Comparing Projections and Outcomes of IMF-Supported Programs

ALBERTO MUSSO and STEVEN PHILLIPS*

"Program numbers" from a sample of IMF-supported programs are studied as if they were forecasts, through statistical analyses of the relationship between projections and outcomes for growth, inflation, and three balance of payments concepts. Statistical bias is found only for projections of inflation and official reserves. Statistical efficiency can be rejected for all variables except growth, suggesting that some program projections were less accurate than they might have been. Nevertheless, most projections are found to have some predictive value. Since several findings are shown to be sample dependent, the full-sample results should be interpreted cautiously. [JEL C5, E3, D8]

Associated with every economic program supported by IMF resources is a set of projections, a quantitative framework setting out a particular scenario of outcomes for policy and other variables. Such a scenario should provide a basis for determining a coherent set of economic policies and financing plans. Moreover, program projections for key macro variables are made public, so they have the potential to convey information useful for private sector decisionmaking. In light of these functions of program projections, it is natural to ask how they tend to compare to actual outcomes.

*Alberto Musso is a Researcher in Economics at the European University Institute, and Steven Phillips is a Senior Economist in the Western Hemisphere Department of the International Monetary Fund. The majority of this work was completed when Alberto Musso was Intern, and Steve Phillips was Senior Economist, in the Policy Review Division of the IMF's Policy Development and Review Department. We thank Mike Artis, Francis Diebold, Timothy Lane, Saul Lizondo, and Massimiliano Marcellino for useful comments on earlier drafts. We are indebted to Patricia Gillett and Sibabrata Das for excellent research assistance.

This question became more prominent following 1997, when the IMF approved financial arrangements with Indonesia, Korea, and Thailand, in support of economic programs calling for continued growth in the near term. In the event, GDP contracted sharply in all three countries. Moreover, in Korea and Thailand, program projections for continued current account deficits were instead followed by large surpluses and capital flight. These large divergences between projections and outcomes drew critical comments, including some from within the IMF, albeit with varying interpretations. Some have suggested that the large "forecast errors" reflected a misunderstanding of these East Asian economies and the crisis they faced, while others have speculated that the program projections announced by the IMF may contain a deliberate optimistic bias.

Lacking from this discussion, however, has been reference to systematic, statistical evidence about differences between projections and outcomes in IMF-supported programs in general. Was the recent experience in East Asia atypical, or in line with a persistent pattern? More generally, do there exist *patterns* in the record of program projection errors, the knowledge of which might be used to improve program projections in the future?

This paper focuses on program projections in three key areas: output growth, inflation, and the balance of payments. We look for systematic patterns in these projections' errors—in the traditional terminology of forecast evaluation, for signs of bias and inefficiency. Where systematic errors seem to exist, we also investigate whether these could be attributable to systematic differences between programs' policy assumptions and policies actually implemented. On the accuracy of program projections (that is, apart from questions of bias or inefficiency), we make the traditional comparison to a naïve forecast, and we also conduct a test of directional accuracy.

The limited scope of this study deserves emphasis. Most importantly, this paper is not a general evaluation of IMF-supported programs, nor by itself can it be a basis for drawing broad policy lessons about the effectiveness, or value, of such programs. Certainly, the accuracy or inaccuracy of program projections does not summarize programs' success or failure.[1] Indeed, beyond presenting the results of a set of standard statistical exercises, the paper refrains from judgments about the causal interpretation or normative evaluation of these results. The reason for this reticence is that it is not clear that IMF "program numbers" are intended to be, or should aim to be, simple *unconditional forecasts.* As discussed in Section II, a variety of strategies for making projections and economic settings could generate patterns in projection errors, and the empirical analysis here cannot, in general, differentiate between these possibilities. Despite these limitations, the statistical facts documented here may be of interest and, ultimately, of practical use, by suggesting ways to increase the accuracy of projections and by hinting at areas for future research in broader studies of Fund-supported programs.

[1]For example, if a program's positive effect on output exceeded projections, this would hardly imply that the program was a failure. Alternatively, a program with outcomes that matched announced projections might still be considered a failure if one believes that another policy approach would have resulted in a superior outcome.

This paper may be distinguished in several ways from two related types of empirical studies. First, a number of studies have systematically analyzed differences between certain IMF projections and actual results, but none has focused on the projections made in the context of IMF-supported programs. Rather, these studies have looked at the projections routinely made—for all IMF member countries—for the twice-annual *World Economic Outlook* (for instance, Artis, 1988 and 1997; Barrionuevo, 1993; Artis and Marcellino, 2001; and Beach, Schavey, and Isidro, 1999). Projections for countries with programs are not analyzed separately; in fact, the analysis refers to regional aggregations of countries.

Second, other empirical analyses of Fund-supported programs have emphasized broader questions of the general experience under, or the estimated effect of, IMF-supported programs.[2] Some of this literature has used before/after comparisons, while another strand has sought to estimate the effect of Fund programs by constructing a counterfactual no-Fund scenario; neither approach deals with program projections. Several studies prepared by IMF staff have included select information on program projections as well as outcomes, but none has concentrated on the purely statistical questions of unbiasedness, efficiency, or accuracy.[3]

Despite the emphasis here on statistical testing, the presentation of this paper is intended to be largely accessible to a non-technical audience. Forecast evaluation methods tend to be based on intuitive ideas, and the associated statistical techniques are mostly straightforward.

I. Interpretational Issues

This first part of this section reviews very basic issues of forecast evaluation; readers familiar with the subject may be more interested by the second part, which discusses some important interpretational issues that arise in the particular case of IMF program projections.

What Makes a Good Forecasting Record?

The question of what makes a good forecasting record has no simple, unique answer, and forecasts may be analyzed along three dimensions:

- Bias—Do forecasts tend to lean one way or the other, so that errors in one direction are larger, and/or more numerous, than errors in the opposite direction?
- Efficiency—Do forecasts take all available information into account? If so, there should be no way to predict the direction or size of projection errors based on available information (no correlation between any variable measured when the projections are formed and the error later observed).
- Accuracy—Do forecasts tend to be near, by some standard, actual outcomes? And do they tend to correctly predict the direction of change?

[2]A summary of this literature can be found in Khan (1990).

[3]To our knowledge, the only analysis with such a focus applied to program projections is a short unpublished note by Ghosh (1996), which referred mainly to transition economies.

This paper asks all three questions,[4] since none alone tells the whole story—each has its own usefulness, as well as its problems of interpretation and other limitations:

- Tests of *bias* are simple, intuitive, and decisive; findings of bias have the potential to be useful in improving forecast accuracy. However, the value of unbiasedness is not always clear: statistical bias is not necessarily a bad thing, and there may be good reasons to produce biased forecasts, as discussed below.[5] Moreover, forecasts that are unbiased could nevertheless suffer from inefficiency, or they could be so inaccurate as to convey no information at all about future outcomes.
- Results of *efficiency* tests can also suggest ways to improve forecast accuracy. Efficiency is a more demanding criterion than unbiasedness; in fact, unbiasedness is a necessary, though not sufficient, condition for efficiency.[6] But, like unbiasedness, efficiency does not guarantee that forecasts will be accurate enough to be useful. Moreover, there is no fully decisive test of efficiency (since failure to detect inefficiency might reflect an inadequate search for neglected information).
- In many contexts, *accuracy* may be what matters most; moreover, it may be the only thing left to worry about, once steps to eliminate bias and identified inefficiencies have been taken. But accuracy is a matter of degree, and there is no objective standard of what is good or bad, or good enough. Accuracy can only be judged relatively, by comparison to some alternative forecasting method. Moreover, measurements of accuracy do not offer guidance on how to improve projections.

Here we will refrain from general judgments of whether IMF program projections are good or bad. The main point is to document facts—such as patterns or the absence of patterns—regarding program projection errors.

Interpretational Issues Specific to IMF Program Projections

Several problems and ambiguities with using conventional statistical techniques to evaluate IMF program projections should be recognized.

First, and most fundamentally, it is not clear that IMF program projections are intended to be forecasts, in the usual sense of being unconditional expectations. Indeed, the term "forecast" seems to be sedulously avoided in IMF public statements about programs, in favor of terms such as "program objective" or "projection." In our experience, private discussions among IMF staff follow this pattern as well, using also such terms as "the baseline scenario" or simply "the program numbers." (Note that this paper uses "projections.")

[4]Note that in our terminology, "accuracy" refers only to the size or direction of projection errors; whereas the statistical literature uses the term more broadly, to also encompass the concepts of unbiasedness and efficiency. See for example Diebold and Lopez (1996).

[5]The statistical terms *bias* and *unbiasedness* used here do not have any direct normative content. Crudely put, it is not necessarily true that a "good" or "honest" forecaster is one who produces unbiased and efficient forecasts. For example, if the costs of errors in one direction are greater than in the other, a competent forecaster might intentionally generate numbers biased in the latter direction.

[6]Intuitively, a biased forecast is also inefficient because it fails to take into account information on its own track record of off-centered forecasts.

Why might program projections be something other than forecasts? To illustrate the possibilities, it is useful to consider a number of projection strategies that might give rise to statistical bias, in the sense of projection errors with a mean other than zero:

An incentive for bias?

This possibility could work in either direction. A recent publication of the Heritage Foundation (Beach, Schavey, and Isidro, 1999) hypothesizes that IMF staff may have an incentive to deliberately err on the side of optimism when projecting the outcomes of programs supported by IMF financial resources. On the other hand, since developments under all such programs are reviewed by the IMF's Executive Board, one might speculate alternatively that IMF staff would have an incentive to suggest too-pessimistic projections, preferring that their job later will be one of explaining results that are "unexpectedly" good rather than bad.

Projections as conditional on assumptions about future policies?

Program projections should perhaps be thought of as conditional expectations, representing only the expected outcome in the event that economic policies in the program country are implemented exactly as negotiated and assumed in the elaboration of the program scenario. For example, outcomes for inflation might systematically deviate from program projections if implementation of monetary or fiscal policy were to *systematically* deviate in a particular direction from program assumptions (and the IMF did not take this pattern into account).

Asymmetric or bimodal distributions of program outcomes?

To produce an unbiased forecast, one must construct (at least notionally) a probability distribution of possible outcomes, then place the forecast at the mean, "expected" value. However, if this distribution is not symmetric and/or has more than one mode, the mean will in general differ from the most likely (modal) value, and from the median value as well. In such situations, it is not obvious that a forecaster should aim at the mean. Where the forecast is placed depends on the so-called loss function; that is, on how the cost of being wrong depends on the direction and size of the projection error.

Why might the distribution of program outcomes not be symmetric and unimodal? *Asymmetric exogenous shocks* are one possibility; for example, a tropical country's output might be subject to occasional but severe hurricane damage, but there may exist no source of a positive shock capable of affecting output as strongly in the opposite direction. Another source of asymmetry may be *nonlinear responses to shocks:* suppose that the amount of rainfall is normally distributed, but agricultural output can suffer not only from drought but also from unusually heavy rains. *Bimodal distributions* of outcomes may relate to so-called multiple equilibria. For example, a bimodal distribution of GDP outcomes might be perceived in the early stages of a balance of payments crisis driven by capital flight; from such a juncture, private sector confidence will either be quickly restored or it will not, so that either almost none or

almost all private capital will flee, and output will either remain about the same or it will crash. If the single program scenario tended to refer to a certain one of these two most likely outcomes, this projection would not likely coincide with the mean.

Projections aiming to influence program outcomes?

The IMF could believe that announcements of program projections themselves (that is, beyond effects of announcements of program policies) can influence private sector expectations and therefore program outcomes. A temptation could then arise to err deliberately on the side of optimism in announcing program numbers. (For such a strategy to be effective, markets must believe the IMF to possess superior information and/or analysis, making its projections more informative than those of others. Moreover, the resulting pattern of too-optimistic projections must go undetected: otherwise, credibility and therefore influence would in time be lost.)

"Agreed" program numbers as compromises?

If program numbers are the outcome of a negotiation, it is possible that they may not represent exactly the IMF's own view but rather a compromise with the authorities' views or interests. Indeed, while it would be implausible to suppose that the Fund and its negotiating counterparts would always find themselves in exact agreement on a program's likely outcome, the practice is to announce only one set of program projections—suggesting that at least one of the two negotiating parties has agreed to a scenario some distance from its own expectations. If one of the parties tends to generate biased projections, compromise projections will likely also be biased.

As these examples illustrate, it is not obvious that IMF program projections should be expected to reflect unconditional expectations and therefore to be statistically unbiased. The more general message is that care needs to be taken in interpreting the results of applying standard forecast analysis procedures to program projections.

Indeed, a complete interpretation of the results would require more information than can be provided in this study. For example, on the question of bias, note that the possible origins of bias discussed above are not mutually exclusive—several could be at work at once. In general, this paper does not attempt to distinguish empirically which channel(s) are at work, though it may provide some clues, along the following lines:

- Conditioning on policy outturns. Where systematic deviations from program projections are found, we investigate whether such patterns could be attributed to systematic deviations of economic policies from programs' assumptions.
- Checking for signs of aiming for the mode (or median) in a setting of asymmetric outcomes. We look for patterns in which the mean error deviates substantially from zero, the distribution is skewed in that same direction, and the median error is much smaller in absolute value (or is close to zero).
- Checking whether bias is limited to, or stronger in, programs in which the possibility of multiple equilibria might be especially relevant and/or boosting private sector confidence might be especially useful. In particular, in Section VIII, the

paper includes a separate focus on programs in which stemming potentially large capital flight was likely to have been a primary concern.

The other main interpretational problem is that conventional methods of forecast assessment are best suited to analyzing a track record generated by the repeated application of a single forecasting procedure to a single, unchanging economic structure or entity. However, IMF program projections are not generated by a single forecasting team, let alone a single forecasting model.[7] Moreover, the pooled cross-section sample of projections analyzed here refers to economies of widely varying structures, and indeed ones especially likely to be changing. Thus it may not be reasonable to think of the projection errors studied here as being drawn from a single distribution. One implication is that the pooled errors may not be normally distributed (even if each itself were drawn from a normal distribution); indeed, projection errors for only one of the five variables studied appear normal. Another implication is that the interpretation of the results is more complicated, and in particular it becomes less clear that any findings of bias or other inefficiency can be used to improve the accuracy of future forecasts.

We will try to counter these problems in several ways. Thus, we will test normality rather than only assume it, and we will use several statistical procedures that do not require normality. We will also be alert to heteroskedasticity. Finally, we will consider results from distinct subsamples, where we suspect that different economic/program structures, or changing structures, could be involved.

II. Data and Sample Specification

Data

The list of program projections that could be analyzed is extensive: IMF-supported programs are typically elaborated in considerable quantitative detail, with the program scenario including projected values of several hundred variables. Of these, values for perhaps two dozen "Selected Economic Indicators" might be announced at a program's outset.[8]

Here, we analyze projections for five variables:

- Output growth: projections for percentage change in real GDP
- Inflation: projections for CPI inflation rates (end-period basis, where available)
- Three balance of payments concepts: the current account balance, net capital account inflows, and the change in official reserves.

These variables are selected for their obvious importance and because both projection and outcome data for them tend to available, both across programs and also in the sense of being public information, in most cases.[9]

[7]For an informal discussion of how, and in what context, such projections are made, see Mussa and Savastano (1999).

[8]See the IMF's web site for examples of press releases announcing new arrangements.

[9]Strictly speaking, however, the data analyzed here are those available to IMF staff, and data for outcomes could in some cases differ from data from other sources. As regards program projections, it is possible that some of the data used here have never been made public; in order to respect any understandings of confidentiality, we will not identify here data specific to any one program.

We also use a set of conditioning variables, to define subsamples and in regression-based tests of the projections' efficiency; these variables are discussed in Sections V and VI.

One of the five projection variables studied, the change in official reserves, has several unique aspects to be kept in mind. First, official reserves are potentially subject to direct policy control. Second, decisions of the IMF itself can also influence reserves outcomes, since its financial support to a program country can deviate from that envisaged at the program's outset. Third, this is the only variable studied for which a "performance criterion" is routinely established in IMF-supported programs.[10] That is, nearly universally, a floor is negotiated at a program's outset, to be compared with subsequent outcomes as part of the program's conditionality for continued financial support. Such a floor may coincide with the level of reserves projected in the program scenario, or it may lie somewhat below this.

Even in this small set of variables, a number of complications arise. As regards inflation, the CPI is preferred over the GDP deflator, since it is often available on an end-year (point-to-point) basis. So-called period average CPI inflation must be used in some cases, however. Also, for some programs, data limitations mean that the current account balance must be defined excluding receipts of official grants. Finally, since program projections for capital account flows are specified with widely varying formats, these had to be measured residually. The upshot is that not all the data are strictly comparable across countries or programs. However, care was taken to ensure comparability, within each program, between the definitions of projections and outcomes.

The precise construction of the projection errors is as follows: for GDP, actual annual growth rate of real GDP minus projected growth rate (both in percentage points); for inflation, actual annual percentage change of CPI *divided* by the projected change;[11] for the three balance of payments measures, actual less projected flows, in US$, then scaled by actual GDP. For this scaling, a PPP-adjusted GDP measure is used.[12]

Sample

The samples analyzed are subsets of the programs approved by the IMF's Executive Board during the five-year period from January 1993 through December

[10]More precisely, the performance criterion is usually defined in terms of reserves net of liabilities to the IMF (thus flows from the IMF do not affect measured performance).

[11]This means, for example, that an inflation outcome of 15 percent when 10 percent had been projected is considered a much larger error than an outcome of 105 percent for a projection of 100 percent. The alternative of treating these errors as identical does not seem reasonable, given the well-known correlation between the level of inflation and inflation variability, and it would have induced severe heteroskedasticity. See also Ghosh and Phillips (1998), who find that the negative empirical association between inflation and growth rates is nonlinear, such that increments in inflation are of less interest where inflation is already high.

[12]For most countries in this sample, PPP-based GDP is three to four times larger than GDP using actual exchange rates. The latter is problematic since in a certain minority of programs such exchange rates seem to have been extremely over- or undervalued. Using PPP-based GDP avoids heteroskedasticity from this source.

1997. Choice of this period is motivated by the desire to capture modern IMF practices inside a sample of decent size, and with final (or at least reasonably-settled) data available for program outcomes.

Data were collected for 69 programs, involving 47 countries, supported by standby or extended arrangements ("SBAs" or "EFFs") with the IMF. Excluded from this set are 22 so-called precautionary stand-by arrangements,[13] as well as all arrangements approved under the IMF's structural adjustment facilities ("SAF" or "ESAF" programs).[14]

A more difficult question is whether programs in transition economies belong in the analysis. The study period 1993–97 was novel for the IMF in that many of the programs involved such countries (34 of the 69 programs noted above). For several reasons, it is difficult to imagine projection errors in these cases, particularly the earlier ones, as being drawn from the same distribution as for other countries.[15] On the other hand, excluding all transition cases would reduce the sample to only 35. As a compromise approach, we consider as our "basic sample" the set of 35 non-transition cases plus the 19 transition economy cases *approved after January 1, 1995*, a total of 54 programs. However, results are also reported for a "narrow sample" including only the 35 non-transition cases, and for an "extended sample" that includes also 15 "early transition" cases (approved in 1993 or 1994), for a total of 69 programs.

In specifying the exact data to be analyzed, various issues involving timing arise. First, program projections are frequently revised, sometimes as often as quarterly, so that there are usually multiple vintages of projections for the same time period. We choose to examine only original program numbers, meaning the projections established at the outset, when an arrangement is first approved by the IMF's Executive Board. Second, numbers for program outturns may also be revised over time; we choose to use latest-available numbers, supposing these to contain less measurement error. Third, we use only annual data, since quarterly data, especially for projections, are often not available. Fourth, although some programs do specify projections more than one year ahead, we analyze projections and outcomes for the "first program year" only.

A further complication arises in defining the program year: rather than a 12-month ahead basis, program projections are always defined on either a calendar- or fiscal-year basis. This means that projections for the "first program year" will

[13]These are arrangements approved with the expectation that the program country most likely would not choose to draw on the resources made available during the arrangement. Our suspicion is that such arrangements tend to occur in less turbulent times than more typical programs, so that errors in their projections may be thought of as coming from a different distribution.

[14]SAF/ESAF-supported programs are excluded because these programs may be of a qualitatively different nature than SBA/EFF cases, and because they occur in low-income economies that may be structurally different (and may tend to have less accurate data).

[15]Clearly, these cases represented a novel economic structure and a highly uncertain environment in general, with such abrupt shocks as economy-wide price liberalization, introduction of new currencies, and the ending of central planning and directed trade. Moreover, many of these programs involved economies that in the recent past had not existed as distinct economic entities, and therefore meant projecting variables for which there were no past data, and which were then being measured for the first time, often with highly uncertain methods.

normally be established sometime within that same year. Consider an example in which a 12-month SBA is approved in mid-May 1996 (typically, the projections would have been set a month or so before the approval date): with all data on a calendar-year basis, we define the first program year as calendar 1996[16] and record the projections' *horizon length* at approximately seven months. Any arrangement approved during the first nine months of the year would be treated analogously. However, for arrangements approved in the last three months of 1996, the first program year would be defined as 1997. In this way we avoid considering projections for absurdly short horizon lengths. Still, horizon length varies considerably across the sample, from a minimum of three to a maximum of 15 months, and this should be kept in mind, for at least two reasons. First, this variation is a potential source of heteroskedasticity. Second, when the horizon length is less than 12 months, the program projections in principle have the advantage of being informed by developments already observed.

III. Empirical Methods

The statistical procedures used here are fairly straightforward, for the most part being standard approaches to forecast evaluation, along the dimensions of unbiasedness, efficiency, and accuracy.[17] Thus, we conduct the usual test of unbiasedness, regressing projection errors on a constant in order to estimate the mean error and testing the hypothesis that this is zero. To test efficiency, we regress outcomes against projections; we also regress projection errors against a set of variables containing other information available in the pre-program year. Finally, to provide a general indication of the accuracy and predictive value of program projections, we use Theil's U statistic to compare their track record with that of an alternative projection, and we also perform tests of directional accuracy (these procedures are elaborated in Section VIII).[18]

While such procedures are standard, we would emphasize the following aspects of the empirical analysis:

- Rather than simply assuming a symmetric, normal distribution, we measure skewness and test the hypothesis of normality, using the Jarque-Bera test.
- We distinguish "median-unbiasedness" from the more standard (mean-) unbiasedness. We test the hypothesis that the median error is zero; given the nature of the data, we use the Wilcoxon signed rank test. This test does not require that the data be drawn from a normal distribution.
- We take into account variation in projections' horizon length, in several ways.

[16]At this arrangement's outset, projections for calendar 1997 might not yet have been set.

[17]See for example Diebold and Lopez (1996) and Diebold (2001). Note that, as in most of the empirical literature on macroeconomic forecast evaluation, most of the tests applied assume that the quadratic loss function is relevant. (The exceptions are the test of "median-unbiasedness" and the test of directional accuracy.) This choice is mainly dictated by the tractability of this loss function, rather than its key properties: that larger errors are penalized more than proportionately than smaller ones, and that over- and underprediction have identical costs. It would be much more difficult to try to identify the most appropriate loss function for each setting. If the relevant loss function is not quadratic, then conventional statistical properties of optimal forecasts (e.g., unbiasedness) are not necessarily valid.

[18]*Eviews 3.1* was used for all computations.

- We assess robustness of the main findings by considering various subsamples, since projection errors may not all be drawn from a single normal distribution.
- We examine whether patterns of bias or other inefficiency are related to deviations of economic policies from those assumed in the program scenario.

IV. Basic Distributional Properties, and the Question of Bias

Table 1 presents basic, mainly qualitative, information on the distribution of projection errors for each of five variables studied, for each of the narrow, basic, and extended samples. (In general, these basic results do not vary across these samples.) The companion Figure 1, which refers to the basic sample only, offers a simple view of the errors' central tendencies, as reflected in median and mean values, and by the range spanned by the inner two quartiles.

From Table 1 and Figure 1, note the following points:

- *Lack of normality.* For four of the five variables, there is statistically significant evidence that the projection errors are not normally distributed. Only for GDP growth is it not possible to clearly reject normality.[19]
- *Mean and median errors tend to have the same direction.* The exception is the capital account, but for this variable both measures are anyway close to zero.
- *Gaps between mean and median errors.* In the basic sample (Figure 1), mean errors exceed median errors, by a factor of two or more, for projections of GDP, inflation, and the current account. In economic terms, the absolute gap between mean and median is especially noteworthy in the cases of inflation and GDP. From the skewness statistics in Table 1, inflation errors are the most skewed.
- *Statistically significant bias for two of the five variables.* For both inflation and the change in reserves, there is strong evidence that projection errors tend to be positive; that is, programs have systematically *underestimated* these variables. On the other hand, for GDP, and for both current and capital account balances, we cannot reject the hypothesis that the projections are unbiased. This same pattern of results applies in all three samples, and to the questions of both mean- and median-biasedness.[20] (The caveat is that the statistical inferences regarding the mean may not be valid, being based on a t-test that assumes normality.)
- *Errors are often large enough to be of interest.* If most errors were in some sense small, questions of bias and efficiency might not be worth studying, since removing bias or increasing efficiency could produce little gain in accuracy. However, for all five variables, the range spanned by the inner two quartiles, containing just 50 percent of the projection errors, strikes us as being large enough to be of practical economic interest, and motivates statistical examination of projections' *accuracy* in Section VII.

[19]Only in the largest sample considered can normality of GDP errors be rejected. Excluding from this sample of 69 programs just one program (an early transition economy case), normality can no longer be rejected.

[20]For an explanation of the Wilcoxon signed rank test, used here to test median-unbiasedness, see, for example, chapter 11 of Hogg and Craig (1995).

Table 1. Program Projection Errors: Central Tendencies and Other Distributional Properties

Null Hypothesis and Sample	GDP Growth	Inflation	Reserves (change)	Current Account	Capital Account
	Sign	Sign	Sign	Sign	Sign
Median (test if 0)[1]					
Narrow (A)	Negative	Positive*	Positive*	Positive	Positive
Basic (A+B)	Negative	Positive**	Positive***	Positive	Positive
Extended (A+B+C)	0.0	Positive***	Positive***	Positive	Positive
Mean (test if 0)[2]					
Narrow (A)	Negative	Positive*	Positive	Positive	Negative
Basic (A+B)	Negative	Positive***	Positive**	Positive	Negative
Extended (A+B+C)	Negative	Positive***	Positive***	Positive	Positive
Skewness (Test of Nonskewness)[3]					
Narrow (A)	Negative	Positive***	Negative	Positive	Positive
Basic (A+B)	Negative	Positive***	Negative	Positive	Negative
Extended (A+B+C)	Negative***[4]	Positive***	Negative	Positive	Negative
Test of Normality[5]					
Narrow (A)		Rejected***	Rejected**	Rejected***	Rejected***
Basic (A+B)		Rejected***	Rejected***	Rejected***	Rejected***
Extended (A+B+C)	Rejected***[4]	Rejected***	Rejected***	Rejected***	Rejected***

***, **, and * refer to statistical significance at 1, 5, and 10 percent levels, respectively.

Subsample definitions: "A" refers to 35 non-transition economy programs, approved 1993-97; "B" refers to 19 transition economy programs, approved 1995–1997; and "C" refers to 15 "early" transition economy programs, approved 1993–94.

[1]Based on the Wilcoxon signed-ranks test statistic (see Conover, 1980).

[2]Based on Student's t-test statistic of the significance of the intercept term in a regression of deviations on a constant.

[3]Based on the Kiefer-Salmon test statistic, as discussed in Davidson and MacKinnon (1993).

[4]This result is not general; if one of the 69 observations in this sample is removed, the null hypothesis cannot be rejected.

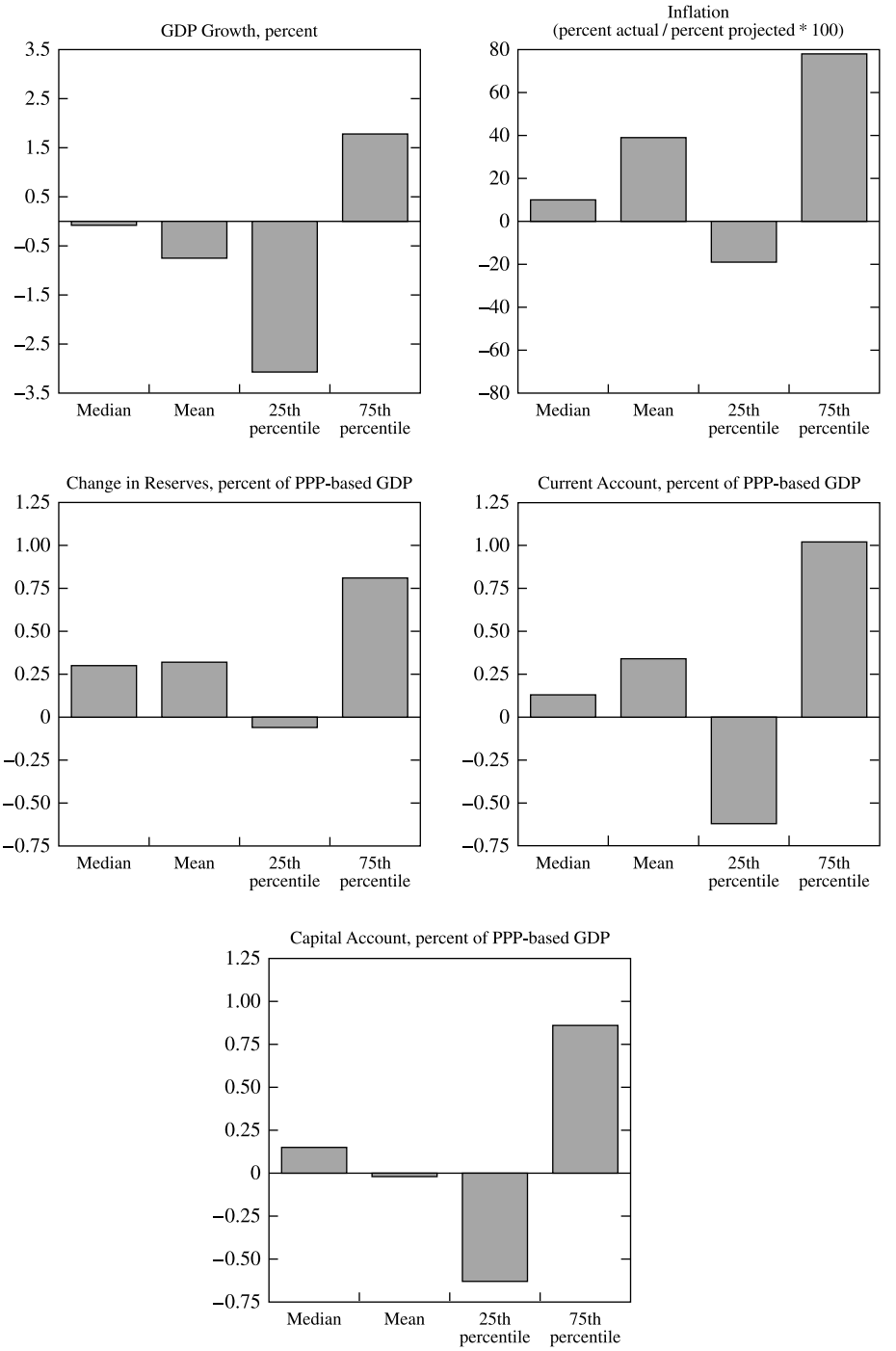
[5]Based on the Jarque-Bera test statistic.

In closing this section of very basic findings, we suggest caution in their interpretation, for several reasons noted in Section I, but in particular because of the concern that these cross-country, cross-program data may not be drawn from the same underlying distribution. This problem cuts two ways. Where bias was not found, it could be that bias nevertheless exists, but only in a certain minority class of programs (and is too diluted to be detected in broader samples). On the other hand, where bias was found, this could be misleading if it reflects only a strong bias in a particular class of programs.

Indeed, investigating a little further demonstrates that some of the results (and nonresults) discussed above are sample dependent. For example, the finding that inflation projections are biased downward turns out to be primarily a reflection of bias in a minority class of programs. Samples restricted to cases with high (above 40 percent) inflation in the pre-program year, or in which a steep[21] disinflation is projected, indicate a sizable, and statistically significant, downward bias in infla-

[21]We consider a steep disinflation to be one in which the projected rate of inflation (expressed in percentage terms) is less than two-thirds of the previous year's inflation rate.

Figure 1. Central Tendencies of Program Projection Errors
(Sample A+B, 54 programs)



tion projections. In samples excluding such programs, deviations from projections are in the same direction, but their mean and median values are much smaller and are not statistically significant.

The opposite situation is encountered in the case of growth projections, for example. Notwithstanding the failure here to find statistically significant evidence of bias in growth projections, a significant upward bias can be found in subsamples constructed to capture programs that are in some sense unusually large. (The concept of a "big program" and the associated results are presented in Section VII.)

V. The Question of Efficiency

We submit the full sample of program projections to two standard statistical tests of efficiency.

The first test is based on a relatively weak criterion of efficiency, making use of a very limited information set. For each of the five variables studied, we regress the level of the program outcome only on a constant term and the level of the program projection. In this test, efficiency is associated with a zero intercept and a slope of one (Mincer and Zarnowitz (1969)). *In general, the null hypothesis of (weak) efficiency cannot be rejected, not even at the 10 percent level.* The single exception involves projections of reserves, for which efficiency can be strongly rejected, at the 1 percent level.

The second test refers to a stronger criterion of efficiency, in which projection errors are uncorrelated with any kind of information available when the projections are made. For each of the five variables studied, we first regress projection errors on an intercept term and a set of 17 candidate regressors. From the results of this general specification, the candidate regressor that is least statistically significant is identified; the regression is then run again excluding this regressor. The process continues until all remaining candidate regressors—if any—are statistically significant at the 5 percent level. (Since, in several regressions, White's test indicated the presence of heteroskedasticity, we choose to judge statistical significance on the basis of White heteroskedasticity-consistent standard errors.[22])

The potential list of such candidate regressors is vast, since it could include any information known in year $(t-1)$. The set of 17 considered here includes such usual suspects as the lagged values of the variables being projected, as well as the values of the projections themselves.[23] In addition, we include dummy variables for certain qualitative types of programs, such as the "early transition economy" cases, those supported by EFFs rather than SBAs, those in low-income countries eligible for ESAF loans, and those that follow on earlier programs, and a measure of the forecast horizon length. We also include a set of four regressors which may capture, or be correlated with, factors relevant to the large projection errors made

[22]That is, without first testing for heteroskedasticity at each stage, we apply White's standard errors. Note that no one test can detect all forms of heteroskedasticity, and it would be tedious to carry out several of these tests for each specification in the general-to-specific procedure.

[23]For gross reserves, both the lagged values and the projected ones refer to levels, the former because of the several missing values for this variable for the $(t-2)$ year (where t is the year of the projection), the latter in order to avoid the problem of perfect multicollinearity.

in the 1997 programs in East Asia: larger access to IMF resources (measured both absolutely and relative to the size of the economy); larger private external debt, and a dummy for programs in countries considered to be emerging markets. While other types of pre-projection information could have been considered, it turns out that this limited set is sufficient to generate statistically significant findings on the question of efficiency.

Following the general-to-specific procedure described above, results for the final regression specification are reported in Table 2. *This time, the general finding is that efficiency of program projections can be rejected, with the exception of growth projections.* For the other four variables studied, at least one regressor is significantly different from zero at the 5 percent level. Thus, with hindsight, it appears that projections for the current account and the capital account could have been made more accurate by taking into consideration the pre-program level of reserves; projections for the change in gross reserves could have been based also on information on the SDR amount of announced access to IMF resources; and CPI projections could have been more accurate using information on whether the program was a followup and on the level of external private debt. Of course, one should not overinterpret the meaning of these particular results—possibly, the variables found significant here could be correlated with other, missing variables that might be much more informative. Even so, the information here could, in principle, be used to increase the accuracy of future program projections.

One indication of potential gain in accuracy comes from the R-squared values shown in Table 2; these suggest, for example, that current account and reserves projections might be made substantially more accurate using readily available pre-program information, whereas the scope for increasing accuracy of CPI and capital account projections in this manner appears relatively smaller. (Of course, greater apparent potential for improvement might have followed from a more ambitious specification search, considering additional regressors or nonlinear treatments.)

VI. Controlling for Policy Outcomes

As discussed in Section I, one reason outcomes might differ from program projections is that economic policies may turn out not to be conducted in the manner assumed when the projections were formulated. Here we examine, in a preliminary and indirect manner, the role of such "policy deviations" in the main findings regarding bias and efficiency.

To try to control for policy deviations, we construct two kinds of summary indicators. One approach considers the amount of IMF financing received—over the entire life[24] of a program—as a share of the amount anticipated at the program's outset; we also construct a dummy variable that takes a value of 1 whenever this "percent drawn" is less than 50 percent.[25] The second approach is similar to that

[24]One flaw in this measure is that the projections we analyze often refer to a time horizon that is shorter than the life of the program.

[25]We thank Miguel Savastano for providing the data (used in Mussa and Savastano, 1999) we used to construct these variables.

Table 2. Program Projection Errors: OLS Regressions
Extended Sample, 69 programs
(with dummy for early transition programs)

	GDP Growth		Inflation		Reserves (change)		Current Account		Capital Account	
	Coeff.	Prob.	Coeff.	Prob.	Coeff.	Prob.	Coeff.	Prob.	Coeff.	Prob.
Intercept	–1.07	0.100*	2.17	0.000***	1.42	0.004**	1.31	0.014**	–0.86	0.095*
Early transition (dummy)										
Log access to IMF resources, SDRs					–0.20	0.039**				
Access / GDP (PPP, $t-1$), percent										
Private debt / GDP (PPP, $t-1$), %			–0.05	0.015**						
Emerging market (dummy)										
EFF dummy										
ESAF-eligible dummy										
Follow-up program dummy			–0.71	0.036**						
H (horizon length, months)										
GDP $t-1$										
log (1+(infl. $t-1$)/100)										
Level of RES $t-1$							–0.41	0.012**	0.39	0.015**
CA $t-1$										
GDP t (proj.)										
log (1+(proj. infl. t)/100)										
Level of RES t (proj.)										
CA t (proj.)										
Diagnostics:										
R-squared	0.000		0.091		0.111		0.107		0.088	
Adjusted R-squared	0.000		0.064		0.098		0.093		0.074	
F-statistic (all coefficients=0)			3.319	0.042**	8.407	0.005***	8.002	0.006***	6.453	0.013**

***significant at 1 percent level.
**significant at 5 percent level.
*significant at 10 percent level.
Cutoff for inclusion of regressors is significance at the 5 percent level.

used in Mecagni (1999), defining an "interrupted program" as one in which the IMF's Executive Board does not act to complete[26] the first scheduled formal review of the program within three months after the date originally scheduled. While neither approach directly measures specific deviations from program policies, it is reasonable to suppose that IMF decisions of whether or not to disburse fully, or to complete a (satisfactory) program review on time, would be related with the degree of adherence to program policy assumptions.

[26]In IMF operational usage, "completion" of a formal program review signals relative satisfaction (since reviews which are initiated but turn out to yield insufficiently satisfactory overall assessments are simply not "completed").

Possibly, programs that are interrupted or have a low percent drawn might be more likely than others to have been hit by some form of bad luck, which may also be reflected in the pattern of projection errors. Negative exogenous shocks might make compliance with originally-agreed policies difficult; if agreement on a modified set of policies cannot be reached, the program could be interrupted. However, the possibility of this interpretation does not undermine the analysis below. Whether interrupted programs reflect mainly bad faith or bad luck, the effect of removing the influence of these programs from the full sample results should be in the same direction. As it happens, we find little sign of such an effect, so the question is moot.

We first reexamine the statistically significant results of Section IV concerning the question of bias. Recall that projections for both inflation and changes in official reserves showed statistically significant bias. Table 3 shows the results of adding, to the simple regressions of projection errors on a constant term, several combinations of regressors based on the indicators discussed above (for the basic sample of 54 programs).

For inflation projection errors, it is interesting to see that the dummy for interrupted programs is statistically significant, with the positive sign one might have expected.[27] The other additional regressors are not significant. The question here, however, is not whether policy deviations influence program outcomes for inflation, but whether the statistical bias in inflation projections reported earlier is mainly driven by such policy deviations. The evidence does not support this suggestion: note from the first two columns of Table 3 that the positive estimate of the intercept is diminished only slightly by the addition of the significant policy regressor, and that its statistical significance is maintained.

As regards errors in projections of changes in reserves, the additional, policy-related regressors are not statistically significant at the 5 percent level, although the dummy for low percent drawn just misses being significant (and has the negative sign one would expect). This time, adding these variables can diminish the statistical significance of the estimated intercept, though the direction of movement in that estimate itself is unclear. Of course, the policy deviations question is somewhat peculiar for the case of reserves projections, since the IMF's own policy disbursement decisions can directly affect outcomes, and therefore projection errors, for reserves. But perhaps the key point is to recall that the bias found for reserves projections was in the direction of *underprojection*—whereas if the formation of program projections ignored the fact that in a certain percentage of programs policies would go off-track, then the resulting pattern of bias would have been expected to be in the opposite direction.

It is also interesting to reconsider the findings regarding efficiency presented earlier in Table 2. Recall that program projections for four of the five variables studied were found to be correlated, statistically significantly, with variables reflecting pre-program information. Possibly, some of the variables that turned out significant in Table 2 did so only because they were somehow predictive of how

[27]That is, if interrupted programs are more likely than others to be characterized by looser-than-programmed monetary and fiscal policies.

Table 3. Inflation and Reserves Deviations: Effect of Controlling for Policy Performance (Basic Sample 54 Programs)

	Intercept		Interruption Dummy		Dummy Percent Drawn 0–50%		Percent Drawn	
	Coeff.	Prob.	Coeff.	Prob.	Coeff.	Prob.	Coeff.	Prob.
Inflation								
Intercept Only	1.393	0.000						
Interruption Dummy	1.218	0.000	0.675	0.043				
Drawn 0–50 % Dummy	1.391	0.000			0.003	0.992		
Percent Drawn	1.636	0.000					0.004	0.542
All	1.393	0.016	0.655	0.058	0.056	0.872	0.002	0.739
Change in Reserves								
Intercept Only	0.321	0.029						
Interruption Dummy	0.314	0.067	0.027	0.936				
Drawn 0–50 % Dummy	0.523	0.004			–0.573	0.055		
Percent Drawn	–0.070	0.866					0.006	0.314
All	0.436	0.43	0.04	0.904	–0.549	0.111	0.001	0.879

policies would later be implemented during the program. But this does not appear to be the case. Adding to the regression specifications shown in Table 2 the dummy for interrupted programs, or the regressor measuring the percentage drawn, yields little change: the regressors previously found to be significant are still significant at the 5 percent level.

These findings are not claimed to be definitive, and in particular it might be useful to reexamine the question using direct measures of policy performance. Bearing in mind that the indirect policy proxies used here have limitations, the message of this section is merely our failure to find signs that policy deviations play a dominant role in the patterns of statistical bias and inefficiency found in Sections IV and V.[28]

VII. Projection Errors in "Big Programs"

In this section, we investigate whether the basic results for central tendencies of projection errors (see Section V) are sample-dependent; in particular, whether these results continue to apply after the data are split into subsamples according to our notion of a "big program." The exercise has two objectives. One is to illustrate concretely the need to interpret the results of Section V with caution—since program projections are not generated by applying a single model to a single economic structure, summary generalizations based on full sample results may not be the most informative.

[28]This of course does not mean that policy deviations do not influence program outcomes, a separate question.

The second purpose is simply to provide a closer look at the track record of program projections in a class of programs which, although a minority, may be the best known: what we call "big programs." Thus we seek to capture programs we suspect may be qualitatively different from others by defining subsamples according to whether the economy in question: (i) is atypically large in terms of its GDP; (ii) is approved to receive IMF financial resources during the program on an unusually large scale, either in terms of SDRs or relative to its economic size; (iii) has developed to the point of being considered an emerging market; and (iv) relatedly, has indebtedness to foreign private creditors that is atypically high. Our informal impression is that these characteristics tend to be found together, and that in these cases IMF-supported programs are conceived at their outset mainly as a response to a crisis driven by capital outflows more than anything else,[29] possibly with a view to avoiding systemic effects.

To investigate the properties of "big" programs' projection errors, we first construct variables to measure, for all programs, the characteristics listed above. Examination of these variables' distributions confirmed that most are skewed toward their high end. We therefore chose threshold levels for each of these variables, arbitrary round numbers which would isolate all positive outliers and a good part of the distribution's right tail (say, the top 20 or 30 percent of the distribution). Details of these variables and the thresholds chosen are found in Table 4, which presents mean and median statistics for subsamples defined by each of these thresholds, for each of the five types of projections we study.

In general, the pattern of results from these subsamples suggests interesting differences in central tendencies of projection errors between "big" programs and the more typical ones. Often, in using Table 4 to compare central tendencies of samples above and below these thresholds, the statistical significance or even the sign of the statistics changes:

- *Growth.* While no significant evidence of bias was encountered in the full sample, there is considerable evidence that growth projections are biased upward in the various big program subsamples—unlike in the complementary subsamples consisting of what might be called "typical" programs.
- *Inflation.* The significant downward bias found in the full sample is not driven by the experience of big programs; on the contrary, evidence of such bias if anything tends to be stronger in the subsamples of more typical programs.[30]
- *Reserves.* Again, the significant downward bias in the full sample is not driven by the big programs; indeed, the bias appears to come entirely from the typical programs.
- *Current Account.* For these projections, no bias is detected in the full samples. Nevertheless, splitting the sample reveals substantial differences: projection errors in the big programs show means and medians that tend to be positive and

[29]For example, as opposed to being prompted by a terms of trade shock or a realization that the external current account and/or fiscal deficit will need to be adjusted, sooner or later, to reach sustainable paths.

[30]But recall from Section V that the full sample findings of downward bias in inflation projections are not robust to excluding other minority classes of programs.

Table 4. Central Tendencies in "Big" and "Typical" Programs (Subsamples of Basic Sample)

Projected Variable and Sample	Observations	Mean	Prob.	Median	Prob.
1. Growth					
Full sample ("Basic")	54	–0.75	0.201	–0.08	0.311
High access – I (>SDR 400 m.)	15	–2.59	0.023**	–3.00	0.036**
All other programs	37	–0.04	0.955	1.02	0.755
High access – II (>1 % of GDP)	19	–2.42	0.052*	–2.20	0.061*
All other programs	35	0.16	0.788	0.10	0.798
High GDP (>US$ 100 b.)	18	–2.49	0.011**	–2.50	0.017**
All other programs	36	–1.38	0.039**	–0.20	0.109
Emerging markets	15	–2.23	0.047**	–3.00	0.065*
All other programs	39	–0.18	0.794	0.10	0.994
High private debt (>10 % GDP)	12	–1.93	0.088*	–1.07	0.170
All other programs	42	–0.41	0.549	–0.03	0.678
2. Inflation (outcome/projection)					
Full sample ("Basic")	54	1.39	0.010***	1.10	0.021**
High access – I (>SDR 400 m.)	15	1.38	0.000***	0.95	0.000***
All other programs	37	1.39	0.000***	1.11	0.000***
High access – II (>1 % of GDP)	19	1.64	0.000***	1.34	0.000***
All other programs	35	1.26	0.000***	1.04	0.000***
High GDP (>US$ 100 b.)	18	1.39	0.098*	1.05	0.200
All other programs	36	1.31	0.000***	1.06	0.000***
Emerging markets	15	1.40	0.164	1.06	0.377
All other programs	39	1.39	0.033**	1.11	0.037**
High private debt (>10 % GDP)	12	1.29	0.000***	1.22	0.004***
All other programs	42	1.42	0.000***	1.08	0.000***
3. Reserves					
Full sample ("Basic")	54	0.32	0.029**	0.30	0.010***
High access – I (>SDR 400 m.)	15	–0.12	0.674	0.05	0.865
All other programs	37	0.49	0.004***	0.39	0.005***
High access – II (>1 % of GDP)	19	0.25	0.480	0.28	0.494
All other programs	35	0.36	0.007***	0.32	0.003***
High GDP (>US$ 100 b.)	18	–0.07	0.786	0.06	0.586
All other programs	36	0.08	0.634	0.08	0.407
Emerging markets	15	–0.15	0.619	0.08	0.820
All other programs	39	0.50	0.003***	0.39	0.004***
High private debt (>10 % GDP)	12	–0.02	0.965	0.07	0.666
All other programs	42	0.42	0.007***	0.33	0.007***

Table 4. *(concluded)*

Projected Variable and Sample	Observations	Mean	Prob.	Median	Prob.
4. Current Account					
Full sample ("Basic")	54	0.34	0.412	0.13	0.366
High access – I (>SDR 400 m.)	15	0.85	0.103	0.43	0.041**
All other programs	37	0.14	0.793	–0.19	0.978
High access – II (>1 % of GDP)	19	0.84	0.401	0.37	0.324
All other programs	35	0.06	0.855	–0.01	0.756
High GDP (>US$ 100 b.)	18	0.72	0.096*	0.40	0.041**
All other programs	36	–0.11	0.783	0.01	0.526
Emerging markets	15	0.78	0.138	0.43	0.132
All other programs	39	0.17	0.575	–0.12	0.878
High private debt (>10 % GDP)	12	–0.59	0.634	0.30	0.969
All other programs	42	0.60	0.141	0.00	0.291
5. Capital account					
Full sample ("Basic")	54	–0.02	0.969	0.15	0.598
High access – I (>SDR 400 m.)	15	–0.97	0.201	–0.27	0.182
All other programs	37	0.35	0.517	0.33	0.237
High access – II (>1 % of GDP)	19	–0.60	0.580	–0.53	0.507
All other programs	35	0.30	0.413	0.33	0.185
High GDP (>US$ 100 b.)	18	–0.78	0.213	–0.17	0.276
All other programs	36	0.19	0.698	0.09	0.706
Emerging markets	15	–0.93	0.222	–0.27	0.244
All other programs	39	0.33	0.536	0.25	0.267
High private debt (>10 % GDP)	12	0.57	0.706	0.12	0.610
All other programs	42	–0.18	0.634	0.15	0.776

Notes: ***significant at 1 percent level; **significant at 5 percent level; *significant at 10 percent level.

much larger than those of typical programs, though only some of these results are statistically significant.

- *Capital Account.* In this case, there is no sign of bias in the full sample, and neither are any of the mean and median values significantly different from zero in either the big or typical program subsamples. Still, it is interesting to note that projection errors' central tendencies in these subsamples tend to go in opposite directions, with the capital account balance tending to fall short of its projected value in the big programs.

In short, our informal concept of a minority class of "big" programs, though roughly defined, often appears to be relevant in analyzing the track record of program projections.

It is interesting to recall here the experience of the 1997 programs in East Asia: as noted earlier, these received wide attention for the deviations that materialized from programs' original projections for output growth as well as the current and capital accounts of the balance of payments. Were these deviations in some sense typical? Compared against the full sample, where no significant bias was found for these types of projections, the answer is no. However, considering only the minority class of "big" programs, there is some evidence of a pattern in which growth turns out lower, the current account balance turns out stronger, and the capital account weaker turns out weaker, than projected. Some but not all of this evidence is statistically significant.

VIII. The Question of Accuracy

The results of the efficiency tests in Section V suggest that program projections for four out of the five variables could have been made more accurate by taking into account available pre-program information. However, those findings say little about the accuracy of program projections (only that accuracy could have been improved).

In this section, we examine the question of accuracy in two ways. First, we consider whether program projections have any predictive value beyond that of a naïve "no change" forecast (that is, forecasts made with the random walk model), using Theil's U statistic. Second, we also perform tests of projections' directional accuracy. (In considering both sets of results, recall from Section III that program projections in principle have the "advantage" of being informed by developments already observed within the projection period.)

Comparison to a Random Walk Forecast

Theil's U statistic represents a comparison of the size of the projection errors that result from two different forecasting methods. Here, the comparison is between the errors in the projections made in IMF-supported programs with the errors that would have resulted if projections for the first program year (t) instead had been set to simply equal the ($t-1$) value. For each set of errors, the root mean squared error (RMSE) is computed; Theil's U statistic is simply the ratio of the RMSE of the actual projection errors to the RMSE of the alternative forecasts. The lower the value of U, the greater the relative accuracy of the IMF set of projections. If the U statistic is equal or close to one, the two forecasting methods perform similarly (in terms of RMSE). If it is lower (higher) than one, then the IMF forecasts are better (worse) than the random walk ones.[31]

Table 5 shows the Theil's U results for all five projection variables studied, for various sample definitions (in the top panel) and forecast horizons[32] (lower

[31]A forecast based on a random walk is often called "naïve," since it utilizes only the information summarized in the $t-1$ value. Still, it is a natural benchmark to consider first, and certain other types of economic forecasts have struggled to outperform it.

[32]Recall that the forecast horizon H is measured as the number of months between the date of the approval of the program and the first month of the year to which the projection refers.

Table 5. Theil's U and Modified Diebold-Mariano Statistics

Sample	Number of Observations	GDP Growth	Inflation	Reserves	Current Account	Capital Account
		Theil's U statistics				
Basic sample (A+B)	54 (51)	0.44	0.10	0.49	0.98	0.93
Extended sample (A+B+C)	69 (65)	0.58	0.15	0.52	0.96	0.89
Of which:						
A. Non-transition	35 (34)	0.47	0.09	0.51	0.96	0.94
B. Transition, 1995–97	19 (17)	0.43	0.68	0.44	1.06	0.85
C. Transition, 1993–94	15 (14)	1.10	3.78	0.64	0.87	0.83
		Modified Diebold-Mariano statistics				
Basic sample (A+B)	54 (51)	–2.62**	–1.50	–1.77	–0.40	–0.31
Extended sample (A+B+C)	69 (65)	–2.31**	–1.48	–2.00**	–0.75	–0.78
		Theil's U statistics				
Within Extended Sample:						
$H \leq 6$	19 (17)	0.32	0.74	0.31	1.57	0.45
$6 < H < 12$	34 (32)	0.65	0.11	0.55	1.01	1.05
$H \geq 12$	16 (16)	0.80	2.66	0.94	0.91	1.09

Notes: The smaller numbers of observations shown in parentheses refer to results for Reserves and Capital Account, due to some missing lagged values. The modified D-M statistics are based on Harvey, Leybourne, and Newbold (1997). They are compared with the critical values from the *t* distribution. In particular, two asterisks (**) means that the null hypothesis of no statistically significant difference between the two sets of forecasts can be rejected at the 5 percent level. "H" refers to length of forecast horizon in months, as discussed in Section III.

panel); it also includes a set of corresponding modified Diebold-Mariano statistics (in the middle panel). This statistic is based on Harvey, Leybourne, and Newbold (1997), who modified the statistic proposed by Diebold and Mariano (1995) in order to test the null hypothesis of no statistically significant difference between two sets of forecasts when the sample is relatively small. Note that results for these tests should be interpreted with some caution because, as recognized by the authors of these statistics, the tests tend to be somewhat oversized, particularly for small samples. For this reason, we report the statistics only for the larger samples.

The results are split strongly, depending on which variable is being projected:

- On the negative side, the last two columns of Table 5 suggest that program projections for both the current and capital accounts of the balance of payments had very little predictive power beyond that of the random walk alternative. Regardless of the sample definition considered in Table 5, the Theil's U values for these two types of projections are close to unity. None of the DM statistics are close to being statistically significant.
- However, the first three columns of Table 5 show that for growth, inflation, and reserves, the program projections in general had considerably smaller errors than did the benchmark alternative. (The exceptions are in the "early transition" subsample.) For the growth and reserves projections, the evidence in favor of the program projections is statistically significant. The DM statis-

tics are not significant for the inflation projections, despite their very low Theil's U value.[33]

For the three variables for which the Theil's U results indicate that the program projections tend to have smaller errors than the random walk alternative, it is interesting to consider the role of horizon length in this finding. Dividing up the extended sample, we find that for two of these variables—growth and reserves—the value of U is higher in the subsamples with longer horizons. This pattern is consistent with some part of the program projections' relative success deriving from being informed by events observed within the projection period. Note that no such pattern is found for the inflation projections.

Directional Accuracy

The previous test of accuracy presumes that the conventional "squared errors" loss function is relevant; here, we conduct a test which avoids this presumption. The analysis of *directional accuracy* asks whether or not the variable being projected actually moved in the same direction, relative to its (provisional) value in year $t-1$, as the movement implied by the program projection. Although the size of the errors is not considered under this test, it may be of interest to know whether a target variable such as inflation, for example, at least moves in the intended direction.

For each variable studied, the number of each of the four possible pairings of projection and outcome directions of change is tallied; this information is then used to construct a test of the null hypothesis of independence (lack of a relationship) between the directions of projected and actual changes. Two alternative test statistics are constructed, one for a chi-squared test and the other the more conservative statistic proposed by Yates (1934).[34] Table 6 presents both statistics for each of the five variables studied, in each of five different sample definitions.

For all but one variable, the results tend to support the directional accuracy of program projections. For GDP growth, inflation, the current account, and the change in reserves, the null hypothesis of independence can be rejected at the 1 percent level, using either test, in both the full sample of 69 programs and the basic sample of 54 programs. (Only in the smaller subsamples are the results not always statistically significant at the 5 percent level.)

On the other hand, for projections of net capital inflows, there is no significant evidence—in any of the samples considered, for either test statistic—that the direction of actual changes is related to program projections at the 5 percent significance level.

[33]We suspect that this pattern reflects considerable variation in the differences between the two inflation projections' errors, combined with a few cases in which the program projections beat the random walk alternative by very large margins.

[34]See Conover (1980). Yates' statistic differs by including a correction intended to compensate for the fact that the continuous chi-square distribution function is used to approximate the discrete distribution function of the test statistic. Since some have argued that this correction is overly conservative, we consider both tests.

Table 6. Directional Accuracy

Variable and Sample	Total Obs.	Of which: DR>0, DF>0	DR>0, DF<=0	DR<=0, DF>0	DR<=0, DF<=0	Percent Correct	Chi-Square	Yates Corrected	Significance (corr.)
GDP Growth Rate									
A. Nontransition, 1993–7	35	18	4	3	10	80.0	11.748	9.428	1% (1%)
B. Transition, 1995–7	19	15	1	1	2	89.5	6.935	3.136	1% (n.r.)
C. Transition, 1993–4	15	11	2	2	0	73.3	0.355	2.935	n.r. (n.r.)
D. A+B	54	33	5	4	12	83.3	19.963	17.198	1% (1%)
E. A+B+C	69	44	7	6	12	81.2	18.688	16.129	1% (1%)
Current Account									
A. Nontransition, 1993–7	35	18	2	3	12	85.7	17.500	14.705	1% (1%)
B. Transition, 1995–7	19	5	5	2	7	63.2	1.571	0.604	n.r. (n.r.)
C. Transition, 1993–4	15	5	2	0	8	86.7	8.571	5.658	1% (5%)
D. A+B	54	23	7	5	19	77.8	16.649	14.487	1% (1%)
E. A+B+C	69	28	9	5	27	79.7	24.798	22.450	1% (1%)
CPI Inflation									
A. Nontransition, 1993–7	35	14	5	4	12	74.3	8.241	6.408	1% (5%)
B. Transition, 1995–7	19	3	1	1	14	89.5	8.872	5.237	1% (5%)
C. Transition, 1993–4	15	2	2	0	11	86.7	6.346	2.757	5% (n.r.)
D. A+B	54	17	6	5	26	79.6	18.261	15.946	1% (1%)
E. A+B+C	69	19	8	5	37	81.2	24.765	22.255	1% (1%)
Changes in Gross Reserves									
A. Nontransition, 1993–7	35	17	8	4	6	65.7	2.333	1.313	n.r. (n.r.)
B. Transition, 1995–7	17	7	1	0	9	94.1	13.388	10.019	1% (1%)
C. Transition, 1993–4	13	6	3	0	4	76.9	4.952	2.633	5% (n.r.)
D. A+B	52	24	9	4	15	75.0	12.956	10.960	1% (1%)
E. A+B+C	65	30	12	4	19	75.4	17.395	15.297	1% (1%)
Net Capital Inflows									
A. Nontransition, 1993–7	35	11	6	7	11	62.9	2.333	1.414	n.r. (n.r.)
B. Transition, 1995–7	17	5	6	2	4	52.9	0.235	0.001	n.r. (n.r.)
C. Transition, 1993–4	13	5	3	2	3	61.5	0.627	0.048	n.r. (n.r.)
D. A+B	52	16	12	9	15	59.6	1.997	1.288	n.r. (n.r.)
E. A+B+C	65	21	15	11	18	60.0	2.675	1.921	n.r. (n.r.)

Notes: “DR” and “DF” refer to the direction of actual and projected changes, respectively. “n.r.” = no rejection of null hypothesis that projected and actual changes are independent.

IX. Concluding Remarks

In this paper we analyzed the statistical relationship between IMF program projections and observed outcomes. We emphasize that this is a limited question, and in particular that statistical analysis of such projections' track record does not itself constitute an evaluation of IMF-supported programs—it does not address such questions as the appropriateness of the design and objectives of these programs, or whether IMF involvement substantially influences economic policies and outcomes.

However, we believe that analysis of the kind provided here can be part of the review of programs' historical experience, and that it may yield insights, or directions for research, that are relevant to overall reviews of IMF-supported programs. If program projections are worth the effort of formulating, and announcing publicly, then the statistical relationship between these projections and actual outcomes is a natural question. We recognize that there exist some potentially important problems of interpretation in applying standard "forecast evaluation" techniques to IMF program projections; these problems should motivate caution in interpretation now, and more in-depth research in the future.

Some readers may have jumped to this concluding section, looking only for a simple bottom line: are IMF program projections good or bad? Having missed Section II, these readers should be reminded that there is no simple, single criterion of a good forecast. This paper has analyzed program projections along three dimensions: bias, efficiency, and accuracy. For three types of projections studied—output growth, and the current and capital accounts of the balance of payments—no statistically significant evidence of bias was found in the main sample considered. However, for both inflation and official reserves, the projections were systematically below outcomes (a finding that confirms the conclusions of previous studies that used a less rigorous approach). Whether there are good reasons for such patterns to exist is a question this study cannot answer. It should also be recalled that some of these results on the question of bias were found to depend on the sample considered—for example, central tendencies in what we call "big" programs were often different from those of more typical programs.

As regards the question of efficiency, the findings depended on the standard considered. In general, the program projections did well against the most basic efficiency criterion. With a stricter criterion, efficiency of the program projections for output growth still could not be rejected. However, for the other four variables studied, statistically significant evidence was found indicating that program projections did not take all available information into account, suggesting that these projections could have been made more accurate (by conventional criteria). Nevertheless, the program projections for several variables were accurate enough to outperform the usual benchmark, a forecast based on a random walk, moreover, for all but one variable, the program projections tended to correctly predict the direction of change. In terms of accuracy, the projections for the current and capital accounts of the balance of payments are the ones that seem to have the most room for improvement, especially in "big" programs.

REFERENCES

Artis, Michael J., 1988, *How Accurate Is the World Economic Outlook? A Post Mortem on Short-Term Forecasting at the International Monetary Fund*, World Economic and Financial Surveys (Washington: International Monetary Fund).

Artis, Michael J., 1997, *How Accurate Are the IMF's Short-Term Forecasts? Another Examination of the* World Economic Outlook, World Economic and Financial Surveys (Washington: International Monetary Fund).

Artis, Michael J., and Massimiliano Marcellino, 2001, "Fiscal Forecasting: The Track Record of the IMF, OECD, and EC," *Econometrics Journal*, Vol. 4, No. 1, pp. S20–S36.

Barrionuevo, José M., 1993, *How Accurate Are the* World Economic Outlook *Projections?* World Economic and Financial Surveys (Washington: International Monetary Fund).

Beach, William W., Aaron B. Schavey, and Isabel M. Isidro, 1999, "How Reliable Are IMF Economic Forecasts?" A Report of the Heritage Center for Data Analysis (Washington: The Heritage Foundation).

Conover, W.J., 1980, *Practical Nonparametric Statistics*, 2nd Edition (New York: John Wiley and Sons).

Davidson, Russell, and J.G. MacKinnon, 1993, *Estimation and Inference in Econometrics* (New York and Oxford: Oxford University Press).

Diebold, F., 2001, *Elements of Forecasting* (Cincinnati: Southwestern College Publishing).

Diebold, F., and J. Lopez, 1996, "Forecast Evaluation and Combination," in Maddala G.S. and Rao C.R., eds., *Handbook of Statistics* (Amsterdam: North-Holland).

Diebold, F., and R. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, Vol. 13, No. 3, pp. 253–65.

Ghosh, Atish R., 1996, unpublished note.

Ghosh, Atish R., and Steven Phillips, 1998, "Warning: Inflation May Be Harmful to Your Growth," *IMF Staff Papers*, Vol. 45 (December), pp. 672–710.

Harvey, D., S. Leybourne, and P. Newbold, 1997, "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, Volume 13, No. 2, pp. 281–91.

Hogg, R.V., and A.T. Craig, 1995, *Introduction to Mathematical Statistics*, Fifth Edition (Englewood Cliffs, New Jersey: Prentice Hall).

Khan, Mohsin, 1990, "Empirical Analysis of Fund-Supported Programs," *IMF Staff Papers*, Vol. 37 (June), pp. 195–231.

Mecagni, Mauro, 1999, "The Causes of Program Interruptions," in H. Bredenkamp and S. Schadler, eds., *Economic Adjustment and Reform in Low Income Countries* (Washington: International Monetary Fund).

Mincer, J., and V. Zarnowitz, 1969, "The Evaluation of Economic Forecasts," in J. Mincer, ed., *Economic Forecasts and Expectations* (New York: National Bureau of Economic Research).

Mussa, Michael, and Miguel Savastano, 1999, "The IMF Approach to Economic Stabilization," IMF Working Paper 99/104 (Washington: International Monetary Fund).

Yates, F., 1934, "Contingency Tables Involving Small Numbers and the Chi-Squared Test," *Journal of the Royal Statistical Society*, Vol. 1, No. 3, pp. 217–35.