IEO
Independent Evaluation Office
of the International Monetary Fund

BACKGROUND PAPER

BP/11/03

Review of IMF Research on Tax Policy

Robin Boadway, Christopher Heady, and Henrik Kleven

May 20, 2011

Address. 700 19th Street, N.W., Washington D.C. 20431, U.S.A. Telephone. +1202 623 7312 Fax. +1 202 623 9990 Email. feedback@ieo-imf.org Website. www.ieo-imf.org

 BP/11/03

IEO Background Paper
Independent Evaluation Office
of the International Monetary Fund

Review of IMF Research on Tax Policy

Prepared by
Robin Boadway, Christopher Heady, and Henrik Kleven

May 20, 2011

Abstract

This study examines the technical quality of a sample of 60 IMF working papers that focus on revenue and tax policy. It found significant variability in the quality of those papers. The papers were generally well motivated and focused on policy issues that were relevant for many countries. The papers were generally well written and mostly set within the context of the relevant literature. But many fell short in the analytical execution of the research, including the formulation of the model, the innovativeness of the approach, and the empirical or theoretical analysis. This resulted in lower scores for value added than for exposition. Fund researchers' reliance in some areas of research on a limited number of established and sometimes dated approaches may reflect an overly inward-looking approach to research. The study offers recommendations for research program management.

JEL Classification Numbers: A1, E6, H2, H3, H6

Keywords: tax policy, fiscal policy

Author's E-Mail Address: boadwayr@econ.queensu.ca

Contents **Page**

I. OVERVIEW

1. To assess the technical quality of tax policy research produced by the IMF over the period 1999–2008, the review panel evaluated a sample of 60 IMF working papers drawn randomly from the 120 such papers that were issued during the period on tax and revenue issues, including fiscal policy papers focusing on tax issues.[1]

2. The panel concluded that the best IMF working papers in tax-policy-related areas are very good by international standards, as reflected by the fact that much of this research is published in strong international journals. At the same time, the top papers by such standards represent only 25 percent of the working papers. There is a large group of average papers in the middle that are mostly unpublished, and some papers that are considerably below publication standards. The top papers come mainly from a small number of researchers, an above-average proportion of whom are from two IMF departments—Fiscal Affairs (FAD) and Research (RES). In the middle and lower categories, there is limited diversity in empirical approaches, and almost no use of techniques that are now becoming commonplace in academic policy-related research. Much of the Fund's tax policy research uses rather mechanical applications of long-established approaches that have limited applicability to real-world policy problems. This suggests a tendency to be too inward-looking in topic selection. There is limited use of outside expertise, which when used often produces fruitful results.

3. We recognize the constraints that IMF researchers face. Much of the research emanates from country missions and advisory work, which means that topics are constrained and so is the time available. These limitations pose challenges for improving the quality of papers in the middle and lower quality ranges.

4. This said, the credibility of the research program would be much improved if the proportion of papers published in refereed outlets were raised significantly and if more of the papers were jointly written with external coauthors. To help Fund researchers keep current with developments in the economics literature, the panel recommends more use of opportunities for training, seminars, and interaction with outside scholars. To foster collaborative research and possibly reduce the cost of empirical research in the longer term, the panel recommends establishment of a data repository for empirical papers.

---

[1] The three international scholars who made up the review panel have diverse backgrounds and experience but share a common interest in tax policy research and analysis. They have published widely in the major academic journals, especially in public economics. Their experience includes posts as editors and associate editors of major journals, as well as experience in tax and fiscal policy research at the Organization for Economic Cooperation and Development. Together, they have skills in public economic theory and policy analysis and modern empirical approaches to tax policy, including recent micro-data-based approaches and natural experiments. Despite their different individual backgrounds, experience, and perspectives, considerable consensus emerged both in their initial evaluations and their overall evaluation of the research.

5. This report is organized as follows. Section II describes the evaluation criteria and how the panel interpreted them in the context of IMF's research mission. Section III is an overview of the evaluation ratings given to the 60 papers by the review panel, and of other quantitative indicators of the papers' quality. Section IV explains the categories into which the papers were grouped for comparison purposes. Section V discusses the concerns that surfaced from the evaluation and Section VI offers conclusions and recommendations. Annex 1 outlines the evaluation procedures used and Annex 2 provides illustrative data.

II. Evaluation Criteria

6. The quality of the research was judged partly by the extent to which it added value—whether by generating new knowledge, broadening the understanding of fiscal policy, or developing new policy frameworks. To assess the value added we rated the soundness of the motivation of the analysis; the relationship of the analysis to the existing literature; the use of state-of-the-art analytical techniques, both theoretical and empirical; the quality and originality of the analysis; the implications of the analysis for policy purposes; and an awareness of the sensitivity of the results to the assumptions and parameters used in the analysis. These are the same criteria as are used to evaluate similar literature found in peer-reviewed outlets, and especially in reputable economics journals that publish public economics and fiscal policy papers.

7. We also considered that the research should be judged against the role and objectives of the IMF. Thus the motivation of the papers and their potential relevance for policy advice and understanding should reflect the Fund's policy interests, and the research should include adequate country and institutional knowledge and an understanding of which research questions are relevant and which policy recommendations can be feasibly implemented. While academic research might be motivated by a search for pure knowledge or innovation, and be rewarded accordingly, IMF research presumably complements the institution's role in evaluating country policies and offering advice. The outcome of that advice has more riding on it than academic output, so IMF researchers must be prudent on that account. This reasonably constrains the choice of topics and the strong orientation toward policy research.

8. Each of the 60 papers had both a primary and a secondary reviewer (see Annex 1). The reviewers rated each paper on ten evaluation criteria using a rating scale of one to five, where the ratings were labeled "superior" (S), "above average" (AA), "average" (A), "below average" (BA), and "unacceptable" (U). As shown by the Research Quality Indicator Form reproduced in Table 1, the evaluation criteria fell into three categories: Framework (referring to the framing of the research question in the context of the literature); Analysis (referring to the theoretical or conceptual framework, the empirical analysis, and the critical discussion and robustness analysis); and Output (referring to the clarity of writing, the value added by

the research, the conclusions, and the policy relevance). In addition, reviewers were asked to provide an overall rating for each paper.[2]

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.

III. Overview of Quantitative Indicators

9. It is useful to begin with a summary of the quantitative evaluations of the 60 papers, based mainly on the ratings given by the primary and secondary reviewers.

A. Overall Ratings

10. The overall rating given by all three primary reviewers to the 60 papers was 3.1, only slightly above an "average" rating. Two of the three reviewers gave average overall ratings of "below average" (2.7 and 2.9), while the third gave a rating "above average" (3.6). The ratings were slightly higher among the secondary reviewers.

[2] The individual reviewers chose how to interpret these ratings, and slight differences in their interpretations are apparent: two of the reviewers tended to judge the papers more positively than did the third reviewer, both in their primary and secondary reviews of the papers (see Annex 2). This likely reflected different interpretations of the letter grades.

11. The primary reviewers gave the 60 papers an overall score slightly above "average" (3.1). They rated 7 of the papers as "superior" overall; 10 as "above average;" 24 as "average;" 17 as "below average," and 2 as "unsatisfactory;" this was a single-peaked distribution in which 40 percent of papers were judged "average" and almost 30 percent were judged "below average." The secondary reviewers gave a similar overall rating of 3.2, but the distribution here was much flatter: 90 percent of the papers were roughly equally distributed across the "above average," "average," and "below average" categories (Table 2).

Table 2. Average Ratings by Primary and Secondary Reviewers Across Evaluation Criteria

Evaluation Criteria	Primary Reviewer	Secondary Reviewer
Question focus	3.7	3.7
Literature review	3.1	3.3
Contribution	3.1	3.2
Framework	3.1	3.1
Data	3.0	3.1
Robustness	2.9	2.9
Clarity	3.5	3.5
Value added	2.9	3.0
Conclusions	3.0	3.1
Policy relevance	3.1	3.2

12. The average judgments mask some differences across the three reviewers. The primary and secondary reviewers largely agreed on which were the top three papers: of the nine papers that were ranked in the top three by primary reviewers, six were also ranked in the top three by secondary reviewers, and two of the other three had similar overall ratings.[3] There was more disagreement about the nine lowest-ranked papers, with very little overlap between those ranked lowest by the primary and secondary reviewers. Nonetheless, the raw overall ratings for lowest-ranked papers were generally similar for both primary and secondary reviewers.

13. Three exceptions—papers to which the primary and secondary reviewers gave widely different overall ratings—highlight some of the differences among reviewers in judging quality. One was an empirical paper that uses panel data to estimate an aspect of government fiscal behavior. One reviewer rated this paper "above average," based on the relevance and motivation for the analysis, the somewhat innovative policy question being posed, and what

[3] In one anomalous case, a primary reviewer ranked a paper in the top three with a rating of "superior," while the secondary reviewer rated it "below average." This was a paper whose contribution lay mostly in devising a statistical technique for revenue-forecasting purposes. It was made difficult to judge by the fact that the only two references it made to the economics literature were two econometrics textbooks, along with two references to documents from public organizations. This failure to place the contribution of the paper in the context of the literature made it difficult to judge its quality and importance.

he considered to be the novel choice of some political explanatory variables. The other reviewer rated it "below average," based on deficiencies in econometric practices—and particularly an inadequate treatment of the potential endogeneity of some of the explanatory variables used, including the political explanatory variables to which the first reviewer attached importance.

14. The second paper was a case study of an administrative aspect of tax policy. One reviewer rated this paper "above average," based on the view that, although it was to a large extent descriptive, it outlined in detail the many policy issues that arise with respect to improving tax administration in a particular developing economy and evaluated the alternatives in light of one country's experience; in that sense, it was a useful input to the policy process in other developing countries. The second reader, who judged it "below average," took the view that despite the paper's policy relevance, the level of economic analysis and value added was limited.

15. The third exception was a paper that provided Monte Carlo simulations of an important policy issue in resource-dependent developing countries. One reviewer rated it "above average," based on the quality and novelty of the statistical analysis. The second reviewer found it "below average," on the basis that the policy objective embedded in the policy analysis was not a reasonable way to manage temporary resource revenues, since it assumed away the need to build up a stock for use by future governments and generations.

B. Ratings by Evaluation Criterion

16. Thus far we have focused on the overall ratings. It is also useful to look at the average ratings for each of the ten evaluation criteria. Using the scale of 1–5 for the ratings U–S, average ratings for each criterion can be calculated. The criteria on which the papers consistently scored highest were criterion 1 (question well posed and clearly focused) and criterion 7 (clear and well organized writing). Primary reviewers gave slightly higher than average ratings to criterion 10 (articulation of policy relevance). Crucially, the lowest ratings tended to be given for criterion 8 (value added) and criterion 6 (critical discussion and robustness analysis). Ratings on the remaining criteria were about average.

C. Ratings by Department

17. Though only one IMF department (FAD) produced half the papers in the sample (the other 30 were produced by 8 other departments), it is instructive to look at the origins of the papers that the panel rated highest and lowest.

18. Of the 9 highest-rated papers, FAD contributed a disproportionately large number: 6 were from FAD (out of the 30 that FAD produced), 2 were from RES (out of the 6 that RES produced), and 1 was from the Asia and Pacific Department (APD) (out of the 2 that APD produced). FAD and RES together produced a significant proportion of the working papers that were eventually published.

19. A further property of the nine highest rated papers is worth noting. One IMF staff member from FAD was an author or coauthor of three of them, while another FAD staff member was coauthor of two. One RES staff member was coauthor of two top papers.

20. Of the nine papers that were ranked lowest, four were from FAD, three were from Western Hemisphere Department (WHD) (out of four it produced), and one each were from the African Department (AFR) and the former Middle Eastern Department (MED). Thus, FAD produced disproportionately few of the lowest-rated papers, while WHD produced significantly more. (None of the papers from RES featured among the lowest ranked.)

D. Other Indicators of Quality

21. Of the 60 working papers reviewed, 15 were eventually published formally in journals and books (some of the more recent working papers might be published and would raise this number). The 15 represent only 25 percent of the working papers—a surprisingly low proportion compared with publication rates in most academic economics departments.

22. Of the papers published, 10 were published in journals outside the IMF,[4] 3 in the internal journal *IMF Staff Papers*, and 2 others in books of readings. Of the 10 journal articles, 5 were produced by FAD (3 of them by 1 author), and 4 by RES (2 of them by 1 author). Three were written jointly with outside authors, and the rest were written wholly by IMF staff members. These results suggest that research by FAD and RES tends to be viewed more positively, by academic standards, than research produced elsewhere in the Fund.

23. Citation counts listed by Research Papers in Economics (RePEc) and Google Scholar (GS) were provided for all 60 assigned papers. Only 5 papers had received more than 10 RePEc citations; all were FAD papers and most of them were published. Nine papers had received no citations at all, and 25 had received between 1 and 3. Among the 12 papers that were rated highest by either a primary or secondary reviewer, 1 had received 36 citations, 1 had received 6, 1 had 5, 3 had 3, and the remainder had received fewer than 3. Not surprisingly, the lowest-rated papers had much lower citation counts, with only 1 having received more than 2 citations. The pattern of GS citations is somewhat similar. Five papers had received more than 100 citations (the same 5 papers from FAD that had received more than 10 RePEc citations). A further 11 had received between 20 and 100 citations, and 31 had received fewer than 10. Among papers in the top-rated category, 1 had received 197 citations (the same 1 that had received 36 RePEc citations), while others had received 58, 33, 29, and 20, and the remainder 12 or fewer. The lowest rated papers showed much lower GS citation counts.

---

[4] Of the outside journals, one is a top general journal (*Economic Journal*), three are top field journals *(Journal of Environmental Economics and Management*, *Journal of International Economics*, and *Journal of Public Economics)*, two are in a good tax policy journal *(National Tax Journal)*, and four are other specialist journals *(Journal of International Money and Finance*, *Journal of Policy Modeling*, *Japan and the World Economy,* and *Review of World Economics)*.

24. It is also useful to look at the average ratings of papers by number and type of authors. External coauthorship seems to produce better papers: the average overall score given by primary and secondary reviewers to the 13 papers produced in collaboration with an external coauthor was 3.5, which is well above the average for all 60 papers. The papers with external coauthors were concentrated in the research category, as distinct from country studies or survey papers. Another 21 papers were single-authored by IMF staff members; they received an average overall rating of 3.0, slightly below the average for all 60 papers.

IV. Characterization of Research in Working Papers

25. The working papers on tax and fiscal policy are of different types, though some of them contain elements of more than one approach. Those of the first type are primarily research-oriented, though usually motivated by fiscal policy concerns. Most of them are empirical and some develop models that are calibrated to country data. Examples include studies of tax reforms or fiscal policy actions, evaluation of development aid, studies of tax smoothing and tax effort, and analyses of stabilization policies, including automatic stabilization. Many of the papers have potential policy applicability to groups of countries such as developing countries, resource-rich countries, transitional economies, or federal economies. They vary in their degree of innovation and value added, but in principle they are candidates for publication in refereed journals.

26. A second type of papers offers policy analysis. These are typically country-specific and use known research tools. Examples include studies of the intertemporal management of natural resource revenues, tax evasion studies, and analyses of country-specific tax reform and fiscal consolidation. They are often related to IMF mission work, and may have been released earlier as selected issues papers (SIPs). They are less likely to be publishable in good journals because they are not innovative, but to the extent that they use state-of-the-art analytical tools, they are valuable as inputs into the policy process of specific countries.

27. Survey and best-practice papers make up a third type. Examples include surveys of resource taxes and intergovernmental transfers, case studies in tax administration and tax reform, and critical discussions of different mechanisms for managing or financing public debt, such as collateralization of future revenues. Many are primarily descriptive, illustrating best practices as they have evolved in policy reforms in various countries, although good ones are also evaluative and emphasize lessons learned from country practices. Some address topics that do not lend themselves easily to standard theoretical or empirical analysis, for example because they involve complex administrative or institutional processes. Survey/ best-practice papers typically do not appear in journals, but such papers are potentially very valuable to policymakers and researchers because they synthesize the relevant literature as it applies to particular policy issues in specific countries or country groupings. They can also be valuable for informing nonspecialists about broad policy areas and indicating alternative policy approaches to a given problem.

28. Each type of paper was judged to be appropriate for the IMF working paper series and was represented in the sample reviewed. Of the 60 papers in the evaluation sample, 27 were determined to be research papers, 21 to be country studies, and the remaining 12 to be survey or best-practice papers. Many of the papers were rated as good as they either advance the boundaries of knowledge or apply good analysis to inform the policy process.

29. On average, the papers that can be thought of as research papers are of significantly higher quality than survey or best-practices papers, which in turn are of higher quality than country studies. Based on the ratings assigned by the first and second reviewers, the average rating of the research papers was 3.5, that of the country studies was 2.6, and that of the survey papers was 2.9 (Table 3). Thus, the research papers were rated distinctly above average and the country studies well below average.[5]

Table 3. Average Ratings by Type of Paper

Type	Number	Average Rating
Research	27	3.5
Country study	21	2.6
Survey/best practices	12	2.9
External coauthor	14	3.5
Multiple internal coauthors	21	3.0

30. Perhaps not surprisingly, many of the papers were based on country-specific policy analysis. Fourteen of the 60 sample papers (about 23 percent) had been issued as SIPs before becoming working papers. Of the former SIPs, 10 were produced by area departments and the rest by FAD; a significant proportion were ranked "below average" by one or both their reviewers. The only one of the former SIPs to be eventually published in a journal received an "above-average" rating.

31. The papers that were formally published, especially those that appeared in the better journals, tended to be research papers, though two of the survey/best practices papers were published in good journals.

32. A number of themes repeated themselves in the research, often along with repetition of analytical approaches and techniques. One example is the management of revenues from nonrenewable resources, where several papers applied versions of the so-called permanent

[5] These relative ratings of the three types of papers were confirmed by classifying the highest and lowest rated papers. Of the 12 papers that were rated among the highest 3 by either the primary or secondary reviewer, 10 were research papers and 1 each was a country study and a survey paper. Of the 16 papers that were rated among the lowest 3 by a first or second reviewer, 9 were country studies, 4 were research studies, and 3 were survey papers. These findings reinforce the differences in quality between the research papers on the one hand and the survey and country studies on the other.

income hypothesis to determining the sustainable level of government spending out of finite-lived resource revenues, as discussed further below. Another was the empirical estimation of tax effort. A third was tax smoothing studies based on representative-agent models.

V. Evaluation of the Research

33. The quality and value added of the research varied considerably. The research was generally well motivated and focused on tax and fiscal policy issues that were relevant for many countries. The papers were also generally very well written, with conclusions carefully drawn where appropriate. Usually, the papers were set within the context of the relevant literature. Where many fell short was in the analytical execution of the research, including the formulation of the model, the innovativeness of the approach, and the empirical or theoretical analysis. This resulted in lower scores for value added than for exposition.

34. The relatively few theoretical papers tended to receive ratings above average. Most of the empirical papers were competent, although many of them shared a number of shortcomings.

35. The best of the papers—which included roughly those rated in the top three by the primary and/or secondary reviewers—comprised about one-quarter of the papers (roughly the number that were published). Almost all of them were research papers, either theoretical or empirical, although one was a best-practice paper dealing with an important issue of tax administration. Some were written in collaboration with outside researchers. All reported original or innovative findings that were almost all of value for policy purposes. Most had a well-articulated theoretical framework and used it either to study alternative policy interventions from a theoretical perspective or to test the predictions of the theory empirically, using good techniques. These papers compare with what one might find in the working paper series of a good research-oriented university, and they typically warranted publication in good, though not the top, international journals (the best of them could aspire to top journals). Compared with a good university working paper series, their orientation is much more toward policy and policy-applicable research results, and the proportion of publishable papers is considerably smaller. Of course, some of the papers that are not suitable for journal publication are nonetheless of some value to the international policy community.

36. The large group of papers in the middle-rated group, comprising at least half the papers, hovered around the "average" rating. Many of them were reasonably solid and well motivated but added rather limited value, and a good number had questionable implications for policy. Relatively few of them were published or publishable, and they would not compare favorably on that criterion with the average papers in working paper series at good research-oriented academic departments.

37. The weakest papers were a very diverse group. Though overall their policy motivation was well conceived, their execution and value added were wanting. They suffered

from different faults. Some were largely descriptive, using statistics to describe fiscal problems that particular countries faced, and offered limited value for policy purposes. Others mechanically applied readily available statistical tools to study a problem faced by a country, with little concern for analytical foundations or causal relationships. A few were theoretical in nature, but used theoretical models that were not well motivated and tended to be somewhat overspecified for the problem being addressed.

38. A few in the weakest group of papers seemed not to be current with theoretical findings in the literature or state-of-the-art empirical approaches. A small number of them were judged to be below the quality one would expect of a research-oriented working paper series, while others might have been made more suitable by some revision. The number of below-the-line papers was not excessively large, although their presence might detract from the overall reputation of the series. Several of the lowest rated papers were single-authored, perhaps revealing the value of collaborative work for quality control. The presence of low-quality papers might also indicate the lack of peer interaction and of oversight by department managers in ensuring that papers that are approved for publication have been developed sufficiently to meet minimal standards.

39. The panel members recognize that comparisons with university research papers may not be wholly appropriate. As noted at the beginning of this review, the goals of the Fund's research agenda differ in some important respects from those of academic research. And, at the same time, IMF researchers tend to have other responsibilities that restrict the time they can take doing research. More important, they may be more sheltered from the cutting edge of academic research and not completely up to date on the most recent techniques and criticisms of conventional empirical approaches. An understandable temptation is to use standard "canned" techniques, especially those that are not too time-consuming. Interaction with outside researchers can help alleviate these problems, as some of the papers illustrate.

40. The panel nonetheless identified a number of specific concerns with the research as a whole, and especially the papers that were judged below the top 20–25 percent. These concerns are discussed below.

A. Concerns About the Choice of Empirical Approaches

41. The breadth and diversity of IMF applied research approaches differs from what may be found in the academic world. In the view of the panel, the gap is greater than it should be.

42. A common problem in the empirical strategies followed by the papers was a neglect of, or failure to deal adequately with, problems of identification. This was a general problem in the many cross-country panel studies, but it also applied to some of the within-country analyses. Explanatory variables were often endogenous—for example due to simultaneity or omitted variables—and although this problem was sometimes acknowledged, it was typically not rigorously discussed and dealt with, and the instruments and identification assumptions chosen were not convincing. Where it was dealt with using instrumental variables techniques,

the instruments chosen were not always appropriate. Examples of empirical studies suffering from important identification problems include work on the determinants of tax revenue ("tax-effort studies"), the effect of foreign aid on growth, the effect of tax policy on income distribution, and the determinants of tax amnesties. In some cases, innovative arguments were used to justify the choice of particular explanatory variables—for example, political economy determinants of policies—but poor treatment of endogeneity concerns undermined the validity of the findings. Related to this, some of the empirical work seemed to be driven by the ready availability of a data set and standard canned techniques for estimation. A more useful approach would be to first think about an important question, and then search for the data and identify the variation needed to shed light on that question.

43. In general, too much emphasis was placed on cross-country studies and macro evidence, at the expense of within-country studies and micro evidence.

44. These features of the papers mean that many of the results reflect correlations rather than behavioral effects or causation. This indicates a need for some caution in inferring policy implications and conclusions from the analysis. Appropriate caution was not always used in drawing conclusions from suspect empirical analysis.

45. The mix of empirical approaches used in the sample papers does not reflect the evolution of empirical research in the economics literature. In most applied research fields, including development economics, public economics, labor economics, international economics, and industrial organization, research in the past two decades has adopted a so-called design-based approach, where the search for identification of causal effects takes center stage. Typically, this involves seeking natural experiments or randomized experiments where causal inference is made by comparing treatment and control groups and applying difference-in-differences estimation and other techniques. The innovation in such research partly involves finding actual experimental situations that allow one to analyze phenomena of interest, and this takes both time and ingenuity. The payoff comes in more reliable estimates of causal effects. But there were no examples of these approaches in our assigned papers.

46. Other approaches can also be used besides natural experiments. Structural estimation can make empirical analysis more persuasive by explicitly identifying behavioral relationships that are suppressed in reduced-form estimation. These approaches are becoming more common in industrial organization, labor economics, and macroeconomics, where structural estimation can be embedded in a calibrated model. They have some advantage in policy evaluation. Structural estimation has drawbacks, but does provide useful evidence complementary to that from other empirical approaches.

47. A focus on identifying variation and causal inference typically requires a micro-approach. It is very hard to identify causal effects of, for example, tax and transfer policy using aggregate data. Micro-data studies have been used increasingly in areas such as public economics, where more aggregated studies would mask important details. Such analyses are

more demanding because they can be time consuming and the needed data are not always readily available, sometimes because of confidentiality problems. While confidentiality problems may be a genuine constraint, increasingly researchers have found imaginative ways to obtain micro-data that might be used for policy analysis. This holds even in the field of development economics where the data limitations are most severe. The IMF may have a comparative advantage in getting access to country-based micro-data.

B. Lack of Diversity of Methodological Approaches

48. As noted above, some modeling themes and approaches tend to be repeated, in many cases quite uncritically, in about 50 percent of the sample of papers we were assigned to review. This tendency might reflect some inward-looking bias and if so, it calls for greater efforts to interact with outside researchers.

49. One example of such repetition is the adaptation of the permanent income hypothesis (PIH)—which was designed originally to explain consumer behavior in the face of lumpy and uncertain income—to the management of revenues from nonrenewable resources to finance government services. The question is what sustainable level of government expenditures, and by implication sustainable fiscal deficits, can be achieved given a finite-lived stream of resource revenues. Several of the working papers address the question of how to spread the benefits of these revenues into the future, especially the distant future when future generations are alive. They adapt the PIH rather simplistically to the problem by suggesting that the present value of revenues should be annuitized to yield a permanent flow of finance for government spending, albeit taking account of some complications like uncertainty and differences in views about the weight to be given to future generations. While there is nothing very wrong with doing such a calculation and the results are suggestive, the approach tends to be accepted rather too uncritically given the many other issues that are relevant—such as the need for public investment and poverty alleviation, as well as the commitment issues a government is likely to face. Some of these papers do address some of the other issues that come with resource wealth, such as the design and management of a resource revenue fund, policies for diverting resource revenues to the public sector, sharing of natural resource revenues between levels of government, and the implications of natural resources for economic development and governance, including well-known resource-curse effects.

50. A second such theme is intertemporal tax smoothing, where representative-agent models inspired by Barro's long-ago analysis are used to study the optimal path of fiscal balances. In representative-agent models, convex deadweight costs of taxation are the main criterion for evaluating the path of fiscal policies in response to shocks; issues of unemployment are completely suppressed; intergenerational redistributive issues are largely suppressed; and the aggregative nature of the models means that components of both the expenditure and revenue sides of the budget cannot be distinguished. From either an

academic or a policy advisory point of view, the value added of these exercises seems rather limited.[6]

51. Fiscal policy analyses based on the PIH approach or the classic tax-smoothing model are largely mechanical applications of existing theory. The use of existing theory is not a bad thing per se. Problems arise when a relatively limited set of theories dominates, so that the overall relevance of the research for informing policymaking is limited, and when the theories used are themselves outdated. For example, studies that rely on the currency-demand approach to estimating tax evasion are unduly restrictive, given that other approaches have superseded this. Again, the Fund researchers' reliance on a limited number of established and sometimes dated approaches may reflect an overly inward-looking approach to research topics.

52. A third example is estimating so-called tax effort using cross-country data. Overall, the tax-effort papers do not add much to understanding of fiscal capacity in developing countries. These studies generally involve the addition of political and institutional explanatory variables into a basic tax-effort regression equation, relating the tax/GDP ratio to variables such as per capita income, share of agriculture in the economy, and economic openness. They do this in a rather ad hoc manner, with no explicit recognition of the processes that the basic equation is intended to capture or of whether the additional variables make sense in this context. The fact that several papers produced in quick succession do almost the same thing—adding slightly different political and institutional variables and adopting different (and sometimes conflicting) identification assumptions—suggests that there could have been some benefit to coordination among the authors.

C. Concern About Ideological Perspective

53. Some, but not all, of the papers seemed to have an ideological predisposition, if only implicitly. Most commonly, this took the form of faith in market outcomes (e.g., with respect to evaluating taxes on financial transactions), especially in macro-based models in which unemployment plays a surprisingly small role, even in those studying responses to shocks. There was often a strong emphasis on efficiency as a criterion of policy, to the exclusion of equity, perhaps partly as the result of the representative-agent models that were used. We have already mentioned the emphasis on tax smoothing as the researchers' main criterion for evaluating dynamic fiscal policy. Sometimes, analyses were done from a normative point of view with limited recognition of governments' ability or interest to implement such policies; for example, commitment issues were often assumed away in dynamic models, including those dealing with the management of nonrenewable-resource revenues. Some papers

[6] The tax-smoothing model is still used in some academic research—an example being political economy models of debt that take account of political processes and constitutional constraints like balanced budget requirements. However, in these studies the emphasis is more on understanding government behavior in realistic public choice settings, and the tax smoothing model is a convenient first approach to evaluating outcomes, as opposed to being an end in itself.

showed a tendency to judge outcomes in non-U.S. economies, including those of the European Union and Russia, using institutional features of the U.S. economy; one example is the use of the U.S. pattern of labor mobility as a norm to judge labor mobility elsewhere. Given these examples, some observers might perceive the research and policy advice of the IMF as disproportionately emphasizing market-based or efficiency-based solutions in policy evaluation and prescription.

54. The risk of this perception should not be overestimated, however, and there are good reasons for not trying to eradicate ideological bias proactively. Normative approaches to policy analysis do in principle form a good basis for policy prescription—even if researchers understand well that normative prescriptions may not be politically palatable—and they provide useful benchmarks against which to judge actual policy outcomes. What is most important is that normative approaches should not neglect important market failures (such as unemployment, limited labor mobility, etc.) that can constrain the choice of optimal policies. Moreover, normative approaches should also be sensitive to the full range of possible normative objectives, considering both efficiency and equity. Given the different views that reasonable people can have about equity objectives and the importance that most governments attach to these objectives, addressing the sensitivity of results and policy recommendations to alternative views is warranted.

VI. Conclusions and Recommendations

55. The research we reviewed has both strengths and weaknesses. The range of types of research is healthy, given the role of research as a complement to the other activities of the IMF, and encompasses policy-oriented applied and theoretical research, policy analysis and evaluation, and survey papers emphasizing best practices. Almost all the papers are very well written, with care taken to relate the research to existing literature and to draw lessons where appropriate for policy. On the other hand, with several exceptions, the value added in terms of innovation and original insights was limited, and this was reflected in the small proportion of the papers that were eventually published in good journals. There was considerable replication, often seemingly uncoordinated, of topics and approaches.

56. Most striking was the limited variety of theoretical and, especially, empirical approaches. Dominant theoretical approaches included the use of the representative-agent model with no room for redistribution or unemployment; tax smoothing as an objective of intertemporal fiscal policy; the permanent income hypothesis approach to managing natural resource revenues; and tax effort studies. Empirical strategies were dominated by panel estimation, often in a cross-country context; time-series estimation, including causation studies; and reduced-form estimation. Problems of endogeneity and causation were often not addressed satisfactorily or convincingly.

57. Working with macro-data will typically not be sufficient for studying the kinds of taxation issues that are of policy importance. The panel found almost no examples of the kind

of empirical techniques that are increasingly common in research elsewhere, such as natural experiments, micro-data based studies, and structural estimation. These approaches are typically more time-consuming than their precursors and require more ingenuity. Researchers must seek out data and situations where causal effects can be isolated by means of quasi-controlled experiments.

58. Given the critical role that policy advice plays at the IMF, quality control is a relevant consideration. Working papers are the face of IMF research to the outside world, and their quality reflects the credibility of the research the Fund is undertaking. The reviewers judged that very few papers were below the line of acceptance for the series, and in that sense, the Fund's vetting process seemed to work. The bigger problem concerns the average quality of most of the working papers: most of the papers were rated as probably not publishable in good journals, and those that were formally published were produced by a small number of departments and researchers.

59. The credibility of the research program would be much improved if the proportion of working papers published in refereed outlets were raised significantly, and if the number of researchers publishing papers were to increase. Mere changes in procedure like pre-refereeing working papers would have limited payoff, although the payoff might be raised if outside reviewers were involved. Encouraging outside participation in research would help, given that academic researchers have strong incentives for publication.

60. The quality of research is intimately related to the quality of researchers in the institution. Presumably, care is taken to hire the best possible new staff, typically in competition with comparable institutions and universities.

61. Even the best trained new researchers need to continue to grow in experience and knowledge, and must keep current with developments in the economics literature at large. Exposure to these ideas can be facilitated by opportunities for training, seminars, and interaction with outside scholars. There are already mechanisms within the Fund, such as the IMF Institute seminars, for keeping researchers current with new developments in economic research, including new methodologies or tools of analysis. The fact that some of the more successful research papers were written in collaboration with outside researchers suggests that these collaborations should be nurtured.

62. In the end, the quality of research is likely to be strongly influenced by the research culture of the institution. To change this culture requires a combination of strong leadership and adequate incentives and rewards for research performance. An atmosphere of collegiality and interaction is also helpful. Some of this is fostered by collaborative research. It might also be helped by such seemingly minor innovations as the establishment of a data repository for empirical papers. This would allow studies to be replicated. It would also encourage internal information sharing and could reduce the cost of empirical research in the longer term.

ANNEX 1. EVALUATION PROCEDURES

The IEO forwarded to the review panel 60 randomly selected IMF working papers produced since 1999 in the general area of tax and fiscal policy. Each of the three panel members was designated as the primary reviewer for 20 of the working papers and the secondary reviewer for another 20. 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 (see Table 1 in main text). Secondary reviewers completed a Research Quality Indicator Form for each assigned paper. The panel members were also asked to indicate the best and worst three papers from both their primary and secondary lists.

The panel visited the IMF on Friday, February 26, 2010. They spent the day at a meeting with members of the IEO team evaluating IMF research, chaired by Hali Edison, Lead IEO Evaluator, and attended by Larissa Leony and Scott Standley. The panel was addressed by Ruben Lamdany, Deputy Director of IEO.

The meeting began with a discussion of the top and bottom rated papers of the panel members, including their strengths and weaknesses. Particular attention was given to cases where primary and secondary reviewers disagreed on top and bottom choices.

This was followed by a discussion of each of the 60 papers in chronological order by year of publication. For each paper, the primary reviewer gave a brief evaluation, and then there was a general discussion of the paper.

This was followed by a general discussion of the quality of the research overall, and a discussion of the emphasis of this report. The mechanics of drafting the panel's report were agreed to.

Staff at the IEO prepared a dossier of background material that included the following items:

- A summary of all research output produced by the IMF in all research areas over the period 1999–2008, of which the largest categories were working papers and selected issues papers (about 40 percent each).

- The numbers of working papers produced by each IMF department and by broad topic over the same period.

- The numbers of fiscal policy working papers by department and by broad topic.

- The numbers of fiscal policy working papers by department, focusing on broad tax policy issues from which papers for this review were selected, consisting of 120 papers of which half were from the Fiscal Affairs Department (FAD); the 120 papers included 101 from the revenue and tax category and the remainder from the general fiscal policy area that focused on tax issues.

- The allocation across departments of the 60 working papers randomly selected from the previous list of 120.

- A summary of which of the 60 working papers selected (a) had originally appeared as selected issues papers; (b) had been formally published; and (c) had been written by IMF authors alone, by external authors alone, or with external coauthor(s).

ANNEX 2. ILLUSTRATIVE DATA

Table A.2.1. Ratings for Papers by Department

	Primary Reviewer						
Department	**Number**	**S**	**AA**	**A**	**BA**	**U**	**Average**
AFR	7	0	0	4	3	0	2.6
APD	2	1	1	0	0	0	4.5
EUR	4	1	0	2	1	0	3.3
FAD	30	4	7	13	6	0	3.3
INS[1]	3	0	1	1	1	0	3.0
MCD/MED[2]	5	0	0	1	3	1	2.0
RES	5	1	1	3	0	0	3.6
WHD	4	0	0	0	3	1	1.8

	Secondary Reviewer							
Department	**Number**	**S**	**AA**	**A**	**BA**	**U**	**Average**	**Overall Average**
AFR	7	0	3	3	1	0	3.3	2.9
APD	2	0	1	0	1	0	3.0	3.8
EUR	4	0	1	1	2	0	2.8	3.0
FAD	30	2	10	9	9	0	3.2	3.2
INS[1]	3	1	0	1	1	0	3.3	3.2
MCD/MED[2]	5	0	2	0	3	0	2.8	2.4
RES	5	3	1	0	1	0	4.2	3.9
WHD	4	0	0	1	3	0	2.3	2.0

[1] IMF Institute.

[2] The Middle Eastern Department became the Middle East and Central Asia Department in 2003.

Table A.2.2. Types of Papers

	Number	**Percent of Total**
SIPs that became WPs	14	23.3
WPs published	15	25.0
WPs with external coauthor	14	23.3
WPs with only external authors	3	5.0
WPs with only IMF affiliation	43	71.7

Table A.2.3. Overall Ratings by Primary and Secondary Reviewers

Rating	Primary Reviewer		Secondary Reviewer	
	Number	**Percent of Total**	**Number**	**Percent of Total**
S	7	12	6	10
AA	10	17	18	30
A	24	40	15	25
BA	17	28	21	35
U	2	3		0
Average[1]	**3.1**		**3.2**	

[1] Based on a scale where S=5 and U=1.