Federal Reserve Bank of St. Louis

REVIEW

JANUARY/FEBRUARY 2005 VOLUME 87, NUMBER 1

REVIEW

Review is published six times per year by the Research Division of the Federal Reserve Bank of St. Louis and may be accessed through our web site: **research.stlouisfed.org/publications/review**. All non-proprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in *Review* also are available to our readers on this web site.

Single-copy subscriptions are available free of charge. Send requests to: Federal Reserve Bank of St. Louis, Public Affairs Department, P.O. Box 442, St. Louis, MO 63166-0442, or call (314) 444-8808 or 8809.

General data can be obtained through FRED (Federal Reserve Economic Data), a database providing U.S. economic and financial data and regional data for the Eighth Federal Reserve District. You may access FRED through our web site: research.stlouisfed.org/fred.

ISSN 0014-9187

FOMC Transparency

William Poole

This article was originally presented as a speech at the Ozark Chapter of the Society of Financial Service Professionals, Springfield, Missouri, October 6, 2004.

Federal Reserve Bank of St. Louis *Review*, January/February 2005, *87*(1), pp. 1-9.

Transparency is at the forefront of many monetary policy debates today. The Federal Open Market Committee (FOMC) has had several formal discussions of communications issues in recent years, and the subject comes up fairly frequently in FOMC meetings and speeches by FOMC members.[1]

It is hardly surprising that central bankers are more talkative than they were just a decade or so ago—and more concerned about how to improve transparency and communication with the market. Perhaps only one issue is settled: Transparency is important but is hard to accomplish because miscommunication is so easy. Clearly, more talk does not necessarily mean greater transparency.

Discussions of monetary policy communication frequently center on three dimensions of transparency: (i) transparency about the objectives of monetary policy, (ii) transparency about current monetary policy actions, and (iii) transparency about expected future monetary policy actions. I'll organize my remarks around these dimensions.

Before proceeding, however, I want to emphasize that the views I express here are mine and do not necessarily reflect official positions of the Federal Reserve System. I thank my colleagues at the Federal Reserve Bank of St. Louis, especially Bob Rasche, senior vice president and director of the Research Division, and Dan Thornton, vice president in the Research Division, for their extensive assistance, but I retain full responsibility for errors.

BACKGROUND

The first formal move in the direction of transparency was initiated by the Reserve Bank of New Zealand, which in 1990 negotiated an agreement with the government of that country, making the Bank responsible for maintaining inflation within a specified range. Hence, the Reserve Bank of New Zealand was the first central bank to be transparent about its policy objective.

The Reserve Bank of New Zealand has been a leader on the other two dimensions of transparency as well. For some time now it has announced its setting of its policy instrument—the official cash rate. The Bank also publishes, on a semiannual basis, its forecasts over a several-year horizon for a number of economic variables, including the 90-day bill rate. Given that the Reserve Bank of New Zealand states that the 90-day rate is closely linked to its official cash rate, these forecasts come

[1] I myself gave a speech on the subject in August 2003: "Fed Transparency: How, Not Whether" The speech was later published in the Federal Reserve Bank of St. Louis *Review*, Vol. 85, No. 6 (November/December 2003).

William Poole is the president of the Federal Reserve Bank of St. Louis. The author thanks colleagues at the Federal Reserve Bank of St. Louis for extensive assistance. Robert H. Rasche, senior vice president and director of the Research Division, and Daniel L. Thornton, vice president in the Research Division, were especially helpful. The views expressed are the author's and do not necessarily reflect official positions of the Federal Reserve System.

very close to projecting a conditional course of monetary policy actions. Thus the Reserve Bank is transparent on all three dimensions of transparency I outlined earlier.

A number of other central banks have followed the lead of the Reserve Bank of New Zealand by adopting and announcing specific numeric inflation objectives. These central banks have become known as "inflation targeters." Currently included in this group are the Bank of Canada, the Reserve Bank of Australia, the Bank of England, and the central banks of Albania, Brazil, Chile, Colombia, the Czech Republic, Georgia, Hungary, Iceland, Israel, Mexico, Norway, Peru, the Philippines, Poland, Serbia, Sierra Leone, South Africa, South Korea, Sri Lanka, Sweden, Switzerland, Tanzania, Thailand, and Turkey. All of these institutions are transparent with respect to their monetary policy objectives. Moreover, all of these central banks announce changes in the settings of their policy instrument (typically a short-term interest rate). Practice differs from institution to institution on the release of forward-looking information such as forecasts for future developments in the economy.

The practice of the European Central Bank (ECB) differs somewhat from that of the "inflation targeters." The ECB offers a degree of transparency with respect to its monetary policy objective—the ECB has an announced goal of keeping the inflation rate close to but below 2 percent per annum "in the medium run." However, the ECB has never announced an explicit definition of the "medium run." The ECB announces changes in its policy rates, but does not disclose forecasts for the European Union economies or minutes of policy discussions.

The Federal Reserve's practice of transparency has evolved over time. I will discuss this evolutionary process with respect to the three dimensions of transparency. I note at the outset that I endorse unconditionally only the first two dimensions of transparency. For reasons I will make clear later, forecasting future policy actions is a complicated issue; without discussing the complexities and the nature of possible policy forecasts, it would be misleading to offer a simple "I support" or "I oppose."

THE EVOLUTION OF TRANSPARENCY AT THE FEDERAL RESERVE

Originally, the minutes of the FOMC meetings were not made public. In response to passage of the Freedom of Information Act, which became effective in 1967, the FOMC divided the minutes into two documents. One was called the *Memorandum of Discussion*, which was released with a five-year lag. The other was a shorter document called the *Record of Policy Actions*, which was released with relatively little delay. The *Memorandum* was a set of complete minutes, identifying speakers, but not in the form of a verbatim transcript. The *Record of Policy Actions* reported the Committee's decisions and provided a summary of the Committee's deliberations. However, the *Record* did not identify by name which FOMC members took which positions.

In 1976, in response to a court suit challenging the legality of delaying the release of the *Memorandum*, the FOMC discontinued publication of that document. The Committee continued to publish the *Record of Policy Actions* but in 1993 changed its name to "*Minutes of FOMC Meetings*." Over time the release lag on the *Record/Minutes* was shortened until, at the present time, the *Minutes* are available two days after the next scheduled FOMC meeting.

In the fall of 1993, members of the FOMC became aware that tape recordings of all FOMC meetings since March 1976 had been preserved. These tapes had been made to assist with the preparation of the *Record of Policy Actions* and to provide accurate information about the Committee's views on current policy to senior staff members. However, it was commonly thought within and without the Federal Reserve that the tapes were destroyed when that process had been completed. Many FOMC members where surprised when they learned that the tapes still existed.

In response to congressional pressure, the FOMC agreed in February 1995 to release, with a lag of five years, verbatim transcripts created from the tapes of FOMC meetings and to transcribe past recordings as quickly as possible. At the present time, published transcripts are available for all FOMC meetings from 1979 through 1998. The transcript is complete, except for redactions of confidential material relating to individual firms and foreign governments and central banks. No other central bank provides such complete and explicit records of its policy deliberations.

The FOMC has not adopted a precise, numerical statement of its monetary policy objectives. The Federal Reserve Act, as amended, requires the Board of Governors and the FOMC "to promote effectively the goals of maximum employment, stable prices and moderate long-term interest rates." The FOMC has interpreted its objective as the responsibility to achieve price stability to promote maximum sustainable economic growth.

In contrast to the inflation-targeting central banks, the FOMC has never associated a value or range of values with "price stability." Chairmen Volcker eschewed quantitative specifications of price stability in favor of a less-specific definition. In a 1983 lecture, Volcker put his position this way:

> A workable definition of "reasonable price stability" would seem to me to be a situation in which expectations of generally rising (or falling) prices over a considerable period are not a pervasive influence on economic and financial behavior. Stated more positively, "stability" would imply that decision-making should be able to proceed on the basis that "real" and "nominal" values are substantially the same over the planning horizon—and that planning horizons should be suitably long.[2]

Subsequently, Chairman Greenspan adopted essentially the same definition of price stability.[3]

A small step toward a more explicit statement of the FOMC's inflation objective was taken in 2003 when, at the May FOMC meeting, the Committee indicated that "the probability of an unwelcome substantial fall in inflation, though minor, exceeds that of a pickup in inflation from its already low level." This statement gives a hint about the view of Committee members of the lower end of a tolerance range of measured inflation. At that time, inflation, as measured by the Committee's preferred "core" personal consumption price index, was approximately 1 percent. To date, the Committee has not addressed the question as to what inflation rate would mark the limit such that a substantial rise in inflation above that rate would be unwelcome.

The transparency of the FOMC with respect to policy actions has improved considerably over the past 10 years. Beginning with the February 1994 meeting, the FOMC issued a press release at the conclusion of every meeting at which a policy action was initiated. In spite of the fact that policy actions had been formulated in terms of a specific quantitative objective for the effective federal funds rate since the 1980s, the FOMC only began including the quantitative funds rate objective (called the "intended federal funds rate") in its formal directive to the Federal Reserve Bank of New York at the August 1997 FOMC meeting.[4]

Beginning with the May 1999 FOMC meeting, the FOMC issued a press release at the conclusion of each meeting at which there were major shifts in the Committee's views about prospective developments. These statements included an indication of the policy "bias," which was widely interpreted in the press and in financial markets as hinting at future policy actions.

After the January 2000 FOMC meeting, the policy "bias" in the press release was dropped in favor of a "balance-of-risks" assessment. The statement following the September 2004 FOMC meeting read as follows: "The Committee perceives the upside and downside risks to the attainment of both sustainable growth and price stability for the next few quarters to be roughly equal." To

[2] Paul A. Volcker, "Can We Survive Prosperity?" Remarks at the Joint Meeting of the American Economic Association and the American Finance Association, San Francisco, December 28, 1983, p. 5.

[3] See for example, Alan Greenspan, "Transparency in Monetary Policy," Federal Reserve Bank of St. Louis *Review*, July/August 2002, *84*(4), p. 6.

[4] However, starting with the meeting in January 1996, the Committee's statement issued after a meeting at which it changed the intended funds rate did indicate the anticipated change in the federal funds rate in quantitative terms: "In a related move, the Federal Open Market Committee agreed that the reduction would be reflected fully in interest rates in the reserve markets. This is expected to result in a reduction in the federal funds rate of 25 basis points, from about 5-1/2 percent to about 5-1/4 percent" (from the statement issued January 31, 1996).

provide guidance on its thinking, the Committee might assess the risk of achieving one or the other, or both, of the goals to be tilted to the upside or downside.

Adoption of the balance-of-risks language reflected the Committee's effort to avoid confusion about the interpretation of the wording of the "bias" statement, which specifically referred to the "intermeeting period." The replacement balance-of-risks statement focuses on providing insight into the Committee's assessment of the outlook for future real growth and inflation, but falls short of providing a full fledged forecast of the economy. Along with the decision to adopt the balance-of-risks language, the Committee adopted the policy of providing a press release after every FOMC meeting.

Another important step toward more-predictable policy was for the FOMC to confine policy actions to regularly scheduled meetings. Since February 1994, policy actions other than at a regularly scheduled FOMC meeting occurred only in unusual circumstances.

Finally, in May 2003 the Committee added an additional sentence to the press release: "In these circumstances, the Committee believes that policy accommodation can be maintained for a considerable period." This language was revised in January 2004 to "the Committee believes that it can be patient in removing its policy accommodation." A second revision occurred in May 2004 to "the Committee believes that policy accommodation can be removed at a pace that is likely to be measured." The first two versions of this sentence were commonly interpreted as placing the Committee on hold with respect to future policy actions; the last revision was widely interpreted as hinting that the intended funds rate would be raised in a succession of 25-basis-point increments.

Most recently, in June 2004, the Committee conditioned its "measured pace statement" with the additional sentence that "the Committee will respond to changes in economic prospects as needed to fulfill its obligation to maintain price stability." To date, the actions of the Committee have been consistent with the public interpretations of these statements; no changes in the funds rate occurred under the "considerable period" and "patient" statements, and there have been three increases of 25 basis points each in the intended federal funds rate under the "measured pace" statement.

WHY TRANSPARENCY?

It is natural to ask why central banks need to be transparent. One answer is that central banks are governmental agencies and as such are accountable to the public for their actions. As laudable as it sounds, the accountability argument only gets you so far. For years, Federal Reserve officials argued that immediate release of policy decisions would make markets more unstable and policy implementation more costly and difficult; creating these effects through disclosure would obviously be inconsistent with the Fed's public responsibilities.[5]

Views on whether immediate release of policy decisions would damage monetary policy have changed. Still, the same basic issue remains: How do we determine what level of transparency serves the public interest? For example, some have suggested that the FOMC should conduct its deliberations in public, perhaps televised on C-Span. Common sense and experience suggest, however, that such a practice would curtail the free and open exchange of ideas that characterize FOMC meetings.

Anything that would diminish the effectiveness of the policy process would be inconsistent with the Fed meeting its responsibilities. Accountability requires only that a central bank be open and honest about its objectives and be held accountable for achieving those objectives. Certainly, the ultimate test is whether disclosure yields better policy outcomes.

The roots of central bank transparency are found not only in the principles of democratic accountability but also in economic theory. The economics of transparency is a subject that can be studied systematically, using all the tools of modern economics. Both economic theory and experience demonstrate that the effects of monetary

[5] The minutes of the FOMC meeting of June 20, 1967, list six reasons for delayed release of information.

policy on the real economy—real gross domestic product, real interest rates, the unemployment rate, etc.—are transient. Monetary policy actions only have a lasting effect on inflation, although uncertainty about policy can increase short-run volatility and, perhaps, damage the economic growth process. In such a world, the role of the market's expectations about the central bank's objective for inflation is the principal reason for central bank transparency.

Here is where the story gets a little complicated, so it is useful to consider some extreme and unrealistic cases to illustrate the point. Consider a world where monetary policy actions have no long-run impact on real variables, such as the unemployment rate, but no short-run impact either. Economic theory predicts this state of affairs if all wages and prices were perfectly flexible. In such a world, economic agents would realize that an easing of monetary policy would result in higher prices. Knowing this outcome, prices would adjust immediately: Policy actions would have no effect on the real economy.

Of course, in the real-world economy, prices are not perfectly flexible. This feature of market behavior means that policy actions have short-run effects on the real economy. A policy problem arises because policymakers do not know exactly how monetary policy actions are translated to the real economic variables; policymakers must estimate, or guess, the magnitude of the response of such variables to policy changes and how long these effects last. The only certainty is that the effects of policy actions on real variables eventually dissipate. "Eventually" may cover a period of several years and may be longer in some circumstances than others. It is worth noting that these hedges on my part reflect ignorance—mine and the profession's—and not obfuscations. We just don't have precise estimates of the magnitudes and durations of effects of monetary policy on real variables.

Given that policy actions have a transient effect on the real economy, but a lasting effect only on prices, and given that the effects on the real economy are uncertain in both magnitude and duration, it is important that the central bank be transparent about both its short-run objective for the real economy and its long-run inflation objective. Transparency should help markets to make the best possible adjustments over time and minimize uncertainty flowing from monetary policy itself.

Consider now the issue of the inflation objective. While there is widespread agreement among policymakers and the profession that rapid inflation—such as the inflation that characterized the 1970s and early 1980s—has damaging consequences for the real economy, particularly the long-term rate of economic growth, there is much less agreement on the rate of inflation that maximizes the long-run rate of economic growth. That is, there is little agreement among economists and policymakers about the "optimal" rate of inflation. At the July 1996 meeting of the FOMC, in response to a question by Governor Yellen about the level at which inflation no longer affects business and household decisions, Chairman Greenspan responded, "zero, if inflation is properly measured."[6]

I agree. Given the known biases in price indices, however, exactly what this definition implies for inflation as measured by the personal consumption expenditure (PCE) price index or the consumer price index (CPI) is uncertain. I am inclined to believe that zero inflation correctly measured translates into about 1 percent inflation for the PCE and about 1.5 percent for the CPI.

There is much less agreement in the profession about how much and how long real economic variables are affected by policy actions. This disagreement is confounded by the fact that the effects of monetary policy on the real economy can be influenced by other developments over which policymakers have no control. For example, a particular policymaker might argue that a given easing of policy will not show in prices for x months if there are no other changes in the economic environment. The same policymaker would likely argue that this period will be longer if the easing in policy is accompanied or followed closely by a marked increase in productivity. If the increase in productivity were permanent, this policymaker might argue that the policy easing

[6] Transcript of the July 2-3, 1996, meeting of the FOMC, p. 51.

may have no effect on the price level: Indeed, the rise in productivity could more than offset the policy actions so that, in the long run, prices decline rather than increase, as they would in an unchanged economic environment.

It is easy to see how uncertainty about the magnitude and timing of the effects of policy actions, combined with uncertainty about how other factors impact the magnitude and timing of these effects, can result in significant differences of opinion about the effects of monetary policy. That means that there may also be significant differences of opinion about the extent to which policy can be used effectively to offset the effects of sudden shocks, or evolving long-run structural changes, to the real economy.

Given these real-world uncertainties, it is important for policymakers to be as explicit as possible not only about the central bank's long-run inflation objective but also about its short-run policy objectives. The more ambiguous policymakers are about these objectives, the more difficult it will be for the public to differentiate policy actions that may reflect a change in the central bank's long-run inflation objective from actions intended only to offset the effects of real shocks on economic activity.

Of course, uncertainty about the inflation objective could be reduced by adopting a specific numerical long-run inflation objective. Real-world experience with announced inflation objectives in other countries shows that the issue is more complicated than it might seem. If an objective is stated as a number, what is the effective range around that number? That is, an inflation objective stated as 2 percent might in practice mean 1 to 3 percent. Is the objective to be met over a time horizon of six months or two years? Might the objective be temporarily modified in the face of special circumstances, such as the 9/11 attacks? Being clear about an inflation objective means being clear, or as clear as possible, about all dimensions of such an objective. I personally believe that it is possible to address these practical concerns and state an inflation objective in an effective way. But that is a subject for another day.

Although the FOMC has not announced a precise inflation objective, it has taken a number of steps to better communicate its objectives. The FOMC has made it clear that it "seeks monetary and financial conditions that will foster price stability and promote sustainable growth in output." This statement clearly indicates that the Committee's price stability objective is consistent with sustainable growth in output. While reasonable people may differ on exactly what this inflation rate is, very few would argue that inflation of 4 percent or higher is consistent with maximum sustainable output growth. Most would choose a much lower rate.

The Committee has yet to form a consensus on the circumstances and extent to which monetary policy can be used to offset shocks to the real economy without endangering its price stability objective. To the extent that it reveals the Committee's sensitivity to short-run objectives of policy, the balance-of-risks statement is beneficial in this regard. The balance-of-risks statement also gives market participants a sense of the Committee's views on what it believes the risks are for its short-run and long-run objectives going forward.

The balance-of-risks language is, however, somewhat ambiguous. For example, one might ask: If the risks are unbalanced, why was policy not adjusted to create balanced risks going forward? One answer is that there is no need that these risks be balanced. The inflation objective is a long-run objective, while other objectives are short-run. There is no economic rationale for balancing such objectives.

The balance-of-risks statement can be misinterpreted because of the prevailing view that employment and inflation necessarily rise and fall together. In fact, employment and inflation, or their changes, are not highly correlated.[7] A scatter plot of the changes in employment and inflation reveals that there is no strong positive relationship between inflation and employment. Sometimes they move together; sometimes they move in opposite directions. Consequently, in my view, an unbalanced balance-of-risks statement should not be interpreted as an indication of a future policy action in a specific direction. Unfor-

[7] William Poole, "Fed Transparency: How, Not Whether," Federal Reserve Bank of St. Louis *Review*, November/December 2003, *85*(6), p. 7.

tunately, it is too often interpreted that way by market participants. By failing to clarify the intent of this statement, the FOMC tacitly shares in this confusion.

STATEMENTS ABOUT FUTURE POLICY

In 2000, the FOMC switched from the "tilt" language to the balance-of-risks language, with the explicit intent to avoid signaling future policy actions. Nevertheless, in August 2003 the Committee added a statement that was intended to give the public some idea of how it believed policy might proceed in the near future.

Of necessity, monetary policy is made with an eye to the future—there is nothing current policy can do about the past. Because of the inherently forward-looking nature of policymaking, policy is made with an expectation of how future events are likely to unfold. Moreover, it is only natural that policymakers assign a higher probability to some events than to others. In so doing, policymakers form judgments about whether additional moves to tighten or loosen policy are likely to be desirable. In our present situation, the issue is whether policy tightening might proceed more slowly or more rapidly than one might otherwise anticipate.

The issue with such statements is that they might be misinterpreted as a firm commitment to proceed in a specific way. At any given time, policymakers might feel more or less certain about the probable direction of policy in coming months, but I think it safe to say that they never believe that future policy should be totally unresponsive to events. No matter how firm a conviction I have about the future direction of policy, I know that things could happen that would make me change my mind. The terrible events of 9/11 illustrate this point dramatically. It would have been irresponsible for the Fed to continue on a preset path, ignoring this event.

Thus, forward-looking Fed policy statements should always be interpreted as conditional on future events. A forward-looking statement is not an ironclad commitment but rather a statement of belief based on what we now know. It is unfortunate whenever such a statement is read as a commitment. The objective or expected path for the intended federal funds rate is set based on all of the currently available information—including expectations of future events. If the future turns out exactly as policymakers anticipated at the time the policy path is set, there will be no need to reset it. Only when new information suggests that the previous setting is no longer consistent with achieving the objectives of policy does the Committee need to adjust the setting.

At any given time, the policy path I anticipate may be held with greater or lesser conviction. Put another way, it may take more or less new information outside the range of what I had anticipated to change my mind on the path. Policy decisions are sometimes close calls and sometimes not. And, of course, different policymakers do not all see things the same way. The communications challenge with respect to future policy is to convey accurately how clear the likely policy direction is. Sometimes the expected policy course might be changed only if major unforeseen events occur and sometimes if an accumulation of smaller bits of new information suggest that a change in policy is appropriate.

Given these ambiguities and the danger of misleading the market when indicating a probable future course for policy, I have generally been opposed to announcing, or hinting at, future policy adjustments. However, this year's situation is unusual. When the current round of policy tightening began last June, the target for the intended federal funds rate was 1 percent. After three adjustments of 25 basis points each, the rate now stands at 1.75 percent. When the process started, there was little doubt in anyone's mind that a 1 percent funds rate was significantly below the long-run equilibrium consistent with price stability. Hence, there was little doubt that, over time, the FOMC would raise the intended funds rate. By saying that the policy tightening could proceed at a measured pace, the FOMC indicated a belief that economic conditions going forward likely would allow steady adjustments of the funds rate toward its long-run equilibrium level.

As the process continues, obviously, the intended rate will in time reach a level such that it is not so clear any more that further increases are in order, or that further increases should continue at the same pace. The measured-pace language remains in the FOMC's most recent policy statement, reflecting the Committee's expectation at its last meeting, in mid-September. What actually happens will depend on economic events that are subject to wide forecasting errors. Hence, it is important the market not interpret this statement as a commitment. It is possible—I would argue, likely at some point—that new information will cause the FOMC to adjust the target at a pace different from what is currently anticipated. The pace could be faster or slower, depending on how the economy evolves. In an attempt to underscore this eventuality, the Committee added a sentence to its June 2004 public statement and reiterated it in August and September. The statement read: "Nonetheless, the Committee will respond to changes in economic prospects as needed to fulfill its obligation to maintain price stability."

I believe that it is important to provide as much information as possible about the rationale for policy actions. It might be useful to provide information about likely future policy on a routine basis, but the difficulties of doing so should not be underestimated.

For one thing, the FOMC will not necessarily agree on the likelihood of a future action. It may be confusing to the public if a policy direction is indicated after some FOMC meetings, when the direction is pretty clear, and not after other meetings, when the probable direction is not clear or is subject to dispute within the FOMC. Even if an agreement could be reached, communicating it to the public would be difficult. Indeed, if the probability of future policy action were sufficiently large, some observers might ask, why wait; why not take the action now?

Moreover, it is important to note that a statement of probable future policy direction may actually be a more important policy decision than the setting of the current intended federal funds rate. How easy would it be for a member to agree to a policy action on the intended federal funds rate but dissent over the wording of the policy statement indicating a probable future direction to policy? The FOMC decision process certainly includes the obligation of FOMC members to dissent when they have a fundamental disagreement with the policy decision; that process is well understood today with reference to the decision on the intended federal funds rate. To maintain the integrity of the dissent process, the public will have to understand that dissents may be in order over the wording of the policy statement, a possibility that has not been widely discussed.

CONCLUDING REMARKS

Let me summarize this discussion. The basic framework for policy is that the FOMC sets the intended federal funds rate and individual members have in mind a probable future course for the funds rate. The probable future course may be pretty clear, or may not be, depending on circumstances. Committee members vote on the intended funds rate at the end of each meeting, but historically have not voted, or even tried to develop a consensus, on the probable future direction of policy. Members understand that, whatever their views about the future, actual policy actions in the future will be conditional on information about the economy that cannot be forecast. What the FOMC does in the future is of necessity determined jointly by the FOMC's policy objectives and economic events as they unfold.

The Committee has an obligation to be clear about its policy objectives and should announce any changes in those objectives. In fact, there is a broad public consensus about these objectives and I would be surprised if the objectives change in any material way in the future. Objectives may be clarified, but I do not anticipate significant change.

With clarity over objectives, the FOMC needs to act in as consistent a way as possible in pursuit of the objectives and to explain the process as clearly as possible. When the process is well understood, it is unlikely that policy actions will take the market by surprise. These policy actions will typically be driven by the arrival of new information, which could not be forecast accurately at the time of previous FOMC meetings.

In instances where the market appears to misinterpret the objectives or the intent of a particular action, the FOMC must endeavor to clarify its intention. But more important than dealing with individual episodes is ongoing discussion about monetary policy. A danger in relying on the FOMC's own forecasts of its policy direction is that the market will focus on these forecasts and not on the underlying rationale. Were that to happen, the market will inevitably be surprised when events require policy actions that differ from the FOMC's own forecasts.

Now that you've heard my argument, I'm sure you will agree that transparency may sound easy, but is not.

The Diffusion of Electronic Business in the United States

Emin M. Dinlersoz and Rubén Hernández-Murillo

The authors provide a recent account of the diffusion of electronic business in the U.S. economy using new data from the U.S. Bureau of the Census. They document the extent of the diffusion in three main sectors of the economy: retail, services, and manufacturing. For manufacturing, they also analyze plants' patterns of adoption of several Internet-based processes and conclude with a look at the future of the Internet's diffusion and a prospect for further data collection by the U.S. Census Bureau.

Federal Reserve Bank of St. Louis *Review*, January/February 2005, *87*(1), pp. 11-34.

The commercial use of the Internet has been diffusing rapidly among consumers and businesses in the United States. As the dust of the shakeout in Internet-based industries settled, both firms and consumers started to increase their understanding of what the Internet is capable of and which Internet businesses are likely to be viable. Partly because of the much-publicized mass withdrawal of many firms from the Internet retail industry during most of 2000 and 2001, the Internet's effect on the retail industry has been the focus of both the popular press and academic research. Internet retailing, however, still represents only a very small fraction of online economic activity. In fact, the volume of business-to-business electronic commerce (e-commerce), representing online transactions within and across firms, is far ahead of the volume of business-to-consumer e-commerce, and it has been transforming the way many business transactions are carried out inside and outside of the firm.

Firms are increasingly finding new uses for the Internet—in the retail, services, and manufacturing industries—ranging from applications at the early stages of production, such as communicating and making transactions with suppliers, to post-sales applications, such as providing online customer service and support. Despite the growing volume of e-commerce in these sectors, little is known about the extent to which the Internet is facilitating various transactions and processes at the individual plant and firm levels. This lack of knowledge can in turn be attributed to a lack of systematic establishment-level data on firms' Internet usage. Earlier reviews of the diffusion of electronic business (e.g., Bakos, 2001, and Lucking-Reiley and Spulber, 2001) have provided excellent accounts of the initial stages of the diffusion. Nevertheless, these studies lack any systematic analysis of data and rely mostly on anecdotal evidence. A more detailed and updated look is required, as changes have taken place rapidly in recent years and several new considerations have become relevant.

In this article, we provide a recent account of the diffusion of the Internet in manufacturing, retail, and services. The data we use come from the U.S. Census Bureau's *E-stats Program* (available online at www.census.gov/estats), which provides the first systematic, albeit limited, coverage of e-commerce activity in various sectors of the economy. For many industries, the data include industry sales from e-commerce, making it feasible

Emin M. Dinlersoz is an assistant professor at the University of Houston and Rubén Hernández-Murillo is an economist at the Federal Reserve Bank of St. Louis. The authors thank Roger Sherman for comments and suggestions. Deborah Roisman provided research assistance.

to quantify the extent of diffusion across sectors. In addition, the dataset includes a large sample of plants from various manufacturing industries for which adoption of several Internet-based processes is documented, allowing us to have a first look at the Internet adoption patterns in U.S. manufacturing at the microeconomic level. In particular, we explore the role of plant size in Internet adoption, in view of the discussion of the Internet's role in small businesses compared with large businesses and the Internet's potential to reduce firm size.

We start with an assessment of the evolution of retail e-commerce, the sector that has drawn the greatest attention in the literature. We first provide some background on the general response and reorganization of industries in the wake of inventions and innovations so that we may put the evolution of this sector into perspective. We also present recent statistics on the growth rate of retail e-commerce and discuss the factors enhancing and impeding the adoption of e-commerce across retail industries. We then consider the services sector and document the extent of the diffusion of e-commerce in this sector. Finally, we investigate the adoption patterns in manufacturing.

We rank manufacturing industries according to their tendencies to adopt Internet-based processes at the plant level. We also highlight the relationship between firm size and adoption rate. Earlier studies have invariably found that firm size is a significant factor in the adoption of new technologies, with larger plants typically adopting at a higher rate than smaller ones.[1] This finding appears to apply broadly to the case of Internet-based processes, although there are some important exceptions. We conclude with a look at the future of the Internet's diffusion and prospects for further data collection by the U.S. Census Bureau.

RETAIL E-COMMERCE

During the past decade, a large number of firms entered the Internet's retail markets and then went out of business. While much has been written in the popular media regarding this mass entry and exit and the path that Internet retailing may follow in the aftermath, more work remains to be done to relate these patterns to the impact of other major innovations on retailing. Looking at this broader picture will help us assess the future prospects of retail activity on the Internet. Some guidance in this direction comes from what we already know about the growth patterns of industries following technological innovations. Many of the possibilities the Internet opens up for retailing are new, but some are only improvements over those that were once provided by other major inventions. In evaluating the Internet's impact, it is important to keep in mind that it is only part of the stream of technological breakthroughs that have gradually transformed retail industries.

Industry Life Cycles and Technological Revolutions

According to the industry life-cycle view, industries are like living organisms: They are born, they grow, and they reach maturity. Figure 1 traces the typical time pattern of the number of firms in an industry, from the commercial introduction of a product to the eventual stable state of the number of firms in the industry. An initial period during which only a few firms are active is followed by an episode of an escalating, and then peaking, number of firms that leads to a period of mass exit, called the shakeout.[2] Eventually, the number of firms stabilizes. This pattern is remarkably regular, and it applies to the evolution of many manufacturing industries as initially observed by Gort and Klepper (1982) and later confirmed by Agarwal (1998) for additional industries and longer time periods. Industry life cycles have also been well recognized in the theoretical literature, and several models have been offered to explain the nonmonotonic path that the number of firms follows.[3]

What initiates the pattern in Figure 1 is a

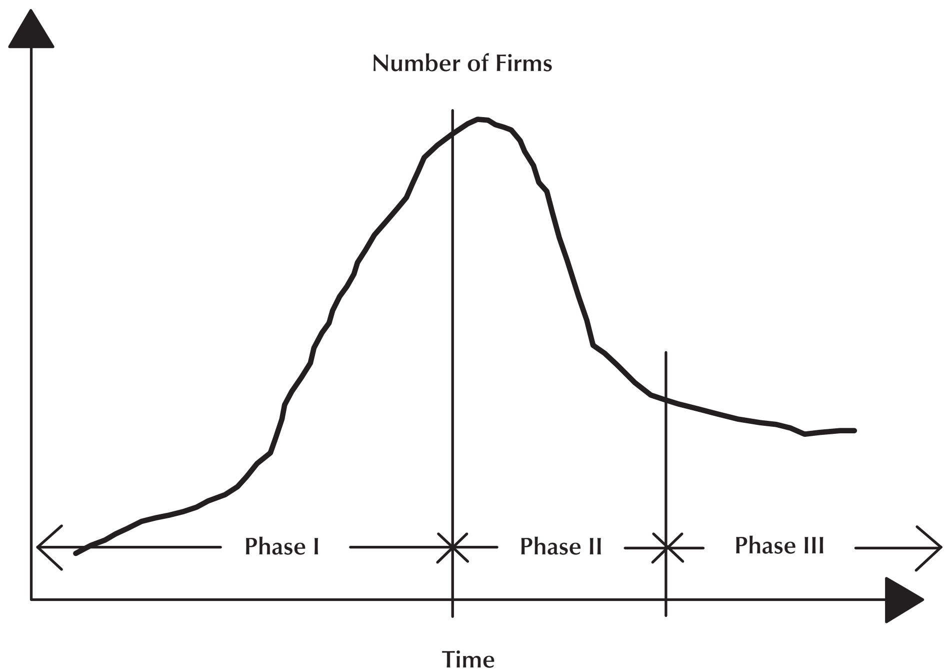
[1] See, e.g., Karshenas and Stoneman (1993), Rose and Joskow (1990), Oster (1982), and Sommers (1980).

[2] There are exceptions to the pattern in Figure 1, as observed by Gort and Klepper (1982). Some industries do not experience a shakeout.

[3] For instance, Jovanovic and MacDonald (1994) consider a model where the shakeout is triggered by an innovation that alters the scale of production.

Figure 1

Evolution of the Number of Firms in an Industry



business opportunity, usually the innovation of a new product or a technological breakthrough that can be exploited commercially. But the life-cycle pattern is not necessarily confined to new manufactured products and also occurs in other industries that experience such breakthroughs.[4]

Following a few first-movers, many firms enter the industry (phase I). However, it is uncertain whether an entrepreneur has the skills to be successful in the new industry, whether the new opportunity is indeed suitable for him, or whether the new product or process will be welcomed by consumers. This uncertainty gradually resolves over time, often when some entrepreneurs realize that the environment is tougher than they expected, or that they overestimated their capabilities. This realization almost invariably triggers the shakeout phase of the life cycle, during which failing entrepreneurs are weeded out and the number of firms declines sharply (phase II). The shakeout ends with the emergence of a set of surviving, successful firms, as the number of firms stabilizes (phase III). At least for manufactured products, total industry output grows throughout the life cycle, even during the shakeout, and the product price falls over time.[5]

In the next subsection, we discuss the diffusion of FM radio broadcasting as an example of the patterns of industry evolution in the wake of technological inventions. For an example of a shakeout that took place on the Internet, see Day, Fein, and Ruppersberger (2003), who consider the case of the shakeout in business-to-business electronic exchanges. As another example, Barbarino and Jovanovic (2003) consider the evolution of the telecom sector in recent years and propose a model of shakeout that embeds the idea of entrepreneurs overshooting the demand in the market by excessively investing in capacity.

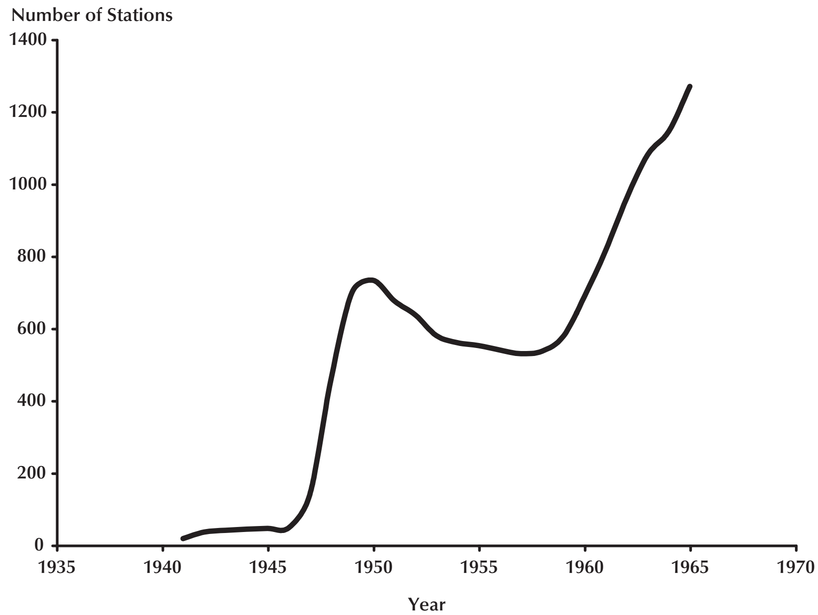
---

4 An example of life-cycle patterns in wholesale trade is given by Fein (1998). More recently, Mazzucato (2002) compares the experience of the personal computer industry to the shakeout episode in the automobile industry.

5 See, again, Gort and Klepper (1982) and Agarwal (1998).

Figure 2

Number of FM Radio Stations: 1941-65



SOURCE: Sterling and Kittross (2002).

The Diffusion of FM Radio

For an example of an industry life cycle generated by technological improvements, consider the commercial diffusion of FM radio broadcasting shown in Figure 2. Much like the Internet, FM technology provided a new medium for broadcasting and opened up a business opportunity for both new and existing radio stations, which could make profits by airing advertisements.

In 1941, the year of the first authorization for commercial FM stations, only five stations were in operation. But the number of stations increased steeply after World War II, peaking in 1950, as the business opportunity was aggressively pursued by both new FM stations and the established AM stations diversifying into FM broadcasting. By 1949, about 85 percent of FM stations were owned by existing AM stations. The AM stations used FM stations frequently as an insurance against a possible demise of the AM technology and at the same time to deter entry by independent FM broadcasters. A shakeout followed between 1950 and 1957, during which 203 stations, about 28 percent of all stations at the peak, shut down. Thereafter, the number of stations rebounded and continued to grow steadily.[6]

A similar pattern of early mass entry and shakeout was observed in the diffusion of AM radio and television stations, but the extents of the entry and the shakeout, their durations, and the reasons driving them were not the same. For example, in the case of AM broadcasting, the main force behind the shakeout was the regulation placed on broadcasting frequencies. In the case of FM stations, the reasons were uncertainty about the future of FM technology, lower-than-expected interest in the new medium from advertisers, competition from AM and television stations, and some conflicts arising from joint ownership of AM and FM stations. Such conflicts were also

[6] In many industries, there is no such post-shakeout growth in the number of firms. The growth in the number of FM stations post-shakeout is probably a consequence of the fact that FM stations are local in nature, and growth in local population over time may have led to an increase in the variety and number of such stations.

pertinent in the early experience of the Internet. That AM stations embraced FM technology to take advantage of synergies, as well as to deter entry by independent FM stations, is similar to the clash between entirely Internet-based retailers and traditional retailers adopting the Internet as a sales channel.

The Evolution of Retail E-commerce

For Internet-based retailers, the business opportunity was clearly not a new product, but rather a new medium through which business could be conducted. Businesses were mainly attracted to this retail medium for (i) its ease of communication between consumers and firms through reduced costs of both advertising and shopping around; (ii) the possibility of eliminating the traditional geographic market boundaries, which allows local entrepreneurs to compete in a wider market; and (iii) the *scale* and *scope* economies made possible by a central warehousing and distribution system that reduces the need for many local facilities and a labor force dispersed across several locations.[7] All of these factors appear to be important considerations for retailing.[8]

The retail industry has benefited from many major innovations, such as the railroad, telegraph, automobile, radio, television, electric elevator, computer, and barcode and scanner technologies. Because doing retail business requires both the flow of goods and the flow of information from one location to the other, any improvement in transportation or communication technologies has an impact on the structure of retail industries. Earlier, the railroad-telegraph combination enlarged the market reach of local retailers and was crucial for the emergence of regional and national department stores and mail-order houses. Automobiles enhanced the physical connection of consumers and retailers, while radio and, later, television further contributed to the emergence of a national market for retailers by increasing the reach of advertising. In this sense, the Internet's effect on retailing is similar to that of other communication technologies, such as newspaper, radio, and television, that help match consumers with firms.

In Internet retailing, we have already witnessed the two phases of the industry life cycle, characterized by the rising and declining number of firms, respectively. What is most interesting about these two phases is that they occurred at a much faster pace than the historical average. A shakeout that spans several years, even decades, in a typical manufacturing industry spanned only a few months in the case of the Internet. Similarly, the initial entry of new firms was much more rapid on the Internet. This can be attributed to easy access to website-design technology that may have reduced entry costs in many, but not all, sectors and to faster diffusion of information about firms' attributes and performance, which probably sped up the demise of inefficient firms and enhanced the dominance of efficient ones.[9]

It appears that the faster pace of these phases is not an entirely new phenomenon, but rather is in line with a gradually decreasing time frame in recent history. The time it takes for additional competitors to enter a new industry in the presence of a few dominant first-movers shrunk throughout the 20th century. Agarwal and Gort (2001) find that this time window decreased from an average of 33 years at the turn of the 20th century down to about 3.4 years for products introduced in the 1967-86 period.[10] Even use of the Internet itself has been diffusing much more rapidly among the U.S. population than major innovations in the past. This appears to be part of a broader trend, that the diffusion of major innovations has been increasingly faster over time.[11]

The adoption of the Internet as a marketing and sales channel proved to be challenging. In the

---

[7] In a single-product firm, *economies of scale* indicates declining per-unit costs as the number of units produced increases; in a multi-product firm, *economies of scope* indicates cost-saving synergies among different product lines.

[8] Dinlersoz and Pereira (2004) provide a theoretical analysis of how these factors may affect adoption incentives for established versus new firms.

[9] See Dinlersoz and Yorukoglu (2004) for an analysis of how improved methods of communication have affected firm and industry dynamics.

[10] See Agarwal and Gort (2001) for potential explanations for this phenomenon.

[11] For instance, it took approximately 45 years for electricity to reach 20 percent of American households, 35 years for the telephone, 25 years for the television, and 15 years for the personal computer.

beginning, the tendency to adopt was quite different for two groups of retailers: existing retailers with established traditional market functions and facilities compared with entirely new entrepreneurs who had no traditional market presence. Even though the website-design technology was available at a low cost to almost anyone who wanted to start a retail business, the cost of investing in warehousing and distribution facilities, which are required for large-scale retail operations, is high in some sectors. Established retailers in such sectors seemed to have an edge with respect to new entrepreneurs, so it is surprising that they were the latecomers.[12]

The reluctance of existing retailers to diversify to the Internet market stemmed partly from the potential problems associated with harmonizing traditional and Internet retail channels, giving rise to *channel conflict*. This conflict comes in many forms, including the resistance of the firm's traditional operations and subunits to the possibility of being replaced by the Internet, the incentives for free riding by traditional market rivals on the product information and related services provided directly on the firm's website, and the possibility that a firm's business on the Internet might compete for its own clientele in the traditional market.[13] Nevertheless, channel conflict currently appears to have lost its role as a major concern in deterring existing retailers from diversifying. Eventually, for well-known traditional retailers, their established names, their ability to raise funding to finance new ventures, and their existing warehousing and distribution facilities allowed them to enter the Internet market strongly. In some product categories, however, the largest online sales today are still made by pure online retailers and by manufacturers selling their products directly, rather than by diversified traditional retailers.[14]

[12] Some Internet-based firms, however, overcame this difficulty by using a method called "drop-shipping," which allowed them to use manufacturers to ship products on their behalf. This reduced the investment needed in warehousing and shipping in some cases.

[13] See, for example, Carlton and Chevalier (2001), Shaffer and Zettelmeyer (2002), and Dinlersoz and Pereira (2004).

[14] For instance, in books, Amazon.com has a much higher share than the traditional retailer Barnes and Noble. See Latcovich and Smith (2001).

During its emergence and early growth, Internet retailing was largely free of regulation. However, one important and persistent policy has been the absence of taxes. Like catalog retailing, Internet commercial activity is free of tax as a result of a moratorium initiated in 1998 that continues to apply. While there has been no other special "infant industry" protection program for Internet retailing, the no-tax environment clearly encouraged the growth of the industry by favoring Internet firms over local firms. Goolsbee (2000) provides preliminary estimates that imposing taxes would have reduced the sales on the Internet by 25 to 30 percent.[15] The evolution of this industry was therefore positively influenced by the absence of taxes. In addition to aiding the growth of Internet retailing, the tax-free environment had some implications for the location of Internet retailers' sales offices and warehouses. Since the shipments within the state where the firm is physically located are subject to local taxes, there are incentives to avoid populous states. However, the tax break neither changed the main course of the industry's evolution nor prevented the shakeout. With taxes, we would have probably observed fewer sales and a smaller number of firms, but no major changes in the trends.

Some Effects of the Internet on Retail Industry Structure

The Internet is a hybrid medium that is capable of combining two basic methods of exchanging information in a market: advertising and shopping around. The reach of the Internet makes these two functions truly global. As a consequence, the location of demand has less influence on retailer location. The geographic separation between the locations of demand and supply can increase the scale and scope of a retailer.

Internet retailers that can dominate the market in a certain category of products are also able to easily expand their operations into other categories. Amazon.com is a good example. Amazon started as a book retailer but now sells many different products. This replicability or expandability,

[15] Also see Ellison and Ellison (2003) for a smaller-scale, but more-recent, analysis of the effects of sales tax on Internet retailing.

in some cases through linkages with traditional retailers, is due to the fact that adding a new product to the existing set of products is probably much easier and cheaper on the Internet. Basically, all that needs to be done is to create digital space for the new product on the website and physical space in the warehouse. Big Internet firms such as Amazon.com have a much wider range of products than traditional big firms, such as Wal-Mart. In addition to the availability of lower prices, the proliferation of varieties on the Internet is a key feature that increases consumer welfare.[16]

Besides enhancing search and advertising, the Internet also offers interactivity. Unlike other media, it allows for a two-way exchange of information between consumers and firms and can also be used to record and store this information—the various steps of this exchange—for future use. This latter feature of the Internet is especially useful for retailing because it makes it possible for firms to learn about consumers' preferences by analyzing their shopping patterns. This type of information extraction works in favor of customization of goods and services to satisfy finer individual tastes. In this respect, the Internet is an advanced form of the scanner technology used at the checkout counter that previously revolutionized retailing by allowing firms to monitor what consumers bought. The Internet also enables firms to target consumers individually or in small groups, unlike other communication tools, such as radio and television, which can at best target large, coarsely defined groups of consumers.

The Internet also offers firms the possibility to monitor rival firms' strategies more closely, especially their prices and promotional efforts, making it easier for firms to respond quickly to changes in rivals' strategies. The costs of pricing products and adjusting prices, referred to as *menu costs*, appear to be much lower on the Internet.[17] This feature is likely to speed up the pace of competition in retail markets.

What will be the main characteristics of retail industries on the Internet in the future? Will the industry structure look more like a competitive industry or a monopolistically competitive one, with many small firms each serving a particular niche in the market? Or will it be more concentrated with a few large firms dominating the market for a particular product type or many product lines simultaneously? It is too early to answer this question convincingly. Clearly, there are features of the Internet that can promote entry, competition, and fragmentation. Initially, it was believed that low entry costs associated with operating a website might foster entry and competition. However, the Internet also provides an environment in which the scale and scope of operations can be expanded at very low cost and information about a firm's attributes can be disseminated easily; it also can give rise to firms that can quickly become large. These features can lead to high concentration.

While some early findings suggest that industry concentration ratios on the Internet were initially much higher than their traditional market counterparts, there is no overwhelming evidence that this is the case. In one of the earlier studies, Latcovich and Smith (2001) find that industry concentration is much higher on the Internet than in the traditional market in the case of book and music retailing. The authors also report that advertising and promotion efforts are more intense on the Internet compared with the traditional market. Thus, post-entry "sunk costs" in the form of investment in advertising and customer loyalty programs may be an important aspect of competition. Such investments have the potential to deter entry and lead to a highly concentrated market structure.[18]

In a more comprehensive study, Noam (2003) also points to high concentration, as measured by the Herfindahl-Hirschman index (HHI), in several industries for the pre-2002 period.[19] He finds that the Internet sector's overall concentration was high, and concentration initially declined

[16] See Brynjolfsson, Smith, and Hu (2003) on the welfare gains to consumers from a high level of variety in online markets.

[17] Brynjolfsson and Smith (2000) estimate that menu costs are substantially lower on the Internet compared with the traditional market. Changing prices of products on the Internet requires simply updating price listings on a website, as opposed to physically marking products on the shelves, which is costly.

[18] For theoretical arguments behind this, see Sutton (1991). Also see Dinlersoz and Yorukoglu (2003) for an alternative analysis of the role of the lower cost of advertising in changing market structure.

[19] The Herfindahl-Hirschman concentration index is defined as the sum of the square of participant firms' output market shares.

in the 1980s and 1990s, but increased toward the mid-1990s. For data starting in 2000, Baye and Morgan (2003) find that the average HHI for 5000 products in their sample initially increased between August 2000 and February 2002, but then exhibited a clear decline until November 2003. The average HHI in their sample, though, is much lower than those in Noam (2003). The authors conclude that differences between the industries analyzed and in the market definitions may be the cause for the discrepancy between the two studies. In some markets, such as for local Internet access providers, there are many competitors for any given town and concentration is low. In other markets, such as for broadband providers in a city, there are only a few competitors and concentration is very high.

Aside from the evidence discussed so far, there is no systematic comparison of concentration levels in traditional versus Internet markets. One of the important issues in such a comparison is the comparability of the industry definitions in U.S. Census Bureau data on traditional retail industries and the data collected independently by individual researchers on Internet industries. The main data source on traditional retail industries, the Census of Retail Trade, provides concentration measures at the four-digit *industry level*, which usually consists of several products. Most of the data privately collected by researchers, on the other hand, are compiled at the *product level*. Unless such product level data are aggregated to the four-digit industry level, compatible with the Census Bureau's industry definitions, a direct comparison of the concentration ratios is not possible. A second issue is the definition of the concentration ratio itself. The Census of Retail Trade reports only *n-firm concentration ratios*, such as a four-firm or an eight-firm concentration ratio.[20] To be comparable with these definitions, independent data collected by researchers must contain enough information to calculate similar ratios. These shortcomings point to a demand for more organized data collection by the Census Bureau, an issue we return to in our conclusion.

The Growth of Retail E-commerce Sales

Despite the shakeout, retail e-commerce sales have been growing at a steady pace over the years, as shown in Figure 3. While the current share of retail sales accounted for by e-commerce is still very low (around 2 percent), its growth rate is very high. As total retail sales grew at an average rate of 1.3 percent quarterly over the sample period, e-commerce sales exhibited an average growth rate of 8.6 percent. The strong seasonality in e-commerce sales is also apparent from Figure 3, with fourth quarters exhibiting exceptional growth, due to the surge in online shopping during holiday seasons.

The sectoral breakdown of the share of retail e-commerce sales is shown in Table 1. In almost all sectors, the share in 2002 was less than 1 percent, and the differences across sectors were not highly perceptible. Table 2 presents the percentage of sales accounted by e-commerce by merchandise line, considering only the firms classified as "electronic and mail-order houses." The electronic and mail-order houses industry includes all catalog and mail-order houses and other direct retailers, many of which sell in multiple channels, as well as pure Internet-based firms and hybrid "brick-and-click" retailers, if the e-commerce group operates as a separate unit and is not engaged in the online selling of motor vehicles.

The diffusion of e-commerce sales was relatively rapid and widespread among electronic and mail-order houses compared with other retail sectors, and differences across merchandise lines in the share of e-commerce are more visible in this industry. In 2001, the highest shares were observed in books and magazines, electronics, and music and videos. Relatively low shares were observed in food, beer and wine, clothing and apparel, and drugs.[21]

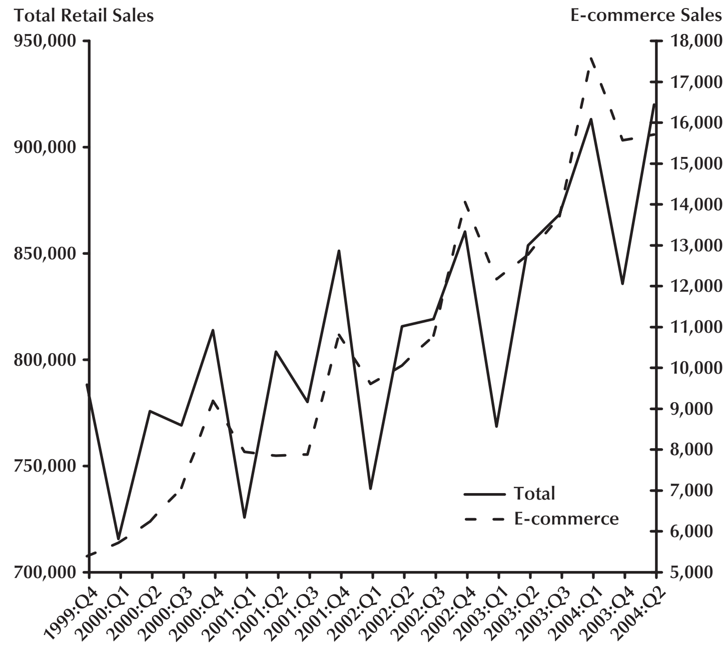
These observations make clear that the nature of the product matters for the extent of the diffusion. However, the differences across categories are expected to vanish over time as both sellers and buyers experiment with various product types

[20] The *n-firm concentration ratio* is defined as the market share accounted for by the *n* largest firms in the market.

[21] Part of the lack of growth observed in beer and wine e-commerce sales is probably related to the restrictions set on interstate shipments of alcohol by many states.

Figure 3

Growth of Total Retail Sales Compared with Growth of E-commerce Sales (millions of dollars)



and discover which products within a category are most conveniently and cost-effectively traded online. Such convergence is already happening to some extent. Some product categories in which e-commerce had little share initially have exhibited strong growth. Examples are food, beer and wine, furniture and home furnishings, and clothing. This growth is likely to be a result of consumers and firms becoming more familiar with the Internet environment and overcoming the concerns they initially had about the medium.

Many other sectors that were once thought of as relatively unsuitable for Internet retailing have been on the rise. A very recent example is jewelry.[22] Mullaney (2004) reports that Internet-based startups are slowly taking over this product category, especially in diamonds. The main reason for the success of Internet-based firms appears to be the substantial cost savings for online retailers in selling diamonds, for which sales traditionally involve several stages before the item reaches the customer. These layers of middlemen, experts, appraisers, and sales force are dramatically reduced for online sellers.[23] As diamond sales on the Internet increase, some traditional retailers that specialize mostly in standard diamond types may lose their market share. On the other hand, some other traditional retailers rely more on image and brand, so that customer loyalty to their name makes them relatively less vulnerable to the effects of increasing online sales. In the meantime, many other small traditional retailers are facing a choice between focusing on more specialized (compared with standardized) diamonds so that they can avoid direct competition with online retailers. This behavior of traditional retailers is just one example of retail industries' reorganization in response to the emergence of e-commerce and is

[22] In April 2004, Amazon.com posted an open letter on its website (signed by the founder, Jeff Bezos) announcing that the company was entering the jewelry market.

[23] It is estimated that a physical chain would need 116 stores and more than 900 workers to match the sales of the leading firm in the Internet market (see Mullaney, 2004).

Table 1

U.S. Retail Trade Sales[1]—Total and E-commerce[2]: 2002 and 2001

NAICS code	Description	E-commerce as % of total sales: Percentage[3] 2002	Percentage[3] 2001	Standard error 2002	Standard error 2001	% Distribution of sales: E-commerce 2002	Total 2002
	Total retail trade	**1.4**	**1.1**	**(Z)**	**(Z)**	**100.0**	**100.0**
441	Motor vehicles and parts dealers	0.9	0.6	(Z)	(Z)	16.3	26.2
442	Furniture and home furnishings stores	(S)	(S)	(S)	(S)	(S)	2.9
443	Electronics and appliance stores	0.9	0.8	0.2	0.1	1.8	2.8
444	Building materials and garden equipment and supplies stores	0.2	0.2	(Z)	(Z)	1.4	9.3
445	Food and beverage stores	(S)	(S)	(S)	(S)	(S)	15.2
446	Health and personal care stores	(S)	(S)	(S)	(S)	(S)	5.6
447	Gasoline stations	(Z)	(Z)	(Z)	(Z)	(Z)	7.6
448	Clothing and clothing accessories stores	0.3	0.2	(Z)	(Z)	1.1	5.3
451	Sporting goods, hobby, book, and music stores	0.8	0.6	0.1	0.1	1.5	2.5
452	General merchandise stores	(S)	(S)	(S)	(S)	(S)	14.0
453	Miscellaneous store retailers	0.7	0.5	0.1	0.1	1.5	3.2
454	Nonstore retailers	18.7	15.0	0.3	0.2	74.8	5.5
454110	Electronic shopping and mail-order houses	28.1	23.0	0.3	0.3	72.7	3.5

NOTE: Reproduced from Tables 5 and 5A in the U.S. Census Bureau's "E-commerce Multi-sector Report." [1]Estimates are based on data from the U.S. Census Bureau, 2002 Annual Retail Trade Survey. Sales estimates are shown in millions of dollars; consequently, industry group estimates may not be additive. [2]Estimates include data for businesses with or without paid employees and are subject to revision. [3]Estimates are not adjusted for price changes. For information on confidentiality protection, sampling error, nonsampling error, sample design, and definitions, see www.census.gov/eos/www/restats.html. (S) Estimate does not meet publication standards because of high sampling variability or poor response quality. Unpublished estimates derived from this table by subtraction are subject to these same limitations and should not be attributed to the U.S. Census Bureau. (Z) Sales estimate is less than $500,000 or percent estimate is less than 0.05 percent.

reminiscent of the way local markets were once reshaped by the entry of Wal-Mart stores and other dominant chains.

SERVICES AND THE INTERNET

Services industries have also been embracing the Internet rapidly, even though the overall share of e-commerce in total revenues is still below 1 percent, as shown in Figure 4. In some ways, the affinity between the Internet and services industries is not very surprising. Services industries in general have been quick in adopting the basic technologies such as computers and Internet access. Moreover, since many service products are essentially information goods that come in digital form, they can be easily traded online. Examples are publishing services, information services, travel reservations, and even mortgage lending and stock trading. Such goods that can be traded in digital form are bound to become dominant categories in online retailing, as argued by Dinlersoz and Pereira (2004), because they can be conveniently delivered and returned by e-mail, they can bypass wholesale and retail layers, they

Table 2

U.S. Electronic Shopping and Mail-Order Houses[1]—Total and E-commerce Sales by Merchandise Line[2]

	E-commerce as % of total sales				% Distribution of sales	
	Percentage[3]		Standard error		E-commerce	Total
Merchandise line	2002	2001	2002	2001	2002	2002
Total electronic shopping and mail-order houses (NAICS 454110)	**28.1**	**23.5**	**0.3**	**0.3**	**100.0**	**100.0**
Books and magazines	46.0	44.9	1.6	1.6	3.5	5.7
Clothing and clothing accessories (includes footwear)	30.5	21.2	0.5	0.5	12.2	13.3
Computer hardware	27.7	25.7	0.5	0.5	18.5	18.2
Computer software	32.8	30.4	1.2	1.4	3.9	4.5
Drugs, health aids, and beauty aids	7.0	5.9	0.8	0.8	18.1	4.5
Electronics and appliances	45.9	39.3	1.4	1.5	3.9	6.3
Food, beer, and wine	34.2	24.2	1.6	1.2	1.6	2.0
Furniture and home furnishings	34.4	25.4	1.3	1.4	6.2	7.6
Music and videos	37.6	32.9	0.9	1.2	3.4	4.5
Office equipment and supplies	40.1	30.0	0.9	0.9	5.3	7.6
Sporting goods	33.9	28.3	3.2	3.1	2.3	2.8
Toys, hobby goods, and games	36.1	31.0	2.0	1.9	3.0	3.9
Other merchandise[4]	24.7	18.4	0.7	0.7	13.7	12.0
Nonmerchandise receipts[5]	45.9	38.2	0.8	0.9	4.3	7.0

NOTE: Reproduced from Tables 6 and 6A in the U.S. Census Bureau's "E-commerce Multi-sector Report." [1]Estimates are based on data from the U.S. Census Bureau, 2002 Annual Retail Trade Survey. Sales estimates are shown in millions of dollars; consequently, industry group estimates may not be additive. [2]Estimates include data for businesses with or without paid employees, are grouped according to merchandise categories used in the Annual Retail Trade Survey, and are subject to revision. [3]Estimates are not adjusted for price changes. For information on confidentiality protection, sampling error, nonsampling error, sample design, and definitions, see www.census.gov/eos/www/restats.html. [4]Includes other merchandise such as collectibles, souvenirs, auto parts and accessories, hardware, lawn and garden equipment and supplies, and jewelry. [5]Includes nonmerchandise receipts such as auction commissions, customer training, customer support, advertising, and shipping and handling.

require neither physical storage space nor transportation, and online demos make product information easy to obtain and product quality easy to verify. Therefore, both firms and consumers stand to gain substantially by trading digital goods online.

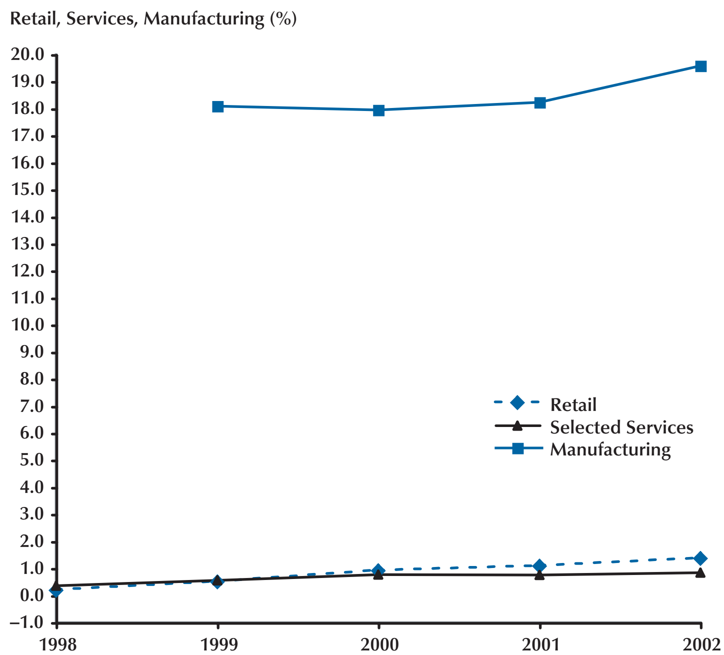
In general, digital products are different from non-digital products, including their pricing and distribution. For such goods, the initial fixed production cost tends to be high, but the marginal cost is generally low. For instance, a computer program may have a substantial development cost, but producing a copy of it is relatively simple and cheap. These peculiar features of digital goods have been the subject of recent research.[24]

Table 3 contains the share of e-commerce sales for various services. Sectors leading in the penetration of e-commerce sales are publishing, online information services, securities and commodity contracts intermediation and brokerage, computer systems design and related services, and travel arrangement and reservation services.

[24] See, e.g., Varian (1995, 2000, 2001).

Figure 4

Revenue Share of E-commerce in Retail, Services, and Manufacturing



Many sectors still have low penetration rates. The data for certain sectors are not of high quality and await further development and refinement in the collection process. Furthermore, some sectors, such as mortgages—a rising sector on the Internet—have not been included.

The travel industry is far ahead of any other industry in the services sector in terms of its share of e-commerce. The importance of a consumer's ability to search and the dynamic nature of travel arrangements make this category very suitable for e-commerce. The demand, capacity, and prices are relatively more volatile and seasonal in this industry, implying that real-time price changes can be monitored by both firms and consumers more easily online than offline. Furthermore, transaction costs are much lower for this industry online than offline, and travel firms are able to pass these cost savings on to consumers in the form of lower prices. Another attractive feature of online travel reservations is that a consumer can select different elements and stages of a trip, such as flight, hotel, car rental, and local tours, in one big bundled reservation. This flexibility in bundling is a source of utility for consumers. This kind of bundling also existed in traditional markets for a long time, but the travel websites make it much easier and much more flexible. Considering all the benefits of online shopping, the travel industry is a prime candidate for becoming the first big industry with the majority of its sales online.

MANUFACTURING AND E-COMMERCE

The Census Bureau's survey of e-commerce activity indicates that industry penetration of the Internet with e-commerce sales has been highest in the manufacturing sector, followed by wholesale, services, and retail. Not surprisingly, manufacturing also leads in terms of the Internet's impact on business-to-business transactions. In fact, the Internet's biggest and most immediate impact has been reduced transaction costs and enhanced efficiency in many ordinary business

Table 3

U.S. Selected Services Revenue[1]—Total and E-commerce[2]: 2002 and 2001

NAICS code	Description	E-commerce as % of total sales: Percentage[3] 2002	Percentage[3] 2001	Standard errors 2002	Standard errors 2001	% Distribution of sales: E-commerce 2002	Total 2002
	Total for selected services industries	**0.9**	**0.8**	**(Z)**	**(Z)**	**100.0**	**100.0**
	Selected transportation and warehousing[4]	**1.4**	**1.2**	**0.1**	**0.1**	**8.3**	**4.9**
484	Truck transportation	1.4	0.9	0.2	0.1	5.8	3.5
492	Couriers and messengers	1.7	2.2	0.1	(Z)	2.2	1.1
493	Warehousing and storage	(S)	(S)	(S)	(S)	(S)	0.3
51	**Information**	**1.3**	**1.2**	**(Z)**	**(Z)**	**26.6**	**18.0**
511	Publishing industries	2.3	2.1	0.1	0.1	12.9	4.7
513	Broadcasting and telecommunications	0.5	0.5	(Z)	(Z)	6.1	10.0
51419	Online information services	5.7	5.7	0.5	0.6	4.4	0.7
	Selected finance[5]	**1.6**	**1.3**	**0.1**	**(Z)**	**10.1**	**5.3**
5231	Securities and commodity contracts intermediation and brokerage	2.5	1.9	0.1	(Z)	9.8	3.4
532	**Rental and leasing services**	**(S)**	**(S)**	**(S)**	**(S)**	**(S)**	**2.1**
	Selected professional, scientific, and technical services[6]	**0.8**	**0.6**	**0.1**	**(Z)**	**15.6**	**17.4**
5415	Computer systems design and related services	2.6	2.0	0.4	0.1	10.3	3.3
	Selected administrative and support and waste management and remediation services[7]	**2.5**	**2.3**	**0.1**	**0.1**	**25.2**	**8.7**
5615	Travel arrangement and reservation services	24.1	23.7	0.8	0.9	15.4	0.5
62	**Health care and social assistance services**	**(S)**	**(S)**	**(S)**	**(S)**	**(S)**	**24.7**
71	**Arts, entertainment, and recreation services**	**(S)**	**(S)**	**(S)**	**(S)**	**(S)**	**2.8**
72	**Accommodation and food services[8]**	**(S)**	**(S)**	**(S)**	**(S)**	**(S)**	**9.4**
	Selected other services[9]	**0.3**	**0.2**	**(Z)**	**(Z)**	**2.6**	**6.7**
811	Repair and maintenance	0.2	0.2	(Z)	(Z)	0.6	2.7
813	Religious, grantmaking, civic, professional, and similar organizations	0.5	0.3	(Z)	(Z)	1.5	2.5

NOTE: Reproduced from Tables 4 and 4A in the U.S. Census Bureau's "E-commerce Multi-sector Report." [1]Except where indicated, estimates are based on data from the U.S. Census Bureau 2002 Service Annual Survey. Revenue estimates are shown in millions of dollars; consequently, industry group estimates may not be additive. [2]Estimates are subject to revision and include data only for businesses with paid employees except for Accommodation and Food Services, which also includes businesses without paid employees. [3]Estimates are not adjusted for price changes. For information on confidentiality protection, sampling error, nonsampling error, sample design, and definitions, see www.census.gov/eos/www/sestats.html. [4]Excludes NAICS 481 (air transportation), 482 (rail transportation), 483 (water transportation), 485 (transit and ground passenger transportation), 486 (pipeline transportation), 487 (scenic and sightseeing transportation), 488 (support activities for transportation), and 491 (postal service). [5]Excludes NAICS 521 (monetary authorities–central bank), 522 (credit intermediation and related activities), 5232 (securities and commodity exchanges), 52391 (miscellaneous intermediation), 52399 (all other financial investment activities), 524 (insurance carriers and related activities), and 525 (funds, trusts, and other financial vehicles). [6]Excludes NAICS 54112 (offices of notaries) and 54132 (landscape architectural services). [7]Excludes NAICS 56173 (landscaping services). [8]Estimates are based on data from the 2002 Annual Retail Trade Survey. [9]Excludes NAICS 81311 (religious organizations), 81393 (labor and similar organizations), 81394 (political organizations), and 814 (private households). (S) Estimate does not meet publication standards because of high sampling variability or poor response quality. Unpublished estimates derived from this table by subtraction are subject to these same limitations and should not be attributed to the U.S. Census Bureau. (Z) Estimate is less than 0.05 percent.

Table 4

Ranking of Manufacturing Industries by Rate of Adoption of Internet-Based Processes

NAICS code	Description	Average rank	Average adoption rate
334	Computer and electronic products	1	0.33
336	Transportation equipment	2	0.29
335	Electrical equipment, appliances, and components	4	0.30
333	Machinery	5	0.26
331	Primary metals	5	0.24
326	Plastics and rubber products	6	0.24
325	Chemicals	7	0.25
323	Printing and related support activities	8	0.27
322	Paper	9	0.23
339	Miscellaneous	10	0.23
332	Fabricated metal products	12	0.22
314	Textile product mills	12	0.21
312	Beverage and tobacco	13	0.21
316	Leather and allied products	14	0.20
324	Petroleum and coal products	14	0.19
315	Apparel	16	0.18
313	Textile mills	18	0.18
311	Food products	18	0.18
337	Furniture and related products	18	0.18
327	Nonmetallic mineral products	19	0.16
321	Wood products	21	0.15

exchanges between firms and within a firm, rather than between firms and consumers. In the next two sections, we document the diffusion of several important Internet-based processes used by manufacturing plants in facilitating stages of production.

Leading Sectors and Processes

To understand the extent and prevalence of manufacturing plants' use of Internet-based processes, we present two simple rankings. Table 4 ranks industries in terms of plants' tendencies to use the Internet for various processes.[25] Here, we assume that a plant in industry i adopts process j with probability p_{ij} independently of other plants. We then compute $\hat{p}_{ij}$, an unbiased estimate of this probability, as the ratio of the number of plants in industry i that adopted process j, n_{ij}, to the total number of plants surveyed in industry i, N_i.[26]

After obtaining estimates $\hat{p}_{ij}$ for each industry i and for each process j, we simply ranked industries according to the rate of adoption of each process and then took the average of these ranks across all processes by industry. We then ranked industries based on this "average rank." The resulting ranking in Table 4 reveals that industries that are generally perceived to be technologically

[25] A shortcoming of the data is that we do not have information on the intensity of usage of a process in a plant. Thus, we only summarize adoption as an all-or-nothing decision, even though firms may have different degrees of usage intensity after adoption.

[26] The estimated standard deviation of $\hat{p}_{ij}$ can be calculated as

$$\hat{\sigma}_{\hat{p}_{ij}} = \sqrt{\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{N_i}}\ .$$

Table 5

Ranking of Internet-Based Processes by Their Rates of Adoption in Manufacturing Industries

Process	Average rank	Average adoption rate
Basic Internet access and degree of access	1	0.84
Access to vendors' products or catalogs	2	0.48
Ordering of materials and supplies	4	0.41
Product descriptions or online catalog for external suppliers	5	0.35
Ordering from vendors	5	0.31
Inventory data for other company units	6	0.30
Ordering by customers	7	0.25
Order status for other company units	8	0.24
Customer support	9	0.22
Product descriptions or online catalog for other company units	10	0.20
Order status for external suppliers	12	0.17
Acceptance of orders for manufactured products	12	0.17
Payment by customers	13	0.14
Product descriptions or online catalog for external customers	14	0.12
Payment to vendors	14	0.11
Outsourcing of research and development	16	0.09
Bidding	18	0.07
Inventory data for external suppliers	18	0.07
Electronic marketplaces linking specialized business buyers and suppliers	18	0.07
Order status for external customers	19	0.06
Inventory data for external customers	21	0.04

advanced, such as *machinery, electrical equipment, computer and electronic products*, and *transportation equipment*, tend to rank high. These industries are also the ones where computers have traditionally been applied in various ways. Industries that are at the bottom of the list are *wood products, nonmetallic mineral products*, and *furniture and related products*.

The second summary, shown in Table 5, is the ranking of Internet-based processes based on their rates of adoption in different industries. As in Table 4, we first ranked all processes for each industry in terms of adoption rate and then calculated the average rank for each process across all industries. The most heavily adopted processes are *basic Internet access and degrees of access, access to vendors' products or catalogs*, and *ordering from vendors*. The least adopted processes are *provision of inventory data for external customers* and *provision of order status information for external customers*.

Somewhat surprisingly, the adoption rates of *online bidding* and *use of electronic marketplaces* are relatively low. These processes are precisely the ones that were initially thought to be revolutionary. Day, Fein, and Ruppersberger (2003) argue that the limited success of these applications can be attributed to the fact that online exchanges did not dramatically alter the existing way firms manage their supply chains. Firms value obtaining the right combination of products at the right time, and coordinating complex production activities is easier with a dedicated, traditional supply chain. The cost savings offered by online exchanges were simply not enough to convince firms to sacrifice other aspects of production, such as timeliness and access to preferred brands.

Plant Size and Adoption Rate

The increasing use of the Internet for transactions within and across firms has also raised the question of whether the rate of usage is closely associated with firm size. A related issue is how adoption of Internet-based processes affects firm size. As Varian (2002) pointed out, it is not clear in which direction firm size will move as Internet-based transactions continue to replace traditional ones. The answer depends on the relative magnitudes of competing forces. If Internet-based transactions reduce the costs of using external markets by more than they reduce internal transaction costs, then firm size can decrease. The data available are not suitable for a full analysis of the Internet's effect on firm size, but they are informative with respect to the role that plant size plays in adoption.

We can estimate the rates at which certain Internet-based processes are adopted by plants of different sizes. For 10 employment size groups, the data contain the number of plants that have adopted a certain Internet-based process at the time the survey was conducted.[27] We can again assume that the population of plants in size group k is generated by a Bernoulli distribution with parameter p_{ijk}, which can be estimated as the ratio of the number of plants in industry i that adopted process j, n_{ijk}, to the total number of plants surveyed in this size group, N_{ik}. In other words, a plant in size group k adopts the process with probability p_{ijk} independently of other plants in the size group and in other size groups.[28]

The sampling procedure used by the census is a *probability-proportional-to-size* sampling scheme in that large plants are sampled with higher frequency and small plants are underrepresented in the sample. Therefore, the standard errors on the estimates for smaller plants are in general higher.[29] As an example, consider the estimated rate of Internet access by plant size in Figure 5. The smallest plant size group has an estimated adoption rate of 48 percent compared with 98 percent for the largest group. For larger size groups, the estimated values are higher and the estimated standard deviations are lower, in part reflecting the sampling scheme mentioned. Consequently, the confidence intervals are narrower for larger size groups and the differences between estimated adoption rates are usually highly significant across size classes, with a few exceptions.

The pattern in Figure 5 is generally applicable to a majority of the processes. In some cases, the standard deviations of the estimates increase with plant size, implying that there is much variation in the adoption rate among large plants, after controlling for the fact that they are represented more heavily in the sample. In the following discussion we will focus on characterizing whether the adoption rate generally exhibits a positive and statistically significant relation to plant size.

For a compact presentation of the patterns, we aggregated the 10 employee size groups into three size classes: small plants (with 1 to 20 employees), medium plants (with 21 to 99 employees), and large plants (with 100 or more employees). Table 6 confirms that in many cases there is a statistically significant increase in the adoption rate as plant size increases. Exceptions occur for some important processes, however. In the case of *placement of orders for materials and supplies*, the adoption rate declines with plant size, as shown in Figure 6. A similar pattern is observed for *acceptance of orders for manufactured products*, as seen in Figure 7. While these exceptions deserve further exploration, lack of plant characteristics prevents us from reaching a definitive conclusion about the adoption rate/firm size relationship.[30] Since larger plants are more likely to be vertically integrated, it is quite possible that these plants rely less on the Internet to access outside suppliers. This explanation may also apply to the case of accepting orders online, albeit to a lesser extent.

---

[27] The size groups are 1 to 4, 5 to 9, 10 to 19, 20 to 49, 50 to 99, 100 to 249, 250 to 499, 500 to 999, 1000 to 2499, and 2500+ employees.

[28] For simplicity's sake, we make the assumption that a plant's adoption decision is independent of the overall adoption rate in the industry. Externalities in adoption are likely to affect the probability of adoption for at least some processes.

[29] The estimated standard deviation of the estimated probability, denoted by $\hat{p}_{ijk}$, can be obtained as

$$\hat{\sigma}_{\hat{p}_{ijk}} = \sqrt{\frac{\hat{p}_{ijk}(1-\hat{p}_{ijk})}{N_{ik}}}.$$

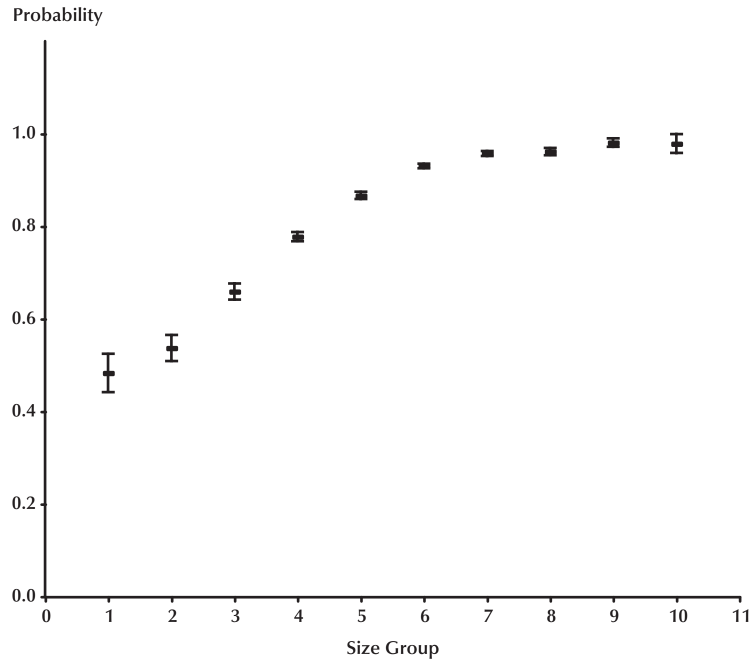
A 95 percent confidence interval for the true adoption probability, p_{ijk}, is then given as

$$[\hat{p}_{ijk} - 1.96\hat{\sigma}_{\hat{p}_{ijk}},\ \hat{p}_{ijk} + 1.96\hat{\sigma}_{\hat{p}_{ijk}}].$$

[30] Plant characteristics are available from the U.S. Census Bureau, but only for on-site usage, as they are classified as confidential data.

Figure 5

Adoption Rates of Internet Access by Manufacturing Plant Size



Two other processes deserve attention. It appears that plant size has little effect on the adoption rate of *online bidding* and *use of electronic marketplaces*, as shown in Figures 8 and 9. While sampling errors may contribute to these two patterns, there does not appear to be a highly statistically significant increase in the adoption rate of these two processes as plant size increases. In fact, both processes are adopted with a rate of less than 20 percent by plants of all sizes. The low adoption rates of these two processes notwithstanding, virtually indistinguishable rates of adoption across a wide range of size classes suggest that large plants may be benefiting from these external market activities as much as small plants are. Obviously, without the intensity of usage of these two processes by plants, a definitive conclusion cannot be reached based on only adoption rates. Nevertheless, one might have expected a priori that small plants adopt these two processes at a higher rate than larger ones, as smaller plants may rely more on these external market activities because of a lack of internal subunits that focus on individual stages of production and procurement.

One of the conjectures about the Internet's impact on the organization of production was that it would lead to more vertical disintegration. Along Coase's (1937) arguments, if the cost of making transactions outside of the firm declines, firms should have increased incentives to carry out these transactions with outside specialists, rather than within the firm. While our results do not offer any direct evidence on the issue, they suggest that, at least for some stages of production, this may be happening to some extent. Most processes are adopted at a higher rate by larger plants. Some of these processes are those that can induce vertical disintegration, such as *placement of orders for materials and supplies online, ordering from vendors, payment to vendors, online bidding, use of electronic marketplaces*, and *outsourcing of research and development*. As such processes are adopted with higher frequency and intensity, plants and firms may reduce the size of internal units undertaking these functions or eliminate them altogether.

Table 6

Adoption Rates of Internet-Based Processes by Plant Size[1]

	Plant size[2]		
Process	Small	Medium	Large
Basic Internet access and degree of access	0.6072 (0.0071)	0.9585 (0.0017)	0.9406 (0.0017)
Product descriptions or online catalog for other company units	0.0759 (0.0039)	0.1368 (0.0029)	0.2717 (0.0033)
Product descriptions or online catalog for external customers	0.0620 (0.0036)	0.1147 (0.0027)	0.1540 (0.0027)
Product descriptions or online catalog for external suppliers	0.2117 (0.0061)	0.3496 (0.0041)	0.4108 (0.0036)
Order status for other company units	0.0927 (0.0043)	0.1622 (0.0031)	0.3127 (0.0034)
Order status for external customers	0.0304 (0.0026)	0.0467 (0.0018)	0.0779 (0.0020)
Order status for external suppliers	0.0896 (0.0043)	0.1410 (0.0030)	0.2192 (0.0030)
Inventory data for other company units	0.1314 (0.0050)	0.2064 (0.0035)	0.3782 (0.0036)
Inventory data for external customers	0.0115 (0.0016)	0.0217 (0.0012)	0.0595 (0.0017)
Inventory data for external suppliers	0.0244 (0.0023)	0.0484 (0.0018)	0.0926 (0.0021)
Access to vendors' products or catalogs	0.6620 (0.0068)	0.8565 (0.0029)	0.9502 (0.0016)
Ordering from vendors	0.2491 (0.0077)	0.2724 (0.0040)	0.3714 (0.0035)
Payment to vendors	0.0558 (0.0041)	0.0666 (0.0022)	0.1292 (0.0025)
Bidding	0.0776 (0.0048)	0.0816 (0.0025)	0.0833 (0.0020)
Electronic marketplaces linking specialized business buyers and suppliers	0.1862 (0.0069)	0.2090 (0.0037)	0.2846 (0.0033)
Ordering by customers	0.0640 (0.0044)	0.0919 (0.0026)	0.1639 (0.0027)
Payment by customers	0.1663 (0.0067)	0.2021 (0.0036)	0.2443 (0.0032)
Customer support	0.0686 (0.0045)	0.0678 (0.0023)	0.0724 (0.0019)
Outsourcing of research and development	0.0658 (0.0044)	0.0818 (0.0025)	0.1159 (0.0024)
Ordering of materials and supplies	0.7371 (0.0129)	0.7517 (0.0063)	0.6551 (0.0051)
Acceptance of orders for manufactured products	0.6174 (0.0154)	0.4572 (0.0077)	0.2036 (0.0045)

NOTE: [1]Standard errors in parentheses. [2]Small: 1 to 20 employees; Medium: 21 to 99 employees; Large: 100 or more employees.

Figure 6

Use of Internet to Place Orders for Materials: Adoption Rate by Plant Size

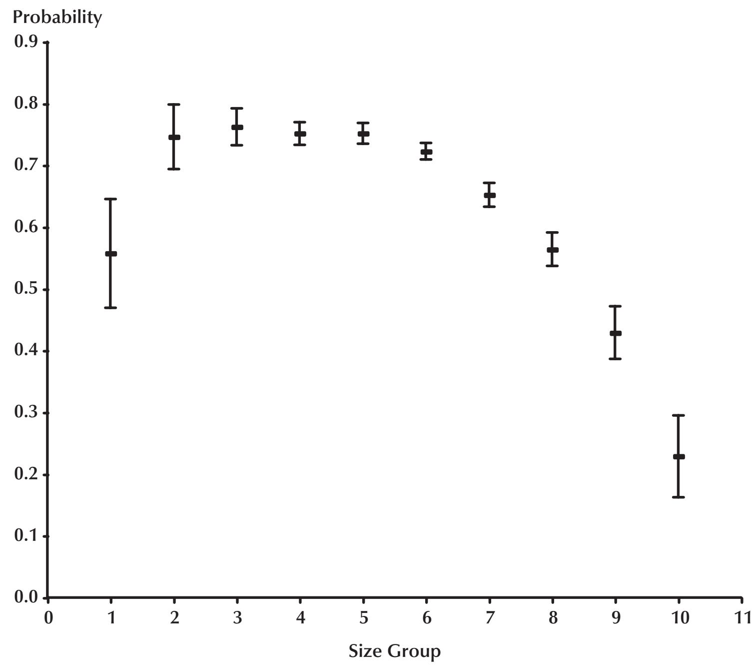


Figure 7

Use of Internet to Accept Orders: Adoption Rate by Plant Size

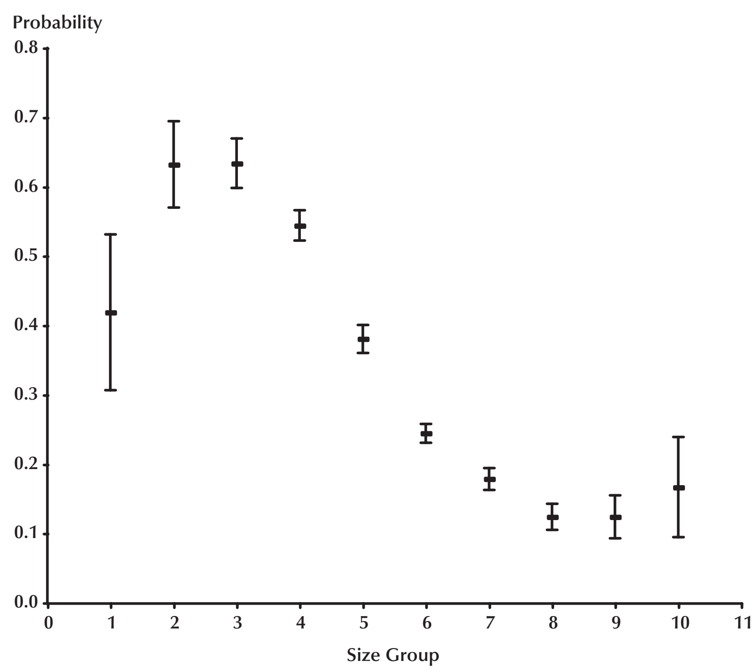


Figure 8

Use of Internet for Bidding: Adoption Rate by Plant Size

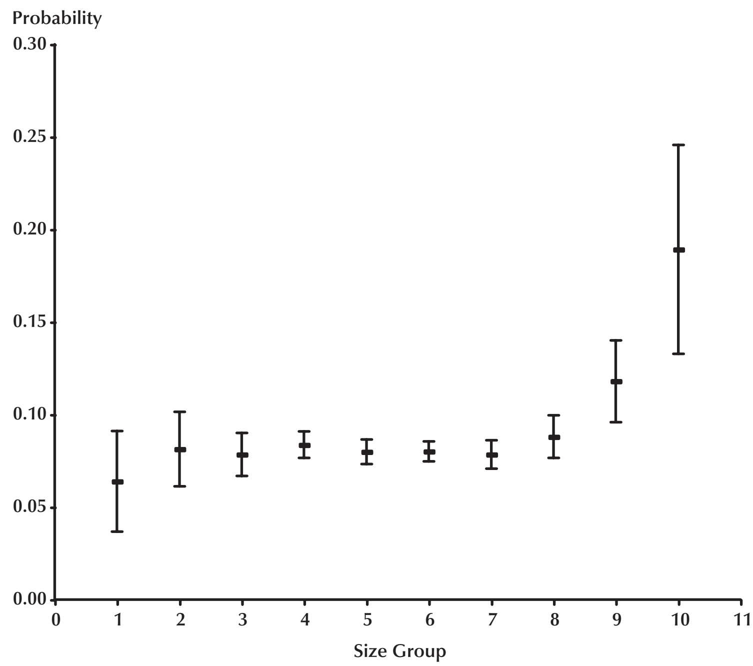
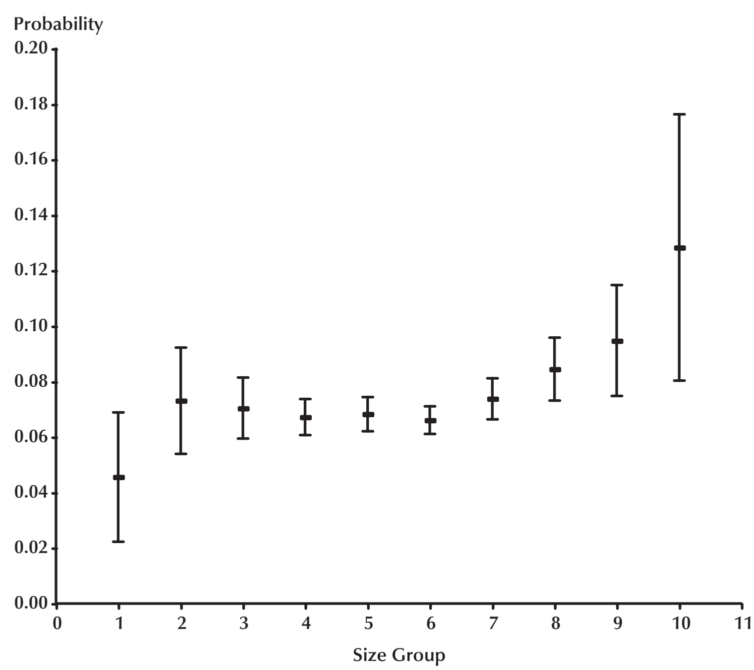


Figure 9

Use of Internet to Access Electronic Marketplaces: Adoption Rate by Plant Size



CONCLUSION

In this paper we have provided a brief account of the diffusion of e-commerce in major sectors of the economy. E-commerce appears to have followed a course of promising growth, much like other industries did in the wake of technological revolutions in the past. Both firms and consumers have learned much, and all parties are now better informed about what to expect in online markets and how to realize these expectations. However, some concerns about faster diffusion of e-commerce persist: for example, improving online security for payments and transactions and improving the quality and speed of transactions.[31]

In summary, some of the important observations presented in this paper are as follows:

- In the retail sector, we have witnessed a rapid development of the two initial phases of the e-commerce life cycle: an initial increase in the number of firms followed by the subsequent shakeout. Although the current share of retail sales from e-commerce is still low, the sector has had high growth rates recently.
- Internet retailers that can dominate the market in a certain category of products seem more capable of expanding operations into other categories, and a vast array of product varieties has proliferated in Internet markets. Patterns observed so far suggest that the variety of goods and services offered on the Internet is bound to increase.
- In the services sector, the travel industry is far ahead of other industries in share of sales accounted for by e-commerce.
- The volume of business-to-business e-commerce transactions far exceeds that of business-to-consumer e-commerce transactions. This is particularly true in the manufacturing sector, where nearly all stages of production have been affected by Internet use.
- Manufacturing industries perceived to be technologically advanced tend to rank high in the adoption of Internet-based processes used to facilitate production.
- Although the most heavily adopted processes include obvious ones (e.g., *basic internet access and degree of access* and *access to vendors' products or catalogs*), other processes initially thought to thrive on the Internet (e.g., *bidding* and *use of electronic marketplaces*) have not been widely adopted.
- Analysis of adoption rates of several Internet-based processes across plant sizes and manufacturing industries reveals that, generally, there is a positive and statistically significant relationship between adoption rates and firms' plant size.

As always, the burden of recording the effects of the ongoing technological revolution rests on the shoulders of data collectors. The steps taken so far by the U.S. Census Bureau are encouraging, but much more remains to be done.[32] In our view, the collection of data pertaining to e-commerce activity should be taken to the mainstream.[33] For instance, new survey questions can be added to the Census of Manufacturers, a quinquennial dataset collected by the Census Bureau that contains information on all active manufacturing plants, to gather detailed information on plants' various uses of the Internet. This practice would allow us to understand the importance of digital inputs in the production processes and how the intensity of usage of such inputs compares with traditional inputs of labor and capital. Any substitution among these various inputs that can take place in the medium- and long-run can then also be detected.

Furthermore, data on the intensity of the use of Internet-based processes should also be collected, rather than just information on whether a process is adopted or not. Several processes investigated in this paper can be measured in a continuous way, rather than with a discrete "adopt versus

[31] Security is still listed as one of the top concerns by consumers. See *The Economist*'s (2004) survey.

[32] Haltiwanger and Jarmin (2000) provide a good list of broad areas in which data collection efforts can be concentrated.

[33] There is also some private effort to collect extensive data, especially on prices. See www.nash-equilibrium.com for an Internet price index tracker.

not adopt" decision. For instance, one could measure the amount of orders received on the Internet versus those received by way of traditional channels. The retail trade surveys, such as the Census of Retail Trade, can be amended to include data on retail e-commerce, especially firm-level data on e-commerce sales. As mentioned earlier, one of the major drawbacks is the absence of e-commerce sales data at the firm level. If such data were collected by the Census Bureau, concentration ratios for electronic markets, as well as statistics on firm-size distribution, could be constructed. These statistics could then be used to fill the void in our understanding of how traditional and electronic markets compare in various dimensions. Existing data do not allow a satisfactory treatment of this issue, partly because comparable data across the two sectors are not easy to obtain, and most data do not provide a comprehensive coverage of one market or the other.

REFERENCES

Agarwal, Rajshree. "Evolutionary Trends of Industry Variables." *International Journal of Industrial Organization*, July 1998, *16*(4), pp. 511-25.

Agarwal, Rajshree and Gort, Michael. "First-Mover Advantage and the Speed of Competitive Entry, 1887-1986." *Journal of Law and Economics*, April 2001, *44*(1), pp. 161-77.

Barbarino, Alessandro and Jovanovic, Boyan. "Shakeouts and Market Crashes." Working paper, New York University, 2003.

Bakos, Yannis. "The Emerging Landscape for Retail E-commerce." *Journal of Economic Perspectives*, Winter 2001, *15*(1), pp. 69-80.

Baye, Michael R. and Morgan, John. "Competition in Internet Industries: Evidence from E-retailing." *OII Internet Issues Briefs No. 1.2*, November 2003.

Brynjolfsson, Erik and Smith, Michael D. "Frictionless Commerce? A Comparison of Internet and Conventional Retailers." *Management Science*, April 2000, *46*(4), pp. 563-85.

Brynjolfsson, Erik; Smith, Michael D. and Hu, Yu. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers." *Management Science*, November 2003, *49*(11), pp. 1580-96.

Carlton, Dennis W. and Chevalier, Judith A. "Free Riding and Sales Strategies for the Internet." *Journal of Industrial Economics*, December 2001, *49*(4), pp. 441-61.

Coase, Ronald. "The Nature of the Firm." *Economica*, November 1937, *4*(16), pp. 386-405.

Day, George; Fein, Adam and Ruppersberger, Gregg. "Shakeouts in Digital Markets: Lessons from B2B Exchanges." *California Management Review*, Winter 2003, *45*(2), pp. 131-50.

Dinlersoz, Emin M. and Pereira, Pedro. "On the Diffusion of Electronic Commerce." Working paper, University of Houston, April 2004.

Dinlersoz, Emin M. and Yorukoglu, Mehmet. "The Impact of Declining Information Costs in a Competitive Industry." Working paper, University of Houston, October 2003.

Dinlersoz, Emin M. and Yorukoglu, Mehmet. "Information and Industry Dynamics." Working paper, University of Houston, 2004.

Economist. "E-commerce Takes Off." May 15, 2004.

Ellison, Glenn and Ellison, Sara F. "Tax Sensitivity and Home State Preferences in Internet Purchasing." Working paper, MIT, 2003.

Fein, Adam. "Understanding Evolutionary Processes in Non-manufacturing Industries: Empirical Insights from the Shakeout in Pharmaceutical Wholesaling." *Journal of Evolutionary Economics*, September 1998, *8*(3), pp. 231-70.

Goolsbee, Austan. "In a World Without Borders: The Impact of Taxes on Internet Commerce." *Quarterly Journal of Economics*, May 2000, *115*(2), pp. 561-76.

Gort, Michael and Klepper, Steven. "Time Paths in the Diffusion of Product Innovations." *Economic Journal*, September 1982, *92*(367), pp. 630-53.

Haltiwanger, John and Jarmin, Ron S. "Measuring the Digital Economy," in Erik Brynjolfsson and Brian Kahin, eds., *Understanding the Digital Economy: Data, Tools and Research*. Cambridge, MA: MIT Press, 2000, pp. 13-33.

Jovanovic, Boyan and MacDonald, Glenn M. "The Life Cycle of a Competitive Industry." *Journal of Political Economy*, April 1994, *102*(2), pp. 322-47.

Karshenas, Massoud and Stoneman, Paul. "Rank, Stock, Order, and Epidemic Effects in the Diffusion of New Process Technologies: An Empirical Model." *Rand Journal of Economics*, Winter 1993, *24*(4), pp. 503-28.

Latcovich, Simon and Smith, Howard. "Pricing, Sunk Costs, and Market Structure Online: Evidence from Book Retailing." *Oxford Review of Economic Policy*, Summer 2001, *17*(2), pp. 217-34.

Lucking-Reiley, David and Spulber, Daniel F. "Business-to-Business Electronic Commerce." *Journal of Economic Perspectives*, Winter 2001, *15*(1), pp. 55-68.

Mazzucato, Mariana. "The PC Industry: New Economy or Early Life-Cycle?" *Review of Economic Dynamics*, April 2002, *5*(2), pp. 318-45.

Mullaney, Timothy. "E-Biz Strikes Again!" *Business Week*, May 10, 2004.

Noam, Eli "The Internet: Still Wide Open and Competitive?" *OII Issue Briefs No.1*, August 2003.

Oster, Sharon. "The Diffusion of Innovation among Steel Firms: The Basic Oxygen Furnace." *Bell Journal of Economics*, Spring 1982, *13*(1), pp. 45-56.

Rose, Nancy L. and Joskow, Paul L. "The Diffusion of New Technologies: Evidence from the Electric Utility Industry." *Rand Journal of Economics*, Autumn 1990, *21*(3), pp. 354-73.

Shaffer, Greg and Zettelmeyer, Florian. "When Good News About Your Rival Is Good for You: The Effect of Third-Party Information on the Division of Channel Profits." *Marketing Science*, Summer 2002, *21*(3), pp. 273-93.

Sommers, Paul. "The Adoption of Nuclear Power Generation." *Bell Journal of Economics*, Spring 1980, *11*(1), pp. 283-91.

Sterling, Christopher H. and Kittross, John M. *Stay Tuned: A History of American Broadcasting*. Third Edition. Mahwah, NJ: Lawrence Erlbaum, 2002, pp. 827-8.

Sutton, John. *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration*. Cambridge, MA: MIT Press, 1991.

Varian, Hal R. "Pricing Information Goods," in *Symposium on Scholarship of the New Information Environment Proceedings*. Cambridge, MA: Research Libraries Group, Harvard Law School, May 2-3, 1995.

Varian, Hal R. "Buying, Sharing and Renting Information Goods." *Journal of Industrial Economics*, December 2000, *48*(4), pp. 473-88.

Varian, Hal R. "Markets for Information Goods," in Kunio Okina and Tetsuya Inoue, eds., *Monetary Policy in a World of Knowledge-Based Growth, Quality Change, and Uncertain Measurement*. New York: Palgrave, 2001.

Varian, Hal R. "If There Was a New Economy, Why Wasn't There a New Economics?" *New York Times*, January 17, 2002.

APPENDIX

DATA

The data used in this article come from two U.S. Census Bureau reports on electronic economic activity. The first is the "E-commerce Multi-sector Report" and the second is the "E-business Process Use by Manufacturers, Final Report on Selected Processes." Both of these reports are available online at www.census.gov/estats/.

E-commerce Multi-sector Report

The data on e-commerce economic activity for the three industries we analyze are collected in three separate Census Bureau surveys. First, data on retail e-commerce sales are collected in the "2002 Annual Retail Trade Survey," a survey of more than 19,000 retailers. More recent data on retail Internet sales (such as those used in Figure 3) are available as part of a quarterly retail e-commerce series. Revenue data on selected services industries are collected in the "2002 Service Annual Survey," a survey of more than 58,000 firms. Finally, data on the value of manufacturing e-commerce shipments are collected in the "2002 Annual Survey of Manufactures," a survey of more than 55,000 manufacturing plants.

The estimates in Figure 3 are reproduced from the August 20, 2004, release, "Retail E-commerce Sales in Second Quarter 2004," produced by the Census Bureau. Estimates are not adjusted for seasonal variation, holiday or trading-day differences, or price changes. For additional details, please see www.census.gov/mrts/www/current.html.

The estimates of e-commerce shares of total sales or revenues (and their standard errors) in Tables 1, 2, and 3 are reproduced from Tables 5 and 5A, 6 and 6A, and 4 and 4A, respectively, in the "E-commerce Multi-sector Report."

E-business Process Use by Manufacturers

This report tabulates the responses of more than 38,000 manufacturing plants to 39 questions about Internet-based processes used at the plant level. These responses were collected in the "Computer Network Use Supplement" to the "1999 Annual Survey of Manufactures."

The estimates of adoption rates of Internet processes reported in Figures 5 through 9 for manufacturing plants were obtained from the authors' own calculations based on the tabulations of the "E-business Process Use by Manufacturers" report. The same tabulations were used to calculate the rates of adoption of Internet processes to rank manufacturing industries in Table 4, to rank Internet-based processes in Table 5, and to contrast the adoption rates of several processes across three aggregate manufacturing plant size classes in Table 6.

Stock Return and Interest Rate Risk at Fannie Mae and Freddie Mac

Frank A. Schmid

Fannie Mae and Freddie Mac are government-sponsored enterprises (GSEs) with the stated objective of promoting home ownership by improving the availability of mortgage financing for private households. These enterprises engage in two separate and distinct lines of business: (i) assembling and marketing pools of mortgages on which they guarantee the timely payments of principal and interest and (ii) purchasing mortgage assets for their own portfolio, mostly funded with debt securities. This article examines the sensitivity of the returns on GSEs' equity shares to realizations of interest rate risk. The study shows that the market value of Fannie Mae's and Freddie Mac's equity is vulnerable to increases in short-term interest rates and changes in the term spread (the difference between the long-term and short-term interest rates).

Federal Reserve Bank of St. Louis *Review*, January/February 2005, *87*(1), pp. 35-48.

This article examines the sensitivity to realizations of interest rate risk of the stock returns of Fannie Mae (Federal National Mortgage Association) and Freddie Mac (Federal Home Loan Mortgage Corporation). The study shows that the market value of Fannie Mae's and Freddie Mac's equity is vulnerable to increases in short-term interest rates and changes in the term spread (the difference between the long-term and short-term interest rates).

Fannie Mae and Freddie Mac are venues for pursuing the public policy objective of furthering home ownership by improving the availability of mortgage financing for medium- and low-income households. These enterprises are organized as government-sponsored enterprises (GSEs), that is, they are privately operated and funded corporations that are chartered by the federal government. Fannie Mae and Freddie Mac, which are competitors, pursue two major lines of business. First, these enterprises purchase mortgage loans, bundle them into mortgage-backed securities (MBS), and sell them to investors. The enterprises guarantee the timely payments of principal and interest on these MBS and collect a guarantee fee in return; this is effectively insurance business. The interest rate risk of these MBS resides with the investors that purchase these securities. Second, Fannie Mae and Freddie Mac purchase mortgage-related securities, including their own MBS, and retain these securities; these purchases are mostly financed with debt securities. In this line of business, Fannie Mae and Freddie Mac take on interest rate risk and, unless these assets are securities issued by Ginnie Mae (Government National Mortgage Association), credit risk.[1] Because the mortgage portfolios of these enterprises are geographically diversified and because, by definition, mortgage loans are collateralized debt, the credit risk is generally held to be small (see the Office of Federal Housing Enterprise Oversight [OFHEO], 2003).[2]

---

1 Securities issued by Ginnie Mae are backed by the full faith and credit of the U.S. government.

2 Fannie Mae and Freddie Mac may raise the credit risk of the retained mortgage portfolios by raising the loan-to-value (LTV) ratios of the mortgages they purchase; for mortgages with an LTV greater than 80 percent, these GSEs have to demand credit enhancement (see OFHEO, 2003).

Frank A. Schmid is a senior economist at the National Council on Compensation Insurance, Boca Raton, Florida. The article was written while Schmid was affiliated with the Federal Reserve Bank of St. Louis. Jason Higbee provided research assistance.

Recent controversies surrounding Fannie Mae and Freddie Mac have concerned the efficacy of subsidizing home ownership through the channel of GSEs and the incentive structure government sponsorship creates at these entities; for an overview of these controversies, see Frame and Wall (2002a) and Van Order (2000). On one hand, Fannie Mae and Freddie Mac are publicly traded corporations. On the other hand, because these enterprises operate with charters issued by the federal government, they enjoy privileges not available to other companies in the private sector. There is concern that government sponsorship generates extra income to the shareholders by establishing a barrier to entry to the market, preventing potential rivals from competing away abnormal profits at Fannie Mae and Freddie Mac (see Hermalin and Jaffee, 1996). Abnormal profits go to the shareholders, the investors that hold the residual income rights. According to a study by the Congressional Budget Office (CBO, 2004), Fannie Mae and Freddie Mac retained about a third of the subsidy that they gathered (see also Passmore, 2003). On the other hand, as pointed out by Frame and White (2004), this surplus is at risk of being eroded through competition from Federal Home Loan Banks and, due to improved risk-based capital requirements laid out in the Basel II regulatory standards, from commercial banks.

In a corporation, the shareholders hold the control rights over the allocation of the assets; this is because bundling control and residual income rights abets the internalization of the consequences of decisionmaking. But these control rights also put the shareholders in a position to behave opportunistically vis-à-vis the debt holders. Remember that the equity of a corporation is a call option on its assets (Merton, 1974). The shareholders may exercise this call by making the promised payments to the debt holders; not exercising this call would entail bankruptcy or, equivalently, the transfer of the control rights over the assets to the debt holders. All else being equal, the value of an option increases with the volatility in the value of the underlying asset; here, the underlying asset is the enterprise's asset portfolio. Put differently, the riskier the firm, the more valuable is the equity; this is why the shareholders have an incentive to behave opportunistically vis-à-vis the debt holders by taking on more risk than originally stated once the debt holders are invested.[3] The shareholders can increase the risk of the firm by choosing an asset portfolio with a greater dispersion of payoffs or by increasing its financial leverage. The Modigliani-Miller theorem implies that financial leverage has no bearing on the value of the firm.[4] Remember that the value of the firm is the market value of the assets, which equals the market value of the financial (debt and equity) claims on these assets. Hence, if the shareholders gain from increased leverage, then the debt holders lose. Anticipating the shareholders' incentives, investors, before underwriting the firm's debt, insist on collateral or restrict through covenants the shareholders' choice set.

At Fannie Mae and Freddie Mac, because of government sponsorship, traditional constraints on shareholder risk-taking do not apply. Generally, when investors underwrite corporate debt, they are buying default-free debt—effectively, government debt—and write a put option to the shareholders, giving the shareholders the right to walk away from the firm; the right to walk away is the privilege of limited liability (Merton, 1974). At Fannie Mae and Freddie Mac, the debt holders do not seem to perceive themselves as writers of put options; in fact, it appears that the debt holders assume that the government writes these options. This perceived government guarantee explains why the credit quality of Fannie Mae's and Freddie Mac's debt is close to default-free. The Federal Deposit Insurance Corporation (FDIC, 2004) states that "if investors were to disregard any implicit guarantee...GSE credit ratings would likely be lowered from the top ratings grades currently issued by major rating agencies. Based on existing studies, we assume that the ratings agencies would lower GSE credit ratings within a range of AA to A."

---

3 Here, risk is total risk, which comprises systematic risk and idiosyncratic risk.

4 If debt is tax-preferred over equity, then financial leverage indeed contributes to the value of the firm. On the other hand, there are bankruptcy costs—the difference between the going-concern value and the liquidation value of the assets accounts for much of these costs. Bankruptcy costs limit the optimal amount of financial leverage.

The evidence of market discipline provided by Seiler (2003) notwithstanding, there is reason to believe that debt holders impose no effective constraint on risk-taking at Fannie Mae and Freddie Mac; see, for instance, OFHEO (2003) and FDIC (2004).[5] Further, the convexity of the enterprises' excess stock returns in the excess market return, as evidenced in Schmid (2004), suggests that there is a conjectural guarantee for the shareholders as well. The assumed option writer—the government and, ultimately, the taxpayer—limits risk-taking at Fannie Mae and Freddie Mac through a regulator, the OFHEO.[6]

What follows is an empirical study of the sensitivity of the stock returns of Fannie Mae and Freddie Mac to draws from interest rate risk distributions. The analyzed time period is May 1991 through December 2003; most of this time window overlaps with the existence of the OFHEO, which began operations in 1993. To be parsimonious, I measure realizations of interest rate risk only in two dimensions, which are changes in the level and the slope (or term spread) of the Treasury yield curve.

The next section offers a brief discussion of retained interest rate risk or, synonymously, balance sheet risk at these enterprises. I then describe the data and the variables employed in the empirical analysis, outline the econometric method, and offer the empirical findings and conclusions. There are two appendixes; one contains information on the data sources and definitions of the variables and another describes the econometric approach.

SOURCES OF INTEREST RATE RISK AT FANNIE MAE AND FREDDIE MAC

Jaffee (2003) offers a detailed analysis of the interest rate risk that emanates from the debt-financed retained mortgage portfolios of Fannie Mae and Freddie Mac. Jaffee distinguishes two potential sources of interest rate risk. First, the cash flow of the mortgage assets over time and across interest rate environments may not match with the cash flow of the debt liabilities. Such a mismatch may arise when these GSEs finance their retained mortgage portfolios with short-term debt. Because short rates are lower than long rates most of the time, the difference between the short borrowing rate and the long lending rate is a source of income, in particular when the term structure of interest rates is strongly upward sloping. This "carry trade" may cause a duration mismatch between the mortgage portfolio and the debt liabilities that finance this portfolio.[7] If, for instance, the weighted average of the times to maturity of the cash flows of the assets is shorter than the weighted average of the times to maturity of the cash flows of the debt liabilities, then there is a negative duration gap. In such a situation, the liabilities are more sensitive to changes in interest rates than are the assets: When interest rates decline, the assets increase in value less than the liabilities and, hence, the value of the equity declines.

The second potential source of interest rate risk may originate in a mismatch of the prepayment options embedded in the mortgage portfolio and the call options embedded in the debt liabilities that finance this mortgage portfolio—this is the prepayment risk. With their retained mortgage portfolios, Fannie Mae and Freddie Mac have a long position in collateralized debt and a short position in call options on this debt; the households that take out these mortgage loans have a long position in the calls. Writing call options is a source of income: The premium of the call contributes to the yield spread between (fixed-rate) mortgages and debt securities of similar duration and default risk. When long rates fall, for instance,

5 Seiler (2003) has shown that the share prices and senior-debt yield spreads of Fannie Mae and Freddie Mac indeed respond to news concerning the enterprises' financial risk and the probability of the government guaranteeing the enterprises' debt.

6 The OFHEO was established under the Federal Housing Enterprises Financial Safety and Soundness Act of 1992.

7 Duration, also known as Macaulay's duration, is a weighted average of the times to maturity of a portfolio's scheduled cash flows. This weighted average is an elasticity that indicates the percentage change in the market value of this portfolio in response to a uniform, 1 percent change in the discount factor for all times to maturity, multiplied by –1. The discount factor for a given cash flow equals $1 + r_t$, where r_t is the interest rate for the remaining time to maturity in question, t. The concept of duration assumes that the discount factors $(1 + r_t)$ change by the same proportion for all t.

then the value of the call options increases, which subtracts from the market value of the mortgage portfolio. The GSEs can hedge their short position in calls by holding call options on their debt.

From these two sources of risk, Jaffee (2003) derives the perfect balance sheet hedge. Interest rate risk is perfectly hedged if the mortgage portfolio is financed with long-term callable debt such that the cash flow of the mortgage portfolio matches the cash flow of the debt in any interest rate environment, regardless of the amount of mortgage loans that is being prepaid.

Fannie Mae and Freddie Mac do not pursue a perfect balance sheet hedge, in part because these enterprises regard risk-taking as a line of business; as Jaffee (2003) has shown, risk-taking is highly profitable for Fannie Mae and Freddie Mac. Jaffee defines as balance sheet risk the fraction of interest rate risk that these GSEs leave unhedged. This author shows that the maturity gap and the short position in call options are significant sources of income at Fannie Mae and Freddie Mac. Jaffee also offers a detailed analysis of the hedging strategies of Fannie Mae and Freddie Mac. It is ultimately an empirical question of how much interest rate risk these GSEs retain and how much they hedge.

DATA AND DEFINITIONS OF VARIABLES

I study the sensitivity of the stock returns of Fannie Mae and Freddie Mac to (good or bad) draws of interest rate risk, as perceived by the marginal shareholder; I control for realizations of market risk (or, synonymously, systematic risk); market risk manifests itself in the covariance with the market return of the stock return of the respective enterprise. I allow these stock-return sensitivities to be time-varying.

In keeping with standard practice, I study the logarithmic excess return, that is, the logarithmic return in excess of an investment in the risk-free asset. I choose the Center for Research in Security Prices (CRSP®) value-weighted stock market index as the market portfolio and a eurodollar money market deposit as the risk-free asset. I measure shifts in the level of the yield curve by changes in the 3-month constant-maturity Treasury yield. I gauge changes in the slope (term spread) of the yield curve by changes in the difference between the constant maturity 10-year and 3-month Treasury yields. The variables are on a weekly basis (Friday through Thursday), to avoid potential autocorrelation of returns due to the weekend effect.[8] The observations run from Friday, May 31, 1991, through Thursday, December 18, 2003. The time period starts when 7-day eurodollar rates became available. For details on the definitions of the variables and the data sources, see Appendix A.

Figure 1 offers a scatter diagram of pair-wise observations of changes in the short rate and the term spread. The scatter diagram shows that these two variables are mildly negatively correlated; the correlation coefficient equals –0.43.

EMPIRICAL METHOD AND RESULTS

In analyzing the sensitivity of the stock returns of Fannie Mae and Freddie Mac to draws from the market risk and interest rate risk distributions, I start out with the nonparametric model

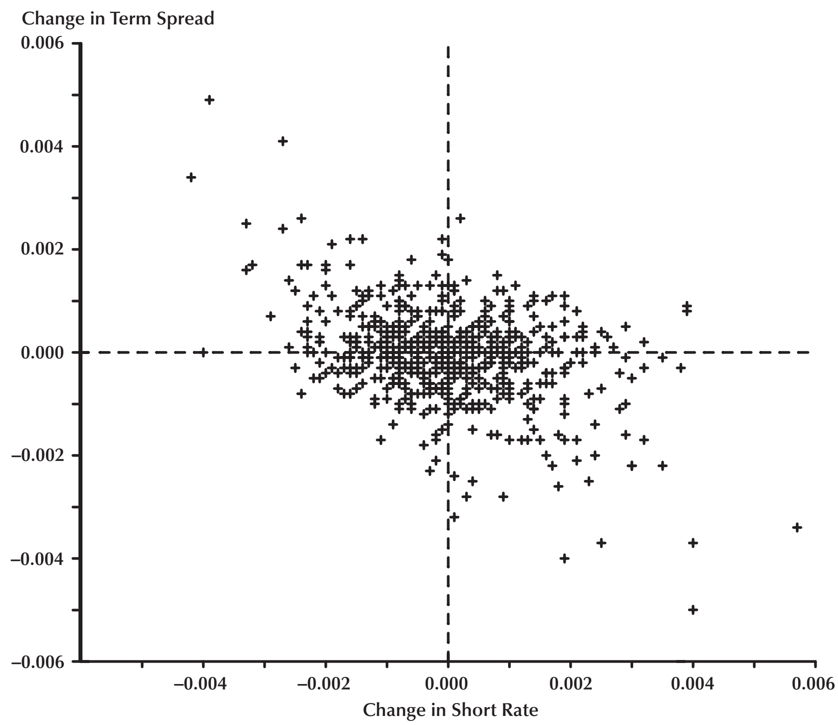
$$y_t = f(\mathbf{z}_t) + \varepsilon_t , \quad (1)$$

where y_t denotes the observation of the dependent variable at time t, the vector $\mathbf{z}_t$ comprises the observations of the explanatory variables at time t, and ε_t is an independently and normally distributed error term with mean 0 and constant, finite variance σ^2. The dependent variable is the 7-day (Friday through Thursday) logarithmic excess stock return of Fannie Mae or Freddie Mac, respectively. The explanatory variables comprise the 7-day logarithmic excess return of the market, the change of the 3-month T-bill yield, the change of the difference between the 10-year T-note and the 3-month T-bill yields during this 7-day period, and a time index. The time index measures the distance of observation t to the first observation

[8] Abraham and Ikenberry (1994) have documented that when Friday's stock return is negative, Monday's return is negative nearly 80 percent of the time, with a mean return of –0.61 percent. Also, when Friday's return is positive, the subsequent Monday's mean return is positive, averaging 0.11 percent.

Figure 1

Scatter Diagram of Interest Rate Risk Realizations



in the studied time period, measured in number of weeks elapsed, plus 1. The functional form $f(\bullet)$ accommodates an intercept. For details on the econometric method, see Appendix B.

I estimate model (1) using the multivariate smoother LOESS (locally weighted regression) as developed by Cleveland and Devlin (1988) and Cleveland, Devlin, and Grosse (1988); for details on the econometric method, see Appendix B. Table 1 offers an analysis of variance for restrictions imposed on model (1) (for details on the calculation of the test statistic, see Appendix B). The first row ("Market") shows that the excess stock returns of Fannie Mae and Freddie Mac covary in a statistically significant manner with the market excess return, as expected. Further, these excess stock returns vary in a statistically significant manner with changes in the short rate and the term spread; these variables are statistically significant individually and as a group (joint test). In conclusion, I can reject the hypotheses that the stock market perceives Fannie Mae's and Freddie Mac's interest rate risks as perfectly hedged.

Table 1 also offers an analysis of variance for the restriction that the influences on the GSEs' excess returns of realizations of market risk (on one hand) and interest rate risk (on the other hand) are additive. Imposing such a restriction on model (1) leads to the following generalized additive model:

(2) $$y_t = f_1(x_t) + f_2(\tilde{\mathbf{z}}_t) + \varepsilon_t .$$

In model (2), the component $f_1(\bullet)$ captures the influence of the log market excess return and the component $f_2(\bullet)$ subsumes the influences of the changes in the short rate and the term spread; both x_t and $\tilde{\mathbf{z}}_t$ include the time index, and both components $f_1(\bullet)$ and $f_2(\bullet)$ provide for an intercept. The test statistic in Table 1 ("Generalized additive model") does not reject the hypothesis that the influences of market risk and interest rate risk are additive; in what follows, this restriction is imposed.

Table 1

Analysis of Variance

	Fannie Mae DDF: 532		Freddie Mac DDF: 500	
	NDF	*F*-statistic	NDF	*F*-statistic
Market	49	4.460*	65	3.579*
Short rate	46	2.848*	61	2.237*
Term spread	54	2.100*	71	1.801*
Time index	29	2.025*	40	1.996*
Short rate and term spread (joint test)	84	2.300*	113	1.884*
Nonconstant explanatory variables (joint test)	116	4.472*	161	3.352*
Generalized additive model	1	0.506	2	0.051

NOTE: *Indicates significance at the 1 percent level. DDF (NDF): Denominator (numerator) degrees of freedom. Number of observations: 615.

Figures 2 (Fannie Mae) and 3 (Freddie Mac) offer quantitative estimates for the sensitivity of the GSEs' excess returns to draws from the interest rate risk distributions as obtained from the component $f_2(\bullet)$ of the generalized additive model (2). The estimates shown in these figures are presented in conditioning plots, as introduced by Cleveland and Devlin (1988). Such plots display the estimated (partial) impact of a selected explanatory variable with the other explanatory variables pegged at chosen values. Because the intercept is not identified in this type of regression, only *changes* in the displayed partial impact (rather than the level itself) can be interpreted in an economically meaningful manner. The variable that varies in a given conditioning plot adopts only values observed in the neighborhood of the pegged explanatory variables. Specifically, when I peg a variable to its median negative (positive) value, only observations for which this variable adopts nonpositive (nonnegative) values are included in the conditioning plot. Similarly, when I peg a variable (such as the changes in the short rate or term spread) at zero, only observations for which this variable lies within the closed interval of the median negative and the median positive values are included in the conditioning plot. A similar principle applies to the time index. (For the time index, substitute 25th percentile for median negative value, median for zero, and 75th percentile for median positive value.) At the bottom of each figure there is a frequency distribution for the variable that varies along the horizontal axis.

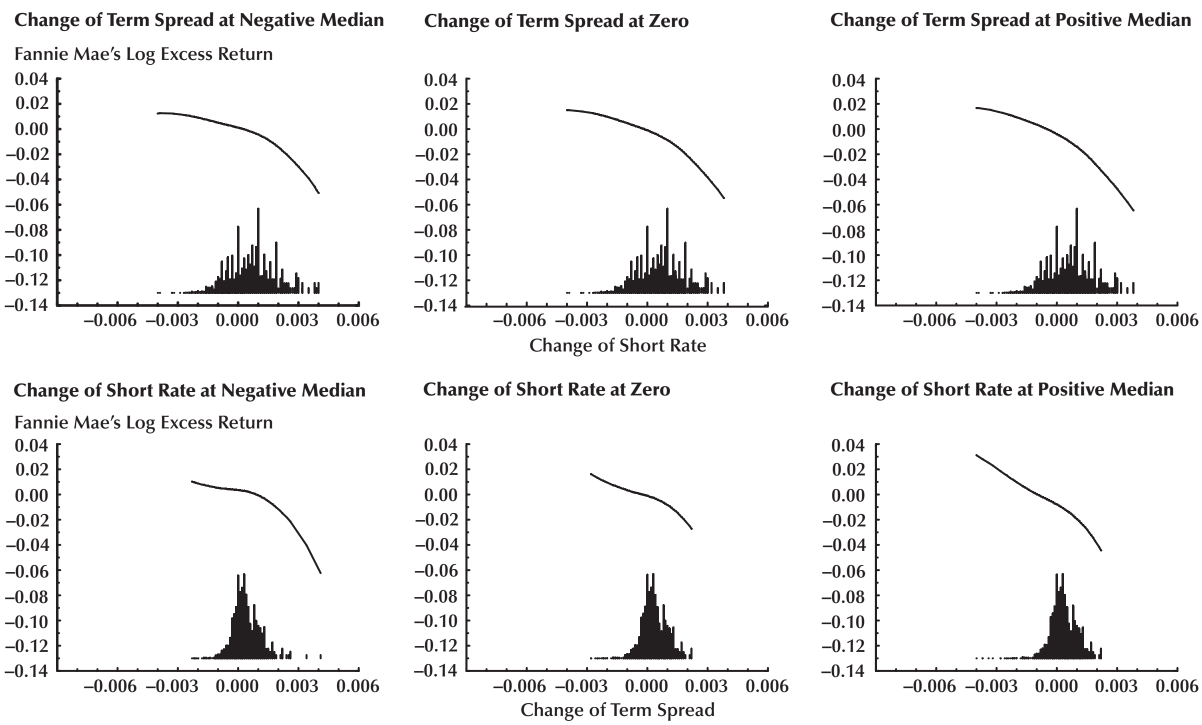
The top rows of each panel (A through C) of Figures 2 (Fannie Mae) and 3 (Freddie Mac) display, for three different values of changes in the term spread, the influence of changes in the short rate on the log excess stock returns. In the leftmost plot of the top row, the change of the term spread is pegged at the median negative value; in the center plot, the change is zero; and in the rightmost plot, this variable is kept at the median positive value. The plots show that the log excess returns of Fannie Mae and Freddie Mac are negatively related to changes in the short rate and that this relation is convex. In Panel A, the time index is pegged at its 25th percentile, which is early October 1994; in Panels B and C, the time index is pegged at its median (mid-February 1998) and its 75th percentile (early July 2001), respectively. Late in the studied time period (Panel C), assuming that there is no change to the term spread, a drop in the short rate by 26 basis points (or 0.0026) boosts the excess return of Fannie Mae by 130 basis points; conversely, a rise in the short rate by 30 basis points depresses this return by 231 basis points.[9] For Freddie Mac, the respective numbers read 105 and 215 basis points.

9 The scatter diagram of Figure 1 shows observations of a 26-basis-point decrease and a 30-basis-point increase in the short rate that come with no or only minute changes in the term spread.

Figure 2

Interest Rate Sensitivity of Fannie Mae's Stock Return

A. Early Period



B. Mid-Period

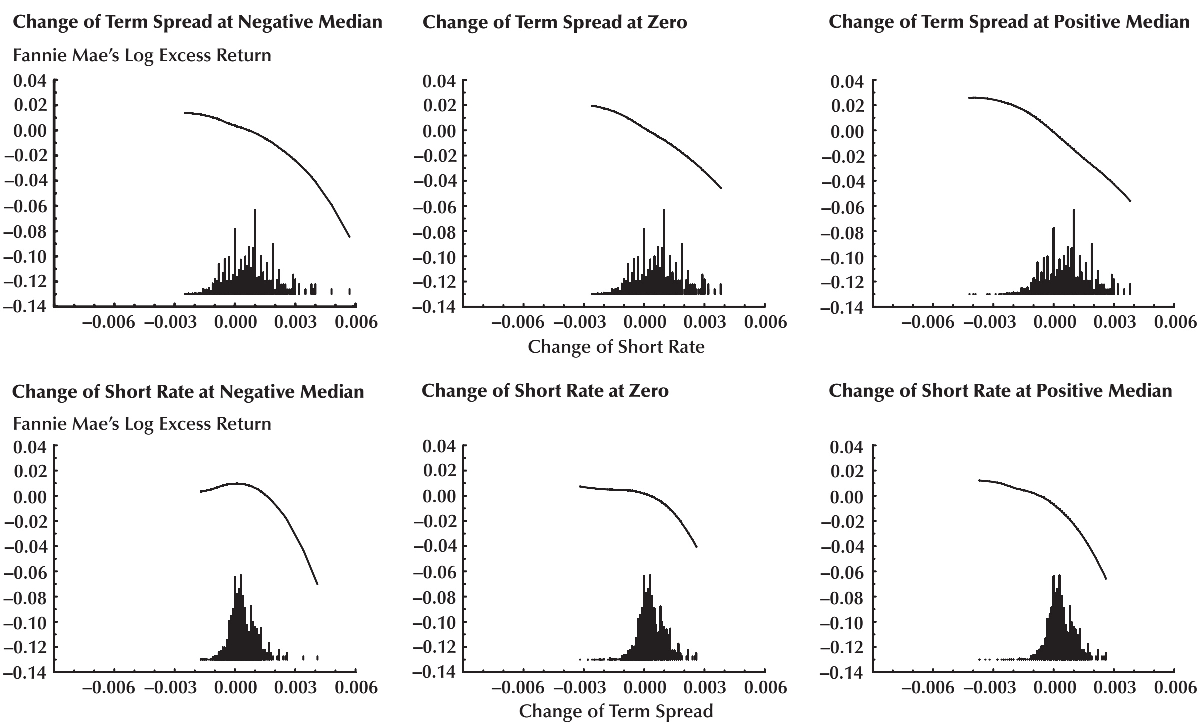
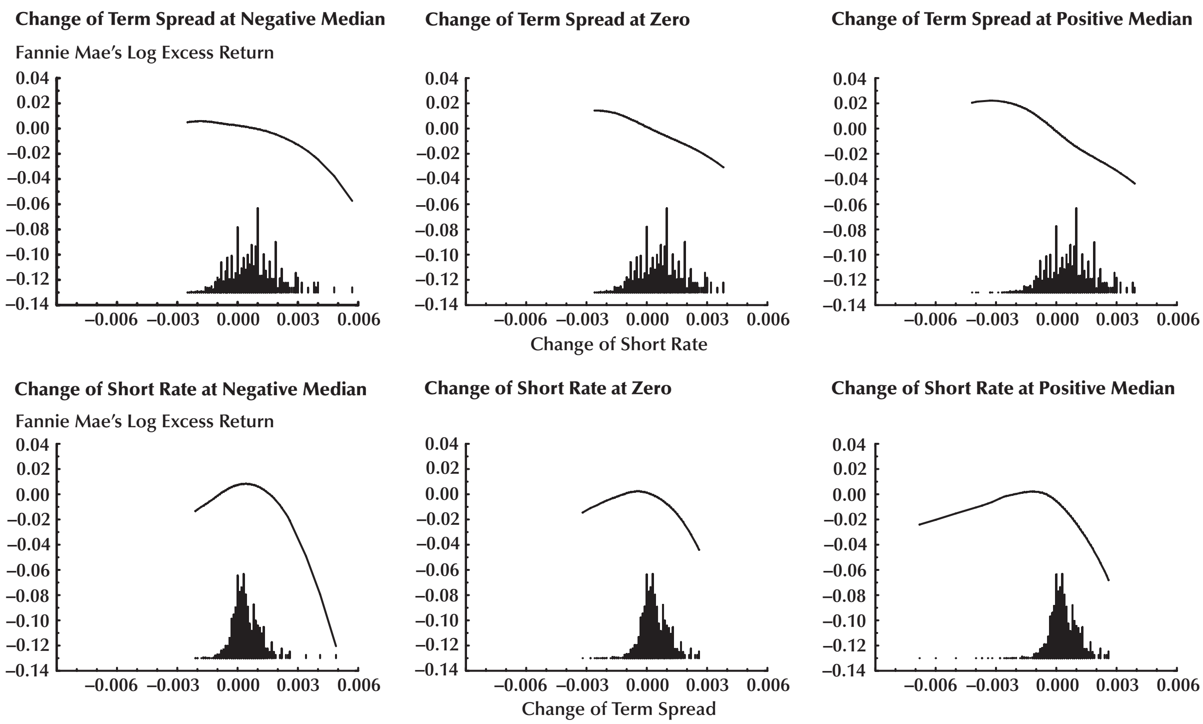


Figure 2 (cont'd)

Interest Rate Sensitivity of Fannie Mae's Stock Return
C. Late Period



The bottom rows of each panel (A through C) of Figures 2 (Fannie Mae) and 3 (Freddie Mac) display, for three different values of changes in the short rate, the influence of changes in the term spread on the log excess stock return. In the leftmost plot of the bottom row, the change of the short rate is pegged at the median negative value; in the center plot, this change is zero; and in the rightmost plot, it is kept at the median positive value. The plots show that the log excess returns of Fannie Mae (late in the studied time period) and Freddie Mac (for the entire studied time period) are "hump-shaped" in changes in the term spread; early in the analyzed time period, the excess return of Fannie Mae is negatively related to changes in the term spread. Late in the studied period (Panel C), assuming no change in the short rate, a drop in the term spread by 32 basis points (or 0.0032) depresses the excess return of Fannie Mae by 158 basis points; correspondingly, a rise in the short rate by 26 basis points depresses this return by 454 basis points.[10] For Freddie Mac, the respective numbers read 189 and 263 basis points.

CONCLUSION

This article offers an empirical analysis of the sensitivity of Fannie Mae's and Freddie Mac's excess stock returns to draws from interest rate risk distributions. The empirical approach allows this sensitivity to be nonlinear and to vary with time, possibly in nonlinear ways. The analysis shows little time variation in the sensitivity of these GSEs' stock returns to changes in the short-term interest rate. This also holds for Freddie Mac's sensitivity to changes in the term spread. But unlike Freddie Mac, Fannie Mae shows, over time,

[10] The scatter diagram of Figure 1 shows observations of a 32-basis-point decrease and a 26-basis-point increase in the short rate that come with no or only minute changes in the term spread.

Figure 3

Interest Rate Sensitivity of Freddie Mac's Stock Return

A. Early Period

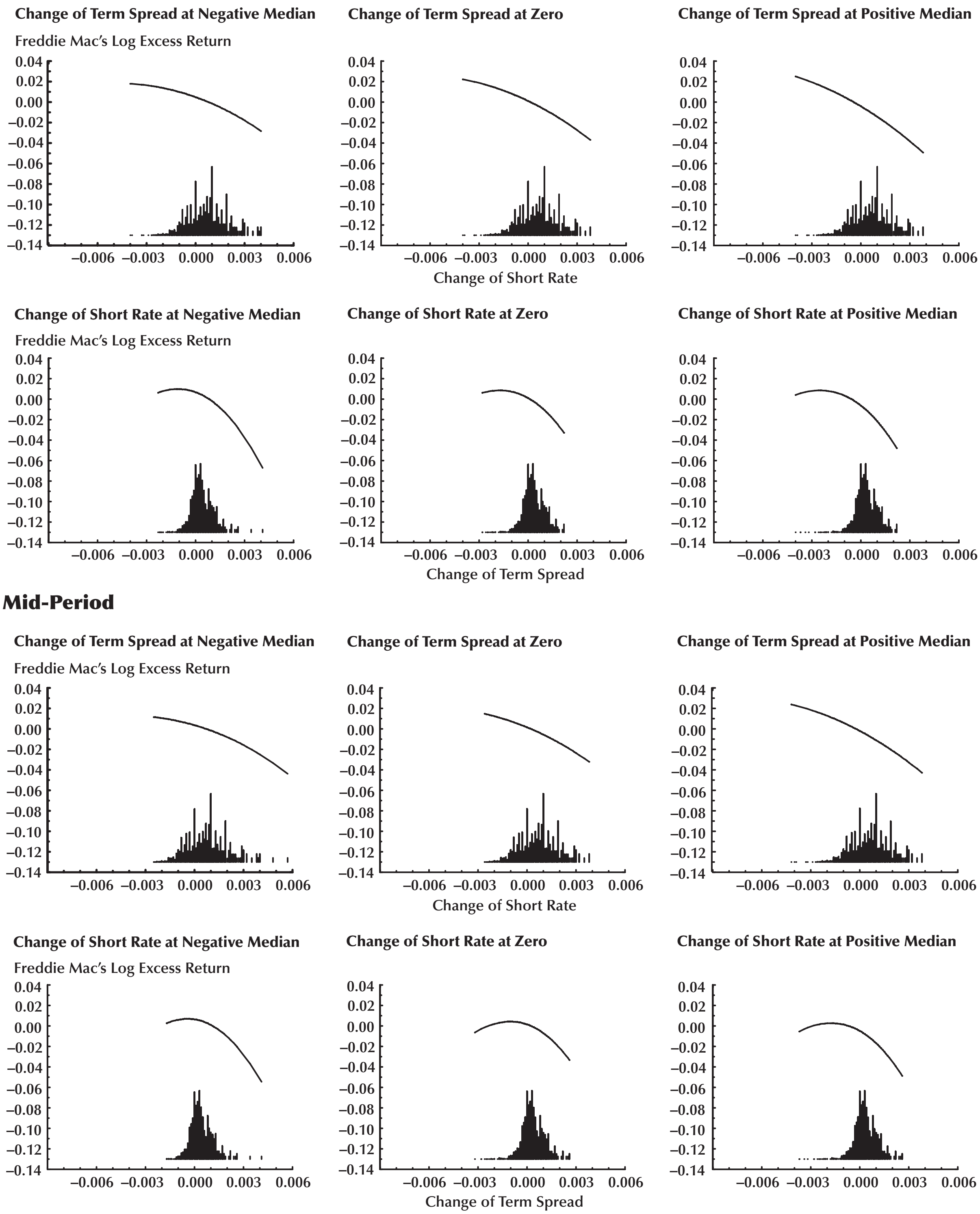
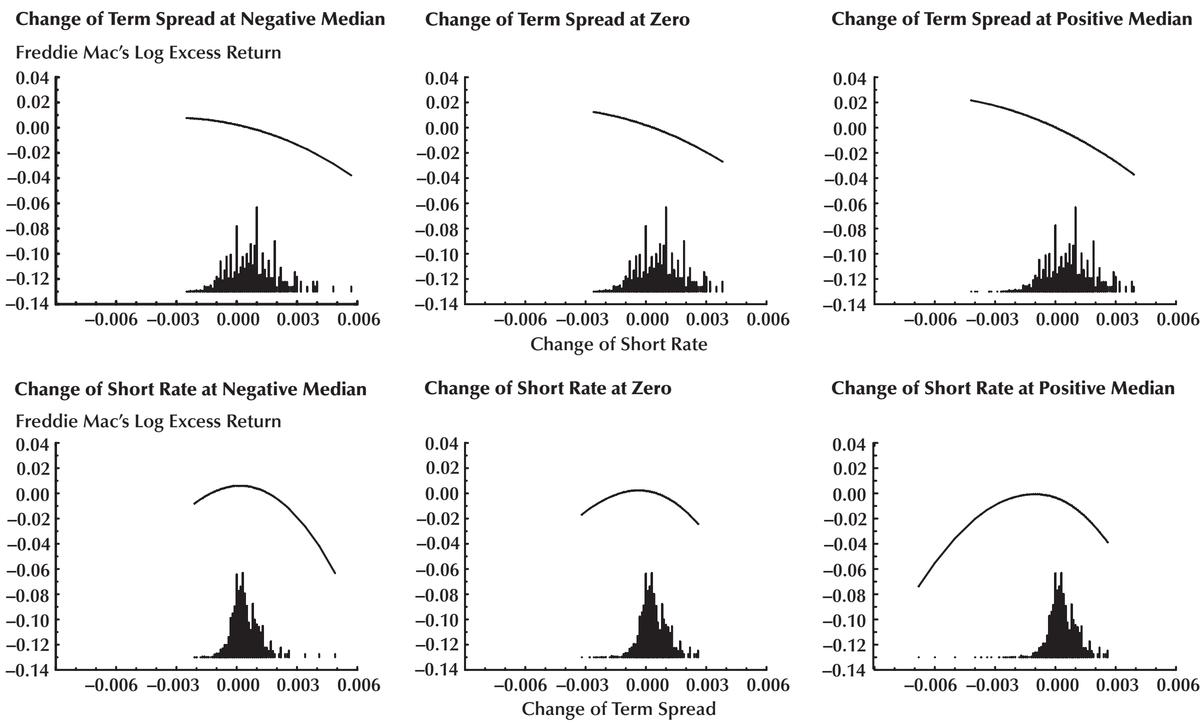


Figure 3 (cont'd)

Interest Rate Sensitivity of Freddie Mac's Stock Return
C. Late Period



a marked change in its stock return sensitivity to changes in the slope of the yield curve. Early in the studied time period, Fannie Mae's stock return varied negatively with the term spread; later, this sensitivity adopted the hump-shaped relation that characterizes Freddie Mac's stock return sensitivity over the entire analyzed time period. Note that the measured interest rate sensitivities are responses above and beyond the variation of the stock return with the market return; this market return itself may be sensitive to interest rate risk. An analysis of variance shows that these two influences—realizations of market (systematic) risk (on one hand) and interest rate risk (on the other hand)—are additive.

It is ultimately a matter of judgment as to whether the measured interest rate sensitivities are considered large (enough to be concerned) or small. Also, remember that the measured sensitivities are "as perceived by the marginal investor."

Information in the public domain on Fannie Mae and Freddie Mac has varied over the studied time period. Following six voluntary initiatives announced in October 2000, the GSEs enhanced the disclosure of their interest rate risk: The enterprises started publishing scenario-based risk measures and the duration gap on a monthly basis. But as Frame and Wall (2002b) point out, the disclosure of interest rate risk is not "as comprehensive as would be desirable."

REFERENCES

Abraham, Abraham and Ikenberry, David L. "The Individual Investor and the Weekend Effect." *Journal of Financial and Quantitative Analysis*, June 1994, *29*(2), pp. 263-77.

Andrews, Donald W.K. "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic

Errors." *Journal of Econometrics*, February/March 1991, *47*(2/3), pp. 359-77.

Cleveland, William S. and Devlin, Susan J. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association*, September 1988, *83*(403), pp. 596-610.

Cleveland, William S.; Devlin, Susan J. and Grosse, Eric. "Regression by Local Fitting: Methods, Properties and Computational Algorithms." *Journal of Econometrics*, January 1998, *37*(1), pp. 87-114.

Congressional Budget Office. "Updated Estimates of the Subsidies to the Housing GSEs." April 8, 2004, ftp://ftp.cbo.gov/53xx/doc5368/04-08-GSE.pdf.

Federal Deposit Insurance Corporation. "Assessing the Banking Industry's Exposure to an Implicit Government Guarantee of GSEs." March 1, 2004, www.fdic.gov/bank/analytical/fyi/2004/030104fyi.html.

Frame, W. Scott and Wall, Larry D. "Financing Housing through Government-Sponsored Enterprises." Federal Reserve Bank of Atlanta *Economic Review*, First Quarter 2002a, *87*(1), pp. 29-43.

Frame, W. Scott and Wall, Larry D. "Fannie Mae's and Freddie Mac's Voluntary Initiatives: Lessons from Banking," Federal Reserve Bank of Atlanta *Economic Review*, First Quarter 2002b, *87*(1), pp. 45-59.

Frame, W. Scott and White, Lawrence J. "Emerging Competition and Risk-Taking Incentives at Fannie Mae and Freddie Mac." Working Paper No. 2004-4, Federal Reserve Bank of Atlanta, February 2004, www.frbatlanta.org/filelegacydocs/wp0404.pdf.

Hastie, Trevor J. and Tibshirani, Robert J. "Generalized Additive Models" (with Discussion). *Statistical Science*, August 1986, *1*(3), pp. 297-318.

Hastie, Trevor J. and Tibshirani, Robert J. *Generalized Additive Models*. New York: Chapman and Hall, 1990.

Hermalin, Benjamin and Jaffee, Dwight. "The Privatization of Fannie Mae and Freddie Mac: Implications for Mortgage Industry Structure," in *Studies on Privatizing Fannie Mae and Freddie Mac*. Washington, DC: U.S. Department of Housing and Urban Development, May 1996.

Jaffee, Dwight "The Interest Rate Risk of Fannie Mae and Freddie Mac." *Journal of Financial Services Research*, August 2003, *24*(1), pp. 5-29.

Li, Ker-Chau. "Asymptotic Optimality for C_P, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set." *Annals of Statistics*, September 1987, *15*(3), pp. 958-75.

Merton, Robert C. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance*, May 1974, *29*(2), pp. 449-70.

Office of Federal Housing Enterprise Oversight. *Systemic Risk: Fannie Mae, Freddie Mac and the Role of OFHEO*, February 2003, www.ofheo.gov/Media/Archive/docs/reports/sysrisk.pdf.

Passmore, Wayne. "The GSE Implicit Subsidy and the Value of Government Ambiguity." Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2003-64, December 2003, www.federalreserve.gov/pubs/feds/2003/200364/200364pap.pdf.

Schmid, Frank A. "Conjectural Guarantees Loom Large: Evidence from the Stock Return of Fannie Mae and Freddie Mac," September 2004, http://frankschmid.com/fanfred.pdf.

Seiler, Robert S. "Market Discipline of Fannie Mae and Freddie Mac: How Do Share Prices and Debt Yield Spread Respond to New Information?" Working Paper 03-4, Office of Federal Housing Enterprise Oversight, December 2003, www.ofheo.gov/media/pdf/workingpaper034.pdf.

Van Order, Robert. "A Microeconomic Analysis of Fannie Mae and Freddie Mac." *Regulation*, Summer 2000, *23*(2), p. 27-33, www.cato.org/pubs/regulation/regv23n2/vanorder.pdf.

APPENDIX A

DATA SOURCES AND DEFINITIONS OF VARIABLES

Data Sources

The stock returns are from CRSP®, Center for Research in Security Prices, Graduate School of Business, The University of Chicago, http://crsp.uchicago.edu. The CRSP® data are used with permission, all rights reserved. The constant-maturity 3-month and 10-year Treasury yields are from the Board of Governors of the Federal Reserve System. The eurodollar rates are from Bloomberg LP. All data are published daily. The variables are on a 7-day basis (Friday through Thursday). The observations run from Friday, May 31, 1991, through Thursday, December 18, 2003. The time period starts when 7-day eurodollar rates became available.

Definition of the Dependent Variable

Fannie Mae (Freddie Mac) log excess return: sum of daily logarithmic total stock returns of Fannie Mae (Freddie Mac) from Friday through Thursday, minus the logarithmic return on a 1-week eurodollar investment undertaken at the beginning of the 7-day investment period (Thursday, close of business). Total stock return assumes daily reinvestment of dividends and capital gains.

Definition of Explanatory Variables

Market log excess return: sum of daily logarithmic total stock returns of the CRSP® value-weighted stock market index from Friday through Thursday, minus the logarithmic return on a 1-week eurodollar investment undertaken at the beginning of the 7-day investment period (Thursday, close of business). Total stock return assumes daily reinvestment of dividends and capital gains.

Change of short rate: constant-maturity 3-month Treasury bill yield as of Thursday at close of business minus the yield observed seven days earlier.

Change of term spread: term spread as of Thursday at close of business minus the spread observed seven days earlier. The term spread is the difference between the constant-maturity yields of the 10-year Treasury note and the 3-month Treasury bill.

Time index: distance of the observation in question to the first observation in the studied time period, measured in number of weeks elapsed, plus 1.

APPENDIX B

ECONOMETRIC METHODOLOGY

I estimate the nonparametric model

(B1) $$y_t = f(\mathbf{z}_t) + \varepsilon_t,$$

where y_t denotes the observation of the dependent variable at time t, the vector $\mathbf{z}_t$ comprises the observations of the explanatory variables at time t, and ε_t is an independently and normally distributed error term with mean 0 and constant, finite variance σ^2. The dependent variable is the 7-day (Friday through Thursday) logarithmic excess stock return of Fannie Mae and Freddie Mac. The explanatory variables comprise the logarithmic excess return of the market, the change of the 3-month T-bill yield, the corresponding change of the Treasury term spread, and a time index; all variables are measured over the same 7-day time period. The term spread is defined as the difference between the constant-maturity 10-year T-note and the 3-month T-bill yields. The function $f(\bullet)$ allows for an intercept. The variables and the data sources are detailed in Appendix A.

I estimate model (B1) using the multivariate smoother LOESS (locally weighted regression) as developed by Cleveland and Devlin (1988) and Cleveland, Devlin, and Grosse (1988). LOESS estimates the functional form in each observation by defining a neighborhood comprising the fraction g of the data points in the population; this fraction of data points is called the smoothing parameter. The data points to be included in the neighborhood are selected and weighted based on their respective Euclidean distance to the observation in question. I employ a tri-cube weight function, as detailed in Cleveland and Devlin.

LOESS smoothes the vector of observations of the dependent variable vector, $\mathbf{y}$, on the matrix of observations of the explanatory variables, $\mathbf{Z}$. The resulting smoother matrix, $\mathbf{S}$, establishes a linear relationship between $\mathbf{y}$ and the estimate $\hat{\mathbf{y}}$:

(B2)
$$\hat{\mathbf{y}} = \mathbf{S} \cdot \mathbf{y}.$$

A restricted version of regression model (B1) is the generalized additive model

(B3)
$$y_t = f_1(\mathbf{x}_t) + f_2(\tilde{\mathbf{z}}_t) + \varepsilon_t,$$

where $\mathbf{x}_t$ comprises market log excess return and the time index and $\tilde{\mathbf{z}}_t$ comprises all explanatory variables included in $\mathbf{z}_t$ as defined in equation (B1), except for the market log excess return. Both $f_1(\bullet)$ and $f_2(\bullet)$ provide for an intercept.

I estimate model (B3) using the backfitting algorithm suggested by Hastie and Tibshirani (1986) (see also Hastie and Tibshirani, 1990). Backfitting consists of alternating the steps

(B4a)
$$\mathbf{f}_1^{(m)} = \mathbf{S}_1 \cdot (\mathbf{y} - \mathbf{f}_2^{(m-1)})$$

(B4b)
$$\mathbf{f}_2^{(m)} = \mathbf{S}_2 \cdot (\mathbf{y} - \mathbf{f}_1^{(m)}),$$

where $m \geq 1$ indicates the stage of the iteration procedure and $\mathbf{S}_1$ and $\mathbf{S}_2$ are the corresponding LOESS smoother matrices for the partial influences of $\mathbf{X}$ and $\mathbf{Z}$, respectively. I start out by smoothing $\mathbf{y}$ on $\mathbf{X}$ (and a vector of ones). The smoothing delivers fitted values for $\mathbf{y}$, $\mathbf{f}_1^{(0)}$. I subtract $\mathbf{f}_1^{(0)}$ from $\mathbf{y}$ and smooth this difference on $\tilde{\mathbf{Z}}$ (which includes a vector of ones), resulting in $\mathbf{f}_2^{(1)}$. I keep alternating the steps (B4a,b) until the vectors of fitted values, $\mathbf{f}_1^{(m)}$ and $\mathbf{f}_2^{(m)}$, stop changing. For the smoother matrix, I can write

(B5)
$$\hat{\mathbf{y}} = \mathbf{S} \cdot \mathbf{y} \equiv \{\mathbf{I} - (\mathbf{I} - \mathbf{S}_2)(\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I} - \mathbf{S}_1)\} \cdot \mathbf{y},$$

where $\mathbf{I}$ is the identity matrix (see Hastie and Tibshirani, 1986, p. 120).

Following Cleveland and Devlin (1988), the F-statistic for testing the statistical significance of the restriction that model (B3) imposes on model (B1)—under the assumptions of normality and the unrestricted model (B1) offering an unbiased estimate of the dependent variable—reads

(B6)
$$\hat{F} = \frac{(\mathbf{y}'\mathbf{R}_L\mathbf{y} - \mathbf{y}'\mathbf{R}_S\mathbf{y}) / \nu_1}{(\mathbf{y}'\mathbf{R}_S\mathbf{y}) / \delta_1},$$

where $\mathbf{R}_L \equiv (\mathbf{I} - \mathbf{L}) \cdot (\mathbf{I} - \mathbf{L})'$, $\mathbf{R}_S \equiv (\mathbf{I} - \mathbf{S}) \cdot (\mathbf{I} - \mathbf{S})'$, $\nu_1 = \text{tr}(\mathbf{R}_L - \mathbf{R}_S)$, and $\delta_1 = \text{tr}\mathbf{R}_S$. The test statistic $\hat{F}$ is approximated by an F-distribution with ν_1^2/ν_2 numerator and δ_1^2/δ_2 denominator degrees of freedom, where $\nu_2 = \text{tr}[(\mathbf{R}_L - \mathbf{R}_S)^2]$ and $\delta_2 = \text{tr}(\mathbf{R}_S^2)$.

I use the *M*-plot method to determine the optimal smoothing parameter, g. *M*-plots, which were suggested by Cleveland and Devlin (1988), offer a graphical portrayal of the tradeoff between the contributions of variance and bias to the mean squared error as the smoothing parameter changes. The expected mean squared error summed over all observations and normalized by the variance, σ^2, reads

(B7) $$M_g = \frac{E(\mathbf{y}'\mathbf{R}_g\mathbf{y})}{\sigma^2},$$

where the subscript g indicates the chosen smoothing parameter. For a sufficiently small smoothing parameter—let us say f—the bias of the vector of the fitted values, $\hat{\mathbf{y}}$, is negligible, resulting in a nearly unbiased estimate of σ^2. In this case then, M_g can be estimated by

(B8a) $$\hat{M}_g = \hat{B}_g + V_g,$$

where

(B8b) $$\hat{B}_g = \frac{\mathbf{y}'\mathbf{R}_g\mathbf{y}}{\hat{\sigma}_f} - \mathrm{tr}[(\mathbf{I}-\mathbf{S}_g)'(\mathbf{I}-\mathbf{S}_g)],$$

(B8c) $$V_g = \mathrm{tr}(\mathbf{S}_g{}'\mathbf{S}_g).$$

$\hat{B}_g$ is the contribution of bias to the estimated mean squared error, and V_g is the contribution of variance. Cleveland and Devlin show that $\hat{M}_g$ can be implemented as

(B9) $$\hat{M}_g = v_1\frac{(\mathbf{y}'\mathbf{R}_g\mathbf{y} - \mathbf{y}'\mathbf{R}_f\mathbf{y})/v_1}{(\mathbf{y}'\mathbf{R}_f\mathbf{y})/\delta_1} + \delta_1 - T + 2\mathrm{tr}\,\mathbf{S}_g = v_1\hat{F} + \delta_1 - T + 2\mathrm{tr}\,\mathbf{S}_f,$$

where $\mathbf{y}'\mathbf{R}_f\mathbf{y}$ is the residual sum of squares when the smoothing parameter is f. Because there is an approximate *F*-distribution for $\hat{F}$, as mentioned, one can derive a probability distribution for $\hat{M}_g$. Cleveland and Devlin argue that the smoothing parameter f, for which the bias of the fitted values is negligible, is "usually in the range of .2 to .4"; I chose $f = 0.3$.

The *M*-plots (not shown) indicate that the largest smoothing parameter for which model (B1) delivers unbiased estimates is $g = 0.5$ for Fannie Mae and $g = 0.35$ for Freddie Mac. An analysis of variance does not reject the generalized additive model (B3); hence, (B3) is the model of choice.

The regression results for the generalized additive model exhibited in Figures 2 and 3 rest on cross-validated smoothing parameters; I use (delete-one) cross-validation, as discussed in Li (1987) and Andrews (1991).

In cross-validation, the following loss function—the cross-validation sum of squares—is minimized (Andrews, 1991):

(B10) $$\frac{(\mathbf{y}-\tilde{\mathbf{S}}\cdot\mathbf{y})'(\mathbf{y}-\tilde{\mathbf{S}}\cdot\mathbf{y})}{T-1},$$

where T is the number of observations. The matrix $\tilde{\mathbf{S}}$ results from the smoother matrix $\mathbf{S}$ by setting the principal diagonal elements of $\mathbf{S}$ equal to zero. The cross-validated smoothing parameters for the generalized additive model (B3) read $g = 0.8$ for Fannie Mae and $g = 1$ for Freddie Mac.

Controlling for Heterogeneity in Gravity Models of Trade and Integration

I-Hui Cheng and Howard J. Wall

This paper compares various specifications of the gravity model of trade as nested versions of a general specification that uses bilateral country-pair fixed effects to control for heterogeneity. For each specification, we show that the atheoretical restrictions used to obtain them from the general model are not supported statistically. Because the gravity model has become the "workhorse" baseline model for estimating the effects of international integration, this has important empirical implications. In particular, we show that, unless heterogeneity is accounted for correctly, gravity models can greatly overestimate the effects of integration on the volume of trade.

Federal Reserve Bank of St. Louis *Review*, January/February 2005, *87*(1), pp. 49-63.

In the 1860s, H. Carey first applied Newtonian physics to the study of human behavior, and the so-called "gravity equation" has since been widely used in the social sciences. In economics, gravity model studies have achieved empirical success in explaining various types of inter-regional and international flows (including labor migration, commuting, customers, hospital patients, and international trade). The gravity model of international trade was developed independently by Tinbergen (1962) and Pöyhönen (1963). In its basic form, the amount of trade between countries is assumed to be increasing in their sizes, as measured by their national incomes, and decreasing in the cost of transportation between them, as measured by the distance between their economic centers.[1] Following this work, Linnemann (1966) included population as an additional measure of country size, employing what we will call the augmented gravity model.[2] It has also been common to instead specify the augmented model using per capita income, which captures the same effects.[3] Whichever specification of the augmented model is used, the purpose is to allow for non-homothetic preferences in the importing country and to proxy for the capital/labor ratio in the exporting country (Bergstrand, 1989).

The gravity model of trade has been used widely as a baseline model for estimating the impact of a variety of policy issues, including regional trading groups, currency unions, political blocs, patent rights, and various trade distortions.[4] Typically, these events and policies are modeled as deviations from the volume of trade predicted by the baseline gravity model and, in the case of regional integration, are captured by dummy vari-

[1] For examples see McCallum (1995), Helliwell (1996), and Boisso and Ferrantino (1997).

[2] For uses of the augmented gravity model, see Oguledo and MacPhee (1994), Boisso and Ferrantino (1997), and Bayoumi and Eichengreen (1997).

[3] Examples of the augmented model with per capita income include Sanso, Cuairan, and Sanz (1993), Frankel and Wei (1998), Frankel, Stein, and Wei (1995, 1998), and Eichengreen and Irwin (1998).

[4] See Aitken (1973), Brada and Mendez (1983), Bikker (1987), Sanso, Cuairan, and Sanz (1993), McCallum (1995), Helliwell (1996), Frankel (1997), Wei and Frankel (1997), Bayoumi and Eichengreen (1997), Mátyás (1997), Frankel and Wei (1998), Frankel, Stein, and Wei (1998), Smith (1999), and Rose (2000).

I-Hui Cheng is an assistant professor at the National University of Kaohsuing, Taiwan. Howard J. Wall is an assistant vice president at the Federal Reserve Bank of St. Louis. The authors thank Martin Sola, Ron Smith, Jim Dunlevy, and Rob Dittmar for their comments and suggestions. Kristie M. Engemann provided research assistance.

ables. The continuing popularity of the gravity model is highlighted by Eichengreen and Irwin (1998, p. 33), who call it the "workhorse for empirical studies of [regional integration] to the virtual exclusion of other approaches."

The perceived empirical success of the gravity model has come without a great deal of analysis regarding its econometric properties, as its empirical power has usually been stated simply on the basis of goodness of fit (i.e., a relatively high R^2).[5] The lack of attention paid to the empirical properties of the model is despite the fact that the strength of any baseline model lies in the accuracy of its estimates. Recently, though, several papers have argued that standard cross-sectional methods yield biased results because they do not control for heterogeneous trading relationships. Because of this, these papers introduced fixed effects into the gravity equation. Fixed-effects models allow for unobserved or misspecified factors that simultaneously explain trade volume between two countries and, for example, the probability that the countries will be in the same regional integration regime (Mátyás, 1997; Bayoumi and Eichengreen, 1997; Cheng, 1999; Wall, 2002, 2003; Coughlin and Wall, 2003).[6] Gravity models with fixed effects have also been used by Glick and Rose (2001) and Pakko and Wall (2001) to estimate the trade effects of currency unions; by Wall (2000) and Millimet and Osang (2004) to estimate the effects of borders on trade; by Egger (2002) to calculate trade potentials; and by Wall (1999) to estimate the costs of protection.

Although the arguments underlying the use of fixed effects as a solution to unobserved heterogeneity are roughly the same in all of these papers, there is little agreement about how to actually specify the fixed effects. For example, Cheng (1999) and Wall (1999) propose two fixed effects for each pair of countries, one for each direction of trade. In Glick and Rose (2001), each pair of countries has only one fixed effect. In Mátyás (1997), each country has two fixed effects, one as an exporter and one as an importer. The purpose of this paper is to evaluate the various fixed-effect specifications in terms of the econometric appropriateness of their underlying assumptions. Specifically, we show (i) how the standard pooled-cross-section specification and other fixed-effects specifications are special cases of the Cheng (1999) and Wall (1999) specification and (ii) that the restrictions to obtain them cannot be supported empirically. To underscore the importance of getting the fixed-effects specification right, we illustrate how the choice of specification has significant implications when estimating the effects of integration on trade volume.

A STATISTICAL OVERVIEW

This section briefly sets out the various forms of the gravity model that have been used to estimate bilateral trade flows. These models are restricted versions of a general gravity model, which has a log-linear specification but places no restrictions on the parameters. In the general model, the volume of trade between countries i and j in year t can be characterized by

$$(1)\quad \ln X_{ijt} = \alpha_0 + \alpha_t + \alpha_{ij} + \boldsymbol{\beta}'_{\mathbf{ijt}}\mathbf{Z}_{\mathbf{ijt}} + \varepsilon_{ijt},\quad t = 1,\ldots,T,$$

where X_{ijt} is exports from country i to country j in year t and $\mathbf{Z}'_{\mathbf{ijt}} = [z_{it}, z_{jt}\ldots]$ is the $1 \times k$ vector of gravity variables (gross domestic product [GDP], population, and distance). The intercept has three parts: one common to all years and country pairs, α_0; one specific to year t and common to all pairs, α_t; and one specific to the country pairs and common to all years, α_{ij}. The disturbance term, ε_{ijt}, is assumed to be normally distributed with zero mean and constant variance for all observations. It is also assumed that the disturbances are pairwise uncorrelated.

Obviously, because (1) has only one observation, it is not useful for estimation unless restrictions are imposed on the parameters. The standard single-year cross-section model (CS) imposes the restrictions that the slopes and intercepts are the same across country pairs, that is, $\alpha_{ij} = 0$ and $\boldsymbol{\beta}_{\mathbf{ijt}} = \boldsymbol{\beta}_{\mathbf{t}}$,

[5] See Sanso, Cuairan, and Sanz (1993) for an examination of the predictive power of various specifications of the augmented gravity model. Also, see Oguledo and MacPhee (1994) for a survey of pre-1990 empirical results.

[6] Soloaga and Winters (2001) also recognize this problem, but their solution is to estimate yearly gravity models and to calculate the effects of integration as the differences in the predicted trade volumes over time.

(CS) $\ln X_{ijt} = \alpha_0 + \alpha_t + \boldsymbol{\beta}'_{\mathbf{t}} \mathbf{Z_{ijt}} + \varepsilon_{ijt}, \;\; t = 1,\ldots,T,$

where α_0 and α_t cannot be separated. Assuming that all the classical disturbance-term assumptions hold, the CS model is estimated by ordinary least squares (OLS) for each year.

The other standard estimation method is a pooled-cross-section model (PCS), which imposes the further restriction on the general model that the parameter vector is the same for all t, $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \ldots = \boldsymbol{\beta_T} = \boldsymbol{\beta}$, although it normally allows the intercepts to differ over time:

(PCS) $\ln X_{ijt} = \alpha_0 + \alpha_t + \boldsymbol{\beta}'\mathbf{Z_{ijt}} + \varepsilon_{ijt}, \;\; t = 1,\ldots,T.$

This is estimated by OLS using data for all available years.

Nearly all estimates of the gravity model of trade use either the CS or the PCS model, which, as we show below, both provide biased estimates. To address this bias, we remove the restriction that the country-pair intercept terms equal zero, although we maintain the restriction that the slope coefficients are constant across country pairs and over time. Specifically, we estimate the fixed-effects (FE) model of Cheng (1999) and Wall (1999):

(FE) $\ln X_{ijt} = \alpha_0 + \alpha_t + \alpha_{ij} + \boldsymbol{\beta}'\mathbf{Z_{ijt}} + \varepsilon_{ijt}, \;\; t = 1,\ldots,T.$

Note that the country-pair effects are allowed to differ according to the direction of trade (i.e., $\alpha_{ij} \neq \alpha_{ji}$). The FE model is a two-way fixed-effects model in which the independent variables are assumed to be correlated with α_{ij} and is a classical regression model that can be estimated using LSDV (least squares with a dummy variable for each of the country pairs).

As mentioned above, others have proposed alternative fixed-effects models to handle country-pair heterogeneity, each of which can be modeled as a restricted version of the FE model above. The symmetric fixed-effects (SFE) model of Glick and Rose (2001) differs from FE only in that it imposes the restriction that the country-pair effects are symmetric (i.e., $\alpha_{ij} = \alpha_{ji}$).

In the Bayoumi and Eichengreen (1997) model, call it DFE, the differences in the dependent and independent variables are used to eliminate the fixed variables, including the country-pair dummies and distance. As with the FE specification, this model allows for the most general fixed effects possible. But rather than estimating the fixed effects using LSDV, it eliminates them by subtracting them out. Specifically,

(DFE) $\Delta \ln X_{ijt} = \gamma_0 + \gamma_t + \boldsymbol{\beta}'\Delta\mathbf{Z_{ijt}} + \mu_{ijt}, \;\; t = 1,\ldots,T,$

where Δ is the difference operator and $\gamma_0 + \gamma_t = \alpha_t - \alpha_{t-1}$. In this model the intercept has two parts: γ_0 is the change in the period-specific effect that is common across years and γ_t is the change that is specific to year t.

When there are no time dummies, such a differencing model yields results identical to a model with dummy variables to control for fixed effects. However, with time dummies it is necessary to impose restrictions on the time effects to avoid collinearity, which in turn makes the DFE estimation a restricted form of the FE estimation. If the collinearity restriction is that the first time dummy in the DFE model is equal to zero, this is equivalent to restricting the common component of the change in the period-specific effects as equal to the difference in the first two period-specific effects (i.e., $\gamma_0 = \alpha_2 - \alpha_1$). If, instead, the collinearity restriction is that the sum of the time dummies in the DFE model is zero, this is equivalent to restricting the common component as equal to the difference between the first and last time dummies (i.e., $\gamma_0 = \alpha_T - \alpha_1$).

Mátyás (1997) proposes

(XFE)
$\ln X_{ijt} = \alpha_0 + \alpha_t + \theta_i + \omega_j + \boldsymbol{\beta}'\mathbf{Z_{ijt}} + \varepsilon_{ijt}, \;\; t = 1,\ldots,T,$

as the correct specification of the gravity model, where the country-specific effect is θ_i when a country is an exporter and is ω_j when it is an importer. Note that in this specification, distance, contiguity, and language are eliminated because they are fixed over time, even though they are not collinear with the country-specific effects. This model is a special case of the FE model in that it has a unique value for each trading pair's intercept, with the restrictions that a country's fixed effect as an exporter or importer is the same for all of its trading partners. This imposes cross-pair restrictions on the intercepts—that is, one of the components of the intercept for Germany-to-Canada trade

Table 1

Regression Results for Models Using Pooled Data

Dependent Variable = Log of Real Exports

	Pooled cross-section	Unrestricted FE model	Restricted FE models		
	PCS	FE	SFE	DFE	XFE
Intercept	6.852* (0.546)	—	—	0.209* (0.028)	—
Origin GDP	0.617* (0.038)	0.122* (0.023)	0.213* (0.025)	0.098* (0.029)	0.122* (0.055)
Destination GDP	0.511* (0.035)	0.208* (0.027)	0.117* (0.024)	0.258* (0.029)	0.208* (0.054)
Origin population	0.141* (0.038)	–0.390 (0.298)	0.935* (0.268)	–0.482 (0.344)	–0.390 (0.565)
Destination population	0.214* (0.038)	2.313* (0.319)	0.989* (0.268)	1.906* (0.344)	2.313* (0.584)
Distance	–1.025* (0.023)	—	—	—	—
Contiguity	–0.125 (0.085)	—	—	—	—
Common language	1.075* (0.072)	—	—	—	—
1987	0.077 (0.067)	0.199* (0.029)	0.199* (0.038)	—	0.199* (0.063)
1992	0.014 (0.068)	0.357* (0.043)	0.357* (0.053)	–0.040 (0.029)	0.357* (0.093)
1997	0.051 (0.064)	0.482* (0.058)	0.481* (0.070)	–0.064* (0.028)	0.482* (0.122)
Observations	3,188	3,188	3,188	2,391	3,188
Parameters	11	804	408	7	63
Log-likelihood	–5,163.27	–1,663.07	–2,863.46	–1,979.64	–4,704.08
$\bar{R}^2$	0.690	0.954	0.916	0.050	0.768

NOTE: All non-dummy variables are in logs. White-corrected standard errors are in parentheses; * denotes significance at the 5 percent level. For the DFE model, all variables are in differences from the previous year.

must be the same as one of the components of the intercept for Germany-to-France trade. These restrictions do not change the coefficient estimates very much but, as we show below, lead to biased and rather large residuals, indicating inaccurate in-sample predictions of trade flows.

STANDARD RESULTS

This section presents regression results for the augmented version of the standard PCS model.[7] The data set is a balanced panel with 3,188 observations (797 unidirectional country pairs in each of four years: 1982, 1987, 1992, and 1997).[8] We included observations of non-zero trade between countries listed in all of the relevant World Bank *World Development Reports* as being upper-middle or high income during these years. Also, we excluded countries that were identified as high-income oil exporters. The result is a manageable data set that is fairly representative of the literature, which typically includes only OECD members or industrialized countries. Descriptions of the data and their sources are provided in the data appendix.

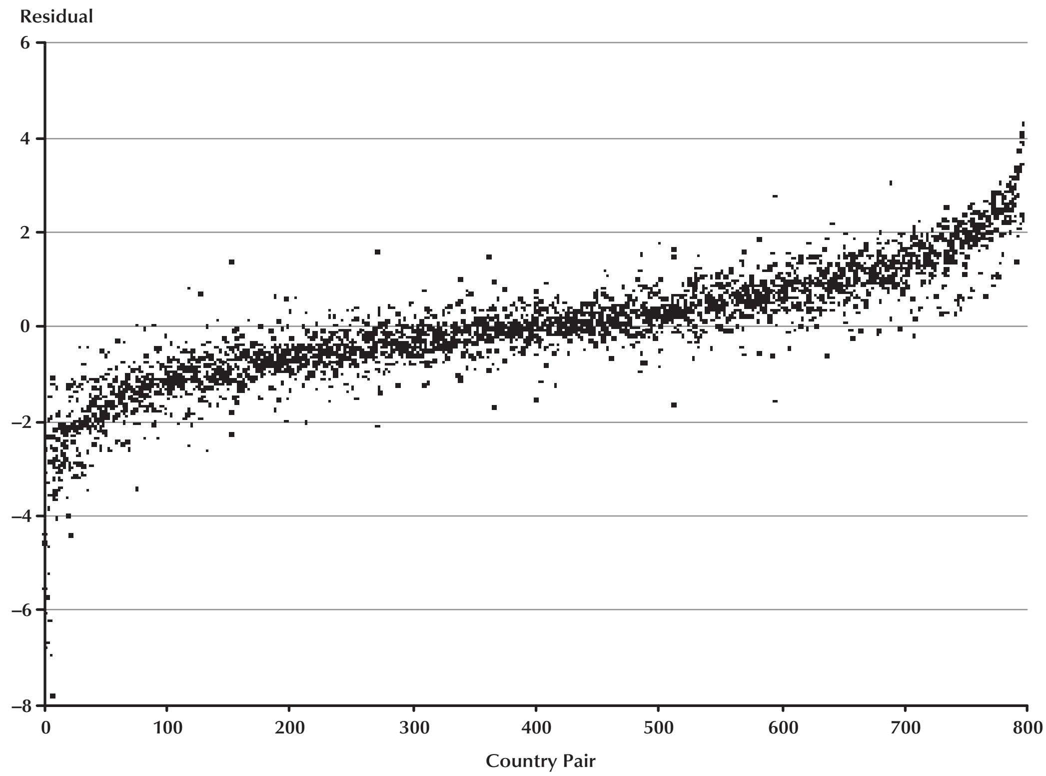
In the augmented version of the gravity model, the gravity variables are the countries' GDPs, their populations, and the distance between them. Thus, the augmented PCS model assumes that, in a given year, trade flows from exporting country

[7] Because the results for the single-year CS model do not differ substantially from those for the PCS model, we do not present them here. However, they are available upon request.

[8] Fixed-effects estimation is sometimes criticized when applied to data pooled over consecutive years on the grounds that dependent and independent variables cannot fully adjust in a single year's time. To avoid this, we left five years between our observations.

Figure 1

PCS Residuals by Country Pair



i to importing country j can be estimated using[9]

(2)

$$\ln X_{ijt} = \alpha_0 + \alpha_t + \beta_1 \ln Y_{it} + \beta_2 \ln Y_{jt} + \beta_3 \ln N_{it} + \beta_4 \ln N_{jt} + \delta_1 \ln D_{ij} + \delta_2 C_{ij} + \lambda L_{ij} + \varepsilon_{ijt},$$

where α_0 is the portion of the intercept that is common to all years and trading pairs, α_t denotes the year-specific effect common to all trading pairs, Y_i and Y_j are the two countries' GDPs, N_i and N_j are their populations, D_{ij} is the distance between them, C_{ij} is a contiguity dummy, and L_{ij} is a common-language dummy. Note that our estimation omits the dummy for 1982 to avoid collinearity.

Because trade flows are expected to be positively related to national incomes, and negatively related to distance, β_1, β_2, and δ_2 are expected to be positive and δ_1 is expected to be negative. The signs expected for population coefficients are not as unambiguous, and the literature has not tended to find a consistent sign for β_3 or β_4.[10] Because L_{ij} is meant to capture cultural and historical similarities between the trading pairs, which are thought to increase the volume of trade, λ is expected to be positive. Finally, we take the time dummies as indicators of the extent of "globalization," which we define as the purported common trend toward greater real trading volumes, independent of the sizes of the economies.

The regression results for PCS are reported in the first column of Table 1. The signs of the coefficients on distance, common language, and the countries' GDPs are as expected and are statistically significant. Only the negative coefficient on

[9] Note that the regression could be suitably rearranged to instead obtain the augmented model with per capita income.

[10] See Oguledo and MacPhee (1994).

the contiguity dummy of PCS is not as expected, although it is not statistically different from zero. Perhaps surprisingly, the coefficients on the time dummies do not indicate a trend toward globalization.

According to the estimates of the PCS model, (i) an increase in a country's GDP will lead to a less-than-proportional increase in its imports and exports and (ii) a country will export 103 percent more to a market that is half as distant as another otherwise-identical market and 108 percent more to a country with the same first language. Finally, we take the fact that the time dummies are not statistically different from zero to mean that globalization, as defined above, was not an important factor in increasing trade over the period.

Despite the supposed empirical success that we have replicated, there is a severe problem with the standard PCS model. This is clear from Figure 1, which plots the residuals for the PCS model for the 797 unidirectional country pairs in our data set, ordered by the pairs' average residuals. If the PCS estimation were unbiased, there would be no discernable pattern in Figure 1 because the average residual for each country pair would be zero. The residuals for 544 of the country pairs, however, always have the same sign. In other words, the PCS model consistently misestimated the volume of trade for at least 68 percent of the country pairs.

THE GRAVITY MODEL WITH COUNTRY-PAIR FIXED EFFECTS

The Model

Standard cross-section estimates of the gravity model yield biased estimates of the volume of bilateral trade because there is no heterogeneity allowed for in the regression equations. With such heterogeneity, a country would export different amounts to two countries, even though the two export markets have the same GDPs and are equidistant from the exporter. This can be because there are historical, cultural, ethnic, political, or geographic factors that affect the level of trade and are correlated with the gravity variables (GDP, population, distance). If so, then estimates that do not account for these factors will suffer from heterogeneity bias.

Some studies using the PCS model have, to some extent, tried to control for this by including things such as whether trading partners share a common language, have had a colonial history, or are in military alliance. However, cultural, historical, and political factors are often difficult to observe, let alone quantify. This is why we control for these factors using a simple fixed-effects model that assumes that there are fixed pair-specific factors that may be correlated with levels of bilateral trade and with the right-hand-side variables. It is in this sense that fixed-effects modeling is a result of ignorance: We do not have a good idea which variables are responsible for the heterogeneity bias, so we simply allow each trading pair to have its own dummy variable.

We assume that the gravity equation for a country pair may have a unique intercept and that it may be different for each direction of trade (i.e., $\alpha_{ij} \neq \alpha_{ji}$). However, we retain the assumptions of the PCS model that the slope coefficients are constant over time and across trading pairs. The Cheng (1999) and Wall (1999) specification of the augmented FE is

(3)
$$\ln X_{ijt} = \alpha_{ij} + \alpha_t + \beta_1 \ln Y_{it} + \beta_2 \ln Y_{jt} + \beta_3 \ln N_{it} + \beta_4 \ln N_{jt} + \varepsilon_{ijt},$$

where α_{ij} is the specific "country-pair" effect between the trading partners. The country-pair intercepts include the effects of all omitted variables that are cross-sectionally specific but remain constant over time, such as distance, contiguity, language, and culture. Using the pooled data described above, we have 797 country-pair intercepts.

Because there is a long-standing problem with determining the appropriate measure of economic distance to capture transportation and information costs (see Head and Mayer, 2001, for a review of the issue), an added benefit of the fixed-effects model is that it eliminates the need to include distance in the regression. The most common method for measuring distance is to do as we have done and simply measure it between the centers (often assumed to be the capital cities) of the two countries. There are problems with this, such as the implicit assumptions that overland transport costs are the same as those over sea and that all overland/oversea distances are equally costly. To provide just one example, Los Angeles is about 1,300 km farther from Tokyo than is Moscow, but the economic distance between Tokyo and Los Angeles is certainly much lower than that between Tokyo and Moscow. Our FE approach eliminates the need to include a distance variable, as it controls for all variables that do not change over time.

Another difficulty with standard measures of economic distance is the common assumption that the capital city, or any other single point in the country, is a useful proxy for the economic center. While this may be useful for small countries with one major city, it is wide of the mark for countries like Canada and the United States, which have major cities thousands of miles apart on different oceans and which serve as centers for trade with completely different countries. By using Washington, D.C., or Ottawa to measure distance between the United States or Canada and its Pacific trading partners is to overstate distance by the entire breadth of the North American continent. As the United States has the highest GDP and the highest volume of trade, the mismeasurement of economic distance can bias the estimation of the coefficients on the other variables in the gravity model.

Another advantage of our approach is that it removes the problem of controlling for contiguity. Although it is potentially important, as a great deal of trade can occur from people crossing the border to make everyday purchases, it is accounted for only sometimes. Even when it is accounted for with a dummy variable, as we do above, it still assumes that all contiguity is equivalent and time invariant in terms of its effect on trade. Considering that Canada and the United States, China and Russia, and Argentina and Chile are all equivalently contiguous pairs, this is difficult to abide by.

The Results

Table 1 reports the estimation results for the augmented version of the FE model. Note that, for comparison with the PCS results, the year dummies are measured relative to that of 1982. Also, the estimates of the country-pair intercepts are omitted for space considerations. According to the results for the FE model, (i) an increase in a country's GDP will lead to a less-than-proportional increase in its imports and exports and (ii) globalization has increased the real volume of trade by 48 percent between 1982 and 1997.

A comparison of the results of the FE and PCS models shows that allowing for trading-pair heterogeneity lowers the estimated income elasticities of trade, greatly increases the absolute value of the coefficients on the countries' populations, and greatly increases the estimated role of globalization. It is obvious from the results that restricting the country-pair effects to zero, as the PCS model does, has statistically significant effects on the results, as is easily confirmed by a likelihood ratio test.[11] Note also that the residuals from the FE estimation across country pairs (Figure 2) have no discernible pattern.

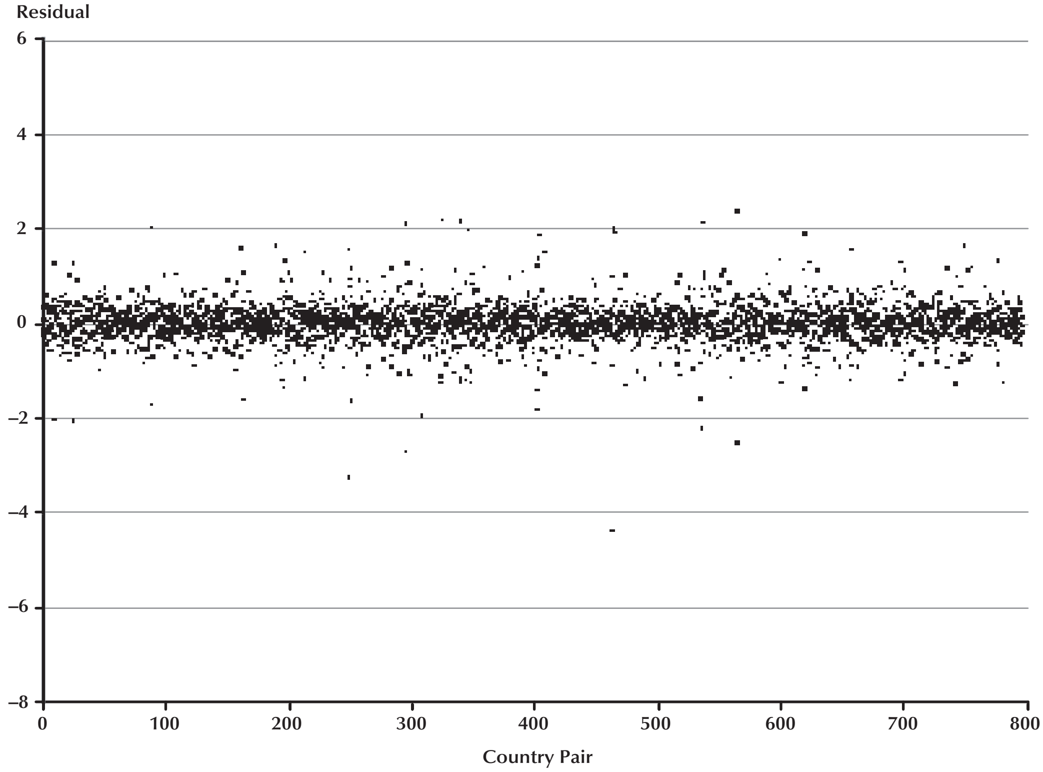
Therefore, because the PCS model is a restricted form of the FE model and the restrictions are not supported statistically, we conclude that the FE model is the preferred specification of the gravity model. In short, there is no statistical support for imposing the parameter restrictions required by the standard procedures for estimating the gravity model of trade. In the absence of any economic arguments for believing that the intercepts of the gravity equation are the same across trading pairs, we conclude that the FE model is the more appropriate specification.

Oddly, Wei and Frankel (1997, p. 125) reject the inclusion of country-pair dummies a priori on the basis that doing so would undermine their efforts at estimating the effects of variables that are constant over the sample period. Presumably, their worry is that, because these variables are subsumed into the country-pair effects, they are hidden from analysis. This is unfounded because the effects of these variables are easily estimated by regressing them on the country-pair effects from the FE model. Specifically, where the estimates of the 797 country-pair effects are denoted as $\hat{\alpha}_{ij}$, and including the log of distance and the contiguity

[11] This is with LR = 7,000.4 and $\chi^2(796) = 862.75$ at the 5 percent level.

Figure 2

FE Residuals by Country Pair



and language dummies as independent variables, we obtain

$$\hat{\alpha}_{ij} = \underset{(0.405)}{10.408} - \underset{(0.049)}{1.236}\ln D_{ij} - \underset{(0.246)}{0.746}C_{ij} + \underset{(0.198)}{1.565}L_{ij}.$$

The numbers in parentheses are White-corrected standard errors and the $\bar{R}^2 = 0.158$. According to these results, all three variables are statistically significant determinants of the country-pair effects. Inexplicably, though, the coefficient on the contiguity dummy is negative, as in the PCS results. Note that these estimates are quite different from those obtained from the PCS model, in which estimates of the effects of time-invariant factors suffer from the same heterogeneity bias as the time-variant factors. So, far from undermining estimation efforts, it is instead *necessary* to control for country-pair heterogeneity to obtain unbiased estimates of the importance of time-invariant factors.

ALTERNATIVE FIXED-EFFECTS SPECIFICATIONS

As discussed earlier, others have used less-general fixed-effects specifications. The first of these, the SFE model of Glick and Rose (2001), is simply (3) with the restriction that $\alpha_{ij} = \alpha_{ji}$. Because our data set does not have the entire set of country pairs for both directions, this does not mean that there are exactly one-half as many country-pair effects as the FE model, although it is close. Our FE estimation had 797 country-pair effects, whereas our SFE estimation has 401. The results of the SFE estimation are in Table 1 and indicate that the symmetry restriction on the country-pair effects has a statistically significant effect on the results. Each of the coefficients on the gravity variables is very different from what we obtain with the FE model, although the coefficients on the year dummies are nearly identical. Also, a likelihood ratio test easily rejects the null hypothesis that the restrictions do not have a statistically significant effect on the estimation.[12] This means that the FE model is preferred statistically to the SFE model.

Taking the time difference of (3), the DFE model of Bayoumi and Eichengreen (1997) is

(4)
$$\Delta \ln X_{ijt} = \gamma_0 + \gamma_t + \beta_1 \Delta \ln Y_{it} + \beta_2 \Delta \ln Y_{jt} + \beta_3 \Delta \ln N_{it} + \beta_4 \Delta \ln N_{jt} + \mu_{ijt},$$

where the intercept is as defined in the statistical overview, $\gamma_0 + \gamma_t = \alpha_t - \alpha_{t-1}$. To prevent collinearity, we set the time dummy for 1987 equal to zero, meaning that other time dummies are measured relative to it. In terms of the more-general FE model, this is equivalent to restricting the common component of the change in the period-specific effects as equal to the difference in the first two period-specific effects (i.e., $\gamma_0 = \alpha_2 - \alpha_1$).[13] The empirical results are presented in Table 1.

The results for the FE and DFE models are similar in terms of the signs and order of magnitude of the coefficients. Nonetheless, the FE and DFE results differ enough to reject the restrictions needed to obtain the DFE model. This can be confirmed easily by a likelihood ratio test. Therefore, given that the restrictions that DFE imposes on the time dummies are not justified on any economic or statistical grounds, our results indicate that they should not be imposed.

The third alternative to the FE model, XFE, is

(5)
$$\ln X_{ijt} = \alpha_0 + \alpha_t + \theta_i + \omega_j + \beta_1 \ln Y_{it} + \beta_2 \ln Y_{jt} + \beta_3 \ln N_{it} + \beta_4 \ln N_{jt} + \varepsilon_{ijt},$$

where the fixed effect when a country is an exporter is θ_i and when it is an importer is ω_j. One way to prevent perfect collinearity in estimating (5) is to impose the restrictions that one of the θs and one of the ωs is zero. Because each θ_i and ω_j comprise part of many α_{ij}s, this is the same as imposing a series of cross-pair restrictions on the α_{ij}s. From the empirical results summarized in the last column of Table 1, it seems that the coefficients are the same as those from the FE model. In fact, the coefficients are not the same, but the differences are so small that they appear only beyond the seventh decimal places provided by STATA. More importantly, though, the standard errors from the XFE model are much larger. Consequently, the FE model is preferred to the XFE model on the basis of any standard goodness-of-fit criteria. As with the other restricted fixed-effects specifications, a likelihood ratio test easily rejects the null hypothesis that the arbitrary restrictions imposed by XFE are not statistically benign.

THE EFFECTS OF INTEGRATION

As we discuss in our introduction, the gravity model has become the primary tool for estimating the effects of regional integration on trade volumes. Up to this point, we have omitted integration variables in order to focus on the importance of controlling for country-pair heterogeneity when estimating gravity models. We now introduce integration into our model and demonstrate the striking effect that heterogeneity bias has on the results. We would also like to alleviate the valid concern that the heterogeneity bias we detected above was due to our implicit assumption that regional integration is uncorrelated with the independent variables.

The most common and straightforward method for estimating the effects of integration in a gravity model is to include dummy variables for each integration regime in place during the sample period (see, for example, Frankel, 1997). Each of these dummies takes the value of 1 for an observation for which the two countries are members of the regime, with the expectation that the coefficients on these dummies are positive. We include five such dummy variables in our model, one each for the European trading bloc, the North American trading bloc, the South American trading bloc (Mercosur), the Australia–New Zealand Closer Economic Relations (CER), and the Israel–United States Free Trade Agreement (FTA).

[12] This is with LR = 2,400.78 and $\chi^2(395) = 442.34$ at the 5 percent level.

[13] The alternative assumption that the sum of the year dummies is zero means that $\gamma_0 = \alpha_T - \alpha_1$ and yields the same results except for the time dummies and the constant.

Although there has been some deepening of trade integration in the European bloc, the primary change over the period was an expansion in the number of countries covered under the customs union. The formation of the European Community (EC) predates our data set, and Portugal and Spain joined in 1986. The 12 countries of the EC renamed themselves the European Union (EU) in 1992, but this had relatively little effect on internal trade policy, as it was already nearly unfettered under the EC. Expansion of the bloc came in 1994 with the European Economic Area (EEA), which extended the free trade zone to include Austria, Iceland, Finland, Norway, and Sweden. To capture the effect of this trading bloc, our European bloc dummy variable takes the value of 1 when trade is between members of the EC or EU for 1982, 1987, 1992 and when trade is between members of the EEA for 1997.

The Canada–United States Trade Agreement of 1988 established a North American trading bloc that included only Canada and the United States. The North American Free Trade Agreement (NAFTA) expanded the free trade zone in 1994 to include Mexico. We ignore NAFTA's relatively mild deepening of Canada–United States integration and focus on it instead as an extension of the free trade bloc to Mexico. Our North American bloc dummy takes the value of 1 for trade between the United States and Canada for 1992 and between Mexico, Canada, and the United States for 1997.

The third significant trade bloc during the period was Mercosur, which came into force in 1995, reducing trade barriers between Argentina, Brazil, Paraguay, and Uruguay. Our Mercosur dummy takes the value of 1 for trade between any two of these countries in 1997. The Australia–New Zealand CER was formed in 1983, so its dummy variable is equal to 1 for trade between the two countries for all years but 1982. Similarly, the Israel–United States FTA entered into force in 1985, so its dummy variable is equal to 1 for trade between the two countries for 1987, 1992, and 1997.

We include these trade bloc dummies in the PCS and FE models and report the empirical results in Table 2. Note that inclusion of these dummies makes little difference for the PCS model. Nonetheless, a likelihood ratio test rejects the null hypotheses that including the trade bloc dummies in the PCS model does not alter the results to a statistically significant extent.[14] Similarly, the results for the FE model are also not dramatically different when the trade bloc dummies are included, although the null hypothesis that the inclusion of these variables has no statistically significant effect on the results is rejected.[15]

Both models find modest effects on trade from the European trade bloc. The PCS estimates say that the bloc had a statistically insignificant effect, but the FE estimates say that it had a statistically significant effect of 8.2 percent ($e^{0.079} - 1 = 0.082$). The larger differences between the two models are in the estimated effects of the other trade blocs. The PCS model suggests a 172 percent increase in trade between North American countries because of their trading bloc, whereas the FE model suggests that the bloc led to only a 34 percent increase in trade. For Mercosur, the PCS model estimates an increase in trade of 23 percent that is far from being statistically significant, whereas the FE model estimates a statistically significant effect of 61 percent. The PCS model also estimates the effects of the Australia–New Zealand CER and the Israel–United States FTA as increases in intra-bloc trade of about 300 and 400 percent, respectively. The FE model, however, finds a statistically significant effect of –12 percent for the Australia–New Zealand CER and a statistically insignificant effect of –7.3 percent for the Israel–United States FTA.

These results highlight how allowing for unobserved or unmeasurable heterogeneity can alter gravity model estimates. Specifically, the fact that the estimated effects of the trade blocs change when country-pair heterogeneity is allowed for means that there are pair-specific effects that are correlated with the level of trade between pairs of countries and with the likelihood that the pair

[14] This is with LR = 23.6 and $\chi^2(5) = 11.07$ at the 5 percent level.

[15] This is with LR = 11.9 and $\chi^2(5) = 11.07$ at the 5 percent level.

Table 2

Regression Results with Integration Dummies
Dependent Variable = Log of Real Exports

	Pooled cross-section	Unrestricted FE model	Restricted FE models		
	PCS	FE	SFE	DFE	XFE
Intercept	6.756* (0.581)	—	—	0.208* (0.029)	—
Origin GDP	0.618* (0.038)	0.125* (0.023)	0.217* (0.026)	0.098* (0.029)	0.176* (0.049)
Destination GDP	0.512* (0.035)	0.212* (0.028)	0.121* (0.025)	0.258* (0.029)	0.263* (0.050)
Origin population	0.138* (0.039)	–0.316 (0.315)	1.009* (0.297)	–0.476 (0.351)	0.884† (0.529)
Destination population	0.210* (0.038)	2.386* (0.339)	1.063* (0.298)	1.912* (0.351)	3.575* (0.561)
Distance	–1.013* (0.028)	—	—	—	—
Contiguity	–0.145† (0.087)	—	—	—	—
Common language	1.050* (0.073)	—	—	—	—
European bloc	0.059 (0.062)	0.079† (0.041)	0.079 (0.051)	0.021 (0.048)	1.196* (0.059)
North American bloc	1.000* (0.165)	0.294* (0.140)	0.295† (0.177)	0.204 (0.227)	2.675* (0.216)
Mercosur	0.203 (0.436)	0.475* (0.172)	0.475* (0.166)	0.365 (0.228)	4.133* (0.246)
Australia–N.Z. CER	1.370* (0.113)	–0.125* (0.044)	–0.124 (0.120)	–0.158 (0.393)	3.981* (0.174)
Israel–U.S. FTA	1.642* (0.102)	–0.076 (0.296)	–0.076 (0.126)	0.024 (0.393)	0.426* (0.168)
1987	0.067 (0.067)	0.189* (0.031)	0.189* (0.040)	—	0.020 (0.062)
1992	0.0005 (0.069)	0.338* (0.047)	0.338* (0.058)	–0.040 (0.029)	0.027 (0.089)
1997	0.022 (0.069)	0.437* (0.069)	0.436* (0.085)	–0.070* (0.029)	–0.220† (0.119)
Observations	3,188	3,188	3,188	2,391	3,188
Parameters	16	809	413	12	68
Log-likelihood	–5,151.48	–1,657.12	–2,860.65	–1,977.80	–4,427.03
$\bar{R}^2$	0.692	0.954	0.916	0.050	0.805

NOTE: All non-dummy variables are in logs. White-corrected standard errors are in parentheses; * and † denote significance at the 5 and 10 percent levels, respectively. For the DFE model, all variables are in differences from the previous year.

will enter a trading bloc.[16] In particular, the lower estimated effect of the Israel–United States FTA using the FE model indicates that there is something special about the relationship between the United States and Israel that makes them trade relatively more with each other than the gravity variables would predict, and which led them to sign a trade agreement. Suppressing this pair-specific effect, as the PCS model does, mistakenly suggests that it is the FTA that is responsible for the high trade volume, rather than the special relationship. Similarly, our results suggest for the Australia–New Zealand CER and the North American bloc that the high levels of intra-bloc trade can be attributed to cultural and geographic proximity not completely captured by the language and distance variables, and not primarily to the blocs themselves.

For the sake of comparison, we also estimated the effects of integration using the three alternative fixed-effects specifications. As shown in Table 2, the point estimates of the effects of the blocs on

[16] We should note that if we regress the estimated fixed effects from this estimation against distance, contiguity, and language, the results do not differ substantially from those obtained above, which used the estimated fixed effects without controlling for regional integration.

trade are nearly identical between the FE and SFE models. Nonetheless, because the standard errors from the SFE estimates are larger, one would conclude from them that the effects of the European bloc and the Australia–New Zealand CER were statistically no different from zero, even though the FE estimates indicate their statistical significance.

Estimates using the DFE model are also not dramatically different from those using the FE model. Again, though, the larger standard errors mean that the estimated effects are further from standard levels of statistical significance. Indeed, the DFE estimates indicate that none of the trading blocs had a statistically significant effect on trade between members. This occurs because the DFE model imposes restrictions on the time dummies, thereby leading to the misestimation of the effects of regional integration regimes, the expansions of which have a significant trend component.

The XFE model provides estimates of the effects of integration that are dramatically different from those provided by any of the other models. Specifically, it suggests that the European bloc led to an increase in trade of 230 percent, that the North American bloc led to a 1,350 percent increase in trade, and that Mercosur and the Australia–New Zealand CER led to increases in trade of greater than 5,000 percent.

CONCLUSIONS

The objective of this paper is to compare ways that heterogeneity has been allowed for when using the gravity model to estimate bilateral trade flows. Our empirical analysis shows first that standard pooled-cross-section methods for estimating gravity models of trade suffer from estimation bias due to omitted or misspecified variables. It also shows that the problem is eliminated using the two-way fixed-effects model of Cheng (1999) and Wall (1999) in which country-pair and period dummies are used to reflect the bilateral relationship between trading partners. The fixed effects capture those factors such as physical distance, the length of the border (or contiguity), history, culture, and language that are constant over the span of the data and that are correlated with the volume of bilateral trade.

We show that alternative fixed-effects models proposed by Glick and Rose (2001), Mátyás (1997), and Bayoumi and Eichengreen (1997) are special cases of our model and that the restrictions necessary to obtain these special cases are not supported statistically. Also, because these restrictions have little or no economic support, we argue that they should not be imposed. As the gravity model has become the "workhorse" of empirical studies of the effects of integration, we also compare the various specifications in this regard. We conclude that the country-pair fixed-effects model is preferred statistically to all other specifications and show that estimates of the effects of integration on trade can differ a great deal across the specifications.

REFERENCES

Aitken, Norman D. "The Effect of the EEC and EFTA on European Trade: A Temporal Cross-Section Analysis." *American Economic Review*, December 1973, *63*(5), pp. 881-92.

Bayoumi, Tamim and Eichengreen, Barry. "Is Regionalism Simply a Diversion? Evidence from the Evolution of the EC and EFTA," in Takatoshi Ito and Anne O. Krueger, eds., *Regionalism versus Multilateral Trade Arrangements*. Chicago: University of Chicago Press, 1997, pp. 141-64.

Bergstrand, Jeffrey H. "The Generalized Gravity Equation, Monopolistic Competition, and the Factor-Proportions Theory in International Trade." *Review of Economics and Statistics*, February 1989, *71*(1), pp. 143-53.

Bikker, Jacob A. "An International Trade Flow Model with Substitution: An Extension of the Gravity Model." *Kyklos*, 1987, *40*(3), pp. 315-37.

Boisso, Dale and Ferrantino, Michael. "Economic Distance, Cultural Distance, and Openness in International Trade: Empirical Puzzles." *Journal of Economic Integration*, December 1997, *12*(4), pp. 456-84.

Brada, Josef C. and Mendez, Jose A. "Regional Economic Integration and the Volume of Intra-Regional Trade: A Comparison of Developed and

Developing Country Experience." *Kyklos*, 1983, *36*(4), pp. 589-603.

Cheng, I-Hui. "The Political Economy of Economic Integration." Ph.D. Dissertation, Birkbeck College, University of London, July 1999.

Coughlin, Cletus C. and Wall, Howard J. "NAFTA and the Changing Pattern of State Exports." *Papers in Regional Science*, October 2003, *82*(4), pp. 427-50.

Egger, Peter. "An Econometric View on the Estimation of Gravity Models and the Calculation of Trade Potentials." *The World Economy*, February 2002, *25*(2), pp. 297-312.

Eichengreen, Barry and Irwin, Douglas A. "The Role of History in Bilateral Trade Flows," in Jeffrey A. Frankel, ed., *The Regionalization of the World Economy*. Chicago: University of Chicago Press, 1998, pp. 33-57.

Feenstra, Robert C. "World Trade Flows, 1980-1997." Center for International Data, University of California, Davis, March 2000.

Frankel, Jeffrey A. *Regional Trading Blocs in the World Economic System*. Washington, DC: Institute for International Economics, 1997.

Frankel, Jeffrey; Stein, Ernesto and Wei, Shang-Jin. "Trading Blocs and the Americas: The Natural, the Unnatural, and the Super-Natural." *Journal of Development Economics*, June 1995, *47*(1), pp. 61-95.

Frankel, Jeffrey A.; Stein, Ernesto and Wei, Shang-Jin. "Continental Trading Blocs: Are They Natural and Supernatural?" in Jeffrey A. Frankel, ed., *The Regionalization of the World Economy*. Chicago: University of Chicago Press, 1998, pp. 91-113.

Frankel, Jeffrey A. and Wei, Shang-Jin. "Regionalization of World Trade and Currencies: Economics and Politics," in Jeffrey A. Frankel, ed., *The Regionalization of the World Economy*. Chicago: University of Chicago Press, 1998, pp. 189-219.

Glick, Reuven and Rose, Andrew K. "Does a Currency Union Affect Trade? The Time Series Evidence." NBER Working Paper No. 8396, National Bureau of Economic Research, July 2001.

Head, Keith and Mayer, Thierry. "Illusory Border Effects: How Far Is an Economy from Itself?" Working Paper, University of British Columbia, 2001.

Helliwell, John F. "Do National Borders Matter for Quebec's Trade?" *Canadian Journal of Economics*, August 1996, *29*(3), pp. 507-22.

Linnemann, Hans. *An Econometric Study of International Trade Flows*. Amsterdam: North-Holland, 1966.

McCallum, John. "National Borders Matter: Canada-U.S. Regional Trade Patterns." *American Economic Review*, June 1995, *85*(3), pp. 615-23.

Mátyás, László. "Proper Econometric Specification of the Gravity Model." *The World Economy*, May 1997, *20*(3), pp. 363-68.

Millimet, Daniel L. and Osang, Thomas. "Do State Borders Matter for U.S. Intranational Trade? The Role of History and Internal Migration." Working Paper, Southern Methodist University, May 2004.

Oguledo, Victor Iwuagwu and MacPhee, Craig R. "Gravity Models: A Reformulation and an Application to Discriminatory Trade Arrangements." *Applied Economics*, February 1994, *26*(2), pp. 107-20.

Pakko, Michael R. and Wall, Howard J. "Reconsidering the Trade-Creating Effects of a Currency Union." Federal Reserve Bank of St. Louis *Review*, September/October 2001, *83*(5), pp. 37-45.

Pöyhönen, P. "A Tentative Model for the Volume of Trade Between Countries." *Weltwirtschaftliches Archiv*, 1963, *90*(1), pp. 93-100.

Rose, Andrew K. "One Money, One Market: The Effect of Common Currencies on Trade." *Economic Policy: A European Forum*, April 2000, *15*(30), pp. 7-33.

Sanso, Marcos; Cuairan, Rogelio and Sanz, Fernando. "Bilateral Trade Flows, the Gravity Equation, and Functional Form." *Review of Economics and Statistics*, May 1993, *75*(2), pp. 266-75.

Smith, Pamela J. "Are Weak Patent Rights a Barrier to U.S. Exports?" *Journal of International Economics*, June 1999, *48*(1), pp. 151-77.

Soloaga, Isidro and Winters, L. Alan. "Regionalism in the Nineties: What Effect on Trade?" *The North American Journal of Economics and Finance*, March 2001, *12*(1), pp. 1-29.

Tinbergen, Jan. *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: The Twentieth Century Fund, 1962.

Wall, Howard J. "Using the Gravity Model to Estimate the Costs of Protection." Federal Reserve Bank of St. Louis *Review*, January/February 1999, *81*(1), pp. 33-40.

Wall, Howard J. "Gravity Model Specification and the Effect of the Canada-U.S. Border." Working Paper No. 2000-024A, Federal Reserve Bank of St. Louis, September 2000.

Wall, Howard J. "Has Japan Been Left Out in the Cold by Regional Integration?" Bank of Japan, *Monetary and Economic Studies*, April 2002, *20*(2), pp. 117-34.

Wall, Howard J. "NAFTA and the Geography of North American Trade." Federal Reserve Bank of St. Louis *Review*, March/April 2003, *85*(2), pp. 13-26.

Wei, Shang-Jin and Frankel, Jeffrey A. "Open versus Closed Trading Blocs," in Takatoshi Ito and Anne O. Krueger, eds., *Regionalism versus Multilateral Trade Arrangements*. Chicago: University of Chicago Press, 1997, pp. 119-39.

DATA APPENDIX

DEFINITIONS OF VARIABLES

Real Exports, measured in millions of U.S. dollars, from *World Trade Flows, 1980-1997* (see Feenstra, 2000). Deflated using CPI-U-RS from the Bureau of Labor Statistics.

Real Gross Domestic Product is in millions of U.S. dollars at market prices from the World Bank's *World Development Indicators 1999* CD-ROM. Deflated using CPI-U-RS from the Bureau of Labor Statistics.

Population in thousands of inhabitants from the World Bank's *World Development Indicators 1999* CD-ROM.

***Distance*,** expressed in kilometers, is the great circle distance between geographic centers, using the Haversine formula. Coordinates from the CIA's *The World Factbook 2000*.

Contiguity is equal to 1 if two trading partners share a border. From the CIA's *The World Factbook 2000*.

Common Language is equal to 1 if two trading partners share a common first language. From the CIA's *The World Factbook 2000*.

European Bloc is equal to 1 when both countries are members of the EC for 1982 or 1987, the EU for 1992, or the EEA for 1997.

North American Bloc is equal to 1 for Canada–United States trade for 1992 and 1997, and for Canada-Mexico and United States–Mexico trade for 1997.

Mercosur is equal to 1 in 1997 for trade between Argentina, Brazil, Paraguay, and Uruguay.

Australia–New Zealand CER is equal to 1 in 1987, 1992, and 1997 for trade between Australia and New Zealand.

Israel–United States FTA is equal to 1 in 1987, 1992, and 1997 for trade between Israel and the United States.

THE 29 COUNTRIES INCLUDED IN THE DATA SET

Argentina, Australia, Austria, Belgium-Luxembourg, Brazil, Canada, Denmark, Finland, France, Germany, Greece, Hong Kong, Ireland, Israel, Italy, Japan, South Korea, Mexico, the Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the United Kingdom, Uruguay, and the United States.

Federal Reserve Bank of St. Louis
P.O. Box 442
St. Louis, MO 63166-0442