pattern of density has led economists to formulate a model of metropolitan spatial structure known as the monocentric city model. In this model, employment is concentrated at the center of the metropolitan area, and the population is spread out around that center as determined by the transportation system. Land is cheaper further from the center because transportation costs are higher, leading to lower demand for land that is further out. This model has explained the data from many

[4] See the article by Alex Anas et al.

[5] See the book by Peter Hall.

[6] It has also been argued that subsidies to transportation have promoted less dense development. See the 2003 article by Jan Brueckner.

[7] See the article by Robert Fishman and the one by Alex Marshall.

[8] The definition of urbanized land used in the studies reported in this article is based on the National Resources Inventory conducted by the U.S. Department of Agriculture. The inventory surveys all land in the country and divides it in into small-area units. Within each area unit the land use is similar. Those areas in which at least 30 percent of the land is covered by man-made features, such as buildings and roads, are classified as urban.

[9] See the article by William Fulton et al.

TABLE 1

Change in Population vs. Change in Urbanized Land in U.S. Regions and Third District States*

	Change in Urbanized Land 1982-1997 Percent	Change in Population 1982-1997 Percent	Change in Population per Acre of Urbanized Land 1982-1997 Percent
United States	47.1	17.0	-20.5
Census Region			
South	59.6	22.2	-23.4
Northeast	39.1	6.9	-23.1
Midwest	32.2	7.1	-19.0
West	48.9	32.2	-11.2
Metro Areas in Third District States**			
Johnstown, PA	53.0	-9.4	-40.8
Sharon, PA	52.5	-5.2	-37.9
Pittsburgh-Beaver Valley, PA	42.6	-8.0	-35.5
Erie, PA	49.9	-0.7	-33.8
Williamsport, PA	53.2	2.0	-33.5
York, PA	77.7	18.1	-33.5
Scranton-Wilkes-Barre, PA	55.0	4.1	-32.8
Altoona, PA	42.0	-4.5	-32.7
Harrisburg-Lebanon-Carlisle, PA	62.4	9.9	-32.4
Allentown-Bethlehem, PA	61.2	13.0	-29.9
Atlantic City, NJ	66.5	22.2	-26.6
State College, PA	55.1	15.2	-25.7
Reading, PA	50.4	15.2	-23.4
Philadelphia-Wilmington-Trenton, PA-NJ-DE	35.6	7.0	-21.1
Lancaster, PA	45.9	23.0	-15.7
New York-Northern New Jersey-Long Island, NY-NJ	20.5	6.1	-15.4

* Using National Resources Inventory urbanized area definition.
** Consolidated and primary metropolitan statistical areas.
Source: Fulton et al.

cities around the world for at least the past two centuries.[10] The model implies that population density declines as the distance from the metropolitan center increases because land is cheaper farther away from the center, and this entices people to consume relatively more land.

Lower density in newly urbanized areas compared with previously urbanized areas is an implication of the monocentric city model because these new areas are farther from the metropolitan center. This type of sprawl is nothing new. As a metropolitan area grows, two things happen. First, population increases, so that the population density of the area within its fixed boundary increases.[11] Second, as the metropolitan area grows, the amount of urbanized land within the area expands, and population density within the newly urbanized land area is lower than the density in the older urbanized land area. The maps (on page 33) show the Philadelphia metropolitan area in 1950 and 2000.[12] As population has risen in the past 50 years, overall density in the metropolitan area has increased, and the urbanized portion of the MSA (shaded areas) has expanded. The urbanized area furthest from the city center remains less dense than the center or the close-in suburbs.

Growth Is Uneven. As metro areas expand, their growth is often uneven. This unevenness — often called leapfrogging, and cited as evidence of sprawl — can occur in the course of development, but it usually does not persist. Newly developing areas tend to be less dense when they first come to public notice, than when they are fully built-up. The observed lack of even growth at a point in time might be simply a failure to account for eventual in-fill development.[13] Much of the land that appears to have been bypassed is eventually developed.

Population Becomes Less Centralized. The growth of metropolitan areas has been accompanied by a trend toward a more even spatial distribution of population. This means that population becomes less concentrated near the center of a metro area as the area expands; in other words, it becomes less centralized, and the decline in centralization is often cited as evidence of sprawl. Centralization is measured by the rate at which population density decreases as distance from an area's center increases — the density gradient. Lack of centralization, which is indicated by a density gradient with a low numerical value, is a measure of sprawl.

The density gradient can reveal the extent to which the spatial structure envisioned by the monocentric city model actually prevails in a given metropolitan area, and changes in the density gradient can reveal how centralization has changed in a given area over time. Estimates of density gradients of metropolitan areas around the world show that centralization has been declining for the past 200 years.[14] There were rapid declines in the decades near the end of the 19th century as railroads were developed. In the 20th century, there was a relatively large decline in the 1920s, a period in which automobile ownership grew substantially, and from the mid-1940s to the mid-1950s, during the post-World War II housing expansion.[15] In more recent years, the decline has continued at a slower, fairly steady rate.[16] Among the 10 largest areas, Philadelphia ranks second, below New York, in centralization, and it is about in line with the large metro areas near it (Table 2).

The history of density gradients indicates that metropolitan area populations have been spreading out and becoming less centralized for a long time. The density gradient will decline if the suburban area's boundary remains fixed and its population grows more rapidly than the population of the central city. This reduces the difference in density between the center and the suburbs. In the more usual case, the suburban area's boundary expands, and the older suburban area becomes more densely populated. (This is illustrated in the Philadelphia area map.) As the suburban area expands, the most recently developed areas are less dense than the previously urbanized area. However, the drop in density between the farthest-out areas and the closer-in areas is not as great

[10] See the articles by William Alonso; Richard Muth; and the 1967 article and 1972 book by Edwin Mills.

[11] The definition of metropolitan area referred to here is the one used by the U.S. Census Bureau in delineating metropolitan statistical areas (MSAs). MSAs are defined along county boundaries, and they do not expand unless the Census Bureau redefines them.

[12] The map uses the Philadelphia MSA definition in effect in 2000. Under that definition the MSA includes the Pennsylvania counties Philadelphia, Bucks, Chester, Delaware, Montgomery, and Philadelphia, and the New Jersey counties Burlington, Camden, Gloucester, and Salem. A new definition was issued in 2003.

[13] See the article by Paul Longley and Victor Mesev. The research reported in the article by Burchfield et al. indicates that most of the development that occurred between 1976 and 1992 took place in areas that were already urbanized in 1976, thus increasing density in areas after they first met the criterion for being considered urbanized.

[14] See the book by Colin Clark.

[15] See Mills's 1972 article.

[16] See the articles by Peter Mieszkowski and Edwin Mills; and Stacy Jordan et al.

TABLE 2

Density Gradients of 10 Largest U.S. Metro Areas (1990)

Ten Largest Metro Areas	Density Gradient*
New York	0.136
Los Angeles	0.067
Chicago	0.095
Philadelphia	**0.117**
Dallas	0.108
Miami	0.109
Washington, DC	0.099
Houston	0.097
Atlanta	0.099
Detroit	0.078
Other Large Metro Areas Near Philadelphia	
Baltimore	0.141
Newark	0.112
Pittsburgh	0.091

*Percent decline in population per square mile for each mile of distance from metropolitan area center.

Source: Jordan et al.

as it was prior to the new development, so the density gradient is lower. As the historical data indicate, this case has been the predominant trend for a long time.

TRENDS IN DISTRIBUTION OF EMPLOYMENT IN METRO AREAS

The monocentric city model is based on the location of business activity at the center of the metro area surrounded by a decreasingly dense residential population. Although the decline in population density from the center outward is an implication of the model, the model does not necessarily imply a decline in the population density gradient over time. One reason for the declining population density gradient is a decrease in the centralization of business activity. The diffusion of employment throughout an area can lead to a loss of orientation toward the center that is represented in the monocentric city model.

Business Activity Spreads Out. Recent data indicate that business establishments, and consequently employment, have been spreading out. Technology has made the distance between business establishments a less important factor in where to locate. Furthermore, congestion costs have risen for businesses operating in densely developed areas, encouraging them to relocate to less dense areas. As part of this spreading out process, employment has grown more rapidly in less dense metropolitan areas — and even in some rural areas — than in denser metropolitan areas.[17] This trend has been especially important for manufacturing and has therefore had more of an effect on reducing employment density in older, more manufacturing-oriented metropolitan areas. Retail and service employment has similarly spread out, but to a lesser degree.[18]

Along with the shift in the share of employment toward less dense metropolitan areas, there has been an increase in the share of employment in farther-out locations within metropolitan areas, and this trend appears to have accelerated in the later decades of the past century.[19] This spreading out has reduced centralization of employment within metropolitan areas. Changes in employment in the city of Philadelphia versus the Philadelphia metropolitan area illustrate this. From 1970 to 2000, the city's share of the area's total employment fell by almost half: from 52 percent to 29 percent.

A major factor in this shift has been a recent trend toward more dispersed service employment and office development. Most new office space built in the last 20 years has been outside downtown central business districts.[20] The spatial distribution of office development is important for two reasons. First, it is associated with service employment, the largest and fastest growing sector of employment. Second, in recent years, it has displayed a sharp difference from the monocentric pattern that characterized metropolitan areas for most of the past 200 years. The recent pattern of office development might be an indication of the future shape of metropolitan areas.[21]

The Rise and Decline of Sub-Centers. The early history of metropolitan expansion, from roughly 1850 to 1950, was characterized by the growth of the downtown business core and a spreading out of residential areas as changes in transportation made commuting feasible over longer distances. In the later half of the 20th

[17] See the article by Gerald Carlino and Satyajit Chatterjee and the 1998 article by Gerald Carlino.

[18] See Carlino's 1983 article, the article by Theodore Crone, and the article by Lawrence Thurston and Anthony Yezer.

[19] In their 2001 article, Edward Glaeser and Matthew Kahn found that the share of employment in the major county of metropolitan areas declined more rapidly from 1970 to 1993 than from 1950 to 1970.

[20] This is a factor in the excess of land development over population growth in some metropolitan areas. See the articles by Robert Lang and Jennifer LeFurgy; and Marcy Burchfield et al.

[21] See Mills's 1988 article.

century, however, alternative centers of employment began to form within a single metropolitan area. This polycentric development became characteristic of growth in all rapidly growing metropolitan areas, leading to the development of city-like areas (so-called edge cities) of office and retail buildings that developed around major freeway intersections in formerly suburban areas.[22] (See *Where's the Edge?*)

The development of sub-centers within the farther reaches of metropolitan areas appears to be a consequence of the increased suburbanization of the residential population. Their location represents a balance between the benefits of a large population from which to draw workers and the need to avoid the congestion cost in the denser, more central portions of a metropolitan area.[23]

While at first glance sub-centers appear to be smaller versions of the traditional monocentric city, there are important differences, and these differences suggest that the agglomeration economies that have historically explained the growth of cities are weakening.[24] Sub-centers or edge cities are primarily employment centers, with more jobs than residents; the jobs are primarily office-based service jobs (especially business services). Sub-centers generally do not have the mix of industries, such as manufacturing, trade, health services, and personal services, historically present in traditional monocentric cities. Consequently, there isn't much, if any, inter-industry agglomeration.

However, the grouping of office buildings in sub-centers does suggest that there might be agglomeration economies for the type of economic activity that takes place in office buildings. In this respect, sub-centers of service industry employment are smaller scale versions of industry clusters: the contiguous location of firms with frequent, mutually beneficial interaction. Studies of employment by industry show that service industries have tended to retain more of a centralized pattern than manufacturing. The location of sub-centers of service employment in suburban areas reflects the joint influence of workers' preferences for lower residential density

Where's the Edge?

In his book, Joel Garreau listed the following edge cities in the three states of the Third District:

- In the New Jersey portion of the New York area: Fort Lee, Paramus-Montvale, Mahwah, the Meadowlands, Whippany-Parsippany-Troy Hills, Bridgewater, Woodbridge, Metropark, and Princeton.
- In the New Jersey portion of the Philadelphia area: Cherry Hill.
- In the Pennsylvania portion of the Philadelphia area: King of Prussia and Willow Grove.
- In the Pittsburgh area: Penn Lincoln Parkway-Airport area.

Of course, the existence and number of edge cities — more commonly called sub-centers — outside the downtown area depend on the definition and size criteria used to identify them. Although there are no definite objective criteria for identifying sub-centers, a variety of measures with varying degrees of complexity have been used to enumerate them in major metropolitan areas. Most definitions of sub-centers no longer include retail development (although such development was included in Garreau's definition) because centers with only office buildings have been increasingly observed.*

Garreau allowed the possibility that some of the edge cities he defined were so dispersed as to lack sufficient centralization on their own to qualify as identifiable places. This lack of centralization has been noted in office development in the years after the concept of the edge city was introduced, and some of Garreau's incipient edge cities are now considered to be areas of dispersed office development. Cherry Hill, NJ, in the Philadelphia metropolitan area, is an instance of this.

* See the references in Daniel McMillen's article.

[22] For more about alternative centers of employment, see Anas et al. For more about edge cities, see the book by Joel Garreau.

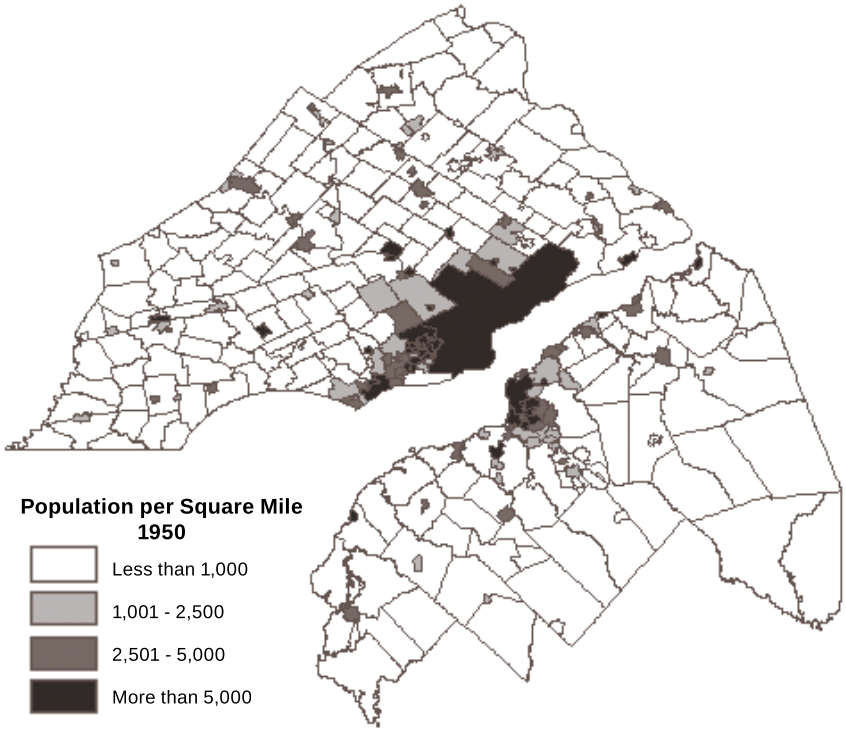
[23] See the articles by Daniel McMillen and Stefani Smith; and Vernon Henderson and Arindam Mitra.

[24] Agglomeration economies are the cost savings of economic activity that result from different activities locating close to one another. For example, a supplier locating close to a major customer may benefit from lower communication and transportation costs, reduction in delivery time and required inventories, and closer collaboration on product design. In the case of consumption activity, agglomeration economies result from retailers locating close to one another, allowing customers to do comparison shopping in less time and at a lower cost and to purchase multiple items in a single shopping trip.

[25] For more on sub-centers and their potential benefits, see the article by Wayne Archer and Marc Smith; the one by Michael Porter; and the 2001 article by Edward Glaeser and Matthew Kahn.

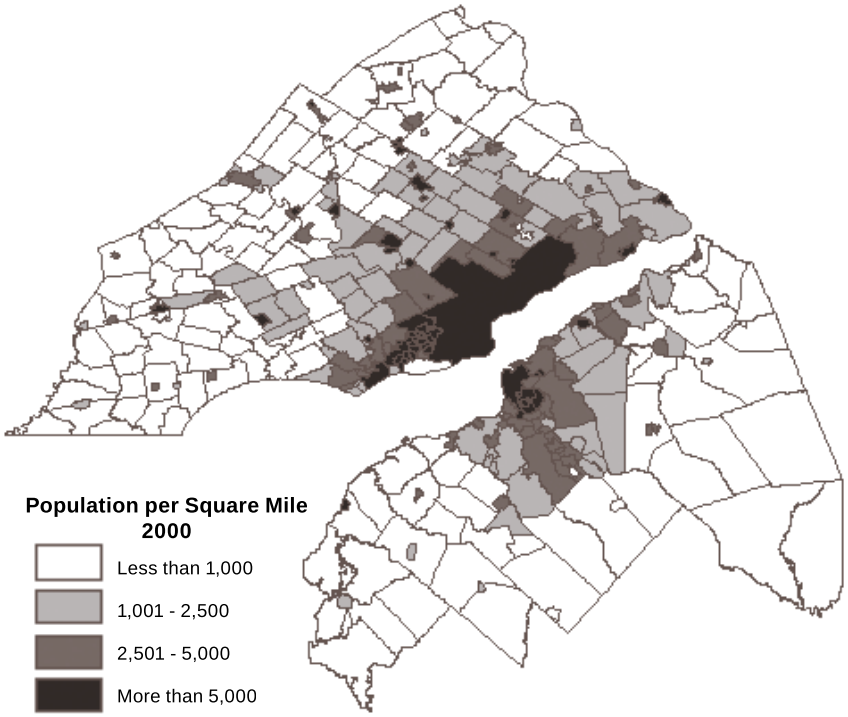
PHILADELPHIA METROPOLITAN AREA

Population Density 1950



PHILADELPHIA METROPOLITAN AREA

Population Density 2000



and service industry needs for close interaction among firms.[25]

Soon after sub-centers or edge cities emerged as the locations of new office development, even more far-flung construction of offices began. In his 2003 study, Robert Lang examined 13 large metropolitan areas and found that more office space was added in smaller, less concentrated office buildings, which he called "edgeless cities," during the 1980s and 1990s than in either central business districts or edge cities. As a result, by 1999, of the total office space in the metropolitan areas that Lang studied, there was very nearly as much space in the edgeless cities as in the primary downtown or central business district (Table 3). In 11 of the 13 areas studied, including Philadelphia, there was more office space in the edgeless cities than in the primary downtown. Edgeless cities accounted for a greater share of total office space in the Philadelphia metropolitan area than in any other area studied with the sole exception of Miami. The data compiled by Lang indicate that 70 percent of the office space built in the Philadelphia area during the 1990s was in edgeless cities, well above the average of 40 percent of 1990s' construction for all 13 cities. Edgeless cities are, by their nature, not identified with specific locations, but encompass areas such as southern New Jersey.[26] In total, edgeless cities in the Philadelphia area account for approximately 54 percent of the area's office space.

Typically, the office buildings in edgeless cities are low rise and include parking lots, two factors that also

[26] The edge cities in the Philadelphia area — King of Prussia and Malvern-Paoli-Wayne in Pennsylvania — together account for approximately 9 percent of the area's office space. Lang classifies all of the office space in the New Jersey portion of the Philadelphia area as edgeless. In contrast to Garreau, who expected Cherry Hill to develop into an edge city, Lang argues that it has not reached the size to qualify for that designation.

contribute to the greater use of space per person in more recent development. Researchers who have examined recent development of dispersed office space believe such development will probably not expand to reach the size of edge cities, nor will major retail space be developed in close proximity to it, because congestion costs set in at a relatively low level of density in low-rise office complexes, where nearly all workers arrive by car.[27]

THE FUTURE STRUCTURE OF METROPOLITAN AREAS: IS THE PAST PROLOGUE?

What about the future? Do the decentralizing tendencies described in this article portend a landscape with lower density and little or no centralized features? Some of the forces that have influenced the spreading out of residential populations and employment and the decentralization of economic activity will continue in the future, particularly declining transportation and communication costs, and increased road mileage.[28]

Forces for Decentralization. Declining transportation and telecommunication costs effectively bring more remote land into the market for urbanization, thus widening the feasible area for the location of jobs and housing. Urban analysts rank the interstate highway system as the main influence on the changes in metropolitan structure since the system was authorized in 1956. Further development of the road system is likely to extend this influence. In particular, the building of beltways around central cities gave rise to the edge cities and less centralized employment and residential development. Further construction of beltways and connectors in the system will extend its decentralizing effect. Increasing telecommunication capabilities (for example, mobile phones, camera phones, and the interconnectivity of voice and data communications equipment) and falling telecommunication costs will likely continue to reduce the need for centralization of work.

Forces for Centralization. Some other influences are working toward increasing, or at least stabilizing, density and centralization: increasing service employment and the rising importance of the intellectual content of work; the aging of the population; and the increasing emphasis on amenities of place. These trends favor continuing centralization and concentration of the residential population and employment in various ways.

Service company activity and the increasing intellectual content of work in such areas as research and development, patenting, and computer applications tend to require more face-to-face contact. This favors the concentrated locations of establishments where this type of work is done.[29] Research indicates that service and high-tech industries currently have a greater tendency toward centralization than other industries. The need for face-to-face contact might account for their relatively greater geographical concentration now, but it is not out of the question that technological advances could reduce this need.[30]

An older population is less mobile and requires more personal services, and serving this population requires frequent person-to-person contact.[31] The aging of the baby boomer generation has already brought about a major change in housing: the development of senior citizen housing and assisted living communities. These types of residential developments typically have greater density than the usual suburban communities, and therefore, they represent a countertrend to declining density. Although these housing arrangements have already begun to influence the newer parts of some growing metropolitan areas, the extent of their effect on development remains to be seen.

Natural amenities, such as ocean views and warm, dry climates, can be provided only by a limited number of places. The growth of metropolitan areas in coastal regions and in warmer, drier parts of the country is a

> Edgeless cities accounted for a greater share of total office space in the Philadelphia metropolitan area than in any other area studied with the sole exception of Miami.

[27] See the 1994 article by Lang and the article by Gary Pivo.

[28] See the articles by Robert Fishman; Jess Gaspar and Edward Glaeser; and Glaeser and Janet Kohlhase.

[29] See Gaspar and Glaeser; and Glaeser and Albert Saiz.

[30] The growth of telecommuting, perhaps the ultimate separation of workers from each other, would seem to be counter to any centralizing influence. But telecommuting seems to be most prevalent in the very industries that are more centralized and does not appear to be affecting residential location patterns yet. However, changes in managerial methods have the potential to reduce the need for face-to-face communication among workers and supervisors. See the articles by Ingrid Ellen and Katherine Hempstead; and Edward Potter.

[31] See Fishman.

TABLE 3

A: Distribution of Office Space in Major Metro Areas (1999)

Metro Area	Primary Downtown	Percent of Office Space Secondary Downtown*	Edge City	Edgeless Cities**
Miami	13.1	4.5	16.6	65.8
Philadelphia	34.2	3.2	8.9	53.6
San Francisco	33.9	8.8	13.9	43.4
Atlanta	23.6	9.9	25.3	41.2
Boston	37.4	4.6	18.8	39.2
Detroit	21.3	N/A	39.5	39.2
Houston	23.0	N/A	37.9	39.1
Los Angeles	29.8	7.8	25.4	37.0
Denver	30.4	4.2	29.4	35.9
Dallas	20.5	4.5	40.3	34.6
Washington	28.6	12.5	27.1	31.8
New York	56.7	7.2	6.2	29.9
Chicago	53.9	N/A	19.5	26.6
Total	37.7	6	19.8	36.5

B: Philadelphia Office Areas

	Percent of Office Space
Downtown	**37.5**
Philadelphia (primary downtown)	34.2
Wilmington (secondary downtown)	3.3
Edge City	**8.9**
King of Prussia	3.9
Malvern-Paoli-Wayne	5.0
Edgeless Cities**	**53.6**

* Office clusters in same city as primary downtown or in smaller cities within same metropolitan area.
** Unconcentrated office development within metropolitan area.
Source: Lang.

feature of post-World War II development. The number of locations suitable for urbanization in these parts of the country is fixed. Consequently, as they become more populated, they will become denser, tending to some extent to offset the general decline in density among metropolitan areas. They will also be subject to more in-fill development, which might reverse or retard the flattening out of their density gradients. Indeed, between 1980 and 1990, density gradients steepened in several drier and warmer cities, such as Oklahoma City, Corpus Christi, Fresno, and San Francisco, but in only two cities without such a climate, Columbus and Madison.[32]

Cultural amenities, such as high-quality museums, live theater, and orchestras, can be supported only in relatively densely populated areas. Among cities nationwide, there appears to be new interest in living closer to city centers where cultural amenities are located. Thus, both physical and cultural place amenities are likely to promote centralization and concentration.[33] If this trend strengthens, it could limit or reverse the decline in centralization of metropolitan areas, although its effect might be operative only in very close-in areas while decentralizing influences retain their force farther away from the center.

SUMMARY

Sprawl, when used to describe the spreading out of the residential population around central cities, is not a new phenomenon. It is the same suburbanization process that has been going on for centuries. Applying the term to the decentralization of employment, though, does seem to describe a more recent phase of metropolitan growth, with a significantly lower centralizing tendency or none at all compared with past development.

Will this trend continue? Although there are both centralizing and decentralizing forces affecting the location of jobs and housing, the influence of the falling costs of transportation and communication currently appears to be dominant, providing impetus to the decentralization trend. As long as that remains the case, it seems likely that we will not see a return to the centralization of population that was prevalent in the past. ER

[32] See the article by Stacy Jordan et al.

[33] See Glaeser's 1999 article and the articles by Glaeser and Jesse Shapiro; and Glaeser, Jed Kolko, and Albert Saiz.

REFERENCES

Defining Sprawl

Abrams, Charles. *The Language of Cities*. New York: Viking, 1971.

Clawson, Marion. "Urban Sprawl and Speculation in Suburban Land," *Land Economics*, 1962, pp. 99-111.

Downs, Anthony. "The Big Picture," *Brookings Review*, 16, 4, 1998, pp. 8-11.

Fulton, William, Rolf Pendall, Mai Nguyen, and Alicia Harrison. "Who Sprawls Most? How Growth Patterns Differ Across the U.S.," Washington, DC: The Brookings Institution Center on Urban and Metropolitan Policy, Survey Series, July 2001.

Geddes, Robert. "Metropolis Unbound: The Sprawling American City and the Search for Alternatives," *American Prospect*, 8(35):40, 1997. http://www.prospect.org/web/printfriendly-view.ww?id-4763 (April 19, 2003).

Mills, David E. "Growth, Speculation and Sprawl in a Monocentric City," *Journal of Urban Economics*, 10, 1981.

Real Estate Research Corporation. *The Costs of Sprawl: Literature Review and Bibliography*. Washington, DC: U.S. Government Printing Office, 1974.

Squires, Gregory D. "Urban Sprawl and the Uneven Development of Metropolitan America," in Gregory D. Squires, ed., *Urban Sprawl: Causes, Consequences, and Policy Responses*. Washington, DC: The Urban Institute Press, 2002.

Trends in the Distribution of Metropolitan Population

Anas, Alex, Richard Arnott, and Kenneth A. Small. "Urban Spatial Structure," *Journal of Economic Literature*, 36, September 1998, pp. 1426-64.

REFERENCES

Berube, Alan. "Gaining but Losing Ground: Population Change in Large Cities and Their Suburbs," in Bruce Katz and Robert E. Lang, eds., *Redefining Urban and Suburban America: Evidence from Census 2000,* Washington, DC: Brookings Institution Press, 2003.

Fishman, Robert. "The American Metropolis at Century's End: Past and Future Influences," *Housing Policy Debate*, Vol. 11, Issue 1, 2000, pp. 199-213.

Hall, Peter. *Cities in Civilization*. New York: Pantheon Books, 1998

Marshall, Alex. *How Cities Work: Suburbs, Sprawl, and the Roads Not Taken*. Austin: University of Texas, 2000.

The Monocentric City Model

Alonso, William. *Location and Land Use*. Cambridge, MA: Harvard University Press, 1964.

Mills, Edwin S. "An Aggregative Model of Resource Allocation in a Metropolitan Area," *American Economic Review*, Vol. 57, 1967, pp. 197-210.

Mills, Edwin S. *Studies in the Structure of the Urban Economy.* Baltimore, Johns Hopkins University Press, 1972.

Muth, Richard F. *Cities and Housing.* Chicago: The University of Chicago Press, 1969.

Metropolitan Growth and Centralization

Burchfield, Marcy, Henry G. Overman, Diego Puga, and Matthew A. Turner. "Sprawl: A Portrait from Space," October 2003, http://emlab.berkeley.edu/users/webfac/quigley/e231_f03/turner.pdf

Clark, Colin. *Population Growth and Land Use*. London: Macmillan, 1967.

Jordan, Stacy, John P. Ross, and Kurt G. Usowski. "U.S. Suburbanization in the 1980s," *Regional Science and Urban Economics*, 28, 1998, pp. 611-27.

Longley, Paul A. and Victor Mesev. "Measurement of Density Gradients and Space-filling in Urban Systems," *Papers in Regional Science*, 81, 2002, pp. 1-28.

Mieszkowski, Peter, and Edwin S. Mills, "The Causes of Metropolitan Suburbanization," *Journal of Economic Perspectives*, 7, 3, Summer 1993, pp. 135-47.

Trends in the Distribution of Metropolitan Employment

Carlino, Gerald. "New Employment Growth Trends: The U.S. and the Third District," Federal Reserve Bank of Philadelphia *Business Review*, September/October 1983, pp. 5-14.

Carlino, Gerald. "Trends in Metropolitan Employment Growth," Federal Reserve Bank of Philadelphia *Business Review*, July/August 1998, pp. 13-22.

Carlino, Gerald A., and Satyajit Chatterjee. "Employment Deconcentration: A New Perspective on America's Postwar Urban Evolution," Federal Reserve Bank of Philadelphia, Working Paper 01-4, 2001.

Crone, Theodore M. "Where Have All the Factory Jobs Gone—and Why?" Federal Reserve Bank of Philadelphia *Business Review*, May/June 1997, pp. 1-16.

Lang, Robert E., and Jennifer LeFurgy. "Edgeless Cities: Examining the Noncentered Metropolis," *Housing Policy Debate*, 14, 3, 2003, pp. 427-60.

Mills, Edwin S. "Service Sector Suburbanization," in George Sternlieb and James W. Hughes, eds., *America's New Market Geography: National, Regional, Metropolis.* New Brunswick, NJ: Rutgers University Center for Urban Policy Research, 1988.

Thurston, Lawrence, and Anthony M. J. Yezer. "Causality in the Suburbanization of Population and Employment," *Journal of Urban Economics*, 35, 1994, pp. 105-18.

Sub-Centers and Edgeless Cities

Archer, Wayne R., and Marc T. Smith. "Explaining Location Patterns of Suburban Offices," *Real Estate Economics*, 31, 2, 2003, pp. 139-64.

Garreau, Joel. *Edge City: Life on the New Frontier.* New York: Anchor Books, 1991.

Glaeser, Edward L. and Matthew E. Kahn. *Decentralized Employment and the Transformation of the American City.* Cambridge, MA: National Bureau of Economic Research, Working Paper 8117, 2001.

Henderson, Vernon, and Arindam Mitra. "The New Urban Landscape: Developers and Edge Cities," *Regional Science and Urban Economics*, 26, 1996, pp. 613-43.

Lang, Robert E. *Beyond the Office Park: A Typology of New Jersey's Business Centers*. New Brunswick, NJ: Rutgers University Center for Urban Policy Research, 1994.

Lang, Robert E. *Edgeless Cities: Exploring the Elusive Metropolis.* Washington, DC: The Brookings Institution Press, 2003.

McMillen, Daniel P. "Identifying Subcentres Using Contiguity Matrices," *Urban Studies*, 40, 1, 2003, pp. 57-69.

McMillen, Daniel P., and Stefani C. Smith. "The Number of Subcenters in Large Urban Areas," *Journal of Urban Economics*, 53, 2003, pp. 321-38.

REFERENCES

Pivo, Gary. "The Net of Mixed Beads: Suburban Office Development in Six Regions," *Journal of the American Planning Association*, 56, 4, 1990, pp. 457-69.

Porter, Michael E. "Location, Clusters, and the 'New' Microeconomics of Competition," *Business Economics*, 33, 1, 1998, pp. 7-17.

The Future of Metropolitan Development

Ellen, Ingrid Gould, and Katherine Hempstead. "Telecommuting and the Demand for Urban Living: A Preliminary Look at White-collar Workers," *Urban Studies*, 39, 4, 2002, pp. 749-66.

Fishman, Robert. "The American Metropolis at Century's End: Past and Future Influences," *Housing Policy Debate*, 11, 1, 2000, pp. 199-213.

Gaspar, Jess, and Edward L. Glaeser. "Information Technology and the Future of Cities," *Journal of Urban Economics*, 43, 1998, pp. 136-56.

Glaeser, Edward L. "The Future of Urban Research: Non-market Interactions," September 1999, http://post.economics.harvard.edu/faculty/glaeser/papers1_00_paper.pdf

Glaeser, Edward L., and Janet E. Kohlhase. "Cities, Regions, and the Decline of Transport Costs," Cambridge, MA: Harvard University Institute of Economic Research, Discussion Paper No. 2014, July 2003.

Glaeser, Edward L. and Albert Saiz. "The Rise of the Skilled City," Federal Reserve Bank of Philadelphia Working Paper 04-2, December 2003.

Glaeser, Edward L. and Jesse M. Shapiro. "Urban Growth in the 1990s: Is City Living Back?" *Journal of Regional Science*, 43, 1, 2003. pp. 139-65.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. "Consumer City," *Journal of Economic Geography*, 1, 2001, pp. 27-50.

Potter, Edward E. "Telecommuting: The Future of Work, Corporate Culture, and American Society," *Journal of Labor Research*, 24, 1, 2003, pp. 73-84.

Density and Efficiency

Altshuler, Alan A., and Jose A. Gomez-Ibanez. *Regulation for Revenue: The Political Economy of Land Use Exactions*. Washington, DC: Brookings Institution, 1993.

Blomquist, Glenn C., and John C. Whitehead. "Existence Value, Contingent Valuation, and Natural Resource Damage Assessment," *Growth and Change*, 26, 1995, pp. 573-89.

Brueckner, Jan K. "Infrastructure Financing and Urban Development: The Economics of Impact Fees," *Journal of Public Economics*, Vol. 66, 1997, pp. 383-407.

Brueckner, Jan K. "Urban Sprawl: Lessons from Urban Economics," *Brookings-Wharton Papers on Urban Affairs*, Washington, DC: Brookings Institution Press, 2001.

Brueckner, Jan K. *Transport Subsidies, System Choice, and Urban Sprawl*. CESifo Working Paper 1090, November 2003.

Ciccone, Antonio, and Robert E. Hall. "Productivity and the Density of Economic Activity," *American Economic Review*, 86, 1, 1996, pp. 54-70.

Glaeser, Edward L. and Matthew E. Kahn. *Sprawl and Urban Growth*. Cambridge, MA: Harvard University Institute of Economic Research, Discussion Paper 2004, May 2003.

Gordon, Peter, and Harry W. Richardson. "Congestion Trends in Metropolitan Areas," in *Curbing Gridlock: Peak-Period Fees to Relieve Traffic Congestion*. Washington, DC: National Academy Press, 1994, pp. 1-31.

Gordon, Peter, and Harry W. Richardson. "Critiquing Sprawl's Critics," Washington, DC: Cato Institute, Policy Analysis 365, 2000.

Ladd, Helen. "Population Growth, Density and the Costs of Providing Services," *Urban Studies*, 29, 2, 1992, pp. 273-95.

Lucas, Robert E., Jr., and Esteban Rossi-Hansberg. "On the Internal Structure of Cities," *Econometrica*, 70, 4, 2002, pp. 1445-76.

Nivola, Pietro S. "Fat City," *Brookings Review*, 16, 4, 1998, pp. 17-20.

Rossi-Hansberg, Esteban. "Optimal Urban Land Use and Zoning," *Review of Economic Dynamics*, 7, 2004, pp. 69-106.

Small, Kenneth A. *Urban Transportation Economics*. Chur, Switzerland:Harwood Academic Publishers, 1992.

U. S. Census Bureau. "2000 Census of Population and Housing, Profiles of General Demographic Characteristics." www.census.gov/Press-Release/www/2002/demoprofiles.html.