インパクト評価事例集

実験デザイン(RCT)から専門家判断まで 18 の評価事例

～よりよい「社会的インパクト評価」の実施のために～

Collection of Impact Evaluation Practices:
18 practices from RCT to Expert Judgement

最近の更新

- 2019/11　祝!! 2019 年度ノーベル経済学賞を MIT「貧困アクションラボ」(J-PAL)の 3 名が受賞。
- 2020/05　フィンランド政府が実施したベーシックインカムのインパクト評価 (2020) を追加。
- 2020/05　「社会的インパクト評価」の解説を追加した。
- 2020/09　コロナ禍における SMS／電話による教育効果のインパクト評価(2020)を追加した。
- 2021/10　エビデンス・ピラミッドの解説を掲載。
- 2022/05　RCT の各事例の「議論」のセクションを加筆した。また、「科学的評価 対 実践的評価」の議論、および「定性的評価 対 定量的評価」の議論の解説を加筆しました。
- 2022/10　統計分析の勘所と学術論文の読み方をまとめたユーチューブの公開を開始しました。
- 2022/12　「インパクト」の 4 種類の使い方」に社会的インパクトの説明を追加しました。
- 2023/04　「ヨルダンにおけるシリア難民への平和の創出に係るインパクト評価」のリンクを掲載。
- 2023/10　教育分野の先行研究に基づく新しい効果サイズ(Effect Size)の基準の提案した。
- 2023/10　効果率(%)(Effect Percent (%))の正式な提案を記載した。
- 2024/08　紹介されている論文の最新の関連論文の情報を追加した。
- 2024/08　回帰分析のためのジェンダー・コード・マトリックスの提案を追加した。

Version 8.71（最終更新日：2024 年 12 月 24 日）

評価学博士

佐々木 亮

Ryo SASAKI, Ph.D.

はしがき

このレポートは、いわゆるインパクト評価のデザインと適用事例を解説したレポートです。筆者がインパクト評価に出会ったのは、1994 年のニューヨーク大学修士課程の在学時で、同課程に、Program Analysis and Evaluation という科目があって履修しました。日本の学部在籍時には、自然科学では実験ができるが社会科学では実験ができない。かわりに、比較や過程分析を使うのだと教えられていました。しかし本当に社会科学では実験ができないのだろうかと思っていたのですが、その科目で使用したテキストには、普通に Social Experimentation（社会実験）の実例が解説されていました。これは目からウロコでした。それが私とインパクト評価の出会いでした。

インパクト評価に魅了された私は、その後、約 30 年にわたってインパクト評価の事例を集め続け、このレポートを公表することになりました。また、インパクト評価のデザインをもっとも単純なデザインからもっとも厳格なデザインまで網羅した解説も追加しました。

昨今の「エビデンスに基づく実践」（EBP：Evidence-Based Practice）や、「エビデンスに基づく政策立案」（EBPM：Evidence-Based Policy Making）の盛り上がりは、私にとってうれしい限りです。エビデンスとは、インパクト評価の結果のことを指します。権力者や有力者や重鎮の意見によって意思決定が左右される社会ではなく、エビデンスによって意思決定がなされる社会になってほしいと思っています。

そしてこれは、たとえ社会的に弱い立場におかれた人々であっても、エビデンスを示せれば、意思決定を主導できることを示しています。つまり、年齢、性別、人種、民族、出身階級、出身国、出身地、障害の有無などの違いから人々を自由にする力を秘めています。

インパクト評価が、人々を自由にして、社会の改善に貢献することを心から願っています。

佐々木亮／Ryo SASAKI, Ph.D.

本書の読み方

本書は以下の構成になっている。最初から読み進めることを想定しているが、必要な部分から読んでも差し支えないように配慮されている。

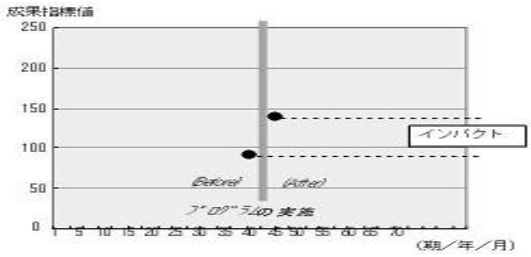
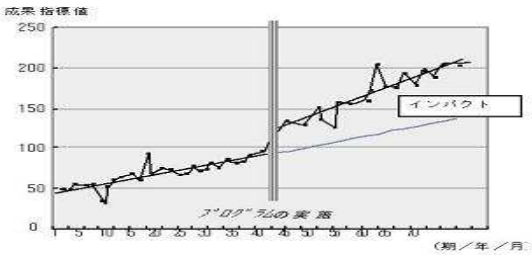
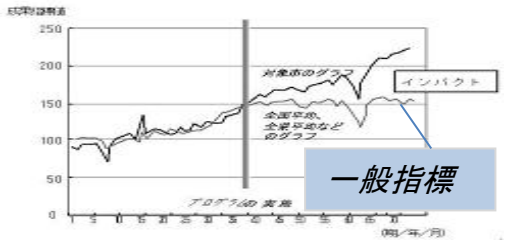
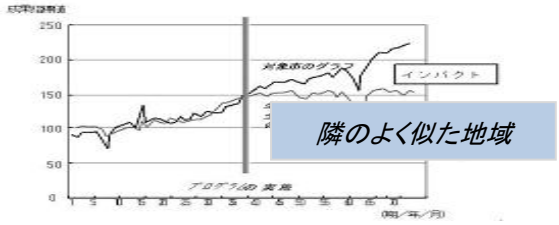
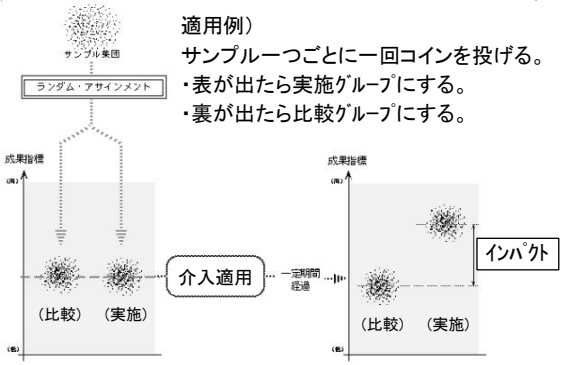
I.基本的な概念の解説（インパクト評価の5つの基本デザイン、インパクトの定義など）
II.5つの基本デザインの考え方と実例の解説
1. 事前・事後比較デザイン （Before-After）
2. 時系列デザイン （Interrupted Time-Series）
3. 一般指標デザイン （Generic Control）
4. マッチングデザイン （Matched control）
5. 実験デザイン（RCT）（Randomized controlled trial）
III.インパクト評価に関連する学術的な議論の紹介
IV.最終章：インパクト評価の起源・現在地点・そしてこれから（有料版のみ）

それぞれの事例や学術的な議論には出所である参考文献リストをすべて掲載している。それらの参考文献はすべて英語であるが、興味のある読者は直接入手して参照することが勧められる。

目　次

Ⅰ. 基本的な概念の解説

はじめに１：インパクト評価の基本デザイン

名称と概念図	説明	
1. 事前・事後比較デザイン (Before-After) 	シンプルに、事前、事後の指標値を比較し、差があれば因果関係があったと推定する。簡便なので広く用いられている。ただし、事前・事後の間に発生した外部要因による影響値をまったく取り除けないので、因果関係の推定の信頼性は低い。 [検定テスト:対応のある(一対の)t 検定]	単純 & 安価
2. 時系列デザイン (Interrupted Time-Series) 	施策介入前の長期的トレンドを導き出し、施策介入後にトレンドが変わっていれば、因果関係の存在を推定する。ただし、長期的トレンド以外の外部要因による影響値を取り除けないので、信頼性はそれほど高くない。 [検定テスト: 回帰分析]	
3. 一般指標デザイン (Generic Control) 	全国平均値、全県平均値などの一般指標値を比較に用いる。外部要因による影響値をある程度除去して考えることができるので(なぜなら対象地域が受けた影響とある程度同じ影響を一般指標値も受けているはずだから)、因果関係の存在の特定に関してある程度の信頼性を確保できる。わりと簡単に用いることができる。 [検定テスト: 目視による判断]	
4. マッチングデザイン (Matched control) 	可能な限り近似のグループを選定して比較に用いる。外部要因による影響は、どちらのグループも(完全に同一ではないが)同程度に受けると考えられるので、因果関係の存在の特定のために高い信頼性を確保できる。 [検定テスト: 独立の t 検定]	
5. ランダム化比較デザイン (Randomized controlled trial (RCT)) 	施策の実施前に、政策適用を無作為割付(ランダム・アサインメント)により、実施グループと比較グループに分ける。成果指標値に現れた違いは、途中の唯一の違いである「介入を適用されたか否か」によって引き起こされたと純粋に判断することができる。なお、外部要因による影響は全く同一になっているので考える必要はない。因果関係の存在の特定に関してたいへん高い信頼性を誇り、これ以上のデザインは存在しない。ただし実際の適用は難しい。 [検定テスト: 独立の t 検定]	厳格 but 高価

(出所)Rossi, Freeman, Lipsay (1999) *Evaluation A Systematic Approach,6th Edition*,Sage Publication,Sage Publication.p261 の表のから筆者が代表的なデザインを選択して掲載した。

はじめに 2：「インパクト」の 4 種類の使い方

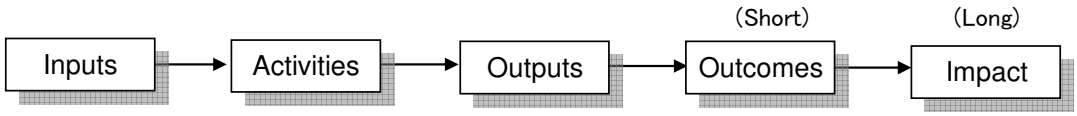
「インパクト」には以下の 4 種類の使い方が観察される。このテキストでは、主流の使い方であるタイプ III の使い方（「介入行為による純粋な変化量」）に従う。

●ODA 評価で頻繁に用いられる DAC 評価 5 項目*のひとつの「インパクト」は、以下のうちタイプ I とタイプ II の双方をカバーする。（*妥当性、有効性、効率性、インパクト、持続性）

●4 番目の「社会的インパクト評価」で言う社会的インパクト（あるいは単にインパクトと呼ぶこともある）が 2010 年代後半から急速に普及してきた。

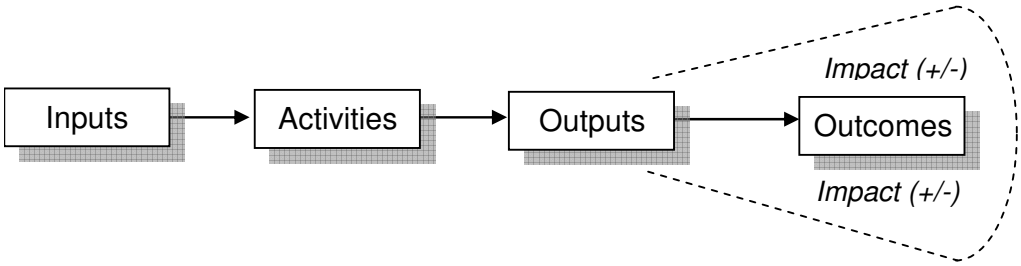
（1）Type I impact : Long-term social/economic impact

（タイプ I インパクト：長期的な社会経済的変化）



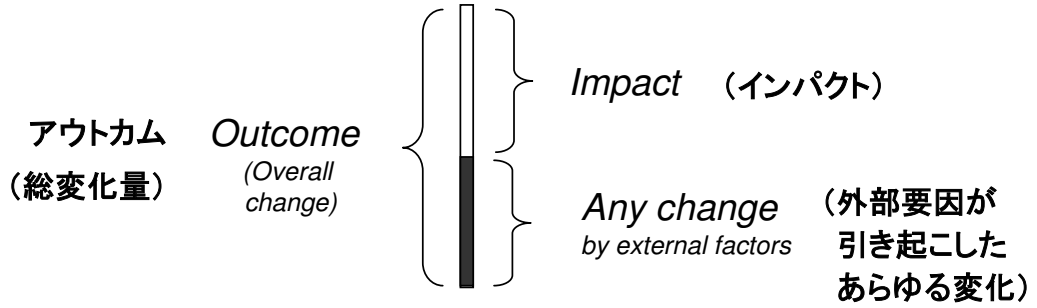
（2）Type II impact : Indirect impact　(either positive and negative)

（タイプ II インパクト：間接的な効果（正負の両方があり得る））



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

（タイプ III インパクト：介入行為による純粋な変化量）



（4）　Type IV impact: Social Impact Measurement

（タイプ IV インパクト: いわゆる「社会的インパクト評価」で言うインパクト）

通常、「当該事業や活動によって生じた短期、長期の変化を含めた社会的、環境的なアウトカム」と定義される。これは上記の（1）～（3）を含む定義となっている。

（出所）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. さらに、TypIV は 2022 年 12 月に佐々木が追加しました。

はじめに３：社会的インパクト評価（Social Impact Measurement）の解説

１．定義

標準的な社会的インパクト評価(Social Impact Measurement)は以下の手続きは以下のとおりです。前の図で解説したType I ImpactとType II Impactを混合した定義になっています。また、Type III Impactの意味を含んで使われる場合もあるようです。

> 自身の活動のおかげで生じる社会経済的な変化を明らかにすること。なお、社会経済的変化には、短期・中期・長期、および直接的・間接的、および意図するものと意図せざるものを含む。

２．具体的な手続き

社会的インパクト評価は以下の手続きで実施します。

社会的インパクト評価の手続き

Step 1：事業の目的を議論して合意する。

Step 2：ロジックモデルを書く。

Step 3：投入・活動・アウトプット・アウトカムを測定する指標を決める。

Step 4：指標値を収集する方法を決める。

Step 5：指標値を収集する。

Step 6：収集された指標値を分析する。

Step 7：分析結果にもとづいて、評価を下す＝＞評価結果を書く。

Step 8：必要ならば、提言を書く。

Step 9：報告書にとりまめる

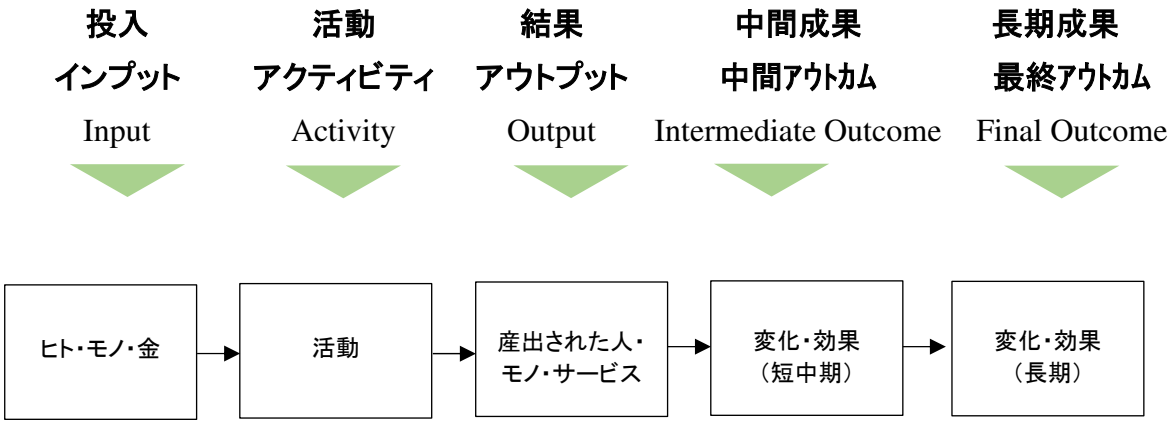
（出所）筆者作成。龍・佐々木(2010,2014)『政策評価の理論と技法』多賀出版を参考にした。

Step 1：事業の目的を議論して合意する

関係者で集まって活動の目的を合意します。意外と**「じつは本当の目的は違うんだよね」**という意見が出たり、**「この面でも役に立っている」**といった話がでますが、関係者で率直に話し合って明示的にメインの目的を合意します。受益者のニーズの実現がメインの目標になります。文書に明記しておくといいでしょう。

Step 2：ロジックモデルを書く

以下のような5段階のロジックモデルを作成します。自分の組織がやっている「活動」から書き始めるとスムーズに書けます。



職業訓練の実施に必要な - 講師 - 教材 - 会場 - 運営資金	・職業訓練の実施（全3か月間）	・職業訓練を修了した人が100人	・そのうち就職した人が30人	・そのうち5年後に起業した人が3人

Step 3：投入・活動・アウトプット・アウトカムを測定する指標を決める

それぞれを測定する指標を決めます。指標はひとつでも複数でも構いません。

- 講師の数 - 教材の数量 - 会場の有無 - 金額	・職業訓練の実施の有無と成功度合（5段階評価）	・修了者数	・就職者数	・起業数

Step 6：収集された指標値を分析する

このテキストで解説するインパクト評価の5つのデザインのどれかを適用してインパクトを特定します。現行の実践では、単純な事前・事後比較が用いられていますが単純すぎます。より上位の厳格なデザインを使いましょう。

Step 7：分析結果にもとづいて、評価を下す＝＞評価結果を書く

「職業訓練を実施して、たいへん良かった/良かった/良くなかった」、あるいは

「職業訓練は、たいへん満足できる/満足できる/不満足である」と結論を書きます。指標値の変化はあくまで事実を特定しただけですので、その変化量を根拠に**「良かった／悪かった」**あるいは**「満足できる／不満足である」**と価値を表現する言葉で結論を書いてください。価値を表現する言葉で結論を書いて初めて、「評価」足りえます。

Step 8：必要ならば、提言を書く

提言は、評価の一部ではないのですが、書くことが常に求められます。(1)政策的な提言：「職業訓練は、来期も実施すべきである/来季はやめるべきである」、(2)事業改善の提言：「募集の仕方を改善すべきである」などの2レベルの提言を書きます。

Step 9：報告書にとりまとめる

Step1からStep8までの情報を報告書にとりまとめます。

3．解説：正式な評価学の観点から

いわゆるインパクト評価と社会的インパクト評価のちがいは以下のとおりです。

「社会的インパクト評価」(Social Impact Measurement)は、評価研究の世界でパフォーマンス・メジャーメント(Performance Measurement)と呼ばれてきたものと同一です。そこで使われている分析のデザインは、事前・事後比較デザイン(Before-after comparison)であり、正式な「インパクト評価」(Impact Evaluation)の理論から見るとたいへん初歩的なデザインです。

社会的インパクト評価を実践する人は、本報告書で解説されるより厳格なデザインを適用することが勧められます。社会的インパクト評価から入って、徐々に本格的なインパクト評価に取り組んでみることを期待しています。

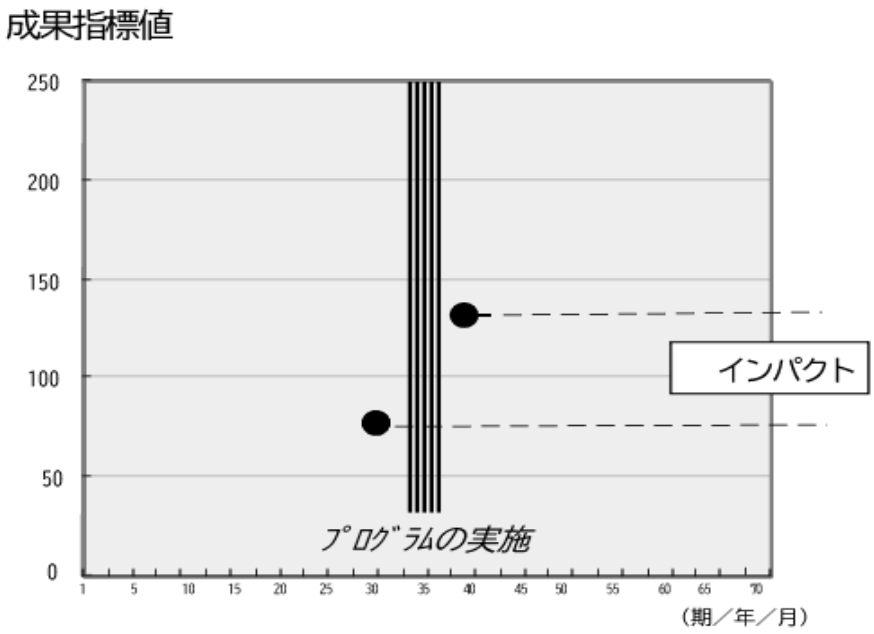
(出所)龍・佐々木 (2010,2014)『政策評価の理論と技法』多賀出版を参考にした。

II. ５つの基本デザインの考え方と実例の解説

1. 事前・事後比較デザイン (Before-After)
2. 時系列デザイン　(Interrupted Time-Series)
3. 一般指標デザイン　(Generic Control)
4. マッチングデザイン (Matched control)
5. 実験デザイン (RCT) (Randomized controlled trial)

1．事前・事後比較デザイン

(Before-After Design)



［説明］

シンプルに、事前、事後の指標値を比較し、差があれば因果関係があったと推定する。簡便なので広く用いられている。ただし、事前・事後の間に発生した外部要因による影響値をまったく取り除けないので、因果関係の推定の信頼性は低い。

［検定テスト］

事前-事後の有意差検定（対応のある t 検定）

Dependent t-test（Paired t-test）

事前・事後比較デザインの適用事例１
初等教育支援プログラム（ガーナ）

世銀がガーナで実施した本件プロジェクトでは、「政策・マネジメントの改善」「物理インフラの改善」に係る支援を行うことにおり、「学校効率性の向上」「教員の教授環境の改善」「関連施設・教材の改善」を実現し、最終アウトカムとして「改善した入学実績と卒業実績」「改善した学力」を実現するとしている。以下の表は本件プロジェクトの事前と事後のテスト平均点を示している。

テスト点数の平均点

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.

“事前” “事後”

（出所）世銀（2004），p. 137

上の表に関して世銀の報告書は次のように結論を記載している。『Table G.2 は、…1988 年と 2003 年のテスト点数の平均点を示している。…表の最後の行は、二つのテスト平均点の間の差に関する t 検定量と p 値を示している。それらは、全ての科目に関して有意な改善を示している。』（The data show a significant improvement in all test scores.）ただしこの分析は単純な事前事後比較であり、当該期間にあったはずの外部要因による影響値や関係する他の介入行為の効果をも含んでしまっているはずだが、その制約に関する記載がないのは残念である。また、「全ての科目に関して有意な改善を示している」というテクニカルな（あるは学術論文で通常用いられる）記載で終わっており、それをもって「初等教育支援プログラムの効果があった」とは断定していない。事前と事後の間が 15 年の長期間となっており外部要因による影響を否定できないことから、世銀の介入だけで何かしらの効果を述べることは危険だと世銀も認識していると推察される。

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

事前・事後比較デザインの適用事例２
井戸改修事業の効果の評価（スーダン）

＜実施前＞業者が川の水を汲んで袋に詰め、ロバで運んで当該地区で売っていた。

＜実施後＞子供たちが井戸に水を汲みに来るようになった。

日本の協力によりスーダンのカッサラ州で「ベーシック・ヒューマン・ニーズに係るサービス提供プロジェクト」（通称：K-TOP）が実施された。その一部として Wad El Helew（地区名）（住民総数：2,000-3,000 ということだが正確には分からない）において、井戸改修工事が実施された（2012 年３月）。そして実施後に、井戸改修工事の「実施前」と「実施後」の状況に関して、サンプルである住民にアンケート調査を実施した。

アンケート用紙は以下のとおりで、シンプルに５つの質問を聞いている。

Questionnaire

Do you use water from Sitit River or the rehabilitated wells?
- Before March 2012: (　　), After March 2012: (　　)

How many hours do your family members spend to fetch water?
- Before March: (　　) hours, After March: (　　) hours

How much do you spend to get water?
- Before March: (　　) SDG a month, After March: (　　) SDG a month

How often do your children go to school?
- Before March: (　　) days a month, After March: (　　) days

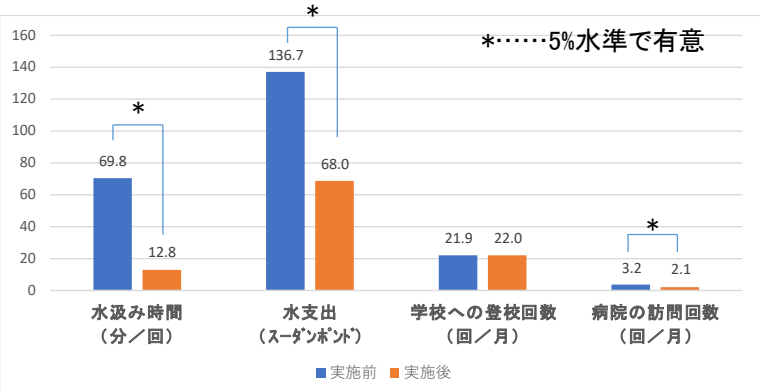
How often does your family go to hospitals?
- Before March: (　　) times a month, After March: (　　) times

水汲みに要する時間は何分？（事前、事後）
水を得るのにいくら支払っている？（事前、事後）
子供は月何回学校に行く？（事前、事後）
病院に月何回行く？（事前、事後）

結果は以下の通りであった。

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



事前・事後の比較により、「水汲みに要する時間」（分/回）、「水に関する支出」（ｽｰﾀﾞﾝﾎﾟﾝﾄﾞ）、「病院の訪問回数」（回／月）がいずれも減少しており、その減少幅は統計学的に有意であった（p＜5%水準）。

一方、「子供の学校への登校回数」（回/月）は、事前、事後とも約 22 回で統計学的に有意とは言えなかった。

総じて、井戸改修事業は、地区の住民の生活を多方面で改善したと判断できる。

（出所）IDCJ（黒田康之・主任研究員の提供資料（「Impact Survey: Has the well rehabilitation project improved the quality of life for residents?」(2014)による）

事前・事後比較デザインの適用事例３

小学校リハビリテーション支援事業（ジブティ）

アメリカ国際開発庁（USAID）が行った小学校リハビリテーション支援事業の簡便な評価では事前の写真と事後の写真が使われている。視覚に訴える効果があるが、恣意的になりやすいという批判は逃れられない。

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.

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

事前・事後比較デザインの適用評価４
厚生サービス強化事業（ペルー）

問題の所在と評価結果

1980 年、1990 年の経済的な苦境の影響で、ペルーの保健セクターは十分なサービスを提供できなくなっていた。この状況を改善するため、ペルー政府は、「保健サービス強化プログラム」を開始した。

１．施策の概要

このプログラムは、次の３つの部分から成り立っていた。（１）事前の研究や調査、（２）保健省の組織強化と分権化の強化、（３）保健医療施設の強化。このうち日本は（３）を支援するため融資を行なった。融資は、総額約 22 億円、金利 3.0%、返済期間 30 年という好条件で、1994 年４月に調印して、1999 年７月まで何度かに分けて実行された。

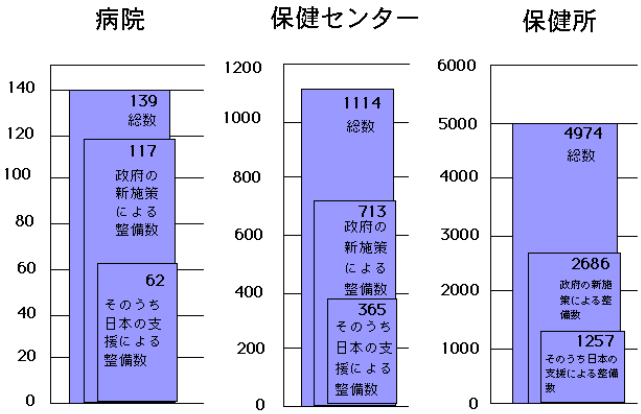
このプログラムによって以下の表のように資機材整備が実現した。また、日本の融資の貢献も表中に表わしたとおりであった。

「保健サービス強化施策」により資機材が整備された病院、保健センター、保健所の数

		総数(a)	政府による「保健サービス強化施策」による整備数(b)	(b)/(a)	そのうち日本の支援による整備数(c)	(c)/(a)
病院	Hospitals	139	117	(84%)	62	(45%)
保健センター	Health Centers	1,114	713	(64%)	365	(33%)
保健所	Health Posts	4,974	2,686	(54%)	1,257	(25%)
合計	Total	6,227	3,516	(56%)	1,684	(27%)

Source)MINSA（ペルー保険省）

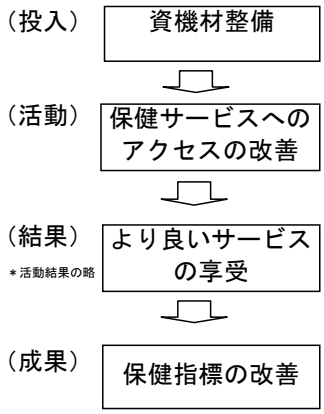
例えば、病院総数 139 軒のうち、本プログラムによって整備されたのは 117 軒で、さらにそのうち日本の融資によって整備されたのは半数近くにあたる 62 軒であった。病院よりも規模が小さいが保健所よりは大きい「保健センター」について見ると、 総数 1,114 軒のうち、本プログラムによって整備されたのは 713 軒で、そのうち日本の融資によって整備されたのは 365 軒で総数の 33%を占めた。最後に保健所について見ると、総数 4,974 軒のうち、本プログラムによって整備されたのが 2,686 軒で、そのうち日本の融資を使って整備されたのは、1,257 軒で総数の 25%であった。表をグラフで表わすと次ページにようになるので、確認していただきたい。



2．評価結果

この施策の効果を評価するにため、本件評価の実施者は、事前・事後比較デザインを用いた。また、日本が融資を実施したのは 1994 年から 1999 年であるが、実際に融資を使って資機材整備が行われるには若干の時間が必要であろうから、1994 年を「事前段階」、2000 年を「事後段階」として、日本の融資のインパクトを評価する。

なお、本プログラムの設計者及び評価者が想定した、本プログラムのインパクト発揮までの因果関係は下の図のとおりである。



施設が整備される→施設の利用者が増加する→国民の健康状態が改善する。そしてそれぞれを評価する指標として、評価者は、施設数、施設の利用者数、各種保健指標を採用した。

そして、まず施設の整備については、１．で解説してそれが実現したことが確かめられたので、次に、施設の利用者数について検証する。以下が施設の利用状況に関するデータである。

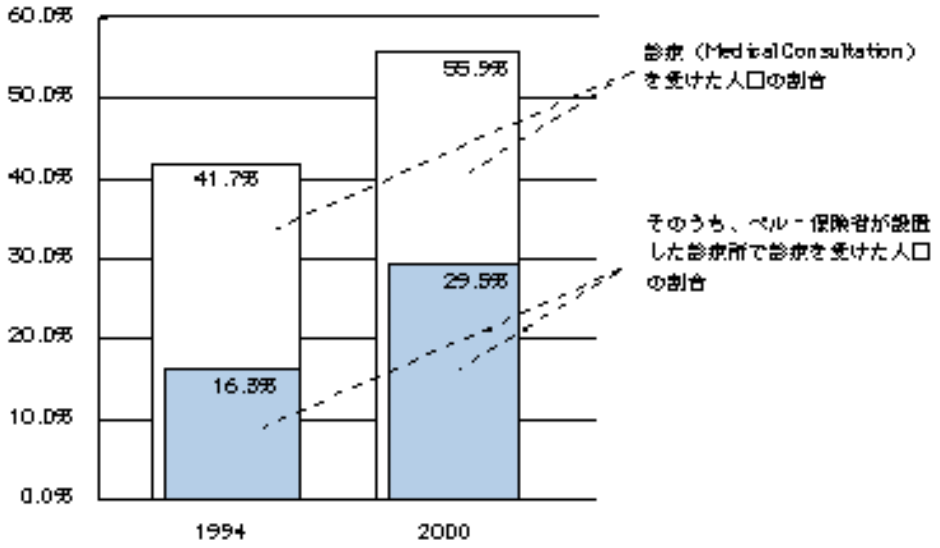
保健サービスの利用度合の比較(1994, 2000)

	1994	2000
（１）診療（Medical Consultation）を受けた人口の割合	41.7%	55.9%
（２）ペルー保険省が設置した診療所で診療を受けた人口の割合	16.3%	29.5%
（参考）ペルー保険省が設置した診療所を利用した比率（=(2)/(1)）	39.0%	52.8%

Source) ENNIV〔The National Standard of Living Survey〕

何らかの診療を受けたペルーの人口の割合を見ると、1994 年には 41.7%であったが、2000 年には 55.9%へと約 14.2%上昇している。一方、ペルー保健省が本プログラムによって新たに整備した診療所で診療を受けたと答えた人口は 16.3%から 29.5%へ 13.2%上昇していることから、ペルーの人口全体に見られた診療受診の改善のほとんどは、本プログラムによって

実現したと言えるわけである。このことをグラフ化すると以下のようになり、全体の押し上げは、ほとんど本プログラムによる押し上げによって実現していることがわかる。



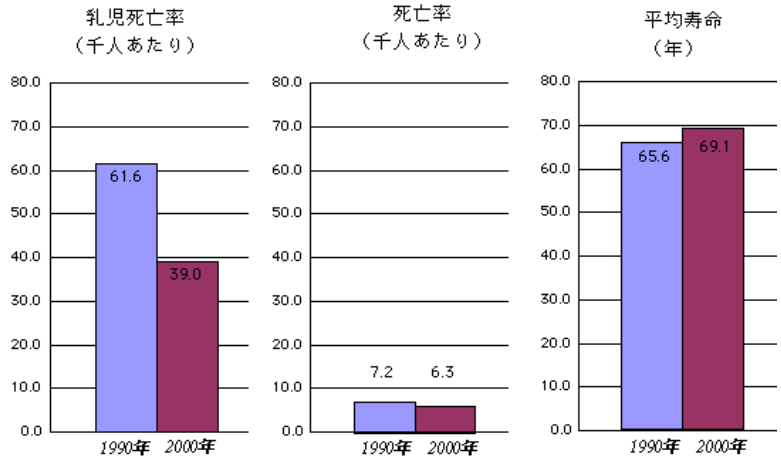
最後に国民の健康状態に及ぼしたインパクトについて、評価者は次のデータを提示している。本来は 1994 年と 2000 年について収集すべきであろうが、実際に収集できたデータは 1990 年と 2000 年のデータであったのでそれを提示している。

保健関連指標の比較(1990, 2000)

指標	Indicators	1990	2000
人口増加率(%)	Annaul Population Growth(%)	1.9	1.7
出生率(千人あたり)	Crude birtth rate (per 1,000 pop.)	29.0	23.7
乳児死亡率（千人あたり）	Infant Mortality (per 1,000 pop.)	61.6	39.0
平均寿命（年）	Life expectancy at birth (year)	65.6	69.1
死亡率（千人あたり）	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.

例えば、乳児死亡率については、千人あたり 29 人(1990 年)だっが、23.7 人（2000 年）に低下している。乳児死亡率は、千人あたり 61.6 人（1990 年）だったが、39.0 人（2000 年）に低下している。また、一般的な死亡率は、千人あたり 7.2 人から 6.3 人に低下している一方で、平均寿命は、65.6 才から 69.1 才へ上昇している。再びグラフで示すと次のようになる。



これらの指標の改善について、評価者は次のように結論している。

他の援助国が支援していくつかのプログラムが並列的に行われていた状況から考えて、日本の融資がペルーの保健分野全体にどれだけの直接的なインパクトを与えたかを特定することは難しい。しかし、資機材整備→保健サービスへのアクセスの改善→より良いサービスの享受→保健指標の改善というあり得べき因果関係に注意を向けることは重要である。そして 1990 年から 2000 年の間に、乳児死亡率、一般的死亡率、その他の指標が改善している。1990 年代のペルーの保健セクターにおいて、日本がおこなった融資がもっとも大口であったことから、日本の融資が保健指標の改善に貢献したと見込むのが安全かも知れない。

３．利点、制約、日本での適用に関する留意点

この方法の利点としては、実施地域だけのデータを参照すればいいという点があげられる。マッチングデザインなどでは、事前段階と事後段階における実施地域と比較地域のデータ（２時点×２地域）が必要だった。事前・事後比較デザインでは、事前と事後の実施地域のデータ（２時点×１地域）である。なお、統計的等化デザインは、事後段階における実施地域と比較地域のデータ（１時点×２地域）が必要だが、実務面から言うと事前・事後で使用する同一地域の事前と事後のデータの方がはるかに入手しやすいのだ。

この方法の制約として指摘されるべきは、これで何かの因果関係を証明しているとは言えないということである。事前と事後で指標値が改善したとしても、それは自分が実施した施策によるとは言い切れない。言い換えれば、この方法は、事前に想定された因果関係（ロジックデザイン）が正しいはずだという一点のみに依って立っていると言える。

日本で適用する際の留意点としては次があげられる。日本ではそもそもロジックモデルの類が検討され明確化されることは少ないのが現状である。例えば道路建設は、所用時間の短縮が目的か、あるいは建設による雇用創出が目的か。ODA は世界の貧困軽減が目的か、あるいは日本の企業進出の基盤を整備するのが目的か。両方なら両方で構わないが、まずロジッ

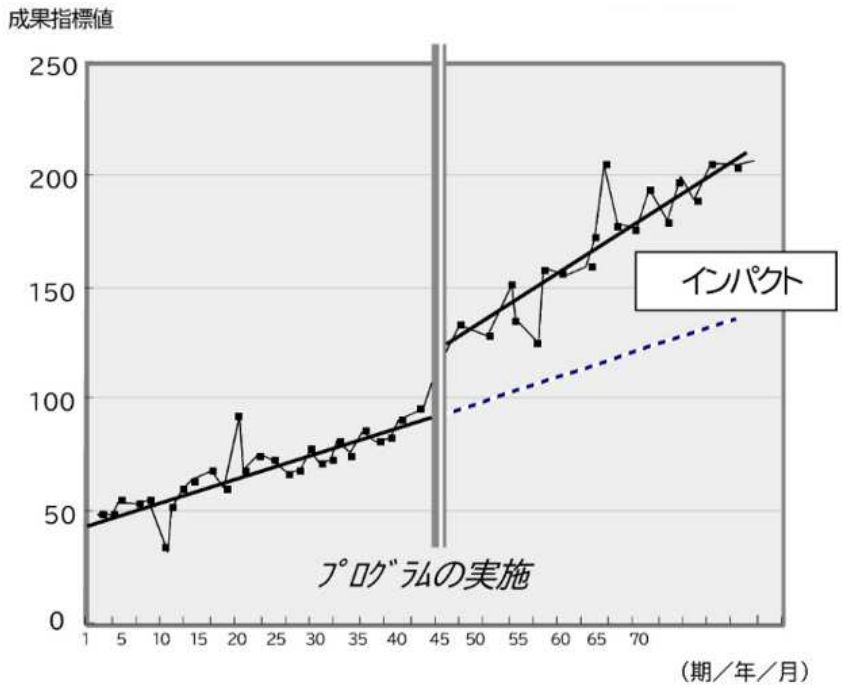
クモデルの作成を通じて関係者間で施策の目的や因果関係について合意するのが、効果を評価する上での大前提である。なお、目的が複数であればロジックモデルも途中から枝分かれし、それに伴って収集すべき指標も複数になる。

（出所）すでに公開されている国際協力銀行(2002)「円借款事後評価報告書 2002」(英文)の記述をもとに筆者が再構成して説明文を作成した。なお本評価の PDF ファイルは以下からダウンロードできる。

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

２．時系列デザイン

(Interrupted Time-Series Design)



［説明］

施策介入前の長期的トレンドを導き出し、施策介入後にトレンドが変わっていれば、因果関係の存在を推定する。ただし、長期的トレンド以外の外部要因による影響値を取り除けないので、信頼性はそれほど高くない。

[検定テスト]

回帰分析

Regression Analysis

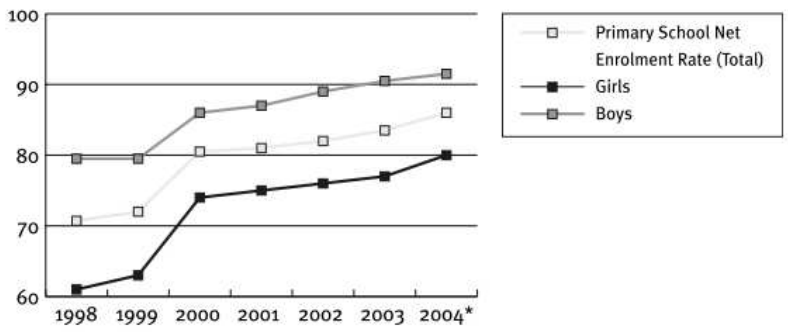
時系列デザインによる評価事例：
初等教育支援事業（ネパール）

問題の所在と施策の概要

ネパールでは、1999 年に「基礎・初等教育プログラム II」が開始された。その目的は、(i) 初等教育の質を改善すること、(ii) 初等教育へのアクセスを増加させること、そして(iii)関係機関の能力向上であった。具体的な内容は、校舎建設、教員養成、カリキュラム改善、教科書配布、関係機関の職員研修など多義にわたっていた。

図 4-1 は、1998 年から 2004 年までの純入学率(Net Enrollment Rate (NER))を示している。なお、同プログラムの開始は 1999 年下半期である。

Figure 4-1: Grades 1- 5, Net Enrolment Ratios 1998-2004

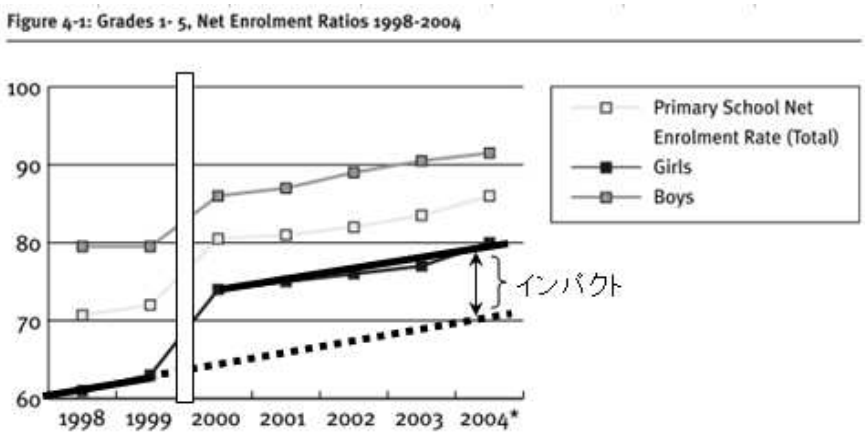


Source: School Level Educational Statistics of Nepal, HMG/N, DOE
* Estimate

評価結果

評価結果は、「総合、男子、女子の３つの指標に関して、1998 年よりも 2004 年の数値が高いので、入学率とカバー率を向上させるパターンが確認された」としている。実際の評価報告書ではその記載しかないが、このグラフに時系列デザインを適用すると次のようにインパクトを推定することができる。女子の場合のインパクトはおよそ 9%であったと推定できる。

Figure 4-1: Grades 1- 5, Net Enrolment Ratios 1998-2004



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

はじめに：この事例は純粋な時系列データの事例ではありません。時系列（20年）ｘ複数地域（全米50州）のデータを使っています（このタイプのデータをパネルデータと言います）。ただし、回帰分析の手続きは、単純な時系列データの事例と全く同一です。

パネルデータの回帰分析による評価事例：
飲酒運転に関する政策変更の効果（アメリカ）

問題の所在と評価結果

飲酒運転に関する規制を強化するとどれくらい交通事故は減るのだろうか？ アメリカでは 2000 年に、連邦交通適正化法案（the 2000 Federal Transportation Appropriations Bill）を巡って大きな議論が起きた。規制強化推進派は、「飲酒運転を撲滅する母親の会」（Mothers Against Drunk Driving）とアメリカ医療学会(American Medical Association)が、民主党のフランク・ローテンバーグ上院議員等を支援して、規制の強化を目指した。対する規制強化反対派は、アメリカ飲料水機構（American Beverage Institute）などのレストランやバーの業界団体で、共和党のウィップ・トム・デレー下院議員等を支援して、規制強化の阻止に回った。

評価の結果、検挙を行う血中アルコール濃度の基準を強化する施策は、交通事故死亡を減らす効果があると結論された。ただし他の政策も交通事故を減少させる効果を発揮していたことが分かった。つまり、他の施策で同程度の効果を上げられる場合があるので、社会全体の利益（厚生）を最大化させるにはどの施策を採用するのがいいのかを今後検討するべきだろうと結んでいる。

１．施策の概要

通常、飲酒運転の規制は、吐く息から、血中アルコール濃度を推定することによって実施される。従来は、検挙の基準を血中アルコール濃度 0.1%以下を基準としていたが、それを 0.08%以下に強化しようというのが法案の内容であった。

2000 年 10 月にクリントン大統領（当時）が署名した妥協案は次のとおりであった。全ての州に 0.8%への規制強化を強制することはしないが、この規制強化を受け入れない州には、高速道路整備のための交付金（Federal highway fund）を 2003 年度は2%減額する。2004 年度は4%、2006 年度は 8%、2008 年度は 10%、それ以降この比率で減らしていく。一方で、もし 2007 年度までに規制強化を受け入れた場合には、それまで減額された交付金全額を一括交付する。この決定以来、かなりの州で 0.08%への規制強化を受け入れたが、2002 年9月現在、15 州では未だに受け入れていない。このばらつきを用いて施策の効果を評価する。

２．事業内容（介入行為）

アメリカでは、1980 年代初頭から、死亡に到る交通事故（以下、「交通事故死亡」と標記）が継続的に減っている。1982 年に成人 1,000 人あたり 2.21 件だった交通事故死亡件数は、2000 年には 1.72 件ま

で減っている。 この傾向が続いている中で、単純な事前と事後の数値の比較を行なえば、今回の規制強化法が実際には効果がなくても効果があったと評価されてしまう。また、なぜ減少傾向が続いているのかと言うと、これ以外に実施されている様々な施策の効果や活動が表れているのではないかと予想された。

それらは、(1)若年層を対象とした段階的な免許付与制度(Graduate Licensing Program)の導入、(2)アルコール販売店規制法(Dram Shop Law)、(3)シートベルト義務付け法(Seat Belt 法)などである。さらに、ビール税の増税による小売価格の上昇も影響しているかも知れない。その他に、血中アルコール濃度の測定機器の精度が上がったことも影響しているかも知れない。未成年者(21 歳以下)は、0.1%か0.08%かを議論する前に、わずかでもアルコール反応が出れば即逮捕であり、年々言い逃れは難しくなっている。また、長く続く景気拡大による雇用状態の改善なども影響しているかも知れない。ただし各州の経済状態の違いによってその影響の程度は違うかも知れない。また、規制強化を受け入れてから今回評価を実施するまでの各州の年数の違いも効果の程度に影響を与えているかも知れない。

これらの様々な要因による影響にも配慮して本件規制強化の効果を評価するため、単純な事前—事後比較ではなく、回帰分析(Regression Analysis)を用いた。

回帰分析に用いた変数は次のとおり。すべての政策変数 (Policy variables)は、はダミー変数で表現した。つまり、その州で政策が実施されていれば「X=1」、実施されていなければ「X=0」と表わす(ビール税と「母親の会」の支部数除く)。

(1)政策変数 (Policy variables)

各種の政策
・飲酒運転の検挙数(血中濃度 0.08%以上)
・飲酒運転の検挙数(血中濃度 0.1%以上)
・未成年者の飲酒運転の検挙
・免許の取り消し
・刑務所への収容日数
・逮捕に先立つ呼吸内アルコール濃度検査
・アルコール販売店規制法
・自動車内でのアルコール飲料開封禁止
・飲酒が許される年齢を 21 歳と定める
・ビール税(Cents 1999)
・シートベルト着用規制
・段階的付与取得
・「飲酒運転を撲滅する母親の会」の支部数

(2)「その他の要素」(Other control variables)

その他の要素	平均値
・州の平均収入(000 $)	24.110
・州の失業率	6.077
・州の運転者の平均年齢	43.070
・州の運転者の平均的運転距離('000miles)	12.070

３．データの入手

データは、全米 50 州のx20 年間 = 1,000 データである。つまり複数地域x複数時点のデータである。データ数が多いので、多数の政策変数（X）を入れることが可能だったと理解できる。

データの入手先は次のとおり。交通事故死亡数は、「全国高速道路安全管理による死亡事故報告システム」（the National Highway Traffic Safety Administration's Fatal Accident Report System (RARS)）から得た。「各種の政策」に関する有無は、「州のアルコールと高速道路安全関連法規ダイジェスト(1982-2000)」（Digest of State Alcohol-Highway Safety Related Legislation (NHTSA, 1982-2000)）、「交通安全情報」（Traffic Safety Facts）, ウェブ、個人的な問い合わせによって情報を得た。「その他の要素」は、政府経済分析局と政府労働統計局、政府統計局のウェブサイトと交通局の年報から情報を得た。

50 州 x 20 年間 = 1,000 データ

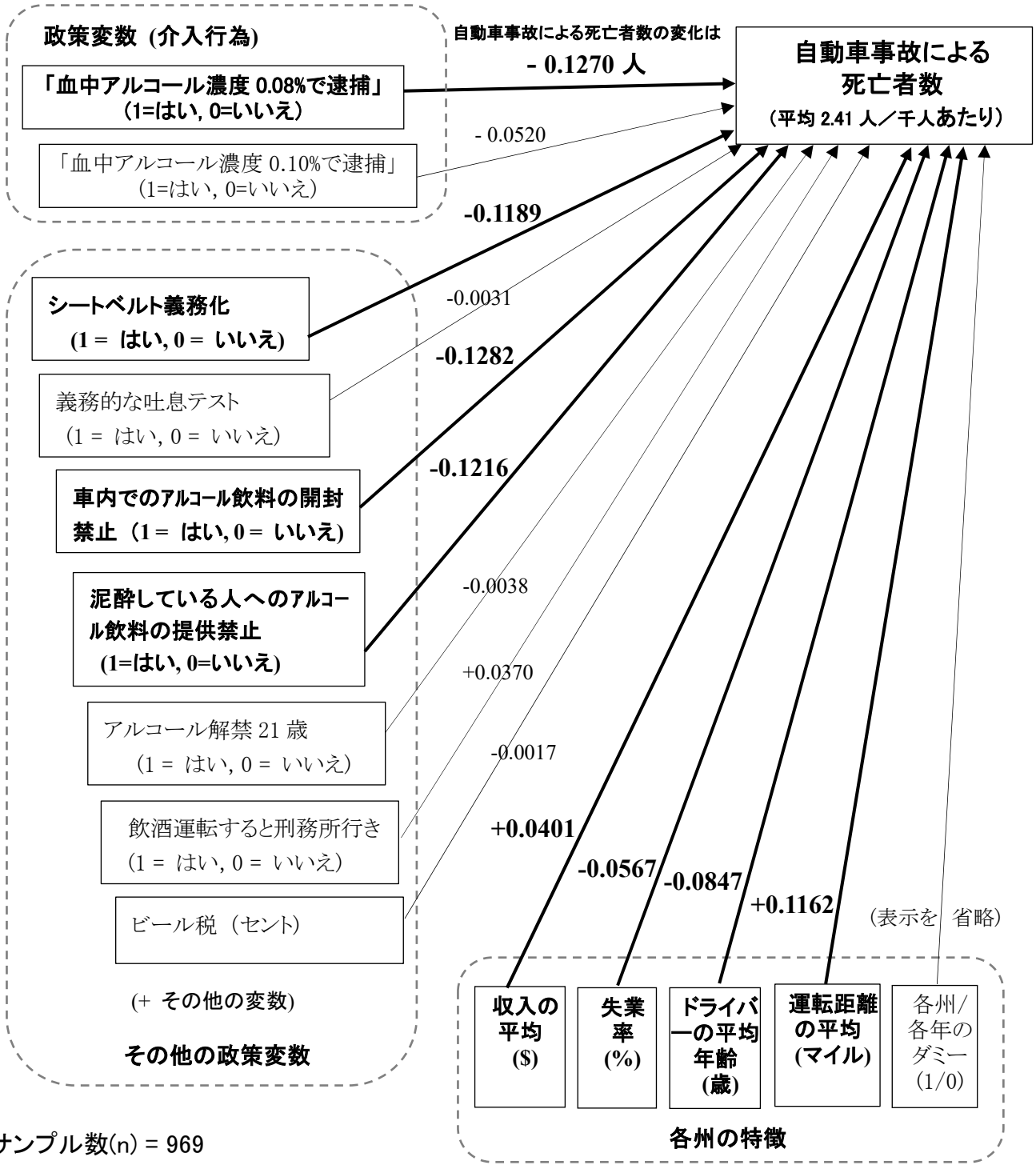
４．分析結果

州ごとの上記のデータを回帰分析の式に投入して、政策ごとの「傾き」を計算した。それぞれの「傾き」が政策の効果の強さを表わす。血中アルコール濃度の規制に関しては、基準として 0.10%を採用している州と 0.08%を採用している州があるので、それぞれの政策の効果が「傾き」として計算されることになる。回帰分析の結果は次のとおり。

回帰分析結果の可視化

（自動車事故による死亡数の全州・全期間の平均値は 2.41 人／千人あたり）

図の見方:それぞれの政策変数の値が 0（いいえ）から 1（はい）に変わると、その傾き（矢印上の値）の分だけ、Y（自動車事故による死亡者数）が変化する。例えば、「血中アルコール濃度 0.08%で逮捕」が0（いいえ）から 1(はい)になると、自動車事故による死亡者数が 0.127 人／千人あたり減る。



サンプル数(n) = 969

説明力=94.5% （R^2 = 0.945）

（Y のばらつきを説明する度合いのこと）

(注) ── 太い矢印 ：5% 水準で有意

── 細い矢印 ：5% 水準で有意とは言えない

（1） **血中アルコール濃度の逮捕の基準を 0.08%とする規制は交通事故死亡を 5.3%減少させる効果がある**と評価された。一方血中アルコール濃度の逮捕の基準を 0.10%とする規制は交通事故死亡を 2.2%減少させる効果がある（ただし、統計学的に有意とは言えない）と評価された。したがって、逮捕の基準を 0.10%以下から 0.08%以下へ規制を強化することにより、交通事故死亡を**さらに 3.1%減少させる**ことができる。（以下の計算になる）

自動車事故による死亡数の全州・全期間の平均値は 2.41 人／千人あたり
（1）「血中アルコール濃度 0.08%で逮捕」の傾き（図の矢印の上の数字）は 0.127 人減少なので、 - 0.127 人÷2.41 人= -0.053 =5.3%減少。
（2）「血中アルコール濃度 0.10%で逮捕」の傾き（図の矢印の上の数字）は 0.052 人減少なので、 - 0.052 人÷2.41 人= -0.022 =2.2%減少。
その差: （1）-（2）= 5.3% - 2.2% = 3.1%

さらに、その他の施策については以下の効果があると評価された。

（2） シートベルト義務付け法は、同様に交通事故死亡を 5.1%減少させたと評価された。そのほか、車内でのアルコール飲料の開封禁止、泥酔している人へのアルコール飲料の提供禁止なども交通事故死亡を減少させたと評価された。

（3）その他に、州の収入の平均、失業率、ドライバーの平均年齢、運転距離の平均なども影響すると評価された。

また、上記の評価のあと、交通事故死亡数を、週末及び夜間だけの死亡事故数、若年層だけの死亡事故数、飲酒していて発生した死亡事故数などいくつか特定の場合の交通事故死亡数のみに限って計算して、上記の各種施策の効果を細かく評価することを試みている。さらに、収入の違い、既婚／未婚の違い、年齢の違い、走行距離の違いによって、上記の施策の効果がどのように違っているかも計算することを試みた。この論文の結論は以下のとおりとされた。

このインパクト評価の結論

血中アルコール濃度の逮捕の基準を 0.1%から 0.08%へ強化する施策は交通事故死亡を減らす効果があったと結論することができる。ただし他の政策も効果を発揮していたことが分かった。規制を強化していない 15 の州でも規制強化を受け入れるべきかどうかについては、他の施策で同程度の効果を上げられる場合があるので、社会全体の更生（Social Welfare）を最大化させるにはどの施策を採用するのがいいのかを、今後、費用対便益分析などを適用して検討するべきだろう。

５．利点、制約、日本での適用に関する留意点

この方法の利点としては、事前の段階から指標値の収集やアンケートの実施をしなくていいという点があげられる。また事後の段階でもアンケートをやらずに既存のデータが利用できる場合が多く、さらに簡便だと言えるだろう。

この方法の制約としては、用意したデータで適切に実態をとらえられているかどうかは分からないという点があげられる。入手できるデータだけが用いられるため、喫煙とがんの発生率の評価と、肥満と死亡率の評価などでほとんど同じデータセットを利用した評価があったりするが両者にはとくに関係がない。単にデータの入手の問題である。それから、過去 10 年分のデータを使うか、15 年分のデータを使うか、あるいは 20 年分を使うかで、評価結果がだいぶ違ってくるが、なぜその期間を選択したかの説明がなされることはまずない。

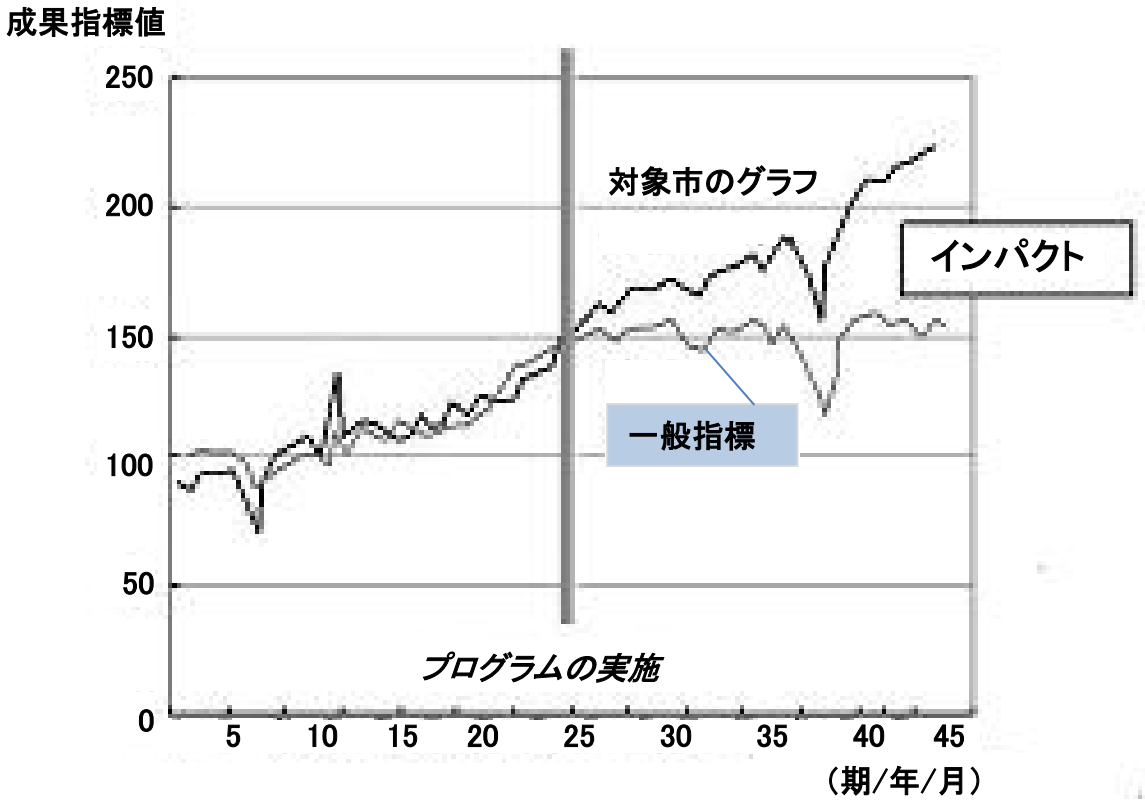
つまり世間一般が抱いている科学的なイメージにも関わらず、回帰分析には恣意性が入る余地がかなりある。しかしそれは回帰分析という手法自体の問題ではなく、多くの場合データの存在の有無や入手可能性に制約があることに常に留意する必要がある。

なお、2020年代になって、**統計分析結果の「可視化」（Visualization）**が叫ばれるようになった。回帰分析も従来のように、**ギリシャ文字による数式と、専門的知識がないと理解できない巨大な表で説明しようとせずに、箱と矢印を使ったシンプルな描画でビジュアルに説明する**べきである。今回の回帰分析が実施された当時は、描画で表示することが一般的でなかったため、解説者（佐々木）が描画して説明した。

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

３．一般指標デザイン

(Generic Control Design)



［説明］

全国平均値、全県平均値などの一般指標値を比較に用いる。外部要因による影響値をある程度除去して考えることができるので（なぜなら対象地域が受けた影響とある程度同じ影響を一般指標値も受けているはずだから)、因果関係の存在の特定に関してある程度の信頼性を確保できる。わりと簡単に用いることができる。

［検定テスト］

目視による判断

Eyeball judgment

一般指標デザインの適用事例：
アルバータ州のビジネスプラン（カナダ）

問題の所在と評価結果

日本では最近、自治体の財政破綻が真剣に議論されるようになってきた。民間企業がつぶれても公共組織はつぶれない、という常識が疑われ、実際に自治体が破綻する可能性が高まっている。こうした財政危機から脱出した自治体の好例として、カナダのアルバータ州の事例がある。しかもアルバータ州は、カナダで最高の公共サービスを最低の税率で提供しながら、財政危機から脱出したのだ。

ラルフ・クレイン知事

アルバータ州の知事ラルフ・クレイン氏は、テレビのニュースキャスター出身で、民間のマネジメント手法を州政府の行政に大胆に導入することを実行した。その発想に基づいて、「アルバータ州のビジネスプラン」を策定した。徹底的な成果主義に基づいて実行された同ビジネスプランでは、インパクトの測定のために、いくつかの戦略目標に関して「一般指標デザイン」を採用した。これにより外部要因による影響値を相当程度取り除いて、アルバータ州政府の施策によるインパクトを評価することに成功している。

1．施策の概要

1993 年に、「アルバータ州のビジネスプラン」が策定された。単一の「使命（ミッション）」のもと、３つの「コアビジネス」が設定され、さらにその下に合計 18 個の個別目標が設定された樹形図上の戦略である。個別目標のひとつとして「13：アルバータ州民の安全を確保し、生活の場として、労働の場として、そして家庭をはぐくむ場所として、アルバータ州が安全な場所であることを保証する」が設定され、具体的には以下の戦略が立案されて実行された。

1）アルバータ州警視庁は、持てる資源（財政的、人的、時間的）を暴力犯罪の防止に集中させる。また地域の防犯活動を促進するとともに、警察活動への地域住民の参加を拡大させる。

2）家族・社会サービス庁は、個人の経済的自立を支援する。子供の安全を保つ。とくに子供に対する犯罪の早期警戒と早期介入、アボリジアニ（筆者注：カナダにもともと住んでいる人々）の生活ニーズに応える、必要に応じて簡易宿泊施設を用意する。

２．評価結果

18 個の個別目標のいくつかの評価方法として「一般指標デザイン」が適用されており、この 13 番目の個別目標にも一般指標デザインが適用された。以下が評価の仕組みの概要である。

個別目標

「13：アルバータ州民の安全を確保し、生活の場として、労働の場として、そして家庭をはぐくむ場所として、アルバータ州が安全な場所であることを保証する」

成果指標

以下の犯罪発生率（２種類)。（さらに、未成年者に限った犯罪発生率も設定している。）

①人口 10,000 人あたりの暴力犯罪（Violent Crime）の被害件数

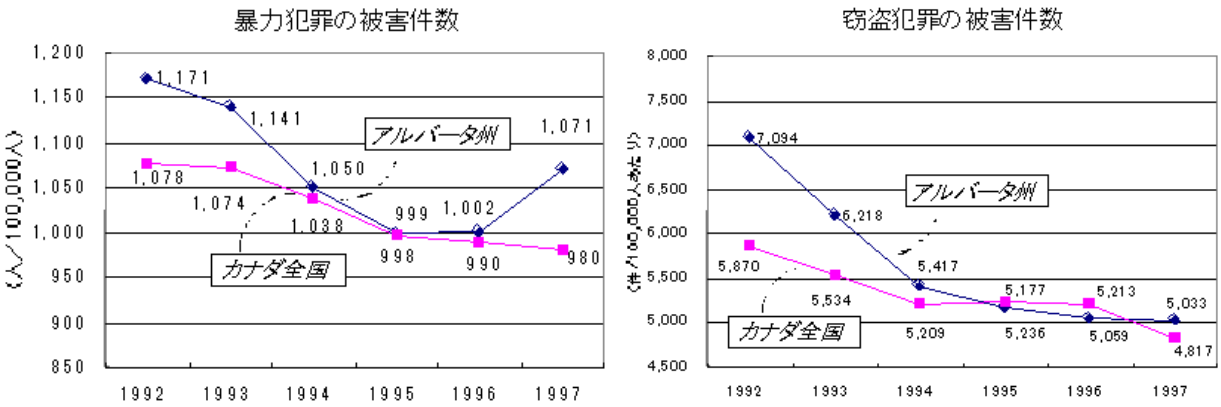
②人口 10,000 人あたりの窃盗犯罪（Property Crime）の被害件数

指標の説明

犯罪発生率は、アルバータ州が安全な場所であるかどうかを直接示す指標である。

数値目標

2000 年までに全国平均以下にする。



戦略期間終了時の評価結果は以下のとおり。

暴力犯罪、窃盗犯罪の率とも、1992 年（基準年）から、全国の改善ペースを上回るペースで順調に改善されている。しかし 1997 年には暴力犯罪が増加に転じたが、アルバータ州政府は、改善傾向に大きな変化はないとしている[1]。

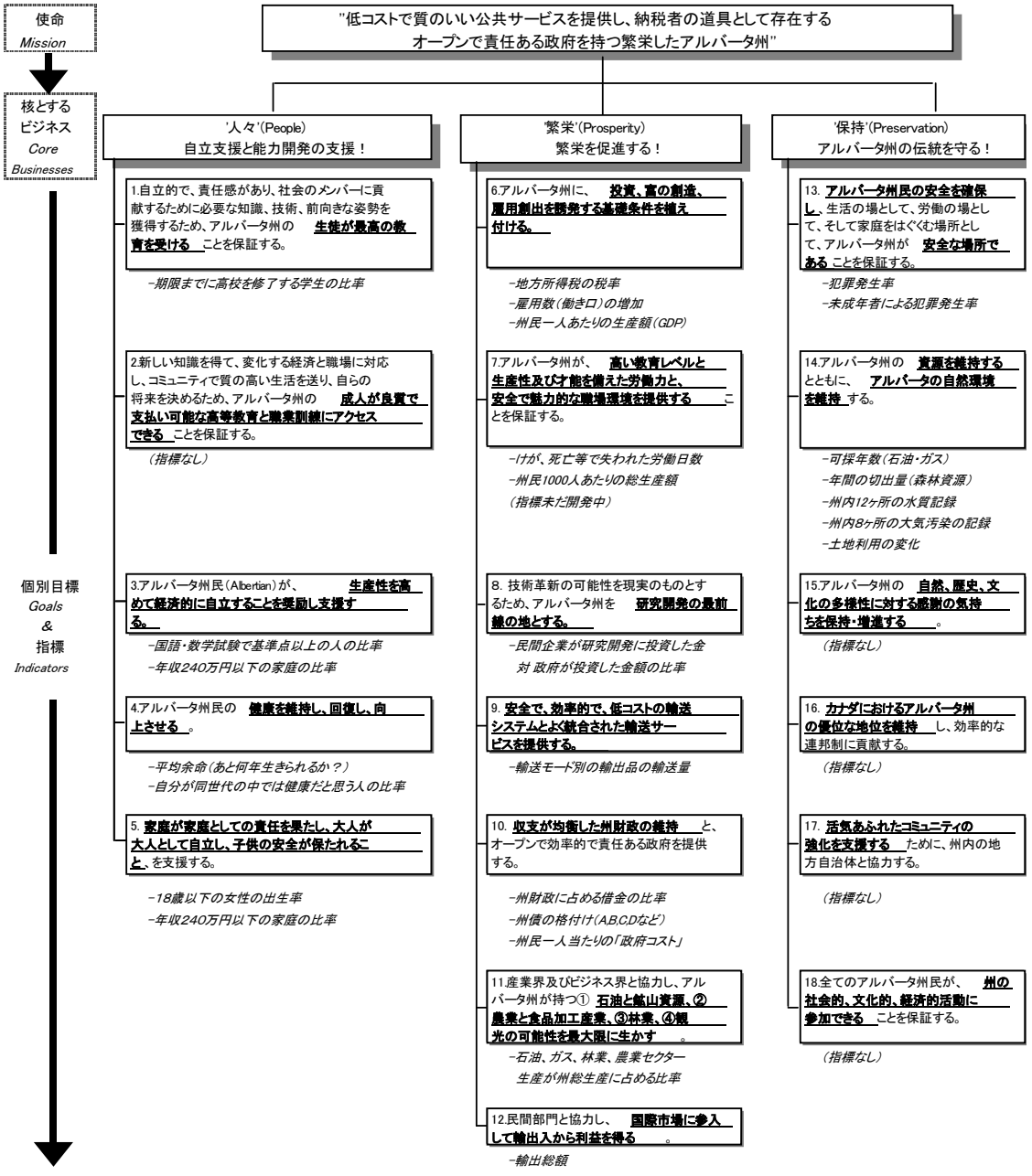
３．利点と制約

一般指標デザインの採用により、外部要因による影響を相当程度除去した上でアルバータ州政府の政策によって効果が発現したのかどうかを比較的純粋に評価することに成功している。世界経済の動向やカナダ連邦政府の政策による全国的な影響などの外部要因がアル

[1] Alberta Treasury, *Measuring Up Report 1999*

バータ州の指標値に影響を及ぼす場合には、全国レベルの指標値にも同程度に影響が及ぶことが想定されるので、**アルバータ州の指標値が全国平均の指標値よりもより改善していれば、それはアルバータ州政府の政策の効果（インパクト）である**とみなすことができる。なお、州のビジネス・プラン（戦略的計画あるいは総合戦略）も参考になるので以下に掲載する。**日本の自治体に見られる「○○の道路を作る」「○○大会を開催する」という「やることリスト」ではなく、住民が幸福を表す指標だけで構成している**のが特徴である。

アルバータ州政府のビジネス・プラン

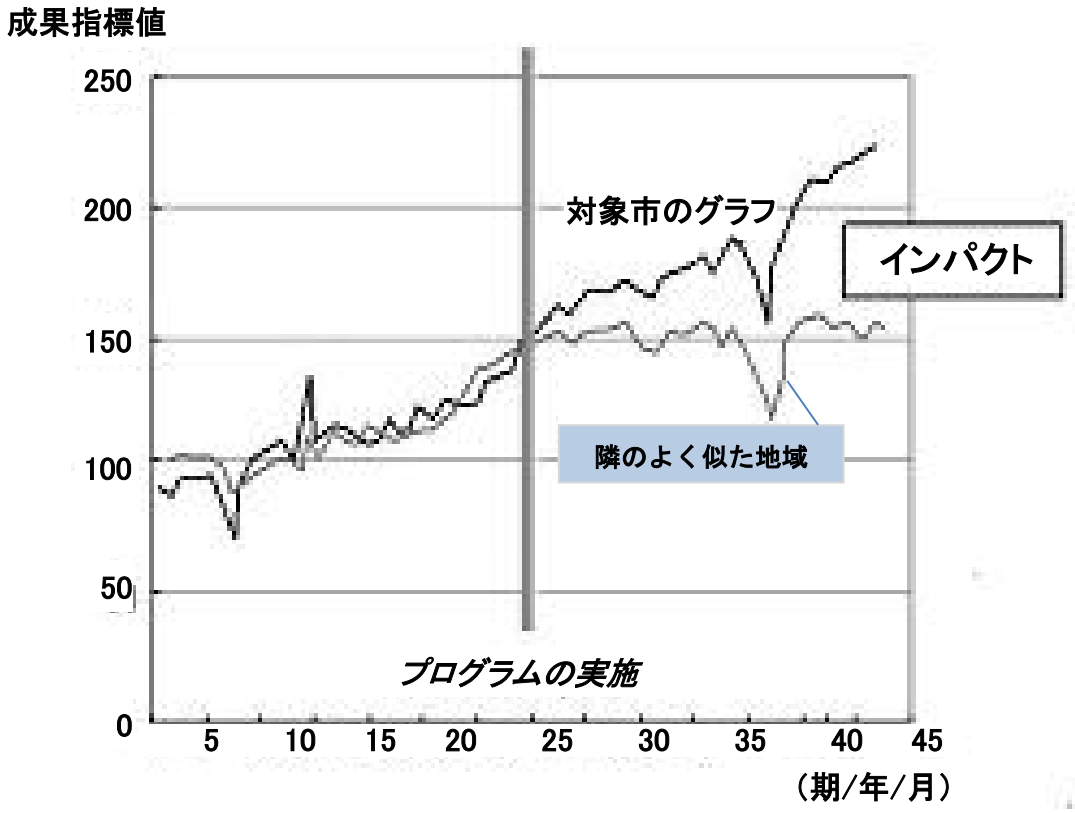


出所）アルバータ州政府ホームページ
http://obm5.treas.gov.ab.ca/comm/perfmeas/measupgu/gra19.gif

1. 事前・事後比較デザイン (Before-After)
2. 時系列デザイン　(Interrupted Time-Series)
3. 一般指標デザイン　(Generic Control)
4. マッチングデザイン (Matched control)
5. ランダム化比較デザイン (RCT)

４．マッチングデザイン

(Matched Control Design)



［説明］

可能な限り近似のグループを選定して比較に用いる。外部要因による影響はどちらのグループも同程度に受けると考えられるので、因果関係の存在の特定のために高い信頼性を確保できる。

[検定テスト]

二群の有意差検定（対応のない t 検定）

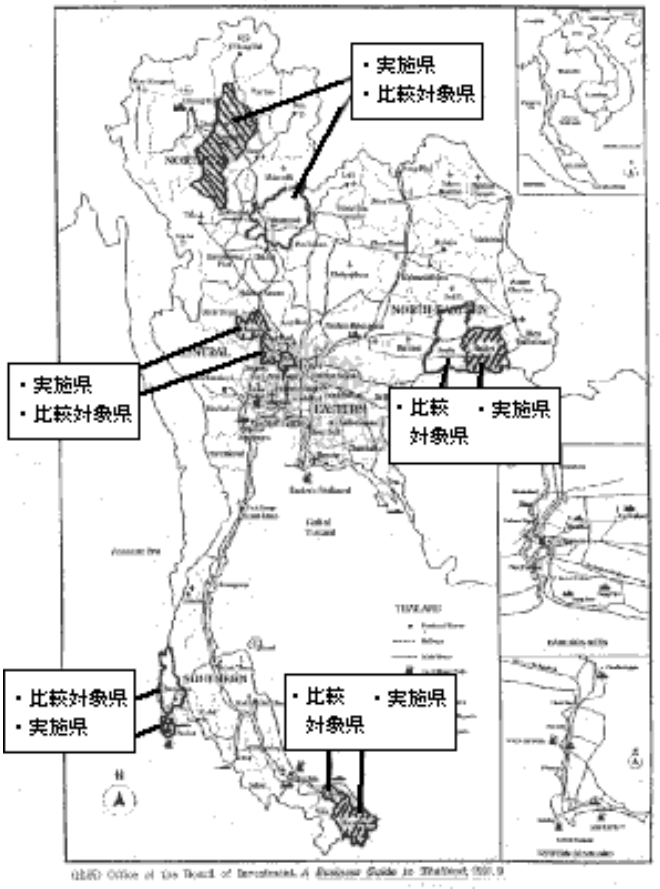
Independent t-test

マッチングデザインの適用事例１：
地方分権化プログラム試行の評価（タイ）

問題の所在と評価結果

タイでは、地方分権化を推進しており、地方分権化プログラムの試行の効果を評価するため、マッチングデザインを用いた評価を実施中である。以下の図のように、面積、人口、産業構造、首都圏からの距離等に基づいて、５つの実施県それぞれに比較対象の県を決めて継続的に指標値をモニタリングしている。

タイの「地方分権化プログラム」の試行に関する
実施県５県と比較対象県５県の位置



利点、制約、日本での適用の留意点

この方法の利点は、簡便だということで、中央政府で整備して公表している県ごとのデータが利用できる場合があり、データ入手可能性が高いということがあげられる。逆に制約としては、施策の適用を受ける以外に実施県と全く同一の条件を備えた近隣の県というのはそもそも存在しないわけで、マッチングするために選定した指標（面積、人口、産業構造、首都圏からの距離等）以外の要因により、効果の指標値の出方に大きな影響が出てくる可能性があるということである。

ただし、日本では、こうした簡便な方法であるマッチングデザインでさえも適用されて来ておらず、現在も一般に用いられてはいない。例えば、**構造改革特区という試みが地域を限定して行われているが、この試みでこそ、施策を実施するという点以外において可能な限り近似した地区を選定して比較対象として用いる**べきであろう。

マッチングデザインの適用事例２：
初等教育に関する４種類の施策の効果（フィリピン）

問題の所在と評価結果

多くの開発途上国において、高い中退率と不十分な学習効果が問題となっている。フィリピンも同様の状況であり、小学校（６学年）を修了する前に約２５％が中退する。また、教えられたことの半分以下しか身についていないという調査結果がある。この状況を改善するため、（１）習熟度別学習教材の無料供与、（２）学校給食の実施、（３）教師と親の連携強化、３種類をそれぞれ組み合せた事業が実施された。

評価調査の結果、小学校の中退率の改善に効果があるのは、「習熟度別学習教材の無料供与」と「教師と親の連携活動」の組み合わせであることがわかった。一方、今回の評価調査が試した施策のなかで中退率の改善に効果が見られなかったのは「給食の実施」であった。また単位コストも計算したところ、「給食の実施」よりも「習熟度別学習教材の無料供与」の方がより安く実施できることがわかったので、「習熟度別学習教材の無料供与」の実施の拡大を提言している。

１．評価の概要

フィリピン政府が、1990～1992年に実施した、中退阻止プログラム（DIP）として、上記音３種類の施策の組み合わせの何通りかの効果が評価された。

中退率の計算は、実施学校におけるプログラム実施前の一年間の退学率からプログラム実施後の一年間の退学率を差し引く（この差が大雑把な改善率ということになる）。さらに、比較学校における同様の率を計算し、その率をさきほどの改善した率から差し引く。残った率が、プログラム実施によって引起こされた純粋な改善率ということになる。（二重引き算法）[2]。

サンプルとなる学校の選定は次の３つの段階を経て行われた。

（１）フィリピンを構成する５つの地方それぞれから、似ていると言える２つの低所得県を選んだ（マッチング）。マッチングの基準は、①教育指標、②保健指標、③住居指標、④失業率、⑤家計支出水準である。

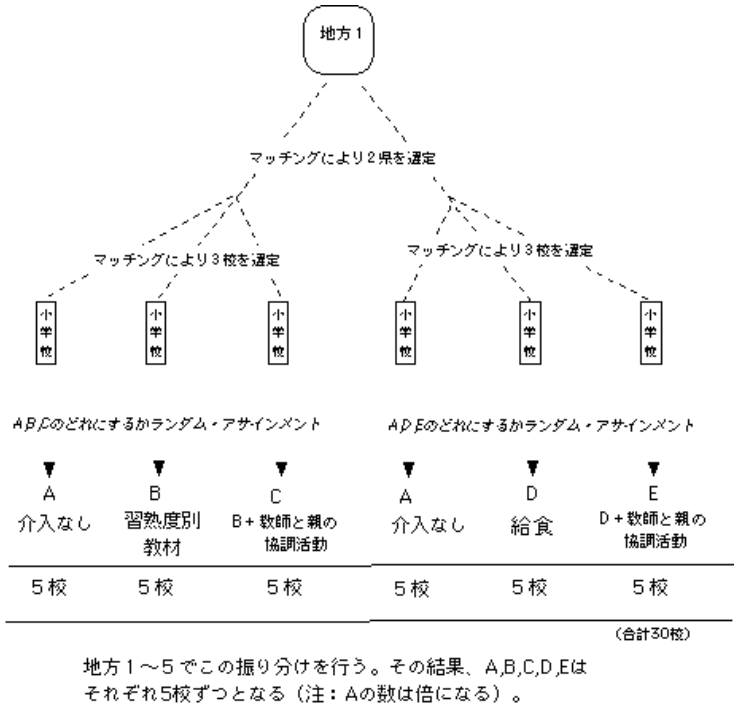
（２）５地方x２県＝10県のそれぞれから、①高い中退率を持つ、②既存の給食プログラムがない等の条件を満たす学校を３つ選定した。（５地方X２県X３校＝30校）

（３）ここから各プログラムの割当である。ある地方から選ばれた２県のうち、片方の県の３校は、A.介入なし（何もしない）、B.習熟度別教材を無料配布、C.習熟度別教材＋教師と保護者の連携活動を実施、のいずれかが割り当てられた。そしてもう片方の県の３校は、A.比較のために何もしない、D.給食を実施、E.給食＋教師と保護者の連携活動を実施、のいずれかに割り当て

さらに、学力成果（Academic Performance）を被説明変数として次の回帰分析を行った。
「学力成果（今期）」＝「学力（前期）」＋「個人特性」＋「家族特性」＋「学習環境」＋「クラス環境」+「プログラムの実施の有無」＋誤差

られた。

この結果、合計30校のうち、B、C、D、Eのプログラムを実施した学校はそれぞれ５校で合計20校、そして何もしない比較のための学校(A)が10校選定された。（下の図を参照）



実施前指標値（ベースラインデータ）の収集は1990‐1991年に実施され、1991‐1992年にプログラムが実施された。その後（1992‐1993）に事後データが収集された。その結果、29学校[3]、180人の教師、約4000人の生徒から詳細なデータを取ることができた。

２．評価結果

プログラムを実施する前の中退率に関するベースライン・データは以下のとおりであった。なお、さらに学力テストの点数もデータもあるがここでは載せていない。事前段階では、E の学校グループをのぞいて、それぞれのグループで差がないことが確認された。

ベースラインデータ (1990-91)

	A 介入なし	B 習熟度別教材	C 習熟度別教材＋教師と親協調活動	D 給食	E 給食＋教師と親協調活動
中退率	9.56	9.29	10.01	8.58	7.02**

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

そして以下が、実施後の指標値である。

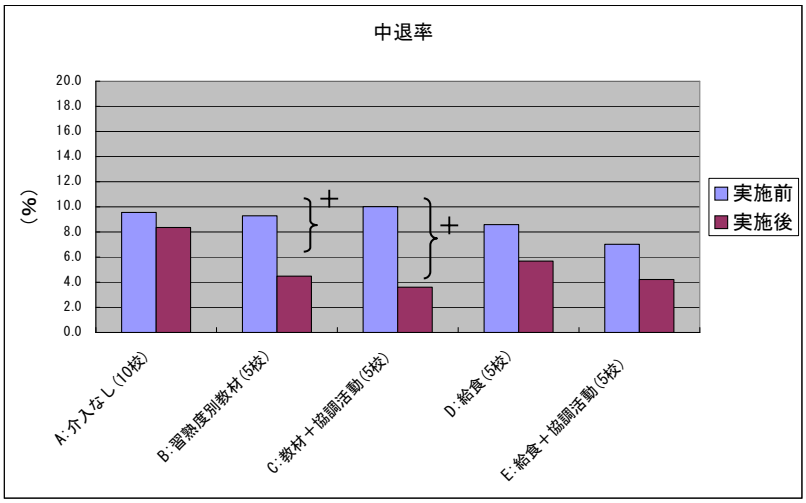
[3] 一校が脱落した理由は、報告書にも記載がなく不明。

実施後データ（1990-91と1991-92の間の変化）

	A 介入なし	B 習熟度別教材	C 習熟度別教材＋教師と親協調活動	D 給食	E 給食＋教師と親協調活動
中退率	8.36	4.49	3.61	5.68	4.22
中退率の変化	-1.2	-4.8	-6.4	-2.9	-2.8
P-value	*0.328*	*0.004****	*0.005****	*0.104*	*0.11*
（Aとの差）	n.a	-3.6	-5.2	-1.7	-1.6

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

これをグラフに表すと次のようになる。



B. 習熟度別学習教材の供与と、C.それと教師と保護者の連携活動の組み合わせ、の二つのプログラムで効果があることが確認された。一方、D.給食の実施は、中退率の改善に貢献しているとは言えないことが確認された。さらに、実施にかかる単位コストを試算しており、効果があることが確認された B.習熟度別教材の供与が、90 ペソ／人、教師と保護者の連携活動が 33 ペソ／人であったのに対して、効果があるとは認められなかった D,給食の実施は、946 ペソ／人と試算された。この評価結果と単位コストの試算から、世銀に対して習熟度別教材の供与の拡大を推進すべきであると評価実施者は提言した。ただし、中退率に関しての提言であり、学力向上を目的とした場合には、この評価調査で試したどの方法もインパクトがあるとは言えなかったので、さらに他のプログラムが試されるべきだとも付け加えている。

なお、評価実施者は次の３点をコメントしている。（１）学校給食に効果が見られなかったという結果はやや行き過ぎで、対象グループをもっと絞ってやればもっといい結果が出るかもしれない。（２）サンプル数が小さかったことが効果の判定にかなり影響したかも知れない。（３）プログラム実施と評価実施の間が極めて短いので中長期間に現れるような効果を測定することは出来なかったかもしれない。

３．利点、制約、日本での適用に関する留意点

この例では、介入なしも含めて５種類の施策組み合わせに関して効果を比較している。これによりどの施策が最も効果があるのかがわかる。また対立する施策案がある場合にこのやり方を利用するのは、行政の意志決定にさらに意味のある情報を提供することになるだろう。

この例に関する留意点としては、マッチングが甘いということが指摘されねばならない。マッチングに用いた指標の数が二つとか三つで少なすぎたのだ。そのため、ベースライン値（実施前指標値）を測定した時点で、E.のグループの成果指標値（中退率）がすでに相違している。もっとマッチングを見る際の指標を多くすべきである。またサンプル数も少ないことは評価実施者自身も指摘しているが、各グループとも最低 25 あるいは 30 欲しいところである。

日本で適用する際の留意点としては、次があげられる。アメリカと違い、日本では全国一律の教育指導要領が適用されていることもあり、施策の評価のために、良好なマッチングを示す学校を比較的容易に準備できるであろう。また、この例のように５つの地方で実施することもなく、ある県で実施すれば、その評価結果は相当程度全国的に適用して問題ないであろう。こうした**日本の状況を踏まえると、一つの県においていくつかの市教育委員会が協力すれば、よく近似した学校が必要数だけ容易に確保できる**可能性がある。

５．議論

個人的な話になるが、この事例を学会で報告した。発表後に質問を受け付けたところ、**『こんな複雑なことをしないと分からないか。こんなことをしなくても専門家が見ればわかるんだ』と一喝**されたことがあった。しかし専門家が効果があると言うことを全国で適用したらじつは効果がなかったという経験が共有されて社会実験が普及しつつあるのだ。専門家を自称することは容易だが、社会実験から得られるエビデンスにはいつも謙虚であるべきである。

資料出所）Tan,J.P.,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.

事後的なマッチングデザイン（統計的等化デザイン）の適用事例：
雇用促進プログラムの効果（チェコ）

問題の所在と評価結果

最近の世界的な市場経済化の流れのなかで、途上国や旧社会主義国では、国営企業の民営化や規模縮小に伴い大量の失業者の発生といった事態に直面している。これに対処するため雇用促進プログラムが政府によって実施されることがあるが、チェコでは５種類の雇用促進プログラムが世銀の融資によって実施された。それらは、（１）新卒訓練プログラム、（２）技能再訓練プログラム（数週間～最大７ヶ月）、（３）公共土木事業の短期雇い、（４）新規採用に対する財政支援、（５）個人による新規事業開始に対する財政支援である。これらは就職率の改善に効果があったのだろうか？

評価結果は、プログラムや参加者グループによって大小の効果が観察されたが、少なくとも「公共土木事業の短期雇い」には就職率向上の効果が全く認められなかったのでそれを廃止すべきで、さらに、その廃止により浮く資金と資源を、各種のプログラムで効果発現の度合いが大きかった若年の女性グループに優先的に投入すべきだと提言している。

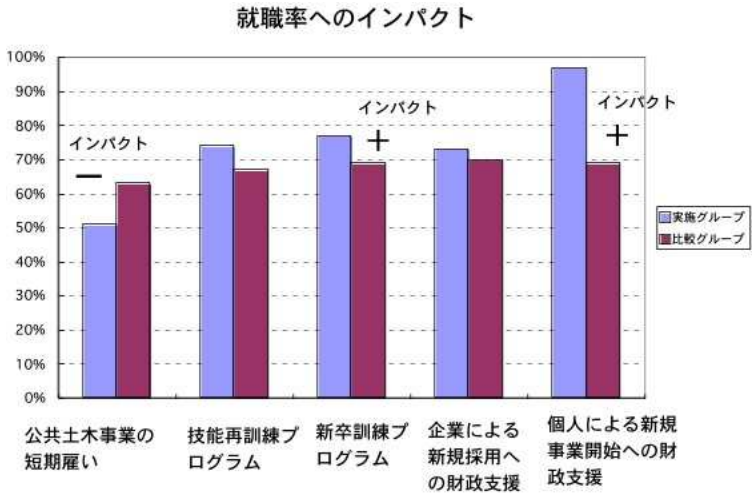
１．施策の概要

まず、職業安定所に登録している失業者から無作為抽出により約25,000人を選んで、アンケート調査への協力依頼状を出した。そのうち4,477人が協力に同意したので、アンケートを送付して回答してもらった。質問は、a) 過去に（１）～（５）のそれぞれの雇用促進プログラムに参加した経験があるか、b) その後実際に雇用されたか、c) 雇用された場合は給与はいくらだったか、である。

回答した4,477人のうち、（１）新卒支援プログラムに参加したことがあると答えた人数は278人であった。次に（１）に参加したことのない人間から同じ特徴を備えた人間を同数選定して、成果指標（雇用されたか否か、給与水準はいくらだったか）の差を計算することになる。プログラムへの参加の有無以外の状況や条件はなるべく同じであることが望ましいわけであり、選定に際しては次の７つのマッチング指標を用いて、可能な限り一致している個人を選定した。それらは、雇用促進プログラムへの参加の有無以外で就職に影響を及ぼすと考えられる要因である１）年齢、２）性別、３）最終学歴、４）失業期間の長さ、５）居住している都市の大きさ、６）既婚／未婚、７）以前の職種であった。278人と同数を選定することを目指したが、結局194人の個人を選定した。以下、（２）～（５）も同様にマッチングによる選定を行って比較グループを形成した。

２．評価結果

評価結果は次のとおり（次ページのグラフ参照）。少なくとも「新卒訓練プログラム」と、「個人による新規事業開始への財政支援プログラム」は、就職率の改善効果があると評価された。逆に「公共土木事業の短期雇用プログラム」は、効果がないばかりか、就職にマイナスの影響が出ていることが分かったので、廃止すべきである。



注）図中の『＋』は統計検定による有為を示している。

公共土木 短期雇い

	実施ｸﾞﾙｰﾌﾟ	比較ｸﾞﾙｰﾌﾟ	インパクト	（判定）
就職率（現在）	51%	63%	-12%	ﾏｲﾅｽ有（強）
96年以降の就職率	77%	72%	5%	無
月給（Kc）	5393Kc	6631Kc	-1238Kc	ﾏｲﾅｽ有（中）

技能再訓練 ﾌﾟﾛｸﾞﾗﾑ

	実施ｸﾞﾙｰﾌﾟ	比較ｸﾞﾙｰﾌﾟ	インパクト	（判定）
就職率（現在）	74%	67%	7%	無
96年以降の就職率	88%	76%	8%	有（強）
月給（Kc）	6536Kc	6636Kc	100Kc	無

新卒訓練 ﾌﾟﾛｸﾞﾗﾑ

	実施ｸﾞﾙｰﾌﾟ	比較ｸﾞﾙｰﾌﾟ	インパクト	（判定）
就職率（現在）	77%	69%	8%	有（弱）
96年以降の就職率	89%	75%	6%	有（強）
月給（Kc）	6500Kc	7844Kc	-1344Kc	ﾏｲﾅｽ有（中）

企業による 新規採用へ の財政支援

	実施ｸﾞﾙｰﾌﾟ	比較ｸﾞﾙｰﾌﾟ	インパクト	（判定）
就職率（現在）	73%	70%	3%	無
96年以降の就職率	91%	80%	11%	有（強）
月給（Kc）	5605Kc	6111Kc	-506Kc	有（中）

個人による 新規事業開 始の財政支援

	実施ｸﾞﾙｰﾌﾟ	比較ｸﾞﾙｰﾌﾟ	インパクト	（判定）
就職率（現在）	97%	69%	28%	有（強）
96年以降の就職率	97%	82%	16%	有（強）
月給（Kc）	6306Kc	7074Kc	1768Kc	無

注）表中の強、中、弱は、統計検定によりクリアした有為水準を示している（1%,5%,10%）

さらに政策変更につながる評価情報を提供するため、（１）～（５）のプログラムそれぞれのインパクトの有無／程度の計算に加えて、性別、年齢、既婚／未婚、以前の職種、以前の企業規

模等で小グループごとに分けてインパクトの有無／程度を測定した。この小グループ化により、どの小グループにより高いインパクトが現れているかが示唆されるわけである。結論は、若年の女性グループで最も高いインパクトが見られた。これらの結論に基づいて評価者は、効果が認められなかったプログラムの廃止により浮く資金と資源を、若年の女性グループに優先的に投入すべきだと提言した。

３．利点、制約、日本での適用に関する留意点

この方法の利点としては、これまでの評価方法のように、事前の段階から指標値を収集しなくていい点があげられる。つまり数年前のベースラインデータ（実施前指標値）が存在しない場合の評価調査には、本件のように事後的なデータを分割して比較する統計的等化デザインが利用できるということである。小グループに分けて効果を計算することは、限りある資源（財政的・人的・時間的）から最も効果のあがるプログラムを設計するためにたいへん有効である。

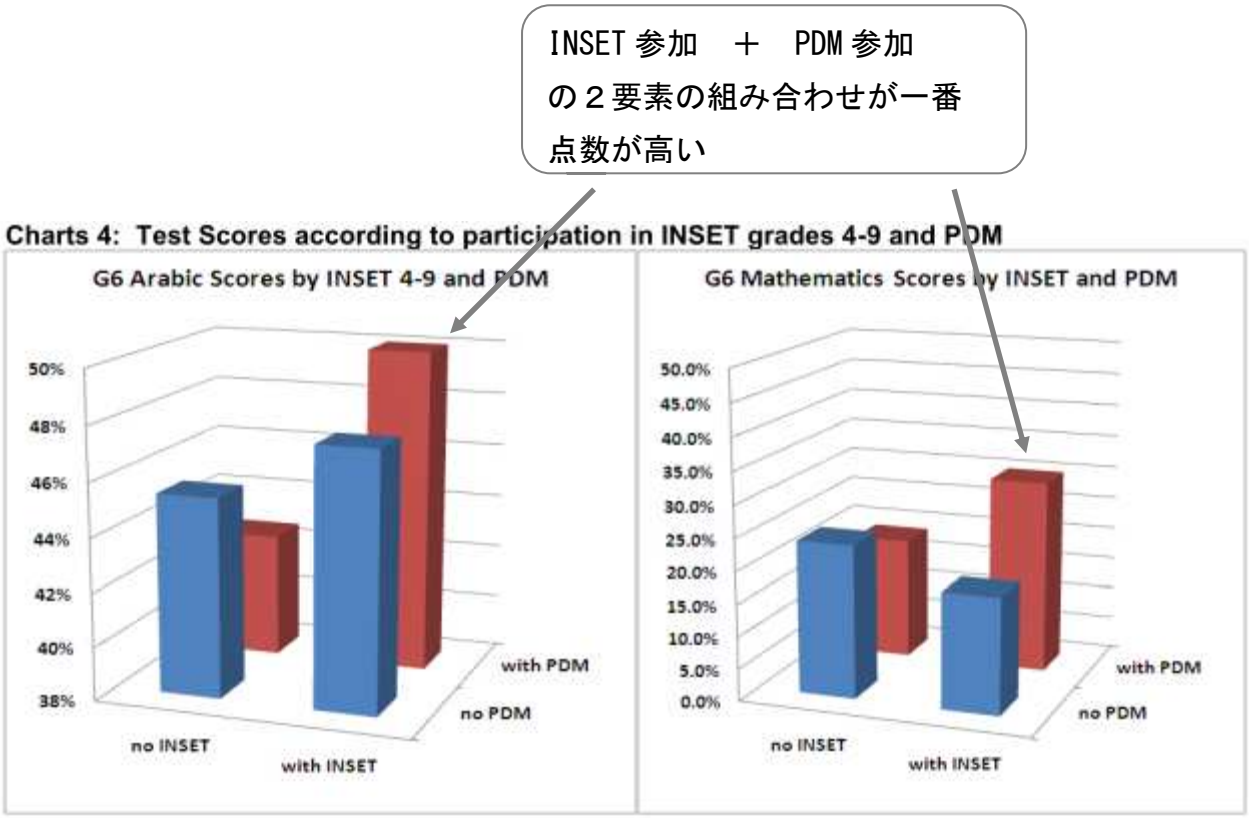
この方法の制約としては、分割をどこまで続けるかが恣意的になる可能性があるということである。２分割、それでだめなら４分割、それでもだめなら８分割、さらに16分割、32分割、64分割、128分割、256分割と、効果が見いだせるまでどこまでも分割を続けることも可能である。事前にどこまでどういう基準で分割するかあらかじめ関係者で決めておくことが勧められる。

評価手法とは直接関係ないことで、この評価事例から日本が学べることがある。もし、この評価を日本で行ったとすると、だいたいの施策は効果が認められたのだから特に何もしなくてもいいが、公共土木の短期雇いが就職率の改善に効果がないというのだから、まさにそれをどう改善すれば効果が出るようになるのかの提言を書くべきだいうことになるだろう。そして改善して続けるべきということになる。つまり**同じ評価結果から、日本ではこの事例とは逆の提言が出される可能性がある**ということである。少なくとも、効果の有無については同じ手法を用いてアメリカなどの評価実施者と同じ結論を出すことはできるだろうが、そこから自動的に提言が出てくる訳ではなく、**提言の選択はたぶんに評価実施者の価値判断や個別の社会状況による部分がある**ということを我々は認識すべきである。

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

統計的等化デザイン（事後的なマッチングデザイン）の適用事例：
現職教員研修(INSET)と専門性開発ミーティング PDM)の効果（イエメン）

GTZ は、イエメンにおいて、（１）合宿方式の現職教員研修(INSET)と（２）教育現場における専門性開発ミーティング(PDM)の二つのプログラムを同時に支援した。その結果は以下の 3 次元のグラフとしてまとめられた。

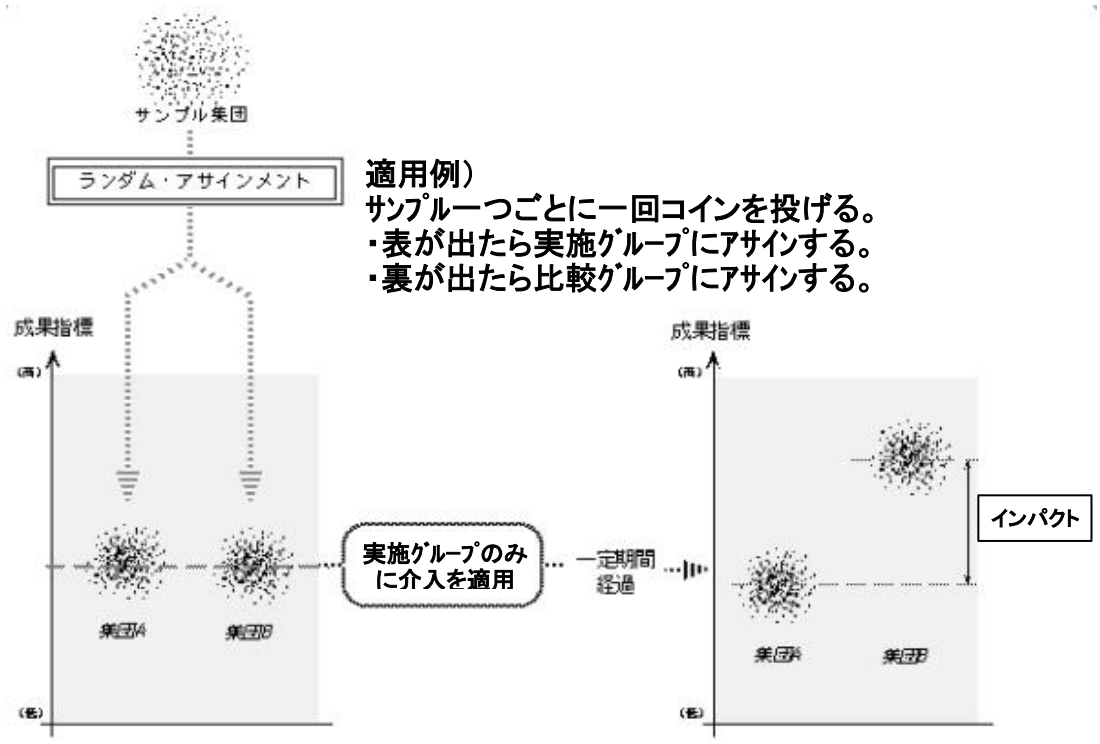


このグラフに基づいて GTZ の報告書は次のように解説している。『PDM に参加した教員が教えるクラスのテスト結果は次の事実に影響されている。それは、PDM と INSET に参加した教員はわずか一人しかいないということである。つまり、研修のコンビネーションを受けた教員はただ一人だということである。しかし、PDM の影響は明らかに見られる。教員が PDM に参加したかどうかによって、アラビア語と数学の双方のテスト結果に関して差が見られる。それは、たとえ、教員が INSET に参加していたとしてもである。数学における改善効果(18.1%から 30%)は、66%である。アラビア語の改善効果 (47.5%から 50%) は、わずか 5%である。』この文章から分かることは、**分析の限界を率直に認めた書きぶりになっている**ということである。今後日本で同様の分析を行う場合にも、このように分析の限界を明記することが勧められる。

（出所）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．ランダム化比較デザイン（実験デザイン）

（Randomized Controlled Trial (RCT), Experimental Design)



[説明]

施策の実施前に、政策適用を無作為割付（ランダム・アサインメント）により、実施グループと比較グループに分ける。成果指標（Outcome indicators）に現れた違いは、途中の唯一の違いである「介入を適用されたか否か」によって引き起こされたと純粋に判断することができる。なお、外部要因による影響は全く同一になっているので考える必要はない。因果関係の存在の特定に関してたいへん高い信頼性を誇る。ただし実際の適用は難しい。

［検定テスト］

二群の有意差検定（対応のない t 検定）

Independent t-test

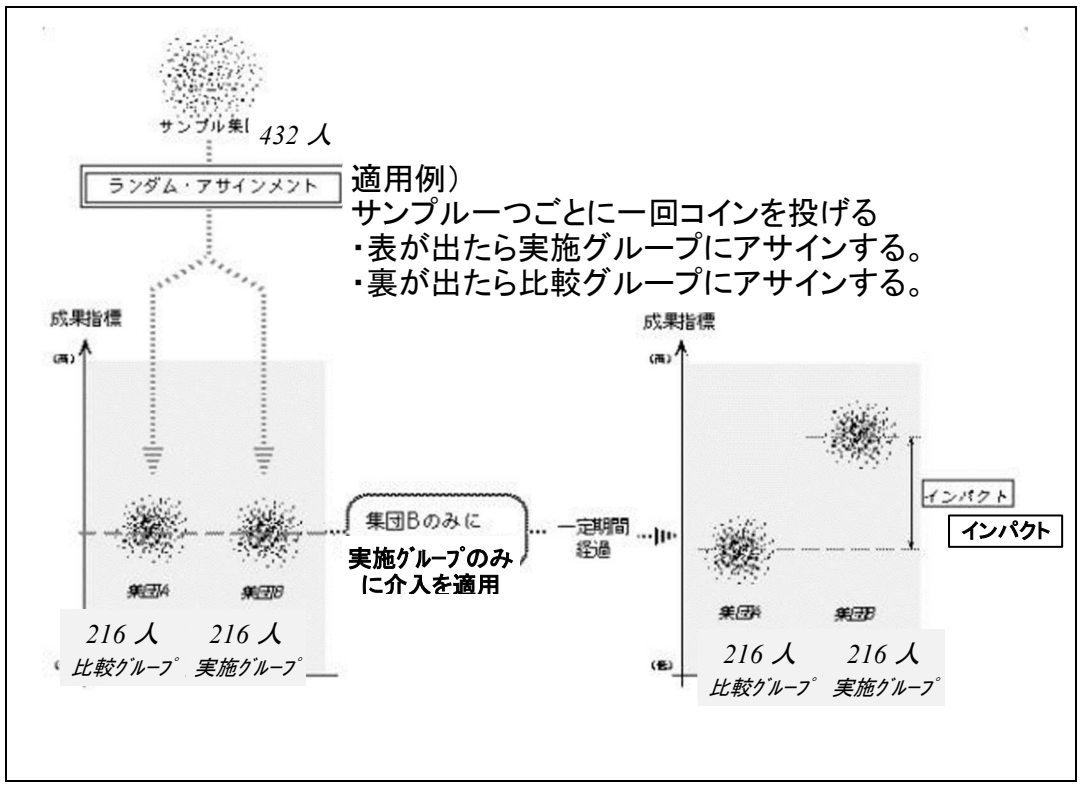
実験デザイン（RCT）の適用事例１：
出所者への小額財政支援施策（アメリカ）

１．問題の所在と評価結果

犯罪の再発を防ぐにはどのような「政策」が有効か？ひとつの考え得る「政策」案は、刑期を終えて出所した者が通常の市民生活へスムーズに移行することを手助けするため､彼ら(彼女ら)に対して小額の財政援助を行うことである。しかし、この「政策」案は本当に効果があるのだろうか？犯罪を犯したうえに現金までもらって、また犯罪を犯すことがないのか？こうした質問に答えるため、メリーランド州ボルチモアでこの「実験」が実施された。その結果、少なくとも「窃盗」に関しては、プログラム実施が意図された効果を持つという結論された。

２．施策の概要と評価デザインの概要

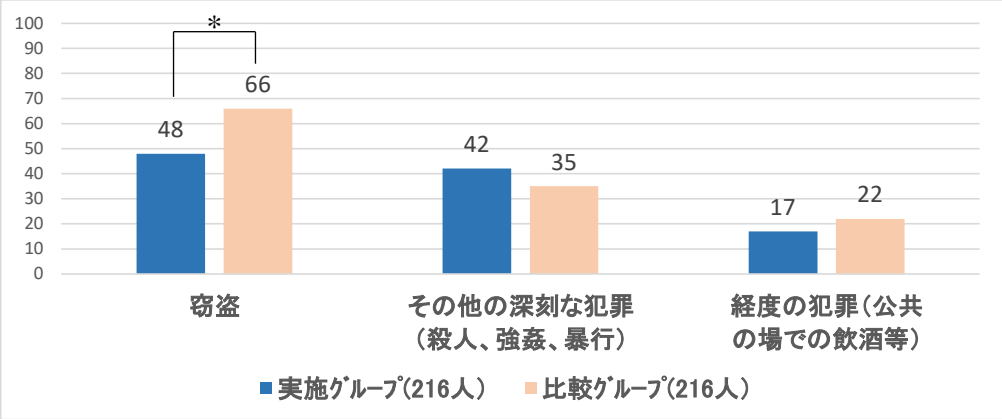
1970 年代の後半に実施された本実験では、メリーランド州立刑務所から出所してボルチモアへ戻った出所者が対象とされた。出所者 432 人に関して、ランダム・アサインメントによって、実施グループになるか比較グループになるかが決定された｡実施グループに割り振られた人(216 人）には、雇用されるまでのあいだ最大 13 週間にわたって毎週$60 が支給された。比較グループに割り振られた人（216 人）には、実験に参加してもらうが支給はないことが伝えられた。



３．評価結果

ボルチモア警察の逮捕記録によって、実験参加者の1年後の逮捕率に関してつぎの結果が得られた。

	実施ｸﾞﾙｰﾌﾟ 216人	比較ｸﾞﾙｰﾌﾟ 216人	差
窃盗	48人 (22.2%)	66人 (30.6%)	-18人 (-8.4%)
その他の深刻な犯罪（殺人、強姦、暴行）	42人 (19.4%)	35人 (16.2%)	7人 (+3.2%)
軽度の犯罪（公共の場での飲酒等）	17人 (7.9%)	22人 (10.2%)	-5人 (-2.3%)



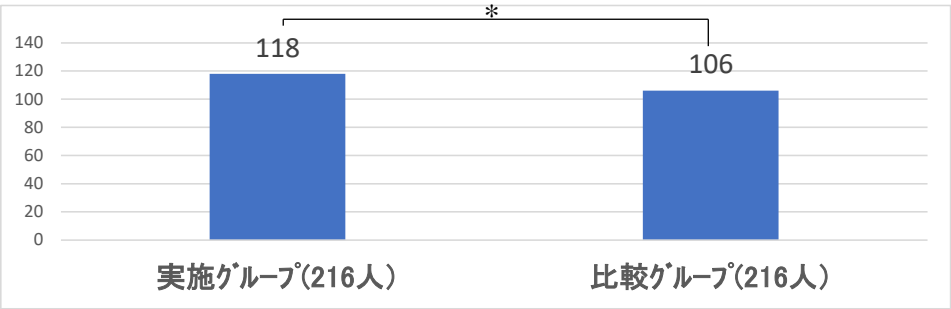
*…統計学的に有意（ただし有意水準の情報は未記載）

「窃盗」に関しては、プログラムを適用された実施グループの方が、比較グループに対して-8.4%低い逮捕率を示した。しかしこの差は、プログラムがなくても偶然に起こり得る程度の差よりも大きい差なのだろうか。この-8.4%は、統計テストをパスした。その他の種類の逮捕率は、統計テストをパスしなかった。言い換えれば、「その他の深刻な犯罪」と「軽度の犯罪」に関する実施グループと比較グループのあいだの差は、偶然に起こりえる程度の差より大きいと判定することはできなかった。

なお、就職率に関しては、以下のとおりの差が測定された。

	実施ｸﾞﾙｰﾌﾟ 216人	比較ｸﾞﾙｰﾌﾟ 216人	差
就職率（フルタイム）	118人 (54.7%)	106人 (49.0%)	12人 (+5.7%)

*ただし第４四半期のみ。第１～３四半期の差は偶然に起こりえるよりも大きいとは判断できなかった。



*…統計学的に有意（ただし有意水準の情報は未記載）

4．結論

評価結果は次のとおり。少なくとも「窃盗」に関しては、プログラム実施が意図された効果を持つと評価された。

さらに、この実験で明らかになった効果は、この施策を大々的に実施するのに十分な値なのであろうか。この質問に答えるために、つぎに費用対便益評価が実施された。アメリカ労働省がその評価を担当した結果、社会全体の見地（from a social perspective）から計算すると、以下のように、便益／費用比率は最も慎重な計算の場合でも 4.02 倍、最も楽観的な計算の場合では 53.73 倍と計算された。

	社会便益	社会費用	便益／費用比率
最も低い計算	$108,565	$27,000	4.02
最も高い計算	$870,431	$16,200	53.73

したがって、この施策によってもたらされる社会便益は社会コストを大幅に上回るという評価結果が出されたので、適用地域を拡大すべきであろう。

なお、本実験によって、「その他の深刻な犯罪（殺人、強姦、暴行）」の再発防止に関しては、別の対策が必要なのだろうということが示唆されたと言える。

5．議論

よく言われるように、**多数決という方法はだれも否定できないが、多数決が事実を明らかにするわけではない**。この例では、まず最初に RCT による実験をしてどの種類の犯罪に効果があるのか明らかにしてから、適用範囲を拡大すべきかどうかを議論して意思決定している。日本では、事実を明らかにせずに、感情に訴える主張が戦わされて、多数決で押し切って本当に実施してしまう場合が見られる。日本でもまずは事実を明らかにする努力がなされるべきである。**エビデンスを明らかにすることは民主主義のため**なのである。

もう一つ言えることは、出所者をコインの裏と表で分けることは倫理に反するという指摘をいただくことがある。これに対する反論は、**効果があるかないかわからないまま全国で適用する方が**

よっぽど被害は大きくなりそれこそ倫理に反するということである。「ゆとり教育」が好例で、いつの間にか始まっていつの間にか終わっていた。その実施期間中にエビデンスと呼べるものを見たことはなく、エビデンスが示されないまま全国の学齢期の子供たち全員が影響を受けたわけである。これは社会実験に反対する人たちがよく理解すべき点である。

(本事例の出所)

（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.'*.

.（文献 2) Greenberg, D. and Shroder, M.,(1997). *The Digest of Social Experiments 2nd edition,* Urban, Institute Press. Pp.217-219. 及び佐々木亮（2003)「政策評価トレーニング・ブック」多賀出版に掲載された記述を参照して加筆及び変更した。

（注）表中の最上段の数字（48 人、66 人）は文献２に明記されている。これは文献１に記載のあるサンプル数と比率から逆算した人数と一致しており、表中の後の２段の数字も同様の計算によって得た数字を記載した。

実験デザイン（RCT）の適用事例２：
出席日数を増加させるには？：小学校における回虫駆除プロジェクト（ケニア）

１．問題の所在

子どもが毎日学校に通うことは、どんな教育効果を考えるにしても最低限必要な条件である。従来は、保護者の意識向上、無料給食の実施、綺麗な校舎への建て替えなど、教育セクターの枠内で対策が考えられてきた。しかしここで少し視点を変えて、寄生虫駆除薬の配布と服用という保健セクターの対策がじつは効果があるという提案がなされた。

２．施策と評価デザインの概要

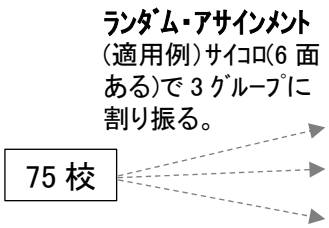
さっそくRCTを適用した評価が実施された。本プロジェクトはケニアのブシア県において、小学生に回虫駆除薬の投与と関連教育を行うことを介入内容として、1998-2002年に実施された。

表1　寄生虫駆除薬に関するRCT適用の概要

対象地域	ケニア・ブシア県
サンプル	ブシア県の75小学校（生徒数　約30,000人）
介入行為	回虫駆除薬の配布。オランダのNGOのInternational Christelijk Steunfounds Africa (ICS)のプロジェクトとして実施された配布を評価した。

同県の75校を、ランダム（無作為）に25校ずつ3つのグループに分けて、以下の年に介入を実施した（本来は3年連続の予定であったが、洪水が発生したので、20000年の分を2001年にずらして実施した)。このように時期をずらして実施することによって結局全ての学校が介入を適用されることになり、通常、実験デザイン（RCT）に関して指摘される倫理的な問題を回避している。

表2　各グループの介入実施年



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

（注）◎は介入実施を表す

３．評価結果

第１年次修了時（1998年末）に時点では、G１を介入グループ、G2を比較グループとして比較できる（G3も比較グループとして利用できるが省略)。同時点のG1（介入グループ）の回虫

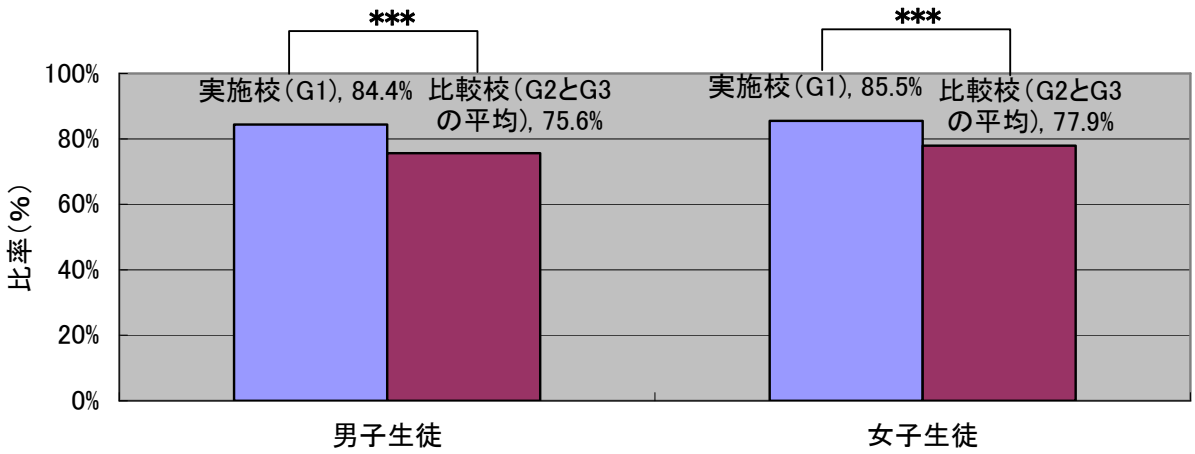
感染率が27%で、G2（比較グループ）の同感染率が52%だったので、その差である-25%が介入の効果であると判断できる。

表3 回虫感染率

	1998年末
G1（25校）：介入グループ	27%
G2（25校）：比較グループ	52%
差（介入の効果）	－25%

さらに、第1年次終了時（１９９８年末）には、寄生虫駆除薬の配布と服用により生徒の欠席日数が約3分の１も減少した（下図の男子の場合-36.1%(=(75.6%-84.4%)/(100%-75.6%)。女子の場合-34.4%(=(77.9-85.5%)/(100%-77.9%))。これを小学校に入学してから卒業するまでの期間に換算するとほぼ１年間分の増加となるほど大きな効果が確認された。また、生徒１人当たり１年間の費用はわずか50セント（=約50円）であり、通常の教育セクターの対策よりも格段に安いと評価された。

図 1 第 1 年次終了時点（1998 年末）の出席率（学校レベル）



有意水準：***1%, **5%, *10%

4．結論

回虫駆除薬を投与するという施策は、出席日数を増加させるという教育面での効果があるだけではなく、その介入費用は伝統的な教育施策よりも格段に安いと結論された。

このインパクト評価の事例の情報は広く世界で共有された。生徒の出席日数を増加させるために、**ケニア、ナイジェリア、エチオピア、インド、ベトナムで国家レベルの政策として採用された。この政策の導入により、世界中で3億人の児童が裨益した**と報告されている(「貧困アクションラボ」のPolicy Actionに掲載された報告による)。**一件の社会実験が世界の教育政策を変えた好例**である。

５．議論

あるセクターの専門家はそのセクターの中でしか考えることができないことが多い。しかし他のセクターの専門家から斬新な解決策が提案されることがあるという事例である。斬新な解決策が提案されたときに、「このセクターについて何も知らない人のたわごとだ」「私こそこのセクターの専門家だ」と言っているだけでは、業界の既得権益を防衛していることにしかならない。**有効かどうかは、どちらがより専門家なのか（つまりどちらが「より深く知っている」か）ではなく、実験によって得られた客観的なエビデンスによって決まる**のだ。それを明確に示した事例であると言える。

(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)

追加の情報

さらに本実験開始から20年後に、20年間の費用便益分析も実施された。

LPS(Life Panel Survey)（生涯追跡調査）をケニア政府が始めて、一人一人のその後の生活状況の調査が行われるようになったのでそのデータを利用した。一点留意すべきことは、純粋な統制群（=非介入群）はすでに存在しないので、G1とG2を介入群、G3を比較群として20年間の生涯追跡調査のデータを比較分析した。

その結果、駆虫により学校欠席が 25% 減少し、対照校への波及効果や、消費支出、時給、都市部に住む可能性 (生活の質向上の代用) が増加したことがわかった。 さらに私たちの 2021 年の論文（*）で評価されているように、これらの利点は、大規模に配布される介入の低コスト (子供 1 人あたり年間 0.50 ドル未満) と合わせて考慮すると、駆虫による収益の増加は、年率換算の**社会的内部収益率（Economic Internal Rate of Return）でEIRR=37%**が推定された。これは著しく高い数字である。

* Hamory, J, E Miguel, M W Walker, M Kremer, S J Baird (2021), “Twenty Year Economic Impacts of Deworming,” Proceedings of the National Academy of Sciences, July, 2024.

（出所）Miguel,E., Ochieng, E. & Walker, M. (2023) Deworming improves lives across generations https://voxdev.org/topic/health/deworming-improves-lives-across-generations

実験デザイン（RCT）の適用事例３：
パンデミック期間中の SMS（ショートメッセージ）と電話による学びの実現：
ボツワナ初等教育におけるローテク支援に関する迅速な RCT による検証

1. 問題の所在

新型コロナ（COVID-19）によるパンデミックは、世界中で教育システムを麻痺させた。ある調査によると、16 億人以上の生徒が、学校から隔離されたとされる（ユネスコ 2020）。このように悪化した学習環境に対処するため、費用対効果の高い方法で世界規模で子供たちの学びを改善できるアプローチが必要であった。

ボツワナでは、早くから予防的な社会的距離措置を導入していた。教育セクターでは、最初の措置として 2000 年 3 月 20 日から 6 か月間学校が閉鎖されており、教育への影響は深刻である。なお同国の初等教育（小学校）の純就学率は約 90％（ユネスコ 2014）と高いが、学習のレベルは低いとされる。

また、新型コロナの問題を別にしても、今までも、他のウイルス（インフルエンザやエボラ出血熱）、教師によるストライキ、地震や自然災害によって世界中で学校の閉鎖が引き起こされてきた。これらへも適用できる一般的なアプローチが必要であると言える。

なお、中低所得国ではインターネットへのアクセスが 15-60 パーセントの家庭に限られる一方で、70-90 パーセントの家庭で少なくとも1台の携帯電話を保有しているという調査がある（Center for Global Development 2020）。この携帯電話を利用することを考える。

2. 介入行為

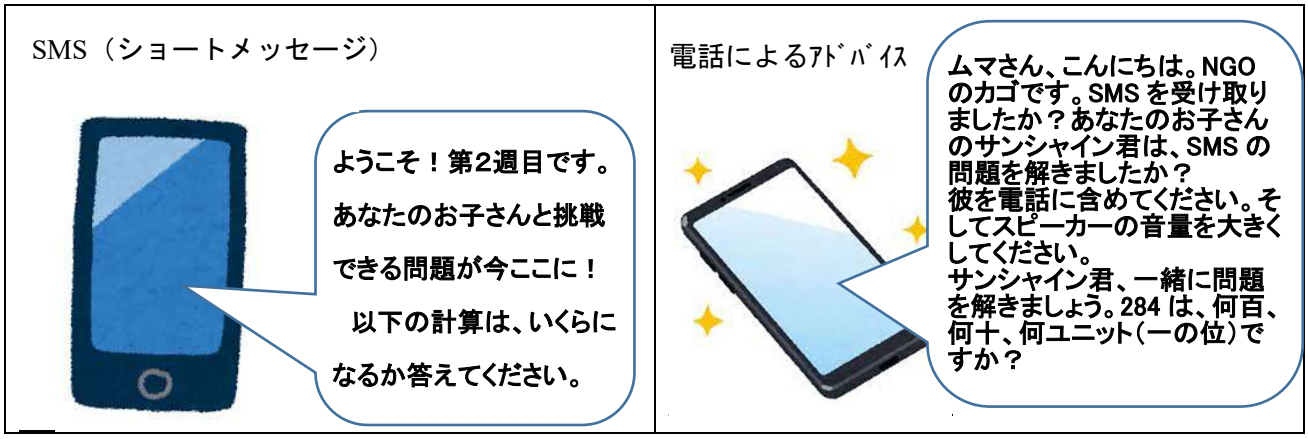
ボツワナ政府が「緊急事態宣言」を出す数日前に、調査チームは小学校から 7,550 の電話番号を入手した。これは、ボツワナにおいて、同国の教育省と協働している活動的な NGO「Young love」が集めていたものである。これらは、同国のほぼ全ての小学校の 3-5 年生から集められていた。

調査チームの 60 人のファシリテーターがそれぞれの番号に電話して、「携帯電話によるリモート支援」（remote learning support via phone）を受けたいかどうかの意思を確認した。ファシリテーターはワッツアップ（WhatsApp）で、どのように発言するかのインストラクションを受け取ってからそれに従って電話した。

「携帯電話によるリモート支援」は、2種類で、(a) SMS のテキストメッセージによる算数の問題の送付、(b)電話による 15-20 分間の生のアドバイスである。なお、親からSMSを送信したり電話してもらうことはその家庭に費用がかかるので、ファシリテーターから SMS を送ったり電話をした。(a)(b)とも、インターネットを利用した高度なウェブサイトによる指導に比べて、「ローテク」(low technology)と言えるだろう。しかしローテクだからこそ、どんな家庭にも届く可能性があると考えた。

1つ目の介入（SMS）	毎週、いくつかのシンプルな算数の問題のSMS を送る。子供自身が携帯電話を持っていることは稀なので、SMSは親の携帯電話に送られた。親はそのまま子供に見せる場合と、一緒になって教えて問題を解く場合があったことが分かっている（いずれも好ましいい）。SMS による解答の返信は求めなかった。あとで答えが送信されたと理解される。
2つ目の介入（電話によるアドバイス）	毎週、週の最初に 15-20 分間、携帯電話でファシリテーターが親に口頭でアドバイスした。電話するたびに、親に子供も一緒にアドバイスを聞くように依頼した。のちの報告によると、子供が算数の問題を解くことができたことに、親としてプライドを感じたと回答している。

46 人のファシリテーターが、それぞれ 24 人の親を担当した。それぞれのファシリテーターは、一日だいたい 6 時間電話した。ファシリテーターは、いつの時間帯が一番都合がいいか定期的に親に質問することにした。50%以上が、家事が終わったあとか、家事をしている時間が都合がいいと回答している。調査の本格的な実施の前に、2週間のトライアルを実施してノウハウを得た。

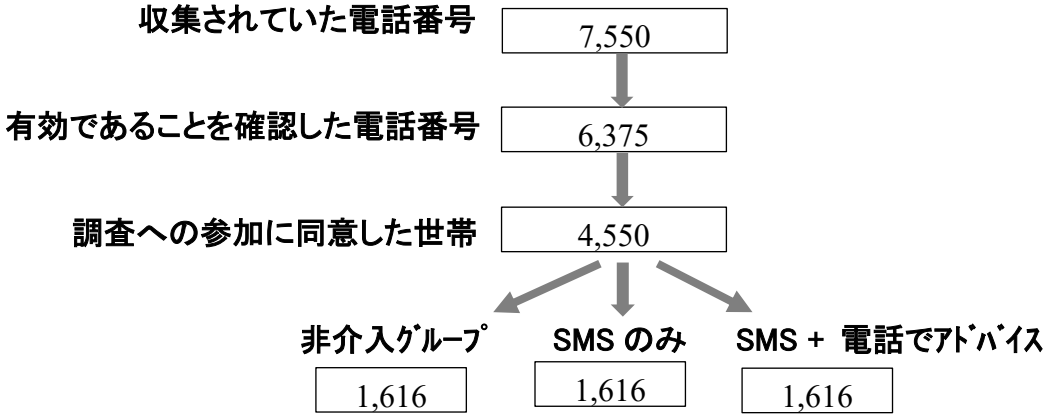


（出所）原典の報告書の p6 のイラストをもとに。筆者が和訳した。

3. 割り付け

ボツワナ全国の小学校から得られた 7,550 の電話番号に電話して、6,375 人に依頼できた。残りは、番号が有効ではない、応答がない、引っ越したなどであった。そして、調査への参加の意思と同意（consent to participate）を依頼して、およそ 71%の 4,550 人が同意した。調査チームは、4,550 人を、ランダムアサインメント（無作為割り付け）によって、同じサンプル数の3つのグループになるように割り当てた。なお、4,550 人を、「子供が一人か」「複数か」で2層に分けて、それぞれの層を3つに割り付ける（stratified）ことにより、それぞれの層の比率が 3 つのグループで同じになるようにした。ここが工夫した点である。

そして4週間経ったときに子供に算数テスト（第一回目）を実施した。その後、正解した子供は次のレベルの問題に進ませた。正解しなかった子供は、同じレベルの問題に留め置いた。10 週間後に、また算数テスト（第二回）を実施して、この「追加的な介入」の効果も測定した（分析はまだである）。



（注）第1段階のデータ収集は3つのグループのそれぞれ半数ずつから、第2段階のデータ収集は全ての世帯から実施した。

（出所） 原典の報告書の p9 のイラストをもとに。筆者が和訳した。

4. データ収集

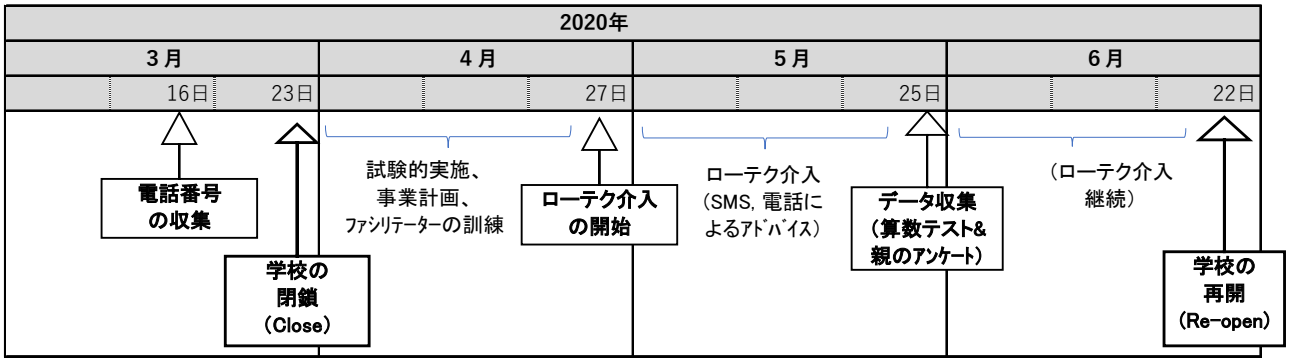
算数テストは、ASERテストを使った。（ASER：Annual Status of Education Report の略とのこと）。ASER テストは、足し算（レベル1）、引き算（レベル2）、掛け算（レベル3）、割り算（レベル4）で構成されている。以下がそのサンプルである。子供に携帯電話で回答してもらったが、側にいる親の介入を防ぐために、1）子供はそれぞれの問題を2分間で解く、2）子供自身に質問して計算の過程を自分で答えられたもののみ正解とした。もちろん完全とは言えないが、最善を尽くした。ASERテストの結果は、0-4 の 5 段階のいずれかである。

介入実施前にASER テストを実施してベースライン値は取っていないが、適切に無作為割り付けしているので、言う

までもなく一致しているはずである。

なお、RCT の計画からエンドラインのデータ収集（算数テストと親へのアンケート）までわずか 2 ヵ月間で終了していることは注目に値する。

RCT の準備・計画・実施と、データ収集・分析の時系列表



（出所）原典の報告書の p9 をもとに筆者が作成した。

5. 介入効果（インパクト）の分析結果

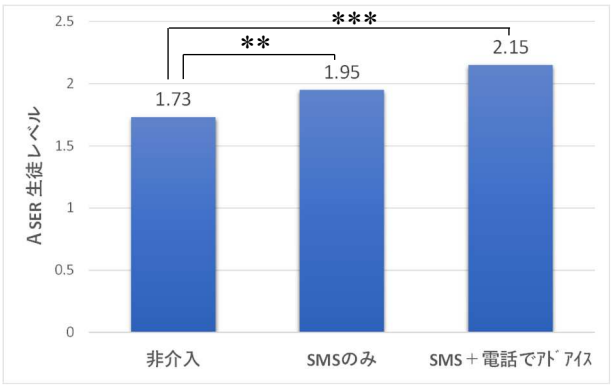
調査チームは、第一段階の介入のあと（＝4週間後）、子供の学びについて大きなインパクトを確認した。

(1) SMS＋電話でのアドバイスのグループで、非介入グループと比べて、24 パーセントの学びの向上があった[4]。

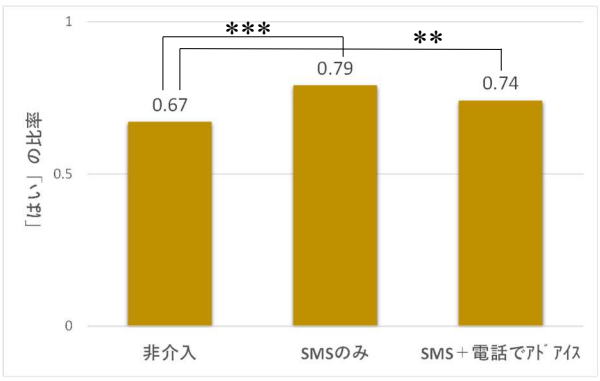
(2) SMS だけのグループは、非介入グループに比べて 13 パーセントの改善であった。それは統計学的に有意であった[5]。

(3) さらに、親に、子供と一緒に関わったか（engaging）したかどうか聞いた。SMS＋電話でのアドバイスのグループでは、非介入グループに比べて7パーセントポイント多く一緒に関わっていた。一方、SMS だけのグループは、非介入グループに比べて 12 パーセントポイント多くかかわっていた。SMS だけの方が親の関わりが多くなることが示されている。

子どもの算数レベル（0～4 段階）



親の関わり（「少なとも時々関わる」=「はい」の比率）



（注）*1%有意、**5%有意、*10%有意**

（出所）原典の報告書の p19, p23, p31 の図表をもとに筆者が作成した。

6. 費用対効果の分析結果

[4] この大きさは、非介入グループの標準偏差の幅1に対して 0.29 であった。いわゆるエフェクトサイズ(Effect size)であり、0.2＝小、0.4=中、0.6 以上=大のインパクトと判断されるものである（Cohen, 1986）。このケースでは、「小～中」にあたる。なお、このように非介入グループの標準偏差を割り算の分母に使うことを Glass, V.は推奨している。

[5] 同じ計算で、エフェクトサイズは 0.16（小）であった。

調査チームは、費用の上限を計算した。その費用は、事業の総費用、関わった個人の時間、(この調査の前にNGOが)電話番号を集めた費用、インフラ、ファシリテーターの訓練の時間、定期的なテスト&親へのアンケートの実施のコストである。SMS だけの場合の総額は 3,200 ドル(約 35 万円)と計算され、子供ひとりあたり $ 2.13 ドル(約 3 ドル)となる。電話でのアドバイスは、$17,800(約 200 万円)であり、一人あたり 14 ドル(約 1600 円)となる。これは、子供一人当たり 1 標準偏差分を向上させるのに、$13.3 ドル(約 1,500 円) と 48.28 ドル(約 5,500 円)となる。なお、1段階はそれぞれ、何もできない("beginner"、レベル 0)、足し算(レベル1)、引き算(レベル2)、掛け算(レベル3)、割り算(レベル4)にほぼ対応しており、それぞれ1段階を引き上げるのに要する費用と解釈することができる。

7. 政策への提言

調査結果は、**SMS と電話によるアドバイスというローテクの介入が、費用対効果に優れ、学校が閉鎖されている期間に、短期的に学びを改善する**ことを示した。調査チームは、ターゲットを絞った電話という次の段階の介入のインパクトも今後分析する予定だし、さらに長期間の影響についても分析する予定である(いずれもまだ実施していない)。

8. 議 論

調査結果は、今までの常識に挑戦する次の3つのことを示唆している。

(1) 薬の効果を確認するのにも使われる、最も厳格なデザインである「ランダム化比較試験」(RCT)は、1年半から2年くらいかかると一般には思われている。しかし、**この事例ではわずか2ヶ月で終了**して、その後評価結果を公表している。じつはRCTはそれほど時間をかけずに介入効果(インパクト)を評価することができることが分かる。日本の公務員の任期の1年目にRCTを実施して、2 年目の政策に利用して自分でインパクトを見ることができる。RCT はもっと気軽に迅速に使われるべきであろう。また、この調査で用いられた程度の統計分析のレベルで十分である。

(2) インターネットを使って遠隔教育というと、1)ウェブサイトを作って、2)ビデオ録画をして貼り付けて、3)回答をクリックできる練習問題を用意して、4)タブレットで見てもらう、ということを考えるであろう。ただしそれは先進国の援助機関発の発想であり、教員対象ならまだしも、個別の家庭レベルでインターネットが引かれていてタブレットやパソコンがある家庭は一般的ではない。一方で、携帯電話なら 70-90%の家庭に普及しているという調査結果があり(冒頭で紹介したとおり)、**ハイテクな遠隔教育ではなくローテクな SMS と電話で学びを支援して教育効果(インパクト)を出せる**ことがわかる。SMS の送信と電話をかける料金はもちろん援助機関持ちである。

(3) 教育効果から考えると、各段に費用が安い。既存の教育事業および援助事業の場合のコストは明示されていないのは、相手政府の教育省や援助機関に気を使っているのかも知れないが、同一の教育効果にかかるコストを計算すると各段に安いことが分かってしまうであろう。教育はテストの点数だけでは計れない、というのはもっともな意見であるが、携帯電話を使って新しい道徳心や新しい倫理観や新しい他人との協力関係を教えることを考えた方がいいかも知れない。そして、これからの世界ではそちらの方が適用範囲が広いかも知れない。日本の教育分野では、**「生きる力」が重要と言うが、「新世代の生きる力」を考える瞬間に教育関係者は立ち会っているのかも知れない。Survivability から New Age Survivability である。** やや決まりすぎの感があるが最後に直言しておきたい。

(出所)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*. Centre 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

実験デザイン（RCT）の適用事例4：
ベーシックインカムは効果があるのか？（フィンランド）

1. 問題の所在

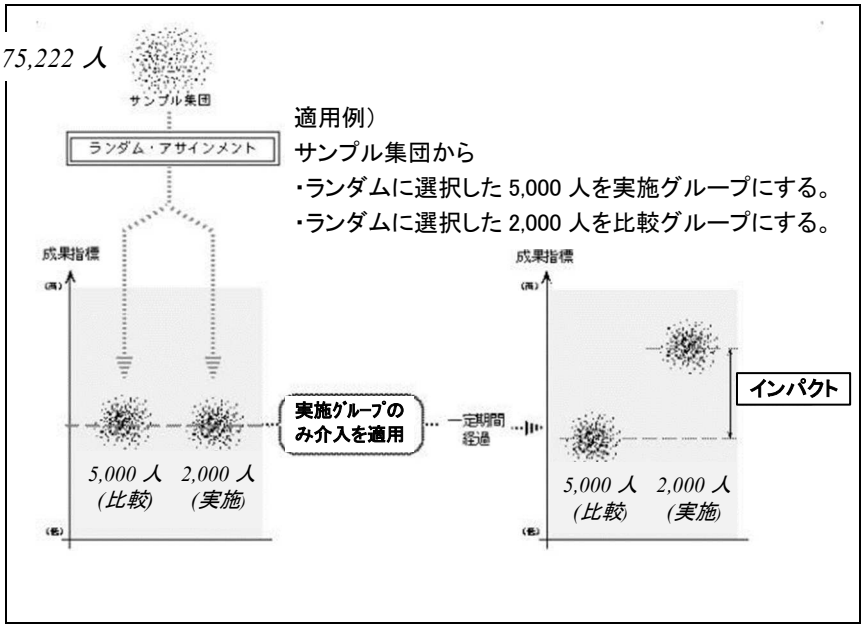
ベーシックインカムは、政府がすべての国民に対して定期的に一定の額の現金を支給するという施策である。これにより、国民が最低限の生活を送れるようになるとされる。一方で、働かなくても現金がもらえるなら人々は働かなくなるはずで、結果的に国全体が貧しくなるという主張も聞かれる。ベーシックインカムという施策は、ITの普及により人間の仕事がなくなって失業者が増大するのではないかという危機への対応策として語られることも多い。実際のところ、ベーシックインカムによって人々の生活はどう変わるのだろうか？

2. 施策と評価デザインの概要

2017-2018年度にフィンランドでRCTを適用した大規模な評価が実施された。サンプルと介入行為の概要は以下のとおり。

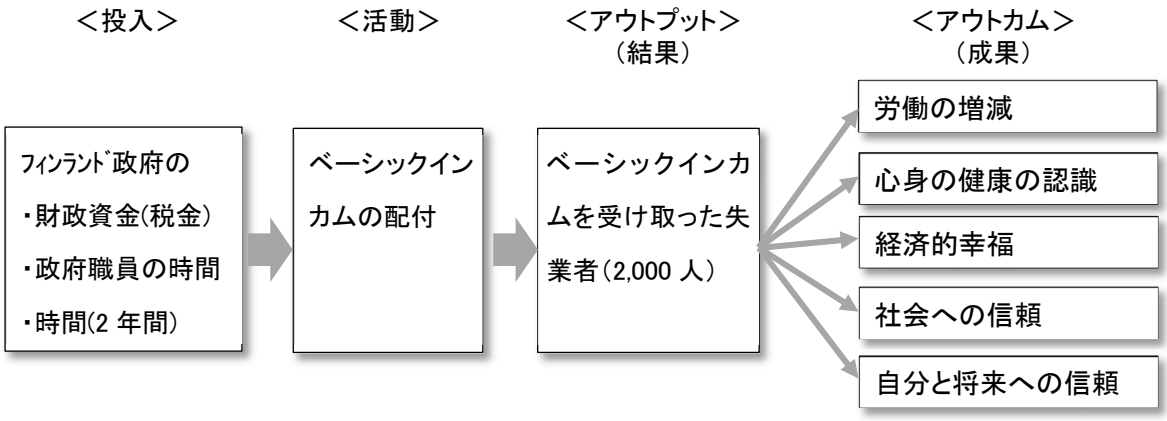
表1　ベーシックインカムに関するRCT適用の概要

対象地域	フィンランド
サンプル	労働市場に登録のあった失業者（25-58 歳）の人口（175,222 人）。 介入グループ：その人口からランダムに選択した 2,000 人。 比較グループ：残りの人口（173,222 人）から 5,000 人。 介入グループには全員に 2 年間ベーシックインカムを給付する。一方。比較グループには、従来の失業施策を継続する。
実施年	2017 年 1 月 1 日-2018 年最終日。2017 年にベースライン調査、2018 年にエンドライン調査。電話インタビューを実施した。回答率は 23.2%（介入群 31.3%、比較グループ 20.2%）。
介入行為	フィンランド政府がベーシックインカムとして毎月 560 ユーロ（€）を支給する。日本円で毎月約 75,600 円（1€＝135 円換算）。



３．評価結果

ベーシックインカムによって実現するアウトカム（成果）としてフィンランド政府は以下の項目を設定した。



（出所）フィンランド政府の報告書の目次をもとに筆者が作成した。

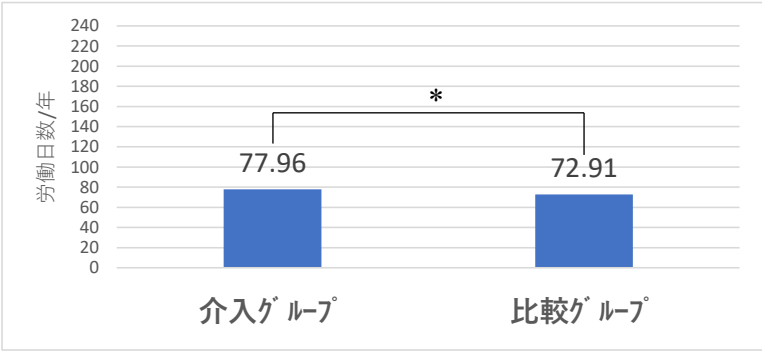
（1） 労働の増減（Employment effects）

ベーシックインカムの配付によって、労働日数は5.05日増大した（表1参照）。さらに、個人の属性の影響を取り除くと6.03日の増大と計算された。ベーシックインカムによって人々は働かなくなるという懸念は当たらず、逆に6～8%ほど労働日数は増加した。ただし、その年に従来の失業施策の変更があったのでその影響も受けた可能性があると述べている。

表1：労働日数への影響（2017年11月1日から2018年10月31日まで（実験の2年目にあたる））

分析の方法	介入ｸﾞﾙｰﾌﾟ	比較ｸﾞﾙｰﾌﾟ	差	標準誤差	P値*
2群の有意差検定	77.96日	72.91日	5.05日	2.84日	0.08
回帰分析（個人の属性の影響**を取り除いたあとの値）		72.91日	6.03日	2.52日	0.02

* p値が0.1 (=10%)未満、0.05 (=5%)未満、あるいは0.01 (=1%)未満だと、両グループの差は偶然では起こり得ないほど大きな差だということになる。
**個人の属性とは。性別、年齢、教育のレベルと分野、母国語、家族種類、診断された病気、市町村のグループ、居住地域、失業給付の種類、失業日数、就業日、仕事と援助からの収入。
（出所）フィンランド政府の報告書p39の表1に、本文の解説を追加して作成した



＊…10 ％ 水準で有意

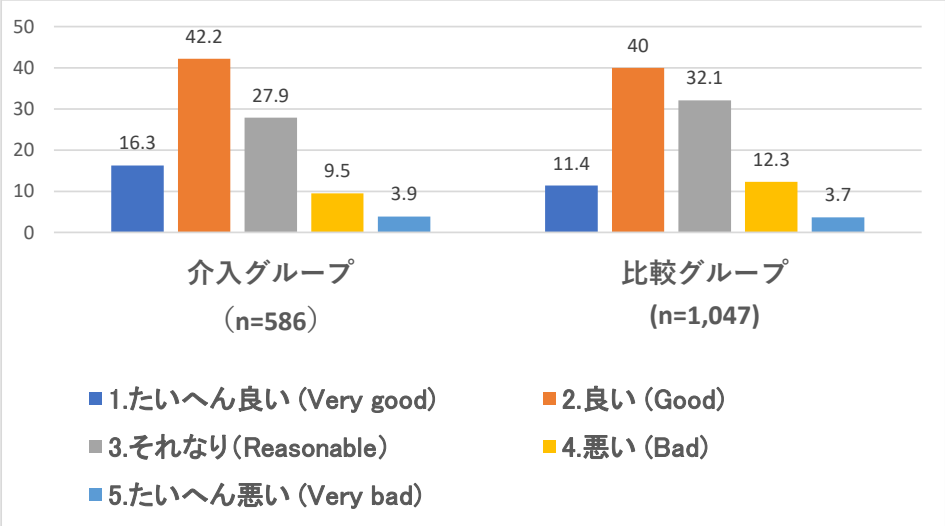
（2） 心身の健康の認識（Perceived health, mental well-being and cognitive performance）

ベーシックインカムの配付によって、健康状態の認識が改善した（表2参照）。さらに、1.病気、身体的疾病、精神障害が日常生活に及ぼす影響の減少、2.保健サービスの利用の減少、3.精神的ストレスの減少、4.認知機能の向上、5.孤独の減少も確認された（追加説明1の別添表1を参照）。

表2：健康状態に関する自己評価

健康状態の認識	介入グループ（n=586）	比較グループ(n=1,047)
▲ 1.たいへん良い（Very good）	16.3	11.4
2.良い（Good）	42.2	40.0
3.それなり（Reasonable）	27.9	32.1
4.悪い（Bad）	9.5	12.3
▼ 5.たいへん悪い（Very bad）	3.9	3.7
		p=0.051

（注） p値が0.1 (=10%)未満、0.05 (=5%)未満、0.01(=1%)未満のいずれかだと、両グループの差は偶然では起こり得ないほど大きな差だということになる。



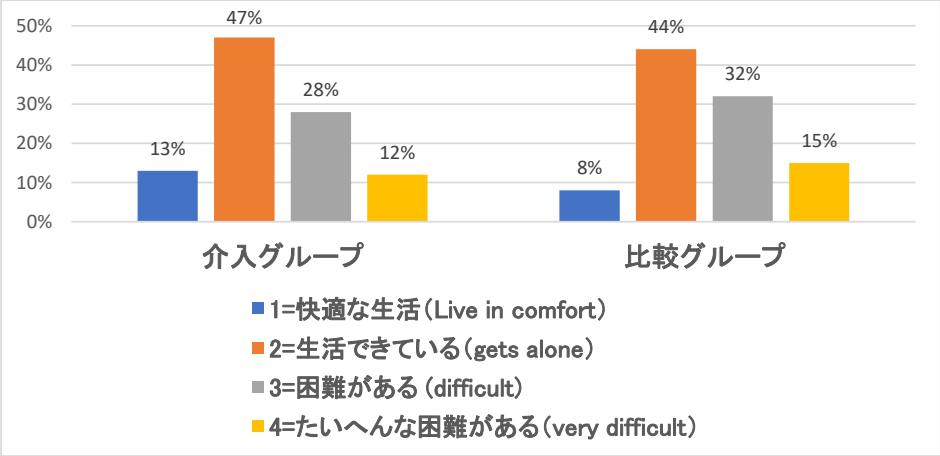
（3） 経済的幸福（Economic Well-being）

経済的幸福とは、(i) 責任をもって経済面の自己管理ができる、(ii)経済的に安定した状態である、(iii)労働生産性がある（と自覚できる）、という状態である。ベーシックインカムを受けた回答者は、経済的幸福のすべての領域で、収入レベルと経済的幸福感が良いと回答した（表3、および追加説明2の別添表2も参照）。

表3：あなたはあなたの世帯の収入で生活できているか

選択肢 / グループ	1=快適な生活（Live in comfort）	2=生活できている（gets alone）	3=困難がある（difficult）	4=たいへんな困難がある（very difficult）
介入グループ	13%	47%	28%	12%
比較グループ	8%	44%	32%	15%

（注） p 値は表示されていない。（出所） フィンランド政府の報告書5.3の記載から作成した。



（4） 社会への信頼、自分と将来への信頼 （Trust on institutions and confidence in oneself ）

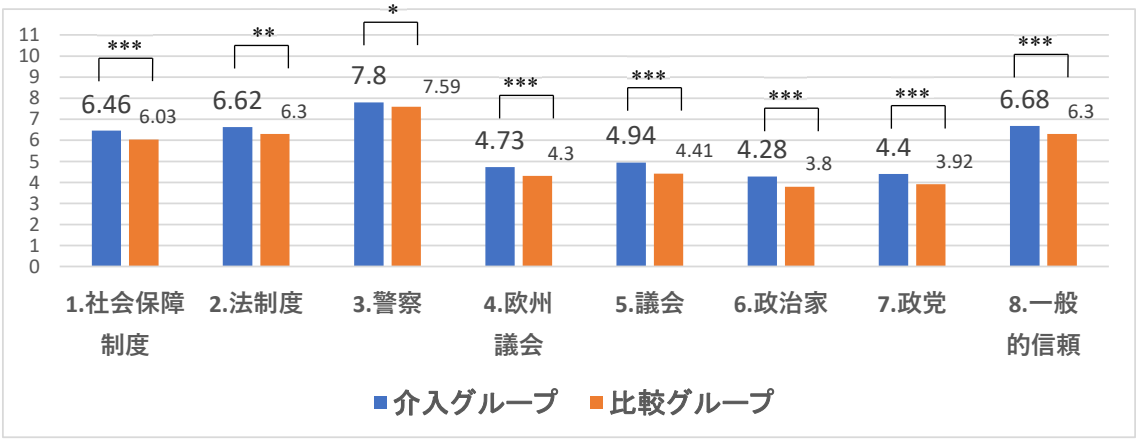
「信頼」がなければ、人生はカオスとなり、私たちは他の人々と有意義な社会的な相互作用を行うことができなくなる。社会機構への信頼は、介入グループの平均値が比較グループの平均値よりも優位に高くなっていることが分かる（表4参照）。同様に、自分自身と将来への信頼は、介入グループの平均値が比較グループの平均値よりも有意に高くなっていることが分かる（追加説明3の別添表3を参照）。

表4：組織への信頼と一般的な信頼（平均値）

グループ / 信頼の対象	介入グループの平均値	比較グループの平均値	両グループの差とｐ値
1.社会保障制度	6.46	6.03	+0.43　p =.001
2.法制度	6.62	6.30	+0.32　p =.018
3.警察	7.80	7.59	+0.21　p =.079
4.欧州議会	4.73	4.30	+0.43　p =.004
5.フィンランド議会	4.94	4.41	+0.53　p =.000
6.政治家	4.28	3.80	+0.48　p =.001
7.政党	4.40	3.92	+0.48　p =.001
8.一般的信頼	6.68	6.30	+0.38　p =.003

（注 1）選択肢は 11 段階：「10＝ 極めて信頼できる」～「0＝ 極めて信頼できない」

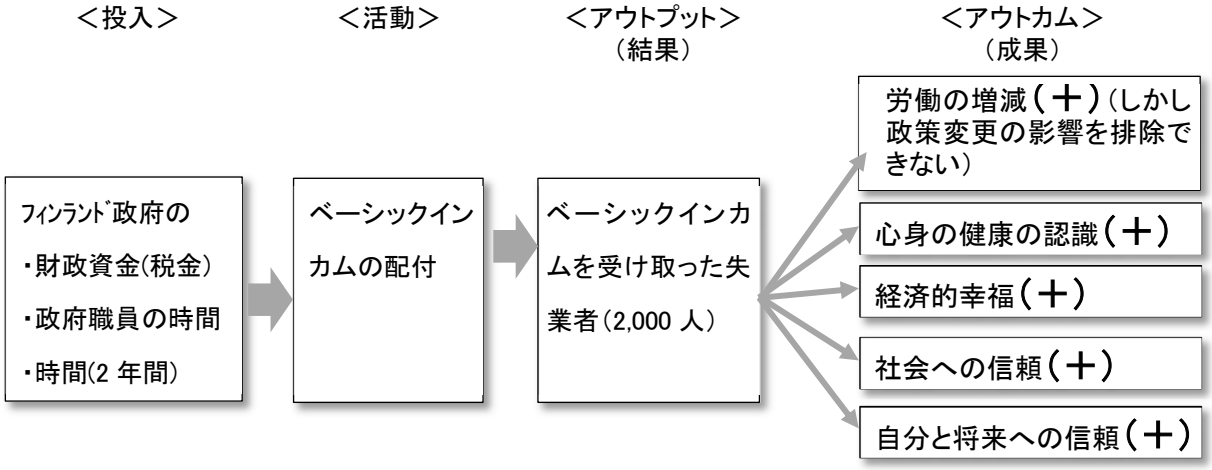
（注2） p値が0.1 (=10%)未満、0.05 (=5%)未満、0.01(=1%)未満のいずれかだと、両グループの差は偶然では起こり得ないほど大きな差だということになる。



***, **, *…1%, 5%, 10%水準で有意

4．結論

ベーシックインカムによって、失業者の福祉は増大したと結論された。また、ベーシックインカムによって労働日数が増加したという結果となったが、実験の途中で従来の失業施策の変更があってその影響を排除できないので明確なことは言えないとしている。しかし少なくとも人々は怠け者（lazy）になるとの懸念は肯定されなかったと言える。



（注）「＋」は改善を示す。「－」は悪化を示す。
（出所）フィンランド政府の報告書の目次をもとに筆者が作成した。

5．議論

この報告書を報じた各国の新聞記事では、結局のところ、「ベーシックインカムが良いとも悪いとも言えなかった」という結論がしばしば見られた。劇的な効果があったとは言えないが、懸念されたほど悪影響があったわけでもなかった。

日本でベーシックインカムが議論されるときには、失業によって最低限の生活費もなくなった人たちを救う政策であり、ITの時代には誰しも失業者になる可能性があるのだ、という文脈で語られることが多い。つまり、最も貧しい脆弱な層にスポットを当てて、ベーシックインカムはその人たちを救うための政策であり、そしてあなたもそうなる可能性がある、と語られることが多い。一方で、それは自己責任であり本人の責任だ、不幸な境遇に負けずにがんばっている人たちだってごまんといるではないか、という話もされる。

フィンランドのベーシックインカムの評価から分かることは、お金の話だけではなく、人々の心身の健康、幸せの度合、自己肯定感、社会への信頼感という効果も併せて見ているということである。**フィンランドの人々の人生観が現れている**のと言える。日本でベーシックインカムの議論をする場合にも、**日本の人々の人生観をまずは考えるべき**というメッセージとして受け取ることができるのかも知れない。

（注1）本報告は、フィンランド政府が実施して公表した報告書（フィンランド語）をGoogle翻訳で英語に翻訳したものを参照している。その際の翻訳が必ずしも正確ではない可能性があることをあらかじめ明記する。
（注2）フィンランド政府の報告書では、数量的なデータの分析だけではなく、参加者への個別のインタビュー結果も分析している。定量手法と定性手法を組み合わせるという「混合手法」（mixed methods）を用いていると言える。ただし、混合手法を用いると結論が曖昧になりやすいという批判があり得る。また、結論が肯定的になりやすいという批判もあり得る。

（出所）

（フィンランド語の報告書）Olli Kangas, Signe Jauhiainen, Miska Simanainen and Minna Ylikännö (eds.). (2020).*Evaluering av basinkomstexperiment.* Social- och hälsovårdsministeriet.

（上記報告書のタイトルの英訳）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

事例の別添（追加説明1、2、3）

追加説明1：「心身の健康の認識」の追加説明

本文で説明したとおり、ベーシックインカムの配付によって、健康状態の認識が改善した。さらに以下の表に示された通り、1.病気、身体的疾病、精神障害が日常生活に及ぼす影響の減少、2.保健サービスの利用の減少、3.精神的ストレスの減少、4.認知機能の向上、5.孤独の減少も確認された。

別添表1：その他の心身の状態に関する自己評価

サーベイの種類	説明	サーベイに使用した選択肢	結果とp値*
1. 有病の状態（病気、身体的疾病、精神障害）	病気、身体的疾病、精神障害が日常生活に及ぼす影響。	1 =はい、多大な影響がある。2 =多少の影響がある。3 =いいえ	影響の減少（p=0.051）
2.保健サービスの利用	看護師（Nurse）、保健センターの医者、巡回診療（houseman）、歯医者、その他の利用頻度。	1=0～2回の利用、2=3回以上の利用、3=言えない。（Nurseのみ有意な差	利用の減少（p=0.070）））
3.精神的ストレスの状態	過去1か月間に感じた緊張状態、回復困難、落ち着きと静けさ、うつ病、幸福感の認識。	1 =常時、2 =ほとんどの場合、3 =ある程度の時間、4 =少し時間、5 =まったくない	ストレスの減少（p=0.003～0.162）
4.認知機能の状態	精神的健康の1つの側面。記憶力、新しいことを学ぶ能力、物事に集中する能力など。	1 =非常に良い、2 =良い、3 =満足できる、4 =悪い、5 =非常に悪い。	認知機能の向上(p<0.001)
5.孤独を感じた経験	精神的健康の1つの側面。	1 =全くない/めったにない、2 =ときどき/しばしば、3 =継続的だ、4 =言えない。	孤独の減少(p=0.032)

* p値が0.1 (=10%)未満、0.05 (=5%)未満、0.01(=1%)未満のいずれかだと、両グループの差は偶然では起こり得ないほど大きな差だということになる。なお、統計検定は、2群のtテストあるいはカイ二乗検定。サンプルサイズは介入グループn=586、比較グループn=1,047。
（出所）フィンランド政府の報告書4.3の記載から作成した。

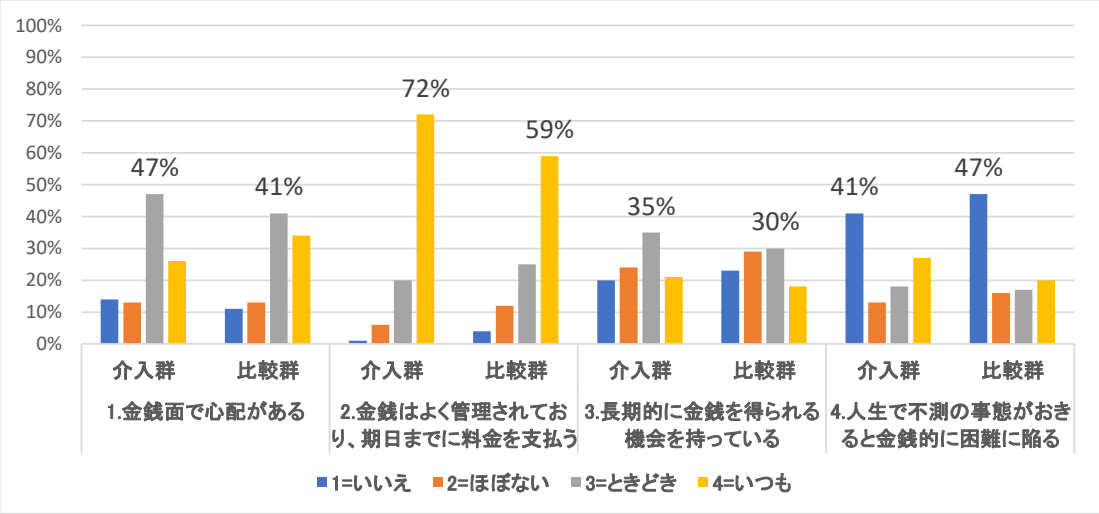
追加説明2：「経済的幸福」の追加説明

ベーシックインカムを受けた回答者は、経済的幸福のすべての領域で、収入レベルと経済的幸福感が良いと回答した。本文で解説した以外に、以下の指標についても質問して回答を得ている。

別添表2：経済的幸福への反応

選択肢 / 指標		1= いいえ	2= ほぼない	3=ときどき	4= いつも	カイ二乗検定のp値
1.金銭面で心配がある	介入グループ	14%	13%	47%	26%	減少
	比較グループ	11%	13%	41%	34%	(p=0.011)
2.金銭はよく管理されており、期日までに料金を支払う	介入グループ	1%	6%	20%	72%	増加
	比較グループ	4%	12%	25%	59%	(p=0.000)
3.長期的に金銭を得られる機会を持っている	介入グループ	20%	24%	35%	21%	増加
	比較グループ	23%	29%	30%	18%	(p=0.059)
4.人生で不測の事態がおきると金銭的に困難に陥る	介入グループ	41%	13%	18%	27%	減少
	比較グループ	47%	16%	17%	20%	(p=0.006)

（出所）フィンランド政府の報告書5.4の表から作成した。



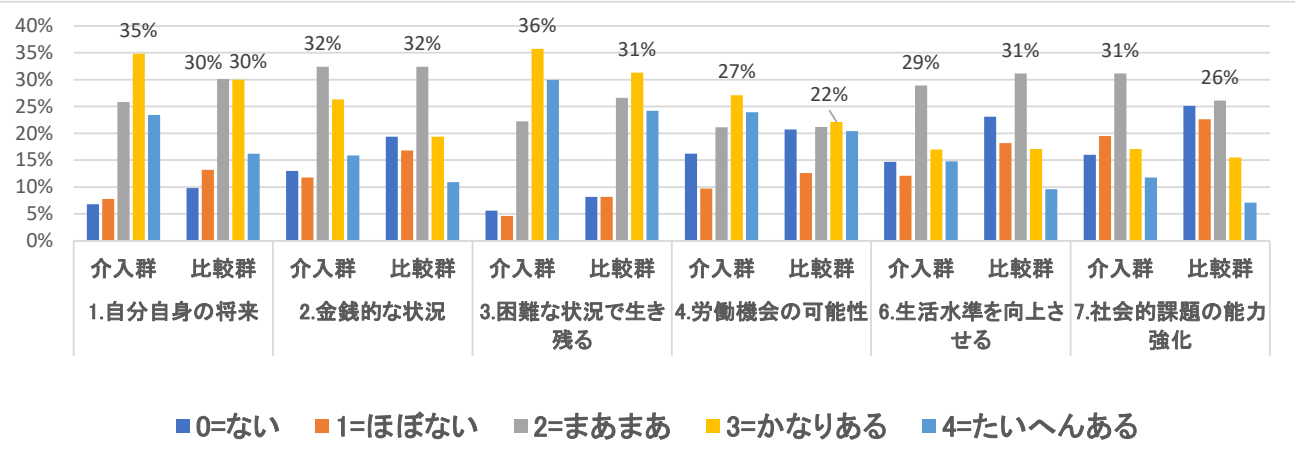
追加説明3:「社会への信頼」、「自分と将来への信頼」への追加説明

「信頼」がなければ、人生はカオスとなり、私たちは他の人々と有意義な社会的な相互作用を行うことができなくなる。本文で説明したとおり、社会機構への信頼は、介入グループの平均値が比較グループの平均値よりも優位に高くなっていることが分かる。さらに、自分自身と将来への信頼(以下の7項目)は、介入グループの平均値が比較グループの平均値よりも有意に高くなっていることが分かる。

別添表3:自分自身と将来への信頼

選択肢 / 指標		0=ない	1=ほぼない	2=まあまあ	3=かなりある	4=たいへんある	言えない	平均	カイ二乗検定のp値
1.自分自身の将来	介入ｸﾞﾙｰﾌﾟ	6.8%	7.8%	25.8%	34.8%	23.4%	1.4%	2.57	差は有意
	比較ｸﾞﾙｰﾌﾟ	9.8%	13.2%	30.1%	30.0%	16.2%	0.7%	2.28	p=0.000
2.金銭的な状況	介入ｸﾞﾙｰﾌﾟ	13.0%	11.8%	32.4%	26.3%	15.9%	0.7%	2.19	差は有意
	比較ｸﾞﾙｰﾌﾟ	19.4%	16.8%	32.4%	19.4%	10.9%	1.1%	1,83	(p=0.000)
3.困難な状況で生き残る	介入ｸﾞﾙｰﾌﾟ	5.6%	4.6%	22.2%	35.7%	29.9%	2.0%	2.76	差は有意
	比較ｸﾞﾙｰﾌﾟ	8.2%	8.2%	26.6%	31.3%	24.2%	1.4%	2.52	(p=0.01)
4.労働機会の可能性	介入ｸﾞﾙｰﾌﾟ	16.2%	9.7%	21.1%	27.1%	23.9%	2.9%	2.29	差は有意
	比較ｸﾞﾙｰﾌﾟ	20.7%	12.6%	21.2%	22.1%	20.4%	3.0%	2.03	(p=0.000)
6.生活水準を向上させる	介入ｸﾞﾙｰﾌﾟ	14.7%	12.1%	28.9%	17.0%	14.8%	6.0%	1.80	差は有意
	比較ｸﾞﾙｰﾌﾟ	23.1%	18.2%	31.1%	17.1%	9.6%	3.1%	1.70	(p=0.000)
7.社会的課題の能力強化	介入ｸﾞﾙｰﾌﾟ	16.0%	19.5%	31.1%	17.1%	11.8%	4.6%	1.80	差は有意
	比較ｸﾞﾙｰﾌﾟ	25.1%	22.6%	26.1%	15.5%	7.1%	3.2%	1.50	(p=0.000)

(出所) フィンランド政府の報告書7.5の表から作成した。「平均」は著者が独自に計算した。



実験デザイン（RCT）の適用事例５：
マイクロファイナスは奇跡か？（インド）

1. 問題の所在

マイクロファイナンスは、貧困削減の切り札として1970年代に登場し、その後急激に普及した。2007年12月の時点で1億5,486万人（うち女性が1億人以上）がサービスを受けていると発表されている（Microcredit Summit Campaign発表)。また、2006年には、グラミンバンクとその創設者のムハマド・ユヌス博士（Dr. Mohammad Yunus）がノーベル平和賞を受賞している。

一方で、マイクロファイナンスが貧困削減に本当に効果があるかどうかは論争が続いている。Pitt and Khandker (1998)は大きな効果があり、特に女性に効果があると結論している。一方で、Morduch (1999)、Rodman & Morduch(2009)は確たる証拠は確認されていないとして一貫して否定的である（高橋2011)。こうした論争に対して確かな証拠を提供すべく、もっとも厳格な手法であるRCTによる検証を行ったのが本例である。

2．施策と評価デザインの概要

対象地域、サンプル、介入行為は次のとおりである。

表1　マイクロファイナンスに関するRCT適用の概要

対象地域	インド、ハイデラバード（アンドラプラデッシュ州の州都）
サンプル	104地区（実施：52地区、比較：52地区）
実施年	2005ベースライン調査、2006-2007事業実施、2007年8月エンドライン調査
介入行為	グラミン銀行のグループ化貸付の手法を採用した Spandana という事業主体がマイクロファイナンス事業を実施。

対象地域は、インドのハイデラバード（アンドラプラデッシュ州の州都）で、同市から104地区を選定して1対1のマッチングを行って52組を形成した。その後、それぞれの組のなかでランダム・アサインメント（無作為割付）を行って、1地区を実施地区（融資実施)、別の1地区を対象地区（融資を実施しない）に分けた。これにより特徴を近似させた実施地区52地区と比較地区52地区を形成した。

融資資格は、(a) 女性、(b) 18-59歳、(c) 同じ地域に1年以上居住、(d)有効な身分証明書を持っている、(e)グループの 80%以上が自宅を所有していること。一方で、グラミン銀行のようにグループに対して訓練は行わない。融資額は10,000-12,000ルピーで金利は12%(年利24%と同等)。

2005 年にベースライン調査を実施して、両グループの経済的な平均値に差がないことを確認した。2006年から2007年にかけて、グラミン銀行のグループ化貸付の手法を採用したマイクロ

ファイナンス銀行である Spandana が融資事業を実施した。2007 年 8 月にエンドライン調査を実施して、両グループ間の指標群の差を測定した。

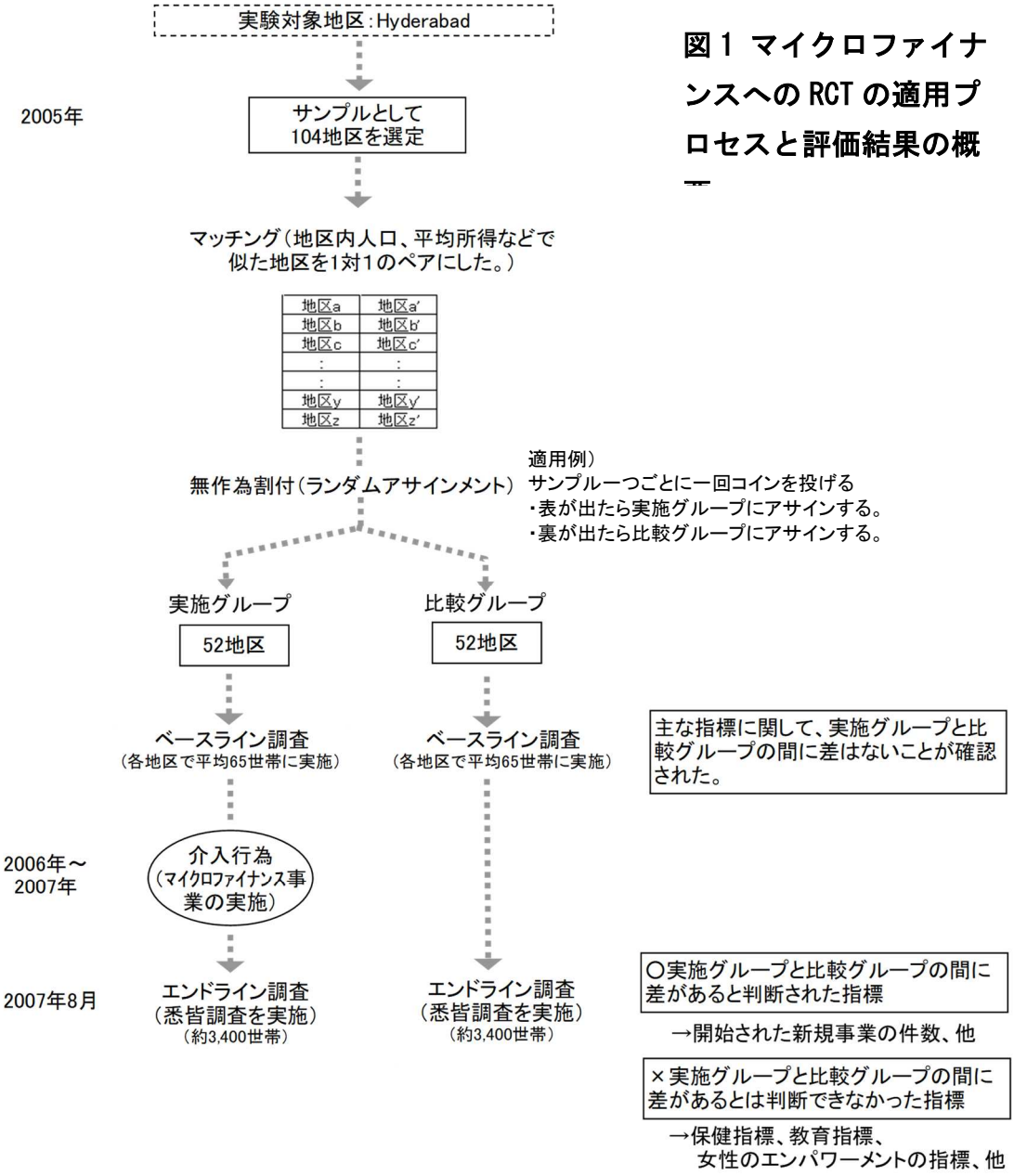
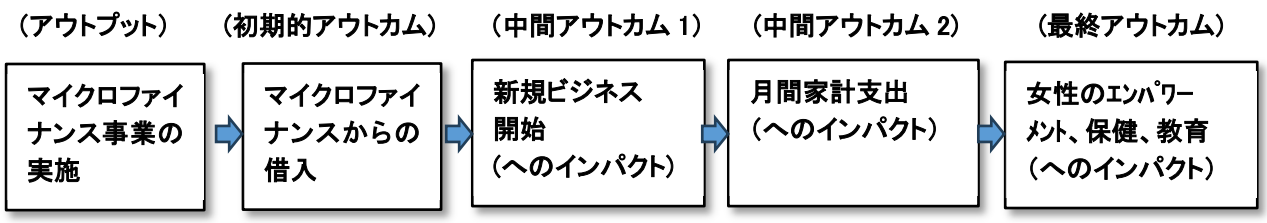


図 1 マイクロファイナンスへの RCT の適用プロセスと評価結果の概要

＜想定されたロジックモデル＞



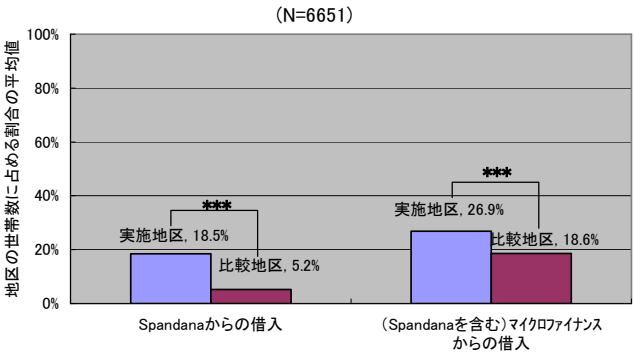
３．評価結果

本件の評価結果は次のとおりであった。（有意水準： ***1%, **5%, *10%）

(1)マイクロファイナンスからの借入（図2）

Spandana から融資を受けた世帯の率は実施地区（52 地区）が 18.5%、比較地区（52 地区）が 5.2%でその差 13.3%だった。比較地区の人もわずかながら Spandana に融資申請して融資を受けていることがわかる。また Spandana を含むマイクロファイナンス機関から融資を受けた世帯の率は実施地区が 26.9%、比較地区が 18.6%だった。したがって、**実施地区の方がより多く融資を受けた**と結論されている。

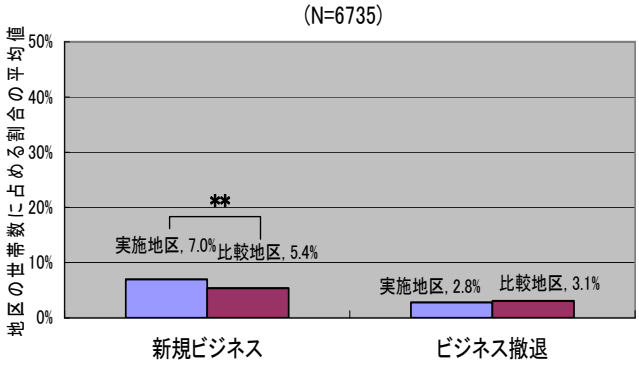
図2 マイクロファイナンスからの借入



(2) 新規ビジネス開始へのインパクト（図表 3）

新規ビジネスを開始した率は、実施地区（52 地区）が 7.0%、比較地区が 5.4%でその差 1.6%だった。これは 5%水準で有意と判定された。一方、新規ビジネスが開始されることにより競争が発生して、ビジネスから徹底するケースが出る可能性も考えられたが、実施地区 2.8%、比較地区 3.1%、その差 0.3%で誤差の範囲内と判定された。したがって、**マイクロファイナンスの実施により、新規ビジネスの開始が増加する**ことと結論された。

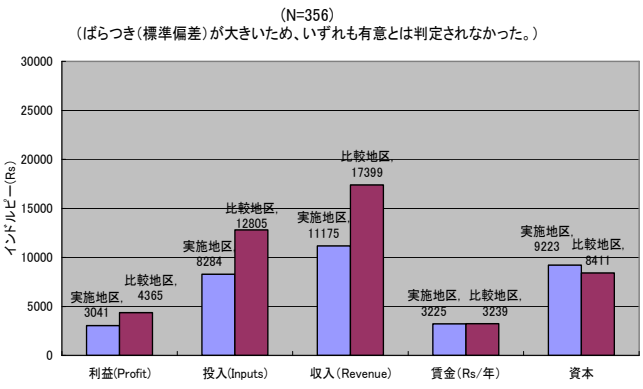
図 3 新規ビジネスへのインパクト



(3) 新規ビジネスにおける効果（図表 4）

実施地区と比較地区で新規ビジネスに限って比較した場合、**利益、投入、収入は、実施地区の平均値の方が低かったがいずれも優位な差であるとは判定されなかった。**賃金、資本に関しても優位な差ではない。これは、一口に新規ビジネスと言っても高収益を上げて一気に規模を拡大したケースからぎりぎりの水準で存続しているケースまで多数のケースがあり、数値のばらつき（＝標準偏差）がおおきくなっているからであると考えられる。また新規ビジネスに限っているのでサンプル数が少なくなっていることも影響していると見られる。

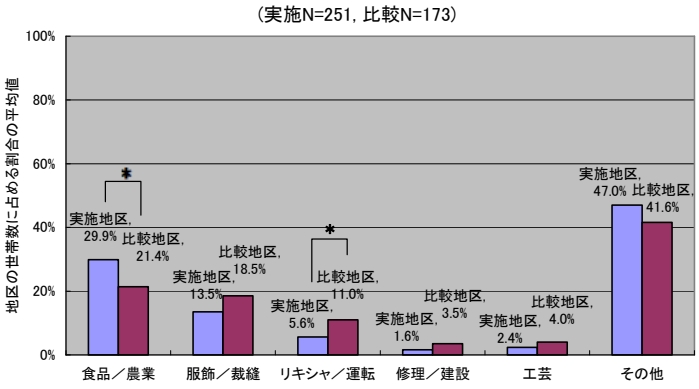
図 4 新規ビジネスへにおける効果



(4) 新規ビジネスの種類(図 5)

実施地区は、**「食品／農業」が多く、「リキシャ／運転」(リキシャはタクシー)が少なくなった。**前者は小資本ですぐに開始できる事業で一方、後者は今回の種類分けでもっとも資本がかかる事業であることが現れていると解説されている。

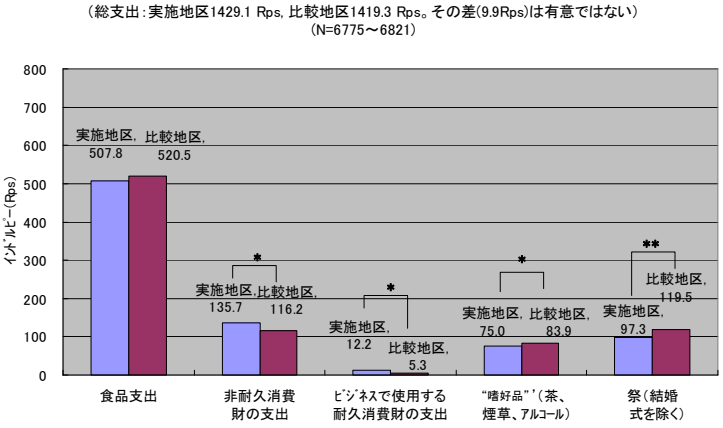
図 5 新規ビジネスの種類



(5) 月間家計支出へのインパクト(図 6)

ビジネスで使用する耐久消費財の支出が増加している一方で、"嗜好品"(茶、煙草、アルコールおよび祭(結婚式を除く))の支出が減少しており、両者の間に支出の移動が見られる。さらに、従来からビジネスをしている世帯、新規ビジネス開始の可能性の高い世帯、新規ビジネス開始の可能性の低い世帯に分割して再集計してみると、新規ビジネス開始の可能性が高い世帯でこの傾向がより顕著であることが観察された。

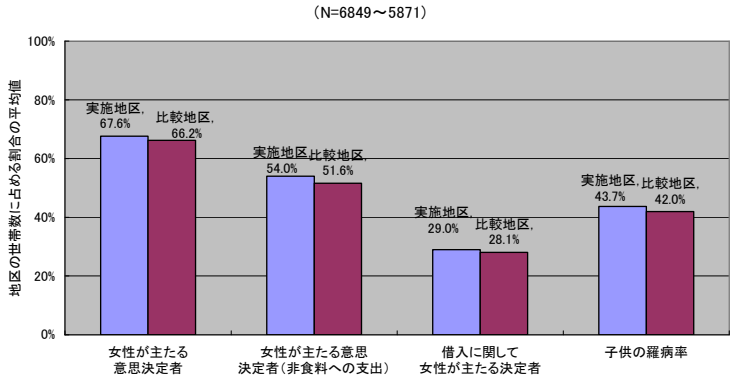
図 6 月間家計支出へのインパクト



(6) 女性のエンパワーメント、保健、教育への効果(図 7)

いずれの指標も、**実施地区の方が比較地区よりも高かったが、その差は統計的に有意であるとは判断されなかった。**(なお、「女性が主たる意思決定者」であると回答した率が 70%近くに達しているのは率直に驚くべきことであり、アンケートのとり方などを再検証する必要があると筆者(佐々木)は考えている。)

図 7 女性のエンパワーメント、保健、教育への効果



４．結論

以上の分析を通じて、マイクロファイナンス事業に関して次のとおり結論された。

マイクロファイナンスは、新規ビジネス開始にある程度の効果がある。また、ビジネス関連を含む耐久消費財への投資の増加と、”嗜好品”（茶、煙草、アルコールなど）とお祭関連の支出の減少をもたらすという効果がある一方で、女性のエンパワーメント、教育、保健への効果は（少なくとも短期的には）確認できなかった。

マイクロファイナンスは、よく主張されるように「奇跡（ミラクル）」ではないかも知れないが、借入、投資、そしてビジネスの拡大を実現することを可能にする。

５．議論

RCTを適用することの利点と懸念・限界についてはすでに多数の論文があるし（例：Bauchet & Morduch 2010)、「貧困アクションラボ」のBanerjeeとの議論をもとに筆者もまとめている（佐々木 2010)。それを繰り返す必要はないので、以下の点のみを述べる。

RCTの適用により、開発援助に関して「何が機能し、何が機能しないのか」が明らかになることが多くなった。今後のSDGs達成に向けた適切な政策選択に寄与することが望まれる。ただしそれは、政策立案者がいかに適切にRCTの評価結果を理解して政策に反映させようとするかの問題でもあり、RCTを用いて評価をする側としては、その政策立案者の努力を継続的に支援していかねばならない。

なお、今回のレビューを通じて、論文がかなり専門的になっていることが懸念された。社会科学系の大学院でひととおり統計学のコースを修了したレベルの知識が要求されるようである（３～４コースの履修が必要であろう）。ただし、**そもそもRCTは二つのグループの平均値を比べるという単純さと分かりやすさが大きな利点**であり、その利点は維持されねばならない。今回レビューした論文でも、せっかくRCTを適用しているのにそのデータを用いて複雑な回帰分析が行われているケースが多数あった。回帰分析では正確な介入効果が分からないからRCTが注目されて普及してきたという経緯があるわけで、その原点に立ち返るべきである。

ただしそうは言っても、RCTを用いた評価結果の論文を適切に理解するには、やはり最低限の統計学の知識は必要であると言わざるを得ない。それは、平均値と標準偏差の計算、2群の有意差検定、データの標準化、重回帰分析の知識である。筆者の経験から言えることは、こうした統計学の知識は、授業を受けて、自ら電卓なりエクセルなりを動かして手計算する訓練を経て、初めて身につくということである。これは、定性的な手法であるインタビュー（キーインフォーマント、フォーカスグループ）、直接観察、参与観察などの「習うより慣れろ」という手法群とは根本的に学び方が違うと言わざるを得ない。開発援助の世界で働く人のために、「統計学のアダルトラーニング」が必要だと思われる。

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

（参考）専門家評価の事例

専門家評価の事例
船員教育（エジプト）

これは、５つのインパクト評価のデザインにあてはまらない非常に簡便なデザインの適用事例ｊになります。まったく勧められませんが、参考までにここで紹介致します。

問題の所在と評価結果

アラブ海運大学校(AMTA)は、1970 年に開催されたアラブ連盟第 12 回運輸・通信理事会における決議に基づいて、1972 年に、連盟加盟国からの拠出金などによってエジプトのアレキサンドリアに設立された。その設立目的は、アラブ連盟諸国の自国産油の自力輸送及び国際収支改善のために自国船隊増強を図るために、外航船乗組員及び陸上勤務者を養成することであった。

１．施策の概要

AMTA では、UNDP などの援助によって 1977 年までの５年間に運営を軌道に載せることを計画していたが、予算不足のために計画に遅延が生じた。1974 年に日本に支援を要請し、1976 年から４年間、AMTA に対して援助実施を行なった。援助は、AMTA の海運訓練センター、航海学部、機関学部において船員養成機構の強化を図った。その後も援助は継続された。

２．評価結果

有識者が現場視察とインタビューを通じて評価を実施する。なお、現場視察に先立ち、事前に日本国内で以下を行なった。

（１）航海訓練船・青雲丸の実地見学（東京湾にて）
（２）航海訓練所本部の訪問と視察（横浜にて）

現地調査を通じて、評価者は次のような評価結果を出した。「AMTA では、24 名の指導者が育成され、協力終了後約 20 年が経過した現在も、その多くが、AMTA の後継組織に勤務している。同組織では、これまでに良質の海運業従事者を多数輩出してきている。（中略）また、ほとんどの研修参加者が帰国後にセミナーや講義などを開催し、研修で修得した技術の再移転・普及を行なっており、効果の拡大も図られている。」したがって、目的とした「外航船乗組員及び陸上勤務者を養成する」ことは長期にわたって実現されてきたと言えるだろう。

３．利点、制約、日本での適用に関する留意点

この方法の利点としては**とにかく簡便だ**ということがあげられる。事前の段階でも事後の段階でもとくにデータを用意しなくてもいいのである。では何を比べるかというと、評価実施者が有する心の中の基準と、事後段階で評価実施者が受けた印象の二つの差である。

そして利点がそのままこの方法の制約である。言うまでもなく、この方法は今まで説明した手

法に比べて極めて曖昧で不安定である。この方法を用いた評価結果の根拠は何かと問われれば、○○大学名誉教授の○○先生がそう言っているのだ、ということのみである。

じつは日本でつい最近まで行われてきた評価とはまさにこの方法を用いた「有識者による評価」が大半だったのだ。この評価を依頼された有識者は新聞記者であったが、今後この方法を用いる際の留意点を以下のとおり指摘している。「『有識者評価調査』をより実効あらしめるものとするため、つまり評価調査自体の質を高めるため、事前に必ず国内で、技術協力の実施にあたり長期専門家などを派遣した派遣元機関の視察を義務づけることを提案したい。**正直なところ、今回東京湾での航海訓練船・青雲丸の実地見学と、横浜の航海訓練所本部の訪問がなければ、現地での満足な調査ができたかどうかは、まったく自信がない。**」比較グループや事前段階のベースラインデータを設定しない専門家評価における「評価の基準」とはその専門家が持つ心の中の基準や経験から導き出す基準であり、その基準が適切に設定できるかどうかにこの手法を用いた評価が成功するかどうかの、ほぼ全てがかかっているのである。

4．議論

この事例は、インパクト評価の5つのデザインのどれにも当てはまらない外側の例である。これが従来一般的だったデザインであり、それは「現場視察＋関係者インタビュー」である。これは簡便であるという利点があるが、なんらエビデンスを明らかにするわけではない。また、この冊子に掲載した「別添1：インパクト評価のデザイン一覧（詳細版）とエビデンス・ピラミッド」の中では最下位に位置づけられるデザインである。ただし、なにも評価をやらないよりはよっぽどよく、さらに事業の実施機関が自ら行う「自己評価」よりもいいと言える。

（出所）すでに公開されている国際協力事業団(2000)「平成12年度事業評価報告書」第3章　事後評価調査III.有識者評価　船員教育エジプトの記載をもとに、筆者が独自に説明文を作成した。なお、原文のPDFファイルは以下からダウンロードできる。

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

III. インパクト評価に関連する学術的な議論の紹介

議論１：インパクト評価のデザイン一覧（詳細版）とエビデンス・ピラミッド

アメリカにおける評価学の標準テキストのひとつである Rossi らの Evaluation: A Systematic Approach では、インパクト評価には３タイプ 12 種類にわたる代表的なデザインがある。本テキストではこのうちの代表的な 5 つのデザインを採用している。

インパクト評価デザインの一覧表

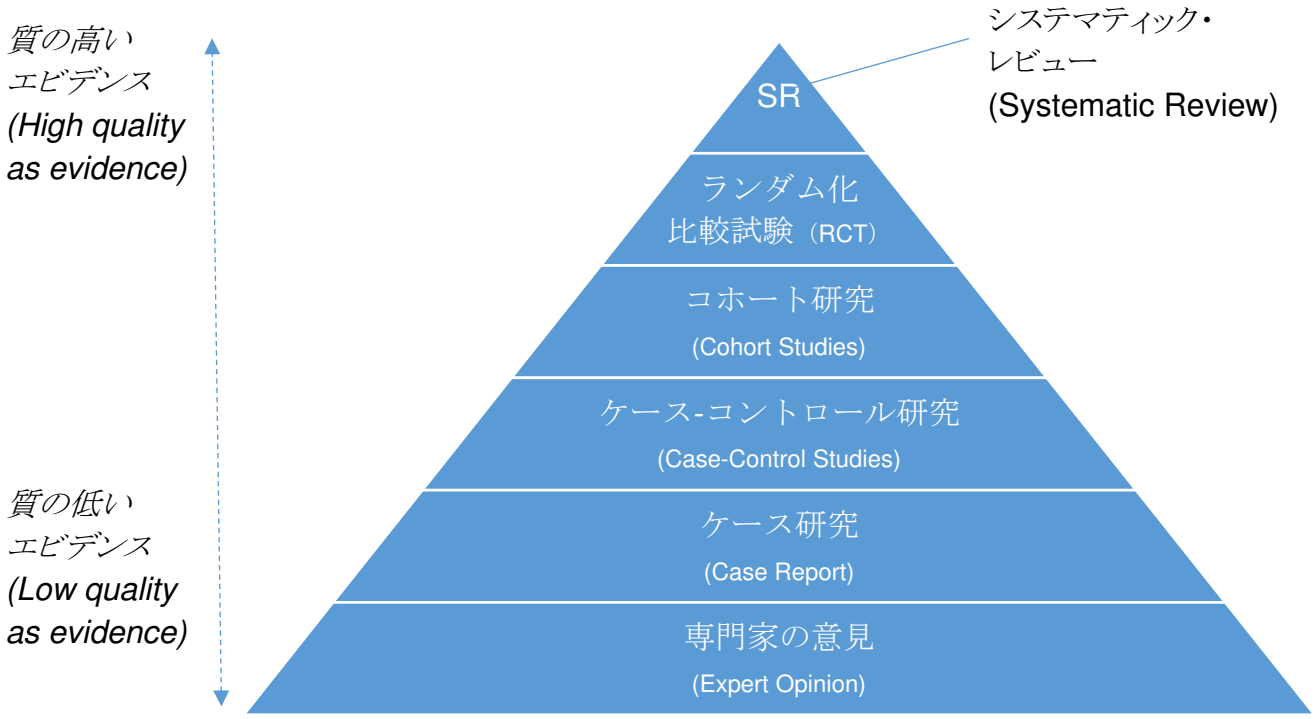
インパクト評価の分類	特徴･制約	客観性/総合ｺｽﾄ/利用難易度
A. 実施-比較グループ両方が存在するケース		
（１）ﾗﾝﾀﾞﾑ化比較ﾃﾞｻﾞｲﾝ	→「政策」の実施前に、政策適用をランダム・アサインメントにより、実施・比較グループを設定する。	極高　極高　極難
（２）準実験デザイン		
① 回帰・分断ﾃﾞｻﾞｲﾝ	→政策実施前に、特定の数値でサンプル集団をふたつに分断して、実施･比較グループを設定する。	高　高　難
② ﾏｯﾁﾝｸﾞﾃﾞｻﾞｲﾝ	→可能な限り近似のグループを選定して比較グループにする。	
③ 統計的等化ﾃﾞｻﾞｲﾝ	→統計処理によりサンプル集団を実施・比較グループに分ける。	
④ 一般指標ﾃﾞｻﾞｲﾝ	→全国平均値、全県平均値等を比較グループのかわりに用いる。	低　低　容易
B. 実施グループしか存在しないケース（E.g.全国対象プログラム）		
（３）ｸﾛｽｾｸｼｮﾝﾃﾞｻﾞｲﾝ	→複数のグループや地域の間のサービス投入量と改善効果の量のばらつきを利用してインパクトを評価する。	高　高　難
（４）時系列ﾃﾞｻﾞｲﾝ	→事前、事後の指標値を長期間にわたって測定して比較する。	
（５）パネルデザイン	→短期間の事前、時中、事後の指標値を比較する。	
（６）事前･事後比較ﾃﾞｻﾞｲﾝ	→シンプルに、事前、事後の指標値を比較する。	低　低　容易
C. 簡便的アプローチ		
（７）ｴｷｽﾊﾟｰﾄ（専門家）評価	→学者や有識者等、いわゆる「専門家」がベースラインを設定する。	低　低　容易
（８）受益者評価	→アンケートやインタビューにより受益者がベースラインを設定する。	
（９）行政官評価	→政策実施を担当した行政官がベースラインを評価する。	極低　極低　極容易

（出所）Rossi, Freeman, Lipsay *Evaluation A Systematic Approach,6th Edition*,Sage Publication,1999,p261 の表の分類を参考して筆者が一部変更した。ただし、「特徴･制約」、「客観性／総合コスト／導入難易度」は著者独自の経験と判断にもとづいて記述した。

さらに、エビデンスの質を一覧にしたものに「エビデンス･ピラミッド」がある。以下にその由来と概略を掲載した。

Evidence-based Policy Making (EBPM)は国際的な研究の潮流であるが、もともとこれは Evidence-based Medicine（EBM）に基づいて提案された。1993 年にアメリカの Agency for Health Care Policy and Research（AHCPR）が臨床研究におけるエビデンスのランクを提案した。この提案を受けて、社会科学系の研究においても、様々な機関や研究者によってエビデンスのランクが提案された。現在提案されているエビデンスのランクの一例は以下の通りであるが、多数の研究者によってさまざまなバリエーションが提案されて、現座に至っている。

図: エビデンス・ランク　Evidence Ranks (Evidence Pyramid)



	アプローチ（Approach）	説明（Explanation）
1a	システマティック・レビュー Systematic Review (SR)	複数の無作為化比較試験のメタアナリシス Meta-analysis of multiple randomized controlled trials
1b	無作為化比較試験 Randomized Controlled Trial (RCT)	プロスペクティブ；無作為化、介入群と対照群の比較 Prospective; Randomization, Comparison between the treatment group and the control group
2	コホート研究 Cohort Studies	前向き（これから介入）、無作為化なし、介入群と対照群の比較 Prospective; No randomization; Comparison between the treatment group and the control group
3	ケース-コントロール研究 Case-Control Studies	後ろ向き（すでに介入済）、無作為化なし、介入群と対照群との比較 Retrospective; No randomization; Comparison between the treatment group and the control group
4	ケースレポート、 ケースシリーズ Case Report, Case Series	記述的な報告 Narrative reviews
5	専門家の意見 Expert Opinion	専門家の意見 Expert opinion

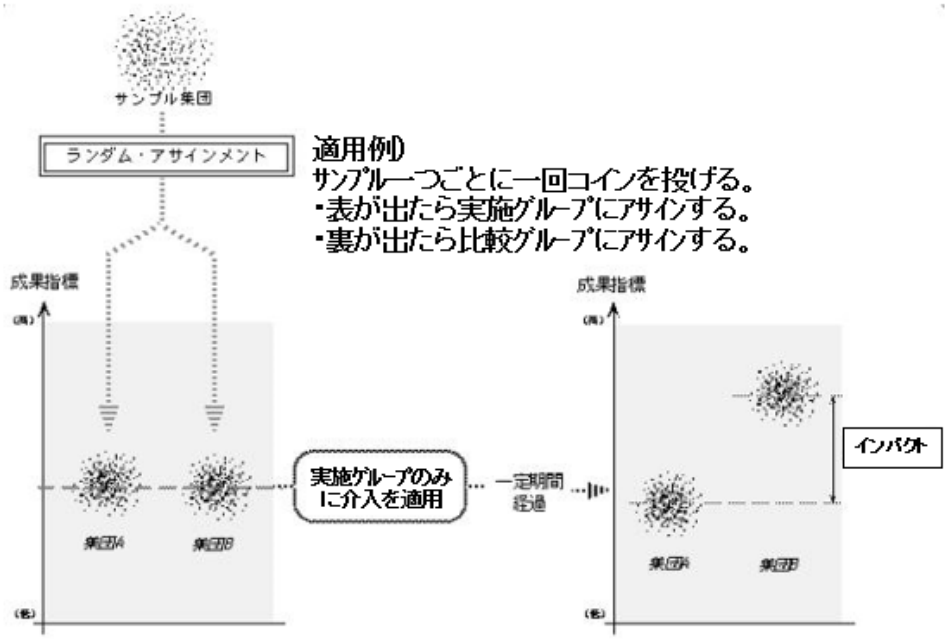
出典：Walden University *Levels of evidence pyramid*. 若干の修正を加えて引用しています。
（https://academicguides.waldenu.edu/healthevidence/evidencepyramid ）

議論２：評価を巡る論争その１
「科学的評価」VS.「実践的評価」

この論争は、長くそして根の深い論争であると言える。また評価の本質を巡る根本的な論争である。決着はついていない。この議論を知ることにより、評価の利点も評価の限界も見えてくるのだろう。

科学的評価 **(Scientific Evaluation)**

キャンベル（Campbell,D.T.）は 1969 年の論文の冒頭で次のような主張をした。「アメリカ合衆国とその他の現代国家は、**社会の改善に向けて実験的アプローチを用いる用意ができなければならない**。ここで言う実験的アプローチとは、特定の社会問題を解決するためにデザインされた新しい施策を実施する際に用いるアプローチであり、このアプローチによって、不完全ながら複数の基準に照らしたときに明らかな効果があったのかどうかを確認し、その確認の結果に基づいて、施策を維持するか、改善するか、あるいは中止するかを決定することになる。」(Campbell,D.T, 1969, p409)



図（上の図):キャンベルが解説する「実験的アプローチ」（通常の RCT です）

実践的評価 **(Pragmatic evaluation (Rossi and Freeman), Practical program**

evaluation (Hatry, Wholey), Practical Evaluation (Patton)など研究者によって英語の呼び方は多様)

これに対して、もともと統計学が専門の**クロンバック（Cronbach, L.J.）**は、1982 年の著書で次のように反論した。「**評価研究をデザインすることは、アートである。・・・**評価の中心的な目的は、基礎的な社会調査とは違う。そして評価は、それぞれ違う制度的及び政治的なコンテクストに適合すべきだ。科学的調査のような長期的な取組みには適するであろう多くの提言は、評価には適さない。さらに、科学的な手法やデザインに関する一般的な論文は、評価実施者には適当ではない。評価に関する一般的な提言も誤解を招く。評価は、ある一つの型にはめ込まれるべきではない。どんな評価でも、たくさんの良い手法(Design)があり得るが、完ぺきな手法というものはあり得ない。」（Cronbach, L.J. , 1982, pp1-2) 。そして次のように言い切っている。**「評価はアートだ（"Evaluation is an art"）**。そして特定の時点の、特定の予算枠の中の、特定のプログラムの研究であったとしても、評価に唯一の最良の計画というものはない。」（Cronbach, L.J. , 1982, pp321)

この両者の議論を、1970 年代から現在まで評価を巡る議論をウォッチしてきた**ロッシ (Rossi, P.H.)** は次のように解説している。

「**'科学的評価' 対 '実用的評価' の態度**（Scientific Versus Pragmatic Evaluation Postures）: たぶん、評価研究の世界で、もっとも影響の大きかった論文とは、キャンベルが 1969 年に発表したものだろう。この論文は、キャンベルが何十年かにわたって主張してきた見方を示している。それは、政策や施策の決定は、社会状況を改善する方法をテストする継続的な社会的実験の結果に基づくべきである。それだけではなく、社会調査の技術は、'実験する社会'を実際に実現するために利用可能だと述べた。そして、キャンベルは、社会心理学において彼が学び、そして実際に適用した手法である実験モデルを、評価調査にも適用することを追求した。彼の後年の著作では、いくぶんその立場を和らげてい

るが、キャンベルは、科学的調査のパラダイムに評価調査をあてはめようとした人物とみなすのがフェアであろう。

一方、キャンベルの立場は、評価のフィールドにおけるもう一人の巨人であるクロンバックによって挑戦されることとなった。調査の手続きと同じ考え方で科学的調査・評価は、使えることもあるかも知れないと断りながら、クロンバックは、評価の目的は、科学的調査の目的とははっきり違うと主張した。彼の見方では、評価は科学というよりもアートであり、全ての評価は、意思決定者や利害関係者のニーズに合うように形作られるべきであるとする。それゆえ、科学的研究が基本的には調査のスタンダードを満たしているかどうかに奮闘する一方で、評価は、政治的環境や、施策の制約、そして利用可能な資源の枠のなかで、意思決定者に最大限に役立つ情報を提供することに貢献すべきである。」（著作(3)）

なおクロンバックと同じ時期(1981 年)に**ハトリー（Harry Hatry）とホーリィ（Joe Wholey)**によって以下の指摘もなされている。

「・・・クラッシックな評価デザインは応用の度合いが限られているし、常識的な考え方を越えて困難さを強いることになっているという認識が年々強まっている。また、評価がどれくらい役に立つか（Usefulness)、そして評価をより役に立つようにするためにはどうしたらいいかに、さらに関心が集まっている。」(Hatry, Winnie & Fisk, 1981, p.ix.)

最後に評価研究に関する最近の著作を見ると、次のような傾向が観察されている。

「近年の評価専門家の間の合言葉（watchword）は、『実用重視の評価』（Utilization-focused evaluation）である。実用重視の評価は、施策を任せられた人達によって挙げられる特定の質問に答えるためにデザインされる評価のことであり、そのおかげで、施策の今後に関する決定に影響を及ぼすことができる。――評価やモニタリングに関して、どんな施策も次の３つの基本的な質問が挙げられるべきである。（１）評価の結果は、施策に関する意思決定に影響を及ぼせるか？（２）評価は、評価結果が必要とされる時点までに終われるか？（３）当該施策は、評価をするだけの重要性があるのか？の３つである」（Wholey, Hatry & Newcommer, 1994, p5）

なおこうした「実用重視の評価」の流れがますます強まった際に、「科学的評価の評価結果の方が長期的に参照されて利用されるので、じつは科学的評価の方がよほど実用的なのだ」という主張もなされている（キャンベルの盟友のクック(Cook, D.)の発言だったと記憶している)。

Campbell,D.T. (1969). *''Reform as Experiments''* American Psycologist, 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.

（画像 1 の出所）https://en.wikipedia.org/wiki/Donald_T._Campbell の画像を参照して著者がスケッチした。
（画像 2 の出所）https://archon.library.illinois.edu/?p=digitallibrary/digitalcontent&id=10865 の画像を参照して著者がスケッチした。
（画像 3 の出所）http://www.columbia.edu/cu/csswp/1995.htm の画像を参照して著者がスケッチした。
（画像 4 の出所）http://www.columbia.edu/cu/csswp/1995.htm の画像を参照して著者がスケッチした。
（画像 5 の出所）http://www.businessofgovernment.org/bio/joseph-wholey/の画像を参照して著者がスケッチした。

（出所）佐々木（2003）から抜粋（pp.20-23）

議論３：評価を巡る論争その２
「定量的評価」対「定性的評価」

これも長くそして根の深い論争。1960～70 年代に定量的評価が広く認知されたあと、定性的評価の唱道者が現われてたびたび定量的評価を批判し、定量的評価の側はその批判に無言で耐えてきた。

定性的評価の側の主張

「今までの評価者（＝定量的手法を用いる評価者）は、改善効果を測定すること及び重要な要因を他の要因から切り離すという、実際の能力以上のことをやろうとしてきた。あげくの果てに、別々の政治的立場に仕える結果となっている、それも不十分に。」（Stake, 1980, p38）

定量的評価の側の主張

「現在主流である定量的手法よりも定性的手法を使うべきだという主張は、ほとんど神秘主義的で、また、改善効果の特定に関しては施策実施者自身の見方を受入れてしまっている。」（Rossi, 1985, p7）

最近の議論（1990 年代末まで）

- 「定性的評価は統計的な厳密さを欠いているという意見がある。しかし、評価に統計的な厳密さを求めるのは適当ではなく、むしろ社会的に弱い立場の人々の関心事を理解するためには定性的評価の方がより適当であるという意見もある。」（Bamberger, 2000,）
- 「定性的評価も定量的評価も長所と短所を有している。両者は代替もできるが両方を同時に使うこともできる。そして同一の評価調査のなかで同時に両方のデータを集めることができる。」(Patton, 1990, p14)
- 「定量的手法と定性的手法を組み合わせるが理想的である。なぜなら、それはプロジェクトの定量的なインパクトを提供するとともに、そのアウトカムを生み出した過程や介入についての説明も提供するからだ。」（Baker, 2000）

つまり、ひとつの評価の中で、定性手法と定量手法を用いる「混合手法」（Mixed Method）が提案されたわけで、これにより 1990 年末頃に一定の合意を見たと言える。これでこの論争は一応の決着に至った。

さらに最近の議論(2000 年以降)

だがしかし、2000 年代に入って新たな動きがあった。それは定量的評価の側の逆襲である。2003 年にマサチューセッツ工科大学（MIT）に「貧困アクションラボ」（Povery Action Lab.）が設立され、同ラボは「実験デザイン（RCT）しか使わない」と宣言した。そして設立以来 20 年で 2,000 件以上のインパクト評価で実験デザイン（RCT）を使ってインパクト評価を実施した。そして 2019 年には設立者の 3 名がノーベル経済学賞を受賞した。定性的評価の唱道者の批判に定量的評価の側は無言で耐えてきたが、ついに定量的評価の側が逆襲に成功したのである。

今後の議論の展望（2020 年以降）

定性的評価と定量的評価の論争は今後も続いていくであろう。そして今後もたびたび両方の手法を同時に使うべきだという結論に達して握手することもあるだろうが、両者とも秘めた信念を絶対に変えないであろう。

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

Rossi,P.H. (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, M.Q. (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.

（出所）佐々木（2003）から抜粋（pp.24-25）。2022 年に加筆した。

議論４：ランダム化比較デザインの是非を巡る考察：
ノーベル経済学賞受賞者 MIT「貧困アクションラボ」の
Abhijit Banerjee（アビジット・バナジー）との議論

2019 年のノーベル経済学賞が、マサチューセッツ工科大学（MIT）のアビジット・バナジー教授とエステル・デュフロ教授、ハーバード大学のマイケル・クレマー教授に授与された。MIT の「貧困アクションラボ」(The Abdul Latif Jameel Poverty Action Lab (J-PAL))が実行し続けている、ランダム化比較デザイン（RCT）を用いて、世界の貧困問題を緩和するために有効な方策を明らかにする取組が評価されたものである。筆者（佐々木）は 2006 年に、「貧困アクションラボ」を訪問して、バナジー教授にインタビューする機会を得て、RCT の適用に関して議論して論文にまとめていたので、その抜粋を以下に掲載する。

（筆者注：文中に「バナージェ」とカタカナ書きが出てくるが、近年は「バナジー」と記載されることが多くなっている。しかし、この抜粋では、論文執筆当時の記載をそのまま維持した。）

（出所）https://www.newssalt.com/30231 の画像を参照して著者がスケッチした。

はじめに：この論文の意義

評価研究の第一人者（Shadish, W. et al, 1991）であり評価研究における唯一の哲学者（Stake, R.E. 1982）と呼ばれることもあるスクリヴェンは、因果関係を証明するとはどういうことかに関して哲学的な研究を重ねてきた（たとえば Scriven, M. 1975 など）。そしてスクリヴェンは、ランダム化実験デザインが因果関係を証明するための最良のデザインであるという主張を一貫して鋭く批判してきた。たとえば、アメリカ教育省の「一人の子供も落ちこぼれにしない法」（No Child Left Behind Act of 2001）においてランダム化実験デザインがもっとも望まれる調査デザインであると明文化されたことに端を発して、それに対する賛否によりアメリカ評価学会が二分された際には、スクリヴェンはランダム化実験デザインを批判する勢力の急先鋒として論陣を張った（詳細は、Davidson,

S.I.& Christie, C.A.（2004）. *The 2004 Claremont Debate* を参照)。スクリヴェンが展開してきた批判を体系的に整理した論文として「因果探索を巡る論理」(Scriven, M.（2007）. *The logic of Causal Investigation*）がある。その中でスクリヴェンは、教育や保健を含む社会施策の評価において、無条件にランダム化実験デザインが最良のデザインであると主張するのは論理的な誤りがあると指摘している。

一方で､経済学者であり貧困アクションラボの所長であるバナージェ（Abhijit Banerjee）は、自身の著作（2007）において、ランダム化実験デザインでなければ本当に効果があるかどうかを判断することはできないと明言している。そして、開発援助分野においてランダム化実験デザインを普及させるために同ラボの活動を開始したと述べている。スクリヴェンとバナージェは専門分野が違うこともあり、二人の間で直接的な論争が交わされた形跡はない。しかし、著者（佐々木）が 2006 年に同ラボを訪問してバナージェと議論する機会 [1] があったので、著者のメンターであったスクリヴェンの主張を直接質した（Sasaki, R.（2006）. *Discussion with MIT's Poverty Action Lab.*）。

以下に、スクリヴェンの批判の概要と、それに対するバナージェの反論を整理した。さらに、開発援助評価を長年にわたって研究対象としてきた著者（佐々木）の考察をそれぞれに記載した。つまり、エビデンスに基づく開発援助評価の 3 つのルーツ（評価研究の系譜、開発援助の系譜、経済学の系譜）に対応する 3 人の意見を整理することにより、議論の結論を得ることを目指した。この整理作業を通じて分かることは、それぞれの主張にはじつはそれほどの違いはないということである。

（1）貧困アクションラボのランダム化実験デザインは盲検法が欠落している

スクリヴェンは、保健医療分野で用いられているランダム化実験デザインは二重盲検法（Double blind design）が適用されているが同ラボのデザインにはそれが欠落していると指摘する。つまり「無盲検」(“Zero” blind）であり、いわゆるホーソン効果が入り込む余地があるので、純粋に介入の効果を明らかにできるとは言えないと指摘する（Scriven M. 2007)。

これに対してバナージェは次のように答えた。その指摘はそのとおりだが、保健医療の実験（Clinical trial）ではないのだから、真薬と偽薬（Placebo）を使うことはできない。かわりに実験グループと統制グループが同一の情報を共有することにより、無盲検である影響を最小化しようとしていると反論している。

結論としてバナージェは自身が用いているデザインの限界を認めたことになると言えるが、社会科学分野にランダム化実験デザインを導入したキャンベルでさえも盲検法の適用を論じていないわけであり、バナージェの言うように、社会施策を対象とする限り「仕方がない」(No other way）と言わざるを得ない。

（2）統計的有意と社会的有意は違う

スクリヴェンは、二群の差が統計的に有意なだけでは不十分であり、社会的あるいは実践的に有意（Socially or practically significant）でなければ介入は効果があったとは言えないが（Scriven M. 2007）、貧困アクションラボは、統計的有意を持って介入は効果があると判断していると指摘する。

これに対してバナージェは、そのアイディアは拒否しないが、統計的に有意でなければ社会的に有意であることもありえないと反論している。つまり統計的有意は、社会的に有意かどうかを検討するために最低限満たされるべき条件である。また同ラボは、地元の人たちの意見を聞いて社会的に意味があるかを判断していると反論しているが、それでは地元の人たちが成功・失敗を決定するのかという問いに対しては、彼らは‘相談される’（Consulted）と答えている。

つまり効果があるかどうかの判断は、まずは専門的な統計分析によってなされるべきでそれをしないで地元の人たちの判断に委ねることは、ランダム化実験デザインの利点である厳格さを損ねると反論していると理解できる。

（3）人々を偶然によって二分することは倫理的問題があるし、その処置に関して事前承認を得ることは困難である

スクリヴェンは、途上国に住む親は、自分の子供が統制グループに入るかも知れない処置を承認しないだろうとして、反倫理の問題および事前承認（Informed consent）の取得の困難さをランダム化実験デザインの避けがたい制約として指摘している（Scriven M. 2007）。

これに対してバナージェは、現場の経験から言えることは、じつはランダム化こそがフェアなのだと反論する。第一に、もともと援助資源は全員をカバーできるほど用意されていることは稀であり、恣意的に適用者を決めることを避けてランダム化（同じ確率に基づく宝くじのアイディア）を適用することはたいへんフェアなのだと指摘する。第二に、今まではランダム化の代わりに「われわれドナーの基準によると」と説明して、幹線道路沿いの村や一日で視察できる村が選定されることが多く、住民としては非常に不平等でアンフェアだと思っていたと聞かされることがある。そうした状況の中でランダム化のアイディアを打診すると歓迎されることが多いと指摘している。

結論として、援助機関側が想定するほどランダム化への抵抗は少なく、ランダム化実験デザインの適用に対する制約にはならないという指摘は説得力がある。実際のところ、援助に携わる者は著者を含めて皆経験しているように、援助側の都合で対象地域や対象者を選んできたわけであり新鮮な指摘である。

（4）「エビデンス」という単語が、定量手法を用いる研究者に独占されている

スクリヴェンは、「エビデンスに基づく実践」という考え方は完全に受け入れられる考え方だが、エビデンスの定義がランダム化実験デザインの適用によって得られた結果のみに限定されていることが問題だと指摘する（Scriven M. 2007）。そして、定性的手法も厳格に適用することによってエビデンスを産出することができるとして、エビデンスという言葉の定義の再検討を要求している。

これに対してバナージェは、第一に、どのような分析手法も否定する意図はないとする。単純に、ランダム化実験デザインがその有用性にも関わらず他の手法に比べて非常にわずかにしか用いられて来なかったことから、現状よりも頻繁に用いられるべきだと主張しているだけだとしている。第二に、定性的手法、特に詳細な観察記述（Rich description（Stake, 1982）は「なぜそれが起こったか」を説明するので補完的な役割を演じることができると指摘している。さらに言えば、ランダム化実験デザインは限られた情報しか提供しないから、詳細な観察記述は、説明の中心的な方法（Center piece of explanation）となっていると指摘している。

スクリヴェンも、クック（Cook, T. 2000）の論文を引用して、定性的手法と定量的手法は相互補完的に利用できるし利用すべきだと結論しており、スクリヴェンとバナージェの間にじつは大きな認識の差はないと言える。

（5）ランダム化実験デザインの適用が適切ではない（あるいは意味がない）介入行為のタイプが存在する

スクリヴェンは、ランダム化実験デザインが時間と資源の制約によりランダム化実験デザインが適用できないタイプの介入があることを指摘することにより、常にランダム化実験デザインが最良であるとは言えないと指摘する（Scriven M. 2007）。

これに対してバナージェは、ランダム化実験デザインが常に最良と言うつもりはなく、今までの経験からランダム化実験デザインの適用が不可能だったり不適切だったりするタイプの介入があったことに同意するとしている。それらは、(i) すでに終わった事業（あるいはすでに開始されている事業）、(ii) 全国を対象とするような大規模事業（ランダム化実験デザインの名が示すとおり実験的な小規模の事業に向く）である。さらに、(iii) 事前に内容が確定しておらず実施しながら決定していくフレキシブルな事業は、ランダム化実験デザインに向かないとしている。そして開発援助にはそうした事業が予想外に多いと指摘する。最後に、(iv) 介入効果が強力に出ることが分かっている事業があえてパイロット事業に選定されていることがあり、これもランダム化実験デザインの趣旨に沿わないと指摘している。

貧困削減ラボの公開資料での宣言と比べて、バナージェは実際には意外と冷めている

と言わざるを得ない。ランダム化実験デザインの適用経験を重ねる中で、ランダム化実験デザインの適用が困難な特定の介入タイプが理解されてきたのだと言える。

結　論

ここまでの議論から分かることは、両者の主張にそれほど違いがあるようには見えないということである。そして、両者が述べる結論はますます違いがなくなる。

スクリヴェンが論文の結論部分において、もともとランダム化実験デザインは定性的手法の助けを借りて運用される混合手法であると言えるからランダム化実験デザインが用いられるべき場合は確かにあると述べている一方で（Scriven M. 2007）、バナージェはランダム化実験デザインによって評価活動が独占されるべきだと考えているのではなく、他の手法に比べてあまりに用いられてこなかったのでもう少し頻繁に用いられるべきだと主張しているだけだと結論している。

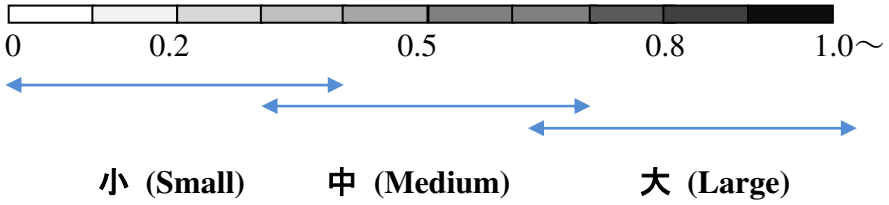
つまり、開発援助評価において、独占的というわけにはいかないが、ランダム化実験デザインが利用できるし利用すべき余地が確かに存在すると言うことである。

[1]（財）国際開発高等教育機構の調査により派遣された際のインタビューによる。（財）国際開発高等教育機構の湊直信・国際開発研究センター所長が調査を指揮した。

（出所）佐々木亮（2010）『エビデンスに基づく開発援助評価―援助評価の歴史、ランダム化比較試験の起源、スクリヴェンとバナージェの考え方の比較』In 「特集：エビデンスに基づく実践の世界的動向と日本における取組」日本評価研究 Vo.10, Np.1, March 2010（編集担当：佐々木亮、大島巌）　なお、文中で引用した参考文献は、以下の PDF に一覧が掲載されています。
http://evaluationjp.org/files/Vol10_No1.pdf

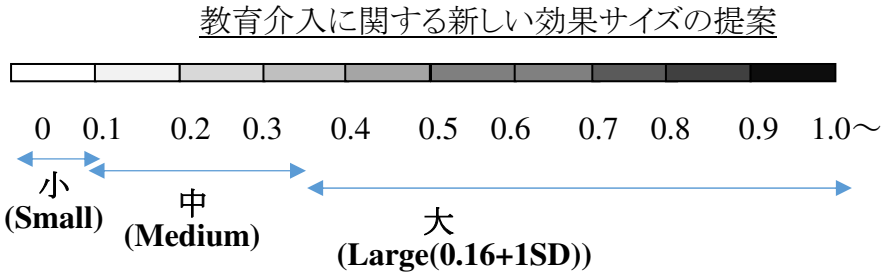
インパクト評価と統計学の発展のための提案１：
教育分野の先行研究に基づく新しい効果サイズ(Effect Size)の基準の提案

Cohen,J.(1988))は、「ES 指数を統計的に表すために、著者は慣例として効果サイズ（Effect Size）の値を「小(small)」、「中 (medium)」、「大 (large)」という質的形容詞の運用上の定義として用いることを提案した。彼の提案は広く受け入れられ、現在でも一般的な基準として使用されている。



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

しかし、この提案の前に、Cohen, J. (1988)は、「彼（＝研究者の意）は、この問いに答えるために理論に助けを求め、さらに助けを求めるために、その分野における先行研究の批判的吟味を求めるべきだ」と書いている。 そして 2023 年になって、私（佐々木）は、教育分野における 96 の RCT のメタ分析という重要な先行研究がついに発表されたと判断する（Evans, D.K. and Yuan, F (2022)）。この重要な先行研究に基づいて、筆者（佐々木）は効果サイズ（Effect Size）の値の「小(small)」、「中(medium)」、「大 (large)」の定義について、新たな基準を以下のように提案する。



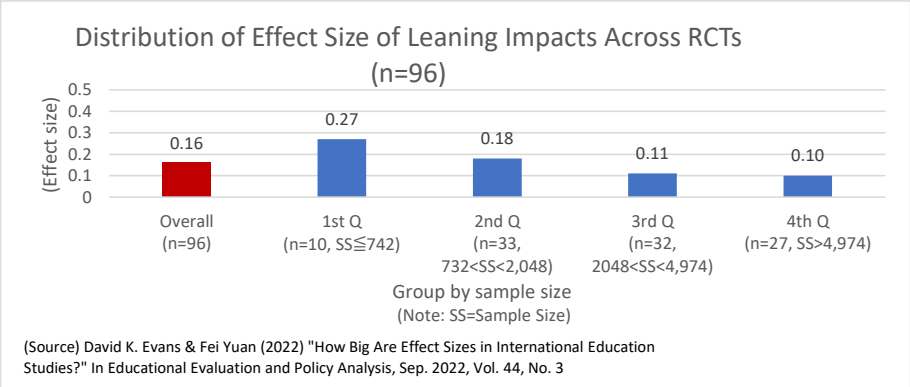
(Source) Sasaki, R. proposed, based on Evan,D. & Yuan,F. (2022)

この新しい効果サイズの提案について、Evans D.K.は筆者（佐々木亮）とのパーソナルコミュニケーションの中で次のように述べているので紹介する。「これは興味深い、そして （私の意見では)) 合理的な経験則です。私たちは、（付録の表にある）テストの種類によって非常に多くのばらつきがあり、1 つの経験則がそれほど役立つとは思えなかったため、この調査では新しいベンチマークを提案しないことにしました。しかし、それが過去の提案者の基準よりも優れていることには同意します。したがって、あなたの提案は進歩です。」

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

SS=Sample size (Unite of Mean and SD: Fffect 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

Reference

(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.

(3) Personal communication with Evans, D.K. (October 5, 2023). “This is an interesting and (in my opinion) reasonable rule of thumb. We made a decision not to propose new benchmarks in our study because there’s so much variation by type of test (in the appendix tables) that I didn’t feel like one rule of thumb would be so helpful; but I’d agree that it’s still better than [the original] benchmarks, so your proposal is an advancement.”

インパクト評価と統計学の発展のための提案2: 効果率（%）

多くの人がこの計算を行っていますが、この計算を正式な方法で提案した人はほとんどいない。これが「効果率(%)」(Effect Percent (%))の初めての正式な提案になる。「効果率(%)」という名称は、著者（佐々木）が教育関係者のために提案するものである。

よく知られたように、効果サイズ(Effect Size) とは「統合された標準偏差に対する 2 群の平均値差」のことである。この定義に従い、「効果率(%)」は以下のように定義する。

効果率（%）： 統制群の平均に対する 2 群の平均の差

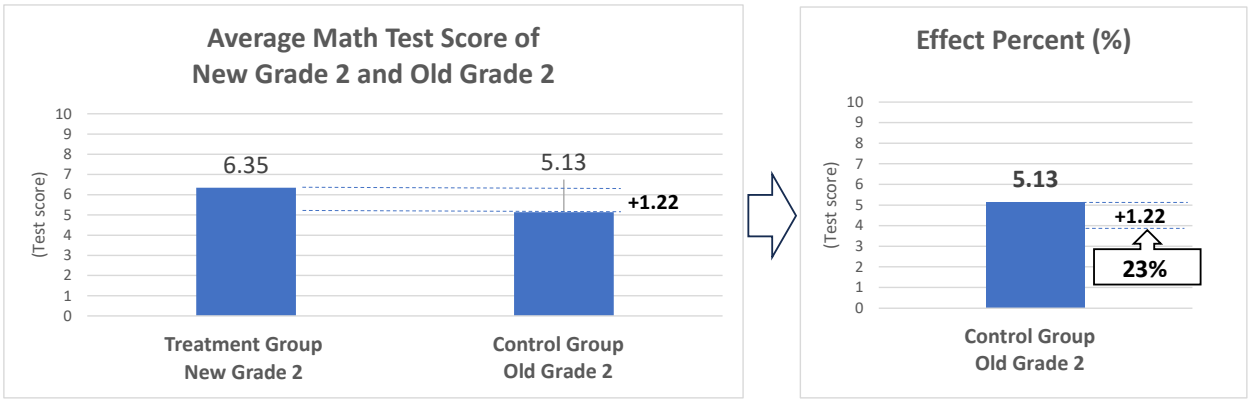
効果サイズ（Effect Size）は統計学の専門家にとっては納得のいくものだが、「小」「中」「大」などの意味は一般の人には明確ではない。一方、効果率（%) (Effect Percent(%))）は、『私たちの介入により、子どもたちの成績が 23%向上した』といった、シンプルですが保護者や生徒を含む一般の人にとって非常にわかりやすい文章となる。効果率(%)の計算式は以下のとおりである。

（効果率（%））＝（2 グループの平均値差）÷（統制群の平均値）

（注）統制群＝非・介入グループのこと

計算例（ミャンマーの教育介入（新しい数学テキスト開発）の効果評価）を紹介する。

23% = 0.23 = 1.22 ÷ 5.13



(注) 小数点以下の計算のため、最終的な%が上図の%と必ずしも一致するとは限らない
(出所) JICA. (2019). Impact Survey of the Project for Curriculum Reform at Primary Level Basic Education (CREATE) in Myanmar

なお、2 つのグループの平均値差を統制群の平均値と比較する必要がある理由は次のとおりである。 Glass G.V. (1976) は、統合された標準偏差の代わりに、2 つのグループの平均値差を統制群

の標準偏差で割ったグラスのデルタを提案した。介入群と統制群の統合は介入行為によって部分的に影響を受けている一方で、統制群のみの標準偏差は、介入行為による影響を受けない純粋な (または自然な) 標準偏差になるのでこちらを使うことが論理的だと言える。同様の考え方により、効果率 (%) (Effect Percent (%)) の計算には、すでに介入行為の影響を受けている統合された平均値または介入群の平均値を使用する代わりに、統制群の平均値を使用することが推奨される。

Glass G.V.は筆者（佐々木亮）とのパーソナルコミュニケーションの中で次のように述べているので紹介する。この効果率(%)（effect percent）の提案について「何を言いたいのかわかります。そして、これが多くの場合に影響を報告する非常に有益な方法であることがわかります。」と述べている。 ただし、（1）テスト問題すべてが easy で全員が満点と取るような場合には機能しない、（2）起点がゼロ(0)が意味を持つ場合に使える、という二つの注意点も指摘している。

Reference

(1) Glass, G.V (Ed.) (1976). *Evaluation Studies Review Annual, Vol. 1.* Beverly Hills: SAGE Publications.
(2) Personal communication with Glass, V.G. (August 8. 2024) . “I see what you are getting at. And I can see how it would be a very informative way of reporting effects in many instances.”

以下はこのミャンマーの結果を報道する新聞記事である。ここで名称を提案した効果率（%）が使用されており、一般読者のためにたいへん分かりやすい記載となっている。

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(59)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

（出所）朝日新聞 2010 年 1 月 6 日朝刊

インパクト評価と統計学の発展のための提案3:
ジェンダー・コード・マトリックス（Gender Code Matrix）

この「インパクト評価事例集」の「はしがき」に次のような記載がある。『(インパクト評価は)年齢、性別、人種、民族、出身階級、出身国、出身地、障害の有無などの違いから人々を自由にする力を秘めています。』 近年の多様性を尊重する時代になって、ジェンダーも男性・女性の二者択一ではなく、多様なジェンダーがあり得るという認識が一般的になってきた。統計学もこの動きを反映させる必要があるだろう。

重回帰分析はY＝aX+bという式を計算する行為である。Yがアウトカム変数で、Xが説明変数である。そしてXがひとつなら単回帰分析(Simple regression analysis)と呼び、Xが複数なら重回帰分析(Multiple regression analysis)と呼ぶ。

この重回帰分析でジェンダーを説明変数(X)の一つとして取り入れる場合には、1=女子、0=男子(あるいはその逆)というコーディングを行うのが一般的である。一方で、民族や年齢グループなど、複数のカテゴリーがあるカテゴリカルデータ(Categorical Data)を取り入れる場合には、複数のXでマトリックスを設定することがあるので、多様なジェンダーを取り入れる場合にもこの方法が使えることになる。

男性・女性のほかに「ノンバイナリー」というカテゴリーも加える。そのほかの呼び方もあり得るし、将来的にはさらに詳細なカテゴリー分けが一般的になっている可能性もある。

ジェンダー・コード・マトリックス１

	女子	男子	ノンバイナリー
Xgender1	1	0	0
Xgender2	0	0	0
Xgender3	0	0	1

あるいは以下のように男子にも1をアサインするマトリックスもあり得る。

ジェンダー・コード・マトリックス２

	女子	男子	ノンバイナリー
Xgender1	1	0	0
Xgender2	0	1	0
Xgender3	0	0	1

ふたつのうちどちらがいいかは分析者の裁量である。ジェンダー・コード・マトリックス１を取り入れると「男子に比べて女子は＋〇点になる」「男子に比べてノンバイナリーと答えた生徒は＋〇点になる」というシンプルな説明となる一方で、Xgender2に傾き（Coefficient）は算出されないことになる。一方で、ジェンダー・コード・マトリックス

2 だと Xgender2 にも傾きが算出されるが、「全体の平均点に比べて、女子は＋〇点、男子は−〇点、ノンバイナリーは＋〇点」という説明になる。ただしジェンダー・コード・マトリックス２だと多重共線性の問題が発生するリスクがあるので注意する必要がある。

こうした複数のジェンダーを取り入れた例が、教育支援事業のインパクト評価でも見られるようになってきた。まずは国際協力の分野で見られるようになったが、今後は各国の国内の事業や施策のインパクト評価や社会的インパクト評価でも普及していくと思われる。

回答選択肢に複数のジェンダーを取り入れた質問の例

（出所）著者（佐々木亮）が参加した教育介入事業のインパクト評価

この提案3の例は、ジェンダーの多様性を回帰分析に取り込むという例であった。

「回帰分析は過去の現実の傾向を再現するだけで、そもそも過去の現実にバイアスがかかっている場合にはそのバイアスも再現してしまう」と言われることがある（注：バイアス＝偏見）。回帰分析がその限界を乗り越えて、人間社会の改善に貢献するためには、人間社会として望ましい価値を分析作業の中に取り込んでいくことが必要である。

今回の提案は、ジェンダーの多様性を回帰分析に取り込むという例であったが、今後も人間社会として望ましい価値を取り込むことにより、回帰分析が発展していくことを期待している。

（謝辞）この提案は次の方々との議論に多くを負っている。この場を借りて御礼申し上げる。ただし提案の責任の一切は著者（佐々木亮）にある。米原あき（東洋大学教授）、村瀬公胤（（一社）麻布教育ラボ）、日本評価学会-社会実験分科会および価値判断のあり方研究分科会の皆様

あとがき

インパクト評価が普及し始めていることはたいへん喜ばしいことです。このレポートに掲載されたインパクト評価のデザインと事例を見て、さらに多くの事例を学んでみたいと思われる方もいらっしゃるでしょう。そのための論文のデータベースを二つ紹介したいと思います。

1番目は、アメリカのマサチューセッツ工科大学（MIT）が運営する The Abdul Latif Jameel Poverty Action Lab （J-PAL）、通称、「貧困アクションラボ」です。同ラボでは、インパクト評価のデザインとして、最も厳格なデザインであるランダム化比較試験（RCT）しか使わないと宣言して、現在までに 2,000 件近くのインパクト評価を実施して、論文を発表しています。このレポートでも、貧困アクションラボの論文からいくつかを採用させていただきました。以下のサイトから、論文を検索してダウンロードすることができます。（https://www.povertyactionlab.org/evaluations）

2番目は、International Initiative for Impact Evaluation、通称、3ie です。RCT だけではなくその他のインパクト評価デザインを用いたインパクト評価の報告書を網羅しており、合計 4,000 以上の論文を検索してダウンロードすることができます。（https://developmentevidence.3ieimpact.org/）

本レポートによって、読者諸氏のインパクト評価の理解と、インパクト評価の新しい取り組みに貢献できれば私にとってたいへんな喜びです。最後に以下の言葉を記しておきます。

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

---評価は社会改善だ。そしてエビデンスなくして社会改善なし。

＊　　＊　　＊　　＊　　＊

最後に、本を出版する時はいつもそうですが、執筆時にリスニングしていた音楽を記載したいと思います。書籍を参考文献にあげるなら、音楽に対しても同じリスペクトがあってもいいでしょう。

Avicii. *Without You ft. Sandro Cavazza; Dear Boy ft. 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

佐々木亮／Ryo SASAKI

著者略歴

佐々木亮／Ryo SASAKI

職　歴

国際開発センター（IDCJ）評価部　主任研究員。
静岡県立大学国際関係学部非常勤講師(2024～)。（特活）ソーシャルバリュージャパン理事(2024～)。
立教大学大学院21世紀社会デザイン研究科兼任講師、
大阪大学グローバルコラボレーションセンター非常勤講師、
名古屋大学大学院法学研究科非常勤講師、
聖心女子大学文学部人間関係学科非常勤講師など歴任。
2023年度は、早稲田大学国際教養学部および国際医療福祉大学でインパクト評価/統計分析の講義を実施。2024年度から、静岡県立大学国際関係学部非常勤講師（援助プログラム評価論）。

学　歴

ウェスタンミシガン大学　評価研究所　評価学博士　（Ph.D.）
ニューヨーク大学　ワグナー公共行政学大学院　公共行政学修士（公共政策分析）（M.P.A.）
日本評価学会　奨励賞（2007）、日本評価学会　功績賞（2023）

著　書

以下の著書があるほか、学術論文多数あり。
『入門評価学』(2014、C.H.ワイス（著）、佐々木（翻訳監修））、前川美湖、池田満　（監訳）（日本評論社）
『サクセスケース・メソッド』(2022、ブリンカーホフ.O.R、佐々木（翻訳）（多賀出版）
『協働評価ステップ・バイ・ステップ』(2022, リリアナ・ロドリゲス・カンポス他（著）、佐々木亮（翻訳）（多賀出版）
『評価論理：評価学の基礎』(2010)（多賀出版）
『政策評価の理論と技法』(2000, 2004)（共著、多賀出版）
『政策評価トレーニング・ブック：7つの論争と7つの提言』(2003)（多賀出版）
『エクセルで政策評価：すごくよくわかる実践的統計マニュアル』(2007)（多賀出版）
『大学の戦略的マネジメント』(2005)（共著：多賀出版）
『戦略策定の理論と技法』(2002)（共著：多賀出版）

最近の仕事

- ネパール　「モニタリング・評価システム強化プロジェクトフェーズ2(SMES2)」（2011-2015、総括）
- ヨルダン　「ヨルダンにおけるシリア難民への平和の創出に係るインパクト評価」（2020-2022、総括／インパクト評価）（報告書：英文、和文、JICAインパクト評価サイトに掲載）
- パレスチナ　「理数科教育　質の改善プロジェクト（本格活動実施フェーズ）」（2021-2023、インパクト評価（理数科学力テスト））
- ミャンマー　「初等教育カリキュラム改訂プロジェクト」（2018-2021、インパクト評価）
- マラウィ（アフリカ）　「村落給水における社会的インパクト調査」（2016、総括／社会的インパクト評価）
- アフガニスタン　「アフガニスタン人道危機対応支援プログラム」(2020)のモニタリング・評価事業のアドバイザー業務

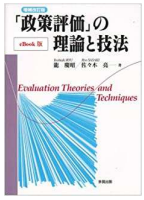
*1～5番目はいずれもJICA委託事業。6番目はJapan Platform (JPF)委託業務。その他に、日本外務省、世界銀行グループIFC、アジア開発銀行(ADB)、国連開発計画（UNDP）など国際機関からの調査委託の経験多数。

関連書籍と研修のご案内

＜学会誌＞

	**日本評価研究　第20巻　第2号（2020年7月）** 『特集：「エビデンスに基づく政策立案（EBPM）」の普及の現状と課題』 論文7本　（編集担当：佐々木亮、正木朋也） http://evaluationjp.org/files/Vol20_No2.pdf
	**日本評価研究　第17巻　第1号（2016年11月）** 『特集：評価における科学性：エビデンスの実践的活用とその方向性』 論文4本　（編集担当：佐々木亮、正木朋也） http://evaluationjp.org/files/Vol17_No1.pdf
	**日本評価研究　第10巻　第1号（2010年3月）** 『特集：エビデンスに基づく実践の世界的動向と日本における取り組み』 論文4本　（編集担当：佐々木亮、大島巌） http://evaluationjp.org/files/Vol10_No1.pdf
	**日本評価研究　第6巻　第1号（2006年3月）** 『特集：エビデンスに基づく評価の試み』 論文4本 （編集担当：佐々木亮） http://evaluationjp.org/files/Vol06_No1.pdf

＜書籍とeBook＞

 龍慶昭、 佐々木亮	**『「政策評価」の理論と技法』(eBook版)**（2000, 2004, 2020eBook版） https://www.amazon.co.jp/gp/product/B08DFN2L91/ref=as_li_tl ・日本における評価理論のベストセラー。 ・セオリー評価、プロセス評価、インパクト評価、費用対効果評価のワンセットで評価することを解説。 ・「ロジックモデル」という用語と概念を日本で初めて解説。 ・社会セクターにおけるRCTの適用事例を日本で初めて紹介。
 佐々木亮	**『評価論理：評価学の基礎』(eBook版)**　（2008,　2020eBook版） https://www.amazon.co.jp/dp/B08D6TM1FY/ref=as_sl_pc_qf_sp_asin_til ・『評価＝事実特定＋価値判断』という基本概念を初めて日本に紹介した評価学の基礎をなす著作。 ・評価における価値観の取扱、価値判断の仕方、客観性の考え方、内的妥当性の概念、外的妥当性の概念等を丁寧に解説。 ・評価学における主要な論争を網羅して解説。
 佐々木亮(翻訳)	**『サクセスケース・メソッド』**(2022、ブリンカーホフ.O.R(著)、多賀出版） **eBook版**と**印刷版**があります。アマゾンで「サクセスケース・メソッド」と検索してください。（https://www.amazon.co.jp/） ・アメリカの企業研修評価のベストセラーの和文翻訳版。 ・SONY、HONDA、アマゾンなどの国際的な大企業、世界銀行などの国際機関、スタンフォード大学などの大学組織が採用しています。 ・「翻訳者から：日本社会での適用に向けて」のなかで『エビデンスに基づく実践とはどういう関係になるか？』を明快に解説。
 佐々木亮(翻訳)	**『協働評価ステップ・バイ・ステップ』**（2022、ロドリゲス=カンポス,R他（著）、多賀出版） **eBook版**と**印刷版**があります。アマゾンで「協働評価」と検索してください。（https://www.amazon.co.jp/） ・評価の最新潮流である「協働評価」を平易に解説した実務書。 ・「補論：日本社会での適用に向けて」のなかで、『エビデンスに基づく実践とはどういう関係になるのか？』『第三者評価とはどういう関係になるのか？』を明快に解説。

＜研修＞

『プロフェッショナル統計分析ワークショップ』
「誰でも必ずわかる」と大好評!!

受講者の声１「本当に”たし算・ひき算・かけ算・わり算”だけで説明しきった。感心した。」
受講者の声２「論文を読めるようになったのがうれしくて、今は読みまくっています。」

【講師】
佐々木亮（ウェスタンミシガン大学評価学博士（Ph.D.））
高木桂一（スタンフォード大学社会学博士（Ph.D.））
他のゲスト講師
【主催】 国際開発センター（IDCJ）評価部
【開催頻度】 ２ヶ月に一回開催（日本語で開催の場合と、英語で開催の場合あり）
【概要説明】
本レポートで解説した統計分析手法が一通り身につきます。 平均値と標準偏差の計算、2群の平均値差の検定、回帰分析、サンプルサイズの考え方、RCTの手順、論文の読み方、実体験に基づく実際の勘所などを解説します。
参加のための事前条件は、足し算・引き算・掛け算・割り算ができることと、エクセルを日常的に使用していることです。学問的な精緻さよりも、実務でどのように使えるかに重点を置きます。本ワークショップの修了者には、修了証が交付されます。
【ウェブサイト】
https://www.idcj.jp/seminar/statistical-analysis-workshop.html

『プロフェッショナル統計分析ワークショップ：応用編』
「誰でも必ずわかる」と大好評!!

上記コースと同様に、（1）例題の解説、（2）原理の説明、（3）簡単な手計算、（4）演習問題の実施、（5）学術論文のディスカッションと解説、の流れで進めます。4コマを用意しております。オンデマンドでのご提供で、手続き後はいつでも受講できます。

【コース】
応用コース1： インパクト評価の最新テクニック（DID, PSM, IV）
応用コース2： 構造方程式モデリング（SEM）
応用コース3： インパクト評価のためのサンプルサイズの計算
応用コース4： メタ分析（システマティック・レビュー）の計算
【開催頻度】 Zoom録画ビデオによるオンデマンド実施（いつでも受講できます）
【インストラクター】
佐々木亮（ウェスタンミシガン大学評価学博士（Ph.D.））
他のゲスト講師
【主催】 国際開発センター（IDCJ）評価部
【ウェブサイト】
https://www.idcj.jp/pickup/grow/statistical-analysis-workshop-advanced.html

『プロフェッショナル統計分析ワークショップ：応用編：STATAによるデータ分析の演習』
テキスト入力ゼロで、メニューバーから選択＆クリックのみで実施します

メニューバーから選択＆クリックのみの操作方法と出力結果の読み方を演習します。
【開催頻度】 ２ヶ月に一度実施
【インストラクター】
佐々木亮（ウェスタンミシガン大学評価学博士（Ph.D.））
【主催】 国際開発センター（IDCJ）評価部
【ウェブサイト】（日米のサイトに開催日時が掲載されます）
（IDCJ のサイト）https://www.idcj.jp/seminar
（米国 Stata 本社のサイト）https://www.stata.com/meeting/short-courses/#japan

無料でご視聴いただけるユーチューブチャンネルのご紹介

IDCJ評価部「プロフェッショナル統計分析ワークショップ」Tipsシリーズ

★広報ビデオ★『Tips1 統計分析のかんどころ』（７分）
https://youtu.be/fu6taNR_jsA
★広報ビデオ★『Tips2 インパクト評価の５つのデザイン』（14 分）
https://youtu.be/neLly_7hv00
★広報ビデオ★『Tips3　論文の読み方１：インパクト評価の論文』（12 分）
https://www.youtube.com/watch?v=OMw7D0vkSIA
★広報ビデオ★『Tips4　論文の読み方２：回帰分析の論文』（12 分）
https://www.youtube.com/watch?v=1gychUeH6zY
★広報ビデオ★『Tips5　最小のサンプルサイズとは』（10 分）
https://www.youtube.com/watch?v=y3IvlvmHvvU
★広報ビデオ★『Tips6　論文事例１ エルサルバドル教育』（14 分）
https://www.youtube.com/watch?v=jZxrC0fMYCg