論文 / 著書情報
Article / Book Information

Title	A Trial of Impact Evaluation Utilizing Effect Size Statistics: Its Application to the Evaluation of Japan ’ s Foreign Student Policy
Authors	Yuriko Sato
Citation	Evaluation, Vol. 18, No. 2, pp. 230-245
Pub. date	2012, 4
DOI	http://dx.doi.org/10.1177/1356389012443163
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Evaluation, Vol.18, No.2, pp.230-245

A Trial of Impact Evaluation Utilizing Effect Size Statistics:
Its Application to the Evaluation of Japan's Foreign Student Policy

Yuriko Sato
(Tokyo Institute of Technology, Japan)

The adoption of the Millennium Development Goals (MDGs) and compilation of the Poverty Reduction Strategy Papers (PRSP) in many developing countries have promoted the collection of indicators related to various aspects of development. These indicators are monitored in order to examine the impact of development interventions conducted by donor organizations, as well as the developing countries themselves. There has been active discussion as to how to measure the impact of the treatments accurately by utilizing outcome indicators: an important question in providing effective development assistance. Clements, Chianca, and Sasaki (2008) point out the existence of the incentive for positive evaluation bias in development agencies.

The Center for Global Development (CGD) set up the Evaluation Gap Working Group. After examining various evaluations of social sector programs, the working group report points out that few truly rigorous impact evaluations have been conducted and calls for collective action by all stakeholders in development assistance to promote more and better impact evaluations supported by evidence (Center for Global Development, 2006).

The most rigorous method for impact evaluation is widely acknowledged to be randomized controlled trial (RCT). According to Lachin, Matts, and Wei (1998), 'RCTs are considered the mostreliable form of scientific evidence in healthcare because they eliminate spurious causality and bias.' However, it is difficult to conduct RCTs in social settings. Ravallion (2003) of the World Bank states, 'Randomization is the theoretical ideal and a natural benchmark for assessing non-experimental (quasi-experimental) method.' While admitting that 'there are sometimes opportunities for randomizing the assignment of an anti-poverty program, possibly on a pilot basis,' he concludes that 'it is frequently the case in practice that randomization is not a feasible option.' In spite of this kind of perception, various efforts have been made by the aid agencies and the network of academics and aid practitioners towards conducting more rigorous impact evaluations (Sasaki, 2006; Aoyagi, 2007; Poverty Action Lab, 2008; Khandker, Koolwal and Samad, 2010; International Initiative for Impact Evaluation, 2010).

Hansen and Rieper (2009) review the practice of evidencebased evaluations by the evidence producing organizations in health, social welfare and education sectors and found that some subscribe to a narrow concept of evidence, only including results from RCT-designed studies, while others subscribe to a broader concept of evidence from a variety of study designs reflecting their epistemological tradition. Radej (2011) proposes a meso approach to synthesize the results of various impacts of a regional development program on the economic, social and natural environment but the magnitude of each impact is not reflected in the synthesis.

The purpose of this paper is to propose a method to measure the effect, impact and efficiency of treatment through using standardized mean difference (SMD) of the outcomes between the treatment and the control groups. We assume that 'standardized effect' expressed by SMD (effect size statistic) is commensurable and gives it a unit name 'effect'. If impact is defined as the total change brought about by the treatment in the population, impact can be calculated by multiplying SMD by the population size. Efficiency can then be calculated by dividing the impact by the total input of the treatment. This population effect approach can facilitate the comparison of impact and efficiency between the different treatments using the same set of indicators.

To facilitate the understanding of this method, the proposed method is applied to the evaluation of Japan's foreign student policy (hereafter 'FSP') by comparing those who had studied abroad in Japan and those who chose to study in the universities in their own country. Since the control group could not be randomly selected due to the limitation of the evaluation design, the result contains a selection bias. An important premise of this method is the accuracy of SMD, so this case is not a good example of applying this method. However, it will help the reader to understand this method better and to utilize this method in a more rigorous way in the future.

The first section of this paper introduces the basic concepts and the steps of the proposed method in comparison with the existing methods of impact evaluation. The second section demonstrates an example of its application in the evaluation of Japan's FSP. The final section concludes with a discussion of this method's merits and limitations.

A Method to Measure the Effect and Impact

In the impact evaluation, ordinary least-squares (OLS) regression is one of the most commonly used methods to measure the impact of treatment (McEwan, 2010). A basic regression model, assuming a linear correlation, can be written as below where Y_i represents the outcome of each individual *i*, *X* stands for the control variables, ε is a residual (error term) and *D* is a dummy variable where *D*=1 represents the treatment while *D*=0 stands for non treatment.

$$Y_i = a + bX_i + cD_i + \varepsilon_i$$

Using OLS, the regression coefficients a, *b*, and *c* can be estimated. In the absence of controls for *X*, the estimates of c would be interpreted as the average difference between the treatment and the non treatment (=control) groups (treatment effect). However, if the treatment and the control groups are not selected randomly, *D* will be an endogenous variable and the treatment effect cannot be measured correctly.

Although RCT is the best solution to this problem, an alternative solution is 'matching' when RCT is not possible. Matching involves pairing treatment and control units that are similar in terms of their observable characteristics. By choosing the appropriate treatment and control groups through matching, the average treatment effect (ATE) can be measured by the difference of the average values of the treatment and control groups. Abadie et al. (2001:2) show the population and sample ATE in the following equation where $Y_i(1)$ is the outcome of the individual *i* when she is exposed to the treatment and $Y_i(0)$ is the outcome of the individual *i* when she is not exposed to the treatment.

$$\tau^{\text{pop.}} = \mathrm{E}[\mathrm{Y}(1) - \mathrm{Y}(0)] \text{ and } \tau^{\text{sample}} = \frac{1}{\mathrm{N}}\left(\sum_{(\mathrm{i}|\mathrm{D}=1)} \mathrm{Y_i}(1) - \sum_{(\mathrm{i}|\mathrm{D}=0)} \mathrm{Y_i}(0)\right)$$

The size of the treatment and control groups do not have to be identical since matching with replacement is possible (Dehejia and Wahba, 2002:154). Based on this idea, sample ATE would be expressed in the following equation as well.

$$\tau^{\text{sample}} = \left(\frac{1}{\mathrm{N}_1}\sum_{(\mathrm{i}|\mathrm{D}=1)} \mathrm{Y_i}(1) - \frac{1}{\mathrm{N}_0}\sum_{(\mathrm{i}|\mathrm{D}=0)} \mathrm{Y_i}\,(0)\right)$$

Utilizing the first equation of OLS, ATE can also be expressed in the following equation.

$$\mathrm{E}[\mathrm{Y}(1) - \mathrm{Y}(0)] = b\mathrm{E}[\mathrm{X}(1) - \mathrm{X}(0)] + c + \mathrm{E}[\varepsilon(1) - \varepsilon(0)]$$

If one could assume that the observable factors are almost the same on average $[\mathrm{X}(1) \approx \mathrm{X}(0)]$ and the unobservable factors are also almost the same on average $[\varepsilon(1) \approx \varepsilon(0)]$ in the treatment and the control groups, estimate of the ATE would be c.

Although only RCT enables the above assumptions, matching is an effort to minimize the difference of the observable factors of the two groups. Propensity score matching (PSM) is one of the most advanced methods of matching. Propensity score is defined as the predicted probability that the individual receives a treatment, given their observed characteristics. Probabilities of the treatment conditional on covariates are estimated using a probit or logit regression. Matching is conducted between the treatment and control units whose propensity scores are sufficiently close. Yet, the causal interpretation of PSM rests on the unverifiable assumption that no unobserved variables are correlated with outcomes and with probability of the treatment (McEwan, 2010).

Using PSM, Abadie et al.(2001) estimate the effect of participation in National Support Work (NSW), a job training program in the USA on individual earnings as $1,903 in 1978 based on 445 observations. ATE expressed in raw units can easily be interpreted as an expression of the impact of the program. However, it is not always easy to examine the impact of programs by comparing ATE in raw units. For example, the magnitude of an ATE of $1,903 would be different between Program A and B, when the average income of the treatment group is $11,903 and that of the control group is $10,000 in the former while that of the treatment group is $101,903 and that of the control group is $100,000 in the latter. One may hope to compare the magnitude of impact by dividing ATE by the average outcome of the control group. However, this is possible only when the outcome is measured on the ratio scales which have a true zero.

Many outcome measures are scaled in arbitrary units and lack a true zero. Effect size statistic expresses the magnitude of effect in a standardized form that makes it comparable across measures that use different units or scales (Rossi, Lipsey, and Freeman (2004: 304). SMD is the expression of effect size when the indicator is a continuous variable.

This characteristic of effect size also makes it easier to calculate the weighted average of the SMD of the different outcome measures to express the synthesized effect as long as all of the outcome measures are continuous variables. The odds ratio is used in the calculation of effect size found in the binary variables. Since SMD and odds ratio are different in their nature, it is not possible to synthesize the effect when the outcome measures consist of both continuous and binary variables.

The calculation of SMD is shown in Box 1. As you can see in the first equation in the box, SMD shows ATE on a scale on which the pooled standardized deviation value of the two groups is expressed as one. It shows the size of ATE relative to the range of the lowest and highest scores of the treatment and control groups. There are a few approaches to the calculation of SMD (Carson, 2008: 3), and the author adopts the Hedge's approach (called Hedge's 'g'), in which pooled standardized deviation (SD) of the treatment and the control groups is calculated being weighted by their sizes. The reason for using Hedge's 'g' is that it is not as influenced by the sample size as Cohen's approach (called Cohen's 'd') (Barnette 2007).

Box 1. Calculation of standardized mean difference (SMD)

$$(\overline{Y_1} - \overline{Y_0})/sd_p = \left(\frac{1}{N_1}\Sigma \quad \mathrm{Y_i}(1) - \frac{1}{N_0}\Sigma \quad \mathrm{Y_i}(0)\right)/sd_p$$

$\overline{Y_1}$ = Mean outcome value of the treatment group

$\overline{Y_0}$ = Mean outcome value of the control group

N_1 = Sample size of the treatment group

N_0 = Sample size of the control group

$$sd_p = \sqrt{\{(N_1 - 1)sd_1^2 + (N_0 - 1)sd_0^2\}/(N_1 + N_0 - 2)}$$

sd_p = the pooled standardized deviation value of the treatment and the control groups

SMD is an expression of the standardized size of effect and commensurable. Based on this the author gives it a unit name 'effect' which shows the size of 'standardized effect'. If we assume that the sum total of the standardized effect in the treatment group shows the whole impact of the treatment in the sample, it is calculated by multiplying the SMD by its sample size as shown in the first equation in Box 2. In the extension of this logic, it would also be possible to calculate the sum total of 'standardized effect' (=SMD) in the population by multiplying SMD by the population size as shown in the second equation in Box 2 given that the sample represents the population properly (the mean and SD of the sample should be almost identical with those of the population). This calculated impact is expressed by the units of 'effect' and the population, for example 'effect*person'.

Box 2. Impact calculated as the sum total of standardized effect

Sum total of the SMD in the treatment group

$$\left((\overline{Y_1}-\overline{Y_0})/sd_p\right)\times N_1 == \left(\tfrac{1}{N_1}\Sigma\ \ \mathrm{Y_i}(1)-\tfrac{1}{N_0}\Sigma\ \ \mathrm{Y_i}(0)\right)/sd_p\times N_1$$

N_1 =Sample size of the treatment group

Sum total of standardized effect (=impact) in the population

$$\left((\overline{Y_1}-\overline{Y_0})/sd_p\right)\times N_1\times\left(P_1/N_1\right)\ =\ (\overline{Y_1}-\overline{Y_0})/sd_p\times P_1$$

P_1 =Population under the treatment

If the impact of the treatment is measured in this way, efficiency can also be calculated by dividing the whole impact by the total input of the treatment expressed in a monetary value. The equations to calculate the efficiency are shown in Box 3. The second equation shows that efficiency is SMD per capita input. The calculated efficiency is expressed by the unit of 'effect' and the unit of the population per monetary unit, such as 'effect*persons/one thousand dollars.'

Box 3. Efficiency calculated from the 'standardized effect'

$$\{(\overline{Y_1}-\overline{Y_0})/sd_p\}/I$$

$$=\{(\overline{Y_1}-\overline{Y_0})/sd_p\}/(I/P_t)$$

I =Total input of the intervention expressed in a monetary value

The important assumption of this method is that the treatment and the control groups are comparable. RCT will minimize this difference of the two groups, called 'selection bias' (Duflo and Kremer, 2003). If RCT cannot be used, other designs such as matching are necessary to minimize the difference of the observable variables of the two groups though they cannot assure the minimization of the difference of unobservable variables between the treatment and the control groups.

We also have to note that a standardized measure of effect size can be misleading in some cases. In raw units, the effect of putting one inch heel shoes on the height of people is constant across all people. In standardized units, however, the effect would depend on whether the sample in a study included people of similar heights or had a wider range of heights. So the efficiency of a one inch heel on height would be different in the different samples even though the 'real' effect (in raw units) is a constant. It is because the standardized effect expressed by SMD shows a comparative size of the effect to the variance of the outcome values of the group(s).

Rossi, Lipsey, and Freeman (2004) point out the difficulty in achieving adequate statistical power (t test with α= .05) in the various combinations of effect sizes (SMD) and sample sizes. Relatively high power (significant difference) is attained only when either the sample size or the threshold effect size is rather large. Achieving both is often unrealistic for impact evaluation (pp.311-313). SMD will be smaller when the outcome measure has any measurement error than the case when the outcome measure contains less measurement error if statistical tests are applied to control the type I error (finding statistical significance even when there is no difference between the two groups).

Case Study: Japan's Foreign Student Policy Evaluation

In this section, an application of the proposed method is demonstrated in the evaluation of Japan's FSP towards Thailand by comparing the Thai people who had studied abroad in Japan and those who have not studied abroad but graduated from a Thai university.

In 1954, nine years after the defeat in World War II, the Japanese Government restarted FSP by introducing the Japanese Government Scholarship Program (hereafter 'JGSP'). Since then, Japan has made efforts to increase the number of foreign students by taking various measures, such as strengthening JGSP and the support program for those studying in Japan at their own expense (hereafter 'SP'), including a system to reduce or exempt their tuition fees and the provision of a moderate amount of honors scholarships. The majority of these foreign students in Japan are from Asian countries.

Based on the analysis of the related policy statements, two main objectives of FSP are identified for the years between 1954 and 2001: to develop human resources in the student dispatched countries and to foster pro-Japanese leaders to promote friendship with these countries (Sato, 2002b: 202). The first objective has been set because FSP has been conducted as part of Japan's official development assistance (ODA).

In 2002 a survey was conducted in Thailand for the evaluation of Japan's FSP. Foreign students from Thailand made up the fifth largest foreign student group in Japan for the year 2002, behind China, Korea, Taiwan and Malaysia. Thailand was chosen as a focus of the research because of the large number of Thai alumni from Japanese higher educational institutions (HEI) and the existence of an alumni association.

A questionnaire was one of the main components of the survey and samples were taken from not only the Thai people who had studied at Japanese HEI between 1954 and 2001 (treatment group), but also those who had not studied abroad but graduated from Thai HEI (control group) and those who studied at American universities for more than one year during this period. The graduates of American HEI were selected as comparison group since USA has attracted the largest number of foreign students from Asia and is considered to be a successful policy model.

According to the UNESCO statistics, the number of Thai students who studied abroad in 1964 was 2,512, which increased to 17,093 or 0.03% of the total Thai population in 1996. The most popular destination has been the USA, and its share was 71.2% of Thai students studying abroad in 1996. According to the recent OECD statistics, however, the percentage of those who study in USA decreased to 43.3% in 2003, followed by Australia (24.6%), UK (11.5%) and Japan (6.0%).

A total of 6,392 Thai students are estimated to have studied at Japanese HEI between 1954 and 2001 from the statistics of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan: 2,993 (46.9%) were those sponsored by JGSP, 3,153 (49.2%) were self financed students with the support of SP, and 246 (3.9%) were sponsored by the Thai government. According to UNESCO and OECD statistics, the number of those who studied at American HEI during the same period is estimated to be around 81,000, about 13 times larger than that of the Japanese HEI graduates.

In response to the questionnaires mailed to 600 members of the Old Japan Student Association, Thailand (OJSAT), who were chosen randomly from the alumni association's name list, 332 replies were collected (response rate was 55.3%). Among the 332 respondents, 165 respondents studied under JGSP, 147 respondents were self financed students under SP, and 20 respondents were supported by the Thai Government Scholarship Program.

In response to the mailed questionnaires to 769 members who have odd numbers on the name list of the American University Alumni Association (AUAA), 223 replies were collected as a comparison group (response rate was 29.0%).

72 replies from the Thai HEI graduates who had not studied abroad were collected as a control group. They were selected from the colleagues of the Japanese HEI graduates.

In order to apply the proposed method, 'matching' was conducted by choosing only those who were born between 1949 and 1982 from the treatment, control and comparison groups. Table 1 shows the main attributes of the three groups after this matching. The attributes of those who studied under JGSP and SP are also shown for the comparison and will be explained later.

Following are the major differences among the groups, which show selection biases.

- The percentage of male is higher in the comparison group than the two other groups.
- The percentage of faculty members of universities is higher in the control group, that of government/semi-government officials is higher in the comparison group, that of company employees is higher in the treatment group, and that of company owners is lower in the control group, than the two other groups.
- The percentage of those who studied as undergraduates is higher in the control group while that of those who studied in master and doctoral programs is higher in the comparison group, and that of those who studied in other programs is higher in the treatment group than the two other groups.

Table 1

Main attributes of the questionnaire respondents

Respondent group		Aumni of Japanese HEI (treatment group)	Alumni of American HEI (comparison group)	Alumni of Thai HEI (control group)	Aumni of Japanese HEI under JGSP	Aumni of Japanese HEI under SP
Number		205	120	63	118	87
Gender		male 51.7%, female 48.3%	male 60.0%, female 40.0%	male 49.2%, female 50.8%	male 48.3%, female 51.7%	male 56.3%, female 43.7%
birth year		1949-1982 mean average 1963	1949-1977 mean average 1962	1949-1980 mean average 1969	1949-1979 mean average 1962	1949-1982 mean average 1965
Year of the start of higher education abroad/in Thailand		1969-1999 mean average 1988	1970-1999 mean average 1985	1964-1999 mean average 1986	1969-1999 mean average 1987	1969-1999 mean average 1988
Major Profession	Faculty member of universities	26.8%	36. 7%	42.9%	37. 3%	12. 6%
	Government and semi-government officials	10.2%	18.3%	3.2%	11.0%	9.2%
	Company employees	67.8%	54.2%	55.6%	65.3%	71.3%
	Company owners	15.1%	12. 5%	1.6%	5. 1%	28. 7%
Enrolled Course	Undergraduate course	29.3%	23. 3%	65.1%	19.5%	42.5%
	Masters course	37.6%	70. 8%	22.2%	54.2%	20.7%
	Doctors course	19.5%	28. 3%	7.9%	28.0%	8.0%
	Other programs	28.3%	5. 0%	3.2%	28.8%	27.6%

Source: Made by the author based on the questionnaire survey conducted in Thailand in 2002.
Respondents who were born between 1949 and 1982 are selected for the matching.

Note 1: % is the percentage of the total respondents of the group.

Note 2: In the item of profession and level of courses, respondents were allowed to choose multiple answers.

Note 3: Other programs include preparatory language courses and exchange programs

Note 4: Alumni of Japanese HEI do not include the Thai people who were sponsored by the Thai government.

The attributes of the Japanese HEI alumni respondents mostly correspond to those of 2,192 members on the OJSAT name list as listed below although the percentages of the faculty members in universities is higher and that of government officials is lower in the sample than in the population.

- Gender: male 51.3%, female 48.7%
- Profession: 54.9% are company employees, 21.0% are government officials and 16.8% are faculty members of universities

In order to see the effect of Japan's FSP towards the objective of 'fostering pro-Japanese leaders to promote friendship between Japan and Thailand', the following questions were posed:

- Applicability of the statement 'I like the Japanese people' 'I like the American people' and 'I think the people who studied in Japan/USA are influential in Thai society'
- Involvement in activities to promote friendship between Thailand and the country of their study

The first group of questions were represented on 5 level Likert scale (where 1 is 'not at all' and 5 is 'very much') while the last question was a choice between 'Yes' or 'No.' The respondents' feedback is shown in Table 2.

Table 2
Affinity with people of the country of study, social influence, participation in friendship activities

		alumni of Japanese HEI (A)		alumni of American HEI (B)		significant difference between A and B	alumn of Thai HEI (C)		significant difference between A and C
		M	SD	M	SD		M	SD	
1	I like the Japanese people	3.61	0.74				3.38	0.58	*
2	I like the American people			3.23	0.71		3.13	0.69	
3	I think the people who studied in the country of my study (Japan/USA) are influential in the Thai society	2.83	0.83	3.45	0.88	**			
4	Are you involved in any activities to promote friendship between Thailand and the country of your study?	Yes 85.9%, No 12.7%		Yes 49.2%, No 49.2%		**			

Source: Made by the author based on the questionnaire survey conducted in Thailand in 2002.
Respondents who were born between 1949 and 1982 are selected for the matching.

Note 1: Items 1, 2, 3 show the applicabitity of the statements shown on 5-level Likert scale (where 1 is 'not at all' and 5 is 'very much').

Note 2: *p<0.05, :**p<0.01

Note 3: % is the percentage of the total respondents of the group

The alumni of Japanese HEI showed more affinity towards the Japanese than the alumni of Thai HEI with a significant difference. The alumni of American HEI showed more affinity towards the Americans than the alumni of Thai HEI but without a significant difference. There was a significant difference between the Japanese and American HEI alumni in their affinity towards the country of their study.

The recognition of social influence of the alumni of Japanese HEI is smaller than that of the alumni of American HEI with a significant difference. An alumnus of Japanese HEI explained the reasons as follows: 'Those who graduated from Japanese universities are minority so the way of thinking (among the Thai) almost seems to be American style'. According to the academic staff data of the Chulalongkorn University, which is one of the most prestigious national universities in Thailand, among the twelve heads of the departments in the Faculty of Engineering, seven heads are the graduates of American HEI and one is the graduate of Japanese HEI. In the Faculty of Science, eight heads are graduates of American HEI and no graduate of Japanese HEI among the fourteen department heads. These figures confirm that the alumni of American HEI are more influential than those of Japanese HEI partly because the former outnumber the latter.

Regarding the participation in the friendship-promotion activities with the country of their study, the percentage is higher among the alumni of the Japanese HEI than the alumni of the American HEI with a significant difference. These questionnaire results seem to show the general tendency of the alumni of Japanese HEI to like the Japanese people more than the average Thai person and to engage in friendship-promotion activities with Japan, though their social influence is much smaller than that of the alumni of American HEI who outnumber them. We also have to bear in mind that those who chose to study in Japan might have had affinity towards the Japanese before their study, which should be counted as a selection bias.

Though the treatment and control groups are not sufficiently homogeneous, the author has chosen to use their data to calculate the standardized effect of Japan's FSP in the attainment of the objective of 'fostering pro-Japanese leaders to promote friendship between Japan and Thailand' in order to show the application of the proposed method.

Table 3

Key Indicator, Effect, Impact and Efficiency of Japan's FSP towards Thailand Regarding the Goal of "Fostering Pro-Japanese Leaders"

	Sample size	Applicability of statement "I like the Japanese " on Likart scale		ATE	Pooled SD	SMD (A)	Population who studied in Japan betwewn 1969 to 1999 (B)	Impact (A*B)	Budgetary input for Thai students between 1969 and 1999 (C) (million yen)	Efficiency (A*B/C)
		M	SD							
Alumni of Japanese HEI	205	3.61	.74	0.229	.71	0.323	4,923	1,589	30,426	0.052
Alumni of Japanese HEI under JGSP	118	3.59	.68	0.212	.65	0.327	2,487	814	26,734	0.030
Alumni of Japanese HEI under SP	87	3.63	.82	0.251	.73	0.344	2,436	837	3,692	0.227
Alumni of Thai HEI (control group)	63	3.38	.58							

Source: Made by the author based on the questionnaire survey conducted in Thailand in 2002.

Note 1: SMD is Hedge's 'g', using the pooled SD of the treatment and the conrol groups being weighted by their sizes

Note 2: The budget of JGSP and SP is calculated by the multiplication of the relevant FSP budget by the percentage of the Thai students of the total foreign students in each category. Each budget is deflated by the Consumer Price Index (CPI) in 2000 before being summed up.

Note 3: Total population who studied in Japan between 1969 to 1999 does not include the Thai people who were sponsored by the Thai government.

As shown in Table 3, standardized effect (SMD between the treatment and the control groups) is calculated to be 0.315 effect. Though SMD contains the error caused by the selection bias, the impact of Japan's FSP is calculated just to demonstrate the proposed method.

The treatment group contains only people who were born between 1949 and 1982 and started their study in Japan between 1969 and 1999 as shown in Table 1. Their impact of Japan's FSP will be calculated from the multiplication of SMD (ATE divided by the pooled SD) by the population (Thai people who started their study in Japan between 1969 and 1999). The calculated impact is 1,549 effect*person as shown in Table 3.

To measure efficiency, the cost of Japan's FSP needs to be ascertained. FSP has been conducted by MEXT and the Ministry of Foreign Affairs (MOFA): MEXT mainly oversees the measures in Japan, such as the education and support for the foreign students and provision of scholarships, while MOFA is responsible for the measures overseas such as the recruitment and selection of JGSP students and the support for the alumni associations.

As Levin and McEwan (2001: 45) point out, the assumption that the budget will contain all the cost information is usually erroneous because 'budgets often do not include cost information on all the ingredients used in the intervention' and miss costs paid by some other agencies. In the case of Japan's FSP, these invisible costs include the extra personnel cost of the academic/administrative staffs in the HEI to guide and support the foreign students who often do not speak/write Japanese well and the cost borne by the HEI and local governments for the education and support of the international students.

However, it is difficult to estimate these personnel and institutional costs correctly since the level of Japanese language proficiency varies among the students and the costs borne by the HEI and local governments for the international students are not often differentiated from those of the Japanese students. It should also be pointed out that the cost of FSP is estimated to be much higher in 1954 when JGSP was started. New educational programs including Japanese language courses, new academic/administrative staffs, and new facilities including dormitories had to be prepared for the newly started program. Inexperienced staffs had to spend much more time to guide and support the international students in the beginning. These costs gradually declined as the education and support of international students was institutionalized and became a part of the regular operation of Japanese HEI. As such it is difficult to estimate the decline of the initial cost correctly without a detailed survey.

Since these costs are difficult to ascertain, the FSP budget of MEXT and MOFA for Thai students will be used in the calculation of efficiency in this paper. We have to note that the cost based on the budget is much smaller than the true cost which contains the expenditures mentioned above.

The FSP budget of MEXT and MOFA used for Thai students between 1969 and 1999 was calculated by multiplying program budget by the percentage of Thai students. The annual budget is deflated by the Consumer Price Index (CPI) in 2000 and then summed up.

The sum total of FSP budget for Thai students from 1969 to 1999 is calculated to be 30,426 million yen. The efficiency of Japan's FSP towards Thailand is calculated to be 0.051 effect*person/million yen as seen in Table 3.

JGSP and SP are the main pillars of Japan's FSP. The Japanese government has used a comparatively large amount of money for JGSP: the JGSP budget used for the Thai students between 1969 and 1999 is calculated to be 26,734 million yen, which was mainly used for the provision of government scholarships, the recruitment and selection of JGSP applicants, and support for their education and living in Japan. The SP budget for the Thai students who study in Japan at their own expense during the same period is 3,692 million yen, which was mainly used for the exemption or reduction of their tuition fees, the provision of honorary scholarships, and the support for their education and living in Japan. The total SP budget for Thai students is about one seventh of the total JGSP budget between 1969 and 1999 as shown in Table 3. Since the budgetary inputs of JGSP and GP are so different, their efficiency needs to be examined.

From the main attributes of the responses under JGSP and SP shown in Table 1, the following differences between the two groups can be seen.

- The percentage of female respondents is higher in the JGSP than in the SP respondents.
- The percentage of faculty members of universities is higher, while that of company employees and company owners is lower, in the JGSP than in the SP respondents.
- The percentage of those who studied in undergraduate course is higher in the SP while the percentage of those who studied in master and doctoral courses is higher in the JGSP respondents than the other group.

Since the OJSAT name list does not differentiate between those under JGSP and SP, it is not possible to compare the characteristics of the samples and the populations.

The mean and SD of the two respondent groups regarding the answer on the affinity towards the Japanese are shown in Table 3. Both the JGSP and the SP groups show higher affinity towards the Japanese than the control group with significant differences. The SP respondents' affinity towards the Japanese is slightly higher than that of the JGSP, though without a significant difference.

When asked about the motivation for their study in Japan, the acquisition of scholarship is the most influential factor for the JGSP respondents, while the interest in Japanese language is stronger for the SP respondents. There are significant differences between the two groups in regard to these responses. In addition, the SP respondents answered higher ability in listening and reading Japanese during their study than the JGSP students with significant differences. The SP students' higher interest and ability in Japanese may have affected their seemingly higher affinity towards the Japanese people.

The standardized effect (SMD) of JGSP to the objective of 'fostering pro-Japanese leaders' is calculated to be 0.327 effect while that of SP is 0.343 effect as shown in Table 3. SP shows higher standardized effect than JGSP though the figures may include the error caused by selection bias.

The impact and efficiency of JGSP and SP sponsored students who started their study in Japan between 1969 and 1999 are calculated based on SMD, population, and the Japanese government budgetary input into the two programs during the same period. For JGSP impact and efficiency were measured to be 813 effect*person and 0.030 effect*person/million yen respectively. For SP impact was 836 effect*person and efficiency was 0.227 effect*person/million yen as shown in Table 3. The efficiency of SP is about 7.6 times higher than that of JGSP.

Though the calculated figures may not be accurate because of the selection bias, it appears that SP has higher efficiency than the JGSP. This indicates a need for the further expansion of SP for the efficient execution of FSP towards Thailand.

Before concluding this section, the role of OJSAT in the promotion of mutual understanding and friendship between the two countries will be introduced as another proof of this evaluation.

OJSAT was established in 1951 as the first alumni association of Japanese HEI in the world. It started one of the most prestigious Japanese language schools in Thailand and has published many books introducing Japanese culture and society. The profits derived from the language school and publishing business have provided necessary funds for their activities such as holding seminars on the latest topics on Japan. OJSAT also has cosponsored Japan Education Fair and Examination for Japanese Universities with the Japanese Embassy in an effort to promote study in Japan.

In the 1970's when Japanese investment drastically increased in Thailand, there was a rise of anti-Japanese sentiment among the Thai people. The most symbolic incident was the visit of Prime Minister Tanaka in 1974, who was met with an anti-Japan demonstration. It was so severe that he was unable to leave his car. However, in the late 1980's when Japanese investment again increased in Thailand, such strong anti-Japanese feeling did not occur among the Thai people. According to the poll survey conducted by the MOFA of Japan (2002) in Thailand, 89 % of the respondents regarded Japan as a friendly country and 69 % replied affirmatively about Japan's role for the development of Asia.

According to an official in charge of public relations at the Japanese Embassy in Thailand, this positive change of the sentiment towards the Japanese has been brought about partly by the efforts of OJSAT, as well as the Japanese companies in Thailand, for the promotion of mutual understanding and friendship. These activities of OJSAT have been recognized by the Japan Foundation, an affiliated organization of MOFA, and the representatives of the OJSAT were awarded its Special Prize by the Crown Prince of Japan in 2002.

In this section, we have examined the calculation of the standardized effect, impact, and efficiency of Japan's FSP, JGSP, and SP regarding the objective of fostering of the pro-Japanese as an example of the application of the proposed method. Though the calculated figures are not accurate because of the selection bias, we can confirm the general tendency of the Thai alumni of the Japanese HEI to have affinity with the Japanese and to engage in friendship-promotion activities with Japan, as represented by the activities of OJSAT. Japan's FSP seems to have had certain positive impact in the attainment of the objective and SP seems to be more efficient than JGSP.

Merits and Limitations of the Proposed Method

In the previous sections an evaluation method to measure standardized effect by using SMD of the outcomes between the treatment and the control groups was introduced. Since SMD is commensurable, the author gives it a unit name 'effect' as an expression of 'standardized effect'. Based on the assumption that impact is the sum total of the change brought about by the treatment in the population, the author asserts that impact can be calculated by multiplying the 'standardized effect' by the treated population given that the sample represents the population. In the extension of this logic, the author also asserts that efficiency can be calculated by dividing the impact by the total input of the treatment expressed in a monetary value.

The method was applied to the evaluation of Japan's FSP towards Thailand though the result contained selection bias. Based on the answers on the Likert scale (continuous variables which lack a true zero) regarding the affinity towards the Japanese between the treatment group and the control group (those who never studied abroad but graduated from Thai HEI), the standardized effects, impacts and efficiencies of JGSP and SP were calculated and compared. SP shows higher standardized effect, impact and efficiency than JGSP. The result implies that SP is more efficient than JGSP in attaining the objective of fostering of pro-Japanese people and promoting friendship with Japan in Thailand.

In the last part of this paper, the merits and limitations of the proposed method will be summarized.

One of the merits of using standardized effect measure is that it is possible to compare outcomes across measures even when the indicators lack a true zero. This characteristic of SMD also makes it easier to measure the synthesized magnitude of the effect by calculating the weighted average of SMD of different outcome measures as long as they are continuous variables.

However, this method has several limitations which must be considered before further application. The first is that 'the statistical effect sizes… are not necessarily good guide to the *practical* magnitude of the program effects they represent. Nor is their level of statistical significance indicative of their practical significance' as pointed out by Rossi, Lipsey and Freeman (2004: 314). Although Cohen (1988: 25-27) states that effect size (in this case, Cohen's 'd') is small when d = .2, medium when d = .5, and large when d= .8 as a common conventional frame of reference, he also makes reservations that these terms are relative.

With that being said, it should be kept in mind that the standardized effect expressed by SMD, is a comparative value, which significance may differ in different cases. One should be careful in comparing them between different treatments which use a different set of outcome measures.

The second reservation is that standardized measure of effect size (SMD) can be misleading in some cases. SMD is harder to interpret than ATE expressed in raw units and is affected by the variance in the treatment and the control groups.

The third point is that the existence of reliable data is a pre-requisite of this method. It is ideal to use the result of RCT. However, if it is not possible, the other methods that minimize the selection bias must be adopted. Cost estimation should also be made carefully considering all the factors of cost which are not included in the budget.

We should also note that this kind of quantitative method, based on statistical data analysis, cannot explain the reasons why certain interventions are effective while the others are not. The reasons for varying performance may be found not only in the treatment itself, but also within the treated population, which may be uncontrollable. Hence, this kind of quantitative method should be conducted in combination with the qualitative analysis to explore these factors further (Bamberger and Fujita, 2003).

Considering these limitations, the proposed method should only be used when the reliable ATE data are available and when it is difficult to compare effect, impact and efficiency in raw units. This method can also be used to show the synthesized effect of different outcome measures if all the outcome measures are continuous variables.

Standardized effect is a relative value and the utility of this method should be tested further by comparing more data with the cooperation of evaluators and development workers.

References

Abadie, A. et al. (2001) 'Implementing Matching Estimators for Average Treatment Effects in Stata', *The Stata Journal* 1(1):1-18.

Aoyagi, K. (2006) 'Trend of Impact Evaluation in International Development Community', in *Issues and Prospects of Evaluations for International Development*, pp.87-153. Tokyo: Foundation for Advanced Studies on International Development (FASID).

Bamberger, M., and Fujita, N. (2003) *Impact Evaluation of Development Assistance*. Tokyo: FASID.

Barnette, J. (2007) 'Using Effect Size and Association Measures', from a handout at a Professional Development Workshopat the 21st Conference of American Evaluation Association in Baltimore.

Carson, C. (2008) 'The Effective Use of Effect Size Indices in Institutional Research', retrieved on August 1, 2008 from http://www.keene.edu/ir/effect_size.pdf

Center for Global Development. (2006) *When Will We Ever Learn? Improving Lives through Impact Evaluation*, retrieved on August 1, 2008 from http://www.cgdev.org/content/publications/detail/7973

Clements, P., Chianca, T., and Sasaki, R. (2008) 'Reducing World Poverty by Improving Evaluation of Development Aid'. *American Journal of Evaluation* 29 (2): 195-214.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences second edition*. NY: Lawrence Erlbaum Associates.

Dehejia, R.H. and S. Wahba (2002) 'Propensity Score-Matching Methods for Nonexperimental Causal Studies', *The Review of Economics and Statistics* 84(1):151-161.

Duflo, E., and Kremer, M. (2003) 'Use of Randomization in the Evaluation of Development Effectiveness', retrieved on August 1, 2008 from http://www.povertyactionlab.org/research/rand.php

Hansen, H.F., and O. Rieper (2009) 'The Evidence Movement: The Development and Consequences of Methodologies in Review Practices'. Evaluation 15(2):141-163.

International Initiative for Impact Evaluation (2010) retrieved on December 15, 2010, from http://www.3ieimpact.org/

Khandker, S.Rl, G.B. Koolwal and H.A. Samad (2010) *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, DC: The World Bank.

Kusek, J.Z., and Rist, R.C. (2004) *Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington, DC: The World Bank.

Lachin, J.M., Matts, J.P., and Wei, L.J. (1988) 'Randomization in Clinical Trials: Conclusions and Recommendations'. *Controlled Clinical Trials* 9 (4): 365-374.

Levin, H. M. and P. J. McEwan (2001) *Cost-Effectiveness Analysis: Methods and Applications, 2nd ed.* California: Sage Publications.

McEwan, P. J. (2010) 'Empirical Research Methods in the Economics of Education', in D.J. Brewer and P. J. McEwan (ed.) *Economics of Education*, pp. 9-14. Oxford: Elsevier.

OECD (2003) 'Education at a Glance 2003 – Tables', retrieved on Oct. 19, 2005, from http://www.oecd.org/document/34/0,2340,en_2649_37455_14152482_1_1_1_37455,00.html

Ministry of Foreign Affairs of Japan (2002) *Nihon ni kansuru ASEAN (6 kakoku) Yoron Chosa (Poll survey on Japan in 6 ASEAN countries),* retrieved on June 19, 2005, from http://www.mofa.go.jp/mofaj/area/asean/yoron.html

Poverty Action Lab (2008) 'About the Abdul Latif Jameel Poverty Action Lab', retrieved on August 1, 2008, from http://www.povertyactionlab.org/

Radej, B. (2011) 'Synthesis in policy impact assessment'. Evaluation 17(2):133-150.

Ravallion, M. (2001) 'The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation', *The World Bank Economic Review* 15(1):115-140.

Ravallion, M. (2003) 'Assessing the Poverty Impact of an Assigned Program', in F. Bourguignon, and L.A. Pereira da Silva (ed.) *The Impact of Economic Policies on Poverty and Income Distribution–Evaluation Techniques and Tools*. Washington, DC: The World Bank.

Ravallion, M. (2005) 'Evaluating Anti-Poverty Programs', retrieved on August 1, 2008, from http://siteresources.worldbank.org/INTISPMA/Resources/383704-1153333441931/Evaluating_Antipoverty_Programs.pdf

Rossi, P. H., Lipsey, M. W., and Freeman, H. E. (2004) *Evaluation: a Systematic Approach, 7th edition*. California: Sage Publications.

Sasaki, R. (2006) 'Challenge of Evidence-Based Evaluation in ODA Sector: Poverty Action Lab'. *The Japanese Journal of Evaluation Studies* 6(1): 43-54.

Sasaki, R. (2007) 'Lessons Learned from the Embedded Institutional Arrangement in Aid Evaluation', from a handout at the multi-paper session on 'Does Aid Evaluation Work?' at the 21st Conference of American Evaluation Association in Baltimore.

Sato, Y. (2002a) 'Impact Study of the Japanese Government Scholarship Policy toward Indonesian Students from the Perspective of Fostering Pro-Japanese Leaders'. *The Japanese Journal of Evaluation Studies* 2(2): 59-78.

Sato, Y. (2002b) 'An Impact of the Japanese Government's Foreign Student Policy toward Indonesia: from the perspective of human resources development'. *Journal of International Development Studies* 11(2):201-219.

Sato, Y. (2003) 'An attempt of Evaluation of Japan's Foreign Student Policy: A Case Study in Thailand'. *Journal of International Students' Education* 8:1-21.

Sato, Y. (2005) 'A Case of Policy Evaluation Utilizing a Logical Framework: Evaluation of Japan's Foreign Student Policy towards Thailand'. *Evaluation* 11, (3): 351-378.

Sato, Y. (2006) 'Quantitative Evaluation Utilizing Standardized Effect Unit: Application to the Evaluation of Foreign Student Policy and Regional Cooperation Program'. *The Japanese Journal of Evaluation Studies* 6(1):103-118.

UNESCO (1963-1999) *Statistical Yearbook*. Paris: UNESCO Publishing and Bernan Press.

World Bank (2006a) *Conducting Quality Impact Evaluations under Budget, Time and Data*

Constraints, retrieved on August 19, 2007, from http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20188174~isCURL:Y~menuPK:384336~pagePK:148956~piPK:216618~theSitePK:384329,00.html

World Bank (2006b) *Impact Evaluation and the Project Cycle: Doing Impact* Evaluation Series No.1., retrieved on August 19, 2007, from http://siteresources.worldbank.org/INTISPMA/Resources/Training-Events-and-Materials/ie_and_projectcycle.pdf