



Title	プログラム評価は教育支援の改善にどう貢献できるか ： 支援プログラム評価のデザインの重要性
Author(s)	佐々木, 亮
Citation	母語・継承語・バイリンガル教育（MHB）研究. 2013, 9, p. 1-20
Version Type	VoR
URL	https://hdl.handle.net/11094/50825
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

≪ 招待論文 ≫

**プログラム評価は教育支援の改善にどう貢献できるか
— 支援プログラム評価のデザインの重要性 —**

佐々木 亮 (国際開発センター)

sasaki.ryo@idcj.or.jp

**How Program Evaluation Can Contribute to Improvement of
Education Support Programs?**

— Importance of Good Evaluation Designs for Education Support Programs —

SASAKI Ryo

キーワード：評価、プログラム評価、価値判断、セオリー評価、インパクト評価

要 旨

本論文では、(1) 評価の基本論理 (評価と調査は何か違うのか? など) を明らかにしたうえで、(2) 評価の基本的作業 (基礎作業と総合化作業、評価の6ステップなど)、(3) 評価の基本項目 (標準的な4評価項目など) について比較的詳細に解説する。そして(4) 評価に関する最近の傾向を解説し、最後に、(5) 「プログラム評価の実践に向けて」と題して、教育支援施策のプログラム評価に関する勘所を解説する。本論文が、教育支援プログラムに関するプログラム評価の普及に貢献することになることを期待している。

Abstract

This thesis review and discuss the following themes: (1) Basic logic of evaluation (what the difference between research and evaluation ? etc.); (2) Basic procedures of evaluation (primary operation and secondary operation, 6 steps of evaluation, etc.); (3) Basic component of evaluation (theory evaluation, process evaluation, impact evaluation and efficiency evaluation, etc.). Then (4) the current discussion in evaluation study is reviewed and discussed. Finally, some tips for practicing program evaluation in education field are indicated. It is hoped that this thesis contribute to dissemination of practice of program evaluation in the field of education support program.

1. プログラム評価ⁱ⁾は何の役に立つ？

教育分野において「プログラム評価」の重要性が増している。プログラム評価は、さまざまな教育施策（プログラム）を体系的な方法と手続きを用いて評価することにより、以下のような質問に答えようとすることを目指していると理解できる。

- あなたの教育支援活動は、論理的だったのか？（論理が飛躍していた部分はなかったのか？）
- あなたの教育支援活動は、当初計画通りに実施されたか？（計画からずれた部分はなかったか？）
- あなたの教育支援活動は、本当に効果があったのか？（他の要因によるのではないだろうか？）
- あなたの教育支援活動は、貴重な税金（科研費を含む）を使って実施するだけの価値があったのか？（最大の効果を最小限の費用で実現することが望ましいということとでいわゆる「費用対効果」（あるいは「費用-効率性」）に関する質問である）

そしてプログラム評価の結果を用いて、一般的には次の目的に貢献することが期待される（龍・佐々木2000, 2004）。

- (1) 大きな意思決定：どのプログラムを継続して、どのプログラムを開始し、どのプログラムを廃止するかを決定する。
- (2) 小さな意思決定：あるプログラムの実施過程を改善する。
- (3) 説明責任の遂行：プログラムの資金を負担する納税者に価値があることを説明する。

これら (1) ～ (3) の目的に貢献することにより、よりよい教育施策の選択と実施および納税者への説明を実現し、ひいては社会の改善に貢献することが期待されている。

本論文では、最初に、評価の基本論理（評価と調査は何が違うのか？など）を明らかにしたうえで、評価の基本的作業（評価の6ステップなど）、および評価の基本項目（標準的な4評価項目など）について詳細に解説する。さらに、評価に関する最近の傾向を解説し、最後に、「プログラム評価の実践に向けて」と題して、教育支援施策のプログラム評価に関する勘所を解説する。本論文が、教育支援プログラムに関するプログラム評価の普及に貢献することになることを期待している。

i) 評価には、政策評価 (Policy evaluation)、プログラム（施策）評価 (Program evaluation)、プロジェクト（事業）評価 (Project evaluation)、人事評価 (Personnel evaluation)、製品評価 (Product evaluation) などの種類があるが、その中で一番メジャーなのがプログラム（施策）評価である。

2. 評価の基本論理

2.1 プログラムの定義

そもそも、プログラム評価の「プログラム」とは何であろうか。プログラムは、「ある社会状況を改善するために、ひとつのあるいはいくつかの目的に向けて組織化された諸資源および行動」ⁱⁱ⁾ と定義されることがある。また、札野 (2011) は Smith (1989) を引用して「特定の、あるいは特定可能な対象者に規定の変化をもたらすために計画された一連の活動」と定義している。したがって、教育支援プログラムとは、「特定の生徒の教育状況の改善を支援するために計画された一連の行動」と定義することができる。また、評価の対象となる教育支援プログラムには、政策レベルの支援と、学校現場で日々行われている実践レベルの支援の双方をも含めて考えることができる。なお、プログラムは日本語では「施策」と翻訳されることが多いことにも注意が必要である。そして、もともとは、教育、保健医療、社会福祉などの特定分野のプログラムを評価対象としてきたが、近年、プログラム評価の概念と手法は対象範囲を広げて次第に一般化し、現在は「プログラム評価」(Program evaluation) と呼称せずに、単純に「評価」(Evaluation) と呼称することも多いⁱⁱⁱ⁾。

2.2 評価の定義

さて次に、プログラム評価の「評価」とは何であろうか？そもそも調査と評価は何か違うのだろうか？じつはこの点に関して、評価研究者の間には現在でも議論が続いている。プログラム評価は長い歴史を有するが、特定の学問領域として認識されるようになったのはアメリカにおいて1960年代になってからである。それ以来、未だに自分たちの研究対象に関して合意を見ていないという驚くべき状況であるが、それでも頻繁に引用される定義がいくつかある。

評価研究の世界で最も売れた本である Rossi & Freeman の Evaluation: A Systematic Approach (邦題:『プログラム評価:体系的アプローチ』) に掲載されている定義は次のとおりである。「評価調査 (Evaluation Research) とは、社会介入施策

ii) Rossi & Freeman (1999), p20; Weiss (1997), p4~6; Wholey, Hatry & Newcomer (1994), p3; OECD(1999), p12 のそれぞれの Evaluation の定義を参考に筆者が作成した定義 (龍・佐々木2000)。

iii) 例えば、アメリカ評価学会が発行する学会誌は、以前、New Direction for Program Evaluation という名称であったが、現在は New Direction for Evaluation と改称されているのがこの状況を端的に表している。

の目的、介入理論、実施過程、効果、効率性を体系的に明らかにするために社会調査手法を使用すること^{iv)}」(p.5)と定義している。

評価研究の世界で2番目に売れた本と言われる Weiss の Evaluation (2nd edition) では次のように定義されており、この定義の方が適切だとする研究者も多い。「評価は、プログラムや政策の実施もしくは成果を、明示的あるいは暗示的な対象と比較しながら、体系的に明らかにすることであり、プログラムもしくは政策の改善に資するものである」^{v)}。

Rossi & Freeman は、目的、介入理論、実施過程など社会施策を構成する要素、および効果、効率性(=純社会便益)など社会施策に関係する事実を明らかにする(assess)のが評価だとしている。これに対して Weiss は、①社会施策がもたらす事実を明記して、②明らかになった事実を何かと比較してうまくいったのかいかなかったのかの判断を下し、③さらに社会施策の改善へのフィードバックがなされるという3段階の要素で評価が構成されるとしている。Weiss の定義は、Rossi & Freeman の定義よりも幅広く、また単なる事実の認定を越えて、何らかの価値判断を下さねば評価足りえないと明言していると言える。

さらに、評価研究の創始者のひとりと目されたり評価研究の唯一の哲学者と言われることもある Scriven は、評価を次のように定義している。「評価は、体系的かつ客観的に、物事の価値を明らかにすること^{vi)}」と定義した(Scriven, 1980)。つまりスクリヴェンは、評価とは価値判断(value determination)を行うことだということを、

iv) Rossi, P.H. and Freeman, H.E. (1993). Evaluation: A Systematic Approach (5th Ed.). p5. "Evaluation research is the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs." そして、Assessment of Program Utilityとして、Impact Assessment と Efficiency Assessment をあげているので (p.36-37)、定義の中の Utility を Impact と Efficiency で置き換えた。ちなみに Rossi は、マサチューセッツ大学の社会学大学院の教授である。

v) Weiss, C.H. (1997). Evaluation (2nd Ed.). p4. "Evaluation is the systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy." ちなみに Weiss は、ハーバード大学教育大学院の教授である。

vi) Scriven, M. (1980). The Logic of Evaluation. preface. (Systematically and objectively determining the merit, worth, and/or significance of something). その定義のなかの、merit, worth and/or significance of somethingを単純に「物事の価値」と訳した。なお、本論文では、merit, worth and/or significance の3種類の価値を総称して、価値(value)という用語を使用した。ちなみに、Scriven は、ウェスタンシガン大学評価研究所の教授だった(現在は、クレアモント大学に所属)。

この定義で端的に示していると言える。

最近の議論をレビューすると、評価はさまざまな方法を取り得るしさまざまな目的に貢献することは認めるが、いずれにせよ対象に関して何らかの価値判断を下すのが評価だというのが近年の合意だと言っていいであろう (Julnes, G. (2012), p.3.; Mark, Henry and Julnes (2000), Shadish, Cook & Leviton. (1991). p.49.)。

2.3 評価の基本構造

ここまでの議論から、調査と評価は本質的に違う点があることが分かる。調査の目的は、因果関係を含む物事の現象を明らかにすることである（＝事実特定）と考えることができる。一方、明らかにされた事実に関して何かしら価値に関する判断（＝価値判断）を下して初めて、評価足りえると言える。つまり、その言葉が示すとおり評価とは、「価値を評する」ということであり、英語の Evaluation もやはり「価値-付ける」(Value-ation) という意味である。以下のように表現されることもある^{vii)}。

評価	=	事実特定	+	価値判断
(Evaluation)		(Factual identification)		(Value determination)

2.4 評価学の未来

評価学は、統計学や倫理学などと同じく分野横断的学問 (Transdiscipline) として成長している (Scriven, 1991)。つまり、独自の学問領域として成立しながら、各分野で横断的に使われる基礎的分析道具として受け入れられつつある。既存の学問では、統計学や倫理学がそうである。

統計学は、教育学、社会学、経済学、医学など幅広い分野で基礎的な分析道具として以前から普及している。同様に倫理学も、教育倫理、社会調査倫理、生命倫理など各分野で必須の教育項目となっている。評価学も、同様に各分野で一般的に用いられる基礎的な道具として普及していくであろう。現に、教育評価、保健医療評価、経済評価などの名前でかなりの普及が見られる。

一方で、統計学や倫理学は自前の学問分野としても確立されて発展を続けている。その発展は、純粋に理論的な発展を追求する能動的な方法論と、各分野からの要求によって反応的に発展する二つの方法論に支えられている。評価学も同様に、能動的な

iv) Scriven (1999)は、Evaluation = Factual premises + Value premisesと表現している。

方法論と反応的な方法論の双方によって発展していくはずである。つまり、評価学は、各学問分野に対して共通かつ基礎的な道具として提供されるとともに、独立した学問分野として発展していくと考えられる。

3. 評価の基本的作業：基礎作業と統合作業

評価の本質は価値判断であると述べたが、価値判断のための作業として、大きく分けると「基本作業」（あるいは一次作業）(Primary operation) と「総合化作業」（あるいは二次作業）(Secondary operation) がある (Davidson 2005。そもそもは Scriven (1991) の解説)。それぞれについて簡単に解説する。

3.1 基礎作業

基礎作業は、あらかじめ特定された価値の側面に関して、価値判断を下す作業である。具体的な手法としては、等級付け (Grating)、点数付け (Scoring)、順位付け (Ranking) がある (Davidson, 2005)。なお、これらの手法は、生徒個人を対象とした評価の手法のように感じられるかも知れないが、より広く、プログラム (施策) を対象とした評価であっても普通に用いられている手法である。

等級付けはいわゆる絶対評価であり、点数付けも適切に用いれば絶対評価の手法として使える。一方、順位付けは、ある集団の中の位置を示すに過ぎない相対評価であり、価値判断の方法としては弱いと言わざるを得ない。以下の表がそれぞれの具体的な説明となる。

なお、等級付けのためには、価値の項目 (Criteria of value) と価値の水準 (Standards of value) で構成される以下のような評価の枠組を取りまとめておくとう分かりやすい。

(i) 価値項目 (Criteria of value)					
(ii) 価値水準 (Standards of value)		妥当性 介入理論の 妥当性	適切性 実施過程の 適切性(計画と 実績の比較)	有効性 社会状況の 改善効果の 実現度合	効率性 投入資源の利用の 効率性(投入資源と 改善効果の比較)
	A (たいへん満足)	(定義文)	(定義文)	(定義文)	(定義文)
	B (満足)	(定義文)	(定義文)	(定義文)	(定義文)
	C (どちらとも言えない)	(定義文)	(定義文)	(定義文)	(定義文)
	D (不満足)	(定義文)	(定義文)	(定義文)	(定義文)
	E (たいへん不満足)	(定義文)	(定義文)	(定義文)	(定義文)

図 1 評価の枠組 (等級付けのための枠組)

表 1 評価の基礎作業（評価の一次作業）

等級付け (Grading)	測定された実績をもとに、評価対象に、価値 (value) に関するカテゴリーを割り当てる作業。最も多く用いられるカテゴリー分けは、A＝優 (excellent)、B＝良 (satisfactory)、C＝可 (acceptable)、D＝不良 (poor)、E＝不可 (unacceptable) である。等級は、相対的評価ではなく、被評価物あるいは被評価人の本来的な価値 (本質) に関する絶対的評価である。
点数付け (Scoring)	測定された業績をもとに、評価対象に、点数を付与する作業。例えば、研究職の応募者に関する点数付けでは、書評 (2点：0～2)、論文 (3点：0～3)、学会発表 (3点：0～3)、出版 (5点：0～5) などの点数を配分して、点数付けをする場合などがこの作業にあたる。これらの点数付けは、多くの場合、絶対評価として用いられる。ただし、評価対象間の順位を決定するためにも散られる場合もある (相対評価となる)。
順位付け (Ranking)	測定された業績をもとに、評価対象を順位付けをする作業。例えば、10人の評価対象に対して、1位～10位の順位付けをする場合などがこの作業にあたる。順位付けは相対評価であり、本来的な価値 (Merit) を明らかにする試みではない。

絶対評価と相対評価を考えるうえで示唆に富む例が日本の教育界に見られる (佐々木, 2010a)。日本の教育界において、従来の相対評価をやめて絶対評価を導入しようという試みがなされている。絶対評価は、上記で解説した「等級付け」を用いるということである。つまり、あらかじめ、それぞれの等級を定義した一覧表を用意して、それぞれの生徒がそのうちのどれに当てはまるかを判断するということである。相対評価は、上記で解説した「順位付け」を用いるということである。つまり順位などで示すので、集団の中の相対的な位置を示すに過ぎない。しかし、日本の教育分野では、上位 30 % は「A」「優」、次の 30 % は「B」(良)、次の 30 % は「C」(可)、最後の 30 % は「D」(不可) といったように、集団の相対的位置に対して等級付けの言葉が用いられてきた。これは等級付けと順位付けの混乱の例である。この状況を改善すべく、等級付けの論理を正確に適用する絶対評価の導入が試みられたわけである。しかし、等級付けを用いる際には、あらかじめ等級を表す定義一覧表が用意されねばならないが、その策定や合意が不十分だったことが、導入直後に見られた現場の混乱につながったと思われる。

3.2 総合化作業

複数の評価項目を設定した場合に、得られた評価結果を総合化 (synthesis) して単一の評価結果を出す必要がある。総合化作業に用いる技法は以下のとおりである

(Davidson 2005)。総合化作業を経て出された評価結果は、「総合評価結果」と呼ばれる。繰り返しになるが、これらの手法は、生徒個人を対象とした評価の手法のように感じられるかも知れないが、より広く、プログラム（施策）を対象とした評価であっても普通に用いられている手法である。

表2 評価の総合化作業（評価の二次作業）

加重・合計 (Weighting) (部分評価→ 総合評価)	部分評価のそれぞれの評価結果を、被評価物や被評価人に関する単一の総合評価へまとめ上げる作業を行う。数量的加重・合計作業 (Numerical Weight and Sum: NWS) は、各小項目に1～5点など数字による重み付けを行って総合化するのに対して、定性的加重・合計 (Qualitative Weight and Sum: QWS) では、「必須」(Critical)、「重要」(Important)、「(あれば) 望ましい」(Desirable) などの言葉による重み付けを行って総合評価を試みる。
最低基準付け (Barring)	部分評価の特定の項目に関して最低点を決めて、それに満たない場合には総合評価結果も自動的に低くする作業。
段階付け (Stepping)	段階付けは、加重よりも洗練された作業である。ある側面に対する加重を一律に3点とするのではなくて、「A」であれば3点、「B」以下であれば2点とする。当該側面に関するA、B、C、Dなどの等級付けの違いに、この段階付けの違いが掛け合わされるため、より洗練された総合評価結果になると考える。

(注) 通常用いられない「倍化検討」(scaling)、側面描写 (profiling) は省略した。

また、総合化作業の特殊な例としてフローチャートを作成して用いるという方法がある。フローチャートは、次の例が示すように、加重・合計と最低基準付けを同時に表現する優れた方法であり、各分野の評価に置いて積極的な利用が勧められると筆者は考えている。

3.3 評価の基本的な6ステップ

基礎作業と総合化作業の考え方を解説したが、それでは実際にはどのような手続きを経て価値判断が行われるべきなのであろうか。佐々木(2010a)は、評価は事実特定＋価値判断である、というScrivenの解説に基づき、以下のように6ステップにまとめている。

- (1) 評価の対象と評価の目的の特定 (Determination of evaluand and evaluation purposes)

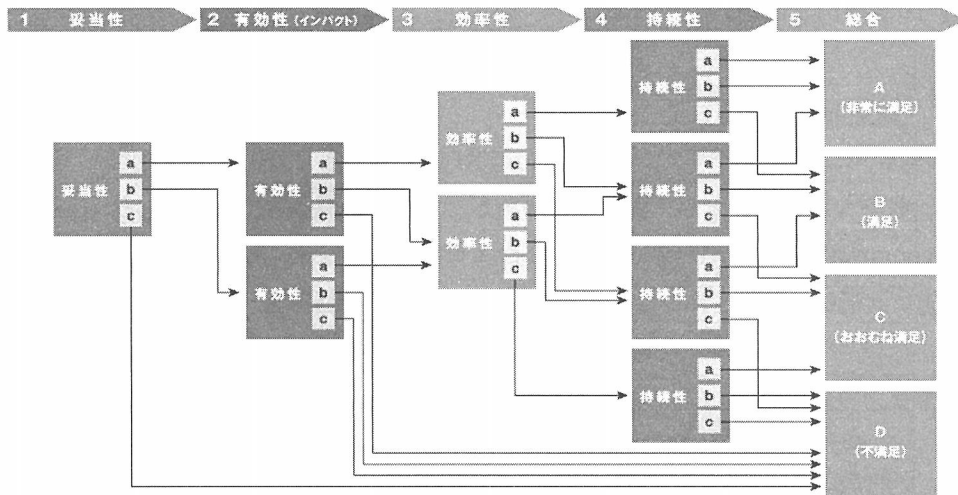


図2 特殊な総合化の手続き：フローチャートの利用（国際協力銀行 2006）

- (2) 価値項目の設定 (Setting of criteria of values)：どのような側面が評価されるべきであろうか？
- (3) 価値水準の設定 (Setting of standards of values)：良くやっているとみなせる水準とはどの程度か？
- (4) 事実の特定 (Factual Identification)：それぞれの価値項目に関する実績 (performance) はどれくらいか？
- (5) 価値判断の実施 (Value determination)：それぞれの価値項目に関してどれくらい良くやっていたか？
- (6) 価値判断の統合化作業の実施：(Synthesis of values)：それぞれの価値項目に沿った価値判断を、ひとつの価値判断に統合する。結局、一言でいうと（その施策には）価値があったのか？

なお、価値項目と価値水準で構成される評価の枠組（等級付けのための枠組）を事実特定の前に設定することが原則ではある。ただし、実際の評価作業においては、事実特定の作業の中で、その枠組が明確になったり変更を余儀なくされることがあり、それは避けられないことである。したがって、一度作成して適用した評価の枠組を、他の評価でも適用して洗練させていき、当該分野で用いられる標準的な枠組とすべく努力していくことが、いずれの分野においても勧められる。

4. 評価の基本項目

4.1 標準的な評価項目

評価の項目は、その時々で個別に十分に検討されるべきであるが、ある程度合意された一般的な評価項目として次の項目がある。なお、これはすでに紹介した Rossi & Freeman (1999)、Weiss (1997) および Wholey, Hatry & Newcomer (1994) などで共通的に解説されている項目である。ただし、Rossi & Freeman はこれ以外に、「ニーズ査定」(Needs assessment)を加えている点に注意が必要である。また、国際協力の分野ではさらに、「自立発展性」(Sustainability)を加えることが多い。

- － セオリー評価 (介入理論の妥当性の評価)
- － プロセス評価 (実施過程の適切性の評価)
- － インパクト評価 (介入効果の有効性の評価)
- － 効率性評価 (費用と有効性の比較による効率性の評価)
- － 総合評価 (上記4項目の総合評価)

ここではこれらに関して若干詳細に解説する。

4.2 セオリー評価 (介入理論の妥当性の評価)

英語では、Program Theory Evaluation と言い、それを短縮して「セオリー評価」と呼称するのが日本では一般的である (龍・佐々木, 2000)。すべてのプログラム (施策) はなんらかの「介入理論」(セオリーあるいはロジック) にもとづいている。セオリーとは、原因と結果が連鎖上に連なる「仮定」である。セオリー評価では、その介入理論 (セオリー) を明らかにして (=事実特定)、その妥当性を評価する (=価値判断)。セオリー評価の結論に使われる等級は、「a: たいへん妥当だった」「b: 妥当だった」「c: どちらとも言えない」「d: 妥当だったとは言えない (論理に飛躍があった)」などである。

セオリー評価では、「風が吹けば桶屋が儲かる」と言ったこじつけや論理の飛躍がなく、首尾一貫して実現の確実性の高い論理構成となっているか (いたか) を検証する。以下のように、いわゆるロジックモデル (Logic model)、別名プログラムセオリー (Program theory) を書いて検証することが勧められる。

ロジックモデルでは、単なる活動結果であるアウトプット (Outputs) と、そのおかげで引き起こされる社会状況の変化であるアウトカム (Outcomes) を峻別して構築することが重要である。例えば、ある教育施策に関して言えば、卒業生数は単なるアウト

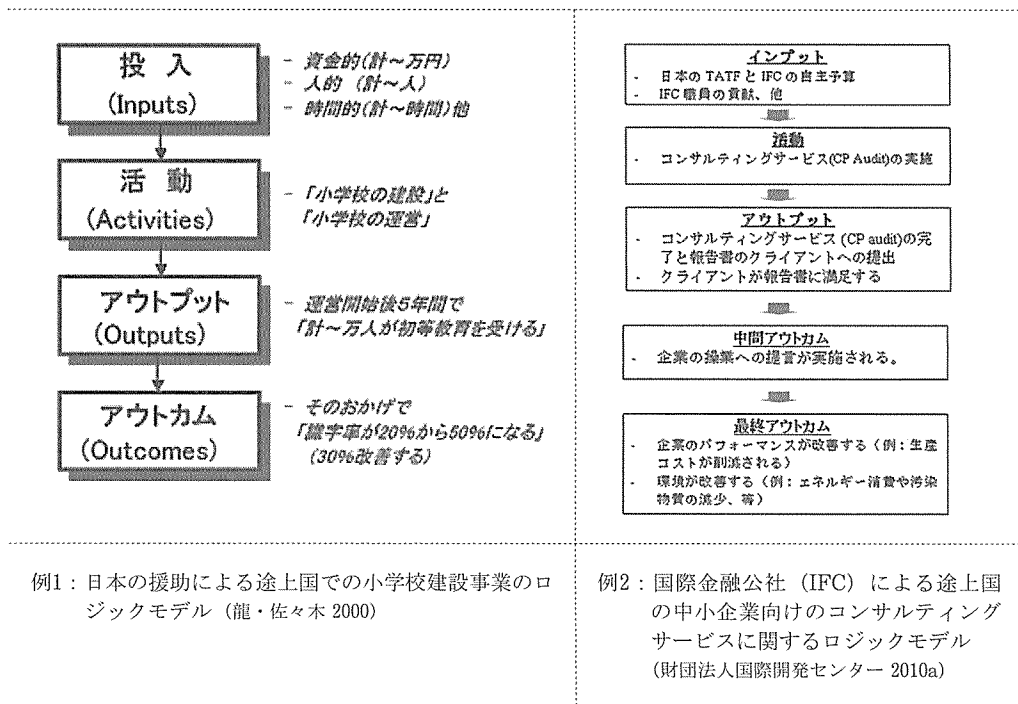


図3 ロジックモデルの例

ットであり、その後、得られた専門技術を適用して仕事をするることにより、卒業生が自己実現を図ったり世の中の状況を改善したりすることがまさにアウトカムということになる。別添1の教育施策の例（アメリカの The Cooperative State Research, Education, and Extension Service (CSREES)）では、アウトカムが、短期・中期・長期の3段階に分けて特定されているがそれは適切である。

4.3 プロセス評価（実施過程の適切性の評価）

プログラム（施策）を、当初の計画どおりに実行するということは、一見あたり前のように思われるが、現実はそのでもない。実際のプログラムの実施はいとも簡単に当初計画からずれる。プロセス評価は、当初計画と実績を比較してその一致度合を明らかにし（＝事実特定）、実施過程がどの程度適切に実施されたかの適切性を評価する（＝価値判断）。プロセス評価の結論に使われる等級は、「a：たいへん適切だった」「b：適切だった」「c：どちらとも言えない」「d：適切だったとは言えない」などとなる。次の図のような比較表を用いることが勧められる。日本で評価と言えば、じつはこのプロセス評価（計画と活動実績の比較）だったこともあり、ここで紹介する評価項目の中で、

じつのところ日本ではたいへん馴染のある評価である。ただし、それに加えて重要なことは、その前段階において介入理論が妥当であったか（セオリー評価）と、その後段階において改善効果が本当にあがったのかどうか（インパクト評価）であり、これらプロセス評価の前後の評価作業にもっと注目する必要がある。

図4 計画値／実績値比較表

	計画値	実績値	一致度合と留意点
投 入 ・資金 ・人員 ・建物・機材			
活 動 ・活動1 [] ・活動2 []			
アウトプット ・アウトプット1 [] ・アウトプット2 []			

4.4 インパクト評価（介入効果の有効性の評価）

プログラム（施策）によって、対象とする人口や対象とする社会状況への「改善効果」（インパクト）があったのかなかったのか、あったとしたらそれは十分だったかという質問に答える評価である。内部妥当性（Internal validity）の検証でもある。教育支援プログラムの場合には、インパクトを測定する指標として、進学率、卒業率、進級率、ドロップアウト率、卒業試験の点数、期末テスト点数などが用いられることが多い。また、高校・大学などの高等教育の場合には、就職率も頻繁に用いられる。インパクト評価は、改善効果（インパクト）の有無・程度を明らかにし（＝事実特定）、それが十分に大きかったかどうかを判断する（＝価値判断）。インパクト評価の結論に使われる等級は、「a：大きな効果があった」「b：効果があった」「c：どちらとも言えない」「d：（効果はなかっただけでなく）逆効果だった」などとなる。本論文で解説する評価項目の中で、事実特定の技法がもっとも発達している項目である。インパクトの有無・程度を明らかにする方法としては、Rossi & Freeman（1999）が3章ほどを割いて詳細に解説している。もっとも厳格と言われる実験デザイン（Randomized

Controlled Trial (RCT)) から、もっとも簡便なデザインと言われる事前・事後比較デザイン (Before-after design) までを手際良くまとめている。Rossi & Freeman の解説をもとに5つの代表的なデザインをまとめた一覧表を別添として掲載したので参照されたい。ただし Rossi & Freeman (1999) は、定性的な手法に関する解説が不十分であり、ケーススタディ法、インタビュー法、直接観察法などの定性的な手法は他の著書の方が詳しい (例えば、Patton (1997)、Stake (1995) などの著書に詳しい)。

インパクトの有無の特定の事例として、単純な事前事後比較による事例を以下に掲載する。教員研修、教材の供与、校舎建設などの複合的な教育支援プログラムを世銀が支援して実施したが、そのインパクトとして、英語、数学の学力テストの点数を集計して事前-事後比較によって分析している例である。ただし、単純な事前-事後比較であり、世銀が支援したプログラム以外の外部要因の影響 (それは日本の教育支援の影響も含む) を全く取り除けていない点は批判を逃れない。

平 均 点 (全体)							
	1998		2003		Difference		t-test
初級英語	6.2		6.6		0.4		3.75***
初級数学	5.5		5.9		0.4		8.16***
上級英語	12.3		13.2		0.9		4.16***
上級数学	8.7		10.1		1.4		6.93***

***:1% significance level

事前

事後

差

有意差検定

図5 インパクトの有無の特定の事例：ガーナにおける基礎教育支援プログラム (世銀支援)
(World Bank 2004:137)

4.5 効率性評価 (費用とインパクトの貨幣価値の比較による効率性の評価)

プログラム (施策) によって実現したインパクト (改善効果) は、限られた資源 (資金的、人的、物的、時間的など) を投入して実施するだけの価値があったかどうかを評価する。効率性評価は、インパクトと資源の双方を貨幣価値に換算し (= 事実特定)、両者の差や比率が十分に大きかったかどうかを判断する (= 価値判断)。効率性評価の結論に使われる等級は、「a: (限られた資源を投入するだけの) たいへんな価値があつ

た」「b：価値があった」「c：どちらとも言えない」「d：(限られた資源を投入するだけの) 価値はなかった」などである(英語では worth-worthless の訳)。日本国内では、「a：たいへんに効率的だった」「b：効率的だった」「c：どちらとも言えない」「d：効率的ではなかった」などの言葉遣いも用いられる。ただし日本では、効率的という言葉を使った途端に、「効率的に敷材が運搬された」「効率的に事業が実施された」など、本来はプロセス評価に属する話に容易に堕することが多いことに、注意が必要である。なお、教育施策や保健施策などの社会施策に関して、インパクト評価までは実施するとしても、効率性評価まで実行すべきか、そして実行可能なのかに関してはアメリカでも議論がある(佐々木, 2002)。

事例として、シンガポールに出稼ぎに来ていたフィリピン女性を支援するための学校である「マイクロビジネススクール」(週末だけの廉価な学校)である Aidha の例を以下に掲載する(佐々木, 2012)。この事例では、Aidha が実現した各種のインパクト(改善効果)を貨幣価値換算して、投資額と比較して、社会投資リターン比率(いわゆる費用便益比率)を 7.48 倍と計算して(=事実特定)、たいへん高い効率性を実現していると結論づけている(=価値判断)。

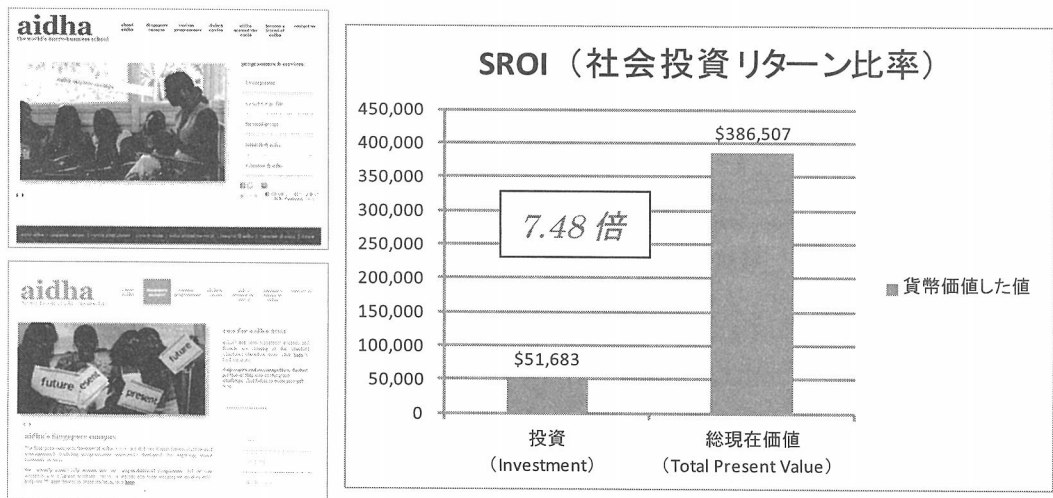


図6 効率性評価の実例：シンガポールの Aidha の教育プログラム
(Social Impact Associates 2008; Aidha のウェブサイト <http://www.aidha.org>)

4.6 総合評価

すでに述べたように、複数の評価項目を設定した場合に、得られた評価結果を総合化 (synthesis) して単一の評価結果を出す必要がある。総合化作業を経て出された評価結果は、「総合評価結果」と呼ばれる。なお、評価のあとに「提言」(Recommendation) も記載すべきかどうかはアメリカの評価研究者の間に議論があることが観察されている (佐々木, 2010a)。Scriven が述べるところによると、適切な提言をなすためには、(i) 確かな評価結果、(ii) 分野に特有の専門知識、(iii) 提言する組織の内部環境に関する知識の3つが必要である。

5. 最近の課題

評価研究の分野では様々な学術的な論争が行われている。代表的な7つほどの論争をまとめた解説もあるが (佐々木, 2002)、最近になって勃興してきた新しい論争として「エビデンスに基づく評価」(Evidence-based evaluation) に関する論争がある。いわゆる実験デザイン (RCT) を用いた事実特定をもっと促進すべきだという主張で、1980年代の実証主義と構築主義の論争以来の熱い論争が交わされている (佐々木, 2010b)。

一方、アメリカ評価学会誌である *New Direction for Evaluation* でつい最近、注目すべき特集号が刊行された (Julnes, G (ed.), Number 133, Spring 2012)。特集号のタイトルが「公共利益の価値判断の促進：価値判断に関して政策立案に情報提供する」(Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation) である。評価の本質は価値判断であるという一定の合意がありながら、具体的な価値判断の方法や実施手続きに関して不十分にしか議論されてきておらず、そうした状況に一石を投じる特集号である。このテーマはまだまだ研究されるべきで、また手法の開発が進められるべきであり、日本でも同種の議論が起きることが期待される。

6. プログラム評価の実践に向けて

最後に、教育支援プログラムに関するプログラム評価の普及に向けたいくつかの勘所 (tips) を解説したい。

6.1 分野横断的な普及の流れをいち早く受け入れる

他の分野に目を向けると、ここ10年の間に、医療、福祉、自治体、コミュニティ活動、NPO活動などさまざまな分野で、プログラム評価が急速に普及してきた。すでに述べた通り、この普及の背景には、統計学や倫理学などと同様に、分野横断的な基礎的道具として評価学が認識されてきたという事情がある。こうした他分野におけるプログラム評価の普及の流れに後れを取らずについていく必要がある。何か施策や事業を実施したら必ずプログラム評価を実施することを奨励するとともに、学部・大学院教育において、本論文で解説したような評価学の基礎を教育していくことが必要であろう。

6.2 小さなプログラムから始めてみる

いきなり規模の大きな教育プログラムを対象としてプログラム評価を実施することは極めて難しい。当初は、小規模・短期間のプログラムを対象として始めてみるのが勧められる。受益者が数名のプログラムを対象としたプログラム評価でも十分に成立する。ただし、繰り返しになるが、研究と評価は違うことに留意することが重要である。数名の対象者に関するケーススタディに留まることなく、明らかにされた事実を踏まえて、何らかの価値判断を下さねばならないことに留意する必要がある。また、ロジックモデルの作成によるそもそもの介入理論の明確化や、就職率や卒業率の厳格な差異分析を伴うインパクト評価を外部要因の影響を取り除いた上で実施すること、そして限られた資金や人材を投入して実施する意義があったのかの効率性評価の実施の普及が重要である。まずは小規模なプログラムの評価において、それらの評価を行って、次第に政策レベルの評価にもそうした評価を適用していくことが望まれる。

6.3 試行を通じて徐々に評価枠組を改善していく

評価は「事実特定+価値判断」とされるが、実際には、事実の確認と価値判断の間を行きつ戻りつして、徐々に結論に達するのが一般に行われている慣行である。そしてその過程を通じて、次に類似のプログラムの評価をする際に適用できる評価枠組を固められればよいといった態度で臨むことが勧められる。

6.4 プログラム評価の究極の目的を常に意識する

最後になるが、なぜプログラム評価を実施するのかの目的を忘れて、技術に走ってしまうことには注意が必要である。Scriven がしばしば述べるように、「評価は社会の改善活動である」(Evaluation is social betterment.) ことを忘れずに、プログラム評

価に取り組んでいかねばならない。外国人学校や私立学校などでバイリンガルの育成を目指している場合、自治体の予算で支援している日本語や母語教育の場合、補習校で行われている継承語教育、それ以前の寺子屋式の継承語教育と、規模や経済基盤が異なるバイリンガル教育機関が存在するが、いずれもプログラム評価の手法を適用することで事業の改善、ひいては社会の改善活動につながると言える。ただし、本論文で解説した諸々の手法をいつでもすべて適用せねばならないというわけではなく、セオリー評価やインパクト評価など一部だけの適用であっても、事業および社会の改善には結びつく可能性が高いわけで、まずは気負わずに、適用可能な部分から適用してみることが勧められる。

注

- 1) 評価には、政策評価 (Policy evaluation)、プログラム (施策) 評価 (Program evaluation)、プロジェクト (事業) 評価 (Project evaluation)、人事評価 (Personnel evaluation)、製品評価 (Product evaluation) などの種類があるが、その中で一番メジャーなのがプログラム (施策) 評価である。
- 2) Rossi & Freeman (1999), p20; Weiss (1997), p4~6; Wholey, Hatry & Newcomer (1994), p3; OECD (1999), p12のそれぞれのEvaluationの定義を参考に筆者が作成した定義 (龍・佐々木2000)。
- 3) 例えば、アメリカ評価学会が発行する学会誌は、以前、New Direction for Program Evaluationという名称であったが、現在はNew Direction for Evaluationと改称されているのがこの状況を端的に表している。
- 4) Rossi, P.H. and Freeman, H.E. (1993). *Evaluation: A Systematic Approach* (5th Ed.). p5. “Evaluation research is the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs.” そして、Assessment of Program Utilityとして、Impact Assessment とEfficiency Assessment をあげているので (p.36-37)、定義の中のUtility をImpact とEfficiency で置き換えた。ちなみに Rossi は、マサチューセッツ大学の社会学大学院の教授である。
- 5) Weiss, C.H. (1997). *Evaluation* (2nd Ed.). p4. “Evaluation is the systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy.” ちなみにWeissは、ハーバード大学教育大学院の教授である。
- 6) Scriven, M. (1980). *The Logic of Evaluation*. preface. (Systematically and objectively determining the merit, worth, and/or significance of something). その定義のなかの、merit, worth and/or significance of something を単純に「物事の価値」と訳した。なお、本論文では、merit, worth and/or significance の3種類の価値を総称して、価値 (value) という用語を使用した。ちなみに、Scriven は、ウェスタンミシガン大学評価研究所の教授だった (現在は、クレアモント大学に所属)。

7) Scriven (1999) は、Evaluation = Factual premises + Value premises と表現している。

【参考文献】

- 国際協力銀行(2006) 『円借款事業評価報告書』 国際協力銀行
- 財団法人国際開発センター(2010) 『日本技術支援信託基金の評価ー環境・社会持続性、ファイナンスへのアクセス、および企業アドバイスの各ビジネスラインに関してー』 国際金融公社 (IFC)
- 佐々木亮 (2002) 『政策評価トレーニングブック～7つの論争と7つの提案』 多賀出版
- 佐々木亮 (2010a) 『評価論理：評価学の基礎』 多賀出版
- 佐々木亮 (2010b) 「エビデンスに基づく開発援助評価ー援助評価の歴史、ランダム化実験の起源、スクリヴェンとバナージェの考え方の比較ー」 『日本評価研究』 Vol.10, No.1 日本評価学会
- 佐々木亮 (2012) 「SROI評価の事例を論じる1」 『CSR 活動における社会投資収益率法 (SROI) を用いた評価』セッション、日本評価学会第9回春季全国大会配布資料
- 札幌寛子 (2011) 『日本語教育のためのプログラム評価』 ひつじ書房
- 龍慶昭・佐々木亮(2000) 『「政策評価」の理論と技法』 多賀出版
- 龍慶昭・佐々木亮(2004) 『「政策評価」の理論と技法 (増補改訂版)』 多賀出版
- Julnes, G.(ed.) (2012). *Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation*, In *New Direction for Evaluation* (Number 133, Spring 2012). Jossey-Bass.
- CSREES-Office of Planning and Accountability (2007). *Document for Higher Education*. CSREES.
- Davidson, J. (2005). *Evaluation methods basics*. Thousand Oaks, CA: Sage.
- OECD.(1999). *Improving Evaluation Practices*. OECD
- Mark, M.M., Henry, G.T. & Julnes, G. (2000). *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*. Jossey-Bass.
- Patton, M.Q. (1997). *Utilization-Focused Evaluation: New Century Text* (3rd ed.). Thousand Oaks, CA: Sage.
- Rossi, P.H. and Freeman, H.E. (1993). *Evaluation: A Systematic Approach* (5th Ed.). Thousand Oaks, CA: Sage.
- Rossi, P. H., Freeman, H. E. & Lipsey, M. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.
- Scriven, M. (1980, 1982). *The logic of evaluation*. Inverness, CA: Edgepress.
- Scriven, M. (1999). "The nature of evaluation part i: relation to psychology." In *Practical Assessment, Research & Evaluation*, 6(11). <http://pareonline.net/getvn.asp?v=6&n=11>
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage.
- Stake, R.A. (1995). *The Art of Case Study Research*. Thousand Oaks, CA: Sage.
- Social Impact Associates (2008). *Aidha Impact Assessment Report*.
- Weiss, C.H. (1997). *Evaluation* (2nd Ed.). Prentice Hall.

- Wholey, J.S., Hatry, P.H. &, Newcomer, K.E.(1994). *Handbook of Practical Program Evaluation*, Jossey-Bass.
- World Bank (2004). Books, Buildings, and Learning Outcomes: An Impact Evaluation of World Bank Support To Basic Education in Ghana. World Bank.
- ulnes, G. (Ed). (2013). " Special Issue: Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation". *New Directions for Evaluation Volume 2012, Issue 133*. Wiley.

別添：インパクトの有無・程度のための基本デザイン		
名称と概念図	説明	
1. 事前・事後比較デザイン (Before-After Design) 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。		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。		
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。		
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。		
5. 実験デザイン (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 (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。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。		↓ 厳格 but 高価

佐々木 (2010a: 61-64)