An Unbiased Appraisal of Purchasing Power Parity
Paul Cashin and C. John McDermott  •  321

Is Africa Integrated in the Global Economy?
Arvind Subramanian and Natalia T. Tamirisa  •  352

The High-Yield Spread as a Predictor of Real Economic Activity:
Evidence of a Financial Accelerator for the United States
Ashoka Mody and Mark P. Taylor  •  373

The Art of Making Everybody Happy:
How to Prevent a Secession
Michel Le Breton and Shlomo Weber  •  403

Rationing Rules and Outcomes:
The Experience of Singapore’s Vehicle Quota System
Ling Hui Tan  •  436

“Big Bang” Versus Gradualism in Economic Reforms:
An Intertemporal Analysis with an Application to China
Andrew Feltenstein and Saleh M. Nsouli  •  458

Structural Vulnerabilities and Currency Crises
Swati R. Ghosh and Atish R. Ghosh  •  481

Index
Volume 50  •  507
The objective of *IMF Staff Papers* is to publish high-quality research produced by IMF staff and invited guests on a variety of topics of interest to a broad audience including academics and policymakers in the member countries of the Fund. The papers selected for publication in the journal are subject to an extensive review process using both internal and external referees. *IMF Staff Papers* also welcomes outside comments, criticisms, and interesting replications of published work. The views presented in published papers are those of the authors and do not necessarily reflect the position of the Executive Board or of the IMF, or any other organization mentioned herein.

Subscription: US$72.00 a volume or the approximate equivalent in the currencies of most countries. Three issues constitute a volume. Single copies may be purchased at $25. Individual academic rate to full-time professors and students of universities and colleges: $46 a volume. Subscriptions and orders should be sent to:

**International Monetary Fund**
Publication Services
700 19th Street, N.W.
Washington, D.C. 20431, U.S.A.

Telephone: (202) 623-7430
Fax: (202) 623-7201
E-mail: publications@imf.org
Internet: http://www.imf.org
EDITOR'S NOTE

The Editor invites from contributors outside the IMF brief comments (not more than 1,000 words) on published articles in IMF Staff Papers. These comments should be addressed to the Editor, who will forward them to the author of the original article for reply. Both the comments and the reply will be considered for publication.

The data underlying articles published in IMF Staff Papers are available on the journal’s website (http://www.imf.org/staffpapers). Readers are invited to use these data to expand on the material in the articles, and the journal will consider publishing such work.

© 2003 by the International Monetary Fund
ISBN 1-58906-203-5
International Standard Serial Number: ISSN 1020-7635

This serial publication is catalogued as follows:

International Monetary Fund
IMF staff papers — International Monetary Fund. v. 1–Feb. 1950–[Washington] International Monetary Fund.

v. tables, diagrs. 26 cm.

Three no. a year, 1950–1977; four no. a year, 1978–

Indexes:
ISSN 1020-7635 = IMF staff papers — International Monetary Fund.
1. Foreign exchange—Periodicals. 2. Commerce—Periodicals.

HG3810.15 332.082 53-35483
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Unbiased Appraisal of Purchasing Power Parity</td>
<td>Paul Cashin and C. John McDermott</td>
<td>321</td>
</tr>
<tr>
<td>The High-Yield Spread as a Predictor of Real Economic Activity:</td>
<td>Ashoka Mody and Mark P. Taylor</td>
<td>373</td>
</tr>
<tr>
<td>Evidence of a Financial Accelerator for the United States</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Art of Making Everybody Happy:</td>
<td>Michel Le Breton and Shlomo Weber</td>
<td>403</td>
</tr>
<tr>
<td>How to Prevent a Secession</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationing Rules and Outcomes:</td>
<td></td>
<td>436</td>
</tr>
<tr>
<td>The Experience of Singapore’s Vehicle Quota System</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Big Bang” Versus Gradualism in Economic Reforms:</td>
<td>Andrew Feltenstein and Saleh M. Nsouli</td>
<td>458</td>
</tr>
<tr>
<td>An Intertemporal Analysis with an Application to China</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural Vulnerabilities and Currency Crises</td>
<td>Swati R. Ghosh and Atish R. Ghosh</td>
<td>481</td>
</tr>
</tbody>
</table>

Index

©International Monetary Fund. Not for Redistribution
This page intentionally left blank
An Unbiased Appraisal of Purchasing Power Parity

PAUL CASHIN and C. JOHN MCDERMOTT*

Univariate studies of the hypothesis of unit roots in real exchange rates have yielded consensus point estimates of the half-life of deviations from purchasing power parity (PPP) of between three to five years (Rogoff, 1996). However, conventional least-squares-based estimates of half-lives are biased downward. Accordingly, as a preferred measure of the persistence of real exchange rate shocks we use median-unbiased estimators of the half-life of deviations from parity, which correct for the downward bias of conventional estimators. We study this issue using real effective exchange rate (REER) data for 20 industrial countries in the post–Bretton Woods period. The serial correlation-robust median-unbiased estimator yields a cross-country average of half-lives of deviations from parity of about eight years, with the REER of several countries displaying permanent deviations from parity. However, using the median-unbiased estimator that is robust to the moving average and heteroskedastic errors present in real exchange rate data reduces the estimated half-life of parity deviations. Using this unbiased estimator, we find that the majority of countries have finite point estimates of half-lives of parity deviation, which is supportive of PPP holding in the post–Bretton Woods period. We also find that the average bias-corrected half-life of parity deviations is about five years, which is consistent with (but at the upper end of) Rogoff’s (1996) consensus estimate of the half-life of deviations from parity. [JEL C22, F31]

Do real exchange rates really display parity-reverting behavior? In summarizing the results from studies using long-horizon data, Froot and Rogoff (1995) and Rogoff (1996) report the current consensus in the literature that the half-life of a shock (the time it takes for the shock to dissipate by 50 percent) to the real exchange rate is about three to five years, implying a slow parity reversion

*Paul Cashin is a Senior Economist in the Macroeconomic Studies Division of the IMF’s Research Department. John McDermott is Chief Economist at the National Bank of New Zealand. The authors thank Peter Clark, Robert Flood, Geoffrey Kingston, Sam Ouliaris, Peter Phillips, Kenneth Rogoff, Miguel Savastano, Peter Wickham, seminar participants at the International Monetary Fund, the Sixth Australian Macroeconomics Workshop (University of Adelaide), and New Zealand Exchange Rate Workshop (Victoria University of Wellington), and an anonymous referee for their comments and suggestions. Nese Erhul and Chi Nguyen provided excellent research assistance.
rate of between 13 to 20 percent per year.\(^1\) Such a slow speed of reversion to purchasing power parity is difficult to reconcile with nominal rigidities and, as pointed out by Rogoff (1996), is also difficult to reconcile with the observed large short-term volatility of real exchange rates.

In earlier work, Meese and Rogoff (1983) demonstrated that a variety of linear structural exchange rate models failed to forecast more accurately than a naive random walk model for both real and nominal exchange rates. If the real exchange rate follows a random walk, then innovations to the real exchange rate persist and the time series can fluctuate without bound. This result is contrary to the theory of purchasing power parity (PPP), which at its most basic level states that there is an equilibrium level to which exchange rates converge, such that foreign currencies should possess the same purchasing power.\(^2\)

Notwithstanding the above-mentioned consensus in the literature on the speed of parity reversion, the conclusion of Meese and Rogoff (1983) has been reached in many subsequent studies of the time-series properties of the real exchange rate. This conclusion has usually been derived using formal statistical tests that failed to reject the null hypothesis of a unit root in the real exchange rate against the alternative of a stationary autoregressive (AR) model. If the unit root model can characterize real exchange rate behavior, then PPP does not hold because there is no propensity to revert back to any equilibrium level.

The empirical literature on testing the existence of PPP has developed in tandem with developments in the unit root econometrics literature, and has taken several paths. First, a standard rationale for the inability of researchers to clearly reject the unit root null, especially in the post–Bretton Woods period, is that unit root tests have low power because of the relatively short sample periods under study. In response, long-run data of a century or more, which span several exchange rate regimes, have been analyzed to improve the power of unit root tests (see Frankel, 1986; and Lothian and Taylor, 1996, among others). Second, panel data methods have been used in an attempt to increase the power of unit root tests (see Frankel and Rose, 1996; and Wu, 1996, among others).\(^3\) Third, as PPP implies cointegration between the nominal exchange rate, domestic price level, and foreign price level, multivariate tests of the null hypothesis of no cointegration between these three variables have been carried out (see Corbae and Ouliaris.

\(^1\)Abuaf and Jorion (1990) use data on bilateral real exchange rates between the United States and several industrial countries during the twentieth century, and find average half-lives of deviations from parity of a little over three years. Frankel (1986) and Lothian and Taylor (1996) use two centuries of annual data on the sterling-dollar real exchange rate in calculating half-lives of about five years. Wu (1996) and Papell (1997) use panel data methods on quarterly post–Bretton Woods data to derive half-lives of between two to three years.

\(^2\)The version of PPP with the longest pedigree is that of relative PPP, which states that the exchange rate will be proportional to the ratio of money price levels (including traded and nontraded goods) between countries, that is, to the relative purchasing power of national currencies (see Wickham, 1993). For earlier surveys on PPP and exchange rate economics, see Isard (1995), Froot and Rogoff (1998), and Sarno and Taylor (2002).

\(^3\)The results of the recent burst of activity in the conduct of univariate tests of PPP have been characterized by Taylor (2001) as either in the “whittling down half-lives” camp (such as Frankel and Rose, 1996; and Wu, 1996) or the “whittling up half-lives” camp (such as Papell, 1997; O’Connell, 1998; and Engel, 2000).
1988; and Edison, Gagnon and Melick, 1997, among others). Both the long-run data and panel approaches have produced results that more frequently reject the unit root null for real exchange rates, while (particularly for post–Bretton Woods data) the results from cointegrating regressions have varied widely in their ability to reject the null of no cointegration (Froot and Rogoff, 1995).

However, one aspect of unit root econometrics that has been largely neglected in the PPP literature is the problem of "near unit root bias," which biases empirical results in favor of finding PPP. An important pitfall in using the autoregressive or unit root model to analyze the persistence of shocks to the real exchange rate is that standard estimators, such as least squares, are significantly downwardly biased in finite samples. This downward bias of least-squares estimates of autoregressive parameters becomes particularly acute when the autoregressive parameter is close to unity—in this case the process is close to being nonstationary and, as the least-squares estimator minimizes the regression residual variance, it will tend to make the data-generating process appear to be more stationary than it actually is by forcing the autoregressive parameter away from unity. As lower values of the autoregressive parameter imply faster speeds of adjustment following a shock, this will also result in a downward bias to least-squares-based estimates of half-lives of shocks. This near unit root bias is likely to be particularly relevant for real exchange rates, as they are often found to be stationary, yet exhibit shocks that are highly persistent.

In this paper we will generate a transformation from some initial estimator to a median-unbiased estimator, in order to correct for the near unit root bias. Median-unbiased estimators for autoregressive models have been proposed by Andrews (1993), Andrews and Chen (1994), McDermott (1996), and earlier by Rudebusch (1992), and Stock (1991). The median-unbiased estimators employed in this paper are based on initial estimators that include Dickey-Fuller, Augmented Dickey-Fuller, and Phillips-Perron unit root regressions. Importantly, this paper is the first to use the initial estimators proposed by Phillips (1987) and Phillips and Perron (1988), which, given they are robust to weakly dependent and heterogeneously distributed time series, are better suited than alternative median-unbiased estimators for modeling real exchange rates.

Conventional analyses of whether real exchange rates are better modeled as stationary or random walk processes typically focus on whether real exchange rate shocks are mean-reverting (finite persistence) or not (permanent). Such tests of the null hypothesis of a unit root in real exchange rates are rather uninformative as to the speed of parity reversion, because a rejection of the unit root null could still be consistent with a stationary model of real exchange rates that has highly persistent shocks. In contrast, this paper concentrates on measuring the duration of shocks to

3While this bias is certainly present in standard (least-squares) estimation of the unit root model, panel-data methods (such as those applied by Wu, 1996; Papell, 1997; and Taylor and Sarno, 1998), which pool cross-country information, will also be subject to the near unit root bias that raises the estimated speed of reversion to parity (see Cermeño, 1999). In addition, the use of panel unit root tests has also been criticized because authors typically assume that rejection of the joint null hypothesis of unit root behavior of the whole panel of real exchange rates implies that all real exchange rates are stationary, when in actuality it only implies that at least one of them is stationary (see Taylor and Sarno, 1998). This bolsters the case for using univariate methods, as is done in the present paper.
the real exchange rate and associated confidence intervals, and undertakes no hypothesis tests as to the suitability of the assumption of a unit root as the process governing the evolution of real exchange rates. Instead of unit root tests, in this paper we characterize the extent of parity reversion in terms of point and interval estimates of the half-life of deviations from purchasing power parity, where the half-life is defined as the duration of time required for half the magnitude of a unit shock to the level of a series to dissipate. Point and interval estimators are useful statistics for providing information to draw conclusions about the relevance of PPP, as unlike hypothesis tests they are informative when a hypothesis is not rejected, and will be used in this paper.

The contributions of this paper are fourfold. First, the median-unbiased estimator of Andrews (1993) and Andrews and Chen (1994) is used to obtain point and interval estimates of the autoregressive parameter in the real exchange rate data. These median-unbiased estimators are generated from a transformation of an initial estimator, and correct for the downward bias in conventional (least-squares) estimation of the autoregressive parameter in unit root models. Second, we follow McDermott (1996) and use median-unbiased estimators that allow for initial estimators that display a wider class of error processes (particularly a moving average error structure) than previously considered in the literature. In particular, we follow Baillie and Bollerslev (1989) and Lothian and Taylor (1996, 2000) and are careful to use heteroskedasticity-robust estimation methods, as it is well known that exchange rate series exhibit both serial correlation and time-dependent heteroskedasticity. In particular, our preferred median-unbiased estimator uses an initial estimator that is based on autoregressive models of Phillips (1987) and Phillips and Perron (1988), which are robust to a wide variety of weakly dependent and heterogenously distributed time series. Third, these unbiased estimates of the autoregressive parameter and associated impulse response functions are used to calculate an unbiased scalar measure of the average duration (in terms of half-lives) and range of typical real exchange rate shocks. Fourth, using Andrews’ (1993) unbiased model-selection rule we can overcome the low power problems inherent in conventional unit root tests of PPP, and be more definitive about our willingness to draw conclusions as to the presence or absence of parity reversion of real exchange rates in the post-Bretton Woods period.

Our main results may be summarized briefly. First, using post-Bretton Woods data on the real effective exchange rates of 20 industrial countries and conventional (least-squares) biased estimation of unit root models, we replicate Rogoff’s (1996) consensus estimate of the half-life of deviations from purchasing power parity (PPP) of between three to five years. Second, we find that serial correlation-robust median-unbiased point estimates of the half-lives of deviations of real exchange rates from PPP in the post-Bretton Woods period are typically longer than the previous consensus allows for, with cross-country average (median) half-lives of parity deviation lasting about eight years. In particular, we find that for at least 5 of the 20 countries in our sample, deviations of the real exchange rate from parity are best viewed as being permanent. Third, notwithstanding this result, when we use the median-unbiased estimator that is robust to the moving average and heteroskedastic errors present in real exchange rate data, we find that the

324
majority of countries have finite point estimates of half-lives of parity deviation, yet with wide confidence intervals around these point estimates. Using an unbiased model-selection rule, for these countries with finite point estimates of half-lives of deviation from parity there is evidence of reversion (albeit slow) of real exchange rates to parity, which is consistent with PPP holding in the post-Bretton Woods period. Fourth, we find that the average heteroskedasticity-robust median-unbiased point estimates of the half-life of parity deviation is about five years, which is shorter in duration than those derived from median-unbiased estimators that fail to account for heteroskedasticity. Such a duration of parity reversion is consistent with (but at the upper end of) Rogoff’s (1996) consensus estimate of the (downwardly biased) half-life of deviations of real exchange rates from parity. In summary, while median-unbiased methods always increase the estimated half-life of deviations from PPP in comparison with those derived from conventional, downwardly biased (least-squares) methods, allowing for heteroskedasticity reduces the bias-corrected estimated half-life of parity deviations.

I. Biased and Unbiased Measures of Half-Lives of Shocks to Parity

The existence of long-run PPP is inconsistent with unit roots (infinite half-lives of parity deviation) in the real exchange rate process. This notion has stimulated the growth of a large literature, using various tests, to resolve whether PPP holds in the post-Bretton Woods period. However, analyses of the trend-stationary or difference-stationary dichotomy of standard unit root tests focus only on whether such shocks are mean-reverting (finite persistence) or not (permanent). For economists, long-run PPP means more than the absence of a unit root—it also means the presence of a sufficient degree of mean reversion in exchange rates (over the horizon of interest) to validate the theoretical predictions of models based on the PPP assumption. For example, using the Dornbusch (1976) overshooting model, which has plausible assumptions about nominal wage and price rigidities, we would expect substantial convergence of real exchange rates to PPP over one to two years. Rather than use unit root tests to evaluate PPP it is preferable to use a scalar measure of the speed of reversion of real exchange rate shocks, and recent papers examining the post-Bretton Woods period have used estimates of the half-life of deviations from PPP to do so (Andrews, 1993; Andrews and Chen, 1994; and Cheung and Lai, 2000b).

To estimate the speed of convergence to purchasing power parity (PPP) most researchers use the first-order autoregressive (AR(1)) model of the univariate time series \( \{q_t; t = 0, \ldots, T\} \), assuming independent identically distributed normal errors. The model considered is

\[
q_t = \mu + \alpha q_{t-1} + \varepsilon_t \quad \text{for} \quad t = 1, \ldots, T, \tag{1}
\]

Biased and median-unbiased point and interval estimates of the half-life of shocks to economic time series have also been used by Cashin, Liang, and McDermott (2000) in modeling the persistence of shocks to world commodity prices, and by Cashin, McDermott, and Pattillo (forthcoming) in modeling the persistence of terms of trade shocks. Earlier, Stock (1991) considered point and asymptotic confidence intervals for the largest autoregressive root in a time series.
where $q_t : t = 0, \ldots, T$ is the real exchange rate, $\mu$ the intercept, $\alpha$ the autoregressive parameter (where $\alpha \in (-1, 1)$), and $\varepsilon_t$ are the innovations of the model. A time trend is usually not included in equation (1), as a trend would not be consistent with long-run PPP (which imposes the restriction that real exchange rates have a constant unconditional mean). This model is the same as that used for testing whether there is a unit root in a time series—consequently, this model is often referred to as the Dickey-Fuller (1979) regression. The half-life, which is the time it takes for a deviation from PPP to dissipate by 50 percent, is calculated from the autoregressive parameter, $\alpha$ (see below for details).

### Problems with Conventional (Least-Squares) Measures of the Half-Life of Shocks to Parity

There are three problems with using these least-squares-based half-lives as evidence of the persistence of PPP deviations: biased autoregressive parameter coefficients; no confidence interval around the half-life measures; and serially correlated and heteroskedastic errors. We discuss each of these problems in turn.

#### Downwardly Biased Autoregressive Parameters

First, it has been known since the work of Orcutt (1948) that least-squares estimates of lagged dependent variable coefficients (such as the autoregressive parameter in the Dickey-Fuller regression) will be biased toward zero in small samples. The literature on the bias of least-squares estimation of autoregressive models is an old one. Marriott and Pope (1954) established the mean bias of the least squares estimator for the stationary AR(1) model, as did Shaman and Stine (1988) for stationary AR(p) models. While least squares will be the best linear unbiased estimator under the Gauss-Markov theorem, in the autoregressive case the assumptions of this theorem are violated, as lagged values of the dependent variable cannot be fixed in repeated sampling, nor can they be treated as distributed independently of the error term for all lags. Marriott and Pope (1954) showed that, ignoring second-order terms, the expected value of the least-squares estimate of the true $\alpha$ in the AR(1) model of equation (1) can be approximated by:

$$E(\hat{\alpha}) = \alpha - (1 + 3\alpha)/N,$$

where $N = T - 1$. Using simulation calculations, Orcutt and Winokur (1969) find that, for $T = 40$ and true $\alpha = 1$, the least-squares mean bias is $E(\hat{\alpha}) - \alpha = 0.129$. Similarly, the simulation calculations of Andrews (1993) reveal that the least-squares median bias of equation (1), again for $T = 40$ and true $\alpha = 1$, is slightly smaller at 0.107. In general, larger the true value of $\alpha$, the larger the least-squares bias, and so the bias is largest in the unit root case. The bias shrinks as the sample size grows, as the estimate converges to the true population value.

The downward bias in least-squares estimates of the autoregressive parameter arises because there is an asymmetry in the distribution of estimators of the auto-

---

*Time trends are sometimes included in tests of PPP in an attempt to control for the Balassa-Samuelson effect, where the failure of PPP to hold can be due to differential rates of productivity growth in the tradable and nontradable sectors.*

©International Monetary Fund. Not for Redistribution
regressive parameter in AR models. The distribution is skewed to the left, resulting in the median exceeding the mean. As a result, the median is a better measure of central tendency than the mean in least-squares estimates of Dickey-Fuller models. The exact median-unbiased estimation procedure proposed by Andrews (1993) can be used to correct this bias. The bias correction delivers an impartiality property to the decision-making process, because there is an equal chance of under- or overestimating the autoregressive parameter in the unit root regression. Moreover, an unbiased estimate of $\alpha$ will allow us to calculate an unbiased scalar estimate of persistence—the half-life of a unit shock.

**Lack of Confidence Intervals Around Point Estimates of Half-Lives**

Second, reporting only point estimates of the half-lives provides an incomplete picture of the speed of convergence toward PPP. To gain a more complete view one should use interval estimates. Fortunately, median-unbiased estimation allows for the calculation of median-unbiased confidence intervals. Moreover, interval estimation addresses the low power problem, usually associated with unit root tests (DeJong and others, 1992), by informing us whether we are failing to reject the null because it is true or because there is too much uncertainty as to the true value of the autoregressive parameter.

**Serially Correlated and Heteroskedastic Errors**

Third, the presence of serial correlation (typical in economic time series) means that the Dickey-Fuller regression will often not be appropriate. In such cases, we can follow Andrews and Chen (1994) and use an AR($p$) model, which adds lagged first differences to account for serial correlation. The AR($p$) model (also known as an Augmented Dickey-Fuller regression) takes the form

$$q_t = \mu + \alpha q_{t-1} + \sum_{i=1}^{p-1} \psi_i \Delta q_{t-i} + \epsilon_t \quad \text{for } t = 1, \ldots, T,$$

where the observed real exchange rate series is $q_t: t = -p, \ldots, T$. Andrews and Chen (1994) show how to perform approximately median-unbiased estimation of autoregressive parameters in Augmented Dickey-Fuller regressions.

While Andrews (1993) assumes iid errors and Andrews and Chen (1994) assume AR($p$) errors, neither approach allows for the possibility of a moving average error structure (unless $p \to \infty$ as $T \to \infty$). If moving average and heteroskedastic errors are present, as is typical for real exchange rate data, then even the Andrews-Chen method may not account for the biases arising from these attributes of the data. One method that can deal with more general error processes than those used in previous work is the semi-nonparametric technique of the Phillips-Perron (1988) unit root regression, which estimates the model of equation (1) and accounts for a wide range of serial correlation and heteroskedasticity using nonparametric methods.

It is well known that selection of the lag truncation point for the spectral density at zero frequency can have a large impact on the estimated spectral density and,
therefore, on the Phillips-Perron test (Andrews, 1991). Monte Carlo studies by Phillips and Xiao (1998) and Cheung and Lai (1997) find that the small-sample properties of the Phillips-Perron unit root test are significantly improved if prewhitened kernel estimation of the long-run variance parameter (following Andrews and Monahan, 1992) occurs prior to the use of any data-determined bandwidth selection procedure (such as that of Andrews, 1991). In particular, with the combined use of these two procedures, the Phillips-Perron test performs better than (or at least as well as) the Augmented Dickey-Fuller test in terms of comparative power and yields tighter confidence intervals. We implement the bias-corrected Phillips-Perron regression in this paper, using the approach initially set out by McDermott (1996).

Bias-Correcting Estimates of the Autoregressive Parameter and the Model Selection Rule

Andrews (1993) presents a method for median-bias correcting the least-squares estimator. To calculate the median-unbiased estimator of $\alpha$, suppose $\hat{\alpha}$ is an estimator of the true $\alpha$ whose median function $(m(\alpha))$ is uniquely defined $\forall \alpha \in (-1,1]$. Then $\hat{\alpha}_u$ (the median-unbiased estimator of $\alpha$) is defined as:

$$
\hat{\alpha}_u = \begin{cases} 
1 & \text{if } \hat{\alpha} > m(1), \\
-1 & \text{if } \hat{\alpha} \leq m(-1), \\
m^{-1}(\hat{\alpha}) & \text{if } m(-1) < \hat{\alpha} \leq m(1), 
\end{cases}
$$

(3)

where $m(-1) = \lim_{\alpha \to -1} m(\alpha)$, and $m^{(-1)}(m(-1), m(1)) \to (-1,1)$ is the inverse function of $m(.)$ that satisfies $m^{-1}(m(\alpha)) = \alpha$ for $\alpha \in (-1,1]$. That is, if we have a function that for each true value of $\alpha$ yields the median value (0.50 quantile) of $\hat{\alpha}$, then we can simply use the inverse function to obtain a median-unbiased estimate of $\alpha$. Intuitively, we find the value of $\alpha$ that results in the least-squares estimator having a median value of $\hat{\alpha}$. For example, if the least-squares estimate of $\alpha$ equals 0.8, then we do not use that estimate, but instead use that value of $\alpha$ that results in the least-squares estimator having a median of 0.8. The extent of the median bias rises with the persistence of the innovations, which is particularly important for near unit root series such as the real exchange rate, which in the literature have previously (using point estimates of $\alpha$) been found to be stationary, yet with shocks that are rather persistent.8

\footnote{8In contrast, the results of De Jong and others (1992)—that the semi-parametric Phillips-Perron unit root test has low power when there is positive serial correlation—were obtained using Phillips-Perron estimators with arbitrarily fixed bandwidth selection and without prewhitening. Choi and Chung (1995) also find that using the Andrews (1991) automatic bandwidth selection procedure for the Phillips-Perron test results in the Phillips-Perron test being more powerful than the Augmented Dickey-Fuller test.}

\footnote{9The size of the bias correction can be large, especially when $\alpha$ is close to one. For example, for a sample size of 60 observations using the AR(1) model of equation (1), a least-squares estimate of $\alpha = 0.80$ would correspond to a median-unbiased estimate of $\alpha = 0.85$.}

\footnote{9Other sources of bias in estimation of unit root regressions have been examined in the large literature on PPP, and will not be examined in this paper. These include large size bias in univariate tests for long-run PPP, due to a significant unit root component in the relative price of nontraded goods (Engel, 2000); size bias in multivariate tests for long-run PPP due to a failure to control for cross-sectional correlation (O’Connell, 1998); and sample-selection bias of the countries analyzed, which biases the results toward understating the general relevance of parity reversion (Cheung and Lai, 2000a).}
Model Selection Rule

The median-unbiased estimator can also be used to derive an unbiased model-selection rule, where for any correct model the probability of selecting the correct model is at least as large as the probability of selecting each incorrect model (Andrews, 1993; Andrews and Chen, 1994). Suppose the problem is to select one of two models defined by $\alpha \in I_a$ and $\alpha \in I_b$, where $I_a$ and $I_b$ are intervals partitioning the parameter space $(-1, 1)$ for $\alpha$, with $I_a = (-1, 1)$ and $I_b = \{1\}$. Then the unbiased model selection rule would indicate that model $I_m$ should be chosen if $\hat{\alpha}_m \in I_m$ for $m = a, b$. This is also a valid level 0.50 (unbiased) test of the $H_0: \alpha \in I_a$ versus $H_1: \alpha \in I_b$.

Importantly, the median-unbiased estimator $\hat{\alpha}_m$ is the lower and upper bounds of the two one-sided 0.5 confidence intervals for the true $\alpha$ when $m(\cdot)$ is strictly increasing (Andrews, 1993, p. 152). These confidence intervals have the property that their probabilities of encompassing the true $\alpha$ are one-half. That is, there is a 50 percent probability that the confidence interval from minus one to $\hat{\alpha}_m$ contains the true $\alpha$, and a 50 percent probability that the confidence interval from $\hat{\alpha}_m$ to one contains the true $\alpha$. For example, if $\hat{\alpha}_m = 0.90$, then the probability that the true $\alpha$ is less than 0.90 is one-half, and the probability that the true $\alpha$ exceeds 0.90 is also one-half.

Based on the median-unbiased estimate of $\alpha$, other tests with different size and power properties can also be constructed. Using the 0.05 and 0.95 quantile functions of $\hat{\alpha}$ we can construct two-sided 90 percent confidence intervals or one-sided 95 percent confidence intervals for the true $\alpha$. These confidence intervals can be used either to provide a measure of the accuracy of $\hat{\alpha}$ or to construct the conventional exact one- or two-sided tests of the null hypothesis that $\alpha = \alpha_0$. In this paper we use such confidence intervals only to provide a measure of the accuracy of $\hat{\alpha}$.

In a Monte Carlo study of the AR($p$) model, Andrews and Chen (1994, p. 194) demonstrate that the unbiased model-selection rule has a probability of correctly selecting the unit root model (when the true $\alpha = 1$) of about 0.5. This is much lower than the corresponding probability for a (two-sided) level 0.10 test or (one-sided) level 0.05 test of a unit root null hypothesis, as the unbiasedness condition does not (unlike the level 0.10 or 0.05 tests) give a bias in favor of the unit root model. The greater size of Andrews’ unbiased model selection rule, in comparison with conventional tests, increases the probability of rejecting the unit root null. This indicates that if the true $\alpha < 1$, then the probability of a type II error (failure to reject the unit root model when it is false) is smaller for Andrews’ model selection rule than for conventional tests, especially for the near unit root case.

The unbiased model selection procedure based on the median-unbiased estimate of the AR(1) model is an exact test, as are its associated confidence intervals. However, the unbiased model selection procedure based on the median-unbiased estimate of the AR($p$) model is an approximate test, as are its associated confidence intervals. This is because the distribution of $\hat{\alpha}_m$ calculated from the AR($p$) model depends on the true values of the $\psi$ terms in equation (2), which are unknown. Andrews and Chen (1994) demonstrate that the approximately median-unbiased point and interval estimates of $\alpha$ in the AR($p$) model are very close to being median-unbiased. A similar unbiased model selection procedure can be invoked for the Phillips-Perron regression, and this is also an approximate test because of the need to estimate the serial correlation correction.
Calculating Half-Lives

Our interest in this paper concerns the persistence of shocks to economic time series. In this connection, the impulse response function of a time series \( \{q_t: t = 1, 2, \ldots\} \) measures the effect of a unit shock occurring at time \( t \) (that is, \( \varepsilon_t \rightarrow \varepsilon_{t+1} \) in equations (1) and (2)) on the values of \( q_t \) at the future time periods \( t+1, t+2, \ldots \). This function quantifies the persistence of shocks to individual time series. For the AR(1) model the impulse response function is given by

\[
IR(h) = \alpha^h \text{ for } h = 0, 1, 2, \ldots
\]

For an AR(\( p \)) model the impulse response function is given by

\[
IR(h) = f_{11}^{(h)} \text{ for } h = 0, 1, 2, \ldots,
\]

where \( f_{11}^{(h)} \) denotes the \((1,1)\) element of \( F^h \) and where \( F \) is the \((p \times p)\) matrix

\[
F = \begin{bmatrix}
\alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{p-1} & \alpha_p \\
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix}
\]

However, rather than consider the whole impulse response function to gauge the degree of persistence, we use a scalar measure of persistence that summarizes the impulse response function: the half-life of a unit shock (HLS). For the AR(1) model (with \( \alpha \geq 0 \)), the HLS gives the length of time until the impulse response of a unit shock is half its original magnitude. and is defined as \( \text{HLS} = \text{ABS}(\log(1/2)/\log(\alpha)) \). Since median-unbiased estimates of \( \alpha \) have the desirable property that any scalar measures of persistence calculated from them (such as half-lives) will also be median unbiased, we can calculate the median-unbiased estimate of HLS by inserting the median-unbiased estimate of \( \alpha \) in the formula for HLS. Similarly, the 90 percent confidence interval of the exactly median-unbiased Dickey-Fuller estimate of the HLS is calculated using the 0.05 and 0.95 quantiles of \( \hat{\alpha} \) in the formula for HLS.

Median-unbiased point and confidence intervals for the HLS are calculated in a similar fashion for the Phillips-Perron estimator of equation (1), under the assumption that the impulse response function can be approximated by an AR(1) process.\(^{11}\) The median-unbiased estimate of the HLS is calculated using the

---

\(^{11}\)In calculating the point and interval estimates of the HLS it is assumed that the nonparametric adjustment of the Phillips-Perron regression removes the serial correlation, and what is left is a pure AR(1) process. Given this assumption holds, the HLS can then be calculated from the “noise reduced” impulse response function. This approximation is required because the true impulse response function is unobservable, as a parametric form of the time-series model does not exist. The results from the Augmented Dickey-Fuller regression suggest that this approach is a reasonable one, because the non-monotonicity of the impulse response function occurs at low lags and the shape of the impulse response is dominated by the AR(1) component.
median-unbiased Phillips-Perron estimate of $\alpha$ (following the approach of McDermott, 1996) in the formula for the HLS. Similarly, the 90 percent confidence interval of the median-unbiased Phillips-Perron estimate of the HLS is calculated using the 0.05 and 0.95 quantiles of $\hat{\alpha}$ in the formula for the HLS.\(^{12}\)

The half-life derived from the values of $\alpha$ assumes that shocks decay monotonically. While appropriate for the AR(1) model, this assumption is inappropriate for an AR($p$) model (with $p > 1$), since in general shocks to an AR($p$) will not decay at a constant rate. The approximately median-unbiased point estimate of the half-life for AR($p$) models (such as the Augmented Dickey-Fuller regression) can be calculated from the impulse response functions of equation (5), with the half-life defined as the time it takes for a unit impulse to dissipate permanently by one-half from the occurrence of the initial shock (Cheung and Lai, 2000b). Similarly, the 90 percent confidence interval of the approximately median-unbiased estimate of the half-life is calculated using the 0.05 and 0.95 quantiles of $\hat{\alpha}$, calculated again as the time it takes for a unit impulse to dissipate permanently by one-half from the occurrence of the initial shock.

As with the estimation of $\alpha$, the median-unbiased half-lives and confidence intervals can be interpreted in two ways. Using the Andrews unbiased model-selection rule, there is a 50 percent probability that the confidence interval from zero to the estimated median half-life contains the true half-life of a shock to any given time series, and a 50 percent probability that the confidence interval from the estimated median half-life to infinity contains the true half-life of a shock to any given time series. Alternatively, we can use the 90 percent confidence interval to indicate the range that has a 90 percent probability of containing the true half-life of a shock to any given time series.

\section*{II. Data and Empirical Results}

In this section we will investigate the persistence properties of the real exchange rate. The theory of relative PPP holds that the exchange rate will be proportional to the ratio of money price levels (including traded and nontraded goods) between countries, which implies that changes in relative price levels will be offset by changes in the exchange rate. By examining the persistence properties of real exchange rates we can determine whether real exchange rates do converge to their equilibrium relative PPP value in the long run, and thus determine whether PPP is consistent with the data.

The data used to estimate the near unit root model are monthly time series of the real exchange rate obtained from the International Monetary Fund’s International Financial Statistics (IFS) over the sample 1973:4 to 2002:4 (the post-Bretton Woods period), which gives a total of 349 observations. The definition of the real exchange rate is the real effective exchange rate (REER) based on consumer prices (line rec), for which 20 industrial countries were selected. As such,\(^{12}\)

\(^{12}\)Both the Dickey-Fuller and Phillips-Perron median-unbiased measures of persistence and associated confidence intervals can be compared with their least-squares counterparts, where the least-squares point and interval estimates will (given they are functions of a downwardly biased $\hat{\alpha}$) tend to understate the actual amount of persistence in shocks to economic time series (see Section II).
we will examine the behavior of REER based on (i) the nominal effective exchange rate, which is the trade-weighted average of bilateral exchange rates vis-à-vis trading partners' currencies; (ii) the domestic price level, which is the consumer price index; and (iii) the foreign price level, which is the trade-weighted average of trading partners' consumer price indices. We analyze effective rather than bilateral real exchange rates as the effective rate measures the international competitiveness of a country against all its trade partners, and helps to avoid potential biases associated with the choice of base country in bilateral real exchange rate analyses.

The REER indices measure how nominal effective exchange rates, adjusted for price differentials between the home country and its trading partners, have moved over a period of time. The consumer price index (CPI)-based REER indicator is calculated as a weighted geometric average of the level of consumer prices in the home country relative to that of its trading partners, expressed in a common currency. The IMF's CPI-based REER indicator (base 1995 = 100) of country \( i \) is defined as

\[
q_i = \prod_{j 
eq i} \left[ \frac{P_j R_j}{P_i R_i} \right]^{W_{ij}},
\]

where \( j \) is an index that runs over country \( i \)'s trade partner (or competitor) countries; \( W_{ij} \) is the competitiveness weight attached by country \( i \) to country \( j \), which is based on 1988–90 average data on the composition of trade in manufacturing, non-oil primary commodities, and tourism services; \( P_j \) and \( P_i \) are the seasonally adjusted consumer price indices in countries \( i \) and \( j \); and \( R_i \) and \( R_j \) are the nominal exchange rates of the currencies of countries \( i \) and \( j \) in U.S. dollars. As shown by McDermott (1996), alternative measures of the real exchange rate, such as real bilateral exchange rates based on consumer prices and the IFS's REER based on normalized unit labor costs, are both highly correlated with the IMF's CPI-based REER index.

The REER data for all 20 countries are set out in Figure 1—an increase in the REER series indicates a real appreciation of the country's currency. Several features of the data stand out. First, a cursory inspection of the REER series indicates that most countries have real exchange rates that appear to exhibit symptoms of drift or nonstationarity. There appear to be substantial and sustained deviations from PPP (that is, nonstationarity in the REER). The evolution of REER appears to be a highly persistent, slow-moving process; for most countries the REER does not appear to cycle about any particular equilibrium value, especially for Japan (the general appreciation of its exchange rate is typical of a process with a unit
AN UNBIASED APPRAISAL OF PURCHASING POWER PARITY

Figure 1. Real Effective Exchange Rate, (base 1995 = 100), Industrial Countries, April 1973–April 2002

Figure 1. Real Effective Exchange Rate, (base 1995=100), Industrial Countries, April 1973–April 2002 (concluded)

AN UNBIASED APPRAISAL OF PURCHASING POWER PARITY

Table 1. Real Effective Exchange Rates: Tests for Serial Correlation and Heteroskedasticity

<table>
<thead>
<tr>
<th>Country</th>
<th>Breusch-Godfrey</th>
<th>White</th>
<th>ARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>21.3*</td>
<td>10.9*</td>
<td>0.8</td>
</tr>
<tr>
<td>Austria</td>
<td>17.8*</td>
<td>21.0*</td>
<td>5.1*</td>
</tr>
<tr>
<td>Belgium</td>
<td>37.6*</td>
<td>3.0</td>
<td>5.7*</td>
</tr>
<tr>
<td>Canada</td>
<td>17.4*</td>
<td>8.7*</td>
<td>4.9*</td>
</tr>
<tr>
<td>Finland</td>
<td>11.7*</td>
<td>1.2</td>
<td>2.6</td>
</tr>
<tr>
<td>France</td>
<td>349*</td>
<td>349</td>
<td>321*</td>
</tr>
<tr>
<td>Germany</td>
<td>30.6*</td>
<td>25.0*</td>
<td>53.6*</td>
</tr>
<tr>
<td>Iceland</td>
<td>9.8*</td>
<td>25.0*</td>
<td>1.8</td>
</tr>
<tr>
<td>Ireland</td>
<td>29.1*</td>
<td>11.8*</td>
<td>0.0</td>
</tr>
<tr>
<td>Italy</td>
<td>29.7*</td>
<td>7.7*</td>
<td>41.4*</td>
</tr>
<tr>
<td>Japan</td>
<td>347*</td>
<td>348*</td>
<td>326*</td>
</tr>
<tr>
<td>Netherlands</td>
<td>31.4*</td>
<td>3.6</td>
<td>49.2*</td>
</tr>
<tr>
<td>New Zealand</td>
<td>30.1*</td>
<td>0.9</td>
<td>16.7*</td>
</tr>
<tr>
<td>Norway</td>
<td>21.2*</td>
<td>2.4</td>
<td>2.3</td>
</tr>
<tr>
<td>Portugal</td>
<td>4.5</td>
<td>1.7</td>
<td>2.4</td>
</tr>
<tr>
<td>Spain</td>
<td>13.3*</td>
<td>0.1</td>
<td>3.5</td>
</tr>
<tr>
<td>Sweden</td>
<td>29.6*</td>
<td>1.8</td>
<td>8.6*</td>
</tr>
<tr>
<td>Switzerland</td>
<td>28.7*</td>
<td>2.3</td>
<td>22.6*</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>40.4*</td>
<td>2.9</td>
<td>11.6*</td>
</tr>
<tr>
<td>United States</td>
<td>28.2*</td>
<td>25.2*</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Notes: Breusch-Godfrey is a Lagrange Multiplier test of the hypothesis of no serial correlation in the residuals of the AR(1) model of equation (1); the test statistic is distributed as a $\chi^2(n)$, where $n$ is the order of the autocorrelations (here $n = 2$, so the 5 percent critical value is 5.99). White’s (1980) test of the null hypothesis of homoskedasticity; the test statistic is distributed as a $\chi^2(s)$, where $s$ is the number of regressors, which here include the square of the regressors only (so $s = 2$, and the 5 percent critical value is 5.99). ARCH is the ARCH LM test for autoregressive conditional heteroskedasticity, where the null is that the coefficient on lagged squared residuals are all zero; the test is distributed as a $\chi^2(q)$, where $q$ is the number of squared residuals (here $q = 1$, so the 5 percent critical value is 3.84). An asterisk denotes significance at the 5 percent level.

Second, sharp movements in the REER during the 1980s and 1990s are a relatively frequent occurrence, especially for countries such as Australia, Italy, New Zealand, the United Kingdom, and the United States. Third, tests carried out on the residuals from the least-squares regression of equation (1) indicate that the majority of REER regressions have residuals that exhibit serial correlation and heteroskedasticity (see Table 1). We now describe the results for our analysis of the persistence of parity deviations for the REER series, using both biased least-squares and median-unbiased estimators.

Biased Least-Squares Estimates of Half-Lives of Parity Reversion

Table 2 sets out the results for the half-life of the duration of shocks to the REER, which are calculated from the least-squares estimates of $\alpha$ in the Dickey-Fuller (DF) regression of equation (1), as set out in Section I. Across all countries, the
### Table 2. Half-Lives of Parity Deviations: Biased Least-Squares and Median-Unbiased Estimation of Dickey-Fuller Regressions

<table>
<thead>
<tr>
<th>Country</th>
<th>Biased Least Squares</th>
<th></th>
<th>Median Unbiased</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>Half-life (months)</td>
<td>90 percent CI</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Australia</td>
<td>0.989</td>
<td>63</td>
<td>[30, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Austria</td>
<td>0.982</td>
<td>38</td>
<td>[23, 104]</td>
<td>0.994</td>
</tr>
<tr>
<td>Belgium</td>
<td>0.995</td>
<td>138</td>
<td>[46, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Canada</td>
<td>0.994</td>
<td>115</td>
<td>[47, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Finland</td>
<td>0.991</td>
<td>77</td>
<td>[32, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>France</td>
<td>0.985</td>
<td>46</td>
<td>[22, ∞)</td>
<td>0.998</td>
</tr>
<tr>
<td>Germany</td>
<td>0.987</td>
<td>53</td>
<td>[25, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Iceland</td>
<td>0.941</td>
<td>11</td>
<td>[7, 23]</td>
<td>0.949</td>
</tr>
<tr>
<td>Ireland</td>
<td>0.981</td>
<td>36</td>
<td>[19, 340]</td>
<td>0.993</td>
</tr>
<tr>
<td>Italy</td>
<td>0.984</td>
<td>43</td>
<td>[22, ∞)</td>
<td>0.997</td>
</tr>
<tr>
<td>Japan</td>
<td>0.990</td>
<td>69</td>
<td>[33, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.987</td>
<td>53</td>
<td>[25, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>New Zealand</td>
<td>0.971</td>
<td>24</td>
<td>[14, 84]</td>
<td>0.980</td>
</tr>
<tr>
<td>Norway</td>
<td>0.969</td>
<td>22</td>
<td>[13, 74]</td>
<td>0.978</td>
</tr>
<tr>
<td>Portugal</td>
<td>0.991</td>
<td>77</td>
<td>[32, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Spain</td>
<td>0.983</td>
<td>40</td>
<td>[22, 200]</td>
<td>0.993</td>
</tr>
<tr>
<td>Sweden</td>
<td>0.993</td>
<td>99</td>
<td>[37, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0.974</td>
<td>26</td>
<td>[16, 76]</td>
<td>0.983</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>0.989</td>
<td>63</td>
<td>[27, ∞)</td>
<td>1.000</td>
</tr>
<tr>
<td>United States</td>
<td>0.994</td>
<td>115</td>
<td>[38, ∞)</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Notes: Least Squares—The results in columns 2–4 of this table are based on least-squares estimation of the Dickey-Fuller regression of equation (1). The half-life is the length of time it takes for a unit impulse to dissipate by half. It is derived using the formula: $\text{HLS} = \text{ABS}(\log(1/2)\log(\alpha))$, where $\alpha$ is the autoregressive parameter. The least-squares estimate of the HLS is calculated using the least-squares estimate of $\alpha$ in the formula for the HLS. The 90 percent confidence intervals (CI) of the half-life of parity deviations are calculated by inserting $\alpha \pm 1.65 \times se(\bar{\alpha})$ into the HLS formula. The half-lives and the 90 percent confidence intervals are measured in months. Median Unbiased—The results in columns 5–7 of this table are based on the median-unbiased estimates of the Dickey-Fuller regression of equation (1), as given by Andrews (1993). The half-life is the length of time it takes for a unit impulse to dissipate by half. It is derived using the formula: $\text{HLS} = \text{ABS}(\log(1/2)\log(\hat{\alpha}))$, where $\hat{\alpha}$ is the median-unbiased autoregressive parameter. The median-unbiased estimate of the HLS is calculated using the median-unbiased estimate of $\hat{\alpha}$ in the formula for the HLS. Similarly, the 90 percent confidence intervals (CI) of the half-life of parity deviations are derived using the 0.05 and 0.95 quantiles of the median-unbiased estimate of $\hat{\alpha}$ in the formula for the HLS. The quantile functions of $\hat{\alpha}$ were generated by numerical simulation (using 10,000 iterations) for $T = 349$ observations. The half-lives and the 90 percent confidence intervals are measured in months.
mean half-life of parity reversion is 60 months and the median half-life is 53 months.15 This result is consistent with Rogoff's (1996) consensus of half-lives of parity reversion of between 36 to 60 months (three to five years).

We also report 90 percent confidence intervals for the least-squares half-life of PPP deviations, in order to gauge the variability of the persistence of shocks to the real exchange rate. The confidence intervals are quite wide and encompass half-lives that are consistent with both PPP holding and not holding in the long run. This indicates that there is a high level of uncertainty about the "true" value of the half-life of PPP deviations. For seven countries the upper bound of the 90 percent confidence interval is finite, indicating that these countries have finitely persistent shocks to their real exchange rates. It is important to note that the confidence intervals used here are formed assuming that the estimated autoregressive parameter has a $t$-distribution, which we know to be incorrect, and further biases hypothesis tests toward rejecting the unit root null.

As an example of how to interpret the table, we take the particular cases of Iceland and Canada. For Iceland, the least-squares (LS) estimate of $\alpha$ is 0.941. Moreover, the time it takes for half of the shock to the REER of Iceland to dissipate is 11 months, while the length of the 90 percent confidence interval for the LS estimate of the half-life of deviations from parity is 7 to 23 months. Accordingly, shocks to the REER of Iceland do not appear to be very persistent, at least relative to the persistence found in other countries' real exchange rates.

For Canada, the LS estimate of $\alpha$ is 0.994. Moreover, the time it takes for half of the shock to Canada's REER to dissipate is 115 months, while the length of the 90 percent confidence interval for the LS estimate of the half-life of deviations from parity is 47 to $\infty$ months. Accordingly, shocks to the REER of Canada do appear to be very persistent, especially since the lower bound of the 90 percent confidence interval indicates that there is only a 5 percent chance that the true half-life of parity reversion of the REER is shorter than 47 months.

The DF regression results presented in Table 2 do not attempt to account for the presence of serial correlation. Tests for serial correlation carried out on the residuals from the least-squares regression of equation (1) indicate that all REER regressions (except Portugal) have residuals with serial correlation (see column 2 of Table 1). Accordingly, least-squares estimates of the Augmented Dickey-Fuller (ADF) regressions, which do account for serial correlation, are set out in Table 3.

In examining for the presence of serial correlation, the general-to-specific lag selection procedure of Ng and Perron (1995) and Hall (1994) is used, with the maximum lag set to 14. For all countries at least one lag ($p=2$) of the first difference of the REER is significant, which ensures that the ADF half-lives will differ from the DF half-lives. The ADF half-lives are typically shorter in duration than those derived from the DF regression, ranging from 11 months (Iceland) to

---

15These half-life results are comparable to those obtained by Cheung and Lai (2000b) using least-squares estimation on monthly bilateral (post-Breton Woods) dollar real exchange rates, which calculated an average half-life of 3.3 years.

16Starting with the maximum lag, first-differences of the logarithm of the REER ($q_t$) were sequentially removed from the AR model until the last lag was statistically significant (at the 5 percent level). At that point all lag lengths smaller than or equal to $p-1$ are included in the AR($p$) regression of equation (2).
### Table 3. Half-Lives of Parity Deviations: Biased Least-Squares and Median-Unbiased Estimation of Augmented Dickey-Fuller Regressions

<table>
<thead>
<tr>
<th>Country</th>
<th>$p$</th>
<th>$\alpha$</th>
<th>Half-life (months)</th>
<th>90 percent CI (months)</th>
<th>$\alpha$</th>
<th>Half-life (months)</th>
<th>90 percent CI (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>2</td>
<td>0.987</td>
<td>55</td>
<td>[29, $\infty$]</td>
<td>0.993</td>
<td>97</td>
<td>[37, $\infty$]</td>
</tr>
<tr>
<td>Austria</td>
<td>8</td>
<td>0.986</td>
<td>52</td>
<td>[29, $\infty$]</td>
<td>0.993</td>
<td>98</td>
<td>[37, $\infty$]</td>
</tr>
<tr>
<td>Belgium</td>
<td>8</td>
<td>0.993</td>
<td>99</td>
<td>[43, $\infty$]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[48, $\infty$]</td>
</tr>
<tr>
<td>Canada</td>
<td>14</td>
<td>0.991</td>
<td>81</td>
<td>[40, $\infty$]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[53, $\infty$]</td>
</tr>
<tr>
<td>Finland</td>
<td>11</td>
<td>0.983</td>
<td>45</td>
<td>[30, 140]</td>
<td>0.989</td>
<td>67</td>
<td>[36, $\infty$]</td>
</tr>
<tr>
<td>France</td>
<td>12</td>
<td>0.984</td>
<td>45</td>
<td>[20, $\infty$]</td>
<td>0.993</td>
<td>96</td>
<td>[35, $\infty$]</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
<td>0.983</td>
<td>43</td>
<td>[24, 214]</td>
<td>0.993</td>
<td>97</td>
<td>[27, $\infty$]</td>
</tr>
<tr>
<td>Iceland</td>
<td>14</td>
<td>0.927</td>
<td>11</td>
<td>[9, 19]</td>
<td>0.936</td>
<td>11</td>
<td>[9, 28]</td>
</tr>
<tr>
<td>Ireland</td>
<td>5</td>
<td>0.979</td>
<td>33</td>
<td>[18, 155]</td>
<td>0.985</td>
<td>47</td>
<td>[21, $\infty$]</td>
</tr>
<tr>
<td>Italy</td>
<td>2</td>
<td>0.981</td>
<td>37</td>
<td>[21, 160]</td>
<td>0.989</td>
<td>65</td>
<td>[24, $\infty$]</td>
</tr>
<tr>
<td>Japan</td>
<td>12</td>
<td>0.986</td>
<td>52</td>
<td>[32, 199]</td>
<td>0.993</td>
<td>95</td>
<td>[38, $\infty$]</td>
</tr>
<tr>
<td>Netherlands</td>
<td>11</td>
<td>0.981</td>
<td>40</td>
<td>[25, 130]</td>
<td>0.989</td>
<td>65</td>
<td>[33, $\infty$]</td>
</tr>
<tr>
<td>New Zealand</td>
<td>3</td>
<td>0.968</td>
<td>22</td>
<td>[13, 64]</td>
<td>0.974</td>
<td>27</td>
<td>[14, $\infty$]</td>
</tr>
<tr>
<td>Norway</td>
<td>2</td>
<td>0.962</td>
<td>19</td>
<td>[12, 44]</td>
<td>0.967</td>
<td>22</td>
<td>[12, 97]</td>
</tr>
<tr>
<td>Portugal</td>
<td>2</td>
<td>0.990</td>
<td>71</td>
<td>[31, $\infty$]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[46, $\infty$]</td>
</tr>
<tr>
<td>Spain</td>
<td>2</td>
<td>0.982</td>
<td>39</td>
<td>[22, 148]</td>
<td>0.989</td>
<td>64</td>
<td>[30, $\infty$]</td>
</tr>
<tr>
<td>Sweden</td>
<td>2</td>
<td>0.990</td>
<td>73</td>
<td>[34, $\infty$]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[43, $\infty$]</td>
</tr>
<tr>
<td>Switzerland</td>
<td>14</td>
<td>0.968</td>
<td>20</td>
<td>[13, 56]</td>
<td>0.980</td>
<td>35</td>
<td>[12, $\infty$]</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2</td>
<td>0.984</td>
<td>45</td>
<td>[24, $\infty$]</td>
<td>0.993</td>
<td>97</td>
<td>[36, $\infty$]</td>
</tr>
<tr>
<td>United States</td>
<td>2</td>
<td>0.991</td>
<td>77</td>
<td></td>
<td>1.000</td>
<td>$\infty$</td>
<td>[46, $\infty$]</td>
</tr>
</tbody>
</table>

**Notes:** 
Least Squares—The results of columns 3–5 of this table are based on least-squares estimation of the Augmented Dickey-Fuller regression of equation (2). The half-life for AR($p$) models is calculated from the impulse response functions (equation (5)), and is defined as the time it takes for a unit impulse to dissipate permanently by one-half from the occurrence of the initial shock. Similarly, the 90 percent confidence interval (CI) is calculated using the 0.05 and 0.95 quantiles, calculated as the time it takes for a unit impulse to dissipate permanently by one-half from the occurrence of the initial shock. The half-lives of parity deviations and the 90 percent confidence intervals are measured in months. In examining for the presence of serial correlation, the general-to-specific lag selection procedure of Ng and Perron (1995) and Hall (1994) is used, with the maximum lag ($p$) set to 14. First-differences of the REER ($q_t$) were sequentially removed from the AR model until the last ($p-1$) lag was statistically significant (at the 5 percent level). The optimal lag length ($p$) is listed in column 2.

Median Unbiased—The results of columns 6–8 of this table are based on the median-unbiased estimates of the Augmented Dickey-Fuller regression of equation (2), as given by Andrews and Chen (1994). The median half-life for AR($p$) models is calculated from the impulse response functions (equation (5)), and is defined as the time it takes for a unit impulse to dissipate permanently by one-half from the occurrence of the initial shock. Similarly, the 90 percent confidence interval (CI) is calculated using the 0.05 and 0.95 quantiles, calculated as the time it takes for a unit impulse to dissipate permanently by one-half from the occurrence of the initial shock. The quantile functions of $\alpha$ were generated by numerical simulation (using 2,500 iterations). The half-lives of parity deviations and the 90 percent confidence intervals are measured in months.
Across all countries, the mean half-life of parity reversion is 48 months and the median half-life is 45 months. While for several countries the 90 percent confidence interval is narrower than for the DF regression (such as New Zealand and Norway), in several cases the variability of shocks to the REER is so wide as to include infinity as the upper bound of the confidence interval (such as Canada and the United States).  

The ADF regressions presented in Table 3 do not attempt to account for the presence of heteroskedasticity, and so will be invalid when there are departures from the maintained hypothesis of AR($p$) errors. Tests for heteroskedasticity carried out on the residuals from the least-squares regression of equation (1) indicate that most REER regressions have residuals with heteroskedasticity (see columns 3–4 of Table 1). Accordingly, given the presence of heteroskedasticity and serial correlation (including moving-average error structures) in the real exchange rate series, the results of Phillips-Perron (PP) regressions, which are valid in the presence of serial correlation and heteroskedasticity, are presented in Table 4. The duration of PP half-lives are typically lower again than those derived from the ADF and DF regressions, ranging from 9 months (Iceland) to 79 months (Canada). Across all countries, the mean half-life of parity reversion from the PP regression is 35 months and the median half-life is 30 months. For those countries with finite upper bounds for the 90 percent confidence interval of the half-lives of deviations from parity, the confidence intervals based on PP regressions are typically tighter than those derived from the ADF and DF regressions, yet continue to encompass a wide range of half-lives. Controlling for the serial correlation present in the data lowers considerably the estimated half-life of parity deviations—the PP regressions detect more serial correlation than the ADF and DF regressions, and thus produce lower estimated half-lives of deviations from parity.

Broadly, the three least-squares results of Tables 2–4 indicate that across all countries the median half-life of parity deviations is finite, with an average length of about four years, and a lower bound on the confidence intervals for the true half-lives of about two years. However, the upper bound of the confidence interval for many REER is greater than ten years (and in many cases is infinity).

**Median-Unbiased Estimates of Half-Lives of Parity Reversion**

The half-lives of PPP deviations calculated above (using the least-squares estimator) are reasonably close to Rogoff’s (1996) consensus of three to five years (36 to 60 months). However, as noted in Section I above, the least-squares estimator of the autoregressive parameter in each of the DF, ADF, and PP regressions is biased downward. As a result, the above calculations of the duration of deviations from PPP are also likely to be biased downward (and in favor of finding that PPP holds in the REER data). Consequently, we remove this bias by

---

17 Consistent with Papell (1997), we find that accounting for serial correlation in the disturbances weakens the evidence against a null hypothesis of a unit root in the real exchange rate series, as the point estimates of the autoregressive parameter are typically lower in the AR($p$) case than for the AR(1) regression.
Table 4. Half-Lives of Parity Deviations: Biased Least-Squares and Median-Unbiased Estimation of Phillips-Perron Regressions

<table>
<thead>
<tr>
<th>Country</th>
<th>$b$</th>
<th>$\alpha$</th>
<th>Half-life (months)</th>
<th>90 percent CI (months)</th>
<th>$\alpha$</th>
<th>Half-life (months)</th>
<th>90 percent CI (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>1.38</td>
<td>0.982</td>
<td>39</td>
<td>[21, 329]</td>
<td>0.995</td>
<td>136</td>
<td>[26, $\infty$]</td>
</tr>
<tr>
<td>Austria</td>
<td>0.65</td>
<td>0.977</td>
<td>30</td>
<td>[18, 79]</td>
<td>0.989</td>
<td>61</td>
<td>[20, 96]</td>
</tr>
<tr>
<td>Belgium</td>
<td>0.14</td>
<td>0.989</td>
<td>61</td>
<td>[27, 264]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[41, $\infty$]</td>
</tr>
<tr>
<td>Canada</td>
<td>0.39</td>
<td>0.991</td>
<td>79</td>
<td>[35, 311]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[52, $\infty$]</td>
</tr>
<tr>
<td>Finland</td>
<td>0.72</td>
<td>0.986</td>
<td>50</td>
<td>[24, $\infty$]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[33, $\infty$]</td>
</tr>
<tr>
<td>France</td>
<td>0.98</td>
<td>0.973</td>
<td>25</td>
<td>[14, 113]</td>
<td>0.984</td>
<td>43</td>
<td>[17, 128]</td>
</tr>
<tr>
<td>Germany</td>
<td>1.32</td>
<td>0.976</td>
<td>29</td>
<td>[16, 143]</td>
<td>0.988</td>
<td>57</td>
<td>[19, 100]</td>
</tr>
<tr>
<td>Iceland</td>
<td>1.14</td>
<td>0.922</td>
<td>9</td>
<td>[6, 16]</td>
<td>0.977</td>
<td>30</td>
<td>[6, 22]</td>
</tr>
<tr>
<td>Ireland</td>
<td>0.43</td>
<td>0.967</td>
<td>20</td>
<td>[12, 64]</td>
<td>0.984</td>
<td>42</td>
<td>[14, 247]</td>
</tr>
<tr>
<td>Italy</td>
<td>1.91</td>
<td>0.972</td>
<td>25</td>
<td>[14, 98]</td>
<td>0.994</td>
<td>123</td>
<td>[25, $\infty$]</td>
</tr>
<tr>
<td>Japan</td>
<td>1.60</td>
<td>0.982</td>
<td>38</td>
<td>[20, 278]</td>
<td>0.988</td>
<td>55</td>
<td>[19, 103]</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1.22</td>
<td>0.976</td>
<td>28</td>
<td>[16, 146]</td>
<td>0.959</td>
<td>16</td>
<td>[9, 74]</td>
</tr>
<tr>
<td>New Zealand</td>
<td>1.95</td>
<td>0.949</td>
<td>13</td>
<td>[8, 99]</td>
<td>0.960</td>
<td>17</td>
<td>[9, 87]</td>
</tr>
<tr>
<td>Norway</td>
<td>1.44</td>
<td>0.951</td>
<td>14</td>
<td>[9, 31]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[40, $\infty$]</td>
</tr>
<tr>
<td>Portugal</td>
<td>0.62</td>
<td>0.989</td>
<td>60</td>
<td>[27, 286]</td>
<td>0.990</td>
<td>68</td>
<td>[21, 91]</td>
</tr>
<tr>
<td>Spain</td>
<td>0.21</td>
<td>0.978</td>
<td>31</td>
<td>[18, 123]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[32, $\infty$]</td>
</tr>
<tr>
<td>Sweden</td>
<td>0.05</td>
<td>0.986</td>
<td>48</td>
<td>[23, $\infty$]</td>
<td>0.968</td>
<td>21</td>
<td>[11, $\infty$]</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1.62</td>
<td>0.958</td>
<td>16</td>
<td>[10, 36]</td>
<td>0.987</td>
<td>52</td>
<td>[18, 108]</td>
</tr>
<tr>
<td>United States</td>
<td>1.53</td>
<td>0.975</td>
<td>27</td>
<td>[15, 160]</td>
<td>1.000</td>
<td>$\infty$</td>
<td>[34, $\infty$]</td>
</tr>
</tbody>
</table>

Notes: Least Squares:—The results of columns 3–5 of this table are based on least-squares estimation of the Phillips-Perron (PP) regression of equation (1). The half-life is the length of time it takes for a unit impulse to dissipate by half from the occurrence of the initial shock. It is derived using the formula: \( HLS = \text{ABS}(\log(1/2)/\log(\alpha)) \), where \( \alpha \) is the autoregressive parameter. The PP estimate of the HLS is calculated using the PP estimate of \( \alpha \) in the formula for the HLS. The 90 percent confidence intervals (CI) of the half-life of parity deviations are calculated by inverting \( \alpha \pm 1.65 \times \text{se}(\alpha) \) into the HLS formula, where \( \text{se}(\alpha) \) is calculated using a long-run variance estimator. The half-lives and the 90 percent confidence intervals are measured in months. To control the amount of serial dependence allowed in the Phillips-Perron regression, the bandwidth parameter needs to be selected—we have used the automatic bandwidth selector of Andrews (1991), where the bandwidth (number of periods of serial correlation included) is indicated by \( b \), and is reported in column 2. Prewhitened kernel estimation of the long-run variance parameter (following Andrews and Monahan, 1992) is used prior to the implementation of the data-determined bandwidth selection procedure. Median Unbiased—The results in columns 6–8 of this table are based on the median-unbiased estimates of the Phillips-Perron regression of equation (1). The half-life is the length of time it takes for a unit impulse to dissipate by half from the occurrence of the initial shock. It is derived using the formula: \( HLS = \text{ABS}(\log(1/2)/\log(\alpha)) \), where \( \alpha \) is the autoregressive parameter. The median-unbiased estimate of the HLS is calculated using the median-unbiased estimate of \( \alpha \) in the formula for the HLS. Similarly, the 90 percent confidence intervals (CI) of the half-life of parity deviations are derived using the 0.05 and 0.95 quantiles of the median-unbiased estimate of \( \alpha \) in the formula for the HLS. The quantile functions of \( \alpha \) were generated by numerical simulation (using 10,000 iterations) for \( T = 349 \) observations. The half-lives and the 90 percent confidence intervals are measured in months. As with the least-squares estimation, both the Andrews (1991) and Andrews and Monahan (1992) procedures were used in the Phillips-Perron regression. Following Andrews and Chen (1994), in calculating the median-unbiased estimates of \( \alpha \), the parameter space is restricted to \((-1,1)\).
calculating median-unbiased point estimates and confidence intervals for the autoregressive parameter in equations (1) and (2).

Median-unbiased estimates of the half-life of PPP deviations for the DF regressions are set out in Table 2. In comparison with the median-unbiased estimates of $\alpha$ in DF regressions, the least-squares estimates of $\alpha$ are biased downward by between 0.005 and 0.013. While this is a small difference in absolute terms, it has important implications for the half-life measures of the persistence of the REER. The median-unbiased point estimates of the half-lives are much greater than their least-squares counterparts for every country, with 11 of the countries having a half-life of infinity. Across all countries, the average (median) bias-corrected half-life of parity reversion is infinity, clearly exceeding the average downwardly biased least-squares AR(1) half-life of 53 months (Figure 2). This implies no parity reversion, rather than the 15 percent per year calculated using biased DF methods. In addition, the 90 percent confidence intervals for the median-unbiased estimates of half-lives of parity reversion are typically much wider than for their LS counterparts, and the REER of all countries in Table 2 (except Iceland) have an upper bound to the confidence interval of the unbiased half-lives that embrace infinity.

The results in Table 2 indicate that, using the Andrews unbiased model-selection rule, 9 of the countries are subject to REER shocks that are finitely persistent, while 11 of the countries experience permanent shocks to their REER series. The interpretation of this rule is that for any given country there is a 50 percent probability that the confidence interval from zero to the estimated median-unbiased half-life contains the true half-life of shocks to its REER, and a 50 percent probability that the confidence interval from the estimated median-unbiased half-life to infinity contains the true half-life of shocks to its REER. Let us again take the examples of Iceland (short-lived half-life) and Canada (infinite half-life). While there is a 50 percent probability that the confidence interval from zero to 13 months contains the true half-life of shocks to the REER of Iceland, there is also a 50 percent probability that the confidence interval from 13 months to infinity contains the true half-life of shocks to the REER of Iceland. For Canada, while there is a 50 percent probability that the confidence interval with a finite upper bound contains the true half-life of shocks to its REER, there is a 50 percent probability that the true half-life of

---

18 Median-unbiased estimates (and confidence intervals) of the half-life of a shock (for $T = 349$ observations (1973:4–2002:4)) were determined using quantile functions of $\hat{\alpha}$ generated by: numerical simulation (using 10,000 iterations), following the method suggested by Appendix B of Andrews (1993) for the DF regression of equation (1) (see Table 2); numerical simulation (using 10,000 iterations), following the method suggested by McDermott (1996) for the PP regression of equation (1) (see Table 4); and numerical simulation (using 2,500 iterations), following the method suggested by Andrews and Chen (1994) for the ADF regression of equation (2) (see Table 3).

19 Our results for the median-unbiased Dickey–Fuller regression are similar to those of Andrews (1993), who calculated point and interval estimates of the half-life of monthly bilateral dollar real exchange rates for several industrial countries over the period 1973 to 1988. He found that, using least squares, the half-life of PPP deviations for each real exchange rate was finite, with an average half-life of about 31 months. However, using the median-unbiased procedure (for an AR(1) model) only three of the eight real exchange rates had finite half-lives of PPP deviations, with an average half-life of about 60 months. The remaining five exhibited permanent parity deviations. Similarly, while all of Andrews’ median-unbiased lower bounds of the 90 percent confidence interval were less than 36 months (as is the case for the majority of countries in the present study), the upper bounds were all infinite (as is the case for all but one country in the present study).
shocks to its REER will be infinite (Table 2). Using the Andrews unbiased model-selection rule, the finite (Iceland) and infinite (Canada) point estimates of the half-lives of deviations from parity indicate that while shocks to Iceland’s REER are transitory, shocks to Canada’s REER are best viewed as being permanent.

Using the alternative loss function implicit in conventional two-sided (level 0.10) hypothesis tests, there is another way to interpret our findings. Again, taking the examples of Iceland and Canada, we find that the estimated 90 percent confidence intervals of the bias-corrected half-life of deviations from parity range from 8 months to 42 months, and from 77 months to infinity, respectively (Table 2). There is a 90 percent probability that the above confidence intervals contain the true half-life of shocks to each country’s REER. Accordingly, there is a 5 percent probability that the confidence interval from zero to 8 months contains the true half-life of shocks to the REER of Iceland, and a 95 percent probability that the confidence interval from 8 months to infinity contains the true half-life of shocks to the REER of Iceland. As found in the biased DF regression, shocks to the REER of Iceland do not appear to be very persistent. In contrast, while there is a 5 percent probability that the confidence interval from zero to 77 months contains the true half-life of shocks to the REER of Canada, there is a 95 percent probability that the confidence interval from 77 months to infinity contains the true half-life of shocks to the REER of Canada.

The median-unbiased estimates of the autoregressive parameter in ADF regressions control for serial correlation, and are reported in Table 3. Again, in comparison with their least-squares counterparts, the median-unbiased half-lives of deviations from PPP are typically much longer, ranging from 11 months (Iceland) to infinity (Canada and the United States, among others). Across all countries, the aver-
age bias-corrected half-life of parity reversion is 96 months, in excess of the average downwardly biased least-squares AR(p) half-life of 45 months (Figure 2). This implies a rate of parity reversion of only 8 percent per year, rather than the 17 percent per year calculated using biased ADF methods. Similarly, the median-unbiased confidence intervals are much wider than their least-squares counterparts. The Andrews unbiased model-selection rule indicates that all but 5 of the 20 countries have finitely persistent shocks to their REER, which is consistent with the reversion of REER to parity. However, of all 20 countries, only Iceland and Norway produced a 90 percent confidence interval for the unbiased half-life of deviations from parity that did not include infinity. Taking the United Kingdom as an example, while there is a 50 percent probability that the confidence interval from zero to 97 months contains the true half-life of shocks to its REER, there is also a 50 percent probability that the confidence interval from 97 months to infinity contains the true half-life. In addition, while there is a 5 percent probability that the confidence interval from zero to 36 months contains the true half-life of shocks to the REER of the United Kingdom, there is a 95 percent probability that the confidence interval from 36 months to infinity contains the true half-life (Table 3). The width of this confidence interval for the half-life indicates there is a great deal of uncertainty as to the duration of the true half-life of parity reversion of the United Kingdom’s REER.

Our bias-corrected ADF regression results accord with those obtained by Murray and Papell (2000), who follow Andrews and Chen (1994) in calculating median-unbiased estimates of half-lives for bilateral dollar real exchange rates. They find that the average bias-corrected half-life is about three years, but with confidence intervals that are typically so large that the point estimates of bias-corrected half-lives from ADF regressions provide virtually no information regarding the true size of the half-lives. However, an important deficiency of the Murray-Papell analysis is the inability of their ADF bias-correction method to account for time-dependent heteroskedasticity, which is a common feature of real exchange rate series. Once allowance is made for a wider class of serial correlation and heteroskedasticity, as is done in this paper, the speed of parity reversion is typically faster than that found with other median-unbiased models (see Figure 2).

The regression that corrects for the least-squares downward bias, and controls for serial correlation and heteroskedasticity, is the median-unbiased PP regression, the results of which are reported in Table 4. Across all countries, the average bias-corrected half-life of parity reversion is 59 months, clearly exceeding the average downwardly biased least-squares PP half-life of 30 months (see Figure 2). This implies a rate of parity reversion of only 13 percent per year, rather than the 24 percent per year calculated using biased PP methods. The broad pattern found in the biased least-squares estimates of half-lives of parity reversion is also found for the median-unbiased estimates of half-lives of parity reversion, with the estimation method that controls for serial correlation and heteroskedasticity (the PP regression) clearly yielding the smallest half-lives of deviations from parity.20

20In the context of biased least-squares estimation, Lothian and Taylor (2000) also find much smaller estimates of the half-lives of shocks to the dollar-sterling real exchange rate when using heteroskedasticity-robust PP regressions rather than ADF regressions.
As shown in Figure 2, implementation of methods of bias correction that do not account for both the serial correlation and heteroskedasticity present in real exchange rate data will tend to overestimate the duration of the half-life of deviations from parity, and erroneously indicate that purchasing power parity does not hold in the post-Bretton Woods period. Interestingly, the average half-life of parity deviations derived from median-unbiased PP methods is very close to the average half-life of parity deviations derived from (conventional) biased DF estimation methods.\(^{21}\) However, the downwardly biased least-squares DF estimates of the half-lives of deviations from parity yield no cases of infinite half-lives; in contrast, 6 of the 20 countries experience non-finite half-lives using bias-corrected PP estimation methods (see Tables 2 and 4).

While the cross-country average half-lives of parity reversion based on the median-unbiased PP regression are slightly longer than the cross-country average based on the biased least-squares DF regression (see Figure 2), the country-specific results display quite a deal of heterogeneity. For those countries that have relatively small (biased) DF estimates of the autoregressive parameter (such as Iceland), the bias-corrected PP estimates of the autoregressive parameter tend to be lower than the biased DF estimates—this is consistent with faster speeds of parity reversion and greater evidence in favor of purchasing power parity (see Figure 3). In contrast, for those countries that have relatively large (biased) DF estimates of the autoregressive parameter (such as Australia), the bias-corrected PP estimates of the autoregressive parameter tend to be greater than the biased DF estimates.

Using the bias-corrected PP regression results, the Andrews unbiased model-selection rule indicates that all but 6 of the 20 countries have finitely persistent shocks to their REER, which is consistent with the reversion of REER to parity (Table 4). For example, the United Kingdom exhibits a median-unbiased half-life of deviation from parity of 52 months. Using the Andrews unbiased model-selection rule, this indicates that the United Kingdom experiences finitely persistent (transitory) shocks to its REER. However, nine countries have 90 percent confidence intervals for their half-lives of deviations from parity that embrace infinity. Taking the United States as an example, while there is a 5 percent probability that the confidence interval from zero to 34 months contains the true half-life of shocks to its REER, there is a 95 percent probability that the confidence interval from 34 months to infinity contains the true half-life (Table 4). Using the Andrews unbiased model-selection rule, the unit root model is the most appropriate representation of the United States REER, and so shocks to its REER are best viewed as being permanent.\(^{22}\)

---

\(^{21}\)The presence of moving average and heteroskedastic errors in the real exchange rate data (as set out in Table 1) is most likely attributable to outliers in the real exchange rate data. The Phillips-Perron regression controls for this form of heteroskedasticity, while the Augmented Dickey-Fuller and Dickey-Fuller regressions do not. As a result, estimated half-lives calculated from the latter two models (derived from both biased least-squares and median-unbiased estimators) will be greater than those derived from the biased and bias-corrected Phillips-Perron model. In addition, the Phillips-Perron regression picks up more serial correlation (particularly of the moving average type), which would also reduce the estimated half-lives of parity deviations.

\(^{22}\)As a check of the robustness of our results, for the same sample of countries as studied in this paper, McDermott (1996) found that using the bias-corrected Phillips-Perron method the persistence properties of shocks to the real effective and real bilateral exchange rate series were very similar. In addition, a comparison of the two versions of the real exchange rate for Canada, Japan, and the United Kingdom (using the bias-corrected Dickey-Fuller method) again found that the median-unbiased persistence results for real bilateral exchange rates are very similar to those derived using the real effective exchange rate series.
Figure 3. Biased Dickey-Fuller (DF), Unbiased Augmented Dickey-Fuller (ADF), and Unbiased Phillips-Perron (PP) Estimates of Autoregressive Parameter

Source: Authors’ calculations.
For the majority of countries, these typically finite half-lives of deviations from parity, and associated slow speeds of parity reversion, are consistent with the findings of Flood and Taylor (1996). They find that evidence in favor of PPP is hard to discern from analyses of the short-run behavior of real exchange rates. However, once the data are averaged over 10 and 20 years, the regression coefficient (in a pooled regression of the exchange rate change on the inflation differential averages) is statistically significant and close to unity, providing stronger evidence of PPP.

In contrast to the above results, Taylor (2001) argues that there are two sources of upward bias in the conventional estimation of the speed of parity reversion: first, temporal aggregation bias, whereby sampling data at low frequencies does not allow one to identify a high-frequency adjustment process; and second, the linear AR(1) specification of the standard (Dickey-Fuller) unit root model, which assumes that reversion occurs monotonically, regardless of how far the process is from parity. However, as the present paper uses monthly REER data, the temporal aggregation bias is likely to be minimal. In addition, our use of AR(p) models allows for shocks to the REER to decline at a rate that is not necessarily constant.23 Taylor finds that when both sources of upward bias are present, then the estimated half-life can be between 1.5 to 2.2 times the true half-life, for the case when monthly averaged data are being used to estimate a monthly (or greater) half-life nonlinear threshold autoregressive process. Taylor (2001, p. 491) also acknowledges that month-to-month variation in nominal exchange rates is likely to dwarf the variation in prices, and that the former are accurately measured in the International Monetary Fund’s IFS data. In comparison, the results from Tables 3 and 4 indicate that even after controlling for serial correlation and heteroskedasticity, the bias-corrected half-life of parity deviations is typically about twice as large as the downwardly biased half-life.24

III. Conclusion

This paper has reexamined whether purchasing power parity (PPP) holds during the post-Bretton Woods period, by investigating the time-series properties of the real effective exchange rate for 20 industrial countries. The theory of relative PPP holds that the exchange rate will be proportional to the ratio of money price levels (including traded and nontraded goods) between countries, which implies that

22While median-unbiased estimation allowing for AR(p) behavior does engender a non-monotonic impulse response function, the shape of that impulse response function will be unaffected by the size of the shock to the real exchange rate. In contrast, methods that are robust to nonlinearity allow for such an effect, yet do not correct for downwardly biased autoregressive parameters (see Taylor, Peel, and Sarno, 2001, for a discussion of these issues).

24In addition to temporal aggregation bias, Imbs and others (2002) demonstrate that cross-sectional aggregation bias raises the persistence of conventionally measured real exchange rate shocks. Cross-sectional aggregation bias arises from the failure of conventional estimation of the speed of parity reversion to take account of cross-sectoral heterogeneity in the dynamic properties of the typical components of aggregate price indices. This failure to allow for the persistence of relative prices to vary across sectors induces an upward bias in aggregate half-life measures, with the bias rising with the extent of cross-sectoral heterogeneity in the speed of parity reversion.
changes in relative price levels will be offset by changes in the exchange rate. In assessing whether real exchange rates do converge to their equilibrium relative PPP value in the long run, we eschew undertaking hypothesis tests of whether real exchange rates follow a unit root process. Instead, we follow Andrews (1993) and use point and interval statistics of the half-life of deviations from parity as our preferred measure of the persistence of real exchange rate shocks.

Univariate studies of the hypothesis of unit roots in real exchange rates have yielded consensus point estimates of the half-life of deviations of real exchange rates from purchasing power parity of between three to five years (Rogoff, 1996). Using conventional (least-squares) biased estimation of unit root models, we replicate the consensus finding in the literature. However, using median-unbiased estimation techniques that are robust to the presence of serial correlation and remove the downward bias of standard estimators, we find that the half-lives of parity reversion are longer than the consensus point estimate, with the cross-country average of unbiased half-lives of deviations from parity lasting about eight years.

We concentrate on the results derived from the regression that allows for the broadest error structure—median-unbiased estimates of Phillips-Perron regressions that are robust to the serial correlation and heteroskedasticity commonly found in real exchange rate series. In the post-Bretton Woods period, the majority of countries are found to have finitely persistent shocks to their real effective exchange rates, which is consistent with the reversion of exchange rate deviations from PPP. Averaging across all countries, the point estimate of the half-life of parity deviations is about five years, which is consistent with (but at the upper end of) Rogoff’s (1996) consensus estimate of the half-life of deviations from purchasing power parity. In summary, while median-unbiased methods increase the estimated half-life of deviations from PPP in comparison with downwardly biased least-squares-based methods, allowing for heteroskedasticity reduces the bias-corrected estimated half-life of parity deviations.

Using conventional two-sided hypothesis tests, the confidence interval of the bias-corrected half-life of deviations from parity is typically extremely wide, indicating that there is a great deal of uncertainty as to the “true” speed of parity reversion. When using the results from the heteroskedasticity-robust Phillips-Perron regression, 11 of the 20 countries in our sample have finite upper bounds to the confidence intervals of their half-lives of parity deviation, indicating that for just under half the countries the associated confidence intervals around the half-lives are too wide to provide much information as to whether PPP holds in the long run. Although the null hypothesis of no PPP is not rejected at the 5 percent level for 9 of the 20 countries (as the upper bound of the bias-corrected confidence interval for the half-life contains infinity), failure to reject the null hypothesis conveys little information as to the validity of PPP, as such a failure may occur either because PPP is true, or because there is too much uncertainty as to the speed of reversion to PPP.

In contrast, using the median-unbiased point estimates of the half-lives of deviations from parity and the Andrews (1993) unbiased model-selection rule, we can be more definitive about our willingness to draw conclusions as to the presence or absence of parity reversion of real exchange rates in the post-Bretton

©International Monetary Fund. Not for Redistribution
Woods period. When using the results from the heteroskedasticity-robust Phillips-Perron regression, we find that 14 of the 20 countries in our sample have finite bias-corrected half-lives of parity reversion. This indicates that for these countries there is a better than even chance that shocks to their real exchange rates are transitory. Consequently, for these countries we can conclude that there is reversion of real exchange rates to parity, and so PPP holds in the post–Bretton Woods period.

Our results confirm Rogoff’s (1996) “PPP puzzle”—that while PPP holds for the majority of countries, the speed of reversion of real exchange rates to parity is, in many cases, rather slow. In comparison with conventional least-squares-based estimators, our use of serial-correlation robust median-unbiased estimators typically doubles the estimated half-life of parity deviations. However, we find that using (biased and median-unbiased) estimators that are robust to the heteroskedasticity typically present in real exchange rate series reduces the estimated half-life of parity deviations. In particular, the cross-country average half-life derived from heteroskedasticity-robust median-unbiased methods is close to the conventionally estimated (biased) average half-life of parity deviations.

While our results increase our understanding of the parity-reverting behavior of real exchange rates over the post–Bretton Woods period, as median-unbiased models are superior to biased least-squares-based models of exchange rates in several respects, there is scope for improvement. In particular, the implications for exchange rate modeling of median-unbiased methods, in the presence of nonlinearities in the adjustment of real exchange rates toward long-run equilibrium, have yet to be explored. This challenge awaits future research.

REFERENCES


Is Africa Integrated in the Global Economy?

ARVIND SUBRAMANIAN and NATALIA T. TAMIRISA

The popular impression that Africa has not integrated into world trade, as suggested by the evolution in simple indicators, has been called into question recently by more formal analysis. This paper refines and generalizes this analysis and lends support to the popular view of disintegration, but only for countries in Francophone Africa. These countries are currently underexploiting their trading opportunities and have witnessed disintegration over time, a trend that is most pronounced in their trade with technologically advanced countries. There is some evidence, on the other hand, that countries in Anglophone Africa are reversing the trend of disintegration, particularly in their trade with advanced countries. [JEL C1, F1, O4]

The state of the current debate on globalization can generally be summarized as: yes, it confers enormous benefits, but it also poses great challenges. In the case of Africa, however, even the first part of this proposition is not uncontested—globalization’s benefits have largely proven elusive for Africa. Reaping these benefits is predicated on embracing globalization in the first place. Has Africa done so—has it globalized or has it been marginalized from world trade? On this question, there seems to be an uneasy tension between two views, with distinct policy implications.

The authors are grateful to Tamim Bayoumi, Hugh Bredenkamp, David Coe, José Fajgenbaum, Markus Haacker, Michael Hadjeri, Gunnar Jonsson, Anne McGuirk, Yadahia Metzgen, Eswar Prasad, Dani Rodrik, Emilio Sacerdotti, Jeffrey Sachs, Hossein Samiei, Antonio Spilimbergo, Alan Winters, Shang-Jin Wei, Bernarda Zamora, and other colleagues at the International Monetary Fund for helpful discussions and comments, and to Rikhil Bhavnani, Nehrunman Pillay, and Vera De Luz for excellent assistance in compiling the dataset. The paper also benefited from the insightful and constructive comments of an outside referee.

In this paper, Africa refers to sub-Saharan Africa. Globalization refers to integration of goods markets through international trade and not to capital market integration.
IS AFRICA INTEGRATED IN THE GLOBAL ECONOMY?

According to the first, popular view, Africa has missed out on the opportunities offered by globalization simply because it has not globalized. The statistic that is commonly invoked in support is a dramatic decline in Africa's share of world exports during the past three decades, representing a "staggering annual income loss of US$68 billion—or 21 percent of regional GDP" (World Bank, 2000). Reviving trade is therefore integral to Africa's economic fortunes, a view that is consistent with the research evidence demonstrating the benefits of integration (Sachs and Warner, 1997; and Collier and Gunning, 1999).

The second view is that Africa did take advantage of trading opportunities in line with the evolution in its income and development. Academic support for this view comes from the spate of evidence that demonstrates that Africa does not trade too little: it is an average trader, trading just as much as can be expected given the underlying determinants of trade, such as income, geography, and size (Foroutan and Pritchett, 1993; Coe and Hoffmaister, 1999; and Rodrik, 1999).²

These views lead to distinct policy implications. The former sees Africa's declining trade as a source of concern and accordingly places considerable emphasis on policy measures to expand trade opportunities (World Bank, 2000; Sachs, 2000). The latter view sees causality running from growth, and other determinants, to trade and hence is less activist toward, or at least sees less urgency in, the need to promote trade (Rodrik, 1999).

The evidence provided by the recent literature, however, has a number of limitations. The literature focuses on selected, rather than all, components of Africa's trade. It is based on a relatively narrow, rather than a general, benchmark for assessing what "average" or "typical" trade is. It treats Africa as a uniform region, failing to distinguish intraregional specificities. Lastly, econometric methodologies employed in estimating Africa's trade could be refined further.

This paper—which focuses on the second of the two strands in the literature described above—seeks to remedy these limitations. It revisits the puzzle of Africa's trade to shed light on the key underlying issues: whether Africa undertrades or overtrades, and how its trading pattern has changed over time.

We find that countries in Francophone Africa are currently underexploiting their trading opportunities and have witnessed disintegration over time, a trend that is most pronounced in their trade with technologically advanced countries. There is some evidence, on the other hand, that countries in Anglophone Africa are reversing the trend of disintegration, particularly in their trade with advanced countries.

A robustness analysis points to two possible explanations for the contrasting experiences of Francophone and Anglophone Africa. Higher trade-related transaction costs, possibly due to greater inefficiencies in key infrastructure services, and currency arrangements in Francophone Africa may have contributed to its relatively inferior trade performance. The results are robust to the inclusion of variables that control for primary-commodity dependence. That is, the results do not reflect the fact that disintegration is due to Francophone African countries being primary commodity exporters.

²Easterly and Levine (1997) suggest that a lot of factors not strictly related to trade, including geography and ethnic divisions, help explain Africa's poor growth performance.
I. Background

Statistics on the evolution in Africa's share of world trade visually suggest that Africa is progressively disintegrating or marginalizing from world trade (Figure 1). The top panel shows that Africa's share of world exports declined from over 4.1 percent in 1980 to about 1.6 percent in 2000, while its share of world imports declined from over 3.2 percent to 1.3 percent over the same period. More disturbingly, the bottom panel suggests that Africa's share of trade in commodities has also declined significantly from about 8 percent in 1980 to about 4.4 percent in 2000. Thus, the disintegration from trade is not, or not just, due to a less-than-average performance in manufacturing, in which Africa may not have comparative advantage.

A series of recent papers have subjected this impression to a formal empirical scrutiny by asking the question of how typical Africa's trade is relative to a pre-selected theoretical benchmark. The salient features of these papers are summarized in Table 1.3

Foroutan and Pritchett (1993) use data on trade, excluding that in primary commodities, for the early 1980s to test whether African trade is unusual. Their sample comprises 53 low- and medium-income countries (with per capita GDP less than US$3,000) as reporting countries and 95 partner countries. Thus, the benchmark of what constitutes typical trade is trade of the countries that are similar to African countries. The gravity model is estimated using the Tobit procedure. Foroutan and Pritchett (1993) find no evidence that African countries trade less with each other than other developing countries. In fact, intra-African trade is higher than expected when trade is measured in terms of exports.

Coe and Hoffmaister (1999) test whether Africa's trade is unusual by examining trade flows between developing and industrial countries during 1970-97. They apply a nonlinear procedure to estimate the gravity model and find that in the 1970s Africa overtraded with the North relative to other countries' trade with the North and that over time this overtrading has declined. In the 1990s, Africa's trade was no different from the average developing country's trade with the North.

The model, however, does not control for a key variable, the preferential trading arrangement between the European Union and Africa under the Lome Convention. Hence, it is difficult to assess whether the Africa dummy is merely picking up the effects of this preferential trading arrangement.4 Indeed, the decline in the magnitude of overtrading with the North is consistent with the decline in preferential margins under the Lome Convention as most-favored-nation tariff rates in Europe have declined and as Europe has entered into other preferential trading arrangements.

This paper encompasses the earlier body of work, yet differs from it in a number of ways. First, the paper explores African trade in its entirety. In other

---

3Rodrik (1999) tests whether Africa's aggregate rather than bilateral trade is unusual, after controlling for size, income, and average distance from the world. The paper does not employ a strict bilateral gravity model, but like other authors, Rodrik finds that Africa's trade is not dissimilar to other countries' trade.

4In Foroutan and Pritchett (1993), the Lome dummy variable has a positive and statistically significant coefficient.
IS AFRICA INTEGRATED IN THE GLOBAL ECONOMY?

Figure 1. Africa's Share of World Trade, 1980-2000

A. Aggregate Trade

B. Trade in Primary Commodities

Sources: The IMF's *World Economic Outlook* for the top panel and the World Bank's *World Development Indicators* for the bottom panel. Primary commodities comprise fuel, ores, metals, and agricultural products. Data on primary commodities are available only for selected years.
words, we test for the typicality of Africa's overall trade, its trade with other African countries, and its trade with developed and developing countries. The earlier studies cited above, in contrast, examine the typicality of a selected component of African trade.

Second, instead of treating Africa as a homogenous region, we disaggregate Africa's trade into that of Central and Western Africa (which we refer to as Francophone Africa) and of Eastern and Southern Africa (referred to as Anglophone Africa). Such a disaggregation appears to be warranted in view of notable differences between these groups of countries in terms of institutions, policies, and the overall approach to regional and global integration, and is validated by our findings.5

Third, the paper uses a global benchmark for assessment. It seeks to answer whether Africa's trade—and all its components—differ from those of a broad

---

Note: NLS is nonlinear least squares.

---

Language is not a criterion for disaggregation in this context. Indeed, in modeling we control separately for commonality of language (and through this partially for historical similarities) among countries.
IS AFRICA INTEGRATED IN THE GLOBAL ECONOMY?

group of countries. The sample comprises 73 industrial and developing countries, of which 16 are in sub-Saharan Africa. Thus, the benchmark for evaluating "average" trade is a general one, unlike in Foroutan and Pritchett's (1993) paper, which asks whether African trade is different from trade of low- and middle-income countries, or in Coe and Hoffmaister's (1999) paper, which examines whether Africa's trade with the North is different from other developing countries' trade with the North. Notwithstanding the above, our framework is flexible enough to permit testing the robustness of results to alternative benchmarks.

Finally, the paper employs nonlinear least squares (NLS) to adequately address the problem of zero-valued observations (similarly to Coe and Hoffmaister, 1999) and relies on bootstrapping to make hypothesis testing valid given the nonnormality of residuals.

II. Methodology

The most commonly used analytical framework for studying bilateral trade flows is the gravity model, and it is well suited for addressing the questions posed in this paper. There are numerous successful empirical applications of the gravity model dating back to the early 1960s.

The gravity model relates a measure of bilateral trade to the economic mass of the two countries and the distance between them:

\[
TRADE_{ij} = (Y_{ij} Y_{ji})^{\alpha} (P_i P_j)^{\theta} D_{ij}^{\beta} e^{\mu_{ij}},
\]

where \(TRADE_{ij}\) is bilateral trade between country \(i\) and country \(j\), \(Y_i\) is nominal GDP in country \(i\), \(Y_j\) is nominal GDP in country \(j\), \(P_i\) and \(P_j\) are population in the two countries, \(D_{ij}\) is geographic distance between country \(i\) and country \(j\), and \(i\) is a time subscript. We expect trade to be positively affected by economic mass (\(\alpha > 0\)); negatively related to the level of population (\(\theta < 0\)), indicating that larger countries tend to be more self-sufficient or, alternatively, that poorer countries—countries with larger populations for a given level of GDP—trade less than richer countries; and negatively related to distance (\(\beta < 0\)). \(\mu_{ij}\) is given by

\[
\mu_{ij} = \kappa + \varphi_{\lambda i},
\]

where \(\kappa\) are fixed effects for trade between African and other countries, \(\varphi_{\lambda i}\) are fixed effects for other potential determinants of bilateral trade (specifically, for membership or participation in the Lomé Convention and the CFA franc zone and...
for countries that share common borders or a common language. We assume that disturbances are independent and identically distributed and enter equation (1) additively.

Recent papers by Deardorff (1998) and Anderson and van Wincoop (2003) emphasize the importance not only of distance (or trade barriers) between two countries, but also of the average trade barriers of the two countries to all their other trading partners. The empirical gravity model literature often includes a remoteness variable, defined in some studies as the weighted distance to all trade partners,9 as a proxy for this:

\[ R_i = \sum_j w_j D_{ij}, \]

for \( i \neq j \) and with \( w_j = Y_j / \sum_i Y_i \) for all \( i \). A similar variable is defined for country \( j \). The more remote a pair of countries is from the rest of the world, the more they will tend to trade with each other.

Thus, the specification we estimate is:

\[ TRADE_{ij} = (Y_i Y_j)^{\alpha} (P_i P_j)^{\beta} (R_i R_j)^{\gamma} e^{\mu_{ij}} + \epsilon_{ij} \]  

(4)

This formulation allows straightforward tests of whether, after controlling for the economic size, distance, remoteness, and other factors, bilateral trade between or within regions in Africa is different from trade of other regions—the test is simply whether the estimated fixed effects (\( \kappa \)) are significant.

The model is estimated for three points in time—1980, 1990, and 2000. This serves as both a comparison with and an update of other work conducted for earlier periods and also facilitates the analysis of evolution in trade over time. Data and their sources are described in Appendix I.

Following Coe and Hoffmaister (1999) and similar to Coe, Subramanian, and Tamirisa, with Bhavnani (2002; hereafter CST, 2002), we employ NLS estimation on a sample that includes zero-valued observations for bilateral trade. Since Africa’s trade is relatively concentrated, the share of zero-valued observations in the dataset is not trivial (about 6-11 percent between 1980 and 2000), and thus the choice of an appropriate methodology critically depends on how a given estimator deals with zero-valued observations. The main advantage of an NLS estimator is that it adequately incorporates the information contained in zero-valued observations by treating them as cases where trade is actually zero rather than negligible or not observed. CST (2002) also confirm the advantage of the nonlinear estimation procedure employed in this paper over the alternatives.

Critical values for hypothesis testing are obtained by bootstrapping with 1,000 replications, since skewness and kurtosis tests indicate that residuals are not

---

9Our specification of remoteness is similar to that in Frankel and Wei (1998).

10Instead of remoteness, Anderson and van Wincoop (2001) propose a “multilateral resistance” term that is a function of equilibrium price indices, which are not observable. Anderson and van Wincoop are able to estimate their model using demanding computational methods. Coe, Subramanian, and Tamirisa, with Bhavnani (2002), however, show that in the nonlinear framework, the specification with fixed effects yields similar effects to that with remoteness.
distributed normally. Hypothesis testing under the assumption of residuals’ normality would be invalid in this case. Point estimates, in contrast, are independent of the distribution of residuals.

There are two alternatives to the methodology we employ. The first is to exclude zero-valued observations (as in Frankel, 1997) from the sample. However, this would be equivalent to nonrandom screening of the data and could bias the results. It would also be unsatisfactory from a conceptual point of view, since zero values in our data set indicate the lack of trade, not missing values. Given our focus on Africa’s trade, which has a disproportionate share of zero-valued observations (about double the share for the entire sample), including zero-valued observations is desirable in this study.

The second alternative is to assign arbitrarily small values to the zero-valued observations and then estimate the model in the logarithmic form. This is the approach adopted in Wang and Winters (1991) and Foroutan and Pritchett (1993). However, using ordinary least squares (OLS) and Tobit estimation procedures on a sample in which zero-valued observations are replaced with small values is not free from problems either. Since the logs of small values are large negative numbers, this approach confers unduly large weights on the adjusted zero-valued observations. We compare below the results obtained from using these alternative methodologies to our results.

III. Africa’s Trade

Africa’s trade is not uniform. In particular, there are important differences in the trade performance of Anglophone and Francophone Africa in the past two decades (Figure 2). After declining during most of the 1980s, Anglophone Africa’s overall trade grew markedly through 2000. Francophone Africa’s trade has grown more steadily, but without the dynamism exhibited by Anglophone Africa since the late 1980s. For Anglophone Africa, the largest increases were recorded in its trade with the South and within the region, while trade with the North grew at a slower pace. While Francophone Africa’s trade exhibited a similar geographic pattern, with trade with the South growing faster than trade with the North, the magnitude of growth rates in trade in each of these markets has been well below that for Anglophone Africa.

In the formal analysis, we represent the different components of African trade by various dummies. (See Appendix II for a list of the countries that are included in the dummies.) AFR-ANG is a dummy for Anglophone Africa and takes on a value of 1 when an Anglophone African country is either a reporting or a partner country. AFR-FRN is the analogue for Francophone Africa. The other dummies are all bilateral. The AFR-ANG (AFR-FRN) dummy represents trade among Anglophone (Francophone) African countries. Similarly, AFRS-ANG (AFRS-FRN) denotes Anglophone (Francophone) African countries’ trade with other developing countries.

AFRNNEU-ANG (AFRNNEU-FRN) denotes Anglophone (Francophone) African countries’ trade with advanced11 countries other than those in the EU that

11 As defined in the IMF’s World Economic Outlook.
Figure 2. Africa’s Trade, 1981–2000
(Billions of U.S. dollars)

A. Africa as a Whole

B. Total Trade and Trade with the North

C. Trade with the South

D. Intraregional Trade

Source: IMF’s Direction of Trade Statistics.

©International Monetary Fund. Not for Redistribution
grant preferences under the Lomé Convention. The rationale for differentiating Africa's trade with EU countries and other advanced countries is related to the effect of the Lomé Convention on trade. The long history of preferential trade embodied in the Lomé Convention has to be controlled for in determining how typical trade is between Africa and the North. If Africa traded more than expected with the North because of preferential arrangements, that would not necessarily shed light on the underlying pattern of trade. A free trade agreement dummy (denoted by FTA) controls for preferential trading relationships.

For the cross-section data, the main findings on whether Africa undertrades are as follows (Table 2). First, currently, Francophone Africa is an undertrader in terms of its overall trade and its trade with the North. The coefficients on the Francophone dummy are negative and significant for 2000; however, Francophone Africa's trade with itself and other developing countries is unexceptional.

Second, and disturbingly, the respective coefficients have become more negative over time, signifying increasing disintegration of Francophone Africa from global trade. For example, Francophone Africa's overall trade, which was normal in 1980, was about 70 percent less than average by 2000 (see columns 1 and 3 in Table 2).

Third, while Francophone Africa is progressively undertrading, the disintegration effect is apparently more pronounced in its trade with the North than with any other group of countries. Between 1980 and 2000, this trade went from being normal to about 80 percent below average (columns 7 and 9 in Table 2). While the coefficient on trade with the South turned negative between 1980 and 2000, it remained insignificant. Only its intra-regional trade shows no clear signs of disintegration. Since technology transfer embodied in capital goods is one of the important channels for trade to enhance growth (see Coe, Helpman, and Hoffmaister, 1997, for example), Francophone Africa's substantial undertrading with its Northern partners, typically the most important suppliers of capital and high-technology goods, raises concerns about respective implications for its growth prospects.

For Anglophone Africa, the results are qualitatively different. In 2000, Anglophone Africa was an average trader in aggregate, with the coefficient on the dummy being negative but insignificant, and also an average trader in terms of the components of trade.

---

12Since the study covers the period from 1980 to 2000, this dummy does not cover recent entrants to the EU.
13Although the non-Lomé industrial countries also grant preferences to Africa under the Generalized System of Preferences, these are less broad in product coverage and subject to greater restrictions and conditions than preferences granted under the Lomé Convention.
14This dummy is time-varying in the sense that it reflects common membership in a preferential arrangement at the time of (and after) its formal inception. Thus, for 1980, FTA includes the following arrangements: European Free Trade Association (EFTA), EU-Turkey agreement, the Andean Pact, Australia-New Zealand agreement, Lomé, and Franc de la Communauté Française d'Afrique (CFA) zone. For 1990, it includes, in addition to the above, the Israel-U.S. free trade agreement. For 2000, it includes, in addition, the Israel-EU free trade agreement. Association of South East Asian Nations (ASEAN) Free Trade Area (AFTA), Southern Common Market (MERCOSUR), North American Free Trade Agreement (NAFTA), the Chile-U.S. free trade agreement, and the EU-Northern Africa (also called the EU-Mediterranean) agreements.
15The extent of undertrading or overtrading is simply the exponential of the coefficient on the dummy minus one; in this case, exp(-1.17) - 1, which is equal to 0.69.
Table 2. Africa's Trade

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Africa’s Trade with the World</th>
<th>Intra-African Trade</th>
<th>Africa’s Trade with the North</th>
<th>Africa’s Trade with the South</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>0.994*</td>
<td>0.887*</td>
<td>0.740*</td>
<td>0.994*</td>
</tr>
<tr>
<td>POP</td>
<td>-0.203*</td>
<td>-0.090</td>
<td>0.113</td>
<td>-0.203*</td>
</tr>
<tr>
<td>DIST</td>
<td>-0.400*</td>
<td>-0.316*</td>
<td>-0.320*</td>
<td>-0.400*</td>
</tr>
<tr>
<td>REM</td>
<td>1.151*</td>
<td>0.871*</td>
<td>0.465**</td>
<td>1.151*</td>
</tr>
<tr>
<td>AFR-ANG</td>
<td>-0.052</td>
<td>0.202</td>
<td>-0.238</td>
<td></td>
</tr>
<tr>
<td>AFR-FRN</td>
<td>-0.068</td>
<td>-0.642</td>
<td>-1.169*</td>
<td></td>
</tr>
<tr>
<td>AFRAFR-ANG</td>
<td></td>
<td></td>
<td></td>
<td>0.501</td>
</tr>
<tr>
<td>AFRAFR-FRN</td>
<td></td>
<td></td>
<td></td>
<td>-2.193</td>
</tr>
<tr>
<td>AFRNNEU-ANG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFRNNEU-FRN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFRS-ANG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFRS-FRN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>0.454*</td>
<td>0.400*</td>
<td>0.516*</td>
<td>0.454*</td>
</tr>
<tr>
<td>LNG</td>
<td>0.275</td>
<td>-0.325</td>
<td>0.033</td>
<td>0.275</td>
</tr>
<tr>
<td>FFA</td>
<td>0.655**</td>
<td>0.782**</td>
<td>0.769*</td>
<td>0.655**</td>
</tr>
<tr>
<td>Adjusted R-squares</td>
<td>0.863</td>
<td>0.873</td>
<td>0.907</td>
<td>0.863</td>
</tr>
<tr>
<td>F statistic</td>
<td>1,518</td>
<td>1,789</td>
<td>2,561</td>
<td>1,518</td>
</tr>
<tr>
<td>Number of observations</td>
<td>2,415</td>
<td>2,593</td>
<td>2,613</td>
<td>2,415</td>
</tr>
</tbody>
</table>

Source: Authors’ estimates.

* (**) indicates significance at the 5 (10) percent level. See Appendix I for definitions of explanatory variables.
Encouragingly, and in contrast with Francophone Africa, Anglophone Africa shows some signs of reversing its disintegration from trade with the North. The coefficient, which was negative and significant in 1980 (signifying undertrading of about 63 percent as column 7 in Table 2 shows) becomes positive by 2000, albeit insignificantly so (column 9 in Table 2). A similar pattern is exhibited in its trade with other developing countries and itself.

Some final remarks can be made on the more general aspects of the results. Coefficients on the standard determinants of the gravity models, such as income, population, and distance, are correctly signed, statistically significant, and yield plausible elasticity estimates broadly in line with those obtained in the literature.

Besides implications for Africa’s trade, the results also shed light on the ongoing process of globalization more generally. To the extent that globalization connotes the decreasing importance of geography, the evidence lends support to this proposition. The elasticity of trade with respect to distance declined by almost 20 percent (from –.40 in 1980 to –.32 in 2000), with all the decline occurring in the 1990s.16 This is consistent with rapid technological progress and wide-ranging liberalization in the trade-related service sectors during the 1990s.

IV. Robustness Tests and Explanations for the Contrasting Trade Performance of Francophone and Anglophone Africa

A number of factors may help explain the dissimilar globalization experiences of Francophone and Anglophone Africa. Differences in the commodity composition of trade and in currency arrangements may play a role in this regard. Likewise, differences in the efficiency of transport and communication sectors could manifest themselves in transaction costs and, thus, trade performance. While a detailed examination of the factors underlying differences in performance of Francophone and Anglophone Africa is beyond the scope of this paper, we can conduct some basic tests of the possible explanations.

Countries in Francophone Africa could be disintegrating from trade because they are primary commodity exporters.17 In this view, African disintegration could merely reflect the decline in its terms of trade that has been evident during the past several decades. To test this, we run regressions including a dummy for primary commodity exporters (PRIM).18 Table 3 (columns 1–3) contains these results. The PRIM dummy is positive and significant for 1980 and 1990, suggesting that being a commodity exporter conferred an advantage in those periods. In 2000, this dummy is insignificant but the sign is still positive. This implies that commodity exporters are not uniquely disadvantaged in trading terms. More important for our purposes, the inclusion of the dummy does not alter the basic results; indeed, they are strengthened. In particular, the Francophone Africa

16See CST (2002) for more details.
17In our sample, four out of six Francophone countries and six out of ten Anglophone countries are primary commodity exporters. Therefore, the set of Francophone countries is not intrinsically more commodity dependent than the set of Anglophone countries.
18Primary commodity exporters are defined based on the IMF’s World Economic Outlook.
Table 3. Robustness Analysis

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Primary Commodities</th>
<th>Transaction Costs</th>
<th>Exchange Rate Misalignments</th>
<th>Panel Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>1.073*</td>
<td>0.899*</td>
<td>0.740*</td>
<td>0.994*</td>
</tr>
<tr>
<td>POP</td>
<td>-0.229*</td>
<td>-0.096</td>
<td>0.113</td>
<td>-0.203*</td>
</tr>
<tr>
<td>DIST</td>
<td>-0.414*</td>
<td>-0.320*</td>
<td>-0.320*</td>
<td>-0.400*</td>
</tr>
<tr>
<td>REM</td>
<td>1.198*</td>
<td>0.867*</td>
<td>0.465</td>
<td>1.150*</td>
</tr>
<tr>
<td>AFR-ANG</td>
<td>-0.575</td>
<td>-0.132</td>
<td>-0.240</td>
<td>-0.051</td>
</tr>
<tr>
<td>AFR-FRN</td>
<td>-0.607</td>
<td>-0.968*</td>
<td>-1.170*</td>
<td>-0.791*</td>
</tr>
<tr>
<td>AFR-CFA</td>
<td></td>
<td></td>
<td></td>
<td>0.330</td>
</tr>
<tr>
<td>DAFR-ANG</td>
<td>-0.008</td>
<td>0.023</td>
<td>-0.026</td>
<td></td>
</tr>
<tr>
<td>DAFR-FRN</td>
<td>-0.009</td>
<td>-0.073</td>
<td>-0.131*</td>
<td></td>
</tr>
<tr>
<td>AFR-ANG-TREND</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFR-FRN-TREND</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>0.481*</td>
<td>0.400*</td>
<td>0.516*</td>
<td>0.454**</td>
</tr>
<tr>
<td>LNG</td>
<td>0.252</td>
<td>-0.330</td>
<td>0.033</td>
<td>0.275</td>
</tr>
<tr>
<td>FTA</td>
<td>0.725**</td>
<td>0.784</td>
<td>0.769*</td>
<td>0.655</td>
</tr>
<tr>
<td>PRIM</td>
<td>0.853*</td>
<td>0.497**</td>
<td>0.002</td>
<td>0.877</td>
</tr>
<tr>
<td>Adjusted R-squares</td>
<td>0.877</td>
<td>0.874</td>
<td>0.907</td>
<td>0.863</td>
</tr>
<tr>
<td>F statistic</td>
<td>1.567</td>
<td>1.637</td>
<td>2.327</td>
<td>1.518</td>
</tr>
<tr>
<td>Number of observations</td>
<td>2,415</td>
<td>2,593</td>
<td>2,613</td>
<td>2,415</td>
</tr>
</tbody>
</table>

Source: Authors' estimates.

* (**) indicates significance at the 5 (10) percent level. See Appendix I for definitions of explanatory variables.

Panel regressions include unreported time dummies.
dummy is negative and significant in both 1990 and 2000. This implies that it is now an undertrader and has experienced trade disintegration over the last 20 years. For Anglophone Africa, the results are as broadly the same with and without the inclusion of the primary commodity dummy.19

Next we consider if high transport and other trade-related costs are a particular obstacle for Africa’s trade (Table 3, columns 4–6). The evolution in these costs would, of course, be affected by certain exogenous factors, such as technological progress, for example. It also crucially depends on domestic policies, which determine the efficiency of certain trade-related service industries, such as transport, port operations, communications, and distribution. Again, the gravity model allows for some preliminary testing of hypotheses about the magnitude of trade-related costs and their evolution over time. As discussed earlier, the distance variable could be considered a proxy for such costs. To test for their effects on African trade, we interacted a dummy for Anglophone and Francophone Africa with the distance variable, denoted in Table 3 by DAFR-ANG and DAFR-FRN, respectively. The results point toward an increase in trade costs for Francophone Africa that decreases trade by about 11 percent between 1980 and 2000.20

Another explanation for the differential performance of Francophone and Anglophone Africa relates to exchange rate misalignments (Table 3, columns 7–9). Countries in the CFA zone have pegged their exchange rate to the French franc.21 The serious and persistent misalignment of the CFA franc until 1994, when it was devalued by 50 percent, is widely acknowledged to have had a debilitating effect on trade performance of the CFA zone countries. To test for such misalignment effects, we redefine the Francophone Africa dummy to exclude the non-CFA zone countries (variable AFR-CFA in Table 3). All the results for Francophone Africa, including the negative and statistically significant coefficient for 2000, broadly carry over to the CFA zone countries. While not necessarily conclusive, the results are generally consistent with the possibility that years of misalignment in the CFA zone might have led its members to undertrading. Future research is needed, however, to substantiate this explanation.

Finally, to confirm that our cross-section estimates for selected years are generally valid, we estimated the same specification on a panel data set comprising annual data for the period 1980–2000. Following CST (2002), we included (unreported) time dummies to capture the effects of changes in prices and exchange rates over time. The results are reported in Table 3 (columns 10 and 11) and are consistent with those obtained for the cross-section dataset.

19 Of course, one explanation for Africa’s disintegration could relate to the increasing vertical specialization that is a more important feature of manufacturing trade than trade in commodities (Hummels, Ishii, and Yi, 2001). Vertical specialization means that goods cross multiple borders in the process of being manufactured, counting as trade each time they do so. Trade thus is a gross rather than a value-added measure. With Africa specializing in commodities, it is excluded from trade-intensive manufacturing transactions. However, as Figure I.B shows, Africa’s disintegration (particularly that of Francophone Africa) appears to be also evident in trade in primary commodities. This is confirmed by the fact that the disintegration results in Table 3 appear to be present, even after controlling for primary commodity dependence.

20 For Anglophone Africa, the distance coefficient is negative in 2000 but insignificant.

21 Since 1999, the peg is to the euro.
In the specification without a time trend, the Francophone Africa dummy is negative and statistically significant while that on Anglophone Africa is insignificant (column 10). In column 11, the Africa dummies are interacted with a time trend to measure the integration/disintegration effect over time. The trend for Anglophone Africa is negative but insignificant while that for Francophone Africa is negative and significant, confirming the trend of disintegration for the latter set of countries.

V. Comparison with Earlier Literature

To complement the robustness analysis, we next examine the key factors driving our results. To test whether our sample selection is the driving factor, we estimate the model using the methodologies employed by the previous researchers and restricting our sample accordingly. The results of this calibration exercise are reported in Table 4, column 1 for the Foroutan and Pritchett (1993) paper and in columns 2 and 3 for Coe and Hoffmaister (1999).

Recall that Foroutan and Pritchett (1993) tested whether intra-African trade was different from trade between other developing countries. Restricting our sample in line with their study and applying their estimation procedure, we find similar results—the intra-African trade dummy for 1980 is positive and significant, albeit at the 10 percent level.

We then replicate Coe and Hoffmaister’s (1999) results, focusing on North-South trade in a panel data context. Replicating their setup and the estimation method, we find that the coefficient for the dummy for Africa’s trade with the North (AFRNEU) is negative and significant without the time trend; including the trend makes the coefficient positive but implies a significant disintegration effect over time. Coe and Hoffmaister (1999) obtain similar results. Thus, we can eliminate the sample as the source of the difference.

Next, to isolate the role of the methodology, we reestimate our basic model (Table 2, columns 1–3) using the methodologies of the previous papers. In this exercise, methodology is the only difference. The results are reported in Table 4. In columns 4–6, for Foroutan and Pritchett (1993), the coefficients on both Africa dummies are negative and significant in 1980 and 1990 and both decline and become insignificant in 2000. Clearly, the use of the Foroutan and Pritchett methodology thus paints an opposite picture compared to our results, with both Anglophone and Francophone Africa reversing the process of disintegration over time.

The application of the Coe and Hoffmaister methodology yields results (columns 7–8) that show that the coefficients on the regional dummies as well as the coefficients of these dummies interacted with the time trend are insignificant. Again, there is a striking contrast with our results, obtained with bootstrapping, where the Francophone dummy and its interaction with the time trend are both significant. We conclude, therefore, that the key factor driving our results is the methodology—nonlinear least squares with bootstrapping on a sample including zero-valued observations—which we consider preferable for reasons explained earlier (and more fully in CST, 2002).
<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Replication of Results</th>
<th>Comparison of Methodologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>2.462*</td>
<td>2.078*</td>
</tr>
<tr>
<td>POP</td>
<td>0.166</td>
<td>-0.536*</td>
</tr>
<tr>
<td>DIST</td>
<td>-2.701*</td>
<td>-1.166*</td>
</tr>
<tr>
<td>REM</td>
<td>2.585*</td>
<td>1.044*</td>
</tr>
<tr>
<td>AFR-ANG</td>
<td>-2.047*</td>
<td>-1.610*</td>
</tr>
<tr>
<td>AFR-FRN</td>
<td>2.413**</td>
<td></td>
</tr>
<tr>
<td>AFR-ANG-TREND</td>
<td>-0.343*</td>
<td></td>
</tr>
<tr>
<td>AFR-FRN-TREND</td>
<td>0.662*</td>
<td></td>
</tr>
<tr>
<td>AFRNNEU</td>
<td>-0.361*</td>
<td></td>
</tr>
<tr>
<td>AFRNNEU-TREND</td>
<td>-0.067*</td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>-0.669</td>
<td>-0.380</td>
</tr>
<tr>
<td>LNG</td>
<td>4.560*</td>
<td>2.631*</td>
</tr>
<tr>
<td>FTA</td>
<td>0.925</td>
<td>1.933*</td>
</tr>
<tr>
<td>Adjusted R-squares</td>
<td>0.818</td>
<td>0.818</td>
</tr>
<tr>
<td>F statistic</td>
<td>0.056</td>
<td>0.197*</td>
</tr>
<tr>
<td>Number of observations</td>
<td>974</td>
<td>2,593</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-3066</td>
<td>2,613</td>
</tr>
<tr>
<td>Wald Chi-squared</td>
<td>376</td>
<td>13,503</td>
</tr>
</tbody>
</table>

Source: Authors' estimates.

* (**) indicates significance at the 5 (10) percent level. See Appendix I for definitions of explanatory variables.

1 Tobit estimation on a sample including South-South trade. Zero-valued observations are replaced with a small positive value.
2 Nonlinear least squares without bootstrapping on a panel sample including North-South trade. Zero-valued observations are included.
3 Tobit estimation on the full sample. Zero-valued observations are replaced with a small positive value.
4 Nonlinear least squares without bootstrapping on the full panel sample. Zero-valued observations are included.
VI. Conclusion

The popular “marginalization-from-trade” hypothesis argues that Africa has not benefited from globalization because it has not globalized in the first place. This view has been challenged recently in a series of papers, which have shown more formally that Africa has not been left behind: Africa trades as much as any other set of traders, given the underlying determinants of trade.

This paper, however, finds support for the “marginalization-from-trade” hypothesis, but only for Francophone Africa. Francophone Africa is an undertrader and, moreover, the degree of its undertrading has increased over time. Anglophone Africa appears to have remained an average trader for the past two decades.

Ominously, Francophone Africa’s trade with the North appears to have suffered most over time: ominous, because trade with the technologically advanced North is one of the more important channels for globalization’s benefits to be disseminated to Africa. Trade with the North also constitutes the largest component of Africa’s overall trade and is hence likely to have a more significant impact on growth. Anglophone Africa, on the other hand, which had undertraded with the North in 1980, has reversed this process and became an average trader with the advanced countries by 2000.

The robustness analysis points to two possible explanations for the contrasting performance of Anglophone and Francophone Africa. Trade-related costs seem to have increased for Francophone Africa. Also, the currency arrangements in the CFA zone may have exerted a depressing effect on trade, owing to persistent exchange rate misalignments. More research is needed in the future, however, to substantiate these findings. The results suggest that primary commodity dependence is not a factor in explaining trade developments for Africa consistent with the decline in Africa’s share of global trade even in primary commodities.

Overall, the results in this paper suggest that the sanguine policy prescription, stemming from the view that Africa trades adequately, may need to be reconsidered. Policy action to assist Africa to better exploit its trade opportunities would seem appropriate. Of course, views differ on the nature of such action—from calls for active government intervention to facilitate export diversification (Sachs, 2000) to the need to maintain competitiveness (World Bank, 2000). At the very least, trade regimes that continue to be highly distorted in a number of African countries need to be liberalized (Subramanian and others, 2000).
# APPENDIX I

## Legend and Data Sources

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition (Source)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trade</td>
<td>Sum of bilateral exports and imports (Direction of Trade Statistics, IMF)</td>
</tr>
<tr>
<td>GDP</td>
<td>GDP of the reporting country times the GDP of the partner country (World Economic Outlook (WEO), IMF)</td>
</tr>
<tr>
<td>POP</td>
<td>Population of the reporting country times population of the partner country (WEO)</td>
</tr>
<tr>
<td>DIST</td>
<td>Geographical distance between capitals of the reporting and partner countries (Fitzpatrick and Modlin, 1986)</td>
</tr>
<tr>
<td>REM</td>
<td>Remoteness is the weighted distance to all trading partners (as defined in the text)</td>
</tr>
<tr>
<td>ADJ</td>
<td>Dummy that takes on a value of 1 when reporting and partner countries share a common border</td>
</tr>
<tr>
<td>LNG</td>
<td>Dummy that takes on a value of 1 when reporting and partner countries share a common language (Coe and Hoffmaister, 1999)</td>
</tr>
<tr>
<td>AFR-ANG</td>
<td>Dummy that takes on a value of 1 when either the reporting or partner country is an Anglophone African country</td>
</tr>
<tr>
<td>AFR-FRN</td>
<td>Dummy that takes on a value of 1 when either the reporting or partner country is a Francophone African country</td>
</tr>
<tr>
<td>AFRS-ANG</td>
<td>Dummy that takes on a value of 1 when the reporting or partner country is an Anglophone African country and the partner or reporting country is a developing country</td>
</tr>
<tr>
<td>AFRS-FRN</td>
<td>Dummy that takes on a value of 1 when the reporting or partner country is a Francophone African country and the partner or reporting country is a developing country</td>
</tr>
<tr>
<td>AFRAFR-ANG</td>
<td>Dummy that takes on a value of 1 when the reporting and partner country are Anglophone African countries</td>
</tr>
<tr>
<td>AFRAFR-FRN</td>
<td>Dummy that takes on a value of 1 when the reporting and partner country are Francophone African countries</td>
</tr>
<tr>
<td>AFRNNEU-ANG</td>
<td>Dummy that takes on a value of 1 when the reporting or partner country is an Anglophone African country and the partner or reporting country is a non-Lome industrial country</td>
</tr>
<tr>
<td>AFRNNEU-FRN</td>
<td>Dummy that takes on a value of 1 when the reporting or partner country is a Francophone African country and the partner or reporting country is a non-Lome industrial country</td>
</tr>
<tr>
<td>AFR-CFA</td>
<td>Dummy that takes on a value of 1 if a country is a member of the CFA currency zone</td>
</tr>
<tr>
<td>FTA</td>
<td>Dummy that takes on a value of 1 when the reporting or partner country is a member of one of the free trade or regional integration agreements listed in footnote 14 of the paper</td>
</tr>
<tr>
<td>PRIM</td>
<td>Dummy that takes on a value of 1 if a country is a primary commodity exporter as defined in WEO</td>
</tr>
<tr>
<td>AFR-CFA</td>
<td>Dummy that takes on a value of 1 when reporting and partner countries are members of the CFA currency zone</td>
</tr>
<tr>
<td>DAFR-ANG</td>
<td>Dummy that takes on a value of the distance variable when the reporting country is an Anglophone African country and zero otherwise</td>
</tr>
<tr>
<td>DAFR-FRN</td>
<td>Dummy that takes on a value of the distance variable when the reporting or partner country is a Francophone African country and zero otherwise</td>
</tr>
<tr>
<td>AFRAFR</td>
<td>Dummy that takes on a value of 1 when the reporting and partner country are African countries</td>
</tr>
<tr>
<td>AFR-ANG-TREND</td>
<td>The Anglophone Africa dummy (AFR-ANG) variable interacted with a time trend</td>
</tr>
<tr>
<td>AFR-FRN-TREND</td>
<td>The Francophone Africa dummy (AFR-FRN) variable interacted with a time trend</td>
</tr>
<tr>
<td>AFRNNEU</td>
<td>Dummy that takes on a value of 1 when the reporting or partner country is an African country and the partner or reporting country is a non-Lome industrial country</td>
</tr>
<tr>
<td>AFRNNEU-TREND</td>
<td>The AFRNNEU dummy variable interacted with a time trend</td>
</tr>
</tbody>
</table>
Appendix II
Lists of Countries

<table>
<thead>
<tr>
<th>Africa</th>
<th>Francophone Africa</th>
<th>Anglophone Africa</th>
<th>CFA</th>
<th>Non-Lome Industrial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameroon</td>
<td>Cameroon</td>
<td>Cameroon</td>
<td>Cameroon</td>
<td>Cameroon</td>
</tr>
<tr>
<td>Congo, Rep. of</td>
<td>Côte d’Ivoire</td>
<td>Côte d’Ivoire</td>
<td>Côte d’Ivoire</td>
<td>Côte d’Ivoire</td>
</tr>
<tr>
<td>Côte d’Ivoire</td>
<td>Ethiopia</td>
<td>Ghana</td>
<td>Cameroon</td>
<td>Cameroon</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>Kenya</td>
<td>Malawi</td>
<td>Congo, Rep. of</td>
<td>Congo, Rep. of</td>
</tr>
<tr>
<td>Ghana</td>
<td>Malawi</td>
<td>Mauritius</td>
<td>Côte d’Ivoire</td>
<td>Côte d’Ivoire</td>
</tr>
<tr>
<td>Kenya</td>
<td>Nigeria</td>
<td>Tanzania</td>
<td>Ethiopia</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>Madagascar</td>
<td>Uganda</td>
<td>Madagascar</td>
<td>Ghana</td>
<td>Ghana</td>
</tr>
<tr>
<td>Malawi</td>
<td>Malawi</td>
<td>Malawi</td>
<td>Kenya</td>
<td>Kenya</td>
</tr>
<tr>
<td>Mauritius</td>
<td>Nigeria</td>
<td>Zimbabwe</td>
<td>Mauritius</td>
<td>Mauritius</td>
</tr>
<tr>
<td>Nigeria</td>
<td>Senegal</td>
<td>South Africa</td>
<td>Nigeria</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Senegal</td>
<td>Tanzania</td>
<td>Tanzania</td>
<td>Senegal</td>
<td>Senegal</td>
</tr>
<tr>
<td>South Africa</td>
<td>Uganda</td>
<td>Uganda</td>
<td>South Africa</td>
<td>South Africa</td>
</tr>
<tr>
<td>Tanzania</td>
<td>Zambia</td>
<td>Zambia</td>
<td>Tanzania</td>
<td>Tanzania</td>
</tr>
<tr>
<td>Uganda</td>
<td>Zimbabwe</td>
<td>Denmark</td>
<td>Uganda</td>
<td>Uganda</td>
</tr>
<tr>
<td>Zambia</td>
<td>Zimbabwe</td>
<td>France</td>
<td>Zimbabwe</td>
<td>Zimbabwe</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>Germany</td>
<td>Greece</td>
<td>Denmark</td>
<td>Denmark</td>
</tr>
<tr>
<td></td>
<td>Iceland</td>
<td>Ireland</td>
<td>France</td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>Japan</td>
<td>Italy</td>
<td>Germany</td>
<td>Germany</td>
</tr>
<tr>
<td></td>
<td>New Zealand</td>
<td>Netherlands</td>
<td>Greece</td>
<td>Greece</td>
</tr>
<tr>
<td></td>
<td>Norway</td>
<td>Portugal</td>
<td>Ireland</td>
<td>Ireland</td>
</tr>
<tr>
<td></td>
<td>Sweden</td>
<td>Spain</td>
<td>Italy</td>
<td>Italy</td>
</tr>
<tr>
<td></td>
<td>Switzerland</td>
<td>United States</td>
<td>Netherlands</td>
<td>Netherlands</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td></td>
<td>Portugal</td>
<td>Portugal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Spain</td>
<td>Spain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>United Kingdom</td>
<td>United Kingdom</td>
</tr>
</tbody>
</table>

(continued)
### IS AFRICA INTEGRATED IN THE GLOBAL ECONOMY?

**APPENDIX II, concluded**

**Full Sample**

| Algeria* | Guatemala | Pakistan |
| Argentina | Guyana* | Paraguay* |
| Australia | Hong Kong, SAR | Peru* |
| Austria | Iceland | Philippines |
| Bangladesh | India | Portugal |
| Bolivia* | Indonesia | Saudi Arabia* |
| Brazil | Iran* | Senegal |
| Cameroon | Ireland | Singapore |
| Canada | Israel | Spain |
| Chile* | Italy | Sri Lanka |
| China | Jamaica | Sweden |
| Colombia | Japan | Switzerland |
| Congo, Repubic of* | Jordan | Taiwan, Province of China |
| Congo, Democratic Republic of* | Kenya | Tanzania* |
| Costa Rica | Korea | Thailand |
| Côte D'Ivoire* | Madagascar* | Tunisia* |
| Denmark | Malawi* | Turkey |
| Egypt | Malaysia | Uganda |
| Ethiopia | Mexico | United Kingdom |
| Finland | Mauritius | United States |
| France | Morocco | Uruguay |
| Germany | Netherlands | Venezuela* |
| Ghana* | New Zealand | Zambia* |
| Greece | Nigeria* | Zimbabwe* |
| | Norway | |

1 Asterisks denote primary commodity exporters.
REFERENCES


The High-Yield Spread as a Predictor of Real Economic Activity: Evidence of a Financial Accelerator for the United States

ASHOKA MODY and MARK P. TAYLOR

Previous studies find that the interest rate term spread predicts real U.S. economic activity. We show that this relationship breaks down for the 1990s and suggest that its earlier success was due to high and volatile inflation. We find, however, that the high-yield spread (HYS) between “junk bond” and government bond yields predicts real activity during the 1990s—especially high levels of the HYS. We also find that the HYS works through both the demand and the supply side of the economy. We interpret our findings as supportive of a financial accelerator mechanism. [JEL E33, E44]

The slope of nominal yield curve, or the term spread, was shown by studies published in the late 1980s and early 1990s to have significant predictive content for future real economic activity, both in the United States and in Europe. Following the early work by Stock and Watson (1989) and Estrella and Hardouvelis (1991), however, confidence in the predictive power of the term spread waned a little when it failed to predict the 1990–91 U.S. recession (see, for example, Dotsey, 1998). Nevertheless, further work by, among others, Estrella and Mishkin (1997) and Plosser and Rouwenhorst (1994) seemed to establish that its power as a leading indicator of real economic activity had not evaporated. This somewhat mixed evidence on the forecasting performance of the term spread remains in the literature.

1Mark P. Taylor is Professor of Macroeconomics at the University of Warwick and a Fellow of the Centre for Economic Policy Research. He was a Visiting Scholar in the Research Department when this paper was written. Ashoka Mody is a Division Chief in the Research Department. We are grateful to two anonymous referees for helpful comments on an earlier version of the paper and also to Lutz Kilian for insightful comments and advice on our forecasting exercises.


2See also Laurent (1988).
Dotsey (1998, p. 50), for example, in a paper that is generally supportive of the view that the term spread does indeed contain predictive information, notes that "that conclusion must be tempered, however, by the observation that over more recent periods the spread has not been as informative as it has been in the past."

Gertler and Lown (1999), drawing on the theory of the financial accelerator (see, for example, Bernanke and Gertler, 1995; Bernanke, Gertler, and Gilchrist, 1999, and the references cited therein), argue that an alternative financial variable should also have predictive power for real economic activity—this is the premium required on less than investment-grade corporate bonds (also referred to as "high-yield" or "junk" bonds) over government debt or AAA-rated corporate bonds. Gertler and Lown (1999) provide some empirical support for this view based on correlation and impulse-response analysis, using data on the high-yield spread and a measure of the U.S. output gap.

We seek in this paper to contribute to this literature in a number of ways. First, we examine the robustness of the term spread as a predictor of economic activity by estimating long-horizon regressions covering broadly three periods—the 1960s, the 1970s and 1980s, and the 1990s. In brief, we find that the term spread does not predict real economic activity well for the most recent period, although it does perform well in this capacity for the 1970s and 1980s. Interestingly, however, we find that the predictive content of the term spread appears to be unique to the 1970s and 1980s, in that we also find it to be only weakly present in the data for the 1960s.

We then move on to examine the predictive content of the high-yield spread. We believe this analysis to be the first using the long-horizon regression, which has been the standard tool for judging the predictive ability of the term spread. We find that the high-yield spread has a high predictive content. In addition, we find evidence of some nonlinearity, in that abnormally high levels of the high-yield spread have significant additional short-term predictive power. Also, despite the low power of out-of-sample tests of forecast accuracy (Inoue and Kilian, 2002), we find that out-of-sample forecasts of movements in economic activity based on the high-yield spread are significantly superior to those produced using the term spread during the 1990s.

Finally, we break down our measure of real economic activity—real industrial production—into temporary and permanent components, using a variant of an econometric technique developed by Blanchard and Quah (1989). After estimating the long-horizon regressions with output purged of, respectively, its permanent (or "supply") and temporary (or "demand") components, we find that the high-yield spread retains its predictive ability. The results suggest that the high-yield spread works through both, which we interpret as further evidence in support of a financial accelerator mechanism for the United States.

Although the predictive content of the high-yield spread is of general interest, we were led to this investigation because of our finding in previous research that the high-yield spread is an important variable in the modeling of capital flows to emerging markets. It has been conventional wisdom, at least since a well-known paper by Calvo, Leiderman, and Reinhart (1996), that U.S. monetary policy through its influence on short-term interest rates has a significant influence on capital flows to developing economies. Recent events raise a question mark over this
stylized fact, however: while U.S. interest rates have fallen sharply, capital flows have at best remained stable. In previous research, we have proxied the influence of conditions in international capital markets by the U.S. high-yield spread rather than by U.S. interest rates. In particular, we find that a rise in the high-yield spread is associated with reduced supply of capital to emerging markets and the inclusion of this spread eliminates the influence of interest rates (Mody and Taylor, 2002a). A rise in the high-yield spread is also associated with increased regional vulnerability as observed in heightened exchange market pressure (Mody and Taylor, 2002b). Thus, movements in the U.S. high-yield spread clearly have an important bearing on emerging market access to international capital markets. In Mody and Taylor (2002a), we speculated that an international financial accelerator mechanism may be at work: high-yield spreads predict higher default rates and hence slower economic activity, which dampens access to credit, further reducing economic activity, and so on (see, for example, Bernanke and Gertler, 1995). Thus, the question we pose in this paper is whether a rise in high-yield spreads does in fact signal slower growth in the U.S., since establishing such a link would provide a firmer basis for understanding the implications of developments in the “North” for access to capital and vulnerability to currency crises in the “South.”

I. Predicting Real Economic Activity: Theoretical Background

The Term Spread as a Predictor of Real Activity

Although several studies have found the term spread to contain information with respect to future economic activity, the theoretical basis for this relationship has remained unclear, as noted, for example, by Plosser and Rouwenhorst (1994) and Dotsey (1998). Thus, Estrella and Hardouvelis (1991), while documenting the predictive ability of the term spread, also cautioned that the relationship could easily wane.

The slope of the yield curve may be influenced by factors such as expected real interest rates, current and expected inflation, and risk or term premiums. A starting point for the link between the term spread and real economic activity could therefore be the theoretical relationship between real interest rates and macroeconomic activity—for example—through consumption and investment (see Taylor, 1999, for a survey). One can use, for example, a simple optimizing model of consumption to derive a theoretical model of the link between future consumption and the real term structure as follows. Consider a representative agent whose real consumption in period \( t \) is \( C_t \), whose instantaneous utility function is \( U(\cdot) \), and whose subjective rate of time preference is \( p \). If the \( j \)-period real interest rate is \( i^{(j)} \), then, making the usual assumptions such as additive separability of preferences, we can derive from the first-order conditions for the agent’s optimal consumption plan Euler equations of the form:

\[
U'(C_t) = \left(1 + i^{(1)}\right)(1 + p)^{-1} E_t U'(C_{t+1})
\]

\[
U'(C_t) = \left(1 + i^{(2)}\right)(1 + p)^{-2} E_t U'(C_{t+2})
\]

I 375
where $U'(\cdot)$ denotes the first derivative of the utility function and hence marginal utility, and $E_t$ denotes the mathematical expectation operator conditional on information at time $t$. The intuition is standard: if the agent is optimizing, then it is impossible to improve the plan by, say, reducing consumption slightly today [at a cost of $-U'(C_t)]$, investing for $j$ periods at the real interest rate $i_{t+j}$, and increasing consumption in period $j$, yielding an expected gain, in period-$t$ present-value terms, of $[(1+i_{t+j})(1+p)^{-j}E_t(U'(C_{t+j}))]$—the cost just offsets the expected gain. From (1) and (2) we can, however, derive a close approximation:

$$
(i_{t}^{(j_2)} - i_{t}^{(j_1)}) = (1 + p) \left[ \frac{E_t U'(C_{t+j_1})}{E_t U'(C_{t+j_2})} \right] - 1.
$$

Equation (3) thus describes a very simple possibility for bow movements in the real yield curve may affect future economic activity. An increase in the slope of the real term structure will induce optimizing agents to take advantage of the better yield available at longer maturities by reducing consumption in the short term and increasing consumption in the long term. With diminishing marginal utility, a rise in $(i_{t}^{(j_2)} - i_{t}^{(j_1)})$ requires a reduction in $C_{t+j_1}$ and an increase in $C_{t+j_2}$. Insofar as movements in the nominal term spread move with the real term spread, therefore, and insofar as increased consumption demand raises economic activity, this framework predicts that rises in the nominal term spread will indeed be associated with increases in future economic activity.

Note, however, that this analysis is based on a consideration of Euler equations rather than proper reduced forms: these are conditions that must hold at the margin, rather than being reduced-form equations. Moreover, the issue becomes complicated when the move is made from considering the behavior of the representative agent to considering the behavior of the economy in aggregate. In fact, the implication of a large empirical literature on consumption is that the statistical link between real interest rates and aggregate consumption is extremely tenuous (Deaton, 1992; Taylor, 1999), suggesting that it is unlikely that the nominal term spread, by acting as a proxy for the real term spread, is predicting future shifts in consumption demand.

A huge amount of empirical work on aggregate investment also concludes that the statistical link between real interest rates and investment demand is weak (Chirinko, 1993; Taylor, 1999), moreover, suggesting that searching for a theoretical link between the term spread and investment is also likely to be fruitless.

Alternatively, we might explore the avenue that the term spread reflects expected future inflation. However, the long-term interest rates typically used in this connection are quite long horizon---of the order of 10 years or more. Given that the term spread appears to have predictive content for at most a few years, it therefore seems unlikely that this forward-looking element plays a large role in this respect. With respect to term or risk premiums, empirical work has typically found no evidence of strong and statistically stable models (see, for example, Taylor, 1992).

There seem to be two other remaining avenues through which the term spread may predict future real activity. First, insofar as a general monetary easing will be reflected in a fall in short-term interest rates and hence a steepening of the yield
curve, the term spread will be positively correlated with future movements in real activity brought about by the expansionary policy.

The second possibility is that the term spread also reflects movements in the current rate of inflation. If a rise in current inflation leads to a rise in short-term interest rates—and, hence, a flattening of the yield curve—and if inflation and real activity are negatively correlated, a fall in the term spread will predict slower future real activity. The proposition that high inflation is likely to be associated with a weakening of economic activity is supported by recourse to standard economic theory. In a simple aggregate supply–aggregate demand framework, for example, reductions in real output brought about by shifts in aggregate supply will be accompanied by a rise in prices and inflation as the economy moves along the aggregate demand schedule. If we introduce nominal wage inertia and a long-run vertical supply curve into such a framework, then aggregate demand shifts may also lead to negative correlation between inflation and growth since, while the short-run effect of a positive demand shock will be to raise output, prices will initially be largely unaffected because of nominal inertia. If the supply curve is vertical in the long run, moreover, then after a few periods the initial rise in output will begin to decline, just as the rise in prices is beginning to feed through, so that inflation and the change in output will tend to correlate negatively. In Appendix I, we set out a formal macroeconomic model with long-run monetary neutrality and nominal wage inertia induced through wage contracting, in which we show that the overall covariance between inflation and output growth may be negative.3

If, further, we allow for the negative effects on economic activity from inflationary uncertainty and other distortions induced by an environment of high and volatile inflation, then a negative correlation between inflation and growth seems even more likely—at least in a period of high inflation.

In Figure 1 we have graphed the 12-month consumer price index inflation rate and the 12-month percentage growth in real industrial production for the United States over the period 1958M1–2001M12 (see Section II for data sources). Two aspects of the graph are particularly striking. First, the rate of inflation is higher and more volatile during the 1970s and 1980s than it was during either the 1960s or the 1990s. Second, there appears to be stronger evidence of negative correlation between inflation and the rate of growth of industrial production during the 1970s and 1980s. In addition, as argued above, while demand shocks may induce, in certain circumstances, negative correlation between output growth and inflation, supply shocks will unambiguously do so, and this effect appears to be particularly marked following the first and second oil shocks of 1974 and 1979. In contrast, the behavior of inflation during the 1990s appears to have much more in common with that of the 1960s, in that it is generally much lower, less volatile, and apparently less negatively correlated with output growth. In light of our discussion of the possible underlying causes of the link between the term spread and future real activity, therefore, this suggests testing for the strength of the link during the 1960s, as well as during the 1990s, to see if the link is in fact significantly weaker during these two periods.

3Fama (1981) argues that the fact that real stock price returns and inflation tend to be negatively correlated may be because inflation and (current and expected) real output may be negatively correlated.
Figure 1. U.S. Inflation and Output Growth, 1958-2001

- 12-month inflation rate
- 12-month growth in output
The High-Yield Spread as a Predictor of Real Activity

The theoretical underpinning of the high-yield spread as a predictor of real economic activity primarily relates to the theory of the financial accelerator (see, for example, Bemanke and Gertler, 1995; Bernanke, Gertler, and Gilchrist 1999; and the references therein). While the details of these models differ, their central features are reasonably uniform and their key elements may be set out informally as follows.

There is some friction present in the financial market, such as asymmetric information or costs of contract enforcement, which, for a wide class of industrial and commercial businesses, introduces a wedge between the cost of external funds and the opportunity cost of internal funds—the “premium for external funds.” This premium is an endogenous variable that depends inversely on the balance sheet strength of the borrower, since the balance sheet is the key signal through which the creditworthiness of the firm is evaluated. However, balance sheet strength is itself a positive function of aggregate real economic activity, so that borrowers’ financial positions are procyclical and hence movements in the premium for external funds are countercyclical. Thus, as real activity expands, the premium on external funds declines, which, in turn, leads to an amplification of borrower spending, which further accelerates the expansion of real activity. This is the basic mechanism of the financial accelerator.

A problem in testing the theory of the financial accelerator empirically in the past has, however, been the lack of any reliable data on a key central variable in the theory—the premium on external funds. This is because firms that are subject to important financial constraints of this kind have typically relied on commercial bank loans as the chief source of external finance, and time series of relevant bank borrowing rates are not available. Moreover, as Gertler and Lown (1999) point out, even if they were, the fact that bank loans typically contain important nonprice terms would render these series very imperfect and noisy signals of the premium on external funds.

Since the mid-1980s, however, the U.S. market for below-investment-grade debt, sometimes referred to as high-yield bonds or, less euphemistically, “junk bonds,” has developed enormously. Gertler and Lown (1999) note that firms raising funds in the high-yield bond market are likely to be precisely those that face the type of market frictions that the theory of the financial accelerator describes. Moreover, since the opportunity cost of internal funding for firms is likely to be close to the “safe” rate of interest such as that on government or AAA rated debt, the spread between high-yield bonds and government debt or AAA rated debt is likely to be a good indicator of the premium on external finance.

If the theory of the financial accelerator works in practice, therefore, one would expect the high-yield spread to be a countercyclical predictor of future real activity.

II. Data

Monthly data for the United States for the period 1964M1–2001M12 were obtained on real industrial production, the consumer price index, the three-month Treasury bill rate, and the ten-year government bond yield from the International Monetary Fund.
Fund’s *International Financial Statistics* database. A monthly series on the high-yield spread was constructed as follows. First, we obtained data from the Merrill Lynch Global Bond Indices database, an index (in annualized yield terms) of the yields on corporate bonds publicly issued in the U.S. domestic market with a year or more to maturity that were rated BBB3 or lower. We then subtracted the ten-year government bond yield from this to construct the spread. Because the market for below-investment-grade debt only developed during the mid-1980s, a reliable series for the high-yield spread could be constructed only for the period of the 1990s.4

**III. Long-Horizon Regressions**

The dependent variable in the basic long-horizon regressions is the annualized cumulative percentage change in real industrial production:

\[
\nabla_k y_{t+k} = \frac{1200}{k} \left( y_{t+k} - y_t \right),
\]

where \( k \) denotes the forecasting horizon in quarters and \( y_t \) is the logarithm of an index of real industrial production at time \( t \). The \( k \)-period change in the logarithm of industrial output is multiplied by \( (1200/k) \) to ensure that the percentage growth rate is expressed in annualized terms, as the interest rates are. The slope of the nominal yield curve is measured by the difference between the yield on ten-year U.S. government bonds \( (R_t) \) and the three-month U.S. Treasury bill rate \( (r_t) \), while the high-yield spread is measured as the difference between the “junk bond” yield \( (Q_t) \) and the ten-year government bond yield \( (R_t) \).

The basic regression equations are therefore of the form:

\[
\nabla_k y_{t+k} = \alpha_k + \beta_k (R_t - r_t) + \eta_{t+k},
\]

for the term spread regressions, and

\[
\nabla_k y_{t+k} = \gamma_k + \delta_k (Q_t - R_t) + \varepsilon_{t+k},
\]

for the high-yield spread regressions, where \( \eta_{t+k} \) and \( \varepsilon_{t+k} \) are the forecast errors.

\[\text{Note that, since the index of yields on below-investment-grade corporate debt is constructed using a range of maturities, while we subtract the ten-year government bond yield to obtain a measure of the high-yield spread, there is a possibility that our measure of the high-yield spread may contain some term structure effects. We would argue that this is not important for our analysis, however, on the following grounds. First, if term structure effects were important in our high-yield spread measure, then we should expect the predictive content of the two series to be similar whereas—as we show below—they behave quite differently. Second, the correlation between the two series is in fact very low: over the period 1991M1–2001M12 the correlation coefficient is –0.11. We also experimented using high-yield spread measures constructed using government bond yields of different maturities and found that this made little difference to the results—which is not surprising given the very high correlation of these alternative measures of the high-yield spread and the measure used in this paper: using daily data from the end of 1987 until the beginning of 2003, we found that the correlation between the high-yield spreads using ten-year and two-year government bond yields was 0.93, while the correlation between the high-yield spreads using ten-year and three-year government bond yields was 0.95.}\]
As is well known, even under the assumption of rational expectations, the fact that the sampling interval is smaller than the forecasting horizon generates a moving average forecast error of order one less than the number of sampling periods in the forecast horizon, because of common "news" items generating successive forecast errors. Hence, the forecast errors may be assumed to have a moving average representation of order \( k - 1 \). This was allowed for by using an appropriate method-of-moments correction to the estimated covariance matrix (Hansen, 1982). As is also well known, however, tests for the significance of long-horizon parameters may also suffer from considerable size distortion in small samples. Accordingly, we constructed empirical marginal significance levels for the asymptotic \( t \)-ratios using bootstrapping techniques that allow for the small-sample empirical distribution of the test statistics to be constructed under the null hypothesis that the term spread or the high-yield spread does not predict economic activity. A brief description of this bootstrapping algorithm is given in Appendix II.\(^5\)

**Term Spread Regressions**

The results of estimating equation (5) for forecast horizons up to 24 months ahead are given in Tables 1, 2, and 3 for various sample periods.

In Table 1, we report the long-horizon regressions for the 1970s and 1980s—i.e., for the sample period 1970M1—1990M12. These results are consistent with those reported in the literature for similar sample periods: the slope coefficient is strongly significantly different from zero, with \( t \)-ratios in every case of the order of around four and empirical marginal significance levels of zero to two decimal places.

In Table 2, we report results for the same regressions applied to data for the mid- to late 1960s—i.e., 1964M1—1970M12. Although the term spread does have some predictive power for real activity during this period, the slope coefficients are significantly different from zero at the 5 percent level only for horizons ranging from six to nine months. The value of the \( t \)-ratios, even for the significant estimated coefficients, is also much lower than for the 1970s and 1980s, ranging between 2.1 and 2.9.

Table 3 shows the results of estimating the long-horizon term spread regressions for the most recent period, 1991M1—2001M12. The estimated slope coefficients are in every case insignificantly different from zero at the 5 percent level. The goodness of fit of the long-horizon regressions has also fallen dramatically at all horizons, relative to the corresponding levels for the 1970s and 1980s.

Overall, therefore, the strong predictive content of the term spread with respect to future real economic activity seems to be largely confined to the period

---

\(^5\)We also tested for the possibility of simultaneity in the long-horizon regressions, since it is feasible that shocks to current output (which enters the \( k \)-period change) may also affect the term spread or the high-yield spread through, for example, a policy Taylor rule or because of an immediate effect of an output shock on the premium for external funds. To test for this we used Hausman’s (1978) specification test, which tests for differences between OLS and instrumental variable estimators (since only the latter are consistent when there is simultaneity), using three lags of each of the change in output, the term spread, and the high-yield spread as instruments. In no case was there evidence of simultaneity at the 5 percent level.
### Table 1. Term Spread Predictions of Industrial Production Growth, 1971M1–1990M12

<table>
<thead>
<tr>
<th>Forecast Horizon $k$</th>
<th>$\beta_k$</th>
<th>$R^2$</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.038 (4.080)</td>
<td>0.071</td>
<td>10.433</td>
</tr>
<tr>
<td>2</td>
<td>2.255 (3.566)</td>
<td>0.120</td>
<td>8.610</td>
</tr>
<tr>
<td>3</td>
<td>2.392 (3.625)</td>
<td>0.165</td>
<td>7.612</td>
</tr>
<tr>
<td>4</td>
<td>2.449 (3.734)</td>
<td>0.200</td>
<td>6.931</td>
</tr>
<tr>
<td>5</td>
<td>2.438 (3.833)</td>
<td>0.223</td>
<td>6.435</td>
</tr>
<tr>
<td>6</td>
<td>2.409 (3.870)</td>
<td>0.243</td>
<td>6.005</td>
</tr>
<tr>
<td>7</td>
<td>2.395 (3.954)</td>
<td>0.269</td>
<td>5.586</td>
</tr>
<tr>
<td>8</td>
<td>2.391 (3.938)</td>
<td>0.294</td>
<td>5.243</td>
</tr>
<tr>
<td>9</td>
<td>2.370 (3.868)</td>
<td>0.316</td>
<td>4.931</td>
</tr>
<tr>
<td>12</td>
<td>2.349 (3.928)</td>
<td>0.384</td>
<td>4.223</td>
</tr>
<tr>
<td>18</td>
<td>2.102 (3.864)</td>
<td>0.445</td>
<td>3.353</td>
</tr>
<tr>
<td>24</td>
<td>1.632 (4.218)</td>
<td>0.375</td>
<td>3.029</td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated co-variance matrix. $k$ is the forecast horizon in months, $R^2$ denotes the coefficient in determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic $t$-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.
### Table 2. Term Spread Predictions of Industrial Production Growth, 1964M1-1970M12

<table>
<thead>
<tr>
<th>Forecast Horizon $k$</th>
<th>$\beta_k$</th>
<th>$R^2$</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.366</td>
<td>0.003</td>
<td>9.790</td>
</tr>
<tr>
<td></td>
<td>(0.428)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.80]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.798</td>
<td>0.010</td>
<td>7.272</td>
</tr>
<tr>
<td></td>
<td>(0.536)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.71]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.703</td>
<td>0.028</td>
<td>6.263</td>
</tr>
<tr>
<td></td>
<td>(0.709)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.58]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.934</td>
<td>0.068</td>
<td>5.652</td>
</tr>
<tr>
<td></td>
<td>(1.027)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.37]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.059</td>
<td>0.119</td>
<td>5.245</td>
</tr>
<tr>
<td></td>
<td>(1.470)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.17]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.018</td>
<td>0.182</td>
<td>4.812</td>
</tr>
<tr>
<td></td>
<td>(2.127)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.04]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6.898</td>
<td>0.252</td>
<td>4.403</td>
</tr>
<tr>
<td></td>
<td>(2.942)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.01]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7.513</td>
<td>0.310</td>
<td>4.125</td>
</tr>
<tr>
<td></td>
<td>(2.893)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.01]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8.070</td>
<td>0.372</td>
<td>3.821</td>
</tr>
<tr>
<td></td>
<td>(2.414)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.02]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>7.424</td>
<td>0.365</td>
<td>3.593</td>
</tr>
<tr>
<td></td>
<td>(1.996)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.06]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5.303</td>
<td>0.284</td>
<td>2.980</td>
</tr>
<tr>
<td></td>
<td>(1.947)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.06]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>3.400</td>
<td>0.175</td>
<td>2.614</td>
</tr>
<tr>
<td></td>
<td>(1.985)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.06]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated covariance matrix. $k$ is the forecast horizon in months. $R^2$ denotes the coefficient in determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic $t$-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.
Table 3. Term Spread Predictions of Industrial Production Growth, 1991M1-2001M12

<table>
<thead>
<tr>
<th>Forecast Horizon $k$</th>
<th>$\beta_k$</th>
<th>$R^2$</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.880</td>
<td>0.025</td>
<td>6.082</td>
</tr>
<tr>
<td></td>
<td>(1.981)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.06]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.015</td>
<td>0.052</td>
<td>4.812</td>
</tr>
<tr>
<td></td>
<td>(2.031)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.06]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.085</td>
<td>0.069</td>
<td>4.423</td>
</tr>
<tr>
<td></td>
<td>(1.888)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.116</td>
<td>0.079</td>
<td>4.212</td>
</tr>
<tr>
<td></td>
<td>(1.729)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.11]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.129</td>
<td>0.090</td>
<td>3.980</td>
</tr>
<tr>
<td></td>
<td>(1.595)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.32]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.164</td>
<td>0.104</td>
<td>3.789</td>
</tr>
<tr>
<td></td>
<td>(1.520)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.216</td>
<td>0.124</td>
<td>3.610</td>
</tr>
<tr>
<td></td>
<td>(1.497)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.17]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.269</td>
<td>0.145</td>
<td>3.438</td>
</tr>
<tr>
<td></td>
<td>(1.479)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.17]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.275</td>
<td>0.158</td>
<td>3.285</td>
</tr>
<tr>
<td></td>
<td>(1.450)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.174</td>
<td>0.155</td>
<td>3.021</td>
</tr>
<tr>
<td></td>
<td>(1.416)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.30]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.870</td>
<td>0.122</td>
<td>2.484</td>
</tr>
<tr>
<td></td>
<td>(1.538)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.908</td>
<td>0.170</td>
<td>2.148</td>
</tr>
<tr>
<td></td>
<td>(1.775)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.11]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated covariance matrix. $k$ is the forecast horizon in months, $R^2$ denotes the coefficient of determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic $t$-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.
of the 1970s and 1980s. For the 1960s, the relationship appears to be much weakened, while for the 1990s the predictive power of the term spread appears to have virtually disappeared.

High-Yield Spread Regressions

Given that the market for below-investment-grade debt only developed in the United States in the mid-1980s, lack of availability of data on the high-yield spread forced us to consider only the most recent of the three sample periods, i.e., 1991M1–2001M12. The resulting long-horizon regressions are reported in Table 4.

The sign of the estimated slope coefficient is negative in every case: a larger spread predicts a slowdown, exactly as suggested by the theory of the financial accelerator. The predictive content of the high-yield spread is, moreover, quite striking. In every case, the estimated slope coefficient is strongly significantly different from zero, with an empirical marginal significance level of virtually zero in every case, with t-ratios ranging in absolute value from around three to around nine. The goodness of fit, as measured by the coefficient of determination, is in nearly every case very much higher than the corresponding \(R^2\) for the term spread regressions, even during the “heyday” of the term spread during the 1970s and 1980s, the two exceptions being at the 18- and 24-month horizons.

Finally, since the theory of the financial accelerator suggests the possibility of nonlinear interactions between financial variables and real activity, we examined whether “unusual” levels of high-yield spreads convey additional information on real activity. Our proxy for “unusual” levels is a spread that is more than 1.5 standard deviations above the sample mean. To test for this, we adjusted the long-horizon regression (6) to:

\[
\Delta_y y_{t+k} = \gamma_k + \delta_k (Q_t - R_t) + \theta_k I_t (Q_t - R_t) e_{t+k},
\]

where \(I_t\) is a dummy variable that takes the value unity if the high-yield spread is more than 1.5 standard deviations above its mean over the sample period. The results of estimating this equation are shown in Table 5 and reveal that such abnormally high spreads do indeed predict an additional slowing down in growth at horizons of one to three months. It should be noted, however, that most of the abnormally large observations of the high-yield spread fall in the last year or so of the data sample. This was why we did not run the long-horizon regressions for horizons greater than 12 months.

IV. Out-of-Sample Predictions

All of the tests of predictability that we have so far employed are in-sample tests in that they employ all of the data sample to estimate the long-horizon parameters. This contrasts with out-of-sample forecasting methods that typically either use a fixed post-estimation sample of data over which to forecast or else employ recursive or rolling (fixed window) regressions to forecast over a moving post-estimation sample. Typically, in macroeconomics and finance, researchers find it much easier
Table 4. High-Yield Spread Predictions of Industrial Production Growth, 1991M1-2001M12

<table>
<thead>
<tr>
<th>Forecast Horizon k</th>
<th>$\delta_k$</th>
<th>$R^2$</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.064 (-8.744)</td>
<td>0.318</td>
<td>5.088</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-1.985 (-7.120)</td>
<td>0.471</td>
<td>3.597</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1.942 (-5.488)</td>
<td>0.524</td>
<td>3.162</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-1.869 (-4.427)</td>
<td>0.522</td>
<td>4.374</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-1.775 (-3.950)</td>
<td>0.518</td>
<td>2.899</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1.685 (-3.620)</td>
<td>0.499</td>
<td>2.834</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1.603 (-3.368)</td>
<td>0.482</td>
<td>2.776</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-1.544 (-3.212)</td>
<td>0.478</td>
<td>2.687</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.492 (-3.105)</td>
<td>0.473</td>
<td>2.600</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-1.345 (-3.097)</td>
<td>0.436</td>
<td>2.470</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-0.852 (-3.859)</td>
<td>0.285</td>
<td>2.089</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>-0.682 (-7.258)</td>
<td>0.297</td>
<td>1.642</td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated covariance matrix. $k$ is the forecast horizon in months. $R^2$ denotes the coefficient in determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic t-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.
Table 5. Threshold Effects in High Yield Spread Predictions of Industrial Production, 1991M1–2001M12

<table>
<thead>
<tr>
<th>Forecast Horizon $k$</th>
<th>$\delta_k$</th>
<th>$\theta_k$</th>
<th>$R^2$</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.312</td>
<td>-0.608</td>
<td>0.334</td>
<td>5.024</td>
</tr>
<tr>
<td></td>
<td>(-2.635)</td>
<td>(-2.144)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td>[0.03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-1.186</td>
<td>-0.630</td>
<td>0.481</td>
<td>3.480</td>
</tr>
<tr>
<td></td>
<td>(-2.483)</td>
<td>(-2.430)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.01]</td>
<td>[0.01]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1.198</td>
<td>-0.516</td>
<td>0.494</td>
<td>3.043</td>
</tr>
<tr>
<td></td>
<td>(-2.699)</td>
<td>(-2.121)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td>[0.03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-1.251</td>
<td>-0.385</td>
<td>0.468</td>
<td>2.903</td>
</tr>
<tr>
<td></td>
<td>(-2.685)</td>
<td>(-1.657)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td>[0.09]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-1.265</td>
<td>-0.307</td>
<td>0.466</td>
<td>2.720</td>
</tr>
<tr>
<td></td>
<td>(-2.621)</td>
<td>(-1.193)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td>[0.23]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1.312</td>
<td>-0.239</td>
<td>0.466</td>
<td>2.608</td>
</tr>
<tr>
<td></td>
<td>(-2.720)</td>
<td>(-0.751)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td>[0.45]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1.342</td>
<td>-0.172</td>
<td>0.456</td>
<td>2.546</td>
</tr>
<tr>
<td></td>
<td>(-2.671)</td>
<td>(-0.514)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00]</td>
<td>[0.60]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-1.222</td>
<td>-0.237</td>
<td>0.460</td>
<td>2.450</td>
</tr>
<tr>
<td></td>
<td>(-2.150)</td>
<td>(-0.650)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.03]</td>
<td>[0.51]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-1.230</td>
<td>-0.211</td>
<td>0.455</td>
<td>2.380</td>
</tr>
<tr>
<td></td>
<td>(-2.007)</td>
<td>(-0.683)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.04]</td>
<td>[0.49]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-1.217</td>
<td>-0.044</td>
<td>0.364</td>
<td>2.357</td>
</tr>
<tr>
<td></td>
<td>(-1.750)</td>
<td>(-0.223)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.07]</td>
<td>[0.82]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated covariance matrix. $k$ is the forecast horizon in months, $R^2$ denotes the coefficient in determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic t-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.
to reject the null hypothesis of no predictability using in-sample tests than they do with out-of-sample tests. This stylized fact appears to have generated a widely accepted belief that in-sample tests are biased in favor of detecting spurious predictability. In fact, as noted in a recent paper by Inoue and Kilian (2002), “This perception has led to a tendency to discount significant evidence in favor of predictability based on in-sample tests, if this evidence cannot be supported by out-of-sample tests.”

Inoue and Kilian go on, however, to challenge this view, through a formal analysis of the size and power characteristics of in-sample and out-of-sample tests of predictability. They demonstrate, inter alia, that although in-sample and out-of-sample tests are asymptotically equally reliable under standard assumptions, in-sample tests may have much greater power to reject a false null hypothesis of no predictability in small samples. Indeed, these authors show that out-of-sample tests may have less than 50 percent of the power of in-sample tests, thus providing an alternative explanation of the observed tendency of in-sample tests to reject the null of no predictability more often than out-of-sample tests.

It is clear, however, that the striking findings of Inoue and Kilian should not be taken as an excuse not to conduct out-of-sample tests but, rather, should be taken into account when interpreting the findings of in-sample and out-of-sample tests of forecast failure. In particular, and especially in cases such as the present analysis where there is not an issue of data mining, the low power of out-of-sample tests means that their failure to reject a null hypothesis of no predictability (or of equal predictive ability of two alternative methods) does not necessarily imply that in-sample evidence of predictability is spurious. On the other hand, if the null hypothesis of no predictability or of equal forecasting ability of methods can be rejected at standard significance levels using out-of-sample tests, then, given their low power, this should be taken as very strong corroborating evidence of the in-sample findings.

Accordingly, we proceeded to construct out-of-sample tests to compare the ability of movements in the high-yield spread to predict economic activity over the 1990s with that of movements in the term spread.

We did this in two ways. In our first investigation of relative out-of-sample predictive ability, we used a fixed post-estimation sample of three years, 1999M1–2001M12, to construct a “portmanteau” test of the equality of the predictive ability of the term spread and the high-yield spread. We used data over the sample period 1991M1–1998M12 to estimate the long-horizon equations (5) and (6) for \( k = 1, 2, \ldots, 36 \). We then obtained two sets of 36 forecast errors over the post-estimation sample, based alternately on the high-yield spread or the term spread in 1998M12, as

\[
\varepsilon^{(1)}_k = \nabla y_{1998M12+k} - \tilde{\alpha}_k - \tilde{\beta}_k (R_{1998M12} - r_{1998M12}), \quad k = 1, 2, \ldots, 36
\]

for the term spread, and

\[\vdots\]

That is, under the null of no predictability, and provided that no data mining has taken place, and in the absence of unmodeled structural change. The authors also show that both in-sample and out-of-sample tests are susceptible to size distortions arising from data mining.
for the high-yield spread, where hats over the parameters denote that they were estimated using data for the period 1991M1–1998M12. We then used these errors to construct measures of the mean square error and mean absolute error over the 36 months, and in turn used these to construct a test for equal predictive ability based on the Diebold-Mariano (1995) test statistic for equality of forecast accuracy. The Diebold-Mariano statistic is defined as

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi f(0)}{N}}}$$

where $\bar{d}$ is an average over $N$ forecast periods of a general loss differential function $d_k$ such as the difference in squared forecast errors ($d_k = [\hat{\tau}^{(1)}_k - \hat{\tau}^{(2)}_k]^2$) or, in absolute errors ($d_k = |\hat{\tau}^{(1)}_k - \hat{\tau}^{(2)}_k|$), i.e.,

$$\bar{d} = \frac{1}{N} \sum_{k=1}^{N} d_k$$

for $N = 36$ in the present application, and $f(0)$ is a consistent estimate of the spectral density of the loss differential function at frequency zero. Under certain regularity conditions, $DM$ will be distributed as standard normal under the null hypothesis of equal forecast accuracy.

Our second set of tests involved assessing the relative predictive ability of the high-yield spread and the term spread by comparing the forecast accuracy of the multi-step-ahead predictions of the long-horizon models (5) and (6) using recursive estimation over the period 1999M1–2001M12 (with initial estimates using data for 1991M1–1998M12) for the same values of $k$ as used in our long-horizon in-sample exercises, i.e., $k = 1–12, 18, 24$. For the long-horizon regression model of horizon $k$, this generates a sequence of 36 $k$-step-ahead forecast errors of the form

$$\tau^{(k,1)}_\tau = N_k y_{1998M12 + \tau} - \hat{\alpha}(\tau - 1)_k - \hat{\beta}(\tau - 1)_k (R_{1998M12 + \tau - k} - R_{1998M12 + \tau - k})$$

for the term spread, and

$$\tau^{(k,2)}_\tau = N_k y_{1998M12 + \tau} - \hat{\gamma}(\tau - 1)_k - \hat{\delta}(\tau - 1)_k (Q_{1998M12 + \tau - k} - Q_{1998M12 + \tau - k})$$

for the high-yield spread.
for the high-yield spread, where \( \tilde{\lambda}(\tau-1)_k \) denotes the estimated value of the long-horizon parameter \( \lambda_k \) using data for the period 1991M1 to 1998M12 + \( \tau - 1 \), for \( \lambda = \alpha, \beta, \gamma, \delta \). We then again used the Diebold-Mariano statistic (10) to test for equality of the forecast accuracy of the models for each value of \( k \), based on the difference in squared errors or in absolute errors \( d_{\text{ts}} = |r^{(k.1)}_\tau|^2 - |r^{(k.2)}_\tau|^2 \) or \( d_{\text{ts}} = |r^{(k.1)}_\tau| - |r^{(k.2)}_\tau| \) and

\[
\bar{d} = \frac{1}{N} \sum_{t=1}^{N} d_t^k
\]

for the \( k \)-horizon model one-step-ahead forecasts).

The results of the out-of-sample portmanteau test of equality of forecast accuracy of the high-yield spread and term spread models are given in Table 6 and show that the high-yield spread regressions easily and significantly dominate the term spread regressions in terms of out-of-sample forecast accuracy. This result is echoed, moreover, in the \( k \)-step-ahead forecast results for each of the individual long-horizon regressions, as shown in Table 7: the high-yield spread model forecasts are in every case superior in terms of either mean square error or mean absolute error, and strongly significantly so in every case except for the \( k = 1 \) and \( k = 2 \) regression models, based on mean square error.

Given our earlier remarks concerning the relatively low power of out-of-sample forecast tests, this is indeed a strong vindication of the superiority of the high-yield spread as a predictor of economic activity over the 1990s relative to the term spread.

V. Supply and Demand Innovations in Real Output

Given the apparent importance of the high-yield spread as a predictor of real economic activity during the last decade or so, we carried out some further investigations as to whether the spread is able to predict the cumulative growth in output when it is stripped of, alternately, its demand-side and supply-side components or, to be precise, its temporary and permanent components.

Permanent and temporary output movements may be variously interpreted according to the underlying theoretical framework employed. In the traditional aggregate demand-aggregate supply (ADAS) model with a long-run vertical supply curve, for example, aggregate demand disturbances result in a temporary rise

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High-Yield Spread</td>
<td>Term Spread</td>
<td>Diebold-Mariano</td>
</tr>
<tr>
<td>Mean square error</td>
<td>4.06</td>
<td>18.65</td>
<td>-4.94</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>1.94</td>
<td>4.25</td>
<td>-6.21</td>
</tr>
</tbody>
</table>

Notes: See Section IV on the construction of these statistics. Figures in parentheses are marginal significance levels to two decimal places.

©International Monetary Fund. Not for Redistribution
Table 7. Out-of-Sample Tests of Equality of Forecast Accuracy over the Period 1999M1-2001M12, k-Step-Ahead Forecasts

<table>
<thead>
<tr>
<th>Long-Horizon Regression Model, k</th>
<th>Mean Square Error from Forecasts Based on High-Yield Spread</th>
<th>Mean Square Error from Forecasts Based on Term Spread</th>
<th>Diebold-Mariano Statistic Based on Mean Square Error</th>
<th>Mean Absolute Error from Forecasts Based on High-Yield Spread</th>
<th>Mean Absolute Error from Forecasts Based on Term Spread</th>
<th>Diebold-Mariano Statistic Based on Mean Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.95</td>
<td>36.47</td>
<td>-1.20</td>
<td>3.72</td>
<td>4.63</td>
<td>-2.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
</tr>
<tr>
<td>2</td>
<td>13.70</td>
<td>18.69</td>
<td>-1.59</td>
<td>2.59</td>
<td>3.69</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.01)</td>
</tr>
<tr>
<td>3</td>
<td>8.96</td>
<td>13.56</td>
<td>-2.02</td>
<td>2.28</td>
<td>3.19</td>
<td>-2.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
</tr>
<tr>
<td>4</td>
<td>7.02</td>
<td>11.66</td>
<td>-2.52</td>
<td>2.15</td>
<td>3.04</td>
<td>-2.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
</tr>
<tr>
<td>5</td>
<td>10.31</td>
<td>10.31</td>
<td>-2.96</td>
<td>1.89</td>
<td>2.91</td>
<td>-3.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>6</td>
<td>5.33</td>
<td>10.30</td>
<td>-3.39</td>
<td>1.87</td>
<td>2.92</td>
<td>-3.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>7</td>
<td>4.86</td>
<td>10.04</td>
<td>-3.84</td>
<td>1.84</td>
<td>2.91</td>
<td>-3.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>8</td>
<td>4.51</td>
<td>10.40</td>
<td>-4.67</td>
<td>1.79</td>
<td>2.96</td>
<td>-4.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>9</td>
<td>3.98</td>
<td>10.16</td>
<td>-5.30</td>
<td>1.69</td>
<td>2.95</td>
<td>-5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>12</td>
<td>2.84</td>
<td>10.54</td>
<td>-6.70</td>
<td>1.41</td>
<td>3.01</td>
<td>-8.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>18</td>
<td>1.84</td>
<td>10.73</td>
<td>-9.55</td>
<td>1.06</td>
<td>3.07</td>
<td>-20.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>24</td>
<td>1.98</td>
<td>9.49</td>
<td>-9.20</td>
<td>1.15</td>
<td>2.89</td>
<td>-23.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
</tbody>
</table>

Notes: See Section IV on the construction of these statistics. Figures in parentheses are marginal significance levels to two decimal places.
in output, while aggregate supply disturbances permanently affect the level of aggregate output.

Blanchard and Quah (1989) use an ADAS framework in their analysis and associate aggregate supply shocks with permanent shocks and aggregate demand shocks with temporary shocks. In this paper we shall follow their taxonomy. While it is possible that demand disturbances may have permanent effects on the real side of the economy, we concur with Blanchard and Quah that shocks having a permanent effect on output are likely to be due mostly, if not wholly, to supply-side factors, while those having only a temporary effect are likely to be due mostly, if not wholly, to demand-side factors. If the permanent long-run effects of demand disturbances are small relative to the long-run permanent effects of supply disturbances, then the Blanchard-Quah taxonomy is a useful organizing principle for empirical purposes. Readers rejecting this taxonomy, however, may simply reinterpret our analysis as investigating whether nominal spreads affect the permanent or the temporary components of real output movements.

Given this taxonomy of permanent and temporary shocks to output, supply and demand shocks to real economic activity can be identified by imposing appropriate restrictions on the Wold representation of time series for real and nominal macroeconomic variables. In particular, consider the Wold representation for annualized percentage changes in the logarithm of output and the logarithm of prices:

\[
\begin{bmatrix}
\nabla_1 y_t \\
\nabla_1 p_t 
\end{bmatrix} = \sum_{j=1}^{\infty} L^j \begin{bmatrix}
\phi_{11j} & \phi_{12j} \\
\phi_{21j} & \phi_{22j}
\end{bmatrix} \begin{bmatrix}
\zeta_{1t} \\
\zeta_{2t}
\end{bmatrix}
\]

(13)

where the \(\phi_{imj}\) are the parameters of the multivariate moving average representation and \(\zeta_{1t}\) and \(\zeta_{2t}\) are white-noise innovations. We can identify \(\zeta_{1t}\) and \(\zeta_{2t}\) as demand and supply innovations in the following way. Write \(\zeta_t = (\zeta_{1t}, \zeta_{2t})^\prime\), and denote the bivariate vector of innovations recovered from the vector autoregressive representation for \((\nabla_1 y_t, \nabla_1 p_t)^\prime\) as \(\nu_t\). Since the vector autoregressive (VAR) representation is simply an inversion of the Wold representation (13), \(\nu_t\) will in general be a linear function of \(\zeta_t\), \(\nu_t = A \zeta_t\), say, where \(A\) is a 2 x 2 matrix of constants. To recover the underlying demand and supply innovations from the VAR residuals then requires that the four elements of \(A\) be identified, which requires four identifying restrictions. Three restrictions can be obtained by normalizing the variances of \(\zeta_{1t}\) and \(\zeta_{2t}\) to unity and setting their covariances to zero. (See Blanchard and Quah, 1989, for a defense of these restrictions.)

The fourth, crucial identifying restriction, which effectively identifies \(\zeta_{1t}\), as the demand innovation (or temporary output innovation), is the requirement that \(\zeta_{1t}\) have no long-run effect on the (log-) level of real output, although it may affect the long-run price level. The latter restriction on the Wold representation (13) may be written:

\[
\sum_{j=1}^{\infty} \phi_{11j} = 0.
\]

(14)
THE HIGH-YIELD SPREAD AS A PREDICTOR OF REAL ECONOMIC ACTIVITY

These four restrictions are then sufficient to recover the underlying temporary and permanent innovations to output, which, as we discussed above, may be interpreted as underlying demand and supply innovations, respectively.9

Having identified the supply and demand innovations, we can then partition the moving average representation for real industrial output into counterfactual series, corresponding to the path that would have obtained in the absence of demand innovations and the path that would have obtained in the absence of supply innovations over the estimation period. We can then utilize these counterfactual series in tests of the predictive power of the high-yield spread.

We applied this method to the monthly series in the logarithm of industrial production and the consumer price index for the whole sample period, 1964M1-2001M12. Preliminary unit root (Augmented Dickey-Fuller) tests on the data (not reported) showed the change in the logarithm of real industrial output and the change in the logarithm of the consumer price index to be stationary processes. There was also no evidence of cointegration between industrial production and prices. This implies that output growth and inflation can be modeled as a bivariate moving average representation, which can be inverted to a pure autoregression not involving error correction terms. We therefore proceeded to estimate a vector autoregressive representation for the vector time series \( (y_t, p_t)' \). The order of the VAR was chosen by sequentially excluding the highest lags of both series, starting from a twelfth-order VAR and testing the exclusion restrictions on the system using a likelihood ratio test. This process was stopped when the exclusion restrictions were jointly significant at the 5 percent level. This led to a choice of lag depth of six. The residuals from the estimated equations were judged to be approximately white noise, using either individual Ljung-Box statistics or Hosking’s multivariate portmanteau statistic. In fact, this choice of lag depth coincided with the lag depth chosen by minimizing the Akaike Information Criterion.

In Figure 2 we have graphed the impulse response functions for the log-levels of output and prices in response to the identified supply and demand innovations. By construction, the long-run impact of demand shocks on real output is zero, but the shape of each of the impulse response functions in each case accords with simple economic priors in that a positive demand shock raises both output (in the short run) and prices (in both the short run and the long run), while a positive supply shock raises output and depresses prices (in both the short run and the long run).10

VI. Counterfactual Analysis

We then used the Blanchard-Quah estimation results to break down the series for U.S. real industrial production into counterfactual series, corresponding to the


10Taylor (2003) notes that, in general, there will be multiple solutions to recursive restrictions of the type suggested by Blanchard and Quah (1989) and that informal or qualitative identifying restrictions of this kind are necessary in order to achieve full identification.
Figure 2. Impulse Responses

Response of Output to Demand Shock

Response of Output to Supply Shock

Response of Prices to Demand Shock

Response of Prices to Supply Shock

path that would have obtained in the absence of demand innovations in the moving average representation and the path that would have obtained in the absence of supply innovations. Effectively, this involves using the estimated VAR to recover the moving average representation (13) and then calculating a counterfactual series for $y$, by alternately holding the identified supply and demand shocks constant at zero over the sample period. We then used this series in estimates of the long-horizon regression using the high-yield spread, the results of which are given in Tables 8 and 9.

The results are interesting and supportive of the financial accelerator theory. We should expect the financial accelerator process to operate through both the supply and the demand sides of the economy. In particular, a positive supply, or productivity, shock that leads to a permanent change in output will increase the collateral value of the future stream of output. The reduced premium for external
funds will, therefore, be associated with future output growth. The results strongly confirm this relationship. When the industrial output series is purged of "demand disturbances," leaving the supply or permanent shocks in, the predictive ability of the high-yield spread remains strong at all horizons (Table 8). Temporary demand shocks can also generate an accelerator. Indeed, following a permanent shock, induced investment demand can generate additional cyclical effects. Interestingly, the high-yield spread does indeed significantly predict the "demand-driven" component of industrial production (Table 9).

VII. Conclusion

Why did the term spread become a much weaker predictor of economic activity in the 1990s? Gertler and Lown (1999) suggest that changes in U.S. monetary policy may have had something to do with this. In particular, a more robust defense of inflation starting in the mid-to late 1980s may have changed private sector expectations with respect to future inflation. While the shift in monetary policy may have been influential, it is relevant that the predictive ability of the term spread was also weaker prior to 1970.

The period during which the term spread was informative with respect to real activity—the 1970s and 1980s—was the period of the two oil shocks and was characterized by high inflation (and possibly, therefore, greater inflation uncertainty) and volatile growth. We should expect in such a period that the negative covariance between current inflation and real activity would be most pronounced. A rise in current inflation, associated with a rise in short-term rates, would lead to a flattening of the term spread and lower real activity. However, in periods when inflation is low—and more predictable—these effects may be weaker.

In contrast, the financial accelerator creates a more robust foundation for the high-yield spread as a predictor of future real macroeconomic activity. That relationship is based on financial frictions that amplify the business cycle. While such frictions, in particular asymmetric information, may decline over time, that seems unlikely to occur in the immediate future. The robustness of this relationship is also suggested by the finding that the high-yield spread captures both the supply-side (or permanent) shocks and the demand-side (temporary) shocks.
Table 8. High-Yield Spread Predictions of Industrial Production Growth with Industrial Production Purged of Demand-Side Disturbances, 1991M1-2001M12

<table>
<thead>
<tr>
<th>Forecast Horizon $k$</th>
<th>$\delta_k$ (t-stat)</th>
<th>$R^2$ (percent)</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.392 (-2.366)</td>
<td>0.070</td>
<td>2.173</td>
</tr>
<tr>
<td>2</td>
<td>-0.474 (-1.923)</td>
<td>0.126</td>
<td>1.956</td>
</tr>
<tr>
<td>3</td>
<td>-0.571 (-1.799)</td>
<td>0.182</td>
<td>1.935</td>
</tr>
<tr>
<td>4</td>
<td>-0.637 (-1.851)</td>
<td>0.222</td>
<td>1.924</td>
</tr>
<tr>
<td>5</td>
<td>-0.687 (-2.026)</td>
<td>0.254</td>
<td>1.864</td>
</tr>
<tr>
<td>6</td>
<td>-0.714 (-2.142)</td>
<td>0.271</td>
<td>1.811</td>
</tr>
<tr>
<td>7</td>
<td>-0.721 (-2.213)</td>
<td>0.277</td>
<td>1.782</td>
</tr>
<tr>
<td>8</td>
<td>-0.730 (-2.293)</td>
<td>0.285</td>
<td>1.767</td>
</tr>
<tr>
<td>9</td>
<td>-0.731 (-2.400)</td>
<td>0.294</td>
<td>1.732</td>
</tr>
<tr>
<td>12</td>
<td>-0.730 (-2.700)</td>
<td>0.311</td>
<td>1.672</td>
</tr>
<tr>
<td>18</td>
<td>-0.679 (-3.966)</td>
<td>0.314</td>
<td>1.567</td>
</tr>
<tr>
<td>24</td>
<td>-0.592 (-2.330)</td>
<td>0.229</td>
<td>1.700</td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated covariance matrix. The series for industrial production has been purged of demand-side disturbances using the Blanchard-Quah method described in the text. $k$ is the forecast horizon in months, $R^2$ denotes the coefficient of determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic t-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.

<table>
<thead>
<tr>
<th>Forecast Horizon $k$</th>
<th>$\hat{\delta}_k$</th>
<th>$R^2$</th>
<th>s.e. (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.785 (-6.189)</td>
<td>0.212</td>
<td>5.255</td>
</tr>
<tr>
<td>2</td>
<td>-1.684 (-5.009)</td>
<td>0.337</td>
<td>3.701</td>
</tr>
<tr>
<td>3</td>
<td>-1.622 (-4.141)</td>
<td>0.398</td>
<td>3.186</td>
</tr>
<tr>
<td>4</td>
<td>-1.534 (-3.620)</td>
<td>0.411</td>
<td>2.960</td>
</tr>
<tr>
<td>5</td>
<td>-1.407 (-3.282)</td>
<td>0.387</td>
<td>2.800</td>
</tr>
<tr>
<td>6</td>
<td>-1.292 (-3.155)</td>
<td>0.356</td>
<td>2.685</td>
</tr>
<tr>
<td>7</td>
<td>-1.168 (-3.167)</td>
<td>0.323</td>
<td>2.589</td>
</tr>
<tr>
<td>8</td>
<td>-1.078 (-3.171)</td>
<td>0.310</td>
<td>2.461</td>
</tr>
<tr>
<td>9</td>
<td>-0.998 (-3.183)</td>
<td>0.302</td>
<td>2.320</td>
</tr>
<tr>
<td>12</td>
<td>-0.858 (-2.994)</td>
<td>0.284</td>
<td>2.094</td>
</tr>
<tr>
<td>18</td>
<td>-0.635 (-2.758)</td>
<td>0.221</td>
<td>1.862</td>
</tr>
<tr>
<td>24</td>
<td>-0.482 (-1.904)</td>
<td>0.142</td>
<td>1.856</td>
</tr>
</tbody>
</table>

Notes: Estimation is by ordinary least squares, with a method-of-moments correction to the estimated covariance matrix. The series for industrial production has been purged of supply-side disturbances using the Blanchard-Quah method described in the text. $k$ is the forecast horizon in months, $R^2$ denotes the coefficient in determination, and s.e. denotes the standard error of the regression. Figures in parentheses below coefficient estimates are asymptotic $t$-ratios and those in square brackets are the bootstrapped empirical marginal significance levels to two decimal places. An intercept term was also included in the regressions.
APPENDIX I

The Negative Covariation of Inflation and Output Growth in a Model with Nominal Wage Inertia and Long-Run Monetary Neutrality

Consider the following simple, log-linear macroeconomic model, which displays long-run monetary neutrality and short-run nominal wage inertia induced through a wage formation equation in which wages are set in a two-period overlapping contracts framework:

\[ y_t = m_t - p_t \]  \hspace{1cm} (A1)

\[ y_t = n_t + \theta_t \]  \hspace{1cm} (A2)

\[ p_t = w_t - \theta_t \]  \hspace{1cm} (A3)

\[ w_t = \omega \left[ E_{t-2} n_t = n^* \right] \]  \hspace{1cm} (A4)

Equation (A1) represents the aggregate demand side of the economy as a function of real balances. The production function (A2) relates output to the level of employment, \( n_t \), and productivity, \( \theta_t \). The price level is shown in (A3) to be a function of the nominal wage and productivity, while in (A4) the nominal wage contract is set two periods in advance at the level expected to generate full-employment level, \( n^* \). The model is closed by assuming that money and productivity are determined by the evolution of demand and supply shocks, \( e_{dt} \) and \( e_{st} \), respectively, as follows:

\[ m_t = m_{t-1} + e_{dt} \]  \hspace{1cm} (A5)

\[ \theta_t = \theta_{t-1} + e_{st} \]  \hspace{1cm} (A6)

Assume that the covariance of \( e_{dt} \) and \( e_{st} \) is zero and that the supply and demand shocks have constant variances.

Solving the model for inflation and output growth as a function of the exogenous demand and supply disturbances yields:

\[ \nabla_1 p_t = e_{dt-2} - e_{st} \]  \hspace{1cm} (A7)

\[ \nabla_1 y_t = e_{dt} - e_{dt-2} + e_{st} \]  \hspace{1cm} (A8)

where \( \nabla_1 \) denotes the first-difference operator. Note that only supply shocks have a permanent effect on output, while both supply and demand shocks can affect long-run prices. Using (A7) and (A8), the covariance of growth and inflation is easily seen to be negative:

\[ \text{Cov}(\nabla_1 p_t, \Delta_1 y_t) = -[\text{Var}(e_{dt}) + \text{Var}(e_{st})] < 0. \]  \hspace{1cm} (A9)
APPENDIX II

A Bootstrap Algorithm for the Long-Horizon Tests

We discuss only the algorithm for the standard long-horizon regression involving the term spread. The algorithms for the other long-horizon regressions are identical except for a change of variable or the addition of an extra regressor. The algorithm consists of four steps:

1. Estimate the long-horizon regression

\[ \nabla_k y_{t+k} = \alpha_k + \beta_k (R_t - r_t) + \eta_{t+k} \]

for \( k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 18, 24 \) by OLS, and in each case construct the test statistic for the null hypothesis \( H_0: \beta_k = 0 \) as the ratio of the estimated value of \( \beta_k \) to its estimated standard error (the latter constructed using a method-of-moments correction to allow for moving average serial correlation up to order \( k - 1 \)); call this test statistic \( \tau \).

2. Since \( \Delta y_t = y_t - y_{t-1} \) and \( (R_t - r_t) \) are each assumed to be stationary, \( I(0) \) processes, by Wold's theorem they will have an invertible joint vector moving average representation, which can be approximated by a vector autoregressive representation of sufficiently high order, say \( p \):

\[
\Delta y_t = \phi_t + \sum_{i=1}^p \kappa_i \Delta y_{t-i} + \sum_{i=1}^p \lambda_i (R_{t-i} - r_{t-i}) + u_t,
\]

\[
\Delta (R_t - r_t) = \phi_t + \sum_{i=1}^p \mu_i \Delta y_{t-i} + \sum_{i=1}^p \nu_i (R_{t-i} - r_{t-i}) + u_t.
\]

Under the null hypothesis that economic activity is unpredictable, however, we have \( \kappa_i = \lambda_i = 0 \) for all \( i \). This VAR model is therefore estimated by generalized least squares with these exclusion restrictions imposed and with the lag order \( p \) chosen using the Akaike information criterion.

3. Based on the fitted model from step 2, a sequence of pseudo observations \( \{ \Delta y_t^* \} \) and \( \{(R_t - r_t)^* \} \) is generated of the same length as the original data series from realizations of the bootstrap data generating process:

\[
\Delta y_t^* = \tilde{\phi}_t + u_t^*,
\]

\[
\Delta (R_t - r_t)^* = \tilde{\phi}_t + \sum_{i=1}^p \tilde{\mu}_i \Delta y_{t-i}^* + \sum_{i=1}^p \tilde{\nu}_i (R_{t-i} - r_{t-i})^* + u_t^*.
\]

where the coefficients are the corresponding estimated parameters from step 2, and the vector pseudo innovation \( (u_t^* u_t^*)^* \) is drawn with replacement from the set of residuals generated in step 2. The process is initialized by setting \( \Delta y_{t-1}^* = (R_{t-1} - r_{t-1})^* = 0 \) for \( t = p - 1, \ldots, 1 \) and the first 500 transients are discarded. This step is repeated 2,000 times.

4. For each of the 2,000 bootstrap replications in step 3, construct the pseudo annualized cumulative percentage change in real industrial production as:

\[
\nabla_k y_{t+k}^* = \frac{1200}{k} \sum_{j=1}^k \Delta y_{t+j}^*
\]

and estimate the long-horizon regression

Note that there is no issue of cointegration in the present application, which would involve using a vector equilibrium correction model rather than a VAR as the data-generating process—see Kilian (2000).
\[ \nabla_k y_{t+k}^* = \alpha_k^* + \beta_k^* (R_t - \gamma) + \eta_{t+k}^*. \]

for \( k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 18, 24 \). For each replication construct the test statistic for the null hypothesis \( H_0: \beta_k^* = 0 \) as the ratio of the estimated value of \( \beta_k^* \) to its estimated standard error (the latter again constructed using a method-of-moments correction to allow for moving average serial correlation up to order \( k - 1 \)); call the test statistic for the \( i \)-th replication \( \tau_{t_i}^* \).

5. Use the empirical distribution of the 2,000 replications of the sequence of bootstrap test statistic \( \{ \tau_{t_i}^* \} \) to determine the marginal significance level of \( \tau \).

REFERENCES


The Art of Making Everybody Happy: How to Prevent a Secession

MICHEL LE BRETON and SHLOMO WEBER

In this paper we examine compensation schemes that prevent a threat of secession by any of a country's regions. We prove that, under quite general assumptions on the distribution of citizens' preferences, there exist transfer schemes that are secession-proof. Moreover, we show that these compensation schemes entail a degree of partial equalization among regions: the gap between advantaged regions and disadvantaged regions has to be reduced but it should never be completely eliminated. We demonstrate that in the case of a uniform distribution of the nation's citizens, the secession-proof conditions generate the 50 percent compensation rule for disadvantaged regions. [JEL D70, H20, H73]

The world political map has undergone dramatic changes since World War II. The number of independent countries in the world almost tripled over the second half of the last century, rising from a mere 74 in 1946 to 193 today; 45 percent of countries that exist today have a population under five million people. The abolition of colonial rule in Africa in the 1960s created 25 new countries. The last decade brought the next major wave of border changes, highlighted by the breakups of the former Soviet Union, Yugoslavia, and Czechoslovakia and the reunification of Germany. Quebec's recent secession bid was defeated by a majority of less than

*Michel Le Breton is Professor of Economics at the Université de Toulouse I, GREMAQ and IDEI, Toulouse, France. Shlomo Weber is Robert H. and Nancy Dedman Trustee Professor of Economics in the Department of Economics, Southern Methodist University, Dallas, Texas, USA, and Fellow, the Center for Operations Research and Economics (CORE), the Catholic University of Louvain-la-Neuve, Belgium. A previous version of this paper was written while the second author was visiting the Department of Fiscal Affairs at the IMF. We wish to thank Ehsisham Ahmad, Francis Bloch, Youngsub Chun, Jacques Drèze, Michel Goemans, Philippe Jéhieh, Michael Keen, Ozgur Kibris, Luc Leruth, Lionel McKenzie, Jean-Charles Rochet, Karl Shell, Partho Shomme, Martin Skutella, Enrico Spolaore, Yuval Weber, an anonymous referee, and an editor of this journal for their valuable comments and suggestions. We are grateful to seminar participants at CORE, IMF, London School of Economics, and the Universities of Brussels, Copenhagen, Rochester, Stockholm, Toulouse, Brown, Georgetown, Kyoto, New York, Peking, and Seoul for their remarks. We would also like to thank Nicholas Gaspard and Monica Sarratt for their help in preparing the manuscript.
percent. One can also point to Belgium’s process of “defederalization” and Scotland’s and Wales’ paths to devolution. Various types of separatist movements are active in Africa (Nigeria, Senegal, Angola, Côte d’Ivoire), Asia (China, India, Indonesia), and Europe (Denmark, Spain, France, Italy, Russia, Yugoslavia), and conflicts over fiscal redistribution, regional power, and autonomy mushroom all over the globe.

The purpose of our paper is to examine distributive policies of the central government that would prevent a threat of secession by any of a country’s regions. The government’s objective is to design a transfer policy that would render the advantages of secession for every region inferior to the benefits of remaining within an integrated country. The analysis of advantages and disadvantages of a secession points out the trade-off between economies of scale in big countries and the costs of heterogeneity in large populations. As Barro (1991) puts it, “We can think of a country’s optimal size as emerging from a trade-off: A large country can spread the cost of public goods over many taxpayers, but a large country is also likely to have a diverse population that is difficult for the central government to satisfy.” Larger political jurisdictions bring about several benefits: the per capita cost of producing public goods declines with the population size of the country; larger countries rely more heavily on more efficient taxes and enjoy economies of scale in the utilization of computer hardware and software systems in their tax collection; the size of a country’s potential market is affected by the size of the jurisdiction in a world with barriers to trade; larger countries are better equipped to absorb uninsurable shocks in different regions; influence and security considerations may also matter. On the other hand being small has its advantages, as relative ethnic, religious, and cultural homogeneity is positively correlated with a country’s institutional efficiency; small countries are usually more open to trade and adjust better to dealing with technological changes in the world markets; and interest groups and unproductive activities play a lesser role in smaller countries.

Since the focus of our paper is the analysis of secessions, we consider a model with one nation and do not address the more general issue of border redrawing. It might be possible that in some situations a secession would necessitate redrawing

---

1 A large population of taxpayers can share the cost of public goods such as roads, a telephone network, defense, civil servants, and education. Alesina and Wacziarg (1998) show that small countries tend to have bigger governments, and bigger government consumption, as a share of GDP. Smaller countries also face substantial costs of maintaining their distinctive language and culture. For example, the economic cost of Iceland’s language is about 3 percent of the country GNP (Economist, 1998).


4 See Friedman (1977), Casella (1992), and Casella and Feinstein (2002).

5 See Persson and Tabellini (1996a,b).

6 See Alesina and Spolaore (1996). In many countries a majority of citizens do not particularly value their country’s political and military might, but in some other countries, particularly China, France, Russia, India, and Pakistan, the citizens do care about their country’s standing and influence in the world. As evidence of this phenomenon, Easterly and Rebello (1993) confirm that large countries spend relatively more on their defense.

7 See Mauro (1995) for an analysis of countries’ language and ethnic diversity.

8 See Alesina, Spolaore, and Wacziarg (2000).

9 See Barro (1996).
of borders in several countries (e.g., the idea of “Greater Albania” consisting of Albania itself, Kosovo, and parts of the former Yugoslavian province of Macedonia; creation of the Basque country from the Basque areas in France and Spain; and the formation of Kurdistan from the Kurdish enclaves in Turkey, Iran, and Iraq). It is quite obvious that the probability of any of those scenarios coming true anytime soon is not very high; and indeed, most secessionist movements and tensions are internal to nations.

In our framework, a national policy consists of two components: a choice of public policy, which can be interpreted as composition of public expenditure, location of the central government, tax rates, immigration quotas, or any other issue of national interest; and a citizen-specific cost allocation designed to cover fixed government costs. This model allows us to examine the trade-off between heterogeneity of citizens’ preferences and increasing returns to scale in larger countries. The heterogeneity of preferences is described by the density function representing a number of citizens of each type, whereas the advantages of size are captured by sharing the fixed government costs among a larger number of citizens.

To address the issue of secession we first examine whether, given heterogeneity of citizens’ preference, it is desirable and socially efficient to maintain the unified country. We consider a notion of efficiency of cooperation among different regions of a country that is the case when all regions are better off under a single national government. The efficiency of cooperation does not necessarily imply that the gains from cooperation can be allocated in such a way that no region can ensure all its citizens a higher payoff than guaranteed by the central government. If no such allocation is feasible, some regions may become secession-prone. We investigate the existence of cost allocations that do not create secession-prone regions. If such an allocation exists, we call it secession-proof and the cooperation would be stable and sustainable. The requirement of stability is, in principle, stronger than that of efficiency. However, our first result suggests that, under quite general assumptions, high government costs eliminate the gap between efficiency and stability, thus reconciling these two notions. Then, whenever the cooperation is efficient, it is also stable, and the unified country is not subject to secession threats.

It is important to mention the way a cost allocation assigned to country’s citizens represents a transfer policy of the central government. The crucial element in designing a secession-proof transfer scheme is the degree of equalization between advantaged and disadvantaged regions of the country. Indeed, some equalization is required in order to support disadvantaged regions that may be vulnerable to

---

10 In our formal analysis, we opt for a cooperative approach to address this issue. The choice of a cooperative versus noncooperative model is a rather delicate task that does not obey very stringent rules. If an interaction among agents is governed by precise rules and protocols, it is appropriate to model it as a strategic form game where all potential moves are described very accurately without room for mistake. Even in this case one incurs the risk of deriving predictions based on a fragile structure of a specific construction. In the absence of a priori protocol for negotiations among parties involved, one may abandon a noncooperative mode in favor of an alternative cooperative approach based only on a surplus available to each coalition of players. We believe that in the context of secessions and monetary compensations, it is worthwhile to adopt the protocol-free cooperative approach of this paper. However, one has to recognize that constitutional constraints on secessions, as in Canada and France, could be modeled as a normal form game. Thus, a mixture of two approaches could be used for an analysis of the issues discussed in this paper.

©International Monetary Fund. Not for Redistribution
secessionist threats in the absence of equalization transfers. Thus, one of the central government's objectives is to design an equalization scheme that would eliminate, or at least reduce, the horizontal imbalances among the regions and deter their threat of secession. Ter-Minassian (1997) argues for a need for transfers to address this issue: "If local jurisdictions must rely on their own revenue sources, the poorer jurisdictions will have less resources than the richer ones. Therefore, they will not be able to finance services at the same level as the richer jurisdictions. Should a country accept this differentiation in the quality of services? If not, there is a need for the transfer of resources from rich to poor jurisdictions." Ahmad and Craig (1997) point out that, indeed, "national governments may wish to ensure that citizens in different regions and localities have access to a certain modicum of publicly provided services." To achieve this goal, horizontal imbalances in fiscal capacity should be addressed by equalization transfers from the center or among regions. Our second result derives a structure of equalization transfers under the requirement of secession-proofness. It establishes the principle of partial equalization, which asserts that

- in order to prevent a threat of secession by disadvantaged regions, they must be subsidized by advantaged regions; and
- in order to deter a threat of secession by advantaged regions, their required contributions should not be excessive.

We specifically determine a degree of partial equalization generated by secession-proofness and show that, in the case of the uniform distribution of citizens' preferences, the equalization rate is exactly 50 percent. We also demonstrate that, in the absence of a redistribution mechanism, no intervention approach may leave disadvantaged areas of the country prone to secession. On the other hand, the Rawlsian transfer scheme, which completely equalizes the fiscal capacities of all regions, would cause advantaged regions to threaten to secede.

The principle of partial equalization suggests that, although the gap between advantaged and disadvantaged regions must be reduced, it should not be completely eliminated.

Before proceeding with the body of our analysis, let us briefly discuss how this work connects to the most closely related literature.11

I. Some Related Literature

First, we would like to compare our analysis with that of Alesina and Spolaore (1997) (AS, henceforth). The only relevant difference between the two definitions of secession-proofness12 is the set of policy instruments available to the country or any of its potentially seceding regions: AS consider the equal share of the government cost, whereas we allow for the full range of compensation schemes. The AS result reconciles stability and efficiency in the sense that, when the number of countries reaches the efficient threshold, no region would want to break away. In our single-country setup this translates into the elimination of a threat of secession

---

11 See also reviews by Bolton, Roland, and Spolaore (1996) and Young (1998).
12 Called C-stability in AS.
in an efficient nation. As an entry-deterring instrument, transfers, therefore, become redundant if the nation and potential seceding regions are limited to the equal-share cost allocations.

The AS result has, however, been derived under the assumption of the uniform density of citizens' preferences. The uniform distribution is a case of a nonpolarized society, and the natural question is whether the irrelevance of equalization schemes as a secession-deterring device would hold under an increased degree of polarization of a country's citizens. Haimanko, Le Breton, and Weber (2003) provide a negative answer to this question and show that if the degree of polarization of citizens' characteristics is sufficiently high, then, in the absence of transfers, an efficient cooperation, in general, is not stable and the efficiency alone does not eliminate a threat of secession by a country's regions. This argument strengthens the need for an examination of equalization schemes as a device against possible secessions in countries where the citizens' preferences exhibit a high degree of polarization and heterogeneity.

The heterogeneity of citizens' preferences over the provision of public goods has been studied by Casella (1992), Casella and Feinstein (2002), Feinstein (1992), Perroni and Scharf (2001), and Wei (1991) who (implicitly or explicitly) utilized the Hotelling location model. Cremer, De Kercbove, and Thisse (1985) develop a model that examines the number and location of public facilities. There is also a literature rooted in the Tiebout tradition (Wooders, 1978; Guesnerie and Oddou, 1981, 1987; Greenberg and Weber, 1986; Weber and Zamir, 1985), where the heterogeneity of preferences among individuals and the impossibility of lump-sum financing of public good provision lead to the formation of small jurisdictions. These papers focus on the existence and the characterization of stable partitions of the individuals into jurisdictions, where equilibrium and stability notions capture various scenarios concerning the mobility of individuals and groups of individuals across jurisdictions and the decision-making process about the level of public good provision. The equilibrium and stability notions used in these papers represent a mix of cooperative and noncooperative concepts. The noncooperative nature comes through the usage of the concept of Nash equilibrium and its refinements, and a partition of individuals into jurisdictions is an equilibrium if no individual would find it beneficial to move to another jurisdiction. The cooperative features are introduced by allowing coalitional considerations, where it is assumed that coalitions can enforce specific feasible plans of action if they desire to do so. The mere knowledge of the payoffs of each coalition is sufficient to make predictions on the likely outcome of the game. Our paper follows this tradition by recognizing the role of coalitions without constraining the rules of their formation.

Another related group of papers focuses primarily on the heterogeneity in income rather than individuals' preferences. The first contribution to this line of research was made by Buchanan and Faith (1987), who explore the limits that the threat of secession puts on the tax burden imposed by the majority (which can be rich or poor). This question is the subject of Bolton and Roland (1997), who

---

13They assume that each jurisdiction decides not only upon the location of its government but also on its size.
Michel Le Breton and Shlomo Weber
develop a model of a two-region nation with different gross income distributions. Their main focus is to examine how the threat of secession determines the choice of a purely distributive taxation rate, under the assumption that if secession takes place, all gross incomes are deflated by a common factor. They show that fiscal accommodation in the union reduces the likelihood of secession, but by no means prevents the breakup of the nation under all circumstances. In addition, fiscal accommodation may, surprisingly, lead to higher taxes. As explained in Persson and Tabellini (1999), the identification of the equilibrium secession-proof tax rate is not straightforward, due to the fact that individual preferences may fail to be single-peaked. The Bolton and Roland model has been extended to allow for mobility across borders (Olofsgard, 1999) and the introduction of region-specific shocks (Fidrmuc, 1999).

Federations may also be considered from a contractual perspective. What sort of arrangements or constitution, including secession clauses, should be considered to promote efficiency? Drèze (1993) investigates how assets and liabilities should be appointed at the time of a secession and argues in favor of distributive neutrality. Bolton and Roland (1997) have interesting insights on the determinants of the most preferred arrangement. Persson and Tabellini (1996a,b) examine a risk-sharing argument under moral hazard considerations. More recently, Bordignon and Brusco (2001) offered an analysis of secession rules, arguing that the absence of explicit secession rules can be seen as a commitment device to increase the stability of the federation. A comprehensive analysis of constitutional provisions on country formation was provided by Jéhiel and Scotchmer (2001).

Finally, our paper also relates to the huge empirical public finance literature on transfers across regions targeted at reducing their horizontal imbalances. Implicit transfers across regions are often generated by taxation systems and the design of public spending programs. Many countries, including Canada, Belgium, Germany, and Switzerland, have also adopted explicit interregional transfer rules that are motivated mostly by equity and solidarity considerations. The literature has focused on whether these rules lead to underequalization or overequalization. Although equalization is not driven by equity considerations in our paper, it turns out that secession-proof transfer schemes will entail some form of partial equalization.

II. The Model

We consider a country whose citizens have preferences over the unidimensional policy space $I$, given by the interval $[0, 1]$ with a mass of 1. Each citizen has symmetric single-peaked preferences over the set $I$ and we identify each citizen with an ideal point. The distribution of all ideal points (and, thus, of all citizens’ preferences) is given by a cumulative distribution function $F$, defined over the space $I$. We assume that $F$ has a density function $f$ that is positive and continuous everywhere on the interval $[0, 1]$.

The country chooses a policy in the issue space $I$. In this paper we adopt a spatial interpretation of our model and identify a policy with a location of the government and we do not distinguish between geographical and preference...
dimensions. The country has to cover the cost of provision of public good $g$, which we will simply call a government cost. We assume that the cost of the government $g$ is fixed, so if a region of the country secedes from $I$, it will have to cover the same cost $g$. For simplicity, we restrict our analysis of possible secessions to those subsets of $I$ that consist of the union of a finite number of intervals and we will use the term region for such a subset of citizens.

Suppose now that an individual $t$ belongs to the set $S$, which could be either the unified country ($S = I$) or a seceding region ($S \subseteq N$), whose government chooses a location $p \in I$, then the disutility or “transportation” cost incurred by the individual $t$, $d(t, p)$, is determined by the distance between $t$ and the location of the government:

$$d(t, p) = \alpha |t - p|,$$

where $\alpha$ is a positive cost coefficient. Denote by

$$D(S) = \min_{p \in I} \int_S d(t, p) f(t) dt$$

the minimal transportation cost of the citizens of $S$.

Let us introduce the notion of an $S$-cost allocation that determines the monetary contribution of each individual $t$ towards the cost of government $g$.

**Definition 2.1:** A measurable function $x$ defined on the set $S$ is called an $S$-cost allocation if it satisfies the budget constraint:

$$\int_S x(t) f(t) dt = g.$$

Since in our setup advantages and disadvantages of a possible secession are common knowledge, we allow for lump-sum transfers and do not restrict the mechanism for reallocation of gains from cooperation within each region $S$. Thus every $S$ would minimize its total cost given by the sum of government and transportation costs:

$$g + D(S).$$

Since the minimization of transportation cost for country $S$ implies the selection of its median as the government location, the cost allocation $x$ would imply that the total disutility of a citizen $t \in S$ would be

$$\alpha |t - m(S)| + x(t).$$

14Since $S$ consists of a finite number of connected regions, there always exists an optimal location of the government and, therefore, the cost function is well defined. It is useful to note that for every set $S$ the total transportation cost is minimized when the government chooses its location at the ideal point of its “median citizen,” $m(S)$, that satisfies $\int_{t \in [1, \leq m(S)]} f(t) dt = \int_{t \in [1, \geq m(S)]} f(t) dt$. If $S$ is an interval, then its median citizen is uniquely defined. However, if $S$ consists of several intervals separated from each other, the median of $S$ is not necessarily unique. To avoid ambiguity, we denote by $m(S)$ the leftmost median of $S$. 

©International Monetary Fund. Not for Redistribution
For notational simplicity we assume hereafter that $\alpha$, the marginal rate of substitution between money and the distance to the location of the government, is equal to 1.\footnote{With an obvious change of the variable, the analysis remains unchanged, with $g/\alpha$ instead of $g$.}

Since the transportation costs incurred by citizens are represented by the distance between their location and the policy chosen by the country or the region to which they belong, it again points to the aforementioned conflict between heterogeneity and increasing returns to size. Indeed, on the one hand, a larger country would require a smaller per capita contribution toward government costs $g$ given by an $S$-cost allocation $x$. On the other hand, the bigger the country, the greater the chance that the government's location is far away from citizens living on the margin. One would expect that higher government costs would strengthen the cooperation, so that increasing returns to size would outweigh secession tendencies created by heterogeneity of citizens' preferences.

To examine this issue formally, we introduce the notions of efficiency and stability of cooperation. Cooperation among the different regions of the country would be efficient if no breakup of the country into smaller parts can provide a total benefit exceeding that generated by the united country. As Wittman (1991) puts it: "Two nations would join together (separate) if the economies of scale and scope and the synergy produced by their union created greater (smaller) benefits than the cost."

Consider all possible partitions of the interval $I$ into several connected or disconnected intervals. A typical partition $P$ of $I$ would consist of a number of smaller regions $\{S_1, S_2, \ldots, S_K\}$, where each individual $t \in I$ belongs to one and only one region in $P$. The following definition in the game-theoretic terminology amounts to super-additivity:

**Definition 2.2:** The cooperation is **efficient** if for every partition $P = (S_1, \ldots, S_K)$ we have

$$D(I) + g \leq \sum_{k=1}^{K} \left[ D(S_k) + g \right].$$

It is useful to point out that if the country is broken up into two parts, $S$ and $T$, the efficiency condition implies that

$$g \geq D(S) - D(T).$$

Let us now turn to stability of cooperation, which requires not only positive gains from cooperation but also a mechanism that will allocate those gains in such a way that no separate region $S$ can generate a higher payoff to all its members than that guaranteed to them by the central government. Given a cost allocation and location of the central government, regions of a country may contemplate the possibility of secession. If a region $S$ can make its members better off than under the central government, then $S$ would be prone to secession:

**Definition 2.3:** Consider a pair $(p; x)$, where $p$ is a location of the national government and $x$ is an $I$-cost allocation. We say that the region $S$ is prone to secession (given $(p; x)$) if
THE ART OF MAKING EVERYBODY HAPPY: HOW TO PREVENT A SECESSION

\[
\int_S (d(t, p) + x(t)) f(t) dt > D(S) + g. \tag{1}
\]

If no region is prone to secession, then the pair \((p; x)\) is called \textit{secession-proof}. The cooperation is called \textit{stable} if there exists a secession-proof allocation.

Since throughout the rest of the paper we deal only with cost allocations defined for the entire interval \(I\), we shall call an \(I\)-cost allocation simply a cost allocation.

We now state an important property of secession-proof allocations. It implies that under secession-proof allocation each region is required to make a nonnegative contribution toward government costs. That is, \textit{secession-proofness rules out direct subsidization}. The reason is obvious: If region \(S\) receives a net transfer via cost allocation, the burden of government costs will fall on the rest of the country \(T = I \setminus S\) that would make region \(T\) prone to secession.

**Lemma 2.4:** Let \(x\) be a cost allocation. Suppose that there exists a region \(S\) such that \(\int_S d(t) f(t) dt < 0\). Then for any location of the government \(p\), the region \(T = I \setminus S\) is prone to secession, and therefore the pair \((p; x)\) is not secession-proof.

To complete this section we would like to point out that gains from cooperation will emerge only if the government cost is sufficiently high. If the government cost is low, then no cooperation would emerge, that is:

**Proposition 2.5:** There is a cutoff value of government costs \(g_e\) such that cooperation is efficient if and only if \(g \geq g_e\).

Similarly, if government cost is low, there is little incentive for different regions to stay together in one country. (In the extreme case where \(g\) is zero, every cost allocation would be secession-prone.) Conversely, if the government cost is prohibitively high, no region would be able to pose a threat of secession.

**Proposition 2.6:** There is a cutoff value of government costs \(g_s\) such that cooperation is stable if and only if \(g \geq g_s\).

As we mentioned above, stability of cooperation requires not only positive gains from being together but also the ability to distribute these gains without creating secession-prone regions. That is, the stability requirement is stronger than the efficiency one:

**Proposition 2.7:** If cooperation is \textit{stable}, it is also \textit{efficient}, i.e., \(g_s \geq g_e\).

In the next section we derive the conditions under which stability and efficiency yield the same cutoff value, \(g_s = g_e\). This would determine the lower bound on government costs yielding a secession-proof allocation.

III. The Main Result

As we have stressed above, the designer of the transfer scheme will have to meet two possibly conflicting objectives. On the one hand, it is important to
Michel Le Breton and Shlomo Weber

identify the conditions under which the mere existence of gains from cooperation yields the possibility of reallocating these gains without creating regions that are prone to secession. It is also much easier to verify whether cooperation is efficient by simply observing the economies of scale rather than examining threats of secessions by every region. Our main result reconciles the two objectives: we establish that, under some general conditions, the bounds on efficiency and stability are the same.

We use two conditions to obtain our equivalence result. The first is

**Symmetry:** $f(\cdot)$ is symmetric with respect to the center, i.e., $f(t) = f(1-t)$ for all $t \in I$.

This assumption is quite standard. It implies that the midpoint of the country, $V_2$, is not only its geographic center but is also the median of the distribution of citizens’ location. In our analysis of secession-proof allocations, we therefore restrict our attention to the situations where the government is located in the middle of the country. Thus, instead of considering a pair $(p, x)$ in Definition 2.3, we focus only on cost allocation, assuming that the national government is always located at the point $V_2$.

To state our second assumption, we need some additional notation. For each $t \in I$, let $L_t$ and $R_t$ be the sets of citizens to the left and right of the point $t$, respectively, i.e., $L_t = [0, t]$ and $R_t = [t, 1]$. For the sets $L_t$ and $R_t$, denote by $l(t)$ and $r(t)$ their respective medians, i.e., $l(t) = m(L_t)$ and $r(t) = m(R_t)$. It is easy to verify that both functions $l$ and $r$ are differentiable and increasing in $t$, with $l(0) = 0$, $l(1) = V_2$, $r(0) = \frac{1}{2}$, and $r(1) = 1$. Moreover, the symmetry of the distribution implies that for every $t \in I$

$$r(t) + l(1-t) = 1.$$  \hfill (2)

Our second assumption is:

**Gradually Escalating Median (GEM):** $l'(t) < 1$ on the interval $[0, 1]$.

This assumption implies that if we increase the length of the interval $L_t = [0, t]$ by a small positive number $\delta$, then the median of the interval $L_{t+\delta} = [0, t+\delta]$ increases by an amount less than $\delta$. Obviously, the symmetry of the distribution represented by (2) immediately implies that if $l'(t) < 1$, then $r'(t) < 1$. The class of distribution functions satisfying the condition of gradual escalation is quite large. In particular, it includes all log-concave functions, \footnote{Log-concavity is a special case of a more general concept of $\rho$-concavity studied in Hardy, Littlewood, and Polya (1934). The applications of log-concavity are relatively novel to economic and political science theory (see Caplin and Nalebuff, 1991, and Weber, 1992). The difference between our setup and the models discussed in Caplin and Nalebuff (1991) is that they impose log-concavity on density functions whereas we consider log-concavity of the distribution function.} i.e., those for which the logarithm of the cumulative distribution function $F$ is concave on the interval $[0, 1]$.

**Remark 3.1:** If the distribution function is log-concave it satisfies GEM.
The log-concavity assumption is satisfied for a wide range of symmetric distribution functions. For example, all symmetric distribution functions that are concave and have an increasing density on the interval $[0, \frac{1}{2}]$ are log-concave. It is important to note, however, that the assumptions of GEM and log-concavity allow for populations exhibiting higher density on the borders than in the center. For example, the distribution function whose density is given by

$$
\begin{cases}
-t + 1.25 & \text{if } t \leq \frac{1}{2} \\
t + 0.25 & \text{if } t \geq \frac{1}{2}
\end{cases}
$$

is log-concave (and therefore satisfies GEM). Moreover, the function whose density is

$$
\begin{cases}
-1.2t + 1.3 & \text{if } t \leq \frac{1}{2} \\
1.2t + 0.1 & \text{if } t \geq \frac{1}{2}
\end{cases}
$$

is not log-concave but nevertheless satisfies GEM and, therefore, also belongs to the class of functions covered by our main result (see Figure 1).

In order to formally state our main result, let us turn to a closer examination of secession-proof allocations. Lemma 2.4 implies that every citizen makes a nonnegative contribution toward government costs. Since we assume the symmetry of the citizens' distribution with respect to the median, it is crucial to examine how the contribution of each citizen is correlated with distance to the location of the government. We have to take into account horizontal imbalances among regions and design an equalization mechanism between advantaged citizens (those close to the center) and disadvantaged ones (those close to the borders). To what extent, if at all, should the more disadvantaged regions be compensated via resulting cost allocation?
For this purpose, consider the cost allocation \( x_g(t) \), which is defined as follows:

\[
x_g(t) = \begin{cases} 
  r(t) + \lambda & \text{if } t \leq \frac{1}{2} \\
  r(1 - t) + \lambda & \text{if } t \geq \frac{1}{2}
\end{cases}
\]

where the value of \( \lambda \) is chosen to satisfy the country’s budget constraint:

\[
\lambda = g - 2 \int_0^{\frac{1}{2}} r(t) f(t) dt.
\]

It is important to note that the assumption of gradually escalating median guarantees that the allocation \( x_g(t) \) satisfies the principle of partial equalization. Indeed, the cost allocation \( x_g(t) = r(t) + \lambda \) is increasing, whereas the total cost \( |t - \frac{1}{2}| + x_g(t) = \frac{1}{2} - t + r(t) + \lambda \) is decreasing on the interval \([0, \frac{1}{2}]\). This means that the closer citizens are to the center, the larger their contribution toward government costs, while the total cost is still higher for those close to the borders. Thus, while some equalization takes place, it is not full. It is interesting to note that in the case of uniform distribution, the equalization rate is 50 percent (see Section 4). Then we have:

**The Main Result:** The symmetry and gradually escalating median assumptions imply \( g_e = g_s \). Moreover, if the level of government costs \( g \) satisfies \( g \geq g_e \), then the allocation \( x_g(t) \) is secession-proof.

To prove this result, we consider a level of government costs \( g \) which guarantees that cooperation is efficient, i.e., \( g \geq g_e \). Then we consider the cost allocation \( x_g \) and show that it is secession-proof. Thus, the cooperation is stable, yielding \( g_e \geq g_s \). Since by Proposition 2.7, \( g_e \leq g_s \), it would imply that \( g_e = g_s \).

Note that Remark 3.1 yields the following:

**Corollary 3.2:** Under symmetry and log-concavity we have \( g_e = g_s \).

Although the complete proof of the main result is relegated to the Appendix, we would like to describe the method of the proof, which is of independent interest.

Let us first indicate the major difficulty with verifying secession-proofness. It stems from the fact that one cannot rule out a possibility of secession-prone regions that consist of disconnected intervals. If we were able to restrict our analysis to connected regions only, we could have used the Greenberg and Weber (1986) result, which yields a stable outcome when only connected, or “consecutive,” coalitions are considered. Unfortunately, it does not hold that if there exists a secession-prone disconnected region, there also exists a connected region prone to secession. The assumption of gradually escalating median plays a major role in removing this obstacle and allows us to consider connected regions as the only ones potentially prone to secession.
We proceed in two steps. First we show that for the set of secession-proof cost allocations, only a specific class of connected regions may be prone to secession. Then we show that the issue of secession-proofness translates into a variational problem. More specifically, we consider a set $X$ of cost allocations that satisfy:

- $x$ is a continuous and nonnegative function on the interval $[0, 1]$;
- $x$ is symmetric: $x(t) = x(1-t)$ for all $t \in [0, 1]$;
- $x$ is increasing on the interval $[0, \frac{1}{2}]$;
- $x(t) - t$ is decreasing in $t$ on the interval $[0, \frac{1}{2}]$.

This simply implies that individuals close to the center make larger contributions toward government cost. However, the total burden, which includes transportation costs, is still heavier for citizens who live close to the borders.

The following lemma, the proof of which heavily relies upon the GEM assumption, plays the central role in our proof:

**Lemma 3.3:** Let $x \in X$ be a cost allocation that is not secession-proof. Then there exists $t \in [0, 1]$ such that either $L_t = [0, t]$ or $R_t = [t, 1]$ is prone to secession.

The intuition is as follows. If a cost allocation entailing some degree of partial equalization is prone to secession by a disconnected region, it is also prone to secession by a connected region that contains at least one of the endpoints of the interval $[0, 1]$. Then we are left with an easier task whose technique consists in examining threats of secession only of connected regions.

Lemma 3.3 implies that for any secession-proof allocation $x \in X$, neither $L_t$ nor $R_t$ are prone to secession for any $t \in I$ or, equivalently, the following two conditions should be satisfied for all $t \in [0, 1]$:

\[
\int_{L_t} \left( x(t) + \left| t - \frac{1}{2} \right| f(t) \right) dt \leq g + D(L_t) \quad \text{and} \\
\int_{R_t} \left( x(t) + \left| t - \frac{1}{2} \right| f(t) \right) dt \leq g + D(R_t),
\]

where, we remember, $D(S)$ denotes the minimum of the aggregated transportation cost of members of $S$. For every $t$, denote by $H(t)$ the aggregated transportation cost of citizens of $L_t$ to the location of the government at $\frac{1}{2}$, i.e.,

\[
H(t) = \int_{0}^{t} \left| t - \frac{1}{2} \right| f(t) dt.
\]

Using the symmetry of the citizens' distribution and rearranging the above inequalities, we obtain the necessary and sufficient conditions for a cost allocation $x \in X$ to be secession-proof, namely, that the inequalities

\[
D(I) - D(R_t) - H(t) \leq \int_{0}^{t} x(t) f(t) dt \leq g + D(L_t) - H(t)
\]

hold for all $t \leq \frac{1}{2}$.

The problem is now reduced to a variational problem: find a cost allocation in $X$ to satisfy the integral conditions in (3). The complete details of the solution
Michel Le Breton and Shlomo Weber

to this problem are presented in the Appendix. In the next section we illustrate this problem by considering a special case of a uniform distribution that will help illustrate both difficulties and their resolution in the general case.

IV. Uniform Distribution

In this section we consider the density function that is uniform on the interval \( I \), i.e., \( f(t) = 1 \) for all \( t \in [0, 1] \).

It is easy to verify that in this case the necessary and sufficient conditions for secession-proofness in (3) turn into:

\[
\frac{t^2}{4} \leq \int_0^t x(s)ds \leq \psi(t)
\]

for all \( t \in [0, \frac{1}{2}] \), where

\[
\psi(t) = \frac{3t^2 - 2t}{4} + g.
\]

This means that in order to obtain a secession-proof cost allocation, one has to find, for a given value of \( g \), a function \( y \) sandwiched between \( \frac{t^2}{4} \) and \( \psi(t) \), with \( y(0) = 0, y(\frac{1}{2}) = g/2, y' > 0 \), and \( 0 < y'' < 1 \). Therefore, the search for a secession-proof cost allocation amounts to finding an increasing and convex (but not "too convex") function that connects 0 and \( A \) and whose graph lies within the shaded area depicted in Figure 2.

Our first result of this section derives the explicit value of the minimal threshold \( g_s \) that guarantees efficiency and stability of cooperation:

**Proposition 4.1:** The efficiency bound \( g_e \) is equal to \( 1/8 \).

Since every secession-proof allocation requires each individual to make a non-negative contribution, the question is how the cost of the government \( g \) is shared among its citizens in the uniform case. Our main result yields a secession-proof allocation \( x_g(t) \) (\( g \geq 1/8 \)), which determines the contribution of a citizen \( t \):

\[
x_g(t) = \begin{cases} 
\frac{t}{2} + g - \frac{1}{8} & \text{if } t \leq \frac{1}{2} \\
1 - \frac{t}{2} + g - \frac{1}{8} & \text{if } t \geq \frac{1}{2} 
\end{cases}
\]

The compensation rate generated by this allocation is 50 percent. This means that for every two citizens \( t < t' < \frac{1}{2} \), the contribution of citizen \( t' \) toward government costs exceeds that of citizen \( t \) by one half the distance between \( t \) and \( t' \). To reinforce the need for a balanced (partial) equalization, we consider two extreme allocations, no-compensation allocation, under which each citizen contributes an equal amount toward government costs (no equalization), and Rawlsian Egalitarian allocation, which assigns equal total contributions (including the transportation cost) to every citizen (full equalization), and show that both are not necessarily secession-proof. The intuitive reasons are that in the case of no-compensation allocation a heavy
burden (a lack of subsidy) is put on the distant regions, which may then be prone to secession. In the full compensation mechanism the burden is almost squarely on the shoulders of central regions, which can make them prone to secession. This intuition is supported by our results.

- No-compensation allocation NC. This allocation assigns an equal burden to all citizens. Thus, for every $t \in I$

$$NC(t) = g.$$ 

We have the following proposition:

**Proposition 4.2:** There exists a level of government costs $g^* = 0.134 > \frac{1}{2}$ such that for all $g$, $\frac{1}{3} \leq g \leq g^*$, the no-compensation allocation NC is not secession-proof.

Note that the left side of inequality (4) is always satisfied for $g \geq \frac{1}{2}$ as $t \leq \frac{1}{2}$. This means that for $t \leq \frac{1}{2}$ no region $R_i$ is prone to secession. The reason is that the citizens in the middle of the country are spared from equalization transfers to support distant regions, which would not wish to secede. The secession-prone candidates are, therefore, the regions $L_i(t < \frac{1}{2})$, which contain border citizens but do not occupy the center of the country. It is interesting to mention that only relatively large regions, containing more than 33 percent of the population, could be prone to secession. Indeed, small regions on the margin are not prone to secession because of a heavy burden of per capita government costs if they wish to go alone.
Figure 3. Graphs of No-Compensation Allocation (NC), Rawlsian Egalitarian Allocation (RE), and Secession-Proof Allocation ($x_g$)

- **Rawlsian Egalitarian allocation RE.** This allocation guarantees an equal total cost (including transportation) for all citizens. Thus, $RE(t) + |t - \frac{1}{2}|$ is the same for all $t$. Then $RE(t) + |t - \frac{1}{2}| = g + \frac{1}{4}$, yielding:

$$RE(t) = \begin{cases} t - \frac{1}{4} + g & \text{if } t \leq \frac{1}{2} \\ -t + \frac{3}{4} + g & \text{if } t \geq \frac{1}{2} \end{cases}$$

(Three cost allocations, NC, RE, and $x_g$, are depicted in Figure 3.)

Hence for the range of government costs $\frac{1}{8} \leq g < \frac{1}{4}$, the values of $RE(t)$ are negative for all $t$ satisfying $0 \leq t < \frac{1}{4} - g$. This implies that for every $t < \frac{1}{4} - g$, the citizens of the region $L_t$ will receive a net subsidy. Lemma 3.4 yields that it would not be acceptable for the rest of the country. Indeed, since full equalization puts a heavy burden on citizens close to the center, the corresponding regions $R_t$ will be prone to secession. This immediately yields the following:

**Proposition 4.3:** For any level of government costs $g$ such that $\frac{1}{8} \leq g < \frac{1}{4}$, the Rawlsian Egalitarian allocation is not secession-proof.$^{17}$

$^{17}$Jacques Drèze pointed out to us that, if the citizens were distributed (uniformly) over the entire real line, rather than the bounded interval $[0, 1]$, the full equalization would be the unique secession-proof compensation scheme. Indeed, in this case, for any level of government costs the gains from cooperation are maximized for a partition of the real line into intervals of equal length.
THE ART OF MAKING EVERYBODY HAPPY: HOW TO PREVENT A SECESSION

Linear Compensation Schemes

We complete this section with a characterization of the set of linear compensation allocations, where the rate of equalization is uniquely determined by a slope of the allocation function. For a given value of government costs $g \geq \frac{1}{2}$, we consider a symmetric linear cost allocation $\tilde{x}(t) = \alpha t + \beta$ for all $t \in (0, \frac{1}{2}]$ and $\tilde{x}(1-t) = \tilde{x}(t)$, where for each $0 \leq \alpha \leq 1$ the value of $\beta$ is chosen to balance the government budget: $\int_0^1 \tilde{x}(s) ds = g$. Thus, the allocation $\tilde{x}$ is actually determined by the equalization rate $\alpha$. In order to stress this link we denote $\tilde{x}(t) = x^\alpha_t(t)$.

Now let us turn to examination of secession-proof linear compensation schemes. Not surprisingly, the range of secession-proof equalization rates crucially depends on the value of $g$. If the value of $g$ declines, it increases a likelihood of secession threats and shrinks the range of secession-proof equalization rates. Specifically, for low levels of $g(0.125 \leq g \leq 0.25)$ this range is quite narrow and for $g = 0.125$ it consists of only one point, namely, the 50 percent equalization rate. Since $\frac{1}{2}$ is always a secession-proof rate, it follows that the 50 percent equalization rule provides the unique rate that is secession-proof for all values of $g$ exceeding 0.125.

For relatively high levels of $g \geq 0.25$ the secession-proofness has no implication whatsoever on the range of equalization. Indeed, if government costs rise, a desire to secede would diminish, and at a certain point, $g = 0.25$, it ceases to affect the equalization rates. Formally,

**Proposition 4.4:** The allocation $x^*$ is secession-proof if the equalization rate $\alpha$ satisfies:

$$\begin{align*}
\psi(g) \leq \alpha \leq 4g & \quad \text{if } g \in (0.125, 0.25] \\
0 \leq \alpha \leq 1 & \quad \text{if } g \geq 0.25
\end{align*}$$

where the function $\psi(g)$ is derived in the Appendix. (The shaded area in Figure 4 represents the secession-proof equalization rates for all possible values of $g$.)

V. Concluding Remarks

In this paper we provide an analytical study of interregional fiscal policies in a country with heterogeneous citizens, when different regions may pose a threat of secession. We focus our analysis on the study of compensation schemes that would make the possible advantages of secession for every region inferior to the benefits of remaining within an integrated country.

To address the issue of secession we first examine whether, given heterogeneity of citizens’ preferences, it is desirable and socially efficient to maintain a unified country. We consider the notion of efficiency in a country when all regions are better off under a single national government. Efficiency does not necessarily imply that the gains from being together in the integrated country can be allocated in such a way that no region can guarantee all its citizens a higher payoff than guaranteed by the central government. We investigate the existence of compensation schemes (cost allocations) across regions that guarantee that no region will
Michel Le Breton and Shlomo Weber

Figure 4. Secession-Proof Rates as $g$ Varies

![Graph showing secession-proof rates as $g$ varies](image)

want to secede and break up the country. Such a scheme is called *secession-proof* and such a country is *stable*.

The first result of the paper is to show that, under quite general distribution of citizens' characteristics, efficiency and stability can be reconciled. That is, if a country is efficient, there is always a secession-proof compensation scheme that guarantees the country stability. We then examine secession-proof transfer schemes and the degree of equalization between advantaged and disadvantaged regions these schemes entail. We establish the principle of *partial equalization*: that the gap between advantaged and disadvantaged regions should be reduced, but it should not be completely eliminated. We determine a degree of partial equalization generated by secession-proofness and show that in the case of the uniform distribution of citizens' preferences, the equalization rate is exactly 50 percent. We also demonstrate that, in the absence of a redistribution mechanism, some disadvantaged areas of the country could be prone to secession and that the *no-compensation allocation* is not secession-proof. On the other hand, the Rawlsian transfer scheme, which completely equalizes the fiscal capacities of all regions, would cause advantaged regions to threaten to secede.

The study presented in this paper can be extended in several directions. The focus of this paper was a study of compensation schemes under unanimity procedures. An analysis of alternative mechanisms, such as majority voting, would be a worthwhile contribution. One could also consider a multidimensional policy in which the citizens are distinguished by more than one parameter, say, income, geography, and ethnicity. This extension poses a difficult theoretical challenge but its resolution would be very promising from both theoretical and empirical points of view. Finally, it would be useful to look at the issue of migration, in particular, given heterogeneous tax burdens and living conditions across regions. All those extensions, as well as some other interesting related issues, are left to future research.

We conclude this section with a brief examination of some empirical evidence and a review of *partial equalization* in the real world. In general, the principle of
Table 1. Provincial Per Capita Notional Revenues Before and After
Equalization, 1990-91, in Canada

<table>
<thead>
<tr>
<th>Provinces</th>
<th>Notional Revenue Yield¹</th>
<th>Equalization</th>
<th>Index of Tax Capacity²</th>
<th>Index of Fiscal Capacity³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newfoundland</td>
<td>2.898</td>
<td>1.686</td>
<td>0.63</td>
<td>0.93</td>
</tr>
<tr>
<td>Prince Edward Is.</td>
<td>2.988</td>
<td>1.595</td>
<td>0.65</td>
<td>0.93</td>
</tr>
<tr>
<td>Nova Scotia</td>
<td>3.517</td>
<td>1.066</td>
<td>0.76</td>
<td>0.93</td>
</tr>
<tr>
<td>New Brunswick</td>
<td>3.295</td>
<td>1.288</td>
<td>0.71</td>
<td>0.93</td>
</tr>
<tr>
<td>Quebec</td>
<td>3.973</td>
<td>610</td>
<td>0.86</td>
<td>0.93</td>
</tr>
<tr>
<td>Ontario</td>
<td>5.085</td>
<td>...</td>
<td>1.10</td>
<td>1.03</td>
</tr>
<tr>
<td>Manitoba</td>
<td>3.737</td>
<td>847</td>
<td>0.81</td>
<td>0.93</td>
</tr>
<tr>
<td>Saskatchewan</td>
<td>4.058</td>
<td>525</td>
<td>0.88</td>
<td>0.93</td>
</tr>
<tr>
<td>Alberta</td>
<td>6.306</td>
<td>...</td>
<td>1.36</td>
<td>1.28</td>
</tr>
<tr>
<td>British Columbia</td>
<td>4.808</td>
<td>...</td>
<td>1.04</td>
<td>0.97</td>
</tr>
</tbody>
</table>


¹ Per capita yield of tax bases at national average tax rates.
² Notional revenue before equalization relative to the national average.
³ Notional revenue yield after equalization relative to the national average.

Partial equalization prevails in many countries, especially developed ones, and, in fact, Australia, Canada, Denmark, and Germany, among others, use horizontal imbalances as a basis for equalization policy among regions.

Canada. The principle of equalization is a part of the Canadian Constitution and receives broad national support. There are a variety of transfer programs that attempt to reduce regional inequalities based on capacity, population, and needs (Krelove, Stotsky, and Vehorn, 1997). Equalization payments were redistributed to only seven relatively poor provinces (in terms of 1990-91 data), while Alberta, British Columbia, and Ontario received no payments from this fund (see Table 1).

Clark (1997) points out the success of equalization efforts in reducing regional inequalities. For example, in 1994-95, Alberta’s fiscal capacity index before equalization was almost twice as high as that of Newfoundland. However, the equalization entitlement reduced this gap by almost 75 percent.

Australia. Here, substantial funds are relocated away from the larger states of New South Wales and Victoria, whereas Southern Australia, Tasmania, and the Northern Territory serve as major recipients (Craig, 1997). (See Figure 5.)

Germany. It is interesting to point out that, due to the heavy economic burden of the unification of West and East Germany, the transfer scheme used there exhibited a degree overequalization. Spahn and Föttinger (1997) (Figure 6) show that the fiscal capacity of poorer former East German provinces increased after the transfer, but the contribution paid by rich former West German states reduced their fiscal capacity below the average.

At first glance, the German case seems to be incompatible with our model, where a threat of secession (or integration as in the German case) rules out a possibility of overequalization. One has to recall, however, that the decision on the German reunification has been made in a manner dictated by the political circumstances of that particular period. Had there been public debate in Germany on the
consequences of reunification, including its enormous cost, followed by national elections or a referendum, the final decision might have been quite different.

The European Union. In their recent paper, Hayo and Wrede (2002) examined the issue of partial equalization in the European Union. Their conclusion was that over the 1986–1997 period the EU equalization scheme did, in general, conform to the principle of partial equalization. Thus, the stability of the European Union could be linked to proper equalization arrangements. Hayo and Wrede mentioned, however, that the weakness of the system was its lack of adjustment to change. A major difficulty is reversing a country’s status from recipient to donor, which may pose a formidable challenge to stability of the EU after the proposed enlargement in 2004.

In some instances, mainly in developing countries, there is still a chasm between policy intentions and the implementation of compensation mechanisms.

Russia. Triesman (1996, 1998) argues that direct financial transfers from the federal budget to regions in Russia in 1992–1994 were a function of the lobbying power of the regions, and the issues of horizontal imbalances were not properly addressed in that period. Indeed, Dabla-Norris, Martinez-Vasquez, and Norregaard (2000) and Dabla-Norris and Weber (2001) demonstrated that the principle of partial equalization was not implemented in Russia over the first part of the last decade and, in fact, the gap between rich and poor Russian regions has even widened. The best-off region (Moscow) in 1993 spent close to 12 times more in capital expenditures than the worst-off regions, with the gap widening to more than 24 times what it was in 1998 (see Table 2).

China. Hu and Tan (1996) have demonstrated that, due to governmental policies, the gap between rich and poor provinces in China declined in the 1980s. One
of the tools cited in this regard is the "anti-mega city" policy that was initiated in 1986. It aimed to achieve a certain equalization across the country by placing restrictions on the largest cities: Shanghai, Beijing, and Tianjin. Ahmad, Li, Richardson, and Singh (2002) argue, however, that "the overall transfer system [in China] continues to be sharply regressive, rewarding wealthy regions with increased transfers." (See Figure 7.)

In fact, Hu and Fujita (2001) show that regional disparity between the coastal and interior provinces in China has increased since the mid-1980s and even accelerated after 1990. (For example, the per capita income in Shanghai is ten times larger than in the province of Guizhou.) Both Ahmad and others (2002) and Hu and Fujita (2001) consider the increasing gap between rich and poor regions a serious problem that needs to be addressed at the national level.

It is important to point out that central governments often use transfer schemes as an appeasement policy toward "target" regions. We have already mentioned the
Table 2. Measure of Horizontal Fiscal Imbalance In Russia: Per Capita Regional Expenditures, 1993–98

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean</th>
<th>Coefficient of Variation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>219.7</td>
<td>0.775</td>
<td>100.6</td>
<td>1,198.2</td>
</tr>
<tr>
<td>1994</td>
<td>959.7</td>
<td>1.191</td>
<td>359.8</td>
<td>8,000.7</td>
</tr>
<tr>
<td>1995</td>
<td>1,904.9</td>
<td>1.023</td>
<td>720.2</td>
<td>13,004.3</td>
</tr>
<tr>
<td>1996</td>
<td>2,835.7</td>
<td>1.067</td>
<td>1,050.3</td>
<td>16,521.1</td>
</tr>
<tr>
<td>1997</td>
<td>3,730.1</td>
<td>1.187</td>
<td>1,336.7</td>
<td>30,543.5</td>
</tr>
<tr>
<td>1998</td>
<td>3,184.4</td>
<td>1.022</td>
<td>1,121.5</td>
<td>22,559.8</td>
</tr>
</tbody>
</table>

Source: Dabla-Norris, Martinez-Vasquez, and Norregaard (2000).

Figure 7. Total Transfers Per Capita and GDP Per Capita in 1998, in China

Source: Ahmad, Li, Richardson, and Singh (2002).

transfer policy of the Russian Federation in the early nineties when “trouble-making” regions received a disproportionately large share of the total transfer budget. In China, Tibet receives a large amount of subsidies from the central government, and its regional GDP growth reached almost 12 percent over the period 1993–1999, thus surpassing the national average for six consecutive years (People’s Daily, 2000). There are also special privileges granted to the Basque country in Spain. The Basque country is allowed to collect its own income tax, corporate tax, and VAT, as well as gasoline, tobacco, and alcohol taxes. Then a previously agreed quota is transferred to the Spanish treasury as compensation for Spanish common expenditures and the cost of running state bodies. Thus, the per capita level of public expenditure in the Basque country is much higher than in the rest of Spain. In fact, it is 80 percent higher than the level of public spending in Catalonia and Galicia (Moreno, 2001).
APPENDIX

Proof of Lemma 2.4: Let a cost allocation $x$ and a region $S$ be such that $\int_S x(t)f(t)dt < 0$. Consider the region $T = I \setminus S$, which represents all the individuals outside of $S$. It will have to contribute more than $g$ toward government costs, i.e., $\int_T x(t)f(t)dt > g$. Suppose that the government is located at point $p$. The total cost of region $T$ is:

$$\int_T (x(t) + |t - p|)f(t)dt.$$

However, since, $\int_T x(t)f(t)dt > g$, and $\int_T |t - p|f(t)dt \geq D(T)$,

it follows that

$$\int_T (x(t) + |t - p|)f(t)dt > g + D(T)$$

and $T$ is prone to secession.

Proof of Proposition 2.5: The cooperation is efficient if for any partition $(S_1, \ldots, S_K)$ the following inequality is satisfied:

$$g + D(I) \leq \sum_{k=1}^K [g + D(S_k)].$$

Since it is trivially satisfied for $K = 1$, we may consider only partitions into $K > 1$ regions. Thus, the last inequality can be rewritten as

$$g \geq \frac{1}{K-1} \left[ D(I) - \sum_{k=1}^K D(S_k) \right].$$

Denote by

$$g_* = \sup \frac{1}{|P|-1} \left[ D(I) - \sum_{S \in P} D(S) \right],$$

where supremum is taken over the set of all partitions $P$ of the nation $I$ into more than one region and $|P|$ stands for the number of regions in $P$. Note that $g_* \leq D(I)/2$ is bounded. Thus, the cooperation is efficient if and only if $g \geq g_*$. 

Proof of Proposition 2.6: First, suppose that the value of government costs $g$ is such that there exists a secession-proof allocation $(p, a)$. We shall demonstrate that for every value $g'$, exceeding $g$, there is also a secession-proof allocation. Indeed, let $g' > g$ and consider a new allocation $(p, a')$, where $a'(t) = (g'/g)a(t)$ for every $t$. Then $\int a'(t)f(t)dt = g'$, and since

$$\int (t - p + a'(t))f(t)dt \leq g + D(S),$$

for every region $S \subset N$, we have

$$0 \leq \int_S (t - p)f(t)dt - D(S) \leq g - \int_S a(t)f(t)dt \leq g' - \int_S a'(t)f(t)dt$$

or

$$\int (t - p + a'(t))f(t)dt \leq g' + D(S).$$
Thus, \((p, a')\) is a secession-proof allocation and there exists a level of government costs \(g_s\) such that cooperation is stable if and only if \(g \geq g_s\). To complete the proof of the lemma, it remains to show that \(g_s\) is bounded. For this purpose, consider an allocation \((\forall S, b)\) with \(b(t) = g\) for all \(t \in I\). It is secession-proof if and only if the inequality

\[
\int_S \left( t \cdot \frac{1}{2} \right) f(t) dt + \int_S b(t) f(t) dt \leq g + D(S)
\]

is satisfied for every region \(S\). Consider an arbitrary region \(S\). We have

\[
\int_S \left( t \cdot \frac{1}{2} \right) f(t) dt - D(S) \leq g \left( 1 - F_S \right),
\]

where \(F_S = \int_S f(t) dt\). Moreover,

\[
\int_S \left( t \cdot \frac{1}{2} \right) f(t) dt - D(S) = \int_S \left( t - \frac{1}{2} - \left( 1 - m(S) \right) \right) f(t) dt \leq \int_S \left( \frac{1}{2} - m(S) \right) f(t) dt.
\]

Since \(f = \min_{t \in I} f(t) > 0\),

\[
\left| \frac{1}{2} - m(S) \right| \leq \frac{1 - F_S}{2 f}.
\]

Thus,

\[
\int_S \left( t \cdot \frac{1}{2} \right) f(t) dt - D(S) \leq \int_S \left( \frac{1}{2} - m(S) \right) f(t) dt \leq \frac{1 - F_S}{2 f}.
\]

Then for every value

\[
g \geq \frac{1}{2 f},
\]

the allocation \((\forall S, b)\) is secession-proof and \(g_s\) is, indeed, a finite number.

**Proof of Proposition 2.7:** Suppose that the value of government costs \(g\) is such that cooperation is stable. Then there exists a secession-proof allocation \((p, a)\). Consider an arbitrary partition \((S_1, ..., S_K)\). Since the inequality

\[
\int_{S_k} \left( |t - p| + a(t) \right) f(t) dt \leq D(S_k) + g
\]

holds for every \(k = 1, ..., K\), we have

\[
\int |t - p| f(t) dt + g = \sum_{k=1}^K \int_{S_k} \left( |t - p| + a(t) \right) f(t) dt \leq \sum_{k=1}^K (D(S_k) + g),
\]

that is, the cooperation is efficient.

**Proof of Remark 3.1:** Let \(F\) be a log-concave distribution function. By the implicit functions theorem,

\[
l'(t) = \frac{f(t)}{2 f(l(t))}.
\]

Since \(F\) is log-concave, we have

\[
\frac{f(l(t))}{F(l(t))} > \frac{f(t)}{F(t)}.
\]
and therefore
\[ t'(t) < \frac{F(t)}{2F'(t)} = 1. \]

**Proof of Lemma 3.3:** Let \( x \in X \) be a cost allocation such that a region \( S \) is prone to secession. Assume, without loss of generality, that \( m(S) \leq \frac{1}{2} \). We shall carry out the proof of the lemma in four steps:

1. \( S \cup [0, m(S)] \) is prone to secession as well: It suffices to show that if there are \( p \) and \( q \) with \( 0 \leq p \leq q \leq m(S) \) and \( S \cap [p, q] = \emptyset \), then \( S_1 = S \cup [p, q] \) is prone to secession. Indeed, by condition (a), we have
   \[ \int_{S_1} \left( x(t) + \frac{1}{2} \right) f(t) dt \geq \int_{p} \left( t, \frac{1}{2} \right) f(t) dt + \int_{q} \left( t, \frac{1}{2} \right) f(t) dt \geq \int_{p} \left( t, \frac{1}{2} \right) f(t) dt + g \]
   Hence, we have
   \[ \int_{S_1} \left( x(t) + \frac{1}{2} \right) f(t) dt + D(S) \leq \int_{S} \left| t - m(S) \right| f(t) dt \geq D(S_1) \]
   and \( S_1 \) is prone to secession.

2. Suppose that \( S \cap [m(S), \frac{1}{2}] \neq \emptyset \) and there are \( p \) and \( q \) with \( m(S) \leq p < q < \frac{1}{2} \) and \( S \cap [p, q] = \emptyset \). Let \( \bar{t} \in [m(S), \frac{1}{2}] \) be such that
   \[ F(t) - F(t(S)) = \int_{S} \left( t, \frac{1}{2} \right) f(t) dt. \]
   Then \( S^2 \setminus [m(S), \frac{1}{2}] \cap [m(S), t(S)] \) is prone to secession: Since the shift from \( S \) to \( S^2 \) is a measure-preserving transformation, it follows that \( m(S) = m(S^2) \) and therefore, \( D(S^2) \leq D(S) \).
   Moreover, condition (b) implies that the difference
   \[ \int_{S^2} \left( x(t) + \frac{1}{2} \right) f(t) dt - \int_{S} \left( x(t) + \frac{1}{2} \right) f(t) dt = \int_{S^2} \left( x(t) + \frac{1}{2} \right) f(t) dt - \int_{S} \left( x(t) + \frac{1}{2} \right) f(t) dt > 0. \]
   Hence, since
   \[ \int_{S} \left( x(t) + \frac{1}{2} \right) f(t) dt > D(S) + g, \]
   it follows that
   \[ \int_{S} \left( x(t) + \frac{1}{2} \right) f(t) dt D(S^2) + g. \]
   Thus, \( S^2 \) is prone to secession.

3. Suppose that \( S \cap [\frac{1}{2}, 1] \neq \emptyset \) and there are \( p \) and \( q \) with \( \frac{1}{2} \leq p < q \leq 1 \) and \( S \cap [p, q] = \emptyset \). Let \( \bar{t} \in [\frac{1}{2}, 1] \) be such that
   \[ F(t) - \frac{1}{2} = \int_{S} \left( x(t) + \frac{1}{2} \right) f(t) dt. \]
Then, $S^3 \cap S[0, V_2] \cap [\frac{V_2}{2}, 1]$ is prone to secession. As in the previous case, the shift from $S$ to $S^3$ is a measure-preserving transformation. Thus, $m(S^3) = m(S)$. Moreover,

$$D(S) - D(S^3) = \int_{S \setminus S^3} (t - m(S)) f(t) dt - \int_{S \setminus S^3} t - m(S) f(t) dt = \int_{S \setminus S^3} tf(t) dt - \int_{S \setminus S^3} tf(t) dt.$$ 

Consider now the difference

$$\int_{S \setminus S^3} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt - \int_{S \setminus S^3} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt.$$ 

It can be presented as

$$\int_{S \setminus S^3} tf(t) dt - \int_{S \setminus S^3} tf(t) dt + \int_{S \setminus S^3} x(t) f(t) dt - \int_{S \setminus S^3} x(t) f(t) dt.$$ 

The property (γ) yields

$$D(S) - D(S^3) < \int_{S^3} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt - \int_{S^3} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt.$$ 

Finally, the inequality

$$\int_{S^3} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt > D(S) + g$$

implies that

$$\int_{S^3} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt > D(S^3) + g$$

and $S^3$ is prone to secession.

4. Suppose that there is $q > \frac{V_2}{2}$ such that $[\frac{V_2}{2}, q] \subset S$ and $S \cap [1 - q, \frac{V_2}{2}] = \emptyset$. Then $S^4 = S \setminus [\frac{V_2}{2}, q] \cup [1 - q, \frac{V_2}{2}]$ is prone to secession: Since $D(S^4) \leq D(S)$, the symmetry property (b) implies that the region $S^4$ is prone to secession.

It is easy to verify that the proof of the lemma follows from steps 1 through 4.

**Proof of the Main Result:** It suffices to demonstrate that if $g \geq g_e$, the allocation $x_e$ is secession-proof. It is useful to recall that $g \geq g_e$ implies that (1) holds.

Recall that Lemma 3.3 yields that for any secession-proof allocation $x \in X$, the following two conditions should be satisfied for all $t \in [0, 1]$: 

$$\int_{L} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt \leq g + D(L),$$

$$\int_{R} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt \leq g + D(R).$$

Note that

$$\int_{L} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt - \int_{S} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt =$$

$$\int_{R \setminus S} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt - \int_{S \setminus S} \left( x(t) + d\left( t, \frac{1}{2} \right) \right) f(t) dt > 0.$$
THE ART OF MAKING EVERYBODY HAPPY: HOW TO PREVENT A SECESSION

Thus, the necessary and sufficient condition for a cost allocation $x \in X$ to be secession-proof is that the inequality

$$D(l) - D(R_i) \leq \int_0^l \left(x(t) + d\left(t, \frac{1}{2}\right) \right) f(t) dt \leq g + D(L_i)$$

holds for all $t \leq \frac{1}{2}$. This inequality is equivalent to (3), given in Section III:

$$D(l) - D(R_i) - H(t) \leq \int_0^l x(t) f(t) dt \leq g + D(L_i) + H(t).$$

To proceed, we need the following lemma:

**Lemma A.1:** For every $t \leq \frac{1}{2}$, we have the following two equations:

$$-(D(R_i) + H(t))' = \left(r(t) - \frac{1}{2}\right) f(t)$$

and

$$(D(L_i) - H(t))' = \left(2t - \frac{1}{2} - l(t)\right) f(t).$$

**Proof of Lemma A.1:** Note that

$$-(D(R_i) + H(t)) = \int_t^\infty s f(s) ds - \int_{l(t)}^t s f(s) ds - \int_0^t d\left(s, \frac{1}{2}\right) f(s) ds$$

and

$$-(D(R_i) + H(t))' = -2r'(t) f(r(t)) - \left(\frac{1}{2} - l(t)\right) f(t).$$

But since

$$r'(t) = \frac{f(t)}{2 f(r(t))},$$

it follows that, indeed,

$$-(D(R_i) + H(t))' = \left(r(t) - \frac{1}{2}\right) f(t).$$

Similarly,

$$D(L_i) - H(t) = -\int_t^\infty s f(s) ds + \int_{l(t)}^t s f(s) ds + \int_0^t d\left(s, \frac{1}{2}\right) f(s) ds$$

and

$$(D(L_i) - H(t))' = -2l'(t) f(l(t)) + tf(t) - \left(\frac{1}{2} - l(t)\right) f(t).$$

Again, since

$$l'(t) = \frac{f(t)}{2 f(l(t))},$$

it follows that

$$(D(L_i) - H(t))' = \left(2t - \frac{1}{2} - l(t)\right) f(t).$$
Denote \( a(t) = r(t) - \frac{1}{2} \) for all \( t \leq \frac{1}{2} \) and let us show that \( a(\cdot) \) satisfies (3). Equation (5) implies that
\[
\int_{0}^{t} a(t)f(t)dt = D(I) - D(R) - H(t)
\]
and (1) yields
\[
\int_{0}^{t} a(t)f(t)dt \leq g + D(L_i) - H(t).
\]
Note that the assumption of gradually escalating median implies that the function \( a \) would be a solution of our problem if it were to satisfy the budget constraint. However, the value of
\[
2\int_{0}^{\frac{1}{2}} a(t)f(t)dt
\]
is not necessarily sufficient to cover the government costs \( g \). Let us, therefore, modify the function \( a \) by adding to each individual a fixed payment \( \lambda \) such that
\[
\int_{0}^{\frac{1}{2}} (a(t) + \lambda)f(t)dt = \frac{g}{2},
\]
or
\[
\lambda = g - 2\int_{0}^{\frac{1}{2}} (r(t) - \frac{1}{2})f(t)dt.
\]
We shall show that inequality (3) would not be violated by the function \( x_g(t) = a(t) + \lambda \), i.e.,

**Lemma A.2:**

\[
D(I) - D(R) - H(t) \leq \int_{0}^{t} x_g(t)f(t)dt \leq g + D(L_i) - H(t).
\]

**Proof of Lemma A.2:** Recall that the left side of (3) was actually an equality for \( a \). To show that it would still hold for \( x_g(t) \), one has to demonstrate that \( \lambda \geq 0 \). By (4),
\[
\lambda = g - 2\int_{0}^{\frac{1}{2}} (r(t) - \frac{1}{2})f(t)dt = g - 2D(I) + 2D(L) + 2H(L) = 0.
\]
Since \( 2H(\frac{1}{2}) = D(I) \) and \( D(L_{\frac{1}{2}}) = D(R_{\frac{1}{2}}) \), we have, by (1),
\[
\lambda = g - D(I) + D(L) + D(R) \geq 0.
\]
Before turning to the right side of (3), consider the expression
\[
g - H(t) + D(L_i) - \int_{0}^{t} x_g(t)f(t)dt.
\]
Its derivative is \( 2t - r(t) - l(t) \), which, by the assumption of gradually escalating median, is increasing in \( t \). But \( 1 - r(\frac{1}{2}) - l(\frac{1}{2}) = 0 \), yielding \( 2t - r(t) - l(t) < 0 \) for \( t < \frac{1}{2} \). That is, the expression \( c - H(t) + D(L_i) - \int_{0}^{t} x_g(t)f(t)dt \) is decreasing on the interval \([0, \frac{1}{2}]\). Thus, to complete the proof of the lemma, it remains to verify that the right side of (3) holds for \( t = \frac{1}{2} \) or
\[
\frac{g}{2} \leq g - H(\frac{1}{2}) + D(L)\}
\]
However, the last inequality is equivalent to
\[
g \geq 2\left(\frac{1}{2}\right) - 2D(L) = D(I) - D(L) - D(R).
\]

430
which follows from (1). This completes the proof of the main result.

**Proof of Proposition 4.1:** For \( t \leq \frac{1}{2} \), we have

\[
D(L) = \frac{t^2}{4}, D(R) = \frac{(1-t)^2}{4}, D(t) = \frac{1}{4}, H(t) = \frac{t - t^2}{2}.
\]

First, note that stability is achieved if the inequality (1) holds for all \( L \) and \( R \), i.e.,

\[
g \geq \frac{1 - t^2}{4}.
\]

Since the maximum of the right-hand side is \( \frac{1}{8} \), it follows that \( g_s = \frac{1}{8} \), i.e., the cooperation is stable and efficient if \( g \geq \frac{1}{8} \).

**Proof of Proposition 4.2:** Consider the right side of (3) for the no-compensation allocation \( NC \), which can be rewritten as

\[
t g \leq \frac{3t^2 - 2t}{4} + g.
\]

Then, for all \( t \leq \frac{1}{2} \):

\[
g \geq \varphi(t) = \frac{2t - 3t^2}{4(1-t)}.
\]

It is easy to see that \( \varphi(\cdot) \) is concave and its maximum given by a solution of the equation \( 3t^2 - 6t + 2 = 0 \), whose root is

\[
r^* = 1 - \frac{1}{\sqrt{3}} = .423.
\]

Thus,

\[
g \geq \varphi(r^*) = 1 - \frac{1}{\sqrt{3}} = .134 > g_s.
\]

that is, for the range of government costs \( g \), satisfying \( .125 < g < .134 \), the inequality (4) is violated. Thus, the allocation \( NC \) is not secession-proof, since as for this range of values of \( g \), there are regions \( L_r \), in particular for \( r = .423 \), that are prone to secession.

**Proof of Proposition 4.4:** It is easy to see that a linear allocation \( x^a \) is determined by \( x^a(t) = \alpha t + g - \frac{\alpha}{4} \). The secession-proof conditions (4) for this allocation are:

\[
\frac{t^2}{4} \leq \frac{\alpha t^2}{2} + gt - \frac{\alpha t}{4} \leq \frac{3t^2 - 2t + 4g}{4}
\]

for all \( t \in [0, \frac{1}{2}] \). Note that the left-hand side of the inequality implies

\[
\frac{t}{4} \leq \frac{\alpha t}{2} + g - \frac{\alpha}{4}
\]

which always holds for \( g \geq \frac{1}{8} \) and \( \alpha \leq 4g \). It remains to consider only the right-hand side of the inequality (4). In other words, the rate \( \alpha \) is secession-proof if and only if:

\[
\frac{\alpha t^2}{2} + gt - \frac{\alpha t}{2} \leq \frac{3t^2 - 2t + 4g}{4}
\]

\[14\] It is easy to verify that a further partitioning of the interval \([0, t]\) into smaller intervals would impact the efficiency bound, and it suffices to check the partitions with two sets only.
or equivalently,
\[
\left( \frac{\alpha}{2} - \frac{3}{4} \right)^2 + \left( g - \frac{\alpha}{4} + \frac{1}{2} \right) t - g \leq 0
\]
for all \( t \in [0, \frac{1}{2}] \). Since \( \alpha \leq 1 \), the left-hand side of the above inequality is a concave function whose maximal value is
\[
\frac{(4g - \alpha + 2)^2}{16(3 - 2\alpha)} - g
\]
obtained at
\[
\hat{t} = \frac{4g - \alpha + 2}{2(3 - 2\alpha)}.
\]
Two cases should be considered:

**Case 1:** \( \hat{t} \leq \frac{1}{2} \). This occurs when \( 4g + \alpha \leq 1 \). Note that since \( g \geq \frac{1}{8} \), it also implies that \( \alpha \leq 4g \).

Simple algebra shows that
\[
\frac{(4g - \alpha + 2)^2}{16(3 - 2\alpha)} - g \leq 0.
\]
or
\[
\alpha^2 + \alpha(24g - 4) + (4 + 16g^2 - 32g) \leq 0.
\]
The last inequality holds if and only if
\[
\psi_1(g) \leq \alpha \leq \psi_2(g),
\]
where
\[
\psi_1(g) = 2 - 12g - \sqrt{128g^2 - 16g},
\]
\[
\psi_2(g) = 2 - 12g + \sqrt{128g^2 - 16g}.
\]
Since \( g \geq \frac{1}{8} \), it is easy to verify that \( \psi_1(g) \leq 1 - 4g \leq \psi_2(g) \).

Therefore, in Case 1, the range of secession-proof values of \( \alpha \) is the interval \([\max(0, \psi_1(g)), \max(0, 1 - 4g)]\).

**Case 2:** \( \hat{t} \geq \frac{1}{2} \). This occurs when \( 4g + \alpha \geq 1 \). Note that the function
\[
\left( \frac{\alpha}{2} - \frac{3}{4} \right)^2 + \left( g - \frac{\alpha}{2} + \frac{1}{2} \right) t - g
\]
is increasing on the interval \((0, \frac{1}{2})\) and its value at \( t = \frac{1}{2} \) is
\[
\frac{1}{4} - \frac{g}{4},
\]
which is always negative. Therefore, the only upper bounds for \( \alpha \) are 1 and 4g. Thus, in Case 2, the range of secession-proof values of \( \alpha \) is the interval \([\max(0, 1 - g), \min(4g, 1)]\).

It is easy to see that the function \( \psi_1 \) decreases and \( \psi_1(g) = 0 \) at
\[
g = 1 - \frac{\sqrt{3}}{2}.
\]
THE ART OF MAKING EVERYBODY HAPPY: HOW TO PREVENT A SECESSION

If
\[
\frac{1}{8} \leq g \leq 1 - \frac{\sqrt{3}}{2},
\]
the range of secession-proof values of \( \alpha \) is the union of the two intervals, \([\psi_1(g), 1 - 4g]\) (generated by Case 1) and \([1 - 4g, 4g]\) (generated by Case 2). Thus, we obtain the interval \([\psi_1(g), 4g]\).

If
\[
1 - \frac{\sqrt{3}}{2} \leq g \leq \frac{1}{4},
\]
the range of secession-proof values of \( \alpha \) is the union of the two intervals, \([0, 1 - 4g]\) (generated by Case 1) and \([1 - 4g, 4g]\) (generated by Case 2). Thus, we obtain the interval \([0, 4g]\). If \( g \geq \frac{1}{4} \), only Case 2 can occur and the range of secession-proof values of \( \alpha \) is the interval \([0, 1]\), that is, the allocation \( x_\alpha(t) \) is secession-proof if the equalization rate \( \alpha \) satisfies:

\[
\begin{cases}
\psi_1(g) \leq \alpha \leq 4g & \text{if } g \in [0.125, 0.14] \\
0 \leq \alpha \leq 4g & \text{if } g \in [0.14, 0.25] \\
0 \leq \alpha \leq 1 & \text{if } g \geq 0.25
\end{cases}
\]

Finally, by setting interval \( \psi(g) = \max[\psi_1(g), 0] \), we complete the proof of the proposition.

REFERENCES


Michel Le Breton and Shlomo Weber


Since 1990, Singapore has sought to control the rate of growth of its motor vehicle population by means of a unique auction quota system. Under the vehicle quota system (VQS), the government fixes the number of new motor vehicles allowed on the road each year, then allocates approximately one-twelfth of this annual quota to the public each month by means of a sealed bid uniform price auction. Prospective motor vehicle buyers first have to obtain a quota license (called a certificate of entitlement) before they are allowed to make their purchase.

There is a long-standing literature on optimal government intervention to achieve noneconomic objectives. This literature concludes that in the presence of the constraint that domestic consumption of a good not exceed a certain level, the social utility maximizing policy is a consumption tax on the good. Assuming that the objective is to limit motor vehicle ownership and assuming that there is perfect competition in the motor vehicle market, an auction quota would be equivalent to an import tariff, which, in turn—given that Singapore has no domestic automobile
manufacturing industry—would be equivalent to a consumption tax. Theoretically, therefore, it could be argued that the VQS is an efficient method of restricting the number of new motor vehicles each year.

In practice, however, the implementation of the VQS involves many rules and restrictions that tend to have highly distortionary effects. This paper highlights two important implementation issues: quota subcategorization and license non-transferability. The first issue refers to the practice of subdividing the overall quota into smaller quotas: under the VQS, motor vehicles are classified into different categories based on type and size, with separate quotas for each category. The second issue refers to the practice of prohibiting resale of quota licenses: when the VQS was first introduced in 1990, quota licenses were transferable across buyers, but after about a year, the quota licenses were made non-transferable. These restrictions—subcategorization and non-transferability—were introduced with the aim of achieving a lower and fairer tax burden; however, as the data will show, the outcomes were not always as expected.

Much has already been written about Singapore’s VQS. However, this literature has largely considered the issue in the wider context of transportation policy and congestion management. The focus of this paper is not on the effectiveness of the VQS in addressing the problem of traffic congestion. Instead, the focus is on the effectiveness of the implementation of the VQS, taking its objective of restricting vehicle ownership as given.

Quota rationing schemes are employed throughout the world to restrict commodities as varied as fishery licenses and taxicab medallions. Auction quotas have been used or considered for allocating pollution permits, import licenses, radio frequencies, and foreign work permits, among other things. Traditionally, little attention has been given to the implementation rules of such schemes, although more recently Krishna and Tan (1997, 1998, 1999) have developed some theoretical models of quota implementation. This paper applies theoretical and empirical analysis to the VQS to demonstrate that quota implementation rules matter a great deal in practice as well as in theory. Thus, the experience with the VQS so far may offer potentially useful policy lessons in other applications.

I. The Vehicle Quota System

The VQS became effective in May 1990. Prior to that, the rate of growth of motor vehicle ownership was controlled primarily through price-based measures, including a road tax, an import duty on motor vehicles, a lump-sum registration fee, and an ad valorem additional registration fee. Both the road tax and the additional registration fee were increased periodically, the latter from 15 percent of the motor vehicle’s open market value in the early 1970s to 175 percent in 1990. From 1975

---

2 See Phang, Wong, and Chia (1996) and Toh and Phang (1997), for example.
3 In that regard, one may argue that it would be more effective to target motor vehicle usage rather than ownership. See Chia, Tsa, and Whalley (2001) for a fuller discussion.
4 See Phang, Wong, and Chia (1996) for a description of the motor vehicle tax structure and policies in Singapore prior to the introduction of the VQS.
5 The open market value is the c.i.f. import price of the motor vehicle. It comprises the manufacturer’s price plus freight and insurance costs.
to 1989, the annual rate of motor vehicle growth averaged 4.4 percent, but with substantial year-to-year fluctuations, with growth ranging from 9.6 percent in 1980 and 1982 to −2.7 percent in 1986. The inability of the pricing mechanism to restrain and stabilize the motor vehicle growth rate was what prompted the Singapore government to introduce a quota system for new vehicles. The quota system operates on top of the tax measures. Its purpose is to ensure that a target number of motor vehicles is maintained annually through fixing the rate of increase of new motor vehicles each year. Thus, the VQS is supposed to limit the volatility in the annual rate of motor vehicle population growth, leaving motor vehicle prices to fluctuate according to the level of demand.

The VQS works in the following way. Each year, the quota for new motor vehicles is determined so as to obtain a target rate of growth in the total motor vehicle population. The quota formula is as follows:

$$\left( \frac{\text{Total motor vehicle quota}}{q_y} \right) = \left( \frac{\text{Motor vehicle population}}{q_y} \right) + \left( \frac{\text{Projected deregistrations}}{q_y} \right) + \left( \frac{\text{Unallocated quota}}{q_y} \right) \quad (1)$$

The subscript \( y \) denotes calendar year and the subscript \( q_y \) denotes quota year (which runs from May to April). The quota is set to allow for \( g \) percent growth in the total motor vehicle population, plus additional quota licenses to cover the number of motor vehicles that will be deregistered during the (calendar) year, plus any unallocated quota licenses from the previous quota year. The target rate of growth, \( g \), was initially fixed at 4.3 percent, then reduced to 3 percent. Initially, projected deregistrations in year \( y \) were simply taken to be equal to actual deregistrations in \( y - 1 \), but from quota year 1999–2000 onwards, the authorities have employed an undisclosed formula to project the number of deregistrations in year \( y \).

At the beginning of each month, approximately one-twelfth of the quota is auctioned to the public. Prospective motor vehicle buyers have to obtain a quota license in the appropriate category before they are allowed to make their purchase. Any unallocated licenses are added to the quota in the next auction.

The quota licenses are sold through sealed-bid, uniform price auctions. Each individual is allowed to submit only one bid. Each bidder is required to leave a deposit equal to half his bid amount. The minimum bid is one (Singapore) dollar, and bids must be in whole dollars. Successful bidders pay the lowest winning bid; the difference between the quota price and the deposit amount is due at the time of registration of the motor vehicle. (If the deposit exceeds the quota price, the difference is applied toward the buyer’s registration fees.) Unsuccessful bidders are refunded their deposits.

Initially, the government planned to hold quarterly auctions of quota licenses: the first auction took place in April 1990 and the quota licenses issued during that

---

6Subsequent to the introduction of the VQS, the additional registration fee was reduced in two steps to 150 percent by February 1991. The motor vehicle tax structure was further rationalized in 1998, following the introduction of electronic road pricing.

7The average exchange rates (Singapore dollars per U.S. dollar) during 1990–2000 were:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.81</td>
<td>1.73</td>
<td>1.63</td>
<td>1.62</td>
<td>1.53</td>
<td>1.42</td>
<td>1.41</td>
<td>1.48</td>
<td>1.67</td>
<td>1.69</td>
<td>1.72</td>
</tr>
</tbody>
</table>
RATIONING RULES AND OUTCOMES

Auction were valid for six months from May 1990 to October 1990, i.e., they had to be used to register a new motor vehicle within that time period. Hence, the quota system is considered to have taken effect from May 1990. After the first auction, the frequency of the auctions was increased to once a month, and the validity period of the quota license shortened to three months. In October 1991, the validity period of the quota license for certain categories was lengthened to six months (see Section IV).

The quota license has a life span of 10 years. At the end of this period, the motor vehicle owner may either deregister the vehicle by exporting or scrapping it, or renew the license for a further 5 or 10 years by paying what is called the “prevailing quota price.” If a vehicle is sold (within the country) before the expiry of its quota license, the quota license will be transferred to the buyer together with the vehicle; the seller will have to bid for a new quota license if he wishes to purchase a new vehicle. If a vehicle is deregistered before the expiry of the quota license, the owner is entitled to a rebate on the quota price paid, pro-rated to the remaining life span of the license.

Under the VQS, motor vehicles are divided into several different categories, with a separate quota for each category. Prior to May 1999, there were seven quota categories:
- Category 1: Small cars with engine capacity of 1,000 c.c. and below;
- Category 2: Medium-sized cars with engine capacity of 1,001 to 1,600 c.c., and taxis;
- Category 3: Large cars with engine capacity of 1,601 to 2,000 c.c.;
- Category 4: Luxury cars with engine capacity of 2,001 c.c. and above;
- Category 5: Goods vehicles and buses;
- Category 6: Motorcycles and scooters; and
- Category 7: “Open.”

Category 7 (“open”) quota licenses may be used to purchase any type of motor vehicle. In May 1999, the number of categories was reduced to five: categories 1 and 2 were merged and redesignated category A; categories 3 and 4 were merged and redesignated category B; and categories 5, 6, and 7 were renamed categories C, D, and E, respectively. Subcategorization is discussed further in Section III.

II. Auction Outcomes: Preliminary Evidence

Has the VQS been successful in controlling the rate of motor vehicle growth? The average annual motor vehicle growth rate during 1975–89 (prior to the introduction of the VQS) was 4.4 percent, with a standard deviation of 4.24 percent. The average annual motor vehicle growth rate during 1990–99 (under the VQS) was 2.9 percent, with a standard deviation of 2.06 percent. Thus it appears that the VQS has been successful in lowering the average annual rate of motor vehicle growth and its volatility.

8The prevailing quota price for a given quota category is computed as a three-month moving average of the quota price of that category. (Prior to November 1998, a 12-month moving average was used.)
9Bidders of motorcycles in the open category paid one-third of the quota price in that category.
There are two points worth noting here. First, the VQS targets the annual growth of the total motor vehicle population, not the growth of new vehicle registrations; the latter has ranged from 22 percent in 1999 to -8.3 percent in 1996, partly because the quota growth rate itself has fluctuated substantially from year to year. Second, the VQS has succeeded only in reducing the volatility in annual motor vehicle growth, not eliminating it. The annual motor vehicle growth rate has ranged from -0.3 percent (in 1992 and 1998) to 5 percent (in 1995). The motor vehicle growth rate is determined by both the number of new motor vehicles registered and the number of motor vehicles deregistered during the year. The quota will miss its target if the projected number of deregistrations is inaccurate (the actual number of deregistrations each year has fluctuated between 22,000 in 1995–96 and 54,000 in 1998–99) or if the quota is underutilized.

The reduction in quantity uncertainty has been replaced with an increase in price uncertainty. Figure 1 shows the movement of the quota prices for the seven categories over time: the most striking feature of the graphs is the volatility of the premiums. Although the quota prices of all categories exhibit a general upward trend, the monthly fluctuations are sizable. Furthermore, the quota prices seem to follow more or less the same general pattern: an initial increase, followed by a dip in the last quarter of 1990, a rebound in the first quarter of 1991, and much higher values thereafter. Category 6 (motorcycles) was a special case where the quota price fell sharply in September 1991 and continued to decline to the minimum bid of $1, at which it remained until March 1994. This was due to the imposition of stricter emission standards effective from October 1991—most of the motorcycles in the market at the time did not meet the standards, and redesigned models were not expected for some time.

III. Subcategorization

As mentioned earlier, separate quotas are specified for different sizes and types of motor vehicle. The subcategorization was introduced to allay fears that the quota system would favor the rich. By holding separate auctions for each category, it was envisioned that lower-income motor vehicle buyers would not have to bid against wealthier motor vehicle buyers for quota licenses. This is particularly the case for cars, which—up to the May 1999 auction—were subdivided into four categories on the basis of engine capacity: small cars (category 1); medium-sized cars (category 2); large cars (category 3); and luxury cars (category 4).

The conventional wisdom holds that subcategorization is an undesirable policy since it can lead to situations where the quota is not binding in certain subcategories and very binding in others, resulting in underutilization of the total quota despite a positive quota price in the binding subcategories. This phenomenon has certainly been observed under the VQS. As noted previously, there was a collapse in the demand for motorcycles during 1992–93 so that the quota for category 6 licenses

During 1991/92 to 1998/99, the average annual quota growth rate was 5.2 percent, with a standard deviation of 35.5 percent. The annual quota growth rate was as high as 57.5 percent in 1992/93 and as low as -54.6 percent in 1994/95.
Figure 1. Singapore: Quota Prices, 1990-2000
(In Singapore dollars)

Source: Singapore, Land Transport Authority.
Note: In May 1999, categories 1 and 2 were merged and redesignated category A; categories 3 and 4 were merged and redesignated category B; categories 5, 6, and 7 were renamed categories C, D, and E, respectively.
Ling Hui Tan

(which represented approximately 20 percent of the total quota) was not binding during that time. As a result, the share of total quota that went unallocated was 6 percent in 1991–92, 34 percent in 1992–93, and fully 51 percent in 1993–94.11 During that time, the maximum quota price in the other categories was as high as $65,000.

Despite this, subcategorization can be (theoretically) desirable under certain conditions, depending on the environment and the objective of the authorities. The rationale for subcategorization in the VQS may be analyzed using a partial equilibrium framework similar to Krishna and Tan’s (1997). For simplicity, consider only two categories: category 1 (small cars) and category 2 (large cars). Assume that: (i) the market for cars is perfectly competitive; (ii) there is no substitution across categories; (iii) all cars are imported; and (iv) Singapore is a price-taker on the world market for each category, so that the supply of each category is horizontal at the given world price for that category. Let \( Q_i \) represent the quantity of category \( i \) cars; \( D_i(Q_i) \) the inverse demand function of category \( i \) cars; \( P_i \) the given world price for category \( i \) cars (inclusive of taxes and other charges), where \( i = 1, 2 \).

Suppose a binding quota of \( V \) units is imposed on both categories combined. The quota will introduce a wedge between the demand price, \( D_i(Q_i) \), that consumers are willing to pay for the restricted cars and the supply price, \( P_i \). This wedge, \( D_i(Q_i) - P_i \), measures the value of the quota license to purchase a category \( i \) car. Left to market forces, arbitrage will ensure that the allocation of licenses between the two categories will be such that at the margin, the value of a quota license for a category 1 car is equal to the value of a quota license for a category 2 car. The equilibrium condition under competitive market allocation is thus: \( D_1(Q_1) - P_1 = D_2(Q_2) - P_2 \), with \( Q_1 + Q_2 = V \). These equations implicitly define the equilibrium allocation of category 1 and 2 licenses under competitive market conditions, subject to the total quota, \( V \). Denote these equilibrium quantities as \( q_1 \) and \( q_2 \), respectively, and the equilibrium quota price as \( L \). This is illustrated in Figure 2 where the number of category 1 cars is measured rightward from the \( O_1 \) axis and the number of category 2 cars is measured leftward from the \( O_2 \) axis, where the distance between \( O_1 \) and \( O_2 \) is \( V \).

But will small car buyers necessarily be squeezed out of the market in the absence of subcategorization? Clearly, if \( D_1(Q_1) - P_1 \) is very low relative to \( D_2(Q_2) - P_2 \), then \( q_1 \) will be very small relative to \( q_2 \); at the extreme, a corner solution could obtain whereby \( q_2 = V \) and \( q_1 = 0 \). To be sure, one would expect that at any given quantity, the inverse demand function for small cars will be lower than that for large cars, i.e., \( D_1(Q_1) < D_2(Q_2) \), since one can think of large cars as being of a higher quality (or providing more “services”) than small cars.12 But one would also expect that the world price of small cars will be lower than

---

11In general, some 1–3.5 percent of the total quota goes unallocated each year due to the fact that no tie-breaking procedure exists for identical bids at the cutoff level. For example, if the quota is 15 and there are 10 bids of $15,000 and 10 bids of $10,000, then 10 licenses will be allocated at the lowest successful bid of $15,000; the remaining 5 licenses will not be allocated but carried over to the next auction.

12Following Swan (1970), the quality of a product may be thought of as the amount of services obtained from its consumption. These services are a homogeneous good with a uniform price. To the extent that two products embody unequal amounts of services, they will differ in quality and, hence, in price.
the price of large cars, i.e., \( P_1 < P_2 \). Hence, a priori there would be no reason to expect \( D_1(Q_1) - P_1 \) to be necessarily lower than \( D_2(Q_2) - P_2 \), and so no reason to expect \( q_1 \) to be necessarily smaller than \( q_2 \). However, it will be true that \( L_1/P_1 > L_2/P_2 \) so the overall quota would be relatively unfair to small car buyers as it would result in a higher tax burden for them compared to large car buyers. By contrast, a fairer outcome could be achieved by subdividing the quota such that: \( D_1(Q_1)/P_1 = D_2(Q_2)/P_2 \), with \( Q_1 + Q_2 = V \). The resulting allocation will be \( q_1 \) and \( q_2 \), as shown in Figure 2, such that \( L_1 < L_2 \) and \( L_1/P_1 = L_2/P_2 \).13

Categories 1–4: Cars

Has quota subcategorization succeeded in achieving the objective of equity? The data indicate that the answer is no. Figure 3 plots the quota prices of categories 1, 2, 3, and 4 on the same axis. If subcategorization worked as it should have, the

---

13The above analysis assumed no substitution between the two car categories. If substitution is possible, then the equilibrium market allocation of category 1 licenses will be less than \( q_1 \) and the equilibrium allocation of category 2 licenses will be greater than \( q_2 \). This is because the overall quota raises the price of small cars relative to large cars, resulting in substitution away from the former toward the latter. In this case, small car buyers are not being squeezed out but are voluntarily upgrading to larger cars. Falvey (1979) analyzes such a case.
Figure 3. Singapore: Quota Prices for Car Categories, 1990-2000
(In Singapore dollars)

Source: Singapore, Land Transport Authority.
line representing category 1 quota prices should lie everywhere below the line representing category 2, which should in turn lie everywhere below the line representing category 3, and so on. This is evidently not the case—as can be seen in Figure 3, the lines intersect at several points.

Of the 106 auctions between May 1990 and April 1999, category 1 premiums ranked the lowest of the four car categories in 86 instances (81 percent of the time); category 2 premiums ranked second lowest in 62 instances (58 percent of the time); category 3 premiums ranked second highest in 52 instances (49 percent of the time); and category 4 premiums ranked highest in 57 instances (54 percent of the time). But the desired outcome of $L_1 < L_2 < L_3 < L_4$ occurred in only 45 of the 106 auctions—in other words, over half of the auctions involved an instance where the quota price for a smaller car exceeded that of a larger car. In 14 of these cases, category 1 quota licenses cleared at a higher price than category 4 quota licenses; in two instances (the November 1990 auction and the October 1998 auction), category 1 quota licenses were the most expensive of all the categories auctioned.

Even in those instances where the quota prices for smaller cars turned out to be lower than those for larger cars, the relative tax burden still fell disproportionately more on small car buyers. For example, in January 1992, the quota price was $10,100 for category 1 cars; $16,602 for category 2 cars; $18,500 for category 3 cars; and $19,666 for category 4 cars. During that period, the open market value averaged around $8,500 for category 1 cars; $13,500 for category 2 cars; $24,500 for category 3 cars; and $70,000 for category 4 cars. Thus, the implicit tax rate was approximately 119 percent for category 1 cars; 123 percent for category 2 cars; 75 percent for category 3 cars; and 28 percent for category 4 cars.

These results highlight the pitfalls of subcategorization. In practice, the shape and position of the demand curves are not known with any degree of precision, so that fixing separate quotas for each category becomes a guessing game. As evidenced by the data, over half of the time one or more of the guesses have been off the mark, with the quotas for small and medium-sized cars set too low and the quotas for large and luxury cars set too high relative to their demands.

Category 7: The Open Category

The rationale for the open category was to introduce flexibility in the motor vehicle mix. Quotas for the different categories are based on their proportion in the total motor vehicle population at the end of the previous (calendar) year. It was thought that by allowing a portion of the total quota to be "open," i.e., usable in any category, there would be some room for deviation from the previous year's motor vehicle mix based on changes in demand.

In practice, the annual quota for vehicle category $i$ is

$$
\left( \text{Category } i \right)_{\text{quota}} = g \left( \text{Category } i \right)_{\text{population}} + \alpha \left( \text{Projected category } i \right)_{\text{deregistrations}} + \left( \text{Unallocated category } i \right)_{\text{quota}}
$$

(2)

These 14 cases occurred between May 1990 and November 1998.
Ung

for \( i = 1, \ldots , 6 \), where the subscripts \( y \) and \( qy \) are defined as before. The target growth rate, \( g \), is the same for all categories; as mentioned earlier, it was 4.3 percent initially, later reduced to 3 percent. The parameter \( \alpha \) was initially set at 70 percent but raised to 75 percent in December 1992. The annual quota for category 7 is simply: \((1 - \alpha)\) (Projected total deregistrations).

The following example illustrates how the quotas evolve over time. Let \( i \) denote vehicle category \((i = 1, \ldots , 6)\); category 7 is the open category. For simplicity, assume that (i) all quotas are fully utilized every year so there is no carryover; (ii) a fraction \( \delta_i \) of the previous year’s population of category \( i \) vehicles is deregistered every year; and (iii) the deregistrations are evenly distributed throughout the year so the quota year is effectively equivalent to a calendar year (denoted by \( t \)). Denote quota by \( V_{it} \), deregistrations by \( R_{it} \), and vehicle population by \( Q_{it} \).

The initial (year 1) quotas will then be: \( V_{it} = gQ_{io} + \alpha R_{it} = (g + \alpha \delta_i)Q_{io}V_{it} \) for categories \( i = 1, \ldots , 6 \), and \( V_{7,1} = (1 - \alpha)R_{1} \) for category 7, where \( R_{1} = \sum_{i=1}^{6} R_{i} \), and \( g \) and \( \alpha \) are defined as above. The total quota is \( V_{1} = \sum_{i=1}^{6} V_{it} + V_{7,1} \). Suppose a fraction \( \lambda_{it} \) of the open quota is utilized in category \( i \), where \( \sum_{i=1}^{6} \lambda_{it} = 1 \). Then at the end of year 1, the population of vehicle category \( i \) will be \( Q_{i1} = Q_{io} + V_{i1} - R_{i1} \). The rate of category \( i \) population growth will be greater than \( g \) if \( \lambda_{i1} R_{1} > R_{i1} \) (i.e., if the number of open category licenses used to register category \( i \) vehicles exceeds the number of category \( i \) deregistrations) and less than \( g \) if \( \lambda_{i1} R_{1} < R_{i1} \). The rate of total vehicle population growth is equal to \( g \). If there is no open quota (\( \alpha = 1 \)), then the rate of population growth will be equal to \( g \) for all vehicle categories, meaning that the composition of vehicles will remain fixed at the year 0 configuration.

In year 2, the quota for category \( i \) will be: \( V_{i2} = (g + \alpha \delta_i)Q_{i2} \), so the rate of quota increase for category \( i \) vehicles will be greater than \( g \) if \( \lambda_{i2} R_{1} > R_{i2} \) and less than \( g \) if \( \lambda_{i2} R_{1} > R_{i2} \). Hence, vehicle categories in which open licenses are heavily used will experience an above-average increase in quota for a given rate of deregistrations; vehicle categories in which open licenses are scarcely used will experience a below-average increase in quota.

But what determines the utilization of the open category licenses, i.e., the \( \lambda_{is} \)? Intuitively, one can think of the open quota as being imposed on the aggregate residual demand for quota licenses. Hence, as long as the open quota is not too large, one would expect that its quota price would be close to the maximum quota price in the other categories and that it would be used in the categories with the highest quota prices (i.e., the categories with the most binding quotas).15 The pricing of open category licenses is considered further in Section IV.

---

15 During 1990–99, the correlation coefficients between the quota prices in category 7 and those in the other categories were as follows:

<table>
<thead>
<tr>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
<th>Category 4</th>
<th>Category 5</th>
<th>Category 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7366</td>
<td>0.9097</td>
<td>0.9627</td>
<td>0.9808</td>
<td>0.9362</td>
<td>0.6456</td>
</tr>
</tbody>
</table>

(The correlation coefficient between category 7 and category 6 takes into account the rule that individuals using a category 7 license to register a category 6 vehicle pay only one-third of the category 7 quota price.)
Data on the usage of category 7 quota licenses are not published, but data on new registrations indicate that the open licenses have been used mainly to purchase large cars. This is consistent with the observation that category 3 or 4 quota prices were the highest in 87 percent of the auctions. On average during 1990–99, the ratio of new registrations to quota level was 95 percent for category 1, 113 percent for category 2, 195 percent for category 3, and 260 percent for category 4. In other words, the number of new category 3 cars that were actually purchased during that period was almost double the amount set by the category 3 quota, and the number of new category 4 cars purchased was over two and a half times the amount set by the category 4 quota. This would have been possible only through the use of the open quota.

The composition of the car population has indeed shifted over the last ten years toward larger cars and away from smaller cars. In 1990, the makeup of the car population was 15 percent category 1 cars; 67 percent category 2 cars; 14 percent category 3 cars; and 4 percent category 4 cars. By 1999 the proportions had changed to 12 percent category 1 cars; 60 percent category 2 cars; 20 percent category 3 cars; and 8 percent category 4 cars. In fact, according to Phang, Wong, and Chia (1996, p. 148), “by 1995, the Mercedes Benz had overtaken the Toyota as the most popular make of car registered in Singapore.” This increasing population of large cars has led to larger quotas for these cars: between 1990–91 and 1998–99, category 1 and 2 quotas declined on average by 6 percent and 1 percent per year, respectively, while category 3 and 4 quotas grew on average by 4 percent and 8 percent per year, respectively.

Therefore, it would appear that the open quota has met its objective of allowing flexibility in the composition of the motor vehicle population. However, this flexibility may be more illusory than real. The mechanism by which the open quota allows flexibility is through price arbitrage across categories—as mentioned above, the open quota will be used in the category with the highest license price, or the greatest residual demand. But the objective of subcategorization was precisely to prevent price arbitrage so as to achieve a more equitable tax burden among the different groups of car buyers. Hence the two rules are inconsistent. As a result, the observed shift in preferences may not reflect an exogenous change in the public’s tastes so much as a response to the quota system itself. Put differently, the shift toward large cars may not have been because the public grew to prefer large cars over small cars and the open quota allowed the system to accommodate this change in preferences; rather, the shift toward large cars may have been caused by the open category, subcategorization, and the quota formula.

An Alternative to Subcategorization: Ad Valorem Bids

The experience with quota subcategorization provides a good illustration of the distortions that come with such a practice. Although social equity is a desirable objective, quota subcategorization is not the best means by which to achieve it. Interestingly, the authorities have so far not considered the possibility of eliminating subcategorization and introducing ad valorem bids.16 Under such a scheme, there

---

16This was first suggested by Koh and Lee (1994). The VQS review committee did consider a suggestion for a single car category with a scaling factor based on the open market value of the motor vehicle to be purchased, but rejected it on the basis that it would make the system "unnecessarily complex" (www.gov.sg/mincom/mincompr/full_text5.htm, p. 3).
would be only one overall quota, and potential motor vehicle buyers would bid in terms of a percentage over the open market value of the motor vehicle rather than in nominal (Singapore dollar) terms. In other words, auction participants would be required to specify the extra ad valorem duty that they would be willing to pay for their desired vehicle (in addition to existing taxes and fees). The equity objective would be better served by this scheme since buyers of expensive motor vehicles would pay the same percentage premium (relative to the price of the motor vehicle) as buyers of less expensive vehicles. Under the current system of quota subcategorization, buyers of expensive motor vehicles usually pay a lower percentage premium (and sometimes even a lower value premium) than buyers of less expensive vehicles.

The idea of ad valorem bids is not unrealistic; Australia's auction quotas for import licenses in the 1980s utilized such a method. It may be argued that ad valorem bids could encourage underinvoicing; however, there is no reason to assume that this would be more likely for more expensive motor vehicles than less expensive ones. Furthermore, such a system would be considerably simpler than the current system of quota subcategorization, both for the general public (by eliminating the need for strategic decisions on which category to place a bid) as well as for the authorities (by eliminating the need for separate auctions and complicated formulas for distributing the quota).

IV. Nontransferability

When the VQS was first introduced in 1990, the quota licenses were transferable: quota licenses could be resold once for a transfer fee of $10, prior to being used for purchasing a motor vehicle. Once a quota license was used to purchase a vehicle, it became "attached" to the vehicle in the sense that the vehicle could not be resold without the license. During the transferable period, there were no penalties on the resale of (license-inclusive) vehicles.

In mid-1991, the local media began reporting that quota prices were at "all-time highs." The public placed the blame on excessive speculative activity in the quota license market and called for additional restrictions. The government initially maintained that transferability was a desirable option as it enabled the market to determine the allocation of rights to purchase motor vehicles according to

17The same effect could be achieved by having a value quota rather than a volume quota, e.g., by auctioning licenses that conferred the right to purchase a given dollar amount's worth of vehicle, so that individuals desiring more expensive vehicles would have to obtain more licenses. However, a value quota would be much harder to implement in the context of the VQS, where the objective is to control the number of motor vehicles rather than their total value.

18Falvey (1979) and Rodriguez (1979) show that unlike quotas or specific tariffs, ad valorem tariffs do not result in a shift in the composition of imports in favor of more expensive items.

19During the 1980s, Australia auctioned import licenses for textiles, clothing, footwear, and motor vehicles. Bidders in these auctions had to specify the category of the items, the quantity (or value) that they were bidding for, and the ad valorem duty rate they would pay above the duty rate otherwise applicable to the item. Unlike the VQS, the purpose of the Australian quota auction was primarily to obtain information on the degree of protection to the import-competing industries and not to restrict consumption; hence a comparison of the two quota systems would not be very meaningful. The point to note here is simply that a quota system with ad valorem bids is feasible. For further information on the Australian quota auctions, see Takacs (1994).
RATIONING RULES AND OUTCOMES

willingness to pay, but eventually acceded to public opinion and placed restrictions on license resale in an effort to lower quota prices.

In October 1991, resale of quota licenses in all categories except 5 (goods vehicles and buses) and 7 (open) was prohibited for a trial period of 12 months. The rule change meant that a prospective motor vehicle buyer now had to bid for a quota license in his own name instead of obtaining it from a motor vehicle distributor or from the secondary market; once a license was allocated, it could only be used to purchase a vehicle by the individual named in the license. At the same time, the validity period of the nontransferable quota licenses was lengthened to six months, i.e., the vehicle purchase had to be made within half a year of buying a license. (The validity period of category 5 and 7 licenses remained at three months.) Transfers of ownership of motor vehicles inclusive of the quota license were still permitted, subject to a transfer fee of 2 percent of the value of the vehicle. However, in April 1995, additional restrictions were introduced to discourage such transfers: transfers of ownership of motor vehicles registered using (nontransferable) quota licenses from categories 1 through 4 (i.e., cars) within three months of registration were disallowed, and transfers of ownership within four to six months from registration were subject to an additional levy.

In the discussion that follows, license nontransferability refers to the inability to resell the quota license before it is used to purchase a motor vehicle. Once a quota license is used to purchase a vehicle, it can technically be transferred (together with the vehicle), subject to the restrictions described above. However, the nature of the transaction will be very different—the sale of a used car versus the sale of a quota license that can be used to purchase a new car—and as such, it will not be the focus of the following discussion.

The rationale for the switch from transferable to nontransferable quota licenses was to eliminate speculation and thereby lower quota prices. As can be seen in Figure 1, the initial effect of the switch was exactly what was desired, i.e., a drop in quota prices across the six categories affected. (The vertical lines in the graphs mark the switch to nontransferability in October 1991.) However, this result was short-lived, as quota prices in all the car categories continued to rise after October 1991, reaching heights well beyond those attained when quota licenses were transferable.20 Despite this, it was decided that the nontransferable categories would remain nontransferable after the trial period was over.

Theoretical Considerations

In order to analyze the effect of (non)transferability on quota license prices, one first has to understand when transferability matters and why. In a world with no uncertainty, where every bidder knows exactly his reservation value of a quota license, the competitive auction would function perfectly in allocating licenses to those who value them most. There would be no scope for resale of licenses after the auction and the secondary market would become redundant.

20The exception was category 6 (motorcycles) mentioned earlier.
When there is some uncertainty surrounding the value of the quota licenses, however, then transferability becomes an important consideration. Purchasing a car in Singapore involves a considerable financial outlay and since a quota license has to be obtained at least one month before the purchase is made, it is conceivable that an individual may be uncertain of his future valuation of the quota license at the time of the auction.

It is often taken for granted that transferability commands a positive premium in the presence of uncertainty; the public’s (and government’s) expectation that the quota prices would fall when resales were prohibited reflect this assumption. Intuitively, one would think that a transferable quota license has an option value in this case, as it gives its holder the option of using it to purchase a motor vehicle, or selling it on the secondary market. In an uncertain world, this option has value that should be reflected in a higher price for a transferable quota license relative to a nontransferable quota license.

However, it turns out that this conventional wisdom does not always hold in theory. Krishna and Tan (1998, 1999) show that when quota licenses are auctioned competitively to bidders who are uncertain about their valuation of a license at the time of the auction, switching from transferability to nontransferability may lower or raise the quota price. If the quota is very restrictive relative to demand, then the transferability premium is positive; but if the quota is not very restrictive relative to demand, then the transferability premium may be negative.

Space constraints preclude a full elaboration of the model in the VQS context, but the following intuition may help to explain its result. Consider the simplest example where bidders have independently and identically distributed valuations; they do not know for certain their valuations at the time of the auction but realize them only after the auction has taken place. Hence, bidders are identical at the time of the auction (when each knows only the distribution of his valuation) but non-identical after the auction (when each realizes his own valuation). If resale is prohibited, then at the time of the auction, each bidder will be concerned only with his own personal valuation and will bid the expected value of the license to himself, regardless of how many licenses are available and how many rivals he has. If resale is permitted, however, the auction price of a license will depend on how much the license can be expected to cost in the secondary market, so each bidder's bid will depend on the others' valuations as well. This is because the licenses can be exchanged after the auction takes place so that if the bidder is successful but his realization turns out to be low, he may be able to sell his license to someone else whose realization is high. Similarly, if the bidder is unsuccessful and his realization turns out to be high, he may be able to purchase the license from a successful bidder whose realization turned out to be low. The successful bidder therefore has the option of using his license to buy a vehicle if his realization is high, or selling his license to someone else if his realization is low. The value of this option, however, depends on the quota size and the number of bidders there are in the market.

Note that the "transferability premium" should not be confused with the "transferable quota price;" the former refers to the quota license price under transferability and the latter refers to the difference between the quota license price under transferability and the quota license price under nontransferability.
If the quota is very restrictive, then this option is very valuable since the license can easily be resold afterwards if the license holder's realization turns out to be low. As the quota increases with a given number of bidders, the possibility of resale in the secondary market in the event of a low realization becomes smaller since more of the demand would be satisfied in the primary auction. The option becomes less attractive in this case. As the quota increases even further with a fixed number of bidders, winning a license may become more of a liability than an asset since in the event of a low realization, it may be difficult to pass it on to someone else without taking a loss. It may then be optimal to put in a low bid and risk having to buy the license on the secondary market. Simply put, if the quota is very restrictive, then the secondary market quota price will be high on average, and this is reflected in a high auction price. If the quota is not very restrictive, then the secondary market price will be low on average, resulting in a low auction price.

**Empirical Analysis**

As noted earlier, the rationale for switching from transferable to nontransferable licenses was to bring about lower quota prices. This reasoning was based on the conventional wisdom that transferable licenses command a positive premium because they can be retracted. However, theory shows that the conventional wisdom is not always right: the transferability premium can be positive or negative, depending on factors such as the restrictiveness of the quota. This section turns to the empirical evidence to determine whether the switch from transferability to non-transferability actually raised or lowered license prices in the affected categories.

Casual observation of Figure 1 suggests that nontransferability raised rather than lowered the quota prices in categories 1 through 4. According to the theory outlined in the previous section, this would imply that the effective quotas for those categories were not restrictive. However, there are other factors that may have affected the quota prices, such as the supply of quota licenses and demand shifts that were unrelated to nontransferability (possible factors may include income growth and road infrastructure development, among others). In fact, Figure 1 shows that the quota prices for category 5 (which remained transferable throughout) were also higher after the third quarter of 1991.

In an earlier study, Koh and Lee (1993) estimate the impact of nontransferability on the quota price by regressing the quota price on a dummy variable for transferability and other variables such as the ratio of bids received to successful bids and the bid range, for categories 1, 2, 3, and 4 separately. They find that nontransferability was associated with a lower quota price in category 1; had no significant effect in category 2; and was associated with a higher quota price in categories 3 and 4.

This paper takes a different approach by looking at license prices in categories 1, 2, 3, and 4 relative to category 5. The rationale for doing this is to control for any exogenous demand-shift factors that were common to all motor vehicles.²³

²³The assumption here is that the fundamentals driving the premium for category 5 are the same as those driving the premiums for categories 1 to 4. Robustness checks indicate that this is not unreasonable: the license price paths of categories 1 to 5 are quite closely related to movements in domestic asset prices in general (i.e., the stock market index).
Category 5 was chosen as a base because it was not affected by the regime switch.\textsuperscript{24}

The regressions were based on the following model. Denote the relative demand for category \( i \) licenses by:

\[
D_{it} = D(L_{it}/L_{5t}, B_{it}/B_{5t}, \text{Dummy})
\]

where \( L_{it} \) denotes the license price (in Singapore dollars) of category \( i \) at time \( t \); \( B_{it} \) denotes a demand shift parameter, such as the number of bids for category \( i \) licenses at time \( t \); and the dummy variable is equal to 0 for the transferability period (1990:9 to 1991:9) and 1 for the nontransferability period (1991:10 to 1999:04).\textsuperscript{25} The relative demand for category \( i \) licenses should be negatively related to the relative price of category \( i \) licenses and positively related to the relative number of bids for category \( i \) licenses, but could be positively or negatively related to the dummy variable.\textsuperscript{26} On the supply side, denote the relative quota of category \( i \) licenses by \( V_{it}/V_{5t} \). Setting demand equal to supply in equilibrium yields a reduced form such as the following:

\[
\ln(L_{it}/L_{5t}) = \beta_0 + \beta_1 \text{Dummy}_t + \beta_2 \ln(V_{it}/V_{5t}) + \beta_3 \ln(B_{it}/B_{5t}) + \epsilon_{it}. \tag{3}
\]

The log transformation was used as a means of removing growth over time of the variance of the data. Separate regressions were run for categories 1, 2, 3, and 4, using monthly auction data from September 1990 to April 1999.

If the switch to nontransferability had the desired effect, the estimated coefficient on the dummy variable \( \beta_1 \) should be negative and significant. The coefficient \( \beta_2 \) is expected to be negative since all else being constant, a larger supply of category \( i \) licenses relative to category 5 should be associated with a lower license price for that category relative to category 5. The coefficient \( \beta_3 \) is expected to be positive since all else being constant, a larger number of bids received for category \( i \) licenses relative to category 5 licenses suggests a greater relative demand for category \( i \) licenses and hence should be associated with a higher license price for that category relative to category 5.

Pre-regression tests indicate that the unit root hypothesis can be rejected for all four relative license price variables—\( \ln(L_{1t}/L_{5t}), \ln(L_{2t}/L_{5t}), \ln(L_{3t}/L_{5t}) \), and \( \ln(L_{4t}/L_{5t}) \)—using both the Augmented Dickey-Fuller (ADF) and Phillips-Perron tests. The unit root hypothesis can also be rejected for the relative demand variables, \( \ln(B_{it}/B_{5t}) \). The unit root tests for the relative quota variables, \( \ln(V_{it}/V_{5t}) \), are less conclusive, although weak evidence of stationarity can be found for all except

\textsuperscript{24} Also, one can reasonably assume no substitution effects between category 5 (goods vehicles and buses) and categories 1–4 (cars). Category 7—the open category—was also unaffected by the regime switch, but, as argued above, the quota price for category 7 is determined jointly with the quota prices of the other categories, so the inverse demand relative to category 7 would be harder to interpret.

\textsuperscript{25} It is possible that transferability/nontransferability affects not only the intercept of the demand function but also the slopes. However, the data are insufficient to allow for this (there are only 14 observations during the transferable period).

\textsuperscript{26} One may argue that the open market value of category \( i \) cars relative to category 5 vehicles should also be included as an independent variable in the inverse demand function for category \( i \) licenses. Unfortunately, while some information is available on these values, no consistent data series exists. This omission is not too serious if the world prices of the different categories of vehicles move in tandem so that their relative prices do not change much over time.
ln(V1 / V5). However, it can be argued in principle that the ratio of quotas should be stationary in the long run and thus the series may be treated as stationary for purposes of finite sample inference.

With this caveat in mind, the regression results are reported in Table 1. Given that nontransferability did not affect category 5, the results indicate that nontransferability lowered the quota price by 85 percent for categories 1 and 2, 80 percent for category 3, and 70 percent for category 4. The coefficients on the other regressors have the expected signs and are statistically significant. Thus it appears that after controlling for license supply and demand shifts (both category-specific as well as those affecting all motor vehicles), the switch to nontransferability in categories 1–4 lowered their quota prices by some 70–85 percent relative to the transferable regime. Although this effect seems substantial, it should be considered in the context of the actual change in license prices. Between the transferable period (1990:05–1991:09) and the nontransferable period (1991:10–1999:04), the average license price rose by 471 percent in category 1; 572 percent in category 2; 556 percent in category 3; and 795 percent in category 4. In other words, all else being constant, the switch from transferability to nontransferability lowered the quota prices by 70–85 percent; but all else was not constant, and the actual change in prices observed after the switch was an increase of about 500 percent or more. The regression results imply that had the switch from transferability to nontransferability not taken place, the license price increase between the two periods would have been 556 percent in category 1; 656 percent in category 2; 635 percent in category 3; and 865 percent in category 4. Furthermore, it must be borne in mind that nontransferability does carry costs that are difficult to quantify. As demonstrated in Krishna and Tan (1998, 1999), welfare—defined as the sum of surplus and quota rent—is generally lower under nontransferability compared with transferability.

Finally, an estimate of the transferability premium associated with the open category license may be obtained by comparing the category 7 quota price against the maximum quota price (excluding category 7) in the same auction. Recall that open quota licenses remained transferable throughout the sample period. It can be shown that when the other categories are also transferable, the open quota price will be equal to the highest quota price of all the categories (assuming the open quota is not large enough for complete arbitrage), whereas when the other categories are nontransferable, the open quota price should exceed the maximum quota price. Intuitively, this may be understood by noting that if the individual purchases a nontransferable—say, category 4—license, his actual surplus may be positive (if his realization turns out to be above what he paid at the auction) or negative (if his realization turns out to be below what he paid at the auction), but if he purchases an open license, his actual surplus cannot be negative since he can always resell the license if his realization turns out to be below what he paid at the auction. Thus in order for him to be indifferent between the two options, the transferable open license will have to cost more than the nontransferable category 4 license.

Following Halvorsen and Palmquist (1980), the percentage effect of the dummy variable on (L1/L5) is calculated as 100(exp(β5)–1).
Table 1. Regression Results

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>$i = 1$</th>
<th>$i = 2$</th>
<th>$i = 3$</th>
<th>$i = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constant</strong></td>
<td>1.009* (0.414)</td>
<td>2.594* (0.401)</td>
<td>1.432** (0.887)</td>
<td>0.980 (0.967)</td>
</tr>
<tr>
<td><strong>Dummy</strong></td>
<td>-1.934* (0.260)</td>
<td>-1.880* (0.328)</td>
<td>-1.578* (0.452)</td>
<td>-1.220* (0.330)</td>
</tr>
<tr>
<td>$(0 = $transferable; $1 = non-transferable)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>$\ln(V_1/V_5)$</strong></td>
<td>-1.703* (0.352)</td>
<td>-1.732* (0.418)</td>
<td>-1.901* (0.771)</td>
<td>-0.967* (0.433)</td>
</tr>
<tr>
<td><strong>$\ln(B_1/B_5)$</strong></td>
<td>0.911* (0.166)</td>
<td>1.340* (0.212)</td>
<td>1.347* (1.271)</td>
<td>0.576* (0.169)</td>
</tr>
<tr>
<td><strong>AR parameters:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AR(1)</strong></td>
<td>0.554* (0.099)</td>
<td>0.583* (0.095)</td>
<td>0.389* (0.110)</td>
<td>0.491* (0.090)</td>
</tr>
<tr>
<td><strong>AR(2)</strong></td>
<td>-0.169** (0.098)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of observations</strong></td>
<td>103</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td><strong>$R^2$</strong></td>
<td>0.711</td>
<td>0.716</td>
<td>0.403</td>
<td>0.529</td>
</tr>
<tr>
<td><strong>Adjusted $R^2$</strong></td>
<td>0.696</td>
<td>0.705</td>
<td>0.378</td>
<td>0.510</td>
</tr>
<tr>
<td><strong>S.E. of regression</strong></td>
<td>0.523</td>
<td>0.485</td>
<td>0.880</td>
<td>0.557</td>
</tr>
<tr>
<td><strong>$Q(4)$</strong></td>
<td>2.578 [0.275]</td>
<td>0.251 [0.969]</td>
<td>2.849 [0.415]</td>
<td>0.578 [0.901]</td>
</tr>
<tr>
<td><strong>$Q(8)$</strong></td>
<td>4.350 [0.629]</td>
<td>5.198 [0.636]</td>
<td>5.307 [0.623]</td>
<td>4.076 [0.771]</td>
</tr>
<tr>
<td><strong>$Q(12)$</strong></td>
<td>7.809 [0.647]</td>
<td>10.798 [0.460]</td>
<td>6.690 [0.824]</td>
<td>5.729 [0.891]</td>
</tr>
</tbody>
</table>

Notes: Figures in parentheses are standard errors. $L$ denotes quota price (in dollars); $V$ denotes quota level (in number of vehicles); $B$ denotes number of bids; subscripts denote license category. Equation (1) was estimated as an AR(2) model; Equations (2)-(4) were estimated as AR(1). $Q(k)$ denotes the Ljung-Box $Q$-statistic with $k$ lags; figures in square brackets are the corresponding $p$-values. * and ** indicate significance at the 5 percent and 10 percent levels, respectively.

A log-linear regression of the open quota price relative to the maximum quota price, $L_7/L_{max}$, on a constant and the transferability dummy (0 for the transferable period; 1 for the nontransferable period) yields the following result:

$$
\ln(L_7/L_{max}) = -0.069 + 0.045 \text{ Dummy}
$$

(4)

(0.053) (0.054)

106 observations; $R^2 = 0.019$; Adjusted $R^2 = 0.009$;
S.E. of regression = 0.114; DW statistic = 1.892;
Standard errors (heteroskedasticity-consistent) in parentheses:
$Q(4) = 1.669$ ($p$-value 0.796); $Q(8) = 4.391$ ($p$-value 0.820); $Q(12) = 7.382$ ($p$-value 0.831)
DF test statistic for $\ln(L_7/L_{max}) = -10.537$; reject unit root at 1 percent level.
RATIONING RULES AND OUTCOMES

The constant is negative but not significantly different from 0, implying that \( L_7/L_{\text{max}} \) is not significantly different from 1 under transferability. Nontransferability (of categories 1–4) is associated with an increase in \( L_7/L_{\text{max}} \), but the increase is not statistically significant. This suggests that the transferability premium on the open quota was negligible. However, this finding may be partly due to the fact that the transferable open category licenses had to be used within a shorter time period than the nontransferable category 1–4 licenses. (As mentioned earlier, the switch to nontransferability for categories 1–4 was accompanied by a lengthening of the validity period of those licenses from three months to six months, while the validity period of the transferable open category licenses remained at three months.)

V. Conclusions and Policy Lessons

Singapore’s experience with the VQS demonstrates that quota implementation can turn out to be quite complicated. The original aim of the VQS was to control the growth rate of the motor vehicle population as efficiently and fairly as possible. Theoretically, one could argue that a quota would be an optimal policy to achieve this aim. However, as this paper serves to highlight, the actual implementation of the quota makes a difference as seemingly rational rules may have unexpected and undesirable consequences. Singapore’s experience with the VQS offers some potential lessons for quota implementation in general.

The first lesson highlighted in the paper is that whereas a reasonable theoretical case may be made for quota subcategorization, in practice the relevant information for setting the individual quotas is often lacking, so that the end result may not be the desired one. In the case of the VQS, the rationale for subcategorization was to ensure social equity in the sense that buyers of small inexpensive cars should not have to pay the same quota price as buyers of expensive luxury cars. But in practice, subcategorization led to a highly regressive outcome, with buyers of inexpensive cars paying more in relative—and, in some cases, absolute—terms than buyers of expensive cars.

A related point is the importance of consistency among the rules. It is logically inconsistent to have subcategorization for social equity together with an open category for flexibility as the aim of subcategorization is to have different quota prices for different categories, whereas the open category works in the opposite direction, through price arbitrage across categories. Hence, the present design of the VQS cannot achieve both social equity and flexibility at the same time.

Switching to a single quota with ad valorem bids would take care of these considerations automatically and greatly simplify the system as well. Although it is somewhat unusual to require that bidders specify an ad valorem tax rate rather than a nominal (Singapore dollar) bid amount, this has been implemented in other countries, notably in Australia’s quota tariffication exercise during the 1980s. Ad valorem bids would encourage the public to think of the quota license more correctly as a tax on the motor vehicle rather than as an asset in its own right. Such a tax would at least be proportional rather than regressive, and doing away with the subcategorization should substantially reduce quota administrative costs.
Another lesson is that making the quota licenses transferable (or nontransferable) has non-obvious implications for the quota price. Although it is often assumed that the transferability premium is positive, theoretically it can be shown that this need not be the case, depending on the restrictiveness of the quota. In the case of the VQS, it appears that after controlling for license supply and demand factors, the switch to nontransferability did have the desired dampening effect on the quota prices of the car categories, although this effect was overwhelmed by other developments that caused an outward shift of the demand for motor vehicle licenses. Further, this effect should be weighed against the disadvantages of nontransferability, namely the loss of flexibility in an uncertain environment and the consequent deterioration in welfare.

As an ongoing experiment in auction quota implementation, the VQS offers many other potential lessons that are worth exploring. The government has recently replaced the sealed bid auction system with “open” bidding whereby potential bidders are able to observe others’ bids before submitting their own. The argument is that the sealed bid system encourages excessively high bids so increased transparency should result in lower quota prices. The issue is worth studying in greater detail when sufficient data become available.

REFERENCES


"Big Bang" Versus Gradualism in Economic Reforms: An Intertemporal Analysis with an Application to China

ANDREW FELTENSTEIN and SALEH M. NSOULI

This paper analyzes issues concerning the speed of adjustment and sequencing of reforms in a transition economy. It presents a dynamic general equilibrium model parameterized with Chinese data. The model is used to generate different policy simulations that highlight the importance of the policy instruments used during the transition period. The simulations consider privatization, tariff reform, and devaluation, as well as alternative speeds of introducing these policies. They show that different speeds of adjustment, as well as sequencing of reforms, will have very different implications for macroeconomic aggregates. [JEL D58, 21]

This paper analyzes the implications of alternative paths of economic reform in the context of an economy with a large public sector that is being transformed to become more market oriented. Two alternative paths to reform can be envisaged. First, the country can move gradually by selectively introducing reforms and spacing them over time. Second, the country can pursue a "big-bang" approach, under which all reforms are immediately and simultaneously introduced.

No general consensus has emerged on whether the "big-bang" approach to reform is superior or inferior to a gradualist approach. Further, the order in which reforms should be undertaken has remained a matter of debate. This paper examines the economic setting in China, the country to which the analysis is applied. We develop a dynamic general equilibrium model that is used to analyze the effects of different speeds and sequencing of reforms. The model is solved numerically, permitting us to carry out simulations for different policies. Finally, we conclude by drawing some policy conclusions from the simulations.

*Andrew Feltenstein is Assistant Director in the IMF Institute and Saleh M. Nsouli is Deputy Director in the IMF Institute. We would like to thank Stanley Black, Ralph Chami, Era Dahan-Norris, Norbert Funke, Mohsin Khan, and Munir Rachd for helpful comments and suggestions.

1For a detailed discussion of the issues and an overview of the literature, see Nsouli, Rachd, and Funke (2002).
The paper focuses on only three types of policies that might be used to implement reforms. These are the privatization of publicly owned capital, the devaluation of a currency, and reductions in tariff rates. We do not address the general issue of just how general reform should be, nor do we consider many possible policies and reforms.

I. The Reform Setting and Model Intent

The model we use is applied to China. China is quite possibly the best example among formerly planned economies of the use of gradualism in introducing economic reforms. In this section, we provide some background information on the Chinese reform process. There is a general theme that connects most of the elements in this process, namely that there has been a move toward the decentralization of economic decision making and toward the opening of the economy.

In the early 1950s, the Soviet model of central planning shaped the structure of the Chinese economy. The central authority exercised direct administrative control over local governments through various mechanisms. The central authorities also directly controlled major enterprises, distributed funds, and supervised fixed investment through a centralized budgetary allocation. At the same time, production was carried on entirely through state-owned enterprises and collectives, the exchange rate was maintained at an artificially overvalued level, and the economy was closed, through a system of quantitative restrictions and prohibitive tariffs.

Concentration of power at the center reduced the initiative of local governments and hindered production, leading, in 1957, to the move to reform by decentralizing. A wave of recentralization, however, began in the early 1960s, when almost all large and medium-sized enterprises were returned to the central authority. A new decentralization movement started in 1964 and continued throughout the Cultural Revolution period. In the 1970s, most central authority over enterprises was transferred to local governments, which were allowed to retain enterprise depreciation funds. At the same time, the gradual movement toward privatization began, as did the slow opening of the economy to foreign trade and the corresponding devaluation of the exchange rate.

Before 1979, China's budgetary policy essentially consisted of generalized tax collection and profit remittances controlled by the central government and then redistributed as needed to the provinces. This system was changed in the 1980 intergovernmental reform, under which different jurisdictions were assigned different expenditure responsibilities and were also made responsible for collecting necessary revenues and managing their own budgets. Regions that raised more revenues than were necessary were permitted to retain the excess, giving them an incentive to increase revenue collection. This ability to retain revenues was especially attractive to newly privatized state-owned enterprises, which now were able to take advantage of locally provided public infrastructure. At the same time, decentralization was supported by the gradual opening of foreign markets and sequenced devaluations of the exchange rate. All of these changes tended to permit newly privatized firms to operate in a
more market-oriented economy than had existed at the beginning of the decentralization process.\footnote{\textsuperscript{2}}

Economic decentralization in the postreform period has explicitly aimed at introducing a free market economy by gradually removing price controls. Decentralized resource allocation allowed an increase in investment in efficient non-state firms, leading to a rise in aggregate economic growth. On the other hand, productivity in the inefficient state sector lagged behind that in the non-state sector.\footnote{\textsuperscript{3}} In order to sustain public welfare, the central government found it necessary to support the ailing state-owned enterprises. The relative inefficiency of state-owned firms implies that they tend to be hurt by tariff relaxation more than do the privatized, or non-state, enterprises. At the same time, they tend to benefit less from devaluations.

Against this background, we will consider three types of policy reforms in our simulation analysis. Although there are many other reforms that can be examined, these three should give some sense of the lessons to be drawn from our model. The focus will be on reform policies relating to state enterprises, exchange rate policy, and external sector liberalization. More specifically, in terms of the model we use, the reforms are introduced as follows:

- **Privatization of capital:** Initially, the government owns capital, which is sector specific. We assume that the private sector is more efficient than the public sector. We allow privatization to be either gradually or immediately introduced.

- **Devaluation:** We start with an overvalued exchange rate. We then explore two devaluation paths. At one extreme, there is an up-front devaluation, while at the other, the devaluation is effected gradually through several discrete steps.

- **Tariff reduction:** We suppose that the economy has operated under a system of high overall rates on import duties. We examine the effects of both gradual and immediate tariff reductions.

We should view our exercises as essentially forward looking for China. That is, the reforms of the past 25 years have been quite different from those that we will simulate. In particular, China has not made active use of exchange rate policy in the past. Nor has there been a significant movement toward trade reform. In addition, despite the privatization that has been carried out so far, most capital still remains in the public sector. Hence, our exercises should be viewed as a quantitative examination of certain possible policies rather than as a description of the past. The main objective is to get a sense of the effects of different speeds and sequences of reforms.


\footnote{\textsuperscript{3} See Groves and others (1994), Dollar (1990), and Jefferson and Rawski (1994) for further discussion of changes in Chinese productivity. The general relationship between fiscal policy and growth is examined in Easterly and Rebelo (1993).}
II. Model Structure

This section develops the analytical structure of our model. Much of this structure is designed in order to permit a numerical implementation. It is also aimed at reflecting certain stylized elements of the Chinese economy. Although we would not claim that this model lends itself to goodness of fit estimations, we will calibrate the model's endogenous macroeconomic outcomes to corresponding Chinese historical data. The comparison of the simulated and historical data should then serve to offer some confidence in both the structure and the parameterization of the model.

Intuitive Background

Let us give a brief intuitive description of our model. This should help clarify the technical description that we will present next. The model has $n$ discrete time periods. All agents optimize in each period over a two-period time horizon. That is, in period $t$ they optimize given prices for periods $t$ and $t+1$ and expectations for prices for the future after $t+1$. When period $t+2$ arrives, agents re-optimize for period $t+2$ and $t+3$, based on new information about period $t+2$. For example, there may have been a change in fiscal parameters, such as tax or tariff rates, or an exchange rate change. Thus the savings decision made in period $t+1$ may not give an optimal allocation when period $t+2$ arrives. We should note that this does not mean that expectations are incorrect. If there are no exogenous parameter changes, then solving the two-period problem will be equivalent to solving the infinite horizon problem.

We wish to avoid having a perfect foresight model since it would tend to underestimate the costs of gradual reform. The reason for this bias comes from the fact that agents today would know about policy changes that might happen far in the future, as would be the case under gradualism. Hence they would adjust today, rather than have a set of imperfect adjustments over time. Thus there would never be any "wrong" decisions, as might occur under a system of gradualism with unknown future policies. In our framework the agents in the model correctly predict prices and quantities for the next period, but do not know what will happen after that. Hence they optimize with perfect foresight for one period into the future, but they base their expectations for the periods thereafter on the past. That is, they use an adaptive expectations formulation.

The model will have certain features that distinguish it from a standard representation of a market economy. In particular, it has production by both the state and private sectors. In general, we will suppose that the private sector profit maximizes, while the public sector has other goals, such as output or employment targets. A key feature of the transition period will be the privatization of public production, via the transfer of capital to private firms.

---

4The dynamic structure of the model is derived from Blejer, Feldman, and Feltstein (2002).
5Thus our structure would tend to generate less favorable outcomes for gradualism than would a perfect foresight model. It is our view, however, that this is a more realistic outcome than would be the case with the perfect foresight results.
Suppose, for simplicity, that there are two firms at the beginning of the transition period. One firm is publicly owned while the other is privately held. Both firms produce current output via identical neoclassical production functions. Hence the two firms differ only in the ownership of their initial capital stocks. We also do not permit loss-making behavior by the publicly owned firm in its current production. Rather, it faces a “hard” budget constraint.

The firms do differ, however, in their investment behavior. The privately held firm invests so as to equate the present value of the anticipated future stream of earnings on new capital to the cost of borrowing needed to finance the investment. This is thus normal market-determined investment that assumes free entry and hence exhaustion of profit. The public firm, on the other hand, invests according to instructions from the government. That is, it invests a nominal amount that is not based on any economic reason but in reality would be based on a governmental decision that could be politically determined. Since there is normally pressure on the public firm to invest, it does so beyond the optimal level, perhaps to support employment. Hence the public firm may find that the returns to its investments may not cover its interest obligations at some time in the future and it may need to be financed by budget transfers from the government.

The government in our model is quite simple. It makes current expenditures on pure public goods, buying inputs of capital and labor from the private sector. It also invests in public capital, such as electricity generation or transportation. This public infrastructure may augment the efficiency of private production. Our government also carries out privatization policies. Although, in reality, privatization of state enterprises may be carried out by the sale of equity, here we make a simplification. We suppose that the government privatizes state enterprises by simply giving publicly owned capital to the private sector. This reflects the notion that there are no developed capital markets in which it would be possible to sell off public enterprises. Finally, the government finances itself by issuing bonds. These bonds are the same as those issued by the private sector, so public and private sectors compete for private savings. Part of the issuance of debt may be monetized by the central bank.

Let us suppose, again for simplicity, that there are two consumer types, one urban and one rural. Rural labor is used in agricultural production, while urban labor is used in all other types of production. Both consumers maximize intertemporal utility functions with transaction demands for money. They save by holding bonds, either domestic or foreign currency-denominated. The rest of the world is represented by a single export equation.

Finally, equilibrium in the model is determined as market clearing for goods and financial markets in each period. Thus our model differs from that of a dynamic market economy in only a few key ways. Capital is both publicly and privately owned, and public investment is determined by political rather than market parameters. Hence the public sector may find itself financing investment by state enterprises. Privatization may be carried out, and is done so by giving the publicly owned capital to the private sector, rather than by carrying out sales of new equity.
Let us now turn to a formal description of our model. We will do so while attempting to highlight the elements of the model that are meant to capture the transition economy.

Production

Private sector

There are eight factors of production and three types of financial assets:

1. Capital types
2. Urban labor
3. Domestic currency
4. Bank deposits
5. Foreign currency
6. Rural labor
7. Domestic currency
8. Land
9. Foreign currency
10. Rural labor

The five types of capital correspond to the major nonagricultural productive sectors from the national accounts. We wish to avoid using a single, perfectly mobile, capital type since it would generate overly rapid sectoral adjustments. The initial ownership of each capital type is divided between the public and private sector. Each of these factors and financial assets is replicated in each period and, accordingly, has a price in each period.

An input-output matrix, $A_{it}$, is used to determine intermediate and final production in the private sector in period $t$. Corresponding to each sector in the input-output matrix, sector-specific value added is produced using capital and urban labor for the nonagricultural sectors, and land and rural labor in agriculture. Agriculture uses land and rural labor, and all other sectors use one of the five capital types plus urban labor.

The specific formulation of the private sector firm's problem is as follows. Let $Y_{Ki}^i$, $Y_{Li}^i$, be the inputs of capital and urban labor to the $j$th nonagricultural sector in period $i$. Let $Y_{Gi}^i$ be the outstanding stock of government infrastructure in period $i$. The production of value added in sector $j$ in period $i$ is then given by

\[ va_{ji} = va_{ji}(Y_{Ki}^i, Y_{Li}^i, Y_{Gi}^i). \]  

We suppose that public infrastructure may act as a productivity increment to private production. Sector $j$ pays income taxes on inputs of capital and labor, given by $t_{Ki}$, $t_{Li}$, respectively, in period $i$.

We suppose that each type of capital is produced via a sector-specific investment technology that uses inputs of capital and labor to produce new capital. Both the public and private sector invest and produce capital. Investment that is carried out by the private sector is entirely financed by domestic borrowing.\(^7\) Let us define the following notation.

Let $C_{Hj}^i$ be the cost of producing the quantity of capital $H_j$ in period $i$. Let $r_i$ denote the interest rate in period $i$. The return to capital in period $i$ is denoted by $P_{K_j}$. The price of money in period $i$ is given by $P_{Mi}$, and $\delta$ represents the rate of depreciation of capital.

\(^7\)We assume that all foreign borrowing is carried out by the government, so that, implicitly, the government is borrowing for the private investor but the debt thereby incurred is publicly guaranteed.
The cost of borrowing must equal the present value of the return on new capital. Hence,

$$C_{j|i} = \sum_{i=2}^{N} \frac{(1-t_{K_{i}})P_{K_{i}}(1-\delta)^{i-2}H_{i}}{\prod_{j=1}^{i-1}(1+r_{j})},$$

(2)

where $r_{j}$ is the interest rate in period $j$.

**Public sector**

We take a very simple view of public sector production. We will suppose that state-owned enterprises have the same production technology for intermediate and final goods as do those firms in the private sector. Hence there are no efficiency gains in current production if production is transferred from the public to the private sector. We make this assumption for essentially data-based reasons. It will not be possible, using Chinese data, to estimate separate production functions for public and private sector firms.

We do, however, assume that public sector investment is different than private sector investment. In particular, public sector firms do not invest in an optimal fashion, as in equation (3). Rather, the government allocates an arbitrary amount of revenues to investment in each sector. Suppose then that the government decides to spend $GINV_{i}$ on public enterprise capital formation in period $i$. Let public enterprise firm $j$ have a Cobb-Douglas investment function with coefficients $\gamma_{1}, 1-\gamma_{1}$.

We suppose that the government allocates $GINV_{i}$ to the different public enterprises according to an arbitrary set of policy weights $\eta_{ij}$ in period $i$. Thus the government spends $\eta_{ij}GINV_{i}$ on sector $j$’s investment in period $i$. Accordingly, sector $j$ uses $\gamma_{j}\eta_{ij}GINV_{i}/P_{k_{ij}}$ units of capital as inputs to investment in period $i$, and $(1-\gamma_{j})\eta_{ij}GINV_{i}/P_{l_{ij}}$ units of labor. The capital thus produced is then available in period $i+1$.

Thus, public investment in public enterprises is determined purely by policy considerations, rather than intertemporal profit maximization. In addition, this capital formation may be financed by taxes or by borrowing, unlike private investment, and it may, in fact, be loss-making over time, even in the absence of shocks. That is, the public sector may overinvest for noneconomic reasons. In addition, the public sector’s investment is not forward looking in the sense of maximizing a stream of profits.

**Privatization**

We will implement a simple form of privatization of public enterprises. We will assume that, when the government privatizes a state enterprise, it simply gives the capital of the state enterprise to the corresponding private firm. This privatization can be partial. In other words, the government gives a portion of the publicly owned capital to the corresponding private firm and retains a fraction for itself. We thus avoid any issue of the marketing and pricing of public capital. As public capital is
allocated to the private sector, there is a corresponding reduction in public capital expenditure on state-owned enterprises.

Consumption

There are two types of consumers, representing rural and urban labor. We suppose that the two consumer classes have differing Cobb-Douglas demands. The consumers also differ in their initial allocations of factors and financial assets. The consumers maximize intertemporal utility functions, which have as arguments the levels of consumption and leisure in each of the two periods. We permit rural-urban migration, which depends upon the relative rural and urban wage rate. The consumers maximize these utility functions subject to intertemporal budget constraints. The consumer saves by holding money, domestic bank deposits, and foreign currency. He requires money for transactions purposes, but his demand for money is sensitive to changes in the inflation rate.

The specification of the consumers’ maximization problem is given in the Appendix.

The Government

The government collects personal income, corporate profit, and value-added taxes, as well as import duties. It pays for the production of public goods, as well as for subsidies. Unlike the government of a market economy, it also pays for investment in state enterprises and collects revenue from the returns to the capital of those enterprises. If the state enterprises have losses, the government subsidizes them. In addition, the government must cover both domestic and foreign interest obligations on public debt. The deficit of the central government in period 1, $D_1$, is then given by

$$D_1 = G_1 + S_1 + r_1 B_0 + r_f F_0 - T_1 - \sum_{j=1}^{5} p_{kj} k_{Gj} (1 - P R I V_{j1})$$

where $S_1$ represents subsidies given in period 1, $G_1$ is spending on goods and services, while the next two terms reflect domestic and foreign interest obligations of the government, based on its initial stocks of debt. $T_1$ represents tax revenues, while the final term represents the income from publicly owned capital that accrues to the government. The term $PRIV_{j1}$ represents the degree to which public capital in sector $j$ is privatized in period 1. Thus if the sector were fully privatized we would have $PRIV_{j1} = 1$. Any partial privatization would be reflected by a value less than 1.

The resulting deficit is financed by a combination of monetary expansion, as well as domestic and foreign borrowing. If $\Delta y_{BG1}$ represents the face value of domestic bonds sold by the government in period 1, and $C_{F1}$ represents the dollar value of its foreign borrowing, then its budget deficit in period 2 is given by

$$D_2 = G_2 + S_2 + r_2 (\Delta y_{BG1} + B_0) + r_f F_2 - T_2 - \sum_{j=1}^{5} p_{kj2} k_{Gj2} (1 - P R I V_{j2})$$

465

©International Monetary Fund. Not for Redistribution
where \( r_2(\Delta y_{BG1} + B_0) \) represents the interest obligations on its initial domestic debt plus borrowing from period 1, and \( e_2T_{F2}(C_{F1} + B_0) \) is the interest payment on the initial stock of foreign debt plus period 1 foreign borrowing. As before, the final term is the revenue from state enterprises after privatization.

The government finances its budget deficit by a combination of monetization, domestic borrowing, and foreign borrowing. We assume that foreign borrowing in period \( i, C_{Fi} \), is exogenously determined by the lender. The government then determines the face value of its bond sales in period \( i, \Delta y_{BGi} \), and finances the remainder of the budget deficit by monetization. Hence,

\[
D_i = P_{Bi}\Delta y_{BGi} + P_{Mi}\Delta y_{Mi} + e_iC_{Fi}.
\]

### The Foreign Sector

The foreign sector is represented by a simple export equation in which aggregate demand for exports is determined by domestic and foreign price indices, as well as world income. The specific form of the export equation is

\[
\Delta X_{wu} = \sigma_1 \left( \frac{\pi_i}{\Delta e_i + \pi_{Fi}} \right) + \sigma_2 \Delta y_{wu}.
\]

The left-hand side of the equation represents the change in the dollar value of exports in period \( i, \pi_i \) is inflation in the domestic price index, \( \Delta e_i \) is the percentage change in the exchange rate, and \( \pi_{Fi} \) is the foreign rate of inflation. Also, \( \Delta y_{wu} \) represents the percentage change in world income, denominated in dollars. Finally, \( \sigma_1 \) and \( \sigma_2 \) are corresponding elasticities.

### Equilibrium

An equilibrium in our model is defined as market clearing in the markets for factors and financial assets, replicated in each time period. Factor markets are capital (five types), urban and rural labor, and land. Financial assets are domestic currency, domestic bank deposits, and foreign currency. We use a solution method that is based on an approximating fixed-point algorithm to solve for the equilibrium.

### III. Data Sources for China

A variety of data sources for China are used to parameterize the model. The technology for intermediate and final production is given by the 1995 Chinese input-output matrix. This is taken from the 1998 China Statistical Yearbook, and represents 1995 technology. The matrix has 17 sectors, which are as follows:

1. Agriculture
2. Mining
3. Foodstuff
4. Textiles
In order to correspond to our different capital types, we have assumed that these 17 sectors are grouped into 5 aggregate groups. These are as follows:

<table>
<thead>
<tr>
<th>Sectors</th>
<th>Capital type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3–5</td>
<td>3</td>
</tr>
<tr>
<td>6–11</td>
<td>4</td>
</tr>
<tr>
<td>12–13</td>
<td>5</td>
</tr>
<tr>
<td>14–17</td>
<td></td>
</tr>
</tbody>
</table>

We derive indirect taxes from the input-output matrix, using the coefficient for Net Taxes on Production. In order to derive import coefficients for the input-output matrix, as well as import tariff rates, we take a somewhat involved approach. This approach is necessary since the Chinese input-output matrix does not include import coefficients. Here, as with all other derived data, we take our figures from 1995 in order to correspond to the input-output matrix. We assume that all inputs are used as intermediate and primary inputs to production, since we lack the information to derive imports used for final consumption. We use Table 16.5 from the 1998 China Statistical Yearbook to obtain sectoral imports for five sectors. These are (1) Agriculture, (2) Mining, (3) Foodstuff, (4) Textiles, and (5) Other Manufacturing. These are given in U.S. dollars, and we use an exchange rate of 8.35 yuan/$ to calculate domestic currency figures. Corresponding input-output (IO) coefficients are then derived by dividing sectoral imports by the total inputs to sectoral production from the IO matrix.

We need to derive the effective rates of direct taxation for enterprises. Table 7.8 of the China Statistical Yearbook gives total revenues transferred to the government by state-owned enterprises (SOE) and collectively owned enterprises (COE). Table 2.10 gives total income from industry, and from this we derive a tax rate of 4.8 percent that is levied on inputs of capital and labor to all nonagricultural sectors. We also need government current and capital expenditures, as percentages of GDP. Nominal expenditure is taken from Table 7.4, while nominal GDP comes from Table 2.13. From these we obtain a figure for capital expenditure of 2.9 percent of GDP, and for current expenditures on goods and services of
8.6 percent of GDP. We should note that this does not include interest payments, which are generated endogenously by the model.

In order to parameterize the consumer’s problem, we need several types of data. We need utility weights for the different consumer demand functions, as well as initial allocations of factors and financial assets. In order to derive utility weights, we use Table 3.18, the final use part of the IO matrix. This gives expenditures on each of the 17 sectors by agricultural and non-agricultural households. From these, we obtain utility weights for the two consumer categories.

Initial allocations of capital are given by the sectoral operation surpluses, that is, returns to capital, from the IO matrix. Similarly, allocations of labor are given by compensation of laborers across sectors. Thus we define a physical unit of capital and labor as that which earned one yuan in 1995. Initial allocations of money are taken from International Financial Statistics (IFS) as M1 for 1994. Initial allocations of bank deposits are also derived from IFS as 1994 holdings of quasi-money. Finally, we assume that there are no holdings by the two domestic consumer types of foreign currency. The initial holding of foreign currency by the rest of the world, that is, the foreign consumer, is taken to be the 1994 value of exports. This, in turn, is taken from Table 16.3 of the China Statistical Yearbook.

IV. Simulations

In this section we will derive certain conclusions about the effect of alternative paths for the economy, corresponding to different assumptions regarding policy changes and reforms.

Baseline Scenario

The baseline scenario assumes no reform actions are taken. We use the period 1990–95 for the simulation in order to make a comparison with theoretical outcomes. Table 1 gives the macroeconomic outcomes over a six-year simulation period.

Under the baseline scenario, real GDP grows at an average annual rate of 7.0 percent over the period of the simulation. At the same time there is a 12.2 percent average inflation rate over the time period. If we compare the baseline scenario for the period 1990–95 with historical Chinese data, the simulated real growth rate is lower than the historical rate of 12.0 percent, while the simulated inflation rate is slightly lower than the historical rate of 12.9 percent. At the same time, the budget over the first four years of the simulation is reasonably close to Chinese historical outcomes. After four years, the simulated budget deficit is higher than historical levels, largely because of our assumption of a fixed real spending by the government. Finally, the simulated interest rate, after the first two years, is broadly in line with historical values. Until the final two periods, our simulated trade balance is higher than the actual levels.

We do not attempt to claim any statistical “goodness of fit” properties for our simulation exercises. That is, there is no econometric comparison between the historical outcomes and the corresponding endogenous outcomes generated by the general equilibrium model. Rather, we wish only to show general similarities
between the simulated and historical time series. There are several reasons for this approach. We lack enough observations to derive meaningful statistical properties. Additionally, Chinese macro data reflect a variety of price and interest rate controls that we do not include (see Feltenstein and Ha, 1991). Also, we do not attempt to incorporate all historical changes in exogenous parameters that actually occurred in China during the period in question. Finally, we are not trying to use the model for predictive purposes. Rather, we wish to be able to make qualitative judgments about the possible effects of counterfactual policies.

The last line in the table represents the utility levels of the two consumers, which are normalized to 100 for the baseline scenario. The utilities are calculated as the present value of the stream of consumption over the time periods of the simulation. The calculation is made ex post: that is, it is made by calculating the value of a utility function of the following form:

\[ U = x_1^{1/(1+\delta)} x_2^{1/(1+\delta)} \ldots x_T^{1/(1+\delta)} \]

where \( \delta \) is the rate of time preference. The values of \( \{x_i\} \), representing vectors of consumption in each period, are given by the solutions to the intertemporal maximization problem over \( T \) periods. The utility function, which is thus time separable, is Cobb-Douglas in each period. That is,

\[ x_i = x_{i1}^{\alpha_1} x_{i2}^{\alpha_2} \ldots x_{iN}^{\alpha_N} \text{, where } \sum \alpha_i = 1. \]

Thus, the consumer maximizes his utility with a two-period time horizon and expectations about the future thereafter. Because unexpected policies may be
introduced over time, his realized consumption may be different from his intended consumption. Hence the value of his utility function, calculated using realized consumption, may also be different than would have been the case had he achieved his planned consumption levels. Accordingly, our utility levels, if this were a single representative agent model, could be thought of as the present value of a real income index.

While we would not claim that our parameterized model offers a statistically significant rendition of Chinese reality, it provides a basis for carrying out the policy simulations for purposes of illustrating the different effects of alternative speeds and sequencing of reforms.

Privatization

Two initial simulations are carried out in which privatization is introduced at different speeds. In the first, there is a gradual process of privatization, while in the second there is complete privatization in the first period. In carrying out privatization, it is assumed that public state-owned enterprise capital is simply given to the private sector, and that privatization is carried out uniformly across sectors.

To simulate gradual privatization, it is assumed that 30 percent of state-owned enterprise capital is given to the private sector in period 1, 30 percent more in period 3, and the final 40 percent in period 5. Thus, in the last two periods of the simulation there is full privatization. The outcomes are given in Table 2.

There are a number of differences compared with the baseline scenario. First, there is a small but uniform increase in the price level in all periods. As the public capital stock is privatized, there is a corresponding decline in the rate of public investment, which is not fully picked up by the private sector. The resulting lower capital stocks cause the general price level to rise. There is an initial decline in real GDP, due to the decline in aggregate investment. Over time, however, there is a more efficient distribution of sectoral investment by the private sector, leading to an eventual rise in real GDP to 1.4 percent above the baseline scenario in period 6. There is also an improvement in the budget position, relative to the baseline scenario, as the loss in public revenue from privatization is more than made up by the reduction in public investment spending.

<table>
<thead>
<tr>
<th>Table 2. China: Gradual Privatization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Price level</td>
</tr>
<tr>
<td>Real GDP</td>
</tr>
<tr>
<td>Budget deficit</td>
</tr>
<tr>
<td>(in percent of GDP)</td>
</tr>
<tr>
<td>Interest rate</td>
</tr>
<tr>
<td>Trade balance</td>
</tr>
<tr>
<td>(in percent of GDP)</td>
</tr>
<tr>
<td>Utility of consumer 1 = 102.3, utility of consumer 2 = 90.5</td>
</tr>
</tbody>
</table>

Source: Authors' simulation results.
The current account deteriorates slightly, compared with the baseline scenario, in line with the increased appreciation of the exchange rate, and the nominal interest rate is higher, as private investment eventually increases in the new environment. The resulting borrowing requirements of the private sector bring about the increase in the interest rate. Finally, the urban consumer is relatively better off than before, while the rural consumer is worse off. This is because the increase in interest rates has created a positive wealth effect for the urban consumer, who owns relatively more financial assets than does the rural consumer. Accordingly, the urban consumer increases his demand, thereby driving up prices. The rural consumer suffers from these higher prices, and hence realizes a lower utility level.

Suppose that, instead of gradual privatization, an immediate full privatization takes place in period 1. Thus, all the capital of the state-owned enterprises is given to the private sector at the beginning of period 1. Table 3 gives the outcome of simulating a full privatization.

A number of interesting observations, compared with a process of gradual privatization, can be made. First, inflation is significantly higher in the initial periods, with the price levels gradually converging under the two scenarios over the six periods. The higher inflation rates, particularly in the earlier periods, reflect the initial drop in capital and real GDP as the government’s cutback on public investment is not picked up initially by the private sector. Second, there is a further decline in real GDP in the initial two periods, because the elimination of public sector investment is not immediately made up for by a corresponding increase in private output. However, by period 3, the more efficient allocation of private, as compared to public, investment leads real GDP to rise beyond the level achieved under the gradual privatization scenario. Indeed, by period 6, real GDP is 3.2 percent higher than under gradual privatization. The budget deficit deteriorates, reflecting the higher interest rates in this case, as compared to the previous case. These higher rates are themselves caused by the fact that all investment is now carried out by the private sector, starting in period 1. Since private investment is entirely financed by borrowing, unlike public investment, which may be partially financed by monetization, the increased borrowing drives interest rates up. There is a further deterioration in the current account balance, as the higher inflation rates lead to a greater overvaluation of the currency under the fixed exchange rate. Both consumers realize higher levels of utility, as the overvaluation of the currency

<table>
<thead>
<tr>
<th>Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price level</td>
<td>112.6</td>
<td>136.2</td>
<td>132.5</td>
<td>161.6</td>
<td>155.2</td>
<td>188.6</td>
</tr>
<tr>
<td>Real GDP</td>
<td>95.2</td>
<td>102.3</td>
<td>116.6</td>
<td>124.3</td>
<td>138.0</td>
<td>146.7</td>
</tr>
<tr>
<td>Budget deficit (in percent of GDP)</td>
<td>3.9</td>
<td>2.5</td>
<td>-1.7</td>
<td>-1.9</td>
<td>-6.1</td>
<td>-5.3</td>
</tr>
<tr>
<td>Interest rate</td>
<td>11.5</td>
<td>14.5</td>
<td>13.5</td>
<td>21.7</td>
<td>15.0</td>
<td>27.7</td>
</tr>
<tr>
<td>Trade balance (in percent of GDP)</td>
<td>9.4</td>
<td>7.8</td>
<td>5.0</td>
<td>3.6</td>
<td>2.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Utility of consumer 1 = 108.5; utility of consumer 2 = 137.3

Source: Authors’ simulation results.
Table 4. China: Gradual Privatization and 5 Percent Annual Devaluation

<table>
<thead>
<tr>
<th>Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price level</td>
<td>104.3</td>
<td>119.9</td>
<td>136.0</td>
<td>165.0</td>
<td>174.3</td>
<td>215.5</td>
</tr>
<tr>
<td>Real GDP</td>
<td>99.0</td>
<td>106.8</td>
<td>115.0</td>
<td>122.9</td>
<td>134.3</td>
<td>143.6</td>
</tr>
<tr>
<td>Budget deficit (in percent of GDP)</td>
<td>1.6</td>
<td>0.5</td>
<td>-1.5</td>
<td>-2.2</td>
<td>-4.1</td>
<td>-4.0</td>
</tr>
<tr>
<td>Interest rate</td>
<td>5.2</td>
<td>7.2</td>
<td>11.2</td>
<td>13.6</td>
<td>12.6</td>
<td>18.9</td>
</tr>
<tr>
<td>Trade balance (in percent of GDP)</td>
<td>11.7</td>
<td>13.7</td>
<td>8.7</td>
<td>7.9</td>
<td>5.5</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Utility of consumer 1 = 101.3, utility of consumer 2 = 99.1

Source: Authors’ simulation results.

has a positive effect on consumption of both consumers. We should note that the deterioration of the trade balance indicates that this higher level of consumption may not be sustainable in the long run.9

The basic conclusion of these two simulations is that, on balance, an immediate privatization has a more positive impact on consumers than a gradual one. However, in both cases, the increased deterioration of the current account relative to the baseline scenario, because of the increasingly overvalued exchange rate, raises questions of policy sustainability in the absence of exchange rate reduction.

Exchange Rate Policy

In view of the results of the two privatization simulations, this section presents the results of simulations combining an adjustment in the exchange rate with privatization. To examine a “gradual-gradual” approach, assume there is a gradual devaluation along with a gradual privatization. Suppose that there is a 5 percentage point devaluation in each period starting with period 1, and that a gradual privatization is implemented consistent with the process shown in Table 2. The results are given in Table 4.

There are a number of differences compared with Table 2. There is a significant increase in the price level, reflecting the effect of the devaluation, as well as a marginal increase in real GDP, due to the expenditure-switching effect of the devaluation. The budget deficit does not change much, as the increased costs in foreign debt are balanced by increased revenues from import duties. As expected, the current account balance improves, as the overvaluation is progressively corrected. Interest rates do not change much in nominal terms and there are no significant changes in the utility levels of the urban and rural consumers. Finally, we should note that the main reason for the relatively small changes in real output in this simulation, as compared to Table 2, comes from the fact that the inputs of imports into domestic production in the Chinese input-output matrix are quite low. Hence there is only a slight impact on domestic output caused by the devaluation.

Would gradual privatization with up-front devaluation be more appropriate? Thus, instead of a 5 percent annual devaluation, assume there is an initial 30 percent devaluation. Table 5 gives the results of the simulation.

---

9 One might analyze the long-run sustainability of the current account by checking running simulations over a considerably longer time period than the six periods in this study.
Compared with Table 4, there is a small boost to real GDP, but, as expected, inflation is initially higher, but tapers off with the price levels under the two scenarios gradually converging. The budget deficits and interest rates do not change much. However, the current account position, at least in the initial periods, improves significantly, but worsens in the last period. Because of the higher price level and the unchanged real GDP, both rural and urban consumers end up being worse off than under the gradual devaluation scenario.

Let us now examine two possible combinations of immediate privatization—with a gradual devaluation and with an up-front devaluation. Table 6 gives the results of a gradual devaluation with immediate privatization.

It is useful to compare Table 6 with Table 4. There is a relative increase in inflation but a relative fall in real GDP in the first two periods, reflecting the fall in public investment. In the last four periods private productivity catches up, reflected in a higher real GDP level, and a dampening in inflation. The budget improves initially, but starts deteriorating, due to the increase in the nominal interest rate. The real interest rate rises as private investment increases, and the current account deteriorates as private consumption also rises. Both rural and urban consumers are better off than under the gradual privatization scenario.

Let us turn to an alternative policy path. Consider a one-step devaluation at the beginning, together with an immediate privatization. The outcomes are given in
Table 7. China: Immediate Privatization Plus 30 Percent Initial Devaluation

<table>
<thead>
<tr>
<th>Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price level</td>
<td>129.8</td>
<td>156.7</td>
<td>155.6</td>
<td>190.6</td>
<td>184.9</td>
<td>226.0</td>
</tr>
<tr>
<td>Real GDP</td>
<td>96.2</td>
<td>103.0</td>
<td>117.4</td>
<td>124.9</td>
<td>138.9</td>
<td>147.5</td>
</tr>
<tr>
<td>Budget deficit (in percent of GDP)</td>
<td>3.1</td>
<td>1.7</td>
<td>-2.2</td>
<td>-2.4</td>
<td>-6.1</td>
<td>-5.5</td>
</tr>
<tr>
<td>Interest rate</td>
<td>12.5</td>
<td>14.3</td>
<td>13.3</td>
<td>19.2</td>
<td>14.2</td>
<td>24.0</td>
</tr>
<tr>
<td>Trade balance (in percent of GDP)</td>
<td>12.6</td>
<td>11.0</td>
<td>7.3</td>
<td>5.8</td>
<td>3.9</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Utility of consumer 1 = 105.9, utility of consumer 2 = 114.2

Source: Authors’ simulation results.

Table 8. China: Gradual Tariff Reform

<table>
<thead>
<tr>
<th>Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price level</td>
<td>100</td>
<td>111.6</td>
<td>123.9</td>
<td>137.8</td>
<td>147.6</td>
<td>178.4</td>
</tr>
<tr>
<td>Real GDP</td>
<td>100</td>
<td>107.6</td>
<td>114.5</td>
<td>123.8</td>
<td>131.2</td>
<td>140.2</td>
</tr>
<tr>
<td>Budget (in percent of GDP)</td>
<td>0.7</td>
<td>-0.2</td>
<td>-3.2</td>
<td>-3.4</td>
<td>-6.6</td>
<td>-6.4</td>
</tr>
<tr>
<td>Interest rate</td>
<td>1.7</td>
<td>3.4</td>
<td>4.6</td>
<td>11.2</td>
<td>9.3</td>
<td>15.2</td>
</tr>
<tr>
<td>Trade balance (in percent of GDP)</td>
<td>11.1</td>
<td>11.7</td>
<td>7.3</td>
<td>8.3</td>
<td>5.0</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Utility of consumer 1 = 99.1, utility of consumer 2 = 100.6

Source: Authors’ simulation results.

Table 7. Compared with the previous scenario (Table 6), we see that the price level is generally higher and real GDP and the budget do not change significantly. Real interest rates are lower, as the devaluation has reduced private investment, thereby reducing borrowing requirements. In addition, the current account surplus improves marginally. The higher price levels, however, are reflected in lower welfare for both consumers.

**Tariff Reform**

Two simulations regarding alternate tariff reform paths are carried out, involving a gradual and an up-front elimination of tariffs. The first simulation (Table 8) supposes that tariff reform is introduced gradually. Assume that, in the first two periods, tariff rates stay at their historical levels. In the remaining four periods, they are reduced by 20, 40, 70, and 100 percent of their initial values. Hence by period 6 they are at 0 percent. Table 8 gives the results of this simulation.

The second simulation assumes the elimination of tariff rates in the first period. The results are given in Table 9.

The outcomes in both simulations are essentially the same as those in Table 1. These suggest that tariff reform, taken alone, appears to have little impact, whether done gradually or in one step. We should, however, qualify our results. The effective
average tariff rate that we have estimated is only 2.7 percent in period 1. Hence the elimination of tariffs would have relatively little impact, at least initially, on prices. At the same time, the coefficients of imports in the Chinese input-output matrix are quite small and, in fact, imports are used as inputs to production in only six sectors. Accordingly, there is little linkage between imports and domestic production.10

The Two Extremes

In this section we consider two cases involving several policy instruments. In both, simulations, privatization, tariff reform, and devaluation are undertaken, with the only difference being in the speed with which these actions are taken.

Table 10 gives the results of a “big-bang” approach involving an up-front full elimination of tariffs, full privatization, and a 30 percent devaluation.

These results provide an interesting contrast to the baseline scenario (Table 1) and give an indication of how the addition of tariff reform in a package affects

---

10Trade barriers in China are incorporated as nontariff barriers rather than as high tariff rates. Hence trade liberalization should really be studied as a reduction in quantitative restrictions. Such simulations are, however, beyond the scope of our current study.
welfare (compared to Table 6). First, compared with the baseline scenario, real GDP is lower in the two first periods, but then rises. The price level is higher throughout. After improving, the current account position deteriorates, as the once-and-for-all effect of the devaluation is gradually eroded. Overall, both consumers are better off, benefiting from the reform package. Second, the welfare effect of up-front tariff reform combined with other policies is somewhat greater than the up-front tariff reform alone.

Table 11 gives the results of a gradual approach to a reform package, involving gradual privatization, tariff reform, and devaluation phased in the same manner as in earlier simulations.

Compared with the big-bang approach, this table indicates that gradualism, although resulting in less of a contraction in real GDP in the first two periods, yields lower real GDP levels in the subsequent periods. Partly because of that, both consumers are distinctly less well off in terms of their welfare than under the big-bang approach. In fact, the gradual approach results in minor welfare improvements relative to the baseline scenario only to the urban consumer.

V. Conclusion

The results of the simulations (summarized in Table 12) illustrate the complexities of the issues involved in deciding on the speed of adjustment and sequencing of reforms. Much depends on the objectives being sought, the time frame, and the sustainability of the macroeconomic situation. Nonetheless, certain conclusions can be drawn from the simulations.

In looking at complete policy packages, the big-bang approach is better from a welfare point of view: both consumers are better off under a package where adjustment and reform policies reinforce each other. Although under the big-bang approach the drop in real GDP is initially greater than under the gradual approach, real GDP rises to higher levels in subsequent periods. However, the current account position remains better under the gradual approach partly because the big-bang approach generates worse budgetary outcomes and higher nominal interest rates for most of the period.
Table 12. China: Summary Table of Simulations1

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Gradual Privatization</th>
<th>Immediate Privatization and Devaluation</th>
<th>Gradual Privatization and Up-Front Devaluation</th>
<th>Immediate Privatization and Devaluation</th>
<th>Immediate Tariff Reform</th>
<th>Immediate Tariff Reform</th>
<th>Gradual Big Bang</th>
<th>Gradual Gradual</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>100</td>
<td>102.3</td>
<td>108.5</td>
<td>101.3</td>
<td>100.0</td>
<td>105.9</td>
<td>106.9</td>
<td>101.2</td>
<td>99.1</td>
</tr>
<tr>
<td>$U_2$</td>
<td>100</td>
<td>90.5</td>
<td>137.3</td>
<td>99.1</td>
<td>89.6</td>
<td>114.2</td>
<td>123.1</td>
<td>101.3</td>
<td>100.6</td>
</tr>
<tr>
<td>Real GDP</td>
<td>140.3</td>
<td>142.2</td>
<td>146.7</td>
<td>143.6</td>
<td>142.7</td>
<td>147.5</td>
<td>148.2</td>
<td>140.4</td>
<td>140.2</td>
</tr>
<tr>
<td>Price level</td>
<td>177.6</td>
<td>187.6</td>
<td>188.6</td>
<td>215.5</td>
<td>222.7</td>
<td>226.0</td>
<td>216.6</td>
<td>179.0</td>
<td>178.4</td>
</tr>
<tr>
<td>Inflation</td>
<td>20.5</td>
<td>21.9</td>
<td>21.5</td>
<td>23.6</td>
<td>22.7</td>
<td>22.2</td>
<td>23.7</td>
<td>23.5</td>
<td>20.1</td>
</tr>
<tr>
<td>Budget</td>
<td>-6.0</td>
<td>-3.9</td>
<td>-5.3</td>
<td>-4.0</td>
<td>-5.5</td>
<td>-5.3</td>
<td>-6.4</td>
<td>-6.4</td>
<td>-6.0</td>
</tr>
<tr>
<td>Interest rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>14.8</td>
<td>20.9</td>
<td>27.7</td>
<td>18.9</td>
<td>17.8</td>
<td>24.0</td>
<td>25.0</td>
<td>16.0</td>
<td>15.2</td>
</tr>
<tr>
<td>Real</td>
<td>-5.7</td>
<td>-1.0</td>
<td>-6.5</td>
<td>-4.7</td>
<td>-4.9</td>
<td>1.8</td>
<td>1.3</td>
<td>7.5</td>
<td>-4.9</td>
</tr>
<tr>
<td>External current account</td>
<td>3.8</td>
<td>2.0</td>
<td>1.1</td>
<td>4.8</td>
<td>3.8</td>
<td>2.6</td>
<td>3.7</td>
<td>3.6</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Source: Authors' simulation results.

1Last period, except for $U_1$ and $U_2$, which refer to the utility of urban and rural consumers, respectively, over the periods simulated.
A piecemeal approach to reform may not only fail to improve overall welfare significantly but may reduce it. A gradual approach to privatization improves marginally the welfare of the urban consumer but leads to a sharp deterioration in the welfare of the rural consumer. Also, a gradual or immediate reduction in tariffs alone may not result in major welfare improvements. Careful sequencing can improve welfare, and improper sequencing can lead to welfare losses. An immediate privatization with a gradual devaluation helps improve welfare more than an immediate privatization and devaluation or a gradual privatization and devaluation.

The objective of improving the current account position over a set number of periods can lead to different results. For example, up-front privatization alone results in a lower current account position in the last period than an overall gradual package, but maximizes the welfare of both sets of consumers. The catch, of course, is that the welfare gains may not be sustainable as the external current account deteriorates further in periods beyond the simulated time frame.

APPENDIX

Consumption

Here, and in what follows, we will use \( x \) to denote a demand variable and \( y \) to denote a supply variable. In order to avoid unreadable subscripts, let us let 1 refer to period \( i \) and 2 refer to period \( i+1 \). The consumer’s maximization problem is thus:

\[
\text{max } U(x), \quad x = (x_1, x_{L1}, x_{L1}, x_2, x_{L2}, x_{L2}).
\]

such that:

\[
(1 + t_i)P_i x_i + P_{Lu} x_{Lu} + P_{Lr} x_{Lr} + P_{Ml} x_{Ml} + P_{Bl} x_{Bl} + e_i P_{BF} x_{BF} = C_i
\]

\[
P_{K2}(1 - \delta)K_0 + P_{A2}A_0 + P_{Lu2}L_{u2} + P_{Lr2}L_{r2} + P_{M2}M_0 + r_2 x_{Bl} + e_2 P_{BF2} x_{BF2} + TR_2 = N_2
\]

\[
P_{K1}K_0 + P_{A1}A_0 + P_{Lu1}L_{u1} + P_{Lr1}L_{r1} + P_{M1}M_0 + r_0 B_0 + P_{Bl} B_0 + e_1 P_{BF} B_{BF} + TR_1 = N_1
\]

\[
C_i = N_i,
\]

\[
\log P_{Bi} x_{Bi} - \log e_i P_{BF} x_{BF} = \alpha + \beta \left( \log r_i - \log \frac{e_i}{e_{fi}} \right),
\]

\[
\log P_{Mi} x_{Mi} = a + b \log \left( 1 + t_i \right) x_i - e \log \pi_i,
\]

\[
P_{B2} x_{B2} = d_0 + d_1 \left( 1 + t_2 \right) P_{B2} x_{B2} + d_2 \frac{r_2 - \pi_0}{1 + \pi_2},
\]

where:

\( P_i \) = price vector of consumption goods in period \( i \).

\( x_i \) = vector of consumption in period \( i \).

\( C_i \) = value of aggregate consumption in period \( i \) (including purchases of financial assets).

\( N_i \) = aggregate income in period \( i \) (including potential income from the sale of real and financial assets).

\( t_i \) = vector of sales tax rates in period \( i \).

\( P_{Lu} \) = price of urban labor in period \( i \).

\( L_{u1} \) = allocation of total labor to urban labor in period \( i \).

\( x_{L1} \) = demand for urban leisure in period \( i \).
The consumer's saving rate for period \( t \) is determined by intertemporal maximization, as well as of financial assets. The next two equations contain the value of the consumer's holdings of capital and labor, as well as the principal and interest that he receives from the domestic and foreign financial assets that he held at the end of the previous period. The equation \( C_t = N_t \) then imposes a budget constraint in each period.

Equation (A3) says that the proportion of savings made up of domestic and foreign interest-bearing assets depends on relative domestic and foreign interest rates, deflated by change in the exchange rate. Equation (A4) is a standard money-demand equation in which the demand for cash balances depends on the domestic rate of inflation and the value of intended consumption.

In period 2 we impose a savings rate based on adoptive expectations, as in equation (A5). The constants \((d_i)\) are estimated by a simple regression analysis, based on the previous periods. Thus if we are in period \( t \), where \( t \) is the end of a two-period segment, then the closure saving rate for period \( t \) is determined by nominal income and the real interest rate. The constants are updated after each two-period segment by running a regression on the previous \( t - 2 \) periods. Thus savings rates are endogenously determined by intertemporal maximization in period \( t \), but are determined by adoptive expectations in period \( t + 1 \). Accordingly, equation (A5) is the terminal condition for the consumer's problem. Combined with the closure rule for investment, described in equation (3), this determines the terminal conditions for our problem.

\[ P_{L_{ti}} \] price of rural labor in period \( i \).
\[ L_{ni} \] allocation of total labor to rural labor in period \( i \).
\[ x_{L_{ti}} \] demand for rural leisure in period \( i \).
\[ a_2 \] elasticity of rural/urban migration.
\[ P_{K_{ti}} \] price of capital in period \( i \).
\[ K_0 \] initial holding of capital.
\[ P_{R_{ti}} \] price of land in period \( i \).
\[ A_0 \] initial holding of land.
\[ \delta \] rate of depreciation of capital.
\[ P_{M_{ti}} \] price of money in period \( i \). Money in period 1 is the numerator and hence has a price of 1.
\[ x_{M_{ti}} \] holdings of money in period \( i \).
\[ P_{B_{ti}} \] discount price of a certificate of deposit in period \( i \).
\[ \pi_t \] domestic rate of inflation in period \( i \).
\[ r_i, r_{Fi} \] domestic and foreign interest rates in period \( i \).
\[ x_{B_{ti}} \] quantity of bank deposits, that is, CDs in period \( i \).
\[ e_i \] exchange rate in terms of units of domestic currency per unit of foreign currency in period \( i \).
\[ x_{B_{Fi}} \] quantity of foreign currency held in period \( i \).
\[ TR_i \] transfer payments from the government in period \( i \).
\[ d_i \] constants estimated from model simulations.

The left-hand side of equation (A2) represents the value of consumption of goods and leisure, as well as of financial assets. The next two equations contain the value of the consumer's holdings of capital and labor, as well as the principal and interest that he receives from the domestic and foreign financial assets that he held at the end of the previous period. The equation \( C_t = N_t \) then imposes a budget constraint in each period.

Equation (A3) says that the proportion of savings made up of domestic and foreign interest-bearing assets depends on relative domestic and foreign interest rates, deflated by change in the exchange rate. Equation (A4) is a standard money-demand equation in which the demand for cash balances depends on the domestic rate of inflation and the value of intended consumption.

In period 2 we impose a savings rate based on adoptive expectations, as in equation (A5). The constants \((d_i)\) are estimated by a simple regression analysis, based on the previous periods. Thus if we are in period \( t \), where \( t \) is the end of a two-period segment, then the closure saving rate for period \( t \) is determined by nominal income and the real interest rate. The constants are updated after each two-period segment by running a regression on the previous \( t - 2 \) periods. Thus savings rates are endogenously determined by intertemporal maximization in period \( t \), but are determined by adoptive expectations in period \( t + 1 \). Accordingly, equation (A5) is the terminal condition for the consumer's problem. Combined with the closure rule for investment, described in equation (3), this determines the terminal conditions for our problem.

\[ ^{11}\text{Since the only information the consumer has about the future is the real interest rate, adoptive expectations is, in this case, equivalent to rational expectations.} \]
REFERENCES


Structural Vulnerabilities and Currency Crises

SWATI R. GHOSH and ATISH R. GHOSH

This paper examines the role of structural factors—governance and rule of law, corporate sector governance (creditor rights and shareholder rights), corporate financing structure—as well as macroeconomic variables in currency crises. Using a technique known as a binary recursive tree allows for interactions between the various explanatory variables. It is found that structural vulnerabilities play an important role in the occurrence of “deep” currency crises (those with a real GDP growth decline of at least 3 percentage points) and that there are complex interactions between these structural vulnerabilities and macroeconomic imbalances. [JEL F31, F41, F47]

There is a growing body of literature that seeks to identify, or even predict, circumstances under which countries may suffer balance of payments crises. Much of this literature, inspired by the theoretical models of Krugman (1979) and Flood and Garber (1984), emphasizes the role of macroeconomic imbalances—large fiscal deficits or excessive rates of credit expansion—as the underlying cause of currency crises (while the proximate triggers may be contagion effects or imprudently low levels of foreign exchange reserves).

Yet the Asian crisis countries, in particular, do not readily fit this mold. Exchange rates in these countries were not especially overvalued, fiscal deficits were small, and macroeconomic performance had generally been exemplary. Rather, structural weaknesses in the corporate and financial sectors appear to have been at play. This paper seeks to complement much of the existing literature on currency crises by examining the role of structural factors and vulnerabilities.1

1Swati R. Ghosh is at the World Bank and Atish R. Ghosh is at the IMF.

In recent, parallel work, Mulder, Perrelli, and Rocha (2001) examine the role of structural factors in currency crises using a probit framework; see also Stone and Weeks (2001).
Swati R. Ghosh and Atish R. Ghosh

At least in East Asia, weak corporate and public sector governance appears to have encouraged an environment of excessive risk taking by the corporate and financial sectors, resulting in highly vulnerable corporate financing structures, with too much reliance on debt rather than equity issuance, and a large fraction of short-term rather than long-term borrowing. But such vulnerable corporate financing structures, while a feature of the East Asian experience, are by no means unique to it and may have been at play in other currency crises as well.

Identifying such structural determinants of currency crises is important for at least two reasons. First, inasmuch as these weaknesses, like macroeconomic imbalances, contribute to the probability of a currency crisis, eliminating them is clearly a priority. Second, if structural factors are at play, then faced by a such crisis, announcing and implementing structural reforms may be crucial in restoring confidence to the markets.

We examine the role of corporate sector vulnerabilities in currency crises using a panel dataset covering some 40 industrialized and emerging market countries, over the period 1987–1999. Our list of crises is taken from Glick and Hutchison (1999), except that, for the bulk of our analysis, we focus on “deep” currency crises—that is, those in which there was an appreciable decline in real GDP growth. In addition to the usual macroeconomic suspects, we consider four broad categories of structural indicators. The first pertains to what might be termed the country’s overall “rule of law,” including ratings on public sector corruption, risk of government expropriation or contract repudiation, and efficiency of the judicial system and legal and accounting standards. The second and third categories concern corporate governance directly, and pertain to the rights and responsibilities of shareholders and creditors, respectively. Weak corporate governance, resulting, in part, from inadequate shareholder and creditor rights, may be manifested in a risky financing structure of corporations (e.g., with an overreliance on short-term debt). As a final category, therefore, we also include corporate debt-equity ratios and maturity structure of debt (medians for a sample of firms in each country).

Much of the literature on currency crises to date has used probit analysis to relate the probability of a crisis to a vector of explanatory variables. Such probit models underlie most of the “early warning systems” for currency crises being implemented in both the private and public sectors. Although the precise explana-

---

2In addition, the lack of deep domestic bond markets in many of these countries contributed to a reliance on bank borrowing as a source of finance.

3For instance, in the sample below, in addition to East Asia, other countries in which the corporate sector had a relatively high proportion of short-term debt include Turkey, Hungary, Greece, Poland, and Peru, while debt-equity ratios were high in the Nordic countries.

4Glick and Hutchison (1999) define a currency crisis as a “large” change (more than two standard deviations from the country-specific mean for the country) of an index of currency pressure, defined as the weighted average of monthly real exchange rate changes and monthly (percent) reserve losses. To focus on cases in which there is an appreciable decline in real GDP growth, we use two cutoffs: those currency crises involving a decline in real GDP growth of at least 3 percentage points, and those involving a decline of at least 5 percentage points (relative to the previous five years).

5Note that some elements of the “rule of law” such as accounting standards and efficiency of the judicial system also have a direct bearing on the exercise of corporate governance.

6See, for instance, the proceedings of the Economic Forum on Early Warning Systems, hosted by the IMF on November 1, 2001 (reported in the IMF Survey, November 12, 2001).
tory variables differ across models, they normally include indicators of macroeconomic imbalances—current account deficits, real exchange rate overvaluation, rapid rates of credit growth, and budget deficits—and, in some recent studies, various indicators of structural vulnerability as well.

These probits give the marginal effect on the probability of a crisis of each of the explanatory variables, holding the others constant at their mean values. While this “other things being equal” (ceteris paribus) assumption is common in economics, it is not the most natural assumption to make when assessing the risk of an event because it does not readily allow for interactions between the various explanatory variables: indeed, in many other contexts, the ceteris paribus assumption would be considered quite odd.

Take, for instance, a doctor diagnosing a patient’s risk of a heart attack, and suppose that both a history of heart problems in the immediate family (hereditary factors) and high (LDL) cholesterol levels are known to be contributory factors. The equivalent of a “probit” approach would be one in which the doctor considers the marginal effect of the patient’s cholesterol level, holding constant his family history at the (population) mean. But no doctor would do this. Rather, it would be much more natural to first ascertain whether there was any history of heart attacks among the patient’s relatives. If the answer was yes, then the “danger” level of cholesterol may be 130, and the patient’s cholesterol assessed in relation to this level. On the other hand, if the answer to a family history of heart attacks was no, a higher level of cholesterol, say 150, may be tolerable.7

In much the same vein, a country may be vulnerable to a crisis because of structural deficiencies but only suffer a currency crisis when macroeconomic imbalances become sufficiently severe. Such “context-dependence” is also reflected in the theoretical literature on currency crises. First-generation models emphasize the inconsistency of policies—governments intent on money-financing their deficit while trying to maintain a fixed exchange rate—as the underlying cause of the crisis. Second-generation models emphasize the cost of maintaining the pegged exchange rate regime, for instance in the face of high unemployment, which under certain circumstances (when the country is within a “zone of vulnerability”) can trigger a speculative attack. Third-generation models are built around potential structural vulnerabilities—especially foreign currency debt exposure of the corporate and financial sectors—leading to a self-fulfilling run on the currency. In any panel dataset, it is likely that each of these various generations (or variants thereof) is represented. If these are simply lumped together, factors that are important in determining one type of crisis may not be identified because they do not help explain the other types of crises.

As a methodological innovation of this paper, therefore, we go beyond standard probit analysis and use a decision-theoretic classification technique known as a binary recursive tree (BRT). This technique is particularly well suited to situations in which there may be “context dependence” and threshold effects. To the extent that there are different types of crises represented in the dataset, for instance, the classification tree can separate these and then examine the interaction

---

7The figures used here are purely illustrative.
of the various variables in determining a currency crisis. Such interactions could
be especially important here because structural factors typically do not change
very rapidly, so their ability to predict crises in a panel (or time series) context may
be limited. Rather, we would expect the interaction of relatively long-standing
structural vulnerabilities and high(er)-frequency movements in macroeconomic
variables to account for currency crises. For instance, the East Asian crisis coun-
tries had structural vulnerabilities for a number of years prior to the onset of the
crisis: there must have been a confluence of events—structural vulnerabilities,
macroeconomic imbalances, and perhaps contagion—to have actually triggered
the crisis. Not only does the binary recursive tree help identify factors that may
trigger (various types) of currency crisis, it can also be used to refine traditional
probit models.

Our main results may be summarized briefly. First, we confirm that macroe-
conomic imbalances, most notably a large current account deficit, are often the
proximate trigger of a crisis. Second, we find that a weak “rule of law” may make
countries particularly vulnerable to the effects of macroeconomic imbalances.
Third, a risky corporate finance structure—high debt-equity ratios and short mu-
turity of corporate debt—is an important determinant of currency crises. When these
debt-equity ratios and maturity composition of corporate debt are included, the
indicators of shareholder and creditor rights figure less prominently, suggesting
that the effect of the latter on the probability of a crisis is manifested mostly
through the financing structure of corporations. Finally, we find that the interac-
tion between structural vulnerabilities and macroeconomic imbalances in deter-
mining crises is often highly complex, highlighting the difficulties of undertaking
effective surveillance and monitoring of countries’ potential vulnerability to crises.

I. Corporate Governance and Structural Vulnerabilities

Although structural factors may have been at play in previous crises, it was the
Asian currency crises at end-1997 and 1998 that brought them to the fore.

A key hypothesis put forward in the context of the East Asian crisis is that the
corporate incentive structure encouraged a rapid pace of investment that was of
increasingly uncertain quality. The rapid pace of investment and, in some cases,
progressively lower returns on these investments made it necessary for firms to
seek financing outside of retained earnings. Given the corporate governance envi-
ronment and a traditional reluctance to dilute family shareholdings, this demand
for outside financing took the form of borrowing rather than equity issuance.

Corporate governance refers to the rules, standards, and organizations that
govern the behavior of corporate owners, directors, and managers and that define
their duties and accountabilities to outside investors (Prowse, 1998). It is thus a
key element in exercising discipline on firms and defining the overall incentive
framework for firms, and is therefore essential for efficient, productivity-driven
investments and safeguards against excessive risk taking.

8See, for example, Alba and others (2000).
9Competition in the product markets is another channel through which discipline is exercised on firms.
Mechanisms that facilitate good corporate governance may be grouped into those that govern the rights of (especially minority) shareholders, those that govern the rights of creditors, and those that facilitate enforcement of these rights as well as monitoring and disciplining.\(^{10}\)

Within the first category, there are measures that strengthen shareholders' rights in general, those that strengthen minority shareholders' abilities to exercise governance, and those that strengthen the rights of "strategic" investors. If the minimum percentage of ownership of share capital required to call an emergency shareholders' meeting is relatively low, for instance, this makes it easier for minority shareholders to organize a meeting to challenge or oust the management. (The percentage varies around the world from 1 percent of share capital in the United States to 33 percent in Mexico). Or, if proxy by mail is allowed, (any) shareholders' ability to exercise their voting rights is considerably facilitated. For strategic investors, the right to hostile takeovers may be an important disciplining mechanism.

Creditor rights are conceptually more complex because creditors exercise their power in several ways. Perhaps the most basic creditor right is the right to repossess and then liquidate—or keep—the collateral (La Porta and others, 1998). Creditor rights are strengthened if, for example, the bankruptcy or reorganization laws stipulate restrictions on reorganization, such as the need for creditors' consent to file for reorganization; or if secured creditors are ranked first in the distribution of the proceeds that result from the disposition of assets of a bankrupt firm. Also important, however, is the incentive structure of creditors or financial institutions themselves, which is shaped not only by their own corporate governance, prudential norms, and regulatory and supervisory framework, but also by the perception of implicit government guarantees. Strong creditor rights without the corresponding good governance of financial institutions will still result in weak disciplining of corporations.

Finally, there is a set of rules and regulations that facilitate monitoring and disciplining, including the legal framework (enforcement and insolvency/bankruptcy or exit mechanisms), accounting standards, transparency, and disclosure, etc.

Weak corporate governance is often reflected in a divergence between "control rights" and "cash-flow rights,"\(^{11}\) in turn, encouraging excessive risk taking in investments. Through control-enhancing mechanisms such as pyramiding,\(^{12}\) cross-holdings, or having a chief executive officer, board chairman, or vice chairman related to the controlling family, cash-flow rights can deviate

---

10 Until recently, the literature on corporate governance focused on the potential conflicts between shareholders and managers. In East Asia, however, major shareholders and managers have tended to be one and the same, giving them almost complete discretionary power to commit company resources. In the East Asian context, therefore, the issue has been of having a few inside owners/managers on the one hand, and outside minority shareholders and financiers on the other. The contention, therefore, is that the mechanisms to exercise corporate governance over the insider shareholders were weak in the region.

11 "Cash-flow rights" refers to the claim on the profits of the corporation associated with the ownership of the stock; however, under certain circumstances, a group owning less than a majority of the cash-flow rights may, nevertheless, exercise "control rights." For instance by stock pyramids, cross-ownership structures, and dual class equity structures.

12 Pyramiding is defined as owning a majority of the stock of one corporation that, in turn, owns a majority stock of another—a process that can be repeated a number of times.
substantially from control rights. Such deviations, if significant, can provide incentives for greater risk taking (both in the nature of investments and in their volume and pace), as corporate owners have less to lose if the project goes wrong (since their cash-flow stake is relatively small), but can benefit if the project is successful—because, through their effective control, they can more easily expropriate the gains. Particularly in East Asia, these perverse incentives were exacerbated because management was generally not separated from ownership control. The combination of concentrated family control rights that exceeded cash-flow rights, and close control of management by family owners, provided corporate owners both greater incentives for risk taking and the means for effecting this.

Excessive risk taking, in turn, can result in a fast pace of investment (often with progressively declining rates of return), necessitating financing outside of firms’ retained earnings. This financing often takes the form of debt, because of the incentive structure of debt, where default allows the borrower to limit the downside risk (particularly in countries where creditors’ recourse to bankruptcy proceedings is limited), while capturing the gains if the project is successful. Beyond this, however, particularly in the East Asian context, there is often a general reluctance by family owners to dilute their share of ownership as well.

Greater integration of the world capital markets allows for easier access to foreign borrowing by domestic corporations—be it directly or indirectly through the intermediation of domestic financial institutions—so that a sizable proportion of this borrowing may be external. Moreover, the macroeconomic policy mix used to deal with capital inflows and attendant macroeconomic overheating, can itself further encourage unhedged short-term borrowing, exacerbating the accumulation of large short-term external liabilities. At the micro level, this can result in highly leveraged corporations and sizable currency and maturity mismatches in the balance sheets of both corporations and financial institutions.

---

14 The macroeconomic policy mix in the East Asian crisis countries tended to encourage the accumulation of external liabilities, mainly on a short-term and unhedged basis. Macroeconomic overheating and capital inflows tended to move in tandem (e.g., Indonesia in 1990 and 1994–96, Korea in 1994–95). In dealing with the overheating, countries often relied on monetary policy as the primary instrument, which reinforced the upward movement of domestic interest rates and provided further incentives to borrow abroad. Since short-term capital flows tend to be most responsive to interest rate differentials, this accumulation was primarily in the form of short-term liabilities. Moreover, given the managed exchange rate system in these countries, the buildup of net foreign assets of the central banks was largely sterilized in order to limit the growth of reserve money and maintain monetary aggregates, which meant that the large interest rate differentials were sustained. And, although the East Asian countries’ fiscal position was sound—entailing low levels of government debt and high savings—the fiscal impulse tended to be positive at a time when domestic demand pressures picked up. Thus the fiscal position added to aggregate demand and interest rate pressures. Finally, their exchange rate policy reduced incentives to hedge the external borrowing. Although, in principle, all countries adopted some form of managed exchange rates (which—to differing degrees across countries—were allowed to fluctuate within bands), in practice, the nominal exchange rates did not tend to vary much, and the exchange rate policy resulted in relatively predictable nominal exchange rates.
II. Methodology

The standard approach in the currency crisis literature is to estimate a probit of the occurrence of a crisis on a set of explanatory variables. Such an approach has the benefit of being familiar, with well-known statistical properties, and of being able to isolate the marginal effect of each individual explanatory variable, holding the others constant at their mean values. This “other things being equal” (or “partial derivative”) assumption is such a standard part of an economist’s toolkit that it is seldom questioned. As noted in the introduction, however, it is not the only approach, and not necessarily the best approach for analyzing crises.

In particular, standard economic analysis implicitly assumes some continuity in the relationships between economic variables. That is, if increasing variable \( x \) elicits a certain response in variable \( y \), then doubling the increase in \( x \) should induce a correspondingly large response in \( y \). Currency crises differ in that they are fundamentally discontinuous: that is, they represent a confluence of factors that trigger a discrete event (the crisis), but only once certain thresholds have been crossed. For instance, in Indonesia and Korea, there were long-standing weaknesses in the financial (and, in Korea, the corporate) sector. It required a particular confluence of events—terms of trade shocks, contagion, and political uncertainty—interacting with these weaknesses to trigger the crisis.

Accordingly, analyzing crises requires a technique that allows both for thresholds in the effects of an independent variable on the probability of a crisis and, moreover, for the thresholds themselves to depend on interactions between the variables.

In principle, it would be possible to capture such interactions within a probit framework by including sufficiently many interactive dummy variables—for instance, estimating the probit with an interaction term between corporate governance indicators and the current account deficit. When there are several explanatory variables, and if they are continuous, however, such an approach soon becomes impractical.

Fortunately, more systematic methods are available. One such technique is known as a binary recursive tree (BRT). Formally, it is a sequence of rules for predicting a binary variable, \( y \), on the basis of a vector of explanatory variables, \( x_j \), . At each branch of the tree, the sample is split according to some threshold value, \( x_j \), of one of the explanatory variables into two sub-branches. The splitting is repeated along the various sub-branches until a terminal node is reached.

To illustrate, let \( y \) be the event of a crisis (equal to 1 if there is a crisis, and 0 otherwise). The sample is randomly separated into a core sample and a smaller test sample, which is used for “out-of-sample” robustness checks. For the core sample, the algorithm searches for sequential splits, each consisting of the explanatory variable and its threshold value, which best discriminates between the groups.

---

15 For probit or regression-based studies of currency crises, see Frankel and Rose (1996); Eichengreen, Rose, and Wyplosz (1996); Sachs, Tornell, and Velasco (1996); Tornell (1999); Kaminsky and Reinhart (1999); and Berg and Pattillo (1999).

16 In the analogy of heart attack risks, the danger threshold for cholesterol level might itself depend on the patient’s family history.

17 See Breiman and others (1984); implementation of the BRT was undertaken using Salford System’s CART program.
Suppose, for example, that a large current account deficit is associated with currency crises and is thus a potentially useful discriminator variable. There will, however, be countries that have a small current account deficit yet suffer a currency crisis (a type I error), and others that have a large current account deficit but (nonetheless) do not have a crisis (a type II error). The algorithm searches over all observed values of the current account deficit in the sample until it finds that threshold value, $x_1$, which best discriminates between crisis and noncrisis countries in the sense of minimizing the sum of the type I and type II errors.18

The minimum sum of errors provides a natural gauge of the ability of the current account deficit variable to predict crises. The same procedure is applied to each of the explanatory variables; then, sorting these variables by their minimum error scores provides a ranking of their ability to discriminate between crisis and noncrisis countries. (To check robustness, the threshold value for each variable is also applied to the test sample, yielding a second error score.) The variable (together with its associated threshold value, $x_j$) that has the lowest error score is used to form the first node of the decision tree. All observations with a current account deficit less than $x_1$ are classified on the left sub-branch of the tree; all observations with a current account deficit greater than $x_1$ go to the right.

For each sub-branch, the algorithm is repeated; once the initial data is partitioned into two subsamples, each part of the tree is analyzed separately, so that the discovery pattern becomes progressively more local. Thus the methodology is very good at discovering local—context-dependent—data structures. In principle, the process of progressive subdivision could continue until every observation has been placed into its own branch. This would be akin to including as many explanatory variables as observations in a standard regression and thus getting a “perfect,” if meaningless, fit. Some termination rule is required. The rule used is roughly the same as an adjusted $R^2$ rule. After each split, the improvement in the overall fit (which, just like the change in the raw $R^2$ on adding another variable in a regression, is always non-negative) is combined with a penalty on the number of branches, which promotes parsimony. If the penalty exceeds the improvement, the branch is terminated at the prior node; otherwise, the branching continues.

Several aspects of the algorithm are noteworthy. First, the algorithm automatically establishes orderings among explanatory variables both globally (toward the top of the tree) and locally (along each of the various sub-branches). Although an explanatory variable that appears toward the top of the tree is more “important” in discriminating between crisis and noncrisis countries, an explanatory variable may appear several times along various sub-branches. To return to the heart attack example, if the critical levels of cholesterol differ across men and women, one branching of the tree might split the sample according to gender, then along each sub-branch the level of cholesterol might be the next discriminator (albeit at different threshold levels). Second, by its very nature, the algorithm captures interactions between explanatory variables. Third, the algorithm is good at capturing threshold effects, which may be particularly important in looking at the effects of structural variables. By the same token, however, if the effect is truly continuous, the algorithm simply

---

18It is also possible to weight the type I and type II errors in the loss function if, for example, it is judged to be more costly to miss predicting a crisis than it is to call a false positive.
tinds the value that best discriminates between crisis and noncrisis countries. For example, if the probability of a crisis increases linearly in the current account deficit, the algorithm would still try to find the best “threshold” value for discriminating between crisis and noncrisis countries. (This is relatively easy to detect, however, because, when the effects are continuous, they tend to show up by repeated branchings by the same explanatory variable along the same branch.19) Fifth, the procedure is very robust to outliers since it splits on an interior threshold (rather like using medians instead of means). Finally, the decision tree is invariant to any monotone transformation of the variables. Again, this is a very important property when looking at structural variables, several of which are rank indexes.

But the methodology is also not without its own limitations. Most importantly, the statistical properties are not yet well known, and formal statistical tests are not available. As such, the only way to assess the model is in terms of its ability to predict crises (more exactly, the likelihood that the model makes either a type I or type II error). Second, as noted above, the procedure is less well suited when the effects are genuinely continuous. Third, at each branch, the procedure picks out the explanatory variable that best discriminates between crisis and noncrisis countries: this is not to suggest, however, that others may not be important (i.e., beaten by only a small margin).20 Fourth, toward the lower branches of the tree, the number of crisis cases may become very small, sometimes leading to counterintuitive results,21 though this can be avoided by more stringent stopping or pruning rules to limit the number of sub-branches.

In our view, these limitations do not preclude the usefulness of this technique, at least as a complement to the more standard probit/regression analysis. As discussed below, the resulting decision trees require careful interpretation but, if nothing else, they make clear that currency crises occur as a result of a complex confluence of factors—an insight that is perhaps lost in the simplicity of the standard probit output. Indeed, the binary recursive tree methodology can be used to refine the probit analysis.

III. Macroeconomic and Structural Data

Our dataset covers 42 industrialized and mainly emerging market countries over the period 1987–99; with missing data, there are 624 observations.22 There are 52 currency crises, of which 19 involve a fall in real GDP growth of at least 3 percentage points, and 14 involve a fall in real GDP growth of at least 5 percentage points.23

19 For instance, a branch would first divide on whether the current account deficit was greater than 2 percent of GDP, and then (at least one) of the sub-branches would divide on whether the current account deficit was greater than 3 percent of GDP, etc.

20 In a standard regression or probit, multicollinearity may imply that individual r-statistics are insignificant; a binary recursive tree, however, simply picks the variable that best discriminates between crisis and noncrisis countries from the vector of explanatory variables (if any at all).

21 For instance, if there is only one crisis observation remaining, and it happens, e.g., to be a country with a current account surplus (and the only country remaining at that node with a current account surplus), the algorithm will—at that node—pick out a current surplus as being a determinant of the country having a crisis.

22 Some of the “rule of law” data were extended from 1995 assuming that they have been constant.

23 The swing is defined as $\frac{g(t)-g_{t-1}}{1+g_{t-1}}$, where g is the growth rate of real GDP.
The traditional currency crisis literature has suggested a smorgasbord of both macroeconomic policy and performance indicators (in addition to "vulnerability indicators" such as the external debt ratio or the level of foreign exchange reserves). Following this literature, but with a view to parsimony, we select five "macroeconomic" indicators: (i) the percentage real exchange rate appreciation over the previous three years (i.e., $t-3$ to $t-1$); (ii) the current account balance as a ratio to GDP, averaged over the previous three years; (iii) the central government balance as a ratio to GDP, averaged over the previous three years; (iv) the growth of the ratio of banking system credit to GDP, averaged over the previous three years; and (v), the ratio of total external debt to reserves. While not exhaustive, this set captures most of the variables that have been identified in the literature as relatively robust predictors of currency crises: external vulnerability, fiscal laxitude, and excessive rates of credit growth.

As noted above, for our structural variables, we include four broad categories, each with a number of separate indicators.

The first category consists of six indicators pertaining to the country's rule of law. These concern both broad governance issues—corruption and property rights (such as risk of expropriation or contraction repudiation by the government, efficiency of judicial system)—and those more narrowly related to the corporate and financial sector such as accounting standards.

The second category consists of eight indicators related to shareholders' rights. These include investor protection (such as whether ordinary shares carry one vote per share) as well as indicators of the ease with which investors can exercise their rights (whether proxy by mail is allowed; whether firms can block shares prior to a general stockholders-meeting; whether minority shareholders have a judicial venue to challenge the decisions of management; whether minority shareholders can name a proportional number of directors to the board). Of the two remaining indicators, one is a composite index of shareholders' rights vis-à-vis company directors, while the other is the percentage of mandatory dividend.

The third category consists of five indicators of creditors' rights. These include the legal requirement for a firm to seek its creditors' consent prior to filing for reorganization; the requirement that management not stay during the period of reorganization (with management in the hands of an official appointed by the court instead); the requirement that secured creditors be paid first in any bankruptcy proceedings; and legal reserve requirements (which can force automatic liquidation before all the capital is wasted or stolen).

Finally, we include the ratio of short-term to total corporate debt, and the ratio of debt to (common) equity for a sample of nonfinancial firms in each country, taken from the WORLDSCOPE database.

In general, within each category, the indicators tend to be correlated across countries. The correlations are greatest for the "rule of law" indicators, ranging...
from 0.6 to 0.9, with a single principal component capturing almost 80 percent of the total variation. The creditor rights variables are somewhat less correlated, but a single principal component captures more than 50 percent of the total variation, while the shareholder variables are the least correlated, with a single principal component capturing only 35 percent of the variation.

The correlation among the various indicators seems intuitive, since countries that have stronger creditor or shareholder rights along one measure are likely to have strong rights along other measures. But it also means that these indicators are subject to multicollinearity, and econometrically it may be difficult to isolate which among them matters (especially since the indicators are qualitative scores along arbitrary scales). Accordingly, in interpreting the results, if it is found that one or more of the indicators (or the first principal component) of a given category is significant, it is perhaps more useful to take this to mean that "shareholder rights" or "creditor rights" broadly construed may be important, and not just the individual indicator that happens to be significant.

There is also the correlation across categories. The first principal component of the shareholder rights category has a correlation of 0.29 (r-statistic: 3.29**) with the (first principal component of) the creditor rights category; it is rather less correlated with the "rule of law" category (correlation = 0.05). Stronger shareholder rights are also (negatively) correlated with higher debt-equity ratios or a larger fraction of short-term debt (with t-statistics of 3.67** and 6.13**). Again, this makes intuitive sense. Countries with strong shareholder rights are also likely to have strong creditor rights and, as a result of the better corporate governance and matching of cash flow and control rights, lower debt-equity ratios and a better maturity of debt. By the same token, however, to the extent that better shareholder and creditor rights affect the probability of a crisis through their effect on corporate debt-equity ratios and financing structure, they are unlikely to be significant in a probit or binary recursive tree once the short-term debt and debt-equity ratios are included directly.

### IV. Empirical Results

#### Probit Results

Since probit analysis is generally familiar, we begin by estimating standard probits for three dependent variables: (i) the occurrence of a balance of payments (BOP) crisis; (ii) a BOP crisis with at least a 3 percentage point growth swing; and (iii) a BOP crisis with at least a 5 percentage point growth swing.

We begin, in panel [1] of Table 1, with only the macroeconomic indicators. These are mostly consistent with intuition: a greater real exchange rate appreciation is associated with a higher probability of a crisis, while a larger current account

\[27\] This is not to suggest that there are not differences within a country between creditor rights and shareholder rights. Indeed, the corporate governance literature often emphasizes these differences in discussing, for instance, the rather different corporate control mechanisms between Germany (creditor-based) and the United States (shareholder-based). Nonetheless, in comparing very diverse countries, one might expect some to be stronger in terms of both shareholder and creditor rights, while others are weak in terms of their corporate governance more generally.
Table 1. Probit Estimation of Probability of a Currency Crisis with Various GDP Growth Swings

<table>
<thead>
<tr>
<th>Balance of Payments Crisis</th>
<th>Crisis with GDP Growth Swing of at Least 3 Percentage Points</th>
<th>Crisis with GDP Growth Swing of at Least 5 Percentage Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>t-stat</td>
<td>Coefficient</td>
</tr>
</tbody>
</table>

[1] Macroeconomic determinants only

| Constant       | -1.541 | -11.01*** | -2.059 | -9.49*** | -2.283 | -9.10*** |
| ΔREER          | 0.027  | 2.30**    | 0.035  | 2.06**   | 0.036  | 1.86*    |
| CAB/GDP        | -0.086 | -3.18***  | -0.119 | -2.79*** | -0.122 | -2.46**  |
| GovB/GDP       | 0.012  | 0.51      | 0.039  | 1.08     | 0.046  | 1.52*    |
| Δ(DC/GDP)      | 0.004  | 0.65      | 0.001  | 0.14     | 0.006  | 0.69     |
| Ext. Debt/Reserves | 0.000 | 0.97  | 0.000 | 0.32  | 0.000 | 0.89 |

Number of observations:
- positive observations: 536
- Loglikelihood: percent correct predicted -152.2 91.0 -67.7 96.0 -49.8 97.0


| Mcorp1 (one share, one vote) | -0.149 | -0.306 | 17.725 | 0.003 | 20.425 | 0.014 |
| Mcorp2 (proxy by mail allowed) | -1.089 | -0.886 | -18.059 | -0.001 | -33.372 | -0.009 |
| Mcorp3 (shares blocked before meeting) | 0.789 | 0.570 | 26.504 | 0.002 | 34.426 | 0.009 |
| Mcorp4 (cumulative voting rights) | -0.333 | -0.255 | -16.770 | -0.001 | -28.712 | -0.007 |
| Mcorp5 (oppressed minority) | -0.331 | -0.306 | -40.733 | -0.004 | -49.679 | -0.015 |
| Mcorp6 (percent share capital to call meeting) | 3.353 | 0.590 | 125.194 | 0.002 | 159.188 | 0.012 |
| Mcorp7 (anti-director rights) | 0.490 | 0.413 | 24.253 | 0.002 | 35.330 | 0.011 |
| Mcorp8 (mandatory dividend) | 0.849 | 0.658 | 31.068 | 0.001 | 7.901 | 0.001 |
| Mcred1 (restrictions on reorganization) | -0.007 | -0.029 | -7.913 | -0.007 | -9.754 | -0.018 |
| Mcred2 (automatic stay on assets) | 0.748 | 1.709* | -13.653 | -0.010 | -14.578 | -0.013 |
| Mcred3 (secured creditors paid first) | 1.004 | 1.824* | 14.885 | 0.001 | 8.357 | 0.003 |
| Mcred4 (management stays in reorganization) | -1.397 | -2.167** | 11.377 | 0.001 | 9.558 | 0.005 |
| Mrule1 (efficiency of judicial system) | -0.340 | -2.096** | -1.937 | -0.002 | -1.970 | -0.006 |
| Mrule2 (rule of law) | 0.161 | 0.685 | 2.668 | 0.000 | 7.114 | 0.006 |
| Mrule3 (corruption in government) | 0.003 | 0.019 | -0.843 | 0.000 | -2.096 | -0.004 |
| Mrule4 (risk of expropriation) | 0.532 | 1.200 | 9.863 | 0.004 | 4.010 | 0.004 |
| Mrule5 (risk of contract repudiation) | -0.545 | -1.366 | -13.077 | -0.006 | -10.912 | -0.009 |
| Mrule6 (accounting standards) | 0.015 | 0.673 | 0.421 | 0.001 | 0.120 | 0.002 |
| Corporate short-term debt ratio | -0.288 | -0.278 | -1.011 | -0.287 | -1.944 | -0.049 |
| Corporate debt-equity ratio | 0.338 | 1.514 | -1.028 | -1.348 | -1.105 | -1.436 |

Number of observations:
- positive observations: 420
- Loglikelihood: percent correct predicted -95.8 91.7 -27.5 96.4 -18.8 97.6
Table 1. (concluded)

<table>
<thead>
<tr>
<th>Balance of Payments Crisis</th>
<th>Crisis with GDP Growth Swing of at Least 3 Percentage Points</th>
<th>Crisis with GDP Growth Swing of at Least 5 Percentage Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3] Macroeconomic and structural determinalts (principal components)</td>
<td>Coefficient</td>
<td>t-stat</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.488</td>
<td>-5.35***</td>
</tr>
<tr>
<td>AREER</td>
<td>0.007</td>
<td>0.37</td>
</tr>
<tr>
<td>CAB/GDP</td>
<td>-0.101</td>
<td>-2.62***</td>
</tr>
<tr>
<td>GovB/GDP</td>
<td>-0.010</td>
<td>-0.28</td>
</tr>
<tr>
<td>Δ(DC/GDP)</td>
<td>0.029</td>
<td>3.52***</td>
</tr>
<tr>
<td>Ext. Debt/Reserves</td>
<td>0.000</td>
<td>1.20</td>
</tr>
<tr>
<td>Shareholder rights principal component</td>
<td>0.078</td>
<td>0.68</td>
</tr>
<tr>
<td>Creditor rights principal component</td>
<td>0.101</td>
<td>0.89</td>
</tr>
<tr>
<td>Rule of law principal component</td>
<td>-0.252</td>
<td>-1.86*</td>
</tr>
<tr>
<td>Corporate short-term debt ratio</td>
<td>0.130</td>
<td>0.19</td>
</tr>
<tr>
<td>Corporate debt-equity ratio</td>
<td>0.408</td>
<td>2.82***</td>
</tr>
</tbody>
</table>

Note: Asterisks denote significance at the 10 (**), 5 (***), and 1 (****) percent levels.
1 Average for years t, t-1, t-2, t-3.
2 End year t-1.

balance (i.e., smaller deficit) is associated with a lower probability of a crisis. A rapid expansion of banking system credit (relative to GDP) and a higher ratio of external debt-to-reserves are positively correlated with crises, but the coefficients are not statistically significant. Counterintuitively, a larger fiscal balance is also positively related to a crisis, although the coefficient is generally not significant.

Next, we turn to the structural factors, augmenting the probit with our structural indicators (Panel [2] of Table 1). While it is possible to obtain sensible parameter estimates for the "occurrence of crisis" dependent variable (column 1)—though most of the coefficients are insignificant—for the crises with growth swings of 3 or 5 percent, there are simply too many individual structural variables to be included simultaneously (the probit estimation does not converge).

One approach is to use the principal components of the three categories of structural indicators: corporate governance, creditor rights, and rule of law; the debt-equity and short-term debt ratios are included directly (Panel [3] of Table 1). Of the macroeconomic variables, the current account balance continues to be highly significant, while the real exchange rate appreciation loses its statistical significance.

The number of principal components to include for each category was chosen to ensure that at least 70 percent of the variation of the underlying indices is captured. Since principal components are orthogonal, excluding a principal component cannot affect the coefficients on those that are included.
The rate of credit growth in the economy, however, now becomes highly significant. Turning to the structural variables, the results are decidedly mixed. The principal components of better shareholder rights and better creditor rights are insignificant, while better rule of law is associated with a lower probability of a crisis. Finally, a higher proportion of corporate short-term debt or a higher debt-equity ratio is correlated with higher probabilities of a crisis (the latter being statistically significant).29

The statistical significance of the latter two variables suggests that they may be masking the effects of stronger shareholder and creditor rights (i.e., the effect of better shareholder and creditor rights on the probability of a crisis happens mostly through the corporate financing structure). Dropping the short-term debt and debt-equity variables confirms this in that the shareholder rights variable now has a negative, and statistically significant, coefficient.30

One interpretation of the results is that structural factors are not very important determinants of currency crises. In a sense, this should be none too surprising, particularly in the context of a panel dataset. Most structural variables change very slowly (some are constant for the country) so they have difficulty in explaining why a crisis occurs when it does, though they may do better at explaining where crises occur (i.e., in a purely, or largely, cross-country dataset). Put differently, Korean corporations have had high debt-equity ratios for a number of years—why did the crisis occur in 1997 and not in 1995? This leads to a second interpretation, however, namely that there may be important interactions between structural vulnerabilities and macroeconomic performance (which can change rapidly) that can explain currency crises. In principle, this can be done within the probit framework by interaction terms between the structural and macroeconomic variables. In practice, deciding which interactions to include in the probit estimation is difficult because there are potentially many. Therefore, we turn next to an approach that allows for such interactions more systematically.

**Binary Recursive Trees**

As discussed above, a binary recursive tree is simply a technique for classifying observations on a binary dependent variable (in our case, the occurrence of a currency crisis). The coefficient on the first principal component becomes $-0.3$ (t-stat. 1.78*) to $-0.4$ (t-stat. 1.89*) for the 3- and 5-percentage-point real GDP growth swings, respectively; for the simple occurrence of a currency crisis, the coefficient is $-0.1$ (t-stat. 1.00).

---

29 Mulder, Perrelli, and Rocha (2001) report probit results using almost identical explanatory variables (but a different definition for the dependent variable) using monthly data for a panel of developing and emerging market countries. Among their corporate vulnerability indicators, a higher ratio of current assets/liabilities and greater profitability are associated with a lower probability of a crisis (and the coefficients are statistically significant). Less intuitively, a larger share of short-term debt in total corporate debt is associated with a lower probability of a crisis, and greater operational cash flow is associated with a higher probability of a crisis (again, both coefficients are significant). Among their legal and institutional variables, strong creditor and shareholder rights are associated with lower probabilities of a crisis, although higher scores on contract enforcement and accounting standards are positively correlated with crises; finally, they find that greater bank and corporate external debt/exports and greater public/total debt to foreign banks ratios are associated with higher probabilities of a crisis, but a higher ratio of bank credit/GDP and a greater ratio of short-term/total external debt is associated with a lower probability of a crisis. Structural variables have also been used in the growth literature: see Barth and others (1999); Levine (1998); Mauro (1995); and in the literature on banking crises. Hutchison (1999).

30 The coefficient on the first principal component becomes $-0.3$ (t-stat. 1.78*) to $-0.4$ (t-stat. 1.89*) for the 3- and 5-percentage-point real GDP growth swings, respectively; for the simple occurrence of a currency crisis, the coefficient is $-0.1$ (t-stat. 1.00).
probability of a currency crisis involving a decline in real GDP growth of at least 3 percentage points. Figures in italics refer to within-node (i.e., conditional) probabilities of a crisis (in percent).

Again, we begin with only the macroeconomic variables. Figure 1 illustrates the resulting binary recursive tree, where the dependent variable is a currency crisis with a 3 percentage point real GDP growth swing, and the explanatory variables are (i) the current account balance; (ii) the real exchange rate appreciation; (iii) the government balance; (iv) the growth in the credit-to-GDP ratio; and (v) the ratio of external debt to reserves.

The tree turns out to be particularly simple as there is a single node, with the current account balance being the explanatory variable at threshold level of about 2½ percent of GDP. The tree branches to the left node when the current account balance is less than −2.6 percent of GDP (i.e., the deficit is greater than 2½ percent of GDP), and to the right node, otherwise. Along the lefthand branch, the probability of a crisis is 7.5 percent: along the righthand branch (countries with current account balances above −2.6 percent of GDP), the probability of a crisis is only 1.1 percent. In other words, countries with current account deficits above 2½ percent of GDP have a seven-fold greater probability of a crisis than countries with smaller deficits.

31 Binary trees for the other cases are available from the authors.
32 Recall that there are 19 crisis observations in the sample, so the unconditional probability of a crisis is only 3 percent.
Among the macroeconomic variables, the algorithm thus picks out the current account balance as the most important variable distinguishing crisis from noncrisis countries. Note that nothing prevents the algorithm from further splitting the tree (using either the current account balance or any of the other potential explanatory variables); however, the improvement in the fit is not sufficient to justify the additional complexity of the tree, given the stopping rule.

Within sample, the tree misclassifies 179 out of 624 observations (about 30 percent); of these, 5 of the 19 crisis observations would have been missed by the tree, and 174 out of 605 noncrisis observations would have been incorrectly called crises. The "out-of-sample" statistics are very similar: 180 out of 624 observations are misclassified, corresponding to 7 crisis observations and 173 noncrisis observations.

**Structural and macroeconomic determinants**

Next, we add the various structural variables (individually, not in terms of their principal components). The resulting tree, again with the conditional probabilities of a crisis at each node, is illustrated in Figure 2.

The first branching of the tree is now the index of public sector governance, with lower scores indicating greater corruption or that "high government officials are likely to demand special payments" and "illegal payments are generally expected throughout lower levels of government in the form of bribes connected with import and export licenses, exchange controls, tax assessment, policy protection or loans." The conditional probability of a crisis in countries that score in the lower half of this governance index (across our sample of countries), i.e., that have worse public sector governance, is 4.8 percent, versus 0.7 percent for countries that score well. Hence, countries with low governance scores are almost seven times more likely to have a crisis than countries with high scores.

Continuing along the lefthand branch of the tree, the second node (2) depends on the current account balance; again with a threshold value of a deficit of about 21/2 percent of GOP. The conditional probability of a crisis in countries with larger deficits is 9 percent compared to 2.2 percent in countries with smaller current account deficits. Continuing along the lefthand branch (node (3)), the next variable is the corporate debt-equity ratio (with a threshold at about 100 percent), and a conditional probability of crisis of 14.7 percent for countries that exceed this threshold (compared to a 3 percent probability of crisis for countries below that threshold). Finally, at node (4), countries that have a high level of total external debt to reserves have a much higher conditional probability of a crisis (though it should be noted that, by this point, there are only two crisis observations remaining).

Returning to the righthand branch of node (2) (countries that score badly on the governance index but have a current account balance greater than 2.6 percent...
of GDP), it is the real exchange rate that matters (node (5)), with a conditional probability of a crisis of 7 percent for countries whose average real exchange rate appreciation exceeds 2 1/2 percent.

Finally, returning to the righthand branch of node (1) (countries that score well on the governance index), it is the corporate debt-equity ratio that matters; those with very high debt-equity ratios have a much higher conditional probability of a crisis.
How should the tree be interpreted?

The algorithm seems to identify two broad groups of crisis countries. For one group, which consists mainly of the advanced industrialized countries and that scores well on the public sector corruption index—probably a proxy for stronger “rule of law” or governance generally—the distinguishing characteristic of countries that suffer currency crises are their banking and corporate sector vulnerabilities (rather than macroeconomic imbalances). Put differently, these countries can better support macroeconomic imbalances, such as current account deficits or real exchange rate appreciations, with relatively less risk of crisis.

The other group, mainly emerging market and developing countries, which tend to score worse on the public sector corruption index, are more vulnerable to macroeconomic imbalances. For this latter group of countries, the most crucial variable is the current account deficit. Even if the current account deficit is modest (less than 21/2 percent of GDP), however, they may still be vulnerable to the effects of real exchange rate appreciations.

Again for this group of countries, when the current account deficit is large, the corporate debt-equity ratio matters, with a cutoff at about 95 percent. Notice that these countries can support a much lower corporate debt-equity ratio (95 percent) compared to the advanced industrialized countries, with good governance, who can support much higher debt-equity ratios (the threshold at node (6) is 380 percent).

How well does the tree perform?

A simple metric of the tree’s performance is the number of misclassified observations (either crisis countries predicted to be noncrisis, or vice versa). Recall that, using only the macroeconomic variables, 179 out of 624 observations are incorrectly classified (in-sample); out-of-sample, 180 out of 624 observations are incorrectly classified. Once the structural variables are added, the number of incorrectly classified observations drops to 130 out of 624 observations, in-sample. In fact, all of the 19 crisis observations are correctly classified (so that all of the 130 incorrect classifications correspond to “false positives”). In the out-of-sample predictions, 156 out of 624 observations are incorrectly classified (of which 10 correspond to crisis cases, and 146 are “false positive” noncrisis cases). Taking the least favorable results, therefore, about 75 percent of all observations are correctly classified, and about one-half of the currency crises that occurred would have been predicted by the tree. The score on predicting crises could, presumably, be improved by weighting type I errors more heavily in the algorithm’s objective function, albeit at the cost of calling more false positives.

Two additional points are worth noting. First, if the short-term debt and debt-equity ratios are dropped from the list of explanatory variables, the resulting tree

Note that since all the macroeconomic variables are lagged (mostly averages over the past three years), while the structural variables are either constant or move very slowly, it would be fair to treat the tree’s out-of-sample results as genuine predictions (rather than just fitted values).
(not shown) again branches on the index of public sector corruption and the current account deficit, but also on accounting standards, and the composite index of shareholders' rights vis-à-vis company directors (MCORP7). As above, this suggests that the impact of corporate governance on the probability of a crisis occurs mostly through the corporate financing structure. Second, if the Asian crisis countries (Indonesia, Korea, Thailand) are dropped from the panel (to see whether the results are being driven by them), structural variables continue to be included in the tree. In particular, node (1) again splits on public sector governance, and node (6) splits on the debt-equity ratio, and a sub-branch of node (5) splits on the corporate short-term debt ratio; however, node (3) no longer exists. As a further robustness test, the upper-income countries were excluded from the dataset to check whether the structural variables are really only picking up the distinction between developing/emerging market countries and industrialized countries. Dropping these countries leaves the tree virtually unchanged, except that the new tree begins with the 21/2 percent of GDP current account deficit threshold (i.e., the first node of this tree corresponds to node (2) of Figure 2, and nodes (1) and (6)—which, as discussed above, reflect the "industrialized country" crises—do not appear in this tree); the tree misclassifies 120 out of 468 observations.

Naturally, different crisis definitions yield somewhat different trees. For instance, if currency crises with the larger swing in real GDP growth is used as the dependent variable instead, the first node no longer splits on the public sector corruption index, simply because there are no observations with such deep crises among the group of advanced industrialized countries that score well on the public sector governance index (i.e., corresponding to node (6) in Figure 2). The tree therefore starts with the current account balance (again picking the threshold of about 21/2 percent of GDP). Among countries with large current account deficits, it is then the structure of corporate short-term debt and the debt-equity ratio that matter, especially in an environment in which credit has been growing rapidly. In-sample, of the 619 noncrisis observations, 10 are misclassified; of the 5 crisis observations, none are misclassified. Out-of-sample, 14 noncrisis observations are misclassified and 3 crisis observations are misclassified.

Alternatively, if the dependent variable is simply the occurrence of a balance of payments crisis, the resulting tree is highly complex, involving some 17 different nodes. The main explanatory variables are the current account balance, the scores on the indices of public sector corruption, risk of government expropriation and contract repudiation, the debt-equity ratio, the short-term debt ratio, the composite index of shareholders' rights vis-à-vis company directors, the minimum percentage of ownership of share capital required to call an emergency shareholders' meeting, and the percentage of mandatory dividends. In-sample, of the

---

35MCORP7 is La Porta and others' aggregate measure of shareholders' rights over company directors formed by adding 1 when (i) the country allows shareholders to mail their proxy vote; (ii) shareholders are not required to deposit their shares prior to a general shareholders meeting; (iii) cumulative voting or proportional representation is allowed; (iv) oppressed minorities mechanisms are in place; (v) the minimum percentage share capital that entitles a shareholder to call for an extraordinary meeting is less than or equal to 10 percent (the sample median in La Porta and others; or (vi) shareholders have preemptive rights that can be waived only by a shareholders' vote (the measure ranges from 0 to 6).
572 noncrisis observations, 92 are misclassified, and of the 52 crisis observations 3 are misclassified; out-of-sample, these become 130 and 25, respectively.\footnote{36These trees are available from the authors}

The precise structure of the trees, therefore, is perhaps of less importance than the general conclusions that emerge. Of these, three bear emphasizing. First, currency crises come in a variety of flavors: occurring both in advanced industrialized countries, with generally sound governance and stronger regulatory frameworks, and in emerging market and developing countries, with much weaker records of governance. Second, there are complex interactions between governance, macroeconomic, and corporate indicators that may contribute to the likelihood of a crisis, and that are not easily captured with the very linear structure of a standard probit. Third, given differences in the overall “rule of law” or governance, countries’ resilience to either macroeconomic imbalances or corporate sector vulnerabilities may differ markedly.

Finally, the tree structure can also be used to refine the probit analysis, allowing for context-dependence within the probit framework. As an illustration, Table 2 reports a probit of a 3 percent real GDP growth swing on the macroeconomic and structural variables identified in the tree above (Figure 2). The first column reports the probit results for the full sample, while the second and third columns report the coefficients for two samples, separated according to the current account balance node (i.e., whether the deficit is greater or less than 2.6 percent of GDP, corresponding to node (2) in Figure 2).\footnote{37Alternatively, the sample could be split at node (1), corresponding to the rule-of-law node. This was not done because the number of crises in the RHS node is too small to obtain useful results using the probit analysis.} Splitting the sample according to the current account node shows that there are qualitative differences between the corresponding crises. For countries with large current account deficits, for instance, the corporate debt-equity ratio becomes highly significant, in contrast to countries with smaller current account deficits.

V. Conclusion

In this paper, we have examined the role of structural factors in currency crises. Given that structural variables typically do not change much, their ability to predict crises—especially in a panel or time-series context—is necessarily limited. Nonetheless, the findings suggest that weak governance may make countries particularly vulnerable to the effects of macroeconomic imbalances and corporate sector weaknesses.

These interactions mean that standard regressions or probits may not be able to identify vulnerabilities arising from a confluence of legal, macroeconomic, and corporate factors. To this end, we have proposed the use of an alternative technique, known as a binary recursive tree, that is better suited to identifying such interactions. While we consider our results to be mostly illustrative and, at best, preliminary, we believe that this approach shows some promise.

The interactions also have implications for monitoring and country surveillance work. In particular, they suggest that assessing countries according to a
Table 2. Macro and Structural Determinants of Currency Crisis

<table>
<thead>
<tr>
<th></th>
<th>Full sample</th>
<th>CA/GDP &lt; -2.65</th>
<th>CA/GDP &gt; -2.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.667</td>
<td>-2.953</td>
<td>-0.782</td>
</tr>
<tr>
<td>t-stat</td>
<td>-3.86***</td>
<td>-2.69***</td>
<td>-1.24</td>
</tr>
<tr>
<td>ΔREER(^2)</td>
<td>0.033</td>
<td>0.052</td>
<td>0.025</td>
</tr>
<tr>
<td>t-stat</td>
<td>1.84*</td>
<td>1.17</td>
<td>1.13</td>
</tr>
<tr>
<td>CAB/GDP(^2)</td>
<td>-0.149</td>
<td>-0.427</td>
<td>-0.157</td>
</tr>
<tr>
<td>t-stat</td>
<td>-2.97***</td>
<td>-2.47***</td>
<td>-1.72*</td>
</tr>
<tr>
<td>Ext.Debt/Reserves(^3)</td>
<td>0.000</td>
<td>-0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>t-stat</td>
<td>-0.60</td>
<td>-2.00**</td>
<td>0.17</td>
</tr>
<tr>
<td>Mn•le3 (corruption in government)</td>
<td>-0.110</td>
<td>-0.308</td>
<td>-0.143</td>
</tr>
<tr>
<td>t-stat</td>
<td>1.90*</td>
<td>-2.02**</td>
<td>-1.58</td>
</tr>
<tr>
<td>Corporate debt-equity ratio</td>
<td>0.256</td>
<td>1.393</td>
<td>-0.232</td>
</tr>
<tr>
<td>t-stat</td>
<td>1.74*</td>
<td>3.30***</td>
<td>-0.87</td>
</tr>
</tbody>
</table>

Number of observations | 469 | 139 | 329 |
Positive observations | 18  | 9   | 9   |
Percent correctly predicted | 96.2 | 95.0 | 97.3 |
Log Likelihood ratio   | -64.4| -20.4| -34.0|

Note: Asterisks denote significance at the 10(*) , 5(**), and 1(***)% percent levels.

1 Currency crisis with at least a three percentage point swing in real GDP growth rate.
2 Average for years t-1, t-2, and t-3.
3 End-year t-1.

given list of vulnerability indicators is unlikely to suffice. Rather, the “danger” thresholds depend very much on the particular combination of institutional, macroeconomic, and corporate governance/financial structure indicators, and each country must be assessed in light of these.

APPENDIX

Structural Indicators

In the text, reference is made to a number of structural indicators. This appendix provides a detailed description of them.

We consider four broad categories of structural indicators: rule of law; corporate governance (shareholder rights); corporate governance (creditor rights); and corporate performance.

1. Rule of Law (MRULE)

MRULE1 AND MRULE2 are measures that pertain to law enforcement. A strong system of legal enforcement could even substitute for weak rules, to some extent, since active and well-functioning courts can step in and rescue investors abused by the management.

MRULE1 This measures the efficiency of the judicial system. The assessment of efficiency and integrity of the legal environment as it affects business, particularly foreign firms, is produced by the country risk-rating agency, Business International Corp. The index is the
average between 1980 and 1983, and the scale ranges from 10 (most efficient) to 0 (least efficient). A higher score indicates a better rule of law.

**MRULE2** This variable is an assessment of the law-and-order tradition or rule of law produced by the international rating agency International Country Risk (ICR). The index is an average of the months of April and October of the monthly index between 1982 and 1995. La Porta and others change the scale of the index from its original range which went from 6 to 0) into one that ranges from 10 (greatest tradition for law and order) to 1 (least tradition for law and order). A higher score indicates a better rule of law.

**MRULE3** and **MRULE4** are variables that reflect how government affects businesses.

**MRULE3** This variable is an assessment by ICR of corruption in government. The scale ranges from 10 to 0 (again La Porta and others changed the original range, which went from 6 to 0), with lower scores indicating greater corruption or that “high government officials are likely to demand special payments” and “illegal payments are generally expected throughout lower levels of government in the form of bribes connected with import and export licenses, exchange controls, tax assessment, policy protection, or loans.” A higher score indicates a better rule of law.

**MRULE4** This is ICR’s assessment of the risk of outright confiscation or forced nationalization, i.e., risk of expropriation. The score ranges from 10 (low risk) to 0 (high risk of expropriation). A higher score indicates a better rule of law.

**MRULE5** This variable is ICR’s assessment of the “risk of a modification in a contract taking the form of a repudiation, postponement, or scaling down” due to “budget cutbacks, indigenization pressure, change in government, or a change in government economic and social priorities” or repudiation of contracts by government. The scale ranges from 10 (lowest risk) to 0 (highest risk). A higher score indicates a better rule of law.

**MRULE6** This is an index of accounting standards created by examining and rating companies’ 1990 annual reports on their inclusion or omission of 90 items. These items fall into seven categories (general information, income statements, balance sheets, funds flow statements, accounting standards, stock data, and special items). A minimum of three companies in each country were studied. These companies represent a cross section of various industry groups: industrial companies represented 70 percent, and financial companies represented the remaining 30 percent. The index ranges from 100 (highest) to 0 (lowest). A higher score indicates better rule of law.

### 2. Corporate Governance: Shareholders’ Rights

**MCORP1** This reflects investor protection. If the law stipulates that ordinary shares carry one vote per share, La Porta and others assign it a value of one. In general, investors are better protected when dividend rights are tightly linked to voting rights, i.e., one share, one vote; when votes are tied to dividends, insiders cannot appropriate cash flows to themselves by controlling only a small share of the company’s cash flows but still maintaining voting control. Equivalently, this variable equals 1 if the law prohibits the existence of both multiple voting and nonvoting shares, and does not allow firms to set a maximum number of votes per shareholder irrespective of the number of shares owned (all of which are ways in which the one share, one vote principle can be circumvented). It is set to zero otherwise. A higher score indicates stronger shareholder rights.

The measures **MCORP2** through to **MCORP5** measure the ease with which shareholders can exercise their voting rights. Because these rights measure how strongly the legal system favors shareholders vis-à-vis managers in the voting process, La Porta and others refer to them as anti-director measures.

**MCORP2** This is assigned a value of one if proxy by mail is allowed and zero otherwise. Clearly proxy by mail facilitates shareholders’ ability to exercise their voting rights. In fact,
when proxy by mail is not allowed, it can render it considerably more difficult and onerous for shareholders to exercise their votes (unless they go through the legal procedure of designating proxies at meetings), especially if companies hold their annual meetings around the same time (as tends to be the case in Japan, where about 80 percent of companies tend to hold their annual meetings in the same week). A higher score indicates stronger shareholder rights.

MCORP3 This is assigned a value of one if the company law or commercial code does not allow firms to block shares prior to a general shareholders' meeting and zero otherwise. In some countries the law requires that shareholders deposit their shares with a company or financial intermediary prior to a shareholder meeting. The shares are kept in custody until a few days after the meeting, which prevents shareholders from selling their shares for several days around the time of the meeting. A higher score indicates stronger shareholder rights.

MCORP4 This is assigned a value of one if the company law or commercial code allows shareholders to cast all their votes for one candidate standing for election to the board of directors or allows for a mechanism of proportional representation in the board by which minority interests may name a proportional number of directors to the board—i.e., if it allows cumulative voting or proportional representation—and is assigned a value of zero otherwise. A higher score indicates stronger shareholder rights.

MCORP5 This is assigned a value of one if the company law or commercial code grants minority shareholders either a judicial venue to challenge the decisions of management (including the right to sue directors as in American derivative suits) or of the assembly, or the right to step out of the company by requiring the company to purchase their shares when they object to certain fundamental changes such as mergers, asset dispositions, and changes in the articles of incorporation. Thus this variable reflects minority shareholders' legal mechanisms against perceived oppression by directors. The variable is set to zero otherwise. (Minority shareholders are defined as those shareholders who own 10 percent of share capital or less.) A higher score indicates stronger shareholder rights.

MCORP6 This is the minimum percentage of ownership of share capital required to call an emergency shareholders' meeting. Clearly, the higher this percentage, the harder it is for minority shareholders to organize a meeting to challenge or oust the management. (The percentage varies around the world from 1 percent in the United States to 33 percent of share capital in Mexico.) A higher score indicates weaker shareholder rights.

MCORP7 La Porta and others construct an aggregate measure of shareholders' rights vis-à-vis company directors, formed by adding one when (i) the country allows shareholders to mail their proxy vote; (ii) shareholders are not required to deposit their shares prior to a general shareholders' meeting; (iii) cumulative voting or proportional representation is allowed; (iv) oppressed minorities mechanisms are in place; (v) the minimum percentage of share capital that entitles a shareholder to call for an extraordinary meeting is less than or equal to 10 percent (the sample median in La Porta and others; or (vi) shareholders have preemptive rights that can be waived only by a shareholders' vote. The index ranges from zero (low protection) to six (high protection). A higher score indicates stronger shareholder rights.

MCORP8 This variable equals the percentage of firms' declared earnings that the company law or commercial code requires them to distribute as dividends among ordinary shareholders, i.e., the percentage of mandatory dividend. Although earnings can, of course, be misrepresented within the limits allowed by the accounting system, it at least prevents declarations of high earnings by firms (which might be needed to raise additional funds) without requiring dividend payouts. The mandatory dividend right (which is slightly different from the other shareholder rights listed above) may be needed when other rights of shareholders are too weak to induce them to invest. The variable is assigned a value of zero when no such requirement exists in the law or commercial code. A higher score indicates stronger shareholder rights.
3. Corporate Governance: Creditors’ Rights (MCRED)

**MCRED1** If the bankruptcy or reorganization laws stipulate restrictions on reorganization such as the need for creditors’ consent to file for reorganization, the variable is assigned a value of one. It equals zero if no such restrictions exist. A higher score indicates stronger creditor rights.

**MCRED2** This variable is assigned a value of one if there is no automatic stay on assets. In some countries, the reorganization procedure imposes an automatic stay on the assets upon filing the reorganization petition. An automatic stay prevents secured creditors from gaining possession of their security. It is assigned a value of zero if the law stipulates an automatic stay on assets. A higher score indicates stronger creditor rights.

**MCRED3** This variable is assigned a value of one if secured creditors are ranked first in the distribution of the proceeds that result from the disposition of assets of a bankrupt firm, i.e., if secured creditors are paid first. In some countries secured creditors are not assured the right to collateral in reorganization (although this is rare). In Mexico, for example, various social constituencies need to be repaid before the secured creditors, often leaving the latter with no assets to back up their claims. The variable is set to zero if nonsecured creditors, such as the government and workers, are given absolute priority. A higher score indicates stronger creditor rights.

**MCRED4** If an official appointed by the court or by the creditors is responsible for the operation of the business during reorganization, this variable is assigned a value of one. Equivalently, the variable is assigned a value of one if the debtor does not keep the administration of its property pending the resolution of the reorganization process, i.e., if management does not stay in reorganization. In some countries management stays pending resolution of the reorganization procedure, whereas in other countries, such as Malaysia, management is replaced by a party appointed by the court or creditors. The threat of dismissal may enhance creditors’ power. The variable is given a value of zero if no such threat exists. A higher score indicates stronger creditor rights.

**MCRED5** This variable is the percentage of total share capital needed to avoid the dissolution of an existing firm as mandated by the corporate law, i.e., the legal reserve requirement. This requirement forces firms to maintain a certain level of capital to avoid automatic liquidation. As with the mandatory dividend in the case of shareholders, the legal reserve requirement protects creditors when they have few other powers in that it forces an automatic liquidation before all the capital is stolen or wasted. The variable takes the value of zero for countries without such a legal reserve requirement. A higher score indicates stronger creditor rights.

4. Corporate Performance

Finally, we construct a set of corporate vulnerability and performance indicators.38

**CorpStDebt** The ratio of short-term corporate debt to total corporate debt and, as such, measures a firm’s vulnerability to liquidity squeeze. It is the median of all (nonfinancial) firms in the WORLDSCOPE database.

**CorpDtEq** The debt to (common) equity ratio, which provides an indication of a firm’s vulnerability to interest rate spikes. It is the median of all (nonfinancial) firms in the WORLDSCOPE database.

---

38Other variables, such as corporate earnings, are also available from WORLDSCOPE. These two were chosen because they better capture the structural vulnerability to a crisis, rather than the (more) endogenous variables, such as corporate earnings, that might be expected to move in response to the crisis.
REFERENCES


Volume 50 Index

Volume 50 (2003) comprises three issues, as follows:

No. 1, 1-156
No. 2, 157-320
No. 3, 321-516

Authors


Boughton, James M. On the Origins of the Fleming-Mundell Model .................. 1


Buffie, Edward F. Tight Money, Real Interest Rates, and Inflation in Sub-Saharan Africa ... 115

Bulíf, Aleš, and A. Javier Hamann. Aid Volatility: An Empirical Assessment ........ 64

Canova, Fabio, and Gianni De Nicoló. The Properties of the Equity Premium and the Risk-Free Rate: An Investigation Across Time and Countries ....................... 222


Dalmazzo, Alberto, and Guido de Blasio. Resources and Incentives to Reform .......... 250

De Blasio, Guido, and Alberto Dalmazzo. Resources and Incentives to Reform .......... 250

De Nicoló, Gianni, and Fabio Canova. The Properties of the Equity Premium and the Risk-Free Rate: An Investigation Across Time and Countries ....................... 222
Goldfajn, Illan, and Poonam Gupta. Does Monetary Policy Stabilize the Exchange Rate Following a Currency Crisis? ................................. 90
Gupta, Poonam, and Ilan Goldfajn. Does Monetary Policy Stabilize the Exchange Rate Following a Currency Crisis? ................................. 90
Hamann, A. Javier, and Aleš Bulif. Aid Volatility: An Empirical Assessment ............. 64
Iwata, Shigeru, Mohsin S. Khan, and Hiroshi Murao. Sources of Economic Growth in East Asia: A Nonparametric Assessment .............................. 157
Izquierdo, Mario, Eduardo Ley, and Javier Ruiz-Castillo. The Plutocratic Gap in the CPI: Evidence from Spain ............................................. 136
Kenen, Peter B. Refocusing the Fund: A Review of James M. Boughton’s 
Silent Revolution: The International Monetary Fund, 1979–1989 . . . . . . . . 291
Khan, Mohsin S., Shigeru Iwata, and Hiroshi Murao. Sources of Economic Growth in East Asia: A Nonparametric Assessment .............................. 157
Le Breton, Michel, and Shlomo Weber. The Art of Making Everybody Happy: How to Prevent a Secession .................................................. 403
Ley, Eduardo, Mario Izquierdo, and Javier Ruiz-Castillo. The Plutocratic Gap in the CPI: Evidence from Spain ............................................. 136
Murao, Hiroshi, Mohsin S. Khan, and Shigeru Iwata. Sources of Economic Growth in East Asia: A Nonparametric Assessment .............................. 157
Pivovarsky, Alexander. Ownership Concentration and Performance in Ukraine’s Privatized Enterprises .......................................................... 10
Ruiz-Castillo, Javier, Mario Izquierdo, and Eduardo Ley. The Plutocratic Gap in the CPI: Evidence from Spain ............................................. 136
Tamirisa, Natalia T., and Arvind Subramanian. Is Africa Integrated in the Global Economy? ......................................................... 352
Taylor, Mark P., and Ashoka Mody. The High-Yield Spread as a Predictor of Real Economic Activity: Evidence of a Financial Accelerator for the United States ................................................... 373
Weber, Shlomo, and Michel Le Breton. The Art of Making Everybody Happy: How to Prevent a Secession ........................................... 403
Xie, Danyang, and Chi-Wa Yuen. A Dynamic General Equilibrium Framework of Investment with Financing Constraint ........................... 274
Yuen, Chi-Wa, and Danyang Xie. A Dynamic General Equilibrium Framework of Investment with Financing Constraint ........................... 274

Titles

Aid Volatility: An Empirical Assessment. By Aleš Bulíř and A. Javier Hamann ................. 64
The Art of Making Everybody Happy: How to Prevent a Secession. By Michel Le Breton and Shlomo Weber ........................................ 403
“Big Bang” Versus Gradualism in Economic Reforms: An Intertemporal Analysis with an Application to China. By Andrew Feltenstein and Saleh M. Nsouli ................. 458
Can Inheritance Alleviate the Fiscal Burden of an Aging Population? By Erik Lueth .......... 178
Does Monetary Policy Stabilize the Exchange Rate Following a Currency Crisis? By Ilan Goldfajn and Poonam Gupta ........................................ 90
A Dynamic General Equilibrium Framework of Investment with Financing Constraint. By Danyang Xie and Chi-Wa Yuen ........................................ 274
Is Africa Integrated in the Global Economy? By Arvind Subramanian and Natalia T. Tamirisa ....................................................... 352
On the Origins of the Fleming-Mundell Model. By James M. Boughton ......................... 1
Ownership Concentration and Performance in Ukraine’s Privatized Enterprises. By Alexander Pivovarsky ........................................ 10

©International Monetary Fund. Not for Redistribution
The Plutocratic Gap in the CPI: Evidence from Spain. By Mario Izquierdo, Eduardo Ley, and Javier Ruiz-Castillo ................................................................. 136


The Properties of the Equity Premium and the Risk-Free Rate: An Investigation Across Time and Countries. By Fabio Canova and Gianni De Nicoló ........................... 222

Rationing Rules and Outcomes: The Experience of Singapore's Vehicle Quota System. By Ling Hui Tan .................................................................................. 436


Resources and Incentives to Reform. By Alberto Dalmazzo and Guido de Blasio ...... 250

Sources of Economic Growth in East Asia: A Nonparametric Assessment. By Shigeru Iwata, Mohsin S. Khan, and Hiroshi Murao .................................................. 157

Structural Vulnerabilities and Currency Crises. By Swati R. Ghosh and Atish R. Ghosh ... 481

Tight Money, Real Interest Rates, and Inflation in Sub-Saharan Africa. By Edward F. Buffie ........................................................................................................... 115


Subjects

To facilitate electronic storage and retrieval of bibliographic data, IMF Staff Papers has adopted the subject classification scheme of the Journal of Economic Literature (Nashville, Tennessee).

B METHODOLOGY AND HISTORY OF ECONOMIC THOUGHT
B3 History of Thought: Individuals
B31 Individuals

On the Origins of the Fleming-Mundell Model. By James M. Boughton ................... I

C MATHEMATICAL AND QUANTITATIVE METHODS
C1 Econometric and Statistical Methods: General

Is Africa Integrated in the Global Economy? By Arvind Subramanian and Natalia T. Tamirisa ................................................................. 352

C14 Semiparametric and Nonparametric Methods

Sources of Economic Growth in East Asia: A Nonparametric Assessment. By Shigeru Iwata, Mohsin S. Khan, and Hiroshi Murao ........................................ 157

C15 Statistical Simulation Methods; Monte Carlo Methods

The Properties of the Equity Premium and the Risk-Free Rate: An Investigation Across Time and Countries. By Fabio Canova and Gianni De Nicoló ................................. 222
VOLUME 50 INDEX

C2   Econometric Models: Single Equation Models
C22  Time-Series Models

C4   Econometric and Statistical Methods: Special Topics
C43  Index Numbers and Aggregation

C6   Mathematical Methods and Programming
C61  Optimization Techniques; Programming Models
   A Dynamic General Equilibrium Framework of Investment with Financing Constraint. By Danyang Xie and Chi-Wa Yuen ........................................ 274

D    MICROECONOMICS
D3   Distribution
D31  Personal Income and Wealth Distribution

D4   Market Structure and Pricing
D44  Auctions
   Rationing Rules and Outcomes: The Experience of Singapore's Vehicle Quota System. By Ling Hui Tan ........................................... 436

D45  Rationing; Licensing
   Rationing Rules and Outcomes: The Experience of Singapore's Vehicle Quota System. By Ling Hui Tan ........................................... 436

D5   General Equilibrium and Disequilibrium
D58  Computable and Other Applied General Equilibrium Models
   Can Inheritances Alleviate the Fiscal Burden of an Aging Population? By Erik Lueth ................................................................. 178
   “Big Bang” Versus Gradualism in Economic Reforms: An Intertemporal Analysis with an Application to China. By Andrew Feltenstein and Saleh M. Nsouli ............. 458
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Author(s)</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>D6</td>
<td>Economic Welfare</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D63</td>
<td>Equity, Justice, Inequality, and Other Normative Criteria, and Measurement</td>
<td>Mario Izquierdo, Eduardo Ley, and Javier Ruiz-Castillo</td>
<td>136</td>
</tr>
<tr>
<td>D7</td>
<td>Analysis of Collective Decision-Making</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D70</td>
<td>General</td>
<td>Michel Le Breton and Shlomo Weber</td>
<td>403</td>
</tr>
<tr>
<td>D72</td>
<td>Economic Models of Political Processes: Rent-Seeking, Elections, Legislatures, and Voting Behavior</td>
<td>Alberto Dalmazzo and Guido de Blasio</td>
<td>250</td>
</tr>
<tr>
<td>D9</td>
<td>Intertemporal Choice and Growth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D92</td>
<td>Intertemporal Firm Choice and Growth, Investment, or Financing</td>
<td>Danyang Xie and Chi-Wa Yuen</td>
<td>274</td>
</tr>
<tr>
<td>E</td>
<td>MACROECONOMICS AND MONETARY ECONOMICS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>General Aggregative Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E12</td>
<td>Keynes; Keynesian; Post-Keynesian</td>
<td>James M. Boughton</td>
<td>1</td>
</tr>
<tr>
<td>E2</td>
<td>Consumption, Saving, Production, Employment, and Investment</td>
<td>Robin Brooks</td>
<td>200</td>
</tr>
<tr>
<td>E21</td>
<td>Consumption; Saving</td>
<td>Ashoka Mody and Mark P. Taylor</td>
<td>373</td>
</tr>
<tr>
<td>E3</td>
<td>Prices, Business Fluctuations, and Cycles</td>
<td>Fabio Canova and Gianni De Nicolò</td>
<td>222</td>
</tr>
<tr>
<td>E4</td>
<td>Money and Interest Rates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E43</td>
<td>Determination of Interest Rates; Term Structure of Interest Rates</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

©International Monetary Fund. Not for Redistribution
### E44 Financial Markets and the Macroeconomy

- Does Monetary Policy Stabilize the Exchange Rate Following a Currency Crisis? By Ilan Goldfajn and Poonam Gupta .................................................. 90

### E5 Monetary Policy, Central Banking, and the Supply of Money and Credit

- Does Monetary Policy Stabilize the Exchange Rate Following a Currency Crisis? By Ilan Goldfajn and Poonam Gupta .............. 90
- Tight Money, Real Interest Rates, and Inflation in Sub-Saharan Africa. By Edward F. Buffie ................................................................. 115

### E6 Macroeconomic Aspects of Public Finance, Macroeconomic Policy, and General Outlook

- Can Inheritance Alleviate the Fiscal Burden of an Aging Population? By Erik Lueth ................................................................. 178

### E62 Fiscal Policy; Public Expenditures, Investment, and Finance; Taxation

- Can Inheritance Alleviate the Fiscal Burden of an Aging Population? By Erik Lueth ................................................................. 178

### E63 Comparative or Joint Analysis of Fiscal and Monetary or Stabilization Policy

- Does Monetary Policy Stabilize the Exchange Rate Following a Currency Crisis? By Ilan Goldfajn and Poonam Gupta .................................................. 90
- Tight Money, Real Interest Rates, and Inflation in Sub-Saharan Africa. By Edward F. Buffie ................................................................. 115

### F INTERNATIONAL ECONOMICS

#### F1 Trade

- Is Africa Integrated in the Global Economy? By Arvind Subramanian and Natalia T. Tamirisa ................................................................. 352

#### F2 International Factor Movements and International Business


#### F3 International Finance

- Structural Vulnerabilities and Currency Crises. By Swati R. Ghosh and Atish R. Ghosh ................................................................. 481
Volume 50 Index

F35 Foreign Aid

Aid Volatility: An Empirical Assessment. By Aleš Bulař and A. Javier Hamann .... 64

F4 Macroeconomic Aspects of International Trade and Finance

F41 Open Economy Macroeconomics

On the Origins of the Fleming-Mundell Model. By James M. Boughton ............... 1

Tight Money, Real Interest Rates, and Inflation in Sub-Saharan Africa. By Edward F. Buffie ................................................................. 115

Structural Vulnerabilities and Currency Crises. By Swati R. Ghosh and Atish R. Ghosh ................................................................. 481

F47 Forecasting and Simulation

Structural Vulnerabilities and Currency Crises. By Swati R. Ghosh and Atish R. Ghosh ................................................................. 481

G FINANCIAL ECONOMICS

G1 General Financial Markets

G12 Asset Pricing

The Properties of the Equity Premium and the Risk-Free Rate: An Investigation Across Time and Countries. By Fabio Canova and Gianni De Nicolò ................. 222

G15 International Financial Markets


G3 Corporate Finance and Governance

G32 Financing Policy; Capital and Ownership Structure

Ownership Concentration and Performance in Ukraine's Privatized Enterprises. By Alexander Pivovarsky ................................................. 10

H PUBLIC ECONOMICS

H2 Taxation and Subsidies

H20 General

The Art of Making Everybody Happy: How to Prevent a Secession. By Michel Le Breton and Shlomo Weber ................................................. 403

H5 National Government Expenditures and Related Policies

H55 Social Security and Public Pensions

Can Inheritances Alleviate the Fiscal Burden of an Aging Population? By Erik Lueth ................................................................. 178
VOLUME 50 INDEX

H7  State and Local Government; Intergovernmental Relations

H73  Intergovernmental Differentials and Their Effects

   The Art of Making Everybody Happy: How to Prevent a Secession. By Michel Le Breton and Shlomo Weber ................................................................. 403

L  INDUSTRIAL ORGANIZATION

L3  Nonprofit Organizations and Public Enterprise

L33  Boundaries of Public and Private Enterprise; Privatization; Contracting Out

   Ownership Concentration and Performance in Ukraine's Privatized Enterprises. By Alexander Pivovarsky ................................................................. 10

O  ECONOMIC DEVELOPMENT, TECHNOLOGICAL CHANGE, AND GROWTH

O1  Economic Development

O19  International Linkages to Development; Role of International Organizations

   Aid Volatility: An Empirical Assessment. By Aleš Bulíř and A. Javier Hamann .... 64

O4  Economic Growth and Aggregate Productivity

   Is Africa Integrated in the Global Economy? By Arvind Subramanian and Natalia T. Tamirisa ................................................................................................................... 352

O47  Measurement of Economic Growth; Aggregate Productivity

   Sources of Economic Growth in East Asia: A Nonparametric Assessment. By Shigeru Iwata, Mohsin S. Khan, and Hiroshi Murao ............................................ 157

O5  Economywide Country Studies

O53  Asia including Middle East

   Sources of Economic Growth in East Asia: A Nonparametric Assessment. By Shigeru Iwata, Mohsin S. Khan, and Hiroshi Murao ............................................ 157

P  ECONOMIC SYSTEMS

P2  Socialist Systems

P21  Planning, Coordination, and Reform

   "Big Bang" Versus Gradualism in Economic Reforms: An Intertemporal Analysis with an Application to China. By Andrew Feltenstein and Saleh M. Nsouli ... 458

P3  Socialist Institutions

P31  Socialist Enterprises

   Ownership Concentration and Performance in Ukraine's Privatized Enterprises. By Alexander Pivovarsky ................................................................. 10
Volume 50 Index

Q  AGRICULTURAL AND NATURAL RESOURCE ECONOMICS

Q3  Nonrenewable Resources and Conservation

Q32  Exhaustible Resources and Economic Development

   Resources and Incentives to Reform. By Alberto Dalmazzo and Guido de Blasio . . . . 250

R  URBAN, RURAL, AND REGIONAL ECONOMICS

R4  Transportation Systems

R48  Government Pricing; Regulatory Policies

   Rationing Rules and Outcomes: The Experience of Singapore's Vehicle Quota System.
   By Ling Hui Tan ......................................................... 436
In statistical matter throughout this issue,

dots ( . . . ) indicate that the data are not available;

a dash (—) indicates that the figure is zero or less than half the final digit shown, or that the item does not exist;

a single dot (.) indicates decimals;

a comma (,) separates thousands and millions;

“billion” means a thousand million; and “trillion” means a thousand billion;

a short dash (—) is used between years or months (for example, 1998–99 or January–June) to indicate a total of the years or months inclusive of the beginning and ending years or months;

a slash (/) is used between years (for example, 1998/99) to indicate a fiscal year or a crop year; and

components of tables may not add to totals shown because of rounding.

The term “country,” as used in this publication, may not refer to a territorial entity that is a state as understood by international law and practice; the term may also cover some territorial entities that are not states but for which statistical data are maintained and provided internationally on a separate and independent basis.

Design: Luisa Menjivar-Macdonald and Sanaa Elaroussi