SADEV
SWEDISH AGENCY FOR DEVELOPMENT EVALUATION

Improving assessments of effectiveness of multilateral organizations

SADEV REPORT 2012:3

Swedish Agency for Development Evaluation
P. O. Box 1902, SE-651 19 Karlstad, Sweden

SADEV REPORT 2012:3
Improving assessments of effectiveness of multilateral organizations

SADEV Reports are available at www.sadev.se
Printed in Karlstad, Sweden 2012

ISSN 1653-9249
ISBN 978-91-85679-41-6

Preface

Sweden has long promoted joined-up approaches to development cooperation. It has conceptualised development as a social transformation process grounded in the expansion of human freedoms. It has recognised reciprocal obligations between developed and developing countries. It has put policy coherence for development concepts to work in its relationships with developing countries. It has emphasised enhanced coordination of development assistance and harmonisation of aid practices. Most of all, it has been a stalwart supporter of multilateral development assistance.

According to the Development Assistance Committee (DAC), Sweden disbursed the equivalent of $2.4 billion to multilateral agencies in 2009. This represents over a third of its overall aid program. Excluding official development assistance through European institutions, Sweden leads other major donors in the share it allocates to multilateral aid (26% vs 18% for the DAC as a whole). Hence, Sweden is vitally concerned with the effectiveness of multilateral development institutions as it seeks to enhance the transparency and 'value for money' of its foreign aid activities.

The Swedish Agency for Development Evaluation (SADEV) was set up to conduct and disseminate evaluations of international development cooperation and thereby contribute to the achievement of the goals of and within Swedish development cooperation. Accordingly, I decided to fund a comparative review of current approaches to the assessment of multilateral organisations under the aegis of a collaborative initiative of senior development evaluators. The initiative, labelled '*Drawing Lessons from Comprehensive Evaluations of International Institutions'*, is led by the Global Environment Fund and the International Fund for Agricultural Development.

The attached report presents the findings of the SADEV-funded study. Based on case studies and an exhaustive literature review, it focuses on the conceptual and methodological challenges associated with development performance assessments across UN agencies, multilateral development banks and global and regional partnerships. These development institutions and collaborative programmes pursue widely different mandates. Yet, they are all committed to implementing the Busan Partnership Document, which stresses the need to reduce aid fragmentation and improve the coherence of multilateral institutions.

Unfortunately, incoherence is the current state of affairs. Typically, multilateral organisations are assessed through *ad hoc* comprehensive evaluations triggered by replenishment exercises. Not all multilaterals are subjected to such reviews. Good practice standards for their design and conduct are not available. The reviews are of mixed quality and they cannot serve as guides for aid allocation purposes across the fragmented aid architecture.

To fill this gap, a cottage industry of comparative reviews and league tables has emerged. However, the quality of these exercises is uneven; their legitimacy is often weak and the conclusions they draw are not always consistent. Against this background the report identifies the overarching issues that need to be tackled by the aid community to improve multilateral assessments. Improved collective action, reduced fragmentation and enhanced quality would result from extensive debate of its recommendations followed by concerted actions by aid donors, their developing countries partners and civil society actors.

Finally, I would like to thank Paul Isenman for carrying out this work with great skill and commitment.

August 2012

Gunilla Törnqvist
Director-General

Acknowledgements

Particular thanks are due to Keith Bezanson, Emily Bosch, and Andrew Rogerson for reviewing a draft of this report and to Paul Balogun, Robert Picciotto, Gunilla Tornqvist, and Rob van den Berg for their continuing help and advice. Thanks are also due for the time and advice of those contacted for their experience and advice, including Alexandra Chevalier, Gerry Cunningham, Thomas Dam, Chris Gerrard, Catherine Gwin, Megan Kennedy-Chouane, Uma Lele, Hans Lundgren, Aira Paivoke, Goberdhan Singh, and Christine Wallich. Thanks are due as well to participants in the workshop at UNESCO on June 14-15, 2012 of the initiative "Drawing Lessons from Comprehensive Evaluations of International Institutions." Those thanked here bear no responsibility for errors of omission or commission.

Paul Isenman, Author, July 2012.

Executive summary

This study examines comparative assessments of multilateral organizations (MOs). It draws also on comprehensive evaluations of individual MOs, a meta-analysis of their coverage (Balogun 2011, 2012), and on a workshop in June 2012 at UNESCO on "Drawing lessons from comprehensive evaluations of international institutions." It aims to draw implications for improving both comparative assessments and comprehensive evaluations. The term MO is used here in a broad sense to cover global and regional partnerships receiving support from bilateral donors and foundations, rather than being restricted to intergovernmental organizations.

Main findings regarding comparative assessments of MOs

- The current approach to comparative assessment of MOs is not a system at all, but is composed of a relatively uncoordinated and fragmented series of joint and individual efforts where the whole is less than the sum of its parts.
- Aid donors and other funders of MOs – emerging economies and foundations – should work together and accept increased accountability for meeting the commitment in the Busan Partnership Document (2011) to reduce fragmentation and "improve the coherence of our policies on multilateral institutions." In doing so they should involve other relevant stakeholders, including developing countries and civil society.
- Aid donors have recently shown increased interest in comparisons of MOs. This is indicated in the publication, for the first time, of comparative assessments of MOs by DFID (2011) and AusAid (2012), as well as in the recent annual series of DAC Reports on Multilateral Aid. There is also relevant comparative analysis in independent assessments, particularly the CGD Quality of ODA Index (QUODA, 2012).
- The analysis here underlines the difficult conceptual and methodological issues in comparing MOs, with their widely differing mandates. Nonetheless, it is vital to pursue efforts to do so in order to improve allocation funds across MOs, to draw lessons to help improve their effectiveness, and to improve the sectoral or overall "aid architecture."

Selective common weaknesses in comprehensive evaluations and comparative assessments

- Governance: Coverage of governance – a key driver of performance — is generally weak and uneven in comprehensive evaluations and even more so in comparative assessments. The same set of governance issues – e.g., relating to strategies, priorities, and accountability — identified for global programs (in the GRPP evaluation) applies to MOs as a whole.

- Aid effectiveness: More generally, assessment of MO compliance with the agreed main principles of the Paris Declaration is uneven. This applies particularly to comprehensive evaluations. Managing for Development Results (MfDR) is the only one of the five broad principles that is covered in more than half the cases.

- Incentives: Both comprehensive evaluations and comparative assessments should give increased attention to MOs' and donors' internal incentives. Internal incentives are far too important as drivers of performance and results to be neglected.

- Results and MfDR: There is a strong tendency in almost all comparative assessments to focus on systems for managing for results rather than on actual achievement of results. Results, particularly sustainable results, are vital even if they are harder to measure. Comprehensive evaluations do better here, although focusing more on efficacy than on cost-effectiveness and efficiency (value for money).

- Fragmentation: The Paris/Accra/Busan process has put increasing emphasis on reducing fragmentation in order to increase efficiency and coherence. In keeping with that process, donors should engage in more joint action and less endeavor to reduce fragmentation by participating in joint comparative assessments and in comprehensive evaluations, rather than doing their own. Those assessments and evaluations should squarely address fragmentation in activities of MOs; in the financing of MOs (less earmarking and more core financing); and in the number of mandates of multilaterals – including whether there should be mergers, closures, or clarifications of mandates.

Improving comprehensive evaluations

- It is important to have agreed principles backed by more detailed guidance on good practice for comprehensive evaluations. That general guidance should be adaptable to the unique mandate and circumstances of each MO as well as the specific objectives of the evaluation. The intended audience for such guidelines would be, importantly, those commissioning comprehensive evaluations as well as those supervising and implementing them.

- One difficult but important issue is the extent to which individual comprehensive evaluations should contribute to comparative assessments. Boards of individual MOs have relatively little incentive to facilitate comparisons. Indeed, the history of MOPAN and COMPAS suggests that MOs generally oppose such comparisons. This is partly an issue of broader governance across MOs. To what extent can boards of individual MOs be induced by their cross-MO stakeholders – donors, developing countries, and broad-mandate CSOs – to do so? An important argument here is that each MO can be more effective if it takes realistic account of its comparative advantage relative to other MOs and if it learns from benchmarking against others – a default practice in the private sector.

- A good basis for developing such principles and guidance is the "Sourcebook for Evaluating Global and Regional Partnerships and Programs" issued by the

Independent Evaluation Group of the World Bank. It is to be updated and supplemented in 2012 by a "guidebook" that takes account of the IEG's 2011 assessment of "GRPPs." Although both are, as their titles indicate, aimed at GRPPs rather than MOs in general, the extent of overlap of issues and approaches is striking. Such guidance (and collaborative work on them) would also be useful in informing DAC (or other) work on comparative assessments.

Issues in comparing MOs

- GPGs, MDGs, and Vertical Funds: GPGs, MDGs and vertical funds all focus on specific rather than general mandates. Comparative assessments tend, unnecessarily, to overweight them relative to the (Paris) principles of aid effectiveness and (except in the case of "pure" GPGs) to other development priorities.

- Normative MOs: Comparative assessments understate the returns to norm-setting MOs, judging them primarily by their effectiveness in delivering aid at country level. Comparative efficiency and allocations:

  × There are difficult analytic issues in comparing returns across MOs. One is, as noted, the limited availability of information on sustainable results. This encourages overreliance on management systems and other aspects of "organizational effectiveness" rather than on "development effectiveness" or value for money. Then there are questions of how to take account of differing mandates of MOs and of different priorities – strategic and commercial as well as developmental — of donors. How does one compare the value of human rights, for example, against the spread of health or education? The MAR and AMA are straightforward about the inevitable arbitrariness of weighting systems and about the need for judgment in determining the overall ratings that then serve as a proxy for development effectiveness.

  × Three additional issues should be addressed to relate overall ratings to donors' priorities:

    - Current comparative assessments do not take scale into account: a given MO would get roughly the same rating whether its rating was double or half its current level.

    - Implicit double counting is frequent as assessment criteria often overlap.

    - Those carrying out comparative assessments do not know the extent to which current criteria, mostly common sense, have already been taken implicitly into account in prior allocation decisions. "Good" is not the same as "more."

- To tackle these difficult issues donors have relied on perception surveys of selected staff – or have simply made executive decisions without much consultation or analysis. In fact, donors have had, in MOPAN, a joint perception survey of MOs for a decade. But they have not until 2012 included any questions that would help address comparative development effectiveness or value for money.

- For the reasons above it will be challenging for donors, individually or jointly, to determine objective zero-based allocation levels. However, it should be feasible to use changes in ratings on agreed criteria to guide incremental changes in allocations. Then as now it will be necessary to use "triangulation" of different methods.

- Each donor will continue to make its own funding and related policy decisions, taking account of its national priorities as well as "burden sharing." But joint comparative analysis by donors would reduce costs and increase the quality of that analysis and would contribute to better-informed joint decision making.

Recommendations for improving comparative assessments:

- Bilateral Assessments: Bilateral donors' past year efforts to engage in multilateral assessment and policy setting are encouraging. If sustained, these efforts could reduce fragmentation as well as add value to decisions on policy, reform, and allocations. It would be highly desirable for these efforts to lead to maximizing joint donor analysis, assessment, and action. This still would leave more than ample room for sovereign decision-making.

- QUODA: QUODA provides a good basis for comparison of behavior of donors and MOs on a broader set of aid effectiveness indicators than those of the Paris Declaration, but it currently does not address results. Like bilateral assessments, it should drop questions that give extra weight to vertical programs simply on the basis of their concentration. QUODA's sustainability will depend on the continuing availability of data from Paris Monitoring Surveys.

- CIDA-EVALNET: The CIDA-EVALNET approach based on review of MO evaluations is a useful complement to other comparative assessments. It addresses similar questions to those of MOPAN with complementary sources of evidence as well as complementary strengths and weaknesses. Impact would increase if existing links to MOPAN were strengthened, including considering the CIDA-EVALNET approach as a third source of "triangulation" under MOPAN's "common approach."

- COMPAS: COMPAS has limited potential for comparison among its MDB members because it is based on self-evaluation that serves to publicize progress as well as to exchange good practice. Its added value depends on what use individual MDBs make of it internally since they oppose its comparative use.

- MOPAN: MOPAN, with 16 donor members, provides tailored assessments based on a common methodology of a rotating series of MOs. Although its website says: "It is not possible to compare multilateral organizations to one another", it is in practice the source of comparative information on MOs most frequently cited by donors (in the 2011 RMA). Its secretariat, now rotating, is planned to be hosted by the DAC Secretariat, while retaining its independent governance. This should facilitate complementarity with the new donor effort on policy as well as with the DAC EVALNET.

  It would be useful for the MOPAN evaluation scheduled for late in 2012 to include, as part of strengthening its methodology and impact, the contentious issue of adjusting its methodology to facilitate comparability among MOs. MOPAN's expert perception surveys can contribute to "triangulation" on the difficult methodological issues of comparative assessments cited above.

In sum, there are important opportunities for improved quality and consistency, as well as for more collective action and reduced fragmentation, of both comparative assessments and comprehensive evaluations of MOs. These are not ends in themselves but need to be designed and implemented so as to maximize their contributions to increased efficiency of multilateral organizations and improved architecture and allocation of funds among them.

Sammanfattning

I den här rapporten görs jämförande analyser av multilaterala organisationer (MO). Den baseras på breda utvärderingar av enskilda MO, en metaanalys av deras innehåll (Balogun 2011, 2012) och en workshop i juni 2012 på UNESCO med titeln ”Drawing lessons from comprehensive evaluations of international institutions”. Syftet är att beskriva konsekvenser, för att förbättra både jämförande analyser och breda utvärderingar. Begreppet MO används här i vid bemärkelse, och omfattar globala och regionala partnerskap som får stöd från bilaterala biståndsgivare och stiftelser. Begreppet är inte begränsat till mellanstatliga organisationer.

Huvudsakliga slutsatser från jämförande analyser av MO

- Det nuvarande sättet att genomföra jämförande analyser av MO är osystematiskt och ger en ganska splittrad och fragmenterad serie av gemensamma och individuella angreppssätt där helheten är mindre än summan av sina beståndsdelar.
- Biståndsgivare och andra finansiärer av MO – nya biståndsländer (tillväxtekonomier) och stiftelser – bör samarbeta och ta ett ökat ansvar för att uppfylla åtagandet i Busan Partnership Document 2011, för att minska fragmenteringen och ”göra politiken om multilaterala institutioner mer sammanhängande”. I denna strävan bör de involvera andra relevanta aktörer, inbegripet utvecklingsländer och det civila samhället.
- Biståndsgivare har nyligen visat intresse för jämförelser av MO. Det anges för första gången i den publicering av jämförande analyser av MO som har gjorts av DFID (2011) och AusAid (2012), samt i den nya årliga serien rapporter från DAC om multilateralt bistånd. Det finns även relevanta jämförande analyser i oberoende undersökningar, särskilt QUODA (CGD Quality of ODA Index, 2012).
- Analysen understryker här de svåra begreppsmässiga och metodologiska problemen med jämförelser av MO, med deras vitt skilda mandat. Det är emellertid ytterst viktigt att sträva efter att göra dessa jämförelser för att förbättra genomslag av anslagna medlen över alla MO. Vidare bör analysen dra lärdomar och förbättra deras effektivitet samt förstärka ”biståndsarkitekturen” rent generellt.

Gemensamma svagheter i breda utvärderingar och jämförande analyser

- Styrning: Styrning som är en viktig drivkraft för att uppnå goda resultat är ofta dåligt och ojämnt belyst i breda utvärderingar och ännu sämre i jämförande analyser. Samma uppsättning styrningsfrågor – som avser strategier, prioriteringar och ansvarsskyldighet – och som har identifierats inför globala program (i utvärderingen av GRPP) gäller också för MO som helhet.

- Biståndseffektivitet: I allmänna ordalag är analysen av MO:s efterlevnad av de avtalade huvudprinciperna i Parisdeklarationen ojämn. Detta gäller särskilt de breda utvärderingarna. MfDR (Managing for Development Results) är den enda av de fem huvud principerna som täcks i mer än hälften av fallen.
- Incitament: Både breda utvärderingar och jämförande analyser bör rikta ökad uppmärksamhet mot MO:s och biståndsgivarnas interna incitament. Interna incitament är alldeles för viktiga som drivkrafter för prestationer och resultat för att försummas.
- Resultat och MfDR: Det finns en stark tendens i nästan alla jämförande analyser att fokusera på ledningssystem för att uppnå resultat i stället för att uppnå faktiska resultat. Resultat, särskilt hållbara resultat, är ytterst viktiga, även om de är svårare att mäta. Breda utvärderingar lyckas bättre här, även om de fokuserar mer på ändamålsenlighet än kostnadseffektivitet och effektivitet (valuta för pengarna).
- Fragmentering: Paris-/Accra-/Busanprocessen har lagt ökad tonvikt vid att reducera fragmenteringen för att öka effektiviteten och sammanhållningen. Enligt denna process bör biståndsgivarna ägna sig mer åt gemensamma åtgärder och sträva mindre efter att minska fragmenteringen genom att delta i gemensamma komparativa analyser och i breda utvärderingar, i stället för att göra egna. Dessa analyser och utvärderingar bör påpeka fragmenteringen i MO:s verksamhet; se över finansieringen (mindre öronmärkning och mer kärnfinansiering); och vad beträffar antalet multilaterala organisationer och deras uppdrag, däribland huruvida samgåenden, avslut eller klargöranden av uppdrag, ska ske.

Förbättrade breda utvärderingar

- Det är viktigt att ha avtalade principer med stöd av mer detaljerad vägledning om god praxis för breda utvärderingar. Denna allmänna vägledning bör anpassas till det unika uppdraget och omständigheterna för varje MO samt utvärderingens särskilda mål. Den avsedda publiken för denna vägledning ska givetvis vara de som beställer breda utvärderingar samt de som övervakar och implementerar dem.
- En svår men viktig fråga är i vilken utsträckning enskilda breda utvärderingar kan bidra till jämförande analyser. Styrelserna i enskilda MO har relativt små incitament att genomföra jämförelserna. Historien bakom MOPAN och COMPAS tyder faktiskt på att MO generellt motsätter sig sådana jämförelser. Det är särskilt en fråga om en bredare styrning mellan MO. I vilken utsträckning kan styrelserna i enskilda MO förmås av sina MO-överskridande intressenter – biståndsgivare, utvecklingsländer och det civila samhällets organisationer med brett uppdrag – att göra det? Ett viktigt argument här är att varje MO kan bli mer effektiv om den tar realistisk hänsyn av sin relativa fördel jämfört med andra MO och om den lär av jämförelser med andra – en normal praxis i den privata sektorn.

- En god grund för att utveckla sådana principer och sådan vägledning är ”Sourcebook for Evaluating Global and Regional Partnerships and Programs”, som har utfärdats av Världsbankens oberoende utvärderingsgrupp. Den ska uppdateras och kompletteras 2012 av en ”guidebok” som tar hänsyn till IEG:s bedömning av ”GRPP” 2011. Trots att båda, som deras titlar antyder, är inriktade på GRPP, och inte MO i allmänhet, är det slående hur många frågor och metoder som överlappar varandra. Sådan vägledning (och samarbetande arbete på dem) skulle också bidra till att informera DAC:s (eller andras) arbete med komparativa analyser.

Frågor vid jämförelser av MO

- Globala allmänna nyttigheter, millennieutvecklingsmål och vertikala fonder: Globala allmänna nyttigheter, millennieutvecklingsmål och vertikala fonder är alla inriktade på specifika snarare än allmänna mandat. Jämförande analyser tenderar att lägga alltför stor vikt vid dem jämfört med Parisprinciperna om biståndseffektivitet och (utom när det gäller ”rena” globala allmänna nyttigheter) andra utvecklingsprioriteringar.

- Normativa MO: Komparativa analyser underskattar betydelsen för MO:s normskapande arbete, och bedömer dem främst för deras effektivitet i leveransen av bistånd på landnivå. Komparativ effektivitet och anslag:

  × Det finns stora analytiska problem vid jämförelsen av insatserna mellan olika MO. Ett är, såsom har noterats, den begränsade tillgången till information om hållbara resultat. Det uppmuntrar till en alltför stor tilltro till ledningssystem och andra aspekter på ”organisatorisk effektivitet” i stället för ”utvecklingseffektivitet” eller valuta för pengarna. Sedan finns det frågor om hur man tar hänsyn till MO:s olika mandat och biståndsgivarnas olika prioriteringar – såväl strategiska och kommersiella som utvecklingsmässiga. Hur jämför man exempelvis värdet på mänskliga rättigheter med spridningen av hälsa eller utbildning? MAR och AMA är öppna kring den oundvikliga godtyckligheten i viktningssystemen och kring behovet av bedömningar vid avgörandet av de övergripande klassificeringar som sedan fungerar som fullmakt för utvecklingseffektivitet.

  × Ytterligare tre frågor bör hanteras för att ställa den övergripande klassificeringen i relation till biståndsgivarnas prioriteringar:

    - Aktuella jämförande analyser tar inte hänsyn till gradindelning: En viss MO skulle i princip få samma betyg oavsett om dess betyg var dubbelt eller hälften så högt som dess nuvarande nivå.

    - Underförstådd dubbel räkning är vanligt, eftersom bedömningskriterierna ofta överlappar varandra.

    - De som utför jämförande analyser vet inte i vilken utsträckning aktuella kriterier, främst sunt förnuft, redan har beaktats underförstått i tidigare beslut om finansiering. ”Bra” är inte detsamma som ”mer”.

× För att hantera dessa svåra frågor har biståndsgivarna förlitat sig på undersökningar av uppfattningar bland utvald personal – eller helt enkelt fattat verkställande beslut utan särskilt mycket konsultation eller analys. Inom MOPAN har man haft en gemensam undersökning av biståndsgivarnas uppfattning om MO i tio år. Det är dock inte förrän 2012 som de har inkluderat frågor som kan bidra till att hantera den komparativa utvecklingseffektiviteten eller valutan för pengarna.

× Av ovan nämnda skäl blir det utmanande för biståndsgivarna att individuellt eller gemensamt fastställa objektiva nollbaserade finansieringsnivåer. Det bör dock vara möjligt att använda förändringar i betyg om avtalade kriterier för att vägleda gradvisa förändringar i finansieringen. Då som nu blir det nödvändigt att använda ”triangulering” av olika metoder.

× Varje biståndsgivare kommer att fortsätta att fatta sina egna finansieringsbeslut och tillhörande politiska beslut, med hänsyn till sina nationella prioriteringar och ”delning av bördan”. Gemensamma jämförande analyser av biståndsgivare skulle dock sänka kostnaderna och öka kvaliteten på denna analys och bidra till ett bättre underbyggt gemensamt beslutsfattande.

Rekommendationer för att förbättra jämförande analyser:

- Bilaterala bedömningar: Bilaterala biståndsgivares arbete under det senaste året för att göra multilaterala analyser och fastställa politik är uppmuntrande. Om detta arbete fortsätter kan det minska fragmenteringen och ge mervärde åt beslut om politik, reformer och anslag. Det vore mycket önskvärt om detta arbete ledde till att de gemensamma biståndsgivaranalyserna, bedömningarna och åtgärderna maximeras. Det skulle ge mer än tillräckligt utrymme för ett oinskränkt beslutsfattande.

- QUODA: QUODA ger en god grund för jämförelser av beteenden hos biståndsgivare och MO utifrån en bredare uppsättning indikatorer på biståndseffektivitet än de som finns i Parisdeklarationen, men den tar för närvarande inte upp resultaten. Precis som bilaterala analyser bör den ta bort frågor som ger extra tyngd åt vertikala program enbart på grund av sin koncentration. QUODA:s hållbarhet kommer att bero på att data från övervakningsundersökningarna från Paris är fortsatt tillgängliga.

- CIDA-EVALNET: CIDA-EVALNET:s strategi, som är baserad på en översyn av utvärderingar av MO, är ett användbart komplement till övriga jämförande analyser. Den tar upp liknande frågor som MOPAN med kompletterande beviskällor samt kompletterande styrkor och svagheter. Påverkan skulle öka om de befintliga länkarna till MOPAN stärktes, vilket innefattar att CIDA-EVALNET:s strategi betraktas som en tredje källa till ”triangulering” enligt MOPAN:s ”gemensamma strategi”.

- COMPAS: COMPAS har begränsad potential för jämförelser mellan sina MDB-medlemmar, eftersom den är baserad på egenutvärdering som syftar till att offentliggöra framsteg samt utbyta god praxis. Dess mervärde beror på hur enskilda MDB använder den internt, eftersom de motsätter sig dess komparativa användning.
- MOPAN: MOPAN har 16 biståndsgivare som medlemmar och organisationen erbjuder skräddarsydda analyser av MO:s baserade på en gemensam metod enligt ett roterande mönster. På dess webbplats står dock följande: ”Det går inte att jämföra multilaterala organisationer med varandra”, men i praktiken är det en källa till komparativ information om MO som oftast anges av biståndsgivarna (i 2011 års RMA). Enligt planen ska dess sekretariat, som nu är roterande, ha DAC-sekretariatet som värd, medan det behåller sin oberoende styrning. Det bör göra det lättare att komplettera biståndsgivarnas nya politiska arbete samt arbetet med DAC EVALNET.

  Det vore bra om den utvärdering av MOPAN som planeras till slutet av 2012, som en del av stärkningen av dess metod och påverkan, kunde innefatta den omstridda frågan om anpassning av metoden för att underlätta jämförelser mellan olika MO. MOPAN:s undersökningar av experters uppfattningar kan bidra till ”triangulering” av de besvärliga metodologiska frågorna i de komparativa analyser som anges ovan.

Totalt sett finns det stora möjligheter till bättre kvalitet och konsekvens, samt mer kollektiva åtgärder och minskad fragmentering, av både jämförande analyser och breda utvärderingar av MO. Det här är inte ändamål i sig, utan måste utformas och implementeras för att maximera deras bidrag till en ökad effektivitet i multilaterala organisationer och en förbättrad arkitektur och fördelning av medel mellan dem.

Acronyms

AMA	Australian Multilateral Assessment
AusAid	Australian Agency for International Development
CGD	Center for Global development
CGIAR	Consultative Group on International Agriculture Research
CIDA	Canadian International Development Agency
COMPAS	Common Performance Assessment System
CSO	Civil Society Organization
DAC	Development Co-operation Directorate
DEREC	DAC Evaluation Resource Centre
DFID	Department for International Development (UK)
FAO	Food and Agriculture Organization of the United Nations
GAVI	Global Alliance for Vaccines and Immunization
GPG	Global Public Goods
GRPP	Global and Regional Partnerships and Programs
IATI	International Aid Transparency Initiative
MAR	Multilateral Aid report
MDG	Millennium Development Goals
MfDR	Managing for development Resources
MO	Multilateral Organization
MOPAN	Multilateral Organization Performance Assessment Network
ODA	Official Development Assistance
RBM	Resluts Based Management
RMA	Report on Multilateral Aid
SADEV	Swedish Agency for Development Evaluation
TOR	Terms of Reference
QUODA	Quality of ODA
UNESCO	United Nations Educational, Scientific and Cultural Organization

Contents

1 Learning from assessments of overall effectiveness of multilateral organizations

1.1 Introduction

This report examines two types of assessments of overall effectiveness of multilateral organizations (MOs) – comparative assessments across a broad range of MOs and (a summary of) comprehensive evaluations of individual MOs. The report seeks to draw implications for improvement in the coverage and quality of both, including taking better account of the demand from donors for better information on value for money among MOs. (The term MO is used here in a broad sense to cover global and regional partnerships receiving donor support.) The report was commission by SADEV on behalf of the initiative "Drawing Lessons from Comprehensive Evaluations of International Institutions." As the October 2011 report of the initiative stated: "The current ad-hoc approach (to comprehensive evaluations) has resulted in evaluations that do not yield optimum value for money." (van den Berg, 2011).

Comparative assessments covered here fall into three categories:

- Multilateral assessments: Multilateral Organization Performance Assessment Network (MOPAN, draft materials for 20012); the CIDA-EVALNET (Pilot Test) approach of a meta-analysis of evaluations; the Common Performance Assessment System of the Multilateral Development Banks (COMPAS, 2010); Second Evaluation of Paris Declaration (2011); and, more generally, the DAC Reports on Multilateral Aid (MRA, 2008-2011);
- Assessments by individual donors: DFID Multilateral Aid Report (MAR, 2011); Australian Multilateral Assessment (AMA, 2012); World Bank IEG Evaluation of Global and Regional Partnership Programs (GRPP Evaluation, 2010) and its underlying Sourcebook (2007);
- Independent assessments: CGD Quality of ODA Index (QUODA, 2012).

These main sources on comparative assessments were chosen on the basis of: relevance –being cited more than once in the survey of donors underlying the 2011 DAC Report on Multilateral Aid (RMA); recentness; and availability. The report was also informed by the analysis done for the Finnish Ministry of Foreign Affairs on the criteria used in a range of comparative assessments. The Appendix provides a "long list" of comparative assessments from which those analyzed here were drawn. For a summary of coverage of comprehensive evaluations of individual comprehensive evaluations, this report draws on Balogun (2011, 2012).

1.2 Differences in coverage

There are substantial differences in the extent and quality of coverage of criteria and issues in both assessments and evaluations. These are illustrated in Table 1: Depth of Coverage of Selected Key Criteria, and Table 2: Key Criteria of Comparative Assessments.

Table 1 Depth of Coverage of Selected Key Criteria

	Institution and its External Environment (Mandate, Relevance)	Governance	Organizational Effectiveness	Effectiveness (Efficacy)	Cost Effectiveness	Efficiency (Value for money)
MOPAN	No	No	Yes	Yes (as of 2012)	No	No
COMPAS (only intended to cover MfDR)	No	No	Partly (re MfDR)	Partly (re MfDR)	No	No
QUODA	No	No	Yes	Yes (focus on allocations)	No	No
MAR	Yes	Yes	Yes	Yes	Yes	Partly (Issue of scale)
AUSTRALIAN AMA	Yes	No	Yes	Yes	No	No
Pilot Phase MI	No	No	Yes	Yes	No	No
2nd Evaluation of Paris Declaration (only intended to cover PD)	No	No	Yes	Yes	No	No
Sourcebook GRPP	Yes	Yes	Yes	Yes	Yes	No
WB Evaluation GRPP 2010	Yes	Yes	Yes	Yes	Yes	No
Coverage in Comprehen-sive Evaluations (average of components in Balogun Note 2)	13/17 (excludes "partnerships")	11/17 (only 4/17 for NGOs and 3/17 for private sector)	10/17	10/17 ("sustainable development impact")		

(Format and coverage of comparative assessments adapted from Balogun Note 2)

As Table 1 shows, comprehensive evaluations generally give emphasis to organizational effectiveness and efficacy (effectiveness in achieving agreed objectives). They are generally much weaker in other areas: how the MO relates to other relevant organizations; governance; and efficiency (value for money). Yet these are areas of strong interest to donors (defined here to include non-official funders), in their allocations across MOs and as participants in governance structures.

Table 2 Key criteria of selected comparative assessments

DAC EVALUATION Criteria and (in parentheses) Principles of Paris Declaration	Balogun (2011,2012)(issues identified for meta evaluation)	MOPAN	DFID MULTILATERAL AID REVIEW	AUSTRALIAN MULTILATERAL ASSESSMENT	QuODA	PILOT TEST	WORLD BANK GRPP SOURCEBOOK
Relevance	Institution re its external environment		Critical role in meeting development objectives	Alignment with Australia's priorities		Relevance	Relevance
			Focus on poor countries	Contribution to multilateral system	(Share of poor countries is part of maximizing efficiency)		
(Neither in DAC criteria nor Paris principles)	Governance						Governance
(Managing for development results)		Strategic management	Strategic performance management	Strategic management and performance			
	Organizational effectiveness	Operational management			Maximizing efficiency (mostly through allocations)	Using evaluation for effectiveness	
			Financial resource management	Cost and value consciousness	Cost and value consciousness		Resource mobilization
(Ownership, alignment, harmonization)		Relational management	Partnership behavior	Partnership behavior	Fostering institutions (harmonization and alignment)		
					Reducing the burden on partner countries (harmonization)		
(Mutual accountability)		Knowledge management	Transparency and accountability	Transparency and accountability	Transparency and learning		
			Likelihood of positive change				
Criteria below relate to measures of results							
Effectiveness	Impact	Demonstrating progress toward results	Contribution to results	Delivering in line with mandate		Objectives achievement (including cross-cutting)	Effectiveness
			Cross-cutting issues				
							Cost effectiveness
Efficiency						Efficiency (includes effectiveness)	Efficiency
Impact	Impact					Impact	
Sustainability						Sustainability	

Table 2 shows a subtler picture. Reading down the columns shows the main criteria for each comparative assessment. Reading across the rows shows whether a given DAC evaluation criterion – or, in parentheses, principle of the Paris Declaration – is covered by a given comparative assessment.

There is, of course, substantial overlap among criteria. In some cases different words are used to describe similar meanings. Nonetheless, there are several areas of no or quite limited coverage of important issues. It is difficult to sustain the argument that these differences in depth and quality of coverage are simply a rational response to differing circumstances. **These weaknesses reinforce the argument made in "Learning Lessons from Comprehensive Evaluations" and the GRPP evaluation for development of agreed guidance on good practice in comprehensive evaluations. Such guidelines should be prepared not as a "cookie cutter" common framework, but to be drawn on and adapted to meet the needs of differing groups of MOs and of specific comprehensive evaluations. The guidelines should aim to inform and advise those commissioning and preparing TORs for comprehensive evaluations, as well as those carrying them out.**[1] **Such guidelines (and collaborative work on them) would be useful in informing comparative assessments as well.**

1.3 Issues for further analysis and action

The main reasons for comparison among MOs are to allocate funds, draw lessons for the improved functioning of individual MOs, and to allocate funds and roles among them. But there are very difficult conceptual and methodological issues, discussed below, in making such comparisons. The most difficult set of issues, relating to the differing mandates of MOs and how to how to compare value for money among them, will be treated last, so that other important issues do not get submerged.

- *Governance:* The importance of governance in determining performance is stressed, regarding global and regional partnerships, in the Sourcebook. The GRPP evaluation (World Bank 2011) and Balogun (2011) find it so important that they add it as a criterion to the traditional five DAC evaluation criteria. The Sourcebook, usefully, takes the OECD Principles of Corporate Governance as a basis for its assessments here. A few of the comprehensive evaluations, including that of the CGIAR and Global Fund, have made it a point of major emphasis. **However, coverage of governance is generally weak and uneven in comprehensive evaluations and even more so in comparative assessments. (See Table 1 re extent of coverage.) This should be addressed in the proposed guidelines. The same set of governance issues identified in the GRPP evaluation – e.g. relating to strategies, priorities and accountability – in reality applies to MOs as a whole.**

  One of the governance issues that merit more attention is the use of earmarked contributions – trust funds – by donors. It was raised as a significant issue in less than half of the comprehensive evaluations. But anecdotal evidence suggests it

[1] It is not feasible to determine whether (the extent to which) areas are neglected for reasons of substance, politics, cost, or lack of knowledge of relevant good practice. The availability of guidelines would at a minimum both provide information on good practice and give support to those involved in commissioning and implementing evaluations who seek to assure that the right questions are asked and in a way (?) likely to lead to useful results.

applies more widely (although not necessarily raised in TORs) and it was emphasized in the GRPP evaluation. The principles of the Paris Declaration – particularly of donor alignment and harmonization – apply here. **Earmarked financing is the equivalent of project financing. It has value in some circumstances. But overall effectiveness of MOs is increased if donors and other constituents debate MO priorities explicitly in the context of governance and priority setting and then support with pooled funding the overall program that ensues.** In effect, the same principles that donors have agreed in the Paris Declaration apply also to their funding of MOs. The DAC Secretariat and the CGIAR have emphasized this and have succeeded in substantially raising the share of core vs. trust-fund support. In the case of the DAC Secretariat, donors at large were faced with the inconsistency of their behavior with what they had just adopted as principles of the Paris Declaration. They adopted a pragmatic solution – trust funds ("voluntary contributions") for the program of the DAC Secretariat at large or for its major subsidiary groups – with the proviso that funds could be shifted among them. The issue of earmarked funding should become a standard part of comprehensive evaluations. (It would be more difficult to use in comparative assessments, where a high share of earmarked funding could be a sign either of donors' seeking to impose their own individual priorities or of lack of trust in the MO; there would need to be a supplementary variable to distinguish between the two.)

- *Principles of the Paris Declaration.* **Coverage of the agreed main principles of the Paris Declaration is quite uneven, as shown in Table 3.** This applies particularly to comprehensive evaluations, where only MfDR is covered in more than half the cases (although the percentage is increasing over time). Paris Declaration principles are by no means a sufficient measure of aid effectiveness. For example, they deal only with process rather than with results. And as the evaluations of the Paris Declaration have confirmed, implementing them has costs. But they arose from the often-painful experience of what happens, in terms of sustainable results, when they are not followed. If there is no country ownership and if donors are financing "hothouse" projects that do not support country ownership and national systems, for example, there is little hope of sustainable results. In addition, donors to MOs and most MOs themselves have committed to them in the Paris/Accra/Busan process. **This is a serious issue, readily addressable in guidelines.**

Table 3 Application of Paris Declaration Principles

	Country Ownership	Harmonization	Alignment	Managing for Results	Mutual accountability
MOPAN	Yes	Yes	Yes	Yes	Yes
COMPAS (not intended to cover PD)	No	No (at country level)	No	Yes	No
QUODA	Yes	Yes	Yes	No (tried but no indicator available)	No
DFID MAR	Yes	Yes (Partnership behavior	Yes (Partnership behavior)	Yes	Yes (Accountability)
AUSTRALIAN AMA	Yes	Yes (Partnership behavior	Yes (Partnership behavior)	Yes	Yes (Accountability)
Pilot Phase MI	Yes	No (in criteria). Partly (cases).	Partly (national goals not systems)	Yes	No
2nd Evaluation of Paris Declaration	Yes	Yes	Yes	Yes	Yes
Sourcebook GRPP	Yes	Limited	Limited	Yes	No
WB Evaluation GRPP 2010	Limited	Limited	No	Yes	No
Comprehensive Evaluations "Yes" (Green light)	8/17	5/17	7/17	11/17	7/17

(Format and coverage of comprehensive evaluations from Balogun, 2012)

- *Incentives.* **There has also been inadequate attention to internal incentives as a driver of improved performance,** whether of the MOs or of the donors interacting with them. For example, incentives are not covered explicitly in MOPAN questions. The importance of internal incentives as a driver was brought out by the Working Party on Aid Effectiveness (2008) and stressed in the Accra Agenda for Action.[2] Interestingly, the Sourcebook deals extensively with incentives as they apply to the role of donors in governance of MOs but does not deal with internal incentives as a driver of performance of either MOs or their bilateral (or other) stakeholders.

- *Results and MfDR.* There is increasingly strong emphasis among donors, as noted, on results. So assessing results of MOs is crucial, as is doing so jointly – both to reduce fragmentation and to facilitate donor focus on "contribution" rather than "attribution." There is a strong tendency, though, in most comparative assessments to focus on systems for managing for results rather than on achievement of results. (MOPAN, COMPAS, CIDA-EVALNET, MAR, AMA, 2nd Evaluation of the Paris Declaration.) The reasons are understandable. Results are harder to measure, particularly further down the results chain (as was learned at high cost, for example, in the Global Fund evaluation). Sustainability and continuing improvement of results, which are linked in part to strengthening national systems and capacity, are that much harder to measure. And there is clearly a link between managing for results and achieving the results. The obvious problem is that the two are not the same and we do not know how high

[2] The inconsistency of donor incentives and behaviour regarding global programs is brought out in Isenman and Shakow (2010).

the correlation is. MfDR systems may look good on paper but not be taken seriously, and even if they are, they are likely to be less important in determining results than, for example, quality of staff or comparative advantage of the MO.

There are, regrettably, no easy answers to how, particularly at reasonable cost, to give appropriate weight to results systems, results themselves throughout the results chain, and sustainability. But this set of issues merits close attention in potential guidelines for evaluations of MOs, as for aid effectiveness in general.

- *GPGs, MDGs, and Vertical Funds*: GPGs, MDGs, and vertical funds are by no means synonymous. But GPG, MDGs, and vertical funds have one important thing in common – focusing on a specific rather than general mandate. Comparative assessments tend to have a bias toward such verticality (see below). There is widespread concern about not giving sufficient attention to GPGs in aid and policies and in evaluation of MOs. This is justifiable, given that GPGs are partly defined by having a free rider problem (non-excludability as well as non-rivalry) and so are underprovided by markets. Similarly, MDGs represent global agreements on areas that need global priority. Most major vertical funds are in areas that fit (to varying extents) as GPGs or MDGs, usually both. The Sourcebook, following the report of the International Task Force on GPGs, has defined GPGs in terms of high value added from global collective action, in effect including MDGs.

  So the issue is not whether GPGs (using the Sourcebook definition) are important. **Rather the key issue is how much extra funding MOs that cover GPGs (whether as vertical programs or included in broader national programs) should get.** To answer this question requires estimating the extent to which a given program or objective counts as a GPG. What is needed is, in effect, the GPG equivalent of what is referred to as the "grant element" (percentage) and "grant equivalent" (amount) of concessionality of aid. But how does one divide the benefits from preventing or curing tuberculosis or from primary education among global public goods, national public goods, and private goods? And how much extra should be allocated for universal primary education, an MDG, over secondary education, which is not an MDG? **As the debate over MDG benefits (primarily mobilization of global support and focus) and costs (primarily distortion and imbalance of country priorities) shows, it is important to avoid simplistic solutions. As Adrian Wood has put it: "MDGs should be taken seriously but not literally." In addition, donor priorities are at times fickle – with a current emphasis on growth and infrastructure relative to the poverty and human development issues that figure importantly in the MDGs.**

  The problem is particularly difficult for allocations. The same donors who feel strongly about priority to GPGs feel just as strongly, if usually on different occasions, about the principles of the Paris Declaration, with their focus on "putting the country in the driver's seat" and not imposing donor choices on developing countries. Yet available comparative assessments that deal with allocations or rankings favor the same vertical funds that donors criticize for

imposing distortions and for fragmenting the "aid architecture." **The AMA admits this bias in favor of specific mandates, stating frankly: "the methodology of assessing organizations against their mandates favors small and specialist organizations." There are similar problems with the MAR and QUODA.[3] These problems should be addressed in joint donor work on MOs as well as in QUODA**. Interestingly, only three bilateral donors reported to the DAC Secretariat that they gave extra credit for GPGs or MDGs in their allocations to MOs

- *MOs with normative mandates:* There are two related problems in dealing with MOs with normative mandates. One is the difficulty in comparing normative mandates – e.g. human rights vs. health or agricultural safety. Weighing them is more an issue of judgment than evidence. The other problem is comparing normative and operational mandates. One might expect that the bias toward specific mandates would extend to these heavily normative MOs. Rather, the opposite is the case, because allocation criteria focus on how aid is delivered at country level. For example, the MAR says: "We therefore clarified that the critical role criterion extended to such (normative) roles, but still required evidence of country-level impact – for example, a multilateral organization involved in setting norms and standards might be helping developing countries to draw up sectoral strategies based on this work." Some field involvement is useful for doing good norm-setting. **But specific interventions at the country level (which raise issues of encouraging fragmentation) are not the right way to judge norm-setting institutions. Comprehensive evaluations can give a more balanced view of MOs with heavily normative functions, as in the case of the FAO.** Comprehensive evaluations can also better address the difficult issue of path dependence – where results of past political decisions (e.g. on governance) have led to inefficiencies that are difficult and take a long time to correct but where mandates remain high priority.

  Fragmentation: **Reducing proliferation should be a key objective of joint donor work on MOs.** The main purpose of establishing MOPAN in 2002 was to curb the proliferation of separate donor assessments of individual MOs. Sixteen DAC donors are now members of MOPAN. MOPAN has plans to do so for its members**.** There are three categories of individual donor reviews to be considered. One is comparative assessments such as the MAR and AMA, which involve direct contact with MOs. What is clear from the 2011 and 2012 DAC Report on Multilateral Aid is that there have been increasing and fragmented donor efforts to do comparative assessments across MOs. This is inefficient in terms of time spent by donors and MOs and of missed opportunities to share information. It is also inconsistent with the donor commitment at Busan to "improve the coherence of our policies on multilateral institutions."

  Another category of donor reviews is that of individual evaluations covering programs of interest to the donor in one or more MOs. The DAC EVALNET database (DEREC) shows very few formal evaluations, but this list is likely to be

[3] In the MAR, MOs that focus on specific MDGs get high marks, by definition, on "critical role in meeting international development and humanitarian objectives," one of only two components of the high-level criterion of "contribution to UK development objectives." Similarly, QUODA has two of its eight indicators for the high-level criterion of "maximizing efficiency" as "focus/specialization by sector" and "support of select global public good facilities."

incomplete, since it is based on voluntary reporting. The third category is that of less formal studies, which are apparently still numerous and add up to a high opportunity cost to MOs. It would be useful to monitor over time the extent of all three categories of fragmentation using simple surveys of MOs and/or donors. (MOPAN is considering doing so, at least in part, for its donor members.)

- *Comparative efficiency and allocations:* "Shareholders and donors are increasingly commissioning … (comprehensive) evaluations to assist decision making regarding resource allocation and funding commitments" for MOs (van den Berg, 2011). Yet with what is produced by comprehensive evaluations, "It is not possible for donors to reach coherent and valid conclusions that would facilitate effective resource allocation…" The same point of not shedding much light on allocations applies strongly to comparative assessments as shown in the last three columns of Table 1– effectiveness, cost-effectiveness, and efficiency.

  Table 1 uses the terminology suggested in the World Bank Sourcebook, which is in turn based on the "DAC Glossary of Evaluation and RBM Terms": effectiveness (or, less ambiguously, efficacy) refers to the extent to which objectives were accomplished; cost-effectiveness refers to the cost of accomplishing them relative to other ways of doing so; and efficiency – roughly synonymous with "value for money" – refers to overall return in comparison with other uses of funds. Unfortunately, the same terms are used for different meanings. [4]

  What are important are the concepts rather than the terms. Making coherent funding decisions across MOs requires a view of value for money across them. It is very difficult but very important to measure, as it is, for example, in efforts by donor and developing countries to allocate funds across different domestic priorities. Even comparison among MOs with substantially overlapping objectives (as to some extent in health) requires a view of comparative cost-effectiveness at least among them.

  Both the MAR and AMA suffer from the difficulty, noted above, of being obliged to use organizational effectiveness, including systems for MfDR, as proxies for results (Faint and Johnson, 2010). And both use their overall ratings as a proxy for efficiency. However, they readily admit the arbitrariness of the weighting systems in their overall ratings and the need for judgment in the choices they have made. They also deal straightforwardly with how they used judgment to take account of differing mandates of MOs as well as of their own national priorities. All these then enter into their overall ratings, which serve as a proxy for efficiency (from the point of view of that donor). There are two other issues, however, that are harder to deal with.

[4] For example, the University of Cambridge website uses two of the same terms very differently (http://www.admin.cam.ac.uk/offices/secretariat/vfm/guide.html): "The definition of the three Es approved by the Value for Money Committee is as follows:

- **Economy** - careful use of resources to save expense, time, or effort.
- **Efficiency** - delivering the same level of service for less cost, time, or effort.
- **Effectiveness** - delivering a better service or getting a better return for the same amount of expense, time, or effort."

The DAC EVALNET website now uses efficiency in the sense recommended by the University of Cambridge rather than that of the DAC Glossary.

- *Taking account of scale:* Let us assume the best feasible "league tables" of comparative efficiency. The next step in the logical chain to determine allocations, whether zero-based or changes from current levels, is to consider the scale of operations (overall budget size) at which a given MO should be operating. To see the importance of scale, consider what difference it would make to the comparative rating of a given organization in QUODA or the MAR if its overall budget were 1/3 bigger or smaller. The answer is: not much. It would still get roughly the same rating and the same up or down signal. The AMA is admirably frank on this. Although it says that a key objective is "to inform decisions on funding allocations," it goes on to mention "the organization's need for additional funding and its capacity to absorb it (which is not assessed in the AMA)."

- *How bad were past allocations?* Then there is a further difficult problem of how to tell the extent to which current actual allocations by a given donor, whether done by present or previous managers, already take implicit account of roughly the same criteria that go into the ratings. "High" is not the same as "more." Approaches such as those of the MAR were designed to be evolutionary – to help decision making by making implicit criteria and choices more explicit, not to dismiss past implicit criteria as wrong or irrelevant.

The point here is not to criticize either the MAR or the AMA, since work over several years suggests that there is no apparent evidence-based answer to these questions.[5] Rather, it is that allocations inevitably involve a good deal of judgment (even leaving aside political or overall budgetary considerations).

- Although individual donors have conducted their own (i.e., fragmented) perception surveys related to aspects of comparative efficiency, they have not used their joint perceptions survey, MOPAN, for this purpose. They have not even, until 2012, assessed perceptions of results achieved (i.e., effectiveness, or efficacy).

- Inevitably, decisions have to use "triangulation" or approximation from different points of view. These include the extent to which current explicit criteria are likely to be different from past implicit ones, the extent to which past allocations seemed generous (or not), and changes in the past year. And then judgment is required as to which MOs should get increases and which should get decreases – and by how much. [6] Adjusting burden-sharing within multilaterally agreed targets, given fungibility, accomplishes little in comparison with reaching agreement with other donors on overall changes in funding or on major reforms that would increase efficiency. Over time, the AMA's proposed annual scorecards, with emphasis on changes rather than just on levels, would provide a better basis for marginal changes in allocations. (MO proposed allocations would need to be on a three-year

[5] See T. Faint and D. Johnson, Multilateral resource allocation: best practice approaches. ODI 2010. http://www.odi.org.uk/resources/docs/6107.pdf.
[6] "Crucially, we also used our best judgment to draw these assessments together into a single evidence-based assessment of performance against the component as a whole." DFID MAR p.12

rolling basis, given the lumpiness of individual, usually three-year, replenishments.)

× Each donor will make its own funding and related policy decisions, taking account of its own national priorities. These may include, for example, regional security issues as well as relative priority of operational or normative mandates. It is no small task, for example, to try to quantify the importance of human rights vs. food safety vs. delivery of a given service. But there are several arguments for joint work leading up to the stage of sovereign decisions: agreeing on replenishments; pooling of knowledge and evidence; increasing consistency for each donor and across donors in being explicit about assumptions; and the Busan commitment to increase coherence and reduce fragmentation.

1.4 Developing good practice and guidelines

The World Bank Sourcebook, although written for evaluations of global programs and partnerships, provides a useful basis for developing guidelines for MOs in general. It would be a challenge to find major points that do not apply to most MOs – including intergovernmental MOs. Adaptations would have to be made in applying guidelines to each type of MO or individual MO, as is already the case in use of the Sourcebook to evaluate global programs.

The World Bank Independent Evaluation Group indicates [7] that, to follow up and update the Sourcebook in the light of the 2011 GRPP Evaluation, it will have draft guidelines (a "guidebook") on evaluation of GRPPs available for review later this year. **It would be useful for the Learning Lessons initiative to engage with the World Bank to see how best to move from those guidelines to those for MOs in general. It would also be useful to have a summary version, including summary criteria, to help increase its impact.**

1.5 Recommendations for improving comparative assessments

- *Bilateral Assessments:* The effort in the past year of bilateral donors to engage on multilateral assessment and policy is encouraging. This effort included a high-level meeting in London in February 2012, which began work to respond to the statement in the Busan outcome document: "We will improve the coherence of our policies on multilateral institutions." This goes well beyond allocations to deal with the full range of issues raised in depth in comprehensive evaluations and drawn on in donor comparative assessments.

The MAR and AMA provide points of reference for joint engagement (as will other bilateral approaches to be surveyed in the MRA). The future work of the donor group will be informed by the forthcoming 2012 DAC Report on Multilateral Aid, which will survey and synthesize a wide range of other bilateral assessments. **EVALNET should consider offering support to this effort.**

[7] Personal communication, Chris Gerrard, May 2012.

Given the importance of the allocations issue for donors, the donor group should put some emphasis on how to treat the challenging issues of cost effectiveness and efficiency – taking account of the issues raised above on scale, verticality, and normative MOs. This does not necessarily mean jointly identifying the multilateral equivalent of "donor orphans" and "darlings", although that would permit more effective joint actions, but it does mean joint analysis on how to identify them.

- *QUODA:* QUODA provides a strong basis for comparison of behavior of donors and MOs on an expanded set of aid effectiveness indicators, going beyond those of the Paris Declaration. It comes from the strongly results-oriented CGD and Brookings Institution but does not include indicators of results (rather than of managing for results), since it says it cannot find appropriate indicators. This is unfortunate but understandable**. Like the MAR and AMA it provides implicit excessive weight to verticality. This should be corrected.** QUODA would increase its impact on the development community if it added a composite index, with whatever weighting of its four major components that Brookings and CGD decide.[8] Future versions of QUODA will depend on when there are data available from monitoring surveys of the Paris Declaration.

This report does not analyze the promising Pilot Donor Transparency Index, since it is incorporated into QUODA.[9] However, transparency continues to get increasing priority for development in general and for donors, particularly with the International Aid Transparency Index (IATI), which had 27 bilateral donors and MOs as signatories as of April 2012. **The Pilot Donor Transparency Index has a highly user-friendly interface, with a "Play with the data" section that permits each viewer to set weights or drop indicators in order to determine a personalized aggregate index. This good practice would be a useful addition to QUODA, as it would be for other comparative assessments.[10]**

- *CIDA-EVALNET APPROACH:* The CIDA-EVALNET (Pilot Test) approach is a complement to and not a substitute for other approaches to comparing MOs. It aims to draw conclusions about the effectiveness of an MO from analysis of a relatively large set (where feasible) of evaluations at country level. The CIDA-EVALNET approach is fairly rapid (taking well under a year for each set of MOs) and moderate in cost; CIDA estimates costs at about $125,000 per MO with low draw on time of MOs or others. It is particularly complementary to MOPAN.

The approach has limitations on use of its results, however. Its comparisons among MOs cannot be precise, given uncertainties of comparability of: ambition of projects/programs being evaluated; rigor and severity of evaluation across MOs; and rigor and severity among those carrying out CIDA-EVALNET studies of different MOs. It deals with policies and decisions taken by the MO a number of years

[8] Although QUODA chooses not to have a composite index, there is in fact one, equally weighted, that can be read from the Y axis of its Figure 1.
[9] Similarly, this report does not analyze "Aid Quality and Donor Rankings," by S. Knack, F. Rogers, and N. Eubank, World Bank Policy Research Working Paper 5290. The reason is, again, that it has been integrated into QUODA.
[10] As of April 2012, Brookings and CGD had the intention to add such a feature to QUODA.

previously. And it is not intended to deal with questions – highly relevant to comparison of MOs – that cannot be gleaned from results of evaluations. These include whether there was a preferable set of activities that could have been undertaken and, more broadly, questions of mandate, comparative advantage, and governance. One weakness that could readily be dealt with in its guidelines is strengthening coverage of the principles of the Paris Declaration on harmonization, alignment, and mutual accountability (See Table 3, above). This would increase utility for donor assessment of MOs, given donor commitment to those principles.

The impact of the CIDA-EVALNET approach would be increased if it were to become more closely linked to MOPAN. This would strengthen existing links, which now include choosing MOs for the CIDA-EVALNET approach from those that MOPAN plans to address. The issues they consider overlap considerably– particularly given MOPAN's recent interest in beginning to cover development effectiveness. With closer links between the two, there could be greater exploitation of complementarity between the CIDA-EVALNET's focus on evidence from evaluations and MOPAN's focus on perception surveys and document review. The CIDA-EVALNET approach could conceivably become a third element in the MOPAN "common approach" (perception surveys and document review). This could be done either with complete integration or with separate governance with evaluators continuing to take responsibility for the CIDA-EVALNET approach.

The future of this approach, which produced its initial assessments (after its "Pilot Test" phase) in May 2012, will depend on the extent to which it is found useful by donors in comparing MOs, whether as an input to MOPAN or on its own. However, it also has potential two-way complementarity with the DAC-UNEG peer reviews of the evaluation function of UN organizations, since the strongest element of the CIDA-EVALNET approach is assessing the functioning of evaluation systems.[11]

- *COMPAS:* It has only limited potential as a source of comparisons among MDBs, although it provides a comparable set of questions across them. This is because: it is self-evaluation, with its value-added depending primarily on what use its MDG members make of it, which the evaluation of COMPAS says is limited; the evaluation raises a set of serious technical problems raised in the evaluation of COMPAS; and there is a strongly held view of MDBs that COMPAS should not be used for comparative purposes other than exchange of good practice among themselves.
- *MOPAN:* MOPAN, with 16 donor members, provides tailored assessments based on a common methodology of a rotating series of MOs. Although its website says: "It is not possible to compare multilateral organizations to one another…," it is the source of comparative information on MOs by far the most cited by donors. (It was cited as a source by 14 out of 23 donors responding to the survey for the 2011 RMA.) It is not that donors do not recognize that the MOPAN methodology and its MO-by-MO assessments are not designed for

[11] This report does not cover the peer review mechanism of the evaluation function of United Nations organizations carried out by the DAC Evaluation Network and the United Nations Evaluation Group. As its website says, however: "A peer review of the evaluation functions of an organization is not in itself an assessment of the effectiveness of that organization. However, it can contribute to the basis for assessing the effectiveness of the organization, by testing the capacity and quality of the organization's own evaluations of effectiveness, and thus the confidence that can be placed in them."

comparisons or that it focuses on organizational rather than development effectiveness. It is rather that donors recognize the need for objective evidence to inform decisions on allocations and other inter-MO policies, and there are so few other sources available.

The MOPAN secretariat, now rotating, is now to be hosted by the DAC Secretariat. This should facilitate complementarity with the new donor effort on multilateral policy as well as with EVALNET. MOPAN will maintain its separate governance structure and its ownership of the assessment methodology and results. It is encouraging to see inclusion of a global program (GAVI) as well as piloting efforts at covering development effectiveness through inclusion of questions on achievement of results in four MOs. There is an evaluation of MOPAN scheduled for later in 2012. **It would be useful for the evaluation to include the following:**

- Whether the methodology of its perception surveys could be strengthened. Perception surveys have limitations. But they can add value in shedding light on key issues: "when key issues are multi-dimensional, there is a mix of both qualitative and quantitative data, and it is not possible to calculate a simple sum of the data points." (MOPAN Common Approach Methodology, March 2012.)

- Whether other aspects of the methodology could be strengthened. (The MOPAN "common approach" uses document review as well as the perception surveys and, as of 2012, consultations with the MO's staff in an effort at triangulation to improve reliability of results.) It might be useful, for example, to compare the (standard deviation of) MOPAN ratings among MOs – both those of the perception surveys and document review – with those of comparable aggregate indicators of the MAR and AMA (and others that become available). This would test whether the current approach – including the reluctance to compare one MO to others – tends to produce relative generosity in ratings of weaker MOs.

- Reducing total costs, including streamlining the questionnaire to sharpen the focus and reduce the opportunity cost of time spent at country level.

- Addressing the contentious issue of whether MOPAN should explicitly include comparison among MOs in its objectives, whether aimed at informing donor decisions on allocations, facilitating benchmarking, or improving the reliability of individual ratings. Comparisons could either be of specific indicators or of aggregate indicators. In either case, each donor would be able to integrate the results as part of its own overall decisions on allocations and policy, including taking account of national priorities.

- Whether and how to take account of the CIDA-EVALNET approach in the MOPAN "common approach."

- Whether coverage of cost-effectiveness should become standard and should be extended to cover efficiency and adequacy of financing as well. This would be one partial contribution to the need for "triangulation" to address these extremely difficult issues. The AMA has suggested significantly expanding the number of annual MOPAN assessments to assist in comparability. An

alternative would be for MOPAN to do a "light" version every two or three years across a much larger group of MOs – with adjustments so that it is specifically aimed at making comparisons.

In sum, there are important opportunities for improved quality and consistency, as well as for more collective action and reduced fragmentation of both comparative assessments and comprehensive evaluations of MOs. These are not ends in themselves but need to be designed and implemented so as to maximize their contributions to increased efficiency of MOs and improved architecture and allocation of funds among them. The focus in this report, as in MOPAN and "Learning Lessons", is on the perspective of donors – given their importance particularly on questions of financing. The current donor approach to comparative assessment of MOs is not a system at all, but is composed of a series of relatively uncoordinated and fragmented series of joint and individual efforts where the whole is less than the sum of its parts. Donors should in their current efforts to work together on MOs accept increased – and monitored – accountability for meeting the Busan commitment to reduce fragmentation and "improve the coherence of our policies on multilateral institutions."

Appendix

Comparative assessments for which criteria are available and assessments analyzed in this report[12]

(Asterisks indicate those analyzed for this report.)

1 *Australian Multilateral Assessment 2012

2 *Common Performance Assessment System of Multilateral Development Banks (COMPAS) 2009-2010

3 *DAC Report on Multilateral Aid 2011

4 Denmark Assessment of Multilateral Organizations 2007

5 European Commission evaluation on the partnership EU-UN 1999-2006

6 Heavily Indebted Poor Countries Capacity Building Program (HICP CBP) Partner Country Evaluation of Multilateral Institutions 2009

7 *CIDA-EVALNET approach (Pilot Test) 2010

8 *Multilateral Organization Performance Assessment Network (MOPAN) 2012

9 Netherlands Multilateral Monitoring Survey and Scorecard

10 Norway's evaluation of 29 multilateral organizations (2011)

11 ODI 2009 survey of partner country perceptions of ME, Multilateral Effectiveness 2009

12 Pathways to Accountability II Framework/One World Trust 2011

13 Pilot Aid Transparency Index/Publish What You Fund 2011

14 *Quality of Official Development Aid (QUODA) 2012

15 Review of Effectiveness of the CIDA Multilateral Channel 2009

16 *Second Evaluation of the Paris Declaration 2011

17 *Sourcebook for Evaluating Global and Regional Partnerships (World Bank IEG) 2007

18 Sweden's evaluations of multilateral organizations 2008-2011

19 *UK 2011 Multilateral Aid Review (MAR)

20 *World Bank's Involvement in Global and Regional Partnership Programs (2010)

[12] This list draws heavily on "Methodology Fact Sheets" kindly provided by the Evaluation Department of the Finnish Ministry of Foreign Affairs. Other bilateral assessment methodologies are being submitted to the DAC Secretariat for its 2012 Report on Multilateral Aid. This appendix and the report do not cover the peer review mechanism of the evaluation function of UN organizations carried out by the DAC Evaluation Network and the United Nations Evaluation Group.

Bibiliography

AusAid (2012) "Australian Multilateral Assessment."

Baastel (Groupe Baastel-Conseil Ltée) (2012) "Case study: Impact of the Fourth Overall Performance Study of the GEF."

Balogun, Paul, (2011) "Comprehensive Evaluations – Note 1."

Balogun, Paul, (2102) "Comprehensive Evaluations – Note 2."

Bezanson, Keith and Paul Isenman (Center for Global Development (2012, forthcoming) "Improving the Governance of Global Multi-Stakeholder Partnerships for Development: Challenges, and Lessons Learned."

Buse, Kent and Sonja Tanaka (2011) "Global Public-Private Health Partnerships: Lessons learned from ten years of experience and evaluation," *International Dental Journal*, 2011: 61 (Suppl. 2) pp 2-10.

COWI, "Independent evaluation of the Cities Alliance: Final Report," COWI, April, 2012.

Fast Track Initiative (2010), Mid-term evaluation of the EFA Fast Track Initiative, Final Synthesis Report, Volume 1 – Main Report.

DFID (2011) "Multilateral Aid Review."

Fourth High Level Meeting on Aid Effectiveness, "Busan Partnership Agreement," 2011.

Gerrard, Chris, Rolf Korte, and Elaine Ooi (2012) "Case Study on the Five-Year Evaluation of the Global Fund to Fight AIDS, Tuberculosis and Malaria."

Isenman, Paul and Alex Shakow (2010) "Donor Schizophrenia and Aid Effectiveness: The role of Global Funds."

ITAD, Ltd., (2012), "A Case Study of the Impact of IFAD's Independent External Evaluation."

Lele, Uma, N. Sadik and A. Simmons (2007) "The Changing Aid Architecture: Can Global Initiatives Eradicate Poverty?", http://www.oecd.org/dataoecd/60/54/37034781.pdf

Markie, John (2012) "Case Study of the Effectiveness of the Independent External Evaluation (IEE) of FAO."

OECD (2009-2012) "DAC Report on Multilateral Aid."

OECD Working Party on Aid Effectiveness (2009) "Improving Incentives in Donor Agencies (First Edition) Good Practice and Self-Assessment Tool."

Williams, Kevin (2012) "Case Study: UNESCO's Independent External Evaluation (IEE)."

World Bank (2004) "Addressing the Challenges of Globalization: An Independent Evaluation of the World Bank's Approach to Partnerships."

World Bank (2011) (Independent Evaluation Group). "The World Bank's Involvement in Global and Regional Partnership Programs: An Independent Assessment."

World Bank (Independent Evaluation Group) (2007) *Sourcebook for Evaluating Global and Regional Partnership Programs: Indicative Principles and Standards.*

SADEV | P.O. BOX 1902 | SE-651 19 KARLSTAD, SWEDEN | PHONE +46 54 10 37 00 | SADEV@SADEV.SE | WWW.SADEV.SE