Federal Reserve Bank of Chicago

Call for Papers

2003 Conference on Bank Structure and Competition

Fourth Quarter 2003

Economic perspectives

Economic perspectives

Economic Perspectives is published by the Research Department of the Federal Reserve Bank of Chicago. The views expressed are the authors' and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Single-copy subscriptions are available free of charge. Please send requests for single- and multiple-copy subscriptions, back issues, and address changes to the Public Information Center, Federal Reserve Bank of Chicago, P.O. Box 834, Chicago, Illinois 60690-0834, telephone 312-322-5111 or fax 312-322-5515.

Economic Perspectives and other Bank publications are available on the World Wide Web at http://www.chicagofed.org.

chicagofed.org

ISSN 0164-0682

Contents

Fourth Quarter 2003, Volume XXVII, Issue 4

Decimalization and market liquidity

Craig H. Furfine

On January 29, 2001, the New York Stock Exchange (NYSE) implemented decimalization. Beginning on that Monday, stocks began to be priced in dollars and cents, and price changes were allowed to be as small as 1 cent.[1] Prior to this change, NYSE stocks were quoted in fractions of a dollar and traded in increments of 1/16, or 6.25 cents. Decimalization of stock markets is relevant for policymakers because it has the potential to affect market liquidity, and therefore the overall functioning of financial markets.

Advocates of the adoption of decimalization argue that the finer gradation of stock prices will benefit investors. This is because the pricing increment dictates the smallest possible bid–ask spread for a given stock. This spread represents the difference between the lowest price an investor can pay for a stock and the highest price an investor can receive for selling the same stock. Prior to decimalization, actively traded stocks often had a spread equal to the minimum price increment, or tick, of 6.25 cents. For instance, an investor might have faced a bid–ask spread of 50 1/2–50 9/16 for shares of stock in XYZ company on Friday, January 26, 2001. That is, abstracting from any transaction fees levied by brokers, an investor wanting to sell a round lot of 100 shares of XYZ could expect to receive $5,050 and an investor needing to buy 100 shares would have to pay $5,056.25. Following decimalization, however, the two investors might face a spread of 50.51–50.53. Thus, the seller of XYZ would receive an extra penny per share and the buyer of XYZ shares would save 3.25 cents per share.

However, decimalization may affect more than a stock's bid–ask spread. To understand this, consider the set of firms and individuals that stand ready to buy and sell the stock of XYZ company. For example, prior to decimalization, a dealer might have been willing to commit to buy 10,000 shares of XYZ company at 50 1/2 and to sell 10,000 shares at 50 9/16. In practice, these commitments may have been made by the dealer placing a limit buy order for 10,000 at 50 1/2 and a limit sell order for 10,000 shares at 50 9/16. If these are the only outstanding orders for XYZ stock at these prices, then the stock will have a bid–ask spread of 1/16 and a so-called depth of 10,000 (at the bid price) by 10,000 (at the ask price) shares. With decimal pricing, the same dealer may have decided not to post limit orders of the same size at the new prices of 50.51–50.53. This is because the profitability of committing to be willing to both buy and sell a given stock, as measured by the bid–ask spread, has declined. Thus, the dealer might only be willing to offer depth of 1,000 by 1,000. From the perspective of a small investor, for example one wishing to trade only a few hundred shares, this reduction in depth at the bid–ask spread is not a concern. Depth at the best available prices will suffice. However, for large traders, for example those wishing to trade several thousand shares, quoted depth at the best-quoted prices may be insufficient to fill the desired order. For such trades, the effective transaction price lies somewhere outside the posted bid and ask.

In this article, I examine how various measures of stock market liquidity changed following decimalization. A stock's illiquidity measures the cost to a buyer or seller of transacting in shares beyond the true underlying value of the security. These costs arise from a lack of an infinite supply of shares that can be purchased and sold at the same price. That is, if investors could buy or sell any number of shares of XYZ stock at 50.52, then one would say that XYZ shares are perfectly liquid at that price. Bid–ask spreads and limited depth represent two departures from this situation and, thus, bid–ask spreads and depth are two measures of

Craig H. Furfine is an economic advisor at the Federal Reserve Bank of Chicago. *The author would like to give special thanks to Bob Chakravorti and Helen Koshy for helpful comments.*

liquidity. Lower spreads and higher depth represent more liquidity. Higher spreads and lower depth signal less liquidity.

I document that decimalization did lead to smaller spreads and lower depth, and thus caused a theoretically ambiguous change to market liquidity. Thus, to empirically address whether and/or to what extent market liquidity was affected by decimalization, one must focus on a liquidity measure that is affected by both a finer pricing grid and lower depth. In this article, I examine the revision to a stock's price that follows a trade as a direct measure of a stock's liquidity. This is known as the price impact of a trade. By definition, perfect liquidity implies that a given trade should not affect the price.[2] With imperfect liquidity, the size of the price revision following a trade is likely positively related to tick size, since any adjustment to prices must be at least as large as the minimum tick. Likewise, the price impact should be negatively related to market depth, since lower depth implies that a given (large) trade may have to travel through more prices in order to be filled.

In this study, I examine the stocks of 1,339 companies that began decimal trading on the NYSE on January 29, 2001. I document what previous studies have found regarding the relationship between tick size, bid–ask spreads, and depth. I then estimate the price impact of a trade, distinguishing trades undertaken before decimalization from those after and further distinguishing large trades from small trades. I find that for both large and small trades, decimalization typically led to an improvement in liquidity as measured by a decline in the price impact of a trade for actively traded stocks. For less actively traded stocks, decimalization led to improved liquidity more often than it led to reduced liquidity. However, the most common empirical finding for infrequently traded stocks was that there was not a statistically significant change in liquidity following decimalization.

Selective literature review

I mention only a few contributions to the extensive literature on the effects of reducing tick sizes on various measures of market performance, including liquidity. In Seppi's (1997) theoretical framework, reduction in tick sizes leads to a reduction in the willingness of both small and large traders to supply liquidity through limit orders (depth). However, because retail investors require less depth to conduct trading, optimal tick size depends positively on typical trade size. That is, institutions that typically trade in large amounts prefer large tick sizes, whereas small investors prefer small tick sizes.

Harris (1994), using data from a time when the minimum tick was 1/8, fits a regression model estimating the frequency at which spreads are at the minimum. Using this relationship, Harris estimates that the impact of reducing the minimum tick size to 1/16 would be accompanied by both lower bid–ask spreads and lower quoted depth. His results are therefore also consistent with the notion that optimal tick size is related to the size of a trade. They indicate that small traders would almost certainly benefit from smaller tick sizes, but that large traders might be hurt if the depth of the market were to fall sufficiently.

Goldstein and Kavajecz (2000) analyze the NYSE's reduction in tick size from 1/8 to 1/16 and address the relationship between minimum tick size, bid–ask spreads, and market liquidity. What is unique about this study is that these authors not only look at the depth reported at the best bid and ask prices, they also collect data on liquidity available at some distance away from the best bid and ask prices. This complete collection of prices and available depth is called the limit order book. They find that not only did depth at the best bid and ask decline, but cumulative depth similarly declined throughout the limit order book following the NYSE's previous reduction in minimum tick size. They found such declines in depth as far as 50 cents from the midpoint of the bid–ask spread. Using implied average price of a trade of a given size derived from the limit order book, these authors find that large traders were not made better off by the smaller tick sizes and were made worse off for infrequently traded stocks.

More recent work has examined changes in market liquidity for NYSE-listed stocks since decimalization. Chakravarty et al. (2001) study a small set of stocks that began trading in decimals as part of an NYSE pilot program in 2000. These authors find that decimalization has led to significantly lower spreads and also lower quoted depth. Bacidore et al. (2001) also study stocks in the decimalization pilot. These authors focus on whether decimalization leads to significant changes in order submission strategies. They find that there is no noticeable change in the use of limit versus market orders, but the size of limit orders has fallen and the frequency of limit order cancellation has increased since decimalization. Smaller limit orders and higher cancellation rates explain how lower depth materialized.

Bessembinder (2003) studies a larger sample of NYSE and Nasdaq (National Association of Securities Dealers Automated Quotation) stocks and documents that following decimalization, bid–ask spreads fell noticeably, with the largest declines seen for the most actively traded stocks. Bessembinder also reports an increase in the frequency of price improvement on the NYSE following decimalization. Price improvement

occurs when a trade is conducted at a price inside the bid–ask spread. Higher rates of price improvement are consistent with the fact that decimalization makes it easier for traders to step in front of the current best bid or ask to take the other side of a market order.

Data, sample selection, and summary statistics

For this study, I extracted stock market trade and quote data from the NYSE TAQ (Trades and Quotes) database, covering the 24 trading days beginning Tuesday, January 16, 2001, and ending on February 16, 2001.[3] Following Hasbrouck (1991), I also impose a minimum price requirement on each company's stock. I require each stock to be trading for at least $5, on average, during the five-week sample period. I similarly impose a maximum price of $200. Also following Hasbrouck (1991), I require a minimum level of trading activity. I limit my sample to stocks that traded, on average, at least every ten minutes over these five weeks. Finally, I eliminate those stocks that were part of the NYSE pilot program and, therefore, traded in decimals prior to January 29, 2001. My final sample consists of 1,339 stocks.

The data are then adjusted according to procedures common in the microstructure literature. Following Hasbrouck (1991), I keep only New York quotes and consider multiple trades on a regional exchange for the same stock at the same price and time to be one trade. Then, I sort the trade data (for each company and day) by time, with the prevailing quote at transaction t defined to be the last quote that was posted at least five seconds before the transaction (Lee and Ready, 1991).

I group the 1,339 stocks into six categories according to their average trade frequency, with Category 1 stocks being the least traded and Category 6 stocks being the most frequently traded. The number of stocks in each category is shown in table 1. I first calculate the narrowest bid–ask spread witnessed by each stock and for each day. I then average these minimums across stocks within the same trading category. These average minimum values are plotted in figure 1. As is apparent from figure 1, minimum tick size strongly influences the extent to which bid–ask spreads can narrow within a day. Every stock in the four most actively traded categories experienced a bid–ask spread equal to the minimum tick size at least once on every day of the sample period. For stocks in Category 2, that is, those trading every one to five minutes, minimum spreads were always within 1/10 of a cent of the minimum tick size, suggesting that nearly all of the 722 stocks in that category experience minimum tick-sized spreads during each day. Even for those stocks trading only every five to ten minutes, average minimum spreads hover around 6.5 cents when the minimum tick size was 6.25 and fall to around 2 cents after decimalization.

TABLE 1

Average trading activity of sample stocks

Category	Average time between trades	Number of stocks
1	5–10 minutes	191
2	1 minute–4 minutes, 59 seconds	722
3	30–59 seconds	237
4	15–29 seconds	108
5	5–14 seconds	70
6	Less than 5 seconds	11

Source: New York Stock Exchange, 2001, Trades and Quotes (TAQ) database, January 16–February 16.

Figure 2 plots the mean spread within a day averaged across stocks in a given trading category. As expected, more frequently traded stocks have lower spreads. For stocks in all categories, decimalization has led to a decline in average bid–ask spreads. This decline in mean spreads is most pronounced for the more actively traded stocks. For example, mean bid–ask spreads for stocks in Category 1 averaged 13.6 cents, about a penny over two ticks, in the two weeks prior to decimalization. Following decimalization, the average mean spread of these stocks fell to 10.5 cents. For the most actively traded stocks, average mean spreads were 11.3 cents before decimalization and fell to 7.0 cents after.

The narrowing of bid–ask spreads following decimalization illustrated in figures 1 and 2 has been accompanied by a decline in average depth at the posted prices. Figure 3 reports the mean of the bid and ask depth posted throughout a day, averaged across stocks in each category. The post-decimalization decline in posted depth was significant, especially for the most actively traded stocks that had previously had a very large number of shares committed to trade at the posted spread. For the least actively traded stocks, posted depth fell by half, from just under 6,600 shares to 3,300 shares on average. The most actively traded stocks experienced depth declines of more than two-thirds, from an average of 18,700 shares to slightly more than 6,000 shares.

Price impact of trading

The statistics reported in the previous section confirm that decimalization has led to both a decline

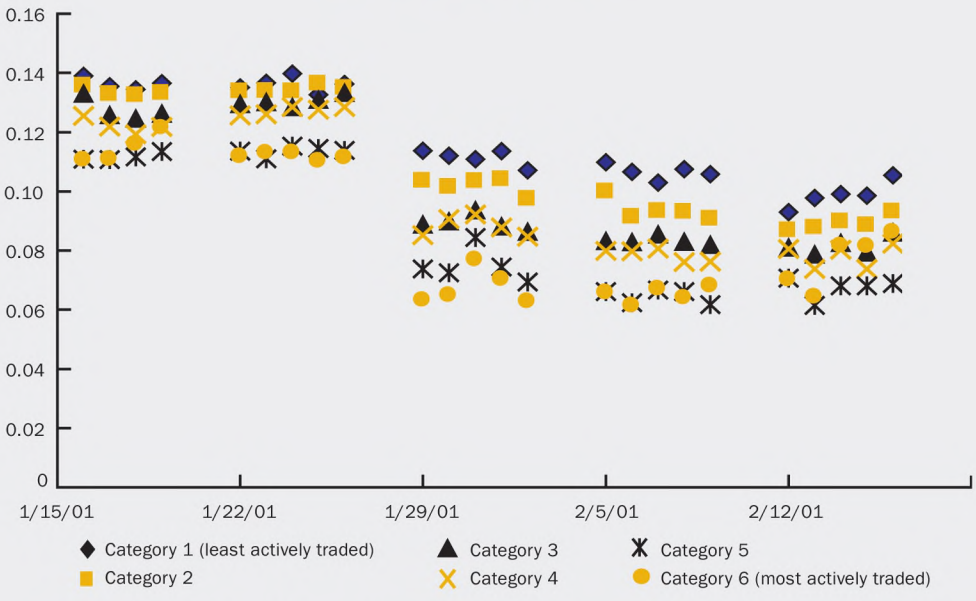
FIGURE 1

Average minimum daily spread

Source: NYSE TAQ database from January 16–February 16, 2001.

FIGURE 2

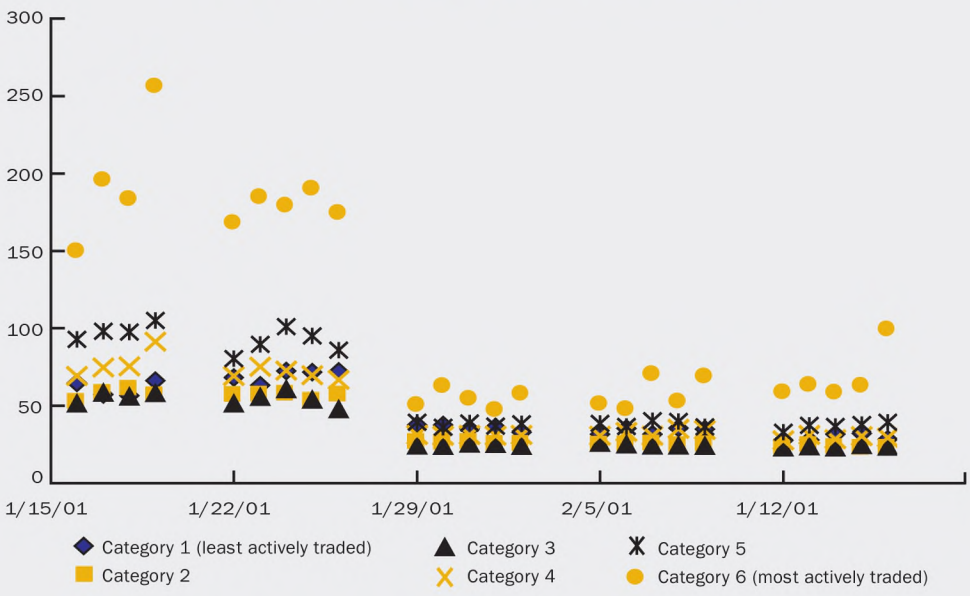
Average mean daily spread



Source: NYSE TAQ database from January 16–February 16, 2001.

FIGURE 3

Average daily quoted depth (in number of round lots)



Source: NYSE TAQ database from January 16–February 16, 2001.

in spreads and a decline in depth. Thus, an investor making a relatively small trade generally faces improved liquidity, since the trade can be executed at a narrower spread. For institutional investors making large trades, however, the lower depth may imply that a large trade must travel through several prices before being fulfilled, and thus decimalization may not necessarily have led to better execution prices. To try to combine the effects of spread and depth in one framework, I examine a liquidity measure called the price impact of a trade, which was first motivated and estimated by Hasbrouck (1991). This is a measure of how much the price of a stock changes following a given trade as estimated in an autoregression framework. While other measures of liquidity are conceivable, price impact has the advantage of being influenced by both smaller price increments and lower depth. In particular, decimalization may cause the price impact to decline since prices can adjust by smaller increments, but it may also cause the price impact to rise since more prices must be exhausted because depth at any given price is lower.

In the price impact framework, the dependent variable of interest is the trade-to-trade return on a given stock. I denote this return r_t and define it formally as the change in the natural logarithm of the midquote of a given stock that follows the trade at time t.[4] That is,

$$1) \quad r_t = 100\left(\ln\left(\frac{bid_{t+1} + ask_{t+1}}{2}\right) - \ln\left(\frac{bid_t + ask_t}{2}\right)\right).$$

The use of midquotes eliminates price changes caused by the bid–ask bounce, that is, the alternating arrival of buys and sells transacting at the ask and bid price, respectively. Following Hasbrouck (1991), I define the variable x_t as an indicator of the direction of the trade occurring at time t. If the trade is initiated by the buyer, the variable $x_t = 1$. If the trade is initiated by the seller, then the variable $x_t = -1$. I assume trades at a transaction price greater than the midquote were buyer-initiated and trades below the midquote were seller-initiated. For trades at the midquote, x_t is assigned to equal zero. Defining D_t as an indicator that equals 1 if trade t occurs during the first 30 minutes of the trading day and dec_t as an indicator that equals 1 if trade t occurs during the decimal period after January 29, 2001,[5] I estimate the regression

$$2) \quad r_t = \alpha D_t x_t + \sum_{i=1}^{5}[\beta_i + \delta_i dec_{t-i}]r_{t-i} + \sum_{i=0}^{5}[\gamma_i + \lambda_i dec_{t-i}]x_{t-i} + \varepsilon_t$$

for each stock in my sample.[6] The price impact of a trade in this framework is equal to the sum of the γ_i coefficients during the pre-decimalization period and equal to the sum of the $\gamma_i + \lambda_i$ coefficients after decimalization. Because purchases should put upward pressure on prices, I expect that γ_i should be positive for some or all of the trade lags i. This prediction follows from traditional microstructure theory. In Glosten and Milgrom (1985), for example, market makers set a positive bid–ask spread as compensation for trades made with counterparties with superior information. As a sequence of sell orders arrives, market makers lower bid prices, incorporating the probability that the order flow implies that better-informed investors believe the previous price was too high. The reverse occurs when a sequence of buy orders arrives. This type of dynamic quote adjustment leads to the prediction of a positive value of the γ_i coefficients. The λ_i coefficients may be either positive or negative depending on whether the stock has become less or more liquid (higher or lower price impact of a trade) since decimalization.

Table 2 summarizes the results from these 1,339 regressions. Each row of table 2 corresponds to stocks in different trade activity categories. The numbers reported in the first column represent the average value of the sum of the γ_i coefficients across all stocks in the given category. These coefficients measure the price impact of a trade in the two weeks prior to decimalization. For instance, the first entry in the first column reports that the average price impact of a trade during this time is 9.2 basis points for stocks in the least actively traded category. Put another way, 11 trades in the same direction move an infrequently traded stock's price by 1 percent. As the numbers in the first column indicate, liquidity as measured by price impact increases with trading activity, since trades of more actively traded stocks move prices less. For example, a trade of a stock of a very actively traded security moves the stock's price by just over 1 basis point. The numbers in the second and third columns summarize the statistical significance of the price impact results. The second and third columns report the fraction of stocks in the given category whose price impact estimate was positive and statistically significant (at the 5 percent level) and negative and statistically significant, respectively. As the numbers in columns two and three indicate, none of the stocks in the sample had a negative and significant price impact of a trade, whereas virtually all of the sample stocks had significantly positive price impacts.

The last three columns summarize the results for the decimalization period. A comparison of the fourth column with the first indicates that for stocks in all categories, the average price impact of a trade declined following decimalization. That is, stocks became more liquid on average. However, the numbers reported in the final two columns suggest that this result is not as uniform as the positive price impact result reported for the pre-decimalization period. For stocks in the least actively traded category, only one-quarter saw a statistically significant decrease in price impact following decimalization. Around two-thirds of the stocks in category two had statistically significant increases in liquidity. For the more actively traded stocks, that is, those that on average trade at least once per minute, more than 95 percent witnessed an increase in liquidity (decrease in price impact) following decimalization. The magnitude of the increased liquidity is fairly large, with the typical decline in price impact following decimalization being close to 40 percent.

TABLE 2

Average price impact of a trade: Before and after decimalization

	Before decimalization			After decimalization		
Trade category	Average price impact (sum of γ_i)	Share of stocks with positive and sig. γ_i	Share of stocks with negative and sig. γ_i	Average price impact (sum of $\gamma_i + \lambda_i$)	Share of stocks with positive and sig. λ_i	Share of stocks with negative and sig. λ_i
1	0.092	0.974	0.000	0.074	0.016	0.251
2	0.070	0.999	0.000	0.045	0.007	0.652
3	0.043	1.000	0.000	0.025	0.008	0.945
4	0.032	1.000	0.000	0.019	0.000	0.954
5	0.023	1.000	0.000	0.014	0.014	0.957
6	0.011	1.000	0.000	0.006	0.000	1.000

Based on the regression equation $r_t = \alpha D_t x_t + \sum_{i=1}^{5}[\beta_i + \delta_i dec_{t-i}]r_{t-i} + \sum_{i=0}^{5}[\gamma_i + \lambda_i dec_{t-i}]x_{t-i} + \varepsilon_t$.
Note: Sig. indicates significant.

The preceding results suggest that decimalization led to increased liquidity for virtually all stocks that trade at least once per minute. Among less frequently traded stocks, the results were more mixed, with a significant fraction of stocks experiencing no statistically significant change in price impact following the move to decimal pricing. My previous discussion of microstructure theory, however, suggests that decreases in tick size and depth may have different implications for trades of different sizes. In particular, small trades may benefit from tighter spreads and correspondingly smaller price increments because they can normally be executed within posted depth. Large trades, however, may have become less liquid following decimalization due to the significant decline in depth. I extend my empirical framework to test for this possibility. I define the variable Big_t as an indicator that equals 1 when trade t is among the largest 10 percent of trades for the given stock and then estimate an extended regression model described by equation 3.

$$3)\quad r_t = \alpha D_t x_t + \sum_{i=1}^{5}[\beta_i + \delta_i dec_{t-i}]r_{t-i} + \sum_{i=0}^{5}[\gamma_i + \lambda_i dec_{t-i} + \mu_i Big_{t-i} + \theta_i dec_{t-i} Big_{t-i}]x_{t-i} + \varepsilon_t.$$

In equation 3, I have extended the previous regression equation by allowing the price impact of a trade, both before and after decimalization, to differ depending on whether the trade is large. Since large trades have been generally found to be less liquid (see Hasbrouck, 1991), one might expect the μ_i coefficients to be positive. That is, larger trades will move prices more than smaller trades. The ϕ_i coefficients allow a direct test of the hypothesis that the liquidity of large trades has been adversely affected by the move to decimal pricing.

Table 3 summarizes the results from estimating equation 3 for all 1,339 stocks in the sample. Table 3 follows a similar format as table 2, only expanded to account for the fact that I now distinguish between large trades (those in the top 10 percent of size within the given stock) and regular trades (all others). The first three columns of panel A in table 3 report statistics for regular trades prior to decimalization. These results are qualitatively similar to those reported in the first three columns in table 2. In particular, I find positive and statistically significant price impacts for virtually all stocks in the sample for regular sized trades. The size of the price impact is lower than that reported for all trades in table 2. This can be explained by the larger impacts found for large trades reported in columns four to six of table 3. For example, a regular trade of a stock in category 6 (most actively traded) moved the stock's price by roughly 0.9 basis points. Large trades of such a stock moved the price by 3.0 basis points. Since I define large trades to be the top 10 percent of the size distribution, one might expect the average price impact of an actively traded stock to be $90\% \times 0.9 + 10\% \times 3.0 = 1.11$ basis points, which is the result reported in table 2.

The results in columns four to six of table 3 also indicate that large trades are not necessarily less liquid for relatively infrequently traded stocks. In particular, I find that larger trades have a larger price impact for only 29 percent of stocks trading every five to ten minutes and for only 50 percent of stocks trading every one to five minutes. Large trades of stocks trading at least every 30 seconds, however, do typically move prices more than other trades. For these more actively traded stocks, large trades move prices between two and three times more than regular trades.

Panel B in table 3 summarizes the regression results related to the period following decimalization. Average price impacts of regular trades fell following decimalization for virtually all stocks trading at least once each minute. For example, the price impact of a regular trade of a very actively traded stock (category 6) fell from 0.9 basis points to 0.5 basis points after decimalization. For stocks trading every 30–60 seconds (category 3), the average price impact of regular trades fell from 3.9 basis points to 2.3 basis points. For stocks trading only every five to ten minutes, decimalization did not generally lead to lower price impacts of regular trades. I find a statistically significant decline in price impact (increase in liquidity) for only 17.8 percent of such stocks.

The final three columns of panel B report my findings regarding the liquidity of large trades after decimalization relative to before. On average, large trades are more liquid in the post-decimalization period. For example, a large trade of a stock in the most actively traded category moved the stock's price by an average of 3.0 basis points before decimalization, but by only 1.9 basis points after decimalization. For all 11 of these stocks, the decline in price impact (increase in liquidity) of large trades was statistically significant. Across stocks in the entire sample, the decline in price impact for large trades was less common than that found for regular trades. For example, I find that the price impact of a regular trade declined post-decimalization for 92.0 percent of stocks that were traded every 30–60 seconds (Category 3). However, the price impact of a large trade fell in only 37.1 percent of these stocks

TABLE 3

Average price impact of a trade: Before and after decimalization, by trade size

A. Before decimalization

	Regular trades			Trades with size in top decile		
Trade category	Average price impact (sum of γ_i)	Share of stocks with positive and sig. γ_i	Share of stocks with negative and sig. γ_i	Average price impact (sum of $\gamma_i + \mu_i$)	Share of stocks with positive and sig. μ_i	Share of stocks with negative and sig. μ_i
1	0.080	0.942	0.005	0.141	0.288	0.000
2	0.062	0.989	0.000	0.115	0.499	0.001
3	0.039	1.000	0.000	0.077	0.789	0.000
4	0.029	1.000	0.000	0.062	0.944	0.000
5	0.019	1.000	0.000	0.049	1.000	0.000
6	0.009	1.000	0.000	0.030	1.000	0.000

B. After decimalization

	Regular trades			Trades with size in top decile		
Trade category	Average price impact (sum of $\gamma_i + \lambda_i$)	Share of stocks with positive and sig. λ_i	Share of stocks with negative and sig. λ_i	Average price impact (sum of $\gamma_i + \lambda_i + \mu_i + \theta_i$)	Share of stocks with positive and sig. θ_i	Share of stocks with negative and sig. θ_i
1	0.067	0.016	0.178	0.126	0.026	0.084
2	0.042	0.008	0.569	0.101	0.008	0.140
3	0.023	0.008	0.920	0.061	0.004	0.371
4	0.018	0.000	0.926	0.044	0.000	0.657
5	0.013	0.014	0.914	0.033	0.000	0.886
6	0.005	0.000	1.000	0.019	0.000	1.000

Based on the regression equation $r_t = \alpha D_t x_t + \sum_{i=1}^{5}[\beta_i + \delta_i dec_{t-i}]r_{t-i} + \sum_{i=0}^{5}[\gamma_i + \lambda_i dec_{t-i} + \mu_i Big_{t-i} + \theta_i dec_{t-i} Big_{t-i}]x_{t-i} + \varepsilon_t$.

Note: Sig. indicates significant.

following decimalization. Most stocks trading less frequently did not witness an increase in liquidity for large trades following decimalization. For example, among stocks trading only every five to ten minutes, only 8.4 percent of the stocks saw a decline in price impact for large trades following decimalization. In fact, 2.6 percent of such stocks actually saw an increase in price impact. Higher price impacts following decimalization, however, were virtually nonexistent for stocks trading at least once every five minutes.

As a robustness check, I reestimate equation 3 after changing the definition of a large trade. The previous definition of large was an absolute one, namely any trade that was among the largest 10 percent of all trades of a given stock during the sample period. To consider more explicitly that the liquidity of a trade is not only related to size, but also to depth, I repeat the analysis defining the variable *Big* to be equal to 1 when the given trade is executed for more shares than posted depth. That is, a buy transaction for more shares than posted ask depth or a sell transaction for more shares than posted bid depth would be defined as a large trade. Whereas 10 percent of trades were previously defined as large, this alternative definition identifies roughly 19 percent of all trades. However, only approximately 60 percent of trades previously identified as large satisfy this new definition. This implies that much of the time, large trades occur when depth is also high.

Table 4 reports results analogous to those in table 3 for this alternative definition of a large trade. The most substantial difference from the earlier results is that trades that satisfy the new definition of big seem more highly correlated with lower levels of liquidity. Whereas less than half of stocks trading less than once per minute (Categories 1 and 2) had higher price impacts for trades in the top 10 percent of the size distribution (table 3), well over half of these stocks experience a greater price impact for trades greater than posted depth (table 4). The magnitude of the price impact for these large trades is now larger under this definition. Whereas the largest 10 percent of trades of the most frequently traded stocks moved price by 3.0 basis points, trades greater than posted depth moved the price of these same stocks by 3.9 basis points before decimalization.

With this new definition of large trades, decimalization is still generally correlated with increased

TABLE 4

Average price impact of a trade: Before and after decimalization, by trade size (alternative definition of a large trade)

A. Before decimalization

	Regular trades			Trades with size > posted depth		
Trade category	Average price impact (sum of γ_i)	Share of stocks with positive and sig. γ_i	Share of stocks with negative and sig. γ_i	Average price impact (sum of $\gamma_i + \mu_i$)	Share of stocks with positive and sig. μ_i	Share of stocks with negative and sig. μ_i
1	0.074	0.942	0.000	0.188	0.476	0.000
2	0.058	0.988	0.000	0.143	0.731	0.003
3	0.035	1.000	0.000	0.091	0.987	0.000
4	0.026	1.000	0.000	0.070	1.000	0.000
5	0.017	1.000	0.000	0.057	1.000	0.000
6	0.008	1.000	0.000	0.039	1.000	0.000

B. After decimalization

	Regular trades			Trades with size > posted depth		
Trade category	Average price impact (sum of $\gamma_i + \lambda_i$)	Share of stocks with positive and sig. λ_i	Share of stocks with negative and sig. λ_i	Average price impact (sum of $\gamma_i + \lambda_i + \mu_i + \theta_i$)	Share of stocks with positive and sig. θ_i	Share of stocks with negative and sig. θ_i
1	0.062	0.010	0.168	0.140	0.010	0.173
2	0.040	0.011	0.553	0.106	0.010	0.325
3	0.020	0.008	0.895	0.062	0.000	0.679
4	0.015	0.000	0.926	0.044	0.000	0.935
5	0.010	0.014	0.957	0.033	0.000	0.957
6	0.004	0.000	1.000	0.019	0.000	1.000

Based on the regression equation $r_t = \alpha D_t x_t + \sum_{i=1}^{5}[\beta_i + \delta_i dec_{t-i}]r_{t-i} + \sum_{i=0}^{5}[\gamma_i + \lambda_i dec_{t-i} + \mu_i Big_{t-i} + \theta_i dec_{t-i} Big_{t-i}]x_{t-i} + \varepsilon_t$.
Note: Sig. indicates significant.

liquidity (lower price impact) of actively traded stocks. Further, more stocks have a statistically significant decline in price impact for large trades under the new definition. With large trades defined as those with size greater than posted depth, virtually all stocks trading at least every 30 seconds and two-thirds of those trading every 30–60 seconds witnessed an increase in liquidity of large trades following decimalization. Thus, according to these two measures of large trade size, decimalization seems to have led to increased liquidity for most actively traded stocks, and in virtually no cases did decimalization lead to less liquidity, even for large trades.

Conclusion

This article has examined the impact of decimalization on the liquidity of NYSE stocks. Analyzing transaction data for a sample of 1,339 stocks listed on the NYSE over a five-week period surrounding the January 29, 2001, implementation of decimalization, I presented evidence of the following: Minimum price increments do seem to have an impact on bid–ask spreads. In particular, for nearly all but the least-traded stocks, bid–ask spreads were equal to the minimum price increment at least once each day during the sample period. Decimalization also led to a narrowing of average bid–ask spreads. The largest declines in spreads were found for the most actively traded stocks, where the average decline in spreads was over 35 percent. I also documented that the observed compression of bid–ask spreads was accompanied by a decline in posted depth. The decline in depth was also most pronounced for the most actively traded stocks.

Because these findings suggest that decimalization had an ambiguous impact on market liquidity using spreads and depth as proxies for liquidity, I then estimated a different measure of liquidity that would be affected by changes in both spreads and depth. Specifically, I estimate the price impact of a trade for each stock in my sample. The price impact measures the percentage change in a stock's price that follows a trade. Larger price impacts, therefore, reflect lower liquidity. Intuitively, one would expect lower price increments to imply lower price impacts but lower depth

to imply higher price impacts. Thus, calculating price impacts before and after decimalization is one measure of stock market liquidity that encompasses changes in both spreads and depth.

Estimating price impact regressions, I find that actively traded stocks generally experienced an increase in liquidity (decrease in price impact) following decimalization. For less frequently traded stocks, the results were mixed. In particular, most stocks in the sample traded once every one to five minutes, and of these, only two in three experienced a statistically significant decline in average price impact following decimalization.

I then expanded my empirical framework to consider explicitly the fact that declines in posted depth may be more important for large trades, defined as those among the largest 10 percent of trades for a given stock. These large trades may be more likely to be executed at prices other than the best bid and ask price. I confirm that price impacts of these larger trades are greater, reflecting their lower liquidity. Even though these large trades have higher price impacts than other trades, I still find that decimalization generally improved the liquidity of large trades for actively traded stocks. However, I find no statistically significant improvement in liquidity of large trades for a wider set of stocks in my sample. For my set of stocks trading every one to five minutes, for example, only one in seven experienced a statistically significant increase in liquidity of large trades following decimalization. The liquidity of large trades of most stocks after decimalization was statistically indistinguishable from their liquidity before. My findings were similar when I defined large trades as those whose size was greater than posted depth.

The findings in this study suggest that decimalization on the NYSE was a positive step because policymakers prefer more liquid markets. Virtually all actively traded stocks had improved levels of liquidity following decimalization using price impact as the liquidity measure, even if this improved liquidity was not realized by those making large trades. Thus, with respect to market liquidity, the lower willingness of market participants to post limit orders seems to be more than offset by the availability of a wider array of prices at which to trade.

In the two years since decimalization, spreads and depth have continued to fall. Actively traded stocks typically have average spreads of around 3 cents, whereas relatively infrequently traded stocks have spreads of around 6 cents. Thus, market developments suggest that it may not be long before one may ask whether minimum price increments of 1 cent should be abandoned. Even before decimalization, prices were not always equal to multiples of a penny, and thus, in principle, prices could be as finely divided as desired by traders. Perhaps, therefore, the optimal tick size is less than a penny.

NOTES

[1]A few months later, the National Association of Securities Dealers did the same for stocks trading on Nasdaq.

[2]Trades can affect prices for reasons of asymmetric information. This will be discussed later in the article.

[3]Monday, January 15, 2001, was a holiday.

[4]I eliminate observations spanning more than one business day.

[5]For the regressions, I do not include the middle week of my sample (that is, the week of January 29, 2001) to control for the possibility that market participants require a period of adjustment to a new quoting system.

[6]I choose five lags following Hasbrouck (1991). The results are robust to adjustments in lag length.

REFERENCES

Bacidore, Jeffrey M., Robert H. Battalio, and Robert H. Jennings, 2001, "Order submission strategies, liquidity supply, and trading in pennies on the New York Stock Exchange," Indiana University, mimeo.

Bessembinder, Hendrik, 2003, "Trade execution costs and market quality after decimalization," *Journal of Financial and Quantitative Analysis*, forthcoming.

Chakravarty, Sugato, Stephen P. Harris, and Robert A. Wood, 2001, "Decimal trading and market impact," Purdue University, working paper.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, "Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders," *Journal of Financial Economics*, Vol. 14, pp. 71–100.

Goldstein, Michael A., and Kenneth A. Kavajecz, 2000, "Eighths, sixteenths, and market depth: Changes in tick size and liquidity provision on the NYSE," *Journal of Financial Economics*, Vol. 56, pp. 125–149.

Harris, Lawrence E., 1994, "Minimum price variations, discrete bid-ask spreads, and quotation sizes," *Review of Financial Studies*, Vol. 7, No. 1, pp. 149–178.

Hasbrouck, Joel, 1991, "Measuring the information content of stock trades," *Journal of Finance*, Vol. 46, No. 1, pp. 179–207.

Lee, Charles M. C., and Mark J. Ready, 1991, "Inferring trade direction from intraday data," *Journal of Finance*, Vol. 46, No. 2, pp. 733–746.

Seppi, Duane J., 1997, "Liquidity provision with limit orders and a strategic specialist," *Review of Financial Studies*, Vol. 10, No. 1, pp. 103–150.

Estimating U.S. metropolitan area export and import competition

William Testa, Thomas Klier, and Alexei Zelenev

This article calculates estimates of the extent to which U.S. cities' manufacturers face competition from foreign producers. Foreign and U.S production can compete in the U.S. domestic market, foreign markets, or both. Accordingly, this article examines measures of metropolitan-area (MSA) level import competition, based on each city's industrial composition and industry-level data on import competition, as well as measures of metro-level export competition, based on U.S. export data. With these measures, we evaluate whether the growth experience of U.S. cities that face high competition from foreign producers substantially differs from that of cities with low competition. Measures of import and export competition at the MSA level may be helpful to metropolitan area residents and policymakers in evaluating their own actions in various arenas such as household movements, investment, and local development.

The International Trade Administration (ITA) tracks exports from U.S. metropolitan areas. However, there are no comparable statistics of actual *imports* into particular metropolitan areas. Nor, even if they existed, would such figures be particularly useful in measuring the degree to which metropolitan area economies (and their local industries) are impacted by import competition. To the extent that the manufacturing sector of a metropolitan area sells much of its output to markets located outside of its own metropolitan area, own-industry imports at the metropolitan area level would not fully measure the degree of competition to this metropolitan area's producers. As an alternative to such a hypothetical measure of local imports, we construct measures of metropolitan areas' exposure to *national* or U.S. market import competition.

In examining trends in imports into the U.S. market over time, we find robust growth in imports of manufactured goods during the 1990s. As one measure, we allocate such imports—good by good—to each metropolitan area based on its own size and mix of manufacturing industries. In constructing these estimates, we find a wide variation across U.S. metropolitan areas in import market share and in the growth of such imports from 1989 to 1999. A rapidly growing share of imports, however, does not necessarily accompany local production decline or stagnation, because rising imports may also be associated with a rising domestic demand for these products. For example, imagine the rapidly rising U.S. imports for pharmaceuticals not necessarily *displacing* domestic production, but simply serving a growing market (perhaps fueled by an aging U.S. population).

A separate and different accounting of import behavior over time, the degree of "import penetration," measures the extent to which the domestic U.S. market for goods is served by foreign sources rather than domestic producers. Here again, we find a wide regional variation, both for the current period and across time. Such evidence of market penetration does not, of course, measure changes in the economic well-being of workers and firms. Imports of capital goods and technology also assist domestic industry to improve and stay competitive in its production and export activity. Indeed, imports are often not final goods but intermediate products used in the production of other goods, which are ultimately sold both domestically and abroad.[1] And importantly, imports of consumer goods presumably improve well-being and quality of life for U.S. individuals and households. Even on its production side, displacement of manufacturing

William A. Testa is a vice president and the director of regional programs, Thomas Klier is a senior economist, and Alexei Zelenev is an associate economist at the Federal Reserve Bank of Chicago. *The authors would like to thank Jeff Campbell and seminar participants at the Federal Reserve Bank of Chicago and express their appreciation to the late senior economist Jack L. Hervey for inspiration and friendship.*

by imports may result in reallocation of workers and capital to higher-valued production, for example, in exports, non-traded goods, or in the service sector. In these ways, enhanced imports can lift domestic production and income rather than retard them.

On the *export* side, we analyze data gathered by the International Trade Administration for large metropolitan areas. These data are reported with several user cautions, the most important of which is that the production locale of exported goods often remains unknown or is misleading, with the reported geography perhaps attributed to the place of final shipment of goods by an intermediary or perhaps to an affiliate of the manufacturer, rather than to its origin of production. To offset the possible slant of these data toward cities where exports are shipped abroad or otherwise affiliated, we construct an alternative, hypothetical measure of exports. This measure allocates U.S. exports by location of similar production activity; in particular, it allocates exports associated with an individual industry in proportion to each MSA's employment share of that same industry in the U.S. Such a measure, imperfect in its own way, is slanted toward production origin of the good rather than toward the place of shipment. Both measures indicate a wide range of openness across individual metropolitan areas. In comparing the two measures of exports, we find significant and systematic differences, suggesting that each measure may reflect a different dimension of metropolitan area exports—both point of production and point of shipment overseas. As evidence, we find that metropolitan areas with large transportation sectors tend to have higher rankings in the ITA's reported MSA export series. The presence of large manufacturing company headquarters, however, does not appear to slant reported ITA export figures in any systematic way.

Given that our import measures are constructs rather than observed data, we would like to test whether these measures lend themselves to a plausible interpretation. We examine the cross-sectional growth behavior of metropolitan areas' net job creation in manufacturing from 1989 to 1999. Using a single-equation ordinary least square (OLS) regression, we regress the growth rate of manufacturing jobs on the growth rates of exports, import market growth, and export and import penetration specific to each U.S. metropolitan area. In this exercise, we find statistically significant regression coefficients that are plausible. That is, trends toward import penetration of an area's local industries are associated with short-term manufacturing job disruption (declines) in a metropolitan area; export growth is associated with manufacturing gain. Yet this simple modeling exercise does not allow firm inferences about causality. For example, increased metro area imports could be a response to a negative technology shock affecting a specific industry in the home country that faces import competition.

In the next section, we begin by looking at some previous studies of imports into the U.S. and then describe our measures of import sensitivity. In the following section, we focus on exports.

Imports

Previous attempts at attributing U.S. imports to regions have been made at broad geographic levels. In their work on the potential impact of the North American Trade Agreement (Nafta), Hayward and Erickson (1995) allocate manufactured imports from Canada and Mexico by individual industry in proportion to each state's share of domestic shipments by that industry. They find much variation among states, and highlight the fact that these trade flows are smaller than most people believe. Hervey and Strauss (1998) allocate manufactured imports for an industrial aggregation at the even-broader "durable" and "nondurable" industry categories, though in the process, they are able to identify imports as coming from 44 individual foreign countries. They attribute high overall import shares to the manufacturing East South Central and East North Central regions. These high import shares are ambiguous in that they might represent either imports into a region or that region's competition for markets served in the remainder of the U.S.

In this section, we improve on these import allocations in two respects. First, we use a much narrower industry breakdown to allocate finely defined U.S. imported goods to particular metropolitan areas. Using employment data for U.S. counties from the *County Business Patterns* (CBP) data, we can identify four-digit Standard Industrial Classification (SIC) based industry definitions for manufactured products.[2] This use of narrow industry definition means, for example, that automotive production (and attendant import competition) need no longer be erroneously attributed to the state of Washington; domestic aircraft production need not be erroneously attributed to Detroit.

A second refinement is that imports can be attributed to metropolitan areas rather than to states and multi-state regions. Metropolitan area economies are more cohesive than state or multi-state economies, in that they share a common work force and transportation infrastructure. In addition, metropolitan areas are not so arbitrarily defined by jurisdictional boundaries, as are state economies, for example.[3]

Import trends

Manufactured imports into the U.S. grew rapidly for most of the 1990s. This is not surprising, given that import growth is strongly influenced by the overall growth of the home country's economy. From 1991, the trough of the previous economic downturn, to 2000, the peak of the expansion, total U.S. imports increased by $722 billion in nominal dollars, and by $616 billion (or 109 percent) as deflated by the general price index for U.S. gross domestic product (GDP) (figure 1). Indeed, import expansion outpaced the more general and robust expansion of the 1990s. As measured against the yardstick of (nominal) GDP, (nominal) imports climbed from an 8.1 percent ratio to GDP in 1991 to over 12 percent in 2000 (figure 2).

Import competition

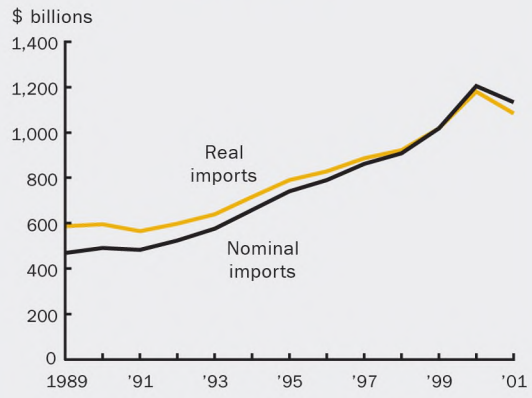
How did the run-up in imports play out across metropolitan areas? In order to link national import growth to the industries of a particular metropolitan area, we examine the industry-by-industry growth of foreign imports in the U.S. market. Therefore, we allocate actual U.S. imports by industry category to individual metropolitan areas according to the metropolitan presence of that same industry. In particular, we use employment data by industry at the county level of geography to construct a local employment share of each national industry for each of 269 metropolitan areas in the United States for 1989 and 1999.[4] Each MSA's employment share of the nation for a particular industry then becomes the metro area's share of national imports for that industry. For each metropolitan area, the sum total of these imports across all industries is taken as the measure of its total import competition.

Thus, import competition in MSA i = Sum over j M_{US}^{ij}, where $M_{US}^{ij} = L^{ij} \times M^{j}$ and L^{ij} = MSA i's share of U.S. employment for good j. M^{j} = U.S. imports of good j.

In examining the 25 most populous metropolitan areas in 1999, the allocated import pattern reveals an approximate but imperfect correspondence with the size of metropolitan population in 1999 (table 1).[5] Places with heavy manufacturing concentrations and large economies—such as Southern California—have an outsized measured share of estimated imports attendant to the region's industrial structure. However, there is much variation in these import allocations owing to varying industry (import) composition. As a yardstick, we can compare allocated imports against the size of each metropolitan economy. In order to do this, we construct estimates of gross metropolitan product (GMP) for each metropolitan area and report imports as a share of GMP.[6] For 1999, we find an estimated average ratio of imports to gross product of 9.48 percent for the 25 most populous metropolitan areas. The Detroit–Ann Arbor area is a leader in this measure with 19 percent. Heavy U.S. imports of automotive products—many of them from nearby Canada—coupled with Detroit's sharp concentration in automotive industries, lie behind the reported import competition. Manufacturing and technology-intensive San Francisco–Oakland–San Jose and Portland–Salem follow behind at 14 percent and 17 percent, respectively. At the other end of the spectrum, the

FIGURE 1

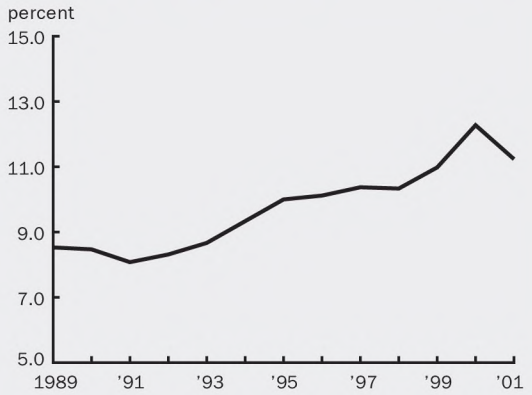
U.S. imports



Note: GDP deflator 1999=100.
Source: U.S. Department of Commerce, International Trade Administration, Trade Development, Office of Trade and Economic Analysis.

FIGURE 2

U.S. imports as percent of nominal GDP



Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Trade Development, Office of Trade and Economic Analysis; Haver Analytics; and U.S. Department of Commerce, Bureau of Economic Analysis.

TABLE 1

Manufacturing imports as percent of GMP (1999)

	MSA (by 1999 population)	Rank by imports	Imports ($billions)	% of GMP	GMP ($billions)
1	New York–Northern New Jersey–Long Island, NY–NJ–CT	2	43.3	5.4	797
2	Los Angeles–Riverside–Orange County, CA	1	51.8	11.2	464
3	Chicago–Gary–Kenosha, IL–IN–WI	6	27.9	8.8	316
4	Washington–Baltimore, DC–MD–VA–WV	31	6.1	2.1	289
5	San Francisco–Oakland–San Jose, CA	3	43.1	14.0	308
6	Philadelphia–Wilmington–Atlantic City, PA–NJ–DE–MD	9	16.6	8.6	193
7	Boston–Worcester–Lawrence, MA–NH–ME–CT	5	28.5	12.5	229
8	Detroit–Ann Arbor–Flint, MI	4	33.7	19.0	177
9	Dallas–Fort Worth, TX	8	17.0	9.1	186
10	Houston–Galveston–Brazoria, TX	7	18.1	11.0	164
11	Atlanta, GA	17	8.4	5.8	145
12	Miami–Fort Lauderdale, FL	40	4.7	5.0	93
13	Seattle–Tacoma–Bremerton, WA	11	10.3	7.8	132
14	Phoenix–Mesa, AZ	16	8.9	10.2	87
15	Cleveland–Akron, OH	18	8.2	9.6	86
16	Minneapolis–St. Paul, MN–WI	14	9.0	8.3	108
17	San Diego, CA	22	7.8	9.4	83
18	St. Louis, MO–IL	24	7.8	10.1	77
19	Denver–Boulder–Greeley, CO	19	8.1	8.5	96
20	Pittsburgh, PA	23	7.8	11.7	67
21	Tampa–St. Petersburg–Clearwater, FL	46	3.5	6.0	59
22	Portland–Salem, OR–WA	10	11.0	17.0	65
23	Cincinnati–Hamilton, OH–KY–IN	28	6.7	11.5	58
24	Kansas City, MO–KS	34	5.6	9.9	57
25	Sacramento–Yolo, CA	64	2.2	4.4	51

Notes: MSA is metropolitan statistical area. GMP is gross metropolitan product.
Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Trade Development, Office of Trade and Economic Analysis; Haver Analytics; and U.S. Department of Commerce, Bureau of Economic Analysis.

Baltimore–Washington, DC area registers only 2 percent on this metric.

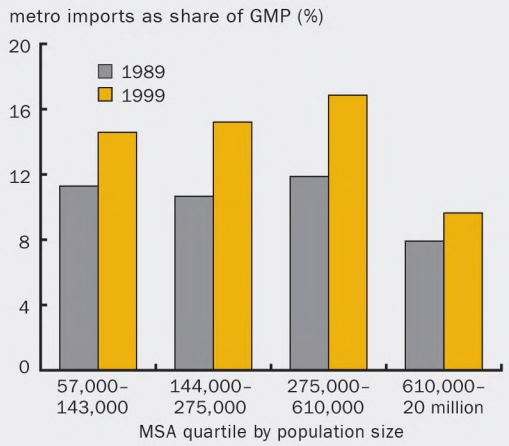
Looking more widely at all metropolitan areas taken as a group, the largest metropolitan areas average less direct import competition in manufacturing than smaller areas (figure 3).[7] In 1999, metropolitan areas greater in population than 610,000 approached 10 percent in import competition versus almost 17 percent for metro areas ranging in size from 275,000 to 610,000. The import competition for smaller metro areas differed little from this.

From 1989 to 1999, import intensity for all metro areas grew significantly, along with the general expansion of imports into the national economy. However, the average differences between large and small metro areas widened significantly, with the third quartile of metro area—with 275,000–610,000 in population—growing the most, from a ratio of 12 percent imports to GMP in 1989 up to almost 17 percent by 1999.

Are large metropolitan areas, then, not "open" economies compared with small metropolitan areas? Large metropolitan areas have increasingly become service economies and less hospitable to many types of manufacturing. Congested highways and high land costs in many large urban areas are not conducive to today's production processes in manufacturing. Nor have the tendencies toward global competition made it any easier for manufacturers in higher cost urban locales. In response, many domestic manufacturing facilities have sought out lower cost locales in small cities and rural areas, often adjacent or with close access to divided highways and the interstate highway system. There are countervailing forces at play, however. A counter-tendency has been the surge in technology and information intensity of the U.S. economy—both manufacturing and services alike. In this regard, urban areas are thought to have an advantage because key inputs to high-tech production—namely information and technology—may be acquired more easily in urban areas. At the same time, high-tech manufacturing industries often feature young firms that require proximity to the wide

FIGURE 3

Import competition by size of metro area, 1989 and 1999



Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Trade Development, Office of Trade and Economic Analysis.

array of specialized business, legal, and financial services that are to be found in large cities.[8]

The high service intensity that is attendant to manufacturing may also mean that data based on manufacturing location alone may belie the actual openness of the largest urban area economies. Embodied in the value of manufactured goods is an increasing service component—be it advertising, design, maintenance, management, marketing, or research and development. The service economy of a large city in America is in this way an unseen portion of international trade in goods. Such considerations are caveats to the traded good measures that we construct, and these caveats are inherent in almost all data on traded goods and their location of value added. More generally, globalization also means that the geography of production is stretched and expanded across wider and wider landscapes, making it more difficult to determine any meaningful and specific location of value added of exports.

Import penetration

An alternative way to measure imports into the U.S. domestic market more directly reflects "competition" to U.S. producers. Import penetration measures the ratio of imports for a particular industry to the sum of imports plus that portion of domestic production that is *not* exported abroad. Varying between zero and one, this measure of import penetration shows the share of domestic sales of a good that is imported rather than domestically produced. We measure an MSA's import penetration as a weighted average of national import penetration for each industry. For each metropolitan area, the weights are its own industry employment shares across all of manufacturing.

Import penetration in *MSAi* = Sum over all industries *j* MP^i, where $MP^i = L^{ij} \times MP^j$ and L^{ij} = *MSAi*'s share of its own manufacturing employment employed in industry *j*. MP^j = U.S. import penetration of good.[9]

Import penetration at the national level is often used to indicate the degree to which domestic sales in an industry have been penetrated or accounted for by imports.[10] For a particular region, we assume that an industry domiciled there tends to sell much of its output across the national domestic market. This assumption is somewhat realistic for U.S. metropolitan area economies because the U.S. market remains the primary market for domestic production plants. Exports as a share of U.S. gross domestic production remain below 8 percent overall. Meanwhile, domestic manufacturing plants sold between 64 percent and 82 percent of production domestically in the year 2000.[11]

We report import penetration estimates for the 25 most populous MSAs for 1999 (table 2). We see a wide range, from an import penetration of 11.7 percent for the Kansas City MSA, to upwards of 24 percent for San Diego. A pattern emerges that seems to suggest that high import penetration alone may not be indicative of local area industrial stagnation. For example, many MSAs known for a concentration in high technology also have high import penetration. These include Boston, the San Francisco Bay area, San Diego, Portland, and Phoenix. Translating metro area import penetration rates to the more familiar state level, we can map the geography of import competition for the entire country (figure 4 on p. 19).[12] One can see that in 1999 most of the states east of the Mississippi River (bold line on map) experienced import competition on par or above the U.S. average (16 percent). Somewhat surprisingly, eight states west of the Mississippi generally not associated with manufacturing report above-average levels of import competition as well.

Increases in import penetration *over time* may be more reflective of industrial competition. Here, the variation in growth of import penetration is again very wide (see table 2). Metro areas such as Miami and Kansas City registered under 30 percent growth in penetration from 1989 to 1999; metro areas as diverse as Seattle and Pittsburgh more than doubled their import penetration over the same period.

TABLE 2

Import penetration, 25 largest metro areas

MSA (by population)		Import penetration *(percent)* 1989	1999	% change 1989–99
1	New York–Northern New Jersey–Long Island, NY–NJ–CT	10.4	17.8	70
2	Los Angeles–Riverside–Orange County, CA	11.1	17.7	60
3	Chicago–Gary–Kenosha, IL–IN–WI	9.6	15.5	62
4	Washington–Baltimore, DC–MD–VA–WV	8.2	12.4	51
5	San Francisco–Oakland–San Jose, CA	16.7	23.7	42
6	Philadelphia–Wilmington–Atlantic City, PA–NJ–DE–MD	9.3	16.5	77
7	Boston–Worcester–Lawrence, MA–NH–ME–CT	14.6	21.7	49
8	Detroit–Ann Arbor–Flint, MI	11.3	16.3	45
9	Dallas–Fort Worth, TX	10.4	16.1	55
10	Houston–Galveston–Brazoria, TX	8.8	13.9	58
11	Atlanta, GA	8.4	12.6	50
12	Miami–Fort Lauderdale, FL	13.2	16.7	26
13	Seattle–Tacoma–Bremerton, WA	7.9	16.7	112
14	Phoenix–Mesa, AZ	14.7	19.3	32
15	Cleveland–Akron, OH	8.8	13.8	56
16	Minneapolis–St. Paul, MN–WI	9.1	14.1	55
17	San Diego, CA	16.8	24.1	43
18	St. Louis, MO–IL	9.6	12.9	35
19	Denver–Boulder–Greeley, CO*	13.2	17.3	31
20	Pittsburgh, PA	8.7	17.8	104
21	Tampa–St. Petersburg–Clearwater, FL	9.5	16.2	70
22	Portland–Salem, OR–WA	13.2	21.7	64
23	Cincinnati–Hamilton, OH–KY–IN	10.0	16.9	68
24	Kansas City, MO–KS	9.2	11.7	28
25	Sacramento–Yolo, CA	7.9	12.8	62

Notes: MSA is metropolitan statistical area. GMP is gross metropolitan product.
Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration Trade Development, Office of Trade and Economic Analysis; National Bureau of Economic Research; and U.S. Department of Commerce, Bureau of the Census, Center for Economic Studies, *Annual Survey of Manufactures*.

Exports

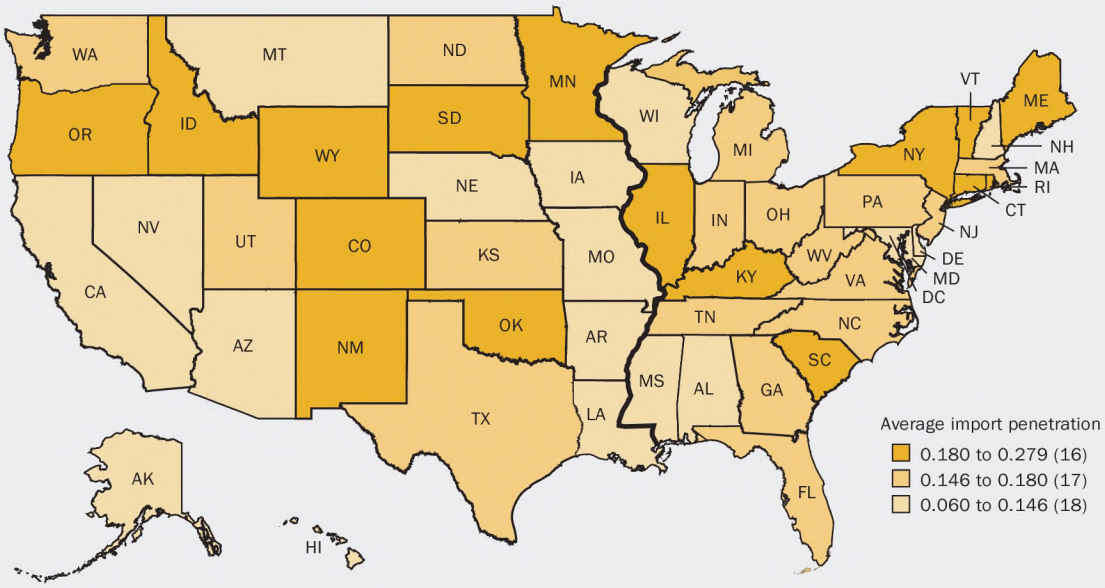
The flip side of import penetration has been the rapid export expansion of U.S. manufacturing. Both economic growth in overseas markets and lower tariff barriers to trade have helped to expand U.S. exports. Until the currency crises beginning in 1997, rapidly developing countries in Asia such as Thailand, Malaysia, Korea, Singapore, and Taiwan led the world in rates of economic growth. Though growth was export-led there, imports of manufacturing goods from developed countries—especially capital goods—grew as well, largely to meet the development needs of manufacturing industries in these nations. The manufacturing sectors of the industrial economies, including the U.S., grew rapidly to meet the demands of both the developing economies and a general worldwide expansion.

The nominal value of manufactured exports attributed to U.S. metropolitan areas was $567 billion for the last reported year, 1999, up from $374 billion in the first reported year, 1993 (see figure 5). Exports began to level off in 1997, coincident with the Asian economic crisis. As measured against the gross domestic product of metropolitan areas, exports declined from a peak of 8.8 percent in 1997 to 7.9 percent by 1999 (see figure 6).

For individual metropolitan areas, export data are telling but not straightforward to interpret. The only publicly reported export figures for MSAs are drawn from information of the U.S. Census Bureau, compiled and reported by the International Trade Administration. In particular, exports are reported by businesses in "export declarations," which identify location using five-digit zip codes. Yet the exporter of record is not necessarily the entity that produced the merchandise, so the data do not fully reflect the production origin of manufactured goods. Instead, the exporter of record is the party "principally responsible for effecting export from the United States."[13] This means that if the exporter of record is a manufacturing company, the location may either be the production plant

FIGURE 4

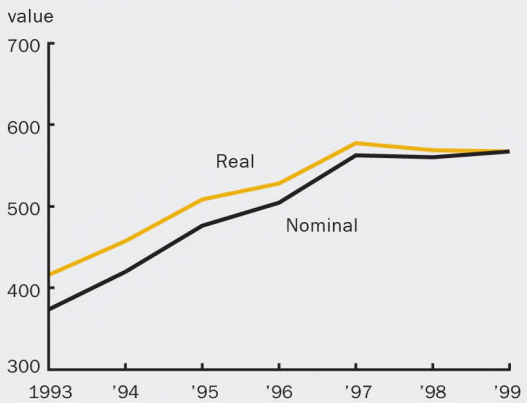
Import penetration 1999: Metro areas by state



or an administrative establishment of the company, such as a corporate headquarters. Similarly, exporters of record can be service companies, typically wholesalers, but also other intermediaries, such as retailers.[14] This means that the wholesaler, headquarters, or marketing arm of manufacturing—to which the export may be attributed by the data—actually tends to be responsible for some significant value added.[15] Yet, the location of export production often tends to be coincident, with wholesalers of a manufacturing product likely to locate in the same region as the producer. The larger metropolitan areas are likely to

FIGURE 5

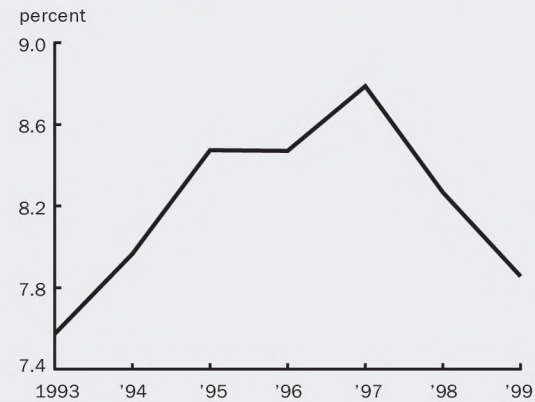
Sum of MSA exports ($ billions)



Notes: GDP deflator 1999=100. Non-manufacturing commodities are subtracted from total exports in 43 MSAs.
Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis and U.S. Department of Commerce, Bureau of Census, *Exporter Location Series*.

FIGURE 6

Metro area exports as share of nominal GMP



Note: Non-manufactured commodities are subtracted from total exports in 43 MSAs.
Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis, *Exporter Location Series*, and U.S. Department of Commerce, Bureau of Economic Analysis.

more accurately represent the location of value-added exports because larger areas are more likely to contain all parties in the transaction—all contributors to value added of the exported good.

Accordingly, we aggregate to the largest possible MSA geographic definition, the so-called consolidated metropolitan statistical area or CMSA. In this way, we minimize the errors inherent in geographical assignment versus site of production. In addition, in reporting on individual metropolitan areas, we focus on the largest MSAs. Wherever possible, we exclude commodities, such as coal and minerals, from our measures of manufacturing exports. Commodities are more likely to be produced outside metropolitan areas, but to be shipped abroad from them.[16]

Export levels tend to correlate with size of the metropolitan area. As a result, a "mega-sized" NY–Long Island–Northern New Jersey consolidated metropolitan area yields very large reported manufactured exports, leading with $46.6 billion in 1999 (table 3). As we might expect from their large size and manufacturing orientation, the greater Chicago and Los Angeles areas round out the top three in value of exports. However, the correspondence between size of economy and level of exports is highly variable across the top 25 largest metropolitan areas. The ratio of exports to gross metropolitan product averaged 8.7 percent in 1999, but the standard deviation was a sizable 5.2. At the low end, service industry and domestically oriented regional areas such as the Washington–Baltimore–Northern VA area reported a low 3.2 percent of regional product. At the top of the spectrum, shipping-oriented and aerospace-intensive Seattle reported a ratio of 25 percent. High-tech San Francisco–Oakland–San Jose (at 14.8 percent) aligns with our high prior expectations for that economy. Auto-intensive Detroit–Ann Arbor's ratio (at 17.2 percent) may be surprising to some, since the automotive sector is not always known as a U.S. export industry. However, the Detroit auto corridor to Ontario ranks among the

TABLE 3

Metro area export intensity, 1999

	MSA (by population)	Exports ($billions)	% of GMP	GMP ($billions)
1	New York–Northern New Jersey–Long Island, NY–NJ–CT	46.6	5.8	797
2	Los Angeles–Riverside–Orange County, CA	34.7	7.5	464
3	Chicago–Gary–Kenosha, IL–IN–WI	21.5	6.8	316
4	Washington–Baltimore, DC–MD–VA–WV	9.4	3.2	289
5	San Francisco–Oakland–San Jose, CA	45.6	14.8	308
6	Philadelphia–Wilmington–Atlantic City, PA–NJ–DE–MD	14.2	7.4	193
7	Boston–Worcester–Lawrence, MA–NH–ME–CT	15.5	6.8	229
8	Detroit–Ann Arbor–Flint, MI	30.5	17.2	177
9	Dallas–Fort Worth, TX	11.8	6.4	186
10	Houston–Galveston–Brazoria, TX	19.3	11.8	164
11	Atlanta, GA	7.2	4.9	145
12	Miami–Fort Lauderdale, FL	13.9	15.0	93
13	Seattle–Tacoma–Bremerton, WA	33.0	25.0	132
14	Phoenix–Mesa, AZ	7.3	8.5	87
15	Cleveland–Akron, OH	7.4	8.6	86
16	Minneapolis–St. Paul, MN–WI	8.4	7.8	108
17	San Diego, CA	8.7	10.4	83
18	St. Louis, MO–IL	4.4	5.7	77
19	Denver–Boulder–Greeley, CO[a]	2.6	2.7	96
20	Pittsburgh, PA	3.6	5.4	67
21	Tampa–St. Petersburg–Clearwater, FL[a]	2.4	4.1	59
22	Portland–Salem, OR–WA	7.9	12.1	65
23	Cincinnati–Hamilton, OH–KY–IN	6.6	11.4	58
24	Kansas City, MO–KS	1.6	2.9	57
25	Sacramento–Yolo, CA[a]	2.6	5.1	51

[a]Exports include nonmanufactured commodity shipments

Notes: MSA is metropolitan statistical area. GMP is gross metropolitan product.

Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis; U.S. Department of Commerce, Bureau of the Census, *Exporter Location Series*; and U.S. Department of Commerce, Bureau of Economic Analysis.

most integrated binational economic relationships in the world.[17]

Exports appear to have added to metro area growth in the 1990s, assuming no displacement. For the 25 most populous regions reported in table 4, estimated export growth added an average of 3.4 percent to the size of metropolitan economies from 1993 to 1999. In comparison, import growth (also measured against GMP) for the same period and sample averaged 8.4 percent. The Asian crisis falloff post-1997 in U.S. exports accounts for some of this difference; U.S. exports flattened out, even while domestic demand (and import purchases) continued to grow robustly.

Export-led growth contributed more to large metropolitan area economies than to smaller ones. Over the 1993–99 period, exports contributed an average of over 3 percentage points to growth in metropolitan areas with over 500,000 in population, versus just over 1 percent in the smallest population size category, 250,000 and less.

Can we generalize about export orientation by size of metro area economy? Figure 7 confirms that larger metropolitan areas are more export oriented. The top quartile, with population of 900,000 and above, report a weighted average of over 8 percent exports in 1999. In contrast, smaller metropolitan areas report smaller average export intensities for 1999, with an average of 6.95 percent for the second largest quartile, 6.36 for the third, and 6.74 for the fourth and smallest quartile. Still, these data suggest that smaller metropolitan areas *do* fully participate in export trade. In this regard, it is noteworthy that export intensity increased across all MSA size classes from 1993 to 1999.

A different explanation for the high degree of export intensity in the San Francisco Bay area is that reported exports may overstate actual exports along

TABLE 4

Prospective export intensity

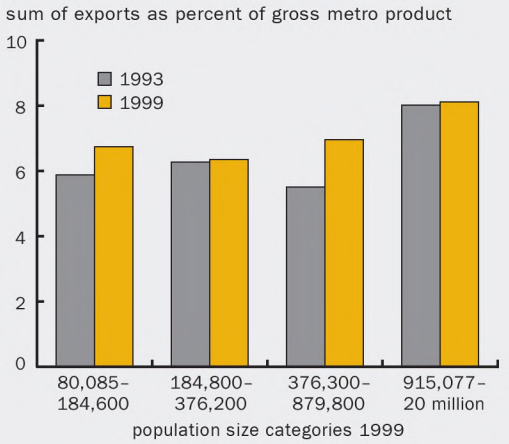
	MSA (by population)	GMP 1993 ($billions)	Exports 1993 ($billions)[b]	Exports 1999 ($billions)[b]	Growth[a] (GMP 1993)
1	New York–Northern New Jersey–Long Island, NY–NJ–CT	561	43.5	41.8	–0.3
2	Los Angeles–Riverside–Orange County, CA	343	25.6	31.1	1.6
3	Chicago–Gary–Kenosha, IL–IN–WI	221	13.8	19.3	2.5
4	Washington–Baltimore, DC–MD–VA–WV	206	8.5	8.4	0.0
5	San Francisco–Oakland–San Jose, CA	190	29.8	41.0	5.9
6	Philadelphia–Wilmington–Atlantic City, PA–NJ–DE–MD	143	9.3	12.8	2.4
7	Boston–Worcester–Lawrence, MA–NH–ME–CT	152	10.0	14.0	2.6
8	Detroit–Ann Arbor–Flint, MI	128	19.7	27.4	6.0
9	Dallas–Fort Worth, TX	112	6.2	10.6	4.0
10	Houston–Galveston–Brazoria, TX	104	12.8	17.3	4.4
11	Atlanta, GA	86	3.6	6.5	3.3
12	Miami–Fort Lauderdale, FL	67	9.0	12.5	5.1
13	Seattle–Tacoma–Bremerton, WA	82	24.0	29.7	7.0
14	Phoenix–Mesa, AZ	48	4.3	6.6	4.8
15	Cleveland–Akron, OH	65	4.7	6.6	2.9
16	Minneapolis–St. Paul, MN–WI	72	6.2	7.5	1.9
17	San Diego, CA	56	4.3	7.8	6.4
18	St. Louis, MO–IL	57	3.0	3.9	1.6
19	Denver–Boulder–Greeley, CO	57	1.3	2.3	1.7
20	Pittsburgh, PA	51	2.6	3.2	1.2
21	Tampa–St. Petersburg–Clearwater, FL	38	1.3	2.2	2.3
22	Portland–Salem, OR–WA	42	3.6	7.1	8.2
23	Cincinnati–Hamilton, OH–KY–IN	40	3.8	6.0	5.4
24	Kansas City, MO–KS	39	1.1	1.5	0.9
25	Sacramento–Yolo, CA	33	1.2	2.3	3.3

[a]Real growth in 1993–99 exports as a percent of 1993 GMP.
[b]Deflated to 1993.
Notes: MSA is metropolitan statistical area. GMP is gross metropolitan product.
Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis; U.S. Department of Commerce, Bureau of the Census, *Exporter Location Series*; and U.S. Department of Commerce, Bureau of Economic Analysis.

some dimensions. In particular, seaports such as Portland and San Francisco may have more exports attributed to them in the reported data than they would under an alternative method that reflected the origin of production. To explore the possible bias in the reported export data further, we construct a second, hypothetical set of export figures (see table 5). We allocate (impute) to metropolitan areas national-level U.S. Department of Commerce, ITA data on exports. Export data at the national level are available for detailed industry classifications. We apportion these exports to particular metropolitan areas according, again, to the area's national share of employment in the corresponding

FIGURE 7

Export intensity by size of metro areas



Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis; U.S. Department of Commerce, Bureau of Census, *Exporter Location Series*, and U.S. Department of Commerce, Bureau of Economic Analysis.

industry. We then develop an overall measure of exports from each metropolitan area by summing across all industries. In this measure, since we are apportioning exports directly by the location of production, the export total will tend to reflect "origin of movement." Thus, this measure will have the opposite bias to the reported exports, which may reflect point of shipment or arrangement to ship. Note that there will be errors in assigning the exports in the new measure due to the fact that not all industries in locales actually have the same propensity to export. In addition, our measure assumes that labor intensity is uniform geographically within each industry.

Looking individually at the most populous metropolitan areas, we see a general tendency for the reported exports to exceed the estimated and hypothetical measures, though this is not the case for nine of the metro regions. More startling is the extent to which the reported metro area figures exceed the production-oriented estimates that are imputed from ITA data in a number of metro areas with significant international ports or trans-shipment industries: In New York, the reported data exceed the imputed estimate by $22.2 billion (91 percent); in the San Francisco Bay metro area by $18.7 billion (69 percent); in Seattle–Tacoma by $18.5 billion (128 percent); in Miami by $11.6 billion (207 percent); in Houston by $8.1 billion (72 percent); and in Detroit by $15.8 billion (108 percent). On the other hand, interior cities such as Denver, Dallas, and St. Louis display a hypothetical export base, according to the production method of allocation, that is greater than the reported export figures for these cities.

To explain the variance between these data series, we ran an OLS regression with the metropolitan export estimates reported by the ITA as the dependent variable (see table 6 on p. 24). We used the sample of all 208 metropolitan areas for which data were reported for 1998 and 1999. We included an independent variable for each observed metropolitan area, "percent of its employment in transportation industries," to test for the effect of shipment rather than production location on manufacturing exports. Even after accounting for the estimated exports based on the location of manufacturing, the transportation variable is positive and significant at the 1 percent level.

Another source of bias in the reported metro export series is suspected to arise from the separate location of a manufacturer's corporate headquarters from its production plants. In particular, exports may tend to be attributed to the location of the headquarters rather than to the location of the production plant. Large companies in particular have a very high propensity to export, and they also tend to have separate headquarters locations. Accordingly, the presence of a single or multiple large company headquarters in an MSA might tend to inflate the reported export figures compared with our estimated (imputed) export figure, the latter being based on the production employment location of industries. When we account for large headquarters in our regression equation, we find no apparent *systematic* relationship between headquarters location and the levels of reported exports. Of course, there may be significant individual instances in which large-scale exports are attributed to company headquarters, and thereby serve to inflate the reported exports of particular MSAs.

TABLE 5

Export sensitivity test: ITA export data versus estimates

	MSA (by population)	1999 ITA exports ($billions)	1999 estimates ($billions)	1999 $X_{ITA} - X_{est.}$ (billions)	1993 $X_{ITA} - X_{est.}$ (billions)
1	New York–Northern New Jersey–Long Island, NY–NJ–CT	46.6	24.4	22.2	23.6
2	Los Angeles–Riverside–Orange County, CA	34.7	32.4	2.3	4.2
3	Chicago–Gary–Kenosha, IL–IN–WI	21.5	17.5	4.0	2.2
4	Washington–Baltimore, DC–MD–VA–WV	9.4	4.5	4.9	5.2
5	San Francisco–Oakland–San Jose, CA	45.6	27.0	18.7	15.6
6	Philadelphia–Wilmington–Atlantic City, PA–NJ–DE–MD	14.2	10.2	4.0	1.7
7	Boston–Worcester–Lawrence, MA–NH–ME–CT	15.5	18.5	–3.0	–0.9
8	Detroit–Ann Arbor–Flint, MI	30.5	14.7	15.8	6.6
9	Dallas–Fort Worth, TX	11.8	14.6	–2.8	–1.7
10	Houston–Galveston–Brazoria, TX	19.3	11.2	8.1	4.0
11	Atlanta, GA	7.2	5.6	1.6	–0.2
12	Miami–Fort Lauderdale, FL	13.9	2.3	11.6	7.3
13	Seattle–Tacoma–Bremerton, WA	33.0	14.5	18.5	15.5
14	Phoenix–Mesa, AZ	7.3	7.9	–0.6	0.4
15	Cleveland–Akron, OH	7.4	6.4	1.0	0.1
16	Minneapolis–St. Paul, MN–WI	8.4	6.5	1.9	1.4
17	San Diego, CA	8.7	6.5	2.2	0.0
18	St. Louis, MO–IL	4.4	8.4	–4.1	–3.0
19	Denver–Boulder–Greeley, CO	2.6	3.5	–0.9	–1.4
20	Pittsburgh, PA	3.6	4.5	–0.9	–0.3
21	Tampa–St. Petersburg–Clearwater, FL	2.4	2.0	0.4	–0.1
22	Portland–Salem, OR–WA	7.9	7.5	0.4	–0.3
23	Cincinnati–Hamilton, OH–KY–IN	6.6	4.9	1.8	0.7
24	Kansas City, MO–KS	1.6	2.6	–1.0	–1.3
25	Sacramento–Yolo, CA	2.6	1.9	0.7	0.4

Notes: The two datasets overlap between 1993 and 1999. In the data reported by the International Trade Administration (ITA), the non-manufactured commodity shipments have been subtracted in 43 metropolitan statistical areas (MSA).

Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis; U.S. Department of Commerce, Bureau of the Census, *Exporter Location Series*; and U.S. Department of Commerce, Bureau of Economic Analysis.

Examining trade and growth

Are the measures of trade—exports, export competition, import growth, and import penetration—meaningful measures that affect the growth and composition of metropolitan areas? One way to explore this question is to examine the relationship between MSA economic growth and these trade-related measurements over time. To do that, we can relate the growth or decline in MSA total manufacturing employment to changes in these variables using multiple regression. We use manufacturing employment as the growth indicator—the dependent variable to be explained—owing to general data availability of these figures at the MSA level of geography over the 1989–99 period. In particular, because MSAs vary in size, we use the percent change in total manufacturing employment as the dependent variable for the 269 U.S. MSAs.

The general estimation strategy is to examine a cross-sectional panel of the percentage change in manufacturing employment—a variable of growth relative to the size of each particular MSA over a ten-year period. The estimation is in "changes" or first difference form. This functional form has the advantage of differencing out variables which presumably remain constant for individual MSAs over the period, but which vary significantly in level across MSAs—the so-called omitted variable problem. A possible downside is that, in first differencing, any measurement errors in the regression variables tend to be magnified—leading to inefficient estimators or large standard errors in the coefficients.

The explanatory variables are also measured in percent changes, so that the coefficients can be read as elasticities. In this, there are two exceptions. One is that the MSA's predominant broad geographic region is entered as a fixed effect. This is intended to pick up the broad inter-regional shifts of economic activity, which have been taking place from Frost Belt regions to Sun Belt and from the Northeast–Midwest to the

TABLE 6

OLS regression: ITA annual manufacturing exports by metro area (natural log)

	1	2	3
Constant	-6.36**	-3.41**	-4.72**
	(.40)	(.35)	(.46)
Year (1998=1, 0 otherwise)	.075	.067	.071
	(.072)	(.07)	(.069)
Estimated exports (log)		.66**	.46**
		(.05)	(.071)
Gross metro product (log)	1.21**	.49**	.73**
	(.041)	(.06)	(.083)
Employment share, transportation (log)	11.40**	14.0**	14.14**
	(2.00)	(2.00)	(1.95)
Employment share, manufacturing (log)	5.27**		2.64**
	(.049)		(.62)
Number of large manufacturing HQs	-.002	.003	.001
	(.005)	(.005)	(.005)
Adjusted R-squared	.80	.81	.82
	N=416	N=416	N=416

** Denotes statistical significance at the 1 percent level. Standard error in parentheses.

Notes: OLS is ordinary least squares. HQs is headquarters. Regressions were also estimated substituting "all manufacturing headquarters" for "large" with different results. The signs on independent variables, including transportation, remained the same. However, statistical significance dropped. "Large headquarters" is defined as worldwide employment of 2,000 or more. The calculations in this table are based upon $Exports_{MSA} = \beta_1(estExports_{MSA}) + \beta_2(GMP_{MSA}) + \beta_3(ShTr_{MSA}) + \beta_4(ShMfg_{MSA}) + \beta_5(HQ_{MSA}) + \beta_6(Year\ Dummy) + \varepsilon_{MSA}$.

Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis; U.S. Department of Commerce, Bureau of the Census, *Exporter Location Series*; U.S. Department of Commerce, Bureau of Economic Analysis; and Compustat.

South and West. Population in particular has been shifting in these directions, which would be especially reflective of regional shifts in nontraded manufactured goods. A second exception is the "size of MSA economy" variable. This measure reflects the fact that production activity in general has been tending to deconcentrate from large metropolitan areas. This trend has been ongoing for 50 years or more as production technology has been shifting from multi-story, railroad-dependent, labor-intensive modes to low-slung, truck-intensive, capital-intensive operations. We measure each of the remaining variables as percent change and define them identically to those discussed in the preceding sections.

In estimating this simple OLS regression, we find statistically significant results that are plausible (table 7). That is, both export and import growth are associated with manufacturing gains. That is consistent with imports having grown faster in industries with a strong expansion in domestic demand.[18] In addition, we find that import penetration is negatively related to manufacturing employment (column 3). Thus, an increase in the level of imports as a share of domestic demand is associated with lower labor usage over time across metro areas. Similarly, we find a negative relationship between a change in export intensity and manufacturing employment. A possible interpretation is that U.S. industries may need to become more efficient in order to compete in and capture foreign markets. Yet, caution is urged as this simple modeling exercise does not allow for confident inferences about causality. For example, increased metro area imports could be a response to a negative technology shock affecting a specific industry in the U.S. that faces import competition.

Conclusion

The process of globalization has moved ahead over the past ten to 15 years, albeit in fits and starts. Some have speculated that metropolitan economies are sure to undergo restructuring and upheaval attendant to globalization. Their industry structure and performance, along with local wages and prices, are thought to be

TABLE 7

OLS regression: Dependent variable as percent change in manufacturing employment by MSA, 1989–99

	1	2	3	4	5	6	7	8
Constant	–0.0562 (0.030)	–0.0504 (0.033)	–0.0003 (0.030)	–0.0154 (0.029)	–0.019 (0.027)	–0.014 (0.027)	–0.017 (0.029)	–0.062 (0.017)*
% Change exports	0.0873 (0.0144)*	0.0881 (0.0145)*	0.1311 (0.014)*	0.1131 (0.01382)*	0.186 (0.017)*	0.178 (0.018)*	0.179 (0.018)*	0.196 (0.018)*
% Change imports	0.0202 (0.00948)*	0.0212 (0.00950)*		0.0492 (0.0097)*	0.021 (0.008)*	0.027 (0.01)*	0.028 (0.010)*	0.035 (0.010)*
% Change import penetration			–0.1208 (0.0241)*	–0.1773 (0.0256)*		–0.039 (0.034)	–0.038 (0.035)	–0.036 (0.035)
% Change export penetration					–0.250 (0.027)*	–0.218 (0.039)*	–0.217 (0.040)*	–0.266 (0.039)*
% Change productivity, 1987–97		–0.0287 (0.0130)*					–0.015 (0.011)	–0.011 (0.012)
Size of MSA by GMP, 1989	–0.0009 (0.0003)*	–0.0007 (0.0003)*	–0.0009 (0.0003)*	–0.0009 (0.0003)*	–0.0006 (0.0003)*	–0.0006 (0.0002)*	–0.0006 (0.00002)*	–0.0007 (0.00002)*
Regional fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Sample size	269	269	269	269	269	269	269	269
R-bar squared	0.350	0.367	0.397	0.451	0.510	0.510	0.517	0.473

* T-stat significant at 5 percent level.

Notes: Standard errors in parentheses. OLS is ordinary least squares. MSA is metropolitan statistical area. GMP is gross metropolitan product. The calculations in this table are based upon $PchMfgEmp_{msa} = \beta_1(pChExports_{msa}) + \beta_2(pChImports_{msa}) + \beta_3(pChImPenetration_{msa}) + \beta_4(pChExportPenetration_{msa}) + \beta_5(pChProductivity) + \beta_6(regional\ dummies) + \upsilon_{MSA}$.

Source: Authors' calculations based on data from U.S. Department of Commerce, International Trade Administration, Office of Trade and Economic Analysis; U.S. Department of Commerce, Bureau of the Census, *Exporter Location Series*, Center for Economic Studies, and *Annual Survey of Manufactures*; and U.S. Department of Commerce, Bureau of Economic Analysis; National Bureau of Economic Research.

under pressure as global integration gives rise to greater trade opportunities and challenges. Many such changes will take place, regardless of direct outflows and inflows of tradable goods. Still, measures of exports and import competition may be important indicators to policymakers and others of the sources and direction of upheaval and change. Both exports and imports of manufactured goods have generally expanded for U.S. metropolitan areas over the past decade or more—imports more than exports—with wide variation across metropolitan areas.

Heretofore, direct insights into trade-related restructuring for U.S. MSAs were sparse, because there is little data reporting directly on international trade at the metropolitan area level of geography. In constructing new estimates, and exploring the properties of existing data, we have suggested that there is a wide variation in the openness to trade in manufactured goods among U.S. metropolitan areas, both in direct exports from these areas and in measures of import competition and penetration for their specific industrial sectors.

Large metropolitan areas do not appear to be experiencing as high a level of import competition for manufactured goods as medium and smaller metropolitan areas. However, this may reflect the more general tendency of production activity to eschew large urban areas in favor of less densely populated areas. In reality, the increasing service intensity of large urban areas may belie their actual trade intensity in the global trade of manufactured goods. Services, many of them originating in large urban areas, are implicitly and increasingly embodied in manufactured goods—imports, exports, and domestically produced goods alike.

When it comes to exports, large metropolitan areas tend to report high export intensity for manufactured goods. Many large urban areas tend to ship or facilitate shipment of exported goods, and they are accordingly cited as the domicile of exports abroad. In one sense, this is misleading in that the origin of production for many of these exports is likely to be in smaller metro areas or rural areas. However, in another sense, it is appropriate to attribute value added to large metropolitan areas because, as mentioned above, services embedded in manufactured goods often originate in these areas.

These first explorations of trade-related data at the metropolitan level remain suggestive and rudimentary. There is much more that we don't know. Trade remains a single but important element contributing to shifting roles and structure of metropolitan areas.

NOTES

[1]See Campa and Goldberg (1997), who identify changes in the use of imported inputs for a set of manufacturing industries from the U.S., Canada, the UK, and Japan. They find manufacturing industries in the U.S. to have experienced a very strong increase in the use of imported inputs. The role of trade in imported inputs in the context of our article will be left for future research. See also Hummels et al. (2001), who show that the internationalization of the supply chain, a phenomenon they refer to as vertical specialization, has accounted for a large and increasing share of international trade over the last several decades. They find that this increase in vertical specialization accounts for a sizable piece of the growth in world trade.

[2]*County Business Patterns* is published by the U.S. Department of Commerce, Bureau of the Census, and contains information on employment, payroll, and number of establishments by industry for every county in the U.S. For 1998 and 1999, the North American Industrial Classification System (NAICS) replaced the Standard Industrial Classification (SIC) system. This new classification maps fairly well into the former SIC system for manufacturing industries. One exception is the auxiliary establishment employment of manufacturing companies (for example, corporate headquarters), which has been shifted to a separate NAICS category in "services."

[3]For example, the Chicago Consolidated Metropolitan Statistical area encompasses the primary metropolitan statistical areas of Chicago, IL, Gary, IN, Kankakee, IL, and Kenosha, WI. Yet, it is possible that the use of subregional entities, such as metropolitan areas, does not sufficiently account for regional economic linkages. Multi-state regions are often highly integrated in their trade between and among industries (see Hewings et al., 1998).

[4]The CBP industry data are available at a four-digit level based on SIC for 1977–97 and a six-digit industry level based on NAICS for 1998 and 1999. Undisclosed or "suppressed" CPB data were estimated by the Center for Public Policy at Northern Illinois University (see Gardocki and Baj, 1985).

The data on U.S. exports and imports by industry contain information on the value of physical goods that have cleared through customs. These were provided to us by the International Trade Administration of the U.S. Department of Commerce. Exports are limited to domestic exports and are valued "free alongside ship," while imports are restricted to goods imported for consumption (not for re-export) and are on a customs value basis.

The ITA data, which we use in our analysis, are classified according to four-digit SIC and six-digit NAICS codes, yet, in their original form, they do not strictly conform to the SIC and NAICS industry definitions. The trade and economic analysts at the ITA mapped original trade data categories from the Bureau of the Census' Foreign Trade (exports and imports) into industry classification codes (NAICS and SIC). They did so based on their knowledge of and familiarity with industry products in international trade into and out of the U.S. Between 2 percent and 3 percent of imports could not be reliably assigned, or were assigned to miscellaneous categories. These are dropped from our analysis.

[5]We construct these measures for 269 metropolitan areas of the U.S. The employment data at the fine level of industry detail are thought to be much more accurate for large metropolitan areas. In such places, the county-level employment by industry is likely to be less subject to errors of imputation. There, employment data will be reported directly rather than imputed due to "disclosure" problems of a thin presence in the number of establishments in any particular industry.

[6]We construct our own estimates of gross regional product. To do so, we allocate nonagricultural gross product for the U.S. to each metropolitan area in proportion to its share of nonfarm personal income.

[7]Metropolitan area figures here are constructed as if the regions were a single region, rather than taking an arithmetic mean with each metropolitan area as an observation. (Either way, the results differ little.)

[8]See Ono (2001).

[9]Import penetration is defined for an industry by the ratio of imports to domestic market sales.

[10]Yet this is not a wholly accurate accounting of the local impact of overseas activity. An unmeasured change in competition or displacement may take place in foreign markets that are now contested between U.S.-domiciled production plants and overseas producers.

[11]The estimated range is derived by taking the value of exports of manufactured goods to total production in the manufactured sector. The larger estimate measures production by "value added in manufacturing." The lower figure uses "value of shipments" in manufacturing as the base. Since "export" value includes value added from nonmanufacturing industries, value of shipments may be an appropriate basis of comparison. However, shipments also include re-shipments, some of which may be exported. Hence, there may be a double counting in the U.S. of shipments data, making value added another, perhaps preferable candidate.

[12]A state's value on the map represents the average of its MSAs' import penetration rates. The MSAs encompassing multiple states and component primary MSAs (PMSAs) are allocated to the state that includes the PMSA.

[13]See U.S. Department of Commerce (1999).

[14]As with other data, there are also flaws in reporting. In this case, the principle problems are that approximately 7 percent of exports do not report a location; and another 3 percent are not allocable to particular metropolitan areas using a zip code basis (the "crossover" or overlap problem). The data are f.a.s. (free alongside ship) basis and include re-exports.

[15]See Dow Jones and Company (2003), which reports that, of the 17 million manufacturing jobs in the U.S., 52 percent are production workers versus 68 percent ten years earlier. For regional perspectives, see Testa (1989). These articles document that the value of manufacturing shipments, exports or domestic shipments, is increasingly composed of both services produced by manufacturing companies and services purchased by manufacturing companies and "embedded" into the value of the final manufacturing shipment.

[16]In particular, for many of the largest MSAs, the reported data break out "commodity" exports, such as agriculture and mining, for metropolitan areas. In our tables listing exports by metropolitan area, we have extracted commodity exports whenever possible.

[17]See Klier and Testa (2002).

[18]See Hine and Wright (1997), who also point out that import and export growth rates tend to be strongly positively correlated.

REFERENCES

Campa, Jose, and Linda S. Goldberg, 1997, "The evolving external orientation of manufacturing industries: Evidence from four countries" *Economic Policy Review*, Vol. 3, July, pp. 53–81.

Dow Jones and Company, 2003, "Manufacturers find themselves increasingly in the service sector," *Wall Street Journal*, February 10, p. A2.

Gardocki, Jr., Bernard C., and John Baj, 1985, "Methodology for estimating nondisclosure in county business patterns," Northern Illinois University, Center for Governmental Studies, April, mimeo.

Hayward D. J., and R. A. Erickson, 1995, "The North American trade of U.S. states: A comparative analysis of industrial shipments, 1983–91" *International Regional Science Review*, Vol. 18, No. 1, pp. 1–31.

Hervey, Jack L., 1999, "A regional approach to measures of import activity" *Chicago Fed Letter*, Federal Reserve Bank of Chicago, No. 147, November.

Hervey, J. L., and W. A. Strauss, 1998, "Foreign growth, the dollar, and regional economies 1970–97," *Economic Perspectives*, Federal Reserve Bank of Chicago, Fourth Quarter, pp. 35–55.

Hewings, Geoffrey J. D., Graham R. Schindler, and Philip R. Israilevich, 1998, "Interstate trade among Midwest economies," *Chicago Fed Letter*, Federal Reserve Bank of Chicago, No. 129, May.

Hine, Robert C., and Peter Wright, 1997, "Trade and manufacturing employment in the United Kingdom," in *International Trade and Labour Markets*, Jitendralal Borkakoti and Chris Milner (eds.), New York: St. Martin's Press, pp. 118–139.

Hummels, David, Jun Ishii, and Kei-Mu Yi, 2001, "The nature and growth of vertical specialization in world trade," *Journal of International Economics*, Vol. 54, pp. 75–96.

Katics, Michelle M., and Bruce C. Petersen, 1994, "The effect of rising import competition on market power: A panel data study of U.S. manufacturing," *Journal of Industrial Economics,* Vol. 42, September, pp. 277–286.

Klier, Thomas, and William Testa, 2002, "The Great Lakes border and economy," *Chicago Fed Letter*, Federal Reserve Bank of Chicago, No. 179a, July.

Ono, Yukako, 2001, "Outsourcing business services and the role of central administrative offices," Federal Reserve Bank of Chicago, working paper, No. 2002–01.

Testa, William, 1989, "Manufacturing's changeover to services in the Great Lakes economy," Federal Reserve Bank of Chicago, Regional Economic Issues, working paper, No. WP 21.

U.S. Department of Commerce, Bureau of the Census, 1999, *County Business Patterns*, 1989–99, Washington, DC.

U.S. Department of Commerce, International Trade Administration, Trade Development, Office of Trade and Economic Analysis, 2003, *U.S. Exports and Imports*, Washington, DC, January.

CALL FOR PAPERS

May 5–7, 2004

40th ANNUAL CONFERENCE ON BANK STRUCTURE AND COMPETITION
FEDERAL RESERVE BANK OF CHICAGO

How Do Banks Compete?

Strategy, Regulation, and Technology

The Federal Reserve Bank of Chicago invites the submission of research and policy-oriented papers for the 40th annual Conference on Bank Structure and Competition to be held May 5–7, 2004, at the Fairmont Hotel in Chicago. Since its inception, the conference has fostered an ongoing dialogue on current public policy issues affecting the financial services industry. In addition to papers related to the conference theme, we are interested in any high-quality research addressing public policy issues affecting financial services and will have a number of sessions on topics unrelated to the conference theme. We welcome submissions on all topics related to financial services and regulation.

The theme of the 2004 conference will address issues related to how banks compete. Over the past two decades, commercial banks have aggressively repositioned themselves to compete under new economic, technological, and regulatory conditions. No longer protected by regulatory entry barriers, and confronted with sea change advances in telecommunications and computer technology, banks are no longer able to rely on traditional banking models. Instead, banks and other financial institutions have invested huge amounts of resources in the search for new competitive strategies. While many of these attempts have been dead ends, the most successful strategic innovations have set new standards for the industry and have changed the way that banks compete. The manner in which commercial banks currently underwrite their loans, finance their activities, grow their franchises, distribute their services, and market their images would be barely recognizable to a banker from the 1970s.

Banks still do not compete in a completely unregulated environment, however, and regulations continue to shape banking strategies. For example, federally insured deposits are a cornerstone of the community bank business strategy. Community Reinvestment Act (CRA) loans are a requirement for any bank that wishes to grow by acquisition. Investment decisions—at least at the margin—are influenced by capital regulations. Our system of multiple regulators and bank chartering agencies can affect the

organizational form that banking companies choose. Terrorist threats and governance scandals have led regulators to make increased informational demands on banks. As banking markets grow more concentrated, anti-trust laws may increasingly limit the scale and scope of bank mergers. At a minimum, regulation is simply a fixed cost that must be borne by banks but does not influence bank behavior. At the other extreme, and perhaps more realistically, regulation can significantly affect banks' strategic choices and influence competition in financial markets. Innovations introduced in the marketplace are often driven by—and in some cases succeed exclusively because of—the existing regulatory environment.

Similarly, commercial banks' competitive strategies are shaped not only by new technologies, but also by the limitations of technology. Retail banking franchises have traditionally been built around paper-based payments, but information and communications technologies have created new strategic possibilities for retail banking. Electronic delivery of financial services can reduce banks' overhead costs, but abandoning bank branch offices can also give rise to disastrous strategic costs. New financial technologies have transformed risk-management at commercial banks, but application of leading-edge techniques may create unforeseen new risks. After generations of technological stasis in the banking industry, the rapid pace of technological change has made "strategic innovation" a viable competitive strategy for some banking companies. In this environment, must all banks become strategic innovators in order to survive, or can some banks remain competitive as strategic followers?

The 2004 Federal Reserve Bank of Chicago's Conference on Bank Structure and Competition will explore these issues. Additional financial topics that we are interested in evaluating include:

- The regulation and riskiness of government sponsored enterprises (GSEs),
- The Basel II Capital Accord,
- Financial industry consolidation,
- Payments innovations,
- Credit access: fair lending, CRA, and predatory lending issues,
- Deposit insurance and safety-net reform,
- Measuring, monitoring, and managing bank risk,
- The viability and future role of community banks, and
- Restructuring of financial regulatory agencies.

If you would like to present a paper at the conference, please submit four copies of the completed paper or a detailed abstract (the more complete the paper, the better) with your name, address, affiliation, telephone number, and e-mail address, and those of any co-authors, by December 22, 2003. Correspondence should be addressed to:

Conference on Bank Structure and Competition
Research Department
Federal Reserve Bank of Chicago
230 South LaSalle Street
Chicago, Illinois 60604-1413

For additional information contact:
Douglas Evanoff at 312-322-5814 (devanoff@frbchi.org),
Robert DeYoung at 312-322-5396 (robert.deyoung@frbchi.org),
or Regina Langston at 312-322-5641 (rlangston@frbchi.org).

Family resources and college enrollment

Bhashkar Mazumder

Introduction and summary

During the 1980s and early 1990s, the U.S. experienced a pronounced increase in income inequality. Associated with the rise in inequality has been a widening gap in earnings between those who have a college degree and those whose schooling ends in high school. According to census data, in 1975 men who completed four or more years of college earned 51 percent more than men who had completed four years of high school. The comparable figure in 2001 was 122 percent.[1] So, on average, college graduates now earn more than double what high school graduates earn.

Why has attending college become so much more important? Many economists argue that as the economy has become more technologically sophisticated, employers simply require a more educated and skilled work force. The rising demand for skilled workers has outpaced the increase in supply, resulting in a sizable premium for college-educated workers.

College attendance is an important issue for other reasons in addition to the growth in income inequality. Clearly, a more educated work force should enhance the productive capacity of the economy and promote faster economic growth (Aaronson and Sullivan, 2001). There are also likely to be important social externalities to promoting greater college attendance, such as greater involvement in the duties and responsibilities of citizenship (for example, higher voting rates). Finally, greater access to college might help foster greater *intergenerational income mobility*, namely a child's ability to achieve economic success irrespective of their parents' economic circumstances. Recent studies have shown that on average, at least 40 percent, and perhaps as much as 60 percent of the earnings differences between families persist from one generation to another (Bowles and Gintis, 2002). Clearly, any policies that might be successful at bridging the divide in educational attainment and, thereby, reduce earnings differences might also help reduce the persistence in income inequality over generations.

For these reasons, policymakers are interested in what determines college enrollment and completion and how best to promote higher education. This is a particularly salient issue now, given the current fiscal problems facing the federal and state governments, which have already led to cutbacks in financial support for higher education.

An analysis of national trends in college enrollment shows that overall college enrollment among young adults has risen steadily over the last 30 years. However, only about 35 percent of 18–24 year olds currently attend college. There is currently a major divide in college attainment by race and ethnic group. In fact, these gaps are higher today than they were 25 years ago. The sharp differences in college enrollment rates suggests that perhaps the key factors underlying these trends are economic variables such as family income and college costs. Indeed, an examination of enrollment levels by income level appears to bear this out. Adolescents from families in the lowest income strata are far less likely to attend college than their better-off peers.

However, the idea that family income and tuition costs largely explain enrollment patterns is not as clear cut as it might appear at first glance. There are many different types of colleges with a wide range of costs, and there are many potential sources of financial aid and loan programs. Indeed, it is not unreasonable to speculate that anyone who truly wants to attend some

Bhashkar Mazumder is an economist in the Economic Research Department and the executive director of the Chicago Census Research Data Center at the Federal Reserve Bank of Chicago. *The author thanks Siopo Pat, Kate Anderson, and David Oppedahl for their research assistance. He also thanks Dan Aaronson and Dan Sullivan for helpful discussions and comments.*

type of college can find a way to finance it. Traditional economic theory suggests that in the absence of market imperfections such as borrowing constraints, those who find it *optimal* to invest in their human capital through postsecondary schooling will in fact do so, irrespective of their family's current income level. The key determinants in this model are the expected financial returns to attending college, the interest rate, and the costs of attending college.

The fact that existing government financial aid and loan programs do not cover the full costs of going to college suggests that the existence of borrowing constraints is certainly plausible (Keane and Wolpin, 2001). Whether individuals actually do not enroll in college because of the inability to borrow is a point of contention in the economics literature. While many studies have found that there is a strong association between family income and college enrollment, Cameron and Heckman (2001) argue that this is because family income captures the *long-run* factors that determine whether an individual has the prerequisite skills to be successful in college. They argue that there is very little role for policies such as college subsidies that influence the *short-term* financing considerations of attending college.

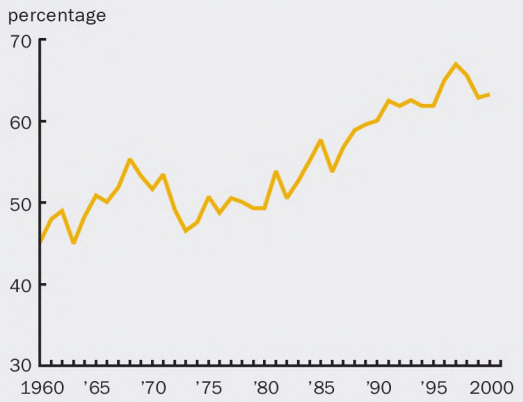
Various other studies (for example, Kane, 1994; Dynarski, 2003) find either that college costs are an important factor or that college subsidies have an important effect on enrollment. While a sensitivity to price is not what economists would call "borrowing constraints," it does imply a potential role for public policy in subsidizing college costs for those on the margin of attending, particularly if there are important social benefits to increasing college enrollments. In fact, there is some common ground in this literature, in that all of these studies find that an increase in college costs of $1,000 in 2001 dollars is typically found to translate into a decline in enrollment of about 4 percentage points. On the other hand, it is not at all clear whether lowering college costs would reduce the *disparities* in enrollment across income or racial groups.

Interestingly, none of the studies in the literature investigate the empirical importance of family *wealth* as opposed to *income* to college attendance. The omission of wealth in the literature is no doubt due to the fact that the survey data used by previous researchers do not contain very good information, *if any*, on families' assets and liabilities. This is an important omission since for many families, a sizable fraction of college expenses are covered by longer-term savings reflected in financial assets. Families with high levels of wealth are much less likely to be borrowing constrained. One might expect that families with more wealth are better able to borrow against their assets. Therefore, data on wealth would seem to be particularly useful for testing the borrowing constraints hypothesis implied by theoretical models. In addition, financial assets are an important part of most financial aid formulas, so higher wealth can potentially lead to higher college costs *net* of this aid and possibly lower enrollment levels, all else equal.

This article begins to address this gap in the literature by using a data source that has highly detailed information on family assets and liabilities, as well as information on the enrollment decisions of adolescents. A preliminary empirical investigation of this data offers some suggestive evidence that income might be an especially important factor for families who have modest amounts of wealth. This may be due to some combination of borrowing constraints and higher actual costs due to lower financial aid. Certainly, this evidence suggests that further investigation of the role of wealth in college enrollment is in order.

FIGURE 1

College enrollment over time



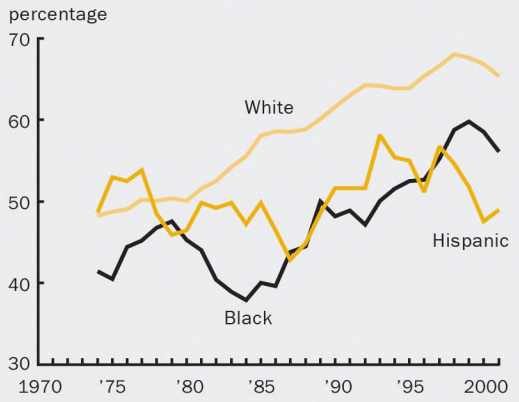
Source: U.S. Department of Education, National Center for Education Statistics (2003), *Digest of Economic Statistics*.

Trends in college enrollment

In recent decades there has been a clear upward trend in the percentage of high school graduates between the ages of 16 and 24 who enroll in college within a year of finishing high school, according to data assembled by the National Center for Educational Statistics.[2] As figure 1 demonstrates, from 1960 until the 1980s, the percentage enrolled in college fluctuated around 50 percent. Since 1980, however, the rate has risen sharply from 49 percent to 62 percent in 2001, reaching a peak of 67 percent in 1997. The rise has been slightly more pronounced among women,

FIGURE 2

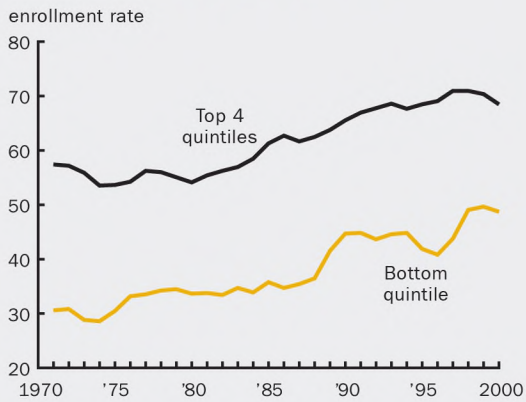
College enrollment rates by race



Source: U.S. Department of Education, National Center for Education Statistics (2003), *Digest of Economic Statistics*.

FIGURE 3

College enrollment, bottom quintile vs. top four quintiles



Source: U.S. Department of Education, National Center for Education Statistics (2003), *Digest of Economic Statistics*.

whose enrollment rate briefly eclipsed 70 percent in the late 1990s.

These figures, however, paint an overly positive picture of college attendance because they only show the rates among those *16–24, who finished high school within the last year*. In contrast, the college enrollment rate among *all 18–24* year olds in 2001 was just 36 percent. While this is still a significant improvement over the 25 percent rate recorded in 1979, despite recent positive trends, college enrollment remains more the exception than the rule.

The gap in enrollment rates between whites and minorities has been a focal point of some recent studies on college enrollment (for example, Kane, 1994; Cameron and Heckman, 2001). Figure 2 shows the enrollment rates among recent high school completers aged 16 to 24 across racial/ethnic groups. (Three-year moving averages are shown so as to reduce the large sampling variance in the survey data.) The difference in the enrollment rates between blacks and whites was only a few percentage points in the late 1970s but surged in the 1980s, reaching a peak of 19 percentage points in the mid-1980s. This sharp rise spurred the debate over the impact of economic factors, such as rising college costs, declining financial aid, and slow real income growth on college enrollments. This was also a motivating factor for studies that used econometric models to understand more broadly the determinants of the propensity to attend college.

In the late 1980s and through most of the 1990s, the black–white gap progressively narrowed, falling back into the single digits. However, since 1998, black enrollment rates have fallen in each year, and the racial gap has begun to widen once more. The enrollment gap between whites and non-white Hispanics is actually larger and has widened considerably since the 1970s.

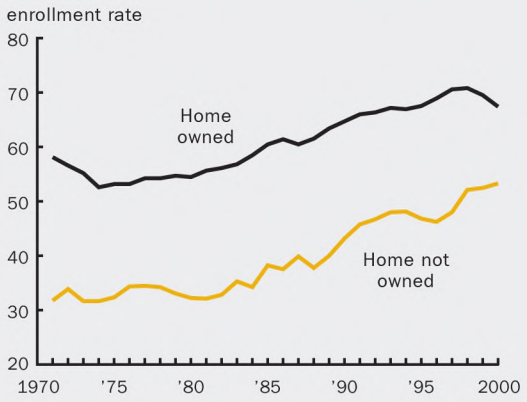
An important question is to what extent these minority enrollment gaps are merely reflecting disparities in enrollment by income level that can be addressed by tuition subsidies targeted to low-income families. Figure 3 compares the enrollment rate of the bottom income quintile versus the top four quintiles using data from the October *Current Population Surveys* (CPS) conducted by the Census Bureau. This chart illustrates that enrollment rates have risen even among families at the bottom of the distribution, but that the gap in enrollment with other families has narrowed only slightly over the last 30 years. This evidence certainly fits a story that emphasizes income differences as a critical factor in college enrollments.

While the CPS surveys typically used by researchers to investigate enrollment patterns do not collect information on wealth, they do collect information on homeowner status. Since housing equity is often the largest share of a family's wealth, tracking enrollment rates by family homeownership might offer a glimpse as to the importance of wealth considerations. Figure 4 shows that historically there has been a large gap in enrollment rates by homeownership status but that this gap has narrowed quite a bit in recent years.

These figures suggest that while progress has been made in achieving higher rates of college enrollment among young adults, the disparities by race and income are wider today than they were 25 years ago. Looking forward, the current fiscal problems facing many state governments are expected to lead to large cuts

FIGURE 4

College enrollment by family home ownership



Source: Author's calculations using October CPS supplements.

in college subsidies and tuition increases at public colleges, which raises the prospect of a further widening of these gaps in higher education. However, these predictions depend critically on the extent to which short-term financial considerations actually influence the propensity to attend college.

Evidence from a sample of econometric studies

The literature on college financing is large and cannot be given a thorough treatment here. I discuss a small sample of recent studies to provide a general sense of how economists have approached this question and their results.

As with many research questions in economics, it is risky to rely exclusively on data that are based on changes over time in aggregate statistics in order to identify behavioral patterns such as those shown in the last section. Economies are constantly in flux, with many variables changing simultaneously. For example, the causal relationship between college costs and enrollment rates may be difficult to discern from "time-series" data. In the 1980s, both variables were increasing, but it is unlikely that an increase in tuition could lead to an increase in enrollment. Aggregate enrollment was probably also influenced by other economic incentives, such as the rising payoff to attending college.

Therefore, economists have estimated econometric models using micro-level data on individuals and their enrollment decisions at a point in time to infer the underlying behavioral relationships that are typically obscured in the national data. These "cross-sectional" studies have generally found that college costs and family income have a statistically significant and economically important effect on enrollment decisions. In a review of a number of studies predating 1990, Leslie and Brinkman (1989) argue that a consensus view is that a $1,000 (2001 dollars) increase in net college costs results in about a 4 percentage point decline in the probability of enrollment.

A more recent study by Kane (1994), which examines the decline and subsequent rise in the black college enrollment rate during the 1980s, uses data from the October CPS and includes a wide range of variables such as parental educational attainment, family income, homeownership, and local labor market conditions. Kane studies the effects of these variables separately for blacks and whites and by income quartiles. He also controls for state "fixed effects," thereby correcting for the potential problem that states with low tuition levels might support enrollment in other ways. Kane concludes that college costs exerted downward pressure on the enrollment rate for blacks in *all* income groups. Kane speculates that the sensitivity of even *high-income* black families to college costs might be explained by the fact that despite their high income, these families have little wealth and, therefore, might also be constrained from borrowing. Given the lack of data on wealth in Kane's sample, he cannot pursue this further.

Overall, Kane finds that a $1,000 (2001 dollars) change in tuition costs lowers the probability of enrollment by around 4 percentage points. However, he finds that these costs explain only about one-third of the drop in enrollment for blacks during the first half of the 1980s and that most of the rest of the decline cannot be explained by his model. One somewhat puzzling finding is that Pell Grant eligibility appears to have a negligible effect on college enrollment. Pell Grants are a federal means-tested program that provides grants to qualified students for postsecondary education. An earlier study based on aggregate time-series data by Hansen (1983) also showed little effect of the program on enrollment levels. Kane speculates that his finding may be due in part to measurement error, since Pell Grant eligibility is estimated based on available survey data. He also suggests that perhaps low-income students are less aware of their eligibility for the program. Nonetheless, the lack of any strong effect of Pell Grants on enrollments is a reason to remain somewhat skeptical about the effectiveness of tuition subsidies.

While cross-sectional studies such as Kane's avoid some of the pitfalls of time-series analysis, they are also subject to other potential deficiencies such as omitted variables and measurement error. The lack of a good measure of scholastic preparedness for college is a particular issue of concern. If the ability to succeed in

college is the key determinant of college enrollment but there is no good measure of this "ability" in the data (for example, test scores) and if family income is highly correlated with ability, then a cross-sectional analysis might mistakenly overemphasize the importance of family income.

This problem and other similar issues have led researchers to pursue alternative approaches to studying the issue. Cameron and Heckman (2001) exploit longitudinal data—repeated observations on the same individuals—to estimate a dynamic model of educational attainment. Through this approach they not only examine college enrollment but also analyze grade transitions prior to college enrollment, where financial considerations ought not to be as important. As part of their statistical model, they also directly incorporate heterogeneous ability. Perhaps most importantly, they use the *National Longitudinal Survey of Youth* (NLSY), a comprehensive dataset that contains not only all of the relevant variables typically used by researchers, but also a measure of scholastic ability, the Armed Forces Qualifying Test (AFQT).

The AFQT is part of the Armed Services Vocational Aptitude Battery (ASVAB) given to applicants to the U.S. military. The ASVAB consists of ten tests. The AFQT score is based on four of the tests that focus on reading skills and numeracy. The AFQT is a general measure of trainability in the military and is a primary criterion for enlistment eligibility. The test was administered to nearly all respondents in the NLSY in 1980 in order to provide new norms for the test based on a nationally representative sample. The AFQT is not viewed by the military or by most researchers as a measure of general intelligence or IQ. Indeed, it is well known that scores rise with additional years of schooling, so researchers typically use scores that are age-adjusted. Cameron and Heckman's sample does not include anyone who took the test after entering college.

Cameron and Heckman estimate their dynamic educational attainment model separately for whites, blacks, and Hispanics and estimate the probabilities of completing ninth grade by age 15; completing high school by age 24; and enrolling in college. The model is run both including and excluding AFQT scores. They use the results of the models to perform the following thought experiment: How much of the white–minority gaps would be eliminated if for each explanatory variable, blacks and Hispanics were assigned the same average values as whites. Using the model results without AFQT scores, they find that equating family income would reduce the expected gap in college enrollments by roughly half. However, they also find that simply equating other family background variables, such as parent education and family size, has an even larger effect on reducing these gaps. When they include AFQT scores, equating this variable alone more than eliminates the entire enrollment gap for both blacks and Hispanics, while income has virtually no independent effect.

Based on this result, they argue that college preparedness is the critical determinant of college enrollment and not any kind of short-term borrowing constraint. This conclusion is also bolstered by their finding that family income has an important effect on grade advancement only at earlier stages in a student's educational career (for example, reaching ninth grade by age 15), when short-term financing issues are presumed to be irrelevant.

While these results appear to be very strong and make a compelling case against the existence of borrowing constraints, they are still not fully satisfying. How is it that white and minority enrollment trends could diverge so rapidly in the early 1980s only to be followed by a period of rapid convergence later in the decade as figure 2 shows? It is possible that there were rapid and sudden shifts in minority college preparedness. But there is no evidence of this in test scores. So while the results appear to present repudiation of the idea that family income during the college-going years matters, the study does not provide a fully persuasive story to explain the trends in the data that motivated the model.

With regard to the broader question of whether public policy ought to subsidize college education, these results actually could be considered to provide some evidence in favor of such a policy. A common criticism of broad-based college subsidies is that they simply subsidize the costs of middle-class families, whose children would have enrolled in college anyway. Cameron and Heckman's results show that college enrollment is sensitive to tuition costs, so that lowering the costs for targeted families might turn out to be an effective policy. This is particularly true for two-year colleges. The authors estimate that a $1,000 (2001 dollars) increase in tuition at two-year colleges lowers black enrollments in two-year and four-year colleges *combined*, by 4 percentage points. For Hispanics, the decline is even larger, at 8 percentage points. Although college enrollments are less sensitive to changes in tuition at four-year colleges, the effect of a $1,000 (2001 dollars) increase in costs at *both* two- and four-year colleges would lower white enrollment by 5 percent—a figure right in line with the results of the previous studies. Finally, the study does not address the possibility that wealth may be a critical factor in determining the likelihood of enrollment, which is the question I turn to in the next section.

Each of the studies so far described exploits the observed variation in a number of variables (for example, enrollment or family income) across a sample of the population to infer the basic statistical relationships under certain simplifying assumptions. This approach can lead to misleading inferences about causality if there are other factors that are not captured by the statistical model. In an ideal setting researchers would prefer to design an experiment where individuals could be *randomly assigned* different levels of family income or tuition costs. Differences in enrollment rates between the treatment and control groups would reveal the behavioral responses. Randomization would eliminate the need to have a full set of control variables. Of course, in the real world, such experiments are close to impossible. In recent years, however, economists have increasingly employed research strategies that take advantage of real world situations that mimic random assignment. These "quasi-experiments" allow researchers to infer behavioral relationships that might otherwise be difficult to identify through standard statistical models.

Dynarski (2003) provides one such example in a study of the effects of a particular tuition subsidy on college enrollment. In 1982, Congress eliminated the Social Security student benefit program that offered monthly financial support to full-time students whose parents were deceased, disabled, or retired. Dynarski uses the NLSY to implement a quasi-experimental design that compares the college enrollments of those who were eligible for the aid due to the death of a parent before the program was eliminated with a later cohort who would have been eligible for the program had it not been eliminated. The enrollment probability of those with a deceased parent fell by more than 20 percentage points compared with a drop of just 2 percentage points for the rest of the sample. Incorporating figures on the size of the program's benefit and the costs of tuition, Dynarski calculates that a $1,000 (2001 dollars) increase in aid increases enrollment by nearly 4 percentage points.

While Dynarski's results are in line with much of the previous literature, the quasi-experimental design of the study makes it more credible than those of standard cross-sectional studies. The quasi-experimental design, however, still has some drawbacks. It is difficult to know if the behavioral response that is estimated from the subgroup of the population affected by the legislative change, generalizes to the broader population.

The findings of Cameron and Heckman and other research not discussed here[3] makes many economists skeptical that borrowing constraints are a critical factor in limiting college enrollments. Indeed in a more recent paper, Carneiro and Heckman (2003) estimate that only about 8 percent of the population faces borrowing constraints to attending college. Still, there appears to be reasonably strong evidence that public policy can influence enrollment levels.

In any case, there are several issues that deserve more attention in future research. The first, which I address below, is examining the role of wealth. It might be the case that, for example, the sharply lower wealth levels of blacks has been a major impediment to college attendance. In fact, the economic literature on consumption has often used levels of wealth to detect the presence of borrowing constraints among low-wealth families (for example, Zeldes, 1989).

A second question, which has not been examined thoroughly, is the extent to which financial resources and costs affect college *completion*.[4] Perhaps the access to college financing is available, but over time financial difficulties overwhelm some families and prevent college completion. Finally, to what extent do financial resources affect the kind of school or quality of school one attends? There is growing evidence that fewer low-income students are attending private universities and four-year colleges (McPherson and Schapiro, 1998). Therefore, there is reason to believe that there is not only a college enrollment gap but there are also likely to be disparities in educational quality.

Wealth and college enrollment

This article begins to address one of the shortcomings in the literature by using a data source that has been neglected in the existing literature. The Census Bureau's *Survey of Income and Program Participation* (SIPP) contains extremely detailed data on assets and liabilities in addition to the full set of variables that have typically been used to study the determinants of college enrollment. The SIPP surveys began in 1984 and are two- to three-year panels that allow for multiple measurements of all the variables of interest. The SIPP surveys approximately 20,000 households every four months on income, labor market activity, and participation in a wide range of federal government programs, such as food stamps and Social Security.

The surveys also ask about school enrollment and sources of financial assistance. Special topical modules once a year collect information on housing equity, vehicle equity, business equity, a range of financial assets, unsecured debt, real estate property, individual retirement accounts (IRA), and other retirement plans. The panel aspect of the data enables one to construct a sample of 11th and 12th graders and determine college enrollment over the next two years.

I estimate a linear probability model (ordinary least squares—OLS) of the likelihood of enrollment.[5] The dependent variable is equal to 1 if a 12th grader begins college by the following school year and 0 otherwise. Similarly the variable is set to 1 if an 11th grader starts college two years later and 0 otherwise. I pool the 1984, 1985, 1986, 1987, and 1990 SIPP panels and use both men and women. The sample for which all the key information, including log wealth, is available is 4,123. Of these, about 37 percent enrolled in college.

The description of the sample is given in table 1. The key explanatory variables that are the focus of this study are family income, tuition costs, and wealth. Family income is averaged over the two calendar years that are available in each of the SIPPs and includes earnings from up to two jobs, two businesses, and any income from other sources. Since ideally I want to measure tuition for those at the margin of attending college, I opt for two-year colleges. Tuition costs are measured by using the average tuition at two-year colleges in the individual's state of residence.[6] Unlike Cameron and Heckman (2001), I cannot measure this at the county level so there is likely to be considerable measurement error. In this analysis, I have not adjusted tuition for Pell Grant eligibility as some previous studies have done.

I use three different wealth variables, since it is not clear *a priori* what the appropriate measure ought to be. First, I consider housing equity, since this is the largest share of wealth for many families. Second, I construct a measure of liquid assets (for example, bank accounts, stocks, bonds) that might better capture the financial resources readily available to the family. The third measure I consider is net worth, which is a summary measure that incorporates a large array of assets and liabilities. A problem with wealth data is that the non-reporting for some variables can be sizable, so many values are imputed by the Census Bureau. As an additional check, I limit analysis to data that is not imputed, though this reduces the sample size.

Figure 5 shows how college enrollment differs by quartiles of family income and the three measures of wealth. It is immediately striking that the wealth measures do not appear to be appreciably different from each other in terms of how they affect enrollment at least unconditionally. Housing equity appears to show the smallest differences across the quartiles. Liquid assets shows the most striking difference between the first and second quartiles, while net worth looks closest to family income. I chose to use net worth, since it is the broadest measure and since the results are not much affected by the alternatives.

To the extent possible, I follow Kane (1994) and Cameron and Heckman (2001) in the choice of other covariates. These include family size, father's years of education, mother's years of education, black indicator, female indicator, indicator for whether a parent has a long spell of unemployment, state, and year effects. For measures of the local labor market, I use the unemployment rate and the average wage for those with a high school degree. Wherever possible these are both measured at the metropolitan statistical area level, otherwise they are measured at the state level. Again, compared with the county level measures used by Cameron and Heckman, my measures are likely to suffer from measurement error.

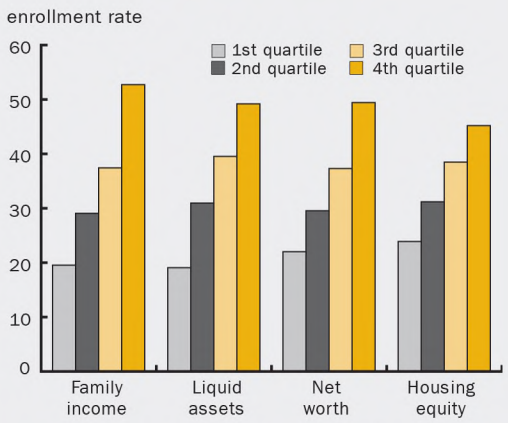
TABLE 1

Summary statistics

Variable	Mean	Standard deviation	Minimum	Maximum
Enrolled in college	0.37	0.48	0	1
Log family income	10.41	0.82	0	12.70
Family size	2.94	0.57	2	6
Father's years of education	10.47	5.84	0	18
No father identified	0.19	0.39	0	1
Mother's years of education	11.19	4.52	0	18
No mother identified	0.10	0.30	0	1
Female	0.51	0.50	0	1
Black	0.11	0.31	0	1
Parent unemployed > 3 months	0.25	0.43	0	1
Local area unemployment rate	0.07	0.02	0	0.19
Local wage for high-school grad (1984$)	7.94	0.70	5.73	10.16
Tuition (1984$)	741	376	30	1,641
Net worth (1984$)	92,463	117,158	38	1,285,442
Housing equity (1984$)	46,562	45,483	-2,385	251,519
Liquid wealth (1984$)	15,853	48,232	0	1,010,100
Sample size	4,123			

FIGURE 5

Enrollment rates by quartiles of family resources



Source: Author's calculations using 1984, 1985, 1986, 1987, 1990 SIPP panels.

The major limitation of the data, however, is that for most years they do not contain information on scholastic ability such as test scores. Therefore, the analysis is subject to Cameron and Heckman's criticism that other variables such as income and wealth may pick up the effects of this omitted variable. On the other hand, this dataset does contain information on the wealth of the parents, which is a critical omission in the NLSY, so the reverse criticism could be made of the existing studies.

There are several hypotheses one might make about how wealth could influence enrollment. First, one might simply imagine that wealth has a direct effect on the probability of attending college. Imagine two families with similar income but one has substantially larger assets to draw from. If we thought that an extra dollar of wealth simply acts the same way as an extra dollar of income, a reasonable first step would be to model wealth the same way as income and assume a linear relationship.

However, there are several reasons to think that the effects of wealth are nonlinear. One reason is that wealth might serve simply as an indicator of borrowing constraints. If there are market imperfections that prevent students from borrowing from their expected future income, they may be forced to rely on parents' wealth either directly or as a form of collateral. In this simple case, we might expect that additional financial resources, either income or wealth, might be important, but only for families below a certain threshold of wealth, for example, the bottom quartile of the wealth distribution.

However, if scholastic ability is a critical factor in determining college enrollment as Cameron and Heckman (2001) show, and if it is correlated with parents' wealth, then the story becomes more complicated. At the low end of the wealth distribution there might be very few families who would actually benefit from greater financial resources due to low levels of academic preparedness. It might be that as we move up the wealth distribution, there are more families for whom additional financial resources might matter. At some point along the wealth distribution, of course, families have sufficient financial resources and the effect might dissipate. In this case financial resources might matter most for families in the middle of the distribution. Corak and Heisz (1999) reported this kind of finding in their study of nonlinearities in intergenerational mobility using Canadian data.

A second reason that wealth might have a nonlinear effect is that it is typically an important variable in financial aid formulas used by colleges and universities, as well as government aid programs. In this case, greater wealth might actually increase the costs of college attendance over a particular range of the wealth distribution. This might produce a more complicated pattern, where income matters the most for families with modest amounts of wealth.

I use two simple approaches to estimate these potential nonlinear effects. First, I simply include indicator variables for quartiles of the wealth distribution. This tests whether the *direct* effects of wealth on enrollment have a nonlinear pattern. It allows us to see whether wealth matters most going from say the bottom quartile to the second quartile. Second, I stratify the sample by levels of wealth to see whether the effects of family income or college costs matter at a particular point of the wealth distribution as hypothesized above. This might help identify whether there is a particular point in the wealth distribution where borrowing constraints might bind and make income particularly important.

The first set of results is shown in table 2. In the first column, the results are shown without including any wealth measures and with no state effects. Here nearly all the coefficients are of the expected sign. The coefficient on log family income is .04 and is highly significant. Parent education is positive and significant. Women are slightly more likely to enroll in college and blacks are about 6 percentage points less likely to enroll even conditioning on these other variables. The one unexpected result is tuition, which has a positive sign. The lack of good geographic detail on tuition is probably the explanation. Local labor market conditions do not appear to be significant. The

addition of state effects appears to make no difference to the results (not shown) and does not improve the performance of the tuition measure.

In column 2, I add log of net worth to the model. This measure of wealth is significant with a coefficient of .02. Adding net worth lowers the coefficient on family income by about one-quarter to .032. Interestingly, most of the difference between whites and blacks is now eliminated.

In column 3, I take a simple approach toward estimating nonlinearities in wealth by using indicator variables for being in a particular quartile of the net worth distribution. I use the first quartile as a basis for comparison. After controlling for other covariates, being in the second quartile of net worth raises the probability of enrollment by only 3 percentage points. The larger jumps take place at the top 2 quartiles. I find a similar pattern when using housing equity or liquid assets instead of net worth (not shown). This provides suggestive evidence of nonlinearities in wealth. It appears from this evidence that having above median wealth is the critical threshold to overcome.

Finally in table 3, I test directly whether family resources are sensitive at particular points in the wealth distribution. Here the exercise is to stratify the sample by quartiles of net worth and compare the coefficients on family income. For quartile 1, the effects of family income are relatively small and only marginally statistically significant. Interestingly, the gap with blacks is small and statistically insignificant while the female enrollment advantage is quite a bit higher. In the second quartile of net worth, there is a dramatic rise in the importance of family income—the coefficient is .07 and highly statistically significant. Income appears

TABLE 2

The effects of adding net worth

Regression results where dependent variable is college enrollment

	1	2	3
Log family income	0.041 (0.008)	0.032 (0.011)	0.029 (0.008)
Family size	–0.036 (0.015)	–0.041 (0.017)	–0.034 (0.015)
Dad's education	0.025 (0.003)	0.025 (0.003)	0.023 (0.003)
Mom's education	0.024 (0.003)	0.024 (0.003)	0.022 (0.003)
Female	0.038 (0.013)	0.035 (0.014)	0.038 (0.013)
Black	–0.058 (0.020)	–0.026 (0.024)	–0.032 (0.021)
Parent unemployed	–0.019 (0.020)	–0.009 (0.022)	–0.014 (0.020)
Local unemployment rate	0.694 (0.378)	0.795 (0.408)	0.753 (0.377)
Local wage for high-school grad	0.011 (0.010)	0.007 (0.010)	0.004 (0.010)
State tuition	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Log net worth	—	0.021 (0.005)	—
Net worth quartile 2			0.034 (0.019)
Net worth quartile 3			0.083 (0.020)
Net worth quartile 4			0.135 (0.021)
Sample size	4676	4123	4676
R-squared	0.125	0.128	0.133

Note: Standard errors in parentheses.

to be twice as important in this range of wealth compared with the sample overall and three to four times as important compared with the lowest wealth quartile. In fact, for this group neither gender nor race appears to have any effect on enrollment rates. For the third quartile, the income effects are similar to what was estimated for the full sample in table 2. For the fourth quartile, as we might expect, income matters much less. In the upper half of the wealth distribution the black–white gap is only marginally significant.

What should we take away from this exercise? The results in table 3 raise the tantalizing possibility that there might, in fact, be a group of families for whom income matters and for whom financial aid or subsidies might promote college attendance. These are not the poorest families, but actually have wealth between the 25th and 50th percentiles. One hypothesis for this finding is that the children of families in the second wealth quartile have sufficient capability to perform well in college but that they do not enroll (at least not right away) because of insufficient financial resources. Under this view, income does not explain the enrollment rate for the poorest group of families (bottom quartile), because they are also the least likely to have children with the capability to succeed, so they would not have enrolled even with additional financial resources.

An alternative explanation for the importance of income for families in the second quartile of the wealth distribution is the extensive use of financial aid formulas in determining college costs. This formula essentially acts as a tax on wealth. Families with little or no wealth are unaffected. However, families with some, but not a lot, of wealth will face higher college costs. Since I do not measure the true *net* costs faced by families, this sensitivity is captured by family income. As we move higher in the wealth distribution, however, the penalty no longer matters since the wealthiest families are ineligible for aid. This makes additional income less important for families in the top two quartiles.

Further analysis

Additional research with other datasets may be necessary to validate these results. It would be useful to know whether this pattern of higher income sensitivity at the second quartile of wealth also affects earlier grade transitions, where we would not expect wealth to matter.

TABLE 3

The effect of income by quartiles of wealth

Regression results where dependent variable is college enrollment
(Samples are stratified by quartiles of the net worth distribution)

	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Family income	0.020	0.074	0.034	0.017
	(0.011)	(0.022)	(0.024)	(0.019)
Family size	–0.042	–0.058	–0.031	–0.012
	(0.025)	(0.029)	(0.035)	(0.037)
Dad's education	0.011	0.017	0.037	0.021
	(0.005)	(0.005)	(0.006)	(0.006)
Mom's education	0.017	0.026	0.014	0.034
	(0.005)	(0.007)	(0.007)	(0.007)
Female	0.072	0.011	0.024	0.037
	(0.024)	(0.026)	(0.027)	(0.028)
Black	–0.023	0.008	–0.086	–0.184
	(0.028)	(0.037)	(0.055)	(0.105)
Parent unemployed	–0.017	0.046	–0.028	–0.072
	(0.035)	(0.038)	(0.042)	(0.051)
Local unemployment rate	0.974	–0.253	0.978	1.598
	(0.738)	(0.736)	(0.740)	(0.823)
Local wage for high-school grad	0.018	–0.014	–0.003	0.018
	(0.017)	(0.020)	(0.021)	(0.020)
State tuition	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Sample size	1153	1169	1169	1159
R-squared	0.064	0.098	0.130	0.104

Note: Standard errors in parentheses.

It would also be interesting to see if these effects still hold up in other datasets where it is possible to control for ability by using test scores. Still, the findings here ought to prompt researchers to consider the possibility that all family resources, including wealth, should be analyzed.

Conclusion

The growing gap in earnings between college graduates and non-graduates has become an important feature of the economy. Promoting greater college enrollment might not only address the current earnings gap but also offer the potential to improve economic mobility for future generations. Other potential societal benefits include a more productive economy and a better-informed citizenry.

To date, economic research has produced only mixed findings for policymakers who wish to promote college enrollment for disadvantaged youth through greater access to financial resources. While there is some skepticism as to whether a large number of families are actually "borrowing constrained," there is more agreement that lower tuition costs and greater financial aid do appear to affect enrollment. Whether these policies will narrow the gaps in enrollment by race, ethnicity or income level is less clear.

Most studies, however, have neglected the potential role of wealth. The preliminary analysis here suggests that incorporating wealth might be a promising avenue for better identifying borrowing constrained families for whom additional financial resources might matter. Income appears to have a very large effect for families in the second quartile of the net worth distribution. Arguably, it is in these families that children are academically prepared for college but for whom additional financial resources make a big difference. This is an especially important area for further analysis, given the vast and growing educational divide.

NOTES

[1]This is based on Census Historical Income Tables, P32 and P35 available at the Census website at www.census.gov/hhes/income/histinc/incperdet.html. These figures are for men aged 35–44 who worked full-time and year round. The figures do not adjust for variables such as hours worked and work experience that are typically used by economists to estimate the "return to education" using a regression model.

[2]This is taken from U.S. Department of Education, National Center for Educational Statistics (2003), table 183.

[3]These include Cameron and Taber (2004) and Keane and Wolpin (2001).

[4]Dynarski (2003) and Carneiro and Heckman (2003) are exceptions to this.

[5]Using probit models produce exactly the same qualitative results. The coefficients from a regression produce results that are easily interpreted at any point of the distribution of the covariates.

[6]Data was provided by the Washington State Higher Education Group.

REFERENCES

Aaronson, Daniel, and Daniel Sullivan, 2001, "Growth in worker quality" *Economic Perspectives*, Vol. 25, No. 4, pp. 53–74.

Bowles, Samuel, and Herbert Gintis, 2002, "The inheritance of inequality," *Journal of Economic Perspectives*, Vol. 16, No. 3, pp. 3–30.

Cameron, Stephen V., and James J. Heckman, 2001, "The dynamics of educational attainment for black, Hispanic, and white males," *Journal of Political Economy*, Vol. 109, No. 3, pp. 455–499.

Cameron, Stephen V., and Christopher Taber, 2004, "Estimation of educational borrowing constraints using returns to schooling," *Journal of Political Economy*, forthcoming.

Carneiro, Pedro, and James J. Heckman, 2003, "Human capital policy," National Bureau of Economic Research, Cambridge MA, working paper, No. 9495.

Corak, Miles, and Andrew Heisz, 1999, "The intergenerational earnings and income mobility of Canadian men: Evidence from longitudinal income tax data," *Journal of Human Resources*, Vol. 34, No. 3, pp. 504–533.

Dynarski, Susan, 2003, "Does aid matter? Measuring the effect of student aid on college attendance and completion," *American Economic Review*, Vol. 93, No. 1, pp. 279–288.

Hansen, W. Lee, 1983, "Impact of student financial aid on access," in *The Crisis in Higher Education*, Joseph Froomkin (ed.), New York: Academy of Political Science.

Heckman, James J., Lance Lochner, and Christopher Taber, 1998, "General equilibrium treatment effects: A study of tuition policy," *American Economic Review*, Vol. 88, No. 2, pp. 381–386.

Kane, Thomas J., 1994, "College entry by blacks since 1970: The role of college costs, family background, and the returns to education," *Journal of Political Economy*, Vol. 102, No. 5, pp. 878–911.

Keane, Michael P., and Kenneth I. Wolpin, 2001, "The effect of parental transfers and borrowing constraints on educational attainment," *International Economic Review*, Vol. 42, No. 4, pp. 1051–1103.

Leslie, Larry L., and Paul T. Brinkman, 1989, *The Economic Value of Higher Education*, New York: Macmillan (for the American Council on Education).

McPherson, Michael S., and Morton Owen Schapiro, 1998, *The Student Aid Game: Meeting Need and Rewarding Talent in Higher Education*, Princeton, NJ: Princeton University Press.

U.S. Department of Education, National Center for Education Statistics, 2003, *Digest of Economic Statistics 2002*, Washington, DC.

Zeldes, Stephen P., 1989, "Consumption and liquidity constraints: An empirical investigation," *Journal of*

An introduction to the WTO and GATT

Meredith A. Crowley

Since its inception in 1995, the World Trade Organization (WTO) has regularly been in the news. There have been optimistic stories of expanding WTO membership that emphasize that freer trade generates numerous benefits for consumers. Newspapers report on the details of WTO entry negotiations for important countries like China and remind us of the gains from trade. At other times, media reports might lead us to believe that disputes among WTO members are about to tear the organization apart. Disagreements between the U.S. and the European Union (EU) over everything from U.S. corporate taxation, to genetically modified organisms, to special steel tariffs make headlines worldwide. Finally, some groups seem unconvinced by and resentful of claims that free trade makes the entire world better off. Huge numbers of people from environmental and labor groups gather at various international meetings of heads of state and government ministers to protest globalization in general and the WTO in particular. Some representatives of developing countries are concerned that they have liberalized their trade and agreed to intellectual property protection for developed country products but have received almost no additional access to agricultural markets in the industrialized world.

What are we to make of all this? What is the WTO? What is it trying to accomplish and why? How does the world trading system function? Why are there so many disputes among countries that belong to the WTO?

This article provides an overview of the General Agreement on Tariffs and Trade, better known as GATT, and the WTO system. In the first section, I present a brief history of GATT and the WTO. In the following section, I discuss the fundamental principles that underlie the post-WWII world trading system and explain how these principles work to increase welfare. In the third section, I describe the numerous exceptions to GATT's requirement of *nondiscrimination*, or equal treatment, and review the economics literature that seeks to explain the rationale for and consequences of these exceptions. Then, I present a short summary of dispute resolution within the WTO.

A brief history of the WTO and GATT

The World Trade Organization (WTO) and its predecessor, the General Agreement on Tariffs and Trade (GATT) have been enormously successful over the last 50 years at reducing tariff and other trade barriers among an ever-increasing number of countries. The predecessor to the WTO began in 1947 with only 23 members; today it has 146 members, comprising approximately 97 percent of world trade.[1] See box 1 for a timeline of GATT and the WTO.[2]

Although the WTO, established in 1995, is relatively young for an international institution, it has its origins in the Bretton Woods Conference at the end of World War II. At this conference, finance ministers from the Allied nations gathered to discuss the failings of World War I's Versailles Treaty and the creation of a new international monetary system that would support postwar reconstruction, economic stability, and peace. The Bretton Woods Conference produced two of the most important international economic institutions of the postwar period: the International Monetary Fund (IMF) and the International Bank for Reconstruction and Development (the World Bank). Recognizing that the *beggar-thy-neighbor* tariff policies of the 1930s had contributed to the environment that led to war, ministers discussed the need for a third postwar institution, the International Trade Organization (ITO), but left the problem of designing it to their colleagues in government ministries with responsibility for trade.[3]

Meredith Crowley is an economist at the Federal Reserve Bank of Chicago. *She thanks Chad Bown, Craig Furfine, and Mike Kouparitsas for detailed comments. Avinash Kaza provided helpful research assistance.*

BOX 1

Timeline of GATT and the WTO

1944: At the Bretton Woods Conference, which created the World Bank and International Monetary Fund (IMF), there is talk of a third organization, the International Trade Organization (ITO).

1947: As support for another international organization wanes in the U.S. Congress, the General Agreement on Tariffs and Trade (GATT) is created. The GATT treaty creates a set of rules to govern trade among 23 member countries rather than a formal institution.

1950: Formal U.S. withdrawal from the ITO concept as the U.S. administration abandons efforts to seek congressional ratification of the ITO.

1951–86: Periodic negotiating rounds occur, with occasional discussions of reforms of GATT. In the 1980s, serious problems with dispute resolutions arise.

1986–94: The Uruguay Round, a new round of trade negotiations, is launched. This culminates in a 1994 treaty that establishes the World Trade Organization (WTO).

1995: The WTO is created at the end of the Uruguay Round, replacing GATT.

2003: The WTO consists of 146 members, accounting for approximately 97 percent of world trade.

By the late 1940s, representatives of the American government had met several times with representatives of other major nations to design a postwar international trading system that would parallel the international monetary system. These meetings had two objectives: 1) to draft a charter for the ITO and 2) to negotiate the substance of an ITO agreement, specifically, rules governing international trade and reductions in tariffs. Although a charter was drafted, the ITO never came into being. By 1948, support for yet another international organization had waned in the U.S. Congress. Without American participation, the institution would have been greatly weakened and, in the event, the effort to create an organization to manage problems relating to international trade was abandoned.

However, although the U.S. Congress would not support another international institution, in 1945 it had given the U.S. president the authority to negotiate a treaty governing international trade by extending the 1934 Reciprocal Trade Agreements Act. This led to the establishment of the General Agreement on Tariffs and Trade (GATT) in 1947—a treaty whereby 23 member countries agreed to a set of rules to govern trade with one another and maintained reduced import tariffs for other members.[4] The GATT treaty did not provide for a formal institution, but a small GATT Secretariat, with a limited institutional apparatus, was eventually headquartered in Geneva to administer various problems and complaints that might arise among members.

Over the next 40 years, GATT grew in membership and in its success at reducing barriers to trade. GATT members regularly met in what came to be known as *negotiating rounds*. These rounds were primarily focused on negotiating further reductions in the maximum tariffs that countries could impose on imports from other GATT members. The success of these rounds is evident (see figure 1). Tariffs on manufactured products fell from a trade-weighted average of roughly 35 percent before the creation of GATT in 1947 to about 6.4 percent at the start of the Uruguay Round in 1986.[5] Over the same time period, the volume of trade among GATT members surged: In 2000 the volume of trade among WTO members stood at 25 times its 1950 volume. This growth in the volume of trade is impressive and appears to have accelerated in recent decades (see figure 2). Comparing the growth of world GDP, expressed as an index number, to the growth of the volume of trade among GATT/WTO members, also expressed as an index number, figure 2 shows that while trade grew more slowly than world GDP in the early years of the GATT/WTO, in recent years it has outpaced GDP growth.

Despite this success, by the 1980s several problems had surfaced with the GATT apparatus. Firstly, the dispute resolution mechanism of GATT was not functioning as effectively as had been hoped. Countries with longstanding disagreements were unable to reach any sort of resolution on a number of issues, ranging from government subsidies for exports to regulations regarding foreign direct investment. Secondly, a number of commodities, most importantly, agricultural products and textiles, were widely exempt from GATT disciplines. Thirdly, it was widely believed that certain forms of administered trade protection—antidumping duties, voluntary export restraints, and countervailing

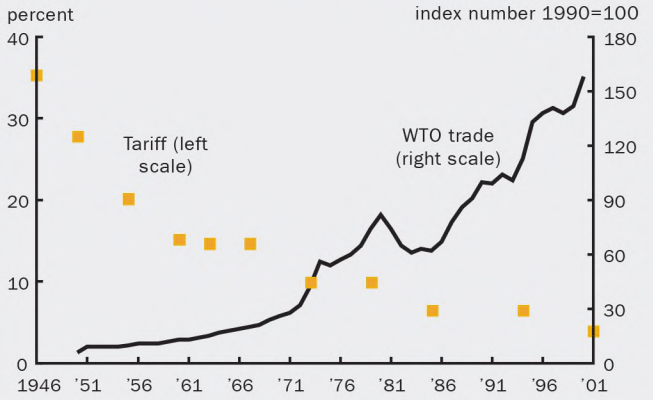
duties—were restricting trade and distorting trade patterns in many important sectors. Fourthly, trade in services was expanding rapidly and GATT had no rules regarding trade in services. Fifthly, countries that produced intellectual property—movies, computer programs, patented pharmaceuticals—were becoming increasingly frustrated by the lack of intellectual property protection in many developing nations. Lastly, the rules regarding trade-related investment measures—for example, domestic purchase requirements for plants built from foreign direct investment—were hotly disputed.

To address these problems, a new round of trade negotiations—the Uruguay Round—was launched in 1986. The goals of the Uruguay Round were far more ambitious than in previous rounds. It sought to introduce major reforms into how the world trading system would function.

The treaty negotiated during the Uruguay Round, the GATT treaty of 1994, established the WTO—the international institution to govern trade that was first visualized by the attendees of the Bretton Woods Conference 50 years earlier. The new GATT treaty provided for an entirely new and different dispute resolution mechanism to eliminate the gridlock of the old system. Furthermore, the Uruguay Round expanded GATT's authority to new areas—agreements regarding trade in textiles, agriculture, services, and intellectual property were major achievements. Finally, new sets of rules regarding administered protection came into effect with the creation of the WTO in 1995.

FIGURE 1

Growth of trade among WTO members and the decline of tariffs, 1946–2001



Sources: Tariff rates— Hoekman and Kostecki (1995); WTO trade— World Trade Organization (2001).

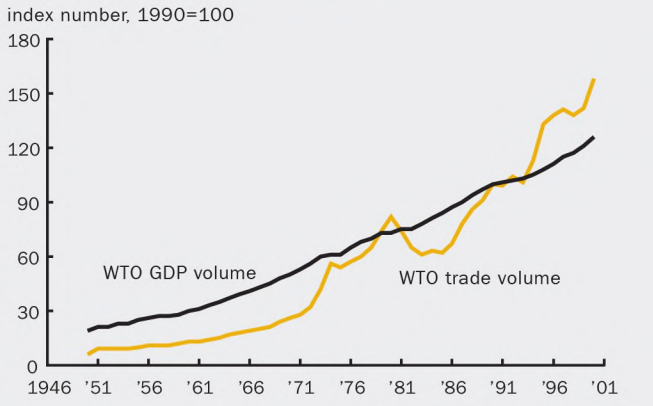
Fundamental principles of the GATT/WTO system

The success of GATT as a dynamic institution that has fostered dramatic increases in worldwide trade lies in its founding principles of reciprocity and nondiscrimination. Reciprocity refers to the practice that occurs in GATT negotiating rounds, whereby one country offers to reduce a barrier to trade and a second country "reciprocates" by offering to reduce one of its own trade barriers. Reciprocity, the practice of swapping tariff concessions, facilitates the reduction of trade barriers. *Nondiscrimination*, or equal treatment, means that if one GATT member offers a benefit or a tariff concession to another GATT member, for example, a reduction in its import tariff for bicycles, it must offer the same tariff reduction to all GATT members. Thus, nondiscrimination extends the benefits of a reciprocal tariff reduction beyond the two parties that initially negotiated it to all GATT members. Papers by Bagwell and Staiger (1999, 2001) argue that, together, these principles work toward increasing the efficiency of the world trading system.

Why is reciprocity important in reducing barriers to trade? Don't countries benefit by unilaterally reducing

FIGURE 2

Growth in volume of trade and total GDP among WTO members, 1950–2000



Source: World Trade Organization (2001).

their tariffs because lower tariffs lead to lower domestic prices? They may, but economic theory teaches us that it depends on the size of the country.

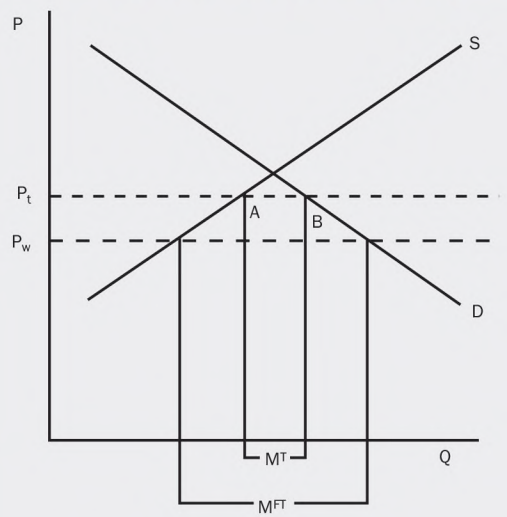
Trade theory teaches us that import tariffs are another type of tax. As a tax, tariffs raise the price that consumers must pay for a good, provide tax revenue to the government, and have the potential to create distortions, or inefficiencies, in consumption and production decisions.

If a country is very small, it will benefit by unilaterally lowering its tariffs, and reciprocity is not an important consideration (see figure 3). This is because small countries are unable to affect the prices of goods on the world market. If a small country suddenly decided to impose a 25 percent tariff on imports of automobiles, this would not affect the worldwide price at which automobiles trade. The tiny decrease in worldwide demand caused by this country's new tariff would be miniscule compared with the demand for automobiles in large markets like the U.S., the EU, and Japan. However, this tariff would make the small country worse off. Although the country's government may now collect more tariff revenue (area A in figure 3), consumers would have to pay a higher price, resulting in a loss of welfare to consumers, and there would be an efficiency loss due to the "consumption distortion" of the tariff (area B in figure 3)—fewer cars would be purchased overall. Thus, the optimal trade policy for small countries is to charge no import tariff. Regardless of the trade policies of its trading partners, a small country should engage in free trade.

The story is a bit more complicated for large countries or customs unions like the U.S. and the EU. Reciprocity is important when large countries are thinking about changing their trade policies (see figure 4). Because import demand in a large country will comprise a large share of worldwide import demand (MD in figure 4), any change in a large country's demand for a good will have an effect on that good's price on the world market. Specifically, when a large country's government imposes a tariff, this reduces the quantity of imports demanded and, consequently, causes the world price to fall. In figure 4, this is reflected in the decline in the world price from P_w to P_t^*. When the price of a country's import good falls on the world market relative to the price of the goods it exports, this is called a terms-of-trade improvement. A terms-of-trade improvement makes a country better off because it can buy imports at a relatively cheaper price on the world market. Although consumers pay a higher price for the imported good than they would under free trade, the importing country's total welfare is higher because

FIGURE 3

Impact of a tariff on a small country



P = Price
Q = Quantity
S = Supply
D = Demand
P_t = Price with tariff
P_w = World price
M^T = Imports under tariff
M^{FT} = Imports under free trade
A = Tariff revenue
B = Consumption distortion

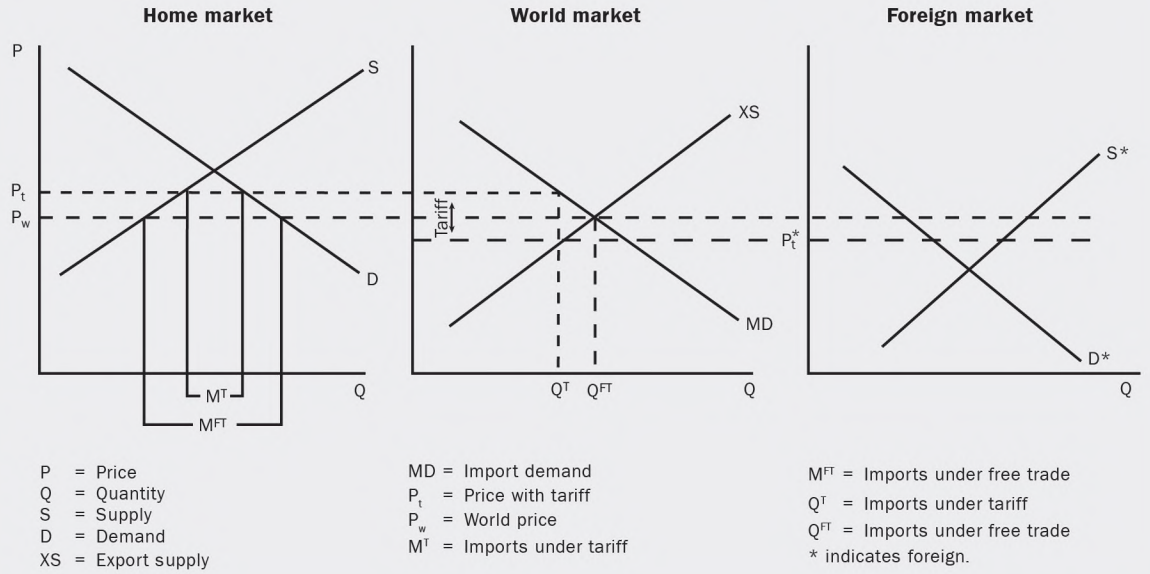
the government earns tariff revenue and because import-competing producers earn higher profits.

Another way to think about a large country's use of tariffs is to focus on the question of who bears the cost of this tax. Although the consumers in a large country must pay a higher final price for the imported good (P_t rather than P_w) when their government imposes a tariff, they don't bear the full tax burden of the tariff. A tariff that causes the world price of a good to fall hurts the foreign exporters that produce that good, because they only receive P_t^* instead of P_w. As a whole, the exporting country loses some of its purchasing power on the world market in this worsening of its terms of trade. In this way, some of the cost of the tariff is pushed onto the foreign producers of the good in the form of the lower price they receive for their product than they would under free trade. Because foreign producers lose out under this import tariff, it is sometimes called a beggar-thy-neighbor policy.

The use of a beggar-thy-neighbor tariff by a large country not only makes the importing country strictly better off and the exporting country strictly worse off, it introduces inefficiencies into the world trading

FIGURE 4

Impact of a tariff on a large country



system that cause the net effect of the tariff to be negative. The import tariff induces inefficient production distortions in both countries. The level of production is too high in the importing country and too low in the exporting country relative to what the levels would be under free trade. However, although the tariff is bad for the world as a whole, it remains a desirable and beneficial policy for the importing country. At the end of WWII, many of the large countries that became the original members of GATT had high tariffs. They found themselves in what economists call a terms-of-trade-driven prisoner's dilemma. The prisoner's dilemma is a famous problem in the field of game theory that describes a situation in which two parties can improve their situations by acting cooperatively, but the individual incentives they face lead them to act noncooperatively.

Figure 5 provides a highly stylized example of the terms-of-trade-driven prisoner's dilemma faced by two large countries—America and a foreign country—at the end of WWII. We can read the figure as follows. The horizontal rows depict the policy options available to America—free trade with the foreign country or charging a beggar-thy-neighbor tariff. The vertical columns represent the policy options available to the foreign country—free trade or a beggar-thy-neighbor tariff. The numerical entries in the four boxes show the payoffs that each country will receive if the different policy options are taken. The first number represents America's payoff; the second, the foreign country's. For example, the box in the lower left-hand corner tells us that if the U.S. imposes a beggar-thy-neighbor tariff and the foreign country practices free trade, the U.S. will receive a payoff of 15 and the foreign country receives nothing. In the upper left corner, if both countries practice free trade, then the worldwide payoff of 20 (the sum of America's payoff and the foreign country's payoff) is higher than under any other set of policy options. However, in this example, both countries want to avoid being the dupe that practices free trade and faces a beggar-thy-neighbor tariff. Thus, in equilibrium, each country charges a beggar-thy-neighbor tariff and receives the low payoff of 5.

As in the stylized example in figure 1, the problem facing countries at the end of WWII was that they knew that they would collectively be better off under

FIGURE 5

The prisoner's dilemma

		Foreign country	
		Free trade	BTN tariff
America	Free trade	10, 10	0, 15
	BTN Tariff	15, 0	5, 5

free trade. Although each country benefited from its own import tariff, it also suffered at the hands of its trading partners' import tariffs. What was needed was a mechanism by which countries could jointly commit to tariff reductions that would reduce the losses due to production and consumption distortions and, through gains in efficiency, make all countries better off.

GATT, through its practice of reciprocal tariff reductions, provided the necessary mechanism for countries to commit to freer trade. Under GATT, large countries that reduced their import tariffs would experience a net gain because their trading partners would simultaneously reduce their import tariffs. In all countries, the reallocation of labor and capital away from protected import-competing firms and toward export sectors would generate real efficiency gains.

It is evident that reciprocity is necessary for two large countries to engage in trade liberalization, but this could have been achieved with a network of bilateral treaties.[6] Why did GATT adopt a multilateral approach with a strict requirement for nondiscrimination?

Nondiscrimination is a convenient way to reduce the complexity of international trading relations. On a purely practical level, it may be easier to negotiate one set of import tariffs than to engage in dozens of bilateral agreements. In fact, Jackson (1997) speculates that when nondiscrimination, or "most-favored-nation," clauses were originally introduced into trade treaties in the sixteenth century, they had a practical benefit—drafters did not have to copy large sections of treaties again and again.

However, while convenience and practicality are important, nondiscrimination would not have become a central feature of GATT if it did not yield real economic benefits. Nondiscrimination in tariff policy, that is, setting the same tariff on imports from all countries, ensures that resources are allocated to their most productive use. On the import side, nondiscrimination ensures that countries purchase imports from the lowest-cost source country. Further, nondiscrimination prevents *trade re-routing*, in which goods are moved through third countries in order to circumvent high tariffs. Lastly, Bagwell and Staiger (2003) show that, on the export side, nondiscrimination protects exporting countries from *bilateral opportunism*.

As an importer, a country can charge a single nondiscriminatory tariff on imports from all countries or it can set different tariffs on imports from different countries. Under a nondiscriminatory tariff, imports will be sourced from the lowest-cost producer in the world. Compare this to a system of discriminatory tariffs in which, for example, the U.S. sets a lower, preferential tariff on T-shirts from Mexico than on T-shirts from China. If China can produce T-shirts more cheaply than Mexico, but the tariff on Chinese T-shirts is so much larger than the tariff on Mexican T-shirts that it is cheaper for Americans to buy T-shirts from Mexico, there is a real loss due to the production distortions caused by the U.S.'s discriminatory tariffs. Resources in Mexico that could have been better employed in some other sector are utilized in its relatively high-cost T-shirt industry. Resources in China that could have been efficiently used to make T-shirts are allocated to another industry. When a country uses a nondiscriminatory tariff, this facilitates the allocation of resources worldwide to their most productive uses.

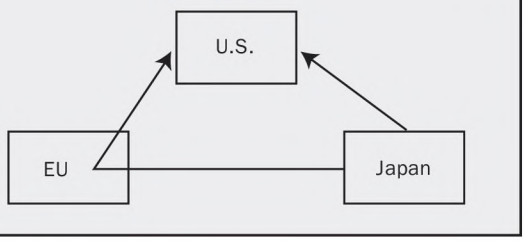
Trade re-routing is a costly practice whereby an exporter ships its goods to a third country, repackages it, and then ships it to a final destination where it will qualify for the third country's lower, preferential tariff rate (see figure 6). In some cases, in order to qualify for the preferential tariff, the product must undergo a *substantial transformation* in the third country. This sometimes leads firms to move a stage of the production process to the third country. When an importing country utilizes a single nondiscriminatory tariff for all imports, there is no need for exporters to engage in the costly process of re-routing.

When two countries bilaterally negotiate tariff concessions, the principle of reciprocity implies that the tariff reductions on various products are balanced in such a way that the terms of trade between the two countries remain unchanged (that is, neither country is "beggaring" the other), while the volume of trade increases to a more efficient level. However, in a world in which both countries remain free to negotiate an additional trade agreement with a third country, the problem of bilateral opportunism arises. For example, if one country were later to offer a lower tariff rate to a third country, this could erode the value of the original tariff concession to the first trading partner. Bagwell and Staiger (2003) have shown that when negotiations utilize the practices of reciprocity and nondiscrimination, the problem of bilateral opportunism is eliminated.

In summary, GATT's founding principles of reciprocity and nondiscrimination facilitate increases in well-being for the countries that belong to the WTO. By coordinating tariff reductions among large countries, GATT makes efficiency gains from trade a reality. By requiring that countries set nondiscriminatory tariffs, GATT facilitates the production of goods in the most efficient location. However, a number of exceptions to GATT's nondiscrimination rule exist. In the next section, I explore these exceptions and why GATT allows the use of discriminatory tariffs in special circumstances.

FIGURE 6

Trade re-routing



Exceptions to GATT's nondiscrimination principle

Although nondiscrimination is an ideal in GATT, in practice a number of exceptions to this general rule exist. *Regional trade agreements*—both free trade areas and customs unions—are allowed. Governments may also use *administered protection*—special tariffs that can be used for particular purposes. Both types of exceptions create both problems and benefits for the world trading system.

Regional trade agreements

In 1947 when negotiators drafted the original GATT treaty, they recognized that from time to time, some countries might want to push ahead with greater trade liberalizations. Although GATT preferred nondiscriminatory tariffs, it did not intend to impede the gains from trade that could be had if only a few members were willing to reduce their tariffs even further. Therefore, it allowed the formation of two types of regional trade agreements—free trade areas and customs unions. In a free trade area, the members maintain their original external tariffs with the rest of the world, but engage in free trade with one another. In a customs union, all member countries set the same external tariff for imports from non-members and eliminate the tariffs on imports from members. When GATT members form a customs union, the common external tariff can be no higher than a weighted average of the tariffs of the member countries before the customs union was formed.

From the beginning, the decision to allow regional trade agreements within GATT was controversial. Viner (1950) framed the question as an essentially empirical one: Were regional trade agreements "trade creating" or "trade diverting?" He coined the term "trade creation and trade diversion" to describe what happens when several countries join together to form a regional trade agreement (RTA). The reduction in tariffs among RTA members leads to trade creation among members. The problem is that the trade that develops between RTA members may not reflect an overall expansion of a country's imports, but rather a diversion of trade away from a non-RTA country to a RTA member. In this case, there may be no worldwide efficiency gains from trade if the non-RTA country is the lowest-cost producer of some good.

Today, the question of whether regional trade agreements are trade creating or trade diverting remains unresolved. In fact, it is almost impossible to answer this question definitively because economists never observe the appropriate benchmark for estimating the amount of trade creation and trade diversion associated with a regional trade agreement. Because economies and trade are always growing, it is hard to construct a counterfactual estimate of how much trade would have grown among RTA members if these countries had not formed a regional trade agreement.

Sampson (1996) argues that the question of trade creation and trade diversion is much less important today than it was 50 years ago because tariffs today are much lower. For the U.S. and the EU, most products face import tariffs of less than 5 percent. Therefore, Sampson argues, although RTA members with these countries do benefit from a 0 percent tariff rate, the size of this *tariff preference*—the difference between the tariff for RTA members and other countries—is so small that it cannot possibly induce much trade diversion.

Sampson's argument finds some support in a number of recent papers that have tackled this question using highly disaggregated data on commodity trade. Both Romalis (2002) and Clausing (2001) examine trade creation and diversion in the context of the North American Free Trade Agreement (Nafta). Both papers find that Nafta created substantial amounts of trade, but Romalis also finds evidence that Nafta may have induced substantial trade diversion when tariff preferences are very large. Prusa (2001) and Bown and Crowley (2003b) focus on deviations from GATT's nondiscrimination rule that arise when the U.S. imposes country-specific antidumping duties. The tariff preference associated with antidumping duties is very large and, thus, these papers find strong evidence of trade diversion. Prusa finds that antidumping duties lead the U.S. to source its imports from countries that face a lower import tariff. Bown and Crowley (2003b) focus on what happens to the exports of a country that faces a country-specific antidumping duty. They find a substantial "trade deflection" effect—exports are diverted to countries with lower import tariffs. Overall, the empirical literature finds evidence that trade diversion occurs. However, the debate over the relative magnitudes of trade creation and diversion continues.

A different but related body of research examines whether regional trade agreements are "building blocks" or "stumbling blocks" (Bhagwati, 1992) on the path to worldwide free trade. A theoretical paper by Bagwell and Staiger (1997a) on free trade areas and papers by Ethier (1998) and Freund (2000) argue that regional trade agreements are building blocks that can facilitate greater multilateral tariff reductions or higher global welfare. However, research on customs unions by Bagwell and Staiger (1997b) and research on regional trade agreements by McLaren (2002), Levy (1997), and Bond and Syropoulos (1996) supports the idea that regional trade agreements are stumbling blocks.

All these papers explore how productive resources are reallocated across countries and/or across sectors within a country when multilateral and regional trade agreements are formed. In the models of Bagwell and Staiger (1997a), Ethier (1998), and Freund (2000), the reallocation of resources that accompanies the formation of an RTA creates a situation where further reallocation under a multilateral agreement is feasible and welfare enhancing for everyone. In contrast, in Bagwell and Staiger (1997b), Levy (1997), McLaren (2002), and Bond and Syropoulos (1996), changes in the economy that result from the formation of a regional trade agreement inhibit further multilateral trade liberalization.

Administered trade protection

While RTAs were permitted by GATT because at least some believed that they could facilitate greater worldwide trade, administered trade protection—temporary tariffs that are usually discriminatory—was allowed for a variety of reasons.

The term *administered protection* refers to trade restrictions that provide protection from imports above and beyond the protection afforded by the tariffs that were negotiated as part of GATT. GATT permits the use of *antidumping duties*, *countervailing duties, safeguard measures*, and tariffs to assist with balance of payments problems.[7] *Voluntary export restraints* are an administered trade barrier that is technically no longer allowed within the WTO but was popular in the 1980s. The use of these trade policies represents a deviation from GATT's principle of nondiscrimination. Antidumping duties, which are imposed at the country or firm level, are probably the most discriminatory. GATT requires that safeguard measures be nondiscriminatory, but in practice many countries apply them in a discriminatory manner.

Economics research that seeks to rationalize the inclusion of the various forms of administered protection in the WTO explores the argument that administered protection either 1) improves worldwide welfare or 2) improves the welfare of politically powerful importing countries and, especially, their import-competing sectors. The first argument is that administered protection can create a net benefit for the world as a whole. The protection may make some countries better off and others worse off, but if we add up the gains and losses to everyone in the world, the sum total is positive. In other words, the gains of temporary trade protection outweigh the losses. The second argument is partly political and partly economic. Some group profits from the use of administered protection. Even though protection may reduce worldwide welfare, it is included in GATT because those who benefit wield enough political power to see that it remains within the agreement. Furthermore, recall from the discussion of reciprocity and the terms-of-trade-driven prisoner's dilemma above that large countries benefit when they unilaterally impose beggar-thy-neighbor tariffs. Although countries may use GATT to arrive at a cooperative welfare-enhancing outcome, they still may be tempted to cheat and reimpose beggar-thy-neighbor tariffs. The different forms of administered protection could provide an avenue for doing this.

Next, I provide some background information on safeguards, antidumping duties, and countervailing duties and review the economic research on these different trade policy instruments.

Safeguards

A safeguard measure is a temporary tariff or quota that is used to protect a domestic industry from "fair" foreign competition.[8] Whereas antidumping and countervailing duties are intended to "level the playing field" when foreigner exporters have an "unfair" advantage over domestic producers, safeguard measures may be used against foreign exporters that have a fair competitive advantage in a product.

The use of safeguards first began in the 1940s when the U.S. began to pursue a liberal trade agenda. Fearing that the lowering of a tariff on some particular good as part of a trade agreement could result in a larger-than-expected import surge that would hurt domestic firms, the U.S. government insisted that a safeguard provision be part of every trade treaty that it signed. Under GATT, when members negotiated reciprocal tariff concessions, they committed themselves to maximum tariffs. These commitments restricted, to a considerable extent, a domestic policymaker's authority to unilaterally raise tariffs at some later date.

To encourage countries to make greater concessions during negotiations, GATT included two provisions

under which countries could reintroduce protective trade policies. Countries remained free to temporarily raise a tariff above the maximum level or introduce a temporary quantitative restriction under the Article XIX safeguard provision. Countries wishing to permanently raise their tariffs could do so under Article XXVIII.

According to GATT's Agreement on Safeguards, safeguard measures should be nondiscriminatory, but in fact countries often use discriminatory safeguards. This practice is contentious and frequently challenged before the WTO's dispute settlement body. For example, the recent U.S. Global Steel Safeguard raised the import tariff on steel for many countries, but granted exemptions for steel imports from many of our free trade partners like Canada and Mexico. The WTO's dispute settlement body recently announced that these exemptions are violations of GATT's rules.[9]

Other GATT rules specify that safeguards should only be used when imports increase unexpectedly or as the result of unforeseen developments. This leads to numerous debates over what developments can be classified as "unforeseen." Prior to the Uruguay Round's revisions to the safeguard rules in 1994, the use of a safeguard measure was subject to measured retaliation. If a country imposed a safeguard on a product, its trading partners that were hurt by the safeguard could retaliate with their own tariff increases on other products. As part of the Uruguay Round reforms, the safeguard rules changed so that safeguards are no longer subject to retaliation for the first three years they are in effect. This rule change was intended to make nondiscriminatory safeguards more attractive for protection-seeking governments relative to discriminatory antidumping duties.

The economics literature provides several different rationales for why the WTO allows the use of safeguards. Perhaps the most widely cited argument for safeguards is that their existence can facilitate greater tariff liberalization by governments during trade negotiations. Because a government has an escape valve if a tariff reduction causes pain to its own producers, it has more freedom to make larger and potentially more risky tariff reductions. Because there are large gains from permanent tariff reductions and relatively small costs from imposing temporary safeguards in a few sectors, the world gains by having safeguards in a trade agreement, even when they are not actually used.

A paper by Ethier (2002) formalizes this basic idea. His central concern is to analyze a trading system like the GATT/WTO, which is characterized by the general practice of negotiating tariff reductions to benefit all members and the occasional use of temporary unilateral tariff increases through safeguards or antidumping duties. He develops a model in which countries grow at different rates. The key insight is that when countries negotiate tariff reductions, they do not know if their growth will be fast or slow. In a trade agreement that does not allow temporary tariff increases, countries fear their growth will be slow and will negotiate only small tariff reductions. When safeguards are added to the trade agreement, countries negotiate large tariff reductions because they know that if they turn out to have slow growth, they can temporarily increase their tariffs.

Klimenko, Ramey, and Watson (2002) arrive at a similar result by examining the question of why the WTO's dispute settlement body (DSB) exists. In their paper, they show that when countries regularly renegotiate their tariffs, as in the WTO's trade rounds, a DSB is necessary for the trade agreement to survive. A DSB makes it possible for countries to punish each other for violations. Because countries want to avoid punishment, they won't violate the trade agreement when it includes a DSB. As an extension to their paper, they also show that if the DSB allows countries to temporarily raise their tariffs (as is the case with safeguard measures) in response to some unexpected change in the economic environment, they will negotiate larger tariff reductions initially.

Although some of the theoretical arguments suggest safeguards help facilitate trade liberalization, other economists arrive at the opposite conclusion. Staiger and Tabellini (1987) show that allowing for safeguard measures could reduce the credibility of a trade agreement. From this perspective, the inclusion of a safeguard measure can weaken the overall agreement.

On the other hand, another economic argument in favor of the inclusion of safeguards is that they act as a form of insurance against fluctuations in the terms of trade. Consider a country that imports a good whose price fluctuates substantially. When prices change, the economic environment can become so different that countries want to pull out of a trade agreement that constrains them to set low tariffs.

Bagwell and Staiger (1990) explore how price fluctuations affect large players in a trade agreement—countries or regions like the U.S., the EU, and Japan with such large markets that their safeguard measures can significantly alter world prices. They argue that due to the self-enforcing nature of the trade agreement, in periods of large import volumes, a safeguard measure acts as a pressure valve to enable countries to sustain cooperation by temporarily raising tariffs. In the absence of a safeguard clause, countries would not be able to sustain cooperation, and the result

would be a costly trade war with high levels of tariff retaliation. Fischer and Prusa (1999) show that even small countries, which cannot affect world prices by imposing a safeguard, can use safeguards to insure themselves against international price shocks.

To date, empirical research in economics hasn't been able to prove or disprove the ideas put forth in the papers mentioned above. In some ways this is an impossible task—how can we prove that countries negotiate lower tariffs when a safeguard is part of a trade agreement when all the trade agreements in existence include safeguards?[10]

Another important area of research argues that the WTO allows the use of safeguards because of concern for the interests of importing countries and their import-competing firms and industries. Safeguards may exist because the agents that benefit from the safeguards are politically powerful. Many of these papers focus on analyzing how the politically powerful agent gains from the safeguard. If one country pursues a policy that benefits itself but harms other countries, economists want to understand how and why the policy creates a benefit so that they can develop alternative policies that create the same or a similar benefit but reduce or eliminate the harm to others. I examine three arguments for why governments use safeguards to assist import-competing industries: to help them catch up to their foreign competitors, to facilitate their exit from the industry, and to reap the gains for a politically preferred sector.

Several theoretical papers (Matsuyama, 1990; Miyagiwa and Ohno, 1995, 1999; and Crowley, 2002) explore how safeguards benefit import-competing firms that are technologically behind their foreign competitors. These papers examine the consequences of using a temporary safeguard to induce domestic firms to adopt newer, more efficient production technologies. Economists have long understood that a government subsidy is better than a tariff for helping a firm adopt a new technology.[11] A direct subsidy can achieve the same result as a safeguard, but because it doesn't increase the price consumers will face, it is less costly to society as a whole. Thus, using a safeguard to facilitate technological improvement is a "second-best" policy at best.

Matsuyama (1990) and Miyagiwa and Ohno (1995) provide theoretical support for the WTO's practice of setting a strict termination date for safeguard protection and allowing exporting countries to retaliate against safeguard measures that extend beyond this limit. Miyagiwa and Ohno (1995) find that safeguards provide an incentive for protected firms to innovate quickly only if the cost of the new technology is falling over time and the termination date for safeguard protection is credibly enforced by foreign retaliation. Crowley (2002) finds a nondiscriminatory safeguard tariff can accelerate technology adoption by a domestic import-competing firm, but will slow down technology adoption by foreign exporting firms. Because a nondiscriminatory safeguard tariff can delay a foreign firm's adoption of new technology, its worldwide welfare costs may exceed its benefits.

Unfortunately, the little empirical evidence on the effect of safeguards on technology adoption is not very encouraging. A 1982 study by the U.S. government's administrative body that reviews safeguard petitions, the U.S. International Trade Commission (USITC), found that most safeguards failed to promote a positive adjustment to import competition. Rather than assisting companies in upgrading their facilities, in most cases safeguards merely slowed an industry's inevitable decline. There are some exceptions; Harley-Davidson, a motorcycle producer, received safeguard protection in 1983 and successfully retooled its plants. However, successful cases are the exception to the rule. A review of U.S. safeguard cases since 1974 shows that some industries seek and receive protection repeatedly—for example, stainless alloy tool steel was granted safeguard protection in 1976 and again in 1983.

Another group of theoretical papers shows how firms in declining industries can utilize political support to maintain protection. Hillman (1982), Brainard and Verdier (1994, 1997), and Magee (2002) all examine the use of tariff protection to allow a dying industry to collapse slowly rather than quickly. Because these papers all assume that there are high costs to quickly scaling back production, they find that a temporary tariff that can slow an industry's decline can improve an importing country's welfare. However, this type of policy also slows the reallocation of capital and labor into other industrial sectors in which they would be more productive. This loss of productivity is an indirect welfare cost on the country imposing the safeguard measure.

In summary, there are a number of potential reasons GATT allows the use of safeguard measures. Most of these papers do not explore the issue of nondiscrimination. The one paper that does, Crowley (2002), finds that a safeguard can only benefit the importing country if the measure is nondiscriminatory.

Antidumping duties

Antidumping duties are a controversial form of temporary trade protection permitted by GATT. An antidumping duty is a tariff that an importing country imposes on imports of a product that have been *dumped*

into its domestic market by some exporting country's firm(s). An importing country may only impose an antidumping duty on a product if there is evidence that foreign firms have sold their products at less than normal value and this has injured the domestic industry.

Historically, antidumping duties have been distinguished from other forms of administered protection by the trade problem they were used to remedy. Antidumping duties were a government's remedy for "unfair" trade and were intended to offset the price undercutting of foreign exporters engaged in anticompetitive practices. In the early twentieth century, the U.S. instituted an antidumping law to protect its domestic firms from German cartels that sold their excess output at low prices in the U.S. market. Although low prices for imported goods improve the well-being of a country and should be welcomed by the government, in some cases they could present a problem. If a foreign firm is engaging in predatory pricing—setting prices low in order to drive competitors out of business—this could lower the welfare of a country in the long run. This can happen if the foreign firm becomes a monopolist and uses its monopoly power to charge consumers extremely high prices. GATT's Antidumping Code allows countries to violate the nondiscrimination rule and impose an additional tariff—an antidumping duty—on imports from a firm that is dumping. Thus, one could view GATT's antidumping rules as an effort to improve worldwide welfare by preventing the harmful practice of predatory pricing. However, although most economists would agree that an anticompetitive practice like predatory pricing is harmful, there is almost no evidence of this type of practice in alleged incidences of *dumping*.

Rather, in almost all modern cases of dumping, foreign firms are either engaging in international price discrimination or temporarily pricing below their average cost of production. In fact, GATT now defines dumping as either international price discrimination—that is, charging different prices for a good in different countries because demand for the good is different in the different countries—or as pricing below the average cost of production. Prior to the Uruguay Round reforms, antidumping duties had no effective time limit. Once an antidumping duty was put in place, it could remain in place for years. Today, antidumping duties are subject to "sunset reviews" every five years. During a sunset review, a duty is removed unless there is evidence that the targeted country continues to dump and this dumping is hurting domestic firms in the importing country.

One peculiarity of GATT's Antidumping Code is that it encourages the use of *price undertakings* in place of an antidumping duty. Economists view price undertakings with suspicion because they look a lot like government-sanctioned collusive pricing. Because collusive pricing—a practice in which several firms agree to simultaneously raise prices and keep them high—hurts consumers, it is surprising that GATT would encourage this type of practice. It is hard to justify on economic welfare grounds.

Although the rhetoric that surrounds the use of antidumping duties focuses on whether foreign firms are behaving "fairly," the important question is not whether dumping is fair but whether dumping is harmful. Thus, to understand how antidumping policy affects the world trading system, economists first ask why firms engage in practices like pricing below the average cost of production.

With antidumping investigations into goods as varied as fresh-cut flowers, semiconductors, and countless varieties of steel, economists have tried to explain the phenomenon of dumping and the government's policy response in terms of the different modes of competition in the markets for dumped goods.

Several papers explain dumping in the context of competitive markets. These papers focus on explaining why competitive firms dump. In Ethier (1982), competitive firms with implicit labor contracts will dump during periods of slack world demand. Essentially, these firms have high fixed costs that lead them to price below their average total cost when demand is weak. Staiger and Wolak (1992) show that a foreign monopolist that faces weak demand in its own market will dump into a perfectly competitive domestic market. In this model, a foreign firm with excess capacity will sell at a price below average total cost in its export market in order to protect its monopoly profits in its own market. Unfortunately, neither Ethier's paper nor Staiger and Wolak's explains how antidumping policy affects the welfare of the importing country. However, another paper that examines dumping in competitive markets, Clarida (1993), finds that antidumping policy reduces an importing country's welfare. In this model, competitive firms dump, that is, sell below average cost, as they learn about their own production technologies during a period of high world demand. The importing country benefits when import prices are low, so the introduction of an antidumping policy that raises prices leaves the country worse off. In summary, the research on dumping in competitive markets suggests that antidumping policy is harmful and cannot provide an economic rationale for its existence.

The literature on dumping in imperfectly competitive markets is somewhat more successful in providing

a rationale for why GATT includes an antidumping provision. Markets where there are a small number of large producers—like automobiles—are said to be imperfectly competitive.

Dixit's (1988) seminal paper on dumping in an imperfectly competitive market shows that, in general, when dumping is defined as international price discrimination, it actually benefits the importing country. More specifically, the benefits to consumers of being able to buy goods at low, dumped prices outweigh the losses to domestic producers. Thus, as a general rule, antidumping policy reduces the welfare of importing countries when markets are imperfectly competitive.

However, two papers, Gruenspecht (1988) and Crowley (2002) utilize an alternative definition of dumping—pricing below the average cost of production—and find that this kind of dumping can hurt an importing country. Thus, antidumping duties can help. These papers provide one explanation for why GATT allows antidumping duties.

Gruenspecht (1988) focuses on dumping by firms in industries with steep learning-by-doing curves in production. That is, he models industries like semiconductors where production costs fall as a firm's experience in making the product increases. He shows that an importing country benefits from an antidumping law. In his model, antidumping duties can improve the welfare of a large importing country by increasing the size of the market available to sales by the home firm. Higher production yields greater productivity gains that improve the home country's welfare.

Crowley (2002) focuses on industries in which firms must pay large sunk costs to install capacity, for example, an industry like steel. She shows that when demand for the good fluctuates, foreign firms will dump their output when demand in their own market is weak. In response to this, the importing country can improve its welfare by imposing a temporary antidumping duty until demand in the foreign country returns to a normal level. In this case, the antidumping duty shifts some of the foreign firm's profits back to the home country.

Finally, the paper by Fischer and Prusa (1999) that I discussed earlier in the context of safeguards also provides a rationale for antidumping law. A small country that faces international price fluctuations can use an antidumping duty as a form of insurance against harmful movements in its terms of trade.

In summary, regardless of the degree of competition in a market, it is hard to rationalize the inclusion of antidumping rules in GATT on economic welfare grounds.

Countervailing duties

Countervailing duties, tariffs used to offset the effect of a foreign government's subsidy, are similar to antidumping duties. Because a foreign government's subsidy to an export good lowers its price in the importing country, in most cases a foreign subsidy benefits consumers in an importing country. Thus, in most cases, there is no economic welfare rationale for a countervailing duty policy within GATT.

However, in markets that are imperfectly competitive, a foreign government's subsidy can reduce the welfare of an importing country. In this case, although consumers in the importing country benefit from the subsidy, the losses to firms in the importing country outweigh the benefits to consumers. Dixit (1988), Spencer (1988), and Collie (1991) all show that in this case, a countervailing duty can prevent the foreign government's subsidization of its export good and improve the welfare of the importing country.

In summary, although countervailing duties are likely to lower an importing country's welfare when markets are competitive, it is theoretically possible for them to improve an importing country's welfare when markets are imperfectly competitive.

Dispute resolution in the WTO

Having reviewed the various exceptions to GATT's nondiscrimination rule, I now turn to the issue of disputes. What happens when a dispute arises between countries over a GATT rule? What power does GATT have to settle disputes? How does GATT enforce its own rules?

GATT is a multilateral trade agreement with the authority to regulate the trade regulations of its member governments. As an international treaty, it has no authority over individuals, private firms, or public corporations. Rather, it governs the interactions of countries that voluntarily agree to abide by its rules.

The WTO mediates and settles disputes among its members. Disputes that cannot be resolved among the members themselves are referred to a panel of three persons who act as judges. When a country is found to be in violation of its GATT obligations, it has three choices. It can appeal and have the case retried before an appellate body, it can amend its laws to bring them in line with GATT, or it can keep its laws as they are and face "measured retaliation" from its aggrieved trading partners. If a country loses an appeal, its options revert to amending its laws or facing retaliation. Measured retaliation is the WTO's main enforcement mechanism. In the simplest case, if one country were to violate its GATT obligations by raising its tariff on some good and this tariff increase caused the volume

of imports from a second country to fall, the WTO could authorize the second country to punish the first by raising its own tariff on something. This retaliation by the second country is "measured," in the sense that it should reduce trade from the offending first country by roughly the same value as the first country's tariff increase.

The practice of measured retaliation is extremely useful in maintaining the smooth functioning of the world trading system. Historically, when one party to a treaty violated one of its terms, the other party could either accept the violation or withdraw for the treaty entirely. Measured retaliation essentially allows both parties to jointly withdraw from some of their treaty obligations while still enjoying the benefits of the rest of the treaty.

In fact, while the recent increase in disputes among WTO members may, on the surface, appear troubling, it could also signal the effectiveness of the dispute resolution system. It could be that countries that have grievances against their trading partners find the dispute settlement system sufficiently effective that they present their disputes to this body rather than seeking some type of resolution outside the WTO.

Conclusion

This article has provided a brief history of the WTO and has suggested that the success of the GATT and WTO system can be attributed to the founding principles of reciprocity and nondiscrimination. I have also reviewed the numerous exceptions to GATT's principle of nondiscrimination. Although the various exceptions may yield benefits, theoretical and empirical research in economics questions whether the benefits of these exceptions are sufficiently large to outweigh the costs.

The WTO is currently engaged in a new round of trade negotiations—the Doha Round. This article's review of the economics literature suggests that it may be time to rethink GATT's rules for administered protection. The proliferation of antidumping duties is costly to both consumers and many exporters. Many countries that belong to the WTO would like to make it more difficult for countries to impose antidumping duties. However, because antidumping protection is popular among import-competing firms in both the U.S. and the EU, it will be politically difficult to achieve meaningful reform of GATT's antidumping rules. There may be more support for modest changes to the Agreement on Safeguards. For example, the discriminatory application of safeguards has been an issue in many WTO disputes. Negotiators to the Doha Round could potentially preempt future disputes over safeguards by closing some loopholes and clarifying the language in the safeguard agreement.

Perhaps the largest gains that could be achieved in the current negotiating round might come from liberalizing trade in agricultural commodities. Developing countries, many of which have a comparative advantage in agricultural production, would like to see developed countries' agricultural markets open up through both tariff and subsidy reductions. The liberalization of trade in agriculture has the potential to generate huge welfare gains for the entire world.

NOTES

[1]See the WTO webpage at www.wto.org.

[2]Jackson (1997) and Hoekman and Kostecki (1995) provide good histories of the post-WWII world trading system.

[3]For definitions of all terms in italics, see the appendix on p. 55.

[4]Under the GATT treaty of 1947, GATT members were technically known as "contracting parties."

[5]Hoekman and Kostecki (1995), p. 20.

[6]This article has emphasized the importance of reciprocity in trade negotiations. However, large countries could engage in trade negotiations for reasons not considered here.

[7]GATT also permits the use of tariffs to assist with balance of payments problems in developing countries with fixed exchange rates. The balance of payments exception is relatively uncontroversial and I do not discuss it here.

[8]This section draws heavily upon Bown and Crowley (2003a).

[9]See WTO (2003).

[10]A related paper is the recent empirical contribution by Staiger and Tabellini (1999), who compare two different policy environments to investigate the question of whether GATT rules help governments make trade policy commitments. They find evidence to support the claim that GATT rules do give governments commitment power. However, their work also provides support for the theory that the inclusion of an escape clause can have damaging effects that erode a government's ability to commit to liberalization.

[11]Dixit and Norman (1980), Caves, Frankel, and Jones (2002), and Krugman and Obstfeld (2000) are a few standard textbooks that make this point.

APPENDIX: GLOSSARY OF TRADE TERMS

Administered protection: Special tariffs, quotas, or other restrictions on imports that are allowed under GATT. The treaty allows the various forms of administered protection for a variety of reasons, including to enable a country to address specific domestic concerns and to promote macroeconomic stability. Policymakers often refer to administered protection as trade remedies.

Antidumping duty: A tariff used to raise the price of a dumped product.

Beggar-thy-neighbor tariff: A tariff, imposed by a large country, that causes the world price of a good to fall. This fall in the world price benefits importing countries and hurts exporting countries.

Bilateral opportunism: The practice by which one country, after negotiating a bilateral trade agreement with a second country, goes on to negotiate a bilateral trade agreement with a third country that undercuts the benefits that the second country expected to receive under its agreement.

Countervailing duties: Tariffs used to offset the advantage foreign exporters have over domestic producers in cases in which foreign exporters receive subsidies from their governments.

Dumping: Selling a product in an export market at a price below its "normal value." GATT defines normal value as the price a good sells for in its home market or a third country's market, or as the average cost of its production.

Measured retaliation: A mechanism to enforce the WTO rules. If one country violates a WTO rule and the violation reduces trade from a second WTO member country, the WTO may authorize the second country to punish the first country by allowing the second country to violate a WTO rule (for example, by raising a tariff). This punishment should reduce trade from the offending first country by roughly the same amount as the trade reduction caused by the original violation.

Negotiating round: A meeting of GATT/WTO members at which members negotiate reductions in tariffs and/or changes to GATT/WTO trading rules.

Nondiscrimination: The policy of treating all of one's trading partners equally. A country is practicing nondiscrimination if it charges the same tariff on imports of a product (for example, 5 percent on shoes) without regard to where the product is made.

Price undertaking: An agreement whereby a foreign firm accused of dumping agrees to raise its price. If the price increase is large enough, the importing country agrees not to impose an antidumping duty.

Regional trade agreement: An agreement among two or more countries in which the tariffs they impose on one another's goods are lower than the tariffs they impose on goods from other countries. These agreements are also known as preferential trade agreements.

Safeguards: Temporary tariffs, quotas, or tariff-rate quotas that protect an industry from fair foreign competition.

Tariff preference: The difference between a country's nondiscriminatory tariff and the tariff applied to imports from a particular country due to participation in a regional trade agreement or application of a special tariff like an antidumping duty.

Terms of trade: The price of a country's exports divided by the price of its imports. An increase or improvement in the terms of trade raises a country's welfare.

Voluntary export restraint: An agreement whereby an exporting country reduces its exports to some importing country. VERs are also known as orderly marketing agreements (OMAs), voluntary restraint agreements (VRAs), and export restraint agreements (ERAs), among other terms.

REFERENCES

Bagwell, Kyle, and Robert W. Staiger, 2003, "Multilateral trade negotiations, bilateral opportunism, and the rules of GATT/WTO," *Journal of International Economics*.

_____, 2001, "Reciprocity, non-discrimination, and preferential agreements in the multilateral trading system," *European Journal of Political Economy*.

_____, 1999, "An economic theory of GATT," *American Economic Review*, Vol. 89, pp. 215–248.

_____, 1997a, "Multilateral tariff cooperation during the formation of free trade areas," *International Economic Review*, Vol. 38, pp. 291–319.

_____, 1997b, "Multilateral tariff cooperation during the formation of customs unions," *Journal of International Economics*, Vol. 42, pp. 91–123.

_____, 1990, "A theory of managed trade," *American Economic Review*, Vol. 80, pp. 779–795.

Baldwin, Robert E., 1985, *The Political Economy of U.S. Import Policy*. Cambridge, MA: MIT Press.

Bhagwati, Jagdish, 1992, "Regionalism versus multilateralism," *The World Economy*, Vol. 15, pp. 535–555.

Bond, Eric W., and Constantinos Syropolous, 1996, "The size of trading blocs: Market power and world welfare effects," *Journal of International Economics*, Vol. 40, pp. 411–437.

Bown, Chad P., and Meredith A. Crowley, 2003a, "Safeguards in the WTO," in *The Kluwer Companion to the World Trade Organization,* A. Appleton, P. Macrory, and M. Plummer (eds.), Dordrecht: Kluwer Academic, forthcoming.

_____, 2003b, "Trade deflection and trade depression," manuscript, August.

Brainard, S. Lael, and Thierry Verdier, 1997, "The political economy of declining industries: Senescent industry collapse revisited," *Journal of International Economics*, Vol. 42, pp. 221–237.

_____, 1994, "Lobbying and adjustment in declining industries," *European Economic Review*, Vol. 38, pp. 586–595.

Caves, Richard E., Jeffrey A. Frankel, and Ronald W. Jones, 2002, *World Trade and Payments: an Introduction*, (ninth edition), Boston: Addison-Wesley.

Clarida, Richard H., 1993, "Entry, dumping, and shakeout," *American Economic Review*, Vol. 83, pp. 180–202.

Clausing, Kimberly A., 2001, "Trade creation and trade diversion in the Canada–United States Free Trade Agreement," *Canadian Journal of Economics*, Vol. 34, pp. 677–696.

Collie, David, 1991, "Export subsidies and countervailing duties," *Journal of International Economics,* Vol. 31, pp. 309–324.

Crowley, Meredith A., 2002, "Do antidumping duties and safeguard tariffs open or close technology gaps?," Federal Reserve Bank of Chicago, working paper, No. WP-2002-13, July.

_____, 2001, "Antidumping policy under imperfect competition: Theory and evidence," Federal Reserve Bank of Chicago, working paper, No. WP-2001-21, December.

Dixit, Avinash, 1988, "Anti-dumping and countervailing duties under oligopoly," *European Economic Review*, Vol. 32, pp. 55–68.

Dixit, A. K., and V. Norman, 1980, *Theory of International Trade*, Cambridge, UK: Cambridge University Press.

Ethier, Wilfred J., 2002, "Unilateralism in a multilateral world," *Economic Journal*, Vol. 112, pp. 266–292.

_____, 1998, "Regionalism in a multilateral world," *Journal of Political Economy*, Vol. 106, pp. 1214–1245.

_____, 1982, "Dumping," *Journal of Political Economy*, Vol. 90, pp. 487–506.

Fischer, Ronald D., and Thomas J. Prusa, 1999, "Contingent protection as better insurance," National Bureau of Economic Research, working paper, No. 6933.

Freund, Caroline, 2000, "Different paths to free trade: The gains from regionalism," *Quarterly Journal of Economics*, pp. 1317–1341.

Gruenspecht, Howard K., 1988, "Dumping and dynamic competition," *Journal of International Economics*, Vol. 25, pp. 225–248.

Hansen, Wendy L., 1990, "The international trade commission and the politics of protection," *American Political Science Review*, Vol. 84, pp. 21–46.

Hillman, Arye, 1982, "Declining industries and political-support protectionist motives," *American Economic Review*, Vol. 72, pp. 1180–1187.

Hoekman, Bernard, and Michel Kostecki, 1995, *The Political Economy of the World Trading System*, Oxford: Oxford University Press.

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

Klimenko, Mikhail, Garey Ramey, and Joel Watson, 2002, "Recurrent trade agreements and the value of external enforcement," University of California, San Diego, mimeo.

Krugman, Paul R., and Maurice Obstfeld, 2000, *International Economics: Theory and Policy* (fifth edition), Reading, MA: Addison-Wesley.

Levy, Philip, 1997, "A political-economic analysis of free-trade agreements," *American Economic Review*, Vol. 87, pp. 506–519.

Magee, Chris, 2002, "Declining industries and persistent protection," *Review of International Economics*, Vol. 10.

Matsuyama, Kiminori, 1990, "Perfect equilibria in a trade liberalization game," *American Economic Review*, Vol. 80, pp. 480–492.

McLaren, John E., 2002, "A theory of insidious regionalism," *Quarterly Journal of Economics*, Vol. 117, pp. 571–608.

Miyagiwa, Kaz, and Yuka Ohno, 1999, "Credibility of protection and incentives to innovate," *International Economic Review*, Vol. 40, pp. 143–163.

_____, 1995, "Closing the technology gap under protection," *American Economic Review*, Vol. 85, pp. 755–770.

Prusa, Thomas J., 2001, "On the spread and impact of anti-dumping," *Canadian Journal of Economics*, Vol. 34, pp. 591–611.

Romalis, John, 2002, "NAFTA's and CUSFTA's impact on North American trade," University of Chicago, Graduate School of Business, mimeo, July.

Sampson, Gary P., 1996, "Compatibility of regional and multilateral trading agreements: Reforming the WTO process," *American Economic Review,* Vol. 86, pp. 88–92.

Spencer, Barbara J., 1988, "Capital subsidies and countervailing duties in oligopolistic industries," *Journal of International Economics*, Vol. 25, pp. 45–69.

Staiger, Robert, and Guido Tabellini, 1999, "Do GATT rules help governments make domestic commitments?," *Economics and Politics*, Vol. 11, pp. 109–144.

_____, 1987, "Discretionary trade policy and excessive protection," *American Economic Review*, Vol. 77, pp. 823–837.

Staiger, Robert, and Frank Wolak, 1992, "The effect of domestic antidumping law in the presence of foreign monopoly," *Journal of International Economics*, Vol. 32, pp. 265–287.

United States International Trade Commission, 1982, "The effectiveness of escape clause relief in promoting adjustment to import competition," USITC publication, No. 1229, investigation, No. 332-115, March.

Viner, Jacob, 1950, *The Customs Union Issue*, New York: Carnegie Endowment.

World Trade Organization (WTO), 2003, "United States—Definitive safeguard measures on imports of certain steel products: Final reports of the panel," Geneva, July 11.

_____, 2001, *International Trade Statistics*, Geneva: WTO.

_____, 1995a, *Analytical Index: Guide to GATT Law and Practice*, Vol. 2, Geneva: WTO.

_____, 1995b, *Uruguay Round Agreement Establishing the World Trade Organization*, Geneva: WTO.

Index for 2003

To order copies of any of these issues, or to receive a list of other publications, please telephone (312)322-5111 or write to: Federal Reserve Bank of Chicago, Public Information Center, P.O. Box 834, Chicago, IL 60690-0834. The articles are also available to download in PDF format from the Bank's website at www.chicagofed.org/publications/economicperspectives/index.cfm.