Collection of Impact Evaluation Practices

18 evaluation exhibits from experimental design (RCT) to expert judgment
~For better "social impact measurement" implementation~

Recent Updates

- Nob. 2 019 Congratulations ! Three researchers of the MIT Poverty Action Lab (J-PAL) were awarded the 2019 Nobel Prize in Economics.
- May 2020_Added basic income impact evaluation (2020) conducted by the Finnish government.
- May 2020_Added explanation of "Social Impact Measurement".
- Sep. 2020_Added Impact Evaluation of Educational Effects via SMS/Telephones during the Coronavirus Pandemic (2020) .
- Oct. 2021_Explanation of the Evidence Pyramid is posted.
- Dec. 2022_Added explanation of social impact to ``Four types of uses of impact".
- April. 2023_Posted a link to "Impact Evaluation of Peace Creation for Syrian Refugees in Jordan".
- Sep. 2023._English translation of this impact evaluation collection is prepared and uploaded.
- Oct 2023. "Effect Size (ES) criteria newly proposed based on prior research in education" is added.
- Oct 2023. "Effect Percent (%)" is newly proposed in formal writing.

Version 8.2 (Last updated : October 12, 2023)

Ryo SASAKI, Ph.D.

sasaki.ryo (a) idcj.or.jp

Foreword

This report explains the design and application examples of so-called impact evaluation. The author first encountered impact evaluation in 1994 when he was enrolled in a master's program at New York University, which included a course called “Program Analysis and Evaluation”. When I was a Japanese undergraduate student, my Professor said people can conduct experiments in the natural sciences, but not in the social sciences. Instead, I was taught to use comparison and process analysis. However, I was wondering if it was really impossible to conduct experiments in social sciences, but the textbook I used for that subject in New York University explained actual examples of social experimentation with no hesitation. This was an eye-opener. That's how I discovered impact evaluation.

Fascinated by impact evaluation, I have continued to collect examples of impact evaluation for about 30 years and decided to prepare this report. I have also added an explanation covering five basic impact evaluation designs, from the simplest to the most rigorous designs.

The recent popularity of "Evidence-Based Practice" (EBP) and "Evidence-Based Policy Making" (EBPM) is very welcomed because “evidence” refers to the results of impact evaluation. I would like to see a society in which decisions are made based on evidence, rather than a society in which decision-making is influenced by the opinions of people in power, influencers, and executives.

And this shows that even socially vulnerable people can take the lead in decision-making if they can show evidence. In other words, it has the power to liberalize people from the differences such as age, gender, race, ethnicity, class of origin, country of origin, place of origin, presence of disability, etc.

It is my sincere hope that impact evaluation will liberate people and contribute to the improvement of society.

Ryo SASAKI, Ph.D.

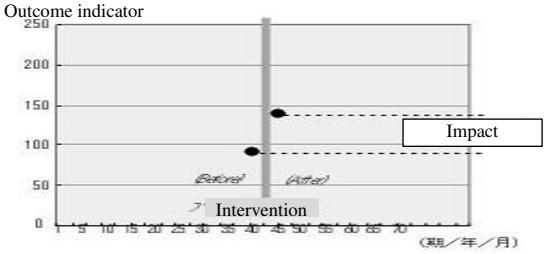
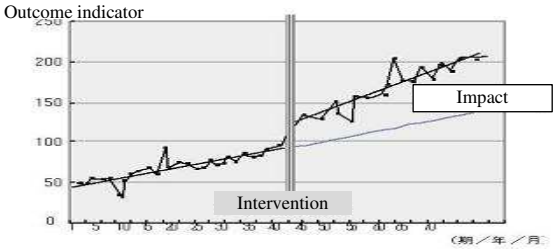
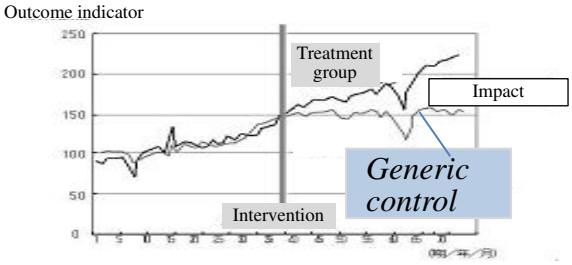
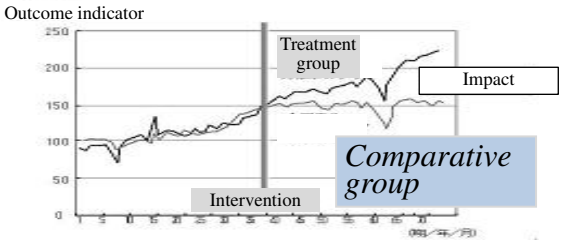
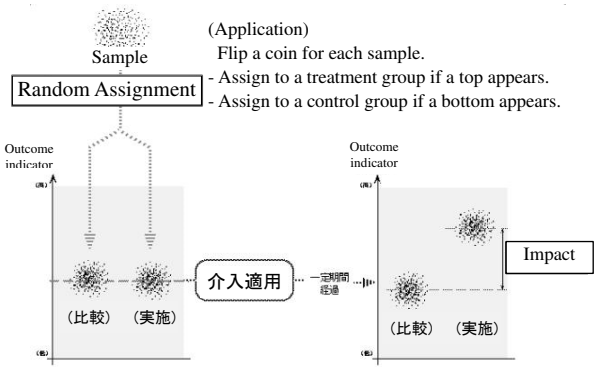
table of contents

Declaration and Disclaimer

This report was initially prepared in Japanese for Japanese audience. Since I received many requests for its English version, I prepared this report. Some materials in this report were captured from the original English report and I translated them into Japanese. Then I again translated Japanese text to English this opportunity. Thus, some expressions in this English report become naturally different from the one in the original English report because the double translation or transformation process from English to Japanese and again to English. I hope readers of this report understand this situation and my best effort. Readers who want to know the original expression should access and download the original English thesis even if they are paid version.

Ryo SASAKI. Sep 19, 2023

Introduction 1: Basic design of impact evaluation

Name and conceptual diagram	explanation	
1. Pre-post comparison design (Before-After) 	Simply compare the indicator's values before and after the event, and if there is a difference, it is assumed that there is a causal relationship. It is widely used because it is simple. However, the reliability of estimating the causal relationship is low because it is impossible to remove the influence value due to external factors that occurred before and after the event. [Test: Paired t-test]	simple & cheap
2. Time series design (Interrupted Time-Series) 	Determine long-term trends before the policy intervention, and if the trends change after the policy intervention, infer the existence of a causal relationship. However, it is not very reliable because it cannot remove the influence of external factors other than long-term trends. [Certification Test: Regression Analysis]	
3. General Control design 	General index values such as the national average value and all prefecture average values are used for comparison. Since it is possible to remove the influence values due to external factors to some extent (because the general index values are likely to be influenced to some extent to the same extent as the target area), there is a certain degree of reliability in identifying the existence of a causal relationship. can be secured. It's fairly easy to use. [Certification test: Visual judgment]	
Four. Matching design (Matched control) 	Select groups that are as similar as possible and use them for comparison. Since both groups are thought to be affected by external factors to the same extent (although not completely the same), a high degree of confidence can be ensured in determining the existence of a causal relationship. [Test: Independent t-test]	
5. Randomized controlled trial (RCT) 	Before implementing a measure, the policy application is divided into an implementation group and a comparison group by random assignment. It can be determined that the differences that appeared in the outcome index values were caused purely by the only intermediate difference, ``whether or not the intervention was applied." Note that there is no need to consider the effects of external factors as they are exactly the same. It boasts extremely high reliability in identifying the existence of causal relationships , and no other design exists . However, actual application is difficult. [Test: Independent t-test]	strict but expensive

(Source) The author selected representative designs from the table in Rossi, Freeman, Lipsay (1999) *Evaluation A Systematic Approach, 6th Edition*, Sage Publication, Sage Publication.p261 .

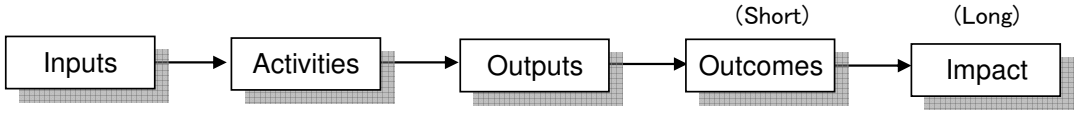
Introduction 2: Four ways to use the word “Impact”

The following four types of usage of "impact" are observed. In this text, we will follow the mainstream usage of Type III (`` pure amount of change due to intervention action ").

- “Impact”, one of the five DAC evaluation items* frequently used in ODA evaluation, covers both Type I and Type II of the following. (*Relevance, effectiveness, efficiency, impact, sustainability)
- Social impact (or sometimes simply called impact) referred to in the fourth category, ``social impact evaluation," has rapidly become popular since the late 2010s .

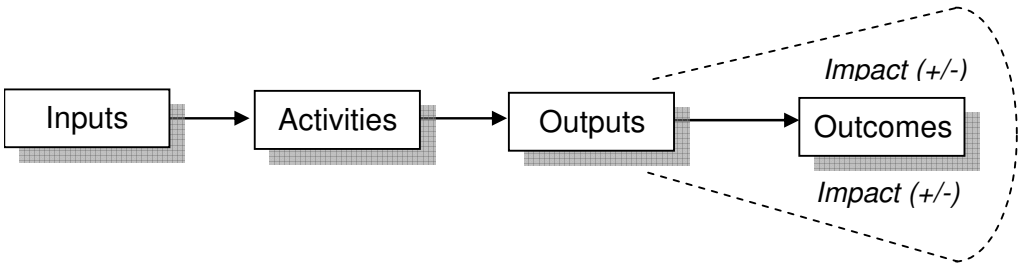
(1) Type I impact: Long-term social/economic impact

(Type I impact: long-term socio-economic change)



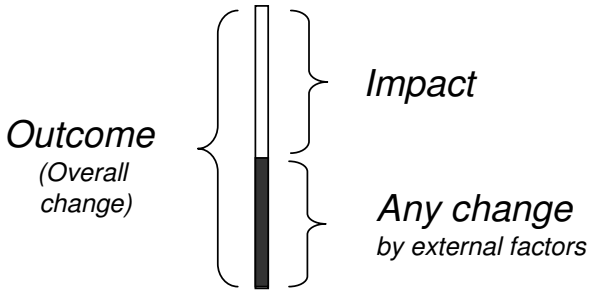
(2) Type II impact: Indirect impact (either positive and negative)

(Type II impact: indirect effect (can be both positive and negative))



(3) Type III impact: Pure change made by intervention

(Type III impact: pure amount of change due to intervention action)



(4) Type IV impact: Social Impact Measurement

(Type I V impact : Impact referred to in so-called "social impact evaluation")

It is usually defined as " social and environmental outcomes, including short- and long-term changes, caused by the project or activity." This definition includes (1) to (3) above.

(Source) Sasaki, R. (2002) *In In-Depth International Comparison of Major Donor Agencies: How Do They Systematically Conduct Country Program Evaluation?* . Journal of Multidisciplinary Evaluation Vol.8 – Number 18. http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/349 .

Furthermore , TypIV was added by Sasaki in January 2022 .

Introduction 3: Explanation of social impact measurement

1. definition

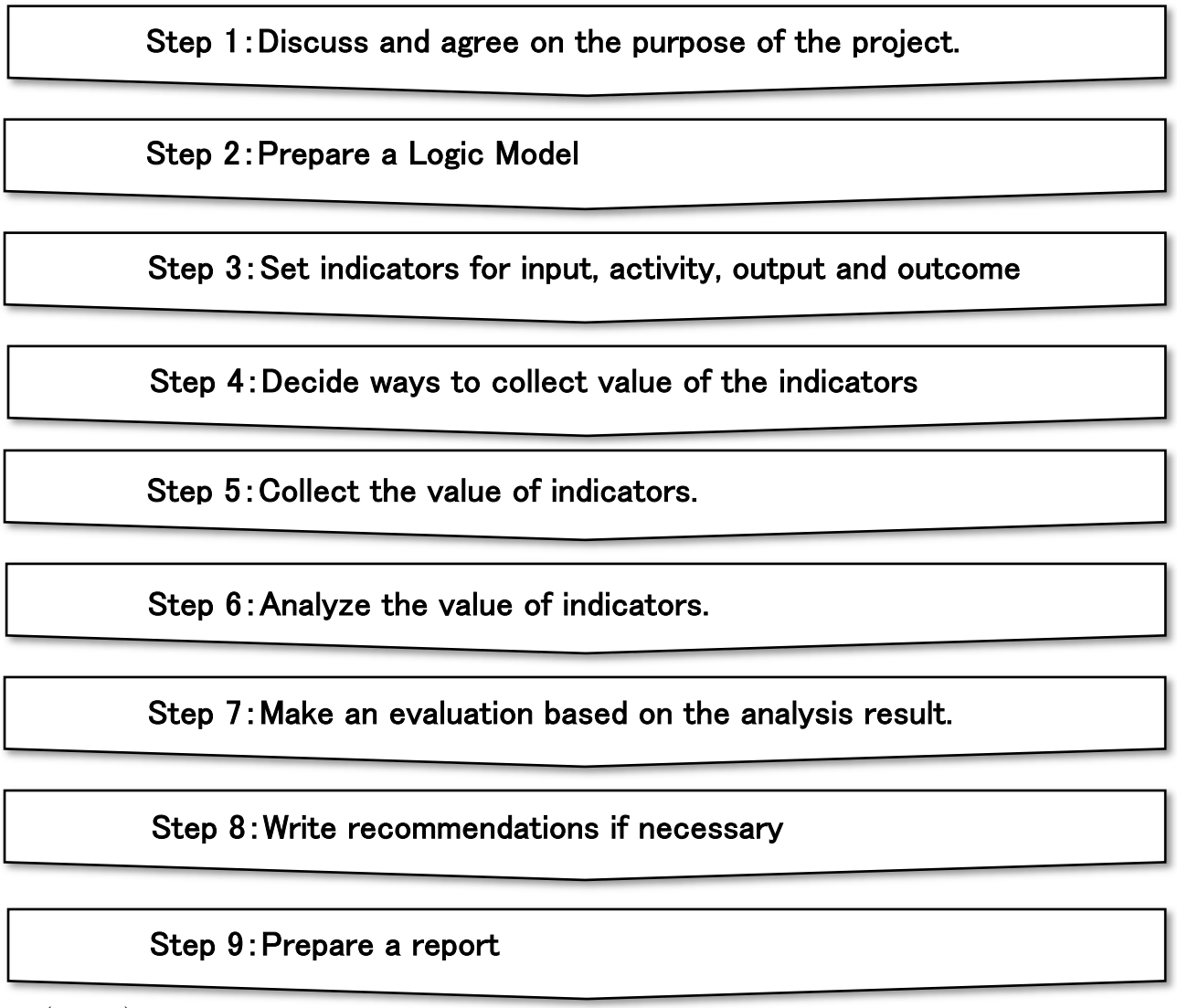
The standard social impact measurement procedure is as follows: The definition is a mixture of Type I Impact and Type II Impact explained in the previous figure . It also seems that it is sometimes used to include the meaning of Type III Impact.

> To reveal the socio-economic changes that result from their activities. Note that socio-economic changes include short-term, medium-term, long-term, direct and indirect, and intended and unintended changes.

2. Specific procedures

Social impact evaluation will be conducted using the following procedure.

Social impact evaluation procedures



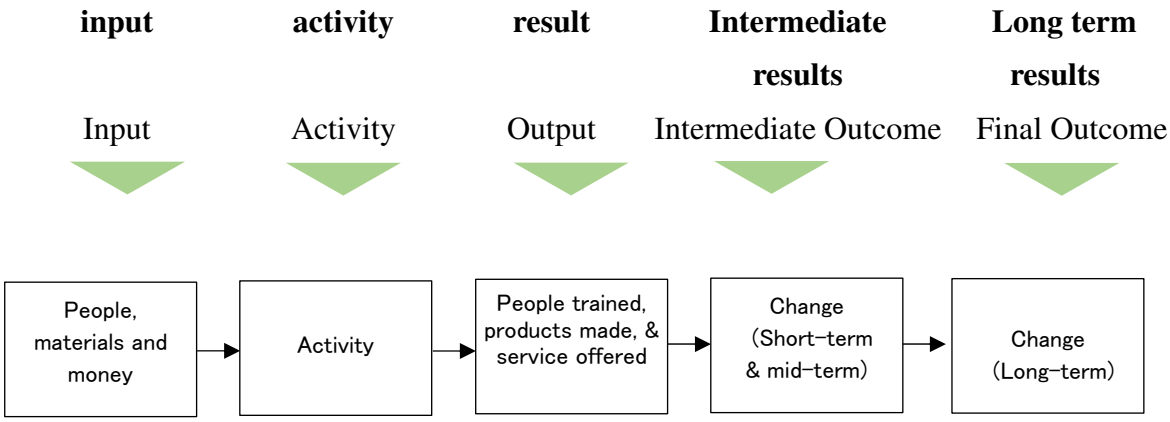
（Source）Ryu,Y and Sasaki.R (2010,2014)

Step 1: Discuss and agree on the purpose of the business

Stakeholders meet and agree on the purpose of the activity. Surprisingly, some people say, ``**Actually, the real purpose is different,'' or ``It's useful in this aspect, too,''** but those involved should have frank discussions and clearly agree on the main purpose. . The main goal is to meet the needs of the beneficiaries. It would be a good idea to specify this in the document.

Step 2 : Write the logic model

Create a 5-step logic model as shown below. It will be easier to write if you start by writing about the activities your organization is doing.



Necessary for implementing vocational training - Teacher - teaching materials - venue - operating funds	Implementation of vocational training (3 months in total)	100 people completed vocational training	30 people found employment.	3 of them started their own businesses after 5 years

Step 3 : Decide on indicators to measure inputs, activities, outputs, and outcomes

Decide on indicators to measure each. There can be one or more indicators.

- Number of instructors - Quantity of teaching materials - Availability of venue - Amount of money	・Whether or not vocational training has been implemented and the degree of success (5-level evaluation)	・Number of graduates	・Number of employed people	・Number of start-ups

Step 6 : Analyze the collected indicator values

Determine impact by applying one of the five impact evaluation designs described in this text. Current practice uses simple ex-post comparisons, which are too simplistic. Use a higher order strict design.

Step 7 : Make an evaluation based on the analysis results => Write the evaluation results

Write a conclusion: “The vocational training was very good/good/not good” or ``Vocational training is very satisfactory/satisfactory/unsatisfactory."

Changes in index values are just identifying facts, so please write your conclusion based on the amount of change using words that express value, such as **"good/bad"** or **"satisfactory/unsatisfactory."** Only when you write your conclusion in words that express value can you give it a sufficient "evaluation."

Step 8 : Write a recommendation if necessary

Although recommendations are not part of the evaluation, they are always required to be written. (1) Policy recommendations: ``Vocational training should be continued next season/should be stopped next season", (2) Business improvement recommendations: ``Recruitment methods should be improved," etc. Write a two-level proposal.

Step 9: Compile all into a report

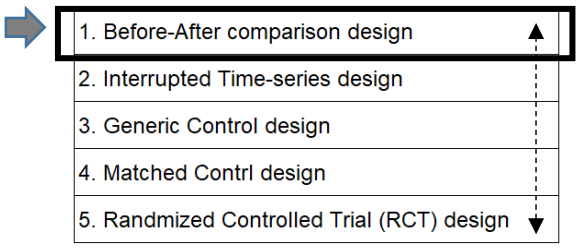
The information from Step 1 to Step 8 will be compiled into a report.

3. Explanation: From the perspective of formal evaluation

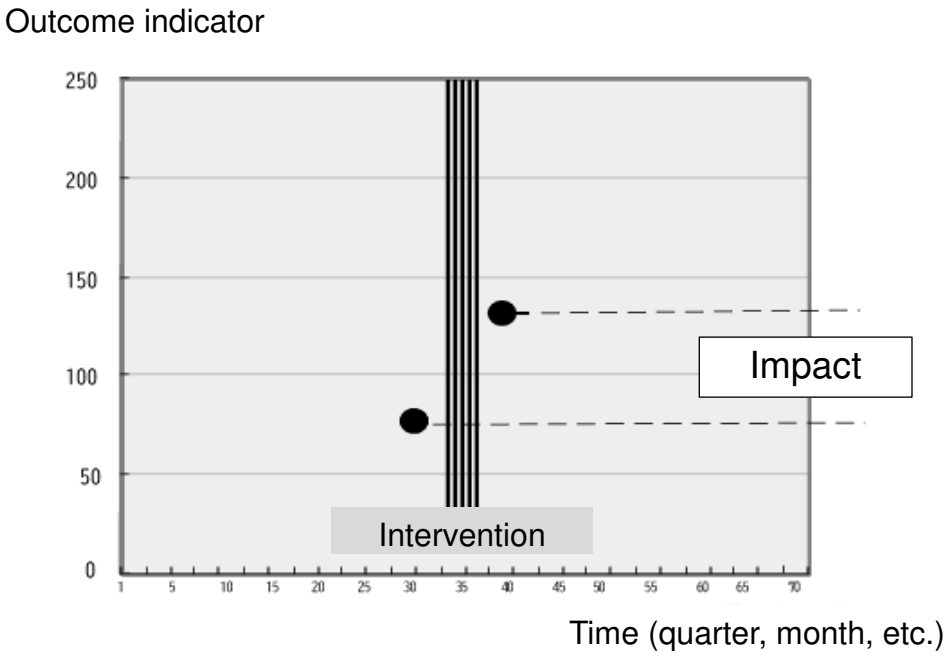
The difference between so-called impact evaluation and social impact evaluation is as follows.

> “ Social Impact Measurement ” is the same thing that has been called as Performance Measurement in the world of evaluation research. The analysis design used here is a before-after comparison design, which is a very basic (simple) design from the perspective of formal `` impact evaluation ' ' theory .
>
> Those practicing social impact evaluation are encouraged to apply the more rigorous designs described in this report. I hope that they will start with social impact evaluation and gradually move on to full-scale impact evaluation.

(Source) Based on Ryu and Sasaki (2010, 2014) “Theories and Techniques of Policy Evaluation” Taga Publishing .



1. Before-After comparison design



[Explanation]

Simply compare the indicator values before and after the event, and if there is a difference, it is assumed that there is a causal relationship. It is widely used because it is simple. However, the reliability of estimating the causal relationship is low because it is impossible to remove the influence value due to external factors that occurred before and after the event.

[Analysis test]

Dependent t-test (Paired t-test)

Application example 1 of pre-post comparison design

Primary Education Support Program (Ghana)

This project, implemented by the World Bank in Ghana, provides support related to "improvement of policy and management" and "improvement of physical infrastructure," including "improvement of school efficiency," "improvement of teacher teaching environment," and "related facilities." ・Improvement of educational materials" and the final outcomes are ``improved admissions and graduation results" and ``improved academic ability." The table below shows the average pre- and post-test scores for this project.

Average test score

Table G.2: Average tests scores: whole sample

	1988	*2003*	*t-stat*	*p-value*
Short English*	6.2	6.6	3.75	0.000
Short math*	5.5	5.9	8.16	0.000
Short local*	...	6.4		
Advanced English	12.3	13.2	4.16	0.000
Advanced math	8.7	10.1	6.93	0.000
Advanced local	...	15.5		
Combined English	17.7	19.2	5.28	0.000
Combined math	14.5	16.2	6.26	0.000
Combined local	...	21.1		

* Corrected for right censoring.

"Before" "After"

(Source) World Bank (2004), p.137

Regarding the table above, the World Bank's report states the following conclusions: ``Table G.2 ... shows the average test scores for 1988 and 2003." ... The last row of the table shows the t-test quantity and p-value for the difference between the two test mean scores. They show significant improvement in all subjects. (The data show a significant improvement in all test scores.) However, this analysis is a simple before-after comparison and does not include the impact values of external factors that would have existed during the relevant period or the effects of other related interventions. Unfortunately, there is no mention of this restriction. In addition, the statement ends with a technical statement (ordinarily used in academic papers) that ``significant improvements have been made in all subjects," which does not conclude that ``the elementary education support program has had an effect." I haven't. Since there was a long period of 15 years between before and after the event, and the influence of external factors cannot be denied, it can be inferred that the World Bank recognizes that it is dangerous to state any effects based solely on the World Bank's intervention. .

(Source) World bank (2004). Books, Buildings, and Learning Outcomes: An Impact Evaluation of World Bank Support To Basic Education in Ghana

Application example 2 of pre-post comparison design

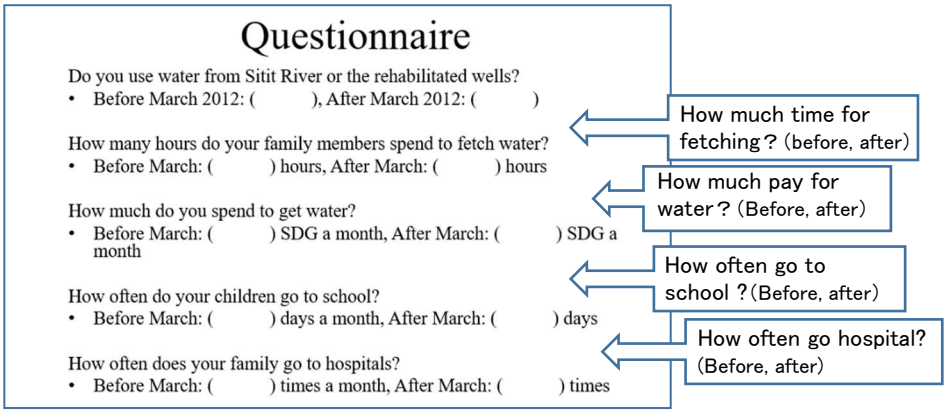
Evaluation of the effectiveness of well rehabilitation projects (Sudan)

<Before implementation>
Businessmen were drawing water from the river, filling it in bags, transporting it on donkeys, and selling it in the area.

<After implementation>
Children started coming to the well to fetch water.

With the cooperation of Japan, the ``Basic Human Needs Service Provision Project" (commonly known as K-TOP) was implemented in Kassala State, Sudan. As part of this , well repair work was carried out in Wad El Helew (district name) (total number of residents: 2,000-3,000, but the exact number is unknown) (March 2012). After implementation, a questionnaire survey was conducted among a sample of residents regarding the conditions before and after the well repair work.

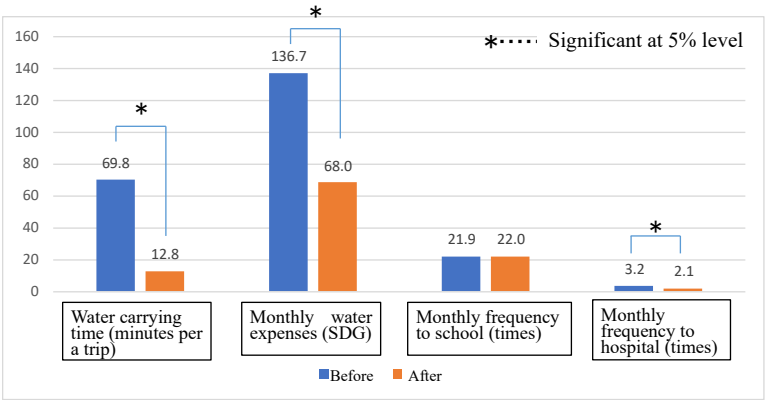
The questionnaire form is as follows, and it asks five simple questions.



The results were as follows.

Surveyed in September 2012 Sample: 62

	Before Mar. 2012 (a)	After (Sep.2012) (b)	Diffrences (c.) = (b) - (a)	T by the paired t test	Remarks
Water carrying time (minutes per a trip)	69.8	12.8	- 56.9	14.74	Statistially significant
Monthly water expenses (SDG)	136.7	68	- 68.7	6.64	Statistially significant
Monthly frequency of commuting to school (times)	21.9	22	0.1	1.00	
Monthly frequency of hospital visits (times)	3.2	2.1	- 1.1	5.80	Statistially significant



By comparing before and after,
"Time required to fetch water" (minutes/times),
“Water-related expenditure” (Sudanpond),
“Number of hospital visits” (times/month)
There was a decrease in both cases, and the decrease was statistically significant. ($p<5\%$ level).

On the other hand, the number of times children attended school (times/month) was approximately 22 times both before and after the survey, which was not statistically significant.

Overall, the well rehabilitation project can be judged to have improved the lives of residents in the area in many ways.

(Source) IDCJ (materials provided by Yasuyuki Kuroda, Chief Researcher (according to “ Impact Survey: Has the well rehabilitation project improved the quality of life for residents? ” (2014))

Application example 3 of pre-post comparison design

Elementary school rehabilitation support project (Djibouti)

A simple evaluation of an elementary school rehabilitation support project conducted by USAID uses before and after photos. Although it has a visually appealing effect, it cannot escape the criticism that it tends to be arbitrary.

Photo: USAID/Leslie McBride

BEFORE Guelleh Batal primary school did not offer an environment conducive to learning. The school lacked a boundary wall or any form of sanitation system, grossly endangering the health and well-being of students and teachers. Classrooms were run down and had minimal school materials and equipment.

Photo: USAID/Leslie McBride

AFTER USAID helped rehabilitate 12 classrooms, replacing doors and windows, repairing the roof, renovating the electrical system and installing new lights and fans. The exterior and interior walls were patched and painted, and the classroom floors were redone. Classrooms were also fully furnished with new equipment. A community outreach program now orchestrates maintenance of the school and its surroundings.

(Source) USAID. “Rehabilitation of Guelleh Batal primary school in Djibouti”.

Application evaluation of pre-post comparison design 4

Welfare service reinforcement project (Peru)

Problem location and evaluation results

1980 and 1990 , Peru's health sector was unable to provide adequate services. In order to improve this situation, the Peruvian government has launched the Health Service Strengthening Program.

1. Summary of measures

This program consisted of three parts: (1) Prior research and surveys, (2) Strengthening the organization and decentralization of the Ministry of Health, and (3) Strengthening health and medical facilities. Of these, Japan provided financing to support (3). The loan was signed in April 1994 and was disbursed in several installments until July 1999 , with favorable terms of a total of approximately 2.2 billion yen, an interest rate of 3.0% , and a repayment period of 30 years.

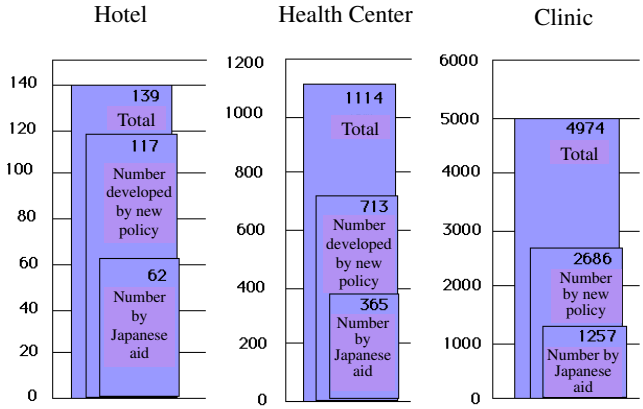
This program has resulted in the provision of equipment and materials as shown in the table below. The contribution of Japanese loans was also shown in the table.

Number of hospitals, heath centers and health posts developed by Japanese aid

	Total(a)	Number developed by the project(b)	*(b)/(a)*	Number developed by Japanese aid(c)	*(c)/(a)*
Hospitals	139	117	*(84%)*	62	*(45%)*
Health Centers	1,114	713	*(64%)*	365	*(33%)*
Health Posts	4,974	2,686	*(54%)*	1,257	*(25%)*
Total	6,227	3,516	*(56%)*	1,684	*(27%)*

Source)MINSA

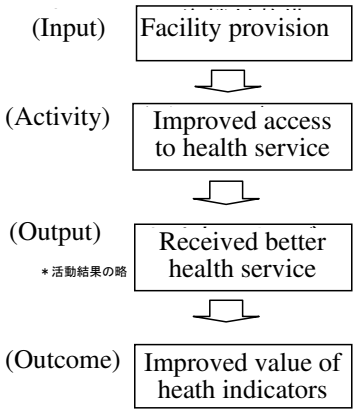
139 hospitals , 117 were improved through this program , and 62 of these, nearly half, were improved with Japanese loans . Looking at "health centers," which are smaller than hospitals but larger than public health centers, Out of a total of 1,114 houses, 713 were developed through this program , of which 365 were built with Japanese loans, accounting for 33% of the total number . Finally, looking at public health centers, out of a total of 4,974 , 2,686 were developed through this program , and of these, 1,257 were developed using Japanese loans, or 25% of the total number . The table can be expressed as a graph on the next page, so please check it out.



2. Evaluation results

In order to evaluate the effectiveness of this measure, those conducting this evaluation used a pre-post comparison design. In addition, although Japan provided loans from 1994 to 1999 , it would take some time for equipment and materials to actually be constructed using the loans, so 1994 was designated as the "preliminary stage ." , to evaluate the impact of Japan's loans, with 2000 as the "post-stage" .

The diagram below shows the causal relationship leading to the impact of this program, as envisioned by the designers and evaluators of this program.



Facilities will be improved → the number of users of the facilities will increase → the health condition of the people will improve. The evaluators adopted the number of facilities, number of facility users, and various health indicators as indicators for evaluating each.

First of all, regarding the maintenance of facilities, 1. Now that we have confirmed that this has been achieved, we will next examine the number of users of the facility. The following is data regarding the usage status of the facility.

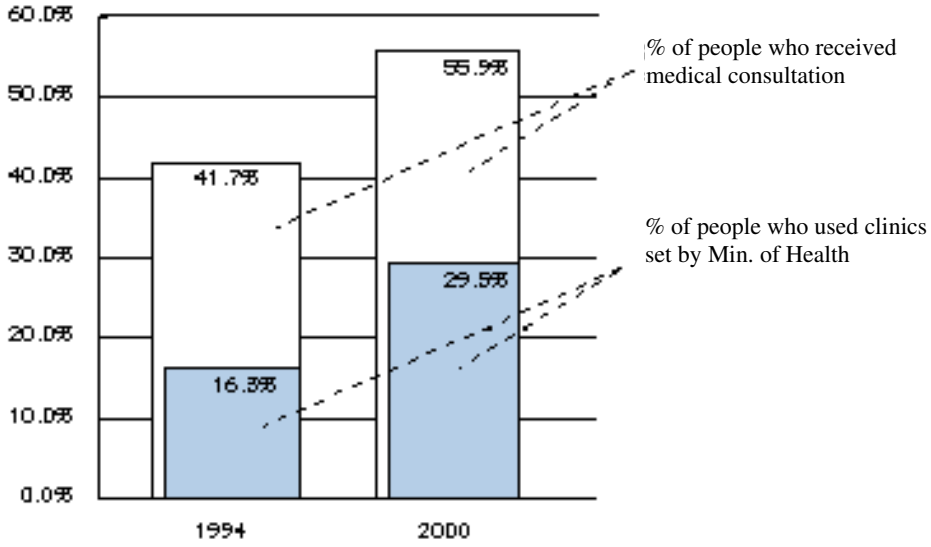
Comparison of use of heath service (1994, 2000)

	1994	2000
(1) % of people who received medical consultation	41.7%	55.9%
(2) % of people who received medical services at clinics set by the Min. of Health	16.3%	29.5%
(Reference) % of people who used clinics set by Min. of Health (=(2)/(1))	39.0%	52.8%

Source) ENNIV (The National Standard of Living Survey)

Looking at the percentage of Peru's population that received some type of medical treatment, it rose from 41.7% in 1994 to 55.9 % in 2000 , an increase of approximately 14.2% . On the other hand, the

number of people who answered that they received medical treatment at clinics newly established by the Peruvian Ministry of Health through this program increased by 13.2% from 16.3% to 29.5 % , indicating that It can be said that most of the improvements were achieved through this program. This is graphed as shown below, and it can be seen that the overall boost is mostly due to the boost caused by this program.



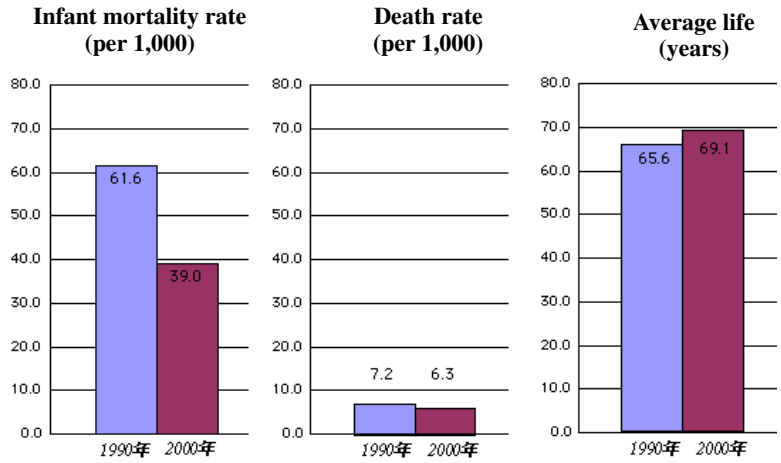
Finally, regarding the impact on the nation's health status, the evaluators presented the following data. Originally, we should have collected data for 1994 and 2000 , but the data we were actually able to collect was for 1990 and 2000 , so we are presenting that data.

Comparison of values in heath indicators

Indicators	1990	2000
Annaul Population Growth(%)	1.9	1.7
Crude birtth rate (per 1,000 po	29.0	23.7
Infant Mortality (per 1,000 pop.	61.6	39.0
Life expectancy at birth (year)	65.6	69.1
Crude death rate(per 1,000pop	7.2	6.3

Source) INEI. Peru: Estimates and Projections of the Population by Calender Year and Basic Age, 1980-2025; Lima: INEI (National Institute of Statistics and Information) 1995.
INEI, Peru: Status of the Peruvian Population 2000. Lima: INEI, 2000.

For example, the infant mortality rate decreased from 29 per 1,000 people (1990) to 23.7 per 1,000 (2000) . The infant mortality rate was 61.6 per 1,000 people (1990), but has declined to 39.0 per 1,000 (2000). Additionally, while the general mortality rate has fallen from 7.2 to 6.3 per 1,000 people, average life expectancy has increased from 65.6 to 69.1 . This is shown as a graph again as follows.



Regarding the improvement of these indicators, the evaluators concluded as follows:

Given that several programs were being implemented in parallel with support from other donor countries, it is difficult to determine the direct impact that Japan's financing had on Peru's health sector as a whole. difficult. However, it is important to pay attention to the possible causal relationship: equipment provision → improved access to health services → receiving better services → improving health indicators. And between 1990 and 2000 , infant mortality rates, general mortality rates, and other indicators improved. Since Japan provided the largest amount of financing in Peru's health sector in the 1990s , it may be safe to assume that Japanese financing contributed to the improvement of health indicators.

3. Advantages, limitations, and considerations regarding application in Japan

The advantage of this method is that it is only necessary to refer to data for the area where the study was carried out. For the matching design, data on the implementation area and comparison area (2 time points x 2 areas) were required in the pre-stage and post-stage. In the pre-post comparison design, the data is from the pre- and post-implementation regions (2 time points x 1 region). The statistical equalization design requires data from the implementation area and comparison area at the ex-post stage (1 time point x 2 areas), but from a practical perspective, it is necessary to have ex-ante and post-data from the same area used ex-post and ex-post. It's much easier to obtain.

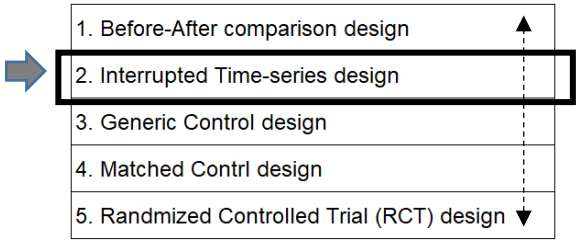
A limitation of this method that should be pointed out is that it cannot be said to prove any kind of causal relationship. Even if the index value improves between before and after, it cannot be said that this is due to the measures that you have implemented. In other words, this method relies solely on the one point that the causal relationship (logical design) assumed in advance must be correct.

The following points should be kept in mind when applying this in Japan. In Japan, the current situation is that logic models are rarely considered or clarified. For example, is the purpose of road

construction to shorten travel time or to create jobs through construction? Is the purpose of ODA to alleviate world poverty, or is it to create a foundation for Japanese business expansion? Both are fine, but a major prerequisite for evaluating effectiveness is to first reach an agreement on the objectives and cause-and-effect relationships among the parties involved by creating a logic model. Note that if there are multiple objectives, the logic model will branch out along the way, and accordingly, the number of indicators to be collected will also be multiple.

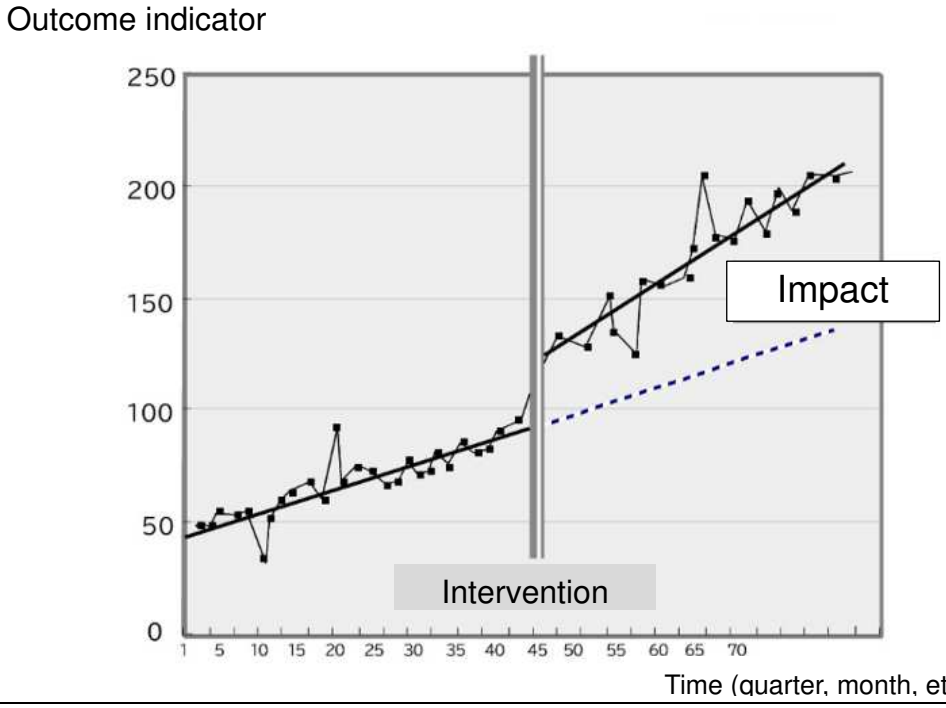
The author reconstructed and created the explanatory text based on the description in the already published Japan Bank for International Cooperation (2002) ``Yen Loan Ex-post Evaluation Report 2002 " (English). The PDF file for this evaluation can be downloaded from below.

http://www.jbic.go.jp/japanese/oec/post/2002/pdf/project_58_alle.pdf



2. Time series design

(Interrupted Time-Series Design)



[Explanation]

Determine long-term trends before the policy intervention, and if the trends change after the policy intervention, infer the existence of a causal relationship. However, it is not very reliable because it cannot remove the influence of external factors other than long-term trends.

[Statistical test]

Regression Analysis

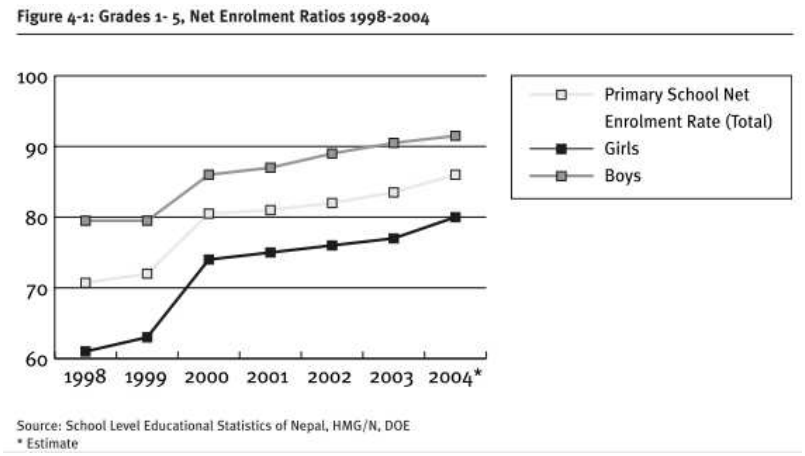
Evaluation example using time series design :

Primary education support project (Nepal)

Outline of the problem and measures

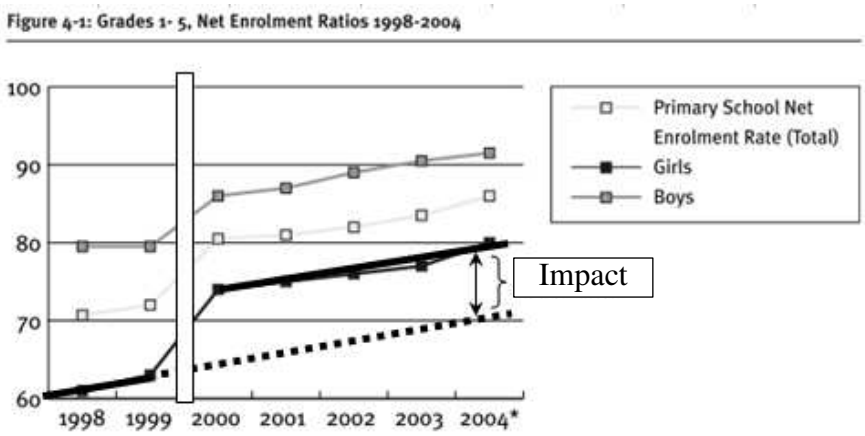
In Nepal, the ``Basic/Primary Education Program II '' was launched in 1999 . The objectives were (i) to improve the quality of primary education, (ii) to increase access to primary education, and (iii) to improve the capacity of relevant institutions. The specific contents were wide-ranging, including school building construction, teacher training, curriculum improvement, textbook distribution, and staff training for related organizations.

Figure 4-1 shows the Net Enrollment Rate (NER) from 1998 to 2004 . The program will begin in the second half of 1999 .



Evaluation results

The evaluation results stated that ``The three indicators of overall, boys, and girls were higher in 2004 than in 1998 , confirming a pattern of improving enrollment rates and coverage rates.'' The actual evaluation report only includes this description, but by applying a time series design to this graph, the impact can be estimated as follows. It can be estimated that the impact for girls was approximately 9 %.



(Source) Danida (2004) Nepal: Joint Government – Donor Evaluation of Basic and Primary Education Programme II

Introduction: This case is not a pure time series data case. We use time series (20 years) x multiple region (50 states in the United States) data (this type of data is called panel data). However, the regression analysis procedure is exactly the same as for simple time series data.

Evaluation example using regression analysis of panel data : Effects of policy changes regarding drunk driving (USA)

Problem location and evaluation results

How much will traffic accidents be reduced by tightening regulations regarding drunk driving? In the United States in 2000, a major debate arose over the 2000 Federal Transportation Appropriations Bill. Supporters of stricter regulations include Mothers Against Drunk Driving and the American Medical Association, which support Democratic Senator Frank Lautenberg and others. I aimed for Opponents of stricter regulations include restaurant and bar industry groups such as the American Beverage Institute, which supported Republican Whip Tom DeLay and others to block tougher regulations.

As a result of the evaluation, it was concluded that measures to strengthen blood alcohol concentration standards for arrests would be effective in reducing traffic accident deaths. However, other policies were also found to be effective in reducing traffic accidents. In other words, it is possible that other measures can have similar effects, so we conclude that we should consider which measures are best to adopt in order to maximize the benefit (welfare) of society as a whole. .

1. Summary of measures

Normally, drunk driving regulations are implemented by estimating blood alcohol concentration from exhaled breath. Previously, the standard for arrest was a blood alcohol concentration of 0.1% or less, but the bill seeks to tighten that to 0.08% or less.

The compromise signed by then-President Clinton in October 2000 was as follows: Although the GOJ will not force all states to tighten regulations to 0.8%, states that do not accept this tightened regulation will receive a 2% reduction in federal highway funding in fiscal 2003. It will be reduced by 4% in fiscal 2004, 8% in fiscal 2006, 10% in fiscal 2008, and at this rate thereafter. On the other hand, if the government accepts stricter regulations by fiscal 2007, the entire amount of the subsidy that was reduced up to that point will be distributed in one lump sum. Since this decision, a number of states have accepted tightening the regulation to 0.08%, but as of September 2002, 15 states have still not accepted it. This variation is used to evaluate the effectiveness of the measures.

2. Business details (intervention)

In the United States, the number of fatal traffic accidents (hereinafter referred to as ``traffic fatalities")

has been decreasing continuously since the early 1980s . The number of traffic fatalities decreased from 2.21 per 1,000 adults in 1982 to 1.72 in 2000 . As this trend continues, a simple comparison of ex-ante and ex-post figures shows that the latest regulatory tightening law was effective even when it actually was not. In addition, it was predicted that the reason for the continuing downward trend may be due to the effects of various other measures and activities being implemented.

These are (1) the introduction of a graduated licensing program targeting young people, (2) the Dram Shop Law, and (3) the Seat Belt Mandatory Law. law), etc. Furthermore, the rise in retail prices due to the beer tax increase may also have an effect. Another factor may be the improved accuracy of blood alcohol concentration measuring devices. Minors (under 21 years of age) are immediately arrested if they show even the slightest reaction to alcohol before even discussing whether it is 0.1% or 0.08%, and it is becoming more and more difficult to get away with it every year. Improvements in employment conditions due to the long-term economic expansion may also be having an effect. However, the degree of impact may vary depending on the economic situation of each state. Additionally, differences in the number of years between each state accepting the tightened regulations and conducting the current evaluation may also have an impact on the degree of effectiveness.

In order to evaluate the effects of the tightened regulations while taking into account the influence of these various factors, we used regression analysis rather than a simple ex-post comparison.

The variables used in the regression analysis are as follows. All policy variables are expressed as dummy variables. In other words, if the policy has been implemented in the state, "X=1", and if it has not been implemented, "X=0" (excluding beer tax and the number of chapters of "mother's association").

(1) Policy variables

Various policies
・Number of arrests for drunk driving (blood concentration 0.08 % or higher)
・Number of arrests for drunk driving (blood concentration 0.1 % or more)
・Arrest of underage drunk driving
・Cancellation of license
・Number of days in prison
・Breath alcohol concentration test prior to arrest
・Alcohol Retailer Regulation Act
・It is prohibited to open alcoholic beverages in the car.
・The legal drinking age is set at 21 years old.
・Beer tax (Cents 1999)
・Seat belt wearing regulations
・Gradual grant acquisition
・Number of branches of "Mothers' Association to Eliminate Drunk Driving"

(2) “Other control variables”

Other elements	Average value
・State average income (000 $)	24.110
・State unemployment rate	6.077
・Average age of drivers in the state	43.070
・Average driving distance of state drivers ('000 miles)	12.070

3. Get the data

The data is 50 states in the United States x 20 years = 1,000 data. In other words, it is data from multiple regions and multiple points in time. It can be understood that because there is a large amount of data, it was possible to include a large number of policy variables (X).

Data can be obtained from: The number of traffic accident deaths was obtained from the National Highway Traffic Safety Administration's Fatal Accident Report System (RARS). For information on "various policies," see "Digest of State Alcohol-Highway Safety Related Legislation (NHTSA, 1982-2000)" and "Traffic Safety Information." (Traffic Safety Facts), information obtained through web and personal inquiries. Information on ``other factors'' was obtained from the websites of the Government Bureau of Economic Analysis, the Government Bureau of Labor Statistics, the Government Statistics Office, and the annual report of the Department of Transportation.

50 states x 20 years = 1,000 data

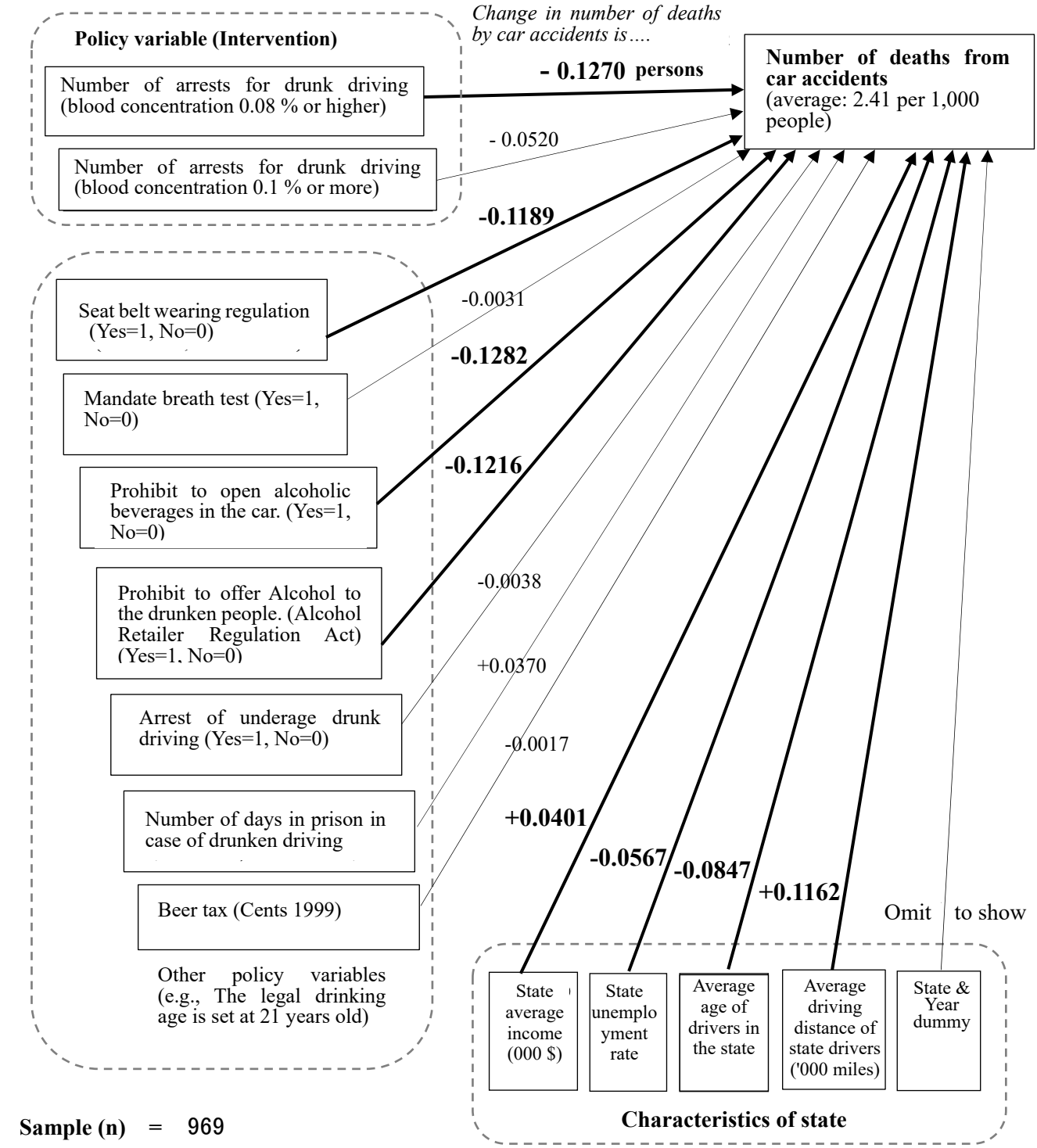
4. result of analysis

The above data for each state was entered into a regression analysis formula to calculate the ``slope'' for each policy. Each ``slope'' represents the strength of the policy effect. Regarding blood alcohol concentration regulations, some states have adopted 0.10% as the standard while others have adopted 0.08% , so the effect of each policy is calculated as a "slope." The results of the regression analysis are as follows.

Visualization of regression analysis results

(The average number of deaths due to car accidents for all states and all periods is 2.41 people /per 1,000 people)

How to read the figure: When the value of each policy variable changes from 0 (no) to 1 (yes), Y (number of deaths due to car accidents) changes by the slope (value above the arrow). For example, if "arrested for blood alcohol concentration 0.08%" changes from 0 (no) to 1 (yes), the number of deaths from car accidents will decrease by 0.127 people/1,000 people.



Sample (n) = 969

R^2 = 0.945

(Note) **Bold line : Significant at 5% level.**

Thin line : Not significant at 5% level.

(1) **Regulations that set the blood alcohol concentration standard for arrest at 0.08%** were evaluated to **be effective in reducing traffic accident deaths by 5.3% .** On the other hand, regulations that set the blood alcohol concentration standard for arrest at 0.10% were evaluated as having the effect of reducing traffic accident deaths by 2.2% (although this cannot be said to be statistically significant). Therefore, by tightening the regulation for arrests from 0.10% or less to 0.08% or less, traffic accident deaths can be **further reduced by 3.1% .** (The calculation is as follows)

The average number of deaths due to car accidents for all states and all periods is 2.41 people/per 1,000 people.
(1) The slope for "arrests with blood alcohol concentration 0.08%" (the number above the arrow in the diagram) is 0.127 people less. **- 0.127 people ÷ 2.41 people = -0.053 =5.3% decrease.**
(2) The slope for "arrests with blood alcohol concentration of 0.10%" (the number above the arrow in the diagram) decreases by 0.052 people, so **- 0.052 people ÷ 2.41 people = -0.022 =2.2% decrease.**
The difference: (1) - (2) = 5.3 % - 2.2% = 3.1%

Furthermore, other measures were evaluated as having the following effects.

(2) Required seat belt laws were similarly assessed to have reduced traffic fatalities by 5.1%. In addition, measures such as banning the opening of alcoholic beverages in cars and banning the provision of alcoholic beverages to intoxicated people were evaluated as having reduced traffic accident deaths.

(3) In addition, the state's average income, unemployment rate, average age of drivers, average driving distance, etc. were evaluated as having an influence.

After the above evaluation, we also calculated the number of traffic accident fatalities in certain specific cases, such as the number of fatalities occurring only on weekends and at night, the number of fatalities occurring only among young people, and the number of fatalities occurring while under the influence of alcohol. We are attempting to evaluate the effects of the various measures mentioned above in detail by calculating only the numbers. Furthermore, we attempted to calculate how the effects of the above measures differ depending on differences in income, married/unmarried status, age, and mileage. The conclusions of this paper were as follows.

Conclusion of this impact evaluation

It can be concluded that the measure to strengthen **the standard for arrest based on blood alcohol concentration from 0.1% to 0.08% was effective in reducing traffic accident deaths .** However, other policies were also found to be effective. Regarding whether or not the 15 states that have not

tightened regulations should accept stricter regulations, other measures may be able to achieve the same degree of effect, so it is important to maximize social welfare for society as a whole . In the future, we should consider which measures are best to adopt by applying cost-benefit analysis.

5 . Advantages, limitations, and considerations regarding application in Japan

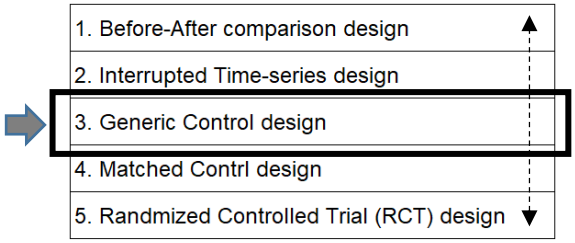
The advantage of this method is that there is no need to collect index values or conduct questionnaires in advance. Furthermore, in many cases , existing data can be used even at the post-event stage without conducting a questionnaire , making it even easier.

A limitation of this method is that it is unclear whether the prepared data adequately captures the actual situation. Because only available data are used, there are evaluations that use almost the same data sets for evaluations of smoking and cancer incidence, as well as evaluations of obesity and mortality, but there is no particular relationship between the two. It's simply a matter of data availability. Also, the evaluation results will be quite different depending on whether you use data from the past 10 years, 15 years, or 20 years, but an explanation was given as to why that period was chosen. It's unlikely.

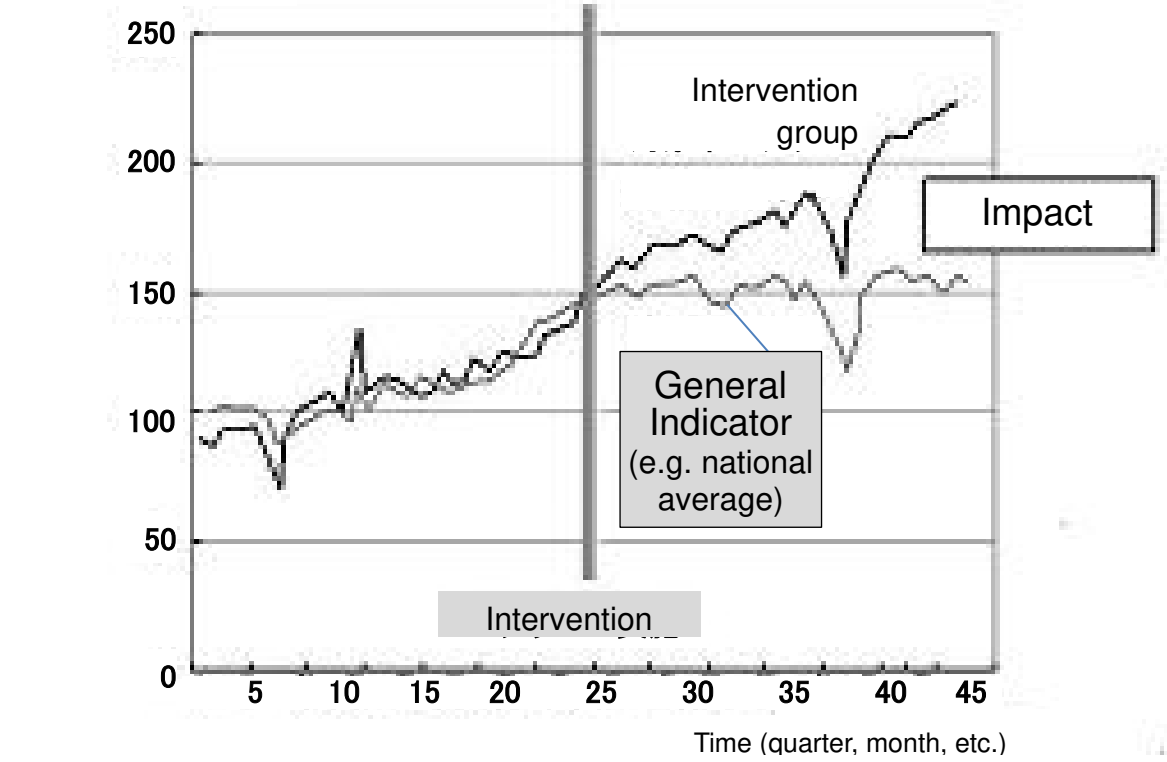
In other words, despite the scientific image that the general public has, there is considerable room for arbitrariness in regression analysis. However, this is not a problem with the regression analysis method itself; it must always be kept in mind that in many cases there are restrictions on the existence or availability of data.

In addition, in the 2020s, **"visualization" of statistical analysis results has become** a hot topic. Regression analysis should also **be explained visually using simple drawings using boxes and arrows, rather than trying to explain it with mathematical formulas written in Greek letters and huge tables that only those with specialized knowledge can understand** . At the time when this regression analysis was conducted, it was not common to display the results in drawings, so the commentator (Sasaki) used drawings to explain the results.

Source: Eisenberg, D (2003). “Evaluating the Effectiveness of Policy Related to Drunk Driving” In *Journal of Policy Analysis and Management* . 22(2):225-248



3. Generic Control Design



[Explanation]

General index values such as the national average value and all prefecture average values are used for comparison. Since it is possible to remove the influence values due to external factors to some extent (because the general index values are likely to be influenced to some extent to the same extent as the target area), there is a certain degree of reliability in identifying the existence of a causal relationship. can be secured. It's fairly easy to use.

[Certification test]

Visual judgment (Eyeball judgment)

Application examples of general indicator design:

Alberta Business Plan (Canada)

Problem location and evaluation results

In Japan, the financial bankruptcy of local governments has recently become a serious topic of discussion. The conventional wisdom that public organizations do not go bankrupt even if private companies go bankrupt is being called into question, and the possibility that local governments will actually go bankrupt is increasing. A good example of a municipality that has escaped such a financial crisis is the case of Alberta, Canada. Moreover, Alberta emerged from fiscal crisis with the best public services in Canada at the lowest tax rate.

Governor Ralph Crane

Alberta's Governor Ralph Crane, a former television newscaster, boldly introduced private-sector management methods into his provincial government. Based on that idea, we formulated the "Business Plan for Alberta." The business plan, which was implemented on a thorough results-based basis, adopted a "general indicator design" for several strategic objectives to measure impact. As a result, we succeeded in evaluating the impact of the Alberta government's policies by removing a considerable amount of influence from external factors.

1. Summary of measures

In 1993, the Alberta Business Plan was developed. It is a tree diagram strategy in which three "core businesses" are set under a single "mission," and a total of 18 individual goals are set below that. One of the individual goals is ``13 : Ensure the safety of Albertans and ensure that Alberta is a safe place to live, work, and raise a family." Specifically, the following strategies were developed and implemented.

1) The Alberta Metropolitan Police Service will focus its resources (financial, human, and time) on preventing violent crime. It will also promote local crime prevention activities and expand the participation of local residents in police activities.
2) The Family and Social Services Agency supports individuals in achieving financial independence. Keep your child safe. In particular, provide early warning and intervention of crimes against children, respond to the living needs of Aboriginal people, and provide temporary accommodation where necessary.

2. Evaluation results

The ``general indicator design'' has been applied as an evaluation method for some of the 18 individual goals, and the general indicator design was also applied to this 13th individual goal. Below is an overview of the evaluation mechanism.

Individual goals

``13 : Keeping Albertans safe and ensuring that Alberta is a safe place to live, work and raise a family."

Performance indicators

The following crime incidence rates (2 types). (Furthermore, the crime rate is set only for minors.)

①Number of violent crimes per 10,000 people

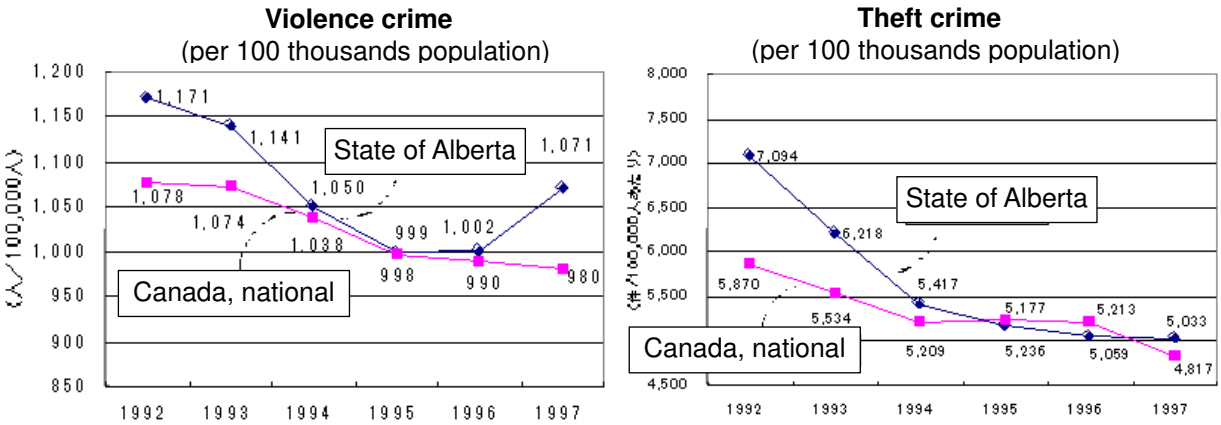
②Theft crime per 10,000 people (Property Crime) number of victims

Indicator description

Crime rates are a direct indicator of whether Alberta is a safe place.

Numerical goal

Reduce it to below the national average by 2000 .



The evaluation results at the end of the strategy period are as follows.

Both violent crime and property crime rates have improved steadily since 1992 (base year) at a pace that exceeds the national rate of improvement. However, although violent crime began to increase in 1997 , the Alberta government says there has been no significant change in the improving trend [1].

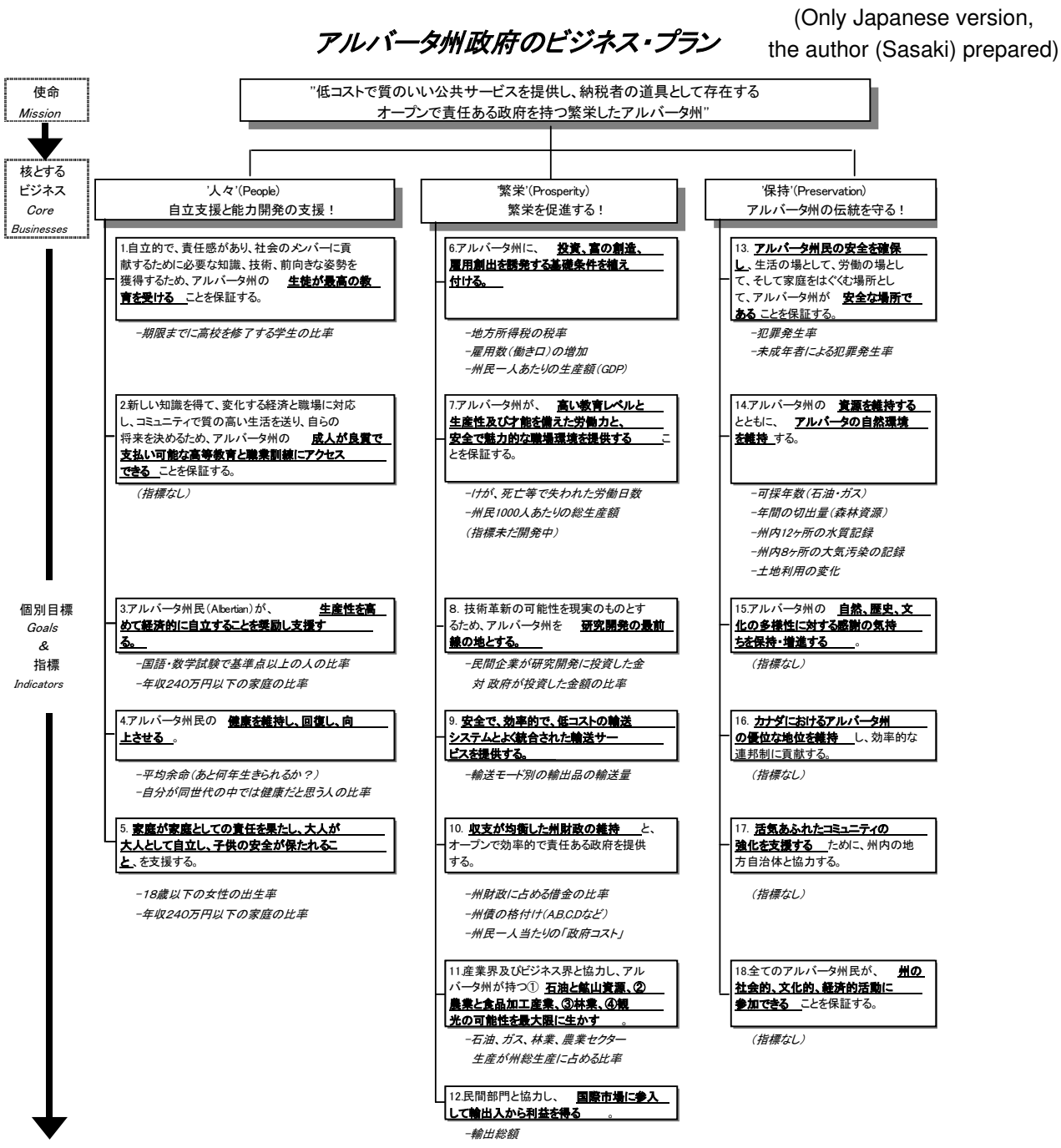
3. Advantages and limitations

By adopting a general indicator design, we were able to remove the influence of external factors to a considerable extent and successfully evaluate in a relatively pure manner whether or not the Alberta government's policies were having an effect. If external factors, such as global economic trends or the national impact of Canadian federal government policies , affect Alberta's indicator values, it is

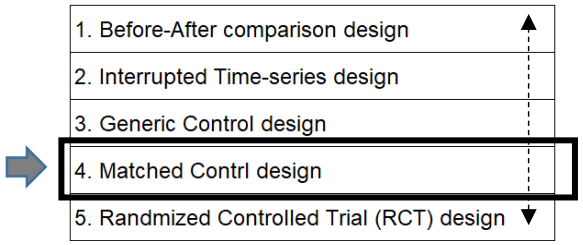
[1] AlbertaTreasury, *Measuring Up Report 1999*

assumed that the national indicator values will be affected to the same extent. Therefore, **if Alberta's indicator values are more improved than the national average, this can be considered an effect (impact) of the Alberta government's policies** . The state's business plan (strategic plan or comprehensive strategy) is also available as a reference and is listed below. Rather than **the ``to-do list'' of ``build XX roads'' and ``host XX tournaments'' seen in Japanese local governments, it is** unique in that it **consists only of indicators that express the happiness of residents.**

(Only Japanese version, the author (Sasaki) prepared)

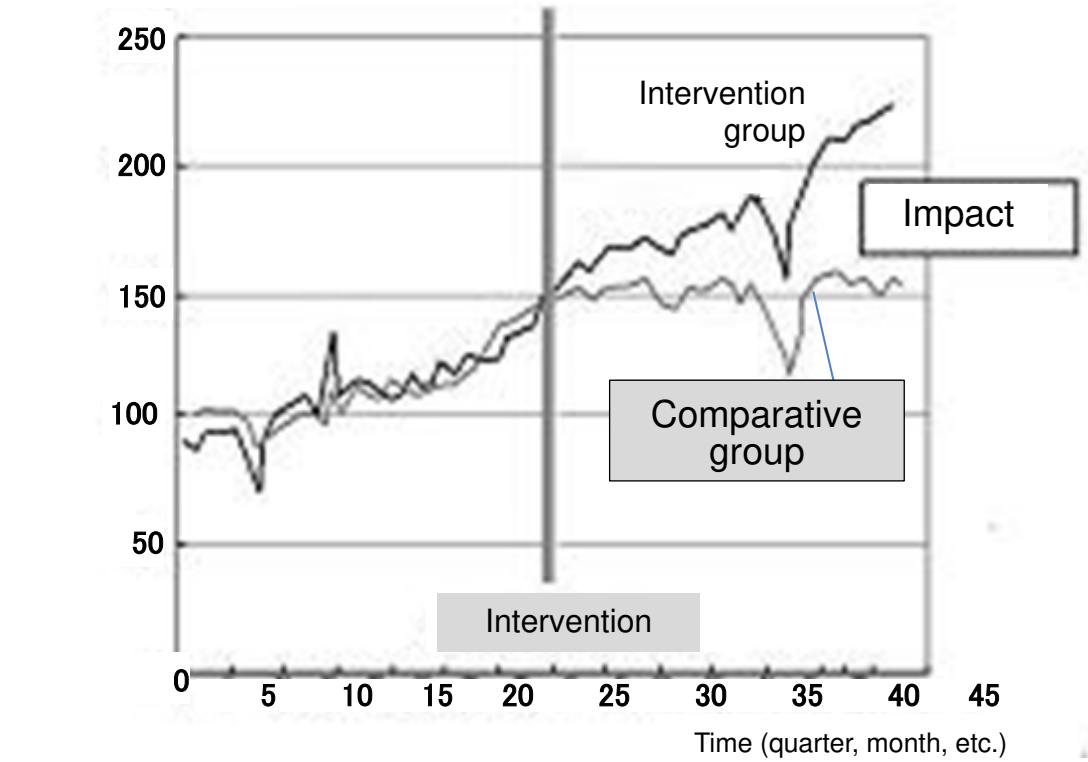


出所)アルバータ州政府ホームページ
http://obm5.treas.gov.ab.ca/comm/perfmeas/measupgu/gra19.gif



4. matching design

(Matched Control Design)



[Explanation]

Select groups that are as similar as possible and use them for comparison. Since both groups are considered to be affected to the same extent by external factors, a high degree of reliability can be ensured to identify the existence of a causal relationship.

[Certification test]

Two-group significance test (unpaired t-test)

Independent t-test

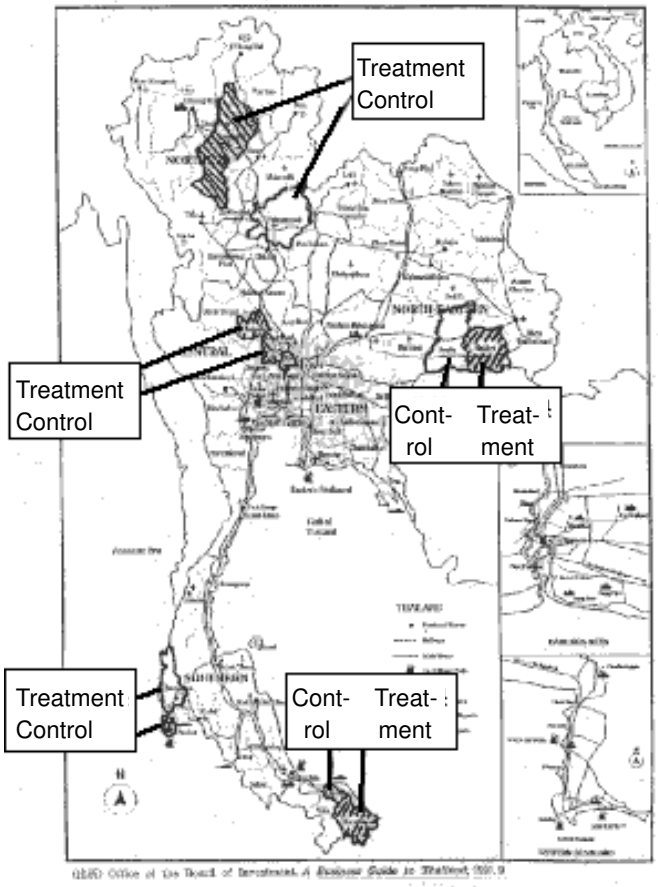
Matching design application example 1:

Evaluation of trial decentralization program (Thailand)

Problem location and evaluation results

Thailand is promoting decentralization and is currently conducting an evaluation using a matching design to evaluate the effectiveness of a trial decentralization program. As shown in the diagram below, we have decided on a comparison target for each of the five implementation prefectures based on area, population, industrial structure, distance from the metropolitan area, etc., and are continuously monitoring index values.

Treatment group (5 districts) and Control group (5 districts) for the decentralization policy trial phase in Thailand



Advantages, limitations, and considerations for application in Japan

The advantage of this method is that it is simple, and data for each prefecture compiled and published by the central government can be used, so data availability is high. On the other hand, a constraint is that there are no neighboring prefectures that have exactly the same conditions as the implementing prefecture other than being subject to the measures, so the indicators selected for matching (area, population, industrial structure, This means that factors other than distance from the metropolitan area (such as distance from the metropolitan area) may have a significant impact on the effectiveness index value.

However, in Japan, even this simple method of matching design has not been applied, and is still not generally used. For example, **attempts have been made to create special zones for structural reform in limited areas, but in this attempt, it would be better to select districts that are as similar as possible in other respects than the implementation of the measures and use them as comparison targets**.

Matching design application example 2:

Effects of four types of measures related to primary education (Philippines)

Problem location and evaluation results

High dropout rates and poor learning outcomes are problems in many developing countries. The situation is similar in the Philippines, where approximately 25% of children drop out before completing elementary school (6th grade). Additionally, research shows that children retain less than half of what they are taught. In order to improve this situation, a project was implemented that combined three types: (1) provision of free learning materials based on proficiency level, (2) implementation of school lunches, and (3) strengthening of cooperation between teachers and parents.

As a result of the evaluation survey, it was found that the combination of ``provision of free learning materials according to proficiency level" and ``cooperative activities between teachers and parents" was effective in improving the dropout rate in elementary schools. On the other hand, among the measures tested in this evaluation survey, the one that was not found to be effective in improving dropout rates was ``implementation of school lunches." In addition, when we calculated the unit cost, we found that it was cheaper to provide learning materials based on proficiency levels than to provide school lunches, so we decided to provide free learning materials based on proficiency levels. We are recommending expansion.

1. Evaluation overview

DIP) implemented by the Philippine government from 1990 to 1992 , the effectiveness of several combinations of the three types of measures mentioned above was evaluated.

The dropout rate is calculated by subtracting the dropout rate for the year after the program from the dropout rate for the year before the program at the implementing school (this difference is a rough estimate of the improvement rate). We then calculate the similar rate at the comparison school and subtract that rate from the improved rate. The rate that remains is the pure rate of improvement caused by program implementation. (double subtraction method) [2].

The selection of sample schools was carried out through the following three stages.

(1) Two low-income provinces that can be said to be similar were selected from each of the five regions that make up the Philippines (matching). The matching criteria are: (1) education index, (2) health index, (3) housing index, (4) unemployment rate, and (5) household expenditure level.

(2) 5 regions x 2 prefectures = Three schools were selected from each of the 10 prefectures that met the following conditions: 1) have a high dropout rate, and 2) have no existing school lunch program. (5 regions

---

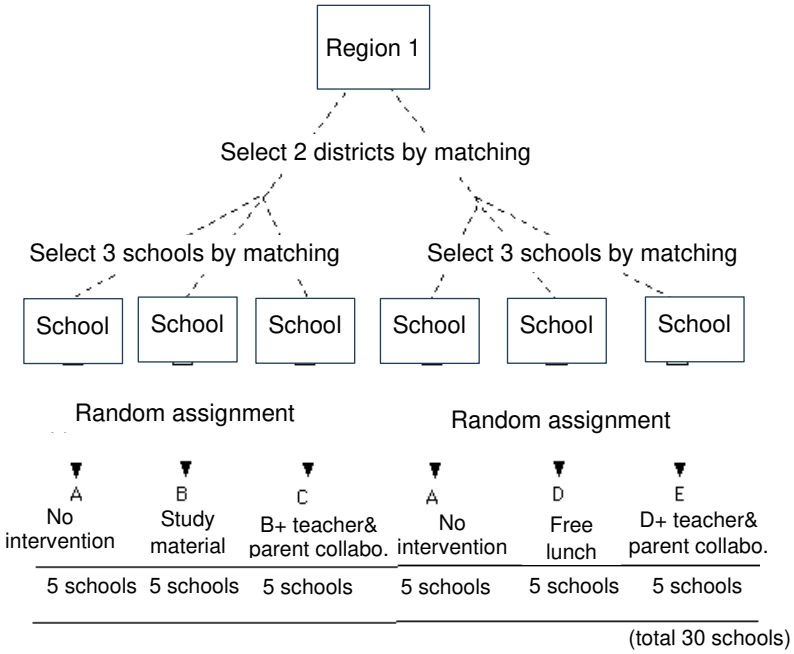
Furthermore, the following regression analysis was conducted using academic performance as the dependent variable.

"Academic achievement (current semester)" = "Academic ability (first semester)" + "Individual characteristics" + "Family characteristics" + "Learning environment" + "Class environment" + "Program implementation status" + Error

x 2 prefectures x 3 schools = 30 schools)

(3) From here, each program is assigned. Three schools in one of two prefectures selected from a certain region: A. No intervention (do nothing) B. Free distribution of proficiency-based teaching materials C. Proficiency-based teaching materials + teachers and parents Implementing collaborative activities, one of which was assigned. The three schools in the other prefecture were assigned to either A. Do nothing for comparison, D. Provide school lunches, or E. Provide school lunches plus collaborative activities between teachers and parents .

As a result, out of a total of 30 schools, 5 schools each implemented programs B , C , D , and E , making a total of 20 schools, and 10 schools were selected for comparison (A) , which did not do anything. (See diagram below)



This assignment for 30 schools were made at five (5) regions.

- implementation indicator values (baseline data) were collected in 1990-1991, and the program was implemented in 1991-1992 . Post - hoc data were collected subsequently (1992-1993) . As a result, we were able to collect detailed data from 29 schools [3], 180 teachers, and approximately 4,000 students.

2. Evaluation results

Baseline data on dropout rates before implementing the program were as follows: Furthermore, there are also academic achievement test scores and other data, but they are not listed here. At the preliminary stage, it was confirmed that there were no differences between each group, except for school group E.

[3] The reason why one school dropped out is unknown as it is not mentioned in the report.

Baseline data (1990-91)

	A No intervention	B Study material	C B+ teacher& parent collabo.	D Free lunch	E D+ teacher& parent collabo.
Dropout rate	9. 56	9. 29	10. 01	8. 58	7. 02**

*Stastistically significant at 10 % level, **at 5 % lrevel, and ***at 1 % level.

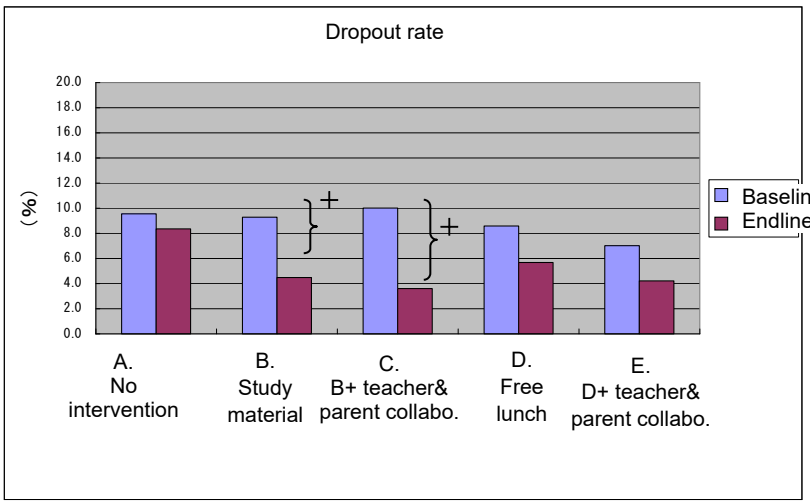
Below are the index values after implementation.

Endline data (with Change between1990-91 and 1991-92)

	A No intervention	B Study material	C B+ teacher& parent collabo.	D Free lunch	E D+ teacher& parent collabo.
Dropout rate	8. 36	4. 49	3. 61	5. 68	4. 22
Change in Dropout rate	−1. 2	−4. 8	−6. 4	−2. 9	−2. 8
P-value	*0. 328*	*0. 004****	*0. 005****	*0. 104*	*0. 11*
(*Difference* with A group)	n. a	−3. 6	−5. 2	−1. 7	−1. 6

*Stastistically significant at 10 % level, **at 5 % lrevel, and ***at 1 % level.

This can be expressed graphically as follows.



Two programs were confirmed to be effective: B. Providing learning materials based on proficiency level, and C. Combining this with collaborative activities between teachers and parents. On the other hand, it was confirmed that D. School lunch implementation cannot be said to have contributed to improving the dropout rate. Furthermore, we calculated the unit cost of implementation, and found that B. Provision of learning materials by proficiency level, which was confirmed to be effective, cost 90 pesos/person, and collaborative activities between teachers and parents cost 33 pesos/person. On the other hand, the cost of providing school lunches (D), which was not found to be effective , was estimated to be 946 pesos/person. Based on the results of this evaluation and the estimated unit costs, the evaluation implementer recommended that the World Bank should promote the expansion of the provision of learning materials based on proficiency levels. However, this is a recommendation regarding the dropout rate, and he added that since none of the methods

tested in this evaluation study could be said to have an impact when aiming to improve academic ability, other programs should be tried. ing.

The evaluator made the following three comments: (1) The finding that there was no effect on school lunches is a bit of an overstatement, and perhaps better results could be obtained if the target groups were narrowed down more. (2) The small sample size may have significantly influenced the determination of effectiveness. (3) Because the period between program implementation and evaluation was extremely short, it may not have been possible to measure effects that would appear over the medium to long term.

3. Advantages, limitations, and considerations regarding application in Japan

In this example, the effects of five combinations of measures, including no intervention, are compared. This will help determine which measures are most effective. Also, using this method when there are conflicting policy proposals would provide more meaningful information for administrative decision-making.

Regarding this example, it must be pointed out that the matching is loose. The number of indicators used for matching was too small, two or three. Therefore, when the baseline value (pre-implementation index value) was measured, the outcome index value (dropout rate) of Group E was already different. There should be more indicators when looking at matching. Also, the number of samples is small, as the evaluators themselves have pointed out, but they would like at least 25 or 30 for each group.

Points to keep in mind when applying this in Japan include the following. Unlike the United States, Japan applies uniform educational guidelines nationwide, so it should be relatively easy to prepare schools that show good matching for the purpose of evaluating policies. Furthermore, if the evaluation is carried out in one prefecture, rather than in five regions as in this example, the evaluation results can be applied nationwide to a considerable extent without any problem. Considering this **situation in Japan, if several city boards of education in one prefecture cooperate, it is possible to easily secure the necessary number of schools that closely resemble each other** .

5. discussion

This is a personal story, but I reported this case at an academic conference. When I took questions after the presentation **, I was asked, ``Don't you have to do something this complicated to understand?''** I was **once scoffed at, saying, ``Even if you don't do this, an expert can tell by looking at it .''** However, social experiments are becoming more popular as people share the experience that what experts say is effective is actually ineffective when applied across the country. It is easy to claim to be an expert, but we should always be humble about the evidence obtained from social experiments.

Source) Tan, JP, J. Lane, and G. Lassibille , 1999, “Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments” *In World Bank Economic Review,* September.

ex-post matching design (statistical equalization design):

Effects of employment promotion programs (Czech Republic)

Problem location and evaluation results

Amid the recent global trend toward market economies, developing countries and former socialist countries are facing the situation of large numbers of unemployed people due to the privatization and downsizing of state-owned enterprises. To deal with this, employment promotion programs are sometimes implemented by the government, and in the Czech Republic five types of employment promotion programs were implemented with loans from the World Bank. These are (1) new graduate training programs, (2) reskilling programs (from a few weeks to up to 7 months), (3) short-term employment in public civil engineering projects, (4) financial support for new hires, and (5) individual This is financial support for starting a new business. Were these effective in improving employment rates?

The evaluation results showed that although large and small effects were observed depending on the program and participant group, at least ``short-term employment in public civil engineering projects" was not found to have any effect on improving employment rates, so it should be abolished. recommends that the funds and resources freed up by its abolition be given priority to young women's groups, where various programs have had the greatest impact.

1. Summary of measures

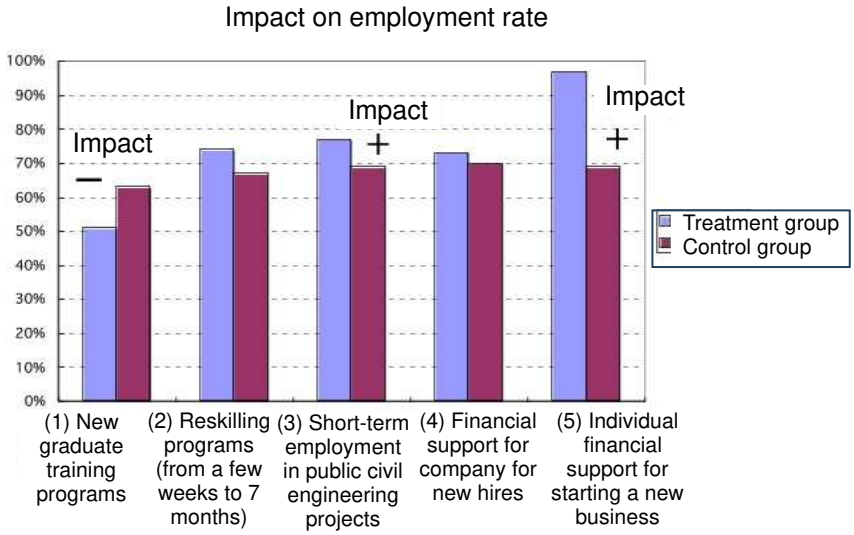
25,000 unemployed people registered with employment security offices were randomly selected and sent letters requesting their cooperation in a questionnaire survey. Of these, 4,477 agreed to cooperate, so we sent them a questionnaire and asked them to complete it. The questions are: a) Have you participated in each of the employment promotion programs listed in (1) to (5) in the past? b) Have you actually been employed since then ? c) If so, how much was your salary? It is.

4,477 people who responded , 278 answered that (1) they had participated in a new graduate support program ; Next, an equal number of people with the same characteristics will be selected from among those who have never participated in (1), and the difference in outcome indicators (whether they were employed or not, and what their salary level was) will be calculated. It is desirable that the situations and conditions other than participation in the program be as similar as possible, so when selecting individuals, we used the following seven matching indicators to select individuals who matched as closely as possible. These are factors that are thought to affect employment other than participation in employment promotion programs: 1) age, 2) gender, 3) final educational background, 4) length of unemployment period, and 5) residence. 6) Married/unmarried status; 7) Previous occupation. We aimed to select an equal number of 278 individuals, but ended up selecting 194 individuals. Hereinafter, in (2) to (5), comparison groups were similarly selected

by matching.

2. Evaluation results

The evaluation results are as follows (see graph on next page). At least the ``New Graduate Training Program'' and the ``Financial Support Program for Individuals Starting New Businesses'' were evaluated as having the effect of improving employment rates. On the other hand, the ``Public Civil Engineering Works Short-Term Employment Program'' has been found not only to be ineffective but also to have a negative impact on employment, and should be abolished.



Note: "+" in the figure indicates significance based on statistical test.

(3) Short-term employment in public projects

	Treat.G	Control G.	Impact	(Judge)	
就職率（現在）	51%	63%	-12%	Sig.(-)	Strong
96年以降の就職率	77%	72%	5%	n.s.	
月給（Kc）	5393Kc	6631Kc	-1238Kc	Sig.(-)	Medium

(2) Reskilling programs

	Treat.G	Control G.	Impact	(judge)	
就職率（現在）	74%	67%	7%	n.s.	
96年以降の就職率	88%	76%	8%	Sig.(+)	Strong
月給（Kc）	6536Kc	6636Kc	100Kc	n.s.	

(1) New graduate training programs

	Treat.G	Control G.	Impact	(judge)	
就職率（現在）	77%	69%	8%	Sig.(+)	Weak
96年以降の就職率	89%	75%	6%	Sig.(+)	Medium
月給（Kc）	6500Kc	7844Kc	-1344Kc	Sig.(-)	Medium

(4) Financial support for company for new hires

	Treat.G	Control G.	Impact	(judge)	
就職率（現在）	73%	70%	3%	n.s.	
96年以降の就職率	91%	80%	11%	Sig.(+)	Strong
月給（Kc）	5605Kc	6111Kc	-506Kc	Sig.(+)	Medium

(5) Individual financial support for starting a new business

	Treat.G	Control G.	Impact	(judge)	
就職率（現在）	97%	69%	28%	Sig.(+)	Strong
96年以降の就職率	97%	82%	16%	Sig.(+)	Strong
月給（Kc）	6306Kc	7074Kc	1768Kc	n.s.	

Note: Strong, medium, and weak in the table indicate significance levels cleared by statistical tests (1%, 5%, 10%)

Furthermore, in order to provide evaluation information that leads to policy changes, in addition to calculating the presence/absence/degree of impact of each program in (1) to (5), The presence/absence/degree of impact was measured by dividing into small groups. This small grouping suggests which small group is having a higher impact. The conclusion was that the highest impact was seen in the young female group. Based on these conclusions, the evaluators recommended that funds and resources freed up by eliminating ineffective programs be prioritized for young women's groups.

3. Advantages, limitations, and considerations regarding application in Japan

The advantage of this method is that there is no need to collect index values in advance, unlike previous evaluation methods. In other words, for evaluation studies where baseline data (pre-implementation index values) from several years ago do not exist, it is possible to use a statistical equalization design that divides and compares ex-post data, as in this case. . Calculating effects by dividing into small groups is very effective for designing the most effective program using limited resources (financial, human, and time).

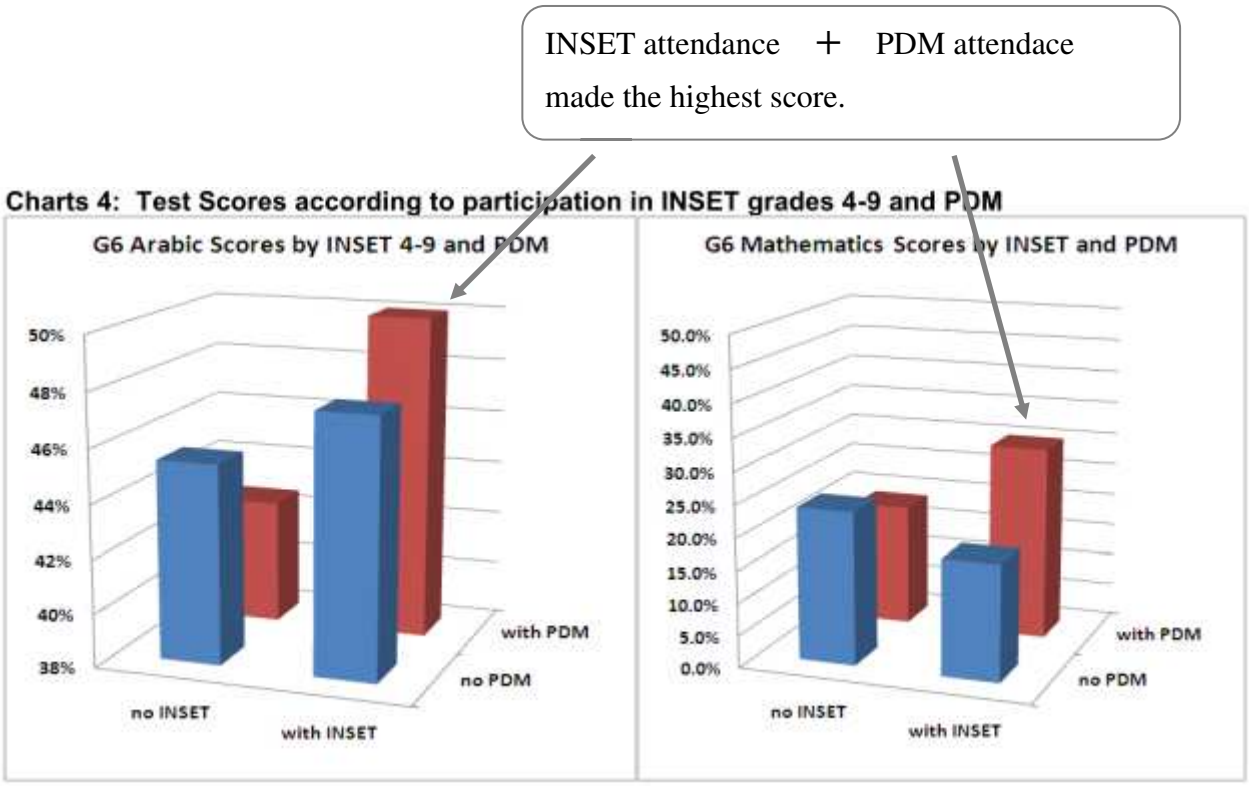
A limitation of this method is that the length of the division may be arbitrary. You can divide it into 2, if that doesn't work, divide it into 4, if that doesn't work, divide it into 8, then 16 , 32 , 64 , 128 , 256 , and so on until you find an effect. It is recommended that the parties involved decide in advance how far and by what criteria the division will be made.

There are things that Japan can learn from this evaluation example that are not directly related to evaluation methods. If this evaluation were conducted in Japan, most of the measures would be found to be effective, so there would be no need to do anything, but short-term employment for public civil engineering workers would be ineffective in improving the employment rate. I guess I should write a proposal on how to improve it to make it more effective. This means that we should continue to improve. In other words, **based on the same evaluation results, there is a possibility that a proposal contrary to this example will be made in Japan** . At the very least, it would be possible to reach the same conclusion as evaluation implementers in the United States and other countries using the same method as to whether or not it is effective, but this does not mean that recommendations will automatically come out, and the selection of recommendations may be difficult **.** We should recognize that this **depends in part on the value judgment of the person conducting the evaluation and on individual social circumstances .**

(Source) Benus, J., Grover N., Jiri, B., Jan, R., 1998, *Czech Republic:Impact of Active Labor Market Programs* . Cambridge, Mass., and Bethesda, Md., Abt Associates.

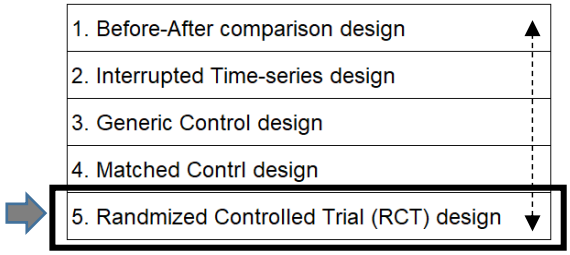
Application examples of statistical equalization design (post-hoc matching design): Effects of in-service teacher training (INSET) and professional development meeting (PDM) (Yemen)

GTZ simultaneously supported two programs in Yemen: (1) training camp-based in-service teacher training (INSET) and (2) professional development meetings (PDM) in educational settings. The results were summarized in the three-dimensional graph below.

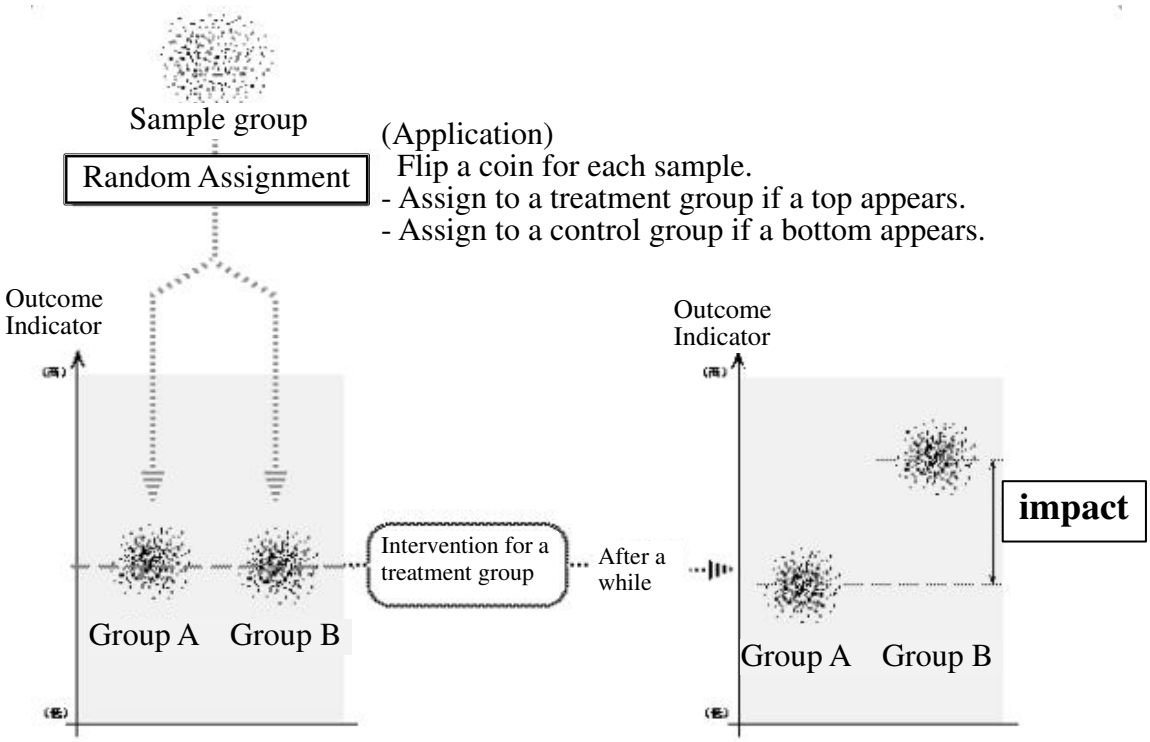


Based on this graph, the GTZ report explains as follows: ``The test results of classes taught by teachers who participated in PDM are influenced by the following facts. That is, only one teacher participated in PDM and INSET. This means that only one teacher received the combination of training. However, the influence of PDM is clearly visible. There are differences in test results for both Arabic and mathematics depending on whether teachers participated in PDM. This is true even if teachers participate in INSET. The improvement effect in mathematics (from 18.1% to 30%) is 66%. The improvement effect for Arabic (from 47.5% to 50%) is only 5%. 'What can be seen from this passage is that it **is written in a manner that frankly acknowledges the limitations of analysis .** If a similar analysis is to be conducted in Japan in the future, it is recommended that the limitations of the analysis be clearly stated in this way.

(Source) GTZ, *Result-Based Management of BEIP-GTZ Interventions in Abyan, Ibb , Hajja and Marib Governorates of Yemen, Schol years 2005/06 and 2006/07 Overall Report.* p.19



5. Randomized Controlled Trial (RCT)
(previously known as Experimental Design)



[explanation]

Before implementing an implementation, the policy application is divided into a treatment group and a control group by random assignment. It can be determined that any differences that appeared in the outcome indicators were caused purely by only ``whether or not the intervention was applied." Note that there is no need to consider the effects of external factors as they are exactly the same. It show extremely high reliability in identifying the existence of causal relationships. However, actual application is difficult.

[Analysis test]

Two-group significance test (independent t-test)

Experimental design (RCT) application example 1:

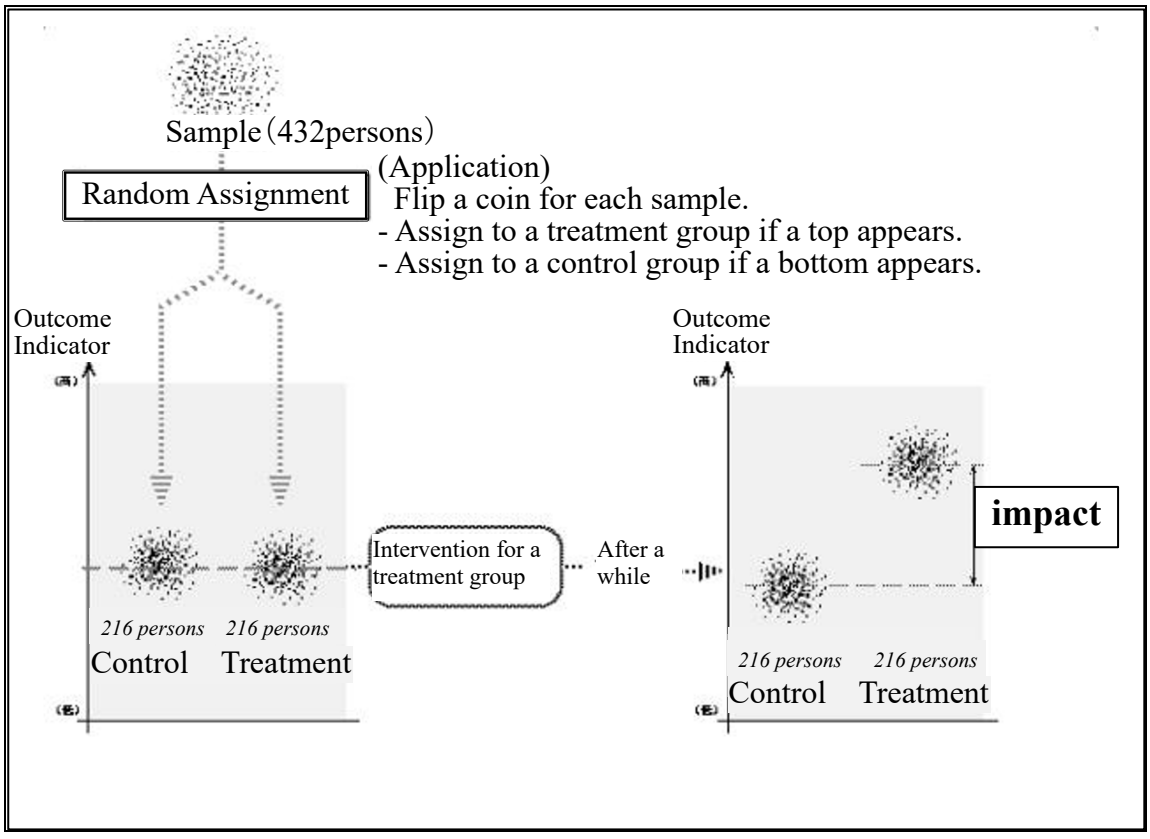
Small financial support measures for people released from prison (USA)

1. Problem location and evaluation results

What kind of "policies" are effective in preventing the recurrence of crime? One possible "policy" idea is to provide small financial aid to those released from prison to help them smoothly transition into normal civilian life. However, is this "policy" proposal really effective? If someone commits a crime and receives cash, will they commit another crime? To answer these questions, this "experiment" was conducted in Baltimore, Maryland. As a result, it was concluded that the implementation of the program had the intended effect, at least with regard to ``theft".

2. Overview of measures and evaluation design

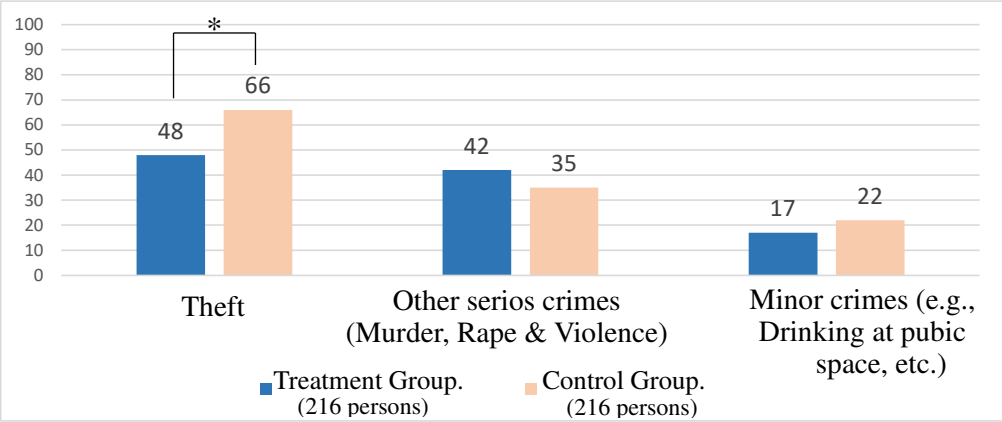
1970s , involved prisoners who had been released from a Maryland state prison and returned to Baltimore. For 432 people released from prison, random assignment determined whether they would be in the implementation or comparison group. Those assigned to the implementation group (216 persons) received $60 weekly for up to 13 weeks until they were hired . Those assigned to the comparison group (216 persons) were told that they would be participating in the experiment but would not be paid.



3. Evaluation results

Based on the arrest records of the Baltimore Police Department, the following results were obtained regarding the arrest rate of the experiment participants one year later.

	Control (216 persons)	Treatment (216 persons)	Difference
-Theft	48 persons (22.2%)	66 persons (30.6%)	-18 persons (-8.4%)
-Other serios crimes (Murder, Rape & Violence)	42 persons (19.4%)	35 persons (16.2%)	7 persons (+3.2%)
-Minor crimes (e.g.,Drinking at pubic space, etc.)	17 persons (7.9%)	22 persons (10.2%)	-5 persons (-2.3%)



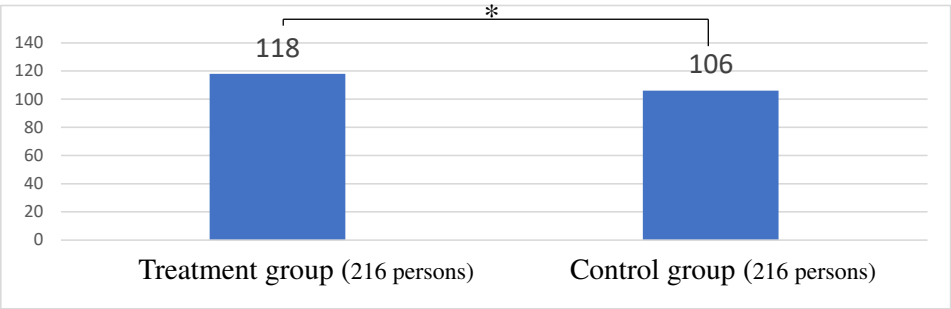
* ... Statistically significant (however, information on significance level is not stated)

Regarding ``theft", the implementation group to which the program was applied had a -8.4% lower arrest rate than the comparison group. But is this difference greater than the difference that could occur by chance without a program? This -8.4% passed the statistical test. Other types of arrest rates did not pass statistical tests. In other words, it could not be determined that the difference between the implementation group and the comparison group regarding ``other serious crimes" and ``minor crimes" was greater than would occur by chance.

Regarding the employment rate, the following differences were measured.

	Treatment group (216 persons)	Control group (216 persons)	Difference
Employment rate (Full-time)	118 persons (54.7%)	106 persons (49.0%)	12 persons (+5.7%)

* Only fourth quarter. The first to third quarters did not show statistically significant difference.



* ... Statistically significant (however, information on significance level is not stated)

4. conclusion

The evaluation results are as follows. At least with regard to ``theft," the program implementation was assessed to have the intended effect.

Furthermore, are the effects revealed in this experiment sufficient to warrant implementing this measure on a large scale? To answer this question, a cost-benefit evaluation was then conducted. The U.S. Department of Labor was in charge of the evaluation, and found that when calculated from a social perspective , the benefit/cost ratio is 4.02 times even in the most conservative calculation, and 4.02 times in the most optimistic calculation. In this case, it was calculated as 53.73 times.

	Social benefit	Social cost	Magnification
Highest case	$108,565	$27,000	4.02
Lowest case	$870,431	$16,200	53.73

Therefore, the evaluation results show that the social benefits brought about by this measure far exceed the social costs, and the areas in which it is applied should be expanded.

Furthermore, this experiment suggests that other measures may be needed to prevent the recurrence of ``other serious crimes (murder, rape, assault)."

5. discussion

As is often said, **no one can deny the method of majority voting, but majority voting does not reveal the facts** . In this example, they first conduct an RCT experiment to find out which types of crimes are effective, and then discuss and make decisions about whether to expand the scope of application. In Japan, there are cases where emotional claims are made without clarifying the facts, and the majority vote is enough to push through and actually implement the measures. In Japan, efforts should first be made to clarify the facts. **Clarifying the evidence** is for democracy.

Another thing I would like to point out is that it is sometimes pointed out that it is unethical to separate people released from prison into two sides of the same coin. The counterargument to this **is that applying**

it nationwide without knowing whether it will be effective or not would cause far more damage and would be unethical . A good example is ``Yutori Kyoiku," which started before we knew it and ended before we knew it. During the implementation period, we have not seen anything that could be called evidence, and all school-age children across the country were affected without any evidence being presented. This is a point that those who oppose social experiments should understand.

(Source of this case study)

（文献 1) Peter H.Rossi, R.A Berk, and K.J.Lenihan (1980), *Money, Work and Crime:Some Experimental Evidence*; New York: Academic Press; Adapted initially as an example in *'Evaluation: A Systematic Approach 6th Edition.'*.

(Reference 2) Greenberg, D. and Shroder , M., (1997). *The Digest of Social Experiments 2nd edition*, Urban, Institute Press. Pp.217-219. and Ryo Sasaki (2003) " Policy Evaluation Training. Additions and changes were made with reference to the description published in "Book" Taga Publishing.

(Note) The numbers at the top of the table (48 people, 66 people) are specified in Document 2. This corresponds to the number of people calculated backward from the sample number and ratio described in Document 1, and the numbers in the second row in the table are also the numbers obtained by the same calculation.

Experimental design (RCT) application example 2:

How to increase attendance days? : Roundworm extermination project in elementary school (Kenya)

1. Where the problem lies

For children to attend school every day is the minimum necessary condition for any educational effect. Until now, countermeasures have been considered within the framework of the education sector, such as raising the awareness of parents, providing free school lunches, and rebuilding school buildings into clean ones. However, from a slightly different perspective, a proposal was made that measures taken by the health sector, such as distributing and taking antiparasitic drugs, were actually effective.

2. Overview of measures and evaluation design

An evaluation using RCT was immediately conducted. This project was implemented from 1998 to 2002 in Busia District, Kenya, with the intervention of administering roundworm drugs to elementary school students and providing related education .

Table 1 Overview of RCT applications regarding antiparasitic drugs

Target area	Busia District, Kenya
sample	75 primary schools in Busia District (approximately 30,000 students)
act of intervention	Distribution of roundworm medicine. Dutch NGO International Christelijk _ We evaluated the distribution carried out as a project of Steunfounds Africa (ICS) .

75 schools in the prefecture were randomly divided into three groups of 25 schools each , and the intervention was implemented in the following years (originally it was planned to be conducted for three consecutive years, but due to flooding, 20000 was shifted to 2001). This staggered implementation ensures that all schools receive the intervention and avoids the ethical issues typically associated with experimental designs (RCTs).

Table 2 Year of intervention implementation for each group

Random assignment
(Application) Allocate to 3 groups with dice (there are 6 sides).

75 schools

	1998	1999	2000	2001
G1 (25 schools)	◎	◎		×
G2 (25 schools)	×	◎		×
G3 (25 schools)	×	×		◎

(Note) ◎indicates intervention.

3. Evaluation results

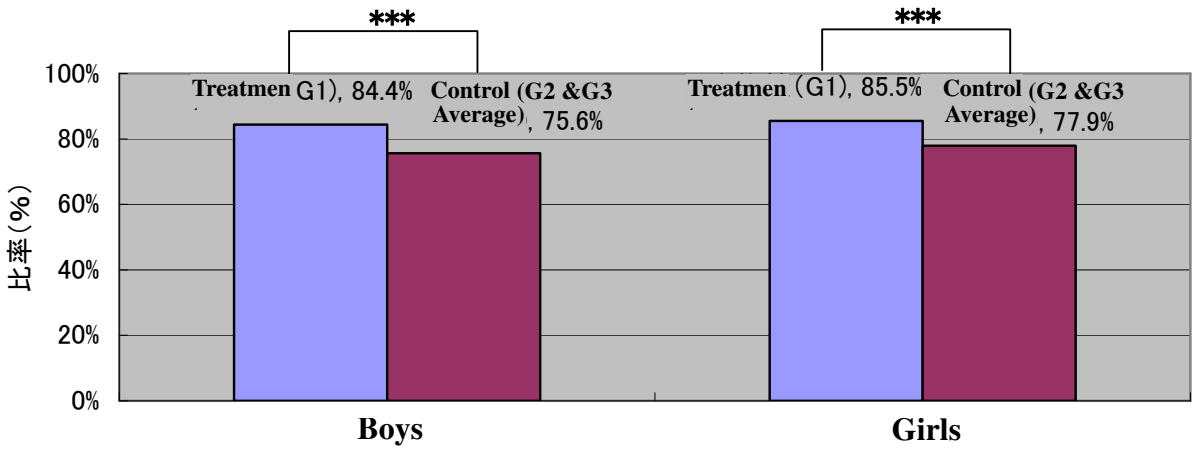
At the end of the first year (end of 1998), it is possible to compare G1 as an intervention group and G2 as a comparison group (G3 can also be used as a comparison group, but is omitted). At the same time, the roundworm infection rate in G1 (intervention group) was 27 %, and the infection rate in G2 (comparison group) was 52 %, so the difference of -25 % can be judged to be the effect of the intervention.

Table 3 Roundworm infection rate

	1998 year end
G1 (25 schools): Treatment Group	27%
G2 (25 schools): Control Group	52%
Difference (Impact)	−25%

Furthermore, at the end of the first year (end of 1998), the number of days students were absent was reduced by approximately one- third due to the distribution and use of antiparasitic drugs (-36.1% for the boy in the figure below) (=(75.6) %-84.4%)/(100%-75.6%) .For girls -34.4%(=(77.9-85.5%)/(100%-77.9%)) . This is calculated from the time they enter elementary school until they graduate. The effect was so large that it amounted to an increase of almost one year when converted into a period of time.Also , the annual cost per student was only 50 cents (approximately 50 yen), which was significantly greater than measures in the regular education sector. It was rated as cheap.

Figure1 Attendance rate at the end of the first year (end of 1998) (school level)



Significance level: ***1%, **5%, *10%

4. conclusion

It was concluded that administering roundworm drugs not only has an educational effect by increasing attendance, but the cost of the intervention is significantly lower than traditional educational measures.

Information on this impact evaluation case study was widely shared around the world. **It has been**

adopted as a national policy in Kenya, Nigeria, Ethiopia, India, and Vietnam to increase student attendance . It is reported that **300 million children around the world have benefited from the introduction of this policy (according to a report published in Policy Action by the Poverty Action Lab).** This is **a great example of how a single social experiment changed education policy around the world .**

5. discussion

Experts in a given sector are often only able to think within that sector. However, this is a case where experts from other sectors sometimes come up with innovative solutions. When a novel solution is proposed, simply saying "this is nonsense from people who don't know anything about this sector" or "I'm an expert in this sector" is defending vested industry interests. It's only going to happen. **Effectiveness is determined not by who is the more expert (that is, who "knows more"), but by the objective evidence obtained through experiments** . This is a case that clearly shows this.

(Source) Kremer, M., and Miguel, E. (2003) *Worms; Education and Health Externalities in Kenya.* Poverty Action Lab, MIT

(Source) Poverty Action Lab. *Policy Action* (https://www.povertyactionlab.org)

Application example 3 of experimental design (RCT) :
Enabling learning via SMS (Short Messages) and phone calls during the pandemic:
Rapid RCT verification of low-tech support in primary education in Botswana

1. Where is the problem?

COVID -19 pandemic has paralyzed education systems around the world . According to one study, more than 1.6 billion students have been isolated from school (UNESCO 2020). To address this deteriorating learning environment , a cost - effective approach was needed to improve children's learning on a global scale .

Botswana was an early adopter of precautionary social distancing measures . In the education sector, schools have been closed for six months since March 20, 2000 as the first measure , and the impact on education has been severe. Although the country's net enrollment rate for primary education (elementary school) is high at approximately 90% (UNESCO 2014) , the level of learning is said to be low.

Additionally, apart from the novel coronavirus, other viruses (influenza and Ebola) , teacher strikes, earthquakes and natural disasters have caused school closures around the world . It can be said that a general approach that can be applied to these issues is needed.

In addition, research shows that in low- and middle-income countries, access to the Internet is limited to 15-60 % of households, while 70-90 % of households own at least one mobile phone (Center for Global Development 2020) . Thinking about using this mobile phone.

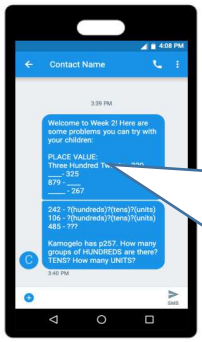
2. act of intervention

Days before the Botswana government declared a state of emergency, the research team obtained 7,550 phone numbers from primary schools. This was collected in Botswana by Young love , an active NGO working with the country's Ministry of Education . These were collected from 3rd to 5th grade students from almost every elementary school in the country.
The research team's 60 facilitators called each number to confirm whether they wanted to receive " remote learning support via phone ." The facilitator received instructions on how to speak on W hatsApp and then called accordingly.
are two types of ``remote support via mobile phone'': (a) sending math problems via SMS text messages, and (b) providing live advice over the phone for 15-20 minutes. Furthermore, since it costs the family money to have their parents send SMS or call them, facilitators sent SMS or made phone calls. Both (a) and (b) can be said to be ` `low technology'' compared to instruction via advanced websites using the Internet . However, because it is low-tech, he thought it could reach any household.

First intervention (SMS)	send an SMS with some simple math problems . The SMS was sent to the parents' mobile phones, as children rarely have their own mobile phones. It has been found that parents sometimes show their children the questions as is, and sometimes they teach them to their children and help them solve the problems (both are preferred). S MS did not ask for an answer. It is later understood that the answer has been sent.
second intervention (Advice by phone)	a facilitator provided verbal advice to parents via mobile phone for 15-20 minutes at the beginning of the week . After each call, parents were asked to bring their children along for advice. According to later reports, parents said they felt proud when their children were able to solve math problems.

46 facilitators served 24 parents each . Each facilitator spent approximately six hours on the phone each day. The facilitator decided to periodically ask parents what time of day was most convenient for them. More than 50 % of respondents answered that it is convenient for them to do housework after finishing their housework or at a time when doing housework is convenient for them. Before conducting the survey in earnest, we conducted a two-week trial to gain know-how.

SMS (Short message)



welcome! It's the second week. Here are some problems you can try with your child!
Please answer how much the calculation below will cost.

Phone advice

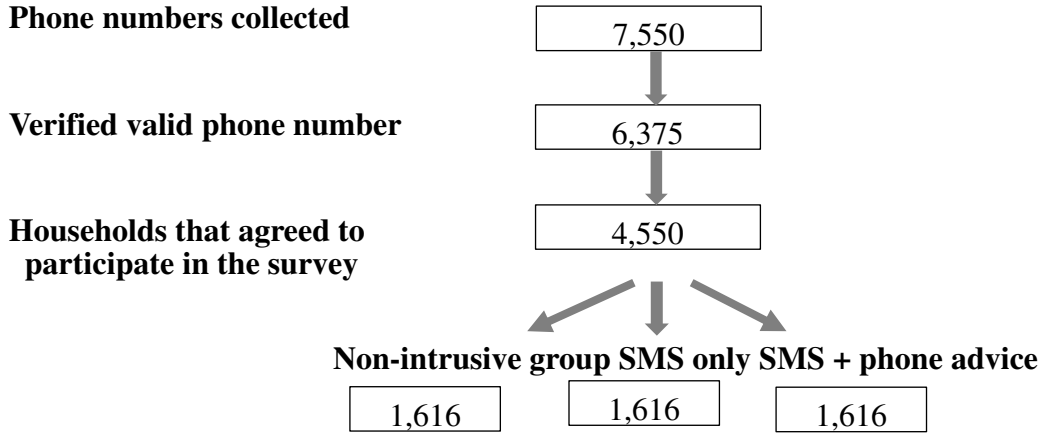
Hello Muma. This is an NGO basket. Did you receive an SMS? Did your child Sunshine solve the SMS problem?
Please include him in the call.
Then increase the volume of the speaker.
Sunshine-kun, let's solve the problem together. How many hundreds, tens, or units (ones place) is 284?

Based on the illustration on page 6 of the original report . Translated into Japanese by the author, and then translated into English again.

3. allocation

By calling 7,550 phone numbers from primary schools across Botswana , we were able to reach 6,375 people. The remaining numbers were either not valid, unanswered, or had moved. We then asked them to consent to participate in the survey , and approximately 71% , or 4,550 people, agreed. The research team assigned 4,550 people to three groups of equal sample size using random assignment. Furthermore, by dividing the 4,550 people into two groups based on whether they have one child or multiple children, and by assigning each group to three groups (s tratified), the proportion of each group is divided into three groups . I tried to make it the same. This is where we improved.

After 4 weeks, the children were given the first arithmetic test. Children who answered correctly were then allowed to move on to the next level of questions. Children who did not answer correctly were kept on the same level of questions. After 10 weeks, a second arithmetic test was administered to measure the effect of this ``additional intervention'' (not yet analyzed) .



(Note) Data collection in the first stage was conducted from half of each of the three groups, and data collection in the second stage was conducted from all households.

(Source) Based on the illustration on page 9 of the original report . Translated into Japanese by the author.

4. Data collection

The ASER test was used for the math test. (ASER: Abbreviation for Annual Status of Education Report). The A SER test consists of addition (level 1), subtraction (level 2), multiplication (level 3), and division (level 4). Below is a sample. I asked the child to answer using a cell phone, but in order to prevent the intervention of the parent nearby, I decided to 1) have the child solve each problem in 2 minutes, and 2) ask the child questions and answer the calculation process

himself. Only those that were answered correctly were considered correct. Of course it's not perfect, but I did my best. The ASER test result is one of 5 levels from 0 to 4 .

Although the ASER test was conducted before the intervention and baseline values were not taken, it goes without saying that the results should be consistent since appropriate randomization was performed.

It is noteworthy that it took only two months from RCT planning to endline data collection (math test and parent questionnaire).

5. Analysis results of intervention effects (impact)

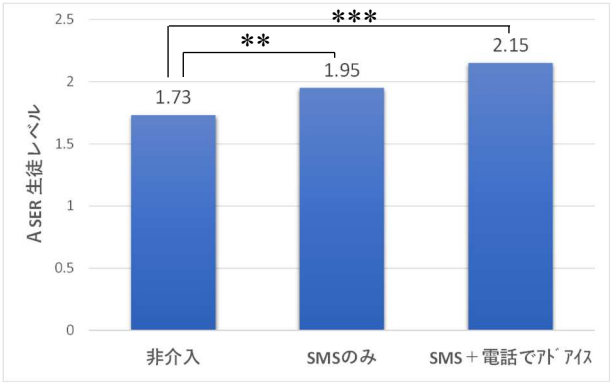
The research team confirmed a significant impact on children's learning after the first stage of intervention (=4 weeks).

（1）There was a 24% improvement in learning [4]in the S MS + telephone advice group compared to the no-intervention group .

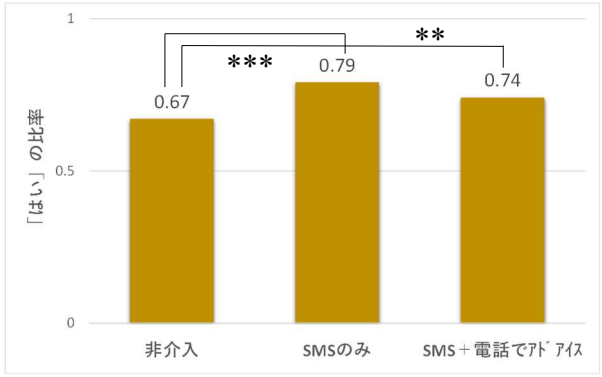
（2）The SMS -only group had a 1.3 percent improvement compared to the no-intervention group. It was statistically significant [5].

（3）Additionally, parents were asked whether they engaged with their children. The S MS + telephone advice group had 7 percentage points more co-involvement than the no-intervention group. On the other hand, the MS -only group was 12 percentage points more involved than the no-intervention group . It has been shown that parents are more involved when using SMS only.

Children's math level (0-4 levels)



Parental involvement (“At least sometimes involved” = percentage of “ Yes ”)



(Note) * **1% significant, * *5% significant, * 10% significant

(Source) Created by the author based on the charts on pages 19, 23, and 31 of the original report .

6. Cost-effectiveness analysis results

The research team calculated an upper limit on costs. The costs include the total cost of the project, the time of the individuals involved, the cost of collecting phone numbers (by the NGO prior to this study), infrastructure, training time for facilitators, regular testing & conducting parent surveys. cost. The total cost for SMS alone is calculated to be $3,200 (about 350,000 yen), or $2.13 (about $3) per child. Telephone advice costs $17,800, or $14 per person. This amounts to an improvement of one standard deviation per child of $13.3 and $48.28, respectively. Each level corresponds to "beginner"

[4]This magnitude was 0.29 per standard deviation width of the non-intervention group. This is the so-called effect size, where 0.2 = small, 0.4 = medium, and 0.6 or more = large impact (Cohen, 1986). In this case, it corresponds to "small to medium". Glass, V. recommends using the standard deviation of the non-intervention group as the denominator in the division.

[5]With the same calculation, the effect size was 0.16 (small).

(level 0), addition (level 1), subtraction (level 2), multiplication (level 3), and division (level 4), respectively. This can be interpreted as the cost required to move up one level.

7. Policy recommendations

The findings showed that **the low-tech intervention of SMS and telephone advice was cost-effective and improved learning in the short term during school closures.** The research team also plans to analyze the impact of the next step of the intervention, targeted phone calls, as well as longer-term effects (none of which have been implemented yet).

8. discussion

The survey results suggest three things that challenge conventional wisdom:

(1) Randomized controlled trials (RCTs), the most rigorous design and also used to confirm the effectiveness of drugs, are generally thought to take between one and a half to two years. However, **in this case, the evaluation was completed in just two months** and the results of the evaluation were made public. In fact, it can be seen that RCT can evaluate the effect (impact) of an intervention without spending much time. You can conduct an RCT in the first year of a Japanese civil servant's term and use it in your policies in the second year to see the impact for yourself. RCTs should be used more easily and quickly. Additionally, the level of statistical analysis used in this study is sufficient.

(2) Distance education using the Internet means 1) creating a website, 2) recording and pasting a video, 3) preparing practice questions with clickable answers, and 4) having students view them on a tablet. I will think about it. However, this idea originated from aid agencies in developed countries, and while it is fine for teachers, it is not common for households that have access to the Internet and have tablets and computers. On the other hand, research shows that mobile phones are widely used in 70-90% of households (as introduced at the beginning), **and learning is supported through low-tech SMS and phone calls rather than high-tech distance learning, and educational effects can be improved. You can see that you can create (impact)** . Of course, aid agencies are responsible for the costs of sending SMS and making calls.

(3) Considering the educational effect, the cost is considerably lower. The costs for existing educational projects and aid projects are not clearly disclosed, perhaps out of concern for the partner government's Ministry of Education and aid agencies, but when calculating the cost for the same educational effect, each You'll probably find it much cheaper. It is a valid opinion that education cannot be measured by test scores alone, but perhaps we should consider using mobile phones to teach new morals, new ethics, and new ways of working with others. do not have. And in the future world, this may have a wider scope of application. In the field of education in Japan, **the ``zest for life'' is said to be important, and educators may be witnessing the moment when they are thinking about the ``zest for life of the new generation.' ' From Survivability to New Age Survivability .** I feel like this is a bit too cliche, but I'd like to say it directly.

(Source) Noam Angrist, Peter Bergman, Caton Brewster, and Moitshepi Matsheng (August 2020) . *Stemming Learning Loss During the Pandemic: A Rapid Randomized Trial of a Low-Tech Intervention in Botswana* . Center of the African Economics. Available at: https://www.povertyactionlab.org/sites/default/files/ research-paper/working-paper_8778_Stemming-Learning-Loss-Pandemic_Botswana_Aug2020.pdf

Application example 4 of experimental design (RCT) :

Is basic income effective? (Finland)

1. Where the problem lies

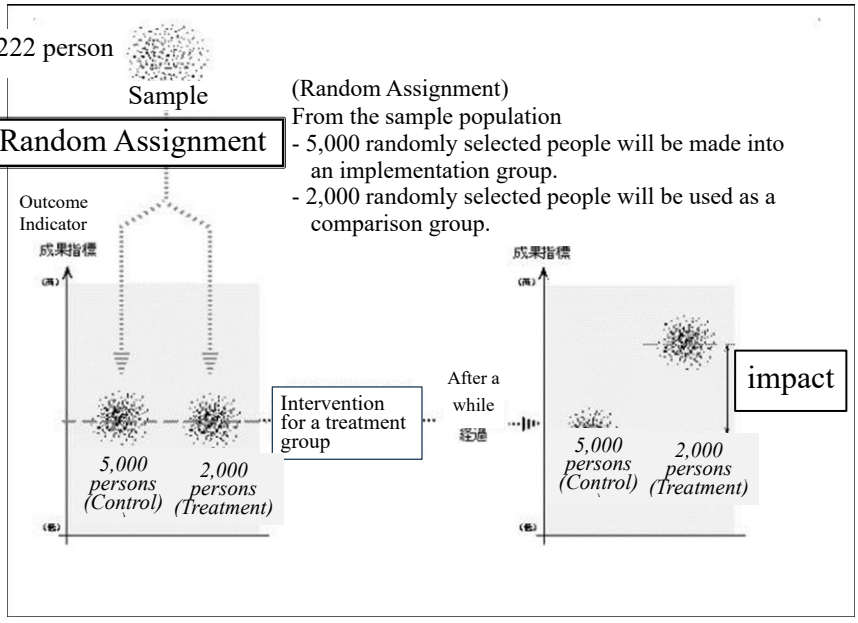
Basic income is a policy in which the government provides a fixed amount of cash to all citizens on a regular basis . It is said that this will enable the people to live a minimum standard of living . On the other hand, some argue that if people could receive cash without having to work, they would stop working, and the country as a whole would become poorer as a result. The basic income measure is often talked about as a response to the threat that the spread of IT will eliminate jobs for people and increase the number of unemployed people. How will basic income actually change people's lives?

2. Overview of measures and evaluation design

2 A large-scale evaluation using an RCT was conducted in Finland in 2017-2018 . The sample and intervention actions are outlined below.

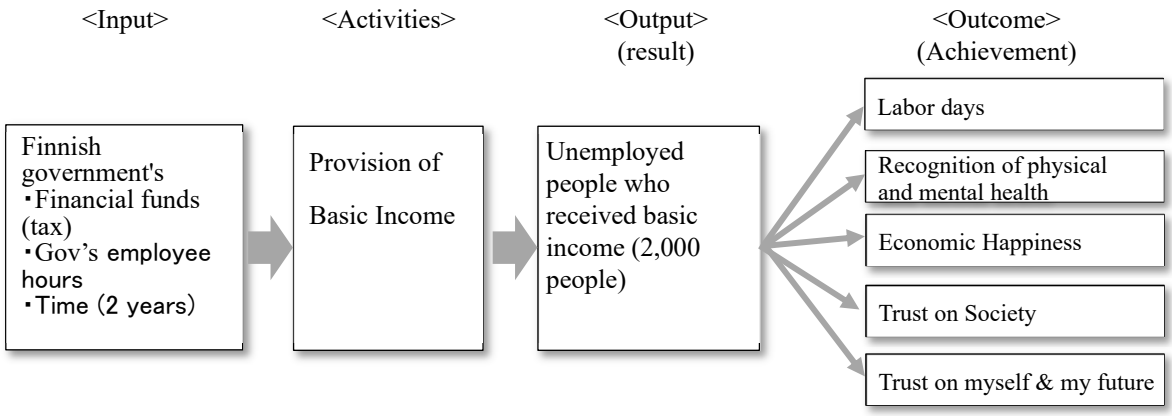
Table 1 Overview of RCT application regarding basic income

Target area	Finland
sample	25-58 years old) registered in the labor market (175,222 people). Intervention group : 2,000 people randomly selected from the population . Comparison group : 5,000 people from the remaining population (1,73,222 people) . Everyone in the intervention group will receive a basic income for two years. on the other hand. For the comparison group, conventional unemployment measures will continue.
Implementation year	2 January 1, 017 - Last day of 2018. Baseline survey in 2017 and endline survey in 2018 . A telephone interview was conducted. The response rate was 23.2% (intervention group 3 1.3% , comparison group 2 0.2%).
act of intervention	The Finnish government provides basic income of 5.60 euros (€) per month. Approximately 75,600 yen per month in Japanese yen (1€ = 135 yen).



3. Evaluation results

The Finnish government has set the following outcomes as outcomes to be achieved through basic income .



(Source) Created by the author based on the table of contents of the Finnish government report.

(1) Employment effects

As a result of basic income distribution, the number of working days increased by 5.05 days (see Table 1). Furthermore, when the influence of personal attributes was removed, an increase of 6.03 days was calculated. Concerns that basic income would cause people to stop working were not true; on the contrary, the number of working days increased by 6 to 8 percent . However, it is possible that this may have been influenced by changes in existing unemployment policies that year.

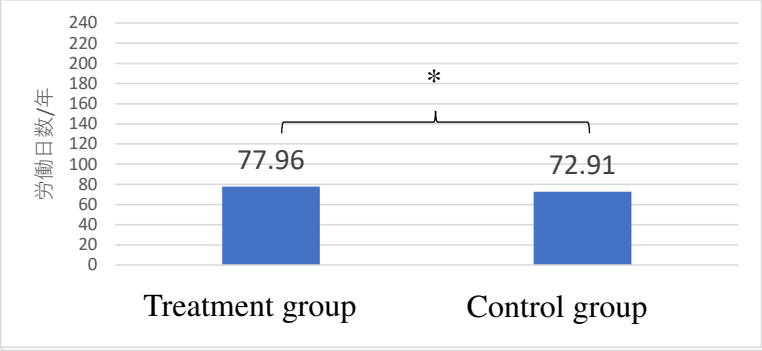
Table 1: Impact on working days (from November 1, 2017 to October 31, 2018 (second year of the experiment))

Method of analysis	Intervention group	Comparison group	difference	standard error	P- value*
Significance test for two groups	7 7.96 days	7 2.91 days	5.05 days	2.84 days	0.08
value after removing the influence of personal attributes * *)		7 2.91 days	6.03 days	2.52 days	0.02

* If the p-value is less than 0.1 (=10%), less than 0.05 (=5%), or less than 0.01 (=1%), the difference between the two groups is too large to be caused by chance.

**What are personal attributes? Gender, age, level and field of education, native language, family type, diagnosed disease, municipal group, area of residence, type of unemployment benefits, number of days of unemployment, days of work, income from work and assistance.

(Source) Created by adding explanations from the main text to Table 1 of the Finnish government report, page 39.



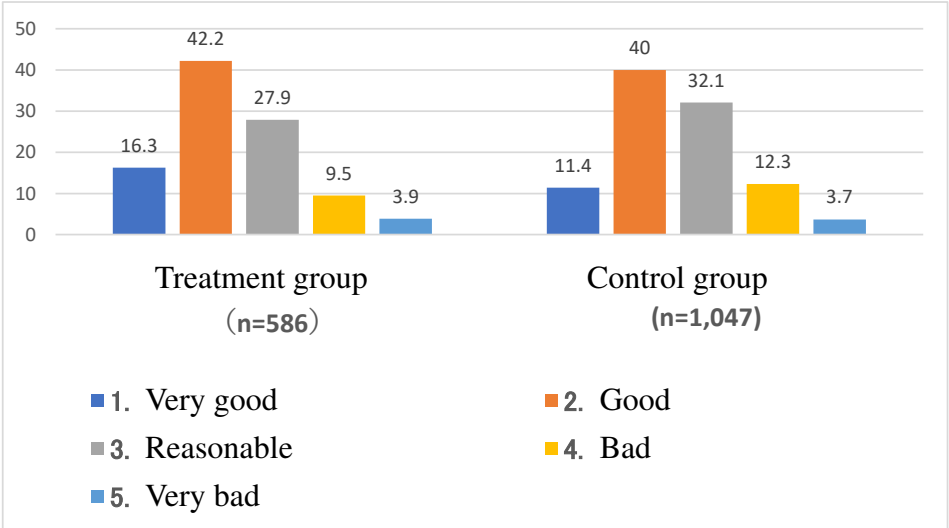
(2) Perceived health, mental well-being and cognitive performance

Basic income distribution improved perceptions of health status (see Table 2). Furthermore, 1. Reduced impact of illness, physical illness, and mental disorders on daily life, 2. Reduced use of health services, 3. Reduced psychological stress, 4. Improved cognitive function, and 5. Reduced loneliness . was also confirmed (see Appendix Table 1 in Additional Explanation 1).

Table 2 : Self-assessment of health status

awareness of health status	Intervention group (n =586)	Comparison group (n=1,047)
▲ 1.Very good _ _	1 6.3	1 1.4
2.Good _ _ _	4 2.2	4 0.0
3. Reasonable _ _ _	2 7.9	3 2.1
4.Bad _ _ _	9.5 _	1 2.3
▼ 5.Very bad _ _	3.9 _	3.7 _
		p=0.051

(Note) If the p- value is less than 0.1 (=10%) , less than 0.05 (=5%) , or less than 0.01 (=1%), the difference between the two groups is such that it cannot be caused by chance. That's a big difference.



(3) Economic Well-being

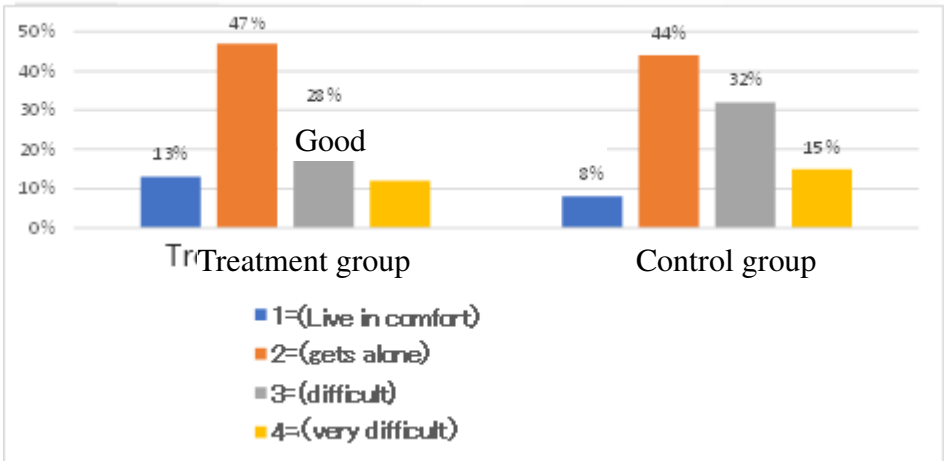
Economic well-being is the state of (i) being able to responsibly manage one's own finances, (ii) being financially stable, and (iii) having (or being aware of) labor productivity. Respondents who received a basic income reported better

income levels and better economic well-being across all areas of economic well-being (see Table 3 and also Appendix Table 2 in Additional Explanation 2).

Table 3 : Are you able to live on your household income?

choices / group	1 = Live in comfort	2 = able to make a living (g ets alone)	3 = difficult _	4 = very difficult
intervention group	13%	47%	28%	12%
comparison group	8%	44%	32%	15%

(Note) P value is not shown. (Source) Created from the description in Finnish government report 5.3 .



(4) Trust on institutions and confidence in oneself

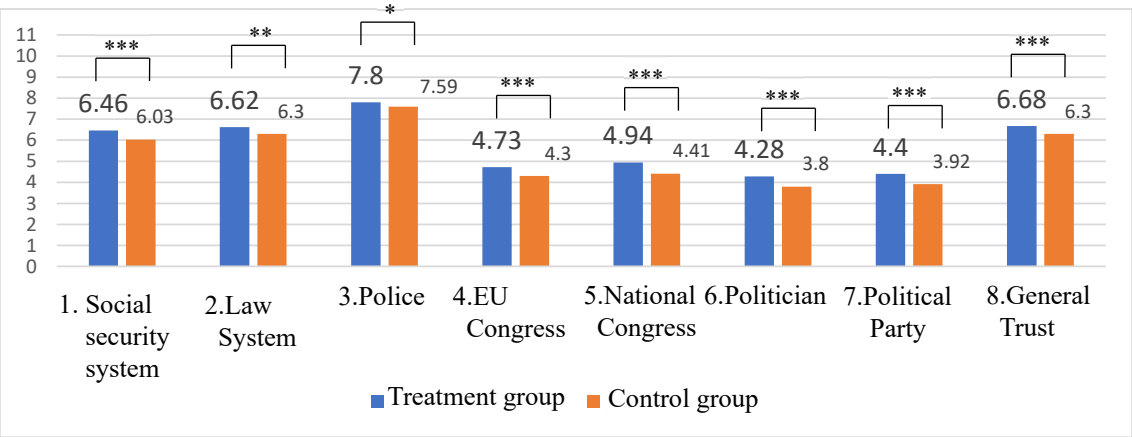
Without trust, life would be chaotic and we would be unable to have meaningful social interactions with other people. Regarding trust in social institutions, it can be seen that the mean value for the intervention group was significantly higher than the mean value for the comparison group (see Table 4). Similarly, for confidence in oneself and the future, it can be seen that the mean value of the intervention group was significantly higher than the mean value of the comparison group (see Appendix Table 3 in Additional Explanation 3).

Table 4: Trust in the organization and general trust (average value)

group / Object of trust	intervention group average value of	comparison group average value of	Difference between both groups and p-value
1. Social security system	6.46	6.03	+0.43 p =.001
2.Legal system _	6.62	6.30	+0.32 p =.018
3. Police _	7.80	7.59	+0.21 p =.079
4. European Parliament	4.73	4.30	+0.43 p =.004
5. Finnish Parliament	4.94	4.41	+0.53 p =.000
6. Politician _	4.28	3.80	+0.48 p =.001
7. Political party	4.40	3.92	+0.48 p =.001
8.General trust _	6.68	6.30	+0.38 p =.003

(Note 1) Choices are 1, 1 level: " 10 = extremely reliable" to " 0 = extremely unreliable"

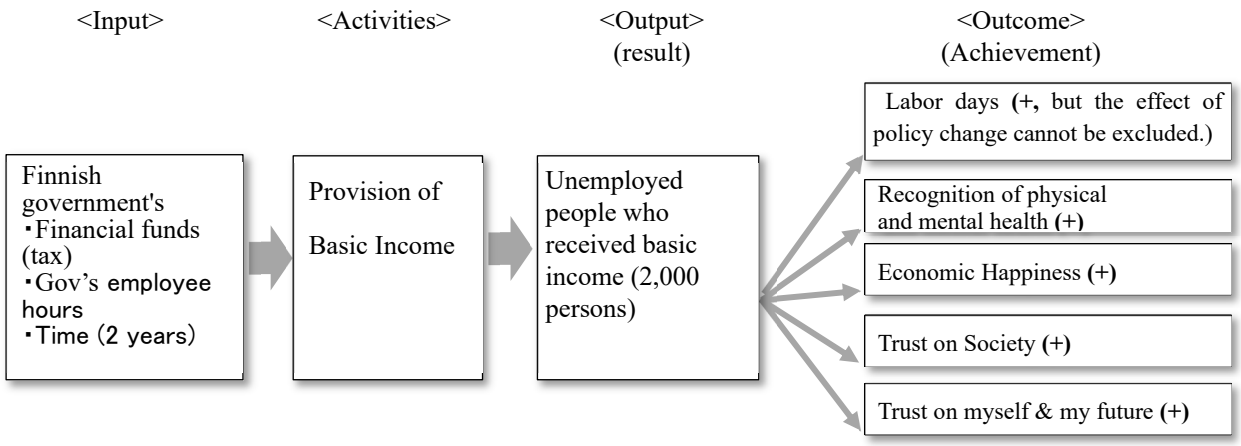
(Note 2) If the p- value is less than 0.1 (=10%) , less than 0.05 (=5%) , or less than 0.01 (=1%), the difference between the two groups cannot occur by chance. That's a huge difference.



*** , **, *… Significant at 1%, 5%, 10% level

4. conclusion

It was concluded that **basic income increased the welfare of the unemployed .** Additionally, the results showed that the number of working days increased due to basic income, but it is not possible to say anything definitive because the existing unemployment policy was changed during the experiment, and the effect cannot be ruled out. However, it can be said that at least the fear that people would become lazy was not confirmed.



(Note) "+" indicates improvement. "-" indicates deterioration.
(Source) Created by the author based on the table of contents of the Finnish government report.

5. discussion

Newspaper articles from around the world that reported on this report often concluded that, after all, basic income was neither good nor bad. Although it cannot be said that it had a dramatic effect, it did not have as much negative impact as feared.

When basic income is discussed in Japan, it is a policy to help people who have lost their basic living expenses due to unemployment, and it is talked about in the context of the fact that in the age of IT, anyone can become unemployed . often. In other words, it is often said that basic income is a policy that focuses on the poorest and most vulnerable people

and that it can help them, and that it can help you too. On the other hand, there are also people who say that it is their own responsibility and their responsibility, and that there are many people who do their best despite their unfortunate circumstances.

What we can see from the evaluation of Finland's basic income is that it is not just about money, but also the effects on people's physical and mental health, happiness, self-esteem, and trust in society. It can be said that **this reflects the Finnish people's outlook on life** . When discussing basic income in Japan, this may be taken as a message that **we should first consider the Japanese people's outlook on life .**

refers to a report conducted and published by the Finnish government (in Finnish) translated into English using Google Translate. It should be clearly stated in advance that the translation may not necessarily be accurate.

(Note 2) The Finnish government's report not only analyzes quantitative data, but also analyzes the results of individual interviews with participants. It can be said that ``mixed methods ' ' are used, which combines quantitative and qualitative methods . However, it can be criticized that the use of mixed methods tends to lead to ambiguous conclusions. Another possible criticism is that the conclusion tends to be positive.

(source)

（フィンランド語のreport） Olli Kangas, Signe Jauhiainen , Miska Simanainen and Minna Ylikännö (eds.). (2020). *Evaluating av basinkomtestperiment* . Social and Ministry of Health .

(English translation of the title of the above report) Olli Kangas, Signe Jauhiainen , Miska Simanainen and Minna Ylikännö (eds.). (2020). *Evaluation of the Finnish basic income experiment.* Ministry of Social Affairs and Health, Finland.

http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/162219/STM_2020_15_rap.pdf?sequence=1&isAllowed=y

Attachment to the case (additional explanation 1, 2, 3)

Additional explanation 1: Additional explanation of " mental and physical health awareness "

As explained in the main text, the distribution of basic income improved perceptions of health status. Furthermore, as shown in the table below, 1. Reduction in the impact of illness, physical illness, and mental disorders on daily life, 2. Reduction in the use of health services, 3. Reduction in psychological stress, and 4. Cognitive function. 5. A decrease in loneliness was also confirmed.

Appendix Table 1: Self-evaluation of other physical and mental conditions

Type of survey	explanation	Choices used in the survey	Results and p-values*
1. Prevalence (illness, physical illness, mental disorder)	The impact of illness, physical illness, and mental disorders on daily life.	1 = Yes, significant impact. 2 = Some impact. 3 = no	Reduced impact (p=0.051)
2.Use of health services	Frequency of use by nurses, health center doctors, housemen, dentists, and others.	1=Used 0-2 times, 2=Used 3 or more times, 3=Can't say. (Significant difference only for Nurse	Decrease in usage (p=0.070))
3. State of mental stress	Perceptions of tension, difficulty recovering, calmness and tranquility, depression, and well-being over the past month.	1 = Always, 2 = Most of the time, 3 = Some of the time, 4 = Some of the time, 5 = Never	Reduced stress (p=0.003-0.162)
4.Cognitive function status	One aspect of mental health. Memory, ability to learn new things, ability to concentrate, etc.	1 = very good, 2 = good, 3 = satisfactory, 4 = poor, 5 = very poor.	Improved cognitive function (p<0.001)
5. Experience of feeling lonely	One aspect of mental health.	1 = Never/Rarely, 2 = Sometimes/Often, 3 = Continuously, 4 = Can't say.	Reduced loneliness (p=0.032)

* If the p- value is less than 0.1 (=10%) , less than 0.05 (=5%) , or less than 0.01 (=1%), the difference between the two groups is too large to be caused by chance. That means. The statistical test was a two-group t-test or chi-square test. Sample size was n = 586 in the intervention group and n = 1,047 in the comparison group .

Created from the description in Finnish government report 4.3 .

Additional explanation 2: Additional explanation of " economic happiness "

Respondents who received basic income reported better income levels and better financial well-being across all areas of financial well-being. In addition to those explained in the main text, we also asked questions about the following indicators and received answers.

Appendix 2: Responses to economic well-being

choices / index		1 = no	2= Almost no	3= time Doki	4 = everytime	p-value for chi-square test
1. I am worried about finances .	**intervention group** **comparison group**	14% 11%	13% 13%	47% 41%	26% 34%	decrease (p= 0.011)
2. Manage your money well and pay your bills on time	**intervention group** **comparison group**	1% Four%	6% 12%	20% twenty five%	72% 59%	increase (p= 0.000)
3. Have the opportunity to earn money over the long term.	**intervention group** **comparison group**	20% twenty three%	twenty four% 29%	35% 30%	twenty one% 18%	increase (p=0.059)
4. Financial difficulties arise when unforeseen circumstances occur in life .	**intervention group** **comparison group**	41% 47%	13% 16%	18% 17%	27% 20%	decrease (p=0.006)

(Source) Created from table 5.4 of the Finnish government report.

Additional explanation for " trust in society " and " trust in yourself and the future "

Without trust, life would be chaotic and we would be unable to have meaningful social interactions with other people. As explained in the main text, it can be seen that the mean value of trust in social institutions was significantly higher in the intervention group than in the comparison group. Furthermore, regarding confidence in oneself and the future (seven items below), it can be seen that the mean values of the intervention group were significantly higher than the mean values of the comparison group.

Appendix 3: Confidence in yourself and your future

choices / index		0= do not have	1= almost none	2= Okay	3= quite a lot	4= very much	I can not say	average	p-value for chi-square test
1.My own future	**Intervention group**	6.8%	7.8%	25.8%	34.8%	23.4%	1.4 %	2.57 _	difference is significant p = 0.000
	Comparison group	9.8%	13.2%	30.1%	30.0%	16.2%	0.7 %	2.28 _	
2. Financial situation	**Intervention group**	1 3.0%	11.8%	32.4%	26.3%	15.9%	0.7%	2.19 _	difference is significant (p= 0.000)
	Comparison group	1 9.4%	16.8%	32.4%	19.4%	10.9%	1.1 %	1,83 _	
3. Survive in difficult situations	**Intervention group**	5.6 %	4.6%	22.2%	35.7%	29.9%	2.0 %	2.76 _	difference is significant (p=0.01)
	Comparison group	8.2 %	8.2%	26.6%	31.3%	24.2%	1.4 %	2.52 _	
4. Possibility of work opportunities	**Intervention group**	1 6.2%	9.7%	21.1%	27.1%	23.9%	2.9 %	2.29 _	difference is significant (p=0.000)
	Comparison group	2 0.7%	12.6%	21.2%	22.1%	20.4%	3.0 %	2.03 _	
6.Improve living standards	**Intervention group**	1 4.7%	12.1%	28.9%	17.0%	14.8%	6.0 %	1.80 _	difference is significant (p= 0.000)
	Comparison group	2 3.1%	18.2%	31.1%	17.1%	9.6%	3.1 %	1.70 _	
7. Capacity building for social issues	**Intervention group**	1 6.0%	1 9.5%	31.1%	17.1%	11.8%	4.6 %	1.80 _	difference is significant (p=0.000)
	Comparison group	2 5.1%	22.6%	26.1%	15.5%	7.1%	3.2 %	1.50 _	

Created from the table in Finnish government report 7.5 . The "average" was calculated independently by the author.

Experimental design (RCT) application example 5:
Is microfinus a miracle? (India)

1. Where the problem lies

Microfinance emerged in the 1970s as a trump card for poverty reduction, and has since become rapidly popular. As of December 2007, it was announced that 154.86 million people (including over 100 million women) were receiving the service (as announced by the Microcredit Summit Campaign). In 2006, Grameen Bank and its founder Dr. Mohammad Yunus were awarded the Nobel Peace Prize.

However, debate continues as to whether microfinance is truly effective in reducing poverty. Pitt and Khandker (1998) conclude that there are significant effects, particularly for women. On the other hand, Morduch (1999) and Rodman & Morduch (2009) are consistently negative, stating that no solid evidence has been confirmed (Takahashi 2011) . In order to provide solid evidence for these controversies, this example uses RCT, the most rigorous method.

2. Overview of measures and evaluation design

The target areas, samples, and intervention activities are as follows.

Table 1 Overview of RCT application regarding microfinance

Target area	Hyderabad, India (capital of Andhra Pradesh state)
sample	104 districts (implementation: 52 districts, comparison: 52 districts)
Implementation year	2005 baseline survey, 2006-2007 project implementation, August 2007 endline survey
act of intervention	Spandana , which adopted Grameen Bank's group lending method, is implementing a microfinance business.

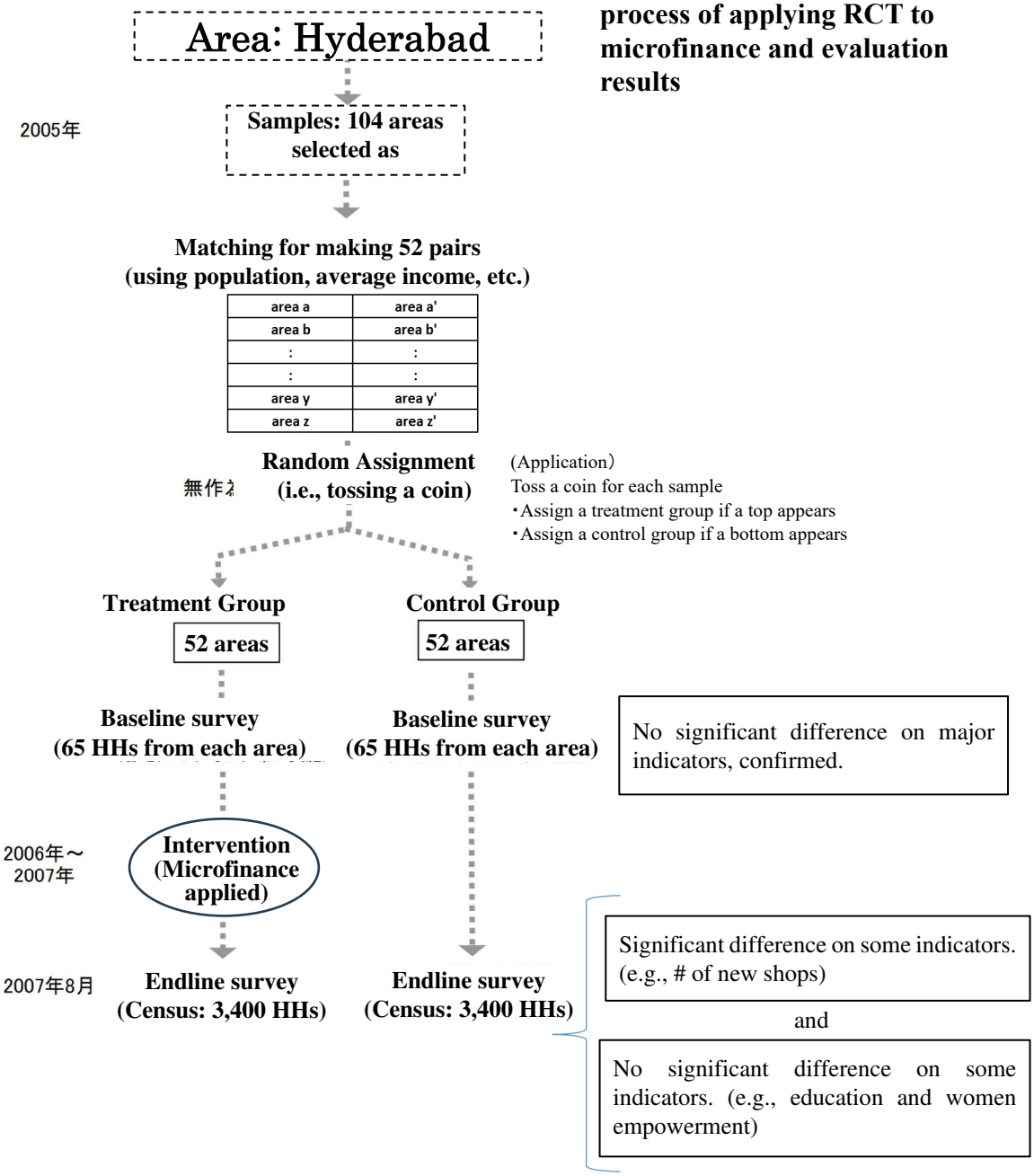
The target area was Hyderabad, India (capital of Andhra Pradesh state), and 104 districts were selected from the city and 52 pairs were formed through one-on-one matching. After that, random assignment was performed within each group, and one district was divided into an implementation district (loaning) and another target district (no financing). As a result, 52 implementation districts and 52 comparison districts with similar characteristics were formed.

Loan eligibility: (a) female, (b) 18-59 years old, (c) resident in the same area for at least one year, (d) have a valid ID, (e) at least 80% of the group own a home; On the other hand, unlike Grameen Bank, it does not provide training to groups. The loan amount is 10,000-12,000 rupees and the interest rate is 12% (equivalent to 24% annual interest rate).

A baseline survey was conducted in 2005 to confirm that there was no difference in the economic averages

of the two groups. From 2006 to 2007, Spandana , a microfinance bank that adopted Grameen Bank's group lending approach, conducted the lending business. An endline survey was conducted in August 2007 to measure differences in index groups between the two groups.

Figure 1 Overview of the process of applying RCT to microfinance and evaluation results



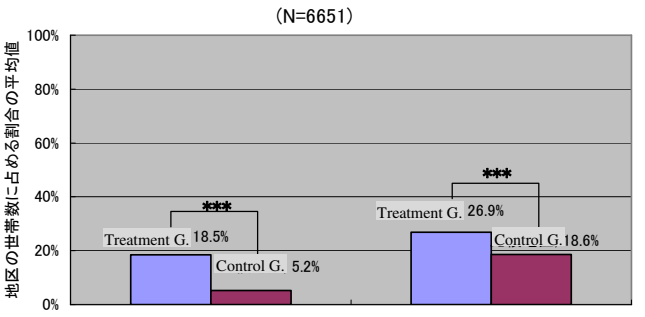
3. Evaluation results

for this case were as follows. (Significance level: ***1%, **5%, *10%)

(1) Borrowing from microfinance (Figure 2)

The percentage of households that received loans from Spandana was 18.5% in the implementation districts (52 districts) and 5.2% in the comparison districts (52 districts), a difference of 13.3% . It can be seen that a small number of people in the comparison areas also applied for and received loans from Spandana . Additionally, the percentage of households that received loans from microfinance institutions, including Spandana , was 26.9% in the implementation area and 18.6% in the comparison area. Therefore, it is concluded that **the implementing districts received more loans .**

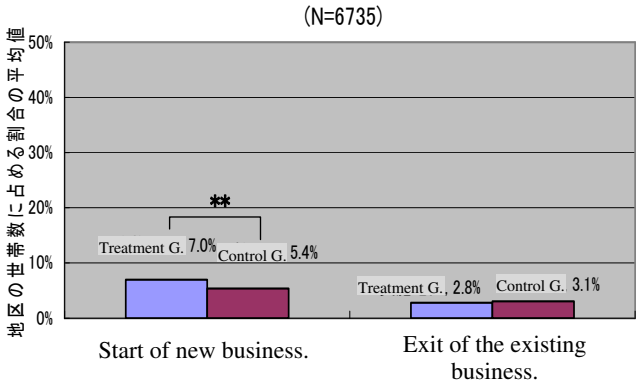
Figure 2 Borrowing from microfinance (incl. Spandana)



(2) Impact on starting new business (Chart 3)

The rate of starting new businesses was 7.0% in the implementation districts (52 districts) and 5.4% in the comparison districts, a difference of 1.6%. This was determined to be significant at the 5 % level. On the other hand, there was a possibility that competition would arise due to the start of a new business, and there would be cases where businesses would be more thorough, but the difference was 2.8% in the implementation area and 3.1% in the comparison area, which is within the margin of error. It was determined that Therefore, it was concluded that **the implementation of microfinance increases the number of new business starts .**

Figure 3 Impact on new business



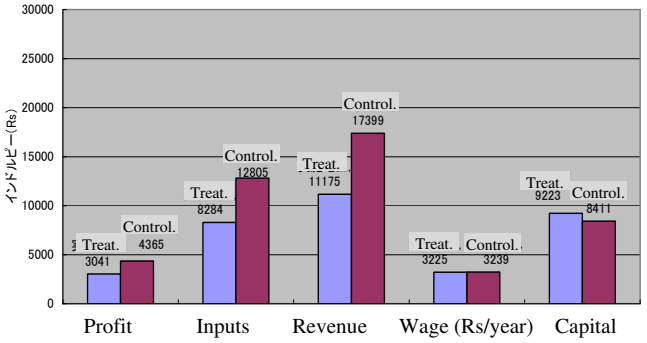
(3) Effects on new business (Chart 4)

When comparing only new businesses between the implementation area and the comparison area, the **average values for profits, inputs, and income were lower in**

the implementation area, but none of the differences were judged to be significant. There is no significant difference in terms of wages or capital either. Although this can be considered a new business, there are many cases, ranging from cases where the business has quickly expanded with high profits, to cases where the business is barely surviving, and the dispersion (= standard deviation) of the numbers is large . This is thought to be because Another factor seems to be that the sample size is small because it is limited to new businesses.

Figure 4 Effect on new business

(N=356)
(All are not statistically significant due to the large standard errors.)

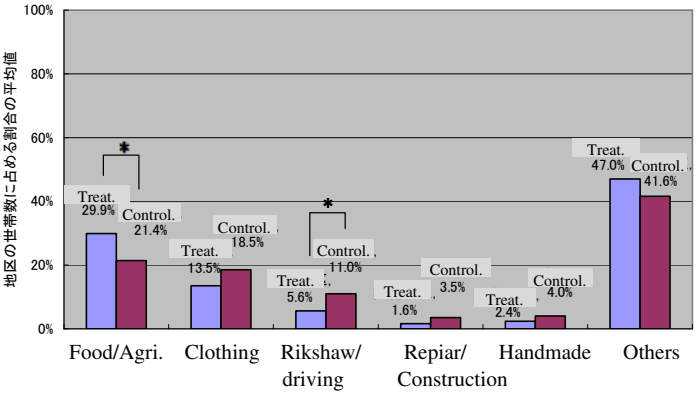


(4) Types of new business (Figure 5)

In the implementation areas, **``food/agriculture'' was more common, and ``rickshaw/driving'' (rickshaws were taxis) was less common .** It is explained that the former is a business that can be started immediately with a small amount of capital, while the latter is a business that requires the most capital in this classification.

Figure 5 Types of new business

(Treatment n=251, Control n=173)

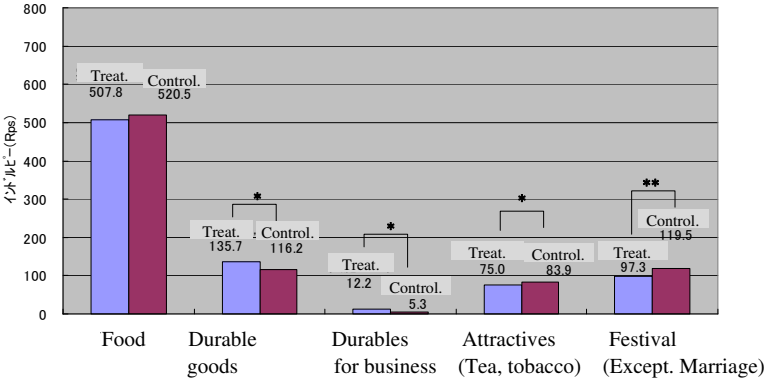


(5) Impact on monthly household expenditure (Figure 6)

While spending on durable consumer goods used in business is increasing, spending on "luxury goods" (tea, cigarettes, alcohol , and festivals (excluding weddings)) is decreasing; Movement can be seen. Furthermore, when we divide the data into households that have been in business for a long time, households that are likely to start a new business, and households that are unlikely to start a new business, we find that households that are likely to start a new business are This trend was observed to be more pronounced.

Figure 6 Impact on monthly household expenditure

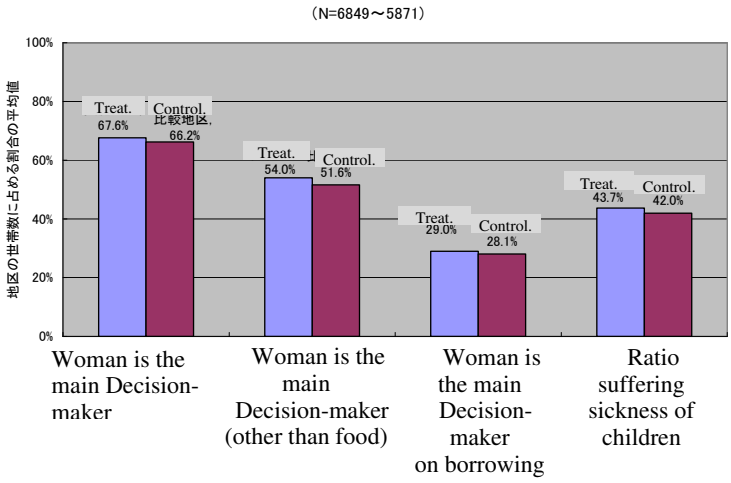
(Total expenditure: TG 1429.1 Rp, CG 1419.3 Rp. The difference (9.9Rp) is not statistically significant.) (n=6775 ~ 6821)



(6) Effects on women's empowerment, health, and education (Figure 7)

All indicators **were higher in the implementation districts than in the comparison districts, but the differences were not judged to be statistically significant.** (The author (Sasaki) is thinking.)

Figure 7 Effects on women's empowerment, health and education



4. conclusion

Through the above analysis , the following conclusions regarding the microfinance business were reached.

Microfinance has some effect on starting new businesses. It also has the effect of increasing investment in durable consumer goods, including those related to business, and decreasing expenditures on "luxury goods" (tea, cigarettes, alcohol, etc.) and festival-related items. No health effects were observed (at least in the short term).

Microfinance may not be a 'miracle' as is often claimed, but it does make it possible to borrow, invest and expand your business.

5. discussion

There are already many papers on the benefits, concerns, and limitations of applying RCTs (e.g., Bauchet & Morduch (2010), which the author also summarized based on discussions with Banerjee of the Poverty Action Lab (Sasaki 2010). There is no need to repeat it, so I will only mention the following points .

Through the application of RCTs, it has become increasingly clear what works and what does not regarding development aid . It is hoped that this will contribute to appropriate policy selection for achieving the SDGs in the future. However , it is also a question of how policy makers appropriately understand the evaluation results of RCTs and reflect them in policy, and as those who use RCTs to make evaluations, it is important for policy makers to continue their efforts. We must provide support.

Furthermore, through this review, we were concerned that the papers were quite specialized . It appears

that the level of knowledge needed to complete a statistics course at a social science graduate school is required (3 to 4 courses may be required) . However, **the great advantage of RCTs is the simplicity and ease of understanding of comparing the average values of two groups** , and these advantages must be maintained. Even in the papers reviewed this time, there were many cases in which complex regression analyzes were performed using RCT data . Since regression analysis does not reveal the exact effect of an intervention, RCTs have attracted attention and become popular, and we should go back to their origins.

However, it must be said that a minimum level of knowledge of statistics is still necessary to properly understand papers based on evaluation results using RCTs . This includes calculating the mean and standard deviation, testing the significance of two groups, standardizing data, and knowledge of multiple regression analysis. From my experience, I can tell you that this kind of statistical knowledge can only be acquired by attending classes and practicing calculating by hand using a calculator or Excel. It must be said that this method of learning is fundamentally different from qualitative methods such as interviews (key informants, focus groups), direct observation, and participant observation, which are all about getting used to learning . It seems that ``adult learning of statistics'' is necessary for those working in the world of development aid.

(Source) Banerjee, A., Duflo , E., Glennerster , R., & Kinna , C. (2010). *The miracle of microfinance? Evidence from a randomized evaluation.* Poverty Action Lab.

(Reference) Examples of expert evaluation

Examples of expert evaluation

Seafarer education (Egypt)

This is an application example of a very simple design that does not fit into the five impact evaluation designs. Although it is not recommended at all, I will introduce it here for reference.

Problem location and evaluation results

The Arab Maritime Academy (AMTA) was established in Alexandria, Egypt in 1972 with contributions from League member countries, based on a resolution at the 12th Council of Transport and Communications of the Arab League held in 1970 . Ta. The purpose of its establishment was to train ocean-going vessel crews and land-based workers in order to strengthen the Arab League countries' own fleets in order to transport their own oil and improve their international balance of payments.

1. Summary of measures

A The MTA had planned to get its operations back on track in the five years ending in 1977 with assistance from UNDP and others, but the plan was delayed due to a lack of budget . Japan requested assistance from Japan in 1974, and provided assistance to AMTA for four years starting in 1976 . The aid aimed to strengthen mariner training institutions at AMTA 's Maritime Training Center, School of Navigation, and School of Engineering. Aid continued thereafter.

2. Evaluation results

Experts will conduct evaluations through on-site inspections and interviews. Prior to the on-site inspection, the following activities were conducted in Japan.

(1) Field tour of the navigation training ship Seiun Maru (in Tokyo Bay)

(2) Visit and inspection of the Navigation Training Center headquarters (in Yokohama)

Through the field survey, the evaluators came up with the following evaluation results. “ AMTA has trained 24 leaders, and even now , approximately 20 years after the end of cooperation, many of them are still working in AMTA 's successor organization. In addition, most of the training participants hold seminars and lectures after returning to their countries to re-transfer and disseminate the skills they acquired through the training, which is expected to expand the effectiveness of the training. Therefore, it can be said that the objective of ``training ocean-going vessel crews and shore personnel'' has been achieved over a long period of time.

3. Advantages, limitations, and considerations regarding application in Japan

The advantage of this method is that **it is simple .** There is no need to prepare any particular data at either

the preliminary or post-event stage. What we should compare is the difference between the standards in the mind of the person performing the evaluation and the impression received by the person performing the evaluation at the post-event stage.

And the advantages are just the limitations of this method. Needless to say, this method is much more vague and unstable than the methods described so far . If you ask me what the basis for the evaluation results using this method is, all I can say is that Professor XX , professor emeritus at XX University, said so.

In fact, until recently, most of the evaluations conducted in Japan were `` evaluation by experts " using exactly this method. The expert who was asked to conduct this evaluation was a newspaper reporter, and he pointed out the following points to keep in mind when using this method in the future. ``In order to make the ``expert evaluation survey'' more effective, that is, to improve the quality of the evaluation survey itself, it is necessary to conduct domestic visits in advance to the sending organizations that have dispatched long-term experts to carry out technical cooperation. I would like to propose that it be made compulsory. .Honestly**, I don't know if I would have been able to conduct a satisfactory on-site investigation if I hadn't had the on-site tour of the navigation training ship Seiun Maru in Tokyo Bay and the visit to the Navigation Training Center Headquarters in Yokohama. I have no confidence at all.''** The ``evaluation standards'' in expert evaluations that do not set a comparison group or preliminary baseline data are the standards derived from the experts' internal standards and experience; The success of evaluation using this method almost entirely depends on whether or not it can be set appropriately.

4. discussion

This case is an outside example that does not fit into any of the five impact evaluation designs . This is the conventionally common design: "on-site inspection + interviews with stakeholders." This has the advantage of being simple, but it does not reveal any evidence. In addition, this design is ranked at the bottom of the "Appendix 1: Impact Evaluation Design List (Detailed Version) and Evidence Pyramid" published in this booklet. However, it is much better than no evaluation at all, and even better than "self-evaluation" carried out by the implementing agency itself.

(Source) Japan International Cooperation Agency (2000) already published " 2000 Project Evaluation Report" Chapter 3 Ex-post Evaluation Survey III. Expert Evaluation The author created his own explanatory text based on the description of Seafarer Education Egypt. The original PDF file can be downloaded from below.

http://www.jica.go.jp/evaluation/general12/pdf/313.pdf

Appendix 1: Impact evaluation design list (detailed version) and evidence pyramid

In Rossi et al.'s Evaluation: A Systematic Approach, which is one of the standard textbooks on evaluation studies in the United States, there are three types and 12 representative designs for impact evaluation. This text uses five representative designs.

List of impact evaluation designs

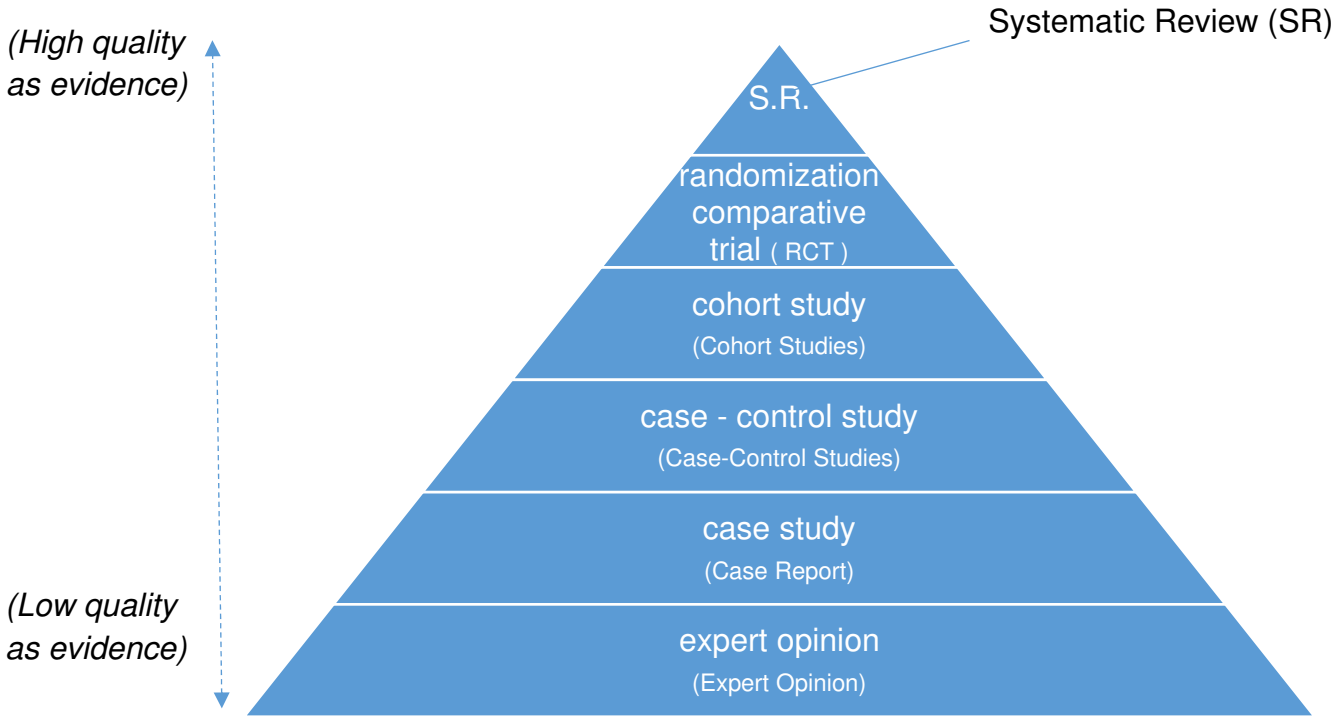
Classification of impact evaluation	Features/Restrictions	Objectivity/total cost/difficulty of use
A. Case where both implementation and comparison groups exist		
(1) Random comparison design	→Before implementing the "policy," set implementation/comparison groups by random assignment of policy application.	Extremely high Extremely difficult
(2) Quasi-experimental design		
① Regression/division design	→Before policy implementation, divide the sample group into two based on specific values and set up an implementation/comparison group.	high high difficulty
② Matching design	→Select groups that are as similar as possible and use them as comparison groups.	
③ Statistical equalization design	→ Divide the sample population into implementation and comparison groups using statistical processing.	
④ General indicator design	→Use national average values, all prefecture average values, etc. instead of comparison groups.	low low easy
B. Case where there is only an implementation group (Eg national target program)		
(3) Cross section design	→Evaluate impact by taking advantage of variations in the amount of service inputs and improvement effects across multiple groups and regions.	high high difficulty
(4) Time series design	→Measure and compare the index values before and after the event over a long period of time.	
(5) Panel design	→Compare the indicator values before, during, and after the short period.	
(6) Pre/post comparison design	→Simply compare the index values before and after.	low low easy
C. Simple approach		
(7) Expert evaluation	→So-called "experts" such as academics and experts set the baseline.	low low easy
(8) Beneficiary evaluation	→Beneficiaries set a baseline through questionnaires and interviews.	
(9) Administrative officer evaluation	→The administrative official in charge of policy implementation evaluates the baseline.	Very low Very low Very easy

(Source) Rossi, Freeman, Lipsay The author made some changes based on the classification in the table in Evaluation A *Systematic Approach, 6th Edition, Sage Publication, 1999, p.261.* However, "characteristics/constraints" and "objectivity/total cost/difficulty of introduction" were described based on the author's own experience and judgment.

Furthermore, there is an "Evidence Pyramid" that lists the quality of evidence. The origin and outline are listed below.

Evidence-based Policy Making (EBPM) is an international research trend that was originally proposed based on Evidence-based Medicine (EBM). In 1993, the American Agency for Health Care Policy and Research (AHCPR) proposed a ranking of evidence in clinical research. In response to this proposal, various organizations and researchers have proposed ranks of evidence in social science research as well. An example of the currently proposed evidence ranking is shown below, but many researchers have proposed various variations that have led to the current status.

Figure : Evidence Ranks (Evidence Pyramid)



	Approach	Explanation
1a	Systematic Review (SR)	Meta-analysis of multiple randomized controlled trials
1b	Randomized Controlled Trial (RCT)	Prospective; Randomization , Comparison between the treatment group and the control group
2	Cohort Studies	Prospective; No randomization; Comparison between the treatment group and the control group
3	Case-control studies	Retrospective ; No randomization; Comparison between the treatment group and the control group
Four	Case Report, Case Series	Narrative reviews
Five	Expert opinion	Expert opinion

Source: Walden University *Levels of evidence pyramid* . Quoted with slight modifications.
(https://academicguides.waldenu.edu/healthevidence/evidencepyramid)

Appendix 1 : Controversy over evaluation part 2

"Scientific evaluation" VS. "Practical evaluation"

This controversy can be said to be a long and deep-rooted controversy. It is also a fundamental controversy over the nature of evaluation. No decision has been made. By understanding this discussion, we will be able to see both the advantages and limitations of evaluation.

Scientific Evaluation

Campbell , D.T. , made the following claim at the beginning of his 1969 paper: "The United States and other modern nations **must be prepared to use experimental approaches to social improvement** . Experimental approaches are new measures designed to solve specific social problems. This is an approach used when implementing measures, and through this approach, it is confirmed whether the measures have had clear effects when compared with multiple criteria, although it is incomplete, and based on the results of that confirmation, whether the measures should be maintained or not. , the decision is whether to improve or discontinue." (Campbell, DT , 1969, p409)

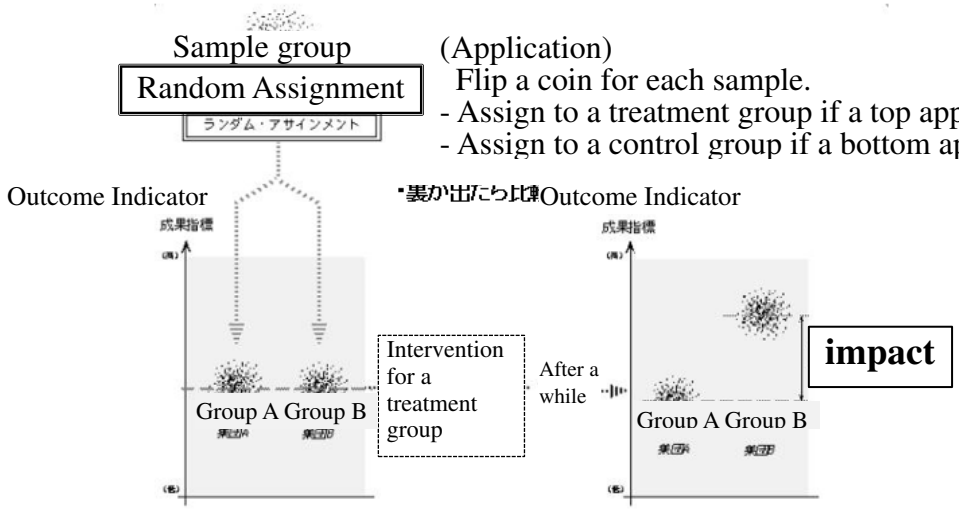


Diagram (above): "Experimental approach" explained by Campbell (= a regular RCT)

Practical evaluation **(Pragmatic evaluation (Rossi and Freeman), Practical program evaluation (Hatry , Wholey), Practical Evaluation (Patton),** etc. are called in English in various ways depending on the researcher)

Cronbach, who originally specialized in statistics, **(Cronbach, LJ)** argued in his 1982 book: " **Designing evaluation research is an art...The** central purpose of evaluation is different from basic social research, and evaluation should fit into different institutional and political contexts. Many recommendations that would be appropriate for long-term efforts such as scientific research are not appropriate for evaluation.Furthermore, general papers on scientific method and design are inappropriate for evaluation practitioners. General recommendations about evaluation are also misleading.Evaluation should not be pigeonholed into one mold.There can be many good designs for any evaluation, but there is no perfect design. Impossible." (Cronbach, LJ, 1982, pp1-2) . And he declares: ``**Evaluation is an art ' '** and there is no single best plan for evaluation, even if it is the study of a particular program at a particular time, within a particular budget. (Cronbach, LJ, 1982, pp321)

Rossi , who has been watching the debate between the two from the 1970s to the present, **(Rossi, PH)** is explained as follows.

`` **Scientific Versus Pragmatic Evaluation** Postures" Perhaps the most influential paper in the world of evaluation research was published by Campbell in 1969. This paper presents the view that Campbell has advocated for decades: that policy and policy decisions are based on the results of ongoing social experiments that test ways to improve social conditions. Not only that, but he also said that the techniques of social research could be used to actually realize an 'experimental society'.Campbell also said that in social psychology he learned and applied Although he softened his position somewhat in his later works, Campbell sought to apply the experimental model, which

is a method based on scientific research, to evaluation research. It would be fair to regard him as a person who

Meanwhile, Campbell's position was challenged by Cronbach, another giant in the evaluation field. While denying that scientific research and evaluation may be used in the same way as research procedures, Cronbach argued that the purpose of evaluation is clearly different from the purpose of scientific investigation. In his view, evaluation is more art than science, and all evaluations should be shaped to meet the needs of decision makers and stakeholders. Therefore, while scientific research fundamentally struggles to meet the standards of research, evaluation is an important part of decision-making within the political environment, policy constraints, and available resources. should contribute to providing the most useful information to the public. ” (Writing (3))

At the same time as Cronbach (1981), **the** following points were also made by **Harry Hatry and Joe Wholey .**


``...There is a growing recognition that the degree of application of classic evaluation designs is limited and that they impose difficulties that go beyond common sense thinking.In addition, there is a growing recognition that the degree of application of classic evaluation designs is limited and that they impose difficulties that go beyond common sense thinking.Also, how useful evaluations are? (Hatry , Winnie & Fisk, 1981, p. ix) .

Finally, looking at recent works on evaluation research, the following trends are observed.

``The watchword among evaluation experts these days is ``utilization-focused evaluation." Utilization-focused evaluation answers specific questions posed by those entrusted with implementing the program. An evaluation designed to help influence decisions about the future of the program.—Three basic questions should be asked of any program when it comes to evaluation and monitoring: (1) Can the results of the evaluation influence decision-making regarding the policy? (2) Can the evaluation be completed by the time the evaluation results are needed? (3) Can the policy be evaluated? (Wholey , Hatry & Newcommer , 1994, p5)

Furthermore, as this trend toward ``practical-oriented evaluation" becomes stronger, ``scientific evaluation is actually much more practical because the evaluation results of scientific evaluation are referred to and used over a long period of time." (I remember that this was said by Campbell's ally Cook , D.) .

Campbell, D. T. (1969). ``*Reform as Experiments"* American Psychologist , April 1969, 24:p.409
Cronbach, L.J. (1982). *Designing Evaluation of Educational and Social Programs* , San Francisco: Jossey-Bass.
Rossi, Freeman and Lipsay. (1999). “Scientific Versus Pragmatic Evaluation Postures” In *Evaluation: A Systematic Approach 6th edition*, pp.29-30, Sage publications

Hatry, Winnie & Fisk. (1981) *Practical Program Evaluation for State and Local Governments, 2nd ed.* Urban Institute,
Wholey, Hatry & Newcomer (Ed.) (1994). ''Meeting the Need for Evaluation'' In *Handbook of Practical Program Evaluation*, Jossay-Bass.

(Source of image 1) Sketched by the author with reference to the image at https://en.wikipedia.org/wiki/Donald_T._Campbell .
(Source of Image 2) Sketched by the author with reference to the image at https://archon.library.illinois.edu/?p=digitallibrary/digitalcontent&id=10865 .
(Source of Image 3) Sketched by the author with reference to the image at http://www.columbia.edu/cu/csswp/1995.htm .
(Source of Image 4) Sketched by the author with reference to the image at http://www.columbia.edu/cu/csswp/1995.htm .
(Source of Image 5) Sketched by the author with reference to the image at http://www.businessofgovernment.org/bio/joseph-wholey/ .

(Source) Excerpt from Sasaki (2003) (pp.20-23)

Appendix 3 : Controversy over evaluation part 2

“ Quantitative evaluation ” vs. “ qualitative evaluation ”

This is also a long and deep-rooted controversy. After quantitative evaluation became widely recognized in the 1960s and 1970s, proponents of qualitative evaluation emerged and often criticized quantitative evaluation, and quantitative evaluation silently endured the criticism.

Claims on the side of qualitative evaluation

“Traditionally, evaluators (those using quantitative methods) have tried to do more than they are actually capable of, measuring improvement effects and isolating important factors from other factors. In the end, they end up serving separate political positions, and insufficiently." (Stake, 1980, p38)

Claims on the side of quantitative evaluation

``The argument that qualitative methods should be used rather than the currently mainstream quantitative methods is almost mystical, and also accepts the views of those implementing the measures themselves when it comes to identifying improvement effects." Rossi, 1985, p7)

Recent discussions (until the end of the 1990s)

- “There is an opinion that qualitative evaluation lacks statistical rigor. However, it is not appropriate to require statistical rigor in evaluation, but rather to understand the concerns of socially vulnerable people. Some argue that qualitative evaluation is more appropriate for this purpose." (Bamberger, 2000,)
- “Qualitative and quantitative evaluations both have advantages and disadvantages. They can be used interchangeably, but they can also be used at the same time. And data from both types can be collected at the same time in the same evaluation study. (Patton, 1990, p14)
- “A combination of quantitative and qualitative methods is ideal because it provides the quantitative impact of the project as well as an explanation of the process or intervention that produced the outcome.” (Baker) , 2000)

method ' ' using qualitative and quantitative methods was proposed in one evaluation , and it can be said that a certain degree of consensus was reached around the end of 1990 . This dispute has now been tentatively settled.

More recent discussions (since 2000)

However, in the 2000s , there was a new movement. This is a counterattack on the part of quantitative evaluation. In 2003 , the Povery Action Lab was founded at the Massachusetts Institute of Technology (MIT) and declared that it would use only experimental designs (RCTs) . In the 20 years since its establishment, it has conducted impact evaluations using experimental designs (RCTs) in more than 2,000 impact evaluations . In 2019 , three of its founders received the Nobel Prize in Economics. The quantitative evaluation side has silently endured the criticisms of the proponents of qualitative evaluation , but finally the quantitative evaluation side has succeeded in

counterattacking.

Outlook for future discussions (after 2020)

The debate between qualitative and quantitative evaluation will continue. And although they may shake hands and come to the conclusion that both methods should be used at the same time, both parties will never change their secret beliefs.

Stake, R. (1981) *The Art of Case Study Methods.* Sage Publication

Rossi, PH . (1985). *Evaluation: A Systematic Approach, 5th ed* . Sage Publication

Bamberger, M. "The Evaluation of International Development Programs: A View from the Front" In *The American Journal of Evaluation (Winter 2000)*

Patton, MQ (1990). *Qualitative Evaluation and Research Methods, 2nd edition* .

Baker, J. (2000). *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners,* The World Bank.

(Source) Excerpt from Sasaki (2003) (pp.24-25) . 2 Added in 022 .

Appendix 4: Consideration regarding the pros and cons of randomized comparative design : Interview at MIT Poverty Action Lab (2006)

The 2019 Nobel Prize in Economics was awarded to Professors Abhijit Banerjee and Esther Duflo from the Massachusetts Institute of Technology (MIT) and Professor Michael Kremer from Harvard University . The Abdul Latif Jameel Poverty Action Lab (J-PAL) at MIT continues to use randomized controlled designs (RCTs) to find effective ways to alleviate global poverty . The project was recognized for its efforts to clarify the measures . In 2006, the author (Sasaki) visited the Poverty Action Lab and had the opportunity to interview Professor Banerjee, discuss the application of RCT, and write a paper about it, an excerpt of which is posted below. .
(Author's note: The word "Banaje" appears in katakana in the text, but in recent years it has been written more often as "Banajee.") However, in this excerpt, I have kept the wording as it was at the time of writing the paper.)

Introduction: Significance of this paper

a leading expert in evaluation research (Shadish, W. et al, 1991) and sometimes called the only philosopher in evaluation research (Stake, RE 1982), has a wide range of questions regarding what it means to prove causality. There have been many philosophical studies (for example, Scriven, M. 1975). And Scriven has consistently been a sharp critic of the claim that randomized experimental designs are the best design for proving causation. For example, the U.S. Department of Education's No Child Left Behind Act of 2001 clearly states that randomized experimental design is the most desirable research design. When the American Evaluation Association was divided into two groups on whether it was for or against randomized experimental design, Scriven was at the vanguard of those criticizing randomized experimental designs (for details, see *The 2004 Claremont Debate* (2004). Davidson, SI & Christie, CA (2004).*).* A paper that systematically organizes the criticisms Scriven has developed is ``The logic of Causal Investigation" (Scriven, M. (2007). The *logic of Causal Investigation).* In his work, Scriven points out that it is a logical fallacy to claim that randomized experimental designs are the best design when evaluating social policies, including education and health.

On the other hand, Professor Dr. Abhijit Banerjee, an economist and director of the Poverty Action Lab, makes it clear in his book (2007) that only randomized experimental designs can determine whether a study really has an effect. are doing. He also stated that he started the activities of the lab to popularize randomized experimental designs in the development assistance field. There

is no evidence that Scriven and Banerjee had any direct exchange, partly because they had different fields of expertise. However, the author (Sasaki) had the opportunity to visit the Lab. in 2006 and have an interview with Banerjee. (Sasaki, R. (2006). *Discussion with MIT's Poverty Action Lab.*).

Below is an overview of Scriven's opinion and the summary of interview responses. Furthermore, the author (Sasaki), whose research has focused on development aid evaluation for many years, has included his thoughts on each topic. In other words, we aimed to reach a conclusion to the discussion by organizing the opinions of three people corresponding to the three roots of evidence-based development aid evaluation (the lineage of evaluation research, the lineage of development aid, and the lineage of economics). What we have learned through this sorting process is that there is actually not that much difference between the various claims.

One notification is the below discussion is summarized by the interviewer (Sasaki) and the summary might not be correct and not appropriately represent what any person intended because of the interviewer's limitation of the subject-matter knowledge and limitation is English ability. Thus, the interviewer (Sasaki) is fully responsible of this article and any other person in this article has no responsibility. If you have any inquiry, please contact the interviewer (Sasaki) (email: sasaki.ryo(a)idcj.or.jp. Please use @ instead of (a)).

(1) Poverty Action Lab's randomized experimental design lacks blinding

Scriven points out that the lab's design lacks a double-blind design, which is often used in the health care field. In other words, it is a "Zero" blind study, which leaves room for the so-called Hawthorne effect to enter, so it cannot be said that it can truly reveal the effects of the intervention (Scriven M. 2007).

The summary response that the interviewer (Sasaki) understand: That point is correct, but since it is not a clinical trial, it is not possible to use a true drug and a placebo. Instead, they argue that they seek to minimize the effects of unblinding by having the experimental and control groups share the same information.

In conclusion, it can be said that it is acknowledged the limitations of the design the interviewee used, but even Campbell, who introduced randomized experimental designs to the social sciences, did not discuss the application of the blind method. As he said, as long as social measures are targeted, there is no other way.

(2) Statistical significance and social significance are different

Scriven argues that it is not enough for the difference between the two groups to be statistically significant, and that the intervention cannot be said to have been effective unless it is socially or

practically significant (Scriven M. (2007), the Poverty Action Lab points out that interventions are judged to be effective with statistical significance.

The summary response that the interviewer (Sasaki) understand: Although he does not reject the idea, it cannot be socially significant unless it is statistically significant. In other words, statistical significance is the minimum condition that must be met in order to consider whether something is socially significant. The lab also counters that it listens to the opinions of local people to determine whether it is socially meaningful, but does this answer the question of whether local people decide whether it is successful or not? When asked, they answered 'Consulted'.

In other words, the judgment of whether it is effective or not should first be made through professional statistical analysis, and leaving it to the judgment of local people without doing so would undermine the rigor that is the advantage of the randomized experimental design. I can understand that you are doing it.

(3) There are ethical issues with dividing people into halves based on chance, and it is difficult to obtain prior approval for such treatment.

Scriven argues that parents living in developing countries will not approve of a procedure that might put their child in a control group, and raises concerns about unethical practices and the difficulty of obtaining informed consent in randomized experimental designs. (Scriven M. 2007).

The summary response that the interviewer (Sasaki) understand: Based on the interviewee's experience in the field, randomization is, in fact, fair. First, it is rare that aid resources are sufficiently prepared to cover everyone, and it is difficult to apply randomization (the idea of a lottery based on the same probability) to avoid arbitrarily deciding who will receive the aid. He points out that it's very fair. Second, in the past, instead of randomization, villages along main roads or villages that could be visited in one day were often selected based on explanations such as ``according to the standards of our donors," which made it extremely difficult for residents to understand. I have been told that I thought it was unequal and unfair. He points out that in such situations, when the idea of randomization is approached, it is often welcomed.

In conclusion, it is persuasive that there is less resistance to randomization than aid agencies assume, and that it is not a constraint on the application of randomized experimental designs. In reality, as everyone involved in aid has experienced, including the author, the target areas and people have been selected based on the donor's convenience, so this is a refreshing point.

(4) The word "evidence" is monopolized by researchers using quantitative methods.

Scriven points out that while the idea of "evidence-based practice" is perfectly acceptable, the problem is that the definition of evidence is limited to results obtained through the application of

randomized experimental designs. . 2007). They argue that evidence can be produced by rigorously applying qualitative methods, and call for a reconsideration of the definition of the word evidence.

The summary response that the interviewer (Sasaki) understand: First of all, that the interviewee has no intention of denying any analytical method. We are simply arguing that randomized experimental designs, despite their usefulness, have been used far less often than other methods and should be used more often than they currently are. It is said that it is. Second, he points out that qualitative methods, especially detailed observational descriptions (Stake, 1982), can play a complementary role because they explain "why something happened." For example, they point out that randomized experimental designs provide limited information, making detailed observational descriptions a central piece of explanation.

Scriven also cited a paper by Cook (T. 2000) and concluded that qualitative and quantitative methods can and should be used in a mutually complementary manner, and there is actually a difference between Scriven and Bannerje. It can be said that there is no big difference in perception.

(5) There are types of interventions for which it is inappropriate (or meaningless) to apply a randomized experimental design.

Scriven points out that randomized experimental designs are not always best by pointing out that there are types of interventions for which randomized experimental designs are not applicable due to time and resource constraints (Scriven M. 2007).

The summary response that the interviewer (Sasaki) understand: The interviewee does not mean to say that randomized experimental designs are always the best, but agrees that experience has shown that there are types of interventions for which randomized experimental designs are either impossible or inappropriate. It is said that it will. These include (i) projects that have already been completed (or projects that have already started), and (ii) large-scale projects that cover the entire country (as the name of the randomized experimental design suggests, small-scale projects that are experimental).). Furthermore, (iii) flexible projects in which the contents are not determined in advance and are determined during implementation are not suitable for randomized experimental designs. He also points out that there are unexpectedly many such projects in development aid. Finally, (iv) projects that are known to have a strong intervention effect are sometimes purposely selected as pilot projects, and it is pointed out that this also does not meet the purpose of the randomized experimental design.

In contrast to the declarations made in the Poverty Action Lab's public materials, it sounds being calm. As we gain experience in applying randomized experimental designs, we have come to understand the specific types of interventions that are difficult to apply.

Conclusion

What we can see from the discussion so far is that there doesn't seem to be much difference between the two claims. And the conclusions that the two sides express become increasingly similar.

In the conclusion of his paper, Scriven states that there are certainly times when a randomized experimental design should be used, since it can be said to be a mixed method operated with the help of qualitative methods. (Scriven M. 2007), the interviewee does not believe that randomized experimental designs should dominate evaluation activities, but rather that they should be used more frequently because they have been used far less than other methods. It is concluded that he is simply claiming that.

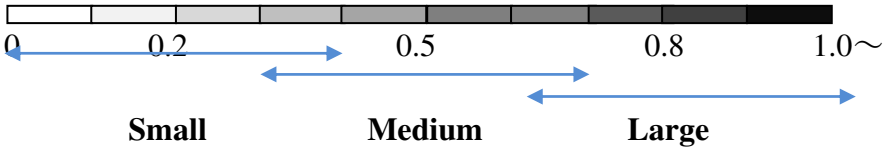
In other words, there is certainly scope, although not exclusive, for randomized experimental designs to be used in development aid evaluation.

[1] Based on an interview conducted by the Japan Institute for International Development and Higher Education. Naonobu Minato, director of the Center for International Development Research at the Institute for Higher Education on International Development, led the investigation.

(Source) Ryo Sasaki (2010) “Evidence-based evaluation of development aid: the history of aid evaluation, the origin of randomized controlled trials, and a comparison of the ideas of Scriven and Banager” In “Special feature: Global trends in evidence-based practice and Japan " Efforts in Japan" Japan Evaluation Research Vol. 10, Np. 1, March 2010 (Editors: Dr. Ryo Sasaki, Dr. Iwao OSHIMA). References cited in the text are listed in the PDF below.
http://evaluationjp.org/files/Vol10_No1.pdf

Appendix 5: Effect Size (ES) criteria newly proposed based on prior research in education

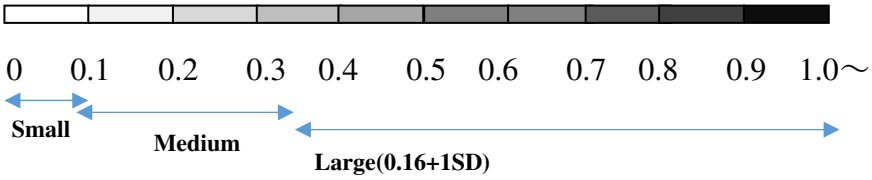
Cohen, J. (1988)) proposed "For statical the of ES index, the author proposes, as a convention, ES (Effect Size) values to serve as operational definitions of the qualitative adjectives "small," "medium," and "large." His proposal has widely accepted and still used as general standard.



(Source) Illustrated based on Cohen, J. (1988)

However, prior to this sentence, Cohen, J. (1988) wrote "He (=a researcher) call upon theory for some help in answering the question and on his critical assessment of prior research in the area for further help." Now I(Sasaki) believe now we have one important prior research, which is meta-analysis of 96 RCTs in education sector (Evans, D.K. and Yuan, F (2022)). Based on this important prior research, I propose a new criteria of ES for definition of "small," "medium," and "large" as follows.

Newly proposed Effect Size Scale for educational interventions



(Source) Sasaki, R. proposed, based on Evan,D. & Yuan,F. (2022)

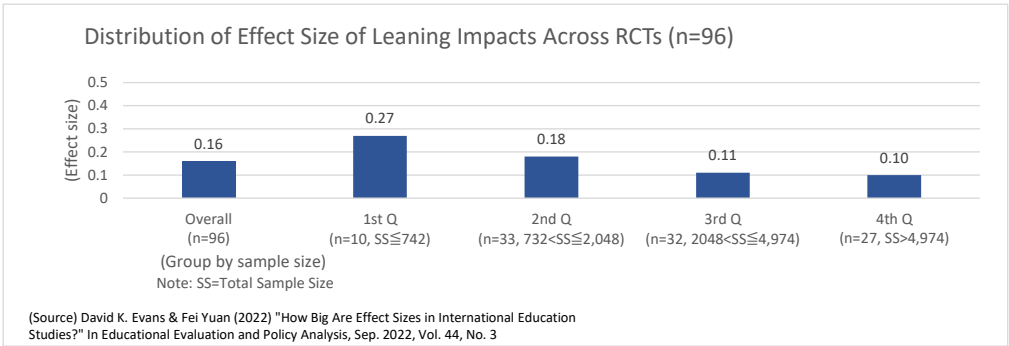
==========.

Annex : The following is the summary of the statistics of 96 RCTs in education sector Evan D. & Yan F. (2022).

Table and Figure : Distribution of Leaning Impacts Across RCTs (n=96)

SS=Total Sample size

	Overall (n=96)	1st Q (n=10, SS≦742)	2nd Q (n=33, 732<SS≦2,048)	3rd Q (n=32, 2048<SS≦4,974)	4th Q (n=27, SS>4,974)
Mean	0.16	0.27	0.18	0.11	0.10
SD	0.25	0.39	0.26	0.15	0.15
No.of effect size	468	81	133	122	132
No. of studies	96	20	33	32	27



(Source) David K. Evans & Fei Yuan (2022) "How Big Are Effect Sizes in International Education Studies?" In Educational Evaluation and Policy Analysis, Sep. 2022, Vol. 44, No. 3

(Source) (1) Cohen, J.C. (1988) Statistical Power Analysis for the Behabvioral Sciences (1988). Lawrence Erlbaum Associates, Publishers.
(2) David K. Evans & Fei Yuan (2022) "How Big Are Effect Sizes in International Education Studies?" In *Educational Evaluation and Policy Analysis*, Sep. 2022, Vol. 44, No. 3
(Download: https://journals.sagepub.com/doi/pdf/10.3102/01623737221079646)

(Proposed by Ryo SASAKI, Ph.D., International Development Center of Japan (IDCJ). 20230728)

Appendix 6: Effect Percent (%)

Although many people conduct it, almost no one has proposed this calculation in formal way. This is the first formal writing of “Effect Percent (%)”. This name is offered by the author (Ryo SASAKI, Ph.D.) for convention of education sector people.

As everybody knows, Effect size is “Two-group mean difference against the consolidated standard deviation”. By following this definition, “Effect Percent (%)” is defined as follows.

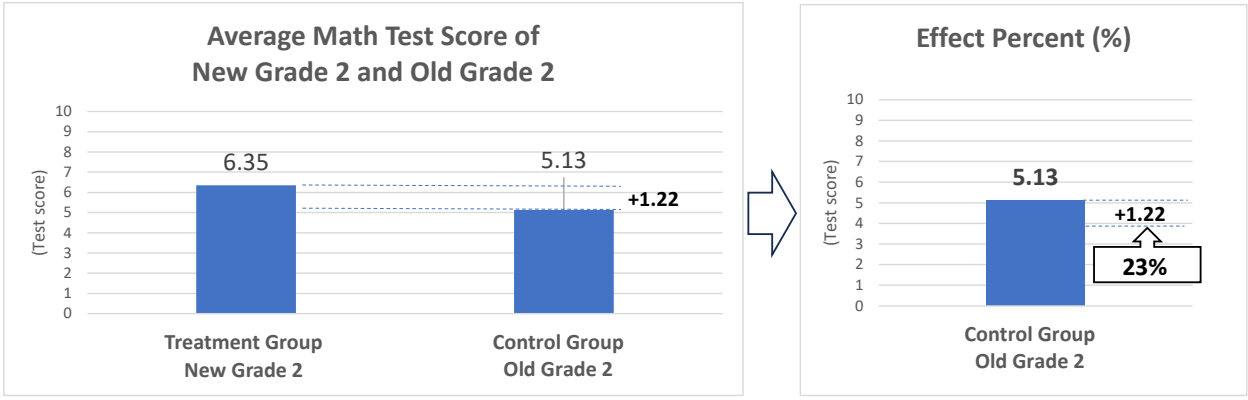
Effect percent (%): Two-group mean difference against the control group mean. (note: mean = average value).

Effect size is plausible for statistics professionals but its meaning, such as “small”-”medium”-“large”, is not clear for general public. On the other hand, Effect percent (%) is simple but very understandable for general public including parents and students, saying the sentence such as “the children’s performance was improved by 23% due to our intervention”. Calculation formula for Effect Percent (%) is as follows.

(Effect Percent (%)) = (Two-group mean difference) ÷ (Control mean)

Here is an example of the calculation (Impact evaluation of education intervention (new math text development) of Myanmar.

23% = 0.23 = 1.22 ÷ 5.13



(Note) Due to calculation of decimal points, the final percentage does not necessarily become the exact percentage in the above figure.

(Source) JICA. (2019). Impact Survey of the Project for Curriculum Reform at Primary Level Basic Education (CREATE) in Myanmar

Here is the reason why a two-group mean difference should be compared with a control mean. Glass V. (1976) proposed Glass's delta, calculating a two-group mean difference divided by a control group's standard deviation instead of the consolidated standard deviation. It is rational for me because consolidation of a treatment group and a control group would have been partially affected by the treatment action. Instead, a control standard deviation would be the pure (or natural) standard deviation without being affected by treatment. By the same thinking, it is recommended that a control mean should be used for calculation of effect percent (%) instead of using the consolidated or the treatment group mean which have been already affected by the treatment action.

The following is the newspaper article in Japanese and the English translation from the original Japanese text is here.

"A major change in education is occurring at elementary schools in Myanmar. This is a shift from an emphasis on rote memorization to interactive classes that stimulate children's excitement. The support was provided by Japanese experts.

Average math score improved by 23%

According to a JICA study, second-grade students who used the new textbooks had an average score increase of about 23% compared to their previous peers. It has shrunk to less than one-third."

ミャンマー 日本流で暗記教育卒業

世界発 2020

ミャンマーの小学校で教育の大変革が起きている。丸暗記重視から、子どもの意欲を引き出す双方向型の授業へ。支援したのは日本の専門家だ。（ヤンゴン＝染田屋竜太）

教師　教科書の絵の人は何をしていますか。

児童　田んぼの水路が見えるので、米づくりをしていると思います。

教師　素晴らしい答えです。みんな、拍手。

ミャンマー最大都市ヤンゴンの公立ヤンキン第13小学校で、1年生が社会科の授業を受けていた。大きな声で発言する児童らが開くのはカラフルなイラストが並ぶ教科書だ。「数年前には考えられなかった」と校長のアウンナインウィンさん(52)は言う。

「子供たち楽しそう」

以前使われていたのは軍政下の1998年に改訂された教科書で、多くがモノクロ印刷で文字ばかりの味気ないものだった。教師が教科書を読み上げ、児童らが繰り返して暗唱する教育は「オウム法」と呼ばれ、問題視されていた。

新教科書では教師と児童の対話を基本に据え、イラストや図で進んで学べるように工夫した。「慣れるのに大変だったが、子どもたちが楽しそうに勉強していて安心した」と教師歴33年のキンメイチャインさん(58)は話す。

新しい小学1年生の理科の教科書（右）は、水について飲み水や生活用水などイラスト付きで説明。旧教科書（左）は簡単な記述にとどまる

双方向の授業・イラスト並ぶ教科書…軍政下から一変

理科実験や郷土史も

改革を支えたのは日本だ。2011年、軍政から民政に移行したミャンマーが教育改革に取り組むにあたり、日本の支援として14年、小学校のカリキュラムを変える国際協力機構（JICA）の事業が始まった。

教師や教員養成校講師が教科ごとにチームをつくり、日本の教科書などを参考に原案を作成。それをたたき台にJICAの専門家と議論した。児童の自発性を生かす授業を目指し、従来なかった理科実験やクラスでの発表などを盛り込んだ。

理科を担当した教育コンサルタントの持仏賢一さんは「自分たちが受けた教育とあまりにも違って戸惑っていたが、次第に呼吸が合ってきた」と振り返る。

地理と歴史を統合し、「社会科」もつくった。軍政時代は、地方の自立への懸念から州などの地域について学ぶ機会はなかった。新しい社会科では「郷土の地理・歴史」を勉強する。教科書がなかった音楽や体育では、児童が取り組みやすいようイラストを多用したものを用意した。

算数平均点23%向上

教え方も改革の対象だ。全教師が受ける約2週間の研修では日本の専門家らが教師役、研修を受ける教師が児童役となり、児童の視線で学び方を考えた。さらに、教師用の指導書を用意。これも日本流だ。「指導書を見れば授業をスムーズにできるようにした」とJICAの岩沢久美子さんは説明する。

理科チームのメンバーでヤンゴンの教員養成学校講師、ソーソーウェさんは「子どもたちはもちろん、現場の先生たちも納得して使える教科書にしたかった」と話した。

JICAの調査では、新しい教科書を使った2年生は以前の同学年の児童よりも算数の平均点が約23%も向上。中でも地方の児童の学力が上がり、都市部との点差はこれまでの3分の1以下に縮まった。

(Source) Asahi Shimbun, 06 January, 2010 朝日新聞 2010年1月6日朝刊

(Source)　Asahi Newspaper, 06 January, 2010.

(Source) Glass, G.V (Ed.) (1976*). Evaluation Studies Review Annual, Vol. 1.* Beverly Hills: SAGE Publications

Afterword

It is great to see that impact evaluation is becoming more popular. After viewing the impact evaluation designs and examples in this report, you may be interested in learning more. I would like to introduce two databases of papers for this purpose.

The first is The Abdul Latif Jameel Poverty Action Lab (J-PAL), commonly known as the ``Poverty Action Lab," run by the Massachusetts Institute of Technology (MIT) in the United States. The lab declares that it will only use the most rigorous design, randomized controlled trials (RCTs) , for impact evaluations, and has conducted more than 1,000 impact evaluations and published papers to date. In this report, we have adopted some of the Poverty Action Lab's papers. I would like to express my sincere appreciation for it. You can search and download papers from the following sites. (https://www.povertyactionlab.org/evaluations)

The second is the International Initiative for Impact Evaluation , commonly known as 3ie . It covers impact evaluation reports using not only RCTs but also other impact evaluation designs, and you can search and download more than 4,000 papers in total. (https://developmentevidence.3ieimpact.org/)

It would be my great pleasure if this report could contribute to readers' understanding of impact evaluation and new approaches to impact evaluation. Lastly, I would like to note the following words.

Evaluation is Social Betterment. No evidence, no social betterment.

＊ ＊ ＊ ＊ ＊

Lastly, as I always do when publishing a book, I would like to include the music I was listening to while writing. If you're going to use books as a reference, you should also have the same respect for music.

Avicii. *Without You ft. Sandro Cavazza ; Dear Boy f t . MØ ; I Could Be The One ft. Nicky Romero.*

Avicii. *Shilloutes ft. Salem Al Fakir "And we will never look back at the faded silhouette."*

Marc Moulin. *Into the Dark. (Karma Fever Mix).*

Oliver Heldens & Shaun Frank. *Shade of Gray ft. Delaney Gene.*

Janet Jackson. *You Want This "Boy, you have to please me." & Someone To Call My Lover.*

And sometimes the music videos speak more than words.

Aloe Blacc - *Wake Me Up (Official)* https://www.youtube.com/watch?v=M_o6axAseak

ONE OK ROCK - *Stand Out Fit In (Official)* https://www.youtube.com/watch?v=IGInsosP0Ac

Clean Bandit - *Symphony (feat. Zara Larsson) (Official)* https://www.youtube.com/watch?v=aatr_2MstrI

Ryo SASAKI

Author biography

Ryo SASAKI, Ph.D.

Work history

Senior Researcher, Evaluation Department, International Development Center (IDCJ).
Development Consultant Intern, the U.N. headquarter (NYC).
Concurrent lecturer at Rikkyo University Graduate School of 21st Century Social Design.
Part-time lecturer at Osaka University Global Collaboration Center, Japan.
Part-time lecturer, Nagoya University Graduate School of Law, Japan.
Part-time lecturer, Faculty of Letters, University of the Sacred Heart, Japan.

Academic background

Ph.D. in Evaluation, the Evaluation Center, Western Michigan University (WMU).
M.P.A. in Public Policy Analysis, Robert. F. Wagner Graduate School of Public Service, New York University (NYU).

Author

In addition to the following books, there are many academic papers.

"Introductory Evaluation" (2014, CH Weiss (author), Sasaki (translation supervision)) , Miho Maekawa, Mitsuru Ikeda (translation supervision) (Nippon Hyoronsha)
"Success Case Method" (2022 , Brinkerhoff.O.R., Translated by Sasaki) (Taga Publishing)
"Collaborative Evaluation Step by Step" (2022, Liliana Rodríguez Campos et al. (author), Ryo Sasaki (translator) (Taga Publishing)
"Evaluation Logic: Fundamentals of Evaluation" (2010) (Taga Publishing)
"Theory and Techniques of Policy Evaluation" (2000, 2004) (co-author, Taga Publishing)
"Policy Evaluation Training Book: Seven Controversies and Seven Recommendations" (2003) (Taga Publishing)
"Policy Evaluation with Excel: A Practical Statistical Manual that is Easy to Understand" (2007) (Taga Publishing)
"Strategic Management of Universities" (2005) (Co-author: Taga Publishing)
"Theory and Techniques of Strategy Formulation" (2002) (co-author: Taga Publishing)

recent work

- Nepal "Monitoring and Evaluation System Strengthening Project Phase 2 (SMES2)" (2011-2015 , summary)
- Jordan "Impact Evaluation of Peace Creation for Syrian Refugees in Jordan" (2020-2022 , Summary/Impact Evaluation) (Report: English , Japanese) Posted on the JICA Impact Evaluation Site (https://www.jica.go.jp/activities/evaluation/impact.html)
- Palestine. "Science and Mathematics Education Quality Improvement Project (Full-scale Activity Implementation Phase)" (2021-2023, Impact Evaluation (Science and Mathematics Impact Evaluation)
- Myanmar. " Primary Education Curriculum Revision Project " (2 018-2021 , Impact Evaluation)
- Malawi (Africa). "Social Impact Measurement on Rural Water Supply" (2016 , Summary/Social Impact Evaluation)

- Afghanistan. Advisor service for monitoring and evaluation project of "Afghanistan Humanitarian Crisis Response Support Program" (2020)

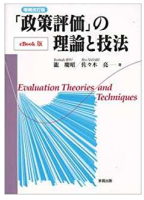
*1st to 5th are all JICA commissioned projects. The sixth is Japan Platform (JPF) contract work. In addition, he has extensive experience in commissioned research from international organizations such as the World Bank Group IFC, the Asian Development Bank (ADB) , and the United Nations Development Program (UNDP).

Related books and training information

<Academia journal>

	**Japan Evaluation Research Vol. 20 No. 2 (July 2020)** "Special feature: Current status and challenges of disseminating "evidence-based policy planning (EBPM)"" 7 papers (editors: Ryo Sasaki, Tomoya Masaki) http://evaluationjp.org/files/Vol20_No2.pdf
	**Japan Evaluation Research Vol. 17 No. 1 (November 2016)** "Special feature: Scientificity in evaluation: Practical use of evidence and its direction" 4 papers (editors: Ryo Sasaki, Tomoya Masaki) http://evaluationjp.org/files/Vol17_No1.pdf
	**Japan Evaluation Research Vol. 10 No. 1 (March 2010)** "Special feature: Global trends in evidence-based practice and efforts in Japan" 4 papers (editors: Ryo Sasaki, Iwao Oshima) http://evaluationjp.org/files/Vol10_No1.pdf
	**Japan Evaluation Research Volume 6 No. 1 (March 2006)** "Special feature: Attempts at evidence-based evaluation" 4 papers (Editing: Ryo Sasaki) http://evaluationjp.org/files/Vol06_No1.pdf

<Books and e- books>

 Yoshiaki Ryu, Ryo Sasaki	**"Theories and Techniques of Policy Evaluation" (eBook version)** (2000,2004,2020eBook version) https://www.amazon.co.jp/gp/product/B08DFN2L91/ref=as_li_tl ・The best-selling evaluation theory in Japan. ・Explanation of evaluation using a single set of theory evaluation, process evaluation, impact evaluation, and cost-effectiveness evaluation. -First explanation of the term and concept of "logic model" in Japan. ・Introducing the first application example of RCT in the social sector in Japan.
 Ryo Sasaki	**"Evaluation Logic: Fundamentals of Evaluation" (eBook version)** (2 008, 2020eBook version) https://www.amazon.co.jp/dp/B08D6TM1FY/ref=as_sl_pc_qf_sp_asin_til ・A work that forms the basis of evaluation science and first introduced the basic concept of ``evaluation = fact identification + value judgment'' to Japan. - Carefully explains the handling of values in evaluation, how to make value judgments, the concept of objectivity, the concept of internal validity, the concept of external validity, etc. ・Comprehensive explanation of major controversies in evaluation studies.
 Ryo Sasaki (translator)	**"Success Case Method"** (2022) Robert Brinkerhoff(author), Taga Publishing) There is an **e- book version** and **a print version .** Search for "success case method" on Amazon. (https://www.amazon.co.jp/) ・Japanese translation of the best-selling American corporate training evaluation. ・It is adopted by large international companies such as SONY, HONDA, and Amazon, international organizations such as the World Bank, and university organizations such as Stanford University. ・In ``From a Translator: Toward Application in Japanese Society'', ``What is the relationship between evidence-based practice?'' " clearly explained.
 Ryo Sasaki (translator)	**"Collaborative Evaluation Step-by-Step"** (2022). Liliana Rodriguez-Campos, et al., Taga Publishing) There is an **e- book version** and **a print version .** Search for "collaborative assessment" on Amazon. (https://www.amazon.co.jp/) -A practical book that explains in simple terms the latest trend in evaluation, ``collaborative evaluation.'' ・In ``Appendix: Toward Application in Japanese Society'', ``What is the relationship between evidence-based practice?'' " "What is the relationship with third-party evaluation?" " clearly explained.

Statistical Training

"Professional statistical analysis workshop"
Very popular as ``Anyone can definitely understand'' !!

Student comments 1 ``He was able to really explain things using just ``addition, subtraction, multiplication, and division.'' I was impressed.''
Participant Comment 2: ``I'm happy that I was able to read the papers, and I'm reading them all the time.''

[Lecturer]
Ryo Sasaki (Ph.D. in Evaluation, Western Michigan University)
PhD in Sociology, Stanford University)
other guest lecturers
[Sponsor] International Development Center (I DCJ) Evaluation Department
[Frequency] Held once every two months (sometimes held in Japanese, sometimes held in English)
[Summary explanation]
You will learn all about the statistical analysis methods explained in this report. We will explain calculations of mean values and standard deviations, testing of mean difference between two groups, regression analysis, concept of sample size, RCT procedures, how to read papers, and practical points based on real experience.
Prerequisites for participation are the ability to add, subtract, multiply, and divide, and to use Excel on a daily basis. The emphasis is on how it can be used in practice, rather than academic sophistication. Those who complete this workshop will receive a certificate of completion.
【website】
https://www.idcj.jp/seminar/statistical-analysis-workshop.html

"Professional Statistical Analysis Workshop: Applied Edition"
Very popular as ``Anyone can definitely understand'' !!

Similar to the above course, the process follows the steps of (1) explanation of example problems, (2) explanation of principles, (3) simple manual calculations, (4) implementation of practice problems, and (5) discussion and explanation of academic papers. Masu. We have 4 frames available. It is offered on-demand, so you can take the course anytime after completing the procedure.

【course】
Applied course 1: Latest techniques for impact evaluation (DID, PSM, IV)
Applied course 2: Structural equation modeling (SEM)
Applied course 3: Sample size calculation for impact evaluation
Applied course 4: Calculation of meta-analysis (systematic review)
[Frequency] On-demand implementation using Z oom recorded videos (you can take the course at any time)
【instructor】
Ryo Sasaki (Doctor of Evaluation (P hD), Western Michigan University)
other guest lecturers
[Sponsor] International Development Center (I DCJ) Evaluation Department
【website】
https://www.idcj.jp/pickup/grow/statistical-analysis-workshop-advanced.html

"Professional Statistical Analysis Workshop: Applied Edition: Data Analysis Exercises with STATA"
No text input required, just select and click from the menu bar

We will practice how to operate only by selecting and clicking from the menu bar and how to read the output results.
[Frequency] Held once every two months
【instructor】
Ryo Sasaki (Doctor of Evaluation (P hD), Western Michigan University)
[Sponsor] International Development Center (I DCJ) Evaluation Department
[Website] (The date and time will be posted on the Japanese and American websites)
(IDCJ website) https://www.idcj.jp/seminar
(U.S. Stata headquarters website) https://www.stata.com/meeting/short-courses/#japan

Youtube channel

Introducing YouTube channels that you can watch for free

IDCJ Evaluation Department "Professional Statistical Analysis Workshop" Tips Series

★Public relations video★ " Tips 1 Key points of statistical analysis" (7 minutes)
https://youtu.be/fu6taNR_jsA

★Public relations video★ " Tips2 5 designs for impact evaluation " (14 minutes)
https://youtu.be/neLly_7hv00

★Publicity video★ " Tips 3 How to read papers 1: Impact evaluation papers" (12 minutes)
https://www.youtube.com/watch?v=OMw7D0vkSIA

★Publicity video★ " Tips 4 How to read papers 2: Regression analysis papers" (12 minutes)
https://www.youtube.com/watch?v=1gychUeH6zY

★Public relations video★ 『 Tips 5 What is the minimum sample size? (10 minutes)
https://www.youtube.com/watch?v=y3IvlvmHvvU

★Publicity video★ "Tips 6 Essay Example 1 El Salvadoran Education" (14 minutes)
https://www.youtube.com/watch?v=jZxrC0fMYCg