Questions soulevées par la comparaison et l'agrégation des résultats des évaluations PEFA à différentes périodes et entre différents pays

Version finale
Secrétariat PEFA
13 mai 2009

TABLE DES MATIÈRES

Page

Résumé analytique

i. Les évaluations PEFA ont pour but d'identifier les points forts et les points faibles des systèmes nationaux de gestion des finances publiques (GFP) pays et d'établir des niveaux de référence pour les performances, qui serviront de base aux programmes de réforme et au suivi de l'évolution des performances dans le temps.

ii. Des données très complètes étant maintement disponibles sur la performance des programmes de GFP dans différents pays, il importe à présent de déterminer comment les utiliser afin de pouvoir effectuer des comparaisons significatives entre les notes, tant au niveau national qu'international, et tant à un moment particulier que sur la durée.

iii. L'approche la moins complexe – et universellement recommandée – à cette fin consiste à comparer les notes relatives à chaque indicateur et à utiliser les explications formulées dans les rapports PEFA (RP-GFP, Rapports sur la performance dans la gestion des finances publiques) pour en tirer des conclusions sur les raisons des écarts entre les notes. Une comparaison de cette nature aide à interpréter les différences entre les pays en fonction de leur contexte national respectif. Quoiqu'une analyse de ce type soit normalement nécessaire pour comparer les performances d'un pays sur la durée, elle devient compliquée et laborieuse lorsque la comparaison porte sur plus de quelques pays. Ni dans un cas ni dans l'autre cependant, cette analyse n'aboutit à des mesures simples aisément compréhensibles par des non-experts et facilement communicables aux décideurs. Ce dernier point est particulièrement préoccupant lorsque la comparaison porte sur un grand nombre d'évaluations.

iv. Les auteurs des rapports d'évaluation, les chercheurs et les autres utilisateurs ont donc cherché à concevoir une mesure agrégée de la performance de la GFP afin de faciliter les comparaisons internationales des notes et, dans certains cas, pour faciliter à terme les progrès de la performance du pays considéré. Du fait de l'accroissement rapide de la quantité de données disponibles sur les indicateurs PEFA, ces comparaisons soulèvent un intérêt considérable. Le programme PEFA s'efforce de mettre ces données à la disposition des utilisateurs potentiels.

v. L'agrégation des résultats de l'évaluation d'un pays fait en général intervenir à la fois la conversion du classement de l'indicateur sur une échelle ordinale en une valeur numérique et l'attribution d'un coefficient de pondération à chaque indicateur. Lorsque les résultats de plusieurs pays sont regroupés, le processus d'agrégation implique également l'attribution de coefficients de pondération à chaque pays.

vi. Cette note présente des exemples de comparaisons entreprises par divers utilisateurs et examine les questions méthodologiques qu'elles soulèvent. Il n'existe pas de méthode d'agrégation et de comparaison scientifiquement juste

pour chacun des trois niveaux d'hypothèses considérés. En conséquence, le programme PEFA ne soutient ni ne recommande aucune méthode d'agrégation particulière.

vii. Le programme PEFA recommande à tous les utilisateurs – dans le cadre de la diffusion des résultats des comparaisons – d'expliquer clairement la méthode d'agrégation et les hypothèses appliquées dans chaque cas, et les raisons de leur choix. Il serait également souhaitable que les utilisateurs réalisent une analyse de sensibilité pour faire ressortir dans quelle mesure leurs constatations restent robustes sous diverses hypothèses d'agrégation.

1. Introduction

1. Des données très complètes étant maintement disponibles sur la performance des programmes de GFP dans différents pays, il importe à présent de déterminer comment les utiliser afin de pouvoir effectuer des comparaisons significatives entre les notes, tant au niveau national qu'international, et tant à un moment particulier que sur la durée.

2. L'approche la moins complexe à cette fin consiste à comparer les notes de chaque indicateur et à utiliser les explications formulées dans les rapports PEFA (RP-GFP, Rapports sur la performance dans la gestion des finances publiques) pour en tirer des conclusions sur les raisons des écarts entre les notes. Cette méthode est laborieuse et ne produit pas de mesure simple aisément compréhensible par des non-experts et facilement communicable aux décideurs.

3. Il est donc tentant de chercher à concevoir une mesure agrégée des résultats pour faciliter les comparaisons internationales des notes, qu'il s'agisse d'une mesure agrégée de la performance globale d'un pays ou d'une mesure pour chacune des six dimensions essentielles du cadre d'évaluation PEFA (« le Cadre » dans la suite du texte).

4. Une série d'études et de rapports d'évaluation ont déjà utilisé les notes des évaluations PEFA pour faciliter ces comparaisons, en calculant des notes chiffrées globales pour chacune des six dimensions essentielles du Cadre et/ou en agrégeant en une mesure unique la performance globale de la GFP.

5. L'objet de la présente étude est d'illustrer les différentes méthodes utilisées pour comparer les performances du système de GFP des pays sur la base des notes des indicateurs des rapports d'évaluation PEFA, et de signaler les éventuels écueils de ces méthodes. Elle examine la comparaison internationale des performances des systèmes de GFP telles que mesurées conformément aux six dimensions essentielles du cadre PEFA[1]. Les questions soulevées intéressent toutefois aussi les utilisateurs des rapports d'évaluation PEFA qui désirent considérer différents sous-ensembles d'indicateurs ou considérer les indicateurs individuellement.

6. Le reste de cette note est structuré comme suit :

   Section 2 : description de l'objet des comparaisons intranationales et internationales ;
   Section 3 : quelques exemples de méthodes utilisées pour la comparaison et l'agrégation
   Section 4 : discussion des problèmes que soulèvent l'agrégation et la comparaison
   Section 5 : conclusions

[1] Les dimensions essentielles du Cadre sont : Crédibilité du budget (PI-1 à 4), Exhaustivité et transparence (PI-5 à 10), Budgétisation basée sur les politiques nationales (PI-11 et 12), Prévisibilité et contrôle de l'exécution du budget (PI-13 à 21), Comptabilité, enregistrement de l'information et rapports financiers (PI-22 à 25), et Surveillance et vérification externe (PI-26 à 28).

2. Objet des comparaisons intranationales et internationales

7. Bien que le cadre PEFA ait été principalement conçu pour suivre les progrès d'un pays donné dans le temps, comparer la performance de la GFP dans des pays différents répond aussi à de nonbreuxautres besoins des parties prenantes. Les parties prenantes recherchent donc de plus en plus une méthode assez commode pour procéder à ces comparaisons (qui ne se limite pas à comparer une par une les notes de tous les indicateurs PEFA). En effet :

.

- Les pouvoirs publics peuvent souhaiter surveiller l'évolution de la performance de leur GFP dans le temps en considérant les dimensions essentielles du cadre PEFA, et comparer leur performance avec celle de pays voisins ou ayant atteint un stade similaire/différent de réforme de la GFP et/ou de développement économique général.

- Les organismes bailleurs peuvent souhaiter surveiller l'évolution dans le temps de la performance de la GFP dans chacun de leurs pays partenaires et comparer lesdites performances afin de suivre les progrès associés à leurs activités d'aide ou bien d'en tirer des enseignements susceptibles de les aider dans leurs activités de conseil et d'appui aux pouvoirs publics pour l'élaboration/la révision de leurs programmes de réforme de la GFP.

- Le regroupement par catégorie des mesures de performance de la GFP des différents pays est également une obligation prévue par la Déclaration de Paris sur l'efficacité de l'aide (indicateurs 2a, 5 et 7) pour le suivi global des progrès de la GFP.

- Enfin, les chercheurs peuvent désirer trouver des explications aux écarts entre les performances de GFP en procédant à une analyse de corrélation entre, d'une part, les différences de performance des systèmes de GFP des pays et, d'autre part, des variables explicatives reflétant leurs caractéristiques nationales, par exemple le revenu par habitant et la taille de la population.

3. Exemples de méthodes utilisées pour procéder à des agrégations et à des comparaisons

3.1 Comparaisons intranationales

8. Le cadre PEFA a été conçu principalement pour mesurer les progrès accomplis sur la durée. En 2008, des évaluations répétées, réalisées pour mesurer ces progrès, ont commencé d'être préparées[2]. Afin de mieux faire comprendre l'évolution de la performance dans le temps, les auteurs des rapports se sont efforcé de présenter les résultats de manière simplifiée et ont retenu un nombre de mesures inférieur à ceux produits par la liste des 28+3 notes attribuées aux indicateurs sur deux ans ; ils ont aussi accompagné leurs résultats d'un texte explicatif de leurs comparaisons. Deux méthodes d'agrégation ont été retenues.

9. Une méthode consiste à compter le nombre et à chiffrer l'ampleur des variations entre les classements et à en faire la somme. Le tableau 1 ci-dessous présente un exemple simplifié de ce système. Les mesures produites par cette méthode sont très simples et faciles à expliquer à un public de non-spécialistes, mais elles dépendent des hypothèses retenues tant pour la conversion de l'échelle ordinale en intervalles numériques que pour l'attribution de coefficients de pondération aux indicateurs ; ainsi, dans l'exemple donné, l'amélioration de PI-11 a-t-elle la même valeur ou est-elle plus importante que la détérioration de PI-14? Faut-il considérer le passage de la mention « non noté» à la note D+ comme une amélioration de la performance ? Et par conséquent, que signifie réellement une amélioration globale de 1 point pour 8 indicateurs ?

Tableau 1 Exemple illustratif des problèmes que pose l'agrégation des variations de performance

Indicateur	Note de 2006	Note de 2009	Variation
PI-11	C+	B	+ ½
PI-12	D	C	+ 1
PI-13	Non noté	D+	n.a.
PI-14	B+	B	- ½
PI-15	C	C+	+ ½
PI-16	C+	D+	- 1
PI-17	B	B	Pas de changement
PI-18	D+	C	+ ½
Variation globale 1/			**+ 1**

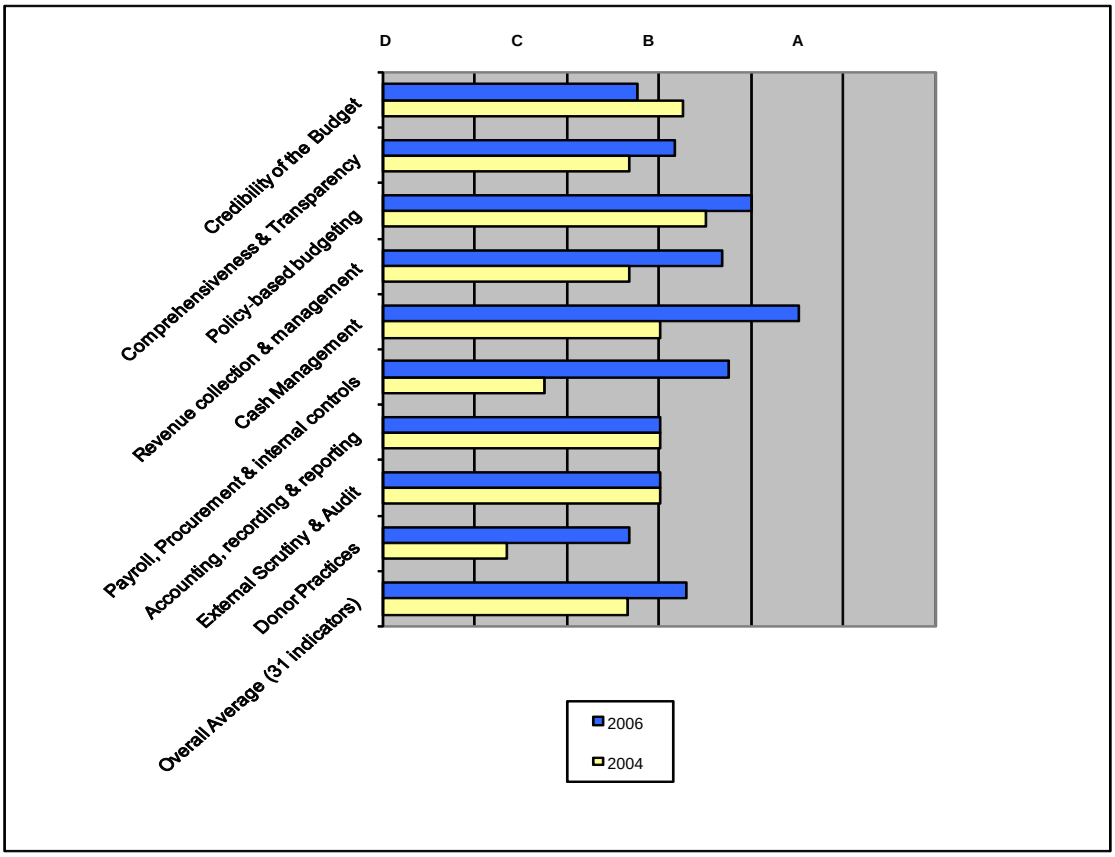
[2] Quelques évaluations répétées ont bien été réalisées mais elles ont été faites trop tôt après l'évaluation de référence pour pouvoir fournir une mesure significative des progrès sur la durée, ou bien elles visaient à établir une référence généralement acceptée lorsque l'évaluation initiale n'avait pas recueilli le soutien unanime des principales parties prenantes dans le pays. Certaines évaluations visaient aussi à mesurer les progrès réalisés depuis l'évaluation de 2004 qui avait servi à tester le projet de cadre PEFA. Comme le cadre définitif a été considérablement modifié par rapport au projet, la comparaison n'est pas possible sauf pour quelques indicateurs spécifiques.

1/ On pose en hyothèse que l'intervalle numérique est le même entre toutes les notes de l'échelle ordinale et que que tous les indicateurs ont les mêmes coefficients de pondération

10. Une seconde technique, qui est de plus en plus utilisée dans les rapports d'évaluation PEFA pour suivre l'évolution de la performance du système de GFP dans le temps consiste à calculer des notes « moyennes » pour les dimensions essentielles. Comme les notes « moyennes » sont beaucoup plus faciles à calculer à partir de chiffres, les chercheurs convertissent généralement les notes alphabétiques en notes numériques puis calculent une moyenne simple, en stipulant la même pondération pour tous les indicateurs ; en d'autres termes, tous les indicateurs ont la même importance.

11. L'évaluation PEFA du Mozambique finalisée et publiée en 2008 illustre cette technique. Les lettres utilisées pour les notes ont été converties en notes chiffrées allant de 1 à 7 (D=1, D+=2) ; ces dernières ont alors été agrégées pour chaque dimension essentielle, puis leur moyenne a été calculée sur la base de l'hypothèse (non expliquée) de coefficients de pondération égaux. La dimension essentielle « Prévisibilité et contrôle de l'exécution du budget » a été subdivisée en trois sous-dimensions essentielles : i) Recouvrement et gestion des recettes : PI 13-15 ; ii) Gestion de la trésorerie : PI 16-17 ; et iii) Gestion des états de paie, passation des marchés publics et contrôles internes (PI 18-21), cette décomposition étant justifiée par le fait que ces domaines sont différents les uns des autres au niveau des concepts. Ces indicateurs ont été complétés par ceux concernant les Pratiques des bailleurs de fonds (D1-D3).

Figure 1: Mozambique, suivi des progrès



Note : les notes agrégées ont d'abord été calculés sur la base de leurs valeurs chiffrées (1, 1,5…4) puis ont été reconverties en scores alphabétiques.

12. Exception faite des questions soulevées par l'hypothèse de pondérations égales (discutée plus en détail ci-après), le rapport sur le Mozambique montre que les progrès peuvent éventuellement être suivis dimension par dimension ainsi que globalement. L'un des principaux objectifs était de suivre les progrès depuis la première évaluation. En traçant le graphe des notes des dimensions essentielles et de la note globale pour chaque évaluation, il est possible de voir clairement les domaines dans lesquels des progrès sont en cours (figure 1). Les différences de longueur des barres relatives à chaque catégorie indiquent l'étendue des progrès ou des reculs. Lorsque l'on ne prend en compte que les écarts enregistrés, le problème de la validité de la méthode d'agrégation est moindre car seules les variations sont considérées, à l'exclusion des niveaux ; en effet certaines des notes des dimensions essentielles pourraient être différentes si elles étaient calculées sur une autre base, par exemple si le processus d'agrégation avait recours à des moyennes pondérées plutôt qu'à des moyennes simples. Cela suppose que les problèmes d'agrégation sont les mêmes d'une évaluation à l'autre, ce qui est sans doute réaliste sur une période de seulement quelques années.

13. La deuxième évaluation PEFA pour la Zambie, conduite vers la mi-2008, utilise la même technique (là encore sans justifier la validité de l'hypothèse de coefficients de

pondération égaux), la seule différence étant que la dimension essentielle de l'exécution du budget est décomposée en deux (Administration de l'impôt et Exécution du budget) et non en trois composantes.

3.2 Comparaisons internationales

14. L'étude intitulée « *Taking Stock: What do PEFA Assessments tell us about PFM systems across countries ?* » (Paulo de Renzio, University of Oxford, avril 2008) utilise les notes PEFA pour déterminer quelles variables caractéristiques de la situation d'un pays (telles que le revenu par habitant et les données démographiques) sont globalement corrélées et pourraient expliquer les différences de performance des systèmes de GFP des différents pays. À cette fin, l'auteur convertit les notes PEFA en chiffres et affecte le même poids aux différents indicateurs et les mêmes coefficients de pondération aux 57 pays de l'échantillon.

15. L'auteur de l'étude admet les problèmes de validité que cela soulève : « L'emploi de valeurs moyennes est critiquable pour un certain nombre de raisons, la principale étant que les diverses dimensions de la méthodologie PEFA mesurent en fait des réalités très différentes, qui ne se prêtent pas nécessairement à des conversions, calculs et analyses quantitatifs. En outre l'utilisation de moyennes repose sur l'hypothèse que tous les indicateurs sont d'importance égale ». Il justifie toutefois l'emploi de ces moyennes par la grande taille de l'échantillon de pays utilisé pour l'analyse et par la rigueur de la méthodologie employée.

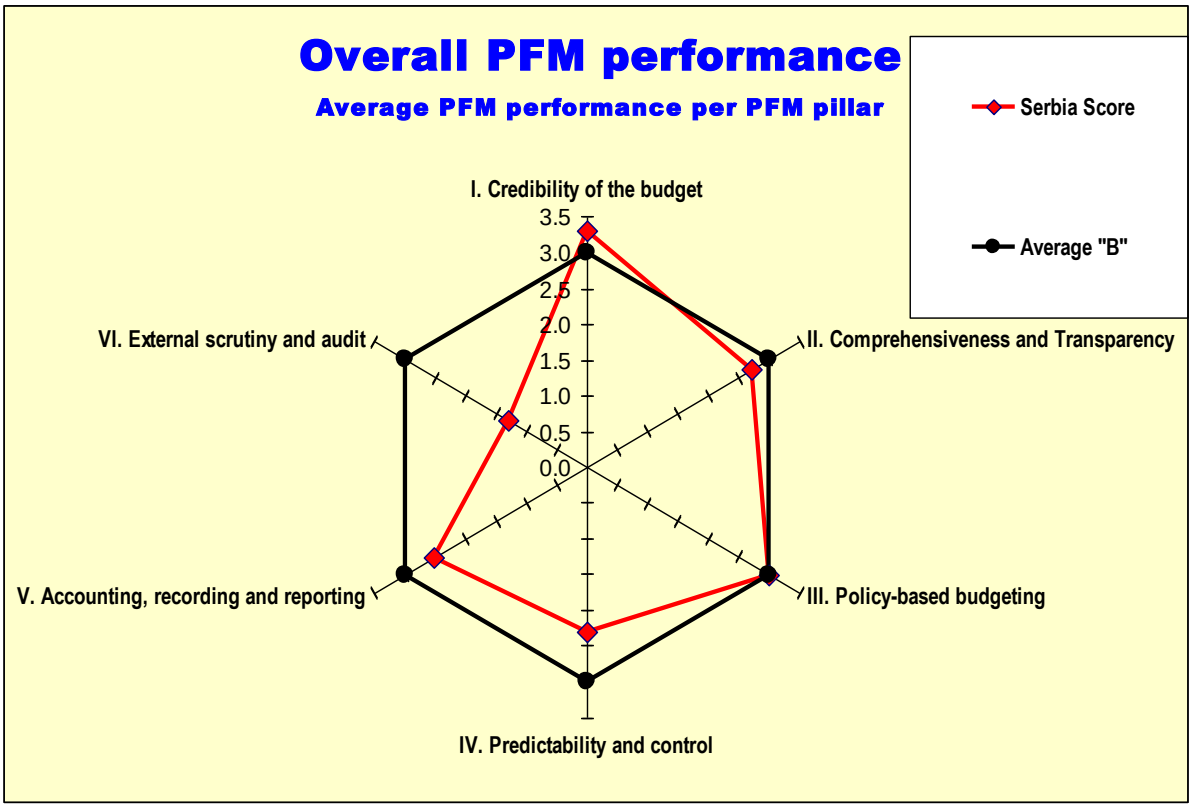
16. Dans d'autres cas, les rapports d'évaluation peuvent comparer la performance du pays évalué avec la performance « moyenne » d'autres pays de la même région, où la méthodologie PEFA a été employée pour évaluer la GFP, en présentant les résultats sous forme de graphe, par exemple, comme dans la figure 2 ci-dessous. Les hypothèses sur lesquelles reposent la conversion des notes et les pondérations sont similaires à celles retenues dans le cadre du rapport, qui sont indiquées ci-dessus.

3.3 Comparaisons à la fois entre pays et dans le temps

17. Seul un petit nombre d'évaluations répétées ont été menées avec la méthodologie PEFA dans le but explicite de déterminer l'évolution des performances dans le temps. Les notes PEFA n'ont donc servi qu'indirectement à comparer les variations de la performance dans le temps au niveau régional ou mondial, ou bien n'ont servi qu'à établir un niveau de référence devant permettre de mesurer l'évolution de la situation lorsque des évaluations répétées auront été réalisées en nombre suffisant[3]. Trois exemples d'utilisation indirecte sont présentés ci-après.

[3] Voir le cadre de résultats de la BAfD.

Figure 2 Exemple de comparaison de la performance d'un pays avec un groupe régionalobale



18. Les rapports CPIA de la Banque mondiale tirent des éléments d'information des évaluations PEFA, principalement pour l'indicateur CPIA n°13 mais aussi dans une certaine mesure pour l'indicateur n°16 (administration fiscale). Il n'y a pas conversion directe des notes PEFA en scores CPIA, mais on s'efforce de maintenir la cohérence entre les deux séries de notations. Les scores CPIA sont utilisés dans le cadre du suivi des engagements pris conformément à la Déclaration de Paris pour mesurer les améliorations globales de la qualité des systèmes de GFP des pays partenaires du Comité d'aide au développement (CAD).

19. L'Évaluation de l'appui budgétaire général en 2006 a utilisé les indicateurs PEFA comme base de mesure de l'évolution dans le temps de la performance de la GFP dans les différentes études de cas nationales qui ont servi à produire des conclusions au niveau mondial. En l'absence d'évaluations répétées, les évaluateurs ont procédé à des évaluations rétroactives des indicateurs dans le cadre d'études sur le terrain et ont appliqué une version simplifiée de la méthodologie de notation PEFA.

20. Dans l'étude intitulée « *Tracking Progress in the Quality of PFM Systems in HIPCs* » *de* P. de Renzio et W. Dorotinsky (2007) (Secrétariat PEFA, novembre 2007), les auteurs cherchent à utiliser les notes PEFA de manière plus directe et détaillée, également dans le but de mesurer les évolutions globales de la performance dans le temps. Du fait de l'absence d'évaluations répétées PEFA, les auteurs ont utilisé les

scores des indicateurs des évaluations et plans d'action des PPTE comme valeurs de référence et les notes PEFA, plus récentes, comme résultats d'évaluations « répétées ». Il leur a donc fallu convertir les notes PEFA en scores PPTE. Les auteurs soulignent également les problèmes méthodologiques et leur incidence éventuelle sur les résultats.

4. Problèmes soulevés par l'agrégation et la comparaison

21. L'agrégation des résultats de l'évaluation d'un pays fait intervenir à la fois la conversion du classement ordinal de l'indicateur en une valeur numérique et l'attribution d'un coefficient de pondération à chaque indicateur. Lorsque les résultats de plusieurs pays sont regroupés, le processus d'agrégation implique également l'attribution de coefficients de pondération à chaque pays. Chacun de ces points est examiné ci-après.

4.1 Agrégation des indicateurs

Les coefficients de pondération des indicateurs

22. Il est nécessaire d'attribuer un poids à chaque indicateur si l'on veut présenter les résultats de l'évaluation sous une forme plus simple que celle de 28(+3) résultats individuels ou d'un texte explicatif. On attribue habituellement des poids égaux aux indicateurs dans le cadre de ces exercices. Mais les conséquences de cette décision sont-elles pleinement prises en compte ? Par exemple, au tableau 1, la variation totale passerait de +1 point sur l'échelle de notation considérée à +1,5 point (une différence significative) si les indicateurs PI-11 et PI-12 recevaient des poids doubles de ceux des autres indicateurs, ce qui pourrait être justifié par le fait que la dimension essentielle de la GFP « Budgétisation basée sur les politiques publiques » ne comporte que deux indicateurs alors que les autres dimensions en comprennent entre trois et huit.

23. L'hypothèse de poids égaux est évidemment très commode car elle facilite l'agrégation. Elle est notamment très pratique pour ceux qui souhaitent agréger des notes d'évaluation PEFA afin de calculer des moyennes à partir de sous-ensembles d'indicateurs qui ne correspondent pas aux six dimensions essentielles PEFA (par exemple dans le cas du Mozambique présenté ci-dessus et dans les notes d'Évaluation des politiques et des institutions nationales – CPIA – de la Banque mondiale). Mais cette hypothèse est-elle valide ?

24. L'hypothèse de poids égaux est peut-être valide mais ce n'est pas certain, et les utilisateurs du Cadre ne doivent donc pas appliquer des poids égaux sans soulever cette question, pas plus qu'ils ne doivent recourir arbitrairement et sans l'expliquer à d'autres méthodes d'agrégation numériques (moyenne pondérée ou maillon faible entre les indicateurs à agréger).

25. Les indicateurs PEFA PI-7, 8 et 9 posent des difficultés particulières. Une note de « D » pour PI-7 (ii), c'est-à-dire la comptabilisation budgétaire de l'aide des bailleurs reçue en numéraire, résulterait en une note de D ou D+ pour l'indicateur dans son ensemble (méthode de notation M1) sans pour autant être particulièrement significative si cette aide ne représente qu'une part minime des dépenses publiques. Une note de D pour les indicateurs PI-8 et 9 pourrait être beaucoup plus significative si les dépenses des administrations infranationales sont une composante importante de

la dépense publique totale. Le calcul d'une moyenne pondérée pour la dimension essentielle « Exhaustivité et transparence » (PI-5 à 10) semblerait donc approprié. Mais il n'existe pas de réponse universelle au problème de la détermination des poids à attribuer.

26. L'hypothèse de poids égaux ne devrait pas non plus être systématiquement retenue pour les indicateurs PI-17 à 21. Des notes basses pour les composantes (i) et (iii) de PI-17 (Suivi et gestion de la trésorerie, des dettes et des garanties) peuvent ne pas être très représentatives du fonctionnement global du système de GFP si la dette ne finance qu'une part minime des dépenses publiques et si les garanties publiques de prêts sont d'une ampleur limitée. PI-18 (Efficacité des contrôles des états de paie) peut être beaucoup plus significatif que les indicateurs PI-19 à 21 (qui concernent le contrôle des dépenses non salariales) si les rémunérations du personnel sont une composante majeure des dépenses publiques tandis que les dépenses non salariales en une composante mineure. Une note basse pour l'indicateur PI-21 (Efficacité du système de vérification interne) ne porte guère à conséquence lorsque le système interne est encore en cours de développement mais que des systèmes robustes de contrôle interne ex-ante sont en place. L'adoption d'une moyenne pondérée pour les indicateurs PI-13 à 21 (Prévisibilité et contrôle de l'exécution du budget) pourrait donc être justifiable.

27. La question n'est cependant pas facile à trancher. Par exemple des insuffisances du contrôle des dépenses non salariales pourraient compromettre la qualité des contributions du personnel à la fourniture de services publics (un contrôle médiocre des achats de livres scolaires pourrait par exemple handicaper le travail des enseignants et, par là, l'efficacité de l'exécution de ce service public).

28. Dans le cas des indicateurs PI-1 à 4 (dimension essentielle « Crédibilité du budget ») du cadre PEFA, on pourrait faire valoir qu'il est justifié d'utiliser la méthode du maillon faible pour calculer une note agrégée pour cette dimension. Les situations peuvent toutefois varier et il est difficile de prescrire de manière définitive la méthode qu'il convient d'employer. Par exemple une note de D pour PI-2 mais de A pour les autres indicateurs met en cause la crédibilité du budget car les ministères et leurs services reçoivent des ressources financières sensiblement différentes de celles qui avaient originellement été budgétisées à leur intention. Une note de D pour PI-4 mais de A pour les autres indicateurs peut signifier que des paiements n'avaient pas été budgétisés, ce qui implique la possibilité de réductions budgétaires compensatrices à l'avenir pour les ministères et leurs services par rapport aux enveloppes budgétaires attendues, ce qui compromet la crédibilité du budget. En revanche, une note de D pourrait être due à l'existence d'importants arriérés dont l'apurement s'effectue de manière graduelle mais selon un calendrier établi.

29. Si l'on part de l'hypothèse que l'agrégation des notes par dimensions essentielles est une procédure valide, peut-donc considérer que procéder à une agrégation globale au niveau de toutes les dimentions et ne retenir qu'une seule lettre (ou chiffre) pour caractériser la performance du système de GFP dans son ensemble est également une opération valide ? La réponse est compliquée parce qu'il existe des liens

bidirectionnels entre les dimensions essentielles et que, par exemple, les indicateurs PI-1 à 4 sont mesurés sur des périodes glissantes de 3 ans et peuvent refléter l'incidence de la performance sur le reste des indicateurs, dont la plupart sont mesurés sur la base du cycle budgétaire annuel le plus récent. Évaluer l'importance relative de chaque dimension essentielle constituerait un exercice difficile et arbitraire. Le choix d'une moyenne simple peut être la seule méthode plausible de procéder à une telle agrégation même si les relations ne sont pas toutes de force égale ; il importe toutefois d'avoir conscience des limitations de cette approche[4].

30. On pourrait soutenir que l'application de poids égaux à toutes les dimensions essentielles n'est pas nécessairement une méthode valide car le nombre d'indicateurs rentrant dans chaque dimension est très variable (il va de deux (PI-11 et 12[5]) à neuf (PI-13 à 21)). Il s'ensuit qu'une moyenne globale obtenue par l'agrégation directe des indicateurs individuels pourrait donner un résultat différent d'une agrégation en deux étapes faisant intervenir une agrégation au niveau des dimensions essentielles (6-7 dimensions essentielles dans le cadre standard, mais 9 dans le cas du Mozambique, comme illustré plus haut).

4.2 Conversion des notes de l'échelle ordinale à une échelle numérique

31. La conversion des notes de l'échelle ordinale PEFA (A, B, C, D) en notes chiffrées est utilisée dans la plupart des exemples de la section 3, sauf lorsque seulement la distribution des notes est utilisée. Les quatre notes ordinales sont en général converties en (4,3,2,1) ou (3,2,1,0)[6], une valeur de ½ point étant attribuée au signe « + » (par exemple B=3, B+=3,5).

32. Le problème essentiel consiste à déterminer la validité de l'hypothèse selon laquelle les différences qualitatives entre les notes sont toujours similaires et les variations des notes par rapport à leur valeur initiale ont toutes la même signification. Comme la plupart des notes sont le reflet de différentes mesures prises pour améliorer, sur le plan qualitatif, les éléments indiqués des processus, des systèmes et de l'utilisation de résultats, il n'existe aucune base scientifique sur laquelle fonder la conversion numérique[7]. Bien qu'elles ne s'attaquent pas au problème des indicateurs « non notés » (voir ci-dessous), les études réalisées par le Groupe de recherche sur le

[4] Il pourrait être utile de calculer des coefficients de corrélation entre les notes moyennes établies pour chacune des dimensions essentielles 2 à 6 (c'est-à-dire la moyenne de la somme des notes pour ces dimensions) et la note relative à la première dimension. Ainsi qu'il est indiqué dans le document du Cadre, la première dimension essentielle (« Crédibilité du budget ») est, sur le plan des concepts, la résultante de l'impact combiné des autres dimensions essentielles. Cependant, un coefficient de corrélation peu élevé n'impliquerait pas nécessairement que les corrélations entre les autres dimensions essentielles sont faibles ; l'obtention de notes élevées pour la première dimension essentielle et de notes basses pour les autres pourrait simplement indiquer la possibilité de risques en ce qui concerne la crédibilité du budget, ce que pourrait clairement faire ressortir une bonne Évaluation résumée.

[5] Il importe de noter que certains éléments connexes sont inclus dans d'autres dimensions, par exemple PI-5 à 7, PI-10 et PI 27

[6] L'emploi de l'une ou l'autre échelle de conversion ne modifie pas les résultats de l'agrégation, sauf s'il est aussi atribué une valeur à la mention « non noté », comme indiqué dans la suite du texte.

[7] Les indicateurs purement quantitatifs PI-1, 2, 3 et D-3 peuvent être considérés comme des exceptions.

développement/gouvernance du secteur public de la Banque mondiale indiquent toutefois que l'apport de changements limités à l'échelle de conversion numérique n'est pas susceptible de produire des conclusions significativement différentes.

Question soulevée par les méthodes de notation MI et M2

33. À première vue, il peut paraître que le calcul de notes chiffrées pour les dimensions essentielles n'est pas valide lorsque certains indicateurs sont notés par la méthode M1 (relation la plus faible) et d'autres par la méthode M2 (moyenne simple). Une note de « D+ » pour un indicateur M1 n'a pas nécessairement tout à fait la même signification qu'une note « D+ » pour un indicateur M2 puisque le procédé de calcul de la note est différent. La question est de savoir si ces différences sont suffisamment importantes pour mériter qu'on y prête attention. Jusqu'à présent aucun utilisateur n'a jugé qu'elles justifient l'adoption d'un autre système de conversion en valeurs numériques.

Questions soulevées par les indicateurs «non notés »

34. La validité de la conversion des notes alphabétiques en notes numériques puis de leur agrégation dépend aussi du nombre d'indicateurs « non notés » parce que les informations nécessaires à cette fin sont insuffisantes ou difficiles à obtenir, ou encore parce que les indicateurs en question ne sont pas considérés pertinents (ou ont été omis délibérément de l'évaluation pour d'autres raisons). Dans ce cas, les procédures d'agrégation et de calcul de la moyenne peuvent produire des résultats erronés. Si un pays manque de données sur des indicateurs dont la note serait probablement élevée (A ou B), la note agrégée sera assortie d'un biais systématique par défaut, et inversement si les données manquantes portent sur des indicateurs pour lesquels les notes seraient probablement basses.

35. Il importe alors de déterminer s'il convient d'attribuer une valeur aux mentions « non noté » dans l'agrégation et si cette valeur doit être prise en compte dans le calcul de la note moyenne ou dans la répartition des notes. La plupart des exemples considérés ont jusqu'ici exclu les indicateurs « non notés » de l'agrégation. On pourrait toutefois soutenir qu'une mention « non noté » due à l'absence des données nécessaires décrit une situation pire qu'une note de D car elle signifie que les informations élémentaires nécessaires pour évaluer la situation ne sont même pas disponibles. Certains exemples concrets d'agrégation prennent en compte les mentions « non noté ». Par exemple la BAfD les inclut dans son cadre de résultats institutionnel et attribue une valeur de 0 (zéro) à l'indication « non noté » sur une échelle dans laquelle D=1 et A=7. La note agrégée qui en résulte est différente des notes obtenues lorsque la mention « non noté » reçoit la valeur de 1 ou n'est pas prise en compte.

36. La base de données sur les notes des rapports d'évaluation, qui est entretenue par le Secrétariat PEFA, compte trois catégories distinctes pour les indicateurs non notés, qui correspondent aux raisons de l'absence de note : NN (non noté en raison de l'absence de données), NA (non applicable dans le contexte propre au pays considéré) et NU (non utilisé pour l'évaluation parce qu'il a été décidé au départ de limiter la couverture de cette dernière). Ces trois catégories doivent être prises en compte de

manière différente dans l'agrégation des résultats de performance, la catégorie des NN étant la seule qui pourrait se prêter à des comparaisons à d'autres notes ou à une procédure d'agrégation.

Distribution des fréquences

37. Une autre méthode d'agrégation, qui évite de procéder à une conversion numérique, consiste à établir un graphique représentant la distribution des fréquences des notes : nombre de « D » en pourcentage du nombre total de notes, nombre de « C » en pourcentage …, nombre de « A ». Les notes A et D représentent habituellement un pourcentage plus faible des notes que les notes B et C, de sorte que ces graphiques comportent généralement une « queue » à chaque extrémité. Il est possible de déterminer les variations dans le temps de la performance du système de GFP d'un pays en comparant les graphiques de la distribution des fréquences, l'objectif recherché étant que la proportion des A et des B augmente avec le temps.

38. Bien que la méthode de la distribution des fréquences permette d'éviter la conversion à une échelle numérique, certains problèmes de validité demeurent. L'hypothèse implicite est que toutes les notes « A » (par exemple) ont la même signification ; en d'autres termes tous les indicateurs ont les mêmes poids et des notes similaires dénotent une performance égale – ou au moins comparable – pour tous les indicateurs. Ce système a un inconvénient par rapport à la méthode de la conversion sur une échelle numérique : il serait difficile de présenter la distribution des fréquences pour chaque dimension essentielle à différents points du temps dans un seul graphique. Ce type de représentation graphique est possible, comme le montre la Figure 1, mais seulement si l'on emploie une méthode de conversion sur une échelle numérique.

39. Si les graphiques des distributions des fréquences sont généralement construits de manière à séparer les notes par des intervalles identiques, ils peuvent aussi être conçus de manière à faire apparaître le lien entre une note de base (par exemple un « B ») et sa contrepartie « + » (B+) en présentant les deux côte à côte et en laissant un intervalle entre ces deux notes et le groupe suivant (par exemple les notes « C » et « C+ »).

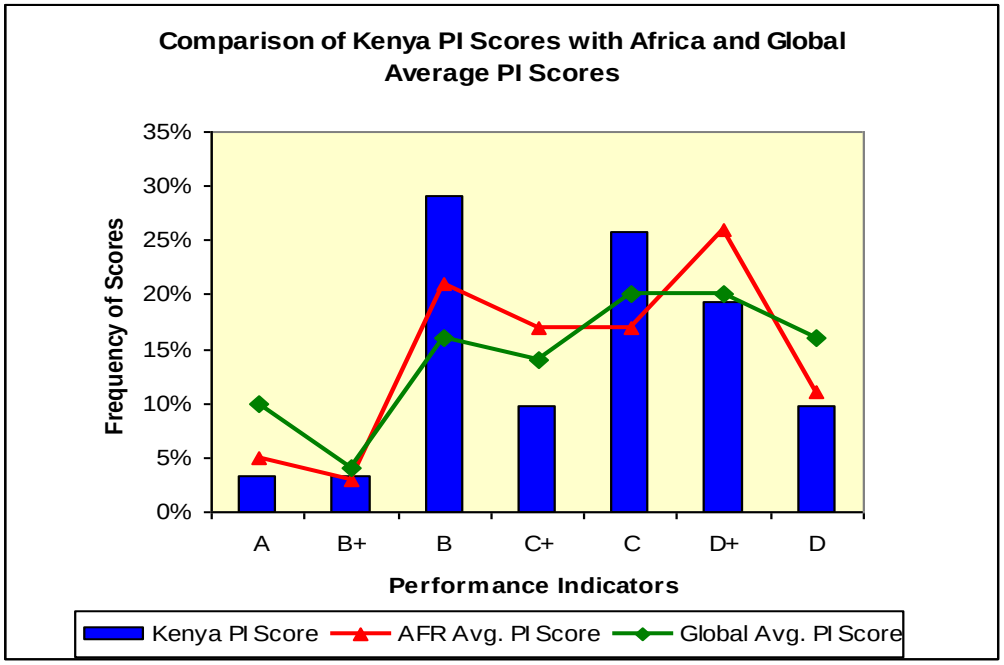
4.3 Agrégation de pays par groupes

40. Les méthodes d'agrégation numérique présentées plus haut peuvent aussi servir à faciliter les comparaisons entre pays, à une date particulière ou sur la durée. Il serait possible de modifier la figure 1 ci-dessus en procédant à une conversion sur une échelle numérique pour présenter des notes globales agrégées pour différents pays ou groupes de pays, chaque barre correspondant à une note. Les groupes peuvent être des régions, ou des groupes de pays dotés de caractéristiques similaires. Il serait aussi possible d'utiliser le graphique pour représenter les notes agrégées globales d'un pays par rapport à une moyenne régionale ou mondiale, ou encore pour présenter les notes moyennes de chaque dimension essentielle pour tous les pays ou pour un échantillon de pays. Il serait cependant difficile de porter les comparaisons entre pays par

dimension essentielle sur un seul graphique ; il est en revanche possible de présenter plusieurs graphiques (un pour chaque note agrégée globale et un pour chaque dimension essentielle) sur une seule page (voir aussi le paragraphe 45).

41. La méthode d'agrégation basée sur la distribution des fréquences peut aussi être utilisée d'une manière très similaire. À titre d'exemple, la figure 3 compare la distribution des fréquences des notes pour un pays d'Afrique (le Kenya) avec la distribution des fréquences des notes pour l'ensemble de l'Afrique et pour les pays du monde entier.

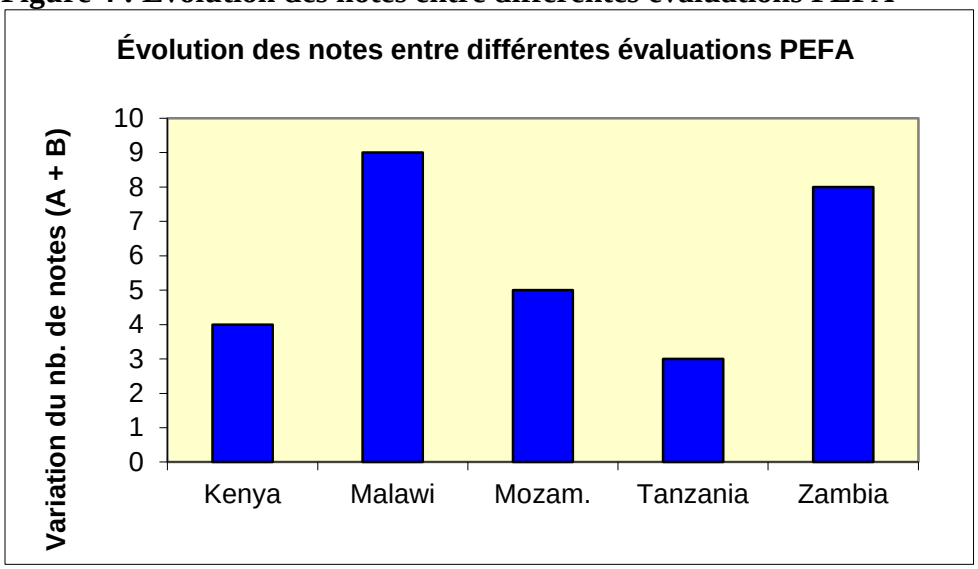
Figure 3: Répartition des fréquences des notes: Comparaison entre les notes du Kenya et des moyennes régionale et mondiale.



42. Il est possible d'inclure les données sur plus d'un pays dans un seul graphique en combinant les notes, c'est-à-dire en regroupant, par exemple, les A et les B, d'une part, et les C, D et NN d'autre part. Cette méthode fait ressortir plus clairement les différences entre la performance des systèmes de GFP des pays considérés.

43. Cette méthode permet aussi de suivre les progrès accomplis dans le temps en présentant la modification du nombre de A et de B indiquée par une évaluation répétée. Par définition, une augmentation du nombre global de A et de B doit être égale à la réduction du nombre global de notes C, D et « Non noté », de sorte qu'il n'est pas nécessaire de représenter aussi ces dernières. La figure 4 illustre ce qui précède.

Figure 4 : Évolution des notes entre différentes évaluations PEFA



i) Dates des évaluations : Kenya ; 2006, Malawi, 2006 et 2008 ; Mozambique, 2006 et 2007 ; Tanzanie, 2006 ; Zambie, 2005 et 2008.

44. Les notes des évaluations répétées PEFA utilisées pour établir la figure 4 sont de deux types. Elles sont les résultats effectifs de ces évaluations dans le cas du Malawi et du Mozambique[8], tandis qu'elles ont un caractère purement illustratif dans le cas du Kénya, de la Tanzanie et de la Zambie. La figure montre que c'est le Malawi qui affiche la plus forte augmentation du nombre de A et de B, suivi par la Zambie et le Mozambique. La figure 4 pourrait aussi servir à comparer la performance de la GFP entre régions, auquel cas chaque barre représenterait une région.

45. La distribution des fréquences des notes peut aussi servir à comparer les notes de chaque dimension essentielle pour un groupe de pays ou même pour des indicateurs spécifiques. Il est possible de présenter huit graphiques sur 1 ou 2 pages pour montrer les progrès de la réforme de manière globale et les progrès de la réforme dans chacune des six dimensions essentielles, ce qui permet d'avoir une idée générale de la contribution au progrès global des progrès réalisés au niveau de chaque dimension essentielle.

Questions soulevées par les comparaisons entre pays

46. Les comparaisons entre pays posent des problèmes dans de nombreux domaines socioéconomiques, et celles des indicateurs de performance de la GFP ne font pas exception. Le risque tient au fait que les comparaisons peuvent porter sur des éléments qui ne sont pas « semblables ». Il convient, par exemple, d'être prudent lorsque l'on compare les estimations du PIB de différents pays parce que les

[8] Rapports finaux du Malawi, du Mozambique et de la Zambie. Les rapports pour le Kenya et la Tanzanie n'étaient pas achevés lorsque les calculs ont été effectués. Lorsqu'ils seront prêts et que l'Ouganda aura été ajouté aux pays ci-dessus (le rapport de l'évaluation répétée conduite fin 2008 est pratiquement achevé), la figure 4 pourrait décrire une situation intéressante.

méthodologies statistiques peuvent être différentes et que des problèmes peuvent se poser au niveau de la collecte des données. L'emploi de définitions différentes peut aussi réduire la validité des conclusions de toute comparaison internationale. Par exemple, pour comparer les dépenses publiques de santé de différents pays, on considère souvent les pourcentages de ces dépenses dans le total des dépenses publiques et du PIB, sans tenir dûment compte des différences qui existent au niveau de la couverture du secteur (par exemple, l'eau et l'assainissement sont inclus dans le secteur de la santé publique dans certains pays, mais dans d'autres non), de la participation du secteur privé à la fourniture des services de santé et des conditions sanitaires.

47. Même si les notes globales établies pour les six dimensions essentielles de la GFP peuvent être estimées de manière valide, les comparaisons internationales des indicateurs PEFA se heurtent à un certain nombre de problèmes, notamment :

- La couverture (administration centrale uniquement, administrations infranationales uniquement, administrations centrale et infranationales) et l'année de l'évaluation peuvent différer d'un pays à l'autre ;

- Les définitions des éléments qui entrent dans la GFP, tels que arriérés de paiement, entités publiques, systèmes de transferts budgétaires, peuvent différer d'un pays à l'autre ;

- Les raisons de l'attribution d'une note identique à un indicateur donné dans différents pays peuvent varier considérablement, car la plupart des indicateurs ont au moins deux et parfois quatre composantes. Par exemple, deux pays pourraient afficher la même note pour PI-8 pour des raisons complètement différentes : le premier pourrait avoir obtenu un A pour la composante 1 et un D pour la composante 3, tandis que le second aurait reçu un D pour la composante 1 et un A pour la composante 3. Les comparaisons internationales peuvent par conséquent être superficielles et exiger une comparaison dimension par dimension, ce qui est un exercice plus complexe.

- La qualité des évaluations peut varier d'un pays à l'autre. Le processus d'assurance qualité du Secrétariat (et d'autres réviseurs) permet d'identifier les faiblesses des rapports d'évaluation PEFA. Les carences notées par le Secrétariat tiennent principalement à l'insuffisance des justifications des notes attribuées aux indicateurs (cas le plus courant) et à l'attribution de notes incorrectes compte tenu des justifications fournies. Les rapports révisés sur la base de commentaires de réviseurs ne font pas nécessairement l'objet d'une nouvelle procédure d'assurance qualité par le Secrétariat[9]. Il s'ensuit que des rapports définitifs peuvent toujours contenir des notes incorrectes.

[9] Secrétariat PEFA : *Report on Early Experience of the Application of the Framework*, novembre 2006 et *PFM Performance Measurement Framework Monitoring Report*, mars 2008, www.pefa.org.

- Les raisons d'être d'une pondération des indicateurs aux fins du calcul de note agrégées pour les dimensions essentielles ou pour la note globale (moyennes simples – dans l'hypothèse de poids égaux – moyennes pondérées ou méthode du maillon faible) sont probablement différentes d'un pays à l'autre, comme expliqué ci-dessus.

- Les proportions des dépenses publiques financées directement par l'État et financée directement par des bailleurs diffèrent d'un pays à un autre. Comparer, par exemple, les notes PEFA d'un pays où 50 % des dépenses publiques sont directement financées par des bailleurs avec celles d'un pays pour lequel cette proportion est nulle est peut-être moins significatif que la comparaison des notes PEFA entre pays où ces proportions sont similaires.

- Le nombre d'indicateurs « Non noté » « Non utilisé » et « Non applicable » peut varier selon les pays, ce qui réduit la comparabilité des notes moyennes ; en effet, les indicateurs notés ne sont pas tous forcément les mêmes ;

- Il importe aussi de déterminer s'il conviendrait de pondérer les notes d'un pays en fonction de la taille de ce dernier représentée, par exemple, par le nombre de ses habitants. Cette question se pose surtout lorsque l'on considère un regroupement régional de pays. Supposons, par exemple, que tous les indicateurs d'un très petit pays soient notés D tandis que ceux des autres pays membres de la région, qui sont beaucoup plus grands, reçoivent principalement des A et des B. Si des poids égaux sont attribués à tous les pays, la note moyenne (calculée par la méthode de conversion sur une échelle numérique) pour la région est réduite du fait de la présence du petit pays. Cette méthode peut donc avoir un impact sur la comparabilité de groupes de pays.

4.4 Comment atténuer les problèmes de comparabilité

Utilisation d'échantillons de taille importante

48. Plus l'échantillon est grand, moindre est le risque de comparaisons invalides. Les analyses pour lesquelles la taille de l'échantillon est la plus importante sont celles dans le cadre desquelles : i) les notes des dimensions essentielles PEFA sont comparées les unes aux autres pour tous les pays où des évaluations PEFA ont été effectuées (en d'autres termes, si des évaluations PEFA ont été effectués dans 50 pays, la note de chaque dimension essentielle est calculée comme la moyenne des notes des 50 pays) ; et ii) la note d'un pays est comparée à la moyenne des notes de tous les autres pays (voir, par exemple, la figure 3 ci-dessus), sous réserve que le pays soit raisonnablement représentatif.

49. Les analyses pour lesquelles la taille de l'échantillon est la plus faible sont celles dans lesquelles les notes des dimensions essentielles PEFA sont comparées pays par pays ; dans ce cas on peut douter de la validité des comparaisons. La répartition des pays en

« groupes » augmenterait la taille de l'échantillon. Cependant les groupes ne contiendront qu'un nombre limité de pays, et les problèmes de validité des comparaisons pourraient subsister[10].

Comparaisons entre pays présentant des caractéristiques similaires

50. On pourrait considérer, notamment, des pays de l'ex-Union soviétique située dans une zone géographique donnée (le Caucase par exemple), les pays africains anglophones d'une région particulière, comme l'Afrique de l'Est, et les pays anglophones de la région des Caraïbes, dont les caractéristiques nationales, juridiques et institutionnelles sont similaires et dont les réformes des systèmes de GFP ont globalement suivi des trajectoires identiques (dans le cadre, en outre, du même programme d'assistance technique financé par des bailleurs).

Ciblage de l'évolution de la performance de la GFP

51. La question de la taille de l'échantillon est également moins importante lorsque la comparaison porte sur l'évolution de la performance de la GFP dans le temps plutôt que sur son niveau à un instant donné. À supposer que les problèmes particuliers qui se posent dans un pays demeurent relativement inchangés dans le temps (hypothèse probablement raisonnable), ils n'ont pas d'impact sur la comparaison de l'évolution de la performance de la GFP. Il importe toutefois que tout rapport présentant ce type de comparaisons internationales démontre que les facteurs propres à chaque pays ne se sont pas modifiés durant la période considérée et que par conséquent les variations des notes dans le temps ont la même signification (en termes d'évolution de la qualité du système de GFP) pour tous les pays intervenant dans la comparaison.

[10] Page 10 de l'article de de Renzio ; les comparaisons sont présentées à la fois sous forme de tableaux et de graphiques à barres.

5. Conclusions

52. La comparaison des notes d'évaluation d'un pays donné à différentes périodes est l'un des principaux objectifs du Cadre PEFA. Lorsqu'elle est effectuée indicateur par indicateur et qu'elle est présentée avec un texte explicatif nuancé, la comparaison ne soulève pas de problèmes méthodologiques, mais peut en revanche ne pas se prêter aisément à la communication de messages simples sur les tendances de la performance à des non-spécialistes. Toute aggrégation de notes d'indicateur en mesures de notation plus simples soulève des problèmes de conversion sur des échelles numériques et de pondération des indicateurs.

53. La comparaison de la performance de la GFP dans différents pays par le biais de la comparaison des notes d'évaluation PEFA de ces pays est valide en principe, mais n'est pas sans risques. Les analystes doivent formellement indiquer les problèmes qui pourraient infirmer la validité de ces comparaisons et justifier toutes les hypothèses faites à cet égard (l'emploi de pondérations égales pour les indicateurs, par exemple).

54. Jusqu'à présent, les rapports d'analyses recourant à des techniques d'agrégation ont tous posé en hypothèse que les indicateurs avaient tous le même poids et que les intervalles numériques correspondant aux notes alphabétiques de l'échelle ordinale étaient d'égale amplitude.

55. Une autre façon de procéder à des comparaisons internationales consiste à comparer la distribution des notes (nombres de A, de B, de C et de D) dans les différents pays à à un instant donné ainsi que sur la durée. Cette méthode produit une analyse qui peut être plus nuancée et évite d'avoir à convertir des notes alphabétiques en notes numériques.

56. Cependant l'une comme l'autre méthode soulèvent des problèmes qui doivent être pris en compte dans les comparaisons internationales. D'autres questions se posent également. Aucun des rapports produits jusqu'ici n'examine clairement les hypothèses retenues, ne justifie leur choix et n'évalue la sensibilité des conclusions à ces hypothèses. Tous les rapports posent en hypothèse que les intervalles entre les notes A, B, C et D sont de même amplitude, que tous les indicateurs ont le même coefficient de pondération et, le cas échéant, que tous les pays figurant dans un groupe de pays ont le même poids. Ces hypothèses ont l'avantage d'être simples, transparentes et réutilisables.

57. Il n'existe pas de méthodes d'agrégation qui soit scientifiquement correcte pour chacun des trois niveaux d'hypothèses considérés. En conséquence, le programme PEFA ni ne soutient le principe de l'agrégation des résultats ni ne recommande aucune méthode d'agrégation particulière.

58. Le programme PEFA recommande à tous les utilisateurs – dans le cadre de la diffusion des résultats des comparaisons – d'expliquer clairement la méthode d'agrégation et les hypothèses appliquées dans chaque cas. Il serait également

souhaitable que les utilisateurs réalisent une analyse de sensibilité pour faire ressortir dans quelle mesure leurs constatations restent robustes sous diverses hypothèses d'agrégation.

59. La validité des comparaisons entre pays augmente :

- *Avec la taille de l'échantillon lorsque les pays sont réunis en groupes.* Plus l'échantillon est grand, moins les problèmes de comparabilité sont nombreux et, donc, importants. Les analyses pour lesquelles la taille de l'échantillon est la plus importante sont celles dans le cadre desquelles les notes des dimensions essentielles ou les notes globales PEFA sont comparées pour tous les pays, ou lorsque la note d'un pays est comparée à la moyenne des notes de tous les autres pays (sous réserve que le pays considéré ne présente aucune caractéristitque pouvant poser un problème de comparabilité). Les analyses pour lesquelles la taille de l'échantillon est la plus faible sont celles dans lesquelles les notes sont comparées pays par pays.

- *Avec la similarité des pays comparés* : il est peut-être plus légitime de comparer les notes PEFA de pays dont les caractéristiques (telles que le niveau des revenus, la taille ou les traditions administratives) sont similaires que de comparer les notes de pays très divers. La validité des comparaisons de pays appartenant à des groupes régionaux homogènes peut donc être supérieure à celle de comparaisons entre différents groupes régionaux.

- *Lorsque c'est l'évolution de la performance dans le temps qui est mesurée, plutôt que son niveau absolu :* comparer les variations des notes d'un pays donné (ou d'un ensemble de pays) peut réduire le risque de comparaisons invalides si les caractéristiques spécifiques du pays ne changent pas significativement dans le temps.

- *Lorsque le nombre et la distribution des indicateurs « non notés » sont similaires d'un pays à l'autre*. Un nombre élevé d'indicateurs « non notés », une distribution variable d'un pays à l'autre des indicateurs « non notés » et, le cas échéant, la vraisemblance des explications justifiant l'attribution de notes élevées ou basses, accroissent la difficulté que pose la réalisation de comparaisons valides entre pays.

60. Le recours à la conversion des notes d'évaluation PEFA sur une échelle numérique pour les comparaisons entre pays est indubitablement la méthode la plus simple, mais elle pose, comme l'indique la présente étude, de nombreux problèmes de validité. Il reste à voir si ces problèmes sont suffisamment importants pour justifier l'emploi des méthodes plus complexes pour procéder à des comparaisons internationales, comme indiqué à la Section 3.

61. Par conséquent, maintenant que la pratique de l'agrégation, à des fins diverses, semble se généraliser parmis les parties prenantes, il pourrait être utile de compléter l'examen théorique présenté ici par une étude des répercussions concrètes des

différentes méthodes d'agrégation. On pourrait ainsi étudier dans quelle mesure les différents facteurs considérés ici produiraient des résultats différents pour les comparaisons entre les pays (ou des groupes déterminés de pays) ou pour l'évolution de la situation d'un pays dans le temps, en partant des données réelles tirées de la base des notes d'évaluation PEFA.

APPENDICE 1 Bibliographie

Banque mondiale : Country Policy and Institutional Assessment (CPIA) : methodology www.worldbank.org

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

Dorotinsky, William et de Renzio, Paulo : *Tracking Progress in the Quality of PFM Systems in HIPCs Countries* ; Secrétariat PEFA, novembre 2007.

Fonds africain de développement : *Results Reporting for ADF-10 and Results Measurement Framework for ADF-11*, document d'information, février 2008.

Groupe indépendant d'évaluation, Banque mondiale : *Public Sector Reforms: What Works and Why? An IEG Evaluation of World Bank Support*, juin 2008.

Kaiser, Kai et Steinhilper, David : *Note on PEFA and Fiscal Flows to Sub-National Governments and Front-Line Providers* (avant-projet) ; Banque mondiale, 2008

OCDE : *Methodology for assessment of national procurement systems*, Version 4, juillet 2006.

Programme PEFA : *Performance Measurement Framework*, juin 2005

Rapports d'évaluation PEFA : Kenya, Malawi, Mozambique, Tanzanie, Zambie ; www.pefa.org.

Secrétariat PEFA : *Common Approach to PEFA Value Added and links to other PFM indicators*, août 2007.

Secrétariat PEFA : *PFM Performance Measurement Framework Monitoring Report*, mars 2008.

Secrétariat PEFA : *Report on Early Experience of the Application of the Framework*, novembre 2006, www.pefa.org.