IEO
Independent Evaluation Office
of the International Monetary Fund

BACKGROUND PAPER

BP/14/02

An Evaluation of Commissioned Studies Assessing the Accuracy of IMF Forecasts

Charles Freedman

BP/14/02

IEO Background Paper

Independent Evaluation Office
of the International Monetary Fund

An Evaluation of Commissioned Studies Assessing the Accuracy of IMF Forecasts

Prepared by Charles Freedman

February 12, 2014

Abstract

This paper examines the studies produced over the years for the IMF Research Department to evaluate the accuracy of forecasts made for the *World Economic Outlook* (*WEO*). The technical statistical quality of these evaluations was very good. Over time, their concerns expanded beyond testing the accuracy of past forecasts to methods of applying statistical test results to improve forecasting. In most cases their authors met with IMF staff to discuss the reports and recommendations, and there are indications that the studies influenced the way in which Fund forecasters approached the task of preparing *WEO* forecasts. The studies were insufficiently documented in some areas: only one had written terms of reference; only one received a formal Management response; and none reviewed the changes made in forecast process and practices made in response to previous reports. Recommendations are made for future such studies, including paying more attention to medium-term forecasts and to evaluating the forecast process in addition to its results.

The views expressed in this Background Paper are those of the author(s) and do not necessarily represent those of the IEO, the IMF or IMF policy. Background Papers report analyses related to the work of the IEO and are published to elicit comments and to further debate.

JEL Classification Numbers: C530

Keywords: Forecast evaluation, IMF, *World Economic Outlook*

Author’s E-Mail Address: Charles_Freedman@carleton.ca

Contents Page

Annexes

ABBREVIATIONS

CD	certificate of deposit
CF	consensus forecast
CFLR	consensus forecast—long-run forecast
CIS	Commonwealth of Independent States
CY	current year
ECB	European Central Bank
GDP	gross domestic product
GNP	gross national product
GPM	Global Projection Model
IEO	Independent Evaluation Office
IMF	International Monetary Fund
LIBOR	London Interbank Offered Rate
MPC	Monetary Policy Committee (Bank of England)
OECD	Organization for Economic Cooperation and Development
RES	Research Department of the International Monetary Fund
RMSE	root mean squared error
TOR	terms of reference
WEO	*World Economic Outlook*
YA	year ahead

I. Introduction and Summary

1. As part of the background to a broader evaluation of IMF Forecasts: Process, Quality, and Country Perspectives, this paper examines the studies by outside experts that were commissioned over the years by RES to evaluate the accuracy of forecasts made for the *World Economic Outlook (WEO)*. There have been four such studies: Artis (1988, 1996), Timmermann (2006), and Faust (2013).[1] In addition, a 1993 study was undertaken by Barrionuevo, an IMF staff member. Because his paper is cited as part of the series in all the subsequent studies, I examine it here as well.

2. The present study:

- assesses the appropriateness of the terms of reference of the five studies;
- reviews the methodology and findings of the studies;
- determines whether and how the main conclusions from these studies have influenced the forecast process subsequent to each report; and
- assesses the overall usefulness of the studies for the *WEO* forecast process and, where pertinent, proposes changes to the terms of reference, frequency, and follow-up process that would enhance the value of future studies.

3. The basis of the analysis is a careful reading of the studies, examination of files where available, and discussions with past and current staff of the Division of RES responsible for the *WEO*.

4. The paper has three sections. Section II evaluates the studies. It assesses their terms of reference, the general approach as indicated in their introductory remarks, the variables analyzed, the sample periods, the countries and groupings of countries covered, the statistical tests used, some highlights of the statistical results, comparisons of the Fund's with other forecasts, treatment of program countries, role of policy assumptions, recommendations, meetings with staff, responses of senior management, implementation of recommendations, effects on the forecasting process, frequency and coverage of evaluations, process of forecasting, and the availability of results. In the context of each of these elements, I make proposals regarding future evaluation studies. Section III draws conclusions and summarizes the recommendations emerging from the evaluation. Annex 1 examines each of the studies in

[1] Two of the studies—the second Artis study and the Timmermann study—were published in two different forms. The second Artis study was issued in August 1996 as an IMF working paper and published in December 1997 as part of the *Staff Studies for the WEO*. Timmermann's study was issued as a working paper in March 2006 and published as an article in *IMF Staff Papers* in June 2007. In each of these cases, I focus on the working paper version. My comments on the Faust (2013) study are based on the version dated February 5, 2013, which was referred to as a draft and was not the final version.

more detail, and Annex 2 reproduces the terms of reference for the Timmermann study, the only study that had TOR.

5. I found that the technical statistical quality of the evaluations was very good from the beginning and became even better as more sophisticated statistical methods were introduced. Over time, the recommendations in the commissioned studies, especially the two most recent, expanded their scope beyond technical statistical suggestions for improving the forecasts. In most cases the authors met with IMF staff to discuss the reports and recommendations. There are indications that that the studies influenced the way in which Fund forecasters approached the task of preparing forecasts for the *WEO*, although it is not always easy to pinpoint the exact effects.

6. The studies were insufficiently documented in a variety of areas. Only one (Timmermann, 2006) had written terms of reference, and even those were not reproduced in the report. Similarly, with one exception, IMF Management did not issue a statement following the publication of the evaluation indicating how it intended to respond to the recommendations. Nor did the studies begin with a discussion of any changes made in the forecast methodology and process as a result of recommendations made in previous reports.

7. More substantive concerns are that the studies paid too little attention to the process by which forecasts are assembled, to the importance of making forecast evaluation results easily accessible to staff and perhaps more widely available, and to the need for periodic reconsideration of the appropriateness of the conventions underlying *WEO* forecasts.

II. Evaluation of the Commissioned Studies

A. Nature of *WEO* Forecasts

8. *WEO* forecasting over the years has been largely based on judgment and is fundamentally in the hands of the country desk economists in the IMF.[2] But while it is mostly a bottom-up process, it has always contained important elements of top-down involvement, including the conditioning assumptions made with respect to such matters as member country policies and oil prices. Moreover, the organizing group in the World Economic Studies Division of the RES essentially aggregates the views of the country desks to ensure coherence and consistency across the entire forecast. For example, ensuring that global exports equal global imports.

[2] In another sense, the *WEO* is an institutional product since it is subject to management and department review. For a detailed description and assessment of the current *WEO* forecast process, see Genberg, Martinez, and Salemi (2014).

9. An interesting recent development has been the increased use of model-based projections as part of the forecasting process. Economists in RES use a model[3] along with their judgment of developments in the world economy to develop projections for the major regions of the world economy. These projections give additional top-down assistance to the country desk economists and are used as a cross-check on their forecasts. Some iteration may follow, too, in response to differences between the country desk forecasts and the model-based projections.

B. Terms of Reference of the Studies

10. Though four of the five studies had no written terms of reference (TOR),[4] members of the *WEO* Division of RES had extensive discussions with the authors at the outset about the purpose of the studies and what areas were to be covered. The introductory section of each study states what was requested of the authors. Broadly speaking, each study was intended to do at least what the previous study had done and to address issues that were on the minds of the senior managers of RES at the time it was commissioned.

11. The first study, Artis (1988), was commissioned by RES in response to IMF Executive Directors' concerns about the possibility of bias in the forecasts. There was no request for recommendations; the study was intended to be a technical analysis and to provide technical tools for looking at forecast accuracy. Artis sought to document the forecast record (which had not been done previously) and to examine whether there was evidence of systematic bias and how the Fund's record compared to that of similar organizations. Barrionuevo (1993) was an internal study intended to update the results of Artis (1988); its author also made some important technical observations on the statistical methods used to evaluate the accuracy of *WEO* forecasts. Artis (1996) was mainly seen as an update of the previous studies and was undertaken partly to examine if the Fund's record had changed over time.

12. The intended readership of these first three studies included the Fund's forecasters, national authorities, and above all the Executive Board. Some Executive Directors were at times quite critical of the forecast record with respect to output growth and inflation, and perhaps worried that the Fund did not do well enough in forecasting these variables. The two Artis studies helped to reassure them that the Fund's record was no worse than that of other

[3] The model currently being used is the six-region version of the Global Projection Model (GPM). The GPM project is designed to improve the toolkit to which economists have access for studying both own-country and cross-country linkages. The model is estimated with Bayesian techniques, which provide a very efficient way of imposing restrictions to produce both plausible dynamics and sensible forecasting properties. See Carabenciov and others (2013).

[4] This is shown by the files, and confirmed by discussions with current and retired senior members of the Research Department who were involved with *WEO* forecasting when the various studies were commissioned.

institutions, and that there often were good reasons why outcomes turned out differently from forecasts.

13. The first three studies focused largely on the accuracy of the forecasts up to the time they were written. The authors of the two most recent studies—Timmermann (2006) and Faust (2013)—were asked to update the assessments of accuracy in light of the longer span of data now available to them, and to recommend ways of improving the forecasts.

14. Even the authors of the two most recent studies do not seem to have been asked for an overall evaluation of the forecasting process. Rather, their studies focused on a narrower set of recommendations that would help to improve the forecasting ability of the country desks. Faust (2013) urged *WEO* forecasters (presumably both the desk economists and the *WES*) to pay much more attention to ongoing structural changes in the economies under study that would affect such variables as potential output, the output gap, and inflation.

15. The TOR[5] for Timmermann (2006) requested the standard analysis of short-term forecasting errors along the lines of Artis (1996) and analysis of some other issues similar to those that had been assessed in the previous evaluations, such as how *WEO* forecasts had fared during the most recent downturn and recovery. But it also set out a number of additional requirements: whether the *WEO* forecasts were too close to consensus forecasts by the private sector, as published by Consensus Economics;[6] whether they adequately reflected international spillovers; why *WEO* forecasts were less accurate for emerging markets than for other country groups; how accurate the medium-term *WEO* forecasts were; and how accurate the forecasts for net oil exporters and importers were. The TOR also raised some issues about elements of the *WEO* process itself, in particular about the way that the process addressed global assumptions and about the nature of the forecast-consistency checks.

16. For the Faust (2013) evaluation, RES discussed the objectives with the author at considerable length. Like his predecessors, Faust applied a standard framework of forecast-efficiency tests to assess whether the *WEO* forecasts were efficient, in the sense of making the best possible use of information available at the time of the forecast. But in his view, the statistical tests should be questioned as much as the *WEO* forecasts themselves. Standard forecast-efficiency tests were designed to shed light on whether a fixed forecasting model (implicit or explicit) was correct or incorrect. Mindful of the recent global economic crisis and the changes that had been taking place in the world economy, Faust argued for a shift in emphasis away from the question "is the model correct?" toward the question "is the model changing appropriately in response to the environment?"

[5] The terms of reference for the Timmermann study are reproduced as Annex 2 below.

[6] Consensus Economics publishes aggregates of private-sector forecasts; it began doing so in 1989 www.consensuseconomics.com/.

17. **Proposals** with respect to terms of reference:

- Develop explicit terms of reference for each future commissioned study.
- Set out the terms of reference in an annex to the study.[7]

C. Introductory Remarks

18. Each of the commissioned studies began with introductory remarks, which were useful in conveying the purpose and the approach taken to the evaluation.

19. Artis (1988) focused on forecasting as a key element in cooperation in international macroeconomic policymaking. After briefly reviewing the Fund's forecasting methods, he indicated that he would provide a detailed analysis of the accuracy of the *WEO* projections for variables and would compare the *WEO* projections with OECD projections. He would also attempt to identify reasons for some of the forecast errors.

20. Barrionuevo (1993) would focus on forecasting accuracy and a qualitative assessment of the way in which various forms of inefficiency were related and how understanding such inefficiencies could help to improve the accuracy of the judgmental forecasts in the *WEO*.

21. Artis (1996) noted that he would assess the accuracy of short-term forecasts for key economic variables for G-7 countries and regional aggregates of developing countries.

22. Timmermann (2006) provided an introductory summary to his main findings. He noted that he would discuss the directional accuracy of forecasts, revisions from the forecasts prepared for the Board to the published forecasts, compare *WEO* and Consensus forecasts, and recommend ways to improve the *WEO* forecasting process.

23. Faust (2013) noted that the *WEO* could probably be systematically improved but, in contrast to the earlier authors, argued that efficiency tests gave at best misleading signals about where any problems might lie and how best to resolve them. He explained that since the recent financial crisis, forecasting for most of the economies in the *WEO* sample had become a matter of evaluating what structural changes had occurred in the economy and what would become the new normal, and determining how quickly the economies might proceed toward it.

[7] Recent forecast-evaluation reports that have published their terms of reference include Meyer and others (2008) for the Bank of Canada, Freedman and others (2011) for the European Central Bank, Meyer and others (2012) for the Bank of Israel, and Stockton (2012) for the Bank of England.

24. **Proposals** with respect to introductory remarks:

- Provide a summary or executive summary in introductory remarks, including main recommendations.

- Discuss early in the report the Fund's follow-up to the recommendations made in the previous study and their effects on the behavior of forecasters and on the forecasting process. In particular, document the management response to the previous report (see proposal in Section N below) and the extent to which the recommendations in the previous report have been implemented (see proposal in Section O below).

D. Variables Analyzed

25. Some of the commissioned studies focused on five variables—output growth, inflation, export and import volumes, and the current account of the balance of payments)—while others limited their analysis to the two principal variables—output growth and inflation.

26. Artis (1988) focused on output growth, inflation, export and import volumes, and the current account of the balance of payments. He also discussed world trade and industrial countries' terms of trade. Barrionuevo limited the data set to output growth and inflation. Artis (1996) and Timmermann (2006) focused on the same five variables used in Artis (1988). Faust (2013) examined only GDP growth and inflation.

27. Artis (1988) established the framework for analysis that was followed to a considerable extent by subsequent studies. He focused on the current-year (CY) and year-ahead (YA) forecasts. More precisely, the current-year forecast was defined as the forecast for the year *t* appearing in the May issue of the *WEO* in the same year. The outturn was specified as the "first available estimate" appearing in the *WEO* of May of the following year. The year-ahead forecast for year *t* was defined as the forecast in the *WEO* issue for October of year *t-1*. The realization was specified as the value published in the *WEO* for October of year *t+1* (first settled estimate). Barrionuevo and Artis (1996) followed this approach.

28. Timmermann (2006) went a step further and made use of two CY estimates for the given year and two YA estimates for the given year, using forecasts published in both the May and October issues of the current year and the previous year.

29. Faust (2013), too, used forecasts made in both Spring and Fall of each year. He differed from his predecessors by specifying the outcome of the forecast as the value of the data as they stood at the time of the Spring *WEO* forecasts two years after the year in question. While his results could therefore not be compared directly with those of the earlier studies, he noted that while some results seemed to depend on the choice of outcome data,

the main results in his report generally did not. Faust also examined the Fund's five-year-ahead forecasts.

30. **Proposals** with respect to variables analyzed:

- Future evaluation studies should focus on the principal variables, typically output growth and inflation, unless at the time the study is commissioned there are specific reasons for going into greater detail.

- Future evaluation studies should examine medium-term forecasts for periods up to five years. More data have become available for this purpose since the *WEO* began to show medium-term forecasts (out to five years) in October 1996 and for the fifth year in tables in April 2008.

E. Sample Time Period

31. The earlier studies began with a mix of unpublished and published data. Subsequent studies extended the data sample forward and began to focus more on comparing the results in subsamples. The two most recent studies dropped the earlier years from the sample, presumably because these had become less relevant in economies whose structure had significantly changed.

32. Artis (1988) used the forecasts in the published versions of the *WEO* starting in May 1980 and similar data from the earlier comparable unpublished documents: for industrial countries from 1971 to 1986 for CY forecasts and from 1973 to 1985 for YA forecasts; and for non-oil developing countries from 1977 to 1986 for current-year forecasts and from 1979 to 1985 for YA forecasts.

33. Barrionuevo (1993) extended the sample to the period 1971 to 1991 for industrial countries, 1977 to 1991 for developing countries, and 1988 to 1991 for non-program developing countries.

34. Artis (1996) extended the period for three more years up to1994. He noted that the increased length of the available series would enable the examination of whether any significant changes had occurred in the IMF's record over time (by analysis of subsamples).

35. Timmermann (2006) covered the period from 1990 to 2003. He did not state his reasons for dropping 1970 to 1989 for most of his analysis, but there were probably at least two. First, since he provided the same type of analysis for advanced economies and developing economies, data availability for the developing economies restricted him to the shorter period. Second, both the advanced economies and developing economies had undergone major structural changes during the longer period.

36. Faust used data for 1990 to 2009.

37. **Proposals** with respect to the sample period:

- Given the longer span of data now available, the practice in the two most recent studies of dropping the data from the earliest years should be continued, especially for emerging and developing economies. The structural changes in these economies over time have been so significant that the data from the earlier period are of little use. Nonetheless, there might be some benefit from continuing to use the longer sample period for the advanced economies or for some subset of the advanced economies (e.g., the G-7 countries).

- Given the longer span of data available, the use of subsets of data to assess whether there has been any change in accuracy of the forecast should continue to be part of the evaluations.

F. Countries and Groupings of Countries

38. The series of studies began with the G-7 countries and the G-7 aggregate, and used only regional groupings for the developing countries. In the more recent studies, forecasts for individual developing countries became part of the data set. Some of the studies also focused on countries that were subject to IMF programs.

39. Artis (1988) examined the accuracy of the forecasts for each member of the G-7, the G-7 as a whole, the aggregates for industrial countries as a group, Europe as a whole, and non-oil developing countries as a group, and for regional groupings of non-oil countries in Africa, Asia, Europe, Middle East, and the Western Hemisphere.

40. Barrionuevo (1993) evaluated forecasts for each G-7 country, the G-7 as a whole, a group of 14 smaller industrial countries, the 21 large and small industrial countries as a group, each of the regional groups of developing countries in Africa, Asia, the Middle East, and the Western Hemisphere, the average of these developing country groups, and the 36 non-program developing countries.

41. Artis (1996) assessed forecasts in the industrial countries group and individual G-7 countries. The analysis of developing countries was confined to regional aggregates as in Artis (1988).

42. Timmermann (2006) examined forecasts for 178 countries in seven economic regions (Africa, Central and Eastern Europe, Commonwealth of Independent States (CIS) countries and Mongolia, developing Asia, the Middle East, Western Hemisphere, and advanced economies).

43. Faust examined the results for 169 countries. He presented much of the detailed information for individual countries in a web appendix.

44. **Proposal** with respect to countries and groupings:

- The evaluation in the text should focus primarily on groupings of countries—industrial countries and regional groupings of emerging and low-income developing economies—and a number of individual very large countries, say the 10 or 15 largest countries in the world. A web appendix should provide statistical results for all countries in the *WEO* database.[8] Assessments of the accuracy of longer-term forecasts might also be considered for some of the countries for which the required data series are available.

G. Statistical Tests

45. By and large, the technical evaluations were state-of-the-art. The studies discussed a limited number of statistical tests in the text, with other tests in appendices. The number of tests was gradually expanded and the tests became increasingly sophisticated. In some cases, the author used specific tests that turned out not to be particularly helpful, and in such cases the tests were dropped in later studies. Of particular importance was the gradual shift from a rather technical assessment of accuracy of past forecasts to an increased emphasis on how to use the test results to improve forecasting.

46. Artis (1988) used the following forecast error summary statistics: mean absolute error of forecast and comparison with mean absolute value of the realized series; root mean square error (RMSE); Theil inequality statistic; and a measure of the rationality of the forecast errors based on the regression of the realization on the forecast. He also assessed whether there was systematic bias in the forecast taken over a long sample period; he evaluated the bias (in an appendix) from the regression of the error series on a constant term.

47. Barrionuevo (1993) carefully analyzed the notions of unbiasedness and efficiency in the forecasts, focusing on necessary and sufficient conditions. For a forecast to be unbiased, a necessary and sufficient condition was that its average error was zero, and this could be tested by regressing forecast errors on a constant. For efficiency, the necessary and sufficient conditions were that the average forecast error was zero, and that the forecast errors were not related to information available at the time the projections were made—including the requirement that the errors be uncorrelated. Barrionuevo defined an accurate forecast as one that is both unbiased and efficient. Unbiasedness was generally regarded as more important than efficiency, because it meant that forecasts were identical to outturns on average and because it was a necessary condition for efficiency.

[8] The proposal for a Web appendix is discussed further in Sections H and T below.

48. Artis (1996) by and large used similar measures to those in Artis (1988), but like Barrionuevo (1993) he placed more emphasis on the bias in forecasting errors and the presence or absence of serial correlation in the errors.

49. Timmermann (2006) presented the standard statistical measures that had been used in earlier studies, including tests for unbiasedness, lack of serial correlation, and efficiency properties. In addition, he evaluated the pattern of forecast revisions, making use of the fact that he was examining the forecasts for CY and YA for both April and September. His evaluation of forecast revisions had the benefit of not requiring a decision to be made on which definition of realization was best. He also examined non-increasing variance of forecast errors as the forecast horizon decreased (i.e., the expectation that the variance of the forecast error should have declined as more information became available). In his analysis of statistical significance, he used bootstrapping to develop measures of the statistical significance of some of the results.

50. Faust (2013) presented more than the usual statistical measures of the data for growth and inflation: mean, median, mode, and standard deviation as well as the 10th, 25th, 50th, 75th, and 90th percentiles of these data. He presented similar measures for the YA forecast errors in GDP growth and inflation for the *WEO* forecasts published in the Spring. He applied the broader set of summary measures to data for all economies as a group and for advanced economies as a group over the sample period as a whole. He also presented the corresponding measures for the decade of the 1990s and for the decade of the 2000s. Use of these measures gave a much broader picture of growth and inflation developments over the period, including the amount of skew in the sample.

51. **Proposal** with respect to statistical tests:

- Studies should continue to make use of advances in statistical literature and take advantage of whatever statistical tests become available to throw light on forecasting accuracy.

H. Some Highlights of Statistical Results

52. An enormous number of results were presented in the five studies. The following discussion highlights some of the key results, especially those related to bias in the forecasts of output growth and inflation.

Output growth in industrial countries

53. Artis (1988) concluded that the accuracy of tests appeared to be fairly satisfactory, with the average absolute errors well below the mean absolute value of the output growth series itself. As might be expected, he found that the CY forecasts were superior to the YA forecasts. He found some bias when the data were pooled, and the tests for bias showed a degree of output optimism in *WEO* forecasts—especially in the second half of the 1970s,

reflecting the fact that the deceleration in growth in many countries took time to be perceived as a break in the trend, rather than as a cyclical downturn.

54. Barrionuevo (1993) noted that the *WEO* year-ahead projections for the whole sample period overstated growth by one half of one percentage point. This bias occurred because the YA forecasts overstated growth in the 1971–82 period. After 1982, however, YA projections of growth were unbiased across the seven major industrial economies. The accuracy of the *WEO* projections for growth improved after 1985, the last year fully analyzed in Artis (1988). This improvement might partly have reflected the more stable world economic environment in the 1980s than in the 1970s.

55. Artis (1996) concluded that the evidence generally indicated that the output growth forecasts were not, on a country-by-country basis, statistically biased. Nonetheless, individual country observations suggested that there might have been widespread output growth optimism, so much so that when country forecasts were pooled the bias turned out to be statistically significant. It appeared that this bias was a feature of the first sub-period (terminating in 1982) and not significant in the later period (1983 to 1994). Artis noted that these conclusions were similar to those in Barrionuevo (1993).

56. Timmermann (2006), looking at forecasts for the advanced countries, found that the mean of the CY forecast error was very close to zero, but that year-ahead forecasts over-predicted the following year's GDP growth by about one half of 1 percentage point in the case of the April projections and about a third of a percentage point in the case of the September projections. He also found serial correlation in the forecast errors.

57. Faust (2013), looking at the Fall YA *WEO* forecasts of output growth for the sample period as a whole, found that the mean forecast error was 0.8 percentage points for all economies and 0.4 percentage points for advanced economies. The corresponding figures for the 1990s were 1.1 percentage points for all economies and 0.1 percentage point for advanced economies, and for the decade of the 2000s they were 0.4 percentage points for all economies and 0.8 percentage points for advanced economies. In all cases, the bias was in the direction of over-optimism. Finally, he found that the recent crisis had resulted in unprecedented forecast errors in growth, with mean and median errors of about 4 percentage points of over-prediction.

Inflation in industrial countries

58. Artis (1988) noted that the track record was somewhat less satisfactory for inflation than that for output, although still highly acceptable overall. The superiority of the CY forecasts over the YA forecasts again stood out. There appeared to be no bias in the inflation forecasts, at least not if the 1974 YA error was excluded.

59. Barrionuevo (1993) noted that the *WEO* YA projections for the whole sample period understated inflation by one-half of 1 percentage point. This bias occurred because YA

forecasts understated inflation in the 1971–82 period. After 1982, however, YA projections of inflation were unbiased across the seven major industrial economies. Barrionuevo found that, like those for growth, the *WEO* projections for inflation became more accurate after 1985.

60. Artis (1996) concluded that the *WEO* inflation forecasts were not, on a country-by-country basis, statistically biased. Inflation forecasts appeared to suffer much more than the output growth forecasts from serial correlation in the errors.

61. The results in Timmermann (2006) for YA inflation forecasts for advanced economies showed relatively small biases in the direction of over-prediction. This suggested that the *WEO* did not fully take into account the disinflation that took place over the sample period.

62. Faust (2013) noted that the inflation forecast errors for the advanced economies were small: 0.1 percent mean over-prediction for the sample period as a whole, 0.3 percentage points over-prediction for the 1990s, and virtually no error on average in the 2000s. Apparently, forecasters did not predict as much progress on disinflation in the 1990s as actually occurred. The drop in inflation associated with the recent global financial crisis was not only unprecedented, it was also not predicted. The mean and median forecast errors were both significantly negative in 2009 (over-prediction), a phenomenon not previously observed in the sample period.

63. Because of the much higher inflation in developing economies in the earlier period, the mean forecast error for all economies in the 1990s was an under-prediction of almost 50 percent, and that in the 2000s was an under-prediction of about 5 percent. The corresponding median forecast errors for all economies were 0.3 percent and 0.4 percent, indicating the importance in the calculations of high-inflation outliers among developing economies.

Developing countries

64. Artis (1988) noted that the summary statistics for output growth and inflation forecast errors clearly reflected a much poorer forecasting track record for developing economies than for industrial countries. His results of directly testing for bias, both on the data for individual regions and on the pooled data for all the regions, suggested a tendency towards output optimism, at least in the YA sample. There was also some bias in inflation estimates for individual regions though this was not significant when the data were pooled.

65. One of the results in Barrionuevo (1993) was that forecast errors for output growth and inflation in developing countries were significant before 1985, but small for growth projections in the 1986–91 period. The average forecast error for inflation in the developing countries as a whole rose significantly between 1976–85 and 1986–91, but this result was dominated by the errors in only a few countries. For the sample of non-program developing

countries, both inflation and real output growth projections were unbiased in the 1988–91 period.

66. According to Artis (1996), *WEO* forecasts for the groups of developing countries were not very accurate. Data for many of these countries were poor and tardy. And some developing economies had been undergoing dramatic structural change. By and large, the results were very similar to those found in Artis (1988).

67. Timmermann (2006) concluded that the *WEO* forecasts in general over-predicted YA GDP growth in developing economies, with the bias being sizable in most regions. He also noted a bias toward under-prediction of inflation in most developing regions.

68. Faust (2013)'s tables did not show results explicitly for developing economies.

69. **Proposals** with respect to presentation of statistical results:

- Tables in the text should focus on the main results, particularly regarding forecasts of GDP growth and inflation. Less, or no, attention should be paid to other variables unless they are of particular interest to the Fund at the time the study is being prepared.

- Statistical results reported in the study itself should be limited to the largest countries and to regional or structural groupings. As proposed in Section F above, an appendix on the web should give comparable statistical results for all countries, and be made available to the country desks and possibly to outside researchers. This appendix should also be available at the meeting with Executive Directors after the completion of the study, as proposed in Section M below.

I. Comparisons with Other Agencies' Forecasts

70. The five studies all used comparisons of *WEO* forecasts with naive and time-series forecasts. The earlier ones also compared *WEO* forecasts to those of public sector agencies while the later ones shifted to a comparison with private sector forecasters as the data from Consensus Economics became available.

71. Artis (1988) compar*ed WEO* forecasts with naive no-change and 10-year trend-change forecasts using the Theil statistic. He also compar*ed WEO* forecasts with OECD forecasts and forecasts by national governments. For output growth in the industrial countries, he found that the *WEO* forecasts outperformed the naive no-change forecast. However, for output growth in non-oil developing countries, a naive prediction of no change would have provided a better forecast than did the *WEO*. Artis compared *WEO* and OECD forecasts of output growth, inflation, and the balance of payments on current account of the G-7 countries, individually and in aggregate, and found little to choose between the two sets of forecasts: roughly speaking, the two organizations tended to make the same errors about

the same variables for the same countries at the same time, and there was little unexploited information in one forecast that could have been utilized by the other. Artis (1988) also found no significant difference overall between the track records of the *WEO* and those of the national forecasting agencies for output and inflation, either in the G-7 or in Europe as a whole.

72. Barrionuevo (1993) also compared *WEO* forecasts with those from time-series models that he developed using autoregressive and moving-average components. He found that while the Theil statistic indicated that the *WEO* projections were superior to random-walk forecasts, the projections from the more sophisticated time-series models were able to outperform the *WEO* forecasts in many cases.

73. Artis (1996) added to the Theil comparison with naive models a comparison of *WEO* forecasts with Consensus forecasts. The most striking point in this comparison was the qualitative similarity in the pattern of errors, with both sets of forecasts making the same type of error in the same years in the same country.

74. Timmermann (2006), too, used a comparison with Consensus forecasts and examined the potential for improving accuracy by combining *WEO* and Consensus forecasts. Overall, the comparison suggested that the performance of the two sets of forecasts was similar, but it highlighted that the timing of the comparison mattered. Moreover, with the possible exception of YA inflation forecasts, there was little systematic evidence that the *WEO* forecasts could be improved by modifying them to account for information embodied in the Consensus forecasts.

75. Faust (2013) also made use of comparisons with Consensus forecasts and examined ways of combining *WEO* forecasts with medium-term Consensus forecast projections. He concluded that it could be particularly useful to compare *WEO* outcomes with other available forecasts. For example, the reporting process used in producing the *WEO* could be expanded to include available Consensus forecasts and governmental forecasts made at the time of the *WEO* forecast and to incorporate comparative information about the relative performance of the various forecasts.

76. **Proposals** with respect to comparison with other forecasts:

- Continue the practice of comparing *WEO* forecasts with other official and private sector forecasts, such as Consensus forecasts.

- One issue that should perhaps be emphasized more in future evaluation studies is that the Consensus forecasts are unconditional while the *WEO* forecasts incorporate policy assumptions.[9]

J. Program Countries

77. The *WEO* forecasts for countries with IMF programs are made in a different way from those for non-program countries, being effectively constrained by the program targets that the Fund has agreed with those countries. For example, if the program agreement requires a country to implement certain policies and thereby to achieve a specific rate of growth in output and a specific level of inflation, then those targets more or less have to be incorporated into the *WEO* forecasts.

78. The possible consequences for forecast accuracy were shown at times of high inflation in some of the South American countries, when there was a systematic downward bias in the country desks' forecasts for inflation in the region. The area department explained that their projections were constrained by the numbers they had agreed in their programs with many of these countries. Naturally, the program forecasts assumed success even if there were doubts about it, and the staff could not have one optimistic forecast in the program and a considerably more realistic one in the *WEO*. While the *WEO* tried to be somewhat more realistic, the nature of the *WEO* forecasts did not improve much until these countries finally brought down the rate of inflation.

79. Artis (1988) made no reference to program countries.

80. Barrionuevo (1993) analyzed the statistical results for non-program developing countries. He did not evaluate the results for program countries explicitly but inferred them from the differences in the statistical results between developing countries taken as a group and non-program countries. In his introductory remarks, he noted that for developing countries in Fund-supported stabilization and structural adjustment programs, the projections assumed that the policies aimed at achieving growth and inflation objectives were adopted and implemented. Thus, deviations between conditional predictions and outcomes might be interpreted as a measure of the extent to which the policies specified in the programs were not fully implemented—or as a reflection of the fact that the assumptions about the international economic environment faced by these countries were not always realized.

81. Artis (1996) noted that some of the forecasts for the developing country group incorporated data from countries under IMF stabilization programs where the program targets were taken as the forecasts.

[9] See Section K for a discussion of policy assumptions.

82. Timmermann (2006) was asked in his TOR to examine whether the forecasts for program countries showed a bias. He found that systematic over-predictions of real GDP growth were prevalent in forecasts for countries with IMF programs. And in inflation forecasts, a large and systematic under-prediction was observed for program countries.

83. Faust (2013) noted that the apparent bias was larger for program than for other countries and recommended that the nature of forecasts in program countries should be clarified. He suggested that for program countries the forecast could be stated to be conditional on successful implementation of the program. He suggested that the IMF might consider simply acknowledging that the forecasts of program countries were driven by a different set of criteria than other forecasts. As one of two program countries to which he devoted special attention, Faust examined the case of Colombia. Colombia experienced substantial disinflation for some time, which the *WEO* forecast did not track very well; the output growth forecast showed consistent over-prediction from about 1995 to 2000 and entirely missed the nearly 5 percent drop that took place in 1999.

84. **Proposal** with respect to forecasts for program countries:

- In evaluations of *WEO* forecasting accuracy, treat the statistical results for program countries separately from those for non-program countries.

K. Role of Policy Assumptions (or Conventions) and How They May Affect Results

85. *WEO* forecasts are conditional on the *WEO* assumptions—sometimes also referred to as “working hypotheses”—and conventions with respect to interest rates, exchange rates, and oil price movements. Thus they differ from the effectively unconditional forecasts issued by Consensus Economics.

86. Over time, the conventions surrounding the Fund’s forecasts have changed in line with economic developments. The first published *WEO* (May 1980) stated that the projections were based on the assumed maintenance of “present policies,” and this has continued to be the approach with respect to fiscal policy developments. The exchange-rate assumptions in the *WEO*’s earlier years were based on either average exchange rates over some period or on the actual exchange rate, but in April 1986, this was changed to assumptions based on average *real* exchange rates. As far as interest rates were concerned, in October 1990 the *WEO* began to be explicit about assumptions with respect to the six-month U.S. dollar LIBOR, and in April 2002 the Fund introduced explicit assumptions about the three-month interbank deposit rate for the euro and the three-month CD interest rate in Japan. Also, over the years, the *WEO*’s exact specifications of future oil prices changed from time to time between U.S. dollar oil prices and real oil prices.

87. In the April 2103 *WEO* it was assumed:

> that real effective exchange rates remained constant at their average levels during February 11–March 11, 2013 … ; that established policies of national authorities will be maintained (for specific assumptions about fiscal and monetary policies for selected economies, see Box A1); that the average price of oil will be $102.60 a barrel in 2013 and $97.58 a barrel in 2014 and will remain unchanged in real terms over the medium term; that the six-month London interbank offered rate (LIBOR) on U.S. dollar deposits will average 0.5 percent in 2013 and 0.6 percent in 2014; that the three-month euro deposit rate will average 0.2 percent in 2013 and 0.4 percent in 2014; and that the six-month Japanese yen deposit rate will yield on average 0.2 percent in 2013 and 2014. These are, of course, working hypotheses rather than forecasts, and the uncertainties surrounding them add to the margin of error that would in any event be involved in the projections. (IMF, 2013, p. ix)

88. Box A1 of the report noted that the:

> short-term fiscal policy assumptions used in the *World Economic Outlook (WEO)* are based on officially announced budgets, adjusted for differences between the national authorities and the IMF staff regarding macroeconomic assumptions and projected fiscal outturns. The medium-term fiscal projections incorporate policy measures that are judged likely to be implemented. In cases where the IMF staff has insufficient information to assess the authorities' budget intentions and prospects for policy implementation, an unchanged structural primary balance is assumed unless indicated otherwise. (IMF, 2013, p. ix)

Specific assumptions with respect to some of the advanced economies were explained following this general statement.

89. In this context, the Fund should periodically reconsider the appropriateness of the nature of the assumptions and conventions that underlie the *WEO*.

90. Another possible source of *WEO* forecast errors is the speed at which the output gap was assumed to be closed over time. Timmermann (2006) attributed at least part of the bias in the forecasts of output growth and inflation to the possibility that the output gap took longer to return to zero than forecasters assumed. If this was the case, it would give rise to over-prediction of output growth and under-prediction of inflation. It is not clear whether *WEO* forecasts actually incorporated such an output gap assumption and, if so, over what period.[10] Constraining the closing of the output gap to a period of five years at most does not

[10] There was no reference to such an assumption in the introductory section of the *WEO* where the assumptions and conventions are laid out.

seem unreasonable in almost all situations of downturns, with the possible exception of the period after the financial crisis.

91. Timmermann's was the only study to focus on this possible issue. He found that, in some cases, accuracy problems appeared to be related to the "standing" *WEO* assumption that the output gap was eliminated after five years. He noted that if this assumption turned out to be incorrect, one would expect the value of the output gap itself to be able to account for forecast errors.[11] For example, if it took longer to eliminate the output gap than assumed in the *WEO*, then the *WEO* projections of output growth would tend to over-predict for countries with large output gaps. His empirical results supported this view. In the case of inflation forecasts, the *WEO* under-prediction of inflation would also tend to be associated with the size of the output gap. Timmermann recommended that the staff review the output gap assumption regularly.

92. **Proposals** with respect to the role of policy assumptions and how they may affect results:

- Periodically, the IMF should reconsider the appropriateness of the policy assumptions and conventions that underlie the *WEO*. In particular, the increased transparency of central banks in recent years may have lessened their sensitivity to basing forecasts on endogenous interest rate and exchange rate scenarios, particularly over the longer-term forecast horizons.

- The Fund should examine whether assumptions about the time period over which the output gap is closed could have led to biases in the results.

L. Recommendations by the Authors of the Evaluations

93. Most of the studies recommend ways to improve the accuracy of the *WEO* forecasts. These fall into three principal groups. The first group involves technical or statistical changes, for example, using the mean error in past YA forecasts of economic growth published in April to adjust the YA forecasts of economic growth in subsequent April *WEO*s. The second group focuses on increasing forecasters' sensitivity to issues that had arisen in the past. For example, one of the main goals of Faust's (2013) evaluation was to sensitize forecasters to structural changes that resulted in a "new normal" following the financial crisis. The third group of recommendations, which has not thus far received much attention, involves suggestions to improve the entire process of forecasting for the *WEO*, including with respect to the relationship between the bottom-up and top-down aspects of the current arrangements. The next few paragraphs list the main recommendations in the various reports.

[11] Alternatively, such results could be explained by structural or trend changes in the economy that were reflected in estimates of the output gap.

94. With respect to whether the *WEO*'s forecast accuracy could be significantly improved, Artis (1988) noted that:

(i) The quality of an internationally consistent exercise in forecasting could be improved by a reduction in the magnitude and especially the volatility of the world current account discrepancy.

(ii) The sensitivity of forecast accuracy to lead time underlined the importance of promptly taking into account any new information that became available.

(iii) Perhaps the accuracy of the *WEO* could be improved by more widespread use of formal, model-based methods, which would reduce processing time and would allow more frequent ad hoc updates of the forecasts.

In any case, a more formal methodology, simply by being more explicit, would more easily allow constructive postmortem analyses of forecast errors, which should help to improve forecast performance over time.

95. Barrionuevo (1993) noted that the results with respect to time-series models suggested that the *WEO* judgmental projections could be made more accurate by using the statistical properties of such model-based methods to incorporate previous years' errors into forecasts of growth and inflation for the current year. This approach roughly corresponded to the error-correction mechanisms present in time-series models.

96. Artis (1996) made no explicit recommendations.

97. Timmermann (2006) made five recommendations:

(i) Timeliness of information is key to forecasting performance. There are systematic gains from using the latest available information. Therefore, staff should update projections just before publication.

(ii) Performance in forecasting should be continuously monitored, particularly at times of structural instability in some of the underlying variables. Given the presence of what appear to be systematic biases in forecasting performance for output growth and inflation, particularly after 1990, the IMF should explore the possibility of instituting real-time indicators of forecasting performance.

(iii) *WEO* forecasters should use bias-adjusted forecasts as guidance. Timmermann recognized that this approach might be too mechanical and might suffer from its own deficiencies (for example the assumption that the bias remains constant through time). Nonetheless, he believed that a comparison of unadjusted forecasts with bias-adjusted forecasts could enhance understanding of the magnitude and direction of any biases that may exist.

(iv) The risk attached to a forecast should be quantified; ideally, a forecast should include the presentation of full probability distributions of key variables over time.

(v) Staff should review the output gap assumption regularly. Also, more frequent reviews of estimates of potential output growth may be needed.[12]

98. Faust (2013) made two broad recommendations:

(i) *Clarify the goals and nature of the forecast.* First, should the forecast be a mean or a modal forecast? Second, the roles and importance of medium-term versus short-run forecasts should be clarified. Third, for IMF program countries the nature of forecasts should be clarified, since the apparent bias is larger on average for these countries than for others.[13] For program countries the forecast could be stated to be conditional on successful implementation of the program. Alternatively, if the IMF wanted unconditional forecasts for program countries, it could make external forecasters responsible for producing them. Or the IMF might simply acknowledge that the forecasts of program countries were driven by a different set of criteria than other forecasts.

(ii) *Implement a standard system of ongoing evaluation*. In a world of ongoing structural change, the forecast process should continually adapt to new conditions. The IMF might investigate how forecasters could monitor on an ongoing basis the emergence of any systematic problems with the forecasts. Ways to do this might include producing reports that reveal patterns of forecast errors and draw attention to the possibility of the longer-run forecasts being affected by structural change. Standard statistical tests, which should always be interpreted with extreme caution, could be used to flag issues for further investigation.

[12] Box 1.3 in the *WEO* of April 2006 phrased the Timmermann recommendations somewhat differently, in the following way: “The report made a number of recommendations to improve the *WEO* forecasting process. These included: (i) *WEO* growth forecasts for some countries could be improved if more attention were paid to important international linkages, particularly with the United States; (ii) the accuracy of the forecasts should be assessed on an ongoing basis by instituting a set of real-time forecasting performance indicators; (iii) IMF forecasters should more carefully consider the historical forecast ‘biases’ when making their forecasts; and (iv) that the forecast process should be broadened to more explicitly consider the risks around the key central projections.”

[13] Faust argued that since the forecast itself would play a role in negotiations over the conduct of policy in program countries, those responsible for the forecast were placed in an untenable situation to be involved both in formulating, negotiating, and implementing a policy and in giving an unconditional, public forecast of success.

99. **Proposal** with respect to the recommendations by the authors of the evaluations:

- Authors of commissioned studies should be explicit in setting out their recommendations.

- All recommendations should be listed together either at the beginning of the report in the executive summary or at the end of the study.

M. Meetings with Staff and Others After Release of Study

100. Artis (1988) was distributed to the IMF Executive Board after its completion in late 1987. It was intended to serve the Board as a background report for future discussions on the *WEO*. The findings were circulated to IMF departments and discussed extensively with area departments. Artis was present to make presentations and discuss the results.

101. Barrionuevo (1993) was reviewed with area departments.

102. The Artis (1996) findings were circulated to other departments and discussed extensively with area departments. Artis apparently was present to make presentations and to discuss the results.

103. Timmermann (2006): Some members of the *WEO* team in RES met with Timmermann to discuss his conclusions and he also made some presentations. He did not meet with the country desk economists who prepared the forecasts, although his results and conclusions were provided to them. A staff member presented the results of the study[14] at the *WEO* Kick-off Meeting on January 5, 2005.

104. **Proposal** with respect to meetings after release of study:

- Authors of future commissioned studies should continue to meet with the management and staff of RES and should also meet and have extensive discussions with the area department economists, particularly those for the major countries.

- A meeting should be held with the Board after the completion of a commissioned study.

[14] In a presentation entitled "*WEO* Forecast Postmortem—Implementing the Recommendations of the Timmermann Report."

N. Response of Senior Management to Recommendations

105. It is becoming best practice for the institution that commissioned a study to respond publicly to the recommendations in the study following its release and then to follow up with actions to implement its response.[15]

106. To my knowledge, the Fund's only public response to the first four studies was to the Timmermann report. Box 1.3 in the *WEO* of April 2006, after setting out four recommendations by Timmermann, noted the following:

> "Internally, the IMF has begun taking steps to implement the first three recommendations. The rest of this box discusses the fourth recommendation—forecast risks—and how these can be incorporated in the *WEO* process."

It went on to discuss the use of fan charts in the *WEO*.

107. **Proposal** with respect to response of IMF Management to recommendations:

- The IMF should respond publicly to the recommendations in the form of a press release or a box in the *WEO*. The response should set out the Fund's views in response to the recommendations and list the steps that it intends to take over time to address those recommendations with which it agrees. It could also give its reasons for not agreeing with other recommendations if that were the case. This response should be described in the introductory section of the subsequent evaluation, when that is prepared.

O. Implementation of Recommendations in the Previous Evaluation

108. While the reports referred to the statistical results, and occasionally to the recommendations, of previous evaluations, they barely discussed the implementation of these recommendations or noted any resulting changes in forecasters' behavior or techniques. Thus readers would have found it difficult to determine whether the earlier recommendations had led to any changes in the *WEO* forecasting process.

109. Best practice suggests that the beginning of each new study should note whether the recommendations in its predecessor have led to changes in of the forecasting process or practice.[16]

[15] See, for example, the response by the Governors of the Bank of England to the Stockton Review of the Monetary Policy Committee's forecasting capability in Bank of England (2012, 2013) and the response by the Bank of Canada to the external review of its economic research activities in Bank of Canada (2008).

[16] An example of such an approach can be found in Section B (entitled "Review of the 2004 Evaluation") of an external evaluation of the European Central Bank's Directorate General, Research b*y* Freedman and

(continued…)

110. **Proposal** with respect to assessments of the implementation of earlier recommendations:

- Every future evaluation should be required to assess the implementation of recommendations made by the most recent previous evaluation. The results of the assessment should be presented near the beginning of the evaluation, with a discussion of what has been done, and what has not been done and why.

P. The Effects on the Forecasting Process

111. This was clearly the most difficult area to evaluate. While some specific changes could be attributed directly to one of the commissioned evaluations (e.g., the introduction of fan charts into the *WEO* and more attention to risk following the Timmermann (2006) study), it was difficult to pinpoint more generally the effects of the various evaluations on the behavior of forecasters and the way they went about their business.

112. In this section, I simply quote some of the comments about the usefulness of the series of commissioned studies that I received from senior Fund officials who had been involved in the *WEO* process. I asked them whether the findings and recommendations of the studies had resulted in changes in the way that the forecasts were prepared and, more specifically, whether changes came about either in the way that the desk officers carried out their responsibilities for the forecast or in the way that staff in RES interacted with the desk officers and coordinated the preparation of the forecasts.

113. Some of the responses are quoted below:

- The answer to this question is definitely affirmative although it is difficult to pinpoint specific or immediate changes to the way the forecasts were prepared coming out of the Artis and the Barrionuevo (1993) studies. But crucially, the studies helped build an internal consensus about the need to increase the frequency of updating the forecasts, especially when it was felt that turning points were occurring. In the course of the 1990s the area departments (including mid-sized country desks) gradually

others (2011). After listing a number of the most important recommendations of the earlier evaluation, the authors assess the results to date of those earlier recommendations: “Many of these recommendations have been implemented since 2004. In particular, the introduction of a point system for weighting publications in academic journals, the reactivation of the Research Coordination Committee, and the publication of the ECB Research Bulletin have all taken place. However, implementation of recommendations on the management and support of research staff has been less satisfactory. In particular, progress on the buildup of a group of internationally recognized senior researchers has been limited, and there has been no major advance in the supervision (by consultants), support (by research assistants) and visibility (by means of accessible personal homepages) of research staff. It should also be noted that Directorate General Research has only seen a small increase in resources devoted to research on financial stability, and almost none in the international and fiscal policy areas. We will review the outstanding issues below in the context of the set of recommendations in this report.”

began to update their forecasts independently of their annual cycle of Article IV consultations. And the ability of staff to make incremental updates for the smaller countries (using a simple forecast adjustment model) was improved. The introduction of mini or mid-term *WEO*s was the most concrete example of how the *WEO* process became more responsive to changes in global economic conditions. Increased use of alternative scenarios also served the same purpose.

- The studies were valuable. They act as one of the ways to keep people "honest." Just the very existence of the exercise would serve that purpose.

- The reports were widely read in RES and in the IMF more generally. They did identify some important issues and they did have an impact. The comparison with Consensus forecasts was an important result of the evaluations. They were helpful for the country desks. For example, after the Timmermann (2006) study, past forecast errors were integrated into the *WEO* submission system.

- The structural change discussed in Faust will likely become more prominent in the thinking of staff.

114. **Proposal** with respect to the effects of the evaluations on the forecasting process:

- To the extent possible, future evaluations should document the effects of the previous study on the forecasting process. As recommended above, the management response to the previous study, and the implementation of its recommendations, should be described in the introductory remarks. Adding a general assessment of the effects of the previous evaluation on the *WEO* forecasting process would also be useful.

Q. Frequency of Evaluations

115. The span of time between Artis (1988) and Faust (2013) was 25 years, making an average of about 6 years between each of the forecast evaluations if one includes Barrionuevo (1993) in the list. This time span seems quite reasonable given that developments and techniques do not change very rapidly in this area, and that the periodic evaluations should not be overly far apart.

116. **Proposal** with respect to frequency of evaluations:

- The Fund should commission an evaluation of forecasting accuracy and the forecasting process every five to seven years.

R. Coverage of Evaluations

117. Medium-term forecasts should receive more attention in future evaluations than hitherto.[17] Medium-term analysis of such issues as fiscal policy and public debt is receiving increasing attention and the underlying base-case scenario for such analysis will be the Fund's medium-term forecast. Moreover, as time passes the enlarging sample could be used to analyze the accuracy of past medium-term forecasts. In this context, it is worth noting that assumptions and conventions are even more important in medium-term than in short-term forecasting.

118. While the accuracy of forecasts should continue to be a key topic of evaluation, increased attention should be paid to finding ways of sensitizing forecasters, in particular country desk economists, to the implications of structural changes in the economy. This message was central to Faust's study and should be an increasingly important element of future evaluations.

119. **Proposals** with respect to coverage of evaluations:

- More attention should be paid to medium-term forecasts.

- Increased attention should be paid to finding ways of sensitizing forecasters to the implications of structural changes in the economy.

S. Process of Forecasting

120. The five studies give relatively little attention to the IMF forecasting process as opposed to its statistical results, and few requests seem to have been made for such a discussion in the arrangements for the evaluations. The terms of reference for Timmermann (2006) included two questions about process:

> "…issues of the postmortem could be related to the current structure of the *WEO* process. Recognizing the limited degrees of freedom with regard to change (e.g., resource constraints, area department primacy for country forecasts), the postmortem could include a review of the following issues."

- Are the set of global assumptions provided to desks adequate and sufficient? Are the forecast procedures for assumptions appropriate? What have been the forecast errors for key assumptions? Are these errors correlated with errors in other variables, such as output, for example?

[17] As mentioned earlier, forecasts out to five years in the future were introduced into *WEO* figures in October 1996 and the values of the forecasts for the fifth year were introduced into *WEO* tables in April 2008.

- Are forecast consistency checks conducted by RES adequate? Could they be extended so as to reduce forecast errors for key variables?

But even Timmermann's paper does little to address these issues.

121. Consideration should be given in future studies to assessing the forecasting process as well as its results. If the Fund decided to move in this direction it should consider having the consultant be present in the Fund for an entire forecasting round. As David Stockton (2012) noted in the introduction of his review of the forecasting capability of the Monetary Policy Committee (MPC) at the Bank of England, "In preparing this review, I studied the materials provided to the MPC in the development of the forecast, and I attended the key meetings held by the staff and MPC during the production of the forecast for the August Inflation Report."[18]

122. With the increasing use of models by RESs, it would appear that there has been an increase in the top-down element of the *WEO* forecast. This strengthens the argument that the assessment of the overall process should be an important part of future evaluations.

123. **Proposals** with respect to the process of forecasting:

- An important task of future evaluations should be assessing the overall process underlying the *WEO* forecast exercise, including the balance of bottom-up and top-down elements of the process and the increasing use of models in the top-down element.

- Consideration should be given to having the author of each future evaluation present in the Fund for an entire forecasting round.

T. Availability of Results

124. The results of future forecast evaluations should be made readily accessible to IMF country desk economists. Statistical tests of forecast accuracy could easily be programmed into a template that could be linked to the *WEO* data base on forecasts and outturns and made accessible to staff (and possibly to others). Any time staff members wanted to see the summary statistics on forecast errors for a country, compare them to those of peer groups, etc., they could access the template and produce the results for whatever time period they desired. A Fund-wide assessment of forecast accuracy could be carried out by the Research Department (RES) once a year (or once every other year) for all countries.

[18] Stockton also interviewed all the current members of the MPC and all former members who had served on the committee since 2007. He interviewed many members of the staff of the Bank of England, and consulted with individuals from the academic community, financial institutions, private consulting firms, and official institutions.

125. The advantage of creating a tool for in-house forecast evaluation would be two-fold. First, it would give staff members (and possibly others) an easy way to find information about forecast performance. Second, it would permit changing the emphasis of future external commissioned studies away from the computation of huge numbers of summary statistics (although explanations of any biases in *WEO* forecasts would still be useful) towards "big picture" issues such as the introduction of new ways of thinking about learning from forecast errors and the evaluation of the forecast process. The commissioned evaluations would thus be able to take a broader overall view of *WEO* forecasts rather than producing summary statistics whose production could easily be automated.

126. **Proposal** with respect to presentation of results:

- The statistical tests applied in the commissioned studies should be programmed into a template that could be linked to the *WEO* data base on forecasts and outturns and made accessible to staff; consideration should also be given to making it available more widely.

- Reports should be prepared for country desk economists that enable them to take into account the results of the evaluations and the concerns that these raise about forecast accuracy and the importance of sensitizing forecasters to structural changes.

III. Conclusions and Summary of Proposals

127. This assessment of the five evaluations of *WEO* forecasts found many positive features. The technical statistical quality of the evaluations was very good. Over time, and especially in the two most recent studies, the recommendations expanded in scope, beyond technical statistical suggestions for improving the forecasts to suggestions for using the test results to improve forecasting. In most cases the authors of the study met with staff to discuss the report and its recommendations. The studies influenced the way in which Fund forecasters approached their task, although it is not always easy to pinpoint the exact effects.

128. The assessment also found that the evaluations were insufficiently documented in a variety of areas. Only one had written terms of reference. And none reported on follow-up actions the IMF had taken in response to previous studies.

129. Future such studies should pay more attention to the forecast process, as well as to the quality of its results. Also, following the recommendation in Faust (2013), increased attention should be paid to finding ways of sensitizing the forecasters to the effects on the forecasts of structural changes in the economy. Ways to do this might include automating the production of regular reports that would quickly reveal patterns of forecast errors.

130. Periodically the IMF should reconsider the appropriateness of the nature of the policy assumptions and conventions that underlie the *WEO* forecasting process.

131. The IMF should commission an evaluation of forecasting accuracy and the forecasting process every five to seven years. Recommendations for these future evaluations are as follows.

Orientation for authors

132. Like their predecessors, the authors of future commissioned studies should meet in advance with the management and staff of the Research Department and should also meet and have extensive discussions with area department economists, particularly those for the major countries. Looking ahead, the IMF should consider having the author of the evaluation present in the Fund for an entire forecasting round.

Content/scope

133. Each future study should:

(i) Have explicit written terms of reference.

(ii) Examine medium-term forecasts (those for periods up to five years) as well as shorter-term forecasts.

(iii) Assess the process underlying the *WEO* forecast exercise, including the balance of bottom-up and top-down elements and the growing use of models in the top-down element.

(iv) Continue the practice of comparing *WEO* forecasts with other official and private sector forecasts, such as those issued by Consensus Economics.

(v) Focus its analysis mainly on groupings of countries—industrial countries, and regional groupings of emerging and low-income developing economies—and on individual very large countries, say the 10 or 15 largest countries in the world.

(vi) Treat the statistical results for IMF program countries separately from those for non-program countries, noting that for countries subject to IMF programs the Fund's forecasts may not be readily comparable with those of other forecasters.

(vii) Given the longer span of data now available on past forecasts, continue the practice in the recent studies of dropping the data from the earliest years, especially for emerging and developing economies. Continue to use subsets of data to assess the accuracy of the forecasts.

(viii) Acknowledge and elucidate the assumptions on which the *WEO* forecasts are based. Examine whether the *WEO* incorporates an assumption about the time period over which the output gap is closed and whether this could lead to biases in the *WEO* forecasts. Perhaps emphasize that the private sector forecasts collated by Consensus

Economics are unconditional while the *WEO* forecasts incorporate policy assumptions.

(ix) Continue to make use of advances in statistical literature and take advantage of whatever statistical tests become available to throw light on forecasting accuracy.

Report format

134. Future studies that evaluate IMF forecasts should:

(i) Set out their agreed terms of reference in an annex to the report.

(ii) Set out explicit recommendations in one section of the report.

(iii) Document, early in the report, the management response to the previous report and the extent of its implementation, and to the extent possible, document the effects of the previous study on the forecasting process.

(iv) Focus the analysis in the text on the principal variables—typically output growth and inflation, unless when the study is commissioned there are additional specific variables on which details are needed.

(v) Provide text tables showing the main results.

(vi) Provide a summary, or an executive summary within the introductory remarks, that includes a statement of the main recommendations.

(vii) Provide a web appendix containing statistical results for all countries in the *WEO* database. This should be made available to the Fund's country desks and possibly to outside researchers, and to Executive Directors at a meeting after the completion of the study.

Follow-up

135. Following the publication of each study:

(i) IMF management should issue a short public statement through a press release or a box in the *WEO* indicating how it intends to respond to the recommendations made.

(ii) A meeting should be held with the IMF Executive Board to discuss the findings and recommendations.

(iii) The statistical tests applied in the study should be programmed into a template that can be linked to the *WEO* data base on forecasts and outturns and made easily accessible to staff and perhaps more widely.

REFERENCES

Artis, M. J., 1988, “How Accurate Is the World Economic Outlook? A Post Mortem on Short-Term Forecasting at the International Monetary Fund,” *Staff Studies for the World Economic Outlook,* World Economic and Financial Surveys, pp. 1-49 (Washington, DC: International Monetary Fund).

————, 1996, “How Accurate Are the IMF’s Short-Term Forecasts? Another Examination of the *World Economic Outlook*,” IMF Working Paper No. 96/89 (Washington, DC: International Monetary Fund).

————, 1997, “How Accurate Are the IMF's Short-Term Forecasts? Another Examination of the *World Economic Outlook*,” *Staff Studies for the World Economic Outlook,* World Economic and Financial Surveys, pp. 1-39 (Washington, DC: International Monetary Fund).

Bank of Canada, 2008, “Research at the Bank of Canada: Response to the External Review of the Bank’s Research.” Available at www.bankofcanada.ca/wp-content/uploads/2011/05/response_research_evaluation.pdf.

Bank of England, 2012, “Court Reviews,” June 15. Available at www.bankofengland.co.uk/about/Pages/courtreviews/default.aspx.

————, 2013, “Response of the Bank of England to the Three Court-Commissioned Reviews,” May 30. Available at www.bankofengland.co.uk/publications/Documents/news/2013/nr051_ courtreviews.pdf.

Barrionuevo, J. M., 1993, “How Accurate Are the World Economic Outlook Projections?” *Staff Studies for the World Economic Outlook*, World Economic and Financial Surveys, pp. 28-46 (Washington, DC: International Monetary Fund).

Carabenciov, I., and others, 2013, “GPM6—The Global Projection Model with 6 Regions,” IMF Working Paper No. 13/87, April 10. Available at www.imf.org/external/pubs/ft/wp/2013/wp1387.pdf.

Faust, J., 2013, “A Report of the Predictive Accuracy of the IMF’s *WEO* Forecast,” Version of February 5 .

Freedman, C., and others, 2011, “External Evaluation of the Directorate General, Research of the European Central Bank.” Available at www.ecb.int/pub/pdf/other/ecbresearchevaluationfinalen.pdf.

Genberg, H., A. Martinez, and M. Salemi, 2014, "The IMF/*WEO* Forecast Process," IEO Background Paper No. BP/14/03 (Washington: Independent Evaluation Office of the IMF).

Meyer, L. H., and others, 2008, *External Review of Economic Research Activities at the Bank of Canada.* February 1. Available at www.bankofcanada.ca/wp-content/uploads/2011/05/ext_review.pdf.

Meyer, L. H., and others, 2012, *Report of the External Evaluation Committee for the Research Department at the Bank of Israel.* August 13. Available at www.boi.org.il/en/NewsAndPublications/PressReleases/Documents/EEC%20Israel%20Report%20081312.pdf.

Stockton, D. , 2012, "Review of the Monetary Policy Committee's Forecasting Capability," Report Presented to the Court of the Bank of England (London: Bank of England, October).

Timmermann, A., 2006, "An Evaluation of the *World Economic Outlook* Forecasts," IMF Working Paper No. 06/59 (Washington, DC: International Monetary Fund).

————, 2007, "An Evaluation of the *World Economic Outlook* Forecasts," *IMF Staff Papers,* Vol. 54 (No. 1), pp. 1-33.

ANNEX 1. INDIVIDUAL COMMISSIONED STUDIES

This appendix examines in greater detail each of the five commissioned studies. The classifications are largely the same as in the earlier overall assessment of the studies as a whole. Because of the detailed nature of the examination, an appreciable amount of material has been quoted directly from the studies. In these quotations, I have attempted to capture the highlights of the studies. Nonetheless, I have been able to incorporate in the quotations only a relatively small amount of the analysis in the various studies since the text, tables, and graphs in those studies are very detailed and cover an enormous amount of material with respect to the results for individual countries as well as for groupings of these countries.

The awkwardness for the reader of having large amounts of text between quotation marks has led me to leave out the quotation marks and to use these introductory remarks to simply attribute the majority of the text directly to the author. Where comments in this part of the study are mine rather than those of the authors, I make that clear in the text.

Artis (1988)

Terms of Reference

There were no formal terms of reference for the first Artis study although there were clearly discussions in advance by members of the Research Department and the author. It was commissioned by the Research Department to document the forecast record (which had not been done previously) and to examine whether there was evidence of systematic bias (a concern of some Executive Directors) and how the Fund's record compared to that of similar organizations. There was no request for recommendations. The study was intended to be a technical analysis and to provide technical tools for looking at forecast accuracy. Put somewhat differently, the purpose of this study was to look at the forecast, which was assumed to be taking place in a reasonable stable world, and make sure that the quality of the forecasts was appropriate in the circumstances.

Introductory comments

In the introductory section of his report, Artis discussed the role of forecasts in the context of the IMF's responsibility in the area of cooperation in international macroeconomic policymaking and economic policy coordination. This role was the principal motivation for the examination of the Fund's forecasting track record based on projections published in the *WEO* and in similar publications circulated internally within the Fund before regular publication began in 1980. Having noted the importance of policy indicators in this area and the necessity of forecasting rather than relying only on current data as indicators, Artis went on to say that for the successful functioning of an indicator system, the degree of forecasting accuracy must be tolerably good, given the alternatives.

Artis emphasized the conditional nature of the forecast as an important characteristic of the forecasting framework. The *WEO* forecast was prepared on the basis of certain assumptions about “exogenous” variables—fiscal and monetary policy, exchange rates, and oil prices being the most important. The basic assumption about policies at the time was that “present policies” would be held unchanged during the forecast period, but this was interpreted to include any currently known announcements about future policy adaptations and might also “encompass certain policy adaptations or changes that seem likely to occur even though they have not been announced by the authorities” (quotation from *WEO* of May 1980). While exchange rates were initially projected at the *nominal* levels prevailing at a recent base date, by the time the report was written, the assumption was that *real* exchange rate levels would prevail over the forecast period. Oil prices were projected as constant in U.S. dollar terms at the time.

The reason for conditionality of the forecasts was the Fund’s desire to draw conclusions for desirable policy adjustment from the analysis of the future outlook. Also, and especially in the case of monetary policy and exchange rate projections, there was a concern that *WEO* projections might move financial markets in a way that would require member governments to react, which was seen as having the potential for embarrassment to the Fund.

A second characteristic of the framework is the relatively informal nature of the forecast. Because it was not based on a model of the sort used in national central banks, it did not permit the decomposition of ex post forecast errors into exogenous variables, judgmental, and model-based errors as could be done in a model-based forecast. Nonetheless, some attempt was made in the study to relate forecast errors to exogenous variable errors (as discussed below).

A final characteristic of the framework was that the *WEO* forecast had a consistency check not shared by national forecasters. Country desk-based forecasts, prepared against the environmental assumptions specified by the Research Department, were aggregated to check for consistency of their trade and balance of payments implications. Any identified discrepancies were then removed by an iterative process in which the country desk forecasts were successively revised until the consistency check was satisfied. This consistency check might not have been fully satisfied because of the discrepancy in the world current account.

Data set

Artis (1988) mainly focused on five key variables—real GNP/GDP growth; inflation; export and import volume growth; and the current account of the balance of payments. It also devoted some attention to the terms of trade.

Horizon of forecast and definition of vintage of realization or outturn data

In evaluating the accuracy of the forecasts using a series of statistical measures based on forecast errors, it was necessary to decide on how the realization or outcome of the variable being forecast should be defined. Artis rejected the use of latest available set of data available to the researcher for a number of reasons. First, these data would not be homogeneous in vintage because the outturns for the later part of the series would be less final than those for the earlier part of the series. Second, since economic series are often rebased, it might not be feasible to reconstruct the data on a consistent base. Third, policy was based on early and not subsequently revised data. While the study used the realization as defined by Artis, he also examined the effects of using other sets of realizations. Appendix B of his report replicated the results using "latest available data" and none of the more general conclusions arrived at in the study appeared to depend on the particular choice of realization series.

The study focused on two horizons. In current year (CY) forecasts, the forecast for year t was that made during year t itself. In year ahead (YA) forecasts, the forecast for year t was that made in year t-1. More precisely, the CY forecasts were made in the earlier part of the year (typically the *WEO* published in April or May but in some cases appearing a month or two later) and the realization or outturn was defined as the first available estimate, which was taken to be the figure reported in the following year's April or May *WEO*. In the case of YA forecasts, the outturn was identified with the "first settled" estimate, that which was available in the *WEO* of the following-year-but-one. That is, the forecast for year t made in October of year t-1 was compared with the data presented in October of year t+1.

Sample time period

The forecasts came from the published versions of the *WEO* (starting in May 1980) and from similar data for the earlier comparable unpublished documents. In the case of the industrial countries, the sample covers the 1971–86 period for the CY forecasts and the 1973–85 period for the YA forecasts. The sample for non-oil developing countries covers the period 1977 to 1986 for the CY forecasts and 1979 to 1985 for the YA forecasts.

Countries or regions covered

The study covered both industrial countries and developing countries. In the former, it evaluated the forecasts for the G-7 as a whole and for each member of the G-7, the aggregate for industrial countries as a whole, and for "Europe" as a group. For developing countries, it evaluated the forecast for developing countries as a whole, and for regional groupings of non-oil producing countries in Africa, Asia, Europe, the Middle East, and the Western Hemisphere.

Statistical measures used

Artis began by noting that absolute measures of forecast accuracy are useless in themselves and that they need to be related, on the one hand, to the standards of accuracy required by the purpose for which they are sought and, on the other, to comparable measures generated by alternative forecasting techniques.

He used the following forecast error summary statistics:

- the mean absolute error of forecast and its comparison with mean absolute value of the realized series.
- the root mean square error (RMSE).
- the Theil inequality statistic, which is the ratio of the RMSE of the forecast under consideration to the RMSE of an alternative forecast (a naive "no change" forecast in the text of the study and a ten-year moving average of the output growth or inflation variable in Artis's Appendix D).

He also examined the rationality of the forecast errors—based on the regression of the realization on the forecast $R(t) = a + bF(t) + u(t)$, where a perfect forecast would have the intercept equal to zero, the slope equal to unity, and a correlation coefficient of unity.

Artis (1988) also focused on whether there was systematic bias in the forecast taken over a long sample period where the bias (average error not equal to zero) was evaluated in his Appendix C from the regression of the error series on a constant term.

Finally, considerable attention was given to the comparison of the *WEO* forecast with those of alternative official forecasters—the OECD and national forecasters.

Statistical results

A. Industrial countries (individual **G-7** *countries;* **G-7** *as a whole; total industrial countries; Europe)*

A1. Output growth

The results appeared to be fairly satisfactory: the Theil statistics were all well below unity, implying that *WEO* forecast outperformed the naive no change forecast, and the average absolute errors were well below the mean absolute value of the output growth series itself. As might be expected, the CY forecasts were superior to the YA forecasts.

There was a finding of some bias when the data were pooled. The tests for bias (presented in Appendix C of Artis (1988) show a degree of output optimism in *WEO* forecasts. Although individual country output forecast errors were not significant, they were predominantly of the same sign so that on pooling a significant amount of bias was suggested—on the order of 0.3 percent in the CY forecasts and even higher in the YA forecasts. The output optimism appeared to have been most pronounced in the second half of the 1970s, reflecting the fact that the deceleration in growth in many countries was only gradually perceived as a break in the trend, rather than as a cyclical downturn.

A2. Inflation

The track record for the forecasts of inflation in industrial countries was marginally less satisfactory than that of output forecasts, although still overall highly acceptable. The superiority of the CY forecasts again stood out. There appeared to be no bias in the inflation forecasts, at least not if the 1974 YA error was excluded.

A3. Export and import volumes

The track record suggested little difference between the CY and YA forecasts. The overall results were reasonably satisfactory on the whole.

A4. Balance of payments

The record for balance of payments forecasts was considerably less reassuring than that for output and inflation. Forecasts were little better than the naive projection. The relative weakness of the balance of payments forecasts was not unexpected and was in line with experience of other forecasters. The problem was evidently related to the sizable fluctuations in the world current account discrepancy, particularly since the late 1970s. Also, the current account is the difference between two large flows, each of which has a volume and a price component. Relatively small forecast errors in any of the underlying volume or price changes could result in relatively large errors in the absolute difference between the nominal flows. Moreover, exchange-rate innovations might at times have contributed to the errors in current account projections.

A5. World trade and industrial countries' terms of trade

The CY forecasts were much more accurate than the YA forecasts. The same was true for terms of trade forecasts. In both cases Artis (1988) found strong evidence of inefficiency.

B. Non-oil developing economies (Africa; Asia; Europe; Middle East; Western Hemisphere; total non-oil developing countries)

The summary statistics for output growth and inflation forecast errors in developing economies clearly showed a much poorer forecasting track record than those for the industrial countries. For example, a majority of the Theil statistics for the YA forecasts for output growth exceeded unity, while half of those for the CY forecasts also did so. This indicated that a naive prediction of no change in output growth would have been a better forecast than the actual *WEO* forecasts. On the other hand, for export and import volume growth, the regional detail for the CY forecasts gave somewhat more reassuring results. The balance of payments forecasts provided some indication of weakness.

The results of directly testing for bias, both on individual regional results and on the pooled data for all the regions, suggested a tendency towards output optimism, at least in the YA sample. There was also some bias in inflation estimates for individual regions though this was not significant when the data were pooled.

Some tests on commodity prices showed that the variability of these prices was notably high and therefore it was not too surprising that the average absolute errors were also rather large. Even so, the forecasts compared well with the naive standard.

C. Summary of the analysis of WEO forecast errors for industrial and developing economies

First, industrial country forecasting appeared to be much better than that for developing countries. Second, among the industrial country forecasts the balance of payments forecasts appeared considerably worse than those for output, inflation, export volumes, or import volumes. Third, the CY forecasts were superior to the YA forecasts for the industrial countries. Fourth, the record appeared comparatively free from inefficiency for country-by-country and region-by-region forecasts, although upon pooling the data there was some evidence of output optimism, which was more pronounced for the YA than for the CY forecasts. Lastly, the relative inferiority of balance of payments forecasts and of the YA forecasts for industrial countries did not carry over to developing countries.

Comparison with OECD forecasts

WEO and OECD forecasts were compared for output growth, inflation, and the balance of payments on current account of the G-7 countries, individually and in aggregate. Overall, the OECD output growth forecasts emerged as slightly superior to those in the *WEO* while for inflation and the balance of payments the evidence suggested the opposite. On the basis of a technique called error triangles, Artis (1988) showed that the *WEO* forecasts were somewhat better than the OECD forecasts. In terms of the Theil statistics, the OECD forecasts dominated the *WEO* forecasts for France and Germany while the *WEO* forecasts for Canada dominated those of the OECD. The overall evidence suggested there was little to choose between the two sets of forecasts. Roughly speaking, the two organizations tended to make the same errors about the same variables for the same countries at the same time. There was little unexploited information in one forecast that could be utilized by the other. As the two groups of forecasters “breathe the same air,” exchange information, and maintain contacts with the same national forecast agencies, Artis did not find these results surprising.

Comparison with forecasts by national forecasting agencies

With respect to output growth, there was a fair measure of similarity between the *WEO* and national agency forecast errors but there were some exceptions, and the correlation coefficients between *WEO* forecasting errors and national agency forecasting errors were lower in nearly every case than those recorded between the *WEO* and OECD country forecasting errors. With 1974 omitted from the comparison, the *WEO* output forecasts generally appeared more accurate than those of the national forecasting agencies in the first subsample from 1973 to 1979 but less accurate in the second subsample from 1980 to 1985.

In the case of inflation, the *WEO* and national forecasts were generally somewhat more highly correlated than those for output growth.

There was no significant difference overall between the track records of the *WEO* and the aggregate national forecasts for output and inflation, either in the G-7 or Europe as a whole.

Artis concluded that, for the most part, *WEO* errors tended to be shared in some fairly large degree by other agencies. They appeared to be general products of the imprecise art of economic forecasting rather than errors purely specific to the *WEO*.

Explanatory factors for forecast errors

Artis (1988) used two approaches to try to explain the factors behind *WEO* forecast errors. First, using a narrative approach, he attempted to identify from an inspection of the error patterns the most significant error episodes, which were then examined in more detail using the *WEO* source documents. Second, an "innovation-accounting" approach was used where the error was attributed to deviations from the *WEO* projection of conventional assumptions about fiscal policy and oil prices.

The narrative analysis focused on four homogeneous sub-periods: the first round of oil price increases and its aftermath (1973 to 1975); the subsequent recovery (to 1978–79); the second round of oil price increases and its aftermath (1978–80 to 1982); and the "dollar shock" and the period up to 1985–86. This pursuit of episodic detail clearly suggested that the major forecasting errors could be associated with the first round of oil price increases. And the second wave of oil price increases also created some obvious problems for forecasters. But some large errors occurred that could not be explained in this way. For example, the effect of collective policies of restraint, especially on the monetary side in the early 1980s, appeared to have been underestimated. Also private sector responses, first to the pressure of high inflation and then to its decline, were similarly incompletely understood. Forecasters' caution, a tendency to miss some turning points, particularly those resulting from novel types of disturbances, or a tendency to understate the strength of the turnaround could also be discerned. These errors appeared to be widely shared by other forecasting agencies, national and international. In any case, most of the errors seemed small and generally were quickly corrected. Just how small is "small" in this context depended of course on the purpose to which the forecasts were put.

The attempt to relate the YA forecasting errors to unexpected changes in fiscal policy and in oil prices led to the following conclusions. Overwhelmingly, oil price changes (especially those of 1973–74) contributed to inflation forecast errors while unexpected oil price changes also accounted for a number of the output and balance of payments errors. By contrast, fiscal policy changes proved relatively unimportant in explaining forecast errors.

General comments and conclusions by Artis

The period since the inception of the *WEO* as a regular forecasting exercise was extraordinarily rich in economic upheavals, which made the odds against forecasting formidable. It should also be recalled that the objective of the *WEO* was not to forecast the most likely outcome but rather to provide conditional estimates of economic developments under the assumption of unchanged policies and exchange rates.

The forecast performance appeared to have been reasonably accurate, particularly for output and inflation, with the industrial country forecasts generally more accurate than those from developing countries. The results also showed that forecast accuracy was quite sensitive to forecast lead time, so the errors typically diminished as more information became available, particularly for the industrial countries.

The forecasts for output growth, both for industrial and for developing countries, appeared to have suffered from a degree of "optimism bias" in the sense that output forecasts were mostly on the high side in relation to realized values. Some of this reflected the fact that the slowdown in growth in the 1971–80 period was only gradually perceived to be a break in trend growth rather than primarily a cyclical phenomenon. Since 1980, output forecast errors were more evenly distributed. There was little evidence of inefficiency in the *WEO* forecasts, in the sense that the forecast errors could not be explained systematically by the level of the forecast itself and were not obviously statistically biased. The *WEO* forecasts also appeared to be efficient in the sense that they were generally incapable of being improved by adding information from the available forecasts produced by the OECD or by national forecasters. Forecasts for the current account of the balance of payments were inferior to those for output and inflation, at least for the industrial countries. The *WEO* forecasts did not generally provide any distinct improvement over those of national agencies in forecasting national output growth and inflation. There was a high degree of common sharing in the principal forecasting errors. Indeed, the largest of these were traceable to the two large oil price increases, especially to the first. Also there were turning point errors outside of these episodes, which appeared to be widely shared by national and international forecasters.

Recommendations by Artis for improvement of the quality of the *WEO* forecast

With respect to the question of whether the *WEO*'s forecast accuracy could be significantly improved, Artis noted that it would probably not be helpful to be overly ambitious. Nevertheless, there might be scope for improvement in several areas:

(i) The quality of an internationally consistent exercise in forecasting could be improved by a reduction in the magnitude and especially the volatility of the world current account discrepancy.

(ii) The sensitivity of forecast accuracy to lead time underlined the importance of promptly taking into account any new information that became available.

(iii) There was a question whether the accuracy of the *WEO* would be improved by more widespread use of formal, model-based methods, since these would reduce processing time and would allow more frequent ad hoc updates of the forecasts.

(iv) In any case, a more formal methodology, simply by being more explicit, would more easily allow constructive postmortem analyses of forecast errors, and thus help to improve forecast performance over time.

Distribution of Artis (1988) report

The report was distributed to the IMF Executive Board after its completion in late 1987. It was intended to serve the Board as a background report for future discussions on the *WEO*. Also, the findings were circulated to other departments and discussed extensively with area departments. Artis was present to make presentations and to discuss the results.

Effects of recommendations

The recommendations apparently had some considerable effect although it is difficult to pinpoint specific or immediate changes to the way the forecasts were prepared coming out of the study. But crucially, this study (and subsequent studies) helped build an internal consensus about the need to increase the frequency of updating the forecasts, especially when it was felt that turning points were occurring. In the course of the 1990s, the area departments (including their mid-sized country desks) gradually began to update their forecasts independently of their annual cycle of Article IV consultations. And the area departments improved their ability to make incremental updates for the smaller countries (using a simple forecast adjustment model). The introduction of mid-term *WEO*s was the most concrete example of how the *WEO* process became more responsive to changes in global economic conditions.

Barrionuevo (1993)

While this study was not commissioned from an outside expert, it has been treated by subsequent researchers as one in the series of studies evaluating *WEO* forecasts. Consequently, I provide a relatively short summary of the document.

Introductory comments

While there were no formal terms of reference for the study, Barrionuevo noted in the introductory remarks that large deviations from anticipated future growth and inflation might prove to be costly in terms of lost output and employment and it would therefore be important to assess whether forecasts were accurate, given the information available when they were made.

He suggested that a useful discussion about forecasting accuracy needed to provide a qualitative assessment of the way in which various forms of inefficiency in the projection were related. This involved distinguishing between unbiasedness and efficiency. He argued that under rational expectations the usual statistical test for efficiency was necessary but not sufficient to ensure efficiency and suggested simple adjustment factors to reduce the inefficiency of forecasts.

Barrionuevo also noted that the relationship between assumptions about policies, oil prices, etc., and deviations from projection outcomes were beyond the scope of his paper.

Data set

The study focused on output growth and inflation, a subset of the variables that Artis (1988) had studied. Barrionuevo also examined the accuracy of growth and inflation projections for the G-7 countries over the business cycles in the sample time period.

Barrionuevo used the same conventions as Artis (1988). That is, CY forecasts were published in the Spring of same year with the outcome published in the following Spring. And YA forecasts were published in the Fall for the following year and the outcome was the estimate published two years later, i.e., in the Fall of the year following the year for which the forecast was made.

Sample time period

Barrionuevo extended the sample period of 1971–86 used by Artis to 1971–91 for industrial countries. The sample period for developing countries was 1977 to 1991 and for non-program developing countries it was 1988 to 1991.

Countries or regions covered

Barrionuevo examined forecasts for each G-7 country, the G-7 as a whole, a group of 14 smaller industrial countries, the average of both large and small industrial country groups, each of the regional groups of developing countries in Africa, Asia, the Middle East, and the Western Hemisphere, the average of these developing country groups, and 36 non-program developing countries.

Statistical measures

Barrionuevo carefully analyzed the notions of unbiasedness and efficiency in the forecast. A forecast was unbiased if its average error was zero. This was a necessary and sufficient condition for unbiasedness and could be tested by regressing forecast errors on a constant. The necessary and sufficient conditions for efficiency were that the average forecast error was zero and that the forecast errors were not related to information available at the time the projections were made. The latter condition included the requirement that the errors be uncorrelated.

Barrionuevo defined an accurate forecast as one that was both unbiased and efficient. Unbiasedness was generally regarded as more important than efficiency because it meant that forecasts were identical to outturns on average and it was a necessary condition for efficiency.

Statistical results

A. **G-7** *countries (each country and pooled projections)*

The *WEO* CY forecasts for growth and inflation for the seven major industrial countries were unbiased for 1971–91. The CY forecasts for growth reflected an important structural change between 1971–82 and 1983–91. In particular, the 1971–82 forecasts of growth were biased upward, whereas those for 1983–91 were biased downward.

The *WEO* YA projections overstated growth and understated inflation by one-half of one percentage point each. This bias for the period as a whole occurred because YA forecasts overstated growth and understated inflation in 1971–82. After 1982, however, YA projections of both growth and inflation were unbiased across the seven major industrial economies.

Only CY forecasts of inflation were efficient. CY and YA forecasts of growth and YA projections of inflation were inefficient in the sense that the projections could be improved by adjusting them on the basis of the statistical properties of the forecast error.

The accuracy of the *WEO* projections for growth and inflation improved after 1985, the last year fully analyzed in Artis (1988). This improvement might have partly reflected a more stable environment in the 1980s than in the more volatile 1970s.

In the 1990–91 recession, the *WEO* projection errors were lower than in the two previous cyclical downturns, and the projections were generally unbiased, which was a distinct improvement over the forecasts for 1974 and 1982. Possible reasons for this difference were that supply shocks did not play a central role in the 1990–91 recession and that this recession was relatively shallow compared with the other two. Nevertheless, the *WEO* projections failed to anticipate the full extent of the 1990–91 downturn.

B. Industrial countries

For all industrial countries, YA forecasts for growth in 1971–91 overestimated actual growth by 0.4 percentage points on average, whereas the YA forecasts of inflation underestimated actual inflation by 0.3 percentage points. And CY and YA projection of growth were efficient for the grouping of industrial countries as a whole.

Theil statistics indicated that *WEO* projections of growth and inflation for industrial countries were superior to random-walk forecasts, except for inflation projections for the smaller countries.

C. Developing countries

There were significant forecast errors for growth and inflation in developing countries before 1985, but they were small for growth projections in the 1986–91 period. Although the economic environment was more stable in the latter period, the improvement in forecast accuracy suggested that policy assumptions had been more frequently met in those years.

The average forecast error for inflation in the developing countries as a group rose significantly between 1976–85 and 1986–91. This result was dominated, however by the errors in only a few countries.

For the sample of non-program developing countries, both inflation and real output growth projections were unbiased in the 1988–91 period.

The growth projections were generally efficient for the developing countries. In contrast, neither the CY nor YA inflation projections were efficient.

For developing countries in Fund-supported stabilization and structural adjustment programs, the projections assumed that the policies aimed at achieving growth and inflation objectives were adopted and implemented. Thus, deviations between conditional predictions of outcomes might be interpreted as a measure of the extent to which policies specified in the programs were not fully implemented, or as a reflection of the fact that the assumptions about the international economic environment faced by these countries were not always realized. Moreover, the economic situation of program countries tended to be, on balance, worse than that of non-program countries, making forecasting more difficult for the group of program countries.

Barrionuevo compared the results of statistical tests for the developing country group as a whole with those for 36 developing countries that were not engaged in Fund-supported programs. The results of the tests performed for the non-program countries were more comparable with the results for industrial countries, suggesting that the forecasting accuracy for program countries was relatively poor.

Indeed, Theil statistics suggested that *WEO* projections for both growth and inflation for the developing countries as a group (including program countries) were inferior to random-walk forecasts. Theil statistics for the pooled sample of non-program countries, however, suggested that the projections for these countries were superior to random-walk forecasts. This also suggested that there were unrealized policy objectives for some program countries.

Time-series forecasts

Barrionuevo developed time-series models with autoregressive and moving average components for output growth and inflation for each of the G-7 economies and for the average of these economies as well as for their pooled data. Typically, both past values of growth and past errors were significant determinants of growth in time-series forecasts for CY. While the Theil statistic indicated that the *WEO* projections were superior to random-walk forecasts, the projections from the more sophisticated time-series models were able to outperform the *WEO* forecasts in many cases.

Recommendations for improvement of quality of forecast

Barrionuevo's above results with respect to time-series models suggested that the accuracy of the *WEO* judgmental projections could be improved by using the statistical properties of such model-based methods to incorporate previous years' errors into forecasts of growth and inflation for the current year. This approach roughly corresponded to the error-correction mechanisms present in time-series models.

Barrionuevo suggested that the *WEO* forecasts could be improved if they were adjusted by making use of the relationship of *WEO* errors and their own past errors. (i.e., the autocorrelation of error terms in the *WEO* forecast). He noted that failure to make adjustments for large errors would reduce significantly the accuracy of a projection.

Effect of such recommendations

While there were benefits from sensitizing forecasters to the autocorrelation of errors and asking them to focus on reasons for the autocorrelation and perhaps adjusting their thinking to take account of biases that were likely to continue over time, there appears to be no evidence that mechanical adjustment of forecasts on the basis of time-series methods was used in subsequent forecasts.

Artis (1996)

Terms of reference

There were no formal terms of reference for the second Artis study although there were clearly discussions in advance by members of the Research Department and the author. The study was commissioned by the Research Department as a follow-up study to document the forecast record, which had originally been done in Artis (1988) and to examine whether there was evidence of improvement in the most recent period.

Introductory comments

Artis (1996) noted that his paper reported results of the examination of the short-term forecasts produced by the IMF and published twice a year in its *WEO*. It followed the precedent of the earlier examination in Artis (1988) subsequently updated and supplemented by Barrionuevo (1993).

Artis added two cautionary notes. First, for many commentators the principal value of the *WEO* might lie in its analysis of the conjuncture, its diagnosis of the situation reached by the world economy and its evaluation of the options available to the world's policymakers—rather than in the fine detail of its short-run forecasts. Secondly, from the perspective of strengthening global economic policymaking and performance in the longer run, the IMF's medium-term projections and scenario analyses were arguably more relevant than the short-term forecasts.

The *WEO* forecast was not produced in a framework of an overall econometric model, so the forecast postmortem methods applicable to model-based forecasting were not appropriate. IMF procedures relied heavily on the provision of forecast information from individual country desk officers. Overall economic consistency was provided in two stages—first, by assuming common global assumptions to which country desks worked and, second, via the aggregation and resultant check for consistency by the Research Department of the individual-country output, trade, and balance of payments projections provided by the country desks. Inconsistencies revealed by the aggregation would result in iterations on the original country forecasts until an acceptable set of forecasts was arrived at. The global assumptions specified to the country desk officers in a *WEO* forecasting round would typically include the values to be assumed for oil prices and assumptions made regarding key monetary and fiscal policy variables and sensitive market variables such as exchange rates. In general, policy variables were taken to be given at current values or at publicly projected values if firm commitments had been made by the governments concerned.[1]

[1] The attempt in Artis (1988) to explain forecast errors by relating those errors to deviations in policy and environmental variables from the values set for the forecast was a difficult procedure and produced no positive results that were not already obvious. It was not repeated in Artis (1996).

Data set

This study used the same forecast horizon as in Artis (1988). The current-year forecast was the forecast for the year t appearing in the May issue of the *WEO* in the same year. The outturn was the "first available estimate" appearing in the *WEO* of May of the following year. The year ahead forecast for year t was found in the *WEO* issue for October of the year t-1. The realization was the value published in the *WEO* for October of year t+1 (first settled estimates).

While *WEO* forecasts were rich in detail, the study focused on GDP growth, inflation, balance of payments, and growth of imports and exports, as in Artis (1988).

Sample time period

The sample covered the period from 1971 to 1994 and was an extension of the series in the earlier study. The increased length of the series available enabled Artis to examine whether any significant change had occurred in the IMF's record over time, particularly in the interval since the previous study.

Countries or regions covered

The most detailed analysis and by far the larger part of the study was devoted to the forecast for the industrial countries group, specifically individual G-7 countries. The analysis of developing countries was confined to regional aggregates, the same as in the earlier study.

Statistical measures used

By and large, the measures used were similar to those in Artis (1988). But there was more emphasis on the bias in forecasting errors and the presence or absence of serial correlation in the errors, similar to the tests in Barrionuevo (1993).

In testing for efficiency, evaluators had generally concentrated on whether forecasters could have improved their forecasts by taking advantage of information from an easily available subset of data. One data source was the forecast variables themselves. This hypothesis was tested by regressions of realizations on forecasts, which were featured in the previous study and again in the current one. Alternative forecasting procedures that could be used to provide a benchmark against which to appraise the performance of the procedures under examination involved naive, or not so naive, time series forecasts. For these comparisons, Artis (1996) presented Theil statistics of the comparisons of *WEO* forecasts with the naive random-walk alternative and with a less naive alternative based on knowledge of the trend of the series, a comparison which had been made in the appendix of his earlier paper and featured in the text of Artis (1996). In addition to point estimates for such comparisons, a more recent extension of this form of testing in Artis (1996) provided significance tests in the form of mean squared error regressions.

Rather than the comparison with official forecasts (OECD and individual national official forecasters) made in his earlier study, Artis (1996) compared *WEO* forecasts with private sector forecasts. More specifically, he used projections published by Consensus Economics, although these began to be available only in the latter part of 1989.

His study included tests of directional accuracy and discussed some aspects of turning point forecasting in the most recent business cycle.

It also examined the extent to which forecast errors were general across the economies of the world. Interdependence between economies could result in synchronization of business cycles, leading individual national forecasters to commit forecasting errors of similar sign. He noted that the IMF should be better placed to internalize international interdependence in its forecasting procedures.

Statistical results

A. Industrial countries

A1. Output growth

Generally, the evidence indicated that these forecasts were not, on a country-by-country basis, statistically biased. The evidence seemed especially strong for CY forecasts of output growth. For CY forecasts Artis found some positive bias in the first sub-period (1971 to 1982) and some negative bias in the second sub-period (1983 to 1994), with no significant bias for the period as a whole. These conclusions were similar to those in Barrionuevo (1993). Nonetheless, all of the point estimates of bias in GDP growth rate forecasts were positive, suggesting that there might be a widespread error of output growth optimism. Indeed, when individual country observations were pooled, the result was a finding that there was significant positive bias in the YA forecasts of just over 0.5 percent a year. It appeared that this bias was overwhelmingly due to experience in the first sub-period. The bias was not significant in the second sub-period.

The output growth forecasts were almost entirely free of serial correlation in the errors.

Using Theil statistics, *WEO* forecasts were found to be superior to the naive alternatives posed. This was true of both naive alternatives—no change (i.e., random walk with no drift, that is, the same rate of growth as last year) and instant mean reversion with value equal to the trend. Also, the performance of CY forecasts was notably better than that of YA forecasts.

A2. Inflation

These forecasts were not, on a country-by-country basis, statistically biased. The evidence seemed especially strong for CY forecasts of inflation. Forecasts for inflation did appear to suffer from serial correlation in the errors far more than output growth forecasts. Also, serial correlation affected errors for the G-7 inflation as a whole. This was especially true for YA forecasts.

Using Theil statistics, *WEO* forecasts for inflation were also superior to both naive alternatives. And the performance of CY forecasts was notably better than that of YA forecasts.

A3. Export and import volumes

These results were comparable to those for output growth

A4. Balance of payments

The results were much less satisfactory than those for output growth and inflation forecasts.

A5. World trade and industrial countries' terms of trade

The data for world trade strongly supported the efficiency of the *WEO* forecasts and they appeared to be superior by a margin to the two naive alternatives. For the terms of trade forecasts, the results were less reassuring. While superior to naive forecasts in RMSE terms, they were strikingly inefficient.

A6. MSE regression tests

These provided a procedure for examining the statistical significance of the difference between alternative forecasts. They therefore supplemented the point value of the Theil statistic. The tests confirmed the handful of particularly weak Theil statistic performances, especially in the balance of payments forecasts.

A7. WEO forecasts over a longer time period

The availability of a longer data set allowed Artis (1996) to address the question of whether the forecast record had improved over time. Although one could answer this question by simply inspecting error statistics, this did not allow for the possibility that the economy might have become easier to forecast. To allow for this, one could make a comparison with alternative forecasts.

Artis halved the entire sample into two subsamples, with a break between 1982 and 1983. For output growth, the mean absolute actual value, the average absolute error, and the RMSE declined in the second half of the sample, while the Theil statistic tended to rise. This might indicate that, with the less volatile economy, the random-walk forecast itself improved. For inflation, there were quite large declines in the mean absolute actual value, the mean absolute error, and the RMSE, while the Theil statistic values displayed little systematic change.

The balance of payments showed increases in the error statistics and in the Theil statistic. Moreover, forecasts of export and import volume growth also showed increases in the Theil statistic in the second sub-period.

In sum, the summary statistics did not afford a basis for a strong verdict either way on whether forecasting errors had fallen over the period. Barrionuevo's conclusion that forecast accuracy had improved throughout the period was based on a data sample that omitted the most recent downturn and, more significant, did not attempt to control for changes in the stochastic structure of the world economy and, thus, in the "ease" or "difficulty" of forecasting.

A8. Directional accuracy

Artis (1996) used a nonparametric method to assess the directional accuracy of the *WEO* forecasts (in which the projections were assessed with respect to their ability to accurately forecast positive versus negative changes in output growth). He concluded that the *WEO* record in CY forecasting was reassuring. The record in YA forecasting was less good. The overall verdict on directional accuracy was therefore somewhat mixed.

A9. Forecasting the cycle

In order to examine the process of recognition of the cycle by forecasters and corresponding revision of forecasts, Artis (1996) examined revisions in a series of successive *WEO* forecasts as the cycle progressed. A systematic turning-point error was defined as an initial underestimate or overestimate of output growth followed by persistence in the same error with accompanying forecast revisions in the same direction. These were uncomfortably pervasive in the data. Thus, the year 1988 for example, a peak year in the real growth cycle almost everywhere, was a year in which the forecast process exhibited systematic underestimation for all G-7 countries. In the subsequent trough year, there was even larger systematic overestimation in most countries. This indicated that the *WEO* forecast took the evolution of the cycle on board too slowly.

The data for inflation typically revealed a pattern of systematic overestimation in the early 1990s, suggesting that forecasters only gradually became convinced about the efficacy of the global policies of disinflation set in place since the early to mid-1980s.

A10. Comparison with private sector forecasts

Consensus forecasts from Consensus Economics became available in late 1989. With so few data points available at the time of Artis (1996), it made little sense to process these data in the same way as the *WEO* data. In the absence of formal tests, Artis used scatter diagrams to compare the record of *WEO* and Consensus forecasts for the G-7 countries. For CY output growth forecasts, the two forecast error records were very similar. For YA output growth forecasts, both had a tendency to overestimate growth over the 1990 to 1994 period, with a somewhat greater propensity on the part of the *WEO* forecasts. For CY inflation forecasts, there was again little difference between *WEO* and Consensus forecast prediction errors and both sets were relatively small and unbiased. The YA forecasts of inflation showed a slight positive bias, with little to note between the *WEO* and Consensus forecast errors. More generally, the most striking point in the comparison was the qualitative similarity in the pattern of errors, with both sets of forecasts making the same type of error in the same years in the same country.

A11. Generality of forecast errors

Artis used the cross-correlation of forecast errors to examine the generality of *WEO* forecast errors across countries. Cross-correlations were perhaps smaller for output growth than might have been expected, although the largest ones reflected strong trading relationships between certain groups of countries.

The prevalence of negative correlations between CY inflation errors was striking. This was less marked in YA forecasts. It was possibly the result of unforeseen exchange-rate movements, but it was not clear why this should not also have been a feature of the YA forecasts.

Nearly all the correlations of export and import growth had a positive sign, indicating that the underestimation or overestimation of the buoyancy of trade as a whole was more important than idiosyncratic error. The prevalence of negative signs in the balance of payments forecasts was as expected because of the closed nature of the world economy as a whole.

While the forecasts of nominal GDP growth outperformed those of real GDP growth and inflation taken separately in the earlier Artis study (because of the negative cross-correlations of forecast errors between output growth and inflation), this result was not as strong for the second sub-period as for the first sub-period. This indicated that the innovations facing forecasters in the first sub-period were predominantly supply shocks and that these were less important in the second sub-period.

B. Developing countries

As in the earlier Artis study, forecasts for the developing countries were analyzed for five regional groupings (Africa, Middle East, Asia, Western Hemisphere, and Europe) and for one functional category—total non-fuel exporters.

WEO forecasts for these groups of developing countries were found not to be particularly accurate. Data for many of these countries were poor and tardy. And in some countries the economy had been undergoing dramatic structural change. Also, some of the forecasts incorporated data from countries under IMF stabilization programs, where the program targets were taken as the forecasts. Moreover, in many countries year-to-year growth and inflation rates could be extremely volatile. By and large, the results were very similar to those in found in Artis (1988).

The suggestion of bias was more widespread in YA forecasts, with positive growth bias in a number of regions and negative inflation bias. There was little evidence that forecast errors were auto-correlated. According to the Theil statistic, forecasts for output growth were little better on average than a random walk and the forecasts were not particularly efficient. The median data were better than the mean data—not a surprising result given some of the outliers in developing countries. Nevertheless the quality of the forecasts continued to leave a good deal to be desired with respect to conformity with weak efficiency desiderata and to conformity with acceptably low Theil statistic. While balance of payments forecasts for industrial countries were notably weaker than output growth and inflation forecasts for those countries, this was not obviously the case for the developing countries.

WEO forecasts of commodity prices (other than fuel) passed statistical tests but their accuracy was not high. The overall conclusion was that the forecasts for developing countries were distinctly weaker than those for the developed industrial group. The findings qualitatively repeated the conclusions of the earlier Artis study.

General comments and conclusions in study

The overall conclusions of Artis (1996) were not dissimilar from those in Artis (1988). His concluding remarks raised the issue of whether any improvement was detectable through time in *WEO* forecasting. He noted a number of reasons why there should be an improvement (accumulation of experience, significant advances in data processing that should improve timeliness, and the competition offered by the increase in economic forecasting practice around the world). At the same time it was clear that there had been important changes in the structure of the world economy.

The earlier study indicated that there was some improvement in the forecast following the second oil shock relative to the first oil shock. It also seemed reasonable to excuse forecasters for not having foreseen the oil price increases. In the second sub-period, the prevalence of supply shocks was not so obvious, and the major world boom towards the end of the decade followed by a deep recession appeared to be endogenous to the development of the economy in a way that provided fewer obvious "excuses" to forecasters. Indeed, the greatest weakness of the subsequent forecast record lay in the failure to anticipate the major world boom towards the end of the 1980s and the subsequent deep recession.

Recommendations for improvement of quality of forecast

There were no explicit recommendations in this study as opposed to the suggestions in the earlier study.

Distribution of report

As was the case in the earlier study, the findings were circulated to IMF departments and discussed extensively with area departments. Artis apparently was present to make presentations and to discuss the results.

Effect of such recommendations

While the two Artis studies may well have had a number of effects, as indicated in the comments on this subject in the notes on the previous Artis study, the lack of explicit recommendations in this study precludes any judgment on this issue.

Timmermann (2006)

Terms of reference

Timmermann (2006) was the only evaluation document that had written terms of reference (TOR). The TOR for Timmermann's study (shown as Annex 2 below) requested the standard analysis of short-term errors along the lines of Artis (1996) and an analysis of some issues similar to those assessed in the previous documents, such as how the *WEO* forecast had fared during the most recent downturn and recovery. But it also set out a number of additional requirements. Specifically, the TOR asked whether the *WEO* forecasts were too close to consensus, whether they adequately reflected international spillovers, why *WEO* forecasts for emerging markets were less accurate, how accurate the medium-term *WEO* forecasts were, and how accurate the forecasts were for net oil exporters and importers. The TOR also raised some issues about elements of the *WEO* process itself, in particular about the way that the process addressed the global assumptions and about the nature of the forecast consistency checks. These questions related to the interaction between the *WES* division in the Research Department and the desk officers.

Introductory remarks

The *WEO* was a key source of forecasts of global economic activity and a key vehicle in the IMF's multilateral surveillance activities. Given the central role of the *WEO* forecasts, it was important that they be evaluated periodically, both to assess their usefulness and to look for ways to improve the forecasting process.

Timmermann (2006) featured three main novel aspects. First, it would analyze forecasts for 178 countries, rather than just regional aggregates for many of these countries. Second, given the substantially longer time series of Consensus forecast data, it would be able to include an extensive comparison between the accuracy of *WEO* forecasts and Consensus forecasts. Third, it would consider the revisions to the forecasts, both over time and within each forecast round. Moreover, the report would look at both CY and YA forecasts in April and September. (The previous studies focused only on April for CY and only on September for YA.)

The introductory section included a summary of the main findings of the report and the recommendations arising from the report. Such a summary is very useful for those readers who do not have the time or inclination to read through the full report.

Data set

This study focuses on the same five variables as in the earlier commissioned studies—real GDP growth, inflation, current account balance, and export and import volume growth.

Sample time period

The coverage for much of Timmermann (2006) was for the period from 1990 to 2003. The longer period, from the early 1970s to 2003, was used in the section on the long-term forecasting performance for G-7 countries. While Timmermann did not explicitly say why he largely excluded the period 1970-89, there appear to have been two main reasons. First, since he provided the same type of analysis for advanced economies and developing economies, data availability for the developing economies restricted him to the shorter period. Second, both the advanced economies and developing economies had undergone major structural changes through the longer period. In the case of the advanced economies, the inflation formation process had clearly changed, in line with the lower expected and actual rate of inflation. In the case of the developing economies and emerging economies, even if the data had been available, the move away from controlled economies to market-based economies meant that information about the accuracy of forecasts in the earlier period would not have been not particularly helpful in improving the accuracy of the forecasts going forward.

Countries or regions covered

Timmermann (2006) analyzed the forecasts for 178 countries in seven economic regions (Africa, Central and Eastern Europe, CIS countries and Mongolia, developing Asia, the Middle East, Western Hemisphere, and advanced economies). This was the first study that included individual developing economies in the analysis—something that could be helpful to the IMF staff at the country desks of those countries.

Timmermann presented the empirical results somewhat differently from those in the earlier studies. He did not make the sharp differentiation between advanced economies and developing economies that was a feature of the earlier studies. Rather, he presented results for the seven country groupings in many of the tables. A limited number of statistics were presented for each of the 178 countries included in the analysis. Longer sample comparisons were made for the G-7 economies because of the availability of *WEO* data for such economies for the longer time period. And the comparison of *WEO* and Consensus forecasts was done for those countries for which Consensus forecasts were available—the G-7, seven Latin American economies, and nine Asian economies—again without distinguishing in the customary way between advanced economies and developing economies.

Statistical measures used

Timmermann presented the standard statistical measures that had been used in earlier studies. These included tests for unbiasedness, absence of serial correlation, and efficiency properties (i.e., no variable in the current information set should be able to predict future forecast errors). In addition, he evaluated the pattern of forecast revisions, making use of the fact that he was examining the forecasts for CY and YA for both April and September. The evaluation of forecast revisions had the benefit of not requiring a decision to be made on which definition of realization was best. The final property under examination was the non-increasing variance of forecast errors as the forecast horizon decreased (i.e., the expectation that the variance of the forecast error should decline as more information became available).

In the analysis of statistical significance, Timmermann used bootstrapping to develop measures of the statistical significance of some of the results. This approach was more reliable than standard test statistics for small samples. He was thus able to make assertions with greater confidence about the systematic tendency of some of the results, e.g., whether the bias results were significant.

Use of all four estimates for any given year (two YA and two CY estimates) allowed Timmermann to address issues such as whether (and by how much) the error in the forecast declined as the time towards the target date was reduced. Also, it allowed him to test another efficiency property, namely, that forecast revisions should themselves be unpredictable.

Statistical results

Overall, the report found that *WEO* forecasts for many variables in many countries met the basic forecasting quality standards in some, if not all, dimensions. However, Timmermann did have some important reservations about the results.

1. Output growth

Generally speaking, *WEO* forecasts for real GDP growth displayed a tendency for systematic over-prediction—that is, predicted growth, on average, tended to exceed actual growth. From a statistical perspective, these biases were most significant in the YA forecasts. The results also indicated that systematic over-predictions of real GDP growth were particularly prevalent in forecasts for countries with an IMF program. This tendency for over-prediction of growth performance was persistent over time.

The evidence suggested that *WEO* forecasts for some countries could be improved if more attention were paid to important international linkages. In particular, forecasts of U.S. GDP growth were positively and significantly correlated with CY forecast errors of output growth in a substantial number of advanced economies. The report also noted that, in some cases, accuracy problems appeared related to the *WEO* assumption at that time that an output gap would be eliminated after five years. In particular, Timmermann pointed to a predominant negative relationship between the output gap and the forecast error in GDP growth, notably for France, Germany, and Italy.

Focusing on the results in somewhat more detail, the mean of the CY forecast error was very close to zero for the advanced economies. Biases in the April CY forecasts were much larger and negative (over-prediction) for Africa, Central and Eastern Europe, CIS and Mongolia, and the Middle East, partly because of large outliers. Not surprisingly, there was a significant reduction in the bias of the September CY forecast relative to the April CY forecast. Timmermann also used more robust statistics such as the median forecast error and proportion of under-predictions to check against the results of the classic mean measures of bias.

The biases in YA forecasts generally exceeded those in CY forecasts. Use of the broader range of statistics also suggested that the *WEO* in general over-predicted YA GDP growth and that the over-prediction was quite sizable. Serial correlation in the forecast errors also appeared to be a problem in some regions.

On average, the September forecast was revised downward when compared to April values. This was consistent with the April and September forecasts both over-predicting GDP growth on average, but with the April forecast being more optimistic than the September value. Timmermann suggested that this information could be used to improve the growth forecasts. Not surprisingly, information arriving between April and September more strongly affected CY than YA forecasts.

2. Inflation

The report noted a bias toward under-prediction of inflation, with this type of bias significant in the YA forecasts for many African, Central and Eastern European, and Western Hemisphere countries. The under-prediction bias was generally found to be weaker in the CY forecasts. Timmermann also used more robust measures to reduce the influence of outliers. Nonetheless, there were was a tendency towards under-prediction of inflation in Africa, Central and Eastern Europe, and CIS and Mongolia.

There was a tendency for both the CY and YA inflation forecasts in the *WEO* to be raised between April and September. Since the September forecasts were generally more accurate than their April counterparts, this suggested that the April *WEO* inflation forecasts could be improved by increasing their value.

3. Export and import volumes

Data on export and import volume in a number of regions were strongly affected by outliers. For the regions not affected by outliers, the bias appeared rather modest. Furthermore, the September CY and YA forecast errors generally had a smaller standard deviation than the corresponding April values, suggesting that information arriving between April and September could be used to improve the forecasts by taking account of the typical change between the two forecasts.

4. Current account balances

There appeared to be fewer problems in the forecasts for current account balances as percentages of GDP, except for April YA forecast errors, which, in some cases, were significantly biased or serially correlated. Moreover, general patterns in the direction of biases were not apparent.

5. Countries with IMF programs

Timmermann found that a potential source of bias in the *WEO* forecasts was whether or not a country was engaged in an IMF program. There were systematic over-predictions of GDP growth in program countries. The upward bias was smallest for the September CY forecasts and largest for the April YA forecasts. Although the bias estimates appeared large, it should be borne in mind that so were the average biases reported for countries in those regions hosting most of the program countries.

For the inflation forecasts, a large and systematic bias was again observed for program countries. However, the results showed that the bias went in the opposite direction relative to that observed for GDP growth, as the inflation rate was under-predicted. Again the largest bias was observed at the longest forecast horizon, i.e., for the April YA forecasts.

Can *WEO* forecast errors be predicted?

The process by which the *WEO* forecasts were produced put considerable emphasis on integrating predictions across countries, regions, and variables in order to produce a coherent and internally consistent projection of current and future economic activities. Timmermann tested for informational efficiency using a range of indicators of global economic activity. The four indicators he used in the empirical application were U.S. GDP growth forecasts, German output growth forecasts, the *WEO* forecast of oil prices, and the global current account discrepancy.

There were only a few cases in which *WEO* predictions of US GDP growth appeared to be correlated with the forecast errors. However, the ones that were found were of considerable interest. For example, U.S. GDP growth forecasts were correlated with a significant number of CY forecast errors in advanced economies.

Timmermann tested for the possibility that the (implicit) convention related to the output gap played an important role in the *WEO* forecasts. If the assumption that the output gap was eliminated after five years turned out to be incorrect, one would have expected that the predicted value of the output gap itself could account for forecast errors. For example, if it took longer to eliminate the output gap than assumed in the *WEO*, then the *WEO* projections about growth would tend to over-predict output growth forecasts for countries with large output gaps.[2] His empirical results supported the view that the assumption in the *WEO* forecast that the output gaps would be reduced too quickly might lead to a prediction of greater output growth and hence to an upward bias in the growth forecast. In the case of inflation forecasts, the *WEO* under-prediction of inflation also tended to be associated with the size of the output gap.

[2] As Timmermann later noted in his section on the advanced economies, an alternative explanation for this outcome was that there was a structural or trend change in the economy.

Directional accuracy

Timmermann (2006) found that *WEO* forecasts were quite successful in predicting the directional change for CY real GDP growth and inflation, but somewhat less so for YA forecasts.

Revisions from Board to published forecasts

The revisions from the February and July Executive Board forecasts to subsequent publications in April and September respectively added considerable informational value, especially for G-7 country forecasts. The average reduction in forecast errors was appreciable for CY forecasts but much less for YA forecasts.

Recent performance of *WEO* forecasts: downturn in 2001 and recovery

WEO forecasts of output growth generally over-predicted growth in 2001 in all regions—which was consistent with the broad patterns among forecasters in earlier downturns. For 2002, the April and September YA *WEO* forecasts prepared in 2001 over-predicted growth in six of the seven regions, although revisions in the April 2002 *WEO* greatly reduced the forecast errors in four regions.

Forecast errors for inflation over the period were more volatile in many of the regions than the corresponding GDP forecast errors. However, the *WEO* inflation forecasts for the advanced economies were very accurate in all years.

Long-run forecasting performance for G-7 economies

Timmermann also examined the longer-term data set for the G-7 economies (from the early 1970s to 2003) and followed Artis (1996) in using the CY forecasts published in April and the YA forecasts published in September. He concluded that forecast accuracy had deteriorated somewhat since Artis (1996).

In particular, *WEO* forecasts systematically and significantly over-predicted economic growth for all the European G-7 economies and Japan during 1991–2003. In contrast, U.S. growth was under-predicted after 1990, although the bias was not found to be statistically significant. Inflation was strongly and significantly over-predicted for Canada, France, Japan, and the United States during the 1990s and 2000s, although it was under-predicted by a significant margin for Italy.

These findings had at least two possible, not mutually exclusive, explanations. One was that output growth and inflation had been subject to structural breaks, such as a break toward higher productivity growth in the United States. Another possibility was that the underlying assumptions—such as the assumption that the output gap would be eliminated over a five-year period—had led to biases.

Longer-term (five-year) forecast

Longer-term forecasts were not further pursued in the analysis because of the rather short data sample, which was unlikely to make a statistical analysis of long-term forecasting performance particularly informative.

Comparison of *WEO* and Consensus forecasts

The comparison of *WEO* forecasts and Consensus forecasts could serve as a yardstick against which *WEO* forecasts could be measured. It also raised the issue of whether a forecaster could do better by using both sets of forecasts. Timmermann (2006) carried out a detailed evaluation of the results of forecasting by the two agencies.

He compared the *WEO* projections to Consensus forecast projections for GDP growth, inflation, and the current account balance for 1990 to 2003. The analysis covered all the G-7 economies, seven Latin American economies (Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Venezuela), and nine Asian economies (China, Hong Kong SAR, India, Indonesia, Korea, Malaysia, Singapore, Taiwan Province of China, and Thailand).

Overall, the comparison suggested that the forecast performance of the *WEO* was similar to that of the Consensus forecasts: the CY *WEO* forecasts of GDP growth in the G-7 economies were generally less biased than the CY Consensus forecasts, but the bias in the YA forecasts was larger in the *WEO* than in the Consensus forecasts across the board. Timmermann highlighted, however, that the timing of the comparison with the Consensus forecast mattered. *WEO* CY forecasts generally performed quite well against CY Consensus forecasts reported in March and performed considerably better against the February Consensus forecasts. However, given the relatively long gestation lag in the preparation, they tended to perform considerably worse against the Consensus forecasts reported in April. With the possible exception of YA inflation forecasts, there was little systematic evidence that the overall *WEO* forecasts could be improved by modifying them to account for information embodied in the Consensus forecasts.

Forecast combinations

While the *WEO* forecasts performed quite well, the results indicated that in some cases the Consensus forecasts could help predict the errors in the *WEO* forecasts and thus in principle could help improve upon the *WEO* forecasts by combining them with Consensus forecast values.

One question that was raised was whether the *WEO* forecasts would improve if they differed more from the Consensus forecasts. For four of the seven G-7 economies, there was evidence that CY *WEO* forecasts of GDP growth could be slightly improved by pushing them further away from the Consensus forecast values. Gains from doing this were very modest, however. There was no evidence that the YA GDP forecasts could be improved in this manner. As far as the YA inflation forecasts were concerned, large gains could be obtained by pulling the *WEO* forecasts strongly towards the Consensus forecast values.

Recommendations for improvement of quality of forecast

Timmermann (2006) made five explicit recommendations.

(i) Timeliness of information is key to forecasting performance. There are systematic gains from using the latest available information. Therefore, staff should update projections just before publication.

(ii) There should be continuous monitoring of forecasting performance. This is particularly important at times of structural instability in some of the underlying variables. Given the presence of what appear to be systematic biases in forecasting performance for output growth and inflation, particularly after 1990, the possibility of instituting real-time forecasting performance indicators should be explored.

(iii) *WEO* forecasters should use bias-adjusted forecast as guidance. Timmermann recognized that this approach might be too mechanical and might suffer from its own deficiencies, for example, the assumption that the bias remained constant through time. Nonetheless, in his view a comparison of unadjusted forecasts with bias-adjusted forecast could help in enhancing understanding of the magnitude and direction of any biases that might exist.

(iv) There should be quantitative indicators of the risk attached to the forecast. Ideally, the forecast should include the presentation of full probability distribution of key variables over time.

(v) Staff need to review the output gap assumption regularly. Also, more frequent reviews of estimates of potential output growth may be needed.

Distribution of report

Some members of the *WEO* team in the Research Department met with Timmermann to discuss his conclusions and he also made some presentations. He did not meet with the country desk economists who actually prepare the forecasts, although his results and conclusions were provided to them. Also, there was a presentation by a staff member of the results of the study entitled "*WEO* Forecast Postmortem—Implementing the Recommendations of the Timmermann Report," at the *WEO* Kick-off Meeting on January 5, 2005.

Effect of such recommendations

Box 1.3 in the *WEO* of April 2006 noted the following:

> The report made a number of recommendations to improve the *WEO* forecasting process. These included: (1) *WEO* growth forecasts for some countries could be improved if more attention were paid to important international linkages, particularly with the United States; (2) the accuracy of the forecasts should be assessed on an ongoing basis by instituting a set of real-time forecasting performance indicators; (3) IMF forecasters should more carefully consider the historical forecast "biases" when making their forecasts; and (4) that the forecast process should be broadened to more explicitly consider the risks around the key central projections. Internally, the IMF has begun taking steps to implement the first three recommendations. The rest of this box discusses the fourth recommendation—forecast risks—and how these can be incorporated in the *WEO* process.

It went on to discuss the use of fan charts in the *WEO*.

Faust (2013)

Broadly speaking, Faust (2013) took a somewhat different approach to the analysis of *WEO* forecasts from those in the other studies, although there were also many similarities. While Faust presented many of the traditional statistical results, he was skeptical about how useful classic, conventional statistical tests are from the standpoint of improving forecasts. His concern focused on structural changes in the economy rather than cyclical movements.[3] By way of example, he pointed out the importance of focusing on such matters as the trend productivity increase in the United States in the late 1990s, the global financial crash in 2007–08 (which almost no one forecast), declines in level or growth rate in potential output related to the financial crash, and the more-than-a-decade-long Japanese deflation. Faust offered a number of recommendations for improving *WEO* forecasts. As he noted in the conclusion to his paper, "major gains in precision will be hard to achieve, but this report provides some suggestions that may help *WEO* forecasters spot and understand errors, and may help in improving this accuracy."

Terms of reference

Similar to all the earlier studies except Timmermann (2006), there were no written terms of reference. However, there was considerable discussion by staff members and Faust about the objectives of the study.

Introductory remarks

In his introduction, Faust noted that the earlier evaluations had taken the approach of assessing the forecast record by applying a standard framework of forecast efficiency tests designed to determine whether the *WEO* forecasters were efficient in the sense of making the best possible use of information available to them at the time of the forecast. Where inefficiency was detected, these reports offered recommendations for corrective action. And he noted that Timmermann was a "tour de force" in its breadth of coverage and in the range of tests applied.

In Faust's view, two broad conclusions stood out from this work. First, the *WEO* was a reasonable forecast in the sense of being broadly consistent with other well-respected forecasts, where they were available. Second, as Timmermann put it, there seem to be "several problem areas where it appears that the *WEO* forecasts can be systematically improved."

[3] From a policy point of view, he may have overstated the point somewhat since macroeconomic policy has an important countercyclical element to it. Also, while it is absolutely essential to do one's best to identify and estimate the size of structural changes, these are very difficult to forecast in advance and even difficult to identify while they are going on.

Faust agreed that the *WEO* could probably be systematically improved, but in contrast to earlier reports he argued that the efficiency tests gave, at best, a misleading signal about where any problems might lie and how best to resolve them. The tests should be questioned as much as the *WEO* forecast.

Faust offered an alternative line of thought regarding evaluation and constructive improvements in the *WEO*. He started with the view that the macroeconomy was imperfectly understood and underwent continuous structural change punctuated by various forms of upheaval. As a consequence, any reasonable real-world forecast process underwent more or less continuous evolution.

Standard forecast efficiency tests were designed to shed light on whether a fixed forecasting model (implicit or explicit) was correct or incorrect. This was the wrong question. Faust wanted to shift emphasis away from the question "is the model right?" toward the question "is the model changing appropriately in response to the environment?" This was especially important at the time the report was written because of the effects of the financial crisis.

Forecasting in many advanced economies had now become a matter of evaluating what would become the *new normal* and determining the pace at which we might proceed towards it. For most of the economies in the *WEO*, these questions—What should we take as "normal"? and Will we go there at some "normal" pace?—had become central.

As Faust noted, his report was limited in scope and objectives. Its objective was to review the record of the point forecast in the *WEO* with an eye to improving its overall quality. Like earlier reviews in the series, his report focused on the forecast record and did not delve deeply into the processes that generated the forecast, which had in fact changed over the last couple of years.

Data set

Faust (2013) evaluated forecasts for GDP growth and inflation. Unlike the earlier studies, it did not examine the forecasts for export or import growth or the current account.

Sample time period and country coverage

Faust extended the time period used by Timmermann, 1990 to 2003, by adding six years so that the end of the sample period was 2009. Like Timmermann, Faust used forecasts made in both Spring and Fall of each year. He also examined not only CY and YA forecasts but also the five-year-ahead forecast.[4] Forecasts and forecast errors were available for 169 countries.

[4] These forecasts began to be included in the *WEO* charts in October 1996 and in the *WEO* data tables in April 2008.

Faust noted that there had been tremendous structural change over the sample period—the breakup of the Soviet Union; the financial crises in Asia as well as in other countries; the gradual move from very high and variable inflation to low and stable inflation in many countries; many institutional changes meant to support and cement this transition; the burst of the asset price bubble in Japan; and the launch of the euro following an extended period of structural change to facilitate convergence under the Maastricht criteria.

The definition of the "outcome" of a forecast differed importantly in Faust's study from that used in the earlier evaluations. Faust took the outcome to be the value of the data as it stood at the time of the Spring *WEO* forecast two years after the year in question. Therefore, his results could not be compared directly with those of the earlier studies. However, Faust noted that while some results seemed to depend on the choice of outcome data, the main results in his report generally did not.

Statistical measures used

Faust presented more than the usual statistical measures of the data for growth and inflation*:* mean, median, mode, and standard deviation as well as the 10th, 25th, 50th, 75th, and 90th percentiles of these data. Similar measures were presented for the YA forecast errors in GDP growth and inflation for the *WEO* forecasts published in the Fall. Also, these summary measures were applied to data for all economies as a group and advanced economies as a group for the sample as a whole, along with corresponding measures for the decade of the 1990s and the decade of the 2000s. Use of these measures gave a much broader picture of growth and inflation developments over the period, including the amount of skew in the sample.

Statistical results

1. Output growth

Faust found that across all economies, both the mean and median growth errors were negative each year through about 2003, meaning that growth came in lower than predicted. On average, the *WEO* growth forecast was about 2 percentage points too high through the early 1990s.

For the sample period as a whole, the mean forecast error was 0.8 percentage points for all economies and 0.4 percentage point for advanced economies, both in the form of over-optimistic forecasts. The corresponding figures for the 1990s were 1.1 percentage points for all economies and 0.1 percentage point for advanced economies. For the decade of the 2000s, the corresponding figures were 0.4 percentage points for all economies and 0.8 percentage points for advanced economies. In both sub-periods, these were also in the form of over-optimistic forecasts.

Finally, the recent crisis resulted in unprecedented forecast errors in output growth, with mean and median errors about 4 percentage points of over-prediction.

Faust presented a more detailed examination of output growth errors later in his study, focusing on the YA *WEO* forecasts published in the Fall. In this evaluation, he examined all economies, advanced economies, and G-7 economies for the full sample (1991 to 2008), two subsamples (1991 to 1999 and 2000 to 2008), and two other sample periods that added 2009 to the full sample and the second subsample. As with essentially all prior work on this topic, the results showed many rejections of the null hypothesis of no bias, with the share of rejections higher for advanced economies and G-7 economies than for all economies. The reason for the latter result is that, as an empirical matter, the narrower groups were more stable over this period and, even if the bias was smaller than for other areas, we might be more likely to reject efficiency. Also, adding 2009 to the sample appeared to have dramatic effects on the outcomes.

2. Inflation

Across all economies, it appeared that very high inflation was very hard to predict with precision. For the sample period as a whole, the mean inflation error for all economies was just over 25 percentage points in the direction of under-prediction. Nonetheless the median forecast error was only 0.3 percentage points, indicating that the very high inflation countries pulled up the mean value by a very considerable amount. The corresponding means for inflation forecast errors for all economies were almost 50 percentage points for the 1990s and five percentage points for the 2000s.

The size of forecast errors for the advanced economies was much smaller—0.1 percent mean over-prediction for the sample period as a whole, 0.3 percentage points over-prediction for the 1990s, and virtually no error on average for the 2000s. Apparently, forecasters did not predict as much progress on disinflation in the 1990s as actually occurred.

Finally, the drop in inflation associated with the recent global financial crisis was not only unprecedented, it was also not predicted. The mean and median forecast errors were both significantly negative in 2009 (over-prediction), a phenomenon not observed before in the sample. In this regard, the advanced economies look very similar to the full set of economies.

In the more detailed examination of inflation errors, there was the same pattern of different results, depending on whether 2009 was added to the sample, as there had been in the growth errors. Also, the first subsample was dominated by disinflation in a large proportion of the economies. Pooling a disinflation sample with a more stable sample made little sense from a statistical standpoint.

3. Two specific examples

Faust used two specific examples—Japan and Colombia—to illustrate the challenges facing forecasters over the sample period and to frame a discussion of their performance.

Japan presented a case of an economy facing fundamental change at the beginning of the sample—change that had persistent implications that were imperfectly understood at the outset. At the beginning of the sample period, there was the bursting of the apparent asset price bubble in Japan, leading to a period of serious economic challenges that continued to the present, with falling inflation at the outset and slow growth. Throughout the period, forecasters of the Japanese economy faced the questions "what will be the new normal?" And "at what pace will we move there?" The growth forecast evolved very gradually downward throughout the entire sample, suggesting that forecasters generally thought, at each point in time, that the new normal was about the same as the old normal. The growth forecast error averaged more than 2 percentage points from 1991 to 2003. The picture was very similar for inflation, where the forecast gradually moved down but averaged about 1 percentage point too high over this period.

Colombia illustrated the case of a country in an IMF program. It experienced substantial disinflation for some time, which the *WEO* forecast did not track very well. The growth forecast showed consistent over-prediction from about 1995 to 2000 and entirely missed the nearly 5 percent drop in 1999. The country faced internal security matters related to insurgencies and the drug trade, along with structural changes in fiscal arrangements in the early 1990s. At the end of the 1990s, the situation began to improve with another round of major structural change. These events were associated with significant volatility in inflation and growth.

Limitations of the conventional forecast-efficiency-testing framework

This was one of the key sections of Faust's study. He began by describing the statistical measures for forecast efficiency, testing forecast errors for bias, serial correlation, and predictability using information in the hands of the forecaster at the time of the forecast. These were the main measures used in earlier studies. He then argued that the standard efficiency-testing framework rested on the assumption that forecasters had full knowledge of the true structure of the economy and that there was nothing left to be learned about the structure.

In contrast, in a changing world in which forecasters had imperfect knowledge of the structure, learning and adaptation would be an essential part of forecasting. In such a case, we would expect the forecast to be biased, have serially correlated errors, and errors associated with variables known to the forecaster. These were the signals that the forecaster would take as evidence of a need to adjust. If forecasters made appropriate adjustments during the sample, failed efficiency tests based on a given sample did not indicate that any mistakes had been made during the sample period.

In short, Faust argued that results from the standard efficiency testing framework as applied to real-world forecasts should always be read with extreme caution. Rejections of efficiency should never be taken at face value and failures to reject efficiency should not provide much comfort.

Persistent forecast errors could arise when agents were forced to learn about the true structure. Faust noted that Timmermann's testing-based recommendations were carefully qualified in line with this thinking. Appropriate learning and adjustment remained a mixture of science and art.

A perspective on forecasting in a world of change

Faust presented a perspective on forecasting in which the main emphasis was on getting the typical value right and that downplayed efforts to exploit regular short-run dynamics to predict how the future may deviate from the typical value.

He began with an example in which the "typical" value (ignoring any short-run factors) was taken from the Consensus forecast (CF) for six to ten years in the future (Consensus forecast long-run forecast or CFLR). He then developed a hybrid forecast that started with the CFLR and added on some portion of any deviation between the CFLR and *WEO* forecasts. The goal here was to assess how much one gained or lost by adjusting the forecast of next year away from the CFLR view of typical. It appeared that treating the CFLR forecast from the previous Spring as the forecast of the coming year competed very favorably with the *WEO* Fall YA forecast. While the *WEO* forecast was often much more variable, the variability did not lead to clear gains in precision.

The main point of this exercise was that in a world of change, focusing on getting the typical *level* of the variable right might be the most important element of forecasting. Business cycle predictability might be extremely limited. There were two exceptions in which this was not the case. The first involved important sources of short-run predictability of a more ad hoc nature, such as natural disasters, crises, and strikes. The second was now-casting or CY forecasting.

Faust broke forecasting into three phases based on the horizon in question—first, long-run (or typical level) forecasting; second, information arriving on short-run factors that might shed light on the deviation between the outcome and the typical level; and third, now-casting.

In the standard perspective, forecasters knew the true structure of the economy and the first phase was straightforward. In the alternative perspective that Faust put forward, determining what should currently be viewed as the normal value was an important part of the first phase of forecasting. Systematic short-run predictability was very limited, so the main element in phase 2 was determining the implications of any known ad hoc factors. He then used what he called “error whisker plots”[5] to explore how forecasts evolved through the three phases, and to examine what these whisker plots would look like in the standard perspective and in the alternative perspective. Effectively, Faust was arguing that issues of persistent change—is there a new normal?—warranted more attention on the part of forecasters.

To implement this approach, Faust suggested that the process of learning might be facilitated by an ongoing system of basic quality control. The goal would be to make it as easy as possible for forecasters to spot problems and to begin to understand their sources. The system of monitoring could be implemented in an efficient low-cost way by developing regular reports that would help to alert forecasters to developing problems. It could be particularly useful to compare *WEO* outcomes with other available forecasts. For example, the reporting process used in producing the *WEO* could be expanded to include available Consensus forecasts and governmental forecasts at the time of the *WEO* forecast and to contain comparative information about the relative performance of the various forecasts. The standard report should also include standard efficiency tests, which could flag issues that might warrant further review.

Recommendations by Faust

1. Clarify the goals and nature of the forecast

First, should the forecast be a mean or a modal forecast? Second, the roles and importance of medium-term versus short-run forecasts should be clarified. Third, the nature of the forecast in program countries should be clarified.

As far as the third issue was concerned, Faust found that the apparent bias for countries on IMF programs was larger on average than that for other countries. He argued that since the forecast itself would play a role in negotiations over the conduct of policy in program countries, those responsible for the forecast were placed in an untenable situation to be involved both in formulating, negotiating, and implementing a policy and in giving an unconditional, public forecast of success. He suggested that for program countries the forecast could explicitly be a conditional forecast, conditioned on some version of “successful implementation” of the program. Alternatively, if the IMF wanted an unconditional forecast for program countries, external forecasters could be made responsible for the forecast. A minimal step that the IMF might consider would be simply to acknowledge that the forecasts of program countries were driven by a different set of criteria than other forecasts.

[5] These error whisker plots were constructed as follows. Consider six forecasts of each target year—two each during the target year and from one and two years ahead. The whisker plots involved collecting these six forecast errors for a given target year and plotting these errors against the date on which the forecast was made.

2. Implement a standard system of ongoing evaluation

Faust noted that in a world of ongoing structural change, the forecast process must adapt on an ongoing basis to new conditions. The IMF might investigate ways for forecasters to monitor on an ongoing basis the emergence of any systematic problems with the forecasts. These might include reports that reveal patterns of forecast errors and draw attention to the possibility of the longer-run forecasts being affected by structural change. Standard statistical tests, which should always be interpreted with extreme caution, could be used to flag issues for further investigation.

Meetings

Not applicable.

Effects

Not applicable.

ANNEX 2. TERMS OF REFERENCE FOR TIMMERMANN (2006)

TOR for the *WEO* Forecast Postmortem[6]

The evaluation should include a standard analysis of short-term forecast errors for *growth*, *inflation*, *current account balances, and export and import volumes along* the lines of Artis (1997) since this is the basis for further analysis and provides standard information that should be available anyway.[7] However, the value added of another postmortem would primarily arise from a focus on a few topical issues, which would be expected to lead to specific suggestions for improvements.

Topical analysis of forecast errors

Some of these issues would be related to the understanding of the factors underlying the forecast errors, as detailed in my earlier memorandum. In particular, the following tentative list highlights key issues that would be of interest to the *WEO* team:

(i) ***How did WEO forecasts fare during the most recent downturn and recovery?*** Did *WEO* forecasts succeed in anticipating the downturn and in predicting the timing of the upswing? Were the length and depth of the downturn and the strength of the subsequent recovery systematically under-predicted, as in previous episodes? If so, forecasts for which components of aggregate demand were particularly inaccurate? How do IMF forecasts compare with consensus forecasts?

(ii) ***Too much consensus?*** IMF forecasts are frequently close to consensus forecasts. This can be a plus given that averaging forecasts across forecasters—the consensus—tends to improve forecast accuracy. However, by being close to consensus, Fund forecasts also risk being trapped in a reputation game, where macroeconomic forecasters copy each other's forecasts in order to avoid being in a strong contrarian position without due attention to fundamentals ("don't be pessimistic when everybody is optimistic"). In this sense, Fund forecasts should perhaps be more contrarian at times, especially when downside risks or imbalances have not yet entered the radar screen of mainstream analysts. The evaluation could assess how IMF forecasts fared in episodes when consensus forecasts were off the mark by large margins and assess whether they were more reactive to risks and imbalances than consensus forecasts.

[6] For this reproduced version, the present author made a few very minor changes to the document to correct some obvious typos.

[7] The standard analysis would include the analysis of bias, serial correlation, efficiency of forecast errors, and directional accuracy.

(iii) ***Do WEO forecasts adequately reflect international spillovers?*** Regional and global spillovers are an important aspect of surveillance. The evaluation could assess whether *WEO* forecasts are efficient in the sense of taking such spillovers into account. For example, it could examine whether forecast errors for growth and inflation for industrial countries are efficient in the sense of being uncorrelated with past forecast errors or with actual values for these variables in other countries.

(iv) ***Why are WEO forecasts for emerging markets less accurate?*** Previous postmortems did not investigate forecast errors for individual emerging market countries in any detail. The evaluation could assess how the quality for emerging markets compares to industrial countries (on a country-by-country basis) and analyze reasons why the forecasts appear to be less accurate. For example, what role do financial crises, which are difficult to predict, play in large and possibly biased forecast errors? Are forecast errors for emerging markets with good and sufficient high frequency data lower and less biased than those with bad data? Is there a program country bias? How do IMF forecasts compare with consensus forecasts for emerging markets?

(v) ***How accurate are medium-term WEO forecasts?*** Area departments provide forecasts beyond the usual short-term horizon reported in the *WEO*. In general, the forecasts cover the current year *T* and the period up to *T+5*. While medium-term forecasts are generally not of immediate interest to the *WEO*, their accuracy is nevertheless important for other areas of Fund work, especially debt sustainability and achievement of Millennium Development Goals. So far, medium-term accuracy has not been investigated systematically, and the evaluation could make a first step in this direction. Besides establishing whether there is a bias in medium-term forecasts, the evaluation could also investigate whether there is inertia in these forecasts, that is, whether there is a tendency to adjust medium-term growth rates and other variables too slowly to changing trends.

(vi) ***How accurate are WEO forecasts for net oil exporters and importers?*** Oil shocks have asymmetric impact and long-run effects on country, depending on the sign of the oil trade balance. For the future assessment of the effects of oil shocks on global growth, it would be useful to have an assessment on the past forecast performance for the two groups of countries.

Analysis of elements of the WEO process

Other topics or issues of the postmortem could be related to the current structure of the *WEO* process. Recognizing the limited degrees of freedom with regard to change (e.g., resource constraints, area department primacy for country forecasts), the postmortem could include a review of the following issues.

(vii) Are the set of global assumptions provided to desks adequate and sufficient? Are the forecast procedures for assumptions appropriate? What have been the forecast errors for key assumptions? Are these errors correlated with errors in other variables, such as output, for example?

(viii) Are forecast consistency checks conducted by RES adequate? Could they be extended such that forecast errors for key variables could be improved?

The results of this analysis could suggest changes in the *WEO* process that may help in improving the forecast quality.