IEO
Independent Evaluation Office
of the International Monetary Fund

BACKGROUND PAPER

BP/11/02

Review of IMF Research on Monetary Policy Frameworks

Kenneth Kuttner, Petra Geraats, and Refet Gürkaynak

May 20, 2011

Address. 700 19th Street, N.W., Washington D.C. 20431, U.S.A. Telephone. +1 202 623 7312 Fax. +1 202 623 9990 Email. feedback@ieo-imf.org Website. www.ieo-imf.org

 BP/11/02

IEO Background Paper
Independent Evaluation Office
of the International Monetary Fund

Review of IMF Research on Monetary Policy Frameworks

Prepared by Kenneth Kuttner, Petra Geraats, and Refet Gürkaynak

May 20, 2011

Abstract

This study examines the technical quality of a sample of 60 IMF working papers on monetary policy frameworks. It found that the quality and value-added of IMF research on monetary policy frameworks varied considerably. Most of the working papers issued in 1999–2008 posed interesting policy-related questions and many were very skillfully executed. Some were cited extensively and made major contributions to the literature. Yet many of the papers were substandard, raising the possibility that some of the IMF's advice might rest on less than rock-solid research. Many of the flaws in the weaker papers stemmed from a lack of proficiency with empirical methods, or from failure to articulate a well-focused research question within the context of a coherent and appropriate theoretical framework, or from less than full and detailed description of the data and methods used to generate the results. The paper offers recommendations on screening, feedback, and documentation that could help address these weaknesses.

JEL Classification Numbers: E42, E52, E58, E61, and F33

Keywords: exchange rate regimes, inflation targeting, monetary policy transmission, policy rules, central banking

Author's E-Mail Address: kenneth.n.kuttner@williams.edu

Contents **Page**

I. OVERVIEW

1. The IMF's research agenda is diverse and ambitious. Even within the papers devoted to monetary policy frameworks, the research poses a wide range of interesting and policy-relevant questions. Many of the papers included deal with conventional topics such as monetary policy transmission, inflation targeting, and policy rules. Others address more idiosyncratic issues such as corruption, bimetallism, West African Economic and Monetary Union, and currency board design. Research methods range from pure theory to data description, and everything in between. Moreover, in line with the Fund's mission, IMF research deals with policy-relevant issues in virtually every country in the world. With its depth of expertise, institutional knowledge, and access to data, the IMF is uniquely situated to do much of this research—and in fact, some of it could have been done *only* at the IMF.

2. This report evaluates the quality of IMF working papers dealing with the general topic of monetary policy frameworks. In performing our evaluation, we were cognizant of the wide range of research topics and methods, the special challenges encountered in doing research on developing countries, and the special role that research plays in the formulation of the IMF's policies.

3. Our review led us to the following conclusions. First, the working paper series contains a number of first-rate papers, comparable in quality to those published in top-tier professional journals—in fact, many of them have been published in prestigious journals. Second, the working paper series contains many papers that, while not suitable for publication in academic journals, are nonetheless worthwhile and well executed. Third, the quality of the analysis varies greatly across papers. While the IMF produces some very good research, a significant proportion of the working papers in monetary policy frameworks suffer from some serious weaknesses. Those focusing on a particular country or region, most of which are based on selected issues papers (SIPs), tended to be among the weaker papers. Fourth, a common problem among the less well written papers is a lack of care in the presentation of quantitative material (charts, tables, etc.). And fifth, data sources and empirical methods are often not carefully documented. Most of these weaknesses are easy to remedy, however, and we believe ample opportunities exist for raising the average quality of the working papers and the research they contain.

4. Our recommendations fall under three headings. The first is quality control: more screening should be applied to the papers submitted for inclusion in the working paper series. We suggest a number of ways to do this: clear articulation of standards, consistent application of these standards to all types of papers, greater selectivity, a greater degree of accountability for supervisors signing off on submissions, and internal or external reviews. Our second recommendation is for more feedback and constructive criticism prior to submission to the working paper series. There is a great deal of research and writing expertise within the institution, and this should be harnessed more effectively to bring the weaker papers up to the IMF's standards. This involves institutionalizing mechanisms for

communication, collaboration, and constructive criticism involving research and country staff. Guidelines for clear exposition and presentation could also be helpful. The third heading is documentation. Because of its important role as an input to policy, IMF research should be as transparent and replicable as possible. Every paper published in the working paper series should therefore contain enough information on data sources and methods to allow other researchers to understand and correctly interpret the results. To that end, the IMF might consider establishing a repository for nonproprietary data at least to allow internal and external researchers to access the data used in the studies.

5. The production of high-quality research is an integral part of the IMF's mission, and the creation of working papers is an important component of the Fund's research output. Although one of our recommendations is to do more screening, ultimately the aim of our recommendations is to raise the quality of research, not just to keep weaker work away from the public eye. IMF working papers should be high quality and representative of the institution's research output.

6. The report is organized as follows. Section II describes the evaluation criteria, and how the panel interpreted them in the context of the IMF's research mission. Section III explains the categories into which the papers were grouped for comparison purposes. Section IV provides an overview of the evaluation ratings given by the panel and of other quantitative indicators of the quality of the 60 papers. Section V highlights a number of features shared by the strongest papers in the sample, and calls attention to some common sources of weakness. Section VI provides conclusions and recommendations. The annex summarizes the procedures used in conducting the review.

II. Evaluation Criteria

7. Each of the 60 papers had both a primary and a secondary reviewer (see Annex 1). The reviewers evaluated each paper according to 10 evaluation criteria, which were grouped into three categories, as shown in Table 1: first, the coherence and focus of the framework; second, the quality of the analysis; and third, the quality of the output. The framework category, for example, included assessments of the paper's contribution to the literature, and whether the research question was clearly focused. The analysis category included ratings of the proficiency with which the empirical and/or theoretical methods were applied, and whether the paper satisfactorily addressed robustness issues. The output category included criteria that included writing style and the degree to which the conclusions were grounded in the analysis and were relevant for policy purposes.

Table 1. Research Quality Indicator Form

Evaluation criteria	Rating[1/]				
	S	AA	A	BA	U
Framework					
1. Question is well posed and clearly focused					
2. Places work within the context of existing literature					
3. Specifies contribution to existing literature					
Analysis					
4. Uses an appropriate theoretical/conceptual framework[2/]					
5. Uses appropriate data and empirical methods proficiently					
6. Includes critical discussion and/or robustness analysis of results					
Output					
7. Writing is clear and well organized					
8. Adds value relative to existing research					
9. Conclusions are firmly grounded on the analysis					
10. Articulates policy relevance of findings					
Overall rating					

[1] The rating scale is as follows: "superior" (S);"above average" (AA); "average" (A); "below average" (BA); "unacceptable" (U).
[2] This includes whether there was excessive use of technique relative to the question being posed.

8. Papers were given marks ranging from "superior" (S) to "unsatisfactory" (U) on each of the ten criteria, with "above average" (AA), "average" (A), and "below average" (BA) falling in between. Each reviewer then aggregated those scores into a single score representing his or her judgment as to overall paper quality. The reviewers were not asked to use a uniform scheme in aggregating the individual scores, but the three categories of evaluation criteria received roughly similar weights.

9. The panel members agreed that the IMF's emphasis on practical policy-oriented research called for emphasizing different aspects of quality than those typically used in evaluating academic research. One dimension we paid close attention to in our review was the inclusion of an appropriate theoretical or conceptual framework. We were not as demanding on this criterion as we might have been if we were evaluating academic work, and we did not insist that every paper include a fully worked-out theoretical model. But in our review it became clear that the most effective papers were those that were able to frame their analysis with reference to a coherent theoretical framework. Those papers that did not include an explicit model invariably benefited from a thorough summary of the relevant theoretical considerations.

10. Technical proficiency is essential to quality research, of course. But some methodological qualities that are highly valued in academic research, such as theoretical novelty or technical innovation, received less weight in our evaluation than they would have in an academic context. We recognized that focused, well-executed research is highly valuable in a policy setting, even when the research is too incremental in nature to be of interest to an academic journal. Technique was valued to the extent that it furthered the analysis, and the use of "excessive" technique was not viewed favorably.

11. In judging the robustness of their analysis, we held IMF working papers to a higher standard than we would have applied to academic research. Our view was that since it is likely to be used as an input into policy decisions, research issued under IMF auspices need to be clear about the limits to the analysis, and what caveats apply. "Overselling" results is never advisable, but is especially inappropriate in policy research. The papers' use in the policy process also makes it essential that their conclusions and policy recommendations be based firmly on the analysis. Papers with conjectural, overgeneralized, or boilerplate conclusions were marked down.

12. We put a premium on clarity of writing and presentation, again reflecting the imperative to communicate the findings to policymakers and nonspecialists. We also expected empirical working papers to fully report their estimation methods and diagnostic tests, and to document the data sources used in the analysis. Enough information should be provided to allow a reader to correctly interpret and replicate the results.

III. PAPER CLASSIFICATION

13. To facilitate comparison, we divided the 60 papers into four categories. The first category consists of 24 empirical papers. Among these, 17 emphasize formal statistical analysis and the other 7 are primarily descriptive. Several of the descriptive papers involve the creation of new data sets, and their main purpose is to document the construction and describe the data set contents and characteristics.

14. Thirteen working papers can be classified as theory papers. Among them, eight can be characterized as "pure" theory, in the sense that they work out the implications of an original model, or develop an extension to an existing theoretical model without empirical testing or simulation of results. Five involve the simulation of calibrated macroeconomic models, most often to evaluate the performance of monetary policy rules.

15. The third category comprises 19 studies of particular countries or regions. Typical issues addressed in this set of papers included the suitability of inflation targeting, or the

choice of the appropriate exchange rate regime. Fourteen of these country studies were adapted from SIPs, and some even retain the SIP formatting and style.[1]

16. All the country studies involve either regression analysis or data description, and therefore could have been included in the empirical category. The reasons for breaking them out as a separate category are threefold. First, the purpose of country studies typically differs from that of more academically oriented research papers, in the sense that they are commissioned to address a specific issue of particular relevance to IMF policymaking. In that sense, they are even more important for the IMF than pure research that addresses questions in the general literature. Second, much country-specific research encounters obstacles such as limited or poor data. And third, because the country studies tend to share a similar set of weaknesses, it is useful to discuss them collectively.

17. The final category consists of four hard-to-classify papers, which we will refer to as "thought pieces" for lack of a better term. The four contain little or no original empirical or theoretical research; all involve some combination of literature survey and commentary.

18. Needless to say, some of the working papers, except for the thought pieces, could have been put into more than one category. A number of the papers develop theoretical models and then proceed to estimate or test the models empirically. These papers were classified according to our judgment about their primary contribution.

IV. Overview of Quantitative Indicators

19. Table 2 shows the panel's overall ratings for the 60 papers by both the primary and secondary reviewer, broken down by evaluation criteria, with each paper receiving two ratings. The overall rating, shown in the last column, was 3.2, slightly above the 3 score corresponding to the "average" rating. The median score fell into the "average" category. Most of the papers on the low end of the distribution were deficient on more than one of the criteria we identified. But even many of the highly ranked papers on our list also fell short on some criteria.

20. The uneven quality of the papers is reflected in the distribution of the scores, and the standard deviation of 0.9 suggests a considerable amount of variability.[2] The dispersion is immediately evident in the histogram of paper scores shown in Figure 1. The good news in the picture is the high proportion of papers scoring 4 ("above average") or higher: 20 papers

[1] The set of papers evaluated included 15 former SIPs. One of these was classified as an empirical paper, rather than as a country study.

[2] In order to eliminate the variance introduced by discrepancies between the ratings of the two reviewers for each paper, the average of the two reviewers' scores was used in calculating the standard deviation and the histogram displayed in Figure 1.

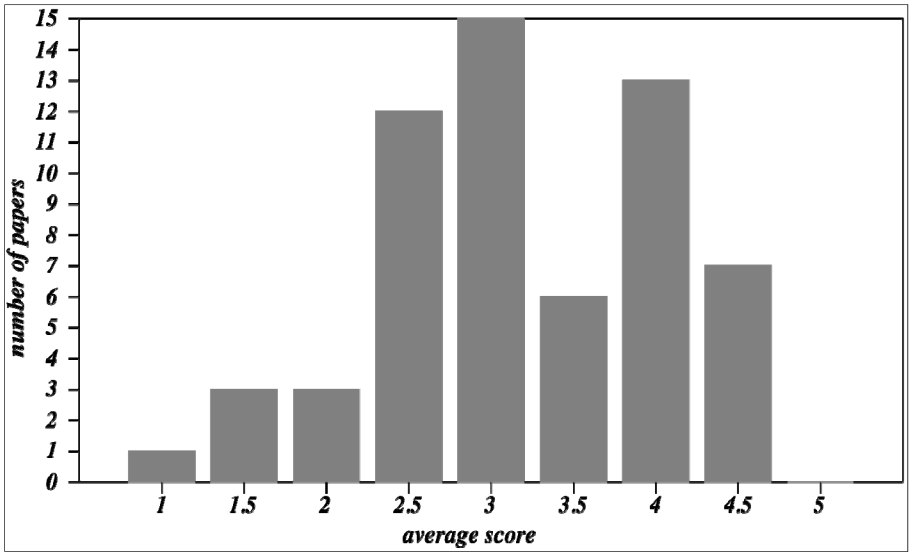
(33 percent) fell into this upper tail. The bad news is that 7 papers (12 percent) scored 2 ("below average") or below.

Table 2. Summary of Overall Ratings

Rating	Empirical	Theoretical	Country studies	Thought pieces	All
Superior	7	4	1	0	12
Above average	18	12	4	1	35
Average	15	8	17	1	41
Below average	7	2	14	4	27
Unsatisfactory	1	0	2	2	5
Total number of papers	**48**	**26**	**38**	**8**	**120**
Average[1]	**3.5**	**3.7**	**2.7**	**2.1**	**3.2**
Standard deviation[2]					**0.9**

[1] Based on a scale where S=5 and U=1.
[2] Standard deviation across papers of the average of the two reviewers' scores.

Figure 1. Distribution of Average Paper Scores



A. Ratings by Category of Paper

21. The breakdown by category reveals some additional insights. The average scores for papers in the theoretical and empirical categories were 3.5 and 3.7, respectively, approximately halfway between the "average" and "above-average" ratings. The median for both groups was "above-average." The country studies tended to score considerably lower. Their average numerical score of 2.7 puts these papers collectively somewhere between the

"average" and "below-average" ranges. Another revealing metric is the number of scores falling into the "below-average" and "unsatisfactory" bins. Among the empirical papers that number is 4 out of 24 (17 percent), and among the theoretical papers, 1 out of 13 (8 percent). Among the country studies, by contrast, 8 out of the 19 ratings (42 percent) were either "below average" or "unsatisfactory."

22. Because the papers in the "thought piece" category contained little or no original empirical or theoretical research, it was difficult to rate them using the same criteria as those used for the other 56 papers. Many of the evaluation criteria listed in Table 1 did apply, however, and in our review we relied more heavily on attributes such as having a clearly focused question and using an appropriate theoretical or conceptual framework. It should also be noted that because this category contains only four papers, our findings may not be representative of the broader population. Having said that, we judged the "thought piece" papers to be considerably weaker than those in the other categories. Our average score was only 2.1, and the median rating was "below average."

23. There was less than complete agreement among the panel members on the strongest and weakest papers among the 60, and relatively little overlap between the lists of papers that were nominated in each category. To some extent, this lack of agreement reflected the different weights that the individual reviewers assigned to specific strengths and weaknesses. But with few exceptions, the papers nominated for the "best paper" category by one reviewer tended to be rated highly by the other reviewer. The same was true for nominees in the "weak paper" category.

24. What is more interesting is the distribution of the "best" and "worst" papers across the 4 categories of paper we identified. Seven on the "best paper" list came from the theory category, 12 came from the empirical category, and 3 were country studies. None was a "thought piece." The country study and "thought piece" categories were overrepresented on the "weakest paper" list, with 7 and 3, respectively. Among the weakest papers, 6 came from the "empirical" category, and only 1 was a theoretical paper, suggesting perhaps that the barriers to entry inherent in this kind of work prevent substandard papers from being written.

B. Citation Counts and Publication Success of the Sampled Papers

25. Publication data and citation counts are often used in academia to gauge research quality. These are imperfect metrics for judging IMF research for the simple reason that they undervalue quality research that might be highly valuable from a policy perspective and yet is insufficiently novel or technically innovative to appeal to the referees and editors of academic journals.

26. The citation and publications statistics do contain some information, however, at least with regard to the visibility of the more academically oriented papers in the sample. Figure 2 is a histogram of citation counts tabulated by Research Papers in Economics (RePEc), and Figure 3 is a comparable histogram using data from Google Scholar. As might be expected,

both are highly skewed to the right. In the RePEc data, the median citation count is two and the mean is six; for Google Scholar, the comparable statistics are thirteen and thirty-two. Still, the numbers suggest that IMF research *is* being cited, and also that a respectable number of papers are getting a great deal of attention.

Figure 2. RePEc Citation Counts

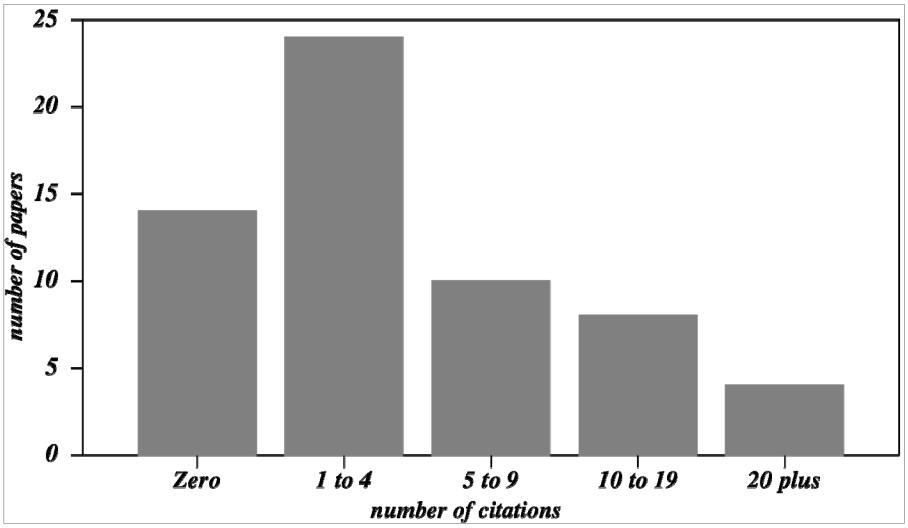
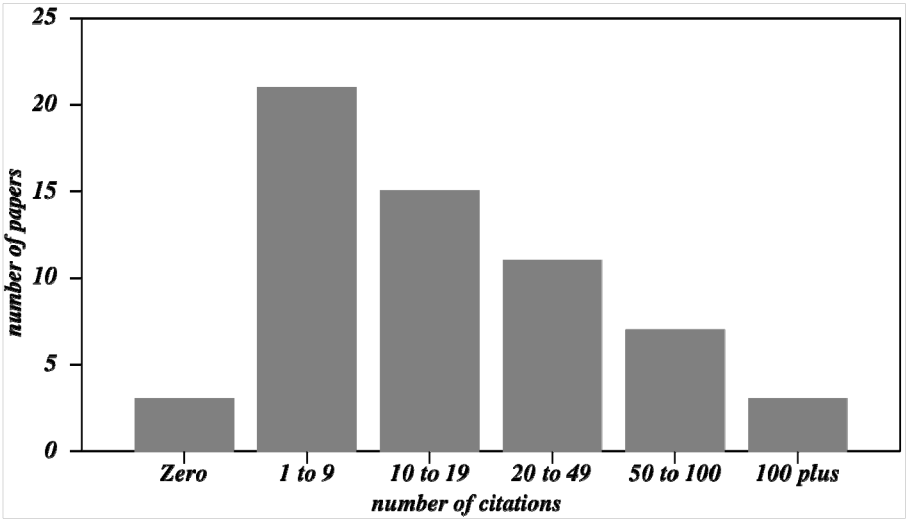


Figure 3. Google Scholar Citation Counts



27. It is also worth nothing that 24 of the 60 papers (including 3 originally issued as SIPs) have been published in journals (including *IMF Staff Papers*) or as book chapters. This may overstate publishing success to some extent, because it includes some non-refereed venues. On the other hand, it may understate the number of publishable papers, because many of the more recently written papers are surely still under review.

28. Another point worth mentioning is that although our paper rankings were intended to measure attributes other than publishability, there is a link between our rankings and the citation counts. To illustrate this, we regressed the citation *ranking* (i.e., the most highly cited paper was a "one") on the overall rankings given by the primary and secondary reviewers, plus a constant, for both the RePEc and the Google Scholar indexes. Both scores had the correct (negative) sign, and were significant at the 5 percent level or better in the RePEc regression.

29. The adjusted R-squared from the RePEc regression was only 0.23. A low R-squared is not surprising, however, given our use of a broad range of quality metrics. A set of criteria that focused more narrowly on academic "publishability" would likely have resulted in a better fit. Indeed, several of the papers scoring highly on our criteria had very modest citation counts. (A few of these were more recent papers, for which low citation counts are to be expected.) One widely cited paper received only average marks.

C. Disaggregation by Department, Affiliation, and CoAuthorship

30. Our ratings suggest some differences in paper quality across IMF issuing departments, although it would be hazardous to make strong inferences based on such small samples. The European Department (EUR) was most highly represented, contributing 12 of the 60 papers. Monetary and Capital Markets (MCM) was next with 11, followed by Research (RES) with 10.[3] Six papers came from the IMF Institute (INS). The area departments (other than EUR) individually accounted for relatively few papers, but they collectively contributed 21.

31. With mean scores of 3.7 and 3.6, respectively, RES and MCM stand out as producers of high-quality research. Papers from EUR also tended to be rated highly, earning a mean score of 3.3. INS received a low mean score of 2.5, but more than the usual amount of caution is warranted in generalizing this number given the small number of papers on which it is based. The area departments (other than EUR) tended to produce the weakest papers, and their average was 2.5. This finding is not surprising, given that most of the low-scoring country studies originate from these departments.

32. Disaggregating the data according to authors' affiliation and coauthorship provides some evidence for a connection between collaboration and paper quality. The 16 papers that were written jointly with at least one outside coauthor received a mean score of 3.6, while those written solely by IMF staff received 2.8. Internal collaboration was also associated with higher paper quality. Of the 40 papers written exclusively by IMF staff, coauthored papers averaged 2.9, while sole-authored papers averaged 2.7. And although the small sample size

[3] The figures for MCM also cover the former Monetary and Exchange Affairs Department (MAE) and Monetary and Financial Systems Department (MFD).

makes it hard to generalize, collaboration with IMF staff may benefit outsiders as well: the 4 papers written by external authors earned a mean score of 3.3, compared with 3.6 for papers involving collaboration between IMF staff and external authors.

V. Strengths and Weaknesses

33. The numeric ratings summarized above conceal as much as they reveal. A more useful critique would involve a comprehensive cataloging of strengths and weaknesses of the papers we reviewed. Naturally, this is hard to do in detail for all 60 papers. But we were able to identify a number of attributes shared by most of the papers we rated highly. Similarly, the weaker papers tended to share many of the same flaws.

A. Common Attributes of Strong Papers

34. **Theoretical framework.** The best papers we reviewed all had a coherent conceptual or theoretical framework. Naturally, this was the case for all the papers in the theory category. The stronger empirical papers were also grounded in economic theory. In some cases, the framework was developed explicitly in an optimizing structural model that could then be applied to the data. In one study, this involved estimating a fully specified dynamic stochastic general equilibrium (DSGE) model. Another paper that was quite successful in this regard used some relatively simple theory to discuss the persistence of inflation under alternative exchange rate regimes, and then tested those implications empirically.

35. Not all empirical analysis is amenable to such a high degree of integration with theory, of course, but it was often possible to bring theory to bear in a looser way. One study successfully used a standard macro model to motivate the identifying restrictions on a Galí-style structural vector auto-regression (VAR). A perfectly standard monetary exchange rate model served as a satisfactory foundation for another paper's empirical analysis. In other cases, an informal but careful reference to existing theory proved effective in motivating and interpreting the empirical analysis. One very good empirical paper on the credit channel of monetary policy transmission framed its analysis with a concise but thorough reference to the conclusions from others' theoretical research. We found that high marks on this dimension tended to carry over into other areas as well. Papers with a well-developed framework tended to be focused, and easier to interpret. The discipline of a model, even an implicit one, usually helped to keep the conclusions firmly grounded in the analysis.

36. **Appropriate empirical methods.** The use of appropriate statistical techniques was another hallmark of the strongest empirical papers in our sample. We emphasized the appropriateness of the method, rather than the sophistication, because in many cases relatively simple econometric methods were perfectly adequate for the question at hand—especially given the data limitations often encountered in IMF work. Many papers using simple ordinary least squares or common panel data methods received high marks on this criterion. In fact, the primary reviewer of one of our most highly ranked papers wrote that it had "not much fancy econometrics," and that it "used the right data and the simplest methods

possible." Another well-regarded paper in the portfolio was awarded high marks more for its development of an original data set than for its econometric analysis.

37. Of course more sophisticated methods were sometimes appropriate, as in the application of panel cointegration techniques to exchange rate behavior, and Bayesian DSGE methods to monetary policy transmission. Similarly, several of the calibrated macro modeling exercises received high marks.

38. **Robustness checks, diagnostic statistics, and caveats.** As mentioned earlier, one criterion to which we paid close attention was whether the paper included enough information for a reader to be able to assess the robustness of the analysis. Good empirical papers provided robustness checks and diagnostic statistics, and took care to acknowledge their limitations. One particularly thorough paper presented two sets of estimates, one based on maximum likelihood and another using the generalized method of moments, along with a Monte Carlo simulation. Others presented structural VAR estimates based on alternative identifying assumptions.

B. Common Weaknesses

39. Not surprisingly, most of the papers on the low end of the distribution were deficient on more than one of the criteria we identified. But even many of the highly ranked papers on our list also fell short on some dimensions. Table 3 gives a breakdown of the ratings by evaluation criteria.

Table 3. Distribution of Ratings by Evaluation Criterion

Evaluation criterion	Distribution by rating (In percent)					Average Rating
	S	AA	A	BA	U	
1. Question is well posed and clearly focused	28	43	17	7	5	3.8
2. Places work within the context of existing literature	20	28	20	27	5	3.3
3. Specifies contribution to existing literature	17	25	28	25	5	3.2
4. Uses an appropriate theoretical/conceptual framework	15	12	30	22	22	2.8
5. Uses appropriate data and empirical methods proficiently	10	24	27	27	12	2.9
6. Includes critical discussion and/or robustness analysis of results	7	18	15	42	8	2.7
7. Writing is clear and well organized	12	27	35	25	2	3.2
8. Adds value relative to existing research	13	28	23	27	8	3.1
9. Conclusions are firmly grounded on the analysis	10	23	27	33	7	3.0
10. Articulates policy relevance of findings	12	23	32	27	7	3.2
Overall distribution[1]	**10**	**29**	**34**	**23**	**4**	
Overall average rating						**3.2**

[1] The distribution by rating criterion is for the primary reviewer only, whereas the overall rating is based on the average of the two reviewers' scores.

40. One of the most common flaws among the empirical papers was the absence of a coherent conceptual or theoretical framework, as reflected in the below-average score of 2.8 on this criterion. The papers scoring poorly on this criterion all involved reduced-form regression analysis. Their choice of variables was sometimes loosely motivated by theory, but the tenuousness of the theoretical link usually rendered their results unintelligible. A paper on equilibrium exchange rates, for example, cited theory to justify the inclusion of half a dozen or more variables in a cointegrating relationship, but the results from this "kitchen sink" regression were impossible to interpret. Another study neglected the likely endogeneity of one of its regressors—a problem that would have been immediately obvious had the analysis been framed with reference to a simple macro model. A third neglected to discuss the theory of optimal currency areas in its assessment of the desirability of moving to a fixed exchange rate regime.

41. A disturbing number of papers contained conclusions that bore at best a tenuous relationship with the core analysis. The mean score for this criterion was 3.0, but the primary reviewer rated one-third of the papers "below average" or below on this dimension. Some papers made little effort to connect their conclusions with the analysis, instead offering a set of unsubstantiated boilerplate conclusions based on conventional wisdom—for example, that fiscal imbalances should be addressed; that the exchange rate should be allowed to float; or that central bank policy should be predictable. One reviewer remarked that one paper's concluding section seemed to have been written first, with the rest of the paper added later. In other cases, the conclusions were overgeneralized, ignored inconvenient findings, or lacked appropriate caveats. In one paper, two of the three policy conclusions that were stated in the abstract could not be substantiated by the results without significant qualifications. Other papers exaggerated or "spun" their policy implications. One study, for example, extrapolated from a reduced-form analysis of trade flows to broader statements about the effects of the credibility of exchange rate pegs. At least two papers presented their findings as bearing on the implications of globalization, though they offered no analysis relating specifically to that issue.

42. Another common problem among empirical papers was a lack of technical proficiency, or the choice of an inappropriate empirical framework. The mean score for this evaluation criterion ("uses appropriate data and empirical methods proficiently") was 2.9, and every paper receiving a "below-average" or "unsatisfactory" overall rating from its primary reviewer received low marks on technical proficiency. Some papers used inappropriate methods given the data problems their authors faced—which in the case of developing economies are often severe. Standard econometric tools, such as VARs and cointegration methods, were sometimes applied mechanically with little attention to their limitations or their suitability to the available data. Issues arising from obvious breaks (due in one case to a civil war) or other anomalies in the data were sometimes inadequately addressed. Issues of trends and nonstationarity generally received insufficient attention.

Endogeneity and sample selection problems were often left unacknowledged and unaddressed, even in otherwise well-regarded papers. An example is a paper that excluded a certain de facto exchange rate regime from its analysis, but failed to discuss the sample selection issues that might have resulted from that decision.

43. Given the emphasis on policy implications, and the data problems inherent in doing this kind of research, our view is that a careful examination of robustness issues is essential. Several papers were marked down on this criterion, including some of those that scored well overall, and the mean score on this criterion was only 2.7. Papers involving model simulations were sometimes prone to this problem. One paper presented simulations based on only one set of parameters, and failed even to consider the possibility that the resulting estimated shock variances may have been exaggerated by measurement error. Another used de jure exchange rate regimes rather than the de facto classifications, leaving unanswered the question of whether similar results would have been obtained with the de facto regimes. Regression-based analyses often lacked robustness checks. Few papers rigorously assessed the results' sensitivity to outliers. For example, one paper claimed to have found evidence for a nonlinearity involving a threshold rate of inflation, ignoring the fact that only four countries in the sample exceeded the threshold. A large number of papers neglected to report the basic diagnostic tests that would have been required to correctly interpret their results.

44. Most of the papers were clearly written, and the mean score on this criterion was 3.2. But some of the weaker papers—and even a few of the good ones—were quite hard to follow. These expositional issues often surfaced in their introductions, which were sometimes long, verbose, or hard to follow. In at least one paper, the symbols used in the model were never defined, leaving the reader to guess their meaning. One paper with more than 50 equations regularly referred to equations several pages back without specifying their equation numbers.

45. A number of papers suffered from slipshod presentation of tables and graphs. Axis labels and units were omitted. Some tables were incomplete, others disorganized. Some papers, particularly those involving model simulations, presented their results as vast numerical tables when a graphical presentation would have been more effective. At least one paper referred to a nonexistent table. Sometimes the description of results in the text was inconsistent with the numbers shown in tables. Another common problem was the inadequate description and interpretation of tables and graphs. For instance, one paper devoted only one short paragraph to the empirical results for 9 out of the 11 countries in its sample.

46. Most papers articulated a well-defined research question, and indeed the mean score on this criterion was 3.8. A few did not, however. Papers lacking a clear focus also tended to be weak on other dimensions, such as the inclusion of a coherent conceptual framework. The "thought piece" papers tended to be quite weak in this regard. Several of the country studies suffered from a vaguely specified research question, and ended up doing little more than

providing a narrative account of a country's monetary and/or exchange rate policies, some reduced-form regression results, and some generic recommendations.

47. Finally, an issue common to many of the papers (again, even some of the highly ranked ones) was inadequate documentation of data and methods. One paper, for example, mentioned that it used a foreign price index in its regression, without specifying exactly what the data represented or where they came from. Another failed to specify the countries included in the sample, the data frequency, and the data source. Replicability is the litmus test for good scientific analysis, and all too many working papers fell short on this criterion.

VI. Conclusions and Recommendations

48. The IMF is clearly doing a lot of things right when it comes to research. Most of the papers pose interesting policy-related questions, and many are executed with great skill. Some have been cited extensively and have made major contributions to the literature.

49. Many of the papers in the working paper series are substandard, however, and the inclusion of these papers in the series reflects badly on the institution. Even more worrisome is the thought that the IMF's advice might rest on less than rock-solid research. But the good news is that much low hanging fruit is waiting to be picked. Some modest interventions would have strengthened most of the papers, and brought them up to respectable standards of quality. Very few were completely unsalvageable.

50. Just getting the basics right will help a great deal. Doing the simple things correctly in 50 percent of the papers would do more to bring up the average than the application of leading-edge methods in 5 percent. A refresher course on basic panel data methods, for example, would yield a bigger "bang for the buck" than a workshop on panel cointegration. In most cases, core first-year Ph.D. level theory would be more helpful in providing empirical papers with the necessary theoretical foundations than the latest advances in DSGE modeling. This is not to say that the more sophisticated methods should be discarded, of course, and it is very important for the IMF to be doing state-of-the-art research. Indeed, even among the strong papers there is scope to move a little closer to the methodological frontier. But in allocating resources optimally, it should be kept in mind that the bulk of IMF's research has and will continue to be practical, applied, policy research relying on tried-and-true econometric methods.

51. With these considerations in mind, our recommendations encompass three elements: screening, feedback, and documentation.

A. Screening

52. Quality control is a serious problem in the working papers on monetary policy frameworks, but it can be remedied with a more systematic screening process. Some of the papers included in the series probably should never have been written. Others merited

inclusion but needed a great deal more work in order to bring them up to a minimal standard of quality. Here are some suggested ways in which the screening process might be improved.

- **Standards.** The IMF should establish a clear standard for quality, perhaps broken down using the same criteria we used for our evaluation. For example, authors should know that a coherent conceptual or theoretical framework is a requirement for publication. Writing and presentation standards should also be clearly articulated. These might include the requirement that all tables and graphs be self-contained, that the relevant diagnostic tests be reported, and that the data and methods be adequately documented.

- **Accountability.** The supervisor signing off on a paper should be responsible for ensuring that the quality standards are met. The signature should not merely be pro forma.

- **Referee reports.** An internal refereeing process should be established to complement supervisory approval. Especially for the more technical papers, it may be more appropriate to have the research evaluated by an in-house expert, not just the immediate supervisor. The European Central Bank sends its working papers out for external referee reports, and the IMF might consider instituting a similar procedure.

- **SIPs.** The IMF should be more selective in turning SIPs into working papers. Not all SIPs will be suitable for inclusion in the working paper series. Those that are should be held to standards comparable to those that apply to other working papers.

- **Nonresearch papers.** The IMF should exclude from the working paper series those papers that contain little in the way of original research. Applying this criterion would weed out the weak "thought pieces" we found in our sample, whose uneven quality reflected badly on the overall quality of the papers we reviewed. Exceptions might be appropriate for survey papers, if these were judged to add value relative to the existing literature.

B. Feedback

53. The second theme in our set of recommendations involves feedback and constructive criticism. Many of the flaws we identified in the weaker papers stemmed from a lack of proficiency with empirical methods, or the failure to articulate a well-focused research question within the context of a coherent and appropriate theoretical framework. This should not happen at an institution with the intellectual depth of the IMF. The expertise required for generating consistently competent papers is already in place. What is needed is to make that expertise more accessible.

- **Collegial feedback.** Our sense in reading the 60 papers is that many would have benefited greatly from the comments of one or two colleagues with relevant expertise.

This does not happen automatically, of course, especially since the people with the relevant expertise may work in a different part of the organization. For that reason, we suggest that the IMF create forums or institutionalize processes that would facilitate peer-level feedback. This could take the form of informal internal seminars—preferably set up in such a way as to cut across organizational boundaries.

- **Editorial feedback.** The editorial process is another possible channel for providing constructive criticism. In addition to (or perhaps instead of) the normal sign-off process, the IMF should consider appointing an editor, or creating an editorial board, for the working paper series. Though one job of the editor or the board would be to weed out inappropriate or unsalvageable papers, an equally important function would be to provide feedback and to make suggestions about any needed changes or improvements. Ideally, the editor would be able to direct the author to someone else within the institution who could give the necessary technical or methodological guidance.

- **Collaboration.** Many of the papers we reviewed showed great depth of knowledge about local and institutional features but were weak in terms of technical proficiency. Others had the opposite problem: strong in methods but weak in local knowledge. This suggests that there are unexploited gains from trade: experts on time series econometrics could work with country authorities to refine the empirical work, for example, or someone who has studied monetary unions in other contexts could be enlisted to work on a project involving the CFA franc zone, strengthening the conceptual framework and facilitating a more comparative analysis.

- **Writing resources.** Judging from the working papers we reviewed, many economists might benefit from some additional resources dedicated to the improvement of writing skills. Simply distributing a style guide for IMF publications could be an effective way to take the rough edges off some of the working papers. This could include templates for formatting regression output and tabular material, and guidelines for annotating and labeling graphs. And because everyone's writing can benefit from the suggestions of an editor, the IMF might consider dedicating a staff member to perform this function. Some of the presentation flaws, such as the failure to define the symbols used, could easily have been detected and corrected by a proofreader.

C. Documentation

54. Our third recommendation concerns documentation. The credibility of any kind of research rests on its reproducibility, which in turn requires a full and detailed description of the data and methods used to generate the results. Credibility is particularly important for an institution like the IMF, whose research forms the foundation for its policy recommendations.

55. Many of the papers we reviewed fell short on this criterion. The IMF is not alone in this regard, of course, and authors of academic publications are often remiss when it comes to documenting data and methods. In an effort to remedy this, authors of IMF working papers should know that adequate documentation is an essential ingredient of good research. Checking for this documentation should be an integral part of the editorial or administrative approval process.

56. The IMF could do even more in this regard. Academic best practice is for authors of published papers to make available online, or by request, the data and programs used in the analysis. This kind of openness is particularly valuable for the IMF's country studies, where locally based economists may have a keen interest in reproducing and extending the IMF's analysis. Undoubtedly, a lot of information is already being shared informally. But as part of a broader effort to improve the documentation of its research, the IMF might consider establishing an online repository for nonproprietary data of potential interest to outside researchers.

Annex. Evaluation Procedures

The IEO forwarded to the review panel 60 randomly selected from the 187 IMF working papers produced since 1999 on the topic of monetary policy frameworks. Each of the three panel members was designated as the primary reviewer for 20 papers and the secondary reviewer for another 20. All papers were therefore read by two panel members. Thus, each paper had both a primary and a secondary reviewer. Primary reviewers prepared a brief report on each assigned paper and completed a Research Quality Indicator Form (reproduced as Table 1 in main text). Secondary reviewers provided a numerical evaluation, and in some cases, a supplementary narrative critique. The panel members were also asked to identify the 4 strongest and the 4 weakest papers from both their primary and secondary lists.

The panel convened on Monday, March 15, 2010 to discuss the paper evaluations. The first substantive topic was an examination of the strongest and weakest papers. Next was a discussion of any major discrepancies between the primary and secondary reviewers' scores. Although the panel members broadly agreed on the papers' merits and demerits, 11 of the papers were assigned scores that differed by 2 or more increments. Our discussions of these papers led in many cases to the convergence of views, and often one of the two reviewers was willing to concede that an important feature of the paper had been overlooked. In only a few cases did a significant divergence of views persist after going over the paper in question, and most of the major rating discrepancies would have been eliminated had there been a formal reconciliation process.

The third task of the panel meeting was to compare notes on each of the 60 papers. This gave the panel an opportunity to discern recurring themes and issues in the body of research as a whole. The meeting concluded with a general discussion of the strengths and weakness of the papers examined, and some preliminary planning for the drafting of the panel's report.