PFM Performance Measurement Framework

Monitoring Report 2010

An analysis of repeat assessments including changes in PFM systems performance measured by means of PEFA indicators

PEFA Secretariat

Final report

May 19, 2011

Table of Contents

The core team members of the report included Helena Grandão Ramos (task coordinator), Clay Wescott and Brandon Lundberg. Technical inputs were provided by Frans Ronsholt, Phil Sinnett, Charles Seibert and Tony Bennett. Peer review of the report was provided by Paulo de Renzio (independent), Rachel Perrin (DFID), Monica Rubiolo (SECO) and Steve Knack (World Bank). The study was supervised by Frans Ronsholt.

List of Abbreviations

AfDB	African Development Bank
AsDB	Asian Development Bank
AusAid	Australian Agency for International Development
CA	Comparative Assessment
CFAA	Country Financial Accountability Assessments
CG	Central Government
CI	Compliance Index
CIFA	Country Integrated Fiduciary Assessment
CN	Concept Note
D-1, 2 or 3	Donor Practice Indicators
DFID	UK Department for International Development
DP	Development Partners
EC	European Commission
ERPFM	External Review of Public Financial Management
Framework	Public Financial Management Performance Measurement Framework
FY	Fiscal Year
GBS	General Budget Support
HLG-1	Predictability of transfers from higher level of government
IADB	Inter-American Development Bank
IMF	International Monetary Fund
MR	Monitoring Report
Norad	Norwegian Agency for Development Cooperation
OECD-BIS	Organization for Economic Co-operation and Development – Baseline Indicator Set for procurement
PA	Previous Assessment
PEFA	Public Expenditure and Financial Accountability
PEMFAR	Public Expenditure Management and Financial Accountability Review
PER	Public Expenditure Review
PFM	Public Financial Management
PFM-PR	Public Financial Management – Performance Report
PFM-PR-SN	Public Financial Management – Performance Report – Sub-National
PI	Performance Indicator
RA	Repeat Assessment
SECO	Switzerland's State Secretariat for Economic Affairs
SNG	Sub-National government
TORs	Terms of Reference
WB	World Bank

Executive Summary

1. One of the main reasons for developing the PEFA Framework was to create a monitoring tool that would enable measuring changes in PFM systems performance over time at the country level. This fourth monitoring report on the roll-out of the PEFA framework looks at whether this objective is being met, and investigates trends in global changes in PFM systems based on an initial batch of repeat PEFA assessments.

2. The report covers the assessment reports that were finalized (final report) or substantially completed (full draft report) by October 6, 2010 and which constituted a 2nd or 3rd generation PEFA assessment in a country (referred to as a repeat assessment or RA). Forty five such repeat assessments have been carried out in 38 countries.

3. Thirty three of those assessments (covering 29 countries) are referred to as "comparative assessments" (CA) because they intended to measure changes over time. Twelve RAs were not comparative since they did not intend to measure changes since an earlier assessment.

Frequency of and drivers behind the RAs

4. Five years and four months after the launch of the PEFA Framework, a total of 206 PEFA assessments had been implemented in 119 countries. In about one third of those countries a repeat PEFA assessment had also taken place. In a quarter of the 119 countries, this repeat assessment constituted a comparative assessment.

5. CAs were carried out an average of 33 months after the previous assessment (PA), ranging from one to four years. This included seven countries that carried out CAs less than 30 months after the previous assessment, i.e. shorter than the recommended 3 year minimum interval. The average for those seven CAs was 22 months.

6. For the 26 countries that carried out CAs 30 months or more after the PA, the most important reasons stated were to measure progress, and to contribute to the design or monitoring of a PFM action plan or reform program, followed by facilitating dialog with donors, and links to ongoing or future GBS. In the cases of shorter intervals, the reasons were basically the same but with more emphasis on future donor assistance being contingent on the CAs: hence the urgency of carrying them out after a short interval.

7. The twelve repeat assessments that were not comparative were carried out an average 30 months after the previous assessment. The main purpose was to prepare a baseline, suggesting that there was insufficient confidence in the previous assessment by major stakeholders.

8. It was not possible to discern any patterns as to variation in the purposes or frequency of CAs by country characteristics, or by sponsoring donor.

9. Five countries have not yet carried out CAs 4 ½ years or more after the PA, but in two cases a repeat assessment is either ongoing or planned. Reasons for a long interval included change of government and delays in launching a government reform program as well as a difficult process for the baseline assessment with little if any buy-in by the government.

10. The recommendation is for a repeat PEFA assessment to take place between three and five years after the baseline assessment. Overall, about 80% of comparative assessments were implemented with an interval of more than 2 ½ years. There are only 5 countries where a repeat assessment has not been implemented within 4 1/2 years of the first assessment. This suggests that repeat assessments are implemented quite consistently across countries and largely within the stipulated interval.

Effectiveness in Measuring Performance Changes

11. To be able to use the PEFA indicator scores for measuring change over time, it is important to be reasonably confident that each pair of scores from the PA and CA for each indicator (and indicator dimension) represents the 'real' performance change. Each of the indicator dimensions for each of the 33 CAs were reviewed to evaluate if the PEFA Secretariat found the evidence adequate for the ratings of both the PA and CA and to identify any other scoring issues (such as provision of new evidence for the PA rating, different definitions used, different sampling used, or different interpretation of similar data).

12. 'No score' in the PA and/or the CA also hinders measuring change over time. But since 'no score' for an indicator dimension is easily detectable when comparing pairs of scores, users of the scoring data will know where and when this is a factor.

13. A comparability level across all indicator dimensions of 80% or more - between PA and CA scores – was considered satisfactory when 'no scores' are excluded from the data set. This robust level of comparability was reached for 25 of the 33 CAs i.e. for 76% of the reports. Poor comparability – less than half of the indicator dimension ratings may be compared with confidence – was found in 3 of the 33 CAs (9%).

14. An analysis of process factors that may have contributed to the robustness of measuring changes over time found that the following factors all appeared important for obtaining a high level of comparability between the PA and CA:

- clarity of CN/TORs with respect to measuring performance changes since the previous PEFA Assessment;
- the use of the same assessment team in both assessments, PA and CA;
- the establishment of a specific structure for government's participation in the assessment;
- undertaking initial and final workshops for the major stakeholders;
- submission of draft reports to the PEFA Secretariat for comments;

15. Of the above factors, only the PEFA Secretariat review of draft reports did not correlate with the degree of comparability for the simple reason that all CAs in the analysis had been reviewed by the Secretariat. However, earlier Monitoring Reports have demonstrated that the Secretariat's reviews contribute significantly to improving the compliance rate and therefore the extent to which indicator scores are appropriately evidenced. The robustness of comparing PA and CA scores are thus impacted by the Secretariat's reviews.

16. An analysis of comparability for each of the 74 indicator dimensions across all CAs, shows that the average comparability is 82% of the observations without 'no scores' (70% if 'no scores' are included). Comparability is above 80% for 51 of the 74 dimensions of the Framework. The lowest level of comparability (61%) is reached for PI-7 (i) '*the extent of unreported government expenditure'*, an indicator notoriously difficult to support with hard evidence. Lower levels of comparability and different ranking of dimensions are obtained when 'no scores' are included in the data sets. This analysis can help to identify indicator dimensions in need of additional clarification and guidance to assessors, including emphasis in training courses.

Trends in PFM Performance

17. Of all indicator dimension ratings, 11% maintained "A" scores which cannot be improved, 21% improved the score, 10% maintained "D" scores which cannot decrease, 9% of scores worsened, 21% maintained "B" or "C" scores, and 30%[1] did not lend themselves to valid measurement of change. Patterns of change were analyzed for *all indicators or dimensions where performance could be validly compared* from one assessment to the next. These changes are thought to mainly represent real performance changes, although it is impossible to rule out that in some cases, new ratings may be based on better information or different judgments or interpretation, which could not be detected during the present exercise.

18. The performance patterns were analyzed using a methodology that categorizes indicators or dimensions as *formal* PFM features where progress can be achieved through adopting a new law, regulation, or technical tool, or focusing on no more than a few agencies, or at an early stage in the budget cycle, and *functional* PFM features where progress requires actually implementing a new law or regulation, or coordinating the work of many agencies, or working downstream in the budget cycle i.e. classifying indicator dimensions according to the related PFM characteristics to distinguish between de jure/de facto, upstream/downstream and actor concentration/deconcentration.

19. The analysis found that formal features are more likely to improve or maintain a highest score, while functional features are more likely to worsen or maintain a lowest score. Both formal and functional features have higher proportions of highest and increasing scores, vs. lowest and worsening scores, although differences between the formal features are greater than between functional features. More specifically,

- Indicators representing actor concentration are showing much higher performance improvements than indicators representing actor deconcentration. In the latter case, hardly any global improvement could be identified.
- Indicators representing upstream and downstream elements of the budget cycle are doing almost equally well
- Indicators of de jure elements show moderately higher degrees of improvement than de facto elements.
- Differences between formal vs. functional scores are greater for the lowest or declining scores than for the highest or improving scores.

20. Seven CAs with intervals of less than 30 months after their respective PAs were analyzed to look for distinctive features in measuring performance. Although the

[1] The total adds up to more than 100 per cent due to rounding. The 30% include 'no scores'.

expectation was that intervals less than three years would be too short to show progress, in fact the proportion of increasing scores was actually higher than average in the short-interval cases, and the proportion of incomparable scores was smaller. A possible explanation for this could be that in all cases of short-interval CAs, a key motivating factor for undertaking the CA was that it was a condition for donor support, creating a possible incentive for showing the PFM system in the best possible light. It may also have been an incentive for both government and donors to carry out an early repeat assessment, if the stakeholders were convinced in advance that a repeat assessment would show significant PFM systems improvement.

Recommendations

21. Based on these findings, a number of recommendations are made, several of which have featured in previous monitoring reports. ***The PEFA Secretariat*** should take the following actions:

- Revise the existing documents with respect to the assessment process ("TOR Checklist", "Good Practices in Applying the PFM Performance Measurement Framework", "Repeat Assessment Guidance Note") in order to highlight the necessity to (i) include in the CN/TOR a specific reference to the Secretariat guidance notes, (ii) plan for additional time and resources for analyzing changes, (iii) submit the CN/TOR to the Secretariat for comments and (iv) ensure transfer of detailed information from the PA to the CA assessors, either by ensuring some overlap in assessors or at least by soliciting collaboration from the PA team leader, and by providing the CA assessors with the comments from stakeholders and the PEFA Secretariat on the PA.
- Revise the existing training material in order to highlight the issues raised in the previous point.
- Examine indicator dimensions with high non-comparability to determine whether difficulties in comparison call for clarification of the framework and guidance to assessors, and to determine if changes to the minimum requirements for the dimension score should be considered.

22. ***Lead agencies*** should ensure that a number of good practices are implemented:

- CN/TORs should be as specific as possible with respect to what is expected from the repeat assessment, as called for in the PEFA-Secretariat "Repeat Assessment Guidance Note".
- CN/TORs need to be sufficiently detailed in specifying how the assessors should incorporate comparison with the PA in the CA report and allow time to verify the basis on which earlier scores have been assigned.
- CN/ TOR must include provisions that the assessment team records all relevant information (e.g. on a CD) in a way that can be understood and be easily accessible by other experts during a later repeat assessment.
- CN/TORs should be included as an annex to the draft/final assessment reports
- CN/TORs should be subject to a quality assurance process (peer-review) by the PEFA Secretariat to ensure that the above points are adequately implemented.

- There should be a practice of providing to the CA assessor team all the relevant information and documents from the previous assessments. These include final report and comments from stakeholders, and from the Secretariat.
- Lead agencies should facilitate the contact between the previous and current assessment team leaders, even if this requires the provision of extra time and implies extra costs.
- Specific reasons should be provided to the Secretariat on its comments to CN/TORs and the assessments if the comments are not agreed to.

23. When carrying out a repeat assessment, ***the Assessors*** should take into account the following:
    - Follow the advice of the PEFA Secretariat's "Repeat Assessment Guidance Note".
    - Request the assessment manager (lead agency) for information on the previous assessment (drafts and final reports and comments from the quality review process).
    - Request the lead agency to establish contact with the previous assessment team.

Further Work

24. The above findings and recommendations provide an 'appetizer' for what the expanding database of PEFA repeat assessments can contribute to further learning about PFM performance changes. Multiple suggestions are made throughout the report on how ***further research*** may strengthen the findings or dig into underlying issues in more detail. Whilst the PEFA Secretariat intends to continue some of the work, particularly as it relates to the need for development and guidance on the PEFA Framework and its application, many other opportunities are available for researchers at large, in particular as regards trends in PFM systems performance.

Chapter 1 – Introduction and Methodology

25. This is the fourth monitoring report prepared by the Secretariat. Measuring progress over time is a primary objective of the PEFA framework. So far, most assessments undertaken have been baseline assessments but repeat assessments (RA) are emerging in significant numbers as baseline coverage is nearly complete in some regions and initial assessments are becoming three or more years old.

26. One of the objectives of an RA is to measure performance since the previous assessment. An RA looks at the specific changes in system performance by verifying what has changed and by how much. When used consistently, stakeholders can expect that the repeated application of the framework will provide evidence of the extent to which country PFM performance is changing. In addition, the PFM performance report will recognize the efforts made by a government to reform its PFM system by describing recent and on-going reform measures, although these may not yet have impacted on PFM performance. It is immediately apparent that for two assessments conducted some time apart to be comparable, they must represent consistent applications of the methodology. Assuming that the initial assessment represents a rigorous application of the methodology, stakeholders would ideally want to know that progress has been measured accurately over time.

27. Drawing on the considerable number of repeat assessments now available, the main purpose of the *Monitoring Report 2010 (MR 10)* is to assess if the PEFA framework is able to provide reliable measurement of performance changes over time[2]. The results of the analysis will be of major importance for determining if there is a need for additional technical and process guidance. The analysis will also enable formulation of guidance on comparing PEFA indicator scores from different years, as well as provide an important input to the overall external evaluation of the PEFA program.

28. The report seeks answers to three main questions: (i) What is the frequency of and drivers behind the repeat assessments? (ii) Does the Framework effectively enable measuring changes, and could change be measured with better validity and reliability? (iii) What trends in PFM performance do repeat assessments reveal?

29. For the purpose of this report the team considered that:

(i) A "*previous assessment*" (PA) is an assessment that measures the country PFM systems and process performance at a certain point in time following the PEFA methodology. The assessment may be carried out in order to set a baseline against which to measure progress in PFM performance as measured by indicators or dimensions[3] over future years.

[2] The study did not undertake an analysis of the entire assessment report as it was not its objective; it focused on specific subjects in the Summary Assessment, section 1 and 2 and in the quality of indicators/dimensions comparability in Section 3.

[3] In this report, dimension scores are used except in cases where an indicator has no dimensions, in which case the indicator score is used.

(ii) A "*comparative assessment*" (CA) is a PEFA repeat assessment that mentions the previous assessment, compares the ratings of the two assessments, and makes some attempt to explain the differences and changes in PFM performance.

30. On the basis of these definitions, the report considered 33 repeat assessments (Annex A), referred to as "*comparative assessments*" because they intended to measure changes over time. Twelve repeat assessments were disqualified because they did not meet the criteria of a CA (see Annex B, with a brief explanation for each case).

31. Chapter 2 analyzes the frequency of and drivers behind the repeat assessments. Chapter 3 evaluates the extent to which the framework effectively enables changes to be measured. Chapter 4 summarizes the trends in PFM performance revealed by repeat assessments and Chapter 5 contains recommendations.

32. The methodology used for Chapters 2, 3 and 4 is as follows:

33. **Chapter 2** considered the reasons to carry out a CA stated in the reports and asks the relevant stakeholders that have funded or undertaken PEFA CAs if there were additional reasons. Consideration was given to drafts, final reports, and the Secretariat's follow-up reviews and comments on summary assessment, section 1 and 2. Interviews were carried out by phone and e-mail with many stakeholders that have funded or undertaken PEFA repeat assessments (representatives of lead agencies and task team leaders). About 70 per cent responded. In some cases, stakeholders interviewed expressed different views from those in the written reports, so some judgment was needed to interpret the actual situation. For example, in many cases, it is difficult to judge whether the donors or country concerned had made the initial request for the CA, and what was the main motivation behind the request. The analysis considered the 33 CAs that are finalized or substantially completed. In cases where countries have carried out more than one CA, the analysis focused on the most recent CA, except in the case of Malawi where two CAs were analyzed.

34. For Chapters 3 and 4, the main basis of the analysis of progress in a CA is an assessment of scored indicators/dimensions. There are 16 possible comparative observations that may be made between scored dimensions. They are presented in Annex C. While performance across dimensions may improve, stay the same, or decline, given the four scoring possibilities (A, B, C or D), the combinations of comparative observations provides a more in-depth look at performance between a previous and a repeat assessment.

35. Dimensions that are not scored (known as a "no score") are also used in some analyses included in this report but since comparison is not possible between two assessments when either assessment has received a no score, they are considered separately.[4]

[4] An indicator/dimension may not have been scored for many reasons. The PEFA Secretariat developed a no scoring methodology based on the observation of three major conditions that emerged: "not rated (NR)," when insufficient information is available to score an indicator, "not applicable (NA)," when an indicator/dimension does not apply to the country being assessed, and, "not used (NU)," when there was no intention of scoring the indicator/dimension.

36. In most parts of the report, the comparative observations are expressed as a percentage of the total number of observations, including those that are not scored. The reason for this is that it is expected that over time, the number of not scored observations will decrease as PFM systems and PEFA assessment techniques improve. This is the pattern observed in the sample, where the percentage of non-scored observations and those with no comparability for other reasons went from an average of 29% for the PAs, to 20% for the CAs.

37. **Chapter 3** examines the comparative assessments to see whether indicators and dimensions can be compared with confidence with those in the previous assessment. As already mentioned, dimension scores are used except in cases where an indicator has no dimensions, in which case the indicator score is used. This meant analyzing the evidence provided for each indicator and dimension assessed to ensure that they are broadly measuring the same thing. Consideration was given to previous and comparative assessment reports, the Secretariat's comments/Secretariat follow-up reviews on drafts and final reports for the previous and the comparative assessment. If a follow–up review did not exist, one was carried out. Five "*tracking issues*" that prevent a comparison with confidence were identified: (a) new evidence collected that was not available to the previous assessment, (b) definition changes, (c) different interpretation, (d) different sampling and (e) previous assessment rescored by the repeat assessment.

38. Consideration was given to the CN/TORs[5] and to the information available from the reports and the Secretariat internal documents on process factors that may affect the quality of measuring changes, namely the use of the same assessment team in assessments, the government involvement in the repeat assessment, initial and exit workshops and submission of draft reports to the PEFA Secretariat for comments. Consideration was also given to the CN/TOR that have been prepared and to the corresponding RA since the PEFA Secretariat published the guidance note *"When undertaking RA"* in February 2010.

39. Based on this analysis, the chapter assesses the extent to which the framework facilitates measuring changes in some respects, and the reasons why it doesn't facilitate such measurement in other respects. The analysis considered 33 comparative assessments of which 22 were finalized (final report) and 11 were substantially completed (full draft report).

40. Chapter 4 examines, where indicator and dimension ratings can be compared with confidence with previous ratings, what is changing and by how much as well as which indicators are changing. In addition, it looked at the patterns of changes across different type of indicators and country characteristics and across regions. The analysis used a methodology (Andrews, 2009; Porter et al, 2010) that categorizes indicators or dimensions in three pairs: de jure and de facto, upstream and downstream, and concentrated and deconcentrated[6].

[5] Less than half of final CN/TORs were available at the Secretariat at the beginning of the MR 10 preparation. The missing CN/TORs were obtained directly from the stakeholders, final reports and in one case from the consultant. Two CN/TORs were not prepared.

[6] This methodology has been used for some widely disseminated research projects. While the PEFA Secretariat does not completely agree with this classification for all of the indicators, the methodology is used here without any changes since a further development of this tool would be time consuming and should involve the original developers of the methodology. Further work could be considered to improve the classification of the indicators and investigate the impact this may have on the results of the analysis.

- The first pair contrasts PEFA dimensions where a C or better score could be earned by a new law, or announcing a new practice, even if it is not implemented (a **de jure** reform) with dimensions for which scores require actual implementation or significant engagement (**de facto**).
- The second pair contrasts PEFA dimensions such as strategic budgeting (multi-year forecasting, strategic planning, investment planning, debt planning), annual budget preparation, legislative analysis of the annual budget, and the structure of formal budget documents on the one hand (**upstream**), and resource management (including cash inflow and outflow management, procurement, payroll); internal control, internal audit and monitoring; accounting and reporting; external audit; and legislative analysis of audit reports on the other (**downstream**).
- The third contrasts PEFA aspects under the control of central, regulatory bodies, like the Ministry of Finance (**concentrated**), with those where multiple agencies or sub-national authorities need to be engaged (**deconcentrated**).

41. Change over time is measured mainly by direction, not by the number of steps up or down the ratings ladder. Changes are then aggregated across indicator dimensions and across countries, giving all dimensions equal weight and all countries equal weight. Sensitivity analysis was not undertaken as part of this report to test results for changes in this approach. Further work in this respect will be useful to confirm the robustness of the findings.

Chapter 2 - Frequency of and Drivers behind the RAs

2.1 Overview

42. This section examines:

- What was the frequency of repeat assessments?
- What are the drivers behind CAs, when done very frequently and when done in line with the recommended interval of 3-5 years?
- Why have some countries not undertaken repeat assessments within the recommended interval?
- What other purposes - in addition to measuring performance changes - have been instrumental in implementing CAs?
- Do the purposes vary by country characteristics? By sponsoring donor?
- Are there any country characteristics associated with whether CAs take place in a timely fashion?

2.2 Findings

43. Between June 2005 and October 6, 2010, forty five repeat assessments have been undertaken using the Framework for both the first and subsequent assessments. They cover 41 national and sub-national governments in 38 countries (three sub-national repeat assessments had taken place in Ethiopia along with a national repeat assessment). Of these 45 RAs, 33 assessments met the criteria for being CAs (listed in annex A) whereas 12 did not (listed in Annex B).

44. The CAs were carried out an average of 33 months after the PA, ranging from one to four years[7]. Seven countries carried out CAs less than 30 months after the previous assessment, for an average 22 months[8]. This is a shorter interval than the recommended guideline of three – five years, set because of the presumed time needed for reforms to demonstrate measurable results. One consideration to keep in mind is that the gap between the dates of the PA and CA may not necessarily reflect the gap in the underlying datasets being used. In the Kenya case (gap between missions 30 months) for example, the PA used data for the 2004/2005 fiscal year, while the CA used data for the 2007/2008 fiscal year; thus the three year recommended interval between assessments was met.

45. With this in mind, the two most common reasons stated for the PEFA in the short-interval cases were the need to facilitate dialog with donors, and the need to measure reform progress. For example, in Afghanistan the World Bank, which funds about 30 per cent of the government's non-military budget, needed a CA as an input to its fiduciary framework. In Dominican Republic, one of the stated purposes was to ensure greater uniformity in the dialog with the international community; however, the government's own concern about monitoring progress in resource management was also crucial. In Mozambique, the CA was considered to provide a robust basis for assessing progress

[7] Based on the dates of the main missions for each assessment. In 3 cases where date of main mission for PA is unknown, the report date is used.
[8] Based on the dates of the main missions for each assessment.

over time. In Tanzania, one of the stated purposes was "…to provide a basis for government/donor dialog on future PFM reforms."

46. Other reasons stated were the need for a PEFA to contribute to the design or monitoring of a PFM action plan or reform program, and as a requirement for ongoing or future budget support, or a future PFM project. For example, the Malawi government was persuaded by the EU to postpone producing an action plan on PFM reform pending the results of the first CA; the action plan was then the basis for a significant increase in GBS. Malawi's most recent CA (its second one) provided the basis for a second PFM Action Plan that was again a condition for receiving GBS from the EU. Successive PEFAs were taken increasingly more seriously by government: the first was not considered well, the second was taken more seriously because of the link to GBS, and the third was fully debated within government, and contributed to useful dialog between government and GBS donors. Stakeholders recognized that the one year interval between the first PA and the first CA was too short to see any progress; the government reportedly had much higher expectations than the donors that the first CA would show significant improvement. Finally, in Kosovo significant changes in the institutional and constitutional framework in the country led both the government and donors to demand an updated assessment taking these changes into account.

47. For the 26 CAs carried out 30 months[9] or more (average 37 months) after the PA, the most important reasons stated were to measure progress, and to contribute to the design or monitoring of a PFM action plan or reform program, followed by facilitating dialog with donors, and links to ongoing or future GBS. For example, Dominica carried out a CA because the PA was considered by the government weak and poorly implemented; the CA would give them a sound basis to draft a new PFM action plan. The CA would also be used to determine continued eligibility for budget support. In Uganda, the CA contributed to an annual review of PFM performance, reduced government transaction costs by facilitating dialog with donors around an agreed pool of information, provided a crucial underpinning for GBS, and a reality check as to whether there was progress in PFM in response to support from a large PFM basket fund. In Timor Leste, a CA was carried out as part of a fiscal transparency assessment against the IMF's Code of Fiscal Transparency, a fiscal ROSC, to evaluate performance and see if a reorientation of PFM efforts was needed. In Kyrgyz Republic, a CA checked progress of PFM reforms, and supported the design of a more strategic and programmatic approach to PFM reforms through a PFM medium-term vision. In Tonga, the Finance Ministry was interested in the independent views of people new to the country on what reforms would be sensible, manageable and offer some benefits, and which would not. In Ghana, a new government wanted to get an accurate picture of the PFM situation, and to encourage aid partners to more fully utilize country systems.

48. In addition, there were 12 examples of assessments following earlier assessments that are not considered CAs because they did not acknowledge the previous assessment, and did not analyze changes in PFM performance. While the reason for this is assessment specific, Table 1 indicates that issues of quality may be particularly pronounced for either the PA or the RA. In these 12 cases there were on average a higher number of unscored dimensions in the PAs and a lower review rate by the PEFA Secretariat.[10] These RAs

[9] Based on dates of main missions.
[10] See annex B for the reasons and full list of RAs excluded.

were carried out an average 30 months[11] after the previous assessment. The main purpose was to prepare a baseline, suggesting that there was not sufficient confidence in the previous assessment. For example, the PA for Bolivia had been a self-assessment covering only part of the Framework, so a complete assessment was needed to produce a full baseline. In the case of Central African Republic, the African Development Bank was preparing institutional support to PFM, and needed an up-to-dated PEFA as a baseline. There were questions raised because the PA had not been validated during a workshop. Further, the African Development Bank and World Bank GBS used PEFA indicators as performance indicators, and needed a current assessment for preparing the completion report. In the case of Montserrat, the 2008 PA was considered dated since it was largely based on DFID's 2006 fiduciary risk assessment. Further, the 2008 PA had not been submitted to the PEFA Secretariat for review. For these reasons, DFID was unable to use it as a basis for its 2010 fiduciary risk assessment (required by internal DFID guidelines) so a new CA was required.

Table 1: Differences between CAs and other RAs as concerns coverage and quality assurance

Repeat Assessments	Average number of unscored dimensions		PEFA Secretariat draft assessment reviewed (%)	
	PA	RA	PA	RA
33 CAs included	10	3	100%	100%
12 RAs excluded	27	3	83%	83%

49. An analysis of correlations was carried out to see if the purposes or frequency of CAs varied by country characteristics, or by sponsoring donor, but no patterns could be discerned. Consideration was given to using multivariate regression to look for possible results. However, because of the small sample size, limited range in the dependent variable (interval between assessments), and the subjective nature of possible explanatory variables (e.g. pressure to demonstrate progress on PFM reforms), it was concluded that such analysis was unlikely to come up with significant results at this stage.

50. In total, 23 countries completed a baseline assessment by the end of June 2006 i.e. during the first year after the Framework's launch. Almost five years later, 18 of those countries (78%) have undertaken a RA. Some of those RAs (6) were not CAs for reasons discussed above. The remaining five assessments (ref. Annex D) were examined in order to understand reasons for not undertaking a RA within the recommended interval. Fiji's PA was finalized in June 2005. The lack of a RA can be explained both on the demand side - the government reportedly was unhappy with the PA - and on the supply side - donors have scaled back assistance since the military regime took power in 2006. Another long gap was for Bangladesh, where the PEFA carried out in 2005 had limited ownership by the Government, so a modified PEFA methodology was used. This in turn raised questions about the validity of the exercise, and so the credibility of scores was not accepted by the Government, and the report not published except for a summary table as an annex to the World Bank's Country Assistance Strategy. Since then, PFM diagnostics have been carried out covering some of the same ground as PEFA (e.g. Government of Bangladesh and DFID, 2007; World Bank, 2006, 2007, 2009; DFID, 2009). A CA is now being undertaken for Bangladesh. In the case of Syria, a partial PEFA assessment was included in an IMF technical assistance report. There was no explicit agreement with

[11] Based on cover dates of reports.

the government on using the PEFA methodology, which the government was not familiar with. Lacking government ownership of the assessment was clearly also an issue for the assessment in Republic of Congo (Brazzaville). In the case of Uganda (local government), an RA is planned to take place later in 2011 and lack of government ownership may not have been an issue in the delay since several RAs at national level have been undertaken in the meantime. Overall, a major reason why an RA was not carried out within the recommended five year period appears to be limited government ownership of the initial PEFA assessment. Several other factors may be at play (such as degree of aid dependency and local politics) but the number of cases is too small to justify a more sophisticated analysis.

51. In conclusion, the vast majority of PEFA assessments are followed by RAs within the recommended 3-5 year interval (or slightly shorter interval). Most of the 33 CAs were carried out at least 30 months after the PA. The most important reasons stated for carrying out the CAs were to measure progress, and to contribute to the design or monitoring of a PFM action plan or reform program, followed by facilitating dialog with donors, and links to ongoing or future GBS. There were 7 CAs carried out less than 30 months after the PA. In most cases, these were motivated by an immediate need to measure progress by donors and/or governments. There were no discernable patterns in the purposes and frequency of CAs linked to country characteristics or by sponsoring donor. The 12 countries that carried out repeat assessments - not considered CAs - did not acknowledge the previous assessment and did not analyze changes in PFM performance. Quality concerns regarding the PA may have played a major role in that respect. Finally, there are five cases where no repeat assessments have been carried out 5-6 years after the PA. An important reason in these cases appears to be weak government ownership of the PEFA process.

Chapter 3 – Effectiveness in Measuring Performance Changes

3.1 Overview

52. This section examines:

- What is the quality of measuring changes?
- For each case, which indicator and dimension ratings can be compared with confidence with previous ratings to determine changes, based on evidence?
- What are the reasons for indicator and dimension ratings that cannot be compared with confidence with previous ratings (e.g. quality of evidence, explanation and extent of questioning earlier scores; clarity of TORs with respect to measuring progress over time)?
- What are the "process factors" affecting the quality of measurement of changes?

3.2 Findings

Context

53. For analysis of performance changes over time, this report considered only thirty three repeat assessments referred to as "comparative assessments" (CAs). By October 6, 2010, 22 CAs were finalized (final report) and 11 substantially completed (full draft report). Of the 22 finalized CAs, 10 were publicly available[12]. The 33 CAs covered 29 countries, of which 14 were in Africa (AFR), six in Latin America (LAC), four in East Asia and Pacific (EAP), four in Eastern Europe and Central Asia (ECA) and one in South Asia. No CAs have been undertaken in the Middle East and North Africa countries (MENA) even though three years have passed since the last PEFA assessment was carried out in some countries (e.g. Syria).

54. The main basis of the analysis of progress is an assessment of scored indicators or dimensions[13]. Comparability is being analyzed on the basis of both a 'vertical analysis' (in terms of comparability of the scores of a single CA with its PA across all indicator dimensions) and by 'horizontal analysis' (in terms of comparability for a particular indicator dimension across all CAs).

55. A "no score" in one or both of the assessments makes a comparison of ratings between two assessments impossible. As mention in Chapter 1, indicators or dimensions that are not scored are also used in the analysis of this report but since comparison is not possible between PA and CA they are considered separately. This is an important issue in explaining non-comparability. Tables 2, 3, and 3a illustrate the impact of the factor "no score" both in the vertical analysis and the horizontal analysis. When "no-scores" are included in the data set, the comparability level across the 74 dimensions is 70%. When 'no scores' are excluded, the comparability level increases to 82%. This means that 30% of non-comparability is caused by 'no scores' which are clearly marked in the assessment reports. A similar difference occurs in the comparability across individual comparative assessments.

[12] Available at www.pefa.org via hyperlink: Ghana, Guinea Bissau, Kenya, Kyrgyz, Kosovo, Moldova, Timor Leste, Trinidad & Tobago, Tonga and Uganda. An additional CA was published in January 2011, Dominican Republic.

[13] Only dimension scores are analyzed. Some indicators have only one dimension in which case dimension scores and indicator scores are the same.

Table 2. Comparability across comparative assessments

Scoring Comparability Band	*Band 80-100%*	*Band 50-79%*	*Band ≤ 50%*	*Total CA*
Number of CAs including "no score" which falls within each band	15	11	7	33
Number of CAs excluding "no score" which falls within each band	25	5	3	33

Table 3. Comparison across indicator/dimensions

% of indicator/dimensions that can be compared with confidence when "no-score" is included	70%
% of indicator/dimensions that can be compared with confidence when "no-score" is excluded	82%

Table 3a. Comparison across indicator dimensions (breakdown by band of comparability and number of indicators)

Scoring Comparability Band	*Band 80-100%*	*Band 60-79%*	*Band 50-59%*	*Band ≤ 49%*	*Total indicators & dimensions*
Number of indicators when "no-score" is included*	19	43	7	5	74
Number of indicators when "no-score" is excluded*	51	23	0	0	74

*HGL-1 three dimensions were removed. See explanation in paragraph 53

Comparability across comparative assessments (vertical analysis)[14]

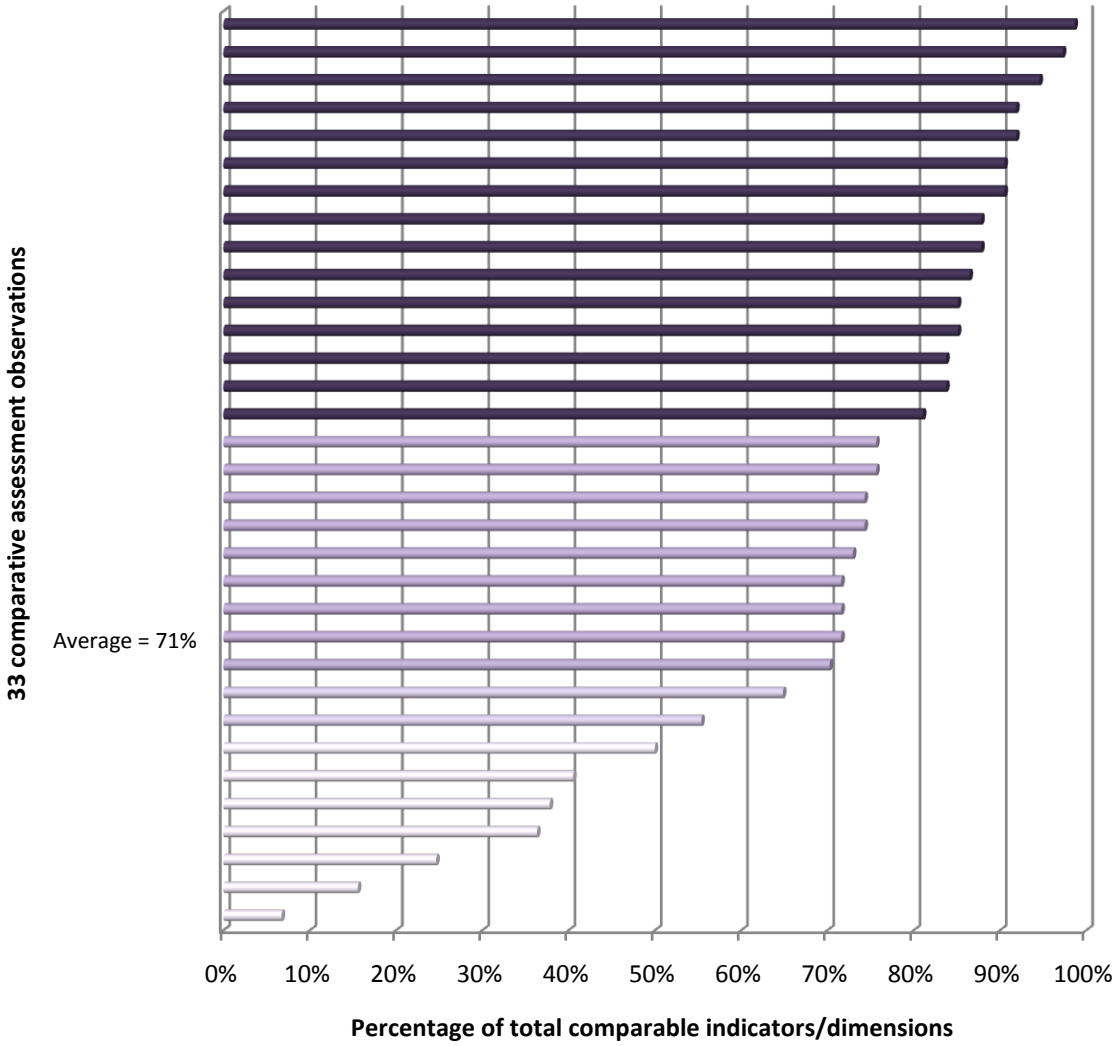
56. Looking across the 33 CAs, data show that in 15 CAs, 80-100 per cent of indicator/dimension ratings could be compared with confidence with the previous ones (see Graph 1)[15]. A further nine CAs show a degree of comparability in the range of 70-80 per cent. Combined, the 24 CAs represent 73 per cent of all the comparative assessments considered in this report .

57. Seven outlier CAs show a low percentage of indicators or dimensions that can be compared with confidence i.e., less than 50 per cent of the ratings could be compared with previous ratings, mainly due to a no score in one or both of the assessments.

Graph 1: Percentage of indicators or dimensions that can be compared with confidence across CAs by assessment ('no scores' included)

[14] Annex E illustrates the percentage of indicators and dimensions that can be compared with confidence across CAs when "no scores" are removed.

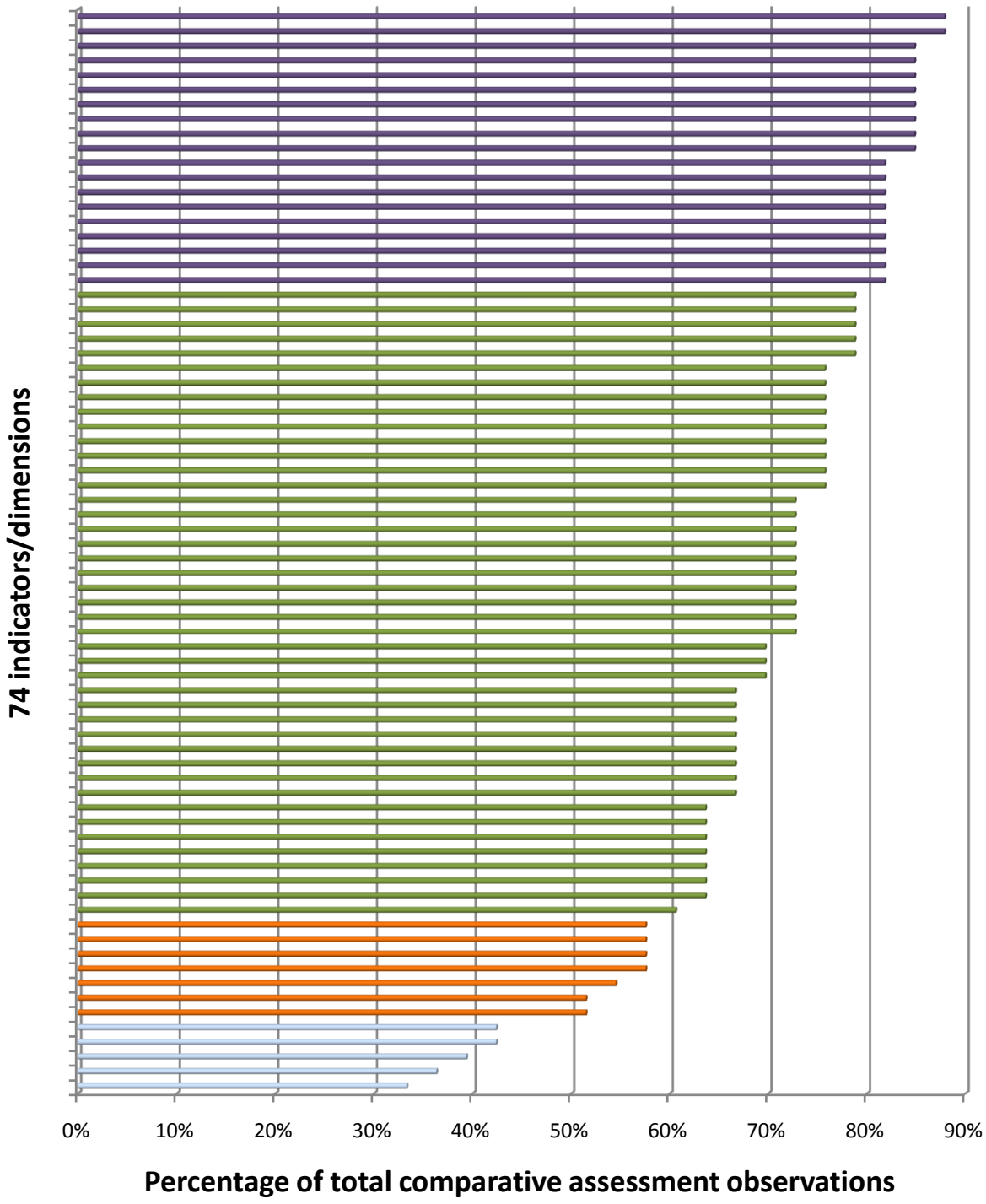
[15] Two CAs that were carried out since after the guidance on repeat assessments was published in February 2010 were among these 15 CAs.



Comparability across indicators or dimensions (horizontal analysis)

58. Looking across all 33 CAs, the data indicate that on average 70 per cent of the indicators or dimensions can be compared with the previous assessment. Around 30 per cent of the indicators or dimensions could not be compared with confidence for lack of evidence, no score in one or both of the assessments, the questioning of earlier scores and other comparability issues. Given that most of the PAs were carried out in the first two years of the Framework application, when assessment quality was significantly lower than it is today, these results are considered satisfactory. Future exercises of this kind would be expected to have higher rates of comparability.

Graph 2: Percentage of scores that can be compared with confidence across CAs for each indicator dimension ('no scores' included)


74 indicators/dimensions
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
Percentage of total comparative assessment observations

Legend[16]**:**

100-80%	PI-10, PI-14 (ii), PI-11 (i), PI-11 (iii), PI-13 (ii), PI-16 (ii), PI-26 (iii), PI-28 (i), PI-28 (ii), PI-28 (iii), PI-6, PI-11 (ii), PI-12 (iii), PI-14 (iii), PI-22 (i), PI-22 (ii), PI-24 (ii), PI-24 (iii), PI-25 (ii)
79-60%	PI-14 (i), PI-16 (i), PI-18 (i), PI-21 (ii), PI-23, PI-3, PI-5, PI-12 (i), PI-15 (ii), PI-21 (i), PI-21 (iii), PI-24 (i), PI-25 (i), PI-26 (i), PI-12 (ii), PI-12 (iv), PI-13 (iii), PI-17 (i), PI-17 (ii), PI-17 (iii), PI-18 (ii), PI-18 (iii), PI-20 (iii), PI-27 (iii), PI-13 (i), PI-25 (iii), PI-27 (v), PI-2, PI-16 (iii), PI-8 (iii), PI-15 (iii), PI-18 (iv), PI-19 (i), PI-19 (iii), PI-20 (i), PI-1, PI-4 (ii), PI-8 (ii), PI-9 (i), PI-26 (ii), PI-27 (i), PI-27 (ii), PI-20 (ii)
59-50%	PI-19 (ii), D-2 (ii), PI-7 (ii), D-3, D-2 (i), PI-8 (i), PI-9 (ii)
≤49%	PI-7 (i), D-1 (i), PI-15 (i), PI-4 (i), D-1 (ii)

Indicators or dimensions with high/low percentage of comparison with confidence

59. Nineteen indicators or dimensions show a percentage of comparability with the previous assessment in the range of 80-100 per cent. On the other hand, five indicators or dimensions show less than 50 per cent of comparability. One indicator, HLG-1 was applied in only three CA (three SNG assessments); this indicator did not exist when the baseline assessment was carried out. As a consequence, comparison would never be possible; this implies a 100% lack of comparability; it was therefore removed from the graph to avoid biasing the conclusion. Graph 2 illustrates the percentages of indicators/dimensions (excluding the HLG-1 dimensions) that can be compared with confidence across the 33 CAs by indicator/dimension.

60. The five indicators or dimensions showing less than 50 per cent of comparability with the previous assessment are:

- PI-4 (i), Stock and monitoring of expenditure payment arrears
- PI-7 (i), Extent of unreported government operations
- PI-15 (i), Effectiveness in collection of tax payments
- D-1 (i) and (ii), Predictability of direct budget support

61. These indicators or dimensions call for quantifiable data as required by the PEFA methodology. This is not the case for the 19 highly comparable indicator/dimensions except for PI-12 (iii) (which requires costing for sector strategies). The problem might be related to data availability (from authorities and development partners) and not necessarily with the ability of measuring performance changes based on the framework. Detailed examination shows that all five indicators or dimensions had considerable levels of 'no score' in one or both the assessments, thus preventing comparison. Another fourteen indicators or dimensions show more than 15 per cent of "no score[17]. Assuming that in the early years it was difficult to measure changes in PFM performance, there is reason to believe that lack of comparability will decrease over time as governments improve their capacity to produce the necessary information and implement reforms. However, further analysis should be carried out in order to find out if difficulties in comparison call for (i) changes in the minimum requirements for the dimension score, (ii)

[16] The indicators are listed in descending order. For example, indicator PI-10 is the closest to 100% while PI-25 (ii) is the closest to 80%.

[17] PI-4 (ii), PI-7 (ii), PI-8 (i, ii, iii), PI-9 (i, ii), PI-12 (ii), PI-15 (iii), PI-17 (i, iii), D-2 (I, ii), D-3

for better "*Clarifications*" to the framework application, or (iii) whether insufficient preparation of PEFA missions may impact on the comparability as data may not be readily available.

62. The graph in Annex F presents the percentage of comparability among the indicators when the factor "no score" is excluded: it shows that three quarters of the total indicators or dimensions fall in the range of 80-100 percent of comparability and one quarter falls in the range of 60-80 per cent; there is no indicator or dimension below 60 percent of comparability. It excludes, as in Graph 2, the indicator sub-national assessment indicator HLG- 1 which was not included in any PAs.

"Tracking issues"

63. Five **"**tracking issues" that prevented comparison were identified: (a) new evidence that was not available (or not used) for the previous assessment, (b) definition changes, (c) different interpretation of the Framework requirements, (d) different basis of sampling, and (e) the comparative assessment rescored the previous assessment to correct palpable errors. It should be emphasized that this represents a very conservative approach to ensuring comparability. For example, when the team found instances of CAs that recalculated earlier scores in order to identify change, the direction of change over time was not identified for lack of comparability. The Guidance on RAs (February 2010) says that indisputable mistakes in PAs should not be re-rated, but should be reported as part of the discussion of changes[18].

64. Looking across the 33 comparative assessments with 2,451 observations, data show that "tracking" issues were identified 175 times, of which (i) 24 per cent related to new evidence collected for previous assessment, (ii) 14 per cent were cases of definition changes, (iii) 45 per cent concerned a different interpretation, (iv) six per cent related to different sampling, and (v) 11 per cent were cases where the comparative assessment rescored the previous assessment. Some "tracking issues" are potentially more avoidable than others (different interpretation, different sampling and rescoring of PA as opposed to new evidence collected for PA and definition changes). The more avoidable ones account for 62 percent of all cases; however, while "different interpretation" is an assessor specific issue, the other two may be considered more as process issues. Follow-up work could be done in order to identify the dimensions more prone to tracking issues and the reasons behind the incidence of such issues and to recommend the actions to be taken. Box 1 illustrates some examples of "*tracking issues".*

Possible process factors affecting the quality of measuring changes

65. To examine the possible process factors affecting the quality of measuring changes, consideration was given to the CN/TORs, the use of the same assessment team in both previous and comparative assessments, the level of government involvement in the assessment, whether initial and final workshops were organized, and the quality control review process.

[18] Note that re-rating will be required for three indicators for which the rating scales have recently been revised. For PI-2 (i), PI-3 and PI-19 (iii), assessment teams will be required to recalculate the earlier score using the new criteria in order to identify the direction of change over time.

Box 1. Examples of "tracking issues"

New evidence collected for previous assessment: the current assessment obtained information that was not available during the previous assessment.
Definition changes: e.g. arrears, domestic arrears, extra-budgetary funds, classification of parastatals, SNG entity versus de-concentrated central government entity.
Different interpretation: e.g. PI-24 (ii): The CA found that during the last year budget execution reports were published 4-7 weeks after end of quarter, with average 5.5 weeks, and gave a 'B' rating. The PA found that performance was exactly the same but gave a 'C' because two reports were more than 6 weeks delayed
Different Sampling*:* e.g. PI-21: The PA was based on information on the state of affairs in the Ministries of Agriculture, Justice and Finance. The CA was based on information from Ministries of Education and Justice. Is direct comparison valid?
Comparative Assessment rescored the previous assessment: e.g (i) wrong assignment of score despite very clear evidence; (ii) assignment of a '+' to a single dimension indicator.

66. *Clarity of Concept Notes/Terms of Reference with respect to measuring changes.* Thirty one comparative assessments CN/TORs were analyzed; Ethiopia Federal and three Ethiopia sub-national government assessments were covered under the same CN/TOR but for statistical purposes counted as 4 CN/TORs. Two comparative assessments did not prepare CN/TORs.

67. Requesting comments from the PEFA Secretariat on the draft CN/TORs is far from being current practice. At the beginning of this Monitoring Report preparation, less than half of the CN/TORs for the assessments considered in this report were available at the Secretariat. They were subsequently obtained from the stakeholders, final reports and consultants. Usually, the Secretariat does not receive the final version after having issued its comments; sometimes CN/TORs are transmitted to the Secretariat but too late to provide comments.

68. The analysis revealed that 24 CN/TORs (73 per cent) required measurement of performance changes since the previous PEFA assessment. The way this requirement is expressed is mainly divided around two categories: (1) a *simple reference* to "measure PFM progress since the last assessment" and (2) a *clear description* of what is expected from the CA.

69. For the 15 CA with 80-100 per cent of indicators or dimensions compared with confidence, 14 CN/TOR required measuring changes since the previous assessment and one was silent. For the seven CA with low percentage of indicators or dimensions comparability (less than 50 per cent), four CN/TOR requested measuring performance changes (three SNG assessments were covered under a CG assessment), two were silent and one was not prepared[19].

70. This confirms the importance of "*ensuring that CN/TOR be sufficiently detailed and clearly understood by all stakeholders*" as recommended by the PEFA Secretariat in the *Good Practices when Undertaking Repeat Assessments* note.

[19] Of the twelve RA not included in this study, in only one case did the CN/TORs require measurement of performance changes over time (Note: three CN/TORs are not available of which one is a self-assessment).

71. Anecdotal evidence indicates that Repeat Assessment CN/TORs should include provision of time to verify the basis on which earlier scores have been assigned. While this may be seen as an implicit assignment when undertaking a RA, the lack of specific provision of time for this task may contribute to prevent or limit a deep analysis of the previous assessment.

72. *The use of the same assessment team in both assessments*. The assessment team composition for both previous and comparative assessments shows that in nine cases the team leader was the same or that some team members have been involved in both assessments (in two cases the previous assessment team leader was a team member in the comparative assessment). The analysis shows that these nine cases fit within the fifteen CAs with high percentages of indicators or dimensions compared with confidence. None of the seven CAs with low percentages of comparison of indicators or dimensions had overlapping assessment teams[20]. The use of the same assessors was perceived in three countries as "undoubtedly an advantage", "the assessors had built confidence of the Government and DPs" and the "Government was confident in the same consultants as before" (in "*Assessing the Impact of the PEFA Framework*" draft Nov 2010,). This highlights an advantageous element in the process of measuring progress over time, but it is often not possible to use the same assessment team for the CA. On the other hand, while the use of the same assessment team may contribute to fewer spurious changes produced by differing judgments or interpretations (or even differing biases) it may be more difficult for the same team to accept that the previous assessment was poorly done in general[21]. In other words, the advantage of using the same assessment team only applies if the previous assessment is considered of good quality.

73. Where it is not possible to ensure that some of the assessors participate in the repeat assessment, an alternative is to establish contact with the team leader of the previous assessment, so that the team for the repeat assessment may obtain detailed information (e.g. from assessors' notes) on the basis for the ratings in the previous assessment and discuss the validity of performance changes identified. The Secretariat has been involved in a few cases of this nature, which helped the current assessors understand the earlier performance levels and avoid conclusions on rating changes based on rough assumptions. This is particularly important where assessors for repeat assessments receive information that questions the validity of the ratings in the earlier assessment.

74. *Government involvement in the CA*. Government involvement in the preparation of the comparative assessment may be categorized into two types: (i) "*cooperation and support to the assessment team*", a more passive role, and (ii) "*driving the process and deep involvement in the process*", a more active role. Twelve countries established a PEFA steering committee or a similar structure (the format varied from country to country), to follow up, supervise and review the assessment: this is considered a very good practice. Another factor that may have influenced the CA quality is the organization of initial and closing workshops: while the first contributes to increase officials' awareness about the PEFA methodology and their engagement in the PEFA exercise, the

[20] None of the twelve RAs disqualified for this study had had the same team leader/team assessors in both assessments.

[21] Though experience shows that assessors are ready to accept that there may have been a few deficiencies in their previous assessment and therefore need to reconsider ratings.

closing workshop provides a forum for discussing the findings and receiving comments. Sixteen CAs organized both initial and final workshops[22].

75. Of the 15 CAs with 80-100 percent of indicators or dimensions compared with confidence, five established a specific structure and five organized both initial and exit workshops; except in one case, these countries do not overlap. In contrast, none of the seven CAs with less than 50 percent indicators or dimensions compared with confidence set up a specific structure to follow up the process[23] and only one organized both workshops.

76. *Quality assurance review*. All repeat assessment draft reports (including the twelve not included in this study) were submitted to the PEFA Secretariat for comments with the exception of two self-assessments - Uganda 2008 and Dominican Republic 2009 (which were not comparative assessments). The quality review by the PEFA Secretariat might be considered an important element in providing the stakeholders with assurance as to the quality of the assessment.

77. **To summarize**, the above process factors may have contributed to the robustness of measuring changes over time, as indicated in Table 4 below; the incidence of good features in the assessment process appears to have contributed to increase the level of comparability; the same does not seem to apply when the assessment process show limited or no 'good process' factors. Table 4 illustrates the observations of process factors across the 33 CAs and those with higher and lower degrees of comparability; percentages are calculated against total number of CAs, total number of CAs comparable at 80-100% and total number of CAs with less than 50% comparability. As already mentioned information with respect to government involvement and workshops is not always provided in the reports and is often vague. More likely, their incidence is higher than the information currently available. In time, when more repeat assessments are available, it would be useful to undertake follow-up work to get a sense of which of these factors matters more for comparability.

78. Of the process factors considered, only the PEFA Secretariat review of draft reports did not correlate with the degree of comparability for the simple reason that all CAs in the analysis had been reviewed by the Secretariat. However, earlier Monitoring Reports have demonstrated that the Secretariat's reviews (as well as reviews by other stakeholders) contribute significantly to improving the compliance rate and therefore the extent to which indicator scores are appropriately evidenced. The robustness of comparing PA and CA scores are thus impacted by the Secretariat's reviews.

79. As recognized in the PEFA Secretariat "Good Practices in Applying the PFM Performance Measurement Framework*"* the assessment process "*has to be well managed in order to help ensure high quality product … the more the Government is involved the more Government's staff will benefit from the exercise … critical to the ability of governments to exert a strong leadership role is their understanding of the Framework*

[22] Information regarding extent of government involvement and workshops is not systematically provided in all the reports; it is sometimes not sufficiently specific to comprehend the degree in which the authorities got involved in the process. Therefore, this number is likely to be higher (e.g. information available shows that 20 initial and 19 final workshops were organized).

[23] For the twelve RAs not included in this study, data show that in two cases the authorities have set up a structure to manage the PEFA evaluation.

methodology and the process of carrying out a PEFA Assessment". Equally important is the fact that unless there is a well designed and implemented process, this may result in successive PEFA assessments, updates and revisions which make monitoring performance changes over time a difficult task and increase the chances of overlooking performance changes that have occurred in between (e.g. in Tanzania, after several updates and revisions carried out since the 2005 PEFA assessment, the 2010 assessment is the first one that explicitly attempts to measure progress).

Table 4: Incidence of process factors in the comparative assessments

Process factors	All CAs	%	CAs comparable 80-100%	%	CAs comparable ≤ 50%	%
Total CA	33	100%	15	100%	7	100%
Clarity of CN/TOR	24	73%	14	93%	4*	57%
Use the same assessment team	9	27%	9	60%	0	0%
Government involvement (specific structure)	12	36%	5	33%	0	0%
Initial & exit workshop	16	48%	5	33%	1	14%
Quality assurance review	33	100%	15	100%	7	100%

*Includes 3 SNG assessments undertaken alongside with a Federal assessment

80. In conclusion, looking across 33 CAs it is observed that comparability is affected by the proportion of no score indicators or dimensions in one or both assessments, which is higher for cases for which there was inadequate evidence. The degree of comparability increases when the no-score factor is excluded from the analysis. This is valid for both vertical and horizontal analysis. "Tracking issues" that precluded comparison were mainly related to different interpretations and new evidence collected for PAs. In addition, some process factors may have influenced the quality of measuring changes over time. The presence of these factors appears to have contributed to increase the level of comparability; however, the same is not observed when its presence is limited, as shown when cross examination is done between CA with high and low degree of comparability.

Chapter 4 – Trends in PFM Performance

4.1 Overview

81. This section examines:

- Which indicators are changing?
- Are there patterns of change across different types of indicator and country characteristics? Across different regions?
- What are possible explanatory factors?

4.2 Findings

82. Patterns of change were analyzed for all indicators or dimensions where performance can be validly compared from one assessment to the next. Thus, these changes are thought to represent real performance changes, and not simply new ratings based on better information or different judgments or interpretation. Of all indicators or dimensions ratings, 11% maintained "A" scores, 21% improved, 10% maintained "D" scores, 9% worsened, 21% maintained "B" or "C" scores, and 30%[24] did not lend themselves to valid measurement of change. The following indicators or dimensions had the best and worst performance:

- PI-24 (ii), *Timeliness of the issue of in-year budget reports*, had the highest percentage (48 per cent) of "A" scores in two successive assessments.
- The next highest percentages were for PI-3, *Aggregate revenue out-turn compared to original approved budget* (45 per cent) and PI-15 (ii), *Effectiveness of transfer of tax collections to the Treasury by the revenue administration* (36 per cent).
- Two dimensions: PI-8 (iii) *Extent to which consolidated fiscal data (at least on revenue and expenditure) is collected and reported for general government according to sectoral categories*, and PI-28 (i), *Timeliness of examination of audit reports by the legislature (for reports received within the last three years)* had the highest percentage (39 per cent) of "D" scores in two successive assessments.
- The next highest percentage of "D" scores was for PI-23 *Availability of information on resources received by service delivery units* (33 per cent).
- PI-13 (ii), *Taxpayer access to information on tax liabilities and administrative procedures* had the highest percentage of cases with increased scores in two successive assessments (48 per cent).
- Three indicators or dimensions had the next highest percentage of cases with increased scores in two successive assessments: PI-14 (iii), *Planning and monitoring of tax audit and fraud investigation programs* (42 per cent), PI-12 (ii), *Scope and frequency of debt sustainability analysis* and PI-14 (ii), *Effectiveness of penalties for non-compliance with registration and declaration obligations* (38 per cent).
- Four indicators or dimensions had the highest percentage of cases with decreased scores in two successive assessments: PI-23, *Availability of information on*

[24] The total adds up to more than 100 per cent due to rounding.

resources received by service delivery units (21 per cent), PI-1, *Aggregate expenditure out-turn compared to original approved budget*, PI-3, *Aggregate revenue out-turn compared to original approved budget*, and PI-17 (ii), *Extent of consolidation of the government's cash balances* (18 per cent).

83. More countries had a higher number of highest or improved scores (23 countries) than lowest or worsened scores (8 countries),[25] indicating a broad and welcome trend of PFM improvement across the countries surveyed. Of the latter, five were small island states with challenges not typical of the larger sample, including high vulnerability to external shocks such as hurricanes and volatile tourism earnings, severe capacity constraints with high emigration of skilled human resources, and political instability. The other three were small African states with many of the same challenges.

84. The overall performance patterns can be further analyzed using a methodology (Andrews, 2009; Porter et al, 2010) that categorizes indicators or dimensions in three pairs (the analysis "*only considers indicators/dimensions PI-5 to PI-28 as indicators PI-1 to PI-4 cover PFM system outcomes and performance and not the quality of PFM systems per se"*)[26].

85. The first pair contrasts PEFA dimensions where a C or better score can be earned by a new law, or announcing a new practice, even if it is not implemented (**de jure**) with dimensions that require actual implementation or significant engagement (**de facto**) [27]. For example, in the case of de jure dimension PI-11 (i), a C score is attained as long as an annual budget calendar exists, even though there may be substantial delays in implementation, with not enough time allowed to budget entities to complete detailed estimates. On the other hand, de facto dimension PI-12 (i) requires that two year forecasts of fiscal aggregates are actually produced on a rolling annual basis.

86. The second pair contrasts PEFA dimensions relating to budget preparation such as strategic budgeting (multi-year forecasting, strategic planning, investment planning, debt planning); annual budget preparation; legislative analysis of the annual budget; and the structure of formal budget documents on the one hand (**upstream**), and dimensions relating to budget execution such as resource management (including cash inflow and outflow management, procurement, payroll); internal control, internal audit and monitoring; accounting and reporting; external audit; and legislative analysis of audit reports on the other (**downstream**). The former deals with the earlier stage of the budget planning cycle, visible to donors and investors, and would be expected to come up against less resistance than the latter aspects, which deal with controls and oversight of actual spending. For example, in the case of upstream indicator PI-5, a C score is attained as long as the budget is classified using GFS or comparable standards: a formal practice that can be monitored by donors and investors. On the other hand, downstream elements deal with more sensitive issues of managing and monitoring actual expenditures. For example, in the case of downstream indicator PI-7, a C score requires that unreported extra-budgetary expenditure be no more than 5-10 percent of total expenditure. This reduces the opportunity for non-transparent slush funds that in many countries are much

[25] In two countries, the number of highest or improved scores was equal to the number of lowest or worsened scores. Highest or improved scores are combined because it's not possible to improve on an A score; lowest or worsened scores are combined because it's not possible to be worse than a D score.
[26] The full list for all coded indicators/dimensions is in Annex H.
[27] Indicators/dimensions were coded independently by three PFM specialists, with any disagreements discussed and reconciled (Andrews, 2009). For a full coding list, see Annex H.

greater than this amount, and may be strongly resisted by well-connected interests benefiting from such arrangements.

87. The third contrasts PEFA aspects under the control of central, regulatory bodies, like the Ministry of Finance (**concentrated**), with those where multiple agencies or subnational authorities need to be engaged (**deconcentrated**). For example, in the case of concentrated dimension PI-12 (ii) a C score is attained as long as a debt sustainability analysis at least for external debt, a technical job that can be done by a small team of technical staff, has been undertaken during the last three years. On the other hand, the deconcentrated dimension PI-12 (iii) requires costed sector strategies for several major sectors, requiring participation of several budget entities.

88. While there is overlap among the three pairs, they broadly illustrate form vs. function. As a general rule, de jure, upstream and concentrated aspects comprise the formal features, while de facto, downstream and deconcentrated comprise the functional features of a system. Static analysis by independent researchers of PEFA scores indicates that C or better scores for the formal features are more commonly achieved than such scores for functional features. The presumed explanation is that formal progress can be achieved through adopting a new law, regulation, or technical tool, or focusing on no more than a few agencies, or at an early stage in the budget cycle; functional progress is more difficult to achieve because it is more difficult to coordinate the work of many agencies, and because reforms actually implemented are more difficult to monitor, and may threaten rents and face greater political and bureaucratic opposition[28].

89. In the same vein, one would expect that formal features would be more likely to maintain top scores or improve, and functional features to maintain low scores or decline. That is in fact broadly the pattern observed here. The formal features are more likely to improve or be at the highest level, while the functional features are more likely to decrease or be at the lowest level. There are many improving and highest scores in functional areas, showing that even in the most difficult areas, high or improving scores are possible. And yet, the high number of lowest or most declining scores in functional areas suggests that advances in these areas could be fragile, with a risk of setback.

90. This methodology can be used to help explain the scoring performance indicated above. The three indicators/dimensions that maintained D scores in the highest percentage of cases were all examples of all three functional features (de facto, deconcentrated and downstream), where progress would be expected to be relatively difficult. In the case of the 2 indicators/dimensions indicating the quality of PFM systems that worsened in the highest percentage of cases, one (PI-23) was also an example of the three functional features, while the other (PI-17 [ii]) was a case where the challenges of working downstream in the budget cycle overwhelmed the advantages of being concentrated and de jure. The two dimensions that maintained A scores in the highest percentage of cases, and four that improved in the highest percentage of cases, were all examples under the centralized control of implementing bodies, where one would expect early stage success, an advantage sufficient to offset the more challenging de facto and downstream elements of some of these features.

[28] The designations "formal" and "functional" refer to the overall tendency of an indicator or dimension, some of which combine both formal and functional features. The designations do not constitute recommendations as to the sequencing of reforms, as functional reforms may need to be started early.

91. Similar patterns can be discerned in other aspects of performance. For example, PI-25ii, *Timeliness of submission of the financial statements,* had the highest percentage of cases improving from D to A (15%), and is under the centralized control of an implementing body, as is the case with PI-13 (ii), *Taxpayer access to information on tax liabilities and administrative procedures*, that had the highest percentage of cases moving from C to A (12%), and with PI-14 (iii), *Planning and monitoring of tax audit and fraud investigation programs*, that had the highest percentage of cases moving from C to B (24%). These well performing dimensions also had some degree of correspondence with de jure and upstream features, although not as consistently. On the other hand, there was one category of good performance, the highest percentage of cases moving from D to C, where one of the two best performing dimensions was an example of three functional features (de facto, deconcentrated and downstream), PI-26 (i), *Timeliness of submission of audit reports to legislature* (18%). The other best performing dimension, PI-12 (iii), *Existence of sector strategies with multi-year costing of recurrent and investment expenditure* (both 18%), was an example of two functional features (de facto and deconcentrated). These examples show that determined governments can make headway on more challenging reforms, perhaps when there is a political opportunity or heightened support from development partners.

92. Table 5 shows the summary number of highest or improving scores, by dimension type. The two are combined because it's not possible to improve on an A score. The formal features on the left do slightly better than the functional features on the right, as it would be expected that aspects under the control of concentrated entities would move faster, that laws and formal rules would need to be in place before functional improvements could take place[29]. However, the differences in percentages are smaller than might be expected, with minimal difference between upstream and downstream. This indicates the complexity of each case where context, donor interventions, patronage networks, and a host of other factors interact to affect performance. There are many highest or improving scores in the three functional areas on the right, showing that even in the more difficult areas, high or improving scores are possible.

Table 5: Highest or improving scores, by indicator/dimension type

Dimension type	% of scores	Dimension type	% of scores
Concentrated	41	De-concentrated	26
De jure	35	De facto	31
Upstream	33	Downstream	32

93. Table 6 looks at lowest or declining scores. Again, the two are combined because it isn't possible to decline beyond a D score, and a similar pattern is evident if one looks at declining scores only. The data show that the functional features on the right are more likely to have lowest or most declining scores than the formal features on the left. This may indicate that advances in functional areas could be fragile, with a risk of setback.

94. A comparison of table 5 and table 6 also indicates there is a much greater proportion of formal feature scores that are highest or increasing than of scores that are lowest or decreasing, as would be expected. The differences between formal and functional scores are greater for the lowest or declining scores than for the highest or improving scores. That is, functional versus formal seems to matter more at the low end

[29] A similar pattern is evident if one restricts the analysis to improving scores only.

of the spectrum, than at the high end. Yet it is surprising that among the functional features, there is a slightly higher proportion of increasing or higher scores than of lowest or decreasing scores, indicating that reforms are attaining results even in the more difficult areas. A possible explanation is that it may be easier to improve things which are working poorly, than things which are already working well (as many of the formal features are). A difference of means test indicates that these results are significant at the 95% level[30].

Table 6: Lowest or declining scores, by indicator/dimension type

Dimension type	% of scores	Dimension type	% of scores
Concentrated	9	De-concentrated	24
De jure	11	De facto	23
Upstream	15	Downstream	19

95. Looking at the lowest and most declining scores gives further reason for concern. Of the four indicators or dimensions that maintained "D" scores or had the greatest decreases, and that can be characterized as processes, all were de facto and deconcentrated, and three out of four were downstream. This further emphasizes the difficulties both of getting out of the starting gate, and of maintaining progress achieved earlier, in these challenging functional areas of PFM.

96. The team also looked at changes in scores of the seven short-interval cases, with results in table 7. Although the expectation was that intervals less than three to five years would be too short to show progress, in fact the number of improving scores was actually higher in the short-interval cases and the number of incomparable scores also smaller. A possible explanation for the surprisingly high percentage of highest and increasing scores could be that in all cases of short-interval CAs, a key motivating factor for the CA was that it was a condition for donor support, creating a possible incentive for showing the PFM system in the best possible light. It may also have been an incentive for both government and donors to carry out an early repeat assessment, if the stakeholders were convinced in advance that a repeat assessment would show significant PFM systems improvement.

97. Kosovo is an example of a short interval country that made rapid progress, maintaining an A score or improving in 47 percent of the indicators/dimensions. According to its CA report, the main reasons included rapid progress in improving the budget execution system through a financial management information system, chart of accounts, and single treasury account. Internal audit and control and external audit also benefited from recent improvements. Mozambique is another such case. According to its CA report, most improvements were driven by ongoing reforms to revenue collection and management, to procurement, and to financial management linked to the ongoing implementation of Mozambique's financial management information system, e-SISTAFE. Some of the improvements resulted from a review of results from the PA, and many were at advanced stages of design when the PA was being carried out. Although

[30] It would be useful at a later stage to add the 12 non-comparative RAs to the dataset, considering the non-comparability issue as merely increasing the random measurement error. This would help determine if the same patterns are evident in the larger sample.

these and other short-interval cases are not typical of the larger sample, they show that rapid progress is possible for certain types of improvements in favorable contexts.

Table 7: Results of short-interval CAs compared to overall averages

	Short interval country average percentage	**Overall average percentage**
"A" scores maintained	8%	11%
Increasing scores	31%	20%
"D" scores maintained	8%	10%
Decreasing scores	8%	8%
Maintained "B" or "C" scores	22%	21%
Incomparable scores	24%	30%

98. Finally, the team looked at possible correlations between measuring performance and country characteristics. A previous study (de Renzio 2009) used multivariate regression analysis to compare static PEFA scores with country characteristics. It found that PFM scores are positively associated with GDP per capita, recent economic growth, and democracy, and negatively associated with natural resource and aid dependency, all confirming expectations.

99. Shifting to measuring performance, one would also expect negative associations between maintaining maximum scores and improving scores on the one hand, and resource and aid dependency on the other. Likewise, one would expect positive associations with trade, GDP per capita, GDP growth and democracy. An analysis of correlations found no significant findings. As already discussed in Chapter 2, it was concluded not to pursue multivariate regression at this stage because of the small sample size, limited range in the dependent variable, and the subjective nature of possible explanatory variables. However, it would be useful at a later stage to carry out a test controlling for the possibility that developed countries don't show more improvement in the CA because their scores were already high in the PA by including the PA scores as an explanatory variable in the regression. One could also analyze factors that differentiate the 32 countries with CAs from the much larger sample of countries without CAS. Explanatory variables could include aid (or more narrowly budget support) volumes, share of aid coming from largest donor, share of aid coming from PEFA partner agencies, country size, democracy and CPIA indicators, etc.

100. A possible reason for the lack of significant findings is that the country characteristics may have two contrary influences: on the one hand richer, more capable, democratic, fast growing, countries open to trade would be expected to make progress on improving PFM, but on the other hand, they may have already achieved for these reasons a high level in the previous assessment, so further progress could be difficult. Likewise, natural resource and aid dependent countries would be expected to have less incentive for improving PFM improvement, but on the other hand, they may have started from such a low level for these reasons in the previous assessment that they were able to show progress.

101. In conclusion, formal PFM features where progress can be achieved through adopting a new law, regulation, or technical tool, or focusing on no more than a few agencies, or at an early stage in the budget cycle are more likely to improve or maintain a highest score than functional PFM features where progress requires actually implementing a new law or regulation, or coordinating the work of many agencies, or working downstream in the budget cycle. The difference is most pronounced for PFM features where progress can be achieved working with one or a few agencies, in comparison with PFM features where many agencies are involved. Likewise, functional features are more likely to worsen or maintain a lowest score than formal features. Both formal and functional features have higher proportions of highest and increasing scores, vs. lowest and worsening scores, although differences between the formal features are greater than between functional features. These results of dynamic patterns of PEFA scores are broadly in line with static results. However, some aspects of the results, such as the higher proportion of increasing or highest scores than for worsening or lowest scores for functional features, are encouraging, showing that even in the more difficult areas, progress is possible.

102. Further analysis was done on short interval cases, showing that rapid progress is possible for certain types of improvements in favorable contexts. Finally, it was not possible to find any statistically significant patterns of change across different types of country characteristics.

103. The reasons behind the trends identified here – i.e. whether trends in PFM were mainly due to the temporary capacity provided by foreign experts, or to changes in the macro environment, or to genuine government reforms – are beyond the scope of this report, but may be studied in connection with the ongoing multi-donor evaluation study of support to PFM reform. The rapidly increasing amount of data from PEFA repeat assessment would be useful for that exercise.

Chapter 5 - Recommendations

104. Based on these findings, a number of recommendations are made, several of which have featured in previous monitoring reports. ***The PEFA Secretariat*** should take the following actions:

- Revise the existing documents with respect to the assessment process ("TOR Checklist", "Good Practices in Applying the PFM Performance Measurement Framework", "Repeat Assessment Guidance Note") in order to highlight the necessity to (i) include in the CN/TOR a specific reference to the Secretariat guidance notes, (ii) plan for additional time and resources for analyzing changes, (iii) submit the CN/TOR to the Secretariat for comments, and (iv) ensure transfer of detailed information from the PA to the CA assessors by including some overlap in team members (or at least soliciting collaboration from the PA team leader) and providing comments from stakeholders and the PEFA Secretariat on the PA.
- Revise the existing training material in order to highlight the issues raised in the previous point.
- Examine indicator dimensions with high non-comparability to see whether difficulties in comparison call for clarification of the framework and guidance to assessors, and to determine if changes in the minimum requirements for the dimension score should be considered.

105. ***Lead agencies*** should ensure that a number of good practices are implemented:

- CN/TORs should be as specific as possible with respect to what is expected from the repeat assessment, as called for in the PEFA-Secretariat "Repeat Assessment Guidance Note".
- CN/TORs need to be sufficiently detailed in specifying how the assessors should incorporate comparison with the PA in the CA report and allow time to verify the basis on which earlier scores have been assigned.
- CN/ TORs must include provisions that the assessment team records all relevant information (e.g. on a CD) in a way that can be understood and be easily accessible by other experts.
- CN/TORs should be included as an annex to the draft/final assessment reports
- CN/TORs should be subject to a quality assurance process (peer-review) by the PEFA Secretariat to ensure that the above points are adequately implemented.
- There should be a practice of providing to the CA assessor team all the relevant information and documents from the previous assessments. These include final report and comments from stakeholders, and from the Secretariat.
- Lead agencies should facilitate the contact between the previous and current assessment team leaders, even if this requires the provision of extra time and implies extra costs.

- Specific reasons should be provided to the Secretariat on its comments to CN/TORs and the assessments if the comments are not agreed to.

106. ***When carrying out a repeat assessment, the Assessors should take into*** account the following:

- Follow the advice of the PEFA Secretariat's "Repeat Assessment Guidance Note".
- Request the assessment manager (lead agency) for information on the previous assessment (drafts and final reports and comments from the quality review process).
- Request the lead agency to establish contact with the previous assessment team.

Annex A: Comparative PEFA Assessments

Region	Country	1st Assessment		1st Repeat Assessment		2nd Repeat Assessment	
		Lead Agency	Date of Report	Lead Agency	Date of Report	Lead Agency	Date of Report
AFR	Burkina Faso	EC	Apr. 07	Govt	Jun. 10		
AFR	Ethiopia	EC	Oct. 07	EC	Oct. 10		
AFR	Ethiopia-Benishangul Region	EC	Oct. 07	EC	Jul. 10		
AFR	Ethiopia-Harari Region	EC	Oct. 07	EC	Jul. 10		
AFR	Ethiopia-Oromiya Region	EC	Oct. 07	EC	Jul. 10		
AFR	Ghana	WB	Jun. 06	EC	EC		
AFR	Guinea Bissau	WB	Jun. 06	EC	May 09		
AFR	Kenya	DFID	Jul. 06	EC	Mar. 09		
AFR	Lesotho	WB	Jun. 07	DFID	Jul. 09		
AFR	Madagascar	EC	May 06	WB	May 08		
AFR	Malawi	EC	Jul. 05	EC	Aug. 06	EC	Jun. 08
AFR	Mozambique	EC	Mar. 06	Norway	Feb. 08		
AFR	Sierra Leone	DFID	Dec. 07	DFID	Sep. 10		
AFR	Swaziland	EC	Jan. 07	WB	May 10		
AFR	Tanzania	WB	May 06	DFID	May 10		
AFR	Uganda	EC	May 06	WB	Jun. 09		
AFR	Zambia	DFID	Dec. 05	Govt	Jun. 08		
EAP	Samoa	EC	Oct. 06	Govt	Apr. 10		
EAP	Timor Leste	EC	Feb. 07	IMF	Jun. 10		
EAP	Tonga	AusAID	Sep. 07	AusAID	May 10		
EAP	Vanuatu	EC	Jul. 06	EC	Nov. 09		
ECA	Kosovo	WB	Mar. 07	Govt	May 09		
ECA	Kyrgyz Republic	DFID	Jan. 06	SECO	Dec. 09		
ECA	Moldova	EC	Apr. 06	WB	Jul. 08		
ECA	Serbia	WB	Feb. 07	Govt	Sep. 10		
LAC	Barbados	EC	Oct. 06	EC	Jul. 10		
LAC	Dominica	EC	Apr. 07	EC	Jun. 10		
LAC	Dominican Republic	EC	Nov. 07	EC	Sep. 10		
LAC	St. Kitts and Nevis	EC	Apr. 07	EC	Dec. 09		
LAC	St. Lucia	EC	Oct. 06	EC	Feb. 10		
LAC	Trinidad and Tobago	EC	Jun. 06	EC	Dec. 08		
SAR	Afghanistan	WB	Dec. 05	WB	Jun. 08		

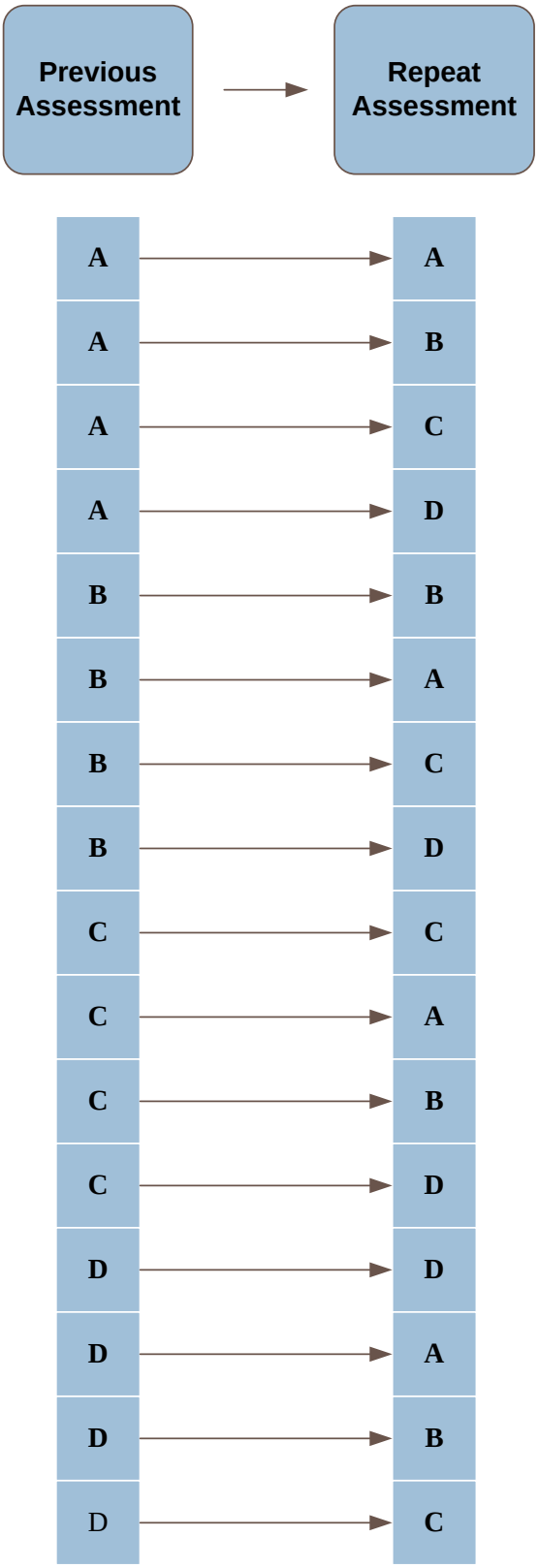
Annex B: PEFA Repeat Assessments not considered 'Comparative'

Region	Country	1st Assessment		2nd Assessment	
		Lead Agency	Date of Report	Lead Agency	Date of Report
AFR	Central African Republic	WB	Jun. 08	EC	Jul. 10
AFR	Ghana	DFID	Sep. 05	WB	Jun. 06
AFR	Sao Tome & Principe	WB	Jun. 07	EC	Jan. 10
AFR	Togo	WB	Jun. 06	EC	Mar. 09
AFR	Uganda	EC	May 06	Auditor General	Mar. 08 (*)
EAP	Lao PDR	EC	May 06	WB	Jun. 10
EAP	Papua New Guinea	WB	Sep. 05	WB	Mar. 09
LAC	Bolivia	Government	Oct. 07 (*)	WB	Aug. 09
LAC	Dominican Republic	EC	Nov. 07	Government	Nov. 09 (*)
LAC	Grenada	EC	Sep. 06	EC	Mar. 10
LAC	Honduras	WB	Jun. 06	EC	Apr. 09
Other	U.K. - Montserrat	EC	Sep. 08 (*)	IMF	Dec. 09

(*) The Secretariat did not receive a request from a stakeholder in the assessment process to undertake a review of the draft PEFA assessment report. Therefore, the PEFA Secretariat did not issue a review and comments to the lead institution.

- CAR - RA does not measure performance changes. Rescoring made in table is not justified.
- Ghana – The June 2006 report made no reference to the September 2005 report.
- St Tomé & Príncipe - RA does not refer to 2006 PA.
- Togo - PA was not a PEFA (sector PERs/CFAA/PEFA). RA does not measure performance changes.
- Uganda - the assessment did not use the 2005 assessment for measuring performance change.
- Lao - Scores of partial 2006 PEFA assessment were not cited or used in the 2010 assessment.
- Papua New Guinea - the report does not make attempt to compare to earlier PEFA ratings.
- Bolivia – RA did not use the 2007 "self-assessment" for measuring performance changes; it mentions the 2007 self assessment but no comparison was envisaged.
- Dominican Republic - is a self assessment, presented as training for Government officials, prior to the 2010 RA. It does not make reference to the prior assessment of 2007.
- Grenada – the RA made no reference to the PA.
- Honduras - PEFA 2009 does not measure performance changes over time.
- UK Montserrat - A PEFA assessment was made in 2008 but there is no reference to the scores in the 2009 assessment.

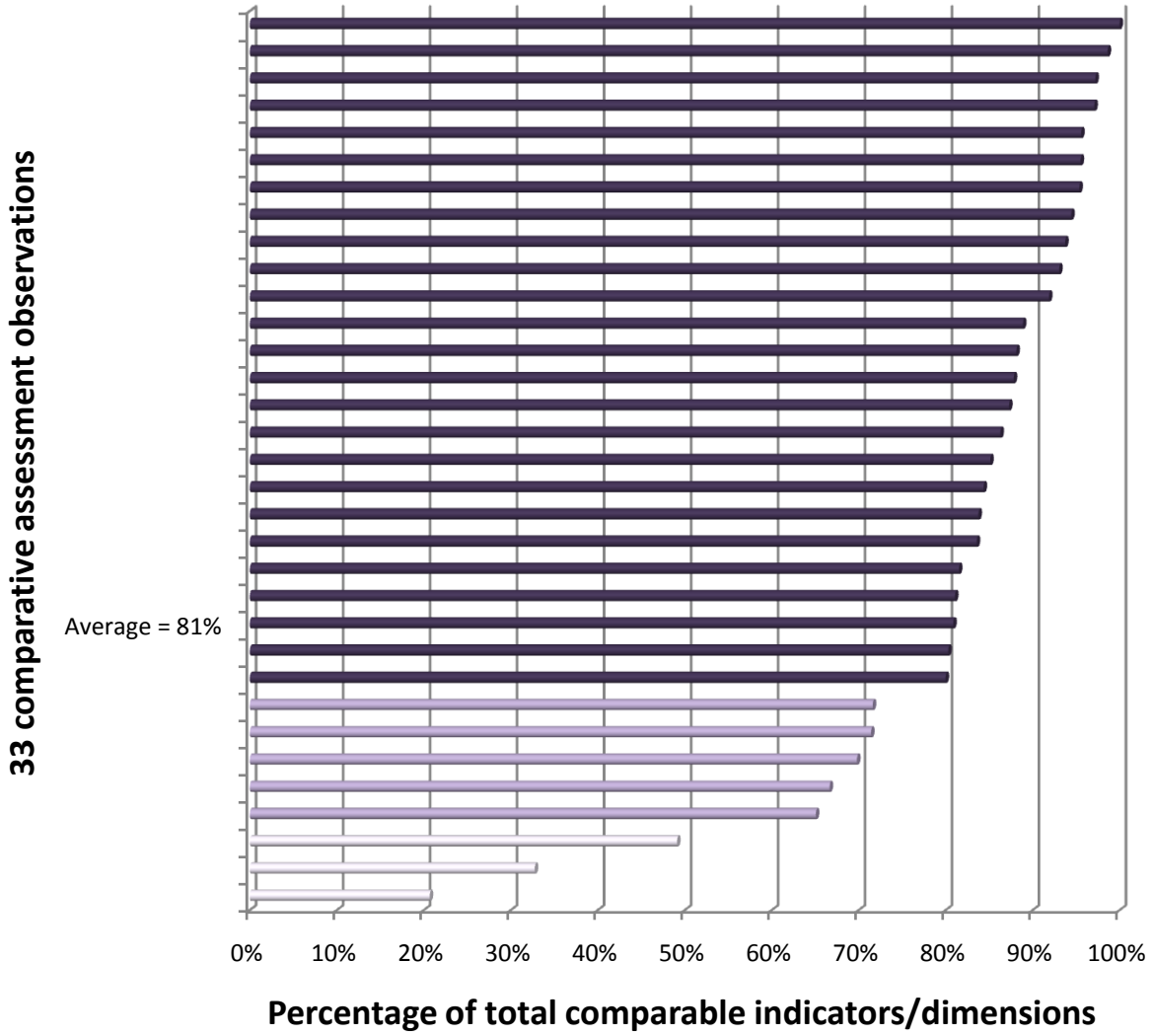
Annex C: Dimension score combinations between a previous and a repeat assessment



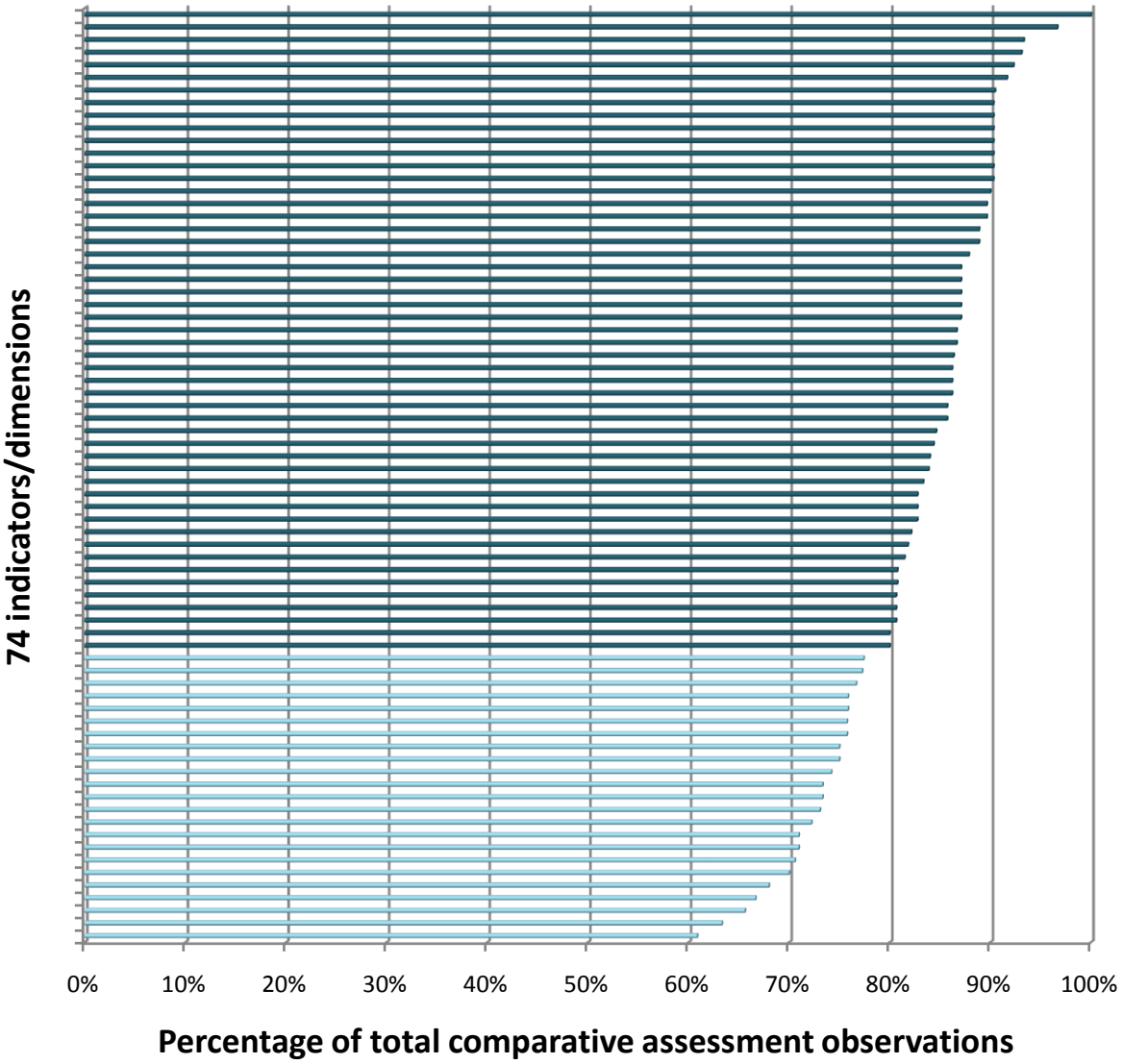
Annex D: Countries with a baseline completed prior to June 30th 2006 (i.e. over 4 ½ years ago) that have yet to complete a repeat assessment

Country	Lead Donor	Published	Status	Date on the cover of the report	Current status of comparative assessment
Bangladesh	WB	Yes	Finalized	Apr-06	Ongoing
Republic of Congo	EC	Yes	Finalized	Mar-06	No information available
Syria	IMF	No	Finalized	Mar-06	No information available
Uganda-Local Government	EC	No	Substantially	Dec-05	Planned
Fiji Islands	WB	No	Finalized	Jun-05	No information available

Annex E: Percentage of indicator dimensions that can be compared with confidence across CAs by assessment, when "no scores" are removed



Annex F: Percentage of scores that can be compared with confidence across CAs for each indicator dimension when "no scores" are removed



Legend[31]:

100-80%	D-1 (i), PI-14 (ii), PI-26 (iii), PI-14 (iii), PI-12 (ii), D-1 (ii), D-3, PI-11 (i), PI-11 (iii), PI-13 (ii), PI-16 (ii), PI-28 (i), PI-28 (ii), PI-28 (iii), PI-25 (ii), PI-18 (i), PI-21 (ii), PI-17 (i), PI-17 (iii), PI-10, PI-11 (ii), PI-12 (iii), PI-22 (ii), PI-24 (ii), PI-24 (iii), PI-14 (i), PI-23, D-2 (ii), PI-15 (ii), PI-21 (i), PI-21 (iii), PI-27 (iii), D-2 (i), PI-8 (iii), PI-22 (i), PI-8 (ii), PI-16 (i), PI-26 (i), PI-17 (ii), PI-18 (ii), PI-18 (iii), PI-27 (v), PI-6, PI-15 (iii), PI-4 (ii), PI-9 (i), PI-12 (i), PI-24 (i), PI-25 (i), PI-13 (iii), PI-20 (iii)
79-60%	PI-12 (iv), PI-9 (ii), PI-25 (iii), PI-18 (iv), PI-19 (i), PI-3, PI-5, PI-27 (i), PI-27 (ii), PI-13 (i), PI-19 (iii), PI-20 (i), PI-7 (ii) , PI-15 (i), PI-2, PI-16 (iii), PI-4 (i), PI-26 (ii), PI-8 (i), PI-20 (ii), PI-1, PI-19 (ii), PI-7 (i)

[31] The indicators are listed in descending order. For example, indicator D-1 (i) is at 100% while PI-20 (iii) is the closest to 80%.

Annex G: Percentage breakdown of comparable, no score and incomparable CAs

Indicator/ Dimension	Compared with confidence	No Score	Incomparable
PI-1	64%	3%	33%
PI-2	67%	6%	27%
PI-3	76%	0%	24%
PI-4 (i)	36%	48%	15%
PI-4 (ii)	64%	21%	15%
PI-5	76%	0%	24%
PI-6	82%	0%	18%
PI-7 (i)	42%	30%	27%
PI-7 (ii)	58%	21%	21%
PI-8 (i)	52%	24%	24%
PI-8 (ii)	64%	24%	12%
PI-8 (iii)	67%	21%	12%
PI-9 (i)	64%	21%	15%
PI-9 (ii)	52%	33%	15%
PI-10	88%	0%	12%
PI-11 (i)	85%	6%	9%
PI-11 (ii)	82%	6%	12%
PI-11 (iii)	85%	6%	9%
PI-12 (i)	76%	6%	18%
PI-12 (ii)	73%	21%	6%
PI-12 (iii)	82%	6%	12%
PI-12 (iv)	73%	6%	21%
PI-13 (i)	70%	6%	24%
PI-13 (ii)	85%	6%	9%
PI-13 (iii)	73%	9%	18%
PI-14 (i)	79%	9%	12%
PI-14 (ii)	88%	9%	3%
PI-14 (iii)	82%	12%	6%
PI-15 (i)	39%	45%	15%
PI-15 (ii)	76%	12%	12%
PI-15 (iii)	67%	18%	15%
PI-16 (i)	79%	6%	15%
PI-16 (ii)	85%	6%	9%
PI-16 (iii)	67%	6%	27%
PI-17 (i)	73%	18%	9%
PI-17 (ii)	73%	12%	15%
PI-17 (iii)	73%	18%	9%

Indicator/ Dimension	Compared with confidence	No Score	Incomparable
PI-18 (i)	79%	12%	9%
PI-18 (ii)	73%	12%	15%
PI-18 (iii)	73%	12%	15%
PI-18 (iv)	67%	12%	21%
PI-19 (i)	67%	12%	21%
PI-19 (ii)	58%	9%	33%
PI-19 (iii)	67%	9%	24%
PI-20 (i)	67%	9%	24%
PI-20 (ii)	61%	9%	30%
PI-20 (iii)	73%	9%	18%
PI-21 (i)	76%	12%	12%
PI-21 (ii)	79%	12%	9%
PI-21 (iii)	76%	12%	12%
PI-22 (i)	82%	3%	15%
PI-22 (ii)	82%	6%	12%
PI-23	79%	9%	12%
PI-24 (i)	76%	6%	18%
PI-24 (ii)	82%	6%	12%
PI-24 (iii)	82%	6%	12%
PI-25 (i)	76%	6%	18%
PI-25 (ii)	82%	9%	9%
PI-25 (iii)	70%	9%	21%
PI-26 (i)	76%	9%	15%
PI-26 (ii)	64%	9%	27%
PI-26 (iii)	85%	9%	6%
PI-27 (i)	64%	15%	21%
PI-27 (ii)	64%	15%	21%
PI-27 (iii)	73%	15%	12%
PI-27 (v)	70%	15%	15%
PI-28 (i)	85%	6%	9%
PI-28 (ii)	85%	6%	9%
PI-28 (iii)	85%	6%	9%
D-1 (i)	42%	58%	0%
D-1 (ii)	33%	64%	3%
D-2 (i)	55%	36%	9%
D-2 (ii)	58%	33%	9%
D-3	58%	36%	6%

Annex H: Percentage breakdown of changes in CAs across indicators and dimensions

Indicator/ Dimension	A to A	Increasing scores	D to D	Decreasing scores	Maintained B or C	No Score/ Incomparable
PI-1	18%	18%	6%	18%	3%	36%
PI-2	12%	12%	15%	9%	18%	33%
PI-3	45%	9%	0%	18%	3%	24%
PI-4 (i)	21%	6%	3%	3%	3%	64%
PI-4 (ii)	6%	21%	6%	12%	18%	36%
PI-5	9%	15%	0%	6%	45%	24%
PI-6	15%	36%	0%	0%	30%	18%
PI-7 (i)	24%	3%	3%	9%	3%	58%
PI-7 (ii)	9%	18%	6%	6%	18%	42%
PI-8 (i)	27%	12%	3%	6%	3%	48%
PI-8 (ii)	18%	15%	6%	15%	9%	36%
PI-8 (iii)	12%	6%	39%	6%	3%	33%
PI-9 (i)	0%	12%	3%	9%	39%	36%
PI-9 (ii)	15%	12%	18%	0%	6%	48%
PI-10	3%	27%	3%	9%	45%	12%
PI-11 (i)	21%	9%	0%	15%	39%	15%
PI-11 (ii)	21%	27%	0%	12%	21%	18%
PI-11 (iii)	24%	21%	12%	12%	15%	15%
PI-12 (i)	0%	24%	0%	6%	45%	24%
PI-12 (ii)	18%	39%	3%	6%	6%	27%
PI-12 (iii)	0%	27%	24%	15%	15%	18%
PI-12 (iv)	0%	24%	18%	9%	21%	27%
PI-13 (i)	9%	12%	0%	6%	42%	30%
PI-13 (ii)	15%	48%	0%	3%	18%	15%
PI-13 (iii)	6%	24%	0%	3%	39%	27%
PI-14 (i)	3%	21%	3%	3%	48%	21%
PI-14 (ii)	3%	39%	0%	15%	30%	12%
PI-14 (iii)	0%	42%	3%	6%	30%	18%

Indicator/ Dimension	A to A	Increasing scores	D to D	Decreasing scores	Maintained B or C	No Score/ Incomparable
PI-18 (i)	21%	18%	15%	15%	9%	21%
PI-18 (ii)	15%	24%	6%	9%	18%	27%
PI-18 (iii)	21%	27%	3%	6%	15%	27%
PI-18 (iv)	0%	21%	6%	12%	27%	33%
PI-19 (i)	0%	30%	27%	6%	3%	33%
PI-19 (ii)	0%	15%	3%	12%	27%	42%
PI-19 (iii)	3%	18%	15%	9%	21%	33%
PI-20 (i)	3%	24%	6%	9%	24%	33%
PI-20 (ii)	0%	18%	6%	3%	33%	39%
PI-20 (iii)	0%	21%	3%	6%	42%	27%
PI-21 (i)	6%	21%	27%	0%	21%	24%
PI-21 (ii)	0%	33%	12%	12%	21%	21%
PI-21 (iii)	3%	18%	27%	12%	15%	24%
PI-22 (i)	6%	33%	6%	6%	30%	18%
PI-22 (ii)	6%	30%	15%	6%	24%	18%
PI-23	6%	9%	33%	21%	9%	21%
PI-24 (i)	21%	18%	0%	12%	24%	24%
PI-24 (ii)	48%	21%	0%	3%	9%	18%
PI-24 (iii)	21%	24%	3%	9%	24%	18%
PI-25 (i)	12%	18%	9%	15%	21%	24%
PI-25 (ii)	30%	27%	9%	6%	9%	18%
PI-25 (iii)	12%	21%	6%	3%	27%	30%
PI-26 (i)	0%	30%	12%	9%	24%	24%
PI-26 (ii)	0%	30%	9%	12%	12%	36%
PI-26 (iii)	3%	24%	15%	9%	33%	15%
PI-27 (i)	9%	15%	0%	6%	33%	36%
PI-27 (ii)	6%	24%	6%	6%	21%	36%
PI-27 (iii)	18%	15%	15%	9%	15%	27%

PI-15 (i)	3%	9%	21%	6%	0%	61%
PI-15 (ii)	36%	21%	0%	3%	15%	24%
PI-15 (iii)	18%	18%	18%	9%	3%	33%
PI-16 (i)	15%	30%	6%	6%	21%	21%
PI-16 (ii)	3%	27%	6%	15%	33%	15%
PI-16 (iii)	12%	9%	3%	6%	36%	33%
PI-17 (i)	9%	21%	0%	3%	39%	27%
PI-17 (ii)	6%	30%	0%	18%	18%	27%
PI-17 (iii)	12%	24%	0%	6%	30%	27%

PI-27 (v)	3%	6%	0%	18%	42%	30%
PI-28 (i)	15%	12%	39%	15%	3%	15%
PI-28 (ii)	6%	24%	30%	12%	12%	15%
PI-28 (iii)	0%	12%	33%	6%	33%	15%
D-1 (i)	15%	12%	6%	9%	0%	58%
D-1 (ii)	3%	12%	9%	9%	0%	67%
D-2 (i)	3%	12%	15%	6%	18%	45%
D-2 (ii)	0%	18%	21%	3%	15%	42%
D-3	0%	9%	39%	6%	3%	42%

Annex I: PFM coding methodology developed by M. Andrews

PEFA Dimension No.	De jure	De facto	Concentrated	Deconcentrated	Upstream	Downstream
PI-5	1	0	1	0	1	0
PI-6	1	0	1	0	1	0
PI-7i	0	1	0	1	0	1
PI-7ii	0	1	0	1	0	1
PI-8i	1	0	1	0	1	0
PI-8ii	0	1	0	1	1	0
PI-8iii	0	1	0	1	0	1
PI-9i	0	1	0	1	0	1
PI-9ii	0	1	0	1	0	1
PI-10	1	0	1	0	1	0
PI-11i	1	0	1	0	1	0
pi-11ii	1	0	1	0	1	0
PI-11iii	0	1	0	1	1	0
PI-12i	0	1	1	0	1	0
PI-12ii	1	0	1	0	1	0
PI-12iii	0	1	0	1	1	0
PI-12iv	0	1	0	1	1	0
PI-13i	1	0	1	0	0	1
PI-13ii	1	0	1	0	0	1
pi-13iii	1	0	1	0	0	1
PI-14i	1	0	1	0	0	1
PI-14ii	0	1	1	0	0	1
PI-14iii	1	0	1	0	0	1
PI-15i	0	1	0	1	0	1
PI-15ii	0	1	1	0	0	1
PI-15iii	0	1	1	0	0	1
PI-16i	1	0	1	0	0	1
PI-16ii	0	1	0	1	0	1
PI-16iii	0	1	0	1	0	1
PI-17i	1	0	1	0	0	1
PI-17ii	1	0	1	0	0	1
PI-17iii	1	0	1	0	0	1
PI-18i	0	1	0	1	0	1
PI-18ii	0	1	0	1	0	1
PI-18iii	1	0	0	1	0	1
PI-18iv	0	1	0	1	0	1
PI-19i	0	1	0	1	0	1

PI-19ii	0	1	0	1	0	1
PI-19iii	1	0	0	1	0	1
PI-20i	0	1	0	1	0	1
PI-20ii	1	0	0	1	0	1
PI-20iii	0	1	0	1	0	1
PI-21i	0	1	0	1	0	1
PI-21ii	1	0	0	1	0	1
PI-21iii	0	1	0	1	0	1
PI-22i	0	1	1	0	0	1
PI-22ii	0	1	0	1	0	1
PI-23	0	1	0	1	0	1
PI-24i	1	0	1	0	0	1
PI-24ii	0	1	1	0	0	1
PI-24iii	0	1	1	0	0	1
PI-25i	0	1	0	1	0	1
PI-25ii	0	1	1	0	0	1
PI-25iii	1	0	1	0	0	1
PI-26i	0	1	0	1	0	1
PI-26ii	0	1	0	1	0	1
PI-26ii	0	1	0	1	0	1
PI-27i	1	0	0	1	1	0
PI-27ii	1	0	0	1	1	0
PI-27iii	1	0	0	1	1	0
PI-27iv	1	0	0	1	1	0
PI-28i	0	1	0	1	0	1
PI-28ii	0	1	0	1	0	1
PI-28iii	0	1	0	1	0	1
Totals	**26**	**38**	**26**	**38**	**16**	**48**

Annex J: References

Andrews, Matt 2009. Isomorphism and the Limits to African Public Financial Management Reform, RWP09-012, Kennedy School, Harvard, accessed May 11, 2011, http://web.hks.harvard.edu/publications/getFile.aspx?Id=340

de Renzio, Paulo, 2009. Taking Stock: What do PEFA Assessments tell us about PFM systems across countries? Working Paper 302. London: Overseas Development Institute, accessed May 11, 2011 http://siteresources.worldbank.org/PEFA/Resources/TakingStockRenzio2007.pdf

Department for International Development, 2009. Fiduciary Risk Assessment. London: DFID.

Government of the Republic of Bangladesh and Department for International Development, 2007. Assessment of the Impact of Financial Management Reforms in Bangladesh 1992 to 2006.

Mackie, Andrew and Caprio, Giovanni, 2010. Assessing the Impact of the PEFA Framework, A study for the PEFA SC, draft report, November.

Porter, Doug, Matt Andrews, Joel Turkewitz and Clay Wescott, 2010. Managing Public Finance and Procurement in Fragile and Conflicted Settings. Washington: World Bank, accessed May 11, 2011, http://wdr2011.worldbank.org/procurement

World Bank, 2006. Bangladesh Country Assistance Strategy 2006-2009. Washington: World Bank, accessed May 11, 2011, http://siteresources.worldbank.org/BANGLADESHEXTN/Resources/CAS_MAIN_BOOK_FINAL.pdf

World Bank, 2007. Public Sector Accounting and Auditing Gap Analysis. Washington: World Bank

World Bank, 2009. Bangladesh Public Expenditure and Institutional Review. Washington: World Bank, accessed May 11, 2011 http://www.spemp.com/admin/download.php?f=Public%20Expenditure%20and%20Institutional%20Review%20(PEIR%20-%20I),%202010.pdf