Journal of MultiDisciplinary Evaluation
Volume 8, Issue 18

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

An In-Depth International Comparison of Major Donor Agencies: How Do They Systematically Conduct Country Program Evaluation?

Ryo Sasaki
International Development Center of Japan / Rikkyo (Saint Paul's) University, Japan

Background: This paper presents an in-depth international comparison of systems and procedures of aid evaluation, focusing on Country Program Evaluation among major donor agencies. The original client of this study is Ministry of Foreign Affairs, Japan (MOFAJ).

Purpose: The purposes of this paper are set as follows: (1) to understand how aid agencies conduct Country Program Evaluation; and (2) to make recommendations for improvement of the current practice of Country Program Evaluation in the aid evaluation community.

Setting: The examined donors include: the World Bank (WB), the Asian Development Bank (ADB), the Inter-American Development Bank (IADB), the United Nations Development Programme (UNDP), the U.S. (USAID), Canada (CIDA), the U.K. (DFID), the Netherlands (IOB), Germany (BMZ), France (Foreign Ministry), and Japan (Ministry of Foreign Affairs (MOFAJ)). In addition, aid agencies conducting respective project evaluation are also examined, and they are JICA (Japan), GTZ and KfW (Germany) and AFD (France).

Intervention: This study presents the result of comparative analysis among those donor agencies in terms of the following viewpoints: (1) evaluation criteria employed; (2) approaches to evaluate "effectiveness" and "impact"; (3) attribution issue; (4) the use of a rating system; and (5) overall evaluative conclusion and integrating methods. All viewpoints are focusing on Country Program Evaluation. One conclusion is that most agencies have been struggling with how to judge the degree and value of their country programs.

Data Collection and Analysis: Mixed methodologies were employed to collect data from the said donor agencies. The analysis was conducted by a systematic procedure consisting of: (i) summarizing information in a comparative table; (ii) trying to make groups/categories based on common characteristics if possible; and (iii) examining and concluding basic thoughts/philosophy which make their differences.

Findings: This study made some new knowledge about how aid agencies conduct Country Program Evaluation and identified several issues remained. Varieties of their practices are observed and it is far from the unified methods agreed. Some remarkable points identified in this study are:(1) Most aid agencies invoke the DAC five evaluation criteria for Country Program Evaluation. (Major exception was USAID); (2) "Strategic relevance" and "coherence/complementarity" are the emerging new criteria; (3) Attribution is still the issue that aid agencies have struggled; and (4) The attitude for introduction of rating system is clearly divided among aid agencies.

Keywords: *effectiveness; impact; attribution; coherence; country program evaluation*

Introduction

Background of This Study

This study conducted an in-depth international comparison of systems and procedures of aid evaluation, focusing on Country Program Evaluation, among major donors. The original study was conducted by International Development Center of Japan (IDCJ) by request of the Ministry of Foreign Affairs, Japan (MOFAJ) in Fiscal Year 2010.[i] MOFAJ has commissioned a study to review the aid evaluation systems and methodologies of other major donors, to compare the results of the review with the current system in Japan, and to provide any useful input to MOFAJ for revising its ODA Evaluation Guidelines and establishing a new system. The report is available at the Ministry's website for the general public (the main report is in Japanese but intensive summary report is in English). Data collected in the study are used for the comparative analysis in this paper with permission of MOFAJ.

Purposes of International Comparison

The purposes of this paper are set as follows.

1. To understand how aid agencies conduct Country Program Evaluation.
2. To make recommendations for improvement of the current practice of Country Program Evaluation in the aid evaluation community.

Target of This Study

This study focuses on so-called "Country Program Evaluation". It is sometimes called Country Assistance Evaluation (CAE) by the World Bank or other similar names. OECD-DAC (2002) defines Country Program Evaluation/Country Assistance Evaluation as "evaluation of one or more donor's or agency's portfolio of development interventions, and the assistance strategy behind them, in a partner country" (p. 19). Country Program Evaluation is categorized as one type of program evaluation by the definition of OECD-DAC (2002).[ii] An important point is Country Program Evaluation is a new challenge that evaluates a set of interventions as a whole in a certain country, and it is essentially different from the traditional and conventional evaluation, commonly known as project evaluation or project-level evaluation.

Table 1
Definitions by OECD-DAC Glossary

Term	Definitions
Program evaluation	Evaluation of a set of interventions, marshaled to attain specific global, regional, country, or sector development objectives. Note: a development program is a time bound intervention involving multiple activities that may cut across sectors, themes and/or geographic areas. Related term: Country Program evaluation/Country Strategy Evaluation.
Project evaluation	Evaluation of an individual development intervention designed to achieve specific objectives within specified resources and implementation schedules, often within the framework of a broader program. Note: Cost benefit analysis is a major instrument of project evaluation for projects with measurable benefits. When benefits cannot be quantified, cost effectiveness is a suitable approach.

(Source) OECD-DAC. (2002). p. 30-31.

The aid agencies compared in this study are total 15 agencies (see Table 2). Figure 1 shows the location of these aid agencies. It can be stated that the selection of agencies has a good balance because it includes both multinational and bilateral agencies whose locations are North America, Europe, and Asia. This study has one good feature: Japan is included. The past similar studies have included only the donor countries that provide English reports and, as a result, excluded Japanese aid agencies (e.g., Cassen, R. (1994); Stokke, O. (1992)), Those studies had serious information imbalances because they omitted the information on the largest donor (in 1980's) or the second largest donor (in 1990's – early 2000's), which is Japan. In contrast, readers can see well-balanced comparative analysis in this report.

Table 2
Aid Agencies of the Study

Multilateral Donors

- the World Bank (WB)
- the Asian Development Bank (ADB)
- the Inter-American Development Bank (IADB)
- the United Nations Development Programme (UNDP)

Bilateral Donors

- the U.S.: United States Agency for International Development (USAID)
- Canada: Canadian International Development Agency (CIDA)
- the U.K.: Department for International Development (DFID)
- the Netherlands: Ministry of Foreign Affairs
- Germany: (1) Country Program Evaluation: Federal Ministry for Economic Cooperation & Development (BMZ); (2) Technical cooperation: Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ);[iii] (3) Loan: Kreditanstalt für Wiederaufbau (KfW))
- France: (1) Country Program Evaluation: Ministry of Foreign and European Affairs(MOFEA); (2) Technical Cooperation: French Development Agency (AFD)
- Japan: (1) Country Program Evaluation: Ministry of Foreign Affairs (MOFAJ); (2) Technical Cooperation/Loan: Japan International Cooperation Agency (JICA)

Figure 1. Location of Aid Agencies Studied

Methodologies for Data Collection and Analysis

Mixed methodologies were employed to collect data from the said donor agencies. The methodologies for data collection include the following:

- Visited the websites of respective agencies and collected (i) evaluation guidelines and (ii) evaluation reports (of Country Program Evaluation).
- Conducted field interviews to collect more information (the World Bank, IADB, the U.S. (USAID), the U.K. (DFID), Germany (BMZ, GTZ) , and Japan (MOFAJ and JICA).
- Conducted telephone interviews (Canada (CIDA)) and mail interviews (UNDP).

The items examined are as follows. The analysis was conducted by a systematic procedure consisting of: (i) summarizing information in a comparative table; (ii) trying to make groups/categories based on common characteristics if possible; and (iii) examining and concluding basic thoughts/philosophy that distinguish donor agencies.

- Evaluation criteria employed
- Approaches to evaluate "effectiveness" and "impact"
- Attribution issue
- The use of rating system
- Overall evaluative conclusion and integrating methods

In addition, the original study by MOFAJ (IDCJ, 2010) examined more comparative analyses in terms of (i) the independency of evaluation departments; (ii) types and expertise of evaluators assigned; (iii) quality control systems; (iv) the utilization of external committees; and (iv) feedback systems of evaluation conclusions and recommendations. It is recommended to access the original report if you are interested in those additional comparative analyses (As stated, the main report is in Japanese but an intensive summary report is in English).

Types of Evaluation

Table 3 shows the number and types of evaluations conducted by four international agencies and seven evaluation offices at headquarters of bilateral donors targeted in this study. All agencies conduct Country Program Evaluations and sectoral/thematic evaluations. Six agencies conduct impact evaluations, while only three or four agencies conduct regional evaluations.

The average number of evaluations conducted per year is 24 to 25 (The average number is 15 to 17, if project evaluations by the Word Bank and ADB are excluded.), though the number varies year by year for some bilateral agencies. The average number of evaluations for bilateral donor agencies is 12 to 15. The average numbers are slightly different year by year.

Evaluation Criteria Employed for Country Program Evaluation

The results of analysis on evaluation criteria employed by target donors can be summarized in Table 4. This table implies some common features as follows.

- Many agencies employ evaluation criteria similar to so called DAC Evaluation Criteria.[iv] However, those criteria were originally recommended for the evaluation of ODA projects; therefore, they do not fully facilitate Country Program Evaluation. Thus, many agencies have struggled to add new criteria or modify a part of DAC Evaluation Criteria as observed.
- An attempt to replace the term "relevance" with "strategic relevance" and an attempt to add the term "coherence" are widely observed (6 out of 11 agencies employ "coherence" for Country Program Evaluation). It suggests that many agencies try to evaluate their implementation in consideration of so called "selection and concentration."
- Results of assistance can be divided into two categories: direct, short-term effects (effectiveness), and indirect, long-term effects (impact). Relatively more agencies regard the former as being able to evaluate satisfactorily, while the latter as being difficult to evaluate, including the evaluation for

degrees of attribution.

- Regarding "efficiency," no agency conducts evaluation for social costs and social benefits induced by the assistance as a whole because they are difficult to be estimated. Instead, many agencies use simple estimation methods for efficiency, such as a comparison of input and output, and a review of the implementation process.

Although those common features are identified, it is also observed that many of the agencies have struggled to add some new criteria (see "other criteria" section) which would fit Country Program Evaluation.

Table 3
Types of Evaluations Conducted by Evaluation Departments at Headquarters

Agency/Country	Total # of Evaluation *1	Policy-Level <----------> Project-Level				New Trend
		Country Program Evaluation	Regional Evaluation	Sector/ Thematic Evaluation	Project Evaluation	Impact Evaluation
World Bank	90 *2	o		o	o	o
ADB	30~33 *3	o	o	o	o	o
IADB	28	o		o	o	o
UNDP	16	o	o	o	o Decentralized*4	
USAID	5~10	o	o	o	o Decentralized*4	o
CIDA	2~7	o		o	o Decentralized*4	
DFID	23	o	o	o	o Decentralized*4	o
The Netherlands	5~11	o		o	Decentralized*4	
Germany (BMZ)*5	9	o		o	o by GTZ, KfW	o
France MOFEA)*6	28	o		o	o by AFD	
Japan (MOFAJ)	11*7	o		o	o by JICA	

*1 The average number of evaluations conducted per year, or the number of the most recent year. Definitions vary as the following remarks.

*2 Of which 70 are project evaluations.

*3 Of which 13 are project evaluations.

*4 Decentralized evaluations include those conducted by local offices, embassies or implementation offices in headquarters.

*5 Information refers to BMZ that is in charge of Country Program evaluations.

*6 Information refers to both MOFEA and AFD（except project evaluations）.

*7 The average number of evaluations conducted between FY2006 and 2009, and it does not include project evaluation (more than a hundred, which is conducted by JICA).

(Source) Prepared by the Study Team based on OECD (2010), evaluation reports by donor agencies, and interviews in field survey.

Evaluation for Effectiveness and Impact in Country Program Evaluation

As stated, effectiveness and impact are two of five major criteria widely used in project evaluation in the aid evaluation community. How do we understand the relationship between them? The followings are the definitions of those terms.

Now this concept is being tried to apply Country Program Evaluation. However, a simple application to Country Programs seems very difficult. Actually, most donor agencies have been struggling and some seems to have reached the conclusion that it is difficult (and better to be abandoned) to evaluate the second criterion (Impact). Three groups are identified as follows.

- Group 1: Agencies that do not

distinguish between effectiveness and impact and evaluate as a whole. (World Bank, Japan (MOFAJ)).
- Group 2: Agencies that evaluate both effectiveness and impact separately just as project evaluation (ADB, Germany (BMZ)).
- Group 3: Agencies that focus on only effectiveness but not impact due to the difficulty in Country Program Evaluation (CIDA, UNDP, the Netherlands).

It is obvious that one of the remaining big issues in Country Program Evaluation is how to appropriately evaluate impact.

In addition, it is widely observed through this study that the use of the word 'impact' is confused among aid agencies, and thus it is very difficult to understand what other agencies are talking about when they say 'impact'.. At least there are three usage of the word 'impact'. Based on the examination during this study, the following types are proposed with the hope of promoting mutual understanding among aid people.

Attribution Issue in Country Program Evaluation

The issue of attribution is also a hot topic in the current Country Program Evaluation practice. First of all, the following is the general definition of attribution by OECD-DAC. However, it should be admitted that the actual usage of the word is not unified and frequently used interchangeably. OECD-DAC (2002) offers a definition of attribution as follows.

> Attribution: The ascription of a causal link between observed (or expected to be observed) changes and a specific intervention. Note: Attribution refers to that which is to be credited for the observed changes or results achieved. It represents the extent to which observed development effects can be attributed to a specific intervention or to the performance of one or more partner taking account of other interventions, (anticipated or unanticipated) confounding factors, or external shocks. (p.17)

It is observed that there is no unified view or approach about the attribution issue. It is more diverse than the issue of effectiveness and impact (See Table 7). The following is a tentatively proposed categorization. These are made by examination of the existing documents and the field interviews at their headquarters.

- Group 1: UNDP has collected several possible approaches, although the actual application seems very limited. However, since UNDP has not given up the idea of assessing attribution, it should be regarded as the most advanced group.
- Group 2: The World Bank, IADB, DFID and France (MOFEA) suggest only viewpoints of attribution, but actual approaches are not clearly proposed.
- Group 3: ADB, CIDA and Japan (MOFAJ) mention that it is difficult to measure attribution but some alternative approaches can be taken.
- Group 4: It is not so meaningful to think about individual attribution, and it is enough to recognize the result as a shared achievement of all actors. (=> logically, this group results in pursuing common-basket/general-budget support).

It is again obvious that the issue of how to appropriately evaluate attribution is another remaining and emerging issue in Country Program Evaluation.

Table 4
Comparison of Evaluation Criteria (Mainly for Country Evaluation)

Agency	Relevance	Effectiveness	Efficiency	Impact	Sustainability	Coherence	Other criteria
Country Program Evaluation							
World Bank	o	o	(o)	(o)	o		Relevance of Bank's Results Achieved; Institutional Development Impact
ADB	o	o	o	o	o		Country Positioning, Contribution to Development Results. ADB Performance
IADB	o	o	o	o	o	o	
UNDP	o	o	o		o	o	Responsiveness, Promotion of UN values; Strategic Partnership
USAID	o	o	o	(o)	o		(According to the 2009 new guidelines)
CIDA	o	o	o		o	o	Management Principles/Adherence to the Paris Declaration, Cross-cutting Issues; Monitoring and Evaluation
DFID	o	o	o	(o)	o	o	Coverage, Harmonization and Coordination
The Netherlands	o	o	o		(o)		
Germany (BMZ)	o	o	o	o	o	(o)	Complementarity
France (MOFEA)	o	o	o	o	o	o	Outcome
Japan (MOFAJ)	o	o	o	(o)			Efficiency is "Appropriateness of the process"
Project Evaluation							
Germany: GTZ & KfW	o	o	o	o	o	o	
France AFD	o	o	o	o	o		AFD's Additionality
Japan JICA	o	o	o	(o)	o		

Note: "(o)" indicates evaluation using that criteria is actually conducted as a part of evaluation using other criteria. (E.g. In Japan (MOFAJ), Impact is evaluated as a part of Effectiveness, as its Guidelines instructed.
（Source）Documents downloaded from donor websites and field interview results.
Adapted from IDCJ (2010), p. 45, translated into English by the author.

Table 5
Definition of Effectiveness and Impact

Term	Definition
Effectiveness	The extent to which the development intervention's objectives were achieved, or are expected to be achieved, taking into account their relative importance.
Impact	Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended.

(Source) OECD-DAC. (2002). p. 20-21, p. 24.

Table 6
Effectiveness and Impact in Country Program Evaluation

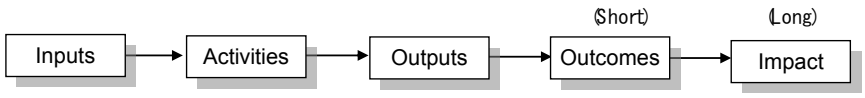
Group	Agency	Explanation
Do not distinguish bewteen effectiveness and impact, and evaluate as a whole.	World Bank	It does not distinguish between effectiveness and impact and evaluates them in one criterion "efficacy".
	Japan, (MOFAJ)	Evaluation of effectiveness includes evaluation of outcomes which includes medium-term and long-term outcomes. Although its guidelines insist that it is very difficult to evaluate outcomes, the actual evaluation reports include evaluation results of outcomes in most cases.
Evaluate effectiveness and impact separately.	ADB	It evaluates effectiveness and impact separately. Effectiveness is the effect directly made by the project, and impact is a longer effect. It is relatively plausible to make clear the cause-effect relationship of effectiveness, but that of a longer effect, which is impact, is difficult to make clear because it is affected by interventions of other donors and change in the environment.
	Germany (BMZ)	BMZ, just as GTZ and KfW, evaluates effectiveness and impact separately. Effectiveness is a direct effect or the effect on the program's objectives, and impact is a diffusive effect or the effect on the overall goal.
Basically do not evaluate impact	UNDP	Effectiveness is evaluated mainly focusing on the outcomes that are set in advance, and it does not include the viewpoint of impact.
	CIDA	Effectiveness in Country Program Evaluation focuses on the relationship between project-level inputs and major outcomes (which is measured quantitatively). Due to the limitations in financial resources for evaluation, impact is usually not deeply examined.
	The Netherlands	It evaluates only effectiveness which is a direct effect on the program's objectives and the degree of contribution to it.

(Source). Documents downloaded from donor websites and field interview results.
Adapted from IDCJ (2010), p. 47, translated into English by the author.

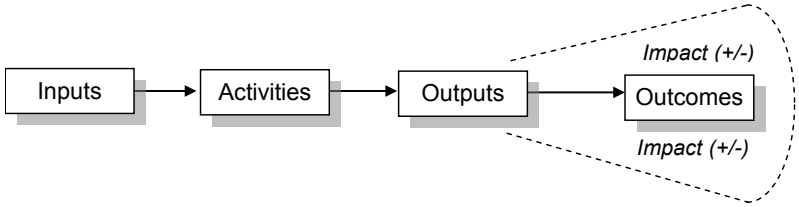
Box 1
Types of 'Impact'

It is observed that donor agencies use the word 'impact' for three different meanings. The following classification is one possible proposal for promoting mutual understanding. This classification is applicable to both Country Program Evaluation and project evaluation.

(1) Type I 'Impact' : Long-term social/economic impact



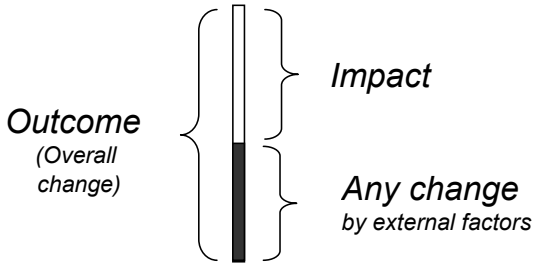
(2) Type II 'Impact' : Indirect impact (both positive and negative, and both intended and unintended)



(3) Type III 'Impact': Impact as pure change made by the intervention

(= Outcome$_{before}$ - Outcome$_{after}$ – any change caused by external factors).

In other word, it is a "RCT-type impact". ("RCT" is randomized controlled trial).



(Source) (1) Davidson, J. (2005). "Impact is often used to refer to long-term outcomes." (p. 241)
(2) OECD-DAC. (1991). "Impact: The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended."
(3) International Initiative of Impact Evaluation (3iE). (December 2011). "Impact: The effect of the intervention on the outcome for the beneficiary population;" "Impact evaluations have either an experimental or quasi-experimental design."

Table 7
Attribution Issue in Country Program Evaluation

Group	Agency	Explanation
		Country Program Evaluation
Propose several possible approaches	UNDP	The Evaluation Office has developed manuals and guidelines, and it holds workshops about this theme. (The guidelines (draft) explain several approaches for assessing attribution)
Only suggest viewpoints of attribution	World Bank	Outcome of the Country Assistance Program is made by the integration of the World Bank, other donors, the host governments and external factors. Thus, degree of attribution of each actor should be measured.
	IADB	The factors that would affect results are: (i) IADB's performance, (ii) the host governments' performance, and (iii) external factors. Although it is difficult to distinguish these three things, descriptive explanation on each factor should be made.
	DFID	Since it is very challenging to make clear the attribution of DFID's intervention, joint evaluation with other partners should be conducted in order to distinguish the attribution of DFID from the attribution of others. Also, attribution is listed as one of the possible evaluation items in Country Program Evaluation, and this suggests DFID's attitude toward evaluating attribution.
	France (MOFEA)	Although MOFEA is interested in identifying its attribution, there is no explanation on how it is actually indentified.
Difficult to measure attribution but some alternative approaches can be taken	ADB	It is impossible to distinguish the attribution of ADB from the attribution of other donors. The only things that are possible are to understand this limitation and to analyze major factors within its limitation.
	CIDA	Since it is difficult to measure its attribution, it regards the ratio of CIDA's aid out of the total aid amount for the host country as its attribution.
	Japan (MOFAJ)	It is very difficult to identify its attribution because various stakeholders, including other donors, multinational agencies, host governments and NGOs, provide their input, and the development results are made out of those various factors. It is generally observed in MOFAJ's country evaluation reports that the ratio of Japan's input out of the total aid for the host country is regarded as the attribution of Japan, like CIDA's approach.
Not so meaningful	Germany (BMZ)	It does not consider the attribution issue because it is very difficult. It is not meaningful at all.
No information available.	USAID	No information is available. It seems it does not pay special attention to this issue.
	The Netherlands	It does not have this criterion.
		Project Evaluation
-	Germany (GTZ, KfW)	Since the "attribution gap" exists between outcomes (= the direct results) and comprehensive development results (= the indirect results), GTZ's result-chain suggests applying (i) before and after comparison and (ii) counterfactual approach.
-	France (AFD)	AFD sets its original criterion, "AFD's additionality", to evaluate attribution,, although how to actually evaluate is not clear.
-	Japan (JICA)	Its guidelines mention a careful examination of attribution. However, the major analysis approach is a simple before-after comparison within its long-term approach.

(Source) Documents downloaded from donor websites and field interview results.
Adapted from IDCJ (2010), p. 50, translated into English by the author.

In the process of reviewing the trials that donor agencies has struggled, several good approaches for evaluating attribution are identified although none of them seems perfect. They are as follows.

- General elimination method (GEM): List all possible factors including own intervention, and then eliminate factors to which are not attributed to development results one by one. If your intervention remains after elimination, it can be judged to a certain degree that your attribution surely exists.
- Application of performance measurement: If the value of indicator goes away from the baseline toward the target value, it is regarded as the assumed cause-effect relationship is appropriate.
- Structured interviews: Ask stakeholders (including governmental people) about (i) the "before" situation and (ii) the "after" situation. Then, ask (iii) how much such difference owes to the intervention.
- Application of the idea of counter-factual: There are various ways of applying the idea of counter-factual to Country Program Evaluation. One example is a comparison with neighboring countries where such interventions have not been applied.
- Joint evaluation: People should stop thinking about the attribution issue. Instead, all parties, including all major donor agencies, host governments and other related stakeholders (e.g., international NGOs), should sit in one table and conduct one single evaluation. Then all parties should agree on the overall "shared" result that all of them have contributed to, and they should stop trying to divide the result for individual advancement.

The Use of a Rating System

The next item to compare is the use of rating system. Recently, rating systems have become popular in project evaluation. On the other hand, the introduction of rating systems into Country Program Evaluation is still not common. Table 8 shows the practice of each donor agency. Also, BOX 2 illustrates the actual application of rating system to Country Program Evaluation.

The donors are divided into four groups according to their views on rating systems (see Table 9). Opinions obtained from interviews include: (i) the introduction of a rating system promotes communication among stakeholders because it is easy to understand (a rating = one word/alphabet/number); and (ii) On the other hand, many donor agencies share their concerns for too much focus on the rating results without considering their background, as well as concerns for the difficulty of utilizing the rating results of Country Program Evaluation.

A rating system makes an evaluation result understandable and communicable because ratings are expressed by very short sentences (e.g., "Highly satisfactory" - "Highly unsatisfactory"), alphabets (e.g., A - E), and numbers (e.g., 4- 1). It is much easier to understand than wordy text explanation. On the other hand, concerns are shared among many donor agencies about focusing on rating results without considering their background, as well as the difficulty in how to utilize the rating results of Country Program evaluations. Some detailed views for the merit of introducing a rating system are as follows.

- A rating system enables people to compare multiple evaluation results. A list of rating results will make people easily conduct comparative analysis. Also people can make pie charts or bar charts using rating results, and, by using numerical rating scales, people can calculate averages and standard deviations of ratings. Such visual and numerical analyses are totally impossible for text explanation.
- A rating system follows the formal evaluation theory. According to the logic of evaluation, evaluation is defined as the determination of merit and worth of things (Scriven,1991; House, 1999; Shadish et al., 1991). For the determination of merit and worth of things, not only "criteria" but also "standards" of merit should be set (Scriven, 1991) (see Figure 2). A rating system serves for the application of this formal theory.

Table 8
Comparison of Rating System

○....introduced; ×... not introduced

Agency	Rating system	Comments
		Country Program Evaluation
World Bank	○ (6 ranks)	"Highly satisfactory", "Satisfactory", "Moderately Satisfactory", "Unsatisfactory", and "Highly unsatisfactory" for each evaluation criterion
ADB	○ (4 ranks)	Different terms are used for each criterion, such as "Highly relevant" - "Irrelevant" for relevancy criterion.
IADB	×	It does not employ a rating system because a descriptive discussion is necessary for evaluation purposes (to provide lessons learned and to increase accountability), and such discussion would be terminated if a rating system is applied.
UNDP	×	UNDP has not introduced a rating system because it is afraid of the possibility of comparison among recipient countries which is not the purpose of its country evaluation. Although UNDP developed a new country evaluation manual which introduces a rating system in 2010, the trial of rating was postponed.
USAID	×	There is no discussion about rating because performance measurement has been applied, which judges the performance of aid based on whether quantitative targets are achieved or not.
CIDA	○ (5 ranks)	CIDA uses a rating with a scale, from "highly satisfactory" to "highly unsatisfactory".
DFID	×	-
The Netherlands	×	-
Germany (BMZ)	×	BMZ does not employ any rating system because it is not suitable for the policies and agenda of BMZ.
France (MOFAE)	×	-
Japan (MOFAJ)	×	MOFAJ does not employ any rating system. But it is now testing a proposed rating system in FY 2011.
		(For reference) Project Evaluation
World Bank	○ (6 ranks)	"Highly satisfactory", "Satisfactory", "Moderately Satisfactory", "Unsatisfactory", and "Highly unsatisfactory" for each evaluation criterion.
ADB	○ (4 ranks)	Different terms are used for each criterion, such as "Highly relevant" - "Irrelevant" for relevancy criterion.
Germany (GTZ, KfW)	○ (6 ranks) ○ (6 ranks)	GTZ： "Very good/Better than expected" to "No good/Situation worsened". KfW：6 ranks ("Very good/Better than expected" to "Completely failed") for relevance, effectiveness, efficiency and impact. 4 ranks for sustainability.
Japan (JICA)	○ (4 ranks)	JICA: Different rating systems have been unified among three aid schemes (Loan, grant, and technical cooperation) in 2009 with a scale from "A (Highly satisfactory)" to "D (Unsatisfactory).

(Source) Documents downloaded from donor websites and field interview results.
Adapted from IDCJ (2010), p. 53, translated into English by the author.

Box 2
Example of ratings in Country Evaluation

The WB's Bank Program Outcome Ratings
Country Assistance Evaluation (1999-2006)

The WB's Independent Evaluation Group (IEG) rated the outcomes of the WB's Country Assistance Program for Cambodia based on its objectives. It is different from both rating Cambodia's development levels and measuring the performance of the WB and the host government. The main question is: To what degree the goals of the WB's program have actually been achieved? Then ratings are assigned to each Strategic Goal of the WB's Country Assistance Strategy. The overall evaluation was "Moderately Satisfactory".

Summary of ratings of the WB's county assistance evaluation (Cambodia)

The WB's Strategic Goal	Rating of the result of WB's program
1. Macroeconomic stability, economic growth, and poverty reduction	Moderately Satisfactory
2. Improvement of social service delivery	Satisfactory
3. Agriculture, rural development, and natural resource development	Moderately Satisfactory
4. Infrastructure recovery, reconstruction and increase in support.	Satisfactory
5. Reform of public administration	Unsatisfactory
Overall rating	Moderately Satisfactory

（Source）World Bank. (2007). *Cambodia: An IEG Country Assistance Evaluation 1999-2006.*
Adapted from IDCJ (2010), p. 54, translated into English by the author.

On the other hand, demerits (constraints) are listed as follows. It should be stated those demerits (constraints) are very severe.

- A descriptive evaluative conclusion includes rich information on the characteristics and backgrounds of respective policy/program/project/intervention that is never the same as others. Although there is no identical program in this world, rating inherently omits this information. There is a risk that people will fail to look such information when a rating system is used.
- On the other hand, since short text/alphabet/number as in ratings are very easy to understand, a rating result will work to develop a life of its own. Due to this risk, stakeholders, especially internal, are sensitive about the result of ratings; therefore, the cost (time and effort) for obtaining internal understanding is high.
- A rating result of project evaluation can be utilized for its improvement. On the other hand, it is difficult to utilize it if a rating is judged for a Country Program. Can we terminate entire aid for the target country?
- Even though a unified standard is shared by multiple evaluators, it is unavoidable to have dispersion among their ratings. Also, the timing of assigning rating affects the results.
- Even though it is not possible to conduct comparison for some cases, there is still a risk of comparing such cases because rating results are in front of us. For example, rating results of country evaluations and those of thematic evaluations are compared, or rating results of larger countries and those of tiny countries are compared, and then wrong decision-making might be made based on those comparisons.
- In some cases, rating is not appropriate because of consideration for diplomacy. There are some cases that aid is an important diplomatic tool because of the historical aspects or geopolitical reasons, even though a

high-level impact cannot be expected at all.

Table 9
Positions of Major Donors on Rating System

1. Agencies that have Introduced a Rating System in Country Program Evaluations	
World Bank, ADB,	Among international agencies, development banks (such as the World Bank and ADB) are forerunners of introducing rating systems. In Country Program Evaluation, a single rating score is calculated as an overall result of the evaluation, which brings to a conclusion. World Bank has introduced a rating system in 1970s. However, pros and cons of applying a rating system for Country Assistance Evaluation (CAE) have continuously been discussed internally until 2010s. They decided to use the rating system partially because the Committee of Development Effects (CODE) of the World Bank considered it as a good tool for drawing management's attentions with the largest impact.
CIDA	CIDA is the only bilateral agency that has introduced a rating system for Country Program Evaluation. Several sample projects are selected from each sector and rated. Then an average rating is calculated according to each sector. However, CIDA does not calculate an overall average rating across multiple sectors in order to avoid that the overall average rating is used as a conclusion of the Country Program Evaluation.
2. Agencies that are considering the introduction of a Rating System in Country Program Evaluation	
UNDP	UNDP started considering the introduction of a rating system in FY2010. Recently, the importance of accountability, especially the importance of evaluation, has been emphasized by the Board of Directors, and the number of Country Program Evaluations has been largely increased. As a result, a standardized methodology for Country Program Evaluations was required: 1) to keep the quality of evaluations constant, 2) to increase the manageability of multiple evaluations, and 3) to make the comparisons of evaluation results possible. Based on these requirements, UNDP created an evaluation manual that included a rating system (ADR Manual, 2010). However, the trial of using this manual scheduled in FY2010 has been postponed. Different from bilateral agencies and development banks, UNDP's activities tend to be conducted jointly with, or under the recognition of, the host governments. If a rating system is introduced into Country Program Evaluations, UNDP would need to seek permission from the host governments to evaluate the policies and performances of the target countries, which is perceived as politically sensitive. Under such circumstances, it is stated that UNDP is deadlocked.
3. Agencies that have not introduced any Rating System in Country Program Evaluation	
IADB	The purposes of evaluation are to learn lessons from experience and to increase accountability. To this end, clear and honest discussions with narrative evaluations are useful and effective. Once a rating system is introduced, people would pay attention only to the rated scores and stop other discussions. IADB has not introduced a rating system for this reason.
Germany (BMZ)	BMZ has not introduced a rating system, because BMZ's Country Program Evaluations or thematic evaluations do not accord with a rating system. The introduction of a rating system has not been discussed in the past. However, because evaluation methodologies are to be discussed by a new external agency for evaluation approved by the Cabinet, there might be some change in BMZ's evaluation policies.
DFID	It is not clear why DFID has not introduced a rating system, or whether related discussions have been made in the past. A possible reason would be: DFID conducts a small number of evaluations annually, such as country-wise, sectoral, and thematic evaluations, which is not enough to apply to coherent rating. Instead, making it easy to understand, DFID employs a "traffic light system", similar to rating, that marks grades to the progress and performances of aid projects in accordance with output indicators using the three colors of traffic lights.
USAID	USAID has not introduced a rating system because it employs the system of Performance Measurement, which brings to a binary judgment on whether the numerical target has been achieved or not.
4. Agencies that have introduced a Rating System in Project Evaluation	
Germany (GTZ, KfW)	The reason that GTZ has introduced a rating system is that it would be an effective tool to increase accountability to taxpayers and congresses of Germany. They said that while searching for a tool for increasing accountability without technical jargon, an idea of using the reporting system in school (like the six grades in Germany) came up, which is familiar to all people. This is why GTZ's rating system has six grades.
Japan (JICA)	JICA applies a rating system to all ex-post evaluations to be accountable for the results by using an easily understandable tool. On the other hand, JICA mentioned that it would not be appropriate to emphasize the results of the rating system. Instead, they should be utilized only as a reference, because the scores eliminate the details of evaluations and do not reflect the overall results of evaluations.

(Source) Documents downloaded from donor websites and field interview results.
Adapted from IDCJ (2010), pp. 89-90, translated into English by the author.

Criteria *of value*

Standards *of value*

	Relevance	Effectiveness	Impact	Efficiency	Sustainability
a. Highly satisfactory	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*
b. Satisfactory	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*
c. Hard to say	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*
d. Unsatisfactory	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*
e. Highly unsatisfactory	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*	*(Definition)*

Figure 2. Criteria and Standards of Values in Evaluation

Overall Evaluative Conclusion and Integrating Methods

Evaluation should provide an overall evaluative conclusion, and it is not enough to provide a sub-evaluation result such as evaluation based on DAC five evaluation criteria. However, in order to generate a single overall evaluative conclusion, some integrating methods are necessary. By examining how each donor agency draws an overall conclusion, three groups are identified (see Table 10). Some remarkable observations are as follows.

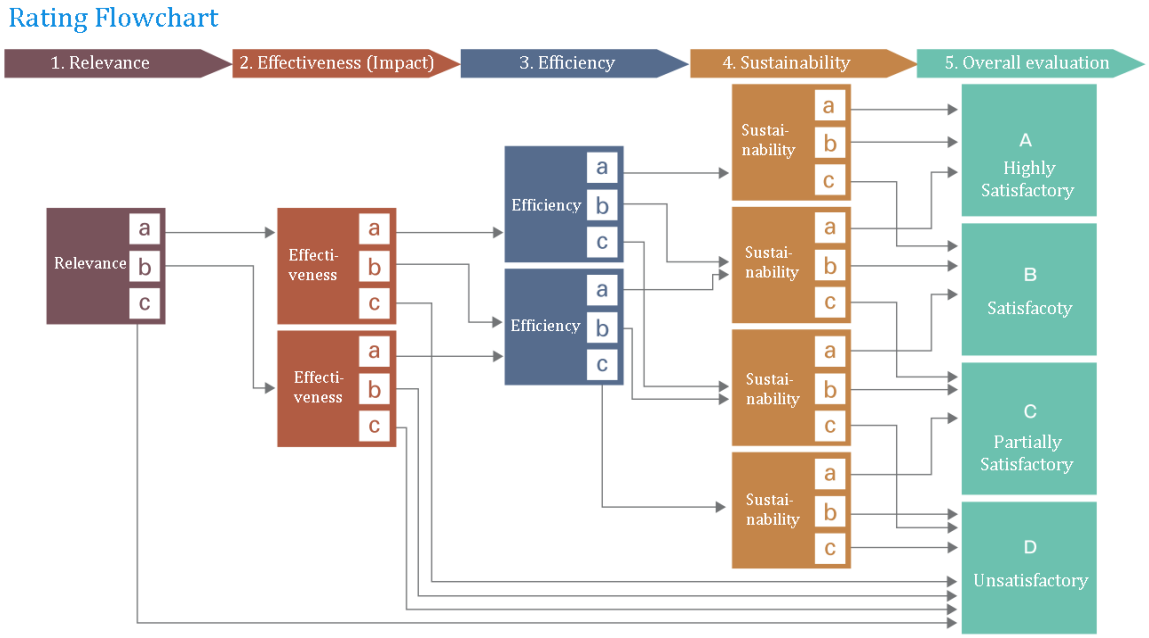
- The World Bank and ADB clearly employ a system of overall evaluative conclusion both for Country Program Evaluation and project evaluation.
- CIDA calculates the average rating from the rating results of multiple sample projects in its Country Program Evaluation, but it declares it does not see and use the average rating as an overall evaluative conclusion.
- Also GTZ, KfW and JICA create synthesized rating for their project evaluation.

One good example of integrating method is to utilize a logical flowchart (See Figure 3) although this kind of flowchart has not been developed for Country Program Evaluation. This flowchart is well based on the logic of evaluation. First, check "relevance", and if the program is evaluated as "not relevant" (= serious flaw of logic of intervention is found), it is a waste of time to continue conducting overall evaluation. It should simply go to the worst overall rating, which is "Unsatisfactory". Also, for "effectiveness", if the program does not produce any result (social/economic favourable change among impactees (beneficiaries)), it is again in vain and the overall evaluative conclusion should be judged as "unsatisfactory" as the following flowchart indicates. This kind of logical flowchart should be developed for Country Program Evaluation.

Table 10
Comparison of Overall Evaluative Conclusion

Approach	Agency	Comments
		Country Program Evaluation
Overall rating	World Bank	• For an overall evaluative conclusion, 5 ranks of the rating scale are "Highly satisfactory", "Satisfactory", "Moderately Satisfactory", "Unsatisfactory", and "Highly unsatisfactory". • Integration is not quantitative but qualitative. The evaluator makes an overall rating by seeing the list of ratings for sub-criteria.
	ADB	• For an overall evaluative conclusion, 4 ranks of the rating scale are "Highly successful", "Successful", "Partly Successful", and "Unsuccessful". • Sector performance points (30 points) made by a bottom-up approach and country points (30 points) made by a top-down approach are added. Thus, full point is 60 points. A rating is assigned by numerical scale using the total point.
Description by text	IADB	The manual indicates an evaluative judgment for each criterion should be written.
	France (MOFEA)	An overall evaluation result is written by text.
	Japan (MOFAJ)	Although it is no mandatory to make an overall evaluation, it is written by text in many country evaluation reports (but not all reports).
No overall evaluative evaluation	CIDA	The average score of eight criteria of sample projects is calculated, but the guideline clearly states this average should not be regarded as the overall evaluative conclusion.
	UNDP	Although evaluation results of respective criterion are written in the conclusion section, UNDP does not make either an overall rating or overall conclusion by text.
	DFID	Although evaluation results of respective criterion are written in the conclusion section, DFID does not make either an overall rating or overall conclusion by text.
	The Netherlands	Overall evaluation had not been made until 2006. The current situation is unknown.
		(For reference) Project Evaluation
Overall rating	World Bank	For an overall evaluative conclusion, 5 ranks of the rating scale are "Highly satisfactory", "Satisfactory", "Moderately Satisfactory", "Unsatisfactory", and "Highly unsatisfactory". Integration is not quantitative but qualitative.
	ADB	For an overall evaluative conclusion, 4 ranks of the rating scale are "Highly successful", "Successful", "Partly Successful", and "Unsuccessful". Integration is quantitative.
	Germany (GTZ, KfW)	GTZ: A rating is assigned as an overall evaluative conclusion. A weighting (3, 2, or 1) is assigned to each criterion, and multiplication and addition is conducted for calculating the overall point. KfW: No weighting is assigned for relevance, effectiveness, efficiency and impact, and the actual weighting is decided case-by-case basis according to the characteristics of each project. However, "unsuccessful" is always assigned if the rating for relevance or effectiveness is low.
	Japan (JICA)	A logical flowchart for integrating ratings is used.

（Source）Documents downloaded from donor websites and field interview results.
Adapted from IDCJ (2010), p. 56, translated into English by the author.



(Source) Japan Bank for International Cooperation. (2006).

Figure 3. Rating flowchart for Making a Overall Evaluative Conclusion

Conclusion

This study made some new knowledge about how aid agencies conduct Country Program Evaluation and identified several issues that remain. Varieties of their practices are observed and it is far from the unified methods agreed. Some remarkable points identified in this study are:

1. Most aid agencies invoke the DAC five evaluation criteria for Country Program Evaluation. (Major exception was USAID).
2. "Strategic relevance" and "coherence/complementarity" are the emerging new criteria.
3. Attribution is still the issue that aid agencies have struggled.
4. The attitude for introduction of rating system is clearly divided among aid agencies.

This study has conducted surveys and interviews with major aid agencies and conducted comparative analysis on their evaluation systems. However, it is true that some issues are left behind as stated, which require more surveys and analyses. Also, it is found that several donor agencies are in the process of a large-scale reform of their evaluation systems. Thus, it is recommended that a further study including a follow-up of outcomes of those reforms should be conducted.

Recommendations

Some recommendations for improving the current practice of Country Program Evaluation based on the findings of this study are as follows.

1. Preparation of a new guideline for Country Program Evaluation is one good idea. (e.g., "DAC Country Program Evaluation guidelines").
2. Introduction of rating system should be more seriously considered with care of its limitation. As identified, the rating system has great merits and fits the formal logic of evaluation.
3. New definition of several emerging words should be prepared and agreed. For example, definitions of Type I, II, and III "Impact" as proposed in this paper.

This study has conducted surveys and interviews with major aid agencies and conducted comparative analysis on their evaluation systems. However, it is true that some issues are left behind as stated, which require more surveys and

analysis. Also, it is found that several donor agencies are in the process of a large-scale reform of their evaluation systems. Thus, it is recommended that a further study including a follow-up of outcomes of those reforms should be conducted.

References

Cassen, R. & Associates. (1994). *Does aid work*? (2nd ed.). Oxford: Oxford University Press.

Davidson, J. (2005). *Evaluation methods basics*. Thousand Oaks, CA: Sage.

Organisation for Economic Co-operation and Development - Development Assistance Committee (OECD-DAC). (1991). *DAC Criteria for Evaluating Development Assistance*. OECD-DAC. Retrieved from http://www.oecd.org/document/22/0,2340,en_2649_34435_2086550_1_1_1_1,00.html

Organisation for Economic Co-operation and Development - Development Assistance Committee (OECD-DAC). (2002). *Glossary of key terms in evaluation and results based management*. OECD-DAC. Retrieved from http://www.oecd.org/dataoecd/29/21/2754804.pdf

Organisation for Economic Co-operation and Development - Development Assistance Committee (OECD-DAC) - Network on Development Evaluation. (2002). *Development evaluation resources and systems: A study of network members*. OECD-DAC. Retrieved from http://www.oecd.org/dataoecd/13/6/45605026.pdf

International Development Center of Japan (IDCJ). (2010). *Study on country program ODA evaluation systems and methodologies*. Ministry of Foreign Affairs, Japan. Retrieved from http://www.mofa.go.jp/mofaj/gaiko/oda/kaikaku/hyoka/pdfs/10_oda_hyoka_tyosa.pdf International

Initiative of Impact Evaluation (3iE). (2011). *Impact Evaluation Glossary – December 2011*. Retrieved from http://www.3ieimpact.org/userfiles/doc/Impact%20Evaluation%20Glossary%20-%20Dec%202011.pdf

Japan Bank for International Cooperation (JBIC). (2006). *JBIC Loan Project Evaluation report*. Japan International Cooperation Agency (JICA).

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.

Ministry of Foreign Affairs of Japan. (2009). *ODA Evaluation Guidelines* . ODA Evaluation Division, International Cooperation Bureau, Ministry of Foreign Affairs of Japan (MOFAJ). Retrieved from http://www.mofa.go.jp/policy/oda/evaluation/guideline.pdf

Scriven, M. (1991). *Evaluation thesaurus (*4th ed.). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Stokke, O. (1992). *Evaluating development assistance: Policies and performance*. London: Frank Cass Publishers.

Full reference of each donor agency is available at Annex 6 of the following report:

International Development Center of Japan (IDCJ) (2010). *Study on Country Program ODA Evaluation Systems and Methodologies*. Ministry of Foreign Affairs, Japan.

[i] The study team consists of Dr. Ryo SASAKI (IDCJ), Mr. Masaharu SHIMIZU (IDCJ), Ms. Mana TAKASUGI (IDCJ) and MS. Mihoko KIKUCHI (IDCJ). Advisors for this study are Professor Ryokichi HIRONO and Professor Masafumi NAGAO.

[ii] The Ministry of Foreign Affairs of Japan (MOFAJ) calls this type of evaluation as "Country Policy Evaluation" and categorizes it as policy-level evaluation (2010, p.43). The Ministry's guidelines (2010) further explain that "A country policy evaluation is conducted on the overall assistance policy for a country, specifically on the Country Assistance Program. In principle, evaluations of this type have the aim of contributing to the formulation and revision of Country Assistance Programs." (p.42) Following this explanation, it should be pointed out that the difference between policy and program is not so clear.

[iii] In 2011, GTZ changed its name to GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit).

[iv] DAC Evaluation Criteria consist of *relevance, effectiveness, efficiency, impact and sustainability*. It should be mentioned that this set is very specific to the aid evaluation community in terms of (i)the separation of *effectiveness* and *impact;* and (ii) the addition of *sustainability*. It should be understood that these features are not common in wider (or mainstream) evaluation community.