

ベストプラクティス・アプローチに基づく評価研究(1)：エビデンスに基づく実践との比較から考える

安田, 節之

(出版者 / Publisher)

法政大学キャリアデザイン学会

(雑誌名 / Journal or Publication Title)

Lifelong Learning and Career Studies / 生涯学習とキャリアデザイン

(巻 / Volume)

16

(号 / Number)

1

(開始ページ / Start Page)

165

(終了ページ / End Page)

175

(発行年 / Year)

2018-11

(URL)

<https://doi.org/10.15002/00021442>

ベストプラクティス・アプローチに 基づく評価研究（1）： エビデンスに基づく実践との比較から考える

法政大学キャリアデザイン学部准教授 安田 節之

本稿の目的は、ベストプラクティス・アプローチ（BPA）に基づいたプログラム評価研究の前提を確認することである。BPAは、成功事例の検討など、保健・医療・福祉・教育など多くの領域で非公式に、しかし日常的に取り入れられていることが多い。他方、そのような最良・最善な実践（ベストプラクティス）を体系的に研究するアプローチや方法論に関しては、少なくとも対人・コミュニティ援助などのヒューマンサービス領域においては、未だ発展途上である。

そこで本稿では、今後のBPAの研究展開に向けたイントロダクションとして、まずベストプラクティスの考え方を概観する。そのうえで、従来のプログラム評価における基本的立場とされるエビデンスに基づく実践（EBP）を詳しくみていく。そして、BPAに基づくプログラム評価研究の今後の方向性を検討する。

ベストプラクティス・アプローチ

BPA¹⁾は、その名が示す通り、「優れた（最良な）実践（best practice）」から多くを学び、実践の質を高めることを基本としている。BPAはもともと、工学分野におけるモノづくりやサービスのマネジメント（経営）領域で発展してきたアプローチである。具体的には、優れた実践を成功事例とし、それを照合枠としつつ、他の実践活

動・プログラムが行われている文脈、組織、場面のニーズに応じて、体系的に導入・実装されることを目的としている。そのためには、照合枠となるベストプラクティスを包括的に検討する一連の手続きが必要となる。それが「ベンチマーキング（benchmarking）」と呼ばれる手法である。

BPAは、優れた実践、つまりベストの中のベスト（best of the best）を探して、その実施過程や手法をベンチマーキングするところに特徴があるが、このアプローチを提案したキャンプ（R. Camp）は、作業仮説（working definition）のなかで「ベンチマーキングはベストプラクティスを探求し、最高の経営を実現することである」とした（Camp, 1989, p. 12; 田尻 訳, 1995, p. 24）。ベンチマーキングは単なる研究開発を目的に始まったのではなく、組織が抱えているマネジメント上の問題をどうしたら早く解決できるのかを学びたいという現場の要請から生まれたものとされている（Watson, 1993）。

そのきっかけとなったのが、1970年代後半の米国ゼロックス社での出来事である。当時、ゼロックス社の複写機の販売が低迷し、日本のキャノンやリコーに首位の座を奪われるなか²⁾、同社は復活をかけてこれらの競合企業の製造工程、製品構成、生産コストに関する包括的ベンチマーキングを行った。さらにゼロックス社の欠点であった製造部品の調達・管理の仕組みを改善する

ために、まったくの異業種であるアウトドアメーカー、L.L. ビーンの在庫管理システムを参考にした。このように組織内・外を問わず、最良とされる実践を体系的に参考にした問題解決を図った結果、見事に業績を回復させたという出来事である (Watson, 1993)。

ヒューマンサービス領域への BPA の応用可能性

BPA は、非営利な活動を基盤とするヒューマンサービス領域においても応用可能な部分が多いと考えられる。BPA の目的は他組織・他領域のベストプラクティスから学ぶこと、即ちニーズ (問題・課題) のある組織が、ある領域で資源 (知識・スキルなど) を持つ組織に出会い、そこでの学習経験によって自組織の活動の改善を目指すところにある (Watson, 1993)。このような BPA は、同業種はもとより、多職種間の連携・協働を特に重視する非営利のヒューマンサービス領域にむしろフィットする可能性が高い³⁾。ベンチマーキングはそのような学習を促進するツールとなる。

BPA の考え方は、従来の社会科学諸領域のアプローチとは異なる点も多い。なかでも、常に科学的な実証結果をベースとして考えていくエビデンスに基づく実践 (EBP) と BPA とは、支援や援助の質向上という共通の目標はあるものの、物事や行動の判断基準などが認識論 (epistemology) のレベルで違ってくる⁴⁾。そこで以下では、EBP の考え方や方法論について詳細に検討し、今後の BPA の研究の方向性を模索する。

エビデンスベースド・アプローチ

エビデンスとは

保健・医療・福祉・心理といったヒューマンサービス領域や教育領域はもとより、さまざまな公共政策の領域においてエビデンスの重要性が指摘されるようになって久しい (例：エビデンスに基づく政策形成)。エビデンスは、説明責任を意味す

るアカウンタビリティとともに、政治や経済を含めたあらゆる場面に登場するようになり、現在では一般的に用いられるようになった。

エビデンスの概念が普及するなかで、その捉え方や用いられ方も多様になってきている。例えば、後に詳しくみていく実験デザインといった科学的手法を用いて根拠を提示することをエビデンスとすることもあれば、定量的なデータ (数値) を総称してエビデンスと呼ぶこともある。さらに言語や映像データといった定性的・質的な情報がエビデンスとして扱われることもある。

このように、エビデンスは様々な手段や目的で用いられる。一方で、より専門的な意味におけるエビデンスには、それを成立させるために必要な条件がある。それが科学的に実証された「因果関係 (cause-and-effect)」である。

因果関係

因果関係が存在するとは即ち、何らかの介入 (実践・支援・活動・プログラムなど) という「原因 (cause)」があり、その原因によって、介入を受けた人々 (利用者・学習者・クライアントなど) に、多くの場合肯定的な「結果 (effect)」が生じるという方程式が成立することを意味している。因果関係 (エビデンス) を提示するとは即ち、原因と結果のメカニズムを示す因果関係の方程式が本当に成り立つのか (‘解’が存在するか) を実証することに他ならない。

古くは哲学者のミル (J.S. Mill) によって、「先行性」「関連性」「唯一性」という因果関係を成立させるための 3 つの条件が提示された。介入 (原因) が効果 (結果) よりも先に起こっていて (先行性)、介入と効果が関連する、つまり両者に説明可能な関係性が存在し (関連性)、さらに効果を生み出している原因が当該介入のみである (唯一性) ことが明らかになってはじめて、因果関係があると結論できるのである (参考：安田, 2011, p.195)。逆にこれらのどれかが欠けると、いくら肯定的な研究結果が得られたとしても、エビデンス (因果関係) が実証できたとは言えない

のである。

ランダム化無作為試行

この先行性・関連性・唯一性を確保し、因果関係（エビデンス）を実証するために、例えば、時系列データを用いた分析や高度な統計モデリングを駆使した分析など、様々なアプローチが模索されている。これらのアプローチは、技術的に難易度が高いものも数多く存在し、またその技術も日進月歩で著しく発展しつづけているのも事実である。

しかし一方で、因果関係を成立させるための条件を比較的シンプルに担保できる方法が存在する。それが「ランダム化無作為試行（RCT：Randomized-Control Trial）」である。RCTを活用した実験を行うことにより、先にみた3つの条件および因果関係成立の妨げとなる外生要因（例：成熟、履歴、テストニングなど）のほぼすべてを統制することが可能となる。RCTが因果関係を証明するための黄金律（gold standard）と呼ばれるのはそのためである。

RCTの構造は容易に示すことが可能である（図1）。主な点としては、介入群の他に、あえて介入を受けない「統制群」を設置すること、そして被験者（利用者・学習者）をどちらかの群にランダムに配置するという「無作為化（ランダム化・無作為配置）」を行うことの2つに集約できる。このような条件をもとに実験を行うところにRCTの特徴がある。

RCTの代表形（無作為化統制群事前事後テストデザイン）は、無作為化を「R」、介入群を「X」、効果測定（事前・事後）を「O」とすると、下記のとおりに表示する：

R	O ₁	X	O ₂
R	O ₁		O ₂

図1. RCTの基本形

例えば、新しい教育プログラム（実験的介入）を実施し、その効果を実験デザインによって確かめるとする。その際には、当該プログラムの他に、

あえて教育的介入を行わない群（統制群）を設置し実験を行う⁵⁾。そして、どちらの群に配置されるかを無作為に決定するのである。このような手順を踏みデータを収集・分析し、新しい教育プログラムが（統制群と比べて）確かに効果がある（例：数値が高い）という、因果関係が認められれば、それを科学的根拠（エビデンス）とすることが可能になるのである（図2）。

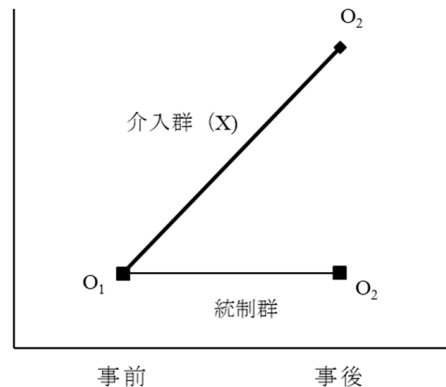


図2. RCTによる結果の例

理論上はスムーズに運用ができそうなRCTであるが、実践面においては、困難な問題を抱えることが多くなる。既述のとおり、RCTは実験デザインであり、その条件として、統制群を設置し無作為化をする必要がある。このような実験デザインを、教育プログラムを受ける学習者や対人・コミュニティ援助サービスの対象となる利用者に対して用いることは、運営面はもとより、倫理面・不公平性（例：介入・支援を受けない群が存在する）といった点を考慮すると、現実的でないことがほとんどとなる。そのため、因果関係の検証を行うにあたっては、RCTを理想形としつつ、それに少しでも近い形での実験・準実験デザイン（例：統制群ではなく従来の教育プログラムを比較群として設置する）を用いた実証研究が行われることになる（参考：安田，2011, p. 192-228; 安田・渡辺，2008, p. 109-149）。

エビデンスに基づく実践

科学的根拠となるエビデンスをもとに介入を行うのが EBP と呼ばれる実践方法である。

EBP を定義し積極的に取り入れたのは医学領域であり、EBP は「研究による最良な科学的根拠に臨床の専門知識・技術および患者の価値（観）を統合したもの」（Institute of Medicine, 2001, p. 147）と定義された⁶⁾。またこの定義を参考に、心理学領域では EBPP（Evidence-based Practice in Psychology）を「研究による最良な科学的根拠に、クライアントの特徴・文化・嗜好を考慮したうえでの臨床の専門知識・技術を統合したものである（APA Presidential Task Force on Evidence-Based Practice, 2006, p. 273）と定義している⁷⁾。さらにこれらの定義を参考にと、エビデンスに基づくプログラム（program）は、「プログラム研究による最適な科学的根拠に、プログラムの実施者を含むステークホルダーの専門知識とマネジメント技術および利用者の価値（観）を統合したもの」とすることも可能であろう。

以上の定義でみた「研究による最良の科学的根拠（best research evidence）」とは即ち、実証研究によって示された治療（介入）と効果との間の科学的根拠（因果関係）を意味している。その

ような“ベストエビデンス”を基軸として、臨床の知と技および患者からの視点（価値）をすり合わせたものが EBP となるのである。

エビデンスのレベル

一口にエビデンスと言っても、その質が高いものやそうではないものがある。このことは、エビデンスの証明方法が複数存在することを踏まえると想像に難くない。それでは、質の高いエビデンスはどのようにして得ることができるのか。

その鍵となるのが、どれだけ方法論に厳格な（methodological rigorous）デザインによって因果関係が検証されたかということである。つまり先にみた RCT といった科学的な手法やデザインを用いることが、質の高いエビデンスを得るための必要条件となるのである。そして、様々なエビデンスが存在するなかで、その質を問う判断基準を定めたものがエビデンス・ヒエラルキー（evidence hierarchy）である。エビデンス・ヒエラルキーは医学領域によって設置されたのが始まりとされ（例：Sackett et al., 1996）、それぞれのエビデンスのレベル（水準・基準）は表 1 に示される：

表 1. エビデンス・ヒエラルキー（Sackett et al., 2000）

1a	複数の RCT の系統的レビュー（i.e., メタ分析）
1b	個別の RCT ⁸⁾
1c	悉無研究 ⁹⁾
2a	コホート研究 ¹⁰⁾
2b	個別のコホート研究または質が低い RCT
2c	アウトカム研究
3a	ケースコントロール研究の系統的レビュー
3b	個別のケースコントロール研究
4	症例研究、質が低いコホート研究、ケースコントロール研究
5	専門家の意見

表1を見ると、エビデンスのレベルが高いもの（1a、1b）はRCTを活用した実証研究によるもので、そのなかでも、複数のRCTをメタ分析（meta-analysis）によって系統的にレビューしたものが最もエビデンスのレベルが高いとされている。

それ以降に、悉無研究（1c）、実験的介入を伴わないコホート研究（2a）、ケースコントロール研究（3a、3b）などが続く。これらは、実験デザインの次に科学的とされる準実験デザインによる研究結果をもとにしたものである。最後に、症例研究（4）や専門家の意見（5）などはエビデンスのレベルが低いとされている。

エビデンスの質の高さは科学的手法とされる実験デザインを用いたものに始まり、徐々に因果推論の証明が確定的ではないデザインにシフトしていき、最終的に一般化可能性が低い症例レベルでの検討や専門家の主観による判断というもので終わっている。一方、ここで最も質の低いエビデンスとされる専門家の意見であるが、それは明確なデータあるいはデータの批判的吟味を欠いた単なる意見、勘や経験のみに基づいた主観的な意見のことを指している（原田、2015）。

このようにみると、質の高いエビデンスを構築しEBPを行うためには、介入と効果との間に存在する因果関係の科学的な実証が是が非でも必要となるのである。したがって、無作為化（ランダム化）や統制・比較群の設置、そして更なる技術的な点として十分なサンプル数の確保、統計的有意性の有無や効果量の多寡など、先にみた厳格な方法論によるデザインと分析方法を用いることが重要となってくるのである。

プログラム評価研究におけるエビデンスの位置づけ

エビデンスの構築およびEBPの運用は、プログラム評価の領域ではどのように捉えられてきたのか。プログラム評価研究は、後にみるように、従来の実験研究や調査研究の焦点と異なる点も多い。実際のところ、これまでみてきたエビデ

ンスの提示方法を活用するとすると、そもそも実践現場にそのようなデザインが受け入れられるか（buy-in）といった運営面に始まり、利用者・学習者への倫理面・不公平性に対する配慮に至るまで、実に様々な課題をクリアしなければならない。そのようななか、評価研究（者）のコミュニティ（“evaluation communities”）は、過去に何度かあったエビデンススペースド・ムーブメントに対してどのような立場をとってきたのか。

ヒューマンサービス領域のプログラム評価研究におけるEBPの主なルーツは、実験・準実験デザインの開発と普及にある（Campbell & Stanley, 1963; Cook & Cambell, 1979）¹¹⁾。これらのデザインを開発した実験心理学者のキャンベル（D. T. Campbell）と共同研究者は1960年代終盤に、当時のいわゆる「何でもあり（“anything goes”）」となっていた教育政策や社会事業の改革に対して、“「実験による社会改革（Reform as Experiments）」”を提唱した（Campbell, 1969）¹²⁾。社会サービスや教育実践といった様々なプログラムを（RCTなどの）実験・準実験デザインを活用した研究によってすべて検証し、科学的根拠があるプログラムのみを社会として受け入れるという“実験社会（Experimenting Society）”の創造である（Campbell, 1991）。

キャンベルは、社会課題の解決を目的としたプログラムをただ単に行うだけではなく、その価値や意義そして維持・継続の可否の判断は実験結果に委ねられるべきであるとしたのである。これはつまり、実験という“試行錯誤（trial and error）”を重ねることによって、真に価値・意義のあるプログラムを探究することこそが、社会問題の解決のあり方であるという信念から生まれた発想である¹³⁾。このような実験・準実験デザインを基本としたアプローチは、ロッシ（P. Rossi）らによって社会調査（実験）という観点からの体系的な評価研究のアプローチとして積極的に取り入れられ、標準的なプログラム評価の方法論として普及していった（Rossi, Freeman, & Wright, 1979; Rossi et al., 2004; Shadish, Cook,

& Leviton, 1991)。

しかし1980年代に入ると、様々な意思決定に携わるステークホルダー（例：政策決定者・政治家）が、科学的には決して妥当とは言えない研究結果を十分なエビデンスを有したものとして受け止めてしまうことや、また逆に、厳格な実験デザインによる実証実験を行えたとしても、信頼度の高いエビデンスが時宜を得て（timely）、有効活用できる形で（useful）ステークホルダーに提供できるケースが少ないという懸念も生まれた（Donaldson, Christie, & Mark, 2009）。

実際、評価研究者のパットン（M. Patton）は、連邦政府レベルで実施されるプログラムの評価にRCTが用いられたなかで、その評価結果（つまりエビデンス）がプログラムに関する何らかの意思決定（維持・継続・廃止）に活用されたケースはほとんどなかったことを明らかにした。そしてその反省から、実用重視の評価アプローチ（Utilization-Focused Evaluation）を開発し、ステークホルダーによって評価結果が確実に活用（use）されるための評価のあり方を模索するに至ったとされる（佐々木, 2010）。

さらに時を同じくして、1980年代にかけてポストモダンの影響を受け、社会科学領域全般に定性的・質的研究アプローチの方向性が模索され始めた。それに伴って、評価研究領域において、量的研究と質的研究のパラダイム論争（“paradigm war”）が起こった（Donaldson et al., 2009, p. 6; Reichardt & Cook, 1979; Reichardt & Rallis, 1994）。そこでは、論理実証主義の考え方に基づき物事（概念）を操作的に定義し、それらの関係性を線形的に、原因系と結果系によって説明しようとする姿勢など、従来の量的アプローチが拠って立つ考え方への批判が起こった（参考：Yasuda, 2017）。

当然、プログラムによる介入と結果とを因果関係の観点から捉えようとするRCTへの批判も生じた。そのような背景がもととなり、自然主義的探究（naturalistic inquiry）（Lincoln & Guba, 1985）そして構築主義のパラダイムを第

四世代¹⁴⁾と位置づけた第四世代評価（Fourth Generation Evaluation）（Guba & Lincoln, 1989）における評価アプローチの検討が行われた。そこでは、RCTに基づく評価アプローチには否定的な立場がとられ、参加者・利用者を始めとした様々なステークホルダーとの対話のなかで、プログラムにおける体験や経験の意味づけや理解の促進に基づく評価形式が模索された（佐々木, 2010）。

RCTとエビデンスベースド・ムーブメント

その一方で、米国におけるエビデンスベースド・ムーブメントは加速しつづけ、2000年代に入ると、主要な教育改革（例：No Child Left Behind Act of 2001）の効果検証・評価研究には、RCTなどの実験デザインを用いることが求められた。実際に、2003年に米国の教育省の機関（U.S. Department of Education's Institute of Education Sciences）は、実験・準実験デザイン主導による研究に対して、優先的に研究費を配分する方針を打ち出した。

これに対して同年の米国評価学会（American Evaluation Association）の年次大会において、教育プログラムの評価においてRCTなどの実験デザインの活用を新たに政策的に後押しするようなことは、評価研究領域においては暗黒時代（“Dark Ages”）に後戻りするようなもので到底認められないという議論が起こった（Donaldson & Christie, 2005, p. 61）。

この議論をもとに、哲学者・評価研究者であるスクリヴェン（M. Scriven）らを中心とし、学会として、「政策・プログラム評価においてRCTまたはそれに準じた実験・準実験デザインだけがエビデンスを提示できるという考え方はミスリードしている」という声明文（The AEA statement, 2003）¹⁵⁾が米国の教育省に提出された。なかでも、教育政策を評価するにあたって「科学的な研究による評価によって介入効果を明らかにすること（“evaluated using scientifically based research...to determine the effectiveness of a

project intervention”）」には賛同できるが、「介入効果の決定（査定）にあたって、実験デザインを最良の評価方法とする（“evaluation methods using an experimental design are best for determining project effectiveness”）」という点に関しては賛同できないとした（Donaldson & Christie, 2005）。

一方、この声明文について学会内で大きな論争が起こった。評価研究者で心理学者のリプセイ（M. Lipsey）や先にみたロッシ（P. Rossi）やビックマン（L. Bickman）、クック（T. Cook）といったメンバーによって今度は、米国教育省の方針つまり実験デザインを推奨する方針を支持する“NOT the AEA statement”という別の声明文が教育省宛に提出された。その主旨は、教育省の打ち出した科学的根拠の実証に対する実験・準実験デザインの認識は正しく、社会科学領域における方法論のスタンダードとして広く

認知されており、なかでも、他の公共政策（e.g., health and medicine, mental health, criminal justice, employment, and welfare）における科学的な効果検証の最も基本となるものがRCTであるという内容であった¹⁶⁾。（参考：Donaldson & Christie, 2005; Donaldson et al., 2009, p. 2-18; 佐々木, 2010）。

以上のような米国評価学会内での教育政策の評価におけるエビデンスの考え方と実験デザインの活用に関する論争がもとなり、両者のスタンスの違いを明確にし、可能であれば何らかの合意を得ることを目的に、翌2004年の年次大会において、リプセイとスクリヴェンのディベートが開催された（The 2004 Claremont Debate: Lipsey vs. Scriven）。詳細は割愛するが、ディベートの結果、下記のような結論が導き出された（Donaldson & Christie, 2005）：

- RCT は、プログラムのインパクト（i.e., 介入効果の因果関係）の査定において最良の方法であり、プログラムのインパクトを判断することは、現代のプログラム評価において主要な条件となるものである（リプセイ、スクリヴェンが合意）
- 一方で、RCT が教育プログラムの評価に成功裏に活用されることはほとんどなく、プログラムの効果を判断するには、別の選択肢がある（スクリヴェン）
- 別の選択肢があるのであれば、具体的な例を挙げる必要がある（リプセイ）

BPA に基づくプログラム評価研究：今後の展開に向けて

以上、EBP なかでも RCT の活用を中心に、これまでのエビデンスベースド・アプローチの歴史の変遷を概観してきた。対人・コミュニティ援助全般において、エビデンスの存在がますます重視されるなかで、EBP を推進しつつも、それが様々な理由で（例：倫理面・運営面）かなわない場合に、いかにそれを補完できるか、何か参考になるアプローチはあるか。その可能性のひとつとして考えられるのが、ベストプラクティスを参照枠にした研究、即ち BPA に基づくプログラム評価研究である。

EBP と BPA には共通のゴールが存在する。それが、現場における実践の改善・質向上である。EBP はそのゴールを、社会科学における実験や調査の視点から実現しようとするものである。しかし、これまでみてきたように、RCT といった実験デザインの活用は、特に対人・コミュニティ援助領域においては、不可能ではないにしても現実的ではないとされるケースがほとんどであることも確認できた。そこで、EBP を補完することを目的に、最良・最善の実践（ベストプラクティス）のベンチマーキングを行い、各プログラムの実践の質向上をめざすのが BPA である（図 3）。

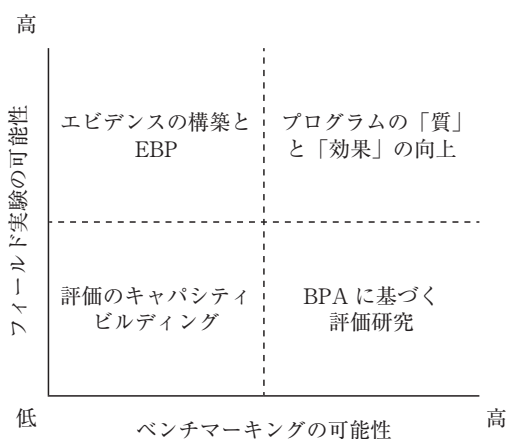


図 3. BPA に基づく評価研究の位置づけ

ここでは、EBP と BPA の共通目標となる『プログラムの「質（クオリティ）」と「効果・成果（アウトカム）」の向上』のために、フィールド実験が可能である場合には EBP を、しかしそれがかなわない場合にはベンチマーキングを行い、BPA に基づく評価研究のルートを模索するという方向性が示されている（参考：安田，2017；Yasuda, 2017）¹⁷⁾。

エビデンスベースドの焦点は、端的には、介入の効果（“アウトカム”）であり、それが妥当か否かを科学的に、つまり因果関係の有無というモノサシで実験・調査（“research”）するところにある。これが先にみたエビデンス構築における方法論の厳格性を重視する視座である。

一方、介入の効果に関する科学的な実証がかなわなかったとしても、補完的にはあるが、介入の質（“クオリティ”）そのものの向上にスポットライトを当てるのが BPA である。言い換えれば、エビデンスベースド・アプローチが介入と効果に関する論理性（logic）を必要条件とし、科学的実証を十分条件とするのに対して、BPA は、少なくとも必要条件となる論理性だけは担保しようとする試みである。そこで、介入活動のプロセスを論理的に明らかにするために工学的（“engineering”）なアプローチが参考になるのである。

ひるがえって「調査」と「評価」の目的の違いは、表 2 のとおりとなる。注目すべきは、調査においては、考察（例：論文でのディスカッション）、結果の一般性の確保（例：統計分析による一般的法則つまり理論の確証）、そして真実（truth）の追究がゴールとなる。

表 2. 調査と評価の違い（Popham, 1993）

	調査	評価
探求の焦点	考察	決定
結果の一般性	高	低
重要性の基準	真実	価値

一方、「評価」では、現場などでの意思決定（例：事業や政策における意思決定）に主眼がおかれる。当然、場面ごと（プログラムごと）の意思決定が必要となってくる。あるプログラムがある現場でうまくいってれば、それが別の現場でうまくいく必要は必ずしもない。そのため、当該実践の分析が主要な関心事となり、結果の一般性の重要度は低くなる。さらに評価の焦点は、実践の価値を「評定する」あるいは「評じ合う」（鵜尾・鴨崎・松本，2015）ところに置かれ、ここでは、当該実践の意義やそれが利用者そして社会にどのような変化をもたらしたのか、が試金石となる。

以上のような評価研究の特徴に着目し、今後、BPA に基づくプログラム評価研究が検討されることが望ましいと言える。

注

- 1) ヒューマンサービス領域では、最新・最善の科学的な研究結果に基づく実践（例：治療・介入）のことをベストプラクティス（BP）と呼ぶことがあるが、本研究で検討するのは、マネジメント研究領域で専門的に発展してきた BP 概念である。前者は、厳密には、エビデンスに基づく実践（EBP）つまりエビデンスベースド・アプローチによる実践である。後者は、本論で検

討するように、最良とされる実践の内容およびプロセスを指標化・ベンチマークし、自身の組織やプログラムの質向上に役立てるというアプローチ、即ちベストプラクティス・ベンチマーキング・アプローチ（BPBA）である。

- 2) 詳細に分析した結果、競合相手の小売価格（キャノンやリコー）が自社（ゼロックス）の製造原価であると分かった（Camp, 1989; 田尻訳, 1995, p. 17）。
- 3) 例えば、ソーシャルワーク実践の領域においてスーパービジョンのシステムの構築が急がれるなか、すでにスーパービジョンの構築が進んでいる心理職領域の実践をベンチマーク調査し、ソーシャルワーク実践のスーパービジョンのシステムを構築していく。あるいは別の例として、高齢者のキャリア支援では、利用者のニーズごとにハローワーク（就職・就労）、シルバー人材センター（生きがい就労）、自治体の生涯学習課・高齢福祉課（生涯学習）と窓口が分散されているが、それを一元化するために、すでにそのようなシステムが存在するワンストップサービス（例：生活保護）をベンチマーキングし、高齢者の支援の質向上を行うなども考えられる。
- 4) そもそも筆者がBPAの可能性を検討し始めたのは、EBPが研究レベル（i.e., エビデンスの構築）あるいは実践レベルにおいて、対人・コミュニティ援助のフィールドでは実現可能であることが極端に少ないことを経験したことによる。そこでEBPを「補完」する（“プランB”の）アプローチとして、BPAを検討するに至った（参考：Yasuda, 2017）。
- 5) 介入群（ X_1 ）と比較群（ X_0 ）による実験デザインの場合は、下記の通りに表せる。

R	○ ₁	× ₁	○ ₂
R	○ ₁	× ₀	○ ₂

- 6) Evidence-based practice is the integration of best research evidence with clinical expertise and patient values (Institute of Medicine, 2001, p. 147).

- 7) Evidence-based practice in psychology (EBPP) is the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences (APA Presidential Task Force on Evidence-Based Practice, 2006, p. 273).
- 8) 信頼区間が狭く信頼度が高いRCTのことである。
- 9) All or none case series. 処方（薬）（Rx）ができる前は特定の病気ですべての患者が亡くなっていたが、現在では一定数の患者が生存しているケース、あるいは、処方（薬）ができる前は一定数の患者が亡くなっていたが、現在では誰もその病気で亡くなることがないケースに関する研究（“Met when all patients died before the Rx became available, but some now survive on it; or when some patients died before the Rx became available, but none now die on it” by Oxford Centre for Evidence-Based Medicine (<https://www.cebm.net/>））。
- 10) 実験的介入を伴わない集団間の比較など。
- 11) 方法論的な詳細については関連図書（Campbell & Stanley, 1963; Cook & Campbell, 1979; Cook, Campbell, & Shadish, 2002）を参照のこと。
- 12) “[t]he United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social programs, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify or discard them on the bases of their apparent effectiveness on the multiple imperfect criteria available” (Campbell, 1969, p. 409).
- 13) 一方、“実験社会”を理想形とする評価研究の展開は、視野狭窄で理想論であるとして、キャ

ンベルの偉大な貢献を称えつつも、評価研究領域の研究者からは受け入れられなかったとしている (Shadish & Luellen, 2013, p. 63; Shadish & Luellen, 2004, p. 83)。

- 14) 第1世代は測定 (measurement)、第2世代は記述 (description)、第3世代は判断 (judgment) とされている (詳細は Guba & Lincoln, 1989)。
- 15) Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality. (中略) RCTs are not always best for determining causality and can be misleading (The AEA statement, 2003)。
- 16) Randomized controlled trials have been essential to understanding what works, what does not work, and what is harmful among interventions in many other areas of public policy including health and medicine, mental health, criminal justice, employment, and welfare (NOT the AEA statement, 2003)。
- 17) そもそもフィールド実験およびベンチマーキングの可能性が低い場合には、評価のキャパシティビルディング (ECB: Evaluation Capacity Building) が行われるが、その焦点や方法論に関しては別稿で検討する。

参考文献

- APA Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology, 61, 271-285.
- Camp, R. C. (1989). Benchmarking: The search for industry best practices that lead to superior performance. Milwaukee, WI: Productivity Press (田尻正滋 訳「ベンチマーキング: 最強の組織を創るプロジェクト」PHP 研究所, 1995)
- Campbell, D. T. (1991). Methods for the experimenting society. Evaluation Practice, 12(3), 223-260.
- Campbell, D. T. (1969). Reforms as experiments. American psychologist, 24, 409-429.
- Campbell, D. & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally.
- Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin Company.
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.
- Cook, T. D., & Reichardt, C. S. (1979). Qualitative and quantitative methods in evaluation. Sage.
- Donaldson, S. I., Christie, C. A., & Mark, M. (Eds.) (2009). What counts as credible evidence in applied research and evaluation practice? Newbury Park, Sage
- Donaldson, S. I., & Christie, C. A. (2005). The 2004 Claremont Debate: Lipsey versus Scriven. Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard? Journal of Multidisciplinary Evaluation, 3, 60-77.
- Guba, E. G., & Lincoln, Y. S. (1989). Fourth generation evaluation. Newbury Park, CA: Sage.
- 原田隆之 (2015) 心理職のためのエビデンス・ベイスト・プラクティス入門: エビデンスを「まなぶ」「つくる」「つかう」金剛出版
- Institute of Medicine (2001). Crossing the quality chasm: A new health system for the 21st century. Washington, DC: National Academy Press
- Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage

- Popham, J. (1993). *Educational Evaluation* (3rd ed.). Boston, MA: Allyn and Bacon.
- Reichardt, C. S., & Rallis, S. F. (1994). The Qualitative-Quantitative Debate: New Perspectives. *New directions for program evaluation*, 61, 1-98.
- Rossi P, Lipsey M & Freeman H (2004) *Evaluation: A Systematic Approach* (7th Ed.). Sage, Thousand Oaks, CA. (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：システマティックな対人 サービス・政策評価の実践ガイド. 日本評論社.)
- Rossi, P. H., Freeman, H. E., & Wright, S. R. (1979). *Evaluation: A systematic approach*. Beverly Hills, CA: Sage.
- Sackett, D., Richardson, W., Rosenberg, W., Haynes, B. (2000). *Evidence-Based Medicine: How to Practice and Teach EBM* (2nd ed.). Edinburgh: Churchill Livingstone.
- Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal* Vol. 312, No. 7023, 71-72
- 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、ランダム化実験の起源、スクリヴェンと バナージェの考え方の比較 日本評価研究, 10(1), 163-173.
- Shadish, W., Cook, T., & Leviton, L. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.
- Shadish, W. & Luellen, J. (2013). Donald Campbell: The accidental evaluator In M. Alkin (Ed.), *Evaluation roots* (2nd ed., pp.61-65). Thousand Oaks, CA: Sage.
- Shadish, W. & Luellen, J. (2004). Donald Campbell: The accidental evaluator In M. Alkin (Ed.), *Evaluation roots* (pp.80-87). Thousand Oaks, CA: Sage.
- Watson, G. H. (1993). *Strategic benchmarking: How to rate your company's performance against the world's best*. Wiley (小林陽太郎 監訳「戦略的ベンチマーキング：組織学習による業績改善の実践手法」ダイヤモンド社, 1995)
- Yasuda, T. (2017). Searching for the Best of Both Worlds: Evidence-Based and Best Practice Approaches for the Delivery of Educational and Social Programs, *Educational Studies*, 59, 117-120.
- 鷗尾雅隆・鴨崎貴泰（監訳）・松本裕（訳）(2015) 社会的インパクトとは何か：社会変革のための投資・評価・事業戦略ガイド [Epstein, M. J., & Yuthas, K. (2014) *Measuring and improving social impacts: A guide for nonprofits, companies, and impact investors*. Berrett-Koehler, Oakland, CA]
- 安田節之 (2017) プログラム評価 臨床心理学（通巻 100 号記念「臨床心理学実践ガイド」）17(4) 554-555
- 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・渡辺直登 (2008) プログラム評価研究の方法 新曜社