FEDERAL RESERVE BANK OF ST. LOUIS

REVIEW

JANUARY/FEBRUARY 2002 | VOLUME 84, NUMBER 1

Review is published six times per year by the Research Division of the Federal Reserve Bank of St. Louis. 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, Missouri 63166-0442, or call (314) 444-8808 or 8809.

Contents

Volume 84, Number 1

wealth-maximization in takeovers are the one share–one vote principle, majority rules, and mandatory tender offers. We analyze the implications of these three institutional arrangements in a simple textbook takeover model. The model helps in understanding the optimal design of a legal environment in which the market for corporate control promotes efficient allocation of capital.

47 **Could a CAMELS Downgrade Model Improve Off-Site Surveillance?**

R. Alton Gilbert, Andrew P. Meyer, and Mark D. Vaughan

The Federal Reserve's off-site surveillance system includes two econometric models that are collectively known as the System for Estimating Examination Ratings (SEER). One model, the SEER risk rank model, uses the latest financial statements to estimate the probability that each Fed-supervised bank will fail in the next two years. The other component, the SEER rating model, uses the latest financial statements to produce a "shadow" CAMELS rating for each supervised bank. Banks identified as risky by either model receive closer supervisory scrutiny than other state-member banks.

Because many of the banks flagged by the SEER models have already tumbled into poor condition and, hence, would already be receiving considerable supervisory attention, we developed an alternative model to identify safe-and-sound banks that potentially are headed for financial distress. Such a model could help supervisors allocate scarce on- and off-site resources by pointing out banks not currently under scrutiny that need watching.

It is possible, however, that our alternative model improves little over the current SEER framework. All three models—the SEER risk rank model, the SEER rating model, and our downgrade model—produce ordinal rankings based on overall risk. If the financial factors that explain CAMELS downgrades differ little from the financial factors that explain failures or CAMELS ratings, then all three models will produce similar risk ratings and, hence, similar watch lists of one- and two-rated banks.

We find only slight differences in the ability of the three models to spot emerging financial distress among safe-and-sound banks. In out-of-sample tests for 1992 through 1998, the watch lists produced by the downgrade model outperform the watch lists produced by the SEER models by only a small margin. We conclude that, in relatively tranquil banking environments like the 1990s, a downgrade model adds little value in off-site surveillance. We caution, however, that a downgrade model might prove useful in more turbulent banking times.

The Controversy Over Free Trade: The Gap Between Economists and the General Public

Cletus C. Coughlin

In contrast to their divergent opinions on many public-policy issues, most economists strongly support free trade policies. Nonetheless, there is substantial public opposition for such policies—from the right as well as the left ends of the political spectrum. Because public opinion affects policy decisions, understanding why this gap exists is a first step in devising strategies to increase public support for free trade.[1] In light of arguments and evidence indicating that free trade yields substantial benefits, attempts to influence public opinion seem warranted.

In the next section I report survey information highlighting the gap between the views of economists and the general public on free trade policies. The primary focus of this paper is on the "whys" of this gap in the United States. After examining why most economists support free trade policies, I explore why free trade is controversial. To ensure that this discussion about controversial issues is of a reasonable length, I focus on trade arguments involving either labor or environmental issues. Next, I examine suggestions for increasing the support for free trade. A summary of key points completes the paper.

DIFFERING VIEWS ON FREE TRADE POLICIES

Surveys have consistently shown strong support among economists for free trade policies. In a 1990 survey of economists employed in the United States, Alston, Kearl, and Vaughan (1992) reported that more than 90 percent agreed generally with the proposition that tariffs and import quotas usually reduce general economic welfare.[2] This consensus mirrored the results of a similar survey in 1976.[3] Obviously, the 1990 results are now more than a decade old, but no compelling reason exists to expect that a similar survey today would yield substantially different results. In fact, Mayda and Rodrik (2001, p. 1) recently stated: "The consensus among mainstream economists on the desirability of free trade remains almost universal."[4]

On the other hand, the general public is not as strongly in favor of reducing trade barriers as economists. Based on answers to a question in a survey by the Chicago Council on Foreign Relations, it is clear that the general public in the United States has major reservations about free trade.[5] In response to a question in 1998 pointing out that the elimination of tariffs and other import restrictions would lead to lower prices but that certain jobs in import-competitive industries would likely be eliminated, only 32 percent of the general public were in favor of eliminating tariffs in this case. Meanwhile, 49 percent were more sympathetic to the argument that tariffs are necessary to protect jobs.[6]

Survey results presented in Scheve and Slaughter (2001a) suggest that Americans recognize both the benefits and costs of international trade. Large majorities of Americans think that freer trade generates benefits in terms of lower prices, increased product variety, and more innovation. On the other hand, a majority of Americans think that trade results in fewer jobs and lower wages for some segments of the labor force. Relative to economists, however, survey respondents tend to emphasize the costs rather than the benefits. For example, the 1999 Program on International Policy Attitudes survey asked whether free trade was a good idea because it could lead to lower prices and faster growth or a bad idea because it could lead to lower wages and lost jobs (University of Maryland, 2000). Survey respondents were nearly evenly divided, with 51

Cletus C. Coughlin is vice president and associate director of research at the Federal Reserve Bank of St. Louis. Heidi Beyer, Sarosh Khan, and Paige Skiba provided research assistance.

[1] See Blendon et al. (1997) for references showing that public opinion influences policy decisions.

[2] A sample of 1,350 economists employed in the United States was used. Each recipient was asked to indicate general agreement, agreement with provisos, or general disagreement with 40 propositions. The number of respondents was 464, a response rate of 34.4 percent.

[3] See Kearl et al. (1979) for details of this earlier survey.

[4] See Krugman (1997) for a similar opinion.

[5] See Rielly (1999).

[6] The remaining 19 percent either did not have an opinion or refused to answer.

percent saying free trade was a good idea and 44 percent saying it was a bad idea. Five percent did not know or refused to answer.

WHY ECONOMISTS SUPPORT FREE TRADE POLICIES

Underlying the consensus among economists on the desirability of free trade is the judgment that nations are better off with free trade than with policies restricting trade.[7] Trade can affect a nation's income and its economic well-being through numerous channels. For example, the reduction of trade barriers allows for gains stemming from (i) specialization and exchange according to comparative advantage, (ii) increasing returns to scale from larger markets, (iii) the exchange of ideas through communication and travel, and (iv) the spread of technology by means of investment and exposure to new goods. Numerous models have been developed that show how a nation benefits from free trade. Rather than discuss numerous models, I examine the key ideas that economists stress when discussing the gains from trade. I complete this section by discussing some studies that measure the gains/losses that are likely to accompany specific trade policies.

The Gains from Trade: A Historical View[8]

The most famous demonstration of the gains from trade appeared in 1817 in David Ricardo's *Principles of Political Economy and Taxation*. In his example, England and Portugal produce the same two goods, wine and cloth, and the only production costs are labor costs. The amount of labor (e.g., worker-days) required in each country to produce one bottle of wine or one bolt of cloth is listed below.

	Wine	Cloth
England	3	7
Portugal	1	5

Because both goods are more costly to produce in England than in Portugal, England is absolutely less productive in producing both goods than its prospective trading partner. Portugal has an absolute advantage in both wine and cloth. Intuitively, one might be inclined to conclude that absolute advantage eliminates the possibility of mutual gains from trade. Thus, a high productivity (i.e., high income) country could not engage in mutually beneficial trade with a low productivity (i.e., low income) country. Productivity is crucial in determining wages. In view of absolute advantage, workers in the country with higher productivity will receive higher wages. However, absolute advantage is irrelevant in whether trade can benefit both countries.

What is crucial is that the ratio of the production costs for the two goods is different in the two countries. In England, a bottle of wine will exchange for 3/7 of a bolt of cloth because the labor content of the wine is 3/7 of that of cloth. In Portugal, a bottle of wine will exchange for 1/5 of a bolt of cloth. Thus, wine is relatively cheaper in Portugal than in England and, conversely, cloth is relatively cheaper in England than in Portugal. Economists say that Portugal has a comparative advantage in wine production and England has a comparative advantage in cloth production.

The different relative prices provide the basis for both countries to gain from international trade. The gains arise from both *exchange* and *specialization*.

The gains from *exchange* can be shown in the following manner. If a Portuguese wine producer sells five bottles of wine at home, he receives one bolt of cloth. If he trades in England, he receives more than two bolts of cloth for five bottles of wine. Hence, he can gain by exporting his wine to England. English cloth producers are willing to trade in Portugal; for every 3/7 of a bolt of cloth they sell there, they receive just over two bottles of wine, which is better than the one bottle of wine they would receive in England. Overall, the English gain from exporting cloth to (and importing wine from) Portugal, and the Portuguese gain from exporting wine to (and importing cloth from) England. Each country gains by exporting the good in which it has a comparative advantage and by importing the good in which it has a comparative disadvantage.

Gains can also arise from *specialization*. Assume initially that each country is producing some of both goods. Suppose that, as a result of trade, 21 units of labor are shifted from wine to cloth production in England and that 10 units of labor are shifted from cloth to wine production in Portugal. This reallocation of labor does not change the total amount of labor used in the two countries; however, it causes the production changes listed on the next page:

[7] Irwin (1996, p. 8) summarizes the history of this consensus as follows: "The case for free trade has endured, however, because the fundamental proposition that substantial benefits arise from the free exchange of goods between countries has not been overshadowed by the limited scope of various qualifications and exceptions."

[8] The bulk of this section appeared in Coughlin, Chrystal, and Wood (1988).

	Bottles of Wine	Bolts of Cloth
England	–7	+ 3
Portugal	+ 10	–2
Net	+ 3	+ 1

The shift of English labor causes cloth production to increase by three bolts and wine production to decline by seven bottles. Meanwhile, the shift of Portuguese labor causes cloth production to decrease by two bolts and wine production to increase by ten bottles. Overall, the production of both goods increases. This increased output of three bottles of wine and one bolt of cloth allows both countries to increase their consumption of both goods. Thus, specialization due to trade based on comparative advantage provides mutual benefits.

The Gains from Trade: Selected Developments Since Ricardo

Not surprisingly, trade theory has progressed since Ricardo. Some of the developments provide alternative explanations of comparative advantage, while others use different explanations of trade flows.

The most well-known alternative explanation of comparative advantage is the Heckscher-Ohlin model of international trade. This model is based on (i) the fact that countries differ from each other in terms of their productive resources (e.g., labor, capital, natural resources) and (ii) the fact that goods are produced using different proportions of those resources.

To illustrate the theory, assume two countries, China and Japan; two productive resources, labor and capital; and two goods, automobiles and clothing. Assume further that China's endowment of labor relative to capital exceeds that of Japan. In this case China is relatively well endowed with labor. Conversely, Japan is relatively well endowed with capital. Thus, one should expect that the price of labor relative to capital would be lower in China than in Japan.

Next, assume that in the production of clothing the use of labor relative to capital is greater than in the production of automobiles. In this case, clothing is produced by relatively labor-intensive methods and, conversely, automobiles are produced by relatively capital-intensive methods.

The Heckscher-Ohlin theory states the following: A country will be able to produce a good at a relatively lower cost if its production requires a relatively larger proportion of a relatively abundant resource in that country. (That is, a relatively abundant resource would be a relatively less expensive factor of production.) In the present example, this implies that China should have a comparative advantage in clothing and Japan should have a comparative advantage in automobiles. As in the Ricardian case, the different relative prices provide the basis for both countries to gain from international trade by means of exchange (i.e., Japan will export automobiles and import clothing and China will export clothing and import automobiles) and specialization (i.e., Japan will increase its production of automobiles and China will increase its production of clothing).[9]

An appealing feature of the Heckscher-Ohlin model is that it can generate insights into the political economy of trade policy. The preceding discussion suggests that allowing for free trade sets in motion a number of price changes. Specifically, the relative prices of goods in the two countries should tend to equalize, as well as the prices of the productive resources. In the two-country, two-good, two-resource world discussed above, the payments to one factor in a specific country will rise and the payments to the other factor will fall.

The Stolper-Samuelson theorem states that free international trade benefits a country's abundant resource and harms that country's scarce resource. In the preceding example, this means that capital will benefit and labor will be harmed in Japan. Meanwhile, labor will benefit and capital will be harmed in China.[10] As a result, it is easy to see why labor in Japan would be opposed to the reduction of trade barriers with China and that capital would support such a change. Later in the paper I use the Stolper-Samuelson theorem in the context of U.S. trade policy.

The Heckscher-Ohlin model focuses on interindustry trade. This trade exists when a country exports goods produced by one industry in exchange for goods produced by another industry in a second

[9] In contrast to the Ricardian assumption of constant opportunity costs, the Heckscher-Ohlin model allows for increasing opportunity costs. In many cases, increasing opportunity costs, which imply that costs per unit increase as more of a good is produced, are more realistic than constant opportunity costs.

[10] The intuition is straightforward. Prior to free trade, labor in Japan is relatively scarce and, thus, wages tend to be high. With free trade, the relative scarcity of labor is reduced by the fact that Japanese consumers can buy the labor-intensive good at a lower price from China. Thus, there is downward pressure on the price of labor in Japan. Similar reasoning can be applied to explain why capital in Japan benefits from free trade.

country. For example, the United States exports machinery to China in exchange for clothing. A common feature of the trade between industrialized countries is that they export and import similar types of products, which is known as intra-industry trade. For example, industrialized countries export and import different models of automobiles. Such trade likely requires explanations other than those based on comparative advantage. One explanation revolves around increasing returns to scale, which are said to exist when an identical percentage increase in the use of each productive input causes an even larger percentage increase in output. For example, if the use of each input were increased by 10 percent, output would increase by more than 10 percent. If increasing returns exist, then the cost per unit for the firm (industry) declines as its output increases.

In a world with increasing returns to scale, benefits from free trade arise because removing trade barriers allows a country to specialize in industries where average costs decline as output expands. Another view of this phenomenon is that productivity in the industry increases as more resources are utilized. These productivity increases are an important source of the gains from trade.

The existence of increasing returns to scale complicates the analysis of international trade by forcing the consideration of market structures other than perfect competition and raises the possibility that both countries do not gain from trade.[11] Overall, however, recent theoretical developments have likely strengthened the case for an open trading system by highlighting three sources of gains from trade. First, as highlighted in the preceding paragraph, as the market potentially served by firms expands, there are gains associated with declining per-unit production costs. A second source of gains results from the reduction in the monopoly power of domestic firms, who face increased pressures from foreign competitors to produce output demanded by consumers at the lowest possible cost. The third gain is that consumers enjoy increased product variety and lower prices.

The Gains from Trade: A Graphical View

Many of the key ideas discussed previously can be illustrated graphically. For space reasons I limit my focus to the static gains from trade by using a partial equilibrium approach.[12] Static gains refer to one-time benefits of reducing trade barriers that arise as national (domestic) prices move closer to global (world) prices. The price changes stemming from the liberalization of trade cause productive resources to be reallocated and consumption patterns to change, which result in the gains from specialization and exchange identified by Ricardo.

The illustration of the static gains from free trade using partial equilibrium analysis assuming perfectly competitive markets is straightforward.[13] As discussed previously, different relative prices for the same good in two countries provide a fundamental reason for international trade. If the price in the United States is higher than the price abroad when no trade is allowed, then the good will be imported into the United States when free trade is allowed.[14] On the other hand, if the price in the United States is lower than the price abroad when no trade is allowed, then the good will be exported from the United States when free trade is allowed. Consequently, two cases—one in which the good is imported into the United States and the other in which the good is exported from the United States—are examined.

In the first case, the price of a hypothetical good abroad is assumed to be lower than that in the United States. In Figure 1 the lines S_{US} and D_{US} are the U.S. supply and demand curves for the hypothetical

[11] A closely related issue that requires the consideration of imperfectly competitive market structures is strategic trade policy. When a small number of firms from different countries compete internationally (e.g., the aircraft industry), the theoretical case for free trade becomes somewhat murky. Government subsidies to a domestic firm can affect the behavior of foreign firms so as to benefit the subsidizing country. However, the theory of strategic trade policy, because many of its policy implications hinge on key assumptions, does not provide a strong enough case to alter substantially economists' opinions about free trade. See Chapter 14 in Irwin (1996) for a summary of this issue.

[12] The static gains from trade are the increases in economic well-being, with fixed levels of productive resources and technology, accruing to a nation as it changes from a policy of allowing no international trade to a policy of free trade. A partial equilibrium approach focuses on how price adjusts to equate quantity supplied with quantity demanded in a single market. The prices of all other goods and resources are assumed to remain unchanged. Alternatively, a general equilibrium approach examines the simultaneous determination of prices and quantities in all markets in an economy.

[13] A market is perfectly competitive if (i) there are many firms producing the good, each with a small market share; (ii) all firms produce a homogeneous product using identical production processes; (iii) all buyers and sellers possess perfect information; and (iv) firms can enter and exit the industry costlessly.

[14] Zero transportation costs are assumed to simplify the analysis. In addition, the foreign good is assumed to be a perfect substitute for the domestically produced good. Such an assumption is unlikely to apply to most traded goods, especially manufactured goods. Assuming foreign and domestically produced goods are imperfect substitutes complicates the analysis but does not alter the basic welfare effects. See Husted and Melvin (2001, pp. 180-82).

Figure 1

The Gains from Trade: The United States as an Importer

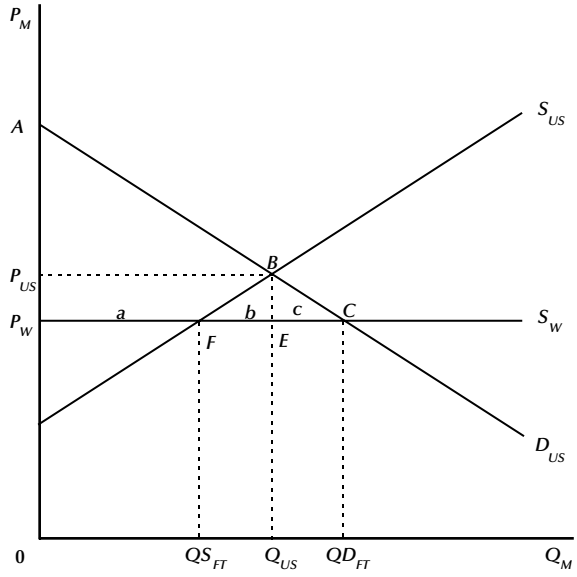
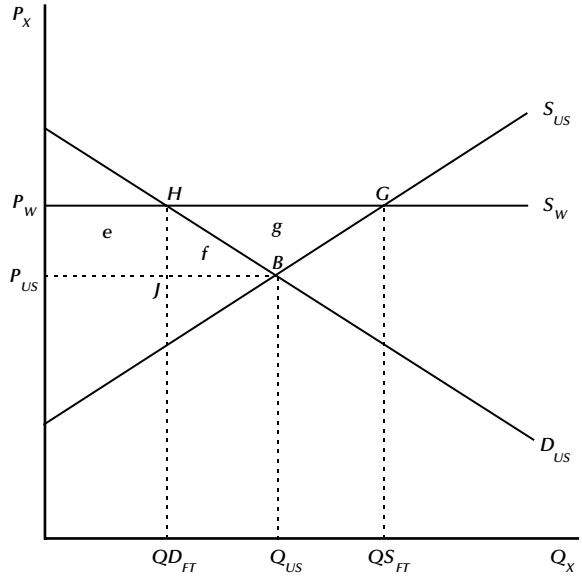


Figure 2

The Gains from Trade: The United States as an Exporter



good. Their intersection at B results in the equilibrium values for price, P_{US}, and quantity, Q_{US}, of the good. Meanwhile, S_W is the supply curve abroad. This curve, represented by a horizontal line, is based on an assumption that U.S. purchases will not affect the price abroad, which in this case is P_W. If one allows for free trade, this lower price abroad has two effects in the United States. First, U.S. consumers will increase their purchases of this good from Q_{US} to the free trade level of QD_{FT}. Second, U.S. producers will decrease their production of this good from Q_{US} to the free-trade level of QS_{FT}. U.S. purchases in excess of U.S. production (i.e., QD_{FT} less QS_{FT}) reflect the quantity of imports.

The lower price simultaneously benefits the U.S. consumers of this product and harms the U.S. producers of this product, a fact that can be used to explain why a free-trade policy is controversial. The magnitude of these gains and losses can be seen in Figure 1 using the concepts of consumer and producer surplus.[15]

First, we look at consumers, who gain in two ways. Prior to free trade, consumers purchased Q_{US} at a price per unit of P_{US}. With free trade, they pay the lower price per unit of P_W for Q_{US}. This gain in consumer surplus is represented by the rectangle $P_{US}BEP_W$. In addition, consumers gain because the lower price induces consumers to increase their purchases from Q_{US} to QD_{FT}. This additional increase in consumer surplus is represented by the triangle BCE. Thus, the total gain for consumers is the area $P_{US}BCP_W$ or, using lower case letters to represent specific areas, $a + b + c$.

Analogously, producers lose because of the lower price per unit they receive for their output, QS_{FT}, and the contraction of production from Q_{US} to QS_{FT}. Thus, the total loss incurred by producers is the area $P_{US}BFP_W$ or a. Overall, the United States gains because the consumer gains exceed the producer losses by $b + c$.

The preceding analysis can also be used for the case when the good is exported from the United States under free trade. The key modification of Figure 1 to create Figure 2 is that the price of the good prior to free trade is higher abroad than in the United States. The horizontal supply curve abroad, S_W, is based on the assumption that U.S. production will not affect the world price. Consequently, if one allows for free trade, the higher price abroad has two effects in the United States. First, U.S. consumers

[15] Consumer surplus is the difference between the amount consumers are willing to pay to purchase a given quantity of goods and the amount they have to pay to purchase those goods. Producer surplus is the difference between the price paid in the market for a good and the minimum price required by an industry to supply the good.

will decrease their purchases of this good from Q_{US} to the free-trade level of QD_{FT}. Second, U.S. producers will increase their production of this good from Q_{US} to the free-trade level of QS_{FT}. U.S. production in excess of U.S. purchases (i.e., QS_{FT} less QD_{FT}) reflects the quantity of exports.

The higher price simultaneously harms the U.S. consumers of this product and benefits the U.S. producers of this product. U.S. consumers lose because with free trade they are paying a higher price per unit, P_W versus P_{US}, for a smaller quantity of the export good, QD_{FT} versus Q_{US}. The reduction in consumer surplus is represented by the area $P_{US}BHP_W$ or $e+f$. Meanwhile, U.S. producers benefit from the higher price they receive for their prior output. In addition, they receive increased producer surplus as they expand production from Q_{US} to QS_{FT}. The total gain for producers is the area $P_{US}BGP_W$ or $e+f+g$. Overall, the U.S. benefits because the producer gains exceed the consumer losses by g.

The preceding partial equilibrium analysis is suggestive of the gains that the United States would generate as it moved from self-sufficiency to free trade. Obviously, the transition from self-sufficiency to free trade would set in motion numerous price changes. A general equilibrium approach allows for the simultaneous determination of prices and quantities in numerous markets. However, this theoretical advantage comes at the cost of increasing complexity in illustrating the gains from free trade; such an approach is not essential in this paper.[16]

The Dynamic Gains from Free Trade

Free trade can also contribute to economic growth, which is another source of gains. Such dynamic gains are potentially more important than the static gains. Most economic models suggest that trade liberalization will have a positive effect on economic growth.[17] An economy grows over time as a result of increases in its productive resources or technological innovation; both developments increase the capacity of an economy to produce goods and services. In addition, reducing trade barriers might increase competitive pressures that would force the efficient use of a nation's resources. Economic theory suggests a number of routes by which freer trade can stimulate growth.

One route is through increased savings that ultimately fund investment spending. Such spending increases the amount of capital. As argued previously, trade raises the level of real income, some of which can be saved. This higher level of savings translates into a greater availability of funds for investment spending. Free trade also allows the possibility for a country to borrow the savings of other countries. When a country imports more than it exports, a country is effectively borrowing funds from the rest of the world. If these funds are being used to finance the imports of capital goods, then a country's capital is increased.

A country, however, need not run trade deficits to import capital goods. When a country imports capital goods in exchange for consumer goods, then its productive capacity increases. This productive capacity allows for subsequent increases in output.

A related idea, stressed by Richardson (2001), is that free trade increases the possibility that a firm importing a capital good will be able to locate a supplier who will provide a good that more nearly meets its specifications. The better the match, the larger is the resulting increase in productivity, which ultimately translates into higher incomes.[18]

International trade may also spur the diffusion of technology by increasing the commercial contacts between employees in firms from different countries.[19] Such interactions serve to transfer information about new products and production processes. Of course, formal transactions may also facilitate the transfer of technology. Licensing is a common practice that allows the international transfer of technology. In addition, technology is embodied in new capital equipment. Thus, freer international trade facilitates the transfer of technology internationally and spurs economic growth.

Another potential route for economic growth results from the competitive pressures associated with international trade. Opening a country's markets to foreign firms tends to reduce the market power of domestic firms. For example, domestic monopolists are subjected to competitive pressures. As a result, the domestic firms are forced to become more effi-

[16] See Husted and Melvin (2001, Chap. 4).

[17] Whether the growth effect is temporary or permanent is closely related to whether an endogenous growth model or a neoclassical growth model is more nearly "true." In both types of models, a trade policy change can affect growth by altering either the accumulation of productive resources or technological progress.

[18] The same reasoning pertains to the gain for a consumer in finding a good that more nearly matches his/her preferences.

[19] Richardson (2001) notes that increased trade generates externalities by producing information about foreign markets and customers that spreads from those involved in international trade to those who are not. Such information can lower the cost of international trade and induce new firms to become involved.

cient or else they perish. Either way, a nation's productive resources will be used more efficiently in producing the goods that consumers desire.

A final route is related to the prior discussion suggesting that, as international trade expands the size of a market that firms face, firms might be able to exploit economies of scale. Recall that increased output at lower per unit cost is a clear-cut gain. Moreover, the larger market size might also spur research and development spending because the spending can be spread over larger levels of output. If successful, the spending would increase the productive capacity of the country.[20]

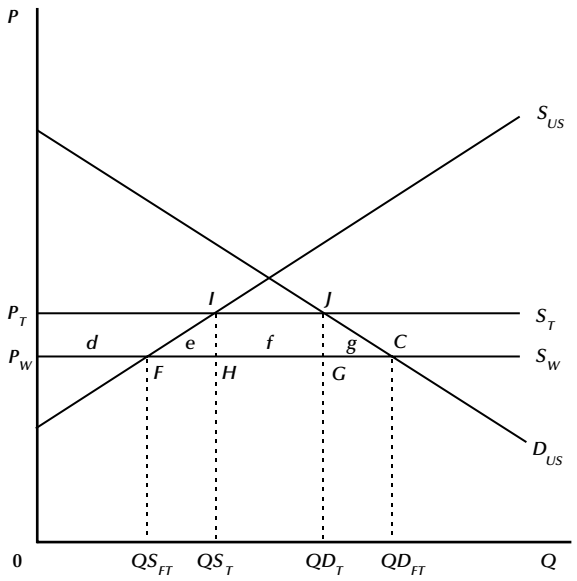
Empirical Studies of the Gains from Trade and the Losses from Protectionist Policies

The preceding discussion of international trade theory provides many reasons why economists support free trade policies. Empirical studies provide additional reasons. As discussed previously, a fundamental proposition is that international trade allows a country to achieve a higher real income than would otherwise be attained. Empirical evidence tends to confirm this proposition. For example, Frankel and Romer (1999) find that the impact of trade on income in 1985 is positive; however, in their study the precise impact is uncertain. Increasing the ratio of trade to gross domestic product by 1 percentage point raises per capita income by between 0.5 and 2 percent. Irwin and Terviö (2000), in an extension of Frankel and Romer, find that the impact of trade on income is positive for various periods in the twentieth century. These results suggest, at a minimum, that policies restricting international trade can result in substantial costs in terms of actual per capita income falling short of potential per capita income.

Additional empirical evidence focused directly on the issue of free trade has also been generated. Numerous estimates of the static and dynamic costs/benefits using partial as well as general equilibrium approaches have been produced assessing the consequences of trade policy changes. Using a partial equilibrium approach, it is easy to illustrate the effects of a trade policy change via supply and demand curves. Figure 3 shows the supply and demand curves for a hypothetical good imported into the United States that is subject to a tariff. Identical to Figure 1 the free trade results reveal, given the free trade price of P_W, U.S. consumption of

Figure 3

The Effects of a U.S. Tariff



QD_{FT}, production of QS_{FT}, and imports equal to the difference between QD_{FT} and QS_{FT}. Assume a tariff is imposed, causing the price in the United States to increase to P_T. The price in the United States now exceeds the price in the world by the amount of the tariff, P_WP_T.

The higher U.S. price causes consumer purchases to decrease from QD_{FT} to QD_T, production to increase from QS_{FT} to QS_T, and imports to decrease from $QS_{FT}QD_{FT}$ to QS_TQD_T. The imposition of the tariff causes consumers to lose $d + e + f + g$, while producers gain d. Thus, domestic producers are protected from foreign competition at the expense of domestic consumers. One complication is that the government collects tariff revenue. This revenue, which can be viewed as a gain for the government, equals the tariff, P_WP_T, times the quantity of imports, QS_TQD_T. This revenue is represented by area f.

Overall, the United States loses because the losses of consumers, $d + e + f + g$, exceed the gains of producers, d, and of government, f. The net national loss is $e + g$. Area e is called a "deadweight production loss" and reflects the loss from inefficient (excessive) production, while area g is called

[20] The automobile industry illustrates many of the routes producing dynamic gains from trade, especially those stemming from the diffusion of technology, competitive pressures, and economies of scale. See Fuss and Waverman (1992).

Table 1

Welfare Effects of Liberalizing Trade in Certain U.S. Industries, 1990 (millions of dollars)

Industry	Tariff or equivalent	Consumer gain	Producer loss	Net national gain	Consumer gain per job lost (dollars)	Net national gain per job lost (dollars)
Ball bearings	11.0%	64	13	1	438,356	6,849
Benzenoid chemicals	9.0	309	127	10	>1,000,000	46,296
Canned tuna	12.5	73	31	10	187,179	25,641
Ceramic articles	11.0	102	18	2	244,019	4,785
Ceramic tiles	19.0	139	45	2	400,576	5,764
Costume jewelry	9.0	103	46	5	96,532	4,686
Frozen orange juice concentrate	30.0	281	101	35	461,412	57,471
Glassware	11.0	266	162	9	180,095	6,093
Luggage	16.5	211	16	26	933,628	115,044
Polyethylene resins	12.0	176	95	20	590,604	67,114
Rubber footwear	20.0	208	55	12	122,281	7,055
Softwood lumber	6.5	459	264	12	758,678	19,835
Women's footwear, except athletic	10.0	376	70	11	101,567	2,971
Women's handbags	13.5	148	16	13	191,462	16,818
Dairy products	50.0	1,184	835	104	497,897	43,734
Peanuts	50.0	54	32	22	136,020	55,416
Sugar	66.0	1,357	776	581	600,177	256,966
Maritime transport	85.0	1,832	1,275	556	415,325	126,049
Apparel	48.0	21,158	9,901	7,712	138,666	50,543
Textiles	23.4	3,274	1,749	894	202,061	55,175
Machine tools	46.6	542	157	385	348,329	247,429

NOTE: Tariffs are the primary protective device for the first 14 industries in the Table. Import quotas are used for dairy products, peanuts, sugar, and maritime transport. Voluntary export restraints are used for apparel, textiles, and machine tools.

SOURCE: Derived from Tables 1.2 and 1.3 in Hufbauer and Elliott (1994).

a "deadweight consumption loss" and reflects the loss from inefficient (too little) consumption.

Hufbauer and Elliott (1994) have generated estimates of the potential net national gains by industry, as well as the consumer gains and producer losses, if the United States were to liberalize trade in 21 industries. Table 1 reveals that the gains for consumers in the apparel industry would exceed $21 billion if protection were removed. Not surprisingly, a substantial portion of this gain would come at the expense of producers whose losses would by nearly $10 billion. The net national gain from liberalizing trade in the apparel industry would be $7.7 billion.

Additional perspective is provided by expressing the consumer and national gains relative to the job losses in the apparel industry resulting from the liberalization. The consumer gain per job lost is $139,000, and the net national gain per job lost is $51,000. What this means is that consumers were effectively paying an average of $139,000 for each job protected in 1990 in the apparel industry, an industry in which the average pay of a production worker was less than $15,000.

Clearly, the net national gains from liberalizing trade in the apparel industry exceed by a large amount the potential gains from liberalizing other

Table 2

Welfare Estimates of Liberalizing Trade in Highly-Protected Sectors, 1996 (millions of dollars)

Sector	Welfare increase
Simultaneous sector liberalization of all significant restraints	12,402
Individual liberalization:	
Textiles and apparel	10,376
Maritime transport (Jones Act)	1,324
Sugar	986
Footwear	501
Dairy	152
Ball and roller bearings, and parts	49
Frozen fruit, fruit juices, and vegetables	28
Costume jewelry and costume novelties	19
Leather gloves and mittens	16
Personal leather goods	14
China tableware	12
Ceramic tile	9
Cutlery	4

SOURCE: U.S. International Trade Commission (1999, Table ES-1).

industries. However, there are large gains that could be realized by liberalizing trade in a number of other industries. Net national gains exceed $500 million dollars in the textiles, sugar, and maritime transport industries. Moreover, the consumer gain per job lost in the sugar industry is $600,000 and the net national gain per job lost is $257,000. It is also noteworthy how much consumers can gain per job lost in other industries. In the benzenoid chemicals industry the consumer gain per job lost exceeds $1 million; in the luggage industry the consumer gain per job lost exceeds $900,000. In the latter case, the net national gain per job lost is $115,000.

A recent study by the U.S. International Trade Commission (1999) uses a general equilibrium approach to explore the consequences of liberalizing trade in industries subject to significant trade restrictions. Based on 1996 data, the simultaneous liberalization of all significant restraints causes a net national gain of $12.4 billion, as shown in Table 2. Given the results in Hufbauer and Elliott (1994), it is not surprising that the elimination of trade barriers in the textiles and apparel sector yields the majority of the gains. Nor is it surprising that the maritime transport, sugar, footwear, and dairy industries are the sources for the majority of the rest of the gains.

The preceding examples reveal the possibility of substantial gains by liberalizing trade in selected industries. Overall, however, U.S. trade policy can be characterized as "open" relative to the policies of other countries. Consequently, estimates of the gains from trade do not reflect a change from the prohibition of trade to free trade, but rather a change from some level of trade restriction to free trade. From this perspective it should not be surprising that, relative to total U.S. economic activity, the static gains from eliminating trade barriers (or the costs stemming from the existing trade barriers) are relatively small.[21] Of course, one might argue that gains exceeding $12 billion are still substantial.

The preceding discussion has been focused on unilateral reductions of U.S. trade barriers.[22] Hertel (2000) provides a quantitative assessment of the

[21] See Zarazaga (1999) for a survey of static models of unilateral trade liberalization. He concludes that the gains range from negligible to moderate.

[22] Unilateral trade liberalization is an alternative to negotiated reductions, which become very complex as the number of goods and services being discussed increases and as the number of countries involved increases. Jackson (1997) noted that 26,000 pages were used to list the results of the Uruguay Round, the most recent multilateral round that lasted more than eight years and involved more than 120 countries.

potential gains from trade liberalization under a new round of multilateral negotiations. Specifically, Hertel analyzed a worldwide, across-the-board elimination of protection in agriculture and in a subset of services—business and financial services and construction services—as well as the elimination of tariffs in manufacturing. He estimated that the gains in real income for Canada, Mexico, and the United States as a whole were 0.37 percent of their income. Because the size of the U.S. economy is substantially larger than either Canada's or Mexico's, it is reasonable to infer that the specific gains as a share of income for the United States are roughly 0.37 percent as well. Such a percentage is consistent with most estimates of the static gains for the United States.

In contrast to the findings concerning the static gains, one finds that the empirical literature assessing the relationship between trade policy and economic growth is far from definitive. Numerous problems with this empirical analysis preclude an unqualified conclusion.[23] Many studies, using different data sets, countries, and methodologies, have found that countries with more-open trade policies (i.e., those closer to free trade) tend to grow faster than countries with less-open trade policies.[24] For example, Sachs and Warner (1995, pp. 35-36) find

> a strong association between openness and growth, both within the group of developing and the group of developed countries. Within the group of developing countries, the open economies grew at 4.49 percent per year, and the closed economies grew at 0.69 percent per year. Within the group of developed economies, the open economies grew at 2.29 percent, and the closed economies grew at 0.74 percent per year.

However, Harrison and Hanson (1999) argue that the openness index used by Sachs and Warner does not measure trade policy only. Harrison and Hanson go on to show that the components most closely linked to trade policy in Sachs and Warner's index are not related to growth.[25] Thus, the results are sensitive to the measurement of trade policy.

To illustrate further this sensitivity in the measurement of trade policy, Harrison and Hanson replace Sachs and Warner's measures of tariffs and quotas with an alternative tariff measure. In this case, openness to trade has a significant impact on growth. A decrease in the tariff rate of 10 percentage points causes an increase in average growth in real per capita gross domestic product of 0.5 to 0.6 percent.

The question of the robustness of the relationship between openness and productivity growth is explored in detail in Edwards (1998). Using data for 93 countries, he found that the more open countries experienced faster productivity growth. This basic finding held up despite the use of different openness indicators, estimation techniques, time periods, and functional forms.[26]

In summary, the empirical literature clearly indicates that liberalizing trade in highly protected industries is likely to yield gains. Whether those gains are large is in the eye of the beholder. The evidence concerning the dynamic gains from trade reveals that economies that are more open are likely to grow faster. If the faster growth is long-lived, substantial increases in well-being can be generated.

WHY FREE TRADE IS CONTROVERSIAL

What Does Research Based on Self-Interested Behavior Reveal?

To understand the opposition to free trade, one must understand the preferences of individuals as they relate to the policy choices available to policymakers. Unfortunately, most economic research does not provide direct evidence on the preferences of individuals. Generally speaking, empirical research on the political economy of trade policy

[23] One problem is difficulty of measuring trade policy. The choice of indicators for "openness" is somewhat arbitrary. A second problem arises because free trade countries may also adopt simultaneously other policies that affect income and growth. Thus, a researcher cannot be certain that the estimated impact of the trade policy measure is capturing solely the impact of trade policy. A third problem is that growth may affect openness just as openness may affect growth. Estimating a single equation in which growth is affected by openness may yield a biased estimate.

[24] For a more thorough discussion of the empirical evidence on the relationship between trade policy and growth—including an extensive bibliography of relevant studies—see a 1997 report by the United States International Trade Commission.

[25] A frequently cited study showing no strong relationship between liberalizing trade and long-run growth is by Levine and Renelt (1992); however, they found a robust, positive relationship between investment and trade share that led them to conclude that trade reform may generate growth through increased capital accumulation.

[26] Empirical evidence on trade policy and growth consists primarily of cross-country analyses. Ideally, one would like to use a dynamic, general equilibrium model for a specific country. Zarazaga (2000) concludes that minimal progress has occurred in constructing and estimating such a model.

focuses on trade policy outcomes. Because representatives do respond to the economic interests of their constituents, these outcomes certainly depend on the preferences of individuals. However, there are a number of other factors that come into play, such as the influence of interest groups, the preferences of policymakers, and the institutional structure of government. These other factors preclude the researcher from making definitive statements about individual preferences.[27]

Nonetheless, the voluminous literature on the determinants of protection does provide some results suggestive of individual preferences. For example, protection received by an industry is higher when it is a labor-intensive, low-skill, low-wage industry. This suggests that individuals are willing to support trade restrictions to improve the job and income prospects of low-income workers.

A recent study by Scheve and Slaughter (2001b) focuses specifically on individual preferences. They find that the lower the skill level of a worker, measured by education or average occupational earnings, the stronger is the worker's support for new trade barriers.[28] This result is consistent with a Heckscher-Ohlin trade model in which the United States is well endowed with skilled labor. Recalling the prior discussion of the Stolper-Samuelson theorem, the movement to free trade would tend to increase the incomes of skilled labor. Meanwhile, the incomes of unskilled labor would fall further behind. Because less-skilled workers have experienced sharp declines in their wages relative to more-skilled workers, Scheve and Slaughter (2001a) argue that the differences in their attitudes toward free trade may reflect the different wage-growth experiences of these groups since the early 1970s.[29] Arguably, the poor labor-market results of low-skilled workers, both absolutely and relative to high-skilled workers, could be due to other factors such as technological changes favoring high-skilled workers.[30] Scheve and Slaughter argue that, regardless of the reasons for their poor labor-market experience, those with relatively less education and skill expect the labor-market results stemming from additional international trade flows to be harmful.

More generally, the public fails to see any broad-based gains from trade. For example, the University of Maryland (2000) survey of public opinion found that Americans viewed the benefits of trade as flowing to business rather than to themselves or to American workers in general. The difficulty of envisioning broad-based gains might simply reflect the difficulty of envisioning any gains. As discussed previously, the static gains for an average individual of implementing free trade for the United States are small. Moreover, it is likely difficult for non-economists to envision how free trade will spur economic growth that will improve their economic well-being. Thus, because they do not see personal benefits, it is easy to see why individuals lack enthusiasm about trade negotiations.

Other Perspectives: The Social Dimensions of Trade

A foundation of economic analysis is self-interested behavior. In the present context, this implies that individuals evaluate trade policy based on how their current well-being is affected without regard for national well-being. However, people act for various reasons, some of which are materialistic and some of which are humanitarian. The allowance for self-interested behavior beyond those satisfying material demands complicates economic analysis. Nonetheless, such motives might well be important in understanding the opposition to free trade policies.

Employment/Income Concerns. The survey information cited previously indicates one of the reasons that the general public remains reluctant to support the free trade policies espoused by most economists: concern about jobs, but not necessarily their own. One might view this reason as reflecting humanitarian motives. Kinder and Kiewert (1979) argue that voters are motivated by collective well-being as well as their own individual well-being. One manifestation of such preferences is reflected in an observation by Krueger (1990). She argued that U.S. residents who stand to gain from trade liberalization may oppose it, nonetheless, when there are identifiable losers.

---

[27] See Rodrik (1995) for further discussion of this and many other issues related to trade policy.

[28] Because economists, on average, are more highly educated than the general public, this finding produces another reason why the free trade views of economists differ from those of the general public.

[29] Actually, the focus of Scheve and Slaughter (2001a) is on globalization, which includes immigration and foreign direct investment as well as international trade. Those with relatively less education and skill expect the labor market results stemming from further globalization to harm their well-being. This interpretation based on economic self-interest is not widely accepted. The standard view is that the opposition consists of a combination of groups with varied interests, not all of which can be connected to their economic self-interest.

[30] See Richardson (1995) for an analysis of the controversy concerning trade and income inequality.

Note how such preferences conflict with the analysis underlying Figure 1. In Figure 1 a given value of losses suffered by producers were netted, dollar for dollar, against the larger value of gains received by consumers. However, it is possible that it is not that simple. For example, assume a change in trade policy that would cause a $105 gain for a high-income individual but a $100 loss for a low-income individual. Despite a net national gain of $5, it is possible that a third party might oppose such a change because the adverse effect for the low-income individual might be viewed as outweighing the beneficial effect for the high-income individual.[31]

In addition, there are short-run adjustment costs stemming from changes in trade policy that might generate opposition. Because some industries will reduce production, some workers will lose their jobs. Being unemployed, regardless of its length, is a noteworthy cost that generates opposition to proposed trade-policy changes from both those likely to be adversely affected and those who sympathize with them.

The sense of community highlighted by Kinder and Kiewert (1979) might well extend beyond U.S. borders. Evidence suggests that U.S. consumers care about the conditions of the workers in developing countries.[32] Elliott and Freeman (2001) concluded that the vast majority of people are willing to pay higher prices for items produced under better working conditions in developing countries. In addition, most Americans favor linking labor standards to trade.[33] The 1999 Program on International Policy Attitudes survey found that 93 percent of respondents felt that as part of international trade agreements countries should be required to maintain minimum standards for working conditions (University of Maryland, 2000). In this same survey, three-quarters of the respondents felt morally obligated to help workers faced with poor working conditions. Moreover, roughly the same percentage reported a willingness to pay $5 more for a $20 garment if they knew it was not made in a sweatshop.[34] Overall, most respondents found the arguments for minimum standards (that harsh conditions are immoral and that standards eliminate cost advantages due to exploitation) to be more convincing than the arguments against standards (that the standards might hinder exports and reduce jobs in developing countries, as well as impinge on national sovereignty).

Note, however, that self-interest might provide a reason for some to argue for the linking of labor standards with international trade. Even when differing labor standards are appropriate given the specific situations of individual countries (i.e., the benefits exceed the costs at the national level), differing labor standards do provide cost advantages to firms in countries with relatively low standards. These advantages cause competitive problems for firms in countries, such as the United States, with relatively high standards. Such competitive problems are especially pronounced for those firms and workers in labor-intensive industries. Thus, higher standards would serve the interests of those being harmed by the imports from low-cost competitors. Not surprisingly, countries with low standards view the proposals to link labor standards with trade measures as protectionist because such proposals would tend to eliminate some of the cost advantages possessed by the firms in these countries.

Environmental Concerns. Similar to linking labor standards to trade, sentiment exists for linking environmental issues to trade. A fundamental concern is that free trade will stimulate economic growth and that this growth will harm the environment.[35] This argument illustrates a basic source for conflict between free traders and environmentalists. Proponents of free trade want to remove governmentally imposed trade barriers so that markets can generate efficient results, while environmentalists see free trade as generating consequences that require additional governmental regulations.

The 1999 Program on International Policy Attitudes survey revealed that 77 percent of respondents felt there should be more international agree-

[31] In theory, the "winner" could compensate the "loser" for his losses and still be better off; however, this is very difficult to implement in the real world. Policies that attempt to reduce the costs incurred by the "losers" are discussed later.

[32] Issues involving child labor have provoked intense controversy. Basu (1999) noted that in 1995 at least 120 million children between the ages of 5 and 14 worked full-time. The number working rises to 250 million when part-time workers are included. Not surprisingly, the incidence of child labor is highest in developing countries and has been so for several decades.

[33] Labor standards are the norms and rules governing working conditions and industrial relations. Standards addressing the freedom of association (i.e., the right of workers to establish and join organizations of their own choosing), the right to organize and bargain collectively, and the abolition of forced labor are commonly viewed as core labor standards.

[34] Of course, the behavior suggested by survey responses need not coincide with actual behavior. Elliott and Freeman (2001) discuss evidence suggesting that people do behave in ways consistent with these survey results.

[35] As discussed later, economic growth does not necessarily lead to environmental degradation.

ments on environmental standards. Underlying this result is a belief by many that environmental problems, such as acid rain and greenhouse gases, are global in nature. Clearly, acid rain and greenhouse gases are international issues that require a solution among governments; however, many economists would argue that many environmental problems are domestic issues that require a national solution. Views of what constitutes a strictly domestic environmental problem and what constitutes an international one can differ.

Some of the concern about the environment, however, can be linked to U.S. jobs. For example, 67 percent of respondents felt that the absence of international environmental standards would threaten U.S. jobs, as well as the environment, because lower environmental standards abroad would make the United States a less competitive location and would induce U.S. companies to relocate. This view that diversity in environmental standards would affect the desirability of maintaining/locating production in the United States tends to make allies of U.S. companies, labor unions, and environmentalists. In terms of trade negotiations, this view requires that environmental regulations must be harmonized with, at least, existing U.S. standards prior to allowing for free trade. Many economists, however, would argue that domestic environmental problems should be handled nationally and that international differences in environmental standards are natural.

Generally speaking, the survey respondents did not support views on environmental issues based on either national sovereignty or fairness. Only 33 percent supported the view that each country should decide how to deal with environmental issues. Only 37 percent supported the view that, because the costs of complying with international environmental standards would vary across countries, such standards would be unfair for countries with relatively high compliance costs. The prevailing views in this survey likely conflict with views that most economists hold. For example, most economists would argue that a national problem requires a national solution and that the costs as well as the benefits of any proposed solution be considered.

Clearly, the protection of U.S. jobs underlies the environmental position of many. Nonetheless, there is evidence that, when faced with a trade-off between protecting the environment and increasing jobs and economic growth, a majority of Americans, 52 percent, chose protecting the environment. Of the remainder, 37 percent chose jobs and 10 percent viewed the environment and jobs as equally important.

BRIDGING THE GAP

Three approaches have been suggested to move public opinion toward supporting free trade. The first approach is to increase economic education on free trade. The second approach reduces the costs borne by those who are harmed by the implementation of free trade policies. In other words, those incurring job losses and wage reductions might be compensated to ameliorate these costs. As a result, those facing job and wage uncertainty related to proposed trade agreements, as well as those concerned about these individuals, might be more inclined to support trade liberalization. The third approach attempts to increase support for free trade by expanding the agenda encompassed by trade negotiations. By addressing additional issues, such as those of concern to labor and environmental interests, support for trade liberalization efforts may be increased.

Education

Because economists find the arguments for free trade to be convincing, they are inclined to think that increased economic knowledge would increase public support for free trade. Some evidence—admittedly sparse—supports this view. Research by Saunders (1980) and Gleason and Van Scyoc (1995) indicates that a college economics course has a lasting impact on the economic knowledge of adults. Walstad (1997) found that economic knowledge was directly related to one's opinion on various economic issues; moreover, the more economic knowledge one had, the more likely it was for the individual to hold an opinion that coincided with the opinion of most economists.

In terms of influencing public opinion, an important issue is how to communicate with those not likely to take an international economics course. Cass (2000) notes that economists' arguments for free trade are often at odds with public discussions. As discussed previously, economists focus on consumption; however, public discussions tend to focus on production. The economist stresses that free trade allows for increases in well-being because consumers can buy more and varied goods at lower prices. Meanwhile, public discussions frequently argue that exports are good, but imports are bad; exports support jobs, frequently well paying ones, but imports destroy domestic job opportunities.

Thus, the economist's view of imports as good rather than evil is ignored by many. Imports provide consumers with increased choices of items that might be of higher quality, lower price, or more suited to one's tastes than would otherwise be available. Exports help us buy imports, but our enjoyment comes from consuming goods rather than from producing goods. To point out the folly of viewing exports as good and imports as bad, nineteenth-century economist Frédéric Bastiat satirically wondered whether the best outcome would be for ships transporting goods between countries to sink.[36] As a result, countries could have exports without imports.

As noted here previously, the nature of the popular discussion tends to strengthen the arguments against free trade in relation to the arguments for free trade. Cass (2000) notes three types of asymmetry. The opposition to free trade is strengthened by its visual appeal. For example, when international trade is identified as the reason for a plant closure or a layoff, a picture of a closed plant can be provided or the consequences for a specific family can be told.[37] Meanwhile, the case for free trade is more difficult to present in concrete terms.

A closely related asymmetry is that the intensity of the argument likely favors the opponents of free trade. The opposition to free trade comes from workers who may lose their jobs. It is easy to see why such a group would be passionately opposed to international trade. Conversely, the beneficiaries of free trade are likely to be more diffuse. Their individual benefits are more likely to be small and frequently hard to identify precisely. Thus, passionate support is unlikely on this side of the argument.

Finally, the arguments against free trade are more readily appreciated than those for free trade. For example, it is relatively easy to understand that competition as a result of imports makes it more difficult for a domestic company to generate profits. Moreover, the competition puts downward pressure on wages and causes layoffs. Arguments in favor of free trade that rely on comparative advantage and the gains from specialization and exchange are not likely to be very convincing, especially in light of the limited knowledge many citizens possess about how markets function.

Given the preceding obstacles of influencing the general public, economists must use approaches and arguments that overcome these obstacles. Roberts (2000) offers a number of suggestions for communicating with the "open-minded skeptic."

Frequently, proponents of free trade suggest that exports create jobs. On the other hand, opponents of free trade stress that imports destroy jobs. It is possible that the focus on jobs distorts one's view of free trade. Recall that the previously discussed survey asked the general public their views about eliminating tariffs by stating that prices would decline, but that certain jobs would likely be eliminated. No mention was made of the fact that jobs would also be created so that the net job effect would likely be negligible. The bottom line is that trade policy does affect the distribution of jobs, but is unlikely to affect substantially the net number of jobs.

Roberts also cautions against stating that free trade is good for everyone. It is not. Despite the argument that the removal of a tariff generates benefits to consumers that exceed the losses of producers, the producers as well as the workers who are adversely affected are not always compensated for their losses. Rather than duck this issue, it should be acknowledged. In addition, policies to assist those incurring losses, which are discussed later, could be stressed.

Because the costs are easier to see than the benefits, Roberts suggests the proponents of free trade attempt to make the gains concrete. Students in economics classes might be convinced of the wisdom of free trade policies using the economic theory and tools that economists find convincing, but the general public would probably ignore such a discussion. A compelling case likely requires an illustration of the gains from trade in the form of specific examples or reasonable hypothetical examples. As discussed previously, many individuals do not see how they gain from free trade or how they are harmed by trade restrictions. Expressing the gains of reducing trade barriers in terms of consumer gains (or national gains) per job lost is one way to argue convincingly. Another specific example is to show how per capita income in the United States would increase over a ten-year period if free trade led to an increased U.S. growth rate of 1 percent per year. In this case U.S. per capita gross domestic product in 2000 would have been more than $3,500 higher than its level of $35,400. Most individuals can appreciate the effect of a roughly 10 percent pay increase. Moreover, stressing the beneficial growth effects of free trade moves the focus from a winners-versus-losers focus to the possibility of everyone sharing in the benefits of increased growth.

However, the benefits of economic growth are

[36] This observation is cited in the *Economist* (2001).

[37] This asymmetry is referred to as an "identity bias" by Krueger (1990).

unlikely to convince some individuals and groups to support free trade. As international trade has become more important, its potential economic and social effects have increased. One consequence is increased demands that trade discussions encompass a broader range of economic and social issues. Moreover, Americans do not see the growth of trade as a key priority. They see international trade as a goal that should be balanced with other goals, such as protecting workers, the environment, and human rights.[38]

Not surprisingly, expanding the range of issues complicates trade negotiations. Resolving social issues is especially difficult because of the tradeoffs that are required to satisfy competing objectives—tradeoffs, in fact, for which policymakers lack precise information. Deardorff and Stern (2000) pose some of the most challenging tradeoffs. For example, although child labor may be deplorable, it is possible that the earnings may be necessary to keep the children alive. A cleaner environment is desirable, but maybe not if the cost pushes the poorest countries further into poverty. Human rights are valuable, but so is national sovereignty. Obviously, disagreements on the "right" balance are inevitable.

In view of this mixing of social issues with trade issues, educational efforts in support of free trade must address the concerns raised by environmentalists and others. In fact, strong arguments can be made that trade liberalization is consistent with the achievement of social objectives.

Bhagwati (1993) has demonstrated that the argument that free trade harms the environment can be handled directly.[39] Growth provides additional revenues for governments to pursue various objectives, including environmental protection.[40] How a specific country decides to spend its additional revenues depends on the relationship between increasing incomes and the demand for a better environment. Generally speaking, the wealthier a country, the greater is its demand for a better environment. However, demand is only part of the story. One must also consider how growth affects the production of pollution. Thus, the net effect on the environment depends on the type of economic growth. Grossman and Krueger (1993) found, using cities throughout the world, that sulfur dioxide pollution fell as per capita income rose beyond $5000. Thus, growth as a result of freer trade should tend to improve rather than harm the environment.

It is also possible to argue that international differences in environmental standards are natural and are not a justification for linking environmental issues with trade negotiations.[41] Different environmental standards for local pollution problems can be justified because they are necessary for economic efficiency.

Economic efficiency requires that pollution be reduced until the point at which the additional benefits of reducing pollution equal the additional costs. Numerous factors, two of which are highlighted, affect the level of environmental quality associated with economic efficiency.[42] Assimilative capacity, which is the capacity of the environment to reduce pollutants naturally, is one factor. Quite possibly, a less-industrialized country has greater assimilative capacity than a more-industrialized country because of less pollution in the past. Thus, it can tolerate a higher level of emissions than an industrialized country without increasing pollution levels.

A second factor likely to affect a country's level of environmental quality is its income level. A low-income country might put a higher value on the production of goods relative to environmental quality than a high-income country. This lower value on environmental quality leads to relatively lower environmental standards in the low-income country.

To summarize, international differences in environmental standards are natural and allow countries to use their productive resources efficiently. Forcing countries to have identical standards is a recipe for economic inefficiency.[43] Economic efficiency, however, might be of little concern to environmentalists. If so, then economic education is unlikely to be effective in convincing environmentalists to alter their opposition to reducing trade barriers. Some argue that the goal of environmentalists is to use trade policy to impose their

---

[38] For example, the University of Maryland (2000) survey of public opinion found 88 percent agreed that increasing international trade is a goal to be balanced against protecting workers, the environment, and human rights, even if the result was a slower growth of trade and the economy in general.

[39] See Butler (1992) for a discussion concluding that free trade and environmental policies can work together to generate worldwide economic growth and environmental quality.

[40] A similar argument can be made concerning child labor. For example, the growth resulting from free trade can provide the resources and opportunities to reduce the participation of child laborers in developing countries.

[41] A similar argument can be made in justifying differences in labor standards.

[42] See Butler (1992) for a more complete discussion of why countries choose different levels of environmental quality.

[43] Note that different regional environmental standards exist within the United States.

values on other countries.[44] In many cases their values are not widely accepted. For example, many environmentalists want to suspend the trading rights of countries that sanction the use of purse-seine nets in tuna fishing and leg-hold traps in trapping. It is clear that different people hold widely different views of the relative importance of, say, dolphins versus the economic livelihood of the Mexican fishing industry. In addition, as Bhagwati (1993) has noted, the inclusion of idiosyncratic values into trade negotiations opens the way for numerous conflicting demands as environmentalists favor dolphins, Indians have sacred cows, and animal-rights activists object to slaughterhouses. Such a scenario would result in dim prospects for reducing trade barriers.

Reducing the Cost for Those Harmed

As highlighted previously, changes in trade policy cause gains for some individuals and losses for others. Generally speaking, high-skilled workers in the United States tend to benefit relative to low-skilled workers when trade barriers are reduced. Those suffering job losses as a result can incur income losses, reductions in health and pension benefits, costs associated with relocating, and the psychological costs of losing a job. The trade adjustment assistance program, which is administered by the U.S. Department of Labor, allows workers who lose their jobs because of increased imports to receive unemployment compensation for an additional period beyond that received by other displaced workers.[45] In addition, trade adjustment assistance recipients can also participate in retraining programs plus receive out-of-area job search allowances and moving expenses.

Among the arguments to justify the trade adjustment assistance program is that the program reduces workers' lobbying efforts against trade liberalization. Even if voters are motivated by their perceptions of collective well-being and not simply their own individual well-being, trade adjustment assistance might increase support for free trade by both those who gain and those who lose. In effect, as Magee (2001) found, trade adjustment assistance payments compensate workers for lost tariff protection.

Despite disagreeing on numerous items, the Democratic and Republican members on the U.S. Trade Deficit Review Commission (2000) agreed that more resources should be allocated to trade adjustment assistance programs. Such a position is consistent with the general public's opinion that the U.S. government should do more to help workers adapt to changes caused by international trade.[46] A more effective trade adjustment program is likely to generate an increased willingness to support trade liberalization.[47]

Another proposal to ameliorate the problems faced by displaced workers and reduce the opposition to trade liberalization is to provide wage insurance.[48] As noted by the U.S. Trade Deficit Review Commission, many displaced workers, especially those with much tenure, suffer not only during the period between jobs but also after they become reemployed. For example, the weekly earnings of all reemployed workers fell 5.7 percent on average during 1995-97. Those displaced from high-tenure jobs experienced a wage decline of over 20 percent. Wage insurance would provide earnings supplements for a set period to workers who become reemployed at a lower wage.

Proponents of wage insurance, such as Kletzer and Litan (2001), argue that it provides an incentive for workers to find a new job quickly as contrasted with unemployment insurance, which provides an incentive to delay looking for work. For younger workers, the quicker reemployment might make it easier for them to acquire training and new skills that will make them more employable and productive over their working lives. For older workers, the wage insurance might allow them to reach retirement without lowering their standard of living or altering their retirement plans. On the other hand, Schoepfle (2000) raises concerns about the potential costs of wage insurance.[49]

---

[44] A similar statement pertains to certain labor groups.

[45] See Schoepfle (2000) for an overview of U.S. Department of Labor programs for dislocated workers and for a history of the trade adjustment assistance program since its passage in 1962.

[46] See University of Maryland (2000).

[47] Despite its political appeal, the effectiveness of the trade adjustment assistance program has been questioned. Decker and Corson (1995), Bohanon and Flowers (1998), and Marcal (2001) study the effectiveness of this program. See Richardson (2000a) for an identification of research relevant to redesigning labor-adjustment programs to increase their effectiveness.

[48] Job displacement can result from technological change, downsizing, restructuring, changes in demand, and changes in public policy (e.g., trade liberalization and environmental regulation).

[49] A proposal by Kletzer and Litan (2001) to provide wage insurance and health insurance subsidies for qualifying displaced workers upon reemployment was estimated to cost less than $4 billion. Obviously, the specifics of the program, such as who qualifies and the benefits provided, will affect the cost.

Expanding the Trade Agenda

During recent years many have argued that policymakers should expand the agenda for trade negotiations occurring under the World Trade Organization (WTO) and other bodies. Prior negotiations have produced substantial reductions in tariff barriers. One result is that the remaining trade barriers are in the most sensitive industries and involve the most complex issues. As discussed previously, sentiment is strong for linking labor and environmental issues with trade negotiations. What is unclear is whether such changes would ultimately increase the prospects for liberalizing trade. Expanding the agenda might provide negotiators with more opportunities for compromise; however, expanding the agenda might also bog down negotiations by introducing issues upon which compromise is very difficult. In fact, many have come to the conclusion that expanding the trade agenda would be detrimental to liberalizing trade in the United States and throughout the world.

A discussion by Brown (2000) highlights some of the challenges of linking labor standards with trade standards in the WTO.[50] The priorities of member countries are unlikely to coincide with each other or with the WTO. For example, the United States argues for rigorously enforcing high labor standards. On the other hand, developing countries desire minimal standards and enforcement because they fear the standards will provide a cover for protectionism. Meanwhile, the WTO may resist enforcing labor standards because they are not related to their original mission of fostering free trade. The bottom line is that such a linkage is not a promising approach for generating gains from trade.

Richardson (2000b) argues that the inclusion of a targeted set of "market-supportive" new issues offers a promising way to propel multilateral trade negotiations. In his view, expanding the negotiations to cover selected competition, technology, and labor policies would increase support by small businesses, technology users, and workers throughout the world. Moreover, such an expansion would increase the effectiveness of the market system.[51] Thus, both market enthusiasts and society "win." In a comment on Richardson's paper, Maskus (2000) raises the fundamental question as to whether the pressures arising from those concerned about the environment, labor rights, the impact of technological change, and globalization can be accommodated in a way that would allow the WTO to be effective.

Irwin (2000) answers this question negatively and, furthermore, expresses fears that both friends and foes of the WTO are pushing for changes in the organization's agenda that will prove detrimental to liberalizing trade. Friends would like to see the WTO expand its scope to set rules on various new trade issues—investment policy, competition policy, and electronic commerce, to name a few. Foes would like to see the WTO deal with labor and environmental regulations.[52] Irwin feels that expanding the agenda is a recipe for inertia and, even worse, will create "an international regulatory bureaucracy in Geneva that will provide full employment for trade lawyers rather than truly open up markets" (p. 355). A far better course would be for the WTO to focus on reducing border measures, especially those disrupting the free flow of agricultural and textile products.

Despite the concerns of Irwin and others, some business leaders in the United States appear to be softening their opposition to embedding social agendas in trade agreements.[53] Cooper (2001) reports that in a January 3, 2001, letter to Charlene Barshefsky, then U.S. Trade Representative, Caterpillar Inc. Chief Executive Glen Barton argued that labor and environmental standards were appropriate topics as part of future multilateral negotiations. Moreover, currently the Bush administration is searching for a way to respond to environmental protection and labor concerns during trade negotiations without allowing these issues to be used for protectionist purposes.

---

[50] See Esty (2001) for a discussion of bridging the gap between free traders and environmentalists.

[51] Richardson's subset of competition policies includes universal commitment to baseline disciplines concerning cartels, mergers, and anti-competitive behavior. The subset of technology policies includes distribution-oriented refinements in the WTO's intellectual property and trade-related investment agreements. The subset of labor policies includes worker agency services, specifically freedom for agents to bargain collectively on behalf of worker associations.

[52] Srinivasan (2000, p. 25) characterizes these opponents of free trade as the "unholy alliance of protectionists." This alliance consists of "industrial labor unions in rich countries, such as the American Federation of Labor and Congress of Industrial Organizations (AFL-CIO), masquerading as champions of the welfare and rights of workers (particularly child and female workers) in emerging countries, naive do-gooders who may be genuinely concerned with the welfare of children, and misguided environmentalists."

[53] Throughout the second half of the 1990s, U.S. involvement in trade negotiations has been hamstrung by Republican and Democratic conflict over linking free trade with labor and environmental standards. This political divide reflects business opposition and labor/environmental group support for linking trade negotiations with social issues.

CONCLUSION

The economic case for free trade is compelling for nearly all economists. Free trade policies enable free market forces to allocate resources to their most productive activities. This allows a nation to maximize the value of the goods and services produced within its borders. Free trade also allows consumers to allocate their incomes to maximize the value of the goods and services that they purchase and consume. Numerous models also suggest that the growth prospects of a nation are improved by using free trade policies. Moreover, the findings of empirical studies reinforce economic theory.

Despite these economic benefits, free trade policies are opposed by a large percentage of the U.S. public. The opposition consists of various groups, such as protectionists, labor unions, environmentalists, human rights activists, and economic nationalists. Clearly, the implementation of trade policies creates winners and losers. Not surprisingly, potential losers oppose free trade policies. Moreover, some oppose free trade because of their recognition that others will lose. This clash suggests that many in the general public differ from economists in how they weigh the costs and benefits of free trade policies. Others oppose free trade because of concerns that free trade contributes to the abuse of workers throughout the world, as well as to environmental degradation. Thus, these individuals will oppose reductions in trade barriers until these issues are addressed.

In view of the potential gains of free trade, an important question is how to reduce the opposition to free trade. A first step would be increased education concerning the benefits of free trade. Such a step is not controversial; however, to date, economists have been only moderately successful in spreading this good news to a large audience. Illustrating the gains from free trade using concrete and personal examples, as opposed to theoretical arguments, is one suggestion for convincing a larger audience.

A second step involves reducing the cost to the losers from free trade. A standard view is that the costs of liberalizing trade fall disproportionately upon less-skilled workers. Trade adjustment assistance is one policy option that has generated much political support. A more controversial policy is wage insurance. Questions about the cost-effectiveness of both policies, especially the latter, have been raised.

The most controversial step is to attempt to increase political support for free trade by expanding the issues covered in trade negotiations. Many Americans have real demands that the well-being of workers be safeguarded in developing countries and that the environment be protected. Whether these demands can be best served by linking them to trade agreements is controversial. Arguably, there are better ways to resolve many of these issues. The inclusion of labor and environmental issues in trade negotiations, as well as other issues, may or may not increase domestic political support. However, even if the inclusion of these other issues generated additional domestic support for free trade, it would not necessarily ensure success in negotiations to reduce trade barriers: foreign opposition to the inclusion of these issues, especially in developing countries, might negate any newly gained domestic support.

The fact that highly controversial steps are being suggested as necessary to propel trade negotiations points to one clear fact. Just as there are no quick fixes for the social issues that are increasingly linked to trade issues, there is no quick fix for generating political support for one of the few things that most economists agree upon—a nation's economic well-being is best served by free trade.

REFERENCES

Alston, Richard M.; Kearl, J.R. and Vaughan, Michael B. "Is There a Consensus Among Economists in the 1990's?" *American Economic Review*, May 1992, *82*(2), pp. 203-9.

Basu, Kaushik. "Child Labor: Cause, Consequence, and Cure, with Remarks on International Labor Standards." *Journal of Economic Literature*, September 1999, *37*(3), pp. 1083-119.

Bhagwati, Jagdish. "The Case for Free Trade." *Scientific American*, November 1993, *269*(5), pp. 42-49.

Blendon, Robert J.; Benson, John M.; Brodie, Mollyann; Morin, Richard; Altman, Drew E.; Gitterman, Daniel; Brossard, Mario and James, Matt. "Bridging the Gap Between the Public's and Economists' Views of the Economy." *Journal of Economic Perspectives*, Summer 1997, *11*(3), pp. 105-18.

Bohanon, Cecil E. and Flowers, Marilyn. "The Unintended Consequences of Trade Adjustment Assistance." *CATO Journal*, Spring/Summer 1998, *18*(1), pp. 65-74.

Brown, Drusilla K. "International Labor Standards in the World Trade Organization and the International Labor

Organization." Federal Reserve Bank of St. Louis *Review*, July/August 2000, *82*(4), pp. 105-112.

Butler, Alison. "Environmental Protection and Free Trade: Are They Mutually Exclusive?" Federal Reserve Bank of St. Louis *Review*, May/June 1992, *74*(3), pp. 3-16.

Cass, Ronald A. "The Trade Debate's Unlevel Playing Field." *Cato Journal*, Winter 2000, *19*(3), pp. 449-57.

Cooper, Helene. "Firms Rethink Hostility to Linking Trade, Labor Rights." *Wall Street Journal*, 2 February 2001, p. A12.

Coughlin, Cletus C.; Chrystal, K. Alec and Wood, Geoffrey E. "Protectionist Trade Policies: A Survey of Theory, Evidence and Rationale." Federal Reserve Bank of St. Louis *Review*, January/February 1988, *70*(1), pp. 12-29.

Deardorff, Alan V. and Stern, Robert M. "Introduction and Overview," in Alan V. Deardorff and Robert M. Stern, eds., *Social Dimensions of U.S. Trade Policies*. Ann Arbor: University of Michigan Press, 2000, pp. 1-18.

Decker, Paul T. and Corson, Walter. "International Trade and Worker Displacement: Evaluation of the Trade Adjustment Assistance Program." *Industrial and Labor Relations Review*, July 1995, *48*(4), pp. 758-74.

Economist. "Economics Focus: Frédéric Bastait," 21 July 2001, p. 64.

Edwards, Sebastian. "Openness, Productivity and Growth: What Do We Really Know?" *Economic Journal*, March 1998, *108*, pp. 383-98.

Elliott, Kimberly Ann and Freeman, Richard B. "White Hats or Don Quixotes? Human Rights Vigilantes in the Global Economy." Working Paper 8102, National Bureau of Economic Research, January 2001.

Esty, Daniel C. "Bridging the Trade-Environment Divide." *Journal of Economic Perspectives*, Summer 2001, *15*(3), pp. 113-30.

Frankel, Jeffrey A. and Romer, David. "Does Trade Cause Growth?" *American Economic Review*, June 1999, *89*(3), pp. 379-99.

Fuss, Melvyn A. and Waverman, Leonard. *Costs and Productivity in Automobile Production: The Challenge of Japanese Efficiency*. Cambridge: Cambridge University Press, 1992.

Gleason, Joyce and Van Scyoc, Lee J. "A Report on the Economic Literacy of Adults." *Journal of Economic Education*, Summer 1995, *26*(3), pp. 203-10.

Grossman, Gene M. and Krueger, Alan B. "Environmental Impacts of a North American Free Trade Agreement," in Peter Garber, ed., *The Mexico-U.S. Free Trade Agreement*. Cambridge, MA: MIT Press, 1993, pp. 13-56.

Harrison, Ann and Hanson, Gordon. "Who Gains from Trade Reform? Some Remaining Puzzles." *Journal of Development Economics*, June 1999, *59*(1), pp. 125-54.

Hertel, Thomas W. "Potential Gains from Reducing Trade Barriers in Manufacturing, Services and Agriculture." Federal Reserve Bank of St. Louis *Review*, July/August 2000, *82*(4), pp. 77-99.

Hufbauer, Gary Clyde and Elliott, Kimberly Ann. *Measuring the Costs of Protection in the United States*. Washington, DC: Institute for International Economics, 1994.

Husted, Steven and Melvin, Michael. *International Economics*. New York: Addison-Wesley, 2001.

Irwin, Douglas. *Against the Tide: An Intellectual History of Free Trade*. Princeton: Princeton University Press, 1996.

_____. "Do We Need the WTO?" *CATO Journal*, Winter 2000, *19*(3), pp. 351-57.

_____ and Terviö, Marko. "Does Trade Raise Income? Evidence from the Twentieth Century." Working Paper 7745, National Bureau of Economic Research, June 2000.

Jackson, John. *The World Trading System: Law and Policy of International Economic Relations*. Cambridge, MA: MIT Press, 1997.

Kearl, J.R.; Pope, Clayne L.; Whiting, Gordon C. and Wimmer, Larry T. "A Confusion of Economists?" *American Economic Review*, May 1979, *69*(2), pp. 28-37.

Kinder, Donald and Kiewert, D. Roderick. "Economic Discontent and Political Behavior: The Role of Personal Grievances and Collective Economic Judgments in Congressional Voting." *American Journal of Political Science*, August 1979, *23*(3), pp. 495-527.

Kletzer, Lori G. and Litan, Robert E. *A Prescription to Relieve Worker Anxiety*. International Economics Policy Brief Number 01-2, Institute for International Economics, February 2001.

Krueger, Anne O. "Asymmetries in Policy Between Exportables and Import-Competing Goods," in Ronald W. Jones and Anne O. Krueger, eds., *The Political Economy of International Trade: Essays in Honor of Robert E. Baldwin*. Oxford: Blackwell, 1990, pp. 161-78.

Krugman, Paul. "What Should Trade Negotiators Negotiate About?" *Journal of Economic Literature*, March 1997, *35*(1), pp. 113-20.

Levine, Ross and Renelt, David. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review*, September 1992, *82*(4), pp. 942-63.

Magee, Christopher. "Administered Protection for Workers: An Analysis of the Trade Adjustment Assistance Program." *Journal of International Economics*, February 2001, *53*(1), pp. 105-25.

Marcal, Leah E. "Does Trade Adjustment Assistance Help Trade-Displaced Workers?" *Contemporary Economic Policy*, January 2001, *19*(1), pp. 59-72.

Maskus, Keith. "Commentary." Federal Reserve Bank of St. Louis *Review*, July/August 2000, *82*(4), pp. 127-9.

Mayda, Anna Maria and Rodrik, Dani. "Why Are Some People (and Countries) More Protectionist than Others?" Working Paper 8461, National Bureau of Economic Research, September 2001.

Ricardo, David. *On The Principles of Political Economy and Taxation*. New York: Penguin, 1971.

Rielly, John E., ed. *American Public Opinion and U.S. Foreign Policy 1999*. Chicago: Chicago Council on Foreign Relations, 1999.

Richardson, J. David. "Income Inequality and Trade: How to Think, What to Conclude." *Journal of Economic Perspectives*, Summer 1995, *9*(3), pp. 33-55.

_____. "Comment," in Alan V. Deardorff and Robert M. Stern, eds., *Social Dimensions of U.S. Trade Policies*. Ann Arbor: University of Michigan Press, 2000a, pp. 126-28.

_____. "The WTO and Market-Supportive Regulation: A Way Forward on New Competition, Technological and Labor Issues." Federal Reserve Bank of St. Louis *Review*, July/August 2000b, *82*(4), pp. 115-26.

_____. "Exports Matter…And So Does Finance," in Gary C. Hufbauer and Rita M. Rodriquez, eds., *The Ex-Im Bank in the 21st Century: A New Approach?* Washington, DC: Institute for International Economics, 2001, pp. 55-79.

Roberts, Russell. "Speaking About Trade to the Open-Minded Skeptic." *Cato Journal*, Winter 2000, *19*(3), 439-48.

Rodrik, Dani. "Political Economy of Trade Policy," in Gene Grossman and Kenneth Rogoff, eds., *Handbook of International Economics*. Volume 3. New York: Elsevier, 1995, pp. 1457-94.

Sachs, Jeffrey and Warner, Andrew. "Economic Reform and the Process of Global Integration." *Brookings Papers on Economic Activity*, 1995, pp. 1-95.

Saunders, Phillip. "The Lasting Effects of Introductory Economics Courses." *Journal of Economic Education*, Winter 1980, *12*(1), pp. 1-14.

Scheve, Kenneth F. and Slaughter, Matthew J. *Globalization and the Perceptions of American Workers*. Washington, DC: Institute for International Economics, 2001a.

_____. "What Determines Individual Trade-Policy Preferences?" *Journal of International Economics*, August 2001b, *54*(2), pp. 267-92.

Schoepfle, Gregory K. "U.S. Trade Adjustment Assistance Policies for Workers," in Alan V. Deardorff and Robert M. Stern, eds., *Social Dimensions of U.S. Trade Policies*. Ann Arbor: University of Michigan Press, 2000, pp. 95-122.

Srinivasan, T.N. "Commentary." Federal Reserve Bank of St. Louis *Review*, July/August 2000, *82*(4), pp. 25-30.

U.S. International Trade Commission. *The Dynamic Effects of Trade Liberalization: An Empirical Analysis*. Washington, DC: U.S. Government Printing Office, October 1997.

U.S. International Trade Commission. *The Economic Effects of Significant U.S. Import Restraints*. Washington, DC: U.S. Government Printing Office, May 1999.

U.S. Trade Deficit Review Commission. *The U.S. Trade Deficit: Causes, Consequences and Recommendations for Action*. November 2000. < http://www.ustdrc.gov/reports/reports.html > .

University of Maryland, Program on International Policy Attitudes. *Americans on Globalization: A Study of*

Public Attitudes. College Park, MD, March 2000. <http://www.pipa.org/OnlineReports/Globalization/contents.html>.

Walstad, William B. "The Effect of Economic Knowledge on Public Opinion of Economic Issues." *Journal of Economic Education*, Summer 1997, *28*(3), pp. 195-205.

Zarazaga, Carlos E.J.M. "Measuring the Benefits of Unilateral Trade Liberalization Part 1: Static Models." Federal Reserve Bank of Dallas *Economic and Financial Review*, Third Quarter 1999, pp. 14-25.

_____. "Measuring the Benefits of Unilateral Trade Liberalization Part 2: Dynamic Models." Federal Reserve Bank of Dallas *Economic and Financial Review*, First Quarter 2000, pp. 29-40.

Not Your Father's Pension Plan: The Rise of 401(k) and Other Defined Contribution Plans

Leora Friedberg and Michael T. Owyang

The number of workers with a 401(k) plan grew from 7.1 million in 1983 to 38.9 million by 1993. The rapid diffusion of this new type of pension plan underscores a broader change in pension structure. Your father's pension plan was designed to give him a fixed income after retirement, but only if he stayed with his employer for 20 or 30 years; if he left early, he ended up with little or nothing. In contrast, your 401(k) or thrift plan is portable; the money accumulated in the account belongs to you when you leave your job, perhaps after a vesting period of a year or two. Consequently, the rise in 401(k) plans may have important implications for job tenure and worker mobility.

Your father's pension plan is called a *defined benefit* pension because the benefit—the money paid out of the pension—is set in advance by a formula that depends on salary and tenure. Employers fund defined benefit pensions by saving money over time, but the amount that they save does not determine the benefit that is paid out. The 401(k) and thrift plans that have become more common today are examples of *defined contribution* pensions. In these plans, the contribution—the money going into the pension—is set in advance, while the final value of the pension is uncertain and depends on the rate of return earned by accumulated contributions.

Pensions can be quite valuable, often worth $200,000 or more in present value at retirement for a worker who has stayed long enough with an employer. Moreover, different types of pensions can have important effects on job mobility, retirement, and saving decisions of workers.

Leora Friedberg is an assistant professor at the University of Virginia and a faculty research fellow at the National Bureau of Economic Research. Michael T. Owyang is an economist at the Federal Reserve Bank of St. Louis. The authors thank Douglas Fore, John M. Lewis, Robert Shimer, and two referees for their useful comments. Abbigail J. Chiodo provided research assistance. This research has been supported by TIAA-CREF, the Federal Reserve Bank of St. Louis, and the University of Virginia's Bankard Fund for Political Economy.

TRENDS IN PENSION COVERAGE

Over the last 20 years, defined contribution (DC) plans have supplanted defined benefit (DB) plans as the typical pension for many workers. Figures 1 through 3 highlight trends in pension coverage from 1983 through 1998.[1] Figure 1 shows that overall pension coverage declined from 67 percent of full-time employees in 1983 to 58 percent in 1998; it also shows trends in the percentage of full-time workers with a DB or DC plan. Figures 2 and 3 show the distribution of workers across pension type in 1983 and in 1998. In 1983, 40 percent of workers with a pension had only a DB plan, while 45 percent had both a DB and DC plan and only 15 percent had a DC plan. Figure 3 shows the dramatic decline in DB pension coverage: 20 percent had only a DB plan and 20 percent had both types, while 59 percent had only a DC plan.

In the rest of this article, we will describe how DB and DC pensions affect incentives of workers and employers. First, we discuss how pensions work and why they exist. Next, we describe the differences between DB and DC pensions, which are also enumerated in Table 1, and we analyze the impact of these differences on workers' incentives to stay in a job. Because of these incentive effects, the switch from DB to DC pensions may alter job tenure, worker mobility, and retirement patterns. Later, we discuss other differences between DB and DC pensions in administrative control and in the distribution of interest rate risk and other risks. These differences may influence saving behavior, stock market participation, and post-retirement consumption patterns.

HOW DO PENSIONS WORK?

The Structure of DB Pensions

A worker who qualifies for a DB pension will get an income flow until his death. The annual benefit is typically a proportion of either the worker's

[1] Pension statistics are reported by individuals in the Survey of Consumer Finances, which is computed for employees working 35 or more hours per week and weighted so that they are nationally representative. The SCF took place every three years from 1983 on, but the questions in 1986 were not asked in the same way, and the 1986 sample is not nationally representative.

Figure 1

Full-Time Employees

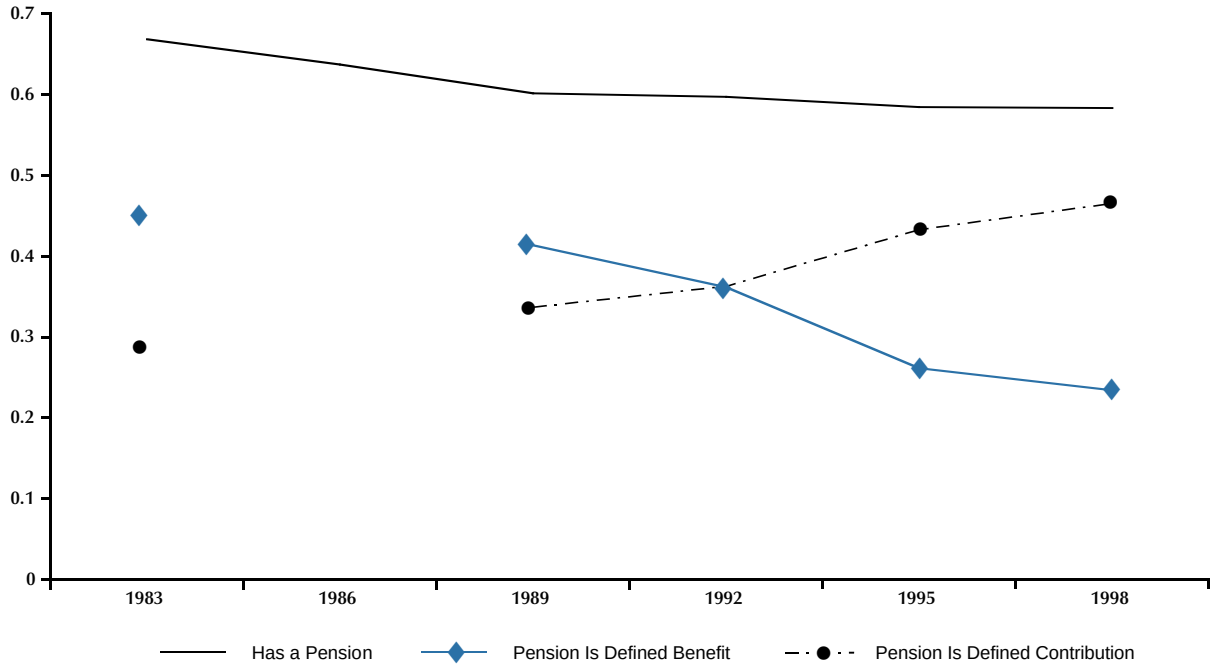
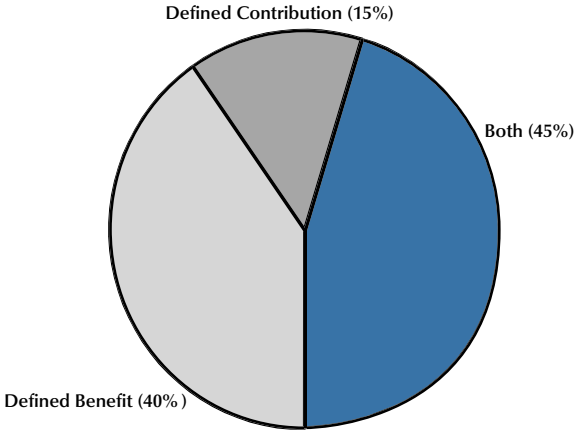


Figure 2

1983 Pension Structure for U.S. Labor Force

Full-Time Employees



NOTE: Includes only those individuals that specified pension type.

Figure 3

1998 Pension Structure for U.S. Labor Force

Full-Time Employees

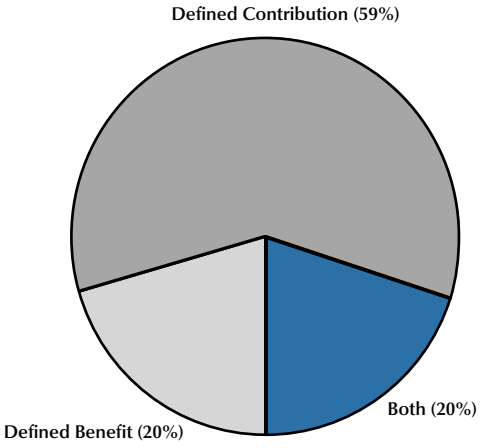
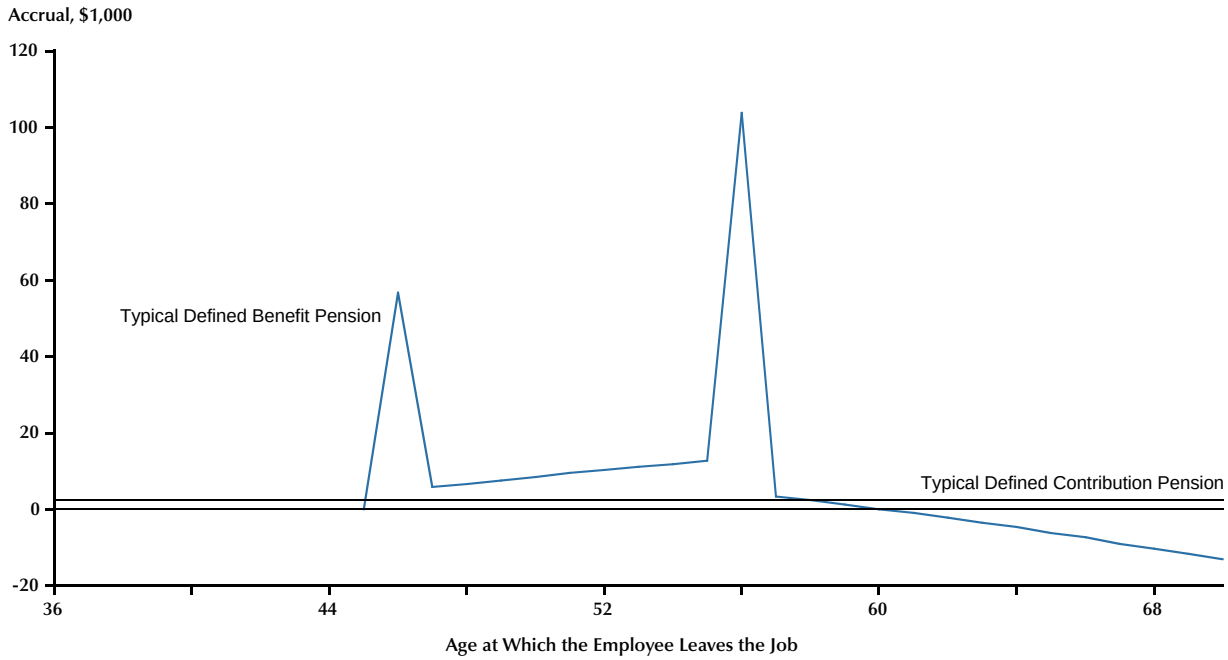


Figure 4

Accrual of Pension Wealth



average or final salary, with the proportion increasing with tenure. We can summarize the flow in present value terms: Pension wealth P_t is defined as the real present value of the worker's expected future pension benefits, actuarially discounted to incorporate uncertainty over the lifespan, if the job ends at time t. Pension wealth accrual is the change in pension wealth,

$$\frac{1}{1+\delta} P_{t+1} - P_t ,$$

discounted at rate $0 < \delta < 1$ if the employee works one additional year and then leaves.

The path of DB pension wealth accrual is characterized by occasional sharp spikes. Figure 4 shows pension wealth accrual in a typical DB pension plan.[2] The first spike, in this case worth about $60,000, occurs when the worker vests, that is, becomes eligible for future benefits. Maximum vesting dates of 10 to 15 years were established in 1974 and have since been lowered to 5 to 7 years. While vesting yields a claim to some future benefits, pension wealth continues to rise as the worker gains tenure.

Another spike, worth over $100,000, occurs when the worker reaches the plan's early retirement date (ERD), often at ages 55 to 60 with at least 20 years on the job. At the ERD a retiree can first begin to receive cash benefits. The early benefit is generally smaller than the full benefit available at the normal retirement date (NRD); if it is significantly smaller, then another spike in pension wealth occurs at the NRD. Frequently, though, the penalty for retiring early is small, as is the case in Figure 4. After the ERD or NRD, pension accruals swing around and turn negative because the worker gets little or no further increase in the benefit level but forgoes income by not retiring.

The Structure of DC Pensions

DC pensions are very simple: Funds go into an account, the worker can choose among a limited number of investment options, and the pension is portable after vesting. Therefore, workers can take

[2] These pension plans are based on information in the Health and Retirement Study (HRS) and have been slightly altered, as described in Friedberg and Webb (2000), to protect confidentiality. The HRS is a nationally representative study of households with at least one member aged 50 to 62 in 1992. The HRS obtained detailed information about pension plans directly from employers of survey respondents. Earlier DB plans were similar or had even sharper spikes; these patterns were documented in a series of papers by Kotlikoff and Wise (1985, 1987, 1989) and Stock and Wise (1990a, 1990b).

their accumulated funds with them when they leave their job. DC pension wealth after vesting is simply

$$P_t^{DC} = P_{t-1}^{DC}(1+r_t)+c_t\,, \tag{1}$$

where r_t is the rate of return earned on assets accumulated through the previous period and c_t is this period's contribution. Most DC pensions have vesting periods that are either immediate or less than two years.[3] Contributions are tax-deductible (as are a firm's contributions to fund a DB pension), and returns accumulate tax-free. Withdrawals from DC pensions, like DB pension benefits, are taxable.

The smooth path of DC pension wealth accrual shown in Figure 4 stands in stark contrast to the path of DB accrual. These differences directly affect incentives to stay in a job. DC pension accruals are largely age-neutral. Compared with a portable DC plan, a DB plan tends to reduce worker mobility for many years after a worker starts a job. Later on, it encourages retirement when pension accruals turn negative, whereas DC pension accruals remain positive and steady.

While the expected rate of return on DC pension wealth in Figure 4 is assumed to be constant, unpredictable changes in the actual return will shift the realized path of pension wealth accrual. For example, the sharp downturn in the financial markets in 2000-01 has reduced the value of DC pensions invested in stocks.[4] This interest rate risk introduces a new element of uncertainty as workers plan for retirement, so the widespread adoption of DC pensions may increase the volatility of retirement rates.

WHY DO PENSIONS EXIST?

Why is part of compensation deferred in the form of a pension? Individuals should prefer cash up front, if all else is equal; pensions exist because all else is not equal. The current theory of pensions was developed in a series of papers summarized in Lazear (1986), when DB pensions were the norm. In Lazear's view, DB pensions alter the incentives for long-term employment. We extend the theory to explain the choice between DB and DC pensions. DC pensions do not offer the same incentives for long-term employment, so they must serve an additional purpose perhaps by encouraging long-term saving. Thus, we focus on the incentives for long-term employment and for long-term saving.

A Stylized View of Pensions

A simplified version of DB pension wealth takes the following form:

$$P_t^{DB} = \begin{cases} 0 & \text{if } t < T \\ \bar{P} & \text{if } t \geq T \end{cases}. \tag{2}$$

A worker gets a fixed payment $\bar{P}$ if she stays in the job until some future date T.[5] DB pensions impose a risk on workers—that their job could end before time T and they would then lose their pension. Portable DC pensions do not impose this severance risk, so DB pensions must be more valuable, at least in expectation, for risk-averse workers to accept them. Thus, $\bar{P}$ can be written as

$$\bar{P} = E_0[P_T^{DC}]+\pi\,, \tag{3}$$

where P_T^{DC} is the value of a DC pension at the same future period T, $E_0[.]$ denotes the expectation at the outset of employment, and $\pi > 0$ is a premium associated with an enduring employment relationship, explained later.

Workers will only accept a DB pension if expected tenure, as well as the DB premium π, are high, relative to the interest rate risk implicit in DC plans. Some evidence on the size of this premium is available from the Health and Retirement Study, a longitudinal survey with detailed pension data for people aged 50 to 62 in 1992. We can use this information to compare DB and DC pension wealth if a worker retires at age 65. As defined earlier and detailed in Friedberg and Webb (2000), DB pension wealth is the present actuarially discounted value of expected future benefits (assuming a 3 percent discount rate and age- and gender-specific survival probabilities), and DC pension wealth is the estimated plan balance. For full-time employees in 1992, median pension wealth was $192,006 for workers with a DB plan and $99,105 for workers with a DC plan. Future workers will have somewhat higher DC pension wealth, as they spend more time in jobs with DC plans. Still, this gives an idea about the relative value of typical DB and DC pensions.

[3] From 30 to 35 percent of DC plans vest immediately and another 20 percent vest in two years or less, while most DB pensions take five years to vest, according to Mitchell (1999).

[4] See, for example, Crary (2001).

[5] In fact, before 1974, many DB pensions vested only at the NRD, according to Ippolito (1988).

The Value of Long-Term Employment

Lazear viewed pensions as a component of an implicit contract. Employers avoid explicit long-term contracts in order to preserve their flexibility, but they may nonetheless wish to encourage workers to stay or to devote greater effort to their job. Several possible explanations lie behind the "implicit contract" theory of pensions.

One reason an employer might encourage longer tenure is linked to the cost of searching for new workers. If searching for a new hire is costly, the decision of whether to search depends on a worker's expected tenure. Also, the relative ease or difficulty of transferring human capital investments can affect the firm's desire to have longer tenured workers. If human capital investments do not easily transfer to other workers or to other jobs, the sooner a worker is expected to leave, the more reluctant the employer will be to train that worker. The expectation of longer tenure then raises the rate of job training and results in higher productivity and profits, which the employer can share with the worker in the form of a DB pension.

Alternatively, in an efficiency wage framework, deferred compensation encourages workers to devote greater effort to their jobs. In some jobs it is difficult or costly for employers to monitor workers, who may shirk their responsibilities. Employers may find it useful in such cases to pay an "efficiency wage," which is higher than the prevailing wage in other jobs. This policy deters shirking, since a worker will lose her high-wage job if shirking is detected. Deferred compensation, in the form of a pension for instance, can also function as an efficiency wage, since a worker who shirks may lose her job before qualifying for that pension.

Pensions and the Incentive to Retire

The most common form of deferred compensation is the implicit promise of future wage increases, which also encourages longer tenure. If a fixed amount of wages are to be paid over some duration, wages can be structured to rise over time by paying a worker less than her marginal product early on and more than her marginal product later.

However, two problems arise with this element of an implicit long-term contract. First, it encourages workers to stay on *too* long. An aging worker will choose to retire when her marginal utility of leisure, which probably increases with age, exceeds her wage; the rising wage profile therefore leads her to retire later than the efficient date. Second, the rising wage profile creates an incentive for employers to violate the implicit long-term contract by firing workers, since employers will receive the benefits of the increased productivity sooner than workers. This credibility problem undermines the implicit contract; workers will not agree to a rising wage profile if they anticipate getting fired when their wages rise.

DB pensions help resolve both of these problems. A DB pension encourages the worker to retire at the "right" age, since the real value of her pension accruals turns negative after a certain point. And that condition, in turn, reduces the incentive of employers to fire older workers, which helps maintain the credibility necessary for the implicit contract. Again in this case, the employer may wish to fire a worker before the major spikes in pension wealth accrual. But, as argued above, that undermines the implicit long-term contract that promised workers a pay-off for long tenure. Furthermore, age discrimination laws and union rules make it difficult to fire older workers systematically.

Evidence for the "Implicit Contract" Theory of Pensions

Several pieces of evidence support the notion that DB pensions function as an implicit contract. For example, workers in jobs with DB pensions are less likely to leave their job. Among workers aged 30 to 54 in the 1998 Survey of Consumer Finances, those with a DB pension had average tenure of 12.5 years, compared with 10.1 for workers with a DC pension. The difference of 24 percent is statistically significant.

In addition, pensions are correlated with the timing of retirement. Using detailed data from both particular firms and national surveys, previous researchers have shown that workers tend to delay retirement until they reach the major spikes in DB pension wealth accrual at the early and normal retirement dates. The evidence suggests that DB pensions affect the timing of retirement by as much or more than Social Security (Stock and Wise, 1990a, 1990b; Samwick, 1998). The spread of DB pensions in the 1950s and 1960s coincided with a substantial decline in the average retirement age (Lumsdaine and Wise 1994). The median retirement age is now age 62, so a significant fraction of workers retire before they are even eligible to receive Social Security benefits. Much of this early retirement may be attributable to DB pensions.

Table 1

Summary of Pension Characteristics

	Defined benefit	Defined contribution
Key pension characteristics		
Determined in advance	Pension benefit	Pension contribution
Encourages longer tenure	Yes	No
Encourages optimal retirement	Yes	No
Encourages long-term saving	Yes	Yes
Contributions are tax-deferred	Yes	Yes
Differences during employment		
Pension design		
Median vesting period	5 years	0-2 years
Timing of pension wealth accruals	Most of pension wealth accrues late in career	Smooth accrual
Portable	No	Yes
Administrative control		
Controls investment of assets	Firm	Worker, firm*
Can borrow against assets†	—	Worker
Bears costs of administration	Firm	Worker, firm
Bears costs of regulatory compliance	Firm	Firm
Risk		
Interest rate risk	Firm	Worker
Underfunding risk	—†	Worker‡
Risk of early severance	Worker	—
Differences after employment		
Pension design		
Form of pension benefit	Annuity	Lump sum
Bequeathable	No§	Yes¶
Administrative control		
Controls investment of assets	Firm	Worker
Bears costs of administration	Firm	Worker
Bears costs of regulatory compliance	Firm	Worker#
Risk		
Interest rate risk	Firm	Worker
Lifespan risk	Firm	Worker

NOTE: *Employers choose which investment options to offer, usually including investment in company stock and several different mutual funds.

†Government regulations constrain both underfunding and overfunding of DB pensions by firms.

‡Contributions to 401(k) plans are voluntary and hence are subject to underfunding risk, but contributions to other types of DC plans are mandatory. Workers can withdraw DC assets in case of financial hardship or when separated from the firm; if they do so before age 59 1/2, they owe a 10 percent penalty to the government. Some firms allow 50 percent of worker contributions to the 401(k) (up to $50,000) to be used as collateral for loans with a term of no more than 5 to 10 years.

§Many DB pensions allow retirees a choice between a larger annual benefit payable until the retiree dies, or a smaller annual benefit payable until both the retiree and his or her spouse die.

¶Individuals are required to make regular withdrawals of assets from their DC plans beginning at age 70. If they do not, they or their heirs face tax penalties, limiting the extent to which DC assets can be saved for a bequest.

#As mentioned previously, individuals owe penalties for withdrawing funds when too young or too old.

Table 2

Rates of Pension Coverage and Job Training: Regression Results by Industry

	Dependent variables among those with a pension		
Independent variables	% in industry with a pension	% in industry with a DB plan	% in industry with a DC plan
% in industry who got job training	1.17 (0.28)	0.792 (0.192)	–0.560 (0.195)
Constant	–0.056 (0.122)	0.208 (0.083)	0.894 (0.085)
Adjusted R^2	0.733	0.728	0.548

NOTE: Standard errors are in parentheses. The sample includes seven one-digit industries (agriculture, mining/construction, manufacturing, retail/wholesale trade, finance/real estate/insurance/business and repair services, transportation/communication/other services, and public administration). Training rates are from the January 1991 Current Population Survey and are weighted to make them nationally representative; the national mean is 0.426 (0.002). Pension coverage rates, from the 1992 Survey of Consumer Finances, are also weighted.

In a similar vein, recent research by Friedberg and Webb (2000) shows that workers with DC plans are retiring later than workers with DB plans because of the differences in pension wealth accrual. The resulting change in the average retirement age is almost two years, controlling for other factors.

Other pieces of evidence are also consistent with the implicit contract theory of pensions. For example, DB pension coverage is more common in industries with high rates of job training. Recall that one reason for employers to encourage longer tenure is to gain more rewards from training their employees. Data on job training rates, aggregated for seven broad industrial sectors, can be matched to pension coverage rates in the 1992 Survey of Consumer Finances.[6] Regression results in Table 2 suggest a link. Industries with high training rates have more pension coverage; a 10 percent higher training rate is associated with an 11.7 percent higher pension coverage rate. Moreover, industries with high training rates also have significantly more DB and less DC coverage. A 10 percent higher training rate is associated with 7.9 percent higher DB coverage and 5.6 percent lower DC coverage, among those with pensions.

Most of the evidence which we have discussed here involves correlations between pension coverage and other variables (tenure, retirement, job training). The correlations do not prove causation, however. DB pensions might cause workers to stay in a job longer when young and retire early when old, for example; or employers might offer DB pensions to attract workers who want to do those things, along the lines suggested in the sorting model of Salop and Salop (1976). In either case, though, DB pensions help employers achieve the desired length of tenure.

Pensions and Personal Saving

The discussion above explains the purpose of DB pensions, but not necessarily DC pensions, which have little effect on tenure. Besides functioning as an implicit contract, deferred compensation obviously alters the path of consumption and saving for workers who face borrowing constraints. This should make pensions less appealing, according to conventional economic theory. However, recent research based on psychological evidence suggests that pensions may help workers save for retirement.

This notion is implicitly tested in most of the existing research on 401(k) plans, which seeks to determine whether people who save in 401(k) plans save more altogether. Conventional theory suggests a small positive response is likely, and a negative response is possible, because people would shuffle their assets and thereby gain a tax break that reduces their need to save. However, comparisons between people whose employers offer 401(k) plans and people whose employers do not suggests that 401(k) eligibility leads to substantial increases in saving.[7]

The magnitude of this response is difficult to

[6] The January 1991 Current Population Survey asked respondents, "Since you obtained your present job, did you take any training to improve your skills?" More information about these data is reported in the notes for Table 2.

[7] Poterba, Venti, and Wise (1995, 1998) and Webb (2001) found similar results in different data sets that covered different time periods. Engen, Gale, and Scholz (1994, 1996) argued, however, that 401(k) savers would have saved more in any case.

explain if people are fully rational. The evidence may be explained if people are irrationally impatient and have trouble saving. Workers with a self-control problem will be better off if they are compelled to save for retirement. Pensions do this, and workers accept them because they recognize their inability to control *their spending*.

This theoretical explanation is supported by extensive psychological evidence and by recent economic analysis in Laibson, Repetto, and Tobacman (1998). These authors used simulation models to show that people who recognize their self-control problems will use 401(k)-like plans to make wealth available to themselves in the future. According to their results, 401(k) plans always raise aggregate private saving because of their tax advantages. Their additional value as a means to commit to a long-term saving plan provides an extra boost of 17 to 60 percent to the aggregate saving rate, if people have self-control problems.[8]

Although the savings debate has focused on 401(k) plans, DB pensions also allow workers to commit to a long-term saving plan. Indeed, earlier evidence suggests that people with DB pensions saved more altogether, as people with 401(k) plans now do. Diamond and Hausman (1984) found that the elasticity of wealth with respect to pension income was –0.141, implying far less than a dollar-for-dollar offset. They also found that Social Security benefits reduced private wealth by less than dollar-for-dollar. Dicks-Mireaux and King (1983) found the same patterns for private and public pensions in Canadian data. Other researchers have suggested that workers prefer rising wage profiles, perhaps because it helps them save (Loewenstein and Sicherman, 1991; Frank and Hutchens, 1993).

It is important to note that 401(k) plans in particular do not entirely solve the self-control problem, since contributions are voluntary and workers can borrow against their 401(k) assets under some circumstances. Other DC pension plans require mandatory contributions.[9] However, any DC pension may be liquidated when a worker changes jobs, subject to a 10 percent penalty before age 59 1/2. These factors raise the risk that some workers will underfund their retirement saving. Chang (1996) found that 401(k) cash-out rates tend to be lower for older workers and for workers with higher balances. Poterba, Venti, and Wise (1999) estimated that cash-outs will reduce the aggregate value of 401(k) assets for workers at age 65 by about 5 percent.

Summary

The existence of pensions and other forms of deferred compensation is puzzling. Pensions constrain workers to save, whether they wish to or not. Existing pension theory suggests that the constraint is accepted because pensions encourage long-term employment, raising productivity and thus overall compensation. Considerable evidence supports this explanation for DB pensions, but the theory fails to account for the use of portable DC pensions.

Therefore, we have proposed a supplementary explanation—that workers value pensions as a vehicle for long-term saving. This explanation is linked to recent economic research that builds on extensive psychological evidence, and it is supported by findings that 401(k) plans, DB pensions, and Social Security all tend to raise personal saving.

POTENTIAL EXPLANATIONS FOR THE EVOLUTION OF PENSION STRUCTURE

A number of factors, both legal and economic, may explain the shift from DB to DC pensions. Legislative changes since 1974 have expanded the flexibility and preferential tax treatment of DC pensions but, at the same time, have boosted the costs of administering pension plans. For example, the government has set increasingly tight standards for maximum benefits, vesting, and eligibility in all types of pension plans, as well as funding requirements in DB plans.[10] Ippolito (1995) reported estimates from the Hay-Huggins Company (1990) that the average administrative costs of DB and 401(k)

[8] The range of increase depends on the particular features of the 401(k). These figures assume a value of one for the rate of relative risk aversion, though saving responds less to the 401(k) if risk aversion is higher. The authors argued that the most careful set of studies support a value of one or less.

[9] In a money purchase plan, the employer's annual contribution is determined by a specific formula, usually either a dollar amount or a percentage of salary. A target benefit plan is designed to provide a specific benefit level, but the benefit is not guaranteed. In a simplified employee pension, the employer contributes to the employee's individual retirement account. The employer distributes its own shares to employees in an employee stock ownership plan, while employees receive an option to purchase shares at a specified price in a stock purchase plan.

[10] Clark and McDermed (1990) provided a detailed explanation of these legal changes, which began with the Employee Retirement Income Security Act (ERISA). Using data for 1980-86, Kruse (1995) found that firms generally offered DC plans alongside existing DB plans, rather than terminating DB plans. Using later data from 1985-92, Papke (1999) found some replacement of both DB and other types of DC plans by 401(k) plans.

plans generally rose at similar rates, although very small DB plans grew relatively more expensive.

While legislative changes can account for some of the shift from DB to DC pensions, they cannot explain other patterns—for example, different rates of diffusion of DC pensions across industries and the movement of workers from types of firms with relatively high rates of DB coverage to types with high rates of DC coverage. A series of papers by Clark and McDermed (1990), Gustman and Steinmeier (1992), Ippolito (1995), and Kruse (1995) showed (i) that DB pensions remain more prevalent in large firms, industries such as manufacturing, and unionized jobs but (ii) that the proportion of workers in such jobs has declined.

Therefore, we are seeking explanations based on the economic theory of pensions outlined previously. DB pensions lose their appeal when the value of long-term employment declines. As with other recent labor market trends, such a change may be rooted in the diffusion of information technologies over the last 20-odd years.[11] Technological change is a leading explanation for the growing demand for skilled workers and consequent rise in earnings inequality between skilled and unskilled workers. It would not be surprising if rapid shifts in skill requirements associated with new technologies have also reduced the value of long-term employment.

In Friedberg and Owyang (2001) we explore this idea. An increasing pace of skill-biased technological change tends to raise the volatility of demand for particular skills. This change will in turn lower the expected duration of employment, and both workers and employers will gain less from the use of DB pensions.

In addition, factors such as technological change that have reduced relative earnings of unskilled workers may also explain their loss of pension coverage. Figure 1 shows the declining rate of pension coverage for all workers; Bloom and Freeman (1992) and Even and MacPherson (2000) have shown that coverage fell substantially more for workers with less education. Thus, it will be important to explore how changes in technology have affected both the level and structure of pension coverage.

OTHER IMPLICATIONS OF THE EVOLUTION OF PENSION STRUCTURE

Other differences between DB and DC pensions, besides those involving portability, are summarized in Table 1. Firms manage DB pension assets and as a consequence bear most of the resulting risks, except the risk of early severance. In contrast, workers manage DC pension assets and bear most of the risks. These differences could have important effects not only on job mobility, but also on consumption and saving before and after retirement and on stock market participation.

Additional Differences Between Pensions During Employment

Many aspects of administrative control and consequent risk are borne by firms when pensions are DB and are borne by workers when pensions are DC. Firms control how DB pension assets are invested and consequently bear the risk of uncertain interest rates, which may leave pensions underfunded or overfunded. Government regulations instituted since 1974 tightly restrict funding of DB pension obligations, however, and thus reduce the extent to which firms can smooth these risks over time.

Workers control how DC pension assets are invested among several options—generally mutual funds and company stock—which employers choose to offer. Consequently, workers bear the risk of uncertain rates of return. Underfunding is a greater possibility, as some employers allow workers to borrow against DC pensions in case of financial need.

Lastly, firms bear all the administrative and regulatory costs of DB pensions but also bear much of the costs of DC pensions. Workers incur costs to the extent that they actively manage their DC pension assets.

Additional Differences Between Pensions Post-Employment

The primary difference post-employment is that DB pension benefits are paid out as an annuity, while DC pension assets are transferred as a lump sum to workers. Consequently, firms bear the risk of the uncertain lifespan of workers who receive DB pensions, while workers bear this risk when they receive DC pensions.

Post-employment, the administrative control and consequent interest rate risk of each pension type generally remain the same as during employment. However, the burden of administrative and

[11] For example, the personal computer was launched by IBM in 1981.

regulatory costs of DC pensions shift from firms to workers upon retirement.

Implications for Consumption and Saving

Consumption and saving patterns are likely to differ for workers with different types of pensions. Workers with DC pensions bear interest rate and lifespan risk, and because individuals are risk-averse, this should induce additional precautionary saving. Similarly, they should be slower to deplete their wealth after retirement. All that, however, depends on individuals making consumption and saving choices rationally. Self-control problems of the type described earlier may be abetted by the lump-sum payout at separation from DC pensions. In order to encourage the preservation of DC pension assets, withdrawals before age 59 1/2 suffer a 10 percent penalty, as we noted earlier.

Implications for Public Policy

Because DC pensions are paid as a lump sum, elderly with DC pensions are more likely to outlive their assets, compared with elderly with DB pensions. This will be exacerbated if self-control problems lead to overconsumption after receiving the lump sum. As a result, the spread of DC pensions may increase take-up of means-tested public programs like Supplemental Security Income (which offers cash benefits), Medicaid (which pays for long-term care), and food stamps.

Medicaid rules dealing with annuitized versus unannuitized wealth may further encourage retirees with DC pensions to spend down their assets.[12] Medicaid only pays for long-term care when income and assets are low enough. Both must be extremely low for single people to qualify. However, the spouse of a married person who qualifies may retain $2,000 in monthly income, $20,000 in assets, and 50 percent of assets between $20,000 and $180,000. Annuitized DB pension wealth is treated as income, while unannuitized DC pension wealth is treated as an asset. Since the asset limit is relatively stricter than the income limit, DC pension wealth is subject to a relatively high implicit tax, in case one spouse applies to Medicaid to pay for long-term care.

Implications for Financial Markets

DC pension plans that do not involve employee stock ownership or stock options give workers some choices over their investment strategy. Thus, pension structure will influence financial markets if firms and workers make different portfolio choices. A growing body of financial research suggests that, even if investors are fully rational, the process by which information diffuses affects both rates of return and volatility in financial markets. For example, a simple model of herding laid out in Banerjee (1992) suggests that investors who have little or no private information rationally follow the behavior of others, which may be highly misleading. Learning models can also lead to herding, as noted in Smith and Sørensen (2000). Individual investors may be more subject to these types of "informational cascades" than institutional investors like pension funds. Another class of models analyzes specific deviations from rationality to which small investors may be more prone; these may explain the equity-premium and other long-standing puzzles involving financial markets.[13]

CONCLUSIONS

In this article, we have reviewed a variety of causes and consequences of the choice of pension structure. It is not surprising, therefore, that the evolution of pensions over the last 20 years has begun to influence many aspects of working and saving. While the spread of defined benefit (DB) pensions in the 1950s and 1960s contributed to the decline in the average retirement age, retirement ages have stabilized in the 1980s and 1990s as defined contribution (DC) pensions have taken hold. Meanwhile, workers at younger ages are changing jobs more frequently.

Although it is too early to tell, post-retirement consumption patterns may also shift. If people correctly evaluate the increased risk of outliving their DC pension resources, they may slow down their consumption and save more. However, access to their entire pension wealth upon retirement may lead some to hasten consumption, ultimately worsening the problems of poverty among widows and the oldest old and increasing the fiscal drain on income support programs for the elderly.

As the age structure of the labor force continues to shift, it will be important to understand the impli-

[12] Medicaid's treatment of DB and DC pension wealth is detailed in Webb (2001).

[13] See, for example, Barberis, Huang, and Santos (1999) and Barberis and Huang (2000). These articles analyze the financial implications of loss aversion and mental accounting, described in Rabin and Thaler (2001).

cations of the ongoing changes in pensions. Future research in this area promises new insights not only about the role of pensions, but more broadly about the behavior of workers and firms in an era of changing expectations and new technologies.

REFERENCES

Banerjee, Abhijit. "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics*, August 1992, *107*(3), pp. 797-818.

Barberis, Nicholas and Huang, Ming. "Mental Accounting, Loss Aversion, and Individual Stock Returns." Working Paper No. 8190, National Bureau of Economic Research, March 2001.

_____; _____ and Santos, Tano. "Prospect Theory and Asset Prices." Working Paper No. 7220, National Bureau of Economic Research, July 1999.

Bloom, David and Freeman, Richard. "The Fall of Private Pension Coverage in the U.S." *American Economic Review Papers and Proceedings*, May 1992, *82*(2), pp. 539-45.

Chang, Angela. "Tax Policy, Lump-Sum Pension Distributions, and Household Saving." *National Tax Journal*, June 1996, *49*(2), pp. 235-52.

Clark, Robert and McDermed, Ann. *The Choice of Pension Plans in a Changing Regulatory Environment*. AEI Studies, No. 509. Washington, DC: The AEI Press, 1990.

Crary, David. "Market Turmoil Clouds Retirement Hopes of Many Investors." Associate Press Newswires, 18 March 2001.

Diamond, Peter and Hausman, Jerry. "Individual Retirement and Savings Behavior." *Journal of Public Economics*, February/March 1984, *23*(1-2), pp. 81-114.

Dicks-Mireaux, Louis-David and King, Mervyn. "Portfolio Composition and Pension Wealth: An Econometric Study," in Z. Bodie and J. Shoven, eds., *Financial Aspects of the United States Pension System*. Chicago: University of Chicago Press, 1983, pp. 399-435.

Employee Benefit Research Institute. "Fundamentals of Employee Benefit Programs." Fourth Edition. Washington, DC: 1996.

Engen, Eric; Gale, William and Scholz, John Karl. "Do Saving Incentives Work?" *Brookings Papers on Economic Activity*, 1994, *0*(1), pp. 85-151.

_____; _____ and _____. "The Illusory Effects of Saving Incentives on Saving." *Journal of Economic Perspectives*, Fall 1996, *10*(4), pp. 113-38.

Even, William and MacPherson, David. "The Changing Distribution of Pension Coverage." *Industrial Relations*, April 2000, *39*(2), pp. 199-227.

Frank, Robert and Hutchens, Robert. "Wages, Seniority, and the Demand for Rising Consumption Profiles." *Journal of Economic Behavior and Organization*, August 1993, *21*(3), pp. 251-76.

Friedberg, Leora and Owyang, Michael. "The Role of Technological Change in Explaining the Evolution of Pension Structure." Unpublished manuscript, 2001.

_____ and Webb, Anthony. "The Impact of 401(k) Plans on Retirement." Discussion Paper No. 2000-30, University of California at San Diego.

Gustman, Alan and Steinmeier, Thomas. "The Stampede Toward Defined Contribution Pension Plans: Fact or Fiction?" *Industrial Relations*, Spring 1992, *31*(2), 361-69.

Hay-Huggins Company. "Pension Plan Expense Study." Final report submitted to the Pension Benefit Guaranty Corporation. September 1990.

Ippolito, Richard. "A Study of the Regulatory Impact of the Employee Retirement Income Security Act." *Journal of Law and Economics*, April 1988, *31*(1), pp. 85-125.

_____. "Toward Explaining the Growth of Defined Contribution Pensions." *Industrial Relations*, January 1995, *34*(1), pp. 1-20.

Kotlikoff, Laurence and Wise, David. "Labor Compensation and the Structure of Private Pension Plans: Evidence for Contractual Versus Spot Labor Markets," in D. Wise, ed., *Pensions, Labor, and Individual Choice*. Chicago: University of Chicago Press, 1985, pp. 55-85.

_____ and _____. "The Incentive Effects of Private Pension Plans," in Z. Bodie, J. Shoven, and D. Wise, eds., *Issues in Pension Economics*. Chicago: University of Chicago Press, 1987, pp. 283-336.

_____ and _____. "Employee Retirement and a Firm's Pension Plan," in D. Wise, ed., *The Economics of Aging*. Chicago: University of Chicago Press, 1989, pp. 279-330.

Kruse, Douglas. "Pension Substitution in the 1980s: Why the Shift Toward Defined Contribution?" *Industrial Relations*, April 1995, *34*(2), pp. 218-41.

Laibson, David; Repetto, Andrea and Tobacman, Jeremy. "Self-Control and Saving for Retirement." *Brookings Papers on Economic Activity*, 1998, *0*(1), pp. 91-172.

Lazear, Edward P. "Retirement from the Labor Force" in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*. Vol. 1. Handbooks in Economics Series. New York: Elsevier Science Publishers, 1986, pp. 305-55.

Loewenstein, George and Sicherman, Nachum. "Do Workers Prefer Increasing Wage Profiles?" *Journal of Labor Economics*, January 1991, *9*(1), pp. 67-84.

Lumsdaine, Robin and Wise, David. "Aging and Labor Force Participation: A Review of Trends and Explanations," in Y. Noguchi and D. Wise, eds., *Aging Issues in Japan and the United States*. Chicago: University of Chicago Press, 1994, pp. 7-42.

Mitchell, Olivia. "New Trends in Pension Benefit and Retirement Provisions." Working Paper No. W7381, National Bureau of Economic Research, October 1999.

Papke, Leslie. "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data." *Journal of Human Resources*, Spring 1999, *34*(2), pp. 346-68.

Poterba, James M.; Venti, Steven F. and Wise, David A. "Do 401(k) Contributions Crowd Out Other Personal Saving." *Journal of Public Economics*, September 1995, *58*(1), pp. 1-32.

_____; _____ and _____. "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence." in D. Wise, ed., *Frontiers in the Economics of Aging*. Chicago: University of Chicago Press, 1998, pp. 23-106.

_____; _____ and _____. "Pre-Retirement Cashouts and Foregone Retirement Saving: Implications for 401(k) Asset Accumulation." Working Paper No. W7314, National Bureau of Economic Research, August 1999.

Rabin, Matthew and Thaler, Richard. "Anomalies: Risk Aversion." *Journal of Economic Perspectives*, Winter 2001, *15*(1), pp. 219-32.

Salop, Steven and Salop, Joanne. "Self-Selection and Turnover in the Labor Market." *Quarterly Journal of Economics*, November 1976, *90*(4), pp. 619-27.

Samwick, Andrew. "New Evidence on Pensions, Social Security, and the Timing of Retirement." *Journal of Public Economics*, November 1998, *70*(2), pp. 207-36.

Smith, Lones and Sørenson, Peter. "Pathological Outcomes of Observational Learning." *Econometrica*, March 2000, *68*(2), pp. 371-98.

Stock, James H. and Wise, David A. "Pensions, the Option Value of Work, and Retirement." *Econometrica*, September 1990a, *58*(5), pp. 1151-80.

_____ and _____. "The Pension Inducement to Retire: An Option Value Analysis," in D. Wise, ed., *Issues in the Economics of Aging*. Chicago: University of Chicago Press, 1990b, pp. 205-24.

Thaler, Richard. "Psychology and Savings Policies." *American Economic Review*, May 1994, *84*(2), pp. 186-92.

Webb, Anthony. "The Impact of 401(k) Plans on Pre-Retirement Saving, Age of Retirement and Post-Retirement Consumption." Ph.D. Dissertation, University of California, San Diego, 2001.

Voting Rights, Private Benefits, and Takeovers

Frank A. Schmid

This article presents a textbook exposition of the effects that institutional design of the firm has on allocation of control over assets. The efficient allocation of control over the assets bundled up in the firm is necessary for the optimal allocation of its resources. Dynamic efficiency in resource allocation presupposes that control over firms will change hands when a given allocation turns suboptimal. The institutional framework within which control changes hands is called the market for corporate control. This market is closely linked to the stock market as control rights over the assets of the firm are linked to voting stock. We analyze how the allocation of shareholder voting rights and other organizational designs of the firm affect the firm's stock market valuation and the allocation of control over its assets.

Transactions in the market for corporate control have increased greatly over the last decade both in number and value. Figure 1 shows that in the United States the number of acquisitions of publicly traded companies quadrupled between a trough in 1991 and a recent peak in 1999. Measured in dollar terms (without inflation adjustment), the rise in acquisitions of publicly traded companies was 30-fold during that period. Among the 50 industries distinguished by Mergerstat (2000, pp. 61-69), "Banking & Finance" was among the seven most active industries in any year in the 1996-2000 period as measured by number of transactions announced. Based on the dollar value offered in announced acquisitions, Banking & Finance was among the six most active industries in that same period and topped the rankings in the years 1997 and 1998.

The mechanics of the market for corporate control are determined by the legal system. Most importantly, the legal system shapes the incentive structure to which the participants in the market for corporate control respond in their actions. Moreover, the incentive structure in place has important efficiency implications. If designed optimally, society's legal system directs the self-interest of economic agents toward the optimal social outcome.

Most significant to the legal framework of the market for corporate control are the firm's articles of association and bylaws. There is also Securities and Exchange Commission (SEC) regulation, and there are the specific rules of the respective stock exchanges (such as the New York Stock Exchange, Nasdaq, and the American Stock Exchange).

Articles of association and bylaws vary across corporations. For instance, corporations may have the choice to amend their articles of association such that unsolicited bidders find it difficult to obtain control over the assets. The legal options that are available to corporations vary across state lines. For instance, a wide variety of anti-takeover amendments exist for Delaware corporations, such as supermajority rules for decisions that pertain to mergers or to the removal of board members.[1] There are also cross-country differences in corporations' articles of association, which become important in cross-border merger and acquisitions transactions.

Acquisitions of publicly traded companies typically involve block trades or tender offers. In a block trade, an investor acquires a block of shares from a large shareholder. In a tender offer, an investor bids for shares that are dispersed across a multitude of mostly small shareholders. Block trades are public transactions, while tender offers are private deals. Both types of transactions might be preceded, accompanied, or followed by acquisitions of shares in the open market. Changes in control that occur through block trades are common on the European continent, where tender offers are rare.[2] In the United States, on the other hand, 27 percent of all acquisitions of publicly traded companies in 2000 were brought about through tender offers (see Figure 2).

Two kinds of value matter for wealth-maximization when control over the firm changes hands. First, there is what is commonly referred to as the public value of the firm, i.e., the market value of its securities. Second, there might be a private value of the firm, through which an investor enjoys some benefit while exercising control over the firm. Private control benefits are most significant for entrepreneurial start-ups, established family-owned businesses, and organizations where personal investors also pursue non-pecuniary goals, such

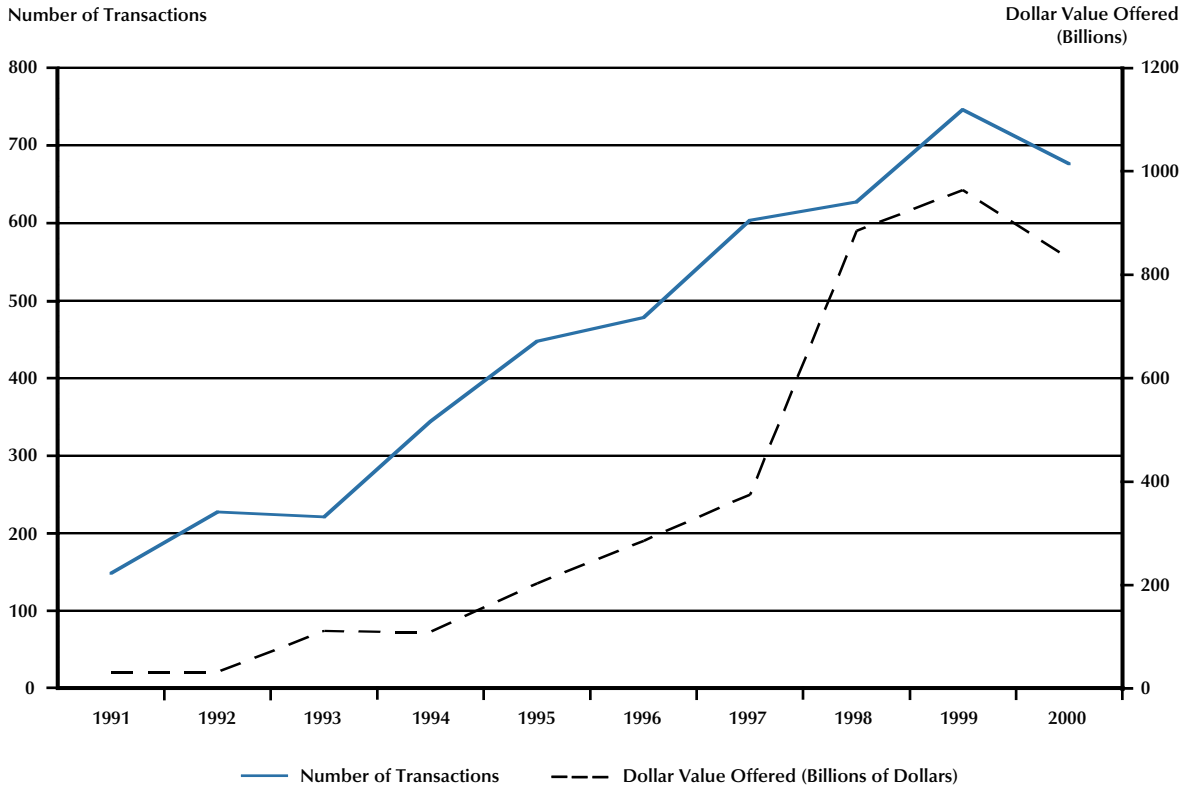
Frank A. Schmid is a senior economist at the Federal Reserve Bank of St. Louis. William Bock and Judith Hazen provided research assistance.

[1] See < http://www.uslegalforms.com/corporations/table13.htm > .

[2] See, for instance, Franks and Mayer (2000), who study control changes in Germany.

Figure 1

Acquisitions of Publicly Traded Companies



NOTE: Annual observations, 1991-2000; completed or pending transactions.
SOURCE: *Mergerstat Review 2001.*

as media groups and professional sports organizations.[3] Maximizing social welfare necessitates that, when control changes, the sum of the public and the private values of the firm assumes its highest value.[4]

The following analysis assumes well-defined property rights for the various stakeholders in the firm, such as labor, bond holders, the tax authorities, suppliers, and customers. Enforceability of property rights precludes opportunistic behavior by the bidder. Absent the enforceability of such property rights, takeovers, even if they destroy value overall, might be worthwhile for the bidder if he succeeds in increasing his wealth at the expense of other stakeholders, such as bondholders and labor.[5]

Three institutional designs in corporate control, according to the finance literature, are most significant to wealth-maximization in takeovers: These are the one share–one vote principle, majority rules, and mandatory tender offers. We analyze the implications of these three organizational designs in a simple textbook takeover model that is drafted along the lines of Grossman and Hart (1988) and Hart (1995). The model helps define the optimal design of the legal environment in which takeovers enhance social welfare.

In the following section we present the framework that we use for analyzing the efficiency implications of institutional design as they apply to corporate control. A brief discussion of the relationship between takeovers and auctions follows. The subsequent analysis of takeovers draws on Hart (1995) but extends his analysis in several ways. We

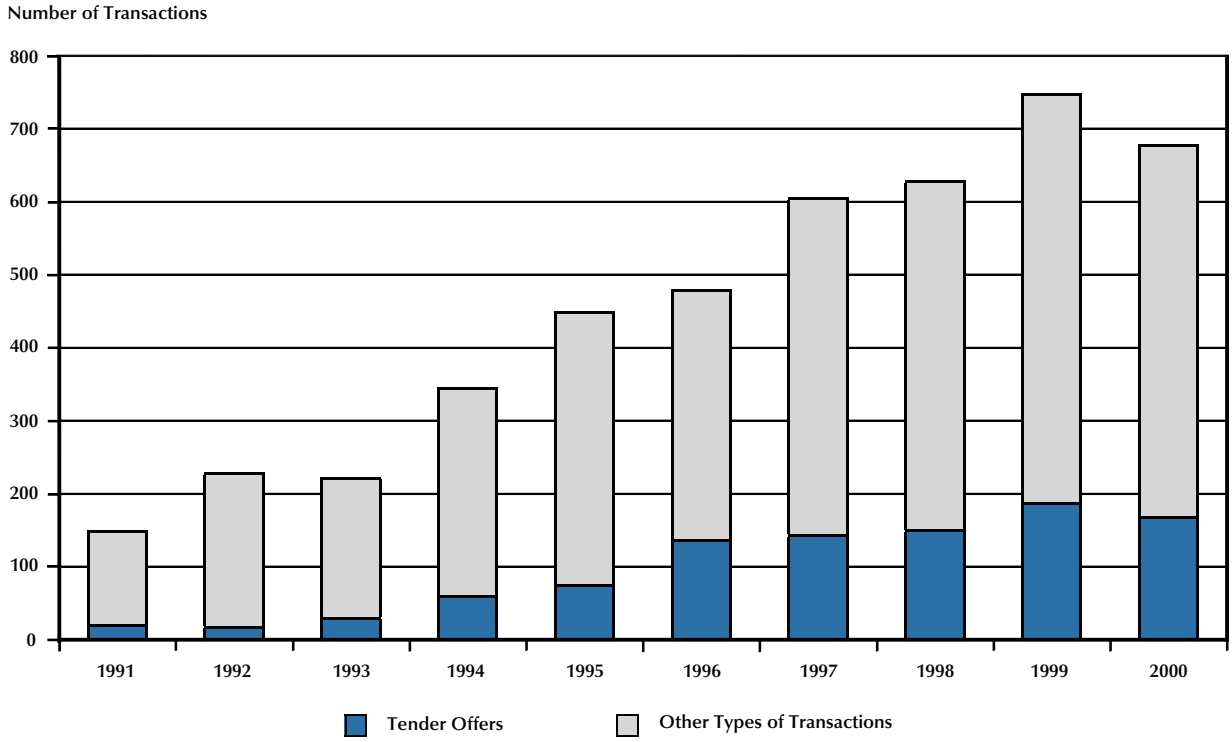
[3] See Demsetz (1983).

[4] See Milgrom and Roberts (1992, pp. 35-38).

[5] See Shleifer and Summers (1988).

Figure 2

Tender Offers Versus Other Types of Acquisitions



NOTE: Annual observations, 1991-2000; completed or pending transactions.
SOURCE: *Mergerstat Review 2001.*

address issues of free-riding in tender offers, stock ownership disclosure regulation, the one share–one vote principle, private control benefits, majority rules, and mandatory tender offers.

CORPORATE CONTROL AND TAKEOVERS: AN ANALYTICAL FRAMEWORK

Grossman and Hart (1988) and Harris and Raviv (1988) analyze the effects that deviations from the one share–one vote principle and the simple majority rule have on the value of the firm. Grossman and Hart take the perspective of the securities holders and restrict themselves to the implications for the public value of the firm. Harris and Raviv, on the other hand, take the perspective of society by looking at both the public and private values of the firm. Both papers allow for private control benefits and present the takeover mechanism as a tender offer to dispersed shareholders. No block trades among big shareholders are considered.

In Grossman and Hart (1988), the subjective probability of the small shareholder being pivotal to the outcome of a takeover attempt is zero. Harris and Raviv (1988), on the other hand, allow for this probability to be positive. This difference between the two studies explains some of the differences in results.

Grossman and Hart (1988) investigate three cases. In their first case, they look at an efficiently operated company with dual-class stock. Half the stock is voting stock, while the other half is non-voting stock. Both classes of stocks are endowed with the same cash flow rights, and the incumbent investor does not enjoy benefits from control. Tender offers are unrestricted, which means that the bidder must take in all shares tendered to him rather than just a fraction of the stock. Grossman and Hart show that, in such a regime, a rival investor who would enjoy private control benefits but operate the firm inefficiently might succeed in obtaining control. The reason is that the investor needs to bid for the

voting stock only. The holders of the nonvoting stock go uncompensated for the loss they suffer from the rival investor's inefficient management. In a one share–one vote regime, on the other hand, the rival investor must acquire all of the company's stock. In a tender offer under this regime, the rival investor would have to pay at least the price at which the stock is trading under the current, efficient management. Thus, the one share–one vote principle maximizes the public value of the firm.

In their second case, Grossman and Hart allow for both the incumbent and the rival investor to enjoy private control benefits. They assume that the investor who enjoys the greatest control benefits is also the investor who operates the company most efficiently. Under this assumption, the company takes on a maximum value if its stock is split into two extreme securities: one class of shares endowed with voting rights only, the other class of shares endowed with cash flow rights only. The two competing investors bid for the voting stock. The investor with the greater control benefits is more willing to pay and consequently wins out. Because this investor is also the one who runs the company efficiently, the owners of the nonvoting stock benefit as well. This scenario maximizes the public value of the firm. (Grossman and Hart point out that the assumptions made in this case are rather restrictive; the results thus cannot be read as a general recommendation for the public firm to deviate from the one share–one vote principle.)

In their third case, Grossman and Hart (1988) allow for restricted tender offers. In a restricted tender offer, the bidder can limit the shares he acquires to a pre-announced fraction. The authors show that a restricted tender offer for 50 percent of the voting stock in a one share–one vote regime is similar to an unrestricted offer in a dual-class stock regime where 50 percent of the cash flow rights are associated with nonvoting stock. Grossman and Hart also analyze the optimality of the simple majority rule, where decisionmaking requires 50 percent plus one vote. They show that the one share–one vote principle and the simple majority rule are optimal when the controlling party's private control benefits substantially exceed the rival investor's private control benefits.

In summary, Grossman and Hart (1988) show that deviations from the one share–one vote rule for public, widely held corporations are likely to be suboptimal. Deviations from the one share–one vote principle might help entrench management by insulating the firm from the market for corporate control. On the other hand, for entrepreneurial companies, issuing nonvoting stock might be optimal because it helps preserve the founding family's private control benefits.

Harris and Raviv (1988) derive results that are less ambiguous than those of Grossman and Hart (1988). As mentioned above, a major difference between the two approaches is that Harris and Raviv do not assume, as Grossman and Hart do, that the shareholder's subjective probability of being pivotal to the success of a tender offer is zero. Also, while Grossman and Hart concentrate on how to maximize the public value of the firm, Harris and Raviv look at social optimality also. Social optimality is achieved when the sum of the public and the private values of the firm takes on a maximum. Harris and Raviv show that, in a regime in which the simple majority rule and the one share–one vote principle apply, the investor that will run the firm most efficiently obtains control. While this regime ensures the socially optimal outcome, it generally does not maximize the public value of the firm. The authors show that, in a dual-class stock regime in which one class of stock has all the voting rights and the other class has all the cash flow rights, the public value of the firm takes on its maximum. This is because, with dual-class stock, the securities holders are able to extract a larger fraction of the rival investor's private control benefits. This finding is similar to an aforementioned result obtained by Grossman and Hart. In summary, Harris and Raviv show that the one share–one vote principle in combination with the simple majority rule is, in general, socially optimal because it maximizes the sum of the public and the private values of the firm. To maximize the public value of the firm, the firm should issue dual-class stock that separates voting rights from cash flow rights.

The present study follows Grossman and Hart (1988) in that we assume that the subjective probability of the small shareholder being pivotal to the takeover success is zero. Like Harris and Raviv (1988), the focus is on maximizing social welfare rather than just the public value of the firm. In contrast to any of these studies, the analysis is not restricted to tender offers, but also allows for block trades. Also, hold-up situations that supermajorities might create are discussed, as is the potentially beneficial role of mandatory tender offers.

TAKEOVERS AND AUCTIONS

Takeovers are typically brought about through successful tender offers or block trades. In both types of transactions, there is at least one bidder extending an offer to the firm's current equity holders. When a rival bidder contests the offer, the takeover resembles an auction. In takeovers, the auctioned object is control over the firm, which is tied to the firm's voting stock.

There are common value and private value auctions. In common value auctions, there are no benefits arising from equity ownership that are not common to all bidders. The common value in takeovers is the present value of the firm's cash flows. In private value auctions, on the other hand, the value of the firm depends on the bidder. Private value matters in takeovers if bidders derive private benefits from exercising control over the firm. In such cases, the firm might then be valued above the present value of its cash flows.

In auctions, the current owners have reservation prices. A reservation price is the price below which the current owner is not willing to trade. For reservations prices, too, private benefits might matter. For instance, if the takeover target is an entrepreneurial firm, it is likely that due to the entrepreneur's private control benefits the reservation price of the seller exceeds the present value of the auctioned firm's cash flows.

Takeover bidding resembles English auctions, which have an ascending bid structure, and the auctioned firm goes to the investor who submits the highest bid. Although this outcome is efficient, the firm might sell for a price that is less than the winning bidder is willing to pay. Such an outcome is possible because all it takes to win the bid is an offer that supersedes, even by the smallest possible increment, the bidder with the next-to-highest willingness to pay.[6]

The Winner's Curse

In a takeover, a bidder might overpay because he overestimates the target firm's present value of cash flows.[7] Even if the bidder has unbiased expectations, a random error in these expectations may cause his willingness to pay to exceed the firm's intrinsic value. This phenomenon is called the winner's curse.[8] Below we illustrate the concept of the winner's curse in two examples. The first example shows that a winning bidder need not overpay even if he overestimates the intrinsic value of the auctioned object. The second example is a case in which a bidder who overestimates does indeed overpay. To keep matters simple, we illustrate the winner's curse for a common value auction.

In this first example, assume that there are two bidders, A and D. Both bidders have unbiased expectations about the present value of the firm's future cash flows, which equals $100. Because of a random element in expectations, D estimates the intrinsic value of the target at $102, while A estimates it at $98.[9] While both bidders' expectations are off the mark, as a group the bidders' expectations are correct (unbiased). In a bidding contest, D will end up with the firm at a price marginally above $98, without overpaying.

In the second example, assume there are two additional bidders, B and C, in addition to bidders A and D from the first example. The expectations of bidders B and C about future cash flows are $99 and $101, respectively. Again, as a group the expectations of the four bidders are unbiased as they average the intrinsic value of the auctioned object. As in the first example, D wins out. This time, D pays marginally more than $101. Although D pays less than he is willing to pay, he nevertheless overpays because the intrinsic value of the auctioned firm is only $100.

In summary, the winner's curse concept rests on estimation errors, although these estimation errors need not be systematic. The winning bidder might, but need not, overpay. As the number of bidders rises, however, the probability that the winning bidder overpays increases, all else equal.

Empirical studies show that in takeover contests all the gains (if any) tend to go to the shareholders of the target firm. On average, the shareholders of the acquiring firm break even. There is no evidence that in takeovers bidders overpay systematically; in

---

[6] In Dutch auctions, the bid structure is descending. The auctioneer calls prices in descending order, and the first bidder to shout "mine" wins out. See Milgrom (1989).

[7] If the bidder is a company that merges the target into its existing operations, the value the bidder assigns to the target is not the present value of the target firm's cash flows on a stand-alone basis. Rather, it is the difference between the present value of the cash flows of the combined firm and the sum of the present values of the cash flows of the two firms when operating on a stand-alone basis.

[8] For an overview on the winner's curse, see Milgrom (1989) and Thaler (1988).

[9] The two bidders' expectations may be viewed as independent draws from the same probability distribution, which is symmetric around the expected value.

the following sections, we make use of this empirical finding.[10] We exclude overpaying by assuming that the intrinsic value of the firm is public knowledge.

TENDER OFFERS AND FREE-RIDING

Tender offers are public bids for stock in which investors can tender their shares in the target firm to a bidder at a certain price within a certain time window. As mentioned above, tender offers are called restricted if this offer applies to a certain fraction of shares only. Otherwise, the offers are called unrestricted. In addition, tender offers may be conditional or unconditional. If the offer is conditional, the bidder is not obliged to acquire the tendered shares if their fraction in the total outstanding stock falls short of a pre-announced minimum.

In tender offers, the shares are typically dispersed among small shareholders. The dispersion of the shares gives rise to a free-riding problem, which might thwart value-enhancing takeovers. In the following, we analyze the free-riding problem in a common value bidding contest. Because the free-riding problem is independent of the number of bidders, one can assume without loss of generality that there is no rival bidder contesting the takeover attempt.

Assume that there is a target company with 2 million shareholders, each holding one share. Before news of the bid reaches the market, the target trades at $1 a share. The bidder is a buyout fund that plans on taking control of the firm and improving the efficiency of the operations. The bidder expects this transaction to add $1 million (or $0.5 per share) to the target's present value of cash flows. We assume that the value added is public knowledge. This is a reasonable assumption for large, traded firms, which are closely followed by financial analysts.

For each shareholder of the target firm, the *objective* probability that his decision is pivotal to the success of the takeover, or to the bidder's decision to better the offer while it is outstanding, is only marginally greater than zero.[11] We follow Grossman and Hart (1988) in assuming that each shareholder's *subjective* probability of being pivotal is zero. This assumption implies that small shareholders are unable to enjoy control benefits.

We assume that the tender offer is conditional (restricted or unrestricted), which means that the bidder acquires the tendered shares only if he succeeds in seizing control. In the absence of supermajority rules, we set the control threshold to 50 percent plus one share.

In the first example, assume that the bidder owns no stock in the target when launching the tender offer. The target firm's shareholders have an incentive not to tender if the bid falls short of the stock's post-takeover value. This response occurs because no shareholder assumes that his decision is pivotal to the outcome of the takeover attempt and each shareholder therefore pursues his best interest. By doing so, however, the shareholders thwart the efficient social outcome. If, for instance, the bidder offers $1.4 a share, the target shareholder will end up with $1.4 if he tenders and the takeover succeeds. If he tenders and the takeover fails, he winds up with $1. On the contrary, if he does not tender, the payoffs in these two situations are $1.5 and $1, respectively. Thus, for the target shareholder, it is optimal not to tender in response to a bid that falls short of the post-takeover share price. On the other hand, for the bidder, a price of $1.5 or higher is unprofitable. Consequently, the takeover—in spite of being value-enhancing—does not materialize.

In the second example, assume that the bidder acquired a toehold in the target firm in the open market at $1 a share before announcing the tender offer. Again, the target shareholder does not tender unless the bid matches the post-takeover share price of $1.5. Because of the toehold, the bidder is able to reap part of the value added even when paying $1.5 a share. If, for instance, the toehold amounts to 5 percent, the bidder retains $50,000 of the value added, while the other $950,000 go to the target shareholders.[12] The takeover materializes, and the efficient outcome obtains.[13]

In the third example, assume that the rival investor is a wealthy individual who enjoys benefits from having control. When announcing the bid, the investor holds no stock in the target firm. The bidder values the private control benefits at $150,000.

---

[10] For a survey on empirical studies on post-takeover performance, see Weston, Chung, and Siu (1998).

[11] Takeover regulation typically requires that when an outstanding bid is bettered, the new price uniformly applies to all tendered shares, including those that have already been tendered.

[12] Many jurisdictions around the world restrict the size of toeholds that investors can accumulate without having to disclose it to the public, the target firm, or the competent stock market supervisory authority. For instance, the threshold for disclosure might equal 5 percent of the company's total equity or the equity within a certain class of stock.

[13] For an extensive analysis on the role of toeholds on the success of takeovers, see Shleifer and Vishny (1986).

Thus, if the takeover succeeds, another $150,000 in value is added because of private control benefits, beyond the $1 million the investor would add through improving the target firm's operations. The total value added through the takeover would amount to $1.15 million. Similar to toeholds, private control benefits help overcome the free-riding problem in tender offers. The target shareholders' payoff matrix is identical to the first two examples. The bidder is willing to pay a price equal to the post-takeover (public) value of the target ($1.5 a share) because the takeover allows the entrepreneur to realize private control benefits equivalent to $150,000.

The above examples assumed the tender offers to be conditional. It turns out that the incentives that prevail in an otherwise identical unconditional tender offer are more conducive to a successful takeover. In the first example discussed above, if the tender offer were unconditional, the target shareholder would receive $1.4 for certain if he tendered, with no change in payoff if he did not tender. This means that the shareholder is more likely to tender if the tender offer is unconditional. For instance, if the small shareholder attaches equal probabilities to the takeover failing or coming to pass, the expected value of the share if not tendered amounts to $1.25, which is short of the risk-free $1.4 paid on the tendered share.

Another instrument for solving the free-riding problem is the two-tier offer. The bidder makes a favorable "front end" offer for the fraction of shares he needs to obtain control and an unfavorable "back end" offer for the remainder. For instance, a bidding corporation might offer cash for the first 50 percent plus one share and newly issued shares for the remainder.[14] After the bidder obtains control through the cash offer, the bidder might find ways of depressing the firm's public value before forging a "back-end" merger under conditions favorable to him. One way to depress the value is to dilute the firm's earnings. Dilution of earnings is possible, for instance, through asset transfers or through transfer pricing of inter-firm trade in intermediate products. Transfer pricing and asset transfers below fair market value violate the arm's-length principle and might be illegal, depending on the jurisdiction.[15] Also, as two-tier offers discriminate between front-end purchase and back-end conversion, the investor might violate duties of equal shareholder treatment. The tender offer is coercive, as the shareholders feel compelled to be in the first tier.

THE ONE SHARE–ONE VOTE PRINCIPLE

Harris and Raviv (1988) have shown that the one share–one vote principle is generally optimal for society and suboptimal for the securities holders. Also, Grossman and Hart (1988) have shown that deviations from the one share–one vote rule might be optimal from the securities holders' point of view. In the following, the implications of violations of the one share–one vote principle for the value of the firm is analyzed in two numerical examples.

Firms might deviate from the one share–one vote rule by issuing preferred stock, which may be either stock endowed with multiple votes or nonvoting stock with preferred cash flow rights. In some jurisdictions, issuing stock with multiple votes is prohibited. Also, legislation might limit nonvoting stock to a certain fraction of the firm's total equity.

We analyze a firm with dual-class stock. Class A stock is common (voting) stock, and class B stock is preferred (nonvoting) stock. Each class of stock is endowed with the same cash flow rights. As above, we allow for two types of investors: block holders and small shareholders. We maintain the assumption that shareholders are investors who each hold one share and attach a subjective probability of zero to being pivotal to the success of a takeover attempt. Class B stock is held entirely by small shareholders—because they do not value control anyway—whereas class A stock might be held by small shareholders or by block holders. We also assume that the marginal investor in class A stock is a small shareholder, which means that both classes of stock trade at the same price.[16] In each class of stock, there are 1 million shares outstanding.

In this first example, we assume that initially all A shares are held by small shareholders. The firm operates efficiently, and class A and class B stocks trade at $1 a share. Assume that there is an investor who—if he were in control—would enjoy private control benefits but would not run the firm

[14] Prorating applies if more than 50 percent plus one share are tendered in the first tier. We assume that there is no supermajority rule in place.

[15] The arm's length principle stipulates that trade among affiliated companies has to be conducted at prices that would prevail in corresponding market transactions with unaffiliated companies.

[16] It is the marginal investor that prices financial assets. Empirically, nonvoting stock may trade higher or lower than voting stock. While the lack of control rights creates a discount on nonvoting stock, preferential cash flow rights generate a premium. Also, for entrepreneurial firms, the float of nonvoting stock frequently exceeds the float of voting stock, which generates a liquidity discount on the voting stock.

efficiently. For instance, with this investor in control, the private value of the firm (which is the monetary equivalent of the control benefits) might equal $150,000 and the public value of the firm might amount to $0.9 a share. The decrease in public value might be due to the fact that the investor employs firm resources to generate private benefits, for instance in the form of luxurious offices and lavish business dinners. If the investor succeeded in taking over the firm, social welfare would decrease by $50,000. This is because the sum of the post-takeover public and private values ($1,950,000) falls short of the company's pre-takeover value ($2 million).

The investor can obtain control over the firm by acquiring a minimum amount of class A shares, which—in the absence of supermajority rules—is 50 percent plus one share. If the investor bids $1.01 a share for the class A shares in a (unrestricted and unconditional) tender offer, the takeover attempt will be successful. This holds in spite of its value-depressing effect on society. At $1.01 a share, the small shareholder tenders. If the small shareholder does not tender, his position is worth $0.9 if the takeover succeeds and remains at $1 if it does not succeed. On the other hand, if the small shareholder tenders, he receives $1.01 for certain. Consequently, all shareholders tender their interests and the takeover succeeds. The investor loses $110,000 on his investment in A shares, but gains $150,000 in private control benefits. The loss to society ($50,000) is the difference between the decrease in public value ($100,000) and the increase in private control benefits ($50,000).

By comparison, the one share–one vote principle generates the efficient outcome by giving the investor no incentive to bid. Under the one share–one vote rule, the investor has to extend a tender offer to all shareholders by bidding $1.01 for each of the 2 million shares. The investor would lose $220,000 on the equity interest but gain only the equivalent of $150,000 in control benefits. A takeover succeeds if (and only if) the total value added is positive, which means that the gain in private benefits must exceed the loss in public value.[17] Note that even a toehold would not help the investor succeed in the takeover attempt.

In the second example, we assume that the firm in question is family-owned. The family holds all of the class A shares, but none of the nonvoting stock. The public value of the firm equals $1.8 million with the B shares trading at $0.9 a share.[18] The family enjoys private control benefits equivalent to $150,000.

In many countries, dual-class stock is a common phenomenon with family-owned companies. Frequently, as entrepreneurial firms grow, the wealth-constrained founding family is unable to maintain its fraction of equity in the firm following public offerings. By issuing nonvoting stock, the family might be able to retain control even after floating equity in the stock market.

Assume that the founding family is in its second generation and that the entrepreneurial skills left the company when the founder left. An institutional investor—a buyout fund, for instance—might be able to run the company more efficiently, without enjoying control benefits. Assume that, if the buyout fund were in control, the public value of the company would amount to $2 million, with B shares trading at $1 a share. Although society would be better off, scoring a net gain of $50,000, in a dual-class stock regime the optimal outcome does not obtain. This occurs because, to the family, the class A equity interest is worth $1,050,000 (private value of $150,000 plus public value of $900,000), which exceeds the post-takeover public value of the class A stock by $50,000.

The takeover can succeed in spite of the presence of dual-class stock if the buyout fund acquires an interest in class B stock before bidding for the family's stake. If the investor—before revealing the takeover plan—accumulated a position in B shares in excess of 50 percent in the open market at a price of $0.9, he would be able to buy out the family. This outcome can occur because the capital gain on the class B equity position would exceed the difference between the family's and the buyout fund's valuations of the class A equity. The buyout fund could pursue this strategy only if there were no stock ownership disclosure rules that would force the investor to reveal the buildup of the class B interest in the early stages of the buyout. Once the buyout fund's intentions leak to the market, B shareholders have an incentive to free-ride.

In the one share–one vote regime the situation is similar. With the family holding 50 percent (plus one share) of the voting stock and the rest being dispersed, the only way a rival investor can seize control is to buy out the family. The same incentives

[17] If the private benefits the investor enjoys are sufficiently low, or the inefficiency the investor causes is sufficiently high, the optimal outcome also prevails in the dual-class stock regime.

[18] Note that A shares do not trade. With B shares trading at $0.9 a share, the shadow price of the family's stake equals $900,000.

that were relevant in the dual-class stock regime apply in the one share–one vote regime. If the buyout fund is unable to accumulate (secretly) a position in excess of 25 percent in the open market at $0.9 a share, the investor has no incentive to bid for the family's equity interest.

The equivalence in outcomes in the two regimes is due to the specific assumptions made in the example. First, the fraction of voting stock was limited to 50 percent of the company's total equity; second, no supermajority rule was in place. Under such conditions, family owners have no incentive to issue nonvoting stock in lieu of voting stock. However, if the law requires supermajorities for certain decisions, the equivalence breaks down and dual-class stock becomes an important tool for protecting the family owner's private control benefits. This is discussed in the next section.

PRIVATE BENEFITS, SUPERMAJORITY RULES, AND DUAL-CLASS STOCK

In some jurisdictions around the world, corporate law mandates that certain decisions at annual meetings require supermajorities of two-thirds or 75 percent of the votes. Among the issues that are typically subject to supermajority rules are changes to the company's equity (e.g., securities offerings or stock repurchases) and major changes to assets (e.g., mergers and major acquisitions). The existence of supermajority rules implies the existence of blocking minority rules. For instance, with a 75 percent supermajority rule in place, a block holder can paralyze a corporation when holding 25 percent plus one vote. A blocking minority interest creates bargaining power vis-à-vis a family owner whose holding might have dropped below the 75 percent threshold due to a binding wealth constraint. Generally, supermajority rules imply that family owners must retain greater fractions of shares to stay in control. In the following we show that the one share–one vote regime is not necessarily optimal when private control benefits exist and corporate law mandates supermajorities for important decisions.

As an example, assume an entrepreneurial firm where the owner family's fraction of voting stock amounts to 60 percent. The remaining 40 percent have been floated in the stock market as the company expanded through public offerings and the family was unable to acquire the additional stock due to its limited wealth. Assume that dual-class stock is prohibited and that the law mandates a 75 percent supermajority for major corporate decisions. The company is run efficiently. The stock trades at $1 a share with a 1.2 million share float, which is dispersed. Effectively, the entrepreneur has command over the necessary supermajority. Dispersed shareholders exercise no control, because they view their probabilities of being pivotal as zero. Also, because the company is run efficiently, the shareholders have no incentive to reject or disapprove of the entrepreneur's operating decisions. The entrepreneur enjoys private control benefits, which he values at $150,000.

Assume that there is a rival investor who attaches a monetary equivalent of $75,000 to the control rights that come with a blocking minority interest in the company in question. The control benefits might emanate from personal pleasure of influencing the business decisions of this particular company or from reduced competition if the investor is a rival.[19] Assume that the rival investor's business goals are at odds with those of the entrepreneur, which paralyzes decisionmaking. The gridlock reduces the present value of the firm's cash flows from $2 million to $1.8 million.

If the rival investor accumulates a block of 25 percent plus one share in the open market at $1 a share and pursues the business strategy outlined above, he will lose a little more than $50,000 on the acquired shares but gain $75,000 in private control benefits. At the same time, the value of the remaining equity (75 percent minus one share) drops by a little less than $150,000. Also, the family loses its control benefits of $150,000 in part or in total. Overall, the net loss to society amounts to at least $125,000 (and at most $275,000). Despite the one share–one vote rule in place, the inefficient outcome prevails.[20]

With dual-class stock, the value-reducing control change can be prevented. Assume that the entrepreneur is allowed to issue nonvoting stock at a maximum of 50 percent of the corporation's

[19] For instance, for certain decisions, the German Stock Corporation Act requires supermajorities at annual shareholder meetings. In Germany, it has repeatedly been observed (in particular in the media industry) that investors take blocking minority interests in competitors, which all but paralyzes these companies before the original owners eventually surrender their stakes.

[20] If the small shareholders anticipate the decrease in the public value of the firm and assume that it will be sustained, they sell to the outside investor at $0.9 a share. This does not affect the change in wealth to society overall, but affects solely the distribution of wealth between the original (small) shareholders and the outside block holder.

total equity. With a 60 percent ownership of total equity, the entrepreneur is able to retain all the voting stock (and also holds 20 percent of the nonvoting stock). The rival investor has no means of seizing control over the firm without fully compensating the family owner. This scenario implies that, if the rival is not able to generate at least as much value as the incumbent, he is unable to gain control. Thus we conclude that in the presence of supermajority rules (i.e., blocking minority rules), the entrepreneurial firm should be allowed to deviate from the one share–one vote rule by issuing nonvoting stock.

BLOCK TRADES

In some of the examples above we have alluded to block trades as a means of transferring control over the firm. An example of a block trade is when a family sells out to a single investor rather than floating the block of shares in the stock market. Block trades are private deals rather than open-market transactions. It has been observed that in block trades the price per share exceeds the going share price in the open market.[21] The concept of the Nash bargaining solution offers a possible explanation for the existence of such block premiums. In a Nash bargaining solution the two parties share the surplus from cooperation evenly.[22]

To illustrate the block premium as it evolves from a Nash bargaining solution, we look at the example from the preceding section where, in a one share–one vote regime, a rival investor paralyzes an entrepreneurial corporation. We assume that, if the rival took full control by buying out the family owner, the present value of the company's cash flows would be back to what it was prior to the rival investor taking a blocking minority interest. This situation implies that the rival has an incentive to pursue a cooperative strategy by bidding for the family's equity stake; in this way he could increase the value of his original position of 25 percent plus one share by a little more than $50,000. Conversely, it may be advantageous for the family to accept the bid. If, for instance, the family has lost all its private control benefits, selling out to the intruder becomes advantageous as it allows the family to reap capital gains on its 60 percent equity stake.

Two cooperative outcomes are conceivable. Either the family sells out to the rival investor, or the rival sells out to the family. If the family buys out the rival, he (or any other investor with similar preferences) would repeat this game ad infinitum. This is because, by selling out, the investor would generate gains from cooperation, which—in a Nash bargaining solution—are shared evenly by the two parties. This means that the rival investor does not only gain when acquiring the blocking minority interest, he also gains when selling it. Thus, the only viable strategy is that the family sells out to the rival (which assumes that the rival investor's wealth constraint is not binding).

In the noncooperative situation (in which the rival investor paralyzes the company), the family's wealth equals $1,080,000 (the equity interest of the family, which has lost all its control benefits). The wealth of the rival investor amounts to a little more than $525,000 (the rival investor's financial position plus his control benefits). Added up between the two parties, total wealth is little more than $1,605,000. If the family sells out, total wealth increases to a little more than $1,775,000 (assuming that the intruder's control benefits remain unchanged). The gain from cooperation equals a little more than $170,000, which is shared evenly between the two parties. Consequently, the block of shares changes hands at a little more than $1,165,000, which implies a price per share of about $0.97. This is $0.07 above the company's share price based on the present value of cash flows in the noncooperative state.

Compared with the situation before the intruder shows up at the company's gates, the family loses (a little less than) $185,000, while the intruder gains (a little less than) $110,000. Society as a whole loses $75,000, which is the difference between the family's control benefits ($150,000) and the intruder's private benefits ($75,000). The intruder winds up with 85 percent of the voting stock.

MANDATORY TENDER OFFERS

With a mandatory offer rule in place, an investor has to make a tender offer for the remaining shares once he has obtained control. The U.K. "City Code" offers the most prominent example of takeover regulation with a mandatory tender offer in place. Control in the U.K. City Code is defined as 30 percent of the voting stock.[23] Once an investor

[21] See, for instance, Franks and Mayer (2000).

[22] For a textbook example of the Nash bargaining solution, see Hart (1995).

[23] For "City Code on Takeovers and Mergers and the Rules Governing Substantial Acquisition of Shares," see < http://www.thetakeoverpanel.org.uk > .

reaches or crosses this threshold from below, he has to make an (unconditional) offer for all remaining shares. In the following we show that mandatory tender offers protect small shareholders against block trades in which the trading parties gain at the expense of the small shareholders.

In this first example, there is no mandatory tender offer rule in place. We look at a company with 2 million shares outstanding. All shares are voting stock. An institutional investor holds the majority of shares (50 percent plus one share), with the remaining shares being dispersed. The incumbent investor operates the corporation efficiently but—because it is an institution rather than an individual—generates no private control benefits. The company trades at $1 a share. We assume that there is no supermajority rule in place.

Suppose there is a private investor who, if he were in control, would enjoy private control benefits but would not run the firm efficiently. For instance, if the investor were to gain control, the private value of the firm would equal $150,000 and the public value of the firm would amount to $0.9 a share. The drop in the company's public value might be due to the investor's funding of perquisites through the company.

If control changed from the institutional to the private investor, society would lose $50,000. We show that, without a mandatory tender offer rule in place, the takeover indeed happens. In contrast, with mandatory tender offer regulation, the efficient outcome prevails.

As outlined above, in a Nash bargaining solution the two parties share the surplus from cooperation evenly. The surplus generated for the two parties when control changes from the institutional investor to the private investor equals a little less than $50,000. Splitting the surplus evenly would imply a price for the 50 percent-plus-one-vote block of a little more than $1,025,000. By selling out, the institutional investor gains a little less than $25,000, and so does the personal investor. Society overall loses $50,000. The small shareholders suffer a loss close to $100,000.

With mandatory takeover regulation in place, an investor does not succeed in taking over an efficiently operated firm if he is unable to add value to society. This is because the investor must acquire all shares. He always pays at least fair market value, be it in block trades or open market operations prior to taking control or in the mandatory tender offer.

While the arguments put forward in favor of mandatory tender offers are strong, it is noteworthy that this rule does not find unanimous support among traded corporations. For instance, Germany introduced a takeover code in 1995 as a voluntary guideline. As of April 11, 2000, only 540 of 933 listed German firms had signed the Takeover Code. Among the companies that have not signed on are BMW AG and Volkswagen AG.[24] A possible reason why companies find the code objectionable is that a mandatory tender offer rule does not allow them to hold minority positions in companies to protect relation-specific investments. Automobile companies frequently take positions in subcontractors to insulate themselves against opportunistic behavior. This position is most important where suppliers also provide part or all of the research and development that pertains to the delivered intermediate products.[25]

CONCLUSION

In a series of numerical examples we analyzed the impact of the one share–one vote principle, the simple majority rule, and mandatory tender offer regulation from the perspective of a socially optimal market for corporate control. Maximizing social welfare means maximizing the sum of the private and public values of the firm, rather than maximizing public value only. While our analysis is too simple to draw policy conclusions, we agree with Harris and Raviv (1988) that the simultaneous presence of the one share–one vote principle and the simple majority rule is generally optimal. At the same time, the analysis lends strong support to prohibiting restricted tender offers and to legalizing mandatory tender offers.

The simple majority rule ensures that the incumbent investor and the rival investor are on equal footing. The one share–one vote principle in combination with a mandatory tender offer regulation forces the rival investor to acquire all the cash flow rights if he wants to obtain control. This prevents value-decreasing takeovers because the rival investor succeeds only if he is able to raise the sum of the private and the public values of the firm beyond the level that comes with the incumbent investor.

[24] For details, see the German Takeover Commission's Web site at <http://www.kodex.de>. The site also posts the list of signatories. As a result of the low acceptance of the Takeover Code, the commission recommended to the legislature to write the code (in a revised form) into law.

[25] For a classic study on subcontracting relations in the automobile sector, see Asanuma (1989), who analyzes Toyota Motor Company.

Mandatory supermajority rules appear to be particularly harmful to society. On one hand, full control requires more than 50 percent (plus one share) of the voting stock, which puts the rival investor at a disadvantage in relation to the incumbent. On the other hand, a minority interest is sufficient to block important decisions. An investor who holds a minority interest can paralyze the firm and expropriate the incumbent investor of his private control benefits. The situation can be avoided with a mandatory tender offer where the offer threshold is set to the blocking minority threshold. Generally, mandatory supermajority rules should come with mandatory tender offers.

REFERENCES

Asanuma, Banri. "Manufacturer-Supplier Relationships in Japan and the Concept of Relation-Specific Skill." *Journal of the Japanese and International Economies*, March 1989, *3*(1), pp. 1-30.

Demsetz, Harold. "The Structure of Ownership and the Theory of the Firm." *Journal of Law and Economics*, June 1983, *26*(2), pp. 375-90.

Franks, Julian and Mayer, Colin. "Ownership and Control of German Corporations." Working Paper, University of Oxford, 2000.

Grossman, Sanford J. and Hart, Oliver D. "One Share–One Vote and the Market for Corporate Control." *Journal of Financial Economics*, January/March 1988, *20*(1/2), pp. 175-202.

Harris, Milton and Raviv, Artur. "Corporate Governance: Voting Rights and Majority Rules." *Journal of Financial Economics*, January/March 1988, *20*(1/2), pp. 203-35.

Hart, Oliver. *Firms, Contracts, and Financial Structure*. Oxford: Clarendon Press, 1995.

Mergerstat. *Mergerstat Review 2001*. Los Angeles: Applied Financial Information LP, 2001.

Milgrom, Paul. "Auctions and Bidding: A Primer." *Journal of Economic Perspectives*, Summer 1989, *3*(3), pp. 3-22.

_____ and Roberts, John. *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice Hall, 1992.

Myers, Stewart C. "Outside Equity." *Journal of Finance*, June 2000, *55*(3), pp. 1005-37.

Shleifer, Andrei and Summers, Lawrence H. "Breach of Trust in Hostile Takeovers," in Alan J. Auerbach, ed., *Corporate Takeovers: Causes and Consequences*. Chicago: University of Chicago Press, 1988, pp. 33-56.

_____ and Vishny, Robert W. "Large Shareholders and Corporate Control." *Journal of Political Economy*, June 1986, *94*(3, Part 1), pp. 461-88.

Thaler, Richard H. "Anomalies: The Winner's Curse." *Journal of Economic Perspectives*, Winter 1988, *2*(1), pp. 191-202.

Weston, J. Fred; Chung, Kwang S. and Siu, Juan A. *Takeovers, Restructuring, and Corporate Governance*. Second edition. Upper Saddle River, NJ: Prentice Hall, 1998.

Could a CAMELS Downgrade Model Improve Off-Site Surveillance?

R. Alton Gilbert, Andrew P. Meyer, and Mark D. Vaughan

The cornerstone of bank supervision is a regular schedule of thorough, on-site examinations. Under rules set forth in the Federal Deposit Insurance Corporation Improvement Act of 1991 (FDICIA), most U.S. banks must submit to a full-scope federal or state examination every 12 months; small, well-capitalized banks must be examined every 18 months. These examinations focus on six components of bank safety and soundness: capital protection (C), asset quality (A), management competence (M), earnings strength (E), liquidity risk exposure (L), and market risk sensitivity (S). At the close of each exam, examiners award a grade of one (best) through five (worst) to each component. Supervisors then draw on these six component ratings to assign a composite CAMELS rating, which is also expressed on a scale of one through five. (See the insert for a detailed description of the composite ratings.) In general, banks with composite ratings of one or two are considered safe and sound, whereas banks with ratings of three, four, or five are considered unsatisfactory. As of March 31, 2000, nearly 94 percent of U.S. banks posted composite CAMELS ratings of one or two.

Bank supervisors support on-site examinations with off-site surveillance. Off-site surveillance uses quarterly financial data and anecdotal evidence to schedule and plan on-site exams. Although on-site examination is the most effective tool for spotting safety-and-soundness problems, it is costly and burdensome. On-site examination is costly to supervisors because of the examiner resources required and burdensome to bankers because of the intrusion into daily operations. Off-site surveillance reduces the need for unscheduled exams. Off-site surveillance also helps supervisors plan exams by highlighting risk exposures at specific institutions.[1] For example, if pre-exam surveillance reports indicate that a bank has significant exposure to interest rate fluctuations, then supervisors will add interest-rate-risk specialists to the exam team.

The two most common surveillance tools are supervisory screens and econometric models. Supervisory screens are combinations of financial ratios, derived from quarterly bank balance sheets and income statements, that have given warning in the past about the development of safety-and-soundness problems. Supervisors draw on their experience to weigh the information content of these ratios. Econometric models also combine information from bank financial ratios. These models rely on statistical tests rather than human judgment to combine ratios, boiling the information from financial statements down to an index number that summarizes bank condition. In past comparisons, econometric models have outperformed supervisory screens as early warning tools (Gilbert, Meyer, and Vaughan, 1999; Cole, Cornyn, and Gunther 1995). Nonetheless, screens still play an important role in off-site surveillance. Supervisors can add screens quickly to monitor emerging sources of risk; econometric models can be modified only after new risks have produced a sufficient number of safety-and-soundness problems to allow re-specification and out-of-sample testing.

At the Federal Reserve, the off-site surveillance toolbox includes two distinct econometric models that are collectively known as SEER—the System for Estimating Examination Ratings. One model, the SEER risk rank model, uses the latest quarterly financial data to estimate the probability that each Fed-supervised bank will fail within the next two years. The other model, the SEER rating model, uses the latest financial data to produce a "shadow" CAMELS rating for each supervised institution. That is, the model estimates the CAMELS rating that examiners would have assigned had the bank been examined using the most recent set of financial

R. Alton Gilbert is a vice president and banking advisor, Andrew P. Meyer is an economist, and Mark D. Vaughan is a supervisory policy officer and economist at the Federal Reserve Bank of St. Louis. The authors thank economists Robert Avery, Jeffrey Gunther, James Harvey, Tom King, Jose Lopez, Don Morgan, Chris Neely, and David Wheelock; bank supervisors Carl Anderson, Kevin Bertsch, and Kim Nelson; and seminar participants at the meetings of the SEER Technical Working Group and the Western Economics Association for their comments. Judith Hazen provided research assistance.

[1] See Board of Governors (1996) for a description of risk-focused examination.

WHAT ARE CAMELS RATINGS?

CAMELS composite rating	Description
Safe and sound	
1	Financial institutions with a composite one rating are sound in every respect and generally have individual component ratings of one or two.
2	Financial institutions with a composite two rating are fundamentally sound. In general, a two-rated institution will have no individual component ratings weaker than three.
Unsatisfactory	
3	Financial institutions with a composite three rating exhibit some degree of supervisory concern in one or more of the component areas.
4	Financial institutions with a composite four rating generally exhibit unsafe and unsound practices or conditions. They have serious financial or managerial deficiencies that result in unsatisfactory performance.
5	Financial institutions with a composite five rating generally exhibit extremely unsafe and unsound practices or conditions. Institutions in this group pose a significant risk to the deposit insurance fund and their failure is highly probable.

NOTE: CAMELS is an acronym for six components of bank safety and soundness: capital protection (C), asset quality (A), management competence (M), earnings strength (E), liquidity risk exposure (L), and market risk sensitivity (S). Examiners assign a grade of one (best) through five (worst) to each component. They also use these six scores to award a composite rating, also expressed on a one-through-five scale. As a rule, banks with composite ratings of one or two are considered safe and sound while banks with ratings of three, four, or five are considered unsatisfactory.

SOURCE: *Federal Reserve Commercial Bank Examination Manual*.

statements and the previous CAMELS rating. Every quarter, analysts in the surveillance section at the Board of Governors feed the latest call report data into these models and forward the results to the 12 Reserve Banks. The Federal Deposit Insurance Corporation (FDIC) and the Office of the Comptroller of the Currency (OCC) also use statistical models in the off-site surveillance of the banks they supervise.[2]

The Federal Reserve employs two distinct models in off-site surveillance to accomplish two distinct objectives. One objective, embodied in the SEER risk rank model, is to identify a core set of financial variables that consistently foreshadows failure. Due to the paucity of bank failures since the early 1990s, the coefficients of the risk rank model were last estimated on data ending in 1991. A fixed-coefficient model, such as the risk rank model, allows surveillance analysts to gauge how much of any change in failure probabilities over time is due to changes in the values of these core financial variables. The second objective is to allow for changes over time in the relationship between financial performance today and bank condition tomorrow. The second half of the SEER framework, the SEER rating model, meets this objective by allowing analysts to reestimate the relationship quarterly, adjusting for any changes in the factors that produce safety-and-soundness problems.

Identifying banks with composite CAMELS ratings of one or two that are at risk of downgrade to a composite rating of three, four, or five is another important objective of the SEER framework, although this relationship is not directly estimated in either SEER model. Supervisors view a downgrade from safe-and-sound condition to unsatisfactory condition as serious because three-, four-, and five-rated banks are much more likely to fail. For example, Curry (1997) found that 74 percent of the banks that failed from 1980 through 1994 held three, four, or five composite CAMELS ratings two years prior to failure. Table 1 contains an update of Curry's

[2] See Reidhill and O'Keefe (1997) for a history of the off-site surveillance systems at the Federal Reserve, FDIC, and OCC.

figures, indicating that 53 of the 58 banks (91 percent) that failed in the years 1993 through 1998 held unsatisfactory ratings at least one year prior to failure. Because of their high failure risk, banks in unsatisfactory condition receive constant supervisory attention. An econometric model designed to flag safe-and-sound banks at risk of downgrade could help allocate supervisory resources not already devoted to troubled institutions. Such a model might also yield even earlier warning of emerging financial distress—warning that could reduce the likelihood of eventual failure by allowing earlier supervisory intervention. Although SEER failure probabilities and "shadow" CAMELS ratings for one- and two-rated banks certainly provide clues about downgrade risks, these index numbers are not the product of a model estimated specifically to flag downgrade candidates.

Even so, the SEER models may produce "watch lists" of one- and two-rated banks that differ little from watch lists produced by a downgrade-prediction model. The CAMELS downgrade model, the SEER risk rank model, and the SEER rating model generate ordinal rankings of banks based on risk. The models differ by the specific measure of overall risk—the risk of failure (SEER risk rank model), the risk of receiving a poor current CAMELS rating (SEER rating model), or the risk of moving from satisfactory to unsatisfactory condition in the near future (downgrade model). The models also differ by the sample of banks used for estimation—the SEER models are estimated on all commercial banks, whereas a downgrade model is estimated only on one- and two-rated institutions. But if the financial factors that explain CAMELS downgrades differ little from the financial factors that explain failures or CAMELS ratings, then all three models will produce similar risk rankings and, hence, similar watch lists of one- and two-rated banks. Only formal empirical tests can determine the potential contribution of a downgrade-prediction model to off-site surveillance at the Federal Reserve.

To answer our title question—could a CAMELS downgrade model improve off-site surveillance—we compare the out-of-sample performance of a downgrade-prediction model and the SEER models using 1990s data. We find only slight differences in the ability of the three models to spot emerging financial distress among safe-and-sound banks. Specifically, in out-of-sample tests for 1992 through 1998, the watch lists produced by the downgrade-prediction model outperform the watch lists produced by the SEER models by only a small margin. We conclude that, in relatively tranquil banking environments like the 1990s, a downgrade model adds little value in off-site surveillance. We caution, however, that a downgrade-prediction model might prove useful in more turbulent banking times.

THE RESEARCH STRATEGY

Our downgrade-prediction model is a probit regression that uses bank financial data to estimate the probability each sample bank will tumble from a composite CAMELS rating of one or two to a composite CAMELS rating of three, four, or five. Specifically, the dependent variable takes a value of one for any bank whose CAMELS rating falls from satisfactory to unsatisfactory in the 24 months following the quarter of the financial data; the dependent variable is zero if the bank is examined but not downgraded in the 24-month window. Although bank failure declined dramatically in the 1990s, CAMELS downgrades were still common, thereby allowing frequent reestimation of the model. (See Table 2 for data on CAMELS downgrades in the 1990s.) The SEER risk rank model is also a probit model, using financial data to estimate the probability that a Fed-supervised bank will fail or see its tangible capital fall below 2 percent of total assets in the next 24 months. The SEER rating model is a multinomial logit regression that uses financial data to estimate a "shadow" CAMELS rating—the composite rating that examiners would have awarded had the bank been examined that quarter. A multinomial logit differs from a standard logit by predicting a range of discrete values (in this case CAMELS composite ratings, which range from one to five) rather than two discrete values (failure/no failure or downgrade/no downgrade).

The explanatory variables for the downgrade-prediction model include a set of financial performance ratios and a bank size variable that all appear in the SEER risk rank model, as well as two additional CAMELS-related variables. Table 3 describes the explanatory variables and the expected relationship between each variable and the likelihood of a future downgrade. The financial performance ratios capture the impact of leverage risk, credit risk, and liquidity risk—three risks that have consistently produced financial distress in commercial banks (Putnam, 1983; Cole and Gunther, 1998). The bank size and CAMELS-related variables capture the impact of other factors that may affect downgrade risk.

The downgrade-prediction model captures leverage risk with total equity minus goodwill as a

Table 1

How Often Did Unsatisfactory Banks Fail in the 1990s?

Year of failure	CAMELS rating at least one year prior to failure	Number of banks in each CAMELS cohort	Number of failures in each CAMELS cohort	Percentage failed in each CAMELS cohort	Percentage of all failures with CAMELS ratings of 3, 4, or 5 one year in advance
1993	1	2,396	1	0.04	91.7
	2	6,549	2	0.03	
	3	**1,877**	**4**	**0.21**	
	4	**762**	**14**	**1.84**	
	5	**218**	**15**	**6.88**	
1994	1	2,508	0	0.00	90.9
	2	6,693	1	0.01	
	3	**1,578**	**0**	**0.00**	
	4	**562**	**5**	**0.89**	
	5	**124**	**5**	**4.03**	
1995	1	3,299	0	0.00	100
	2	6,469	0	0.00	
	3	**916**	**0**	**0.00**	
	4	**303**	**2**	**0.66**	
	5	**56**	**3**	**5.36**	
1996	1	3,759	0	0.00	75.0
	2	5,995	1	0.02	
	3	**587**	**1**	**0.17**	
	4	**158**	**1**	**0.63**	
	5	**39**	**1**	**2.56**	
1997	1	4,041	0	0.00	100
	2	5,472	0	0.00	
	3	**400**	**0**	**0.00**	
	4	**91**	**1**	**1.10**	
	5	**23**	**0**	**0.00**	
1998	1	4,328	0	0.00	100
	2	4,941	0	0.00	
	3	**329**	**0**	**0.00**	
	4	**57**	**1**	**1.75**	
	5	**15**	**0**	**0.00**	

NOTE: This Table shows that banks with composite CAMELS ratings of one or two were less likely to fail in the 1990s than were banks with composite ratings of three, four, or five. The number of failed banks that were classified as unsatisfactory banks (CAMELS three, four, or five composite ratings) at least one year prior to failure are shown in bold. Supervisors recognized that these banks were significant failure risks and, therefore, monitored them closely. Because supervisors do not monitor CAMELS one- and two-rated banks as closely, they are interested in a tool that can identify which of these institutions is most likely to encounter financial distress.

Table 2

How Common Were CAMELS Downgrades in the 1990s?

Year of downgrade	CAMELS rating at beginning of year	Number of banks	Number of banks downgraded to unsatisfactory status	Percentage of banks downgraded to unsatisfactory status	Total number of downgrades to unsatisfactory status
1990	1	2,182	38	1.74	728
	2	5,572	690	12.38	
1991	1	2,189	34	1.55	698
	2	5,475	664	12.13	
1992	1	1,959	22	1.12	424
	2	5,275	402	7.62	
1993	1	2,289	7	0.31	182
	2	5,976	175	2.93	
1994	1	2,910	9	0.31	162
	2	5,717	153	2.68	
1995	1	3,091	8	0.26	102
	2	4,885	94	1.92	
1996	1	3,260	10	0.31	126
	2	4,487	116	2.59	
1997	1	3,223	7	0.22	123
	2	3,719	116	3.12	
1998	1	3,006	19	0.63	153
	2	3,090	134	4.34	

NOTE: This Table demonstrates that downgrades from safe-and-sound to unsatisfactory status were common in the 1990s, thereby making it possible to reestimate a downgrade-prediction model on a yearly basis. Specifically, the far right column shows the number of sample banks rated as safe and sound (CAMELS one or two) at each year-end that were downgraded to unsatisfactory status (CAMELS three, four, or five) within the following year. Note that two-rated banks were much more likely to slip into unsatisfactory status than one-rated banks. Note also that the percentage of banks suffering downgrades to unsatisfactory status fell as overall banking performance improved in the mid-1990s, but the trend reversed in the late 1990s.

percentage of total assets (NET WORTH) and net income as a percentage of total assets (or, return on assets [ROA]). Leverage risk is the risk that losses will exceed capital, rendering a bank insolvent. We expect higher levels of capital (lower leverage risk) to reduce the likelihood of CAMELS downgrades. We include ROA in the leverage risk category because retained earnings are an important source of additional capital for many banks and because higher earnings provide a greater cushion for withstanding adverse economic shocks (Berger, 1995). We expect that higher earnings reduce the risk of a future downgrade.

The downgrade-prediction model captures credit risk with the ratio of loans 30 to 89 days past due to total assets (PAST-DUE 30), the ratio of loans over 89 days past due to total assets (PAST-DUE 90), the ratio of loans in nonaccrual status to total assets (NONACCRUING), the ratio of other real estate owned to total assets (OREO), the ratio of commercial and

Table 3

What Factors Help Predict Downgrades to Unsatisfactory Condition (CAMELS Three, Four, or Five)?

Independent variables (risk proxies)	Symbol	Hypothesized relationship
Leverage risk		
Total net worth (equity capital minus goodwill) as a percentage of total assets	NET WORTH	–
Net income as a percentage of average assets (return on average assets)	ROA	–
Credit risk		
Loans past due 30-89 days as a percentage of total assets	PAST-DUE 30	+
Loans past due 90+ days as a percentage of total assets	PAST-DUE 90	+
Nonaccrual loans as a percentage of total assets	NONACCRUING	+
Other real estate owned as a percentage of total assets	OREO	+
Commercial and industrial loans as a percentage of total assets	COMMERCIAL LOANS	+
Residential real estate loans as a percentage of total assets	RESIDENTIAL LOANS	–
Liquidity risk		
Book value of securities as a percentage of total assets	SECURITIES	–
Deposits >$100M (jumbo CDs) as a percentage of total assets	LARGE TIME DEPOSITS	+
Non-financial variables		
Natural logarithm of total assets, in thousands of dollars	SIZE	?
Dummy variable equal to 1 if bank has a CAMELS rating of 2	CAMELS-2	+
Dummy variable equal to 1 if the bank's management rating is worse than its composite CAMELS rating	BAD MANAGE	+

NOTE: This Table lists the independent variables used in the downgrade-prediction model. The signs indicate the hypothesized relationship between each variable and the likelihood of a downgrade from satisfactory status (a CAMELS one or two composite rating) to unsatisfactory status (a CAMELS three, four, or five rating). For example, the negative sign for the net worth ratio indicates that, other things equal, higher net worth today reduces the likelihood of a downgrade to unsatisfactory status tomorrow.

industrial loans to total assets (COMMERCIAL LOANS), and the ratio of residential real estate loans to total assets (RESIDENTIAL LOANS). Credit risk is the risk that borrowers will fail to make promised interest and principal payments. The model contains six measures of credit risk because this risk was the driving force behind bank failures in the late 1980s and early 1990s (Hanc, 1997). We include the past-due and nonaccruing loan ratios because banks charge off higher percentages of these loans than loans whose payments are current.[3] We include other real estate owned, which consists primarily of collateral seized after loan defaults, because a high OREO ratio often signals poor credit risk management—either because a bank has had to foreclose on a large number of loans or because it has had trouble disposing of seized collateral. PAST-DUE 30, PAST-DUE 90, NONACCRUING, and OREO are backward-looking because they register asset quality problems that have already emerged (Morgan and Stiroh, 2001). To give the model a forward-looking dimension, we add the commercial-and-industrial-loan ratio because, historically, the charge-off rate for these loans has been higher than for other types of loans. We also employ the residential real estate ratio because, historically, losses on these loans have been relatively low. With the exception of the residential loan ratio, we expect a positive relationship between the credit risk measures and downgrade probability.

The downgrade-prediction model captures liquidity risk with investment securities as a per-

[3] In bank accounting, loans are classified as either accrual or nonaccrual. As long as a loan is classified as accrual, the interest due is counted as current revenue, even if the borrower falls behind on interest payments.

centage of total assets (SECURITIES) and jumbo certificates of deposit (CDs) as a percentage of total assets (LARGE TIME DEPOSITS). Liquidity risk is the risk that a bank will be unable to fund loan commitments or meet withdrawal demands at a reasonable cost. A larger stock of liquid assets—such as investment securities—indicates a greater ability to meet unexpected liquidity needs and should, therefore, translate into a lower downgrade probability. Liquidity risk also depends on a bank's reliance on non-core funding. Core funding—which includes checking accounts, savings accounts, and small time deposits—is relatively insensitive to the difference between the interest rate paid by the bank and the market rate. Non-core funding—which includes jumbo CDs—can be quite sensitive to interest rate differentials. All other things equal, greater reliance on jumbo CDs implies a greater likelihood of a funding runoff or an interest expense shock and, hence, a future CAMELS downgrade.

The downgrade-prediction model also includes variables that capture the impact of asset size, bank heterogeneity, and management competence on downgrade risk. We add the natural logarithm of total assets (SIZE) because large banks can reduce risk by diversifying across product lines and geographic regions. As Demsetz and Strahan (1997) have noted, however, geographic diversification relaxes a constraint, enabling bankers to assume more risk, so we make no prediction about the relationship between size and downgrade probability. We include a dummy variable equal to one if a bank's composite CAMELS rating is two; we do this because two-rated banks tumble into unsatisfactory status more often than one-rated banks. (See Table 2 for data on the downgrade rates for one- and two-rated institutions.) Finally we employ a dummy variable (BAD MANAGE) equal to one if the management component of the CAMELS rating is higher (weaker) than the composite rating. In these cases, examiners have registered concerns about the quality of bank management, even though these problems have yet to produce financial consequences.

After estimating the downgrade-prediction model, we use all three models to produce rank orderings, or "watch lists," of one- and two-rated banks. With the downgrade model, the list ranks safe-and-sound banks from the highest probability of tumbling into unsatisfactory condition to the lowest. With the SEER risk rank model, the list ranks safe-and-sound banks from the highest probability of failing to the lowest. With the SEER rating model, the list ranks safe-and-sound banks from the highest (weakest) shadow CAMELS rating to the lowest. Although each model produces a different index number, they all may produce similar ordinal rankings. Supervisors could use the SEER framework to monitor safe-and-sound banks by focusing on the riskiest one- or two-rated banks as identified by either the rating or failure-prediction model. Again, only a formal test of out-of-sample performance can gauge the value added by a customized downgrade-prediction model. Out-of-sample tests—which use an evaluation period subsequent to the estimation period—are crucial because supervisors use econometric models this way in practice.

We compare out-of-sample performance of the watch lists by examining the type-one and type-two error rates associated with each list. Type-one errors are sometimes called false negatives; type-two errors are false positives. Each type of error is costly to supervisors. A missed downgrade—a type-one error—is costly because an accurate downgrade prediction gives supervisors more warning about emerging financial distress, and early intervention reduces the likelihood of failure. A type-two error occurs when a predicted downgrade does not materialize. An over-predicted downgrade is costly because it wastes scarce supervisory resources on a healthy bank. Type-two errors also impose unnecessary costs on healthy banks because on-site examinations disrupt day-to-day operations.

Following Cole, Cornyn, and Gunther (1995), we generate power curves for the three watch lists that indicate the minimum achievable type-one error rates for any desired type-two error rate. (These curves are illustrated in Figures 1 and 2.) Power curves allow comparison of each list's ability to reduce false negatives and false positives simultaneously. A more theoretically appealing approach would minimize a loss function that places an explicit weight on the benefits of early warning about financial distress and the costs of wasted examination resources and unnecessary disruption of bank activities. The relative performance of the watch lists could then be assessed for the optimal type-one (or type-two) error rate. Unfortunately, the data necessary to pursue such an approach are unavailable. Without concrete data about supervisor loss functions, we opt for power curves that make no assumptions about the weights that should be placed on type-one and type-two errors. This approach also allows supervisors to use our results to compare model performance over any desired range of error rates.

For example, the SEER risk rank power curve shows the type-one and type-two error rates when an ordinal ranking based on failure probability is interpreted as a rank ordering of downgrade risk. We trace out the curve by starting with the assumption that no one- or two-rated bank is a downgrade risk. This assumption implies that all subsequent downgrades are surprises, making the type-one error rate 100 percent. In this case, the type-two error rate is zero because no banks are incorrectly classified as downgrade risks. We obtain the next point by selecting the one- or two-rated bank with the highest failure probability. If the selected bank suffers a subsequent downgrade, then the type-one error rate for the SEER risk rank watch list decreases slightly. The type-two error rate remains at zero because, again, no institutions are incorrectly classified as downgrade risks. If the selected bank does not suffer a downgrade, then the type-one error rate remains at 100 percent and the type-two error rate increases slightly. By selecting banks in order of their failure probability and recalculating type-one and type-two error rates, we can trace out a power curve. At the lower right extreme of the curve, the entire failure probability rank ordering is considered at risk of a downgrade. At this extreme, the SEER risk rank watch list posts a type-one error rate equal to zero percent and a type-two error rate equal to 100 percent.

The area under the power curves provides a basis for comparing the out-of-sample performance of each watch list. A smaller area implies a lower overall type-one and type-two error rate and a more accurate model. We express the area for each watch list as a percentage of the total area in the box. A useful benchmark is the case in which downgrade risks are selected at random. Random selection of one- and two-rated banks, over a large number of trials, produces power curves with an average slope of –1. The area under a "random" watch list power curve equals, on average, 50 percent of the area of the entire box.

THE DATA

We exploit two data sources for our analysis—the Federal Financial Institutions Examination Council (FFIEC) and the National Information Center of the Federal Reserve System (NIC). We use income and balance sheet data from the Reports of Condition and Income (the call reports), which are collected under the auspices of the FFIEC. The FFIEC requires all commercial banks to submit quarterly call reports to their principal supervisors; most call report items are available to the public. We rely on CAMELS composite and management ratings from the NIC database. This database is available to examiners and analysts in the banking supervision function of the Federal Reserve System but not to the public. We also draw on the NIC database for the SEER failure probabilities and "shadow" CAMELS ratings.

To ensure an unbiased comparison of the models, we exclude any bank with an operating history under five years from the estimation sample for the downgrade-prediction model. The financial ratios of these start-up, or *de novo*, banks often take extreme values that do not signal safety-and-soundness problems (DeYoung, 1999). For example, *de novos* often lose money in their early years, so their earnings ratios are poor. These extreme values distort model coefficients and could compromise the relative performance of the downgrade-prediction model. Another reason for excluding *de novos* is that supervisors already monitor these banks closely. The Federal Reserve conducts a full-scope on-site examination every six months for a newly chartered state-member bank.[4] Full-scope exams continue on this schedule until the *de novo* earns a one or two composite CAMELS rating for two consecutive exams.

As an additional safeguard, we use a timing convention for estimating the downgrade-prediction model that corresponds to the timing convention used to estimate the SEER risk rank model. Specifically, we estimate the downgrade model six times—each time using financial data for one- and two-rated institutions in the fourth quarter of year t and downgrade status (1 = downgrade, 0 = no downgrade) in years $t + 1$ and $t + 2$. For example, to produce the first downgrade equation (reported as the "1990-91" equation in Table 4), we use a sample of banks rated CAMELS one or two as of December 31, 1989. We then regress downgrade status during 1990 and 1991 on fourth quarter 1989 data. A bank that is examined but maintains a one or two rating during the entire two-year period is classified as "no downgrade." A bank that is examined and suffers a downgrade to a three, four, or five composite rating anytime in the two-year period is classified as "downgrade."

Finally, when comparing out-of-sample performance of the models, we note biases that result from

[4] The Federal Reserve supervises bank holding companies and state-chartered banks that belong to the Federal Reserve System. The FDIC supervises state-chartered banks that do not belong to the Federal Reserve System. The OCC supervises banks chartered by the federal government.

using revised call report data rather than originally submitted call report data. Supervisors sometimes require banks to revise their call report data after an on-site examination. Indeed, some economists have argued that this auditing function is the principal value of examinations (Berger and Davies, 1998; Flannery and Houston, 1999). Revisions of fourth quarter data tend to be particularly large because banks strive to make their year-end financial reports look as healthy as possible (Allen and Saunders, 1992). Gunther and Moore (2000) have found that early warning models estimated on revised data outperform models estimated on originally submitted data. Because of this evidence, estimation and simulation of an early warning model with the original data, rather than the revised data, would provide a more appropriate test of the value of a model for surveillance. The original data, however, are not available for all banks and all periods. Hence, we estimate the downgrade model on revised rather than original call report data. The coefficients of the SEER risk rank model were estimated using revised call report data, and we apply these coefficients to revised call report data to generate failure probability rankings. Because the SEER risk rank model and the downgrade-prediction model are estimated with revised data, our performance comparisons do not favor either model ex ante. But because the SEER rating model was estimated on originally submitted call report data, out-of-sample comparisons favor the downgrade-prediction model over the rating model. Data limitations do not allow us to correct for this bias, so we bear it in mind as we interpret the power curve evidence for these two models.

IN-SAMPLE FIT OF THE DOWNGRADE-PREDICTION MODEL

As noted, we estimate the downgrade-prediction model six times—first regressing downgrade status in 1990 and 1991 on fourth quarter 1989 financial data, then regressing downgrade status in 1991 and 1992 on fourth quarter 1990 data, and so on, up through regressing downgrade status in 1995 and 1996 on fourth quarter 1994 data. The results of these regressions appear in Table 4.

Overall, the downgrade-prediction model fits the data relatively well in-sample. For each of the six regressions, the log-likelihood test statistic allows rejection of the hypothesis that all model coefficients equal zero at the 1 percent level of significance. The pseudo-R^2, which indicates the approximate proportion of the variance of downgrade/no downgrade status explained by the model, ranges from a low of 14.9 percent for the 1993-94 equation to a high of 22.4 percent for the 1991-92 equation. These pseudo-R^2 numbers may seem low, particularly when viewed against the figures for failure-prediction models—the pseudo-R^2 for the SEER risk rank model is 63.2 percent—but CAMELS downgrades are less severe than outright failures and, therefore, much more difficult to forecast. In this light, the pseudo-R^2 figures look more respectable. The estimated coefficients on eight explanatory variables—the jumbo-CD-to-total-asset ratio, the net-worth-to-total-asset ratio, the past-due and nonaccruing loan ratios, the net-income-to-total-asset ratio, and the two CAMELS dummy variables—are statistically significant with the expected sign in all six equations. The coefficient on the size variable has a mixed-sign pattern, which is not surprising, given the theoretical ambiguity in the relationship between bank size and risk. The coefficients on the other four explanatory variables are statistically significant with the expected sign in at least three of the six equations.

The in-sample fit of the downgrade-prediction model does deteriorate slightly through time. The log-likelihood statistic declines monotonically from the 1991-92 equation through the 1995-96 equation. Indeed, the psuedo-R^2 averages 20.7 percent for the first three equations (1990-91, 1991-92, 1992-93) and 16.5 percent for the last three equations (1993-94, 1994-95, 1995-96). The number of statistically significant coefficients with expected signs also declines slightly over the estimation years. For instance, the coefficients on the commercial-and-industrial-loan-to-total-asset ratio are statistically significant with the expected sign in the first three equations but in only one of the last three equations (1995-96). The monotonic deterioration in model fit reflects the decline in the number of downgrades. In the first three regressions, the average number of downgrades per year was 500; in the last three regressions, the average dropped to 127 downgrades per year.

OUT-OF-SAMPLE PERFORMANCE COMPARISONS OF THE SEER RISK RANK MODEL, THE SEER RATING MODEL, AND THE DOWNGRADE-PREDICTION MODEL

With a timing convention that mimics the way supervisors use econometric models in surveillance,

Table 4

How Well Did the CAMELS Downgrade-Prediction Model Perform In-Sample?

	Years of downgrades in CAMELS ratings		
Explanatory variables	1990-91	1991-92	1992-93
Intercept	–2.053*** (0.232)	–0.923*** (0.249)	–0.284 (0.290)
COMMERCIAL LOANS	0.010*** (0.003)	0.013*** (0.003)	0.012*** (0.003)
RESIDENTIAL LOANS	–0.005** (0.002)	–0.003 (0.002)	–0.004 (0.003)
LARGE TIME DEPOSITS	0.017*** (0.003)	0.018*** (0.003)	0.014*** (0.004)
NET WORTH	–0.053*** (0.008)	–0.050*** (0.010)	–0.049*** (0.011)
PAST-DUE 90	0.396*** (0.038)	0.304*** (0.039)	0.232*** (0.045)
PAST-DUE 30	0.100*** (0.021)	0.136*** (0.021)	0.151*** (0.025)
NONACCRUING	0.227*** (0.027)	0.201*** (0.030)	0.188*** (0.035)
ROA	–0.242*** (0.031)	–0.330*** (0.038)	–0.104*** (0.038)
SECURITIES	–0.015*** (0.002)	–0.017*** (0.002)	–0.014*** (0.002)
OREO	0.212*** (0.030)	0.210*** (0.032)	0.021 (0.033)
SIZE	0.076*** (0.016)	–0.029* (0.017)	–0.128*** (0.022)
CAMELS-2	0.622*** (0.060)	0.542*** (0.067)	0.577*** (0.081)
BAD MANAGE	0.488*** (0.050)	0.405*** (0.053)	0.429*** (0.058)
Number of observations	8,927	8,636	8,361
Pseudo-R^2	0.218	0.224	0.179
–2 log likelihood testing whether all coefficients (except the intercept) = 0	5,909.617***	5,020.667***	3,476.658***

NOTE: This Table contains the estimated regression coefficients for the downgrade-prediction model. The model regresses downgrade status (1 for a downgrade and 0 for no downgrade) in calendar years $t+1$ and $t+2$ on explanatory variables from the fourth quarter of year t. See Table 3 for the definitions of the explanatory variables. Standard errors appear in parentheses next to the coefficients. One asterisk denotes significance at the 10 percent level, two asterisks denote significance at the 5 percent level, and three asterisks denote significance at the 1 percent level. Shading highlights coefficients that were significant with the expected sign in all six years. The pseudo-R^2 gives the approximate proportion of the total variance of downgrade status explained by the model. Overall, the downgrade-prediction model predicts in-sample downgrades well. Eight of the 13 regression variables are significant with the predicted sign in all six years, and all of the variables are significant in at least some years. Note that, by most measures of in-sample fit, the model declines in power over time, primarily due to the decrease in the number of downgrades.

we conduct six separate tests of the out-of-sample performance of the downgrade-prediction model. As noted, the first downgrade-prediction model regresses downgrade status in 1990 and 1991 on year-end 1989 financial data. By the end of 1991, supervisors would have had coefficient estimates from that regression. Our first out-of-sample test applies those coefficients to year-end 1991 financial ratios to compute downgrade probabilities for each sample bank. We then use the ranking of downgrade probabilities to construct power curves for type-one and type-two errors over the 1992-93 test window. To ensure compatibility between the in-sample and out-of-sample data, we limit the first out-of-sample test to banks with five-year operating histories, with CAMELS ratings of one or two as of year-end 1991, and with at least one full-scope examination in 1992 or 1993. The next five out-of-sample tests of the downgrade-prediction model—for the 1993-94, 1994-95, 1995-96, 1996-97, and 1997-98 windows—employ the same timing convention and the same sample restrictions.

Our out-of-sample tests of the SEER risk rank and the SEER rating models use the same timing convention as the out-of-sample tests of the downgrade-prediction model. Specifically, we apply the fixed SEER risk rank coefficients to year-end 1991 data and rank the one- and two-rated banks by their estimated probabilities of failure. We then derive a power curve reflecting the type-one and type-two

Table 4 cont'd

How Well Did the CAMELS Downgrade-Prediction Model Perform In-Sample?

	Years of downgrades in CAMELS ratings		
Explanatory variables	**1993-94**	**1994-95**	**1995-96**
Intercept	0.340 (0.358)	–0.809** (0.379)	0.069 (0.425)
COMMERCIAL LOANS	0.005 (0.005)	0.007 (0.005)	0.013** (0.005)
RESIDENTIAL LOANS	–0.005 (0.003)	–0.002 (0.003)	–0.013*** (0.004)
LARGE TIME DEPOSITS	0.018*** (0.005)	0.023*** (0.005)	0.021*** (0.005)
NET WORTH	–0.094*** (0.014)	–0.025* (0.013)	–0.034*** (0.012)
PAST-DUE 90	0.329*** (0.058)	0.286*** (0.063)	0.324*** (0.073)
PAST-DUE 30	0.169*** (0.032)	0.113*** (0.034)	0.162*** (0.035)
NONACCRUING	0.148*** (0.046)	0.183*** (0.045)	0.146*** (0.050)
ROA	–0.137*** (0.040)	–0.252*** (0.050)	–0.162*** (0.038)
SECURITIES	–0.007*** (0.002)	–0.003 (0.003)	–0.010*** (0.003)
OREO	0.080* (0.041)	0.193*** (0.044)	0.154*** (0.052)
SIZE	–0.171*** (0.027)	–0.149*** (0.030)	–0.210*** (0.034)
CAMELS-2	0.444*** (0.095)	0.625*** (0.103)	0.590*** (0.102)
BAD MANAGE	0.453*** (0.066)	0.406*** (0.073)	0.515*** (0.078)
Number of observations	8,600	9,169	9,200
Pseudo-R^2	0.149	0.153	0.193
–2 log likelihood testing whether all coefficients (except the intercept) = 0	2,248.122***	1,911.719***	1,628.444***

errors of this ordinal ranking, assuming that a higher failure probability at year-end 1991 indicates a higher downgrade probability in 1992 and 1993. For each year of the sample, we repeat this procedure, applying the fixed SEER risk rank model coefficients to the end-of-year call report data for one- and two-rated banks. Because SEER rating model estimates are not available for 1991 and 1992, we start out-of-sample testing of this model with "shadow" CAMELS ratings based on year-end 1993 data. We derive a power curve for the ordinal ranking of shadow CAMELS ratings based on the assumption that higher (weaker) estimated ratings indicate higher downgrade risk in 1994 and 1995. We use the same timing convention for the remaining three out-of-sample tests of the rating model (1995-96, 1996-97, and 1997-98).

Using any of the three models to flag downgrade candidates markedly improves the results compared with randomly selecting one- and two-rated banks. Panel A of Table 5 presents the results of the out-of-sample performance tests of the downgrade-prediction model and the two SEER models. Figures 1 and 2 offer the same information in visual form. Over the four test windows that include both SEER models—1994-95 through 1997-98—the average area under the power curves for the three models is 20.78 percent, substantially less than the 50 percent area under the power curve for random selection. Over all six test windows—1992-93 through 1997-98—the average of the area under the downgrade-prediction power curve and the SEER risk rank power curve equals 21.41 percent. Across individual models and individual years, the areas range from a high of 26.59 percent for the SEER risk rank model in flagging 1994-95 downgrades to a low of 15.14 percent for the downgrade-prediction model in flagging 1996-97 downgrades.

Overall, the downgrade-prediction model slightly outperforms the two SEER models in the out-of-sample performance comparisons. Over four tests covering the years 1994 through 1998, the downgrade-prediction model produces an average power curve area of 18.48 percent, whereas the two SEER models, on average, produce an area of 21.93 percent. Over six tests covering 1992 through 1998,

Table 5

How Did the Out-of-Sample Performance of the Downgrade-Prediction Model and the SEER Models Compare?

Panel A: Area under power curves

Downgrade years	Downgrade model (%)	SEER risk rank model (%)	SEER rating model (%)
1992-93	21.01	22.06	NA
1993-94	22.64	25.54	NA
1994-95	22.31	26.59	21.88
1995-96	17.40	21.45	22.13
1996-97	15.14	19.09	19.35
1997-98	19.08	24.62	20.30
Mean over all years	**19.60**	**23.23**	**20.92**

Panel B: Area under power curves below 20 percent type-two error rate

Downgrade years	Downgrade model (%)	SEER risk rank model (%)	SEER rating model (%)
1992-93	12.03	12.32	NA
1993-94	12.14	12.61	NA
1994-95	11.87	12.70	11.95
1995-96	10.72	11.66	11.82
1996-97	10.28	11.24	11.52
1997-98	10.92	12.28	11.73
Mean over all years	**11.33**	**12.14**	**11.76**

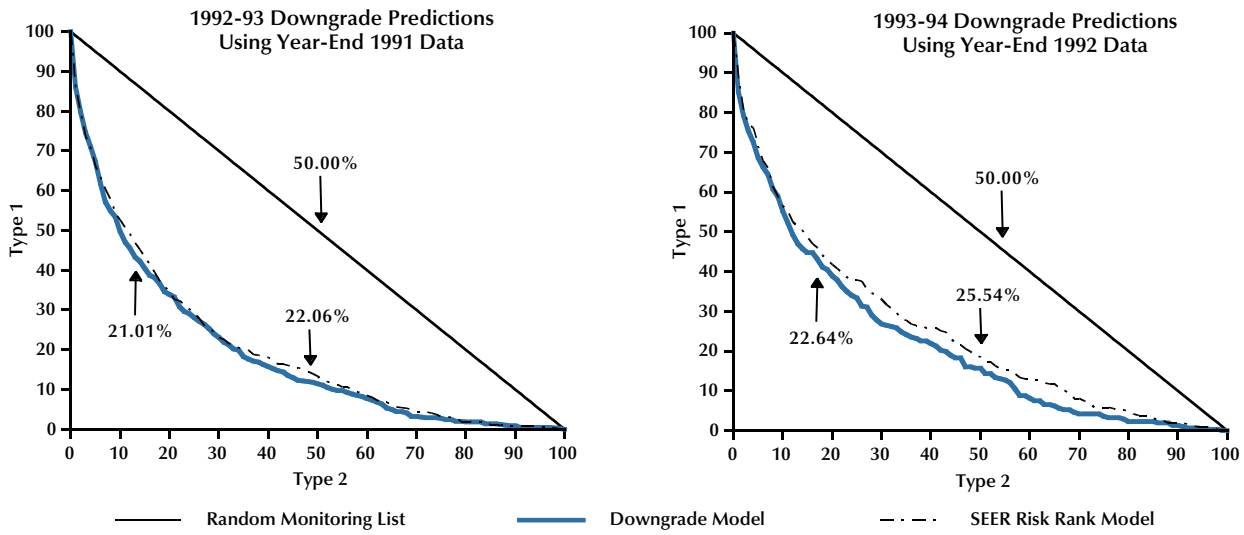
NOTE: This Table contains the areas under each model's power curve for each two-year test window. Each power curve reveals the trade-offs between type-one errors (missed downgrades) and type-two errors (over-predicted downgrades) for a particular model. We assess relative performance by comparing areas under the curves; smaller is better because smaller areas imply simultaneous reduction of both types of errors. The SEER rating model data were not available before 1993, so the Table contains no shadow CAMELS areas for the 1992-93 and 1993-94 test windows. When comparing areas, we bear in mind that the area produced by a randomly generated watch list equals, on average, 50 percent. Although all three models improve considerably over random selection of downgrade candidates, the downgrade-prediction model does not materially outperform the two SEER models (Panel A). When the maximum allowable type-two error rate is 20 percent, the results are virtually identical (Panel B). We use this cut-off as representative of model comparisons when supervisors insist on small watch lists.

the downgrade-prediction model generates an average area of 19.60 percent; the SEER risk rank model generates an average area of 23.23 percent. In each of the six test windows, the downgrade-prediction model outperforms the SEER risk rank model, the difference in area ranging from 1.05 percentage points for the 1992-93 window to 5.54 percentage points for the 1997-98 window. The downgrade-prediction model outperforms the SEER rating model in three of the four test windows, with area differentials ranging from 1.22 percentage points for the 1997-98 test to 4.73 percentage points for the 1995-96 test. Only in the 1994-95 test did the SEER rating model outperform the downgrade model—by an area differential of 0.43 percentage points.

Still, the difference between the out-of-sample performance of the downgrade-prediction model and the SEER models is quite small. On average over all tests, the downgrade model outperforms the SEER models by an area differential of just 2.48 percentage points. Over the restricted area (with less than a 20 percent type-two error), the area differential is even smaller—only 0.62 percentage points. At the same time, on average, the two SEER models outperform random selection by an area differential of 27.93 percentage points. Moreover, the out-of-sample tests are biased in favor of the downgrade-prediction model in two respects: the

Figure 1

How Well Do the Models Predict Out-of-Sample CAMELS Downgrades?



NOTE: This Figure shows that the SEER risk rank model and the downgrade model have similar type-one vs. type-two tradeoffs for most of the range of errors for 1992-93 and 1993-94 downgrades. The downgrade model slightly edges out the SEER failure model by 21.01 percent to 22.06 percent for the 1992-93 downgrades, and by 22.64 percent to 25.54 percent for the 1993-94 downgrades. The SEER rating model numbers were not available before 1993, so a SEER rating model power curve does not appear in the Figure.

This Figure depicts the trade-off between the type-one error rate and the type-two error rate for the SEER risk rank model, SEER rating model, and the downgrade model. The type-one error rate is the number of missed downgrades (false negatives) divided by the total number of CAMELS one- and two-rated banks; the type-two error rate is the number of incorrectly flagged downgrades (false positives) divided by the total number of CAMELS one- and two-rated banks. The area under each curve, divided by the total area in the box, offers a convenient way to compare the performance of each model. Smaller areas imply lower levels of *both* types of errors and, hence, better model performance. The 50 percent line indicates the type-one and type-two error rates associated with random selection of one- and two-rated banks.

coefficients on the SEER risk rank model have been fixed since 1991, and the SEER rating model is estimated on originally submitted call report data. The small difference in performance, particularly when viewed in light of these potential biases, suggests that the SEER models and our customized downgrade-prediction model flag downgrade candidates equally well.

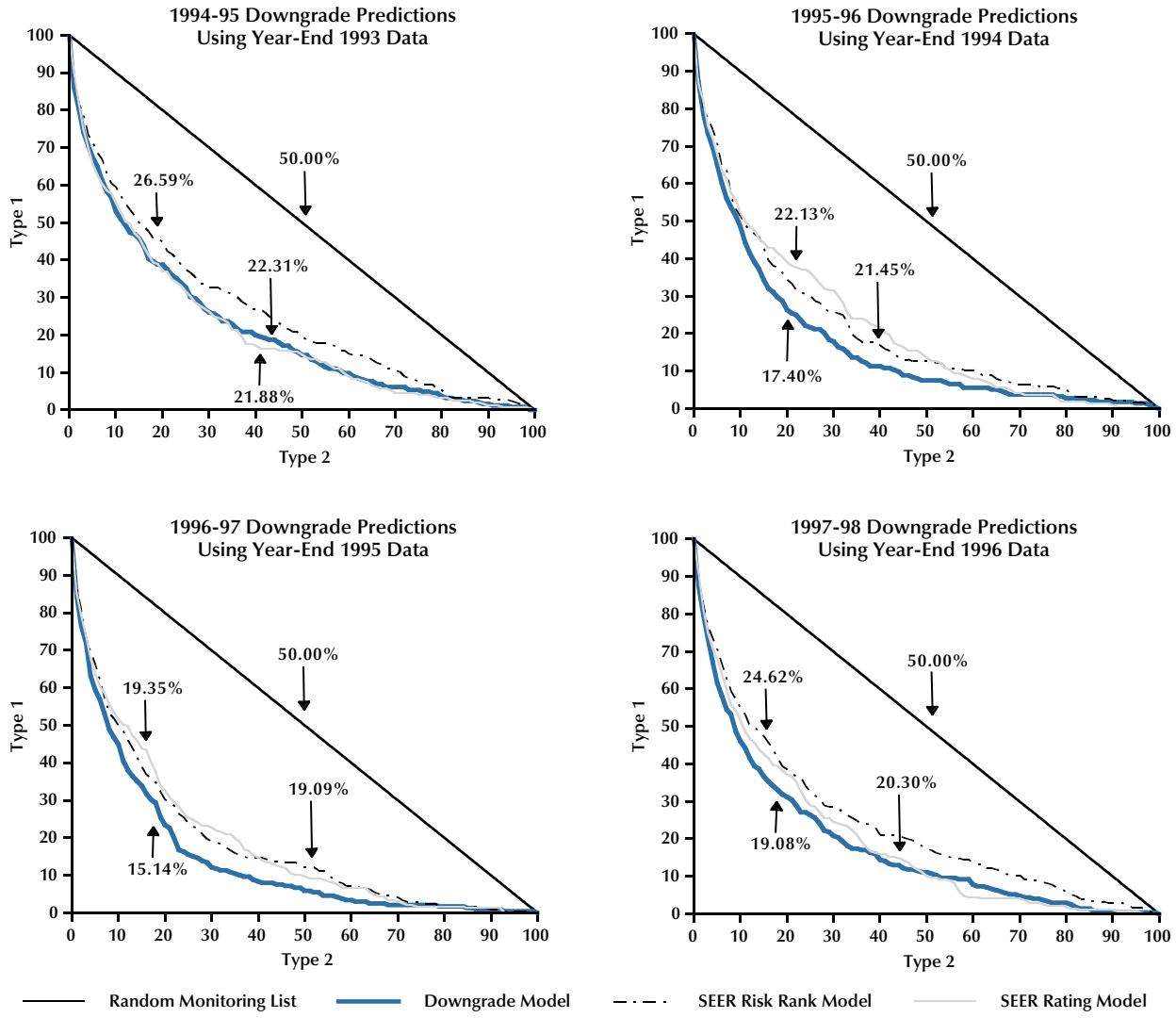
Analyzing a region with low type-two error rates confirms that the out-of-sample performances of the downgrade-prediction model and those of the SEER models are comparable. If monitoring healthy banks were costless, then supervisors would want a watch list long enough to catch all downgrade risks—a list that produced a zero type-one error rate. But because monitoring healthy banks is costly, supervisors would prefer a watch list that is reasonably sized. Panel B of Table 5 contains the areas under the power curves for all three models when the maximum allowable type-two error rate is 20 percent. Over this restricted region, the difference in the performance of the downgrade-prediction model and the two SEER models—again, expressed in terms of average areas under power curves over multiple tests—is less than 1 percentage point. Although the 20 percent threshold is arbitrary, it conveys a larger message—that the small difference between the performance of the downgrade model and the SEER models becomes even smaller when the comparison focuses on regions where supervisors are likely to operate.

ROBUSTNESS CHECKS

To check the robustness of our findings, we experimented with a "fresh" set of explanatory variables for each of the six downgrade-prediction

Figure 2

How Well Do the Models Predict Out-of-Sample CAMELS Downgrades?



NOTE: This Figure shows that the downgrade model, the SEER risk rank model, and the SEER rating model produce similar type-one vs. type-two tradeoffs for most of the range of errors for 1994-95 downgrades. The downgrade and SEER rating models do slightly outperform the SEER risk rank model, largely because the coefficients of the risk rank model are fixed. The downgrade model slightly outperforms both the SEER risk rank model and the SEER rating model as a tool for flagging downgrade candidates in 1995-96, 1996-97, and 1997-98.

This Figure depicts the trade-off between the type-one error rate and the type-two error rate for the SEER risk rank model, SEER rating model, and the downgrade model. The type-one error rate is the number of missed downgrades (false negatives) divided by the total number of CAMELS one- and two-rated banks; the type-two error rate is the number of incorrectly flagged downgrades (false positives) divided by the total number of CAMELS one- and two-rated banks. The area under each curve, divided by the total area in the box, offers a convenient way to compare the performance of each model. Smaller areas imply lower levels of *both* types of errors and, hence, better model performance. The 50 percent line indicates the type-one and type-two error rates associated with random selection of one- and two-rated banks.

regressions. If the factors driving downgrades change through time, then the out-of-sample performance of a model with a fixed set of explanatory variables should decay over a sequence of tests, even if new coefficients for the fixed set of variables are obtained each year. To combat this bias, we compiled an expanded list of candidate variables based on a review of the early warning literature.[5] Next, we identified the best subset of explanatory variables in each year based on in-sample fit of the model.[6] Specifically, our variable selection technique resembled stepwise and backward-elimination variable selection but improved upon these methods by considering all possible combinations, rather than adding or subtracting explanatory variables sequentially. Because our technique is most effective when the explanatory variables are not highly correlated, we started by grouping candidates into clusters based on correlation.[7] For example, we grouped all the nonperforming asset ratios in one problem-loan cluster. Then, for each year, we identified the variable in each cluster that was least correlated with the variables in the other clusters. Finally, we added this variable to the final set of explanatory variables for that year's downgrade-prediction model.

As an additional robustness check, we shortened the forecast horizon for all three models. In our previous analysis, we compared each model's ability to forecast downgrades two years into the future. Because the SEER rating model regresses this quarter's ratings on last quarter's financial data (i.e., uses a one-quarter lag), out-of-sample performance comparisons over a two-year horizon may be biased against the shadow CAMELS. To correct for this potential bias, we regressed downgrade status in the first quarter of year t on financial data from the fourth quarter of year $t-1$. Then, we applied the coefficients from that model to financial data from the fourth quarter of year $t+1$ to generate downgrade probabilities. Next, we traced out power curves for the downgrade-prediction model—and for the two SEER models—using the first quarter of year $t+2$ as a test window. Finally, we compared the areas under each model's power curve four times with all three models (first quarter 1994 through first quarter 1997) and six times for the SEER risk rank model and the downgrade-prediction model (first quarter of 1992 through the first quarter of 1997).

Both robustness checks confirmed our principal empirical result—that the downgrade-prediction model does not improve significantly over the SEER models as a tool for flagging downgrade candidates. In the first robustness check, we found that respecifying the CAMELS downgrade model annually did not improve its out-of-sample accuracy. Indeed, the resulting power curves were nearly identical to those obtained with the original downgrade-prediction model. In the second robustness check, we found that shortening the forecast horizon did improve the out-of-sample performance of all three models, presumably because predicting near events is easier than predicting more distant ones. For example, the average area under the downgrade-prediction power curve improved 4.32 percentage points (six tests), the average area under the SEER risk rank power curve improved 3.22 percentage points (six tests), and the average area under the SEER rating model power curve improved by 4.55 percentage points (four tests). Still, average areas produced by each model were fairly close: 15.28 percent for the downgrade-prediction model, 20.01 percent for the SEER risk rank model, and 16.37 percent for the SEER rating model. When viewed against the random selection benchmark, these performance differences seem economically insignificant.

CONCLUSION

The Federal Reserve's off-site surveillance system includes two econometric models that are collectively known as the System for Estimating Examination Ratings (SEER). One model, the SEER risk rank model, uses the latest financial statements to estimate the probability that each Fed-supervised bank will fail within the next two years. The other model, the SEER rating model, uses the latest financial statements to produce a "shadow" CAMELS rating for each supervised bank. Banks identified as risky by either model receive closer supervisory scrutiny than other Fed-supervised banks.

Because many of the banks flagged by the SEER models have already tumbled into poor condition and, hence, receive considerable supervisory attention, we developed an alternative model to identify safe-and-sound banks headed for financial distress. Such a model could help supervisors allocate scarce

[5] In addition to the papers we already cited, we drew on Cole and Gunther (1995), Hooks (1995), Wheelock and Wilson (2000), and Estrella, Park, and Peristiani (2000).

[6] See Lawless and Singhal (1978) for details.

[7] See Jackson (1991).

on- and off-site resources by pointing to banks not currently under scrutiny that need watching. Specifically, we estimated a model to flag banks with composite CAMELS ratings of one and two that are likely to receive downgrades to composite ratings of three, four, or five in the next two years. We then compared the out-of-sample performance of the model and the SEER models as tools for identifying downgrade candidates.

Over a range of two-year test windows in the 1990s, we found that the CAMELS downgrade model outperformed the SEER models by only a small margin. Our evidence suggests that, during relatively tranquil banking times such as the 1990s, a downgrade-prediction model contributes little to the Federal Reserve's off-site surveillance framework. Our evidence also indirectly validates the performance of the current SEER framework as a tool for supporting on-site examinations by the Federal Reserve.

Our evidence does not imply, however, that downgrade-prediction models have no role to play in off-site surveillance. Our sample period is marked by the longest economic expansion in U.S. history. During this period, the U.S. banking industry enjoyed robust profitability and healthy asset quality. Indeed, downgrades to unsatisfactory status as well as outright failures dropped off considerably in the 1990s relative to the 1980s. A possible interpretation of our findings is that one early warning model is as good as another when financial distress in the banking industry is relatively rare. The downgrade-prediction model could materially outperform the SEER models in a different economic climate—for example, the early stages of a contraction in which downgrades are frequent but failures still relatively rare. Only a series of out-of-sample tests that span the business cycle can conclusively determine the value added by a CAMELS downgrade model.

REFERENCES

Allen, Linda and Saunders, Anthony. "Bank Window Dressing: Theory and Evidence." *Journal of Banking and Finance*, June 1992, *16*(3), pp. 585-623.

Berger, Allen N. "The Relationship Between Capital and Earnings in Banking." *Journal of Money, Credit, and Banking*, May 1995, *27*(2), pp. 432-56.

_____ and Davies, Sally M. "The Information Content of Bank Examinations." *Journal of Financial Services Research*, October 1998, *14*(2), pp. 117-44.

Board of Governors of the Federal Reserve System. "Risk-Focused Safety and Soundness Examinations and Inspections." SR 96-14, 24 May 1996.

Cole, Rebel A.; Cornyn, Barbara G. and Gunther, Jeffrey W. "FIMS: A New Monitoring System for Banking Institutions." *Federal Reserve Bulletin*, January 1995, *81*(1), pp. 1-15.

_____ and Gunther, Jeffrey W. "Separating the Likelihood and Timing of Bank Failure." *Journal of Banking and Finance*, September 1995, *19*(6), pp. 1073-89.

_____ and _____. "Predicting Bank Failures: A Comparison of On- and Off-Site Monitoring Systems." *Journal of Financial Services Research*, April 1998, *13*(2), pp. 103-17.

Curry, Timothy. "Bank Examination and Enforcement," in *History of the Eighties: Lessons for the Future*, Vol. 1. Washington, DC: Federal Deposit Insurance Corporation, 1997, pp. 421-75.

Demsetz, Rebecca S. and Strahan, Philip E. "Diversification, Size, and Risk at Bank Holding Companies." *Journal of Money, Credit, and Banking*, August 1997, *29*(3), pp. 300-13.

DeYoung, Robert. "Birth, Growth, and Life or Death of Newly Chartered Banks." Federal Reserve Bank of Chicago *Economic Perspectives*, Third Quarter 1999, *23*(3), pp. 18-35.

Estrella, Arturo; Park, Sangkyun and Peristiani, Stavros. "Capital Ratios as Predictors of Bank Failure." Federal Reserve Bank of New York *Economic Policy Review*, July 2000, *6*(2), pp. 33-52.

Flannery, Mark J. and Houston, Joel F. "The Value of a Government Monitor for U.S. Banking Firms." *Journal of Money, Credit, and Banking*, February 1999, *31*(1), pp. 14-34.

Gilbert, R. Alton; Meyer, Andrew P. and Vaughan, Mark D. "The Role of Supervisory Screens and Econometric Models in Off-Site Surveillance." Federal Reserve Bank of St. Louis *Review*, November/December 1999, *81*(6), pp. 31-56.

Gunther, Jeffrey W. and Moore, Robert R. "Early Warning Models in Real Time." Financial Industry Studies Working Paper No. 1-00, Federal Reserve Bank of Dallas, October 2000.

Hanc, George. "The Banking Crises of the 1980s and Early

1990s: Summary and Implications," in *History of the Eighties: Lessons for the Future*, Vol. 1. Washington, DC: Federal Deposit Insurance Corporation, 1997, pp. 3-85.

Hooks, Linda M. "Bank Asset Risk: Evidence from Early-Warning Models." *Contemporary Economic Policy*, October 1995, *13*(4), pp. 36-50.

Jackson, J. Edward. *A Users Guide to Principal Components*. New York: Wiley, 1991.

Lawless, J.F. and Singhal, K. "Efficient Screening of Nonnormal Regression Models." *Biometrics*, 1978, *34*, pp. 318-27.

Morgan, Donald P. and Stiroh, Kevin J. "Market Discipline of Banks: The Asset Test." Working Paper, Research Department, Federal Reserve Bank of New York, 2001.

Putnam, Barron H. "Early Warning Systems and Financial Analysis in Bank Monitoring: Concepts of Financial Monitoring." Federal Reserve Bank of Atlanta *Economic Review*, November 1983, pp. 6-13.

Reidhill, Jack and O'Keefe, John. "Off-Site Surveillance Systems," in *History of the Eighties: Lessons for the Future*, Vol. 1. Washington, DC: Federal Deposit Insurance Corporation, 1997, pp. 477-520.

Wheelock, David C. and Wilson, Paul W. "Why Do Banks Disappear? The Determinants of U.S. Bank Failures and Acquisitions." *Review of Economics and Statistics*, February 2000, *82*(1), pp. 127-38.

COMING IN THE NEXT ISSUE OF *REVIEW*

The High-Tech Investment Boom and Economic Growth in the 1990s: Accounting for Quality

Michael R. Pakko

Why Are Stock Market Returns Correlated with Future Economic Activities?

Hui Guo

Why the Fed Should Ignore the Stock Market

James B. Bullard

The Monetary Policy Innovation Paradox in VARs: A "Discrete" Explanation

Michael J. Dueker

All nonproprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in *Review* are available to our readers on our Web site: <**www.stls.frb.org/publications/review**>. Also, you may request data and programs on either disk or hard copy: **Research Department**, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, Missouri 63166-0442. Please include the author, title, issue date, and page numbers with your request.

These data and programs are also available through **Inter-university Consortium for Political and Social Research (ICPSR)**. Member institutions may request data through the CDNet Order facility. Nonmembers may write to ICPSR, Institute for Social Research, P.O. Box 1248, Ann Arbor, Michigan 48106-1248; call 734-998-9900; or e-mail <**netmail@icpsr.umich.edu**>.

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: <**www.stls.frb.org/fred**>.

Federal Reserve Bank of St. Louis
Post Office Box 442
St. Louis, Missouri 63166-0442