学校改善モデルの有効性に係わる科学的基準に関する検討

― 米国の Comprehensive School Reform プログラムに着眼して ―

桐村　豪文

京都大学大学院教育学研究科紀要 第58号
2012

学校改善モデルの有効性に係わる科学的基準に関する検討
―米国の Comprehensive School Reform プログラムに着眼して―

桐村　豪文

1．課題設定

米国における教育行政では、「児童生徒の結果を改善することを目的とした、教育プログラム（whole school reform 等）や製品（教科書、カリキュラム等）、実践（学年を超えたグルーピング等）、また政策（クラスサイズの縮小等）」[1]を介入（intervention）と呼ぶ。教育行政は教育実践に対して、一定の目的・目標を達成するため、何らかの介入を用いて直接的／間接的に操作的であろうとする。本稿は、実践に対する介入の操作性から導かれる「介入の有効性」について、米国連邦教育政策の Comprehensive School Reform（CSR）プログラムで活用される学校改善モデル（以下、CSR モデル）を事例に、科学哲学の観点から検討するものである。ただし、介入の有効性それ自体（「介入 X は有効である」の真偽）を対象とするのではなく、有効性について語られるものの真理性を保証する基準、方法論（何が「介入 X は有効である」を真（より真）とするのか）を対象とする。多くは前者を問うことにこそ研究的意義が認識される中で、なぜ後者の思弁的な問題を問わなければならないのか、これを問うことでどういった受肉した問題が顕在化してくるか、その問うことの意味について、本題に入る前に示したいと思う。

介入の有効性については、「介入 X は結果 Y に対して有効である（か）」といった具合に仮説が設定され、特に統計的手法を用いて検証が行われることが多い。その典型例を示せば、志水宏吉は、「各学校の子ども達の学力」を従属変数、「学力向上に関する各学校の取り組み」「各学校が置かれている社会経済的背景」を独立変数とし、重回帰分析を通じて「取り組み」の影響力を分析している[2]。その分析結果はさておき、こうした高度な分析手法を通じて介入の有効性に関する立派な研究結果が、とりわけ米国においては Hanushek に代表される計量経済学の分野において多く産出されてきたわけだが、この立派さを支える土台が意外にもろいことには無自覚的であることが多い。そしてさらに重要なことは、自らの土台の脆弱さに無自覚的で無反省的である限り、どれほどその研究が立派に見えようと、論理的に共有されえないし、その後の建設的な研究的進展に寄与し得ないということである。1 つのありうる物語を語っているに過ぎない。

この土台を検討する上で問題としなければならないのは、(1) 有効性に関する研究結果「介入 X は結果 Y に対して有効である」が真であると言うこと（結論の決定性、信頼性）自体の意味、(2)「介入 X は結果 Y に対して有効である」が真であることを保証する基準である。本稿の直接の対象は (2) であるが、(1) は常に (2) の問題を論ずる上で考慮しなければならない問題なのである。以下ではまず、CSR モデルの有効性を検証する際に用いられる基準の中身を確認する（いかなる基準で「モデル X は有効である」と判定しているのか）。その後、(1) (2) の問題を詳細に示した上で、その基準の信頼性（なぜその基準がより十分な土台と言えるのか）を検討する。

表1　CSR プログラムの 11 つ構成要素[3]

1.	**科学的基盤をもつ研究（scientifically based research）に基づき、有効性が立証された（proven）方法と方略―** **包括的学校改善プログラムは、科学的基盤をもつ研究や効果的実践例に基づき、そして学校において首尾よく再現されてきた、児童生徒の学習や教科指導、学校運営に関する、その有効性が立証された方略や方法を活用する。** **第 1 構成要素は、学校が包括的改善プログラムを設計する上で、主要教科、特に数学とリーディングにおいて科学的基盤をもつ研究に根ざし、有効性が立証された方略や方法を採用する必要性を強調する。学業成績へ焦点化しつづけること、それを支える包括的プログラムを構築すること、授業において「何が機能するか」を重視すること、これらは成功する包括的デザインの重要な要素である。**
2.	包括的デザイン― 効果的に学校を機能させるための包括的デザインは、教科指導、評価、クラス運営、専門職開発、保護者参加、学校運営を総合化するものである。学校のニーズアセスメントを通じて同定されたニーズに取り組むことによって、包括的デザインは、カリキュラム、テクノロジー、専門職開発を、スクールワイドな変革のための計画の中に組み入れる。このデザインの最終目標は、すべての児童生徒が州の定める、内容及び学業成績に関する挑戦的基準を満たすことができるようにすること。
3.	専門職開発― プログラムでは、質の高い、そして継続性を持った、教職員のための専門職開発及び訓練を提供する。専門職開発は、その有効性が立証され、革新的で、費用効果が高く、容易に利用可能な方略を含むもので、そして教員が州のアセスメントと州の定める内容に関する挑戦的スタンダードを用いて、教科指導の実践と児童生徒の学業成績を改善させることが可能となるようにするものである。（略）
4.	測定可能な目標及びベンチマーク― 包括的学校改善プログラムは、児童生徒の学業成績に関する測定可能な目標を含み、その目標を達成するためのベンチマークが設定される。連邦教育省は、地方学区がこれらの目標を、ESEA Section 1111(b)(2)の下、州の定める AYP の定義と連結させることを奨励する。
5.	学校内の支援体制― 学校を通じて教員、校長、管理職その他の職員は、CSR 学校におけるプログラムを支援するものである。教職員は、この支援を、学校の包括的改善プログラムを理解し、喜んで応ずることにより、そして授業での指導の継続的改善を重点的に取り扱うことにより、そして専門職開発へ参加することにより、この支援があることを立証すること。
6.	教員、校長への支援　（Added in 2001）― CSR プログラムは、教員、校長、管理職その他の職員に対し、共有されるリーダーシップと、改善努力に対する応答責任の広範な基礎を構築することによって、支援を提供するものである。CSR プログラムは、チームワークと功績を祝賀することを奨励する。これらを含む支援の方法は、学校の包括的デザインの役目である。
7.	保護者と地域住民の参加― プログラムは、学校改善活動を計画、実施、評価する上で、保護者や地域住民の有意味な参加を認めるものである。この構成要素を扱う際には、学校は Title I, Part A（ESEA Section 1118）の保護者参加の要件と矛盾しない方略を構築する。学校は、保護者が参加する能力を形成することに特別の注意を払い、保護者が教科指導プログラムの話し合いの場に参加でき、自分の子どもの学業成績へ貢献できるような方法を設計するものである。
8.	外部の技術支援及び援助― プログラムは、スクールワイドな学校改善における専門的知識や経験を有した主体（高等教育機関を含む）から、質の高い外部支援及び援助を活用する。CSR の法律上においては、補助金を受けたプログラムが、成功を収めてきた過去の業績や、財政的安定性をもち、そして改善実施期間中、質の高い道具、学校職員のための専門職開発、現地支援を提供する能力をもった、資格を有する技術援助プロバイダーによって支援を受けることを州教育当局が保証するよう、要求している。
9.	毎年の評価― 学校改革の実施や児童生徒の学力結果を評価する毎年の方策を立てること。（略）
10.	資源の調和― 他の活用可能な資源（連邦、州、地方、民間）がどのようにして、学校の改革をサポートし継続させるためのサービスを学校が調和させることができるか、その方法を同定する。（略）
11.	**学業成績を改善させる方略　(Added in 2001)―** **プログラムは、以下の要件のうち 1 つを満たさなければならない。** **■ プログラムは、科学的基盤をもつ研究（scientifically based research）を通じて、参加する児童生徒の学業成績を改善させることについて、有意な研究結果をもつものであること。** **■ プログラムは、参加する子どもの学業成績を改善させることについて、有意な強力なエビデンス（strong evidence）をもつことが判明していること。**

2．Comprehensive School Reform プログラムの科学性

2．1．Comprehensive School Reform プログラムの概要

CSR プログラムの前身 Comprehensive School Reform Demonstration（CSRD）プログラムは民主党下院議員 David R. Obey と共和党下院議員 Edward Porter の先導の下、1997 年に制定された FY 1998 Appropriations Act を受け、1998 年から開始された。その後、2002 年に制定された No Child Left Behind Act of 2001（NCLB 法）によって、初等中等教育法（Elementary and

Secondary Education Act）の Title I, Part F として位置づけられ、Demonstration の文言がはずされ、Comprehensive School Reform（CSR）プログラムと変更された。

CSR プログラムの目的は、すべての子どもが、とりわけ低い成果しかあげられず高い貧困度を持つ公立学校に在籍する子どもが、州の定める挑戦的な、内容及び学業成績に関するスタンダードを達成することができるよう、科学的基盤をもつ研究や効果的実践例に基づいて、包括的な学校改善（comprehensive school reform）を実施することを連邦が支援することにより、児童生徒（K-12）の学業成績を改善することにある。CSR プログラムでは、包括的なデザインの中へ組み合わされ、一体化された改善戦略は、互いに独立して実施される同様の戦略よりもより良く機能するという前提に立脚している。学校は、表 1 に示す 11 の要素を総合化しなければならない[4]。

２．２．CSR プログラムの構成要素における科学性の要求

CSR プログラムの補助金を受け、各学校で実施される包括的学校改善プログラムは、表 1 に示す 11 つの構成要素のすべてを満たさなければならない。これらの要素の中で、ここで特に注目されたいのは 1 番目と 11 番目の要素である。特に 1 番目の要素は、1997 年時点の 9 つの構成要素にも含まれていた要素であり[5]、ロバート・スラヴィン（Robert E. Slavin）によると、連邦教育補助金と有効性に関するエビデンスとを直接結び付けた歴史上初めての試みであるという[6]。また NCLB 法では 111 回に亘って「科学的基盤をもつ研究（scientifically based research、SBR）」という言葉が登場し、教育実践、プログラムに対して SBR に基づくことが要求された。CSR プログラムも NCLB 法を受けて、構成要素の中に SBR の文言が組み込まれた。

構成要素の第 1 番目は、教科指導の方略を対象に要求されるものである。そこでは SBR のみが認められている。第 11 番目の構成要素は、残りの構成要素（教職員の職能開発や保護者等の学校参加等）に対して要求されるもので、そこでは SBR 若しくは強力なエビデンス（strong evidence）が認められている[7]。CSR プログラムを実施する学校は、SBR 若しくは強力なエビデンスによって裏付けられたプログラム若しくは実践を採用することが要求されているのである。この要求により、CSR プログラムにおいては外部開発された CSR モデルを活用することが、義務ではないが、関係が非常に強いのである。次節では、その外部開発された CSR モデルの有効性に関する評価について、その評価が基づく基準、方法論を考察、検討する。

３．CSR モデルの有効性の検証に用いられる科学的基準

３．１．CSR モデルの有効性に関する評価とその問題の指摘

米国においては、現に多くの学校現場において科学的なモデルが活用され、実践されているし、その円滑な活用のために、CSR モデルに関する有効性の検証を行う組織及び支援が確立されている。「連邦教育省は、教育プログラムに関する研究を総合化するためのいくつかの努力に対して支援を行ってきた。その代表的なイニシアチブは What Works Clearinghouse（WWC）であるが、他の主要なイニシアチブとして、Comprehensive School Reform Quality Center（CSRQ） や Best Evidence Encyclopedia（BEE）がある」[8]。そしてプログラムの有効性を確実に抽出するよう、検証の手続も厳格に規定しようとする先進的試みが為されている。

しかしこの先進的な試みにおいても問題は指摘されている。「問題は、これらの総合化で用いられる方法が各々、基本的な点で異なっており、それがゆえに、有効性に関する頑強なエビデン

スをどのプログラム・実践がもっており、もっていないのかについて、結論に一貫性を欠く事態を招いている、ということである。この不画一性は、エビデンスに基づく改革にとって潜在的に深刻な問題である。なぜなら、教育者や政策立案者が特定の事業全体に置く信頼の土台を侵食してしまう可能性があるからである。学術的な不同意は健康的（であり不可避）であるが、少なくとも、問題を理解し、プログラム評価の総合化のための基盤的なルールについて同意を得ることは重要である」[9]。この指摘のとおり、プログラムの有効性を評価するにおいて、その土台を確立すること（土台があること＋土台が頑丈であること）は、極めて重要なことなのである。以下、CSR モデルの有効性を検証する組織の 1 つである CSRQ について、その検証で用いられた検証手続、方法論（土台）について考察し、その信頼性（頑丈性）について検討する。

３．２．CSRQ の検証で用いられる科学的基準[10]

３．２．１．CSRQ の検証の手続

CSRQ の報告書は、Quality Review Tool (QRT)を用いて作成される。QRT は、モデルに対する独立した公平で信頼あるレビューを行うための基準を提供するものである。QRT には 3 段階のプロセス（Part 1、Part 2、Part 3）があり、所定の基準に基づき、モデルの評価が行われる。

QRT の Part 1 は、質的データの収集段階である。QRT の Part 1 では 4 つの段階が用意されている。まず、①規格化されたフォームを用いてモデルの説明事項を記述する。説明事項には、当該モデルの設定する目的、歴史、マーケットシェア、費用等が含まれる。そして、②CSR モデルのプロバイダーと電話を通じて、①で収集した情報が正しいかどうかを確かめる。また、③当該モデルを活用する学校 3 校の校長とも電話を通じて、①②で収集した情報の正しさを確認する。最後に、④以上の質的データを総合化するため、規格化されたフォーム Model Description Form-Complete を用いて CSR モデルの説明事項を完了する。

QRT の Part 2 は、量的データの収集段階である。そこでは、CSR モデルの有効性に関する文献について体系的レビューが行われ、因果的妥当性（causal validity）の情報をコード化する。注意すべきは、QRT の Part 2 では、モデルの有効性それ自体を検証するのではなく、そのモデルの有効性を論じる個々の研究について、その研究が基づく研究デザインの厳格性（rigor）を検証するということである（土台が土台たりうるための頑強性の基準）。QRT の Part 2 では 5 つの段階が用意されている。まず、①委嘱された研究者の下で徹底した文献調査が行われる。研究者はまず、教育に関わるデータベース (JSTOR, ERIC, EBSCO, Psychinfo, Sociofile, NWREL, DAI 等）やその他ウェブツールを通じて各モデルの検索を行い、また CSR モデルに関する重要で包括的な研究 Herman et al.（1999）[11]と Borman et al.（2002）[12]も参照する。以上のデータソースから収集した、22 のモデルに関する約 800 の論文のサマリーを第 1 回目のスクリーニング（initial screen）にかける。このスクリーニングを通過するためには、いくつかの基準を満たさなければならないが、これについては３．２．２．(a) で述べる。

次に、②規格化されたフォーム Study Description Outcome Form (SDOF)を完成させる。SDOF を用いて研究者は、各論文が含む研究デザイン、結果（outcome）の変数、人口統計に関する情報をコード化し記録する。コーディングの段階で最も重要な焦点は、信頼性のある研究デザインであるか否かスクリーニングするということにある。研究デザインに関する所定の基準を満たし、スクリーニングを通過し、そして児童生徒の学業成績の変数を結果（outcome）に含む

論文は、次の段階におけるフルレビューの資格が与えられる。そのスクリーニングの基準についても３．２．２．(b) で述べる。

③第 2 の規格化されたコーディングとして、Complete the Quality Indicators Form (QLIF) がある。このコーディングの目的は、各論文のフルレビューを行い、各々の研究の質を検証し、統計的情報を収集することにある。ここで検証される研究の質、収集される統計的情報については３．２．２．(c) で言及する。なおこのフルレビューは、2 名の研究者が各々独立して行い、続く段階では、④2 名の研究者の間で、コーディングした各項目について同意に至るよう、調整が行われることとなる。

最後に、⑤すべての因果的妥当性のスコアに関して、各論文を格付ける。この段階においては、各研究を因果的妥当性について点数をつけるために設計された規定に基づき、QLIF から得られる情報を、「非決定的（inconclusive）」、「示唆的（suggestive）」、「決定的（conclusive）」のいずれかに体系的に格付ける。ここで「示唆的」または「決定的」に格付けされた研究は、研究デザインの厳格性の基準を満たすものである。その基準については、３．２．２．(d) で言及する。

QRT の Part 3 では、各 CSR モデルの格付けを行うため、エビデンスの領域を 5 つのカテゴリーに分けて、以上に収集した質的・量的データを総合化する。先の Part 2 ではモデル自体ではなく、モデルの有効性を論じるエビデンスの厳格性を検証し格付けを行う段階であったが、Part 3 はそのモデル自体の有効性の格付けを行う段階である。なお、5 つのカテゴリーとは、①児童生徒の学業成績に対する有効性、②その他の結果（児童生徒の規律、出席、学校環境、留年率、教員の満足度等）に対する有効性、③保護者、家庭、地域住民の参加に対する有効性、④モデルのデザインと研究との間のつながり、⑤首尾よく実施しうるために学校に対して提供されるサービスや支援、である。ここではカテゴリー①に特化する。その理由は、CSR モデルの多くが、その焦点を主として学業成績の向上に当てており、それゆえ CSR モデルの有効性を評価する側においてもまた、このカテゴリーをその評価の目的の筆頭に挙げているからである。

カテゴリー①においては、各モデルの評価を行うため、QRT の Part 2 で収集された量的情報を用いる。これら量的情報を用いて、児童生徒の学業成績に対するモデルの有効性について格付けがなされる。この判定基準については３．２．２．(e) で言及する。

３．２．２．CSRQ のスクリーニングの基準

（a）第 1 回目のスクリーニング

QRT の Part 2 の第 1 回目のスクリーニングの基準は、各モデルに関する論文について、①1980 年から 2005 年 4 月の間に発表されたものであること、②調査対象とする CSR のモデルのうち少なくとも 1 つのモデルについて検証しているものであること、③量的方法を用いていること、④フルテキストの研究論文として報告されていること（パワーポイントのプレゼンテーションや要約ではないこと）である。

（b）研究デザインに関するスクリーニング

QRT の Part 2 の Study Description Outcome Form (SDOF)における研究デザインに関するスクリーニングの基準は、まず消極的基準として、①結果に関する量的データを有さずモデルを支持し、ただ理論を評価するもの、②十分厳格でない研究デザイン（1 つのグループの事前 - 事後テスト等）を用いるもの、これらは排除される。また積極的基準として、実験デザインや、事前

- 事後テストを用いて、統制グループと CSR モデルのグループとを比較して評価する準実験デザイン、そしてテストに必要な多様な段階をもち縦断的（longitudinal）かつコホート（cohort）のデザインであるもの、これらはこの段階を通過する。

（c）検証される研究の質と収集される統計情報

QRT の Part 2 の Complete the Quality Indicators Form (QLIF)で、検証される研究の質には、内的妥当性（internal validity）、外的妥当性（external validity）、外観妥当性（face validity）、心理測定的妥当性（psychometric validity）やその他の研究の質に関する指標が、そして収集される統計的情報には、効果量（effect size）や生の統計情報といったものが含まれる。

（d）因果的妥当性の観点からの研究の質の判定

QRT の Part 2 の最後の段階において、因果的妥当性の観点から各研究を「非決定的」、「示唆的」、「決定的」のいずれかに格付けするのだが、その判定基準として重要となるのが、「妥当性に対する重大な脅威 (critical threats to validity)」と「妥当性に対する重大でない脅威 (Noncritical threats to validity)」である。

「妥当性に対する重大な脅威」には、例えば脆弱な外観妥当性や信頼性、不十分なプログラムの忠実性、処置グループ（treatment groups）と統制グループ（control groups）の非同質性（nonequivalence）、適切な基準値（baseline）の欠如、結果測定のタイミング（CSR モデル実施後 1 年未満、事前テストと事後テストとの間の経過時間が 1 学年未満）といったものが含まれる。「妥当性に対する重大な脅威」をもつ場合、その論文は「非決定的」に格付けされる。

他方、「妥当性に対する重大でない脅威」には、歴史的出来事、新奇性・逸脱性効果 (novelty and disruption effects)、器具の使用の変化、成熟（maturation）、選択バイアス（selection bias）、統計的回帰といったものが含まれる。

「示唆的」と格付けされる論文は、「妥当性に対する重大な脅威」は持たないが、「妥当性に対する重大でない脅威」を 3 つ以上もつものである。長期的かつコホートの研究デザインを含め、統制グループを持たない研究は、分析技術がより高水準の厳格性を産み出すもの[13]でないならば、「示唆的」であると格付けされる。

「決定的」と格付けされる論文は、さらに高水準の厳格性を持つものであり、「妥当性に対する重大な脅威」は持たず、「妥当性に対する重大でない脅威」も 2 つ以下である、実験デザイン、準実験デザインである。「決定的」な因果的妥当性の格付けをもつ研究に限り、嘱託研究者によって効果量（effect size）が報告され、あるいは生の統計データから算出される。

（e）モデルの格付け

格付けを下す際に着目する要素は 2 つである。1 つは、研究デザインの持つ因果的妥当性に基づいたエビデンスの頑強性、もう 1 つが報告されたモデルの影響力、効果の大きさである。

1 つ目のエビデンスの頑強性は以下の 3 つの要素に依存する。つまり、(a) 研究デザインの厳格性とそこから産出されたエビデンスの信頼性、(b) モデルごとに提示された研究のエビデンスの量、(c) プラスの結果を教えるエビデンスの一貫性である。また 2 つ目のモデルの影響力の大きさを測定するには、ここでは効果量が算出される。効果量は、グループ間の差異を標準化したもので、異なる結果（例. 異なるテストの成績）でもその影響力を比較することを可能とする指標である。以上 2 つの要素に着眼し、表 2 に示すように、7 つのレベルの格付けが判定される。

表2　モデルの有効性に関する格付け

格付け	厳格性の基準を満たす研究	「決定的」に格付けされた研究	統計的有意（p ≦.05）を示す効果量	効果量の全体の平均
非常に強力（very strong）	10 以上	5 つ以上（もしくは全研究の 50%以上）	結果の 75%が統計的有意を示す＋の効果量	+0.25 以上
適度に強力（moderately strong）	5 つから 9 つ	3 つ以上（もしくは全研究の 50%以上）	結果の 51%から 74%が統計的有意を示す＋の効果量	+0.20 以上 +0.24 以下
普通（moderate）	2 つから 4 つ	1 つ以上（もしくは全研究の 50%以上）	結果の 26%から 50%が統計的有意を示す＋の効果量	+0.15 以上 +0.19 以下
制限的（limited）	1 つ	—	結果の 1%から 25%が統計的有意を示す＋の効果量	+0.14 以下
ゼロ（zero）	0	—	結果の 0%が統計的有意を示す＋の効果量	—
否定的（negative）	10 以上	5 つ以上（もしくは全研究の 50%以上）	結果の 75%が統計的有意を示すーの効果量	0 未満
格付けなし（no rating）	モデルが 1 つも研究を持たない場合（例．エビデンスが入手できない）			

4．CSR モデルの有効性の検証に用いられる科学的基準の信頼性の検討

4．1．介入の有効性に関する研究の決定性、信頼性

本節では、以上に確認した CSRQ において CSR モデルの有効性を検証する際に用いられる基準、方法論の信頼性を検討しなければならない。そのためにはまず、冒頭に言及した（1）有効性に関する研究結果「介入 X は結果 Y に対して有効である」が真であると言うこと（結論の決定性、信頼性）自体の意味、(2)「介入 X は結果 Y に対して有効である」が真であることを保証する基準、について説明を加えなければならない。語ることの困難を乗り越えるには、それに十分な準備が必要とされる。語ることの困難とは、存在論、認識論、方法論の問題である。

（1）の問題は存在論の問題である。ここでは決定性と自由性との対立が焦点となる。この問題は、介入の有効性を安易に語ることを許さない。「介入 X は結果 Y に対して有効である」が含意するところは、介入 X と結果 Y との間にプラスの効果をもった因果関係が実在するということ（介入 X によって結果 Y が生じる）である。これは実態においては、介入 X が結果 Y に直接関係する人間集合 P の行為に対する操作を介した結果として結果 Y が獲得される、ということを意味する。さらには、「介入 X は結果 Y に対して有効である」という有効性の結論は多くの場合、人間集合 P（より厳密には、時間 T における集合 P）を超えた事象にも適用しうることを期待する（外的妥当性）。この明らかな論理的飛躍には、また因果関係の抽出には、十分な土台が必要となるわけだが、その土台づくりにおいて問題となるのが決定性と自由性との対立である。

ここで言う「決定性」とは、物理学に代表される自然科学においては当然要求される一般性、汎用性、再現性である。それに対して「自由性」とは、操作対象である実践がもつ（もっているはずの）有機性、不可逆性、文脈依存、もっと広く言えば人間の自由意志である。

高度な統計手法や計量経済学の手法を用いる「科学的」研究の多くは、インプットとアウトプットに設定された変数間の因果関係（介入 X → 結果 Y）の獲得が目的として先行しているため、その変数間に生きた人間の実践が介在していること、そしてその実践に対する操作性が実在しない限り、因果関係も実在し得ないことを捨象してしまう。また「介入 X」と「結果 Y」により明瞭な変数が設定されるために、実は検証の対象が因果関係「→」や「によって」という直接観取し得ない、奇妙で不明瞭なものであることも意識されない。そして「→」や「によって」を間に挿入することがいかに難しいことであるかも、特段意識されない。

それに対し、実践の自由性やあるいは専門性といった側面を強調する者においては、介入の有効性について（肯定的であれ否定的であれ）語られたものの一般性、汎用性、再現性には限界があると反論を呈するわけである。例えば、「有効性が発見されたとしてもそれはその事例についてだけ言えることだ」や「教育という営みは生きた子どもたちを相手にしている以上、その営み自体有機的で流動的だから、一定の有効性や有効性を持ったモデルなんてものは本来存在し得ない」といった批判が突きつけられるわけである。したがって、いくら立派な「科学的」分析を通じて有効性を語ったとしても、この対立を乗り越えない限り、不毛な論議に帰着するのがオチである。

４．２．介入の有効性に関する研究の決定性、信頼性を保証する基準

介入の有効性について、「介入Ｘは結果Ｙに対して有効である」が真であると言うことには、本来的に、以上の決定性と自由性との対立が不可避に含意されており、それゆえ「介入Ｘは結果Ｙに対して有効である」が真であると言うことがいかに困難なことであるかが分かる。その困難の壁を乗り越えるためには、十分な頑丈さと高さをもった土台が必要となる。「介入Ｘは結果Ｙに対して有効である」と言えるための土台づくりである。さらには、その頑丈さと高さの基準ももちろん問題としなければならない。それが（2）の認識論及び方法論の問題である。

（2）で問うていることは、「何をもってエビデンス（土台）とするか」という問いである。「介入Ｘは結果Ｙに対して有効である」と言えるための土台が、本当に壁を乗り越えるに十分な土台であることを保証するための基準が必要となるのである。当然のことながら、その基準は恣意的に設定してはならない。では一体何によって基準を設定、選択すべきなのか。それは、後述する第２値における比較の作業を通じて、である。

ところで、なぜこのような愚問を問う必要があるのかと怪訝に思う者もいるかと思う。適正な調査等で収集、蓄積された確固たる「事実」によって研究結果の信頼性は保証される、と。つまり客観的な「事実」であるならば十分な土台になるのだ、という主張である。しかしこれは愚直にすぎる。特に「介入Ｘは結果Ｙに対して有効である」といった因果的説明を目的としている場合においては、「事実」は（必要だが）十分な土台になりえないことは、すでに科学哲学で多く指摘されてきたことである。例えばデュエム＝クワインテーゼや決定不全性（underdetermination）の問題によって、新たな事実の発見により理論が反証され、別の理論が新たに確立するといった理想は描き得ないこと、また事実の蓄積によって理論が１つに決定されるという理想も描き得ないこと、といったことが明らかにされている。こうした指摘は、従前の「事実によって語りは規定される」という古典的科学観を覆し、語りは事実によって一定制限されるけれども、それでも依然自由を有しているということを明白にするものである。語る対象の自由性（４．１．）に加え、語る主体の自由性の発見により、「介入Ｘは結果Ｙに対して有効である」と言えるための十分な土台づくりは、より厄介になるわけであり、厄介なものとして対峙しなければならないのである。

ここで「介入Ｘは結果Ｙに対して有効である」の命題に戻ると、語る主体の自由性も踏まえるならば、この命題の真偽判定に際して必要となる土台の十分性（基準）を考えるためには、「介入Ｘは結果Ｙに対して有効であると言える」という命題に問い改めなければならない。「有効か否か」ではなく「有効と言えるか否か」を問題とするのである。両命題は決定的に意味を異にし、後者の命題は、「介入Ｘは結果Ｙに対して有効である」という本命の命題真偽に向けて無闇に突き進むのではなく、その道を地道に建設、舗装していく作業を行う機会を与えるのである。

わかりやすい例を示せば、「神は存在する」と「神は存在すると言える」という命題である。前者は存在論的問いであり、その真偽の答えは「存在する」か「存在しない」かの 2 値である。それに対して後者は、認識論的あるいは言語論的問いであり、その真偽の答えは「存在すると言える」か「存在するとは言えない」かであるが、「存在するとは言えない」には程度の幅があり、「存在しないと言える」か「存在するとも存在しないとも言えない」となる。つまり後者において真理値は 3 値（真と言える、真とも偽とも言えない、偽と言える）となるわけである。ここでは「神」という教育研究とはやや縁遠い存在を例に用いたが、しかし有効性を検証しようとする研究の多くは「神」と同様、直接観取し得ない「因果関係」や「構造」といったものを発見しようとする。その点では全く同様である。

なぜこの極めて思弁的な手続を踏まなければならないかと言えば、①因果的説明の多くはこの「真とも偽とも言えない」領域（以下、第 2 値）に帰着すること、②第 2 値においては真偽判定を直接の目的とはできないがため、別の目的、そしてそのための別の基準が必要となること、③その目的、基準の下では、不毛な水掛け論ではなく、漸進的な進歩が可能となるからである。

先述の決定不全性によって、1 つの理論（因果的説明 X）が決定し得ない結果、その理論の真偽が帰着するのはこの領域である。注意すべきは、決定不全性は、因果的説明 X が真とは言えないことを決定付けるものだが、しかしそれが偽であることを意味するのではないということである（「因果的説明 X は真である」という命題は偽であるが）。つまり、事実に基づき構築されたその因果的説明は、実は真であるかもしれないが、しかし確実に真であるとは言えない、言うには不十分であるという判定である。因果的説明の多くはこのあやふやな判定領域に帰着する。

「真とも偽とも言えない」この第 2 値の領域において、それでもなお因果的説明に優劣をつける動機の存在は不可避である。この問題を克服することが次なる重大かつ革新的な課題である。つまり、科学の直接の目的を「真偽の判定」ではない別の目的に据える必要があるわけである。しかしその目的は「真偽の判定」と同じ道の上になければならない。ラリー・ラウダン（Larry Laudan）は、科学の目的を「問題解決活動」と捉えることの必要性を提起する[14]。そこでは、更新される理論が、実際には真かも偽かも知れないがそれは分かりえないので、先行理論よりもより多くの問題を解決するとき「進歩している」と呼ぶのである。本研究ではその進歩を消極的に「より真と言える」と呼ぶ。しかし単に「より多くの問題を解決する」ことを目的とするだけならば、例えば「神によって全てが決定される」という説明は優れた説明と判定されてしまう。そこで重要な点は、①経験的問題（理論と観察の調和）と②概念的問題（理論内部、理論間の調和）である。この①②の問題を基準として、「より真と言える」ものとそうでないものの優劣が判定されるのである。特に②は、「自由意志」等の他概念・理論との整合／不整合を問題にするものであり、「より真と言える」ためには整合性もまた重要となるのである（①は専ら重視されるが、②はしばしば捨象されてしまう。４．３．で言及する「自己確証」は②に関するもの）。

科学の目的を「問題解決活動」に位置づけ、①②をその進展の判定基準に据えることによって、「より真と言える」かどうかの比較可能性が用意されるのである。またそれは基準（方法論）の選択においても同様で、ラウダンは、方法論の選択の問題は経験的証拠で片が付くと論じている[15]。そしてこの比較可能性を前提にすることで、「真偽の判定」を純粋に目的とする場合に陥る不毛な事態を回避できるのである。例えば、因果関係を発見する方法として（本稿も対象とする）ラン

ダム化比較試験の有用性を論ずる場において、批判者は「それは説明的に完全でないため」としてその有用性を棄却する。それに対して擁護者は、そう言う批判者もまた「そうした（不完全な因果的）言明もしばしば真であるため有用であると考えることに同意している」と反駁する[16]。両者は共に、複雑かつ完全な世界（存在論的な真偽判定）を知りたいという願望をもつことにおいては共通している。しかし擁護者は、言語の合理性ゆえに知りうるところは不完全でしかないということを認めている。ただその不完全性（真とも偽とも言えない領域）と完全性を求める願望との狭間で、いかなる道筋をつけるべきかを知らない。従来の基準（方法論）の選択においてはこのように、言語の限界を超えたユートピアに直接の知的願望をもつからこそ、不毛な論議（道の構想・破壊の繰り返し）に陥るのである。対照的に、ラウダンの提起する目的の下では、「より真と言える」か否かの比較を通じて、慎重かつ漸進的な道の建設、舗装作業が行えるのである。（目的地が不明なままに進み方だけは整備するという一見矛盾する方針は、身体の有限性ゆえに僅かに許された方針である。）

以上の前提を踏まえた上で、次項では、先に確認した、CSRQ で CSR モデルの有効性を検証する際に用いられた基準、方法論の土台としての信頼性を検討する。これは、介入の有効性について「より真と言える」結論を導く「より良い」方法論を探る地道な作業の一部である。

４．３．なぜランダム化比較試験が「土台」としてより優れていると言えるのか

CSRQ の評価では、「モデル X が結果 Y に対して有効である」の命題に答えるために、３．２．に示した厳格かつ詳細な手続きが取られており、またそうした厳格性をもってしかモデルの有効性をより真として語りえない、ということである（それに引き換え、我が国で蔓延する「実証的」な教育政策研究は、「政策科学」と称してどれほど厳格性を欠いていることか。本研究はその点で従来の政策科学への根本的なアンチテーゼでもある）。この CSRQ の格付け作業の中で、エビデンス（土台）の信頼性（頑丈性）に関わる最も重要な要素を抽出するとすれば、「決定的」と格付けされる研究を含むことである。そして研究が「決定的」と格付けされるための基準は、「妥当性に対する重大な脅威」は持たず、「妥当性に対する重大でない脅威」も 2 つ以下である実験デザイン、準実験デザインである。ここではいくつか重要なポイントがあるだろうが、本稿では、昨今の米国のエビデンスに基づく政策の潮流において「gold standard」として専らもてはやされている「実験デザイン」、ランダム化比較試験（Randomized Controlled Trial、RCT））という方法に着眼する。「ランダム化比較試験は、特定の介入の効果を評価するため、無作為に個人を介入グループまたは統制グループに割り当てる研究である」[17]。なぜ RCT は、モデルの有効性を保証する上で、より優れた基準であると言えるのだろうか。以下、その根拠をナンシー・カートライト（Nancy Cartwright）の理論を用い、本節第 1 項、2 項に立ち返りながら検討する。

カートライトは「信頼性（credibility）」を「ある政策を採用することを支持するエビデンスがそれ自体非常に真であるらしいこと（very likely to be true）」と定義した上で、信頼性を保証する 2 つの基準を提示する。1 つは要因の確定（clinchers）、もう 1 つが自己確証（self validating）である。カートライトは、2 つの基準を用いて、計量経済学の方法に比べ、RCT がより優れていることを示している[18]。

要因の確定（clinchers）は、検証の結果を因果的説明の結論に演繹することができることであ

る。RCT においては、「理想的な RCT の下、統制グループよりも、処置 T を行う介入グループの方が、結果 O が生じる確率がより高い」ならば、「特定の実験条件の下、特定の実験母集団において、T によって O が生じた」と演繹されるわけである。しかしこれは計量経済学においても、仮説演繹法を用い、特定の分析技術を用いることで、要因の確定がなされる（例. 回帰モデルを用いて係数が推定される）。

RCT が計量経済学の方法に比べてより優れている点は後者の自己確証（self validating）にある。理想的な RCT は計量経済学の方法にはない特徴を持つ。それは交絡要因（confounding factors）を介入グループと統制グループに等しく配分しているという点である。そしてその手続きを適切に遂行するために様々な方策がリストされているのである。例えば無作為抽出（randomization）、四重盲検法（quadruple blinding）[19]などがそれである。計量経済学にはそうしたリストは存在しない。ただしここでリストの有無を土台（方法）の優劣に繋げるのは、単に倫理的感情から（RCT の方が努力しているから）ではない。多くの手続によって実験の条件をより統制しているからこそ、「T によって O が生じた」がより真であるとして語れるのである。

例えば、単純な事前事後テスト（pre-post test）を事例に示すと、A 学校に介入 T（例. スペシャリスト P 氏、少人数学級）が投入された。そして 1 ヵ月後テストを行ったところ、結果 O（例. 学業成績の向上）が見られた。これにより「T によって O が生じた」と言える。否。もちろん言えるわけがない。実際には T によって O がもたらされたかもしれないが、ここでの問題は第 4 節第 1 項、2 項でも述べたとおり、「によって」を「言える否か」である。T による O に対する操作性が確証されない限り、「T によって O が生じた」とは言えない。第 1 項で示した「対象の自由性」により、「によって」を挿入することは容易ではない。ここで P 氏が安易に「によって」を挿入するとすれば相当に高慢なスペシャリストである。この場合「T によって O が生じた」という命題は「真とも偽とも言えない」が解である。

第 2 項で論じたように、「真とも偽とも言えない」第 2 値の領域においては、「真偽の判定」というユートピアではない別の目的「問題解決活動」の下、①経験的問題と②概念的問題を基準として、比較という地道な作業を通じて科学の漸進的進展が遂行される。そしてしばしば捨象されるがゆえに特に問題とすべきは②の問題である。自由意志の対立（第 2 項で示した）を含めた他概念・理論との整合性を保つことの重要性がここに示されているのである。そしてこの整合性の問題ゆえに、多くの手続きによって実験の条件をより統制し、自己確証の特性をもつ RCT は、「T によって O が生じた」を「より真と言える」ものとして語ることを可能にするのである。

5．課題

本稿は、介入の有効性に関する議論の土台をより頑丈にし、その議論をより建設的にするべく、まず不毛な議論に陥る要因である存在論的、認識論的（言語論的）問題を踏まえた上で、建設的議論の土台構築への道筋を明示した。そしてその道筋に沿い、漸進的に頑丈な道を建設、舗装するべく、CSR モデルの有効性に関する評価で用いられた方法 RCT の相対的優位性を示した。

最後に次なる課題を示す。本稿で優位性を示した RCT もまた課題を持つ。実験の場（field）という極めて限定された条件の下で導出された因果関係を、いかにして実験の場以外の場に応用していくか、ということである。これについて典型的な主張は、実験と同質の条件を整える必要

がある、というものである。連邦教育省の発行する SBR に関するガイダンスにおいても「あなたの学校あるいはクラスでエビデンスに基づいた介入を実施する際、その介入の詳細を、忠実に遂行することの重要性は、十分に理解されていないことが多い。実施の詳細は時に・・・介入の効果において大きな相違を生むことがありうる」[20]と、同様の内容が示されている。

しかし現実に同質性の要求を実現する（同じような子どもが存在し、同じように行為し、同じように反応をし、同じような結果を出す）ことがどれだけありうるのだろうか。この問題に取り組むに際しては、「そもそも因果関係をどのように捉えるか」という認識論的問題が関わってくる。RCT で発見された因果関係のみが「本当の」因果関係であるか、他の方法においても因果関係は発見しうるのか。これについてはカートライトがまた答えているところだが、次の課題にしたい。

---

[1] The What Works Clearinghouse, Glossary of terms（http://ies.ed.gov/ncee/wwc/help/glossary/）

[2] 志水宏吉「学校の力―「効果のある学校」の量的分析」『「力のある学校」の探求』大阪大学出版会、2009 年。

[3] U.S. Department of Education, Comprehensive School Reform (CSR) Program Guidance, 2002

[4] CSR プログラムの補助金は、毎年競争的資金として、各学校 3 年契約で最低 5 万ドルが与えられる。補助金は一度州教育当局へ割り当てられ、その後、地方教育当局が学校の代わりに州教育当局へ補助金の再割当を申請する、という形式をとっている。なお、CSRD の際は、構成要素は 9 つであった。

[5] CSRD の構成要素の際は「革新的方策や立証された方法（Innovative strategies and proven methods）―信頼性のある研究や効果的実践に基づき、その有効性が立証された方法（proven method）や革新的方策、また多様な特徴を持った学校間で再現性をもってきたと立証される方法や革新的方策」という表記であった。

[6] Robert E. Slavin, Evidence-Based Education Policies: Transforming Educational Practices and Research, *Educational Researcher*, Vol.31, No.7, 2002, p. 15.

[7]「SBR」とは、①体系的及び経験的方法②厳格なデータ分析③信頼ある妥当なデータ収集④強力な研究デザイン⑤再現を可能にする詳細な結果⑥結果が慎重な審査を受けることの 6 つの基準すべてを満たす「質の高い研究」から成る最も水準の高いエビデンスのこと。「強力なエビデンス」とは、6 つのうち 5 つの基準を満たす「妥当な質の研究」と「質の高い研究」が混合して成るエビデンスのこと。（U.S. Department of Education, Scientifically based research and the Comprehensive School Reform (CSR) Program, 2002）

[8] Robert E. Slavin, What Works? Issues in Synthesizing Educational Program Evaluations, *Educational Researcher*, Vol.37, No.1, 2008, p.5.

[9] ibid., pp.5-6.

[10] The Comprehensive School Reform Quality Center, American Institutes for Research, CSRQ Center Report on Elementary School Comprehensive School Reform Models, Washington, DC, 2005.

[11] Herman, R., Aladjem, D., McMahon, P., Masem, E., Mulligan, I., O'Malley, A. S., et al., An educators' guide to schoolwide reform. Arlington, VA: Educational Research Service, 1999.

[12] Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S., Comprehensive school reform and student achievement: A meta-analysis, Baltimore: Center for Research on the Education of Students Placed At Risk, 2002.

[13] 例えば、過去の事実に関する interrupted time series designs は、縦断的研究や縦断的コホート研究よりもより厳格であると考えられるという。

[14] ラリー・ラウダン（戸田山和久訳）『科学と価値』勁草書房、2009 年、243 頁。（Larry Laudan, *Science and Values: The Aims of Science and Their Role in Scientific Debate*, The Regents of the University of California, 1984.）

[15] 同上、231-234 頁。

[16] Cook, T. D., and Payne, M. R., Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch（Eds.）, *Evidence Matters: Randomized Trials in Education Research*, Washington, DC; Brookings Institution Press, 2002.

[17] William R. Shadish, Thomas D. Cook, Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Wadsworth, Cengage Learning, 2002, p.12.

[18] Nancy Cartwright, Evidence-based policy: what's to be done about relevance?, the 2008 Oberlin Philosophy Colloquium, Springer, 2008.

[19] 被験者、調査者、評価者、データ分析者の 4 者に対して盲検法をとるもの。

[20] U.S. Department of Education, Coalition for Evidence-Based Policy, Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide, 2003.

（日本学術振興会特別研究員　比較教育政策学講座　博士後期課程 3 回生）

（受稿 2011 年 9 月 2 日、改稿 2011 年 11 月 25 日、受理 2011 年 12 月 26 日）

What Enables US to Say What Works? The Methodology Used to Evaluate Models' Effectiveness: Focusing on the Models Used in the Comprehensive School Reform Program in the USA

KIRIMURA Takafumi

It may be true that improving student outcomes is the ultimate purpose of all educational policymakers and educational researchers. Improving the effectiveness of the intervention, which involves an educational program, a product, a practice, or a policy aimed at improving student outcomes is a typical purpose to be accomplished. However, it should be noted that this study addressed the question "how and why can you say that intervention X caused the outcome Y?" or "what gives assurance that intervention X caused the outcome Y?" and not "did intervention X cause the outcome Y?" That is, the purpose of this study was not to discuss the effectiveness itself but pave the way for discussing the effectiveness. The target of this paper is the methodology used in the Comprehensive School Reform Quality Center report, which evaluated and rated the effectiveness of models used by the schools in the Comprehensive School Reform program. The characteristic of the methodology is a Randomized Controlled Trial (RCT), which is the "gold standard" for evidence-based policies in the USA. This paper demonstrates why the methodology is superior to the econometric methodology based on the theory of Cartwright.