研究委員会企画チュートリアルセミナー 1

教育・臨床実践のための介入効果の解析と評価

Analysis and Evaluation of Intervention Effects in Educational and Practical Studies

Tomoyuki YASUDA AND Takahiro HOSHINO

話題提供 安田 節之 (西武文理大学)
企画司会 星野 崇宏 (東京大学)

企画の趣旨

星野 崇宏

知覚心理学や生理心理学などの基礎心理学分野はもちろんのこと、教育心理学研究においても研究協力者をランダムに各条件に割り当て、条件間の比較を行う「無作為割り当てを伴う実験研究」を行った研究の方が、実験を伴わない調査研究や観察研究、あるいは介入研究よりも“高度”であり“より正確な結果を与える”という一種の思い込みが存在するようである。実験状況で生じるホーソン効果などの生態学的妥当性の問題はもちろんのこと、実験研究は往々にして本来研究の対象となる人々とは異なる人々（例えば大学生）をターゲットとしたものになりがちであり、アナログ研究としての問題点（対象者が異なることによって生じる外的妥当性の問題点）を有している。

また基礎心理学では錯覚やプライミングなど短期間しか効果が持続しないものを対象にする場合が多いが、教授法や発達などでは長期的な効果が存在する要因について関心があり、これについて無作為割り当てによる実験研究を行うことは倫理的にも難しく、また実施したとしても指示した条件に従わない不服従や、従うのに難しい条件群では脱落が起きるなどの問題があり、無作為割り当てをしても通常の群間差による比較では要因の効果を推定することができない場合も多い。

また、そもそも教育実践や臨床実践を志向する研究や実践の現場で行う研究では、対象者に特定の治療条件を強制することは倫理的な理由から難しく、対象者がどの条件を受け入れるかを選択するという場合が多い。教育心理学研究において多く生じる、実験ができない調査観察研究から介入効果を評価するために、様々な準実験デザインとそこでの分析方法やデータ取

得の注意点などについて企画者が紹介した。話題提供者からはより広い意味でのプログラム評価の考え方と方法論について、ロジックモデルの重要性を中心に、コミュニティ介入研究などの実例を交えながらご講演頂いた。

講演：調査観察研究での研究デザインと解析の基礎

星野 崇宏

無作為割り当てが行えない調査・調査研究では教授法や学習方略の効果など、観察された群間差が背景要因の群間差による見かけ上の効果か、本当に関心のある要因によって生じているのかを判別することが難しいとされる。一方、実験研究では背景要因の群間差は標本誤差によるものになるため、群間差を関心のある要因によるものとみなせるとされる。このような理由から「相関と因果は異なる」「実験研究を行わないと因果関係を議論できない」という言説が無批判に繰り返されたが、これは実践的な研究関心を持つ多くの研究者に質的研究を志向させる原因の一つになったと考えられる。

国際的な動向としては無作為割り当てが行えない調査研究や観察研究から介入効果やプログラム評価を行う場合でも、質的研究ではなくあくまで量的な研究の枠組みでの研究が行われることが多く、この観点からは国内の教育心理学研究が世界的な潮流から大きく取り残されている感が強い。

また因果関係の理解に関する科学哲学的な議論や疫学研究での因果関係の評価基準などからも、相関と因果は質的に異なるものではなく連続的なものであり、また実験研究でなくても因果関係は議論できること、更には実験研究には企画趣旨にも述べたように様々な問題点を有しえることが分かっている。

非実験研究でも因果関係を議論できることを説明するために、まずは現代的な統計的因果推論の枠組みであり、初めて提案されたのは米国の教育心理学研究誌

でる Rubin の因果モデル (Rubin, 1974) を説明する。このモデルは複数の条件ごとに「もしその条件に割り当てられたときの従属変数の値」という潜在的な結果変数を想定して議論を行う、いわゆる“反実仮想モデル”といわれる。

具体的に低学力層に対してチューターを付けたり、放課後に補習を実施する特別な教育プログラムの因果効果を考えるとする(教育プログラムの代わりに早期教育や教授法、ライブイベントなどを考えれば、様々な研究に一般化可能であることが分かるだろう)。

ここで特別プログラムに参加する生徒は親や教師、または生徒自身にとってプログラムを受ける動機がある、ニーズがあるような生徒であり、通常事前の成績が低く、家庭での学習環境がよくない場合が多いと考えられる。従ってこのプログラムへの割り当ては無作為ではない。

さて、ここでこのプログラムを1年間実施するとし、1年後のテスト得点への因果効果を考える際には、潜在的な結果変数として2つの変数 y_1 「特別プログラムを1年受けた後でのテスト得点」、 y_0 「通常の教育を1年受けた後でのテスト得点」を考える (Figure 1)。 $y_1 - y_0$ は同じ個人での2つの条件の差であるため、この母集団での平均(期待値)を Rubin の意味での教育プログラムの因果効果と呼ぶ。しかし実際には特別プログラムに参加した人については y_0 が、通常の教育を受けた人では y_1 が観測されておらず、 $y_1 - y_0$ をすべての人について計算することはできない。

もし両群が同質な群であれば、 y_1 の全体での平均は

特別プログラム群での平均と等しくなることが期待され、同様に y_0 の全体での平均も通常教育群でのそれと等しくなる。つまり両群を同質にする無作為割り当てによって、通常の群間差が因果効果の正しい推定値になる。しかしこの特別プログラムのように、無作為化されていない場合には両群は事前の成績など様々な観点で異なるため、単純な群間差によって因果効果を推定できない。

このような場合に因果効果を推定する方法として、近年様々な背景要因(共変量や交絡要因と呼ばれる)の影響を除去する様々な手法が開発されている。ここでは紙面の制約の都合上詳しく説明しないが(詳しくは星野, 2009などを参照)、因果効果推定法の開発はこの20年ほどの間に統計学において急速に進展した研究領域である。そのルーツは教育心理学にあるだけに、本来的には教育心理学研究と高い親和性を有する。メタ分析などと同様に、応用研究への普及は医学疫学研究で先行していたが、この10年間で社会科学での調査観察研究でも必須とされつつあり (APA の Journal Article Reporting Standards, 2008など)、国内でも近年やっと認知されるようになった。

本講演では、調査観察研究から因果効果を推定するための様々な研究デザインと、因果効果を推定する際に必要な背景要因(共変量と呼ぶ)の選択方法、およびそれを利用した統計的な調整法について、数理的な知識を要求せずにその本質と応用研究での利用のあり方について説明を行った。

* 網掛け部分は実際は得られないデータの部分
無作為割り当てなら無視してよいが、観察研究では無視できない

Figure 1 欠測データと Rubin の因果効果

講演：心理教育プログラムの効果実証に向けた プログラム評価の導入方法

安田 節之

プログラム評価とは

プログラム評価におけるプログラムとは「何らかの問題解決や目標達成を目的に人が中心となつて行う実践的介入」のことを指し、そのプログラムを評価することを一般にプログラム評価と呼ぶ。プログラム評価は主に、欧米の心理学、教育学、社会学などの社会科学領域において発展を遂げてきた。

プログラム評価の発展初期である 1950 年代から 1970 年代には、実験心理学の手法を用いたプログラム評価の方法論が整備され、特に実験・準実験デザイン (experimental and quasi-experimental design) を用いたデータに基づく教育プログラム及び社会・公共サービスの変革が目指された (例：Campbell, 1969；Shadish, Cook., & Campbell, 2002)。

このような流れのなかで、様々な評価のアプローチや方法論の体系化が行われた。なかでも 1979 年の初版出版から版を重ねてきたロッシ (P. H. Rossi) らの「Evaluation: A systematic approach」では、「評価とは、社会調査の方法を活用し、社会プログラム等による介入の効果を体系的に研究することである。これらのプログラムは、政策的・組織的な文脈・環境において用いられるものであり、社会状況を改善するためのソーシャルアクションの情報源となるものである」と定義され (Rossi, Lipsey, Freeman, 2004, p. 16)、社会調査をベースとしたプログラム評価の理論が提唱された。

一方、実験・準実験デザインの方法論に偏ったプログラム評価は、政治・社会経済的な要因の影響や倫理的配慮の必要性が存在するフィールド (実践現場) においては、非現実的かつ視野狭窄であるという見方が専門家の間で広まっていった (例：Shadish & Luellen, 2013)。このようななかで、実用重視型の評価 (utilization-focused evaluation) を提唱したパットン (M. Q. Patton) は、「プログラム評価とは、プログラムの活動、性質、アウトカムの情報を体系的に収集し、当該プログラムについて何らかの判断を下し、プログラム介入による効果の改善を行い、将来のプログラムについての決定を行うことである」とし (Patton, 1997, p. 23)、調査活動よりも踏み込んだ、実践的・実用的な評価理論を体系化した。

さらに、理論主導型の評価 (theory-driven evaluation) を体系化するなかでチェン (H. T. Chen) は、「プログ

ラム評価とは、プログラムを体系的に査定し、プログラムの計画、実施、効果を向上するために、評価のアプローチ、テクニック、知識を応用することである」 (Chen, 2005, p. 3) とした。

以上の定義をみるとプログラム評価は、人々の意識・態度・行動などを明らかにしようとする通常の実社会調査や実験データによって仮説検証を行う心理実験とは違い、アクション志向が強い実践的なアプローチであることが分かる。

このような流れを鑑みつつ、安田・渡辺 (2008) はプログラム評価を「特定の目的を持って設計・実施されるさまざまなレベルの介入活動およびその機能についての体系的査定であり、その結果が当該介入活動や機能に価値を付与するとともに、後の意思決定に有用な情報を収集・提示することを目的として行われる包括的な探究活動」と定義し (p. 5)、評価研究に基づく介入・実践の改善や質向上のあり方を検討した (例：安田, 2011)。

プログラムの可視化・構造化

プログラムを実施する目的は様々であり、その規模 (small-/medium-/large-scale) はもとより、介入のレベル (個人・組織・地域コミュニティ)、介入方法 (直接的介入・間接的介入)、評価主体 (自己評価・外部評価)、評価目的 (説明責任・実践の質向上・価値判断・評価研究等) も多種多様である。プログラムやサービスは、“モノ” と違いその存在が確認しにくく、その効果も見えにくい。また、得られる効果も人それぞれ違うことが多い。

さらにプログラムを評価する際には、評価方法や評価基準だけでなく、様々なステークホルダーの存在からの影響を受けることになる。研究者は何らかの研究データに基づくプログラムの評価を望むものの、一般にプログラムの価値の査定や意思決定は、社会状況や政治的判断から市場 (マーケット) の影響まで、実に多くの要因が関連してくる。政策レベルの大規模なプログラムの場合は、研究データ以外の影響が特に大きくなる。

このような特徴をもつプログラムの評価を行うには、まずプログラムの可視化・構造化が必要になる。具体的には、1) プログラムの目的を確認・再確認するためのゴールの構造化を行い、2) プログラムの介入によって参加者・利用者・クライアントにどのような効果・成果がどのような道筋で及ぼされるのかをインパクト理論などにより確認し、3) 期待される効果・成果をもたらすためにどのようなプログラムの運営を行うのかをロジックモデル (W. K. Kellogg Foundation,

Figure 2 ロジックモデルの概要

2004) を用いて可視化することが有用となる (Figure 2)。以上の点をもとに、4) 評価の焦点や優先順位を絞り、研究活動におけるリサーチクエストと同様に、評価活動における評価クエストを設定し、実際のデータ収集・分析等を行うことになる (安田, 2011; 安田・渡辺, 2008)。

これらの4つの基本ステップは、「評価実践」において重要な視点であると同時に、プログラムの実践者や評価者を含むステークホルダーの“評価マインド”の涵養そして評価のキャパシティ・ビルディング (evaluation capacity building) にもつながる視点である。さらに、「評価研究」のための方法論や「評価教育」のあり方を検討するうえでも、これらのステップは実用的である。

エビデンスに基づく実践とプログラム評価

エビデンスに基づく実践 (evidence-based practice) とは「研究による最適な科学的根拠に臨床の専門的知識・技術と患者の価値尊重を統合したもの」(Institute of Medicine, 2001, p. 147) のことを指し、心理学領域では「研究による最適な科学的根拠にクライアントの特徴・文化・嗜好を考慮した上での臨床の専門的知識・技術を統合したもの (APA Presidential Task Force on Evidence-Based Practice, 2006, p. 273)」と定義されている。

ここでの“研究による最適な科学的根拠 (best research evidence)”とは、つまるところ、ランダム割付か否か、比較群が存在するか、サンプル数は十分であるか、統計的有意性があると同時に効果量の程度は高いか、といった方法論の厳格性 (methodological rigor) に拠るところが大きい。しかしながら、ランダム化無作為試行 (RCT: randomized control trial) に代表される方法論に厳格な実験デザインを適用したプログ

ラム評価は、ヒューマンサービス領域では数少ない。利用者やクライアントへの倫理的配慮や現場の理解 (“buy-in”), そして現場への方法論の技術移転 (technology transfer) が十分でない現状を踏まえると、実験デザインのような厳格な評価方法を用いた評価に堪えるプログラムはごくわずかと予想される。

特に、地域や組織で実施される小規模・中規模なプログラムを評価するにあたっては、以上のようなアプローチは事実上、不可能である場合が多いのが現状である。そこで、より現実的な評価研究の視点や方法が必要となってくる。ここではその一つの試みとして、既述のロジックモデルの枠組みを用いたプログラム (地域での健康増進プログラム) の評価事例についてみていく (安田, 2013, 2014)。

まず安田 (2013) では、プログラムを簡易的かつ体系的に捉えるために、当該プログラムのステークホルダー (N=36) であるプログラムの運営者と指導者からの情報をもとにしたベンチマークを開発し、活動の改善・質向上につなげることが目的とされた。ステークホルダーは、1) プログラムを実施するうえで重要だと考えていること (significance), 2) 活動が参加者または地域に対してどのような貢献をしていると思うか (contribution) の2点について自由記述による回答を求められた。そこでの回答をロジックモデルの視点 (インプット・アクティビティ・アウトプット・アウトカム・インパクト) から分類しベンチマーク指標が作成され、これらの指標への回答データをもとにプログラムの改善点が検討された。ロジックモデルの各要因がさらに正確に測定できるようになれば、例えば、インプットに対するアウトカムの割合である「有効性」(アウトカム/インプット) やインプットに対するアウトプットの

割合である「効率性」(アウトプット／インプット)に関する評価研究がスムーズに行えると言える。

一方、ロジックモデルで示されるプログラムの「機能(function)」の他に、そのプログラムの効果を高める要因と考えられる組織の「キャパシティ(capacity)」に着目したのが安田(2014)である。ここでは、プログラムの「関係機関および地域とのパートナーシップ」や「社会生態学的コンテキスト」を踏まえたアクションモデル(Chen, 2005)を参考に、プログラム実施にあたっての「レディネス」「リーダーシップ」「エンパワメント」「コミュニティ感覚」がロジックモデルのどの要因に影響を及ぼすかが検討された。その結果、プログラムや実施組織に対するつながりを意味する「コミュニティ感覚」とロジックモデルの各要因(アクティビティ・アウトプット・アウトカム)との関連性が認められた。このようにプログラムの実施組織のキャパシティ(構造や次元など)の検討は、効果的なプログラムを開発・実施するうえで重要な視点となると考えられる。

プログラム評価領域における今後の課題

プログラム評価を導入するうえでより発展が期待されるものに、1) プログラムのプロセスを可視化するツールの充実、2) 評価や効果測定スキルを持った人材の育成、3) プログラムの評価を行った後のアクションの質向上が挙げられる。

まず、プログラムのプロセスの可視化には、前述のゴールの構造化やロジックモデルの活用が考えられる。今後、これらの枠組みに沿ったツール等の開発がさらに進むことにより、効果的なプログラムの開発・実施につながると言える。同時に、様々なステークホルダーへのアカウントビリティ向上が期待できることになる。

2点目の評価スキルを持った人材に関しては、評価教育・研修によって人材育成を行い、フィールドで評価の専門家として評価実践を行う通常のアプローチが考えられる。しかし、プログラムの質を高めるための自己評価に限定すれば、すでに評価スキルを持った研究者がプログラムの実施者・実践者に対して関連する評価の方法論(例:プログラムの可視化・構造化や簡単なデータ収集・分析の方法)の技術移転(technology transfer)を行い、その後、研究者が必要に応じて技術支援(technical assistance)を行うというアプローチも考えられる。実際、このような視点は、プログラムの実践者の評価のキャパシティ・ビルディングを行ううえでも有用になると言えよう。

最後に、プログラム評価の後のアクションについては、評価結果の利用・活用(use)に着目することが重要となる(例: Patton, 1997)。これは、エビデンスの活用方法の検討にもつながる視点である。プログラムの効果に関するデータや情報は、一般に、論文や報告書などの形で研究者側に対して発信されることが多い。しかし、プログラムの実践者に対してエビデンスを公開・提供する情報センター(clearinghouse)などを充実させ、プログラムの効果に関する情報やデータを一元化し公開することにより、エビデンスの構築を行う研究活動とそれを活用してよりよいプログラムを実施する実践活動とのよい相互作用が生まれることになる。その結果、より望ましい形でのエビデンスに基づいた実践につながると言える。

まとめと議論

2つの講演を踏まえ、参加者を含めた議論が行われた。そのうちのいくつかの重要な観点について記載したい。

1: 実践場面での介入効果とプログラム評価のための研究デザインと解析法の開発及び理解の促進の重要性

日本のアカデミック環境が厳しくなり、また社会環境の変化から、教育心理学分野の研究者も学校での実践研究や臨床研究にとどまらず、社会連携活動や自治体によるプログラムなど幅広い場面における実践的な研究や介入法の評価を行うことが要請されるようになっていく。

そのような現状を踏まえ、教育心理学分野の研究者が理想とする介入評価法と実践場面でのそれが大きく乖離していることが指摘された。例えば無作為割り当てが難しいというのはもちろん、対照群を置いた比較ができない場合や、事前事後の測定が十数人に対してしか行えないなど、研究デザインやサンプルサイズの制限の中でいかによりよい意思決定を可能にする介入効果やプログラム評価の在り方が可能かについては非常に関心が高いことが分かった。

2: 教育心理学分野ならではの貢献のために必要な観点

参加者からは、教育心理学分野の研究者が実践研究やプログラム評価に加わる場合、心理学的構成概念をアウトカム又は説明変数、中間変数として測定することが本来必要と考えられる場面が多いにもかかわらず、実践的な場面では諸事情により妥当性や信頼性が担保された尺度や観測評定項目を測定することが困難な場面が多いという難しさがあることが指摘された。これ

に対しては、近年盛んになっている実証的な経済学研究（例えば労働・教育経済学や行動経済学など）では例えばIQや性格特性などで既存の尺度との相関が高いことが確認されている3項目程度の短縮尺度が盛んに利用されていることが紹介され、国内の教育心理学研究でもより実践的な場面に有用な短縮尺度や評定法の開発も重要な研究分野であるのではとの意見があった。

引用文献

- APA Presidential Task Force on Evidence-Based Practice (2006). *Evidence-based practice in psychology*, **61**, 271-285.
- APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, **63**, 839-851.
- Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, **24**, 409-429.
- Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.
- 星野崇宏 (2009). 調査観察データの統計科学—因果推論・選択バイアス・データ融合— 岩波書店
- Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.
- Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.
- Rossi, P., Lipsey, M., & Freeman, H. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage. (ロッシ, P.・リプセイ, M.・フリーマン, H. 大島 巖・森 俊夫・平岡公一・元永拓郎 (監訳) (2005). プログラム評価の理論と方法—システムティックな対人サービス・政策評価の実践ガイド— 日本評論社)
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton Mifflin.
- Shadish, W. R., & Luellen, J. K. (2013). Donald Campbell: The accidental evaluator. In M. C. Alkin (ed.). *Evaluation roots: Tracing theorists' views and influences* (2nd ed.) (pp. 61-65). Sage.
- W. K. Kellogg Foundation (2004). *Using logic models to bring together planning, evaluation and action: Logic model development guide*. Battle Creek, MI: Author.
- 安田節之・渡辺直登 (2008). プログラム評価研究の方法 新曜社
- 安田節之 (2011). プログラム評価—対人・コミュニティ援助の質を高めるために— 新曜社
- 安田節之 (2013). ステークホルダー情報に基づいた予防介入プログラムのアセスメント—ベンチマーキングによる評価研究の予備調査として— 西武文理大学サービス経営学部紀要, **23**, 27-36.
- 安田節之 (2014). コミュニティ介入の効果を高める組織特性の検討—臨床心理地域援助における評価研究の試みとして— 臨床心理学, **14**, 402-411.