IEO
Independent Evaluation Office
of the International Monetary Fund

BACKGROUND PAPER

BP/11/04

# An Examination of the Quality of a Sample of 60 Selected Issues Papers

Marcelo Selowsky and Marko Skreb

May 20, 2011

BP/11/04

# IEO Background Paper

Independent Evaluation Office
*of the* International Monetary Fund

# An Examination of the Quality of a Sample of 60 Selected Issues Papers

Prepared by Marcelo Selowsky and Marko Skreb

May 20, 2011

## Abstract

This study reports on an assessment of the technical quality of a sample of 60 Selected Issues Papers (SIPs), which were prepared as part of IMF Article IV consultations. To be effective, these papers need to address policy issues in a way that can be understood by the economic community in the country in question. About one-third of the evaluated papers were found to be better than satisfactory by both readers; they included very good and excellent papers. Good papers addressed well-defined and relevant questions and exhibited knowledge of country context—they made intuitive use of economics and the technique used matched the question. Approximately half of the papers were judged as satisfactory but exhibited specific elements of weakness. Finally, 12 percent of papers were judged to be unsatisfactory by both readers. SIPs prepared for advanced countries were typically found to be better than those for low-income countries. Common factors were identified in weak papers: they had a cryptic definition of the issue to be addressed and the relevance to the country was often not convincing; they showed a weak knowledge of basic institutional country context and often lacked the minimum data needed to address the issue; they exhibited an excessive eagerness to apply a quantitative technique without a good explanation of the economics behind the technique; and they seemed to be prepared with little time and with authors having not spent enough time in the respective country. The evaluation offers some recommendations to improve the management of SIPs.

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

JEL Classification Numbers: A1, F5

Keywords: Selected issues papers, IMF bilateral surveillance, quality comparison, research evaluation

Author's E-Mail Address: mselowsky@imf.org

**Contents** **Page**

# I. INTRODUCTION[1]

1. This background paper for the IEO evaluation of IMF research examines the technical quality of the IMF's selected issues papers (SIPs), based on a sample of 60 such papers that were issued during 2004–08. SIPs serve as analytical background for key policy issues discussed during Article IV consultation missions, which are part of the Fund's surveillance mandate. The analysis contained in SIPs plays an important role in the provision of policy advice and therefore differs somewhat from research contained in working papers. This difference is an important element in shaping the judgments of this assessment. The introduction describes how the evaluation sample was drawn, the key dimensions of quality being assessed, and the evaluation criteria. Section II presents the overall results. Section III discusses the patterns of strength and weakness in the sample papers; and Section IV offers recommendations.

## A. Sample Selection

2. The sample of papers that were evaluated was drawn from a universe of 1,110 SIPs issued by the IMF during 2004–08. The final sample excludes the SIPs for 21 countries that are featured in the in-depth case studies for the IEO evaluation of IMF research. The universe of papers was classified into three groups: low-income countries (LIC), 290 papers; middle-income countries (MIC), 472 papers; and high-income countries (HIC), 348 papers.

3. To draw the evaluation sample, 12 countries were selected from each of the first 2 country income groups and 6 from the third group, making sure that the selection included countries covered by different area departments in the IMF and deliberately overrepresenting the LIC and MIC groups. For each selected country, those SIPs issued in 2004–08 were identified. Using a random-number generator, the SIPs for each country were ranked and the first 2 were selected. The evaluation sample thus contained 24 SIPs from each of the first 2 country groups and 12 from the third country group, yielding a total of 60 SIPs.

## B. Evaluation Criteria

4. The template reproduced in Table 1 shows the seven evaluation criteria, which fall into three categories. The first category of criteria refers to how clearly the question addressed in the paper is posed, and how well its relevance—particularly its policy relevance to the country in question—is explained. The paper needs to convince the reader of the importance of the question, particularly for the authorities and think tanks/academics in the country. Criteria in the second category examine how the question is addressed. Does the analytical framework and data being used match the question? Here the analytical framework is defined to include the technique being used as well as the economic reasoning behind it;

[1] We wish to thank Hali Edison, Ruben Lamdany, and Peter Montiel for useful inputs and suggestions.

technique alone does not suffice—the underlying economics reasoning must be clear and intuitive. A discussion of limitations and robustness is important to make the results credible and usable by the economic community in the country in question, and also to encourage further work on the topic. The third category of evaluation criteria examines how the conclusions are delivered. Are they clearly presented? Do they follow logically from the earlier analysis and are their implications for policy well articulated?

Table 1. SIP Quality Assessment

| **Evaluation criteria** | **E** | **VG** | **S** | **MU** | **U** |
|---|---|---|---|---|---|
| **The question** | | | | | |
| Question is clearly posed and its relevance to the country well articulated | | | | | |
| **Analysis** | | | | | |
| Uses an appropriate theoretical/conceptual framework | | | | | |
| Uses appropriate data and empirical methods proficiently | | | | | |
| Includes critical discussion and/or robustness analysis of results | | | | | |
| **Output** | | | | | |
| Writing is clear and well organized | | | | | |
| Conclusions are firmly grounded in the analysis | | | | | |
| Articulates clearly the policy relevance of findings for the country in question | | | | | |

5. Five ratings were used to assess the papers on each evaluation criterion: “excellent” (E), with a score of 5; “very good” (VG), with a score of 4; “satisfactory” (S), with a score of 3; “moderately unsatisfactory” (MU), with a score of 2; and “unsatisfactory” (U), with a score of 1. The average score of the paper is the simple average of the scores given to each of the quality dimensions. Each paper was read and scored independently by the two authors (readers).

6. In scoring the papers, judgments were made that take into account the particular role and major audience of SIPs. These papers accompany Article IV consultations and ought to address issues of high policy relevance. Their basic audience is in the country where the consultation is taking place and consequently SIPs need to address policy issues in a way that can be understood by the economic community in the country. Authors of SIPs need to be sensitive to the ability of economists in the country to absorb technical material—these economists themselves will probably not come forward and voice such problems.

## II. Results

7. On the basis of the seven scores given to each paper each reader computed an average score for each paper ranging between 1 and 5. Using that average score, each reader independently ranked the 60 papers. As expected, the specific ranking of papers tended to differ between the readers. However, in ranking, the two readers independently identified

three common sets of papers: a top group of papers considered above “satisfactory” by both readers, including “very good” and “excellent” papers; a middle group of papers that ranged from “satisfactory” to “moderately unsatisfactory,” namely having some weaknesses that reduced their potential; and a bottom set of papers considered “unsatisfactory.” The groups are described below:

- ***The best 20 papers (33 percent of the sample)***. Both readers agreed independently that about one-third of the papers reviewed could be considered above “satisfactory.” Papers in this group passed the test of being in the best third of papers of each reader. These papers had scores higher than 3, namely those either “satisfactory,” “very good,” or “excellent.” Furthermore, both readers agreed independently on the top 9 papers (15 percent of the sample). These papers were judged to be “very good” or “excellent.”

- ***The middle group (55 percent of the sample)***. This group included papers with scores that according to both readers ranged between 2 and 3, namely those between “moderately unsatisfactory” and “satisfactory.” Readers often differed on the ranking of specific papers within the group: Sometimes one reader believed the paper to be below “satisfactory,” while the other reader believed the paper to be above “moderately unsatisfactory.”

- ***The bottom group of 7 papers (12 percent of the sample). These papers were considered “unsatisfactory.”*** Scores in this group ranged from 1.5 to 2.0.

**Some patterns**

8. A further examination of the groups yields some interesting insights:

- ***There were large differences in quality of papers across income groups.*** The HIC papers were significantly more represented in the top group, with 8 papers. These 8 papers account for two-thirds of the HIC group. The MIC group was also represented with 8 papers, but these 8 papers account for only 40 percent of the MIC group. The results were significantly less favorable for the LIC group. Only 4 of 20 papers in the LIC group were included in the top one-third of papers.

- ***A clearly posed and relevant question was an important feature of the top one-third of the papers.*** Of the 40 combined scores assessing this dimension in the top 20 papers 36 scores were 4 or above. An important feature of the good papers was a clear articulation of the issue being addressed and its policy relevance.

- ***The unsatisfactory papers were rated “unsatisfactory” in basically all the criteria of quality***. Out of a total 98 possible grades given to this group (7 papers, 7 quality dimensions, 2 reviewers), 89 were scored 2 and below.

- ***On the other hand, the top 9 papers tended to be rated "very good" or "excellent" in most criteria of quality.*** Almost 90 percent of the 126 possible grades given to this group (9 papers, 7 quality dimensions, 2 reviewers) were above 4.

## III. Qualitative Issues

9. This section discusses some problems that explain differences in the quality of papers, in particular the differences between the papers in the middle group and the papers in the top group. Weaknesses in addressing some common themes are also discussed.

### The lack of precision in identifying the issue or why a question needs be addressed seemed adversely to influence the later stages of the paper

10. Sometimes many objectives were specified and it was not obvious why they were critical, for example, the impact of a relatively small policy change today on many (endogenous) variables over both the short and medium term, which is obviously an ambitious task. Sometimes it was not obvious why a problem was a welfare problem, be it a small deviation in inflation rates with respect to the target or a small correction needed in the fiscal balance. This lack of precision in defining the rationale for the problem being addressed made it difficult to assess the robustness of results and the policy conclusions.

### Too much eagerness to quickly apply a particular technique with which the author was familiar; in some instances the question being addressed seemed to be an afterthought

11. Several of the weaker papers had a very cryptic definition of the problem to be addressed and its relevance, but then they moved quickly into specifying a complex econometric model (usually of a reduced form) with little intuitive explanation of its structure and why that model was critical to the problem at hand. Some papers attempted to derive a set of results for a large (ambitious) number of endogenous variables that were not intuitively linked to the question. Other papers in this category did not go through a process of explaining why that technique was critical to the question and what the economic forces were behind that reduced form.

### Tendency to estimate country-wide/multi-sector econometric models when the issue could be addressed by a simpler model but with better and richer country-specific data

12. Some papers tried to estimate the impact of developments or shocks in a very large country (the U.S. or the rest of the world) on a smaller country (where the Article IV consultation was taking place) which is basically a "price taker" from the larger country (or the rest of the world). These papers tried to model fully *both* countries, using very aggregate data and many parameter assumptions. As a consequence, it is difficult for readers to assess the robustness of the results. An alternative could have been to simply simulate a specific shock emanating from the larger country/rest of the world on the small country, for example, taking the shock from the large country as given. In this case the modeling could

focus on the smaller country and concentrate the effort in further disaggregation and collection of country-specific data (that the author could have collected during his/her mission).

13. In several cases there appears to be an inclination to assess the impact of a disturbance on an excessively large number of variables. As a consequence many parameter assumptions need to be made. Why not focus on a few outcomes for which a simpler but more robust model could be estimated? One gets the impression of an eagerness of authors to go to "general equilibrium" too quickly, without first carefully posing the need for such an approach given the problem at hand.

**Very quick use of cross-country econometric estimates then applied to the country in question**

14. Some papers estimated cross-country relationships with very aggregate cross-country data and then applied the results (sometimes only at the end of the paper and rather cryptically) to the country in question (where the Article IV consultation was taking place). This was done without first examining whether such a relationship could have been estimated within the country by using time series data. Authors should make extra effort to improve their country database during their visits based on the specific hypotheses they want to test. There is nothing wrong with cross-country analysis but a chance should be given to time series country data analysis, at least for an initial exploration of country hypotheses.

**Sometimes papers showed few efforts to collect data/hypothesis within the country; sometimes it seemed as if authors had seldom visited the country**

15. The quick use of aggregate cross-country data for econometric analysis rather than sharpening specific hypotheses and disaggregating the analysis within a country was a major syndrome found in some papers. Many papers used very aggregate data/indicators collected by other institutions. One gets the impression authors did not try to sharpen the quality of country data or get better data during their country visits. This may call for lengthier country visits or for stronger collaboration with local researchers. For example, several papers discussed the need to reduce obstacles for the private sector and improve the business environment. Published data from business indicators from the World Bank and UN were used. One would have expected that authors would have sharpened such indicators during their visits, unbundling them more, or would have deepened their understanding of what ultimate factors were behind those obstacles. They should have focused on some specific and critically important constraints and examined why they had not been addressed in the past. For example, papers elaborate very little on the political economy of reform: if reforming such obstacles was so beneficial why had these reforms not taken place in the past? There seems to be insufficient institutional knowledge and one is sometimes left with the impression that authors seldom visited the country.

### Weaknesses in addressing common themes

16. Some papers tried to estimate *output gaps* on the basis of past time series and using different methodologies. Presumably the purpose of the exercise was to provide the authorities with a forward-looking tool that would allow modulating fiscal/monetary policies to the possible emergence of such a gap in the outer years. However, these estimates were sometimes explained too cryptically. Estimates should have been more qualified and the robustness of results more carefully assessed. Papers should provide examples on how to use such estimates in a forward-looking context and discuss the implications of possible structural breaks. The policy implications of these results should be better discussed.

17. Several papers estimated *real exchange rate (RER) equations* based on time series data, relating RER levels to economic fundamentals. Presumably, this would allow judging possible RER misalignment situations in the future. Sometimes these papers did refer to future sustainability/risks issues but without explicitly linking these topics to the equation being estimated. Future sustainability of specific RER levels presumably depends on the future sustainability of the fundamentals—fiscal deficits, foreign aid, terms of trade, etc. The exercises usually do not put these estimates in "future motion," discussing sustainability in terms of future developments in the fundamentals. For example, the estimated coefficients could be used to assess how RER will change according to future aid flows—this may be an important area in aid dependent countries. In deciding what past data on RER to use (in estimating these equations), it is important to be aware of periods where the exchange rate (ER) regime has been characterized by multiple exchange rates or segmented ER markets. This calls for careful decisions on what constitute the "binding" ERs in building the appropriate time series.

18. A topic of several SIPs was the analysis of *sources of growth*. The papers emphasized the importance of increases in education and skills. However, these papers showed a weak familiarity with the well-known literature on how to build a labor skills index based on changes in the educational distribution of the labor force in order to measure the contribution of education (Griliches and Jorgenson, 1967).[2] In much of the literature such correction has reduced the "residual" and explained an important share of growth in developed and developing countries. Authors could have easily built such indices from wages by educational levels and data on changes in the educational distribution of the labor force.

19. *Clearer separation between once and for all effects in price levels from sustained inflation trends*. Several papers tried to analyze the impact of shocks (world food or energy prices, increased political risk) on inflation, but they did not spell out the mechanism by which these impacts might translate into sustained periods of inflation. Again, this was the

[2] Zvi Griliches and Dale W. Jorgensen (1967), "The Explanation of Productivity Change," *Review of Economic Studies,* Vol. 34, No.3, pp. 249–83.

result of using reduced form equations without a clear discussion of the underlying causal mechanism. Several papers were not sensitive enough to the difference between trends in inflation and once and for all changes in the price level.

**Common attributes of strong papers**

20. Many of the good papers relied on simpler analytical tools or simpler models incorporating good economic analysis. All these papers had very specific, well-defined, and convincing welfare questions. The authors seemed to be familiar with the country context and country data, and in many cases they also made special efforts to collect relevant data. Some of these papers focused on fiscal issues such as institutional aspects and reforms of the fiscal framework, fiscal rules, debt sustainability issues, competitiveness in the context of expansions in external flows, specific issues in the banking systems, and specific issues of tax reform.

## IV. Conclusions and Recommendations

21. During the period covered by the evaluation, the Fund produced about 220 SIPs a year, accounting for about 40 percent of total research output.[3] Among the Fund's research outputs, SIPs probably have the most significant potential to enhance the policy advice of the Fund as part of its Article IV surveillance mandate. These papers also carry important reputational responsibilities for the institution. Furthermore, in the lower income countries, SIPs are particularly important given the scarcity of policy-oriented research. Given the resources and talent available to the institution, weak papers could easily be improved to become as good as the top one-third of papers identified by this evaluation.

**The following are specific recommendations to help improve SIPs:**

(i) *The Fund should better clarify the exact objective and function of SIPs* and how they differ from other analytical outputs of the institution such as working papers, technical assistance activities, etc. If the objective is to assist country authorities and the economic community in addressing major policy areas, the institution should make this clear and draw its implications (below). SIPs should not be vehicles whose main purpose is to test techniques or specific models. Such objectives should be served by other research outputs.

(ii) *The Fund should pay special attention to how SIP topics are selected.* A lack of consultation on topics to be addressed by future SIPs was a major finding that emerged from the country visits that were undertaken for the research evaluation. Such consultations should improve the relevance of the issues addressed and increase

[3] IEO Background Document II: "IMF Research: Taking Stock" will be available at: www.ieo-imf.org.

the sensitivity of IMF staff to the way SIPs are written and disseminated in order to be more effective. Collaboration with local institutes in low-income countries could be important to learn about the country context and institutions as well as to improve the quality of data.

(iii) *The Fund should focus on areas of examination where it has expertise and a mandate*. Too wide a scope of topics has the risk of lowering the quality of SIPs and opens questions of overlap with other institutions.

(iv) *The Fund should pay more attention to the structure of a SIP and how to make it more "reader friendly."* Some suggestions are given below:

- First, clear executive summaries are needed. Authors should keep in mind that they are addressing policymakers (among others) with limited time.

- Second, econometric or modeling techniques should not occupy the center of attention in SIPs, but should be included in appendices or reserved for the working papers series. SIPs could then extract from such working papers the most intuitive economic analysis and conclusions.

- Third, SIPs should be careful in giving policy advice as many emerging/low-income countries have limited institutional capacity for reforms. SIPs should try to be focused and specific in their recommendations and avoid being vague in their advice (like strengthening tax revenues, improving financial stability, etc.). Stronger collaboration with authorities and institutions will allow a better understanding of specific conditions in a country.

- Fourth, many papers were based on weak data and authors should make explicit the implications for the robustness of results. Again, this may be particularly important for low-income countries. If the topic to be selected is critical but the required data is poor, Fund staff should work with researchers in the country to improve such data.

(v) *A major conclusion of the evaluation is that in order to produce a good SIP more time should be spent at all levels of the process*. This would include more time with authorities in identifying issues, more collaboration in the field to increase country knowledge of context and institutions, and stronger efforts to obtain a minimum quality of data. In the context of a given resource envelope this could be achieved by reducing the number of SIPs per year and focusing more effort on low-income countries. For example, producing only one SIP per country every two years may allow for a much better crafted paper. This could be accompanied by a stronger quality control process within the institution.