Issues in Comparison and Aggregation of PEFA Assessment Results Over Time and Across Countries

Final
PEFA Secretariat
May 13, 2009

TABLE OF CONTENTS

Page No.

Executive Summary

i. PEFA assessments are carried out in order to identify strength and weaknesses of a particular country's PFM systems and to establish performance baselines, used for formulation of reform programs and for monitoring performance progress over time.

ii. With comprehensive data now available on PFM performance in several countries, the question arises how to use the data to make meaningful comparisons of scores over time within countries and cross-country comparisons both at a given point of time and over time.

iii. The most basic – and universally recommended - approach to do this would be by comparing scores for each indicator and using the narratives in the PEFA reports (PFM-PRs) to draw conclusions on reasons for differences in scores. A comparison for this nature helps interpretation of country difference in relation to the country context of the respective countries. Whilst an analysis of this nature would be a standard requirement for a comparison of performance over time in one country, it may be complicated and time-consuming in connection with cross-country analysis of more than a few countries. In neither case, however, does the analysis lead to simple measures that are easily understandable to non-experts and easily transmitted to decision-makers. The latter is a particular concern when a substantial number of assessments are compared.

iv. Assessment report authors, researchers and other users, therefore, have sought an aggregated measure of PFM performance to facilitate cross-country comparisons of scores, and in a few cases, to facilitate performance progress over time in a given country. Substantial interest in such comparisons has emerged on the basis of the rapidly increasing availability of PEFA indicator data. The PEFA program strives to make such data available to potential users.

v. Aggregation of the results of a country assessment typically involves both the conversion of ordinal indicator ratings to numerical value and allocation of weights to the individual indicators. Where country results are clustered, aggregation also requires assignment of weights to each country.

vi. This paper presents examples of comparisons undertaken by various users and discusses the related methodological issues. There is no scientifically correct method on how aggregation and comparison should be done for each of those three levels of assumptions. Consequently, the PEFA program neither supports nor recommends any particular approach to aggregation.

vii. The PEFA program recommends that any user - as part of the dissemination of results from comparison - clearly explains the aggregation method and assumptions applied in each case, and the reasons for the choice. It would also be advisable that users undertake sensitivity analysis to highlight the extent to which their findings are robust under alternative aggregation assumptions.

1. Introduction

1. With comprehensive data now available on PFM performance in several countries, the question arises how to use the data to make meaningful comparisons of scores over time within countries and cross-country comparisons both at a given point of time and over time.

2. The most basic approach to do this would be by comparing scores for each indicator and using the narratives in the PEFA reports (PFM-PRs) to draw conclusions on reasons for differences in scores. This is time-consuming and does not lead to a simple measure that is easily understandable to non-experts and easily transmitted to decision-makers.

3. It is tempting, therefore, to seek an aggregated measure of results to facilitate cross-country comparisons of scores; either an overall measure of aggregate country performance or for each of the six core dimensions in the PEFA Framework (hereinafter termed as Framework).

4. A range of research papers and assessment reports have already used PEFA assessment scores to facilitate such comparisons, either by calculating aggregate numerical scores for each of the six core dimensions in the Framework and/or by aggregating into one measure of overall PFM performance.

5. The purpose of this paper is to illustrate methods applied in comparing PFM system performance between countries, using indicator scores from PEFA assessment reports, and to point out the potential pitfalls of using such methods. The paper is oriented towards cross-country comparison in terms of the performance of PFM systems as measured according to the six core dimensions of the PEFA Framework.[1] The issues raised, however, also apply to users of the PEFA assessment reports who want to use different subsets of indicators or use indicators on an individual basis.

6. The rest of this note is structured as follows:
   Section 2: a description of purposes of within and cross-country comparisons;
   Section 3: some examples of methods used for comparison and aggregation
   Section 4: a discussion of issues arising aggregation and comparison
   Section 5: the conclusions

[1] These are: Credibility of the Budget (PIs 1-4), Comprehensiveness and Transparency (PIs 5-10), Policy-based budgeting (PIs 11-12), Predictability and Control in Budget Execution (PIs 13-21), Accounting, Recording and Reporting (PIs 22-25), and External Scrutiny and Audit (PIs 26-28).

2. Purposes of Within- and Cross-Country Comparisons

7. Whilst the PEFA Framework was primarily intended for the purposes of tracking progress over time in a particular country, the use of the PEFA Framework to compare PFM performance across countries serves a number of stakeholder interest. Therefore, stakeholders are increasingly seeking a reasonably manageable way to undertake such comparisons (i.e. not just by comparing scores for each and every PEFA indicator). Stakeholder interests include:

.

- Governments may want to monitor their PFM performance overtime in terms of the core dimensions of the PEFA Framework and compare their performance with those of neighboring countries or countries at similar or different stages of PFM reform and/or economic development in general.

- Donor agencies may want to monitor PFM performance over time within their partner countries and compare PFM performance across their partner countries in order to monitor progress related to their aid operations or to draw lessons that might help in the provision of advice and assistance to governments in the drawing up/revising of their PFM reform programs.

- Clustering PFM performance measures across countries is also a requirement under The Paris Declaration on Aid Effectiveness (indicators 2a, 5 and 7) in terms of global monitoring of progress in PFM.

- Finally, academics may want to seek explanations for variance in PFM performance by correlating differences in PFM performance across countries with possible explanatory variables reflecting country characteristics, for example per capita incomes and population size.

3. Examples of Methods Used in Aggregation and Comparison

3.1 Within Country Comparison

8. The PEFA Framework was created with the primary purpose of measuring progress over time. During 2008, repeat assessments that track such progress began to emerge[2]. In order to more effectively convey messages about performance changes over time, report authors have tried to simplify the results into fewer measures than the listing of 28+3 indicator ratings in two years, combined with a narrative discussion of the comparison. Two methods of aggregation have been noted.

9. One method has been to count the number and magnitude of rating changes and add them up. A simplified example is given in table 1 below. The resulting measure is very simple and easy to convey to non-specialist audiences, but are subject to assumptions about both conversion from the ordinal scale to numerical intervals and to the issue of assigning weights to indicators e.g. in the example, has the improvement in PI-11 the same or a higher value than the deterioration in PI-14? Should a change from a 'no score' to a D+ be considered a performance improvement or not? And therefore, what does an aggregated one point improvement for 8 indicators actually mean?

Table 1 Illustrative example on issues in aggregating performance changes

Indicator	2006 score	2009 score	change
PI-11	C+	B	+ ½
PI-12	D	C	+ 1
PI-13	No score	D+	n.a.
PI-14	B+	B	- ½
PI-15	C	C+	+ ½
PI-16	C+	D+	- 1
PI-17	B	B	Nil
PI-18	D+	C	+ ½
Overall change 1/			**+ 1**
1/ assuming equal numerical interval between all ratings on the ordinal scale and equal weight of indicators			

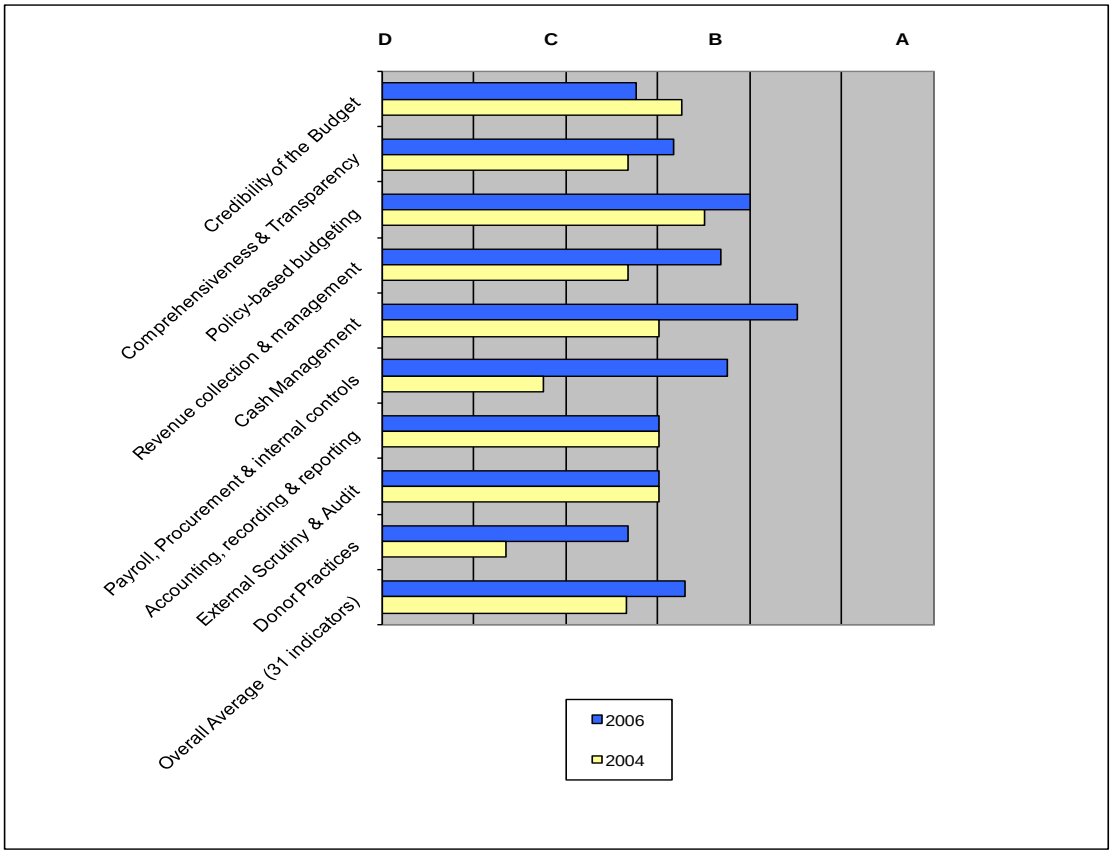
10. A second technique increasingly being used in PEFA assessment reports for monitoring PFM performance over time has been to compute "average" scores for the core dimensions, As "average" scores are much easier to compute from numbers, researchers have tended to convert alphabetic scores to numerical scores and then

[2] A few earlier repeat assessments do exist but were either done too soon after the baseline to provide meaningful measuring of progress over time, or were in fact an attempt to create a commonly agreed baseline where the initial assessment did not have full support of the main stakeholders in the country. Some assessments also attempted to measure progress since the 2004 assessment used to test the draft PEFA Framework. As the final framework changed significantly since the draft, a comparison is not possible except for a few selected indicators.

calculate a simple average, assuming equal weights for each indicator; all indicators have equal importance.

11. The Mozambique PEFA assessment finalized and published in 2008 is an example. Letter scores were converted to numerical scores from 1 to 7 (D=1, D+=2) and then aggregated for each core dimension and averaged, assuming (without explanation) equal weights. The core dimension covering Predictability and Control in Budget Execution was broken down into three core sub-dimensions: (i) Revenue collection and management: PIs 13-15; (ii) Cash management: PIs 16-17; and (iii) Payroll, procurement and internal controls (PIs 18-21), the logic being that these areas are conceptually different from each other. Donor Practices were added (D1-D3).

Figure 1: Mozambique, Tracking Progress



Note: The aggregate scores were first calculated numerically (1, 1.5…4) then converted back into letter grades.

12. Notwithstanding any issues concerning the assumption of equal weights (discussed in more detail below), the Mozambique report demonstrates that progress can potentially be tracked on a core dimension and overall basis. A key objective was to track progress since the first assessment. Charting the core dimension and overall scores for each assessment, it is possible to see clearly where progress is being made (Figure 1). The differences in the height of each bar indicate the extent of progress or regress. By looking at the change only, the issue of the validity of the aggregation

method diminishes as only the change is being looked at rather than the levels; i.e. some of the core dimension scores might be different if they were calculated on a different basis, for example aggregating according to weighted averages instead of simple averages. This assumes that the aggregation issues are the same for each assessment, probably realistic over a period of only a few years.

13. The second Zambia PEFA assessment, conducted during mid-2008, uses the same technique (also no explanation of the validity of assuming equal weights), the only difference being that the budget execution core dimension is divided into two (tax administration and budget execution) and not three components.

3.2 Comparison across countries

14. The paper "Taking Stock: What do PEFA Assessments tell us about PFM systems across countries?: Paulo de Renzio, University of Oxford, April 2008" uses PEFA scoring data to find country characteristics (such as per capita incomes and population size) that globally correlate and possibly explain difference in PFM performance among countries. The paper uses a numerical conversion of PEFA ratings with equal weights given to every indicator and equal weights also given to all countries in the sample of 57 countries.

15. The paper acknowledges the validity issues that may arise: "The use of average values could be criticized on a number of grounds, the main one being that the various dimensions of the PEFA methodology actually measure very different things, which are not necessarily amenable to quantitative conversions, calculations and analysis. Also, the use of averages is based on the assumption that all indicators are equally important". The author, however, justifies averaging on the basis of the large sample of countries used for the analysis and the rigorous methodology used.

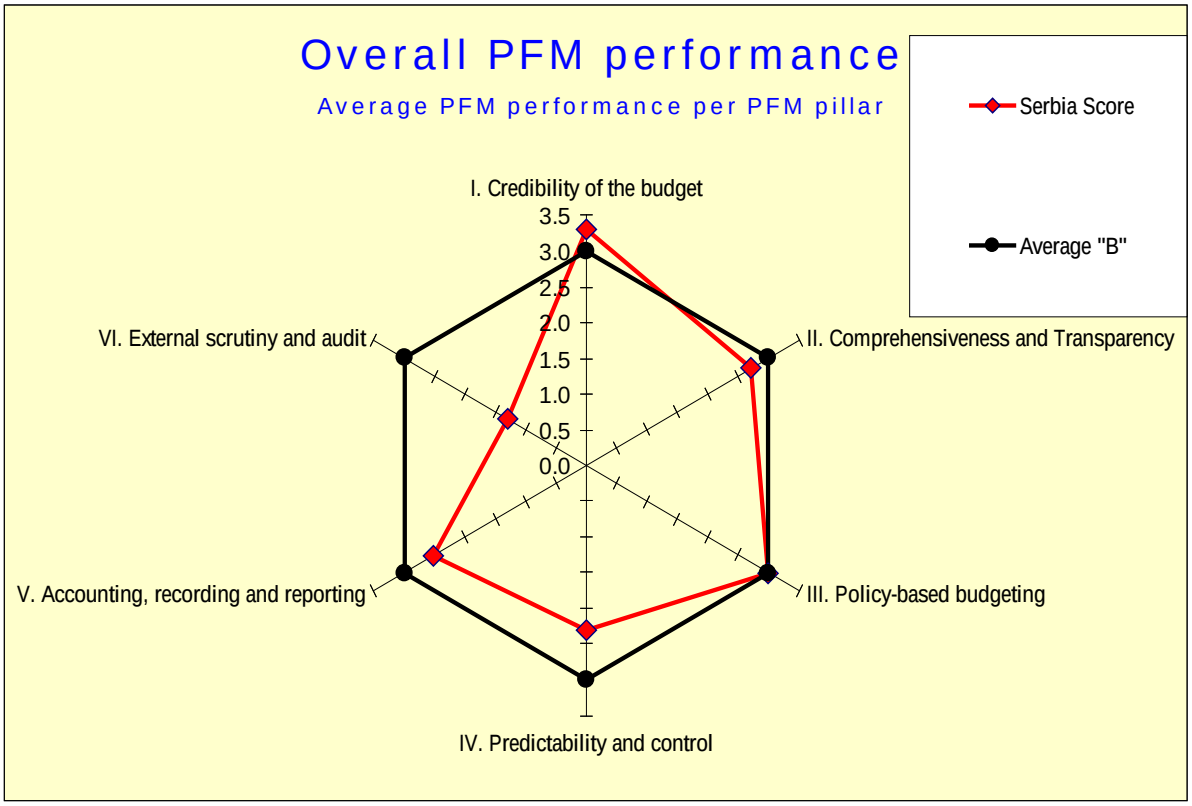
16. In other cases, assessment reports may compare the performance of the assessed country to the 'average' performance of other countries in the same region, where the PEFA methodology has been used for a PFM assessment, e.g. presented in graph as in figure 2 below. The assumptions made regarding score conversions and weights are similar to the approach in the report described above.

3.3 Simultaneous comparison across countries and over time

17. Only a handful of repeat assessments have been completed under the PEFA methodology with the explicit purpose of tracking progress in performance over time. The use of PEFA ratings for comparison of performance changes over time at a regional or global level has therefore been indirect only, or only a baseline has been established from where to measure change when the eventually the number of repeat assessments have increased sufficiently[3]. Three examples of indirect use are the following.

[3] Ref. AfDB's results framework.

Figure 2 Example of comparing performance in a country to a regional benchmark



18. The World Bank's CPIA ratings make use of PEFA assessments to inform the scoring of primarily CPIA indicator #13, but also to some extent #16 (on tax administration). There is no direct conversion from PEFA to CPIA scores, but efforts are made to ensure consistency between the two sets of ratings. The CPIA ratings are in turn used in the monitoring of the Paris Declaration to measure global improvements in the quality of PFM systems among DAC partner countries.

19. The Evaluation of General Budget Support in 2006 used PEFA indicators as a basis for measuring changes in PFM performance over time in the selected country case studies that were used for generating conclusions at the global level. The lack of repeat assessments led the evaluators to undertake retroactive indicator assessments as part of field studies and to apply a simplified version of the PEFA scoring methodology.

20. The paper "Tracking Progress in the Quality of PFM Systems in HIPCs: P. de Renzio and W. Dorotinsky (2007), PEFA Secretariat, November 2007" attempts a more direct and detailed use of PEFA scores, in a similar attempt to measure global trends in performance over time. Due to the lack of PEFA repeat assessments, the paper used HIPC AAP indicator scores as baseline and more recent PEFA scores as 'repeat' assessments. A conversion of PEFA scores to HIPC scores were then required. The report also highlights the methodological issues and their potential impact on the results.

4. Issues Arising in Aggregation and Comparison

21. Aggregation of the results of a country assessment involves both the conversion of ordinal indicator ratings to numerical value and allocation of weights to the individual indicators. Where country results are clustered, aggregation also requires assignment of weights to each country. Each of these issues are discussed in turn below.

4.1 Aggregating across indicators

Issue of Indicator Weights

22. Assigning weights to the individual indicators is necessary if the assessment results are going to be presented in simpler terms than 28(+3) individual results or a purely narrative description. Equal weights are typically given to the indicators in such exercises. But are the implications fully appreciated? For instance, in table 1 the overall change would shift from +1 point on the rating scale to +1.5 points (a significant difference), if the indicators PI-11 and PI-12 were given double the weights of the other indicators; the latter potentially justified by having only two indicators under the core PFM dimension of 'policy-based budgeting' compared to 3 to 8 indicators under the other core PFM dimensions.

23. The assumption of equal weights is obviously very convenient in terms of facilitating aggregation. It is also very convenient for users of PEFA assessment scores who may want to aggregate and average according to sub-sets of indicators that are different from the PEFA six core dimensions (e.g. the Mozambique example above and the World Bank's Country Policy and Institutional Assessment – CPIA – ratings). But is this assumption valid?

24. Assumption of equal weights may be valid, but not necessarily so and thus users of the Framework should not uncritically assume equal weights or intransparently and arbitrarily use other numerical aggregation methods (weighted average or weakest link among the indicators to be aggregated).

25. PEFA PIs 7, 8 and 9 are particularly contentious. A 'D' score under PI-7 dimension (ii) covering fiscal reporting of donor aid received in cash would result in a D or D+ score for the indicator as a whole (M1 method of scoring), but may be of little significance if such aid represents only a minor component of government expenditure. D scores under PIs 8 and 9 may be of much greater significance if SNG expenditure is a significant component of total government expenditure. A weighted average score for the Comprehensiveness and Transparency core dimension (PIs 5-10) might therefore seem appropriate. But there is no universal answer to what those weights should be.

26. The assumption of equal weights for PIs 17-21 should also not be automatic. Under PI-17 (cash and debt management), low scores for dimensions (i) and (iii) may be of little significance to the workings of the PFM system as a whole if debt finances only

a minor part of government expenditure and government loan guarantees are of only small magnitude. PI-18 (payroll management) may be of much greater importance than PIs 19-21 (concerning controls over non-wage expenditure) if personnel emoluments are a major component of government expenditure and non-wage expenditure a minor component. A low rating for PI-21 (Internal Audit) may be of little significance if development of an internal audit function is still in process and strong ex-ante internal control systems are in place. A weighted average approach for PIs 13-21 (Control and Predictability in Budget Execution) might therefore have merit.

27. The issue is not clear-cut, however. For example, weaknesses in controls over non-wage expenditure might undermine the quality of personnel inputs to public service provision (e.g. weak controls on purchases of text books could undermine the effectiveness of teachers and thus efficient service delivery)

28. In the case of PEFA PIs 1-4 (credibility of the budget core dimension), it could be argued that a weakest link approach might be appropriate for deriving an aggregate score for this core dimension. The circumstances may vary however, and it is difficult to stipulate definitively what method should be used. For example, a D grade for PI-2 but A grades for the other indicators undermines budget credibility.as MDAs are receiving significantly different financial resources than originally budgeted for them. A D grade for PI-4 but A grades for the other indicators may indicate payments that have not been budgeted for, implying the possibility of offsetting lower future budgets for MDAs relative to what they had been expecting, thus damaging budget credibility. On the other hand, a D grade might relate to large historical arrears that are gradually, but predictably being paid off.

29. Assuming that aggregation of scores according to core dimensions is valid, is it valid to derive an overall aggregation, so that just one letter (or number) indicates overall PFM performance? The answer is complicated due to the two-way linkages between the core dimensions and the fact that e.g. PIs 1-4 are measured over a rolling 3-year period and may reflect the impact of performance on the remaining indicators, most of which are measured on the basis of the most recent one-year budget cycle. Assessing the relative importance of each core dimension would be a difficult and arbitrary exercise. Simple averaging may be the only plausible method, if such an aggregation is attempted, even if the linkages are not of equal strength, but one should be aware of its limitations[4].

30. It could be argued that applying equal weights to each core dimension is not necessarily valid as the number of indicators in each core dimension varies widely

[4] A useful exercise might be to calculate correlation coefficients between average scores for core dimensions 2-6 grouped together (i.e. the average of the sum of the scores for these dimensions) and the first dimension. As shown in the Framework document, the first core dimension (credibility of the budget) is conceptualized as the outcome of the combined impact of the other core dimensions. However, a low correlation coefficient would not necessarily imply weak correlations between the other core dimensions; high scores for the first core dimension and low scores for the other core dimensions may simply indicate potential threats to the credibility of the budget. A good Summary Assessment would point this out.

between two (PIs 11-12[5]) and nine (PIs 13-21). This could mean that an aggregation directly from individual indicators to an overall average would lead to a different result than an aggregation in two steps via the core dimensions (6-7 core dimensions in the standard framework, but 9 in the case of Mozambique as illustrated above).

4.2 Converting from an Ordinal to a Numerical Scale

31. Conversion of PEFA's ordinal scale ratings (A, B, C, D) to numerical scores is used in most of the examples in section 3, except where score distributions only are used. Typically the four ordinal scores are converted to (4,3,2,1) or (3,2,1,0)[6] with '+' score given ½ point (e.g. B=3, B+=3.5).

32. An overriding issue is whether it is valid to assume that quality differences between each grade are the same and that changes in scores mean the same for all initial scores. As most of the scores are based on different steps in improving qualitatively described elements of processes, systems and output use, there is no scientific basis on which to convert to numerical value[7]. Leaving aside the issue of 'no scores' (ref. below), research conducted by the Development Research/Public Sector Governance unit at the World Bank indicates, however, that moderate variations in the numerical conversion scale are unlikely to result in significantly different findings.

Issue of MI and M2 scoring methods

33. At first sight, it may also appear invalid to derive numerical core dimension scores when the indicators are scored according to both M1 (weakest link) and M2 (simple average) methods. A 'D+' score for an M1 indicator does not necessarily mean exactly the same thing as a 'D+' score for an M2 indicator, as the process of calculating the score is different. The question is whether such differences are large enough to merit much attention. No users have so far felt that the issue was sufficiently important to merit different conversion to numerical values.

Issue of "No Scores"

34. The validity of converting letter scores to number scores and then aggregating is also affected by the incidence of "no scores" due to lack of adequate information or problems in accessing it, or indicators not rated at all because they are deemed to be inapplicable (or otherwise deliberately omitted from the assessment). In this case, aggregating and averaging may produce misleading results. If a country is missing data on indicators that are likely to score high (A or B), the aggregate score would have a downward bias, and vice versa if data is missing on indicators that are likely to score low.

[5] Note that related aspects are included in other dimensions, for example PIs 5-7, PI-10 and PI 27

[6] Whether the former or the latter conversion is used make no difference to aggregation results, unless 'no score' is also assigned a value, ref. below.

[7] The purely quantitative indicators PI-1, 2, 3 and D-3 could be considered the exceptions.

35. The question then arises whether a 'no score' should be given a value in the aggregation and should contribute to the average score or the score distribution. Most examples so far have kept the 'no scores' outside the aggregation. It can be argued, however, that 'no score' due to lack of adequate data represents a situation worse than a D rating as it reflects that not even the basic information to assess the situation is available. Some actual examples of aggregation include 'no scores'. E.g. the AfDB includes it in its corporate results framework giving 'no score' a 0 (zero) value compared to a scale that sets D=1 and A=7. This results in an aggregate score that is different from a score where 'no score' is assigned a value of 1 or not counted at all.

36. The database of assessment report scores, kept by the PEFA Secretariat, has introduced a distinction between the different reasons for 'no score' by subdividing it into NR (not rated due to insufficient information), NA (not applicable in the specific country context) and NU (not used for the assessment, being an ex-ante decision to limit the scope of the assessment). These three categories would need to be incorporated differently in an aggregation of performance results, with NR being the only category that could conceivably be compared to or aggregated with other ratings.

Frequency distributions

37. An alternative method of aggregation, that avoids a numerical conversion, is to compile frequency distributions of scores in chart form: number of 'D's as a percentage of total scores, number of 'C's,…number of 'A's. A scores and D scores usually comprise a smaller percentage of scores than B and C scores, so that such charts would tend to have a "tail" at each end. Changes in PFM system performance over time in a country can be assessed by comparing frequency distributions on one chart, the proportion of As and Bs hopefully increasing over time.

38. Although use of the frequency distribution method avoids numerical conversions, some of the validity issues remain the same. The implicit assumption is that all 'A' scores (for example) have the same value; i.e. all indicators have the same weights and that performance at a particular rating level are equal - or at least comparable - across the indicators. A specific disadvantage relative to the numerical conversion method, is that showing frequency distributions for each core dimension at different points in time on one chart would be difficult; Figure 1 above shows that this is possible, but only if using a numerical conversion method.

39. Whilst presentations of frequency distributions have tended to show the scores as if they have equal value intervals, the presentation of score frequencies may also be used to illustrate the link between a base score (e.g. a 'B') and its '+' counterpart (B+) by showing the two back-to-back with an interval to the next group (say 'C' and 'C+' scores).

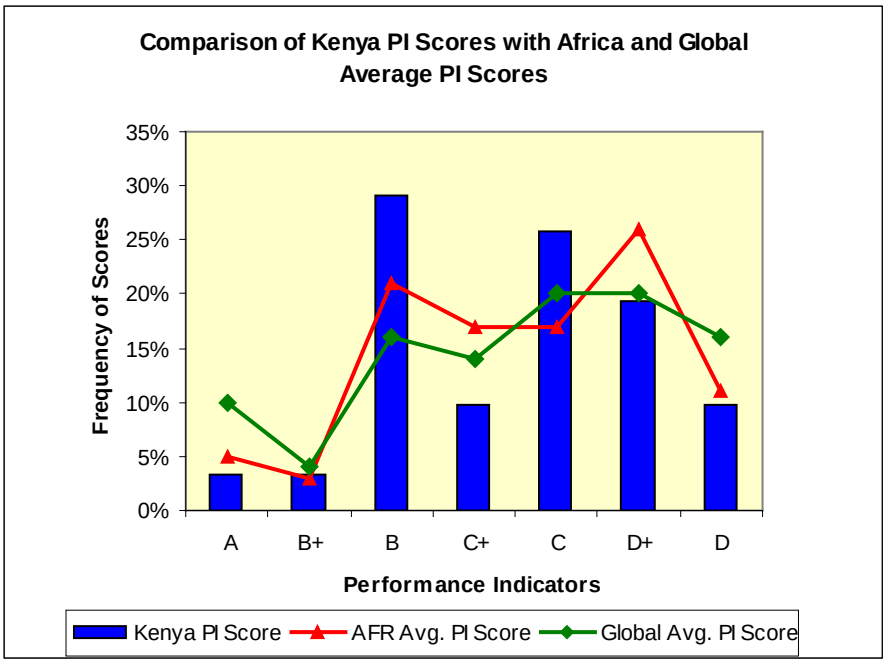
4.3 Clustering of Countries

40. The numerical aggregation methods outlined above can also be used to facilitate cross-country comparisons, both at a point in time and over periods of time. Using the

numerical conversion method, Figure 1 above could be modified to show overall aggregate scores for different countries or clusters of countries, each represented by a bar. Clusters could represent regions, or groups of countries with similar characteristics. Or it could be modified to show overall aggregate scores for one country against a regional or global average. Another option would be to show average scores of each core dimension for all or a sample of countries. It would be difficult, however, to show on one chart inter-country comparisons by core dimension, though a number of charts (one for overall aggregate and one for each core dimension) could be shown on one page (also see paragraph 45).

41. The frequency distribution aggregation method can be used in much the same way. As an example, Figure 3 compares the frequency distribution of scores for an African country (Kenya) with the frequency distribution of scores for Africa as a whole and globally for all countries.

Figure 3: Frequency Distribution of Scores: Kenya compared with Global and Regional Average.

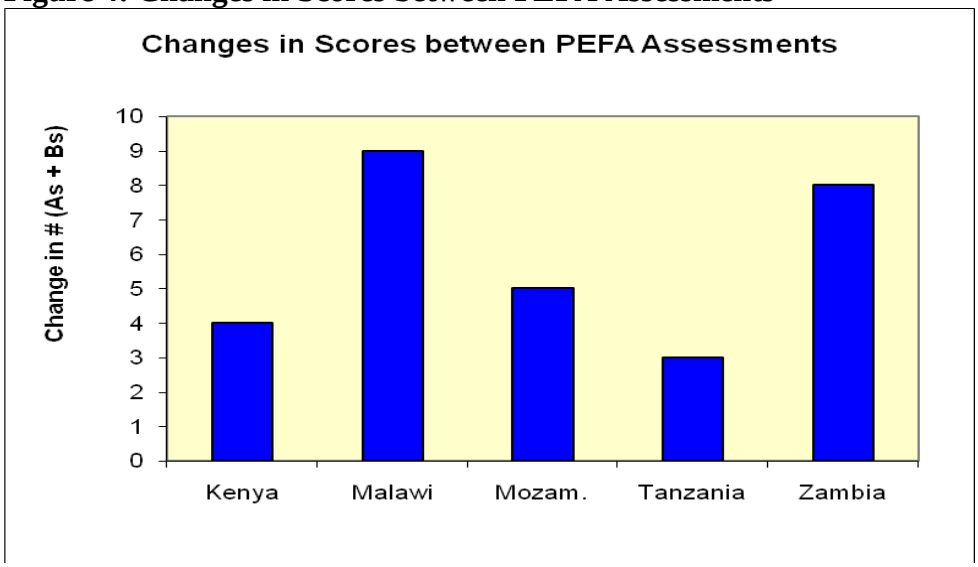


42. More countries can be shown on one chart by combining scores, for example by showing the number of As and Bs together and the number of Cs, Ds and NR scores together. This highlights more clearly the difference in PFM performance between countries.

43. Progress can also be tracked over time using this method, by showing the changes in numbers of As and Bs combined in terms of a repeat assessment. By definition, an increase in the numbers of As and Bs combined is identically equal to the reduction in

the numbers of Cs, Ds and "Not Rated" scores combined, so it is not necessary to also show the latter. Figure 4 provides an example.

Figure 4: Changes in Scores between PEFA Assessments



i) Dates of assessment(s): Kenya; 2006, Malawi, 2006 and 2008; Mozambique, 2006 and 2007; Tanzania, 2006; Zambia, 2005 and 2008.

44. For Figure 4, actual PEFA assessment scores have been used for the PEFA repeat assessments conducted for Malawi and Mozambique. [8]. Purely illustrative scores for the latest assessments have been used in the cases of Kenya, Tanzania, and Zambia. The figure shows that Malawi has made the largest gains in terms of increases in As and Bs, followed by Zambia and Mozambique. Figure 4 could also be used to compare PFM performance between regions, each bar representing a region.

45. The frequency distribution of scores can also be used to compare scores for each core dimension for a group of countries or even for specific indicators. Eight charts on 1-2 pages can be compiled, showing overall progress in reform and progress of reform in each of the six core dimensions. In this way, it is possible to see roughly what is the contribution of progress in reforms under each core dimension to overall progress.

Issues Arising in Cross-Country Comparisons

46. Cross-country comparison issues arise in many areas of socio-economics and PFM performance indicators are no exception. The risk is that "like" will not be compared with "like". Differences in statistical methodology and problems in data collection mean, for example, that care needs to be taken in comparing GDP estimates between countries. Differences in definitions may also preclude meaningful cross-country

[8] Based on the final reports of Malawi, Mozambique and Zambia. The reports for Kenya and Tanzania were not completed at the time of calculation. Once these have been finalized and Uganda added to the above (the repeat assessment conducted in late 2008 is close to being finalized), Figure 4 might show an interesting picture.

comparisons. For example, government expenditure on health services is often compared between countries in terms of percentages of expenditure and GDP, without due regard to differences in definitions of the health sector between countries (e.g. in some countries, water and sanitation is included, in others it is excluded), differences in the extent that the private sector is involved in providing health services, and differences in health conditions.

47. Even if aggregate scores for the six core PFM dimensions can be validly estimated, there will be a number of "like with like" issues in cross-country comparison of PEFA indicators, , including:

- The scope (central government only, SNG only, combined) and year of the assessment may not be the same for the countries being compared;

- Definitions of PFM terms may differ across countries, for example, payments arrears, government entities, fiscal transfer systems.

- The reasons for the same scores across countries for each indicator may vary considerably, as most indicators have at least two and sometimes four dimensions. For example, two countries may obtain the same score for PI-8 but for completely different reasons: one country may receive an A score for dimension 1 and a D score for dimension 3, while the other receives a D score for dimension 1 and an A score for dimension 3. Cross country comparisons may therefore be superficial and require dimension by dimension comparison, a more complicated exercise.

- The quality of assessments may vary across countries. The quality assurance process provided by the Secretariat (and other peer reviewers) identifies weaknesses in PEFA assessment reports. In the case of the Secretariat, the weaknesses consist mainly of indicator scores not backed up by sufficient evidence (most common) and incorrect scores assessed on the basis of the evidence provided. It is not necessarily the case that reports that have been revised following peer reviewer comments are sent to the Secretariat for further review.[9] Thus finalized reports may still contain scoring errors.

- The case for assigning weights to indicators in deriving an aggregate score for the core dimension or overall (simple averages - assuming equal weights - weighted averages or weakest link approaches) will most likely differ between countries, as explained above.

- Differing proportions between countries of public expenditure financed directly by the government and financed directly by donors. For example, comparison of PEFA scores between a country where, for example, 50 percent

[9] PEFA Secretariat: Report on Early Experience of the Application of the Framework, November 2006 and PFM Performance Measurement Framework Monitoring Report, March 2008, www.pefa.org.

of public expenditure is financed directly by donors and a country where zero percent is financed by donors is perhaps less meaningful than comparisons of PEFA scores between countries with similar such percentages.

- The incidence of "Not Rated", "not used" and "not applicable" indicators may differ across countries, thus lessening the comparability of averaged scores as not all of the same indicators may have been rated;

- Another issue is whether country scores should be weighted according to size, population for example. This issue is most likely to arise within a regional cluster of countries. A very small country within the region may score all Ds, reducing, if equal weights are used, the average score (using the numerical conversion method) for the region if the other much larger countries are scoring mainly As and Bs. The method chosen might affect the comparability between country clusters.

4.4 Ways to mitigate against comparability issues

Large sample size

48. The larger the sample size, the less the risk of invalid comparisons. The largest sample sizes would arise: (i) where PEFA core dimension scores are being compared to each other across all countries where PEFA assessments have been carried out (i.e. if the number is 50, the score for each core dimension is calculated as the average of the scores for 50 countries); and (ii) where one country is compared to the average of all other countries (e.g. see Figure 3 above), assuming the country is a reasonably representative one.

49. The smallest sample for comparing PEFA core dimension scores would be the case where the core dimension scores are compared between each individual country; in this case doubts might arise about the validity of such comparisons. Dividing countries into "clusters" would increase the sample size. However, the clusters will contain only limited numbers of countries, and validity of comparison issues may still arise.[10]

Comparisons between countries with similar characteristics

50. Examples include Former Soviet Union countries in a particular geographic region (e.g. Caucasus), African Anglophone countries in a particular region, for example East Africa, and the English speaking Caribbean countries, that have similar country, legal and institutional characteristics and have broadly followed the same path of PFM reform (even under the same donor-supported technical assistance program).

Emphasize changes in PFM performance

[10] Page 10, de Renzio paper; comparisons are shown both in tabular and bar chart form.

51. The issue of sample size also becomes less important if changes in PFM performance over time are compared, rather than the level of PFM performance at a point in time. Assuming that country-specific issues are relatively unchanged over time (probably a reasonable assumption), the comparison of changes in PFM performance eliminates the influence of such issues. However, any report in which such cross-country comparisons are being made would need to demonstrate that country-specific factors have not changed over time and thus that changes in scores over time have equal significance (in terms of the change in the quality of the PFM system) for all countries being compared.

5. Conclusions

52. Comparison of assessment ratings over time in a particular country is one of the main objectives of the PEFA framework. When done indicator by indicator and accompanied by a nuanced summarizing narrative, the comparison does not raise methodological issues, but may on the other hand not lend itself to easy conveying of simple messages about performance trends to non-specialists. Any aggregation of indicator ratings into simpler rating measures raises issues of numerical scale conversion and indicator weights.

53. Comparison of PFM performance across countries by comparing PEFA assessment scores is valid in principle, but is potentially hazardous. Analysts should explicitly highlight issues that may detract from the validity of such comparisons and justify any assumptions made in this regard (e.g. use of equal weights for indicators).

54. The reports using aggregation techniques - so far encountered - have all assumed that all indicators have equal weight and that the numerical intervals between the alphabetic scores on the ordinal scale are of equal magnitude.

55. Another way of making cross-country comparison is to compare the distribution of scores (numbers of As, Bs, Cs and Ds) across countries at a given point in time and over time. This method provides a potentially more nuanced analysis and avoids conversion of alphabetic scores into numerical scores.

56. However, both methods raise issues that should be taken into account in making cross-country comparisons. Other issues also arise. None of the reports so far encountered have clearly discussed the assumptions made, their rationale and the sensitivity of findings to those assumptions. All reports have assumed that the intervals between the A, B, C and D scores are of equal magnitude, that all indicators have equal weight and - where applicable – that all countries carry equal weight in clustering of countries. These assumptions have the advantage of being simple, transparent and replicable.

57. There is no scientifically correct method on how aggregation should be done for each of those three levels of aggregation. Consequently, the PEFA program neither supports aggregation of results in general, nor any particular aggregation method.

58. The PEFA program recommends that any user - as part of the dissemination of results from comparison - clearly explains the aggregation method applied in each case. It would also be advisable that users undertake sensitivity analysis to highlight the extent to which their findings are robust under alternative aggregation assumptions.

59. The validity of cross country comparisons increases according to:

- *The size of the sample when clustering countries.* The larger the sample, the less important the “like with like” issues become, as these issues are limited in

number. The largest sample arises when core dimension scores/overall scores are being compared for all countries or where one country is being compared with the average of all other countries (provided that the country being compared does not have any country-specific circumstances that complicate comparability). The smallest sample arises when one country is being compared with another.

- *The greater the similarities of countries being compared*: Comparison of PEFA scores for countries with similar characteristics (such as income levels, size or administrative heritage) may be more legitimate than comparison of scores for countries that are very different. Thus, countries in homogeneous regional groups can be compared with more validity than regional groups being compared with other regional groups.

- *For measuring change over time, rather than the absolute performance level:* Comparing changes in scores with a specific country (or for a set of countries) may reduce the risk of invalid comparisons, if country-specific factors do not change significantly over time.

- *The greater the extent that the incidence and distribution of "No Scores" is similar between countries*. High incidence of No Scores, differences between countries in the distribution of No Scores across the indicator set, as well as the likelihoods of the evidence, if had been available, justifying high scores or low scores, complicate valid cross country comparisons.

60. Cross-country comparisons through numerical conversions of PEFA assessment scores is undeniably the simplest way of making comparisons, but, as detailed in this paper, pose many validity issues. It remains to be seen whether the importance of these issues is great enough to warrant the use of the more complicated methods of enabling cross-country comparisons, as outlined in theory in Section 3.

61. Thus, as the practice of aggregation appears to spread among various stakeholders for various purposes, it may be useful to follow-up on this paper's theoretical discussion by investigating the practical impact of different approaches to aggregation. This could be done by conducting a study of the extent to which the considerations in this paper would result in different conclusions in cross-country comparison or comparing progress in a country (or a fixed set of countries) over time, when applied to actual data drawn from the database of PEFA assessment scores.

APPENDIX 1 Bibliography

African Development Fund: Results Reporting for ADF-10 and Results Measurement Framework for ADF-11, Background Paper: February, 2008.

Dorotinsky, William and de Renzio, Paulo: Tracking Progress in the Quality of PFM Systems in HIPCs Countries; PEFA Secretariat, November 2007.

De Renzio, Paulo: Taking Stock: What do PEFA Assessments tell us about PFM systems across countries? University of Oxford, April 2008.

Independent Evaluation Group, World Bank: Public Sector Reforms: What Works and Why? An IEG Evaluation of World Bank Support, June 2008.

Kaiser, Kai and Steinhilper, David: Note on PEFA and Fiscal Flows to Sub-National Governments and Front-Line Providers (draft report); World Bank, 2008

OECD: Methodology for assessment of national procurement systems, Version 4, July 2006.

PEFA Program: Performance Measurement Framework, June 2005

PEFA Assessment Reports: Kenya, Malawi, Mozambique, Tanzania, Zambia; www.pefa.org.

PEFA Secretariat: Common Approach to PEFA Value Added and links to other PFM indicators, August 2007.

PEFA Secretariat: Report on Early Experience of the Application of the Framework, November 2006, www.pefa.org.

PEFA Secretariat: PFM Performance Measurement Framework Monitoring Report, March 2008.

World Bank: Country Policy and Institutional Assessment (CPIA): methodology on www.worldbank.org