

日本評価研究

Japanese Journal of Evaluation Studies

Vol. 25, No. 1, March 2025

巻頭言

日本評価学会の評価は

齊藤 貴浩

ミニ特集：エビデンスに基づく実践の世界の最新動向

特集に寄せて「エビデンスに基づく実践の世界の最新動向」

佐々木 亮 田辺 智子

エビデンスに基づく政策立案（EBPM）における
質的研究の役割

田辺 智子

日本のEBPM（エビデンスに基づく政策形成）はなぜ機能しないのか？
－日米比較とEvidence-Support Systemから考察－

小林 庸平

「EBP的思考」で教師や学校はようになるか

森 俊郎

インパクト評価の発展の可能性：
望ましい社会的価値の取り込み

佐々木 亮

ミニ特集：日本における業績測定の開拓とその意義の 再評価－ハリー・ハトリの業績と関連づけて－

特集に寄せて「日本の業績測定はハトリから何を学べるか」

田中 啓

Harry P. Hatry's Contribution to Performance Measurement:
A Review of 55 Years of His Work

Hiraki P. Tanaka

米国の業績測定の実践から我が国取組が学ぶべきこと
－日本の行政評価の取組を振り返る－

左近 靖博

業績測定に何を求めるか
－EBPMとマンネリズムの間隙を巡って－

小野 達也

研究論文

Examining the Appropriateness of Using Year-Over-Year Accident
Numbers to Evaluate the Effectiveness of Product Safety
Administration

Shunsuke Saito, Hironobu Kawamura

総額型研究開発税制が研究開発費の促進に与える影響

河瀬 豊

実践報告・調査報告

地方自治体におけるEBPMの実態と課題
－福島県市町村調査からの示唆－

加藤 穂高 鈴木 あい 高森 智嗣

第25回全国大会開催報告

『日本評価研究』編集委員会
Editorial Board

編集委員長
Editor-in-chief

米原 あき（東洋大学）
Aki YONEHARA

編集副委員長
Deputy Editor-in-chief

村上 裕一（北海道大学）
Yuichi MURAKAMI

編集委員
Editors

大島 巖（東北福祉大学）
Iwao OSHIMA

窪田 好男（京都府立大学）
Yoshio KUBOTA

佐々木 亮（国際開発センター）
Ryo SASAKI

新藤 健太（日本社会事業大学）
Kenta SHINDO

西野 桂子（関西学院大学）
Keiko NISHINO

橋本 昭彦（日本女子大学）
Akihiko HASHIMOTO

牟田 博光（東京工業大学）
Hiromitsu MUTA

山谷 清志（同志社大学）
Kiyoshi YAMAYA

特定非営利活動法人日本評価学会
編集委員会事務局
E-mail: journal@evaluationjp.org

日本評価研究

第25巻 第1号 2025年3月

目次

巻頭言

齊藤 貴浩

日本評価学会の評価は 1

ミニ特集：エビデンスに基づく実践の世界の最新動向

佐々木 亮 田辺 智子

特集に寄せて「エビデンスに基づく実践の世界の最新動向」 3

田辺 智子

エビデンスに基づく政策立案（EBPM）における質的研究の役割 5

小林 庸平

日本のEBPM（エビデンスに基づく政策形成）はなぜ機能しないのか？

—日米比較とEvidence-Support Systemから考察— 21

森 俊郎

「EBP的思考」で教師や学校はどうなるか 37

佐々木 亮

インパクト評価の発展の可能性：望ましい社会的価値の取り込み 51

ミニ特集：日本における業績測定の開拓とその意義の再評価

—ハリー・ハトリの業績と関連づけて—

田中 啓

特集に寄せて「日本の業績測定はハトリから何を学べるか」 55

Hiraki P. Tanaka

Harry P. Hatry's Contribution to Performance Measurement:

A Review of 55 Years of His Work 57

左近 靖博

米国の業績測定の実践から我が国取組が学ぶべきこと

—日本の行政評価の取組を振り返る— 71

小野 達也

業績測定に何を求めるか

—EBPMとマンネリズムの間隙を巡って— 83

研究論文

Shunsuke Saito, Hironobu Kawamura

Examining the Appropriateness of Using Year-Over-Year Accident Numbers to Evaluation the Effectiveness of Product Safety Administration	99
--	----

河瀬 豊

総額型研究開発税制が研究開発費の促進に与える影響	109
--------------------------------	-----

実践報告・調査報告

加藤 穂高 鈴木 あい 高森 智嗣

地方自治体におけるEBPMの実態と課題 —福島県市町村調査からの示唆—	123
--	-----

第25回全国大会開催報告

開催の報告とお礼	137
プログラム詳細	139
基調講演報告	143
共通論題セッション報告	144
自由論題セッション報告	152

【巻頭言】

日本評価学会の評価は

齊藤 貴浩

日本評価学会副会長／大阪大学教授

これまで様々な活動に対して各種の評価が行われるとともに、効果的な評価を行うために試行錯誤が繰り返されている。今後、更に各種評価を包含するような評価の体系化が行われるならば、個々の評価ニーズに合った評価方法が整理され、評価活動の向上につながる事が期待できる。評価の体系は、それを構成する各種評価手法の研究の進展により、一層充実したものとなる。加えて、評価手法の実践に際しては、情報処理を含め、各種の技術の活用によって、評価の効率が向上することなども期待される。また、評価の方法・手法は、単なる研究に留まることなく実務として実践されることで、その真価が発揮される（後略）

以上は、『日本評価学会設立趣意書』（2000年9月25日）事業内容の項の抜粋です。

日本評価学会は設立から四半世紀を迎えようとしています。それにもかかわらず、設立趣意書に書かれている言葉は今現在の評価の状況をよく表していると言えるでしょう。常に新しい課題や評価手法が出てきているからなのか、日本の評価に進展がないからなのかはわかりませんが、発起人の方々に先見の明があったことについては評価が一致することでしょう。

他方、私たち日本評価学会自体も、コロナ禍による混乱と前後して多くの改革が進められてきました。まずはコロナ禍でもオンラインで全国大会が開かれ続けたこと自体が素晴らしいことと言えるでしょう。学会の組織としては、2021年度より委員会の活動に会員が積極的に参加して会務が遂行される体制が構築され、それに伴う規程も整備されました。学会のMLが整備され、2021年10月には学会報『JES News』の第1号が発刊され、それ以降定期的に発行されています。2022年8月には中期計画が策定され、同年12月の総会で共有され、学会としての方向性が定められました。2023年10月には学会のホームページがリニューアルされ、効果的な情報共有と情報提供が可能になりました。同年12月には会員管理システム(SMOOSY)が導入され、業務の効率化が図られました。同12月の第24回大会では対面での全国大会が復活し、2024年の第25回全国大会では懇親会も開かれました。会期中の総会では、正会員の会費の減額が正式に決定され、またシニアの方が職を離れても会員として活動ができるよう、シニア会費が設定されました。その間、学会誌は毎年2回発刊され、電子ジャーナル化も進み、評価士養成講座の研修事業も大きく発展いたしました。

これらの改革は、これまでの会長、副会長、事務局、委員会、さらには事業ごとのタスクフォースによって進められてきた取り組みの成果であるとともに、学会の趣旨に賛同して様々な活動にご参加くださった会員の皆様のご協力の賜物であり、心より感謝申し上げます。

さて、本稿のタイトルは「日本評価学会の評価は」という曖昧な表現にしてみました。私たちが評価をするときには、政策、施策、事業の目的・目標をもとに、その達成について、あるいは事業の進捗について様々な観点を設定して評価を行います。日本評価学会も、組織として目的を有し、会員や多くの方々の資源をいただいて事業を進めています。日本評価学会としてこれまでの各種体制整備の進捗を確認し、ま

た日頃の学会活動を再認識し、評価してはいかがでしょうか。という意味での「日本評価学会の評価は」…となります。「評価」というと緊張感が高まってしまうかもしれませんが、評価の専門家の集まりにはその心配は無用でしょう。いや、逆に議論になってしまうという意見もあるかもしれませんが、その辺は実務的にざっくりと整理することにして、日本評価学会が何をしてきたかという事実を把握し、次のステップを考えていくことは、学会としての意義ある活動になると思います。

最後に私の専門分野の一つである大学評価の動向について触れたいと思います。大学評価は1991年の大学設置基準の大綱化に伴う自己点検・評価の努力義務化に始まるとされます。その後、認証評価、法人評価、予算配分のための評価などが導入されてきましたが、正直なところ、大学の評価が効果的に機能しているとは言えません。そのような状況で、昨今注目されているのが社会的インパクトです。本学会でも社会的インパクト評価分科会などで概念整理をしようとしているところですが、また新しい意味合いが出てきてしまうかもしれません。導入の意図としては、「大学は社会に役に立っている」ということを示す必要がより強くなったということと考えています。新しいプロジェクトであれば比較的容易であっても、定常的な活動で社会的インパクトを示すことは、理論上は可能であっても現実的にはなかなか難しい評価活動です。本学会の会員には大学に勤務している人が多くいます。日本評価学会が大学評価の在り方についても考える場になり、評価の体系化と評価活動の質的向上につながればよいと考えています。

【特集に寄せて】

エビデンスに基づく実践の世界の最新動向

佐々木 亮

国際開発センター（IDCJ）

田辺 智子

早稲田大学

日本評価学会-社会実験分科会は2001年に設立された。その設立趣意書には次のように記載されていた。なお、この文章が書かれた当時は「ランダム化比較試験」を意味する用語としてRCTという略語の呼び方はまだなく、「実験デザイン」（Experimental Design）という用語が使用されている。

テレビや雑誌の暴力描写は、子供たちの犯罪を誘発しているのだろうか？死刑制度は、犯罪抑止の効果があるのだろうか？選挙前の世論調査の公表は、投票結果に影響を与えているのだろうか？小学校からの英語教育は、大人になったときの英語能力を向上させるのだろうか？これら私たちが普段疑問に思っていることに、実験デザインという評価手法を適用して、政策決定のためのひとつの判断材料を提供することを最終的な目標とします。なお、本分科会では、評価実施に対する誠実さと高い倫理観を保持することを常に奨励しあうことを心がけたいと思います。

この文章が書かれたのは、マサチューセッツ工科大学（MIT）の貧困アクションラボ（Poverty Action Lab.）が2003年に設立される前であるが、同ラボの設立者のアビジット・バナジーとエステル・デュフロらがその後ノーベル経済学賞を受賞した時の受賞理由は、「グローバルな貧困削減のための実験的アプローチに対して」であった。この受賞理由のなかの実験的アプローチとはまさに実験デザイン、つまり別名RCTのことである。

日本でも「エビデンスに基づく政策立案」（EBPM: Evidence-Based Policy Making）が2017年から内閣府を中心に政府を挙げて推進されているが、エビデンスという用語の本来の意味は、RCTをはじめとする厳格なデザインによるインパクト評価の結果のことを指すと理解される。

社会実験分科会は、2001年の設立以来5年おきに学会誌特集号を企画・編纂して、世界的な議論を日本に導入して議論してきた。特集号のたびに学会の全国大会や研究報告会で報告を行っており、現在のEBPMを含む日本の各分野における「エビデンスに基づく実践」（EBP: Evidence-Based Practice）の普及と議論にいくらかでも貢献できていたとすれば望外の喜びである。

さて、社会実験分科会の設立から四半世紀が過ぎ、今回の特集が5回目の特集となる。今回も最新の議論が行われ、次の四半世紀に向けて従来の議論の枠組みを超えた新たな提案がなされているといえる。これらの議論が日本における「エビデンスに基づく実践」のさらなる普及への貢献となることを願っている。

田辺智子会員の論文では、EBPMにおいて質的研究をどのように活用できるかを議論している。海外のエビデンス関連機関として、コ克蘭、キャンベル共同計画、英国国立保健医療研究所（NICE）、EPPIセンターを対象に質的研究の利用状況を調査し、各機関のガイドラインに質的研究の利用方針が明記され、質的研究を含むエビデンスレビューが実施されていることを明らかにした。

小林庸平会員の論文では、EBPMに関する先進的な取組が進むとされる米国と比較し、日本のEBPMが

機能しにくくなっている要素を明らかにしている。日本のEBPMの課題として、エビデンス需要の喚起が弱いこと、エビデンスの需給調整メカニズムが弱いこと、内部における専門人材が不足していることを指摘している。

森俊郎会員の論文では、EBP的思考に基づき教育データの活用に先進的に取り組む学校を調査対象とし、学校組織や教師に生じた変容を明らかにしている。その結果、変容においては、EBP的思考のリフレクション、EBPメンターの存在、EBP的思考の醸成等が重要であることが示唆された。

最後の佐々木亮会員の企画記事では、今後のインパクト評価の展開の方向性として、評価は価値中立ではありえないので、積極的に社会的な価値の判断基準を導入していく方法を考えるべきことを提案している。

最後に、本特集号は、編集委員会のご支援のほか、多数の査読者の尽力によって成立しており、関係者の皆様に深く感謝申し上げます。エビデンスに基づく実践の研究者の層が25年前と比べて分厚くなったことが実感される。日本における今後のエビデンスに基づく実践の研究と実践の普及・深化に期待したい。

参考文献

内閣府（2024）『内閣府におけるEBPMの取組』ウェブサイト（最終アクセス2025年1月18日）

<https://www.cao.go.jp/others/kichou/ebpm/ebpm.html>

行政改革推進本部（1997）『統計改革推進会議最終取りまとめ』（最終アクセス2025年1月18日）

https://www.kantei.go.jp/jp/singi/toukeikaikaku/pdf/saishu_honbun.pdf

日本評価学会（2024）日本評価学会-分科会活動ウェブサイト（最終アクセス2025年1月18日）

<https://evaluationjp.org/activities/subcommittee-meeting-activity#02>

The Abdul Latif Jameel Poverty Action Lab (J-PAL)(2024). About us.（最終アクセス2025年1月18日）

<https://www.povertyactionlab.org/about-us>

The Nobel Prize (2019). Press Release.（最終アクセス2025年1月18日）

<https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/>

【総説】

エビデンスに基づく政策立案（EBPM）における質的研究の役割

田辺 智子

早稲田大学

s.tanabe@waseda.jp

要 約

本稿では、EBPMにおいて質的研究をどのように活用できるかを、主として文献調査を基に明らかにした。関連文献のレビューからは、EBPMにおける質的研究の利用形態は、(1) 量的な有効性評価の改善のための利用、(2) 有効性以外の問いに答えるための利用、(3) 有効性評価に直接的に貢献する利用、の3種類があることが示された。海外のエビデンス関連機関として、コクラン、キャンベル共同計画、英国国立保健医療研究所（National Institute for Health and Care Excellence: NICE）、EPPIセンター（The Evidence for Policy and Practice Information Centre）を対象に質的研究の利用状況を調査したところ、各機関のガイドライン等に質的研究の利用方針が明記され、質的研究を含むエビデンスレビューが実施されていた。質的研究に求められる情報としては、介入が有効となるプロセスの理解、施策の実施に関わる情報、関係者の認識・経験等がある。質的研究の利用の背景には、政策決定に当たっては有効性評価に加え質的研究の情報が必要となり、英国政府などの政策決定機関が質的研究の情報を求めていることがある。

キーワード

エビデンスに基づく政策立案、EBPM、エビデンス、質的研究、系統的レビュー

1. はじめに

エビデンスに基づく政策立案（Evidence-Based Policy Making: EBPM）では、科学的な研究で有効性が示された施策を採用することが主眼とされる。この背景には、公共政策として採用されている施策の有効性が必ずしも評価されていないという現実がある。実社会における政策立案は多様なステイクホルダーが参加する政治プロセスであり、決してエビデンスのみで意思決定ができるわけではないが、そこで用いられる情報の一つとしてエビデンスがあるべき、というのがEBPMの考え方といえる。

EBPMの重要なポイントとして、有効性の評価に用いられる方法には様々なものがあり、それに応じてエビデンスの質に違いがある点がある。これを表現したエビデンス階層（表1）では、ランダム化比較試験（Randomized Controlled Trial: RCT）やその系統的レビュー（Systematic Review: SR）が最上位とされる。

一方で、社会科学の研究方法は、表1に含まれないものも含め多様である。その区分として一般によく使われるのが、量的研究（quantitative research）と質的研究（qualitative research）の2類型である。前者を定量的研究、後者を定性的研究と呼ぶ場合もある。量的研究では、研究対象を数

値化し、記述統計や推測統計を用いて分析を行う。質的研究では、インタビュー調査等で得られた主として言語によるデータを対象に分析を行う。こうした質的研究は、エビデンス階層には明確に表れないか、表れても下位に位置づけられる。

表1 エビデンス階層の例

	研究デザイン
1	体系的レビューとメタ分析
2	信頼性の高いランダム化比較試験 (RCT)
3	信頼性の高い非ランダム化比較試験
4	コホート研究
5	症例対照研究
6	クロスセクション調査
7	症例報告

(出所) Petticrew and Roberts (2003), p.527を筆者が和訳。

しかし、実は海外ではEBPMと質的研究について多くの議論が重ねられてきた経緯があり、近年では質的研究のSRや、質的研究と量的研究の混合SRが増加するなど、活用が進みつつある。では、エビデンス階層では下位にあるはずの質的研究が、EBPMにおいて、どのように活用されているのだろうか。

日本では、EBPMと質的研究の関係については、ほとんど議論されてこなかった。EBPMも視野に入れつつ質的研究の今日的意義を再検討した文献として、井頭 (2023) がある。しかし、同書では質的研究と量的研究とを対比した理論面・方法面での検討が中心となっており、実務としてのEBPMにおける質的研究の具体的な利用については検討されていない。

本稿では、EBPMにおいて質的研究をどのように活用できるのかを、主として文献調査を基に明らかにする¹。まず第2章で質的研究の性格、量的研究との関係を整理した上で、第3章ではEBPMの関連文献で質的研究の利用がどのように想定されているかを整理する。その上で、第4章では、海外のEBPM関連機関における質的研究の具体的な位置づけと様態を明らかにする。

「エビデンス」という用語は、人により異なる含意で使われやすい傾向があり、政策の因果効果に限定し狭義に用いる例も多い。しかし、本稿で参照したコクランやキャンベル共同計画の文献で

は「質的エビデンス統合」(Qualitative Evidence Synthesis: QES) という用語が使われるなど、エビデンスを質的研究も含め広く捉える傾向が見られるため、本稿では「エビデンス」を政策過程で参照される各種情報という意味で幅広く定義する。ただし、可能な限り「エビデンス」という用語を用いず、「質的研究」「量的データ」等、対象を具体的に指し示す用語を用いる。

なお、あらかじめ筆者の立場を明らかにするならば、筆者はEBPMにおける量的分析の必要性を否定する意図はまったく持ち合わせていない。RCTを始めとした定量的な有効性評価はEBPMの核心であり、それ無くしてEBPMの実践はありえないと考える。また、日本では質の高い有効性評価が不足しているが、だからといって質的研究のみでEBPMが実現できるとも考えていない。しかし、EBPMを多面的に理解し、その政策過程での実装を進めていく上で、質的研究の役割を正しく認識することは重要であり、その点に貢献することが本稿の意義と考えている。

2. 質的研究の性格

(1) 質的研究とは何か

質的研究とは、半構造化インタビュー、参与観察、フィールドワーク等の調査法でデータを収集し、研究対象者を深く詳細に記述・分析することで、その経験・認識・価値等を理解しようとする研究方法である。例えば、不登校児の数や分布は量的研究で明らかにできるが、質的研究では不登校となった過程や要因、本人や家族の認識・経験等を明らかにしようとする。多くの場合、既存理論の検証よりは帰納的に新たな概念や理論を構築する性格を持つ。

量的研究が、世界は客観的に明らかにできるという実証主義パラダイム (positivist paradigm) に立つのに対し、質的研究の多くは、世界への認識が主観とは切り離せないと捉える解釈主義パラダイム (interpretative paradigm) に立つ (大谷 2019, p.31)。さらにその内実は多様であり、研究目的や社会・人の捉え方について異なる思想的系譜を持ち、エスノグラフィ、グランデッドセオリー、ラ

イフヒストリー、ナラティブ・アプローチ等の様々な方法を包含している。こうした多元性が質的研究を統一的に定義することを難しくしており、しばしば「量的研究ではないもの」、つまり量的研究の残余として定義されたり、量的研究と対比する形で定義されたりする。

本稿では、質的研究を、質的データを用い、分析においても質的方法を用いた研究と定義する。質的データとは、数値化されていない文字・音声・映像等によるデータであり、主として言語による文字情報が用いられる。データ分析の方法としては、主題分析、エスノグラフィ分析、ナラティブ分析、内容分析等がある（佐藤 2008；メリアム 2004）。

なお、学問分野でいえば、質的研究は社会学、政治学、医学、評価学等のあらゆる分野で用いられてきた。そして、分野によって質的研究の位置づけはそれぞれに異なる。例えば、政治学では、少数の国を対象とした比較歴史分析等で因果推論を行ってきた経緯があり、量的研究でも質的研究でも因果推論を行いうるという立場に立つ（久米 2013）。社会学では、古くは社会現象を集合的に捉える量的分析が優位だった時代があるが、1960年代後半からパラダイム見直しが進み、マイノリティなど研究対象が多様化する中で質的研究が増加していった（町村 2023）。医学において質的研究への関心が高まったのは比較的新しいが、1996年に質的研究の解説書である *Qualitative Research in Health Care* の初版が刊行され、現在では第4版を重ねている（Pope and Mays 2020）。評価学では、特定のステークホルダーのニーズに合致した評価をするために適切な方法が選択される。ただし、評価者は、一方では量的・質的研究を含め適切な方法を選ぶべき、他方では因果効果の問いが最も重要でありRCTが他の手法より優れている、という対立する二つの主張に直面していると指摘されている（Patton 2015, p.93）。

（2）量的研究との関係

質的研究の性格は、量的研究と対比することで理解が容易となる。井頭（2024）は、質的研究と量的研究を区別する観点として、①測定尺度の違い（名義尺度・順序尺度までが質的データ）、②

標本の数（10～20を境界線とする場合が多い）、③統計的検定を行うか、④個々の事例について広く浅い分析をするか濃密な分析をするか、という4点を挙げている。これは後述するKKV論争で採用された区別と同じである。

標本の抽出について補足すると、質的研究では量的研究のような確率抽出法ではなく、調査目的に合致し、最も多くを学ぶうる調査対象が意図的に選択される（メリアム 2004, p.90）。以上から明らかなように、質的研究はそもそも一般化を目的としておらず、調査対象を、その置かれた文脈から切り離さず、ありのままに理解し解釈することが目指される。こうした質的研究の性格は、量的測定になじまないもの、量的測定からこぼれ落ちるものを分析しようとする（大谷 2019, p.23）ものだと指摘される。

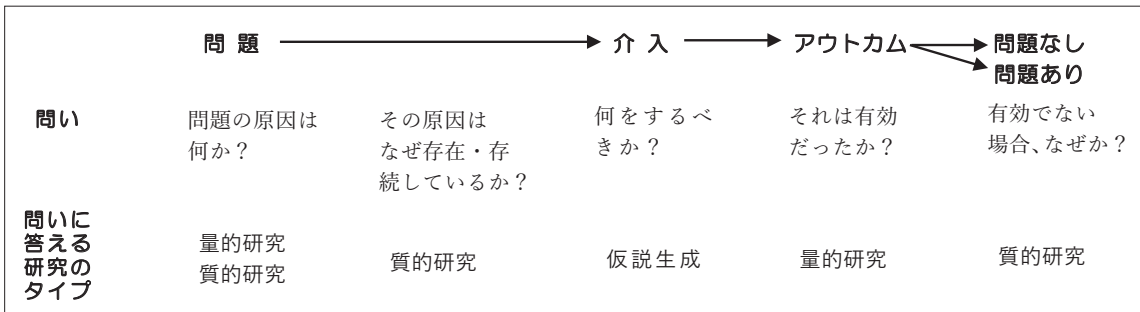
近年では、両者を補完的なものとして組み合わせる混合研究法（mixed methods research）が発展し、2000年代には専門誌も刊行されるようになった（Denzin et al. 2024, p.7）。併用の仕方は各種あるが、例えば、質的研究を基に仮説を構築し、量的研究で検証するといった形がある。ただし、混合研究法では、質的研究が補助的な位置づけにされがちという指摘も見られる（Torrance 2024, p.520）。

（3）量的研究者と質的研究者の論争

質的研究者と量的研究者の間では、そのよって立つパラダイムが異なること、量的研究者から質的研究が「科学的でない」と見られがちであったことから、しばしば激しい対立が生じてきた。Denzin et al.（2023）によれば、1970年代以降に少なくとも3回のパラダイム論争があり、その最後に当たるのが2005年以降現在まで続くエビデンス・ベースドの方法論を巡る論争である。そこでは、RCTがゴールドスタンダードとされ、質的研究は「厳しく抑圧」された（Torrance 2024, p.519）。Patton（2015, p.94）は、研究資金の提供者や世界銀行が何百万ドルという多額の資金をRCTに振り向け、質的研究を含め他の研究デザインが適する場合にもRCTが強制された様子を報告している。

こうした動きに対しては多くの反論が起り、米国や欧州の評価学会等が、よりバランスの取れた研究デザインを推奨する声明を発表した

図1 質的・量的研究の関係



(出所) Gray (2001), p.161 を筆者が和訳。

(American Evaluation Association 2003；European Evaluation Society 2007)。これが次章に見る質的研究の復権の一つの背景となっている。

なお、質的研究と量的研究に関わる重要な対立として、いわゆるKKV論争がある。これは、国際関係・政治学の研究者であるキング、コヘイン、ヴァーバによる古典的テキスト、『社会科学のリサーチ・デザイン：質的研究における科学的推論』（キングほか2004。原典は1994年刊行。筆者3名の頭文字を取ってKKVと略称される）を巡る論争である。そこでは、事例数の少ない質的研究でも一般化や因果推論を行う方法が提案されたが、筆者らの質的研究への本質的な理解が不十分であったこと、回帰分析の方法論を前提に質的研究について論じたことから、多くの批判が生じた（井頭2023、p.3）。この論争については既に多くの論考があるためここでは立ち入らないが、EBPMとの関係で補足するならば、近年のEBPMを巡る議論で焦点となっているのはRCTと質的研究の対比であり、KKVが前提とした回帰分析との対比とは論点が異なるという点である。表1のエビデンス階層においても、回帰分析は上位とはなっていない。

3. EBPMにおける質的研究の役割の類型

では、EBPMの文脈では、質的研究はどのように位置づけられるのだろうか。図1は、エビデンスに基づく医療政策の定評ある解説書であるGray (2001)による図である。政策形成の流れに沿って、

各段階で量的・質的いずれの研究が利用できるかが示されている。同書によれば、質的研究は主に健康医療サービスの働きを理解するために利用され、意思決定者に重要な情報を提供する（Gray 2001, p.163）。そして、質的研究は量的研究を補完したり、組み合わせたり、仮説を生成したりするために利用されると説明している（Gray 2001, p.160）。ここでは、他の各種文献を基に、この図に示された質的研究の利用形態を、以下の3種類に整理する。

(1) 量的な有効性評価の改善

まず、量的な有効性評価と組み合わせで実施し、有効性評価を改善するという利用がある。RCT等による量的な有効性評価は、図1でいえば右から2番目の「それは有効だったか？」というボックスに当たる。有効性評価自体は量的研究で行うが、それを改善するために、その前後のボックスに当たる部分で質的研究を用いるのがこの利用形態である。

Popay (1998, p.34) は、このような利用を「強化 (enhancement) モデル」と呼んでいる。質的研究の役割は量的研究と強くリンクしており、仮説生成、変数の特定、予期せぬ結果の説明等に質的研究が用いられる。

Torrance (2024, p.521) は、RCTは特定の問いに答えることはできるが、問いそのものを立てることはできず、問いを立てるためには事前の質的研究等による探索が必要だと述べている。また、RCTは何かが生じるか・生じないかの説明は

表2 混合研究法において介入前・介入中・介入後に質的データ収集を追加する選択肢

		選択肢
介入研究前	1	適切な手段がない場合に手段を開発する
	2	募集／同意の手続を開発する
	3	介入が機能するように参加者、状況、環境を理解する
	4	介入の必要性を文書化する
	5	包括的なベースライン評価をまとめる
介入研究中	1	参加者の声を代表する質的データを使用して量的アウトカムを検証する
	2	研究実施前に特定した量的アウトカムを超えて、介入が参加者に与えるインパクトを理解する (例: 促進または阻害要因)
	3	介入研究中の予期せぬ患者経験を理解する
	4	介入研究中にアウトカムに影響を与える可能性のある状況を特定する (例: 社会文化的環境の変化)
	5	介入が機能するために重要なリソースを特定する
	6	実験群の経験を理解する
	7	データ収集手順の忠実性を確認する
	8	媒介・調整要因を特定する
介入研究後	1	参加者が介入の結果をどのように見ているかを理解する
	2	介入をどのように修正するかについて参加者からフィードバックを得る
	3	研究アウトカムを説明する
	4	介入研究後に介入の持続可能性を判断する
	5	理論モデルからメカニズムを理解する
	6	介入研究のプロセスが治療の忠実性を持っていたかを判断する
	7	アウトカムとベースラインデータの比較を行う際の状況を評価する
	8	後続の研究のためのパイロットデータを収集する

(出所) Fetters and Molina-Azorin (2020), p.136 を筆者が和訳。

できず、そこにも質的研究が利用できるとしている。実際にそうした併用がされた例としては、米国で1990年代に実施されたニューホープという貧困者向け施策が挙げられる。そこではRCTに加えて詳細なインタビュー調査が継続的に行われ、なぜ給付金が大きな効果をもたらさなかったのが分析された (Grissmer 2016)。なお、Torrance (2024, p.521) は、RCTに関連する研究活動の連鎖として図1に類似したフローチャートを示した上で、RCTはその一部に過ぎないことを指摘している²。

Garbarino and Holland (2009) は、国際開発プロジェクトのインパクト評価での質的研究の利用についてまとめており、①測定面の改善（調査票の質問改善等）、②連続的な利用（質的研究に基づく仮説を量的研究で検証する等）、③提言の改善（プログラムからの脱落理由を質的研究で明らかにし対応策を含める等）、の3種類の併用があると説明している。

混合研究法も、この利用形態の一つに位置づけ

うる。Fetters and Molina-Azorin (2020) は、教育、保健、社会科学の分野で介入研究を実施する際は混合研究法を用いるべきと主張しており、質的データ収集を追加する際の選択肢として表2を示している。彼らは、介入研究の欠点を補い、研究助成の利益を最大化するために、研究の資金提供者は特段の理由がない限り混合研究法に優先的に資金を提供すべきだと指摘している。

(2) 有効性以外の問いへの適用

続いて、有効性以外の問いに答えるために、質的研究を用いるという利用形態がある。(1) との違いは、量的分析と直接的に関連付けない形で質的研究を利用する点にある。例えば、図1左端の「問題の原因は何か？」という段階で、関係者へのインタビュー調査等の探索的な研究を行うことによって、問題の所在や程度を明らかにする場合が該当する。

この利用形態が認識されるようになった背景に

は、EBPMを実践する中で、政策決定者が必要とする情報が有効性にとどまらないことが明らかになったことがある。Head (2020, p.84) によれば、英国ではEBPMが導入されてほどなく、厳格な量的アプローチの推進派が多面的な知識の必要性に気づき、中央政府が信頼性の高い質的研究の重要性を認識するようになった。

ロンドン大学のEPPIセンターは、英国政府と連携してエビデンス統合を研究する機関であるが、その統合チームの研究者達は、政策立案のためのSRは、医療の臨床現場で利用されるSRよりも幅広いものになると指摘している (Sutcliffe et al. 2015)。そこでは、介入の効果量を特定するよりは、ある状況や人々に有効となりうる施策を特定することが重視され、「問題はあるか」、「それに対して何ができるか」、「人々はどのような経験をしているか」、「ある介入をこの状況で実施できるか」といった有効性を越えた問いに対処する必要がある、そのために質的研究を含む様々な研究が必要だと述べられている (Sutcliffe et al. 2015, p.2)。

以上のようなEBPMの実践を通じた気づきに加え、表1で示したようなエビデンス階層への批判からも、質的研究の役割拡大が提唱された。例えば、有効性評価の信頼性は研究デザインだけで決まるのではなく個々の研究の質に依存すること、エビデンス階層は優れた観察研究を過小評価する傾向があること、階層化により多くの重要なエビデンスが除外されること等が指摘された (Nutley, Powell and Davies 2013 ; Petticrew and Roberts 2003)。特にRCTに焦点を当てた批判も多く、RCTが有効性評価のゴールドスタンダードであることは認めつつも、薬剤の服用のように介入と結果に直接的な因果関係がある場合には有効でも、関連する要素が多い複雑な介入には向かないこと、誰にどんな状況で有効かを明らかにできないこと、実験の統制された状況がスケールアップした場合に必ずしも再現できないこと等が指摘された (American Evaluation Association 2003 ; European Evaluation Society 2007 ; Schorr and Farrow 2011)。

こうした批判の一つに、政策過程で必要になる研究上の問いには様々なものがあり、それに応じて適した研究デザインは異なるという指摘があっ

た。しばしば誤解される点であるが、エビデンス階層はもともと政策に関連するあらゆる研究を対象としたものではなく、あくまで有効性評価のための研究デザインを階層化したものである。しかも、人を対象とした介入に限定される。林 (2024) が適切に指摘しているように、気候変動の計算機シミュレーション研究などは、たとえ政策の有効性の要素を含んでいたとしても、この階層に当てはめることはできない。

では、有効性以外の問いには、どのような研究デザインを使うべきか。これについては、表3や表4のようなマトリックスが複数提案されている。表3はエビデンスに基づく医療政策の解説書であるGray (2001) に掲載されたもの、表4は経営学におけるエビデンスに基づく実践を推進するNPOによるものである。いずれも、研究上の問い (リサーチクエスション) のタイプに応じて、適する研究デザインが記載されている。研究デザインには質的研究が含まれ、表3では「サービス提供プロセス」、「重要性」、「安全性」、「適切性」、「サービス満足度」、表4では「態度・意見」、「経験・認識・感情・ニーズ」、「探索・理論構築」に質的研究が適するとされている。こうした位置づけの質的研究は、量的研究の限界を補う役割を果たすと指摘されている (Velti and Miller 2014)。

(3) 有効性評価への貢献

第三に、有効性評価にも質的研究が貢献できるという議論がある。前述のように、政治学では少数事例の質的研究に基づく因果推論が検討されてきた (久米 2013)。以下の議論も、そうした政治学の方法の流れを汲む側面がある。

EBPMに関わる文脈で、質的研究の因果推論への貢献について論じた文献としてGreen (2022) がある。単一事例を集中的に研究する事例研究は、EBPMではエビデンス階層の最下位に位置づけられ因果推論には役立たないと考えられているが、因果メカニズムについてのエビデンスや因果関係の実証を提供しようと論じている。具体的には、①関係するアクター自身による因果関係の説明、②因果関係の実証 (単一の事例であっても一般的に信じられている仮定が成り立たないことを示せる場合がある等)、③因果メカニズムについての

表3 エビデンスのタイプ分けの例（子供への社会的介入の場合）

研究上の問い	質的研究	サーベイ	ケースコントロール研究	コホート研究	RCT	準実験	非実験的評価	系統的レビュー
有効性 これは機能するか？ これを行うことは、あれを行うより効果的か？				+	++	+		+++
サービス提供プロセス どのように機能するか？	++	+					+	+++
重要性 それは重要か？	++	++						+++
安全性 利益が害を上回るか？	+		+	+	++	+	+	+++
受容性 子供／親は提供されるサービスを受け入れるか？		+			+	+	+	+++
費用対効果 サービスは費用に値するか？					++			+++
適切性 子供達にとって適切なサービスか？	++	++						++
サービス満足度 受益者・提供者・その他のステイクホルダーはサービスに満足しているか？	++	++	+	+				+

（出所） Petticrew and Roberts (2003), p.528 を筆者が和訳。

表4 方法の適切性：どの問いにどのデザインか？

目的	例	RCT	非RCT 前後比較	前後比較	コントロール研究	クロスセクション研究	質的研究
有効性・インパクト	AはBに有効か／インパクトがあるか？ Aにとって重要な成功要因は何か？ Bに影響する要因は何か？	A	B	C	C	D	na
予測	AはBに先行するか？ Aは経時的にBを予測するか？	A	A	A	D	D	na
関連性	AはBに関連するか？ AはしばしばBとともに起こるか？ AとBは共変するか？	A	A	A	A	A	na
相違	AとBには違いがあるか？	A	A	A	A	A	na
頻度	Aはどの程度頻繁に起こるか？ Aを好む人はどれだけいるか？	na	na	na	na	A	na
態度・意見	人々のAへの態度はどうか？ 人々はAに満足しているか？ 人々はAに同意するか？	na	na	na	na	A	C
経験・認識・感情・ニーズ	人々のAへの経験はどのようなものか？ 人々のAへの感情は？ 人々のAへの認識は？ 人々はなぜAをしたい／使いたいのか？	na	na	na	na	B	A
探索・理論構築	なぜAは起こるのか？ なぜAはBと違うのか？ Aはどのような状況で起こるか？	na	na	na	na	B	A

（注） na：not applicable（適用できない）。

（筆者注） A、B、C、Dの順に適する。

（出所） Center for Evidence-Based Management. (n.d.) を筆者が和訳。

エビデンス提供、④因果メカニズムが機能する条件の解明、⑤複雑なシステムでの因果関係の説明、の5点について、事例研究が貢献しうると主張している。これらは因果推論そのものとはいえないように見えるが、因果推論の理解を助ける情報だと理解できる。

さらに、有効性評価に直接的に利用できる質的研究方法として、EBPMで適用が研究されているものに質的比較分析 (Qualitative Comparative Analysis: QCA) がある。QCAは、複数の要素で構成される介入について、一つ一つの要素の効果を見るのではなく、どのような要素の組合せが効果をもたらすかを特定する手法である。社会的な施策はしばしば複雑な構成要素から成り、望ましい結果をもたらす経路が複数ある場合がある。例えば、A～Dの四つの構成要素がある場合、A・Bの組合せがあると有効だが、同時にA・C・Dの組合せも有効、といった場合である。QCAはRagin (1980) が体系化した比較的新しい手法であるが、その後急速に普及し、現在では経営学、公衆衛生、環境等の多くの分野で適用が進み、社会科学における標準的な研究方法の一つとなっている (藤田 2023, p.76)。QCAでは、通常の量的分析で必要とされるより少ないデータ数から、複数条件のうち、のぞましい結果をもたらす必要条件や十分条件を特定することができる。

これに加え、複雑な介入の場合、RCT等で介入の構成要素が十分詳細に報告されていない場合があることから、エビデンス統合の際にそれを補う質的分析を行う介入要素分析 (Intervention Component Analysis: ICA) という方法が考案され (Sutcliffe et al. 2015)、QCAとも併用されている。

なお、本章で述べた (1) から (3) の3種類の質的研究の利用形態は、量的な有効性評価との関係性に着目して筆者が整理したものであり、各類型は相互に排他的ではないことに注意されたい。(1) はRCT等の有効性評価と強く関連付けて用いられる利用形態であるが、そこでの問いは (2) で示した表3や表4に包含されうる。(3) で見た事例研究についても同様である。量的研究との関係でいえば、(1) より (2)、(2) より (3) のほうが、量的研究から独立して利用される度合いが相対的に高くなっている。

4. 海外のEBPMにおける質的研究利用

本章では、海外のエビデンス関連機関において、実際にどのように質的研究が利用されているかを確認する。対象としては、SR等のエビデンス統合を実施し、エビデンスと政策現場をつなぐ役割を果たす機関を対象とした。質的研究がエビデンスとして利用される際は、そうした機関での利用が顕在化しやすいと考えられるためである。

対象としては、まずコクランとキャンベル共同計画を取り上げる。コクランは医療の臨床現場のためのSRに加え、保健医療政策のためのSRも実施しており、質的研究の統合をリードしてきた。また社会科学分野のキャンベル共同計画はコクランと連携して活動している。これに加え、NICEとEPPIセンターを取り上げる。後述のように、質的研究の統合では英国の研究者が中心的役割を果たしていると判断できるためである。

(1) コクランとキャンベル共同計画

コクランは保健医療分野のSRを提供する非政府組織である。介入の有効性に関するSRが中心であるが、質的研究の統合や、質的研究と量的研究を統合する混合SRも行われている。コクランのSR手順に関するハンドブック (Higgins and Thomas 2024) では、1章が質的研究に割かれ、詳細な解説が行われている。そこでは、質的エビデンスが意思決定に潜在的に寄与することは実証されており、質的研究を統合することで、介入の複雑さについての理解、実施する文脈による違い、ステークホルダーの意向・経験についてのエビデンスが得られると説明している。質的研究のSRは質的エビデンス統合 (QES) と呼ばれており、2013年に最初のQESとしてGlenton et al. (2013) が発表された。その後、2021年9月までの時点で32件のQES及び統合プロトコルが発表されており、QESへのニーズが高まっているとされる (Glenton et al. 2022)。

コクランにはSRの方法面について検討・助言を行う方法論グループが設置されているが、質的研究については2004年に質的研究法グループ (Cochrane Qualitative Research Methods Group) が設置され、2013年に質的・実施方法グループ

（Cochrane Qualitative and Implementation Methods Group: CQIMG）に改組された。QESの方法論は2010年代に急速に発展し、QESで用いられる統合方法は30以上、QESを量的なSRと統合する方法も10程度あるとされる（Noyes et al. 2017）。

コクランのQES発展の背景として、国際保健機関（World Health Organization: WHO）がQESを求めているという事情がある。WHOは公衆衛生に関する様々なガイドラインを作成しており、そこでは介入の有効性に関するエビデンスが最重要であるものの、具体的な推奨を作る上では質的研究を含む様々なエビデンスが必要とされる（Oxman et al. 2006）。コクランとWHOは提携関係にあるが、最初のQESであるGlenton et al.（2013）はWHOの依頼によるものであり、母子保健へのアクセス改善のための補助的医療従事者（lay health worker）プログラムの実施における促進・阻害要因についてのレビューであった。コクランでQESを最も積極的に実施しているのは効果的实践・ケア組織レビューグループであるが、WHO等の関連機関が当該分野のQESを求めていることが理由の一つとされる（Glenton et al. 2022）。

なお、有効性のSRではRCTが中心であるものの、レビューグループによって含める研究の方針が異なり、健康促進・公衆衛生介入タスクフォースでは、公衆衛生の場合は有効性に関する問いのすべてに単一の方法で答えることはできないとして、RCTから事例研究まで幅広い研究デザインを利用する可能性があるとしている（Jackson and Waters 2005）。

キャンベル共同計画については、コクランほど明確に質的研究の活用が表明されていない。キャンベル共同計画のSRのガイドラインでは質的研究を取り上げた章はなく、SRに含める研究についての説明の中で、質的研究その他の記述的研究は、介入が効果を生む理由等について詳細を理解するのに役立つため、利用可能な場合はそうした研究を要約形式でレビューに組み込むべきと記載している（Campbell Collaboration 2021, p.10）。また、SRに含める研究としては、RCTが基本であるが、RCTや強力な比較試験が難しい介入の場合や、強力なエビデンスが無い場合には、単一事例研究を含む他のデザインによる研究で質が十分高いもの

を含める場合がありえるとしている（Campbell Collaboration 2021, p.10）。

直近の動きとしては、コクランのCQIMGとキャンベル共同計画のQESワーキンググループが共同で、QESのためのハンドブックを作成している（Cochrane Qualitative and Implementation Methods Group 2023）。現在は完成した章が第1.0版としてウェブ公開されており、2025年に刊行予定とされている。全21章からなる大部の資料であり、QESの様々な方法とともにQCAの方法も解説されている。なお、コクランのCQIMGはメンバー10名中8名、キャンベル共同計画のワーキンググループは12名中7名が英国の研究者であり、米国の研究者は含まれていない³。このことから、質的研究の統合については、英国の研究者が主導していることが伺える。

（2）国立保健医療研究所（NICE）

NICEは、英国の公衆衛生ガイダンス、医療経済評価、医療関係者向けの臨床ガイドライン等を提供する機関である。英国政府がEBPMを推進するために各政策分野で設置したエビデンスセンター（What Works Centre）の一つに位置づけられている。ここでは、EBPMと特に関連の深い公衆衛生ガイダンス作成のためのガイドラインを確認する（NICE 2012）。

このガイドラインでは、エビデンスを非常に広く捉えている点に特徴がある。ガイダンス作成に当たっては複数の情報源からのエビデンスが必要であり、それには科学的エビデンスと口語的エビデンス（colloquial evidence）が含まれる。口語的エビデンスは科学的エビデンスの補完等に用いられるもので、専門家の証言、利害関係者のコメント等が含まれる（3.2章）。

科学的エビデンスは、文脈フリーのエビデンスと文脈依存のエビデンスに区分されており、前者は介入の有効性、SR、メタ分析等の客観的測定可能なもの、後者はその他の量的・質的研究、調査、費用対効果分析等の情報だと整理されている。後者の文脈依存のエビデンスは、文脈フリーのエビデンスを補完し、特定の文脈でより具体的で実用的な推奨を作るために利用される。例としては、職業・学歴等による介入の有効性の違い、10代の

親になる予測因子等が挙げられている (3.2.1章)。

ガイドンス作成に当たってはエビデンスレビューを基にするが、その際はRCTを最上位とする標準的なエビデンス階層に頼るのではなく、幅広い研究デザインを利用し、必要な問いに答える必要があると説明されている。その理由としては、RCTは介入の有効性に関する問いに限定されること、RCTは理想的な状況下で実施され推定値が過度に膨らむ傾向があること、公衆衛生では介入のもたらす効果に相互作用がありその複雑性を反映する必要があることが挙げられている (3.3章)。

このガイドンスでは、表3に類似した表で、どのような問いに対してどの研究デザインを用いたレビューを行うかが示されている。対象人口や医療提供者の認識・経験についての問いには質的エビデンスレビューが採用される。なお、NICEと並ぶエビデンスセンターの一つであるEEF (Education Endowment Foundation) においても、内部資料として類似の表を使用している⁴。

(3) EPPIセンター

EPPIセンターは、ロンドン大学に属する研究機関であり、信頼性の高い新たなエビデンス統合方法の開発を使命としている。初期の研究では有効性についてのSRを対象としていたが、政策立案者等のレビュー利用者からは介入の必要性、適切性、受容性、介入実施に影響する要因等の問いも投げかけられたため、質的研究を含む幅広い研究を含むようになった (Thomas and Harden 2008)。EPPIセンターは英国政府やエビデンスセンターと連携して活動しており、コクランでQESの方法論を扱うCQIMGにも所属メンバーが参加している。

ここでは、質的なエビデンス統合の研究例をいくつか取り上げる。まず質的研究と量的研究の双方を統合した混合SRとしては、子供の健康的な栄養摂取の促進・阻害要因についてのレビューがある (Thomas et al. 2004)。このレビューでは、量的な有効性評価のSR、質的研究のSR、さらにそれらを統合したSRという3種類の統合が行われた。子供自身の栄養への認識を分析した質的研究8件のSRからは、①果物と野菜を同様に勧めるべきではない、②健康に関するメッセージは最小限に抑えるべき、という2点の推奨が示された。量的な

SRに含まれた研究で評価対象となった介入が、これら質的研究の推奨と合致するかどうか、また合致する場合の効果量はどうかを分析した結果、①の推奨については合致する介入がなく、研究にギャップがあることが示された。②の推奨については合致するものがあり、サブグループ分析からは、効果量が最も大きい2研究は健康メッセージを強調しておらず、推奨と一致する結果が示された。

QCAを適用したレビューも複数実施されており、最近の例としてSuen et al. (2024) による高齢者居住施設における転倒予防のための運動介入のレビューがある。運動介入は高齢者の転倒予防に有効なことが明らかになっているが、高齢者居住施設での試験では結果にばらつきがあり、何が有効性を左右するかを特定する必要があった。QCAによる分析の結果、中程度または軽度のグループ運動、歩行可能な独立居住者の場合は週1時間を超える運動という2要素によって有効性の有無を説明できることが示された。

5. まとめと考察

本稿では、EBPMにおいて質的研究をどう活用できるかを検討した。第3章の関連文献のレビューからは、EBPMにおける質的研究の利用形態は、(1) 量的な有効性評価の改善のための利用、(2) 有効性以外の問いに答えるための利用、(3) 有効性評価に直接的に貢献する利用、の3種類があることが示された。

第4章で見た海外のエビデンス機関では、各機関のガイドライン等で質的研究を利用していく方針が明記されていた。質的研究の統合についても研究が進んでおり、実際に質的研究のSRであるQESや、量的研究との混合SRが実施されていた。

第3章で見た質的研究の利用形態は、一次情報としての有効性評価に関連する利用であるため、第4章で見た二次情報としてのエビデンス統合と直接的な対応関係があるわけではないが、(2) の有効性以外の問いへの利用、(3) の有効性評価への貢献に対応する事例が観察された。(2) については、NICEのガイドラインで、RCTを最上位とす

る標準的なエビデンス階層に頼らず、幅広い研究デザインを利用して必要な問いに答える方針が明確に示されていた。(3) については、コクランの健康促進・公衆衛生介入タスクフォースやキャンベル共同計画において、RCTによる有効性評価が難しい場合等に、質が十分高い事例研究を使う可能性があるとしていた。また、EPPIセンターでは、既存の有効性評価に一貫性がない場合に、有効性を決定づける要素を特定する目的で質的研究法の一つであるQCAが用いられていた。

こうした質的研究の利用の背景には、英国政府やWHOなど政策に関する意思決定を行う機関が、量的な有効性評価に加え、質的研究を求めているという状況があった。WHOの場合は、公衆衛生ガイダンスを具体化する上では有効性の情報だけでは不十分だという認識があり、英国政府もまた、EBPMを進める中で多角的な知識の必要性に気づいたという事情があった。コクランでは質的研究へのニーズが増大していると報告されており、コクランとキャンベル共同計画が共同で質的研究のハンドブックを作成中であるなど、現在進行形で質的研究の適用が拡大しつつあることが伺える。

質的研究に求められる情報としては、介入が有効となるプロセスの理解、施策の実施に関わる情報、関係者の認識・経験等が挙げられている。例えば、有効性評価の実施前に、政策の対象となる問題の形態や人々の意向・経験について探索的に理解し、仮説や介入を考案するという利用がある。有効性評価を実施し、ある介入の有効性が確認された後も、それを特定の文脈で実施可能か、どのように実施するか、介入が有効となる人々は誰か等の情報が質的研究に期待される。

質的研究の必要性を説明する際に、しばしば使われるのが「文脈（context）」という用語である。介入の有効性そのものは外的妥当性の高い方法で検証され、対象を問わず適用できるとしても、それを特定の地域・人々に対する施策として実施する上では、その地域・人々に固有の文脈について質的研究で補うことが必要だと考えられている。例えば、ワクチンの有効性自体はRCTで明らかにできたとしても、それをある地域で、必要な人々に漏れなく効率的に接種するためには、当該地域の健康・医療を取り巻く状況や、接種を人々が受

け入れるか、受け入れない場合はその理由等についての情報が利用される。

もう一つ、よく使われるのが「複雑性（complexity）」という用語である。「複雑な介入（complex intervention）」という用語は介入研究の世界で一般的に利用されている（Sutcliffe et al. 2015, p.2）。社会的な介入・施策はしばしば複数の要素の組合せで成り立っており、そうした介入の有効性を評価する際にRCTでは限界があることから、それを補うためにQCA等の適用が検討されている。

近年の質的研究への関心の高まりは、EBPMの取組が始まった当初に、RCTの重要性が強調されすぎたことへの反動という側面もある。研究資金がRCTに優先的に振り向けられたことで質的研究が予算減少などの逆風に直面し、質的研究者と量的研究者の論争が激化した。しかし、質的研究を推奨する人々も、有効性評価におけるRCTの有用性を否定しているわけではない。行き過ぎた振り子を戻し、質的研究と量的研究の本来あるべきバランスを模索しているのが、現在の状況だといえるだろう。言い換えれば、EBPMを実社会で推し進めていく上では、単に「何が有効か」を知るだけでは不十分であり、それを質的研究等で補う必要がある、というのが現時点のEBPM実践の到達点といえるのかもしれない。

6. おわりに

最後に、日本への含意を記載したい。EBPMの核心が有効性評価にあるという理解からすれば、日本の現状はEBPMにはまだ遠く、EBPMに向けた準備段階にあるといえる。そうした状況では、本稿で見たような海外の質的研究の利用が直接参考にできるわけではないだろう。まずは有効性評価を蓄積し、それを活用していくべき段階といえる。

しかし、目指すEBPMの最終形においては、有効性評価と合わせて質的研究を適切に利用する必要があるという点は、認識しておいてよいものと考えられる。そしてそれは、単なる「質的情報」ではなく、体系的で質の高い「質的研究」である

必要がある。質的情報といっても、単一のエピソードや特定の専門家の意見のみが政策に反映されるのであればEBPMに逆行している。施策対象者等のステイクホルダーの見解も、科学的で透明性の高い質的研究に落とし込み、そうした研究が無ければ見落とされるような、本当に施策を必要としている人々の小さな声まで特定し、政策過程での情報の一つとして利用するのが、そのあるべき姿となるだろう。なお、紙幅の関係で割愛したが、EBPMで質的研究を利用する上で不可欠となる、研究の質や信頼性の評価方法についても研究が進められている。

今後の研究課題としては、本稿で扱えなかった政府・自治体など、政策の意思決定を担う機関において、質的研究がどのように求められ、利用されているかを明らかにすることが挙げられる。米国など他の諸国での状況も明らかにしていく必要があるだろう。

謝辞

本稿は、早稲田大学特定課題研究助成費（課題番号：2024E-010）による研究成果の一部である。ヒアリング調査にご協力いただいたEPPIセンター、EEFの関係諸氏に御礼申し上げる。

注記

1. 文献調査を補うため、EPPIセンター、EEFへの聞き取り調査を2024年9月に行った。
2. 図1と異なる点として、右端に「より広範な実施・改良」、「大規模な普及・政策実施」が含まれている。介入研究にとどまらず、施策の実施プロセスでも質的研究が利用可能であることを反映したものと考えられる。
3. コクラン (<https://methods.cochrane.org/qi/contact-us>)、キャンベル共同計画 (<https://www.campbellcollaboration.org/methods/other-groups>) ウェブサイトによる。
4. 2024年9月16日の訪問調査による。

参考文献

- 井頭昌彦 (2023) 『質的研究アプローチの再検討：人文・社会科学からEBPsまで』、勁草書房
- 大谷尚 (2019) 『質的研究の考え方』、名古屋大学出版会
- キング, G. ほか (2004) 『社会科学のリサーチ・デザイン：定性的研究における科学的推論』、勁草書房（真淵勝監訳）
- 久米郁男 (2013) 『原因を推論する』、有斐閣
- 佐藤郁哉 (2008) 『質的データ分析法：原理・方法・実践』、新曜社
- 堂免隆浩 (2023) 「政策における質的なエビデンスの活用可能性」、『計画行政』、46 (4)：21-26
- 藤田泰昌 (2023) 「多様で複雑な因果をどう捉えるか：質的比較分析 (QCA) の方法論的発展は何をもたらしたのか」、『経営行動科学』、34 (3)：75-93
- フリック, ウヴェ (2011) 『新版 質的研究入門』、春秋社（小田博志監訳）
- 町村敬志 (2023) 「社会学における「素朴な折衷主義」はなぜ（あえて？）「失敗」してきたのか」、井頭昌彦『質的研究アプローチの再検討：人文・社会科学からEBPsまで』、勁草書房、281-302
- メリアム, S. B. (2004) 『質的調査法入門』、ミネルヴァ書房（堀薫夫ほか訳）
- American Evaluation Association (2003). *American Evaluation Association Response to U. S. Department of Education*. <https://www.eval.org/Policy-Advocacy/Policy-Statements/Response-To-Department-of-Education> (2024年9月28日閲覧)
- Campbell Collaboration (2021). *Campbell systematic reviews: Policies and guidelines*. Ver. 1.8. https://onlinelibrary.wiley.com/pb-assets/assets/18911803/Campbell%20Policies%20and%20Guidelines%20_May3%202022-1653054593497.pdf (2024年9月28日閲覧)
- Center for Evidence-Based Management. (n.d.) *What are the levels of evidence?* <https://cebma.org/resources/frequently-asked-questions/what-are-the-levels-of-evidence> (2024年9月28日閲覧)
- Cochrane Qualitative and Implementation Methods Group (2023). *Cochrane-Campbell Handbook for Qualitative Evidence Synthesis*. Version 1.0. <https://training.cochrane>.

- org/cochrane-campbell-handbook-qualitative-evidence-synthesis (2024年9月28日閲覧)
- Denzin, N. K. et al. (2023). Introduction: The discipline and practice of qualitative research. In Denzin, N. K. et al. (Eds.), *The Sage handbook of qualitative research*. 6th ed. Sage publications.
- European Evaluation Society (2007). *EES Statement: The importance of a methodologically diverse approach to impact evaluation - specifically with respect to development aid and development interventions*. <https://europeanevaluation.org/wp-content/uploads/2020/03/EES-Statement-on-methodological-diversity.pdf> (2024年9月28日閲覧)
- Fetters, M. D., and Molina-Azorin, J. F. (2020). Utilizing a mixed methods approach for conducting interventional evaluations. *Journal of Mixed Methods Research*, 14(2), 131-144.
- Garbarino, S., and Holland, J. (2009). *Quantitative and qualitative methods in impact evaluation and measuring results*. Governance and Social Development Resource Centre.
- Glenton, C. et al. (2013). Barriers and facilitators to the implementation of lay health worker programmes to improve access to maternal and child health: a qualitative evidence synthesis. *Cochrane Database of Systematic Reviews*, (10).
- Glenton, C. et al. (2022). Cochrane effective practice and organisation of care (EPOC) qualitative evidence synthesis, differences from reviews of intervention effectiveness and implication for guidance. *International Journal of Qualitative Methods*, 21, 1-14.
- Green, J. et al. (2022). Case study research and causal inference. *BMC medical research methodology*, 22(1), 307.
- Grissmer, D. W. (2016). *A guide to incorporating multiple methods in randomized controlled trials to assess intervention effects*. 2nd ed. American Psychological Association. <https://www.apa.org/ed/schools/teaching-learning/randomized-control-guide.pdf> (2024年9月28日閲覧)
- Higgins, J. and Thomas, J. (2024). *Cochrane Handbook for Systematic Reviews of Interventions*. Ver. 6.5. <https://training.cochrane.org/handbook/current> (2024年9月28日閲覧)
- Jackson, N., and Waters, E. (2005). Criteria for the systematic review of health promotion and public health interventions. *Health Promotion International*, 20(4), 367-374.
- Gray, J. A. M. (2001). *Evidence-based healthcare: How to make health policy and management decisions*. 2nd ed. Churchill Livingstone.
- National Institute for Health and Care Excellence (2012). *Methods for the development of NICE public health guidance*. 3rd ed. NICE process and methods [PMG4]. <https://www.nice.org.uk/process/pmg4/chapter/introduction> (2024年9月28日閲覧)
- Noyes, J. et al. (2018). Cochrane Qualitative and Implementation Methods Group guidance series—paper 1: introduction. *Journal of Clinical Epidemiology*, 97, 35-38.
- Nutley, S. M., Powell, A. E., and Davies, H. T. O. (2013). *What counts as good evidence? Alliance for Useful Evidence*. https://research-repository.st-andrews.ac.uk/bitstream/handle/10023/3518/What_Counts_as_Good_Evidence_published_version.pdf?sequence=1 (2024年9月28日閲覧)
- Oxman, A. D., Schünemann, H. J., and Fretheim, A. (2006). Improving the use of research evidence in guideline development: 7. Deciding what evidence to include. *Health Research Policy and Systems*, 4, 1-7.
- Patton, M. Q. (2015). *Qualitative research & evaluation methods*. 4th ed. Sage publications.
- Petticrew, M., and Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology and Community Health*, 57(7), 527-529.
- Popay, J., and Williams, G. (1998). Qualitative research and evidence-based healthcare. *Journal of the Royal Society of Medicine*, 91(35_suppl), 32-37.
- Pope, C. and Mays, N. (2020). *Qualitative Research in Health Care*, 4th ed. Wiley-Blackwell.
- Sutcliffe, K. et al. (2015). Intervention Component Analysis (ICA): a pragmatic approach for identifying the critical features of complex interventions. *Systematic Reviews*, 4, 1-13.
- Schorr, L. B. and Farrow, F. (2011). *Expanding The Evidence Universe: Doing Better by Knowing More*. Center for the Study of Social Policy. <https://cssp.org/wp-content/>

uploads/2019/04/Expanding-Evidence-the-Evidence-Universe_Doing-Better-by-Knowing-More_December-2011.pdf (2024年9月28日閲覧)

Suen, J. et al. (2024). Qualitative Comparative Analysis of exercise interventions for fall prevention in residential aged care facilities. *BMC Geriatrics*, 24(1), 728.

Thomas, J., and Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 1-10.

Thomas, J. et al. (2004). Integrating qualitative research with trials in systematic reviews. *BMJ*, 328(7446), 1010-1012.

Torrance, H. (2024). Science, evidence, and the development of policy and practice. In N. K. Denzin et al. (Eds.), *The SAGE handbook of qualitative research*. 6th ed. SAGE Publications, 519-533.

Veltri, G. A., Lim, J., and Miller, R. (2014). More than meets the eye: The contribution of qualitative research to evidence-based policy-making. *Innovation: The European Journal of Social Science Research*, 27(1), 1-4.

(2025.1.14 受理)

The Role of Qualitative Research in Evidence-Based Policy Making (EBPM)

Satoko Tanabe

Waseda University

s.tanabe@waseda.jp

Abstract

This paper clarifies the ways in which qualitative research can be utilized in EBPM, primarily based on a review of the literature. The review of related literature demonstrates that there are three applications of qualitative research in EBPM: (1) to enhance quantitative effectiveness evaluations, (2) to address questions beyond effectiveness, and (3) to directly contribute to the evaluation of effectiveness. A review of the literature on international organizations, including Cochrane, Campbell Collaboration, the National Institute for Health and Care Excellence (NICE), and the EPPI Centre, reveals that each organization's guidelines explicitly outline their policies on the use of qualitative research, and they have conducted evidence reviews incorporating qualitative research. Qualitative research provides insights into the processes by which an intervention works, information related to program implementation, and the perceptions and experiences of stakeholders. The rationale for incorporating qualitative research is that in policymaking, information from qualitative research is necessary in addition to evaluating effectiveness, and policy-making organizations, such as the UK government, are seeking insights from qualitative research.

Keywords

Evidence-Based Policy Making, EBPM, evidence, qualitative research, systematic review

【研究論文】

日本のEBPM（エビデンスに基づく政策形成）は なぜ機能しないのか？ —日米比較とEvidence-Support Systemから考察—

小林 庸平

三菱UFJリサーチ&コンサルティング/経済産業研究所/東京財団政策研究所

y.kobayashi@murc.jp

要 約

日本の中央政府がエビデンスに基づく政策形成（Evidence-Based Policymaking：EBPM）に本格的に着手してから8年余りが経過した。さまざまな取組が進む一方で、ロジックモデルづくりや指標設定・測定が「目的化」「作業化」してしまい、形式的な取組が増えただけで政策改善につながっていないといった指摘もなされている。そこで本稿では、EBPMに関する先進的な取組が進むとされる米国と比較することで、日本のEBPMが機能しにくくなっている要素を明らかにする。その際、Global Commission on Evidenceが提示するEvidence-Support Systemの枠組みを用いることで、日本の特徴を炙り出す。そのうえで、日本のEBPMの課題は、エビデンス需要の喚起が弱いこと、エビデンスの需給調整メカニズムが弱いこと、内部における専門人材が不足していることを指摘する。

キーワード

エビデンスに基づく政策形成（EBPM）、アメリカ、
Evidence-Support System、エビデンス需要の喚起、エビデンスの需給調整メカニズム

1. はじめに

日本の中央政府がエビデンスに基づく政策形成（Evidence-Based Policymaking：EBPM）に本格的に着手したのは2017年頃からとされ（越尾 2020；小林 2020；杉谷 2022等）、8年余りが経過した。さまざまな取組が進んでいる一方で、ロジックモデルづくりや指標設定・測定が「目的化」「作業化」してしまい、形式的な取組が増えただけで政策改善につながっていない、EBPMに取り組む行政側のインセンティブがない、活用可能なデータが不

十分である等、課題も多く指摘されている。

EBPMに関する先進的な取組が進むとされるのが米国である（津田・岡崎 2022；小林 2024b等）。海外事例は日本とは前提条件や仕組みが異なるため、そのまま日本に適用可能なことは多くない。しかしながら、同じような目的のもとでEBPMを推進しており、比較分析を通じて日本の現状や今後の在り方を考えるうえで指針が得られる部分は大きい。そこで本稿では、日米における中央政府のEBPMの取組を概観したうえで、両国の比較分析を行い、類似性と差異を導き出す。そのうえで、

Global Commission on Evidence to Address Societal Challenges (2024) の提示するEvidence-Support System (ESS) をベースに、日本のEBPMを機能させるための条件を導出する。

本稿の構成は以下の通りである。第2節では、日本における近年のEBPM進展のきっかけや主な取組を整理する。第3節では、米国について同様に、近年のEBPM進展のきっかけや概要を整理する。第4節では、日米のEBPMの比較を行う。第5節では、日米のEBPMの違いが生まれる背景を、ESSの枠組みを用いて論じ、第6節では日本のEBPMを機能させるための要素を整理する。第7節はまとめである。

2. 日本のEBPMの現在地

本節では、先行研究¹および政府公表資料を基に日本のEBPMの現在地を整理する。

(1) 2017年以降のEBPM進展のきっかけ

近年、日本でEBPMが進展するきっかけとなったと考えられるのが、2015年10月16日に開催された経済財政諮問会議である²。同会議において消費統計の正確性について論戦が交わされたことを受けて、統計改革が重要な政策テーマとして浮上した。EBPMの推進は、統計ユーザー側のニーズを反映した統計改善に資すると考えられることから、統計改革に関する議論とEBPMの検討がセットで進み、2017年には統計改革推進会議が発足した。当初は、統計の精度向上や統計データに対するアクセシビリティの向上が議論の中心だったが、徐々に政策効果の測定や各省庁での推進体制強化の方策等、EBPMの議論に発展していった。

統計改革推進会議は2017年5月に最終とりまとめを公表した。最終とりまとめでは、GDP統計の精度向上や統計調査の利活用促進等といった統計改革に加えて、EBPM推進統括官（現：政策立案総括審議官）や政府横断的なEBPM推進委員会の設置等、政府全体としての推進体制の青写真が描かれた。

(2) 主な取組

統計改革推進会議最終とりまとめ以降の日本のEBPMに関する主な取組を、脚注1の先行研究およびEBPM推進委員会の資料に記載されている取組を中心に整理したものが表1である。最終とりまとめでは、政策・施策・事業の各レベルにおける推進体制も整理された。政策レベルについては、経済財政諮問会議事務局が、施策レベルは総務省行政評価局が、事業レベルは行政改革推進本部事務局が担うこととされた³。なお表1の2列目には「分類・階層」という項目があるが、これは各取組がどの分類・階層に位置付けられるかを整理したものである。「枠組」と記載されているのは、政府全体のEBPM推進にかかわるものであり、それ以外は、政策・施策・事業の各階層を表す。以下ではレベルごとに主な取組を整理する。

①政策レベル

経済財政諮問会議事務局が推進する政策レベルについては、2020年度に、同会議経済・財政一体改革推進委員会の下にEBPMアドバイザーボードが設置され、2021年度からはエビデンス整備プランが策定され、以降アップデートされている。経済・財政一体改革推進委員会 EBPMアドバイザーボード（2021）は「エビデンス整備プランは、新経済・財政再生計画改革工程表の改善に向けて、①改革工程表の「政策目標」・「KPI」等のつながり（ロジック）の検証、②必要なエビデンスの構築、のための取組・工程等を示すもの。」と定義しており、エビデンスを将来に向かって（Prospectiveに）構築するという注目すべきものである。しかしながら実際のエビデンス整備プランを見ると、KPIの設定やその測定に必要なデータの整理に留まっているものが多く、力点は①の業績マネジメントにある。また、施策や事業レベルとの連動も明確ではない。

②施策レベル

総務省行政評価局が推進する施策レベルについては、「政策効果の把握・分析手法の実証的共同研究」が2018年から行われている⁴。これは関係府省と共同し、外部有識者の参画を得ながら、ロジックモデルの作成から調査設計、データ収集、統計分析まで、政策の効果検証を支援する仕組みであり、EBPMのリーディングケースの創出を目指し

表1 日本のEBPMに関する主な取組

年度	分類・階層	項目	内容
2017	枠組	統計改革推進会議最終とりまとめ	・各省へのEBPM推進統括官（現：政策立案総括審議官）設置や、政府横断的なEBPM推進委員会の設置等、日本のEBPM推進体制の青写真を提示。 ・またGDP統計の精度向上や統計調査の利活用の促進、報告者負担の軽減についても提言。
	枠組	EBPM推進委員会発足	・政府全体としてEBPMを推進するために、内閣官房副長官補を会長、各省の政策立案総括審議官（EBPM統括責任者）等を構成員として発足。
	事業	行政事業レビューにおけるEBPMの試行的検証	・ロジックモデルや統計・データ等を用いた行政事業レビューにおける試行的検証を実施。
2018	枠組	政策立案総括審議官の設置	・各府省においてEBPM推進のための局長級のポストとして政策立案総括審議官が設置。
	事業	実例創出	・行政事業レビューを中心として、EBPMの実例創出開始。
	施策	政策効果の把握・分析手法の実証的共同研究開始	・総務省行政評価局が、関係府省や学識経験者と連携して、政策効果の把握・分析手法の実証的共同研究を開始。以降、毎年実施されている。
2020	政策	EBPMアドバイザーボード設置	・経済財政諮問会議 経済・財政一体改革推進委員会の下にEBPMアドバイザーボードを設置。
2021	政策	エビデンス整備プラン策定	・EBPMアドバイザーボードにおいてエビデンス整備プランが策定され、以降アップデートされている。
2022	事業	EBPM支援制度の創設	・EBPM補佐官制度 ・伴走型支援ネットワーク ・政策設計ラボ
2023	事業	行政事業レビューへのEBPM導入	・政府が実施する約5,000の全事業について、レビューシートにおけるロジックモデルの作成や、指標・目標設定を義務付け。 ・レビューシートの予算編成過程での活用。 ・レビューシートシステムを稼働。

（出所）統計改革推進会議最終とりまとめ、EBPM推進委員会資料、経済・財政一体改革推進委員会 EBPMアドバイザーボードの資料等を基に筆者作成

ている⁵。過去の事例をみると、「競争政策における広報（公正取引委員会）」や「刑務所における受刑者の就労支援希望の申し出促進（法務省）」でクラスターランダム化比較試験を用いた効果検証が行われていたり、「＃7119（救急安心センター事業）の導入効果（総務省消防庁）」や「在外教育施設に派遣された教師に係る派遣効果（文部科学省）」では差の差分分析を用いた効果検証が行われたりしており、因果推論に基づくエビデンス構築がなされていることが分かる。ただし個別の共同研究事例を見ると、必ずしも施策レベルをカ

バーするものではなく、事業レベルのものも多い。また、実証的共同研究の目的から分かる通り、目指しているのはあくまでもリーディングケースの創出であり、施策レベルのEBPMの仕組みづくりまでは射程に入っていない。

③事業レベル

行政改革推進本部事務局が推進する事業レベルについては、表1でも分かるように、行政事業レビューがその中心的役割を果たしてきた。2017年度から、ロジックモデルの作成やデータを用いた指標の設定といった試行的検証が開始された⁶。そ

の後、試行的検証や実例創出が積み重ねられた上で、2023年度からは全ての予算事業（約5,000）のレビューシートにおいて、ロジックモデルの作成や、アウトプットやアウトカムの指標・目標設定が義務付けられている⁷。

こうした取組は、政策レベルや施策レベルの取組とは異なり、すべての省庁の政策立案・運営・評価に影響を与えるものである。その意味で、行政事業レビューへのEBPMの組み込みが、日本における主要な取組だと言える。

ただしこれらの取組についても、エビデンスを参照して政策形成を行ったり、新しいエビデンスを構築したりしているというよりも、業績マネジメントの役割に力点が置かれていると言える。

3. 近年における米国のEBPMの進展

本節では、津田・岡崎（2022）や小林（2024b）を参考にしながら、米国のEBPMの進展を整理する。

（1）近年における米国のEBPM進展のきっかけ

米国のEBPMがどこから始まったかを特定することは難しいが、最近の進展は2012年の政府閉鎖から始まったと言える⁸。

民主党オバマ政権下の2010年3月に医療保険制度改革（いわゆるオバマケア）が成立したが、同年11月の中間選挙において民主党は大幅に議席を減らし、上院は多数派を維持したものの、下院は共和党が多数派になった。それによって予算審議に影響を及ぼすこととなった。民主・共和両党間における2012年の予算交渉にあたったのが、上院民主党の責任者のパティ・マレーと、下院共和党の責任者のポール・ライアンだった。両者は多くの面で対立したものの、政策の運営状況や効果について知ることと、政策立案に役立つようなデータ活用の必要性について合意した⁹。

この合意をベースに、2016年に法律に基づいて創設されたのが「エビデンスに基づく政策形成委員会」（Commission on Evidence-Based Policymaking: CEP）である。CEPは翌2017年に、データインフラやプライバシー、政府内におけるエビ

デンス構築の在り方について報告書（Commission on Evidence-Based Policymaking 2017）を取りまとめた¹⁰。

（2）米国のEBPMの基本的枠組み¹¹

米国においてEBPMの基本的枠組みを規定しているのが、CEPの提言を受けて2019年に成立したエビデンス法（Foundation for Evidence-Based Policymaking Act of 2018）である。エビデンス法は各省に対して、EBPMの責任者である評価官（Evaluation Officer）の任命や、ラーニングアジェンダや年間評価計画とよばれるエビデンス構築計画の策定・公表およびそれに基づくエビデンス構築、エビデンス構築活動のためのキャパシティアセスメントを求めている。

これらのなかで、EBPMの中心的位置を占めているのがエビデンス構築計画である。エビデンス構築計画とは、政策立案上明らかにしたい問い（リサーチクエスション）と、それを解決するための具体的な方法を示した文書のことである。エビデンス構築計画の策定にあたって、はじめに行うのは、政策立案上の問いを明らかにすることである。分析や評価を用いてこうした問いに答えを出していくための活動を文書としてまとめたものがエビデンス構築計画である。ラーニングアジェンダとは、エビデンス構築の中期計画（基本的には4年間）であり、年間評価計画は単年度の短期計画である。

もう一点重要なのが、こうしたエビデンス構築計画上の問いが、目標設定・測定による政府の業績マネジメントを定めたGPRAMA（Government Performance Results Act-Modernization Act）と連動していることである¹²。GPRAMAの下では、各省は優先的に解決すべき政策課題やゴールの設定をする必要があるが、エビデンス構築計画の策定はそれらと連動している。つまり、体系的な業績マネジメントによって、各省は目指すべき目標を設定し、そのために効果的な政策を実施することが要請される¹³とともに、そうした過程で生まれる重要な問いを、エビデンス法上で求められているエビデンス構築計画と紐づけることによって、政策の改善につなげている。

こうしたエビデンス構築活動の責任者が、各省の評価官であり、各省の取組を束ねているのが行

表2 EBPMのきっかけ・青写真・推進体制の日米比較

項目	日本	米国	類似性
きっかけ	統計データの不整合 (2015年10月)	政策がどのように運営され、政府の資金が効果的に使われているかを知る必要がある。 意思決定のために、タイムリーで関連性があり、信頼できるデータの不足。(2012年)	中
委員会設置	統計改革推進会議 (2017年2月)	Commission on Evidence-Based Policymaking (2016年3月～2017年9月)	高
報告書	統計改革推進会議最終とりまとめ (2017年5月)	The Promise of Evidence-Based Policymaking (2017年9月)	高
法制度	なし	エビデンス法 (Foundation for Evidence-Based Policymaking Act of 2018) (2019年1月)	低
責任者	政策立案総括審議官	評価官 (Evaluation Officer)	高
政府横断推進組織	行政改革推進本部事務局	行政管理予算局エビデンスチーム	高
政府横断推進会議体	EBPM推進委員会	Evaluation Officer Council	高

(出所) 筆者作成

政管理予算局 (Office of Management and Budget : OMB) のエビデンスチームである。エビデンスチームは、EBPM推進に関するガイダンスの作成・執行を行ったり、各省の評価官で構成される会議体であるEvaluation Officer Councilを利用してその普及やニーズの共有などを図っている。

4. 日米のEBPM比較

(1) 日米のEBPMの類似性

以上を踏まえて、日米のEBPMのきっかけ・青写真・推進体制を比較したものが表2である。

EBPM推進のきっかけは、日本では統計データの不整合だったが、米国では政策が効果的に実施されているかを把握するための信頼できるデータの不足だった。つまり問題意識に多少の違いはあるものの、データの信頼性や不十分さがEBPM推進のきっかけだった点はある程度共通している。

その後日本では、2017年に統計改革推進会議が最終とりまとめを公表し、政府全体のEBPMの推進体制が描かれた。米国でもほぼ同時期の2016年にCEPが設立され、2017年に最終報告書が出されEBPM推進の青写真が示されている。これらの点は、時期と目的が極めて類似している。ただし米

国ではCEPの報告書を受けて2019年にエビデンス法が制定されているものの、日本ではそうした法的枠組みは存在しておらず、その点は異なっている。

具体的な推進体制についてみると、日本ではEBPM推進の責任者として政策立案総括審議官が各府省に設置され、行政改革推進本部事務局が政府横断の推進組織となり、EBPM推進委員会が政府横断の会議体となっている。米国についても、EBPMの責任者として評価官がエビデンス法に基づいて各省に設置されており、行政管理予算局が政府横断の推進組織であり、各省の評価官によって構成されるEvaluation Officer Councilが推進の会議体になっている。

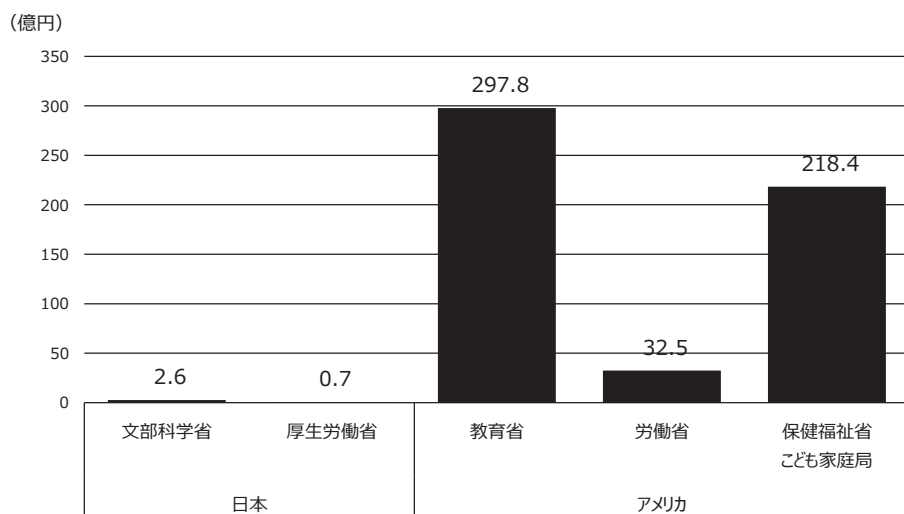
このように日米のEBPMのきっかけ・青写真・推進体制を比較すると、きっかけや法制度の有無等に違いはあるものの、表2の最右列に示されているように両国のEBPMは実はよく似ていると言える。

(2) 日米のEBPMの予算・人員¹⁴

しかしながら、EBPMの予算・人員を具体的にみると、日米の差は小さくない。

図1は、EBPMを適用しやすい教育・労働・福祉関連省庁について日米のEBPM部局の予算額を比

図1 日米の教育・労働・福祉関連省庁のEBPM部局の予算額



(注) 2022年度のデータ。1ドル140円で円換算。

(出所) 内閣官房「令和4年度EBPMに関するリソース状況」(令和4年2月21日)、Results for America (2022) を基に筆者作成。

較したものである¹⁵。日本のEBPM部局が有する予算額は、2022年度で、厚生労働省は約0.7億円、文部科学省は約2.6億円である。それに対して米国のEBPM部局の予算をみると、労働省は約33億円（1ドル140円で換算、以下同様）、保健福祉省子ども家庭局は約220億円、教育省は約300億円となっている。日本の調査研究費は、EBPM部局以外から拠出されているケースも多いため単純比較には留意が必要だが¹⁶、全体として100倍程度の差がついている。

EBPM部局の人員については、日本のEBPM部局の人員の詳細は公表されていないが、各省庁ともに2～5名程度配置されていることが多く、調査分析の専門家や博士号取得者が配置されていることはかなり少ない¹⁷。それに対して米国労働省は18名のスタッフがあり、うち7名が博士号取得者¹⁸、保健福祉省子ども家庭局は77名のスタッフがあり、うち52名が博士号取得者¹⁹、教育省は30名のスタッフがいる（内訳は非公表）²⁰。日本では政策立案総括審議官がEBPM推進の局長級のポストだが、研究・評価のスペシャリストが就任す

ることは少なく、2年程度で異動するケースがほとんどである。米国の評価官は基本的に研究・評価のスペシャリストであり、在任期間も一般的に長い。評価官以下のスタッフについても、博士号取得者が多くを占めていることから分かるように、大半は研究・評価のスペシャリストである。また、米国のEBPM担当者は10～15年程度同じポジションに留まることも多い²¹が、日本の場合は数年で人事異動することが多い。

(3) 日米のEBPMの具体的な取組内容

次に日米両国の具体的な活動についてみてみよう。

①米国

米国ではエビデンス構築活動がEBPMの中心になっていることは前述した通りだが、その際に重視されているのが、政策担当者との協働である。エビデンス構築活動や研究・評価は、政策担当者にとっては抽象的で意義を見出しにくい。そこでEBPM部局では、エビデンス構築計画の策定に際して、こうした活動は政策改善に寄与するもので

あることへの理解を求めたうえで、政策担当者とのコミュニケーションを通して、何が政策立案上重要な問いなのかを明らかにすることに労力を割いている²²。こうした活動を支えるために、前述のような充実した人員がEBPM部局に配置されている。

また、米国労働省のラーニングアジェンダをみると^{23、24}、リモート学習の効果に関する実験研究や、再雇用プログラムの効果検証等、具体的なデータ分析・統計分析が数多く行われていることが分かる。また教育省や保健福祉省でもランダム化比較試験をはじめとした数多くの効果検証や、データ分析が実施されている²⁵。

もうひとつ重要な点は、エビデンス構築活動が、過去の政策の評価ではなく未来に向かった政策改善になっていることである。実施済みの政策の点検・評価だと、政策担当者から拒否反応が示されるケースが多い。しかしながら、政策改善に向けた問いに答えを出す形であれば、政策担当者との協力関係を構築しやすい。また、因果関係に関するエビデンスを明らかにするためには、過去のデータだけでは不十分なことも多く、前向き研究を増やすことの意義も大きい。信頼できるエビデンスを構築するためには、過去の分析をするだけでなく、未来に向かったエビデンス構築が重要となる。

②日本

日本のEBPMにおいて、もっとも労力が割かれて実施されているのが、行政事業レビューにおける取組である。しかしながら前述の通り、行政事業レビューにおけるEBPMは、ロジックモデルの作成や指標の設定・測定といったものが中心となっている。

こうした日本のEBPMに対して批判的な声も聞かれる。政策現場からの声はなかなか表に出てこないが、数少ない例外が法務省（2023）である。法務省（2023）は、EBPMにおけるロジックモデルづくりが「作業化」してしまっていることや、既に導入されている政策評価等と重複感があり、メリットが少なく負担が多いと指摘している。また青柳（2018）は統計整備やロジックモデルの作成が行われてもEBPMが実現する訳ではないと批判している。小林（2020）や関沢（2023）は、日

本のEBPMが本来的な意味から乖離していることを指摘している。

5. 何が日米を分かつのか？

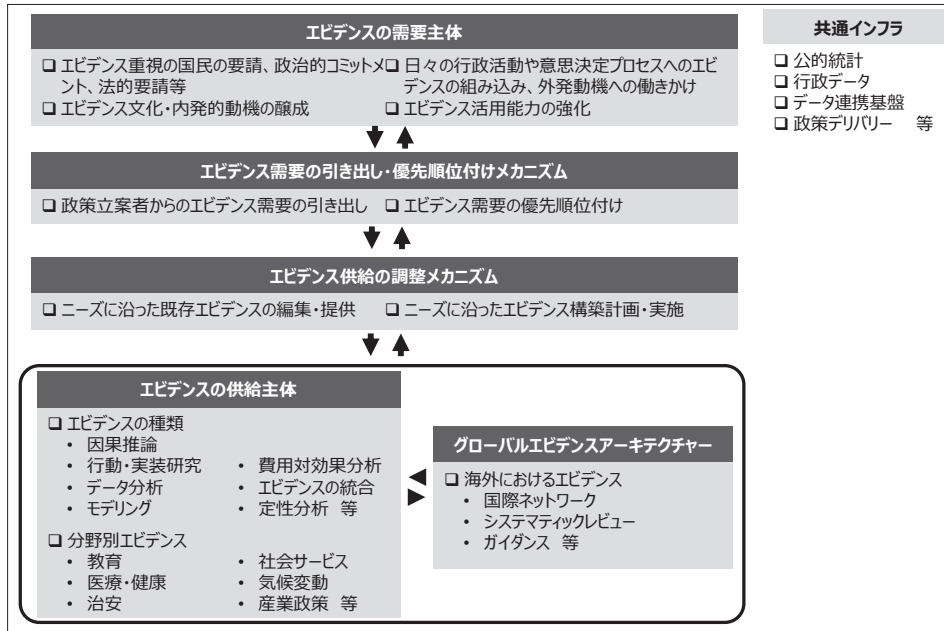
このように、一見すると日米のEBPMのきっかけや推進体制は類似しているものの、予算・人員・具体的な取組を見ると、その差は小さくない。また、どちらのEBPMがより効果的に機能しているかを客観的に測定することは難しいが、日本のEBPMは政策改善に寄与しきれていないと考えられる。そこで本節では、こういった要素が日米のEBPMの差を生み出すのかを整理する。

(1) Evidence-Support System

日米のEBPMを比較するにあたって、Global Commission on Evidence to Address Societal Challenges（2024）が提唱するEvidence-Support System（ESS）の枠組みを用いる。Global Commission on Evidenceは、エビデンスに基づく医療（Evidence-Based Medicine）や、コクラン共同計画やキャンベル共同計画といった国際的な枠組みに依拠しながら、研究エビデンスの利用改善を目的として2021年に発足した取組であり、世界約20か国、約80の政府・研究機関・NGO等が参画している。関係機関と連携し、研究だけでなく実践も進めている組織である。そうした世界各国の関係者への膨大なインタビューを通じて、エビデンスを活用した政策・実践を実現するための条件を取りまとめたレポートがGlobal Commission on Evidence to Address Societal Challenges（2024）である。ESSは世界各国の関係者のコンセンサスが得られている枠組みであり、EBPMをどう機能させるかを検討する上で有用なものだと言える。

図2はESSを和訳したうえで、日本では公的統計や行政データといった共通インフラ部分がEBPMのきっかけとなっているため、その部分を加筆したものである。ESSの最上段にあるのがエビデンスの需要である。EBPMの文脈では、エビデンスの需要主体は政策立案主体とほぼ同義であり、立法府や行政府等が該当する。エビデンスに基づく意思決定が行われるためには、需要主体がエビデ

図2 Evidence-Support System



(出所) Global Commission on Evidence to Address Societal Challenges (2024) に、公的統計や行政データといった共通インフラ部分を加筆して改訂したもの。

ンス活用意欲や能力を持っていることが必要条件になる。

一方で、最下段にあるのがエビデンスの供給である。エビデンスの供給主体としては、例えば大学、シンクタンク、コンサルティングファーム等が該当する。エビデンスに対する需要があったとしても、それを充足する供給がなければ、EBPMは絵に描いた餅となるため、供給主体に厚みがあることも不可欠となる。また、コ克蘭共同計画やキャンベル共同計画といった国際的なネットワークもエビデンス供給主体に成り得る。

そしてEBPMを機能させるために重要となるのが、中段にあるエビデンス需要の引き出し・優先順位付けメカニズムと、エビデンス供給の調整メカニズムである(この2つをあわせて、以下では「需給調整メカニズム」と呼ぶ)。一般に、エビデンスの需要主体である政策立案者は、研究・評価の専門家ではない。そのため政策立案者は、漠然とした問題意識を持っていたとしても、それが研究・評価によってどのように解決し得るのかを理解で

きていないケースが多いと考えられる。リサーチクエスションが明確になっていないと言い換えることもできる。エビデンスの需要引き出し・優先順位付けメカニズムは、こうしたエビデンス需要主体のニーズを引き出すとともに、解決すべき重要なリサーチクエスションが何なのかを優先順位付けする。

引き出されたエビデンス需要に適合するように、エビデンス供給を意思決定に利用可能な形にするのがエビデンス供給の調整メカニズムである。例えば、需要主体から発せられたリサーチクエスションに対して既存のエビデンスが答えを与えていたとしても、それが使いやすい形で提供されていなければ意思決定には活用できない。また、リサーチクエスションに答える既存のエビデンスがない場合、それに答えるためのエビデンス構築計画を立て、供給主体と連携しながらエビデンスを構築する必要がある。

ESSの枠組みから分かるのは、こうした各要素が有機的に機能して初めてEBPMが実現するとい

うことである。

(2) Evidence-Support Systemに基づく日米比較

ESSの枠組みを用いて、日米のEBPMの現状を比較したものが表3の2列目および3列目である。白いセルはあまり進捗が見られない項目、薄い網掛けのセルはある程度進捗が見られる項目、濃い網掛けのセルはかなり進捗している項目である。

米国では、予算編成が直接的にエビデンスに基づいて実施されているわけではないため²⁶、その部分の進捗は中程度だが、それ以外のエビデンスの需要主体の取組、需給調整メカニズム、供給主体の取組はかなり進展していると言える。特に、エビデンス構築計画の策定プロセスは、エビデンスに対する政策担当者の内発的な動機を喚起しながら、政策的に重要なリサーチクエストを引き出し、それを研究計画に落とし込んでいる。また各省のEBPM部局には研究・評価の専門家が数多く配置されており、外部機関と連携してエビデンスの編集・提供・構築に取り組んでいる。その意味で、米国の取組はESSの条件を満たしながら進んでいると共に、とりわけ中段の需給調整メカニズムが充実していると言える。

日本については、EBPM推進の総理指示や骨太方針、行政事業レビューにおけるEBPMへの組み込みを通じて、エビデンスの需要を喚起しようという試みは一定程度進捗している。また大学、シンクタンク、コンサルティングファームといったエビデンスの供給主体も一定程度存在している。その意味で、需要と供給は一定程度存在していると言える。

ただし米国のような体系的な業績マネジメントの仕組みがないため、政策を改善したり効率化したりする動機付けは弱いと言える。図3はEBPM本格化前後における日本の業績マネジメントの導入状況を政策体系（政策・施策・事業）に沿って整理したものである。2002年の政策評価法によって、施策レベルについては目標管理型評価に基づく業績マネジメントが義務付けられるようになった。また、行政事業レビューにEBPMの要素が採り入れられたことで事業レベルにも業績マネジメントが導入され、エビデンス整備プランによって政策レベルでもKPIが設定されるようになった。しか

しながら、政策体系全体の業績マネジメントは相互にあまり連動しておらず、業績マネジメントの結果が政策立案段階で活用されている程度も弱い。つまり業績マネジメントを通じた政策の改善・効率化インセンティブがあまり働いていないと言える²⁷。加えて、政策立案の根拠として、エビデンスが必ずしも要請されているわけではない。

そして日本の大きな課題は、需給調整メカニズムである。総務省行政評価局が実施する実証的共同研究は、政策の効果検証を支援する仕組みであり、エビデンス供給の調整もしくは強化に資する仕組みだが、年間数本程度の実施であり、政策形成全体に与えるインパクトは小さい。また、EBPM補佐官制度や伴走型支援ネットワークといった支援制度も、あくまでも存在しているエビデンス需要への支援策であり、需要の引き出し・優先順位付けメカニズムは弱いと言わざるを得ない。

6. 日本のEBPMに求められる要素

以上の比較を踏まえて、日本のEBPMを機能させるために今後求められる要素を整理したものが表3の右側2列である。

第一に、エビデンスに対する需要の喚起を継続して行うことが欠かせない。米国では、GPRAMAに基づく業績マネジメントによって、目標設定と進捗管理が行われているため、政策を効率化させるインセンティブが働いている。日本でも、予算編成プロセスや法律においてエビデンスの参照や評価計画を求めたり、体系的な業績マネジメントを整備することは重要になる。また米国では、予算・立法措置のなかでエビデンス構築を要請していたり、資金配分メカニズムのなかにエビデンスの要素が採り入れられており²⁸、そうした形でのエビデンス需要の喚起も重要だろう。

第二は、エビデンスの需給調整メカニズム、とりわけ需要の引き出し・優先順位付けメカニズムを整備することが重要になる。政策立案者の困りごとや課題を引き出し、エビデンス構築計画等に落とし込むとともに、それらを外部のエビデンス供給機関とつなぎ、調査・分析を未来に向けた政

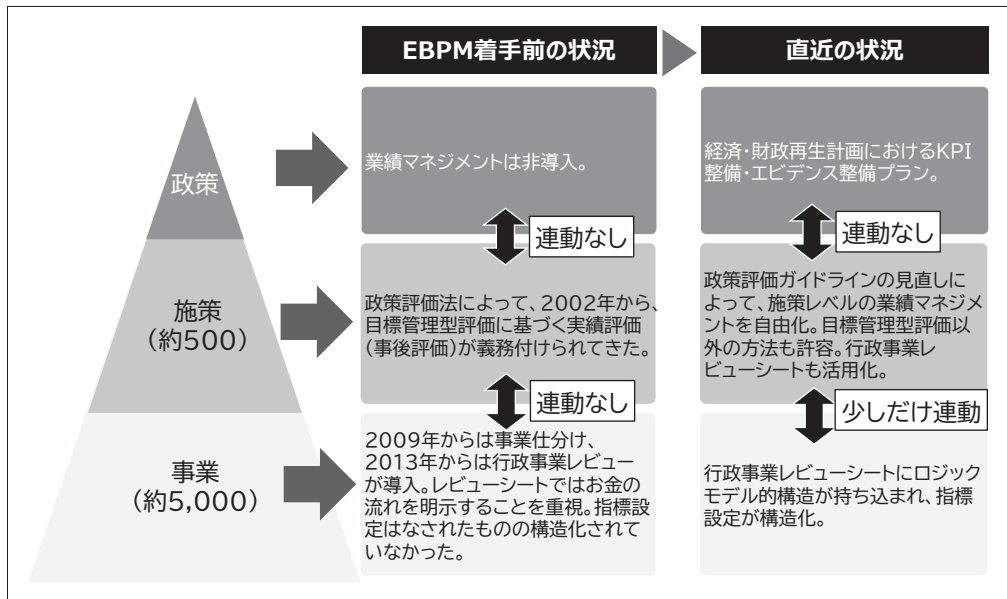
表3 Evidence-Support Systemに基づく日米のEBPM比較および日本の今後の検討事項

項目		米国	日本		
		現状	現状	今後求められる要素	
				個別	共通
エビデンスの需要主体	国民的・政治的合意、コミットメント	・超党派合意 ・大統領令 ・Evidence Act等の法律上の要請	・総理指示 ・骨太方針	・国民・政治家からのEBPMに対する継続的な要請 ・予算・立法におけるEBPMの要請	・政府内部における専門家登用・キャリアパスの確立 ・官民の流動的な労働市場の構築 ・公的統計や行政記録情報の活用 インフラ整備
	日々の行政活動・意思決定プロセスへの組み込み、外発的動機への働きかけ	・業績マネジメントとEBPMの連動 ・予算との緩やかな連動 ・ホワイトハウスからのガイダンス等の発出 ・補助金執行等を通じたインセンティブ付け	・エビデンス整備プランの策定 ・行政事業レビューへのEBPMの組み込み	・予算編成等の政策立案プロセスにおいて、エビデンス活用や構築を促進 ・政策・施策・事業をカバーする体系的な業績マネジメントの導入	
	エビデンス文化・内発的動機の醸成	・EBPM部局による内発的動機の醸成	・研修実施 ・ガイドブック等の作成	・継続的な研修実施 ・政府内のネットワーク形成 ・内部専門家による文化・内発的動機の醸成	
	エビデンス活用能力強化	・専門職公務員		・政策立案プロセスにおけるエビデンス活用研修・ガイダンス	
エビデンス需要の引き出し・優先順位付けメカニズム		・EBPM部局の内部専門家によるエビデンス構築計画策定		・EBPM部局や内部専門家によるエビデンス需要の引き出し、エビデンス構築計画策定	
エビデンス供給の調整メカニズム		・EBPM部局の内部専門家による既存エビデンスの編集・提供 ・EBPM部局からの外部機関へのエビデンス供給依頼		・EBPM部局や内部専門家による既存エビデンスの編集・提供 ・外部機関と連携したエビデンス供給依頼	
エビデンスの供給主体		・大学、シンクタンク、コンサルティングファーム等が充実	・大学、シンクタンク、コンサルティングファーム等が一定程度存在	・公共政策大学院等による人材供給強化 ・AI等を活用した迅速なエビデンス統合 ・エビデンスデータベースの構築	
グローバルエビデンスアーキテクチャー		・Global Commission on Evidenceへの参加		・Global Commission on Evidence等との連携強化	

(注) 日米の「現状」で、白いセルは進捗の遅いもの、薄い網掛けのセルは進捗がある程度あるもの、濃い網掛けのセルはかなり進捗しているもの。

(出所) 筆者作成

図3 日本におけるEBPM本格化前後の業績マネジメント導入状況



策改善の営みに活かす機能が必要となる。また業績マネジメントの過程で生まれる問いに基づいてエビデンス構築活動が行われており、業績マネジメントがエビデンス需要を引き出す下準備になっている。

第三は、行政内部での専門家の登用・キャリアパスの確立である。予算編成プロセスや法律においてエビデンスの参照や評価計画の策定を求めたとしても、そうした業務全体を外部機関に委ねることは難しい。また、エビデンス需要の引き出しや優先順位付けは、非常に漠然としたフェーズからスタートすることが多いため、これも外部機関に委ねることは難しい。米国では、こうした業務は、研究・評価の専門家がジョブ型の雇用の下で政府内で実施しており、専門性を行政内部に組み込まない限り日本でも実施は難しいだろう。またデータ分析や効果検証の実施には一定の時間を要することも少なくないが、日本のように公務員の異動のサイクルが短いこともEBPMの足枷になる。

この点の詳細な検討は公務員制度に議論が及ぶため本稿の射程を超えるが、以下のような選択肢

があるだろう。第一が、公務員制度のなかに専門人材の採用・キャリアパスを構築することである。現在でも、日本には技官の国家公務員がいるものの、ジェネラリスト公務員が引き続き政策立案における中心的役割を担っている。技官制度を再定義して、ジョブ型・スペシャリスト型人材の採用・キャリア形成を構築することが考えられる。第二が、任期付き任用もしくはインハウスコンサルタント等の拡大である。現在でも、専門的な知識を持つ人材を雇用する仕組みとして任期付き任用がある。また、政府組織のなかでも、コンサルタントを内部に常駐させているケースがある。こうした仕組みを拡大して、EBPM機能を担わせることも考えられる。第三が、専門組織の設立である。かつて経済企画庁には多くの官庁エコノミストが所属し、経済計画の策定や経済白書の執筆といった専門業務を担っていた。また英国政府において行動科学を推進するために設立されたBehavioural Insights Teamには、行動科学や政策効果検証の専門家が数多く所属し、各府省に対してサービスを提供する機関として設置された。つまり、日本の

メンバーシップ型公務員制度を前提にしたとしても、旧経済企画庁のように、スタッフの全員が同じ専門性を有しているのであれば、メンバーシップ型雇用のなかでも専門性を取り入れることが可能となる²⁹。

7. おわりに

本稿では、日米のEBPMの進展を整理したうえで、日本のEBPMが形式的な取組が多くなっている現状を指摘した。そのうえで、Evidence-Support System (ESS) の枠組みを用いて、日米のEBPMのどこに差があるのか、日本のEBPMを機能させるためには何が必要なのかを論じた。

日本のEBPMの課題および今後求められる要素の第一は、需要の喚起である。骨太方針や行政事業レビューにおけるEBPMへの導入を通じて、エビデンス需要の喚起が一定程度進んでいるものの、米国のように法律や予算編成を通じた需要の喚起はほとんど行われていない。業績マネジメントの仕組みが不十分なことも、政策の効率化インセンティブを弱めている。予算編成プロセスや法律においてエビデンスの参照や評価計画を求めたり、体系的な業績マネジメントを整備することは重要になる。

第二は、需給調整メカニズムである。政策立案サイドに潜在的なエビデンス需要があったとしても、それを引き出し、供給と適合させるメカニズムがなければ、エビデンスを活用した意思決定は難しい。

第三は、行政内部における専門家の活用である。エビデンスの供給と比較して、需要の喚起や需給調整は外部人材に委ねることが簡単ではない。それらを機能させるために、専門人材を組織内部に融合していくことが不可欠となる。

謝辞と付記

本稿の中間報告を、2024年6月29日に開催された日本評価学会 社会実験分科会で行い、津富宏氏および青柳恵太郎氏から貴重なご指摘を頂いた。また査読プロセスにおいて、匿名の査読者から幅

広い観点のご指摘を頂いた。記して感謝申し上げたい。本稿における全ての責は筆者に帰するものであり、筆者の所属する機関の意見を代表するものではない。

注記

- 1 越尾 (2020)、小林 (2020)、杉谷 (2022) 等。
- 2 小林 (2020)。
- 3 統計改革推進会議 (2017) p.6 ~ 7。
- 4 総務省行政評価局 (2019)。
- 5 総務省 政策評価ポータルサイト https://www.soumu.go.jp/main_sosiki/hyouka/seisaku_n/seisaku_ebpm.html#/。(2024年11月26日アクセス)
- 6 内閣官房行政改革推進本部事務局 (2017)。
- 7 内閣官房行政改革推進本部事務局 (2023)。
- 8 Hart and Yohannes (2019)。
- 9 Hart and Yohannes (2019)。
- 10 津田・岡崎 (2022)、小林 (2024b)。
- 11 本小節は小林 (2024b) の3.2節を要約しつつ現地インタビューを踏まえて加筆したものである。
- 12 2024年2月に実施した、教育省教育科学研究所 Matthew Soldner氏のインタビューより。多くの省は業績マネジメントとエビデンス構築計画を連動させているが、やり方は各省に柔軟性がある。
- 13 業績マネジメントによって、各省は政策改善が動機づけられている。ブッシュ (子) 政権時代の政治任用職員や、2024年時点における各府省のEBPM部局の職員など、匿名を条件に2024年に実施した複数の現地インタビューより確認。
- 14 本小節は小林 (2024a) を基に加筆修正したものである。
- 15 大竹他 (2022) は潜在的な政策対象者が多く、実施プロセスへの関与可能性が高い政策ほど、EBPMの適用可能性が高いとしており、教育、児童福祉・子育て支援、医療・介護・健康、就労支援などを具体的に指摘している。また、Results for America (2022) はEBPMが進んでいる連邦政府省庁を格付けしているが、教育、労働、保健福祉、社会保障関連の省庁が多く含まれており、米国においてもこれらの政策領域はEBPMの先進分野だと言える。
- 16 ただし米国についても、EBPM部局以外から調査研究費が支出されているケースは多い。例えば労働省

- の場合、図1に示されているのはChief Evaluation Officeの予算額のみだが、労働省には傘下に多数の庁（Sub-agencies）があり、各庁にもEBPM部局があり独自の予算が確保されている。
- 17 筆者の実務経験のなかでの実感値。
- 18 労働省ウェブサイト。
- 19 保健福祉省子ども家庭局ウェブサイト。
- 20 教育省Institute of Education Sciencesウェブサイト <https://ies.ed.gov/>（2024年11月26日アクセス）。
- 21 2023～2024年にかけて、連邦政府職員に対して匿名を条件に実施した複数の現地インタビューより確認。
- 22 2023～2024年にかけて、連邦政府職員に対して匿名を条件に実施した複数の現地インタビューより確認。
- 23 Department of Labor “Evidence-Building Plan Fiscal Years 2022-2026”（労働省のラーニングアジェンダ）
- 24 労働省の取組はエビデンス法作成の際に参考にされており、その意味で連邦政府内でも代表的なリーディングケースだと言える。
- 25 2023～2024年にかけて、連邦政府職員に対して匿名を条件に実施した複数の現地インタビューより確認。
- 26 予算編成をエビデンスに基づいて実施しようとしている国としては英国がある。詳細は小林他（2024）参照。
- 27 ただしHatry（1967）が指摘するように、政策の有効性を主張するためにパフォーマンス情報や業績マネジメントの情報を活用することは簡単ではないため、副作用には留意が必要となる。
- 28 小林（2024b）。
- 29 この点は、小峰隆夫氏との議論のなかで示唆を頂いた。記して感謝申し上げたい。
- 参考文献**
- 青柳恵太郎（2018）「経済学および国際開発の視点から」、第2回エビデンスに基づく実践と政策セミナー 報告スライド
- 大竹文雄・内山融・小林庸平（2022）「EBPMとは何か」、大竹文雄・内山融・小林庸平編著『EBPM エビデンスに基づく政策形成の導入と実践』、日本経済新聞出版
- 経済・財政一体改革 EBPMアドバイザーボード（2021）「経済・財政一体改革エビデンス整備プラン（令和3年8月30日）」
- 越尾淳（2020）「政府におけるEBPMはどのような一歩を踏み出しているか」、大橋弘編『EBPMの経済学』、東京大学出版会、43-59
- 小林庸平（2020）「日本におけるエビデンスに基づく政策形成（EBPM）の現状と課題—Evidence-Basedが先行する分野から何を学び何を乗り越える必要があるのか—」、『日本評価研究』、20（2）:33-48
- 小林庸平（2024a）「米国に学ぶEBPM 未来志向で政策改善を支援」、日本経済新聞経済教室、2024年3月21日
- 小林庸平（2024b）「アメリカ連邦政府におけるエビデンスに基づく政策形成の実際」、独立行政法人経済産業研究所 Policy Discussion Paper 24-P-009
- 小林庸平・内山融・川瀬仁志・尾谷祐樹（2024）「英国におけるEBPMの深化—政策評価タスクフォースを中心として—」、独立行政法人経済産業研究所 Policy Discussion Paper 24-P-008
- 杉谷和哉（2022）『政策にエビデンスは必要なのか』、ミネルヴァ書房
- 関沢洋一（2023）「EBPM登場の経緯と和風EBPM」、独立行政法人経済産業研究所 EBPMコラム
- 総務省行政評価局（2019）「EBPM推進に係る行政評価局の取組状況—実証的共同研究の実施—」、第4回EBPM推進委員会（2019年9月9日開催）提出資料
- 津田広和・岡崎康平（2022）「米国におけるEBPM」、大竹文雄・内山融・小林庸平編著『EBPM エビデンスに基づく政策形成の導入と実践』、日本経済新聞出版、111-144
- 統計改革推進会議（2017）「最終取りまとめ」
- 内閣官房行政改革推進本部事務局（2017）「EBPM推進のための「行政事業レビュー」の取組について」、第1回EBPM推進委員会（2017年8月1日開催）提出資料
- 内閣官房 行政改革推進本部事務局（2023）「EBPM推進委員会説明資料」、第1回EBPM推進委員会（2023年4月13日開催）提出資料
- 法務省（2023）「法務省におけるEBPM 推進に向けた取組」、こども家庭庁 第2回EBPM 研究会
- Commission on Evidence-Based Policymaking（2017）.
Promise of Evidence-Based Policymaking. Commission

on Evidence-Based Policymaking.

Global Commission on Evidence to Address Societal Challenges (2024). *Global Evidence Commission update 2024: Building momentum in strengthening domestic evidence-support systems, enhancing the global evidence architecture, and putting evidence at the centre of everyday life*. Hamilton: McMaster Health Forum.

Hatry, H. (1967). "Criteria for Evaluation in Planning State and Local Programs" Senate Subcommittee on Intergovernmental Relations of the Committee on Government Operations.

Hart, N. & Yohannes, M. (2019). *Evidence works: Cases where evidence meaningfully informed policy*. Bipartisan Policy Center.

Results for America (2022). *2022 Invest in What Works Federal Standard of Excellence*.

(2024.12.21 受理)

Why Evidence-Based Policymaking Falls Short in Japan: A Comparative Study with the U.S. and Insights from the Evidence-Support System

Yohei Kobayashi

Mitsubishi UFJ Research & Consulting
y.kobayashi@murc.jp

Abstract

Over eight years have passed since Japan's central government embarked on evidence-based policymaking (EBPM). Despite progress, there are concerns that efforts like logic model creation and indicator measurement have become formalities, failing to improve policy outcomes. This paper compares Japan's EBPM with the more advanced approach in the United States to identify obstacles hindering its effectiveness in Japan. Using the Evidence-Support System framework from the Global Commission on Evidence, the paper highlights Japan's challenges, including weak demand for evidence, inadequate mechanisms for balancing the supply and demand of evidence, and a shortage of internal experts.

Keywords

Evidence-Based Policy Making (EBPM), United States, Evidence-Support System, Stimulating-Evidence Demand, Mechanisms for Balancing the Supply and Demand of Evidence

【研究ノート】

「EBP的思考」で教師や学校はどうか

森 俊郎

名古屋大学大学院博士課程・岐阜県養老町立東部中学校

m072132@gmail.com

要 約

近年、学校現場において、教育データの利活用が推進されている。これまで、教育データから導き出されたエビデンスを実践の改善につなげようとするEvidence Based Practiceの在り方について、その功罪について議論されてきた。しかし、様々な諸要因が絡み合う学校現場において、実践の示唆が乏しい。そこで、本稿では、「EBP的思考」を展開する学校を調査対象とし、何がきっかけで、何が起こり、学校組織や教師個人がどのように変容したのか事例的解明に取り組んだ。複数職員へのインタビュー調査の結果、「探求心や受動的きっかけ→EBP的思考の共有→チーム形成→実践（教育データ会議・授業実践）・EBPメンター→リフレクション→EBP的思考の定着」といった一連の変容が確認された。本研究を通じて、EBP的思考のリフレクション、EBPメンターの存在、EBP的思考の醸成の意義が示唆された。

キーワード

教育データ、Evidence Based Practice、EBPメンター、教師

1. 本研究の背景

近年、教育分野では、教育データの利活用が目指されている（文部科学省 2024）。全国学力・学習状況調査やタブレット端末での学習の実施によって得られた学力データ、校務支援システムによって得られた児童生徒の保健、学籍に関する生活データ等、近年のデジタル革新によって、各種教育データが容易に収集しやすくなったためである。そして、教育データから導き出された質の高い知見、所謂、エビデンス（科学的根拠）（津富 2008）を活用することによって、いじめ問題等、学校教育の諸課題の解決につなげようとするEvidence Based Practice（EBP）の動きがある（佐藤 2016：9）。このEBPとは、患者の特徴、文化、

嗜好等の文脈において、入手可能な最良の研究結果を臨床技能に統合することである（APA2006）。EBPに関する先行研究として、教師を取り巻く社会的な背景（松下 2015）や、エビデンスに基づく教育についての可能性や限界について論じたもの（杉田・熊井 2019）がある。これらは、EBPによって、「今まで気付かなかった点に気付く」良さや、エビデンスに縛られて「実践的思考の自由を束縛される」懸念といった、「エビデンスと教師」の在り方について、その功罪について論じている（内田 2015）。このような指摘は、教育分野におけるEBPの在り方を理論的に考察する際の重要な示唆となる一方、実際の学校現場においては、理論的に示されるような良さや懸念がそれぞれ独立して存在しているわけではないため、実証性に乏しい

(鈴木ら 2018)。また、EBPに関する様々な議論は、「エビデンス・ベースド」という言葉から誤解や誤用を招きやすい(原田 2015: 48)。そのため、EBPによって、教師や学校がどうなるのかといった説明が進んでおらず、実践的示唆が乏しい。

確かに、EBPの議論において、エビデンスは、狭義・広義、質の差、多義性という視点で論じられてきた(原田 2015: 37; 米原 2024)。これらを概観すると、単なるデータはエビデンスとはいえない。しかし、学校現場においては、これらの違いや、その影響の違いを明らかにすることや、一つ一つの教育データの詳細を取り扱うことは、匿名性の確保のため、明記・明文化が非常に難しい。そこで、本研究では、学力データや生活データ等の総称を「教育データ」、教育データから導き出された知見や先行研究等、より質の高い情報を実践に統合しようとする思考過程を「EBP的思考」として、操作的定義を行った。

この上で、本研究では、教師のEBP的思考による教師や学校の変容に着目する。このことで、実証性が乏しく、誤解・誤用の多いEBP議論を乗り越え、日本の教育分野におけるEBP事例の現象学的検証を試みる(杉谷 2022: 231; 森 2019: 130)。

具体的には、先進的な学校を対象にして、その学校のEBPについて、【何がきっかけで、何が起こり、どうなったのか】を時系列で追う。そして、各教諭がどのような認識変容を辿ったのか、さらにその要因は何かといった実態解明を試みる。これらは、教師個人、学校組織におけるEBPの実装可能性を高めるだけでなく、先行研究で指摘されている「エビデンスと教師」の在り方やその功罪について、より実践的示唆を得ることができると考える。

2. 目的と方法

(1) 目的

本研究のねらいは、EBP的思考による学校の変容を事例的に検証することである。具体的には、EBP的思考に取り組んだ先進校において、【何がきっかけ】で、【何が起きて】、【どうなったのか】をインタビュー調査から明らかにする。

(2) 調査対象と調査方法

インタビュー調査の対象となったのは、X県Y市にある公立Z学校である。Z学校は、教育行政から特別指定を受け、学校全体として、積極的に教育データの活用に先進的に取り組んできた。例えば、Z学校では、学力向上のための教育データ会議を月に1回行ったり、各種アンケート調査の結果を用いたりして、児童への学習支援の在り方について、検討を重ねてきた。教育行政及び学校から、本研究への協力及び参加の同意を得ることができたこと、多くの実践資料を有していたこと、年間を通して組織的实践に取り組んできたこと等により、本研究の調査対象とした。インタビュー調査対象となった教諭は、Z学校の在籍歴の異なるA教諭(男性:30代:教職14年目)、B教諭(女性:30代:教職9年目)、C教諭(女性:40代:教職19年目)の3名である。各教諭は、いずれも学級担任である。Z学校の組織的な変容を明らかにするため、A・B・C教諭とは別に、ミドルリーダーである教務主任(女性:40代:教職22年目)と校長(男性:50代:教職27年目)にもインタビュー調査を行った。インタビューは、ICレコーダーで記録した。インタビュー調査に加え、事前に、本研究に関する質問事項をウェブアンケートにて各々の回答を得た。インタビュー調査当日に、ウェブアンケートにて回答のあった内容を踏まえ、半構造化面接を行った。インタビュー調査は、1月末日に実施し、EBP的思考による教師や学校の変容について回答を得た。

(3) データの取り扱いについて

日本の学校においてEBP事例は非常に少ないため、インタビュー調査の実施にあたって、個人情報保護に最大限配慮した。固有名詞(学校名・都道府県名等も含む)を匿名化したり、対象となる児童に関する情報は、全て取り扱わないようにしたりした。このような研究実施上の配慮について、名古屋大学大学院の倫理審査委員会の承認(24-2144)を得た。記録した全ての発話は、テキストデータに起こした。分析にあたっては、MAXQDAのAIアシスト、自動コーディングツールやサブコードの提案を適時利用して、叙述を教職経験18年の筆者が何度も再整理した。さらに、博

士号をもつ外部研究者1名に、質的分析の妥当性を確認した。とくにインタビュー対象者の一貫した発言、各インタビュー対象者の共通する発言内容、各教諭による相互作用や発話ストーリーを確認した。分析の結果は、インタビュー対象者や所轄行政機関に送付し、内容の過度な切片化がないか確認した。このような分析は、日本語版 Standards for Reporting Qualitative Research のガイドライン (Kayama 2022) に基づき、分析の質を高めた。また日本評価学会社会実験分科会において、口頭発表を行い、内外の研究者から意見を求め、本研究の妥当性と正確性を高めた。なお、本論文に関して、開示すべき利益相反関連事項はない。

(4) Z学校の概要

Z学校は、各学年2クラスの中小規模校である。教育データの利活用に関して、X県では、県独自で、児童生徒の学力の伸びをより検証できるようにした学力テストを毎年実施している。Y市では、教育委員会主導で、EBPを目指し、勘と経験のみの教育からの脱却を目指している。この方針を踏まえ、Z学校では、独自に学力分析チャートを開発したり、特別支援の外部専門家を招いて定期的に研修会を実施したりしてきた。教育データ会議を昨年度より取り入れ、個別最適な学びの実現に取り組んできた。

3. EBP的思考による学校の変容

(1) 何がきっかけか？

学校のみドルリーダーである教務主任によると、Z学校が、学校全体で取組始めたのは、毎年行われる学力調査がきっかけだった。教務主任は、自校の学力テストの結果から、学力向上計画を立案し、学校全体をリードする立場にあった。これまで、学力データの分析を行い、改善策を立て、教師も児童も一緒になって取り組んできたが、思うような結果が出ていなかった。なぜ自校の学力は向上しないのか？このことが教務主任のEBP的思考の始まりとなった。

教務主任：学力調査分析と学力向上の取組か

ら改善策を立てて取り組んでも、次の年にその効果検証をしていないため、グッドプラクティスを探せていませんでした。

教務主任は、学力テストの結果が、向上しないことの原因について、主に2つあると考えていた。一つ目は、効果検証である。学力調査は、全国的に年に1回実施する。そのため、同一集団を対象とした実践前後の比較分析を行うことができない。このため、教務主任にとっては、一定の学力向上の実践を行っても、それが学力向上に影響したのかどうかといった効果検証が不十分だと感じていたのである。ここでいう教務主任の語る効果検証は、自分たちの教育実践に対するアウトカムがどうであったのかという意味を指す。2つ目はグッドプラクティスの特定である。教務主任は、様々な取組を実施する上で、本当にどれがよかったのかが分からなかった。教務主任の、「本当に効果があったのはどういう実践だったのか」を明らかにしたいという思いは他の場面でも一貫して語られた。

教務主任：こういう手立てがあったから伸びたっていうところを見つけれることが教育データの目的。いろいろ数撃ちゃ当たるんですけど、でも一番効果があるのを知りたいじゃないですか。

教務主任の「知りたいじゃないですか」と、強く発言された語りには、次への実践の改善につなげようとする探求心がみられる。このような実践上の探求心は、教務主任以外にもみられた。以下は、A教諭の語りである。

A教諭：個別最適な学びを始める際に、当時は実践例が多くなく、独学で始めました。自分なりの方法でやっていったので、それが正しいのかどうか、正解はないと思いますが、見極めるために、教育データの活用をしようと思いました。自分が行った学習方法や手立てが適切かどうか検証し、さらにブラッシュアップしていくために、教育データの活用をしています。

教職経験14年の男性A教諭（30代）は、Z学校で先進的に個別最適な学びの学習指導に取り組んできた。個別最適な学びとは、Z学校が3年程前から校内の授業改善として、目指してきたテーマである。この個別最適な学びは、学習指導要領に明記され、近年の学校現場での授業改善の大きなキーワードになっている。この個別最適な学びに対して、A教諭は、学習指導要領の方針に共感し、この具現化に取り組もうとした。しかし、実践事例が身近になく、自分なりの方法で調べて、試行錯誤の実践が続いた。その中で、試行錯誤の実践が有効かどうかを見極めるために、教育データの利活用に取り組んだという。

具体的には、A教諭は、自分なりに考えたり、調べたりして取り組んだ個別最適な学びの授業について、教育データ会議で積極的に実践事例を報告した。とくにA教諭が語ったのは、個別最適な学びを実現するための授業改善の一環として取り組んだということである。A教諭にとって、教育データを活用することそれ自体が目的ではなく、あくまで目指す授業の振り返り、専門職としての飽くなき探究心といった意味合いが強い。

教務主任やA教諭のように、実践の改善、自己の専門職としての力量形成の一助としてEBPに積極的に取り組む事例は、教育分野以外の看護分野でも、同様に報告されている（Melnysら 2022）。様々な手立ての中で、「何が効果のか？自分の実践は有効だったのか？」という探究心は、EBP的思考のきっかけだと考えられる。

一方、A教諭が積極的に教育データの利活用に取り組むようになったきっかけは、所属の管理職に言われてやるようになったとも正直に語っている。

A教諭：まあ、管理職にやってみてくれとも言われたので、まずは試験的に始めました。

A教諭のきっかけは、管理職に言われたことでもあると語った。専門職としての探求心と同時に、管理職に言われてやるようになったという受動的なきっかけの両方をA教諭はもつ。そして、A教諭は、当初、教育データについて、否定的な感情を抱いていた。

A教諭：最初は難しいもの。詳しく分析し、活用しなければならぬもの。教育データだけで、学びの向上につながるのか不安だった。

Z学校は、県内で独自の学力テストを行ったり、行政の主導で、EBPの方針を示されていたりした。そのため、A教諭はじめ、所轄の各学校は、どうしても上から教育データを使わなければならない状況でもあった。そして、それは、A教諭にとって、「学びの向上になるのか不安だった」という語りの通り、EBPの効果に対する不安を抱く結果となっていた。

ただ、A教諭は、一定の不安を抱きながらも、個別最適な学びを実現しようとする自身の実践の課題や疑問に答えるための一つの道具として、受動的なきっかけと探究心の両方をもちEBPに取り組んだ。

教務主任とA教諭とは対照的に、より受動的なきっかけを語ったのは、Z学校在籍歴の少ないB教諭とC教諭である。

B教諭は、調査該当年度、Z学校に転入した。その際に、教育データ分析を担当する教育データ会議の担当責任者になった。そこでは、学校長の方針や教務主任から「教育データを読みながら授業改善をしていきなさいね」と言われたという。

B教諭：データって何がわかるんだよみたいな、ただの数字じゃんみたいな感じでずっと生きてきて、動物的な感情と感覚で生きてきた人間なんですけど、この学校に来て、特に教育データ会議の担当になって、そこでやっぱりデータを読みながら授業改善をしていきなさいねというか、ある意味あの、言われたっていう、そういう会議やっていくよっていうふうに言われて、使ってみたら、やっぱりこの授業の時は点数が良かった。

C教諭は、該当年度より教育データ会議に参加するメンバーとなった。C教諭のきっかけは、この会議に参加することになったという受動的なものであった。B教諭とC教諭は、受動的なきっかけだけでなく、教育データそのものに対して、非常に否定的な印象をもっていた。

B教諭：読み取って分析するのが面倒だ。数字だけで、何かがわかるとは思えない。

C教諭：データの分析や見方が難しそうだと思っていました。

B教諭やC教諭からは、受動的なきっかけと共に、教育データそのものに対する当初の否定的な印象が同時に語られた。これは、前述のA教諭と同様である。

ここまでの発話を踏まえると、EBP的思考のきっかけについて、主に2つの現象がみてとれる。一つは、教師としての探究心である。教務主任の学力向上に対する様々な取組に対する効果検証、A教諭の自分の個別最適な学びの授業改善の有効性を見極めるためのものとしてEBPのスタートがあった。もう一つは、受動的なきっかけである。これらは「データは冷たいもの」、「人の感情はエビデンスでは分からない」といった教育データそのものに対する否定的な感情と共に語られた (Andersen 2020:2)。

これら2つが、本事例においては、EBP的思考のきっかけになったと推察される。

(2) 何が起こったか

では、次に、Z学校では何が起こったのだろうか。Z学校の教育データのEBPに関する取組について、Z学校の実践資料から、以下のようにまとめた。

Z学校の実践概要

- ・児童一人一人の教育データアンケート調査を、複数種類、複数回実施した。
- ・教育データを一元化管理したダッシュボードで共有化した。
- ・教育データ会議を月に1回30分実施し、取組の効果検証を行った。
- ・学校全体で、肯定的な行動支援を行う応用行動分析に取り組んだ。
- ・月に1回、外部専門家に指導を得た。

上記のような活動は、組織として取り組んできた客観的な記録である。本研究の目的である学校の変容を明らかにするためには、組織の中で、何が起こったのかといった当事者の解釈が必要にな

る。本研究では、各教諭の解釈を通して、学校全体としての変容と、その要因を明らかにする。

まず、学校全体をリードする教務主任の語りをみていく。教務主任は、本年度当初、Z学校の日課に教育データ会議の時間を位置付けたり、チームを組織したりした。

教務主任：今年度から日課表の方にもデータを見るようなミーティングを位置付けたりですとか、データを活用するようなプロジェクトを組織の中に入れてたりとかっていうふうにしながら、少しずつシステムを、組織を作っていました。

こうして組織されたチームの担当者にB教諭がなり、C教諭は、そのチームに入るようになった。このチームをEBPチームと本研究では命名する。校長もこのEBPチームについて「民間企業の組織のチーム制と同じような部署を作った」と語っている。

EBPチームを組織した背景には、インタビュー対象となった校長と教務主任の教育データに対する思いがあった。

校長：多くの学校って、行政からしてそうなんだけど、だいたい平均点で語るんですよ。なんで個別の伸びるものを入れてるのに、なぜ平均点で、何パーセントで上だとか下だとかいう話ですか。で、結局うちの学校もずっとそういうふうに平均でやっていて、平均的にここのところが弱いからやりましょうって言うと、みんな一斉にそのプリントとかなんとか全部やる。そういうことをやって、絶対計算が弱いとか、漢字ができないとか言ったって上がらないですよ。実は、この個別最適な学びやデータ活用っていうのは、一人一人を見ていく学びに変えていこうっていうふうに示しているんですよ。

校長の語りは、教育データを活用する際に、行政の取組やZ学校のこれまでの取組を振り返って、児童一人一人の学びに着目して授業改善をしていこうという学校経営理念と、「平均点のみに着目

する教育データの利用は、児童一人一人への対応が見えなくなる」という意味でのEBP的思考が垣間見える。

このような校長の一人一人の学びに着目した授業スタイルの重要性和、平均点のみに着目する教育データの利用への懸念は、教務主任にも同様の語りがみられた。

教務主任：やっぱり何のためにそのデータを使うのか、何のためにその調査をするのかっていう、まず目的をしっかりと教職員で共通理解を図らないと、本当にただ調査をして終わってしまっているっていうものがたくさんありました。本校は最上位目標である学校生活が楽しい、学びが楽しいっていう、子どもたちが本当にそういう子どもたちになっているのかっていう効果検証のためにこの調査を使うとか、学級が本当に満足するような過ごしやすい場になっているのかっていうことで、質問紙調査を使うというように、目的を明確にしてデータを紐付けていったっていうイメージです。

教務主任は、繰り返し、学校の教育方針を共有することがEBPにおいて重要であると語っていた。校長、教務主任ともに、学校の教育方針や教育データの目的といった理念の共有、学力テストの平均点のみを捉えた一斉指導への懸念をもっていった。つまり、学校リーダー層の「EBP的思考の共有化」がみられたのである。

このようなEBP的思考の上に、教育データを扱うEBPチームが組織されたのである。この「チーム形成」という組織的な動き出しによって、Z学校のEBP的思考が広がっていったと推測できる。

では、実際に、教育データ会議の場で、その時、何が起こったのだろうか。

注目すべきが、教育データそのものを否定的に捉えていた各教諭の教育データ会議での語りである。A教諭、B教諭、C教諭の3名は、上述のように、教育データそのものを「難しいもの」「ただの数字」「面倒なもの」として、否定的に捉えていた。しかし、Z学校の年間の取組を経て、全ての教諭が教育データに対する見方が「変わった」とウェブ

アンケートにて答えた。

実際、教育データに対して、否定的な印象をもつ学校現場の教諭は多い（Andersen 2020:2）。その中で、教育データを扱う場である教育データ会議での出来事を明らかにすることで、今後のEBPの在り方への重要な示唆があると考えられる。各教諭の語りを確認していく。

A教諭は、本年度より本格的に行われるようになった教育データ会議について以下のように語った。

A教諭：教育データ会議でテストの結果と結びつけて、検証してみました。自由進度学習とかが、もうほぼ私が話さないような授業っていうか、そういうのをやらせてみて。最初はテストの結果が変わらなかったで、じゃあ自分（教師）が話さなくていいじゃんって思って、その一斉授業じゃなくていいじゃんってところから、個別最適な学びをちょっともう少しやってみようかなと思っていました。

A教諭：この学年3回目、3連続なので、これまでのテストの結果と比べたりとかして、あんまり自分が喋らなくても単元テストの結果変わんないじゃん。結構、力入れてやった一斉授業なのに、子ども主体でやらせたやつは全然変わらなかったみたいなところがありました。

A教諭は、これまでの一斉授業スタイルから子ども主体の授業スタイルに挑戦していた。そして、教育データ会議にて、これまでの一斉授業でのテストの結果と比較分析した。その結果、これまで教師が丁寧に説明する一斉授業と比べて、テストの結果が変わらなかった。これまでA教諭は、該当学年の担任を3連続で3回経験している。その担任経験の中で、自分が力を込めて説明したにも関わらず、子どもたち主体の学習を生み出せなかったのである。このような経験を通して、A教諭は、以下のように語った。

A教諭：テストの結果が、子ども主体でやったとしても点数が変わらなかったり、逆に

そっちの方が点数が良かったりするので、何かあるのかな？って思いました。

これまでの一斉授業スタイルを変えて、子ども主体の授業実践に取り組んだ結果、テストの点数が変わらなかったり、逆に良かったりした経験を通して、A教諭は、「何かあるのかな？」と疑問をもった。この語りは、一斉指導のスタイルで行ってきたA教諭にとって、これまでの経験を転換する疑問が生まれた瞬間である。これは、教育データ会議において、授業実践の改善に取り組んだA教諭の強い気付き：リフレクションとなった。ドナルド・ショーン（2001）によれば、教師のリフレクションは、専門職において重要な役割を担う。このA教諭のリフレクションは、教育データを分析したからこそ引き起こされたものである。これはA教諭のEBP的思考のリフレクションであるといえよう。

このようなリフレクションが起こる現象は、実はB教諭やC教諭にもみられた。

B教諭：（教育データを）使ってみたら、やっぱりこの授業の時は点数が良かったとか、この授業をやった時は点数がダメだったっていうことが本当に表れるんだなと。今まで感覚でつかんできたものが数字でも表れるというか、その裏付けになるっていうか、そういうことを経験しましたね。

B教諭は、当初、「データって何がわかるんだよみたいな、ただの数字じゃん」と、強い否定的な感情を抱いていた。そのB教諭が、教育データ会議で、教育データを使うことの良さに気付いたという。教育データ会議にて、教育データを正確に見ていく作業を通して、今まで感覚でつかんできた児童の実態が数字で表れることを感じたという。このリフレクションをB教諭は、「裏付け」と表現している。この「裏付け」という表現は、B教諭の発言の中で何度も語られた。

B教諭：定期的にデータ取るんですけど、本当に今、あ、ここがダメだなんて私が思ったところがやっぱり如実にへこんでいたり

かするので、そういうやっぱ裏付けにもなるし。それが正確に分かったら、どうしていけばいいのかなっていう対策を子どもたちと一緒に考えられる。

このようなB教諭の「裏付け」は、学習面以外の場面での指導の支えになったという。

B教諭：なんかアセスメントって細かく、なんだろう、なんか運動のここが苦手ですみたいなのが出てるんで、そのデータを見た時は、ああ、これが苦手だったから、これができなかったわけねみたいなのを、納得した記憶があります。

「裏付け」と表現したB教諭のEBP的思考は、自分の「納得」と結びついた。教育データ会議の場で、教師としての感覚を支え、次の指導の方針に役立つB教諭の「納得」が生まれた。このB教諭の「納得」は、児童の姿を、教育データを通して分析的にみようとするEBP的思考のリフレクションであろう。

C教諭：データ、学力って言ったときに、全部のデータを見て、あ、この子低いから支援が必要だとか思ってたんですけど、そうじゃなくて、この子、普通の授業の中では全然できてないけど、ここは得意なんだみたいな、そういう細かいところが見れるようになりました。

C教諭は、主に、児童の学力データを見て、児童の実態を、さらに「細かいところ」を捉えることができるようになったと語る。これまでC教諭は、平均的な全体の学力データを見て、学習支援が必要な児童を見極めて支援してきた。しかし、教育データ会議にて、詳細な学力データをみて、これまで学習支援が必要だと捉えていた児童の得意な学力領域を見出したのである。単に学習支援の必要な児童だということこれまでのC教諭の指導観から、学習内容によって、学力の実態や特性が異なるという新しい認識に変わっていった。その結果、C教諭は、児童の学力の実態を、より詳細に

看取ることの大切さを認識したという。この「細かいところまで見れるようになった」ということが、C教諭のEBP的思考のリフレクションといえる。

ここまでA教諭・B教諭・C教諭の教育データ会議を通じた出来事を示した。教職経験等異なる背景をもつA教諭・B教諭・C教諭の語りには、いくつかの共通点を見出すことができる。

まず、EBP的思考のリフレクションが起きたことである。A教諭の「何かあるのかな?」、B教諭の「裏付け」や「納得」、C教諭の「細かいところまでみれるようになった」という語りは、教育データ会議の場で生まれた。A教諭は、これまでの一生懸命に取り組んだ授業とテストの結果が変わらなかった、もしくは、上がったことから、少し悲壮的な語り、B教諭は、自分自身の感覚を裏付けたということから、喜びのような語り、C教諭は、児童への見方を細かくしたこと、冷静な語りであった。いずれの語りも各教諭の情緒が強く表れていた。このため、EBP的思考のリフレクションは、教師の指導観の再形成という点で重要な意味をもつのではないだろうか。

(3) 変化の要因

では、どのようにして、上述のようなリフレクションが起こったのだろうか。本研究では、教育データ会議において、メンター役となっていた教務主任の働きかけに着目した。

EBPにおいて、メンターの存在の重要性は、先行研究の中で指摘されている。熊井・杉田（2019）では、EBPメンター役の教師の働きかけにより、学校全体のEBP的思考の広がりにつながったことを示唆した。本事例におけるEBPメンターは、教務主任に他ならない。教務主任は、校長の学校教育方針やEBP的思考の共有化を図り、チームの組織作りに携わった。EBPメンターとしての、教務主任の働きかけが、先述のリフレクションを起こした可能性が高い。この点から教務主任の働きかけについての語りから、学校の変容の要因を探る。

でにありました。ただ、土壌としてはできていたんですが、教育データを実際に見て、もっと正確に見ていこうって言ってきたのは、昨年度から始まりました。

このように、教務主任は、児童の実態を全ての教諭で見えていこうとする雰囲気があったことを認める一方、教育データをより正確に読み取っていく各教諭の力量に課題があることを感じていた。そして、教務主任は、教育データの正確な読み取りを促した。教務主任は以下のように語った。

教務主任：（指導の結果に関する教育データは）なんとなく見ちゃう。お疲れみたい、そんな感じに、今まではなっていました。なんとなくじゃなくて、80人しかできなかったのが120人になったってことは、すごいその取組良かったんだねっていう、そういうところですね。

教務主任は、ある生活面での児童の行動について、80人しかできなかった実態から、ある指導方法によって、120人に増えたということは、一定の児童の変容や指導方法の効果があったこと等、効果検証について述べている。このような教育データの見方や解釈について、教育データ会議において、各教諭に伝えてきたという。この教育データをどう解釈していくのかという力量は、Sackett（1997:5）の指摘するEBPの「批判的な吟味」である。「批判的吟味」とは、エビデンスをただ鵜呑みにするのではなく、収集したエビデンスを注意深く読み解き、思慮深く吟味・解釈することである。つまり、教育データ会議において、教務主任は、同僚教諭の教育データの細かな読み取りや解釈する力量を高めていくEBP的思考を意図的に高めようとしていたのである。

もう一つ、この教育データ会議の場にて、教務主任が意識的に、働きかけたことが、「無意識の自覚化」である。

教務主任：本校は、多様な子どもたちがもうすでに在籍していたので、まずみんな子どもたちを見ていこうっていう雰囲気はもうす

教務主任：言葉にすると、私こんなことやってたんだって、さっきの意識化させることができるので、「どうですか?」って聞くと、いっ

ばい出てくるんですね。やってるじゃんみたいな感じになるんですけど、多分、その時間で整理されるんだと思うんですね。そこがすごく大事なんだなっていうのは思いますね。

教務主任は、教育データ会議の場で、教育データの解釈を伝えながら、同時に、各教諭に、「どうですか？」と問う働きかけを意識的にした。そこで、教務主任がねらっていたのは、教育データを基にして、教師の無意識を意識化させることだという。この教務主任の「どうですか？」という問いによって、自分の実践を「整理する時間」が生まれた。前述のA教諭、B教諭、C教諭に起こったリフレクションは、まさに、この時、教務主任の発する無意識を意識化させる「問い」によって、起こったのだと考えられる。

さらに特徴的なのが、教務主任のEBP的思考が常に職場の雰囲気作りと一緒に語られていたのである。

教務主任：教育データ会議は、笑って終われるようにしようって言ってます。これもデータ活用になりますけど、本当にデータを見てもらえと、たった一人の子の行動をどうにかしようっていうのを色々な人としゃべる。それもまたすごくいい感じになっているんです。

教務主任は、同僚教諭が「やってよかった」と思ってもらえるような教育データ会議を目指した。「すごくいい感じ」とは、どのようなものかについて、さらに、教務主任は語った。

教務主任：生徒指導での教育データ会議も完全にそういう感じです。一つの問題、ちょっと気になるなって子に対していろんなメンバーが、担任以外の人たちが本当にこうでもない、ああでもないって言って、誰も暗くならないっていうか、最後笑って終わるみたいな。

教務主任の語る「誰も暗くならない会議、最後、

笑って終わる会議」は、学校現場で一般的に行われる問題行動や生徒指導場面での生徒指導会議（多くの生徒指導会議の雰囲気は暗くなる）と、大きく異なる。このような雰囲気になるのは、教務主任による教育データ会議において、EBP的思考を広げようとする組織作りの工夫であると考えられる。EBP的思考そのものが組織を高めていく潤滑油となることは、Ploegほか（2014:6）が指摘してきた。本事例の「いい感じ」が生まれる理由は、教務主任の「潤滑油としてのEBP的思考」だと考えられる。そして、このような教育データ会議を通して、EBP的思考は当たり前であるという職場の風土が広がっていったと教務主任は語った。

この点について、先述の熊井・杉田（2019）の研究に再度注目できる。熊井・杉田（2019：121）では、小学校において、EBPメンターの働きかけにより、若手とベテラン問わず、EBPが、「面白い」「楽しい」といった情緒的な評価を得ていたことを明らかにした。本事例は、熊井・杉田らの取り上げた事例と、全く異なる時期の異なる場での事例でありながら、EBPメンターの存在意義を見出すことができるのである。両事例の詳細な比較分析は匿名性の観点から難しいが、いずれにしても、本事例において、A、B、C教諭のリフレクションが生まれた事実、そして、その要因には、教務主任のEBPメンターとしての働きかけが観察されたのである。

(4) どうなったのか

このようなEBPメンターの働きかけが継続的に行われたZ学校は、その後、どうなったのだろうか。教務主任、校長は、以下のように語る。

校長：自分たちも先生たちがすごくそれを意識して、しかもそこに抵抗が全然なくなってきた。

教務主任：教育データを活用することが当たり前になっちゃって、もうすごいですね。必ずそれはやっていますね。データをとるのが前提。変容を見るっていうところがね。

校長：他の色々な研修会とかなんとかとかも、みんなデータ見ながらやっているよね。なんか当たり前にそういう感じ。

校長も教務主任も教育データを活用することが当たり前になっている組織文化を感じていた。教育データを正確に捉え、児童一人一人を支援していくという校長、教務主任のEBP的思考が組織文化として定着していったことが読み取れる。実際、各教諭は、教育データそのものに対する意識の変化を一貫して語っていた。

A教諭：今、この学校だとそのデータ分析して終わりじゃなくて、それからどうしていくか、個別最適までつないでいく方が学級のパワーアップにつなげていくのかとかっていうのになっている。

A教諭は、「この学校だと…」と表現し、Z学校のEBP的思考が学校の文化であるという認識を示した。

B教諭：定期的データ取るんですけど、それが正確に分かったら、どうしていけばいいのかなっていう対策を子どもたちと一緒に考えられる。

B教諭は、定期的にデータをとり、今後の指導の方針を子どもと一緒に考える材料になったという。

C教諭：テストの点数や児童の意欲等、必要な情報を選択して活用することができるようになった。

C教諭は、学力や生活面での意欲等、幅広く、必要な情報を自ら活用するようになったと語った。これらA教諭、B教諭、C教諭の語りは、EBP的思考の「定着」と捉えることができる。このようなことから、EBP的思考が組織に文化として広がり、「定着」する一連の流れを確認することができた。

4. インタビュー調査のまとめ

本インタビュー調査でのZ学校の年間の変容を

まとめると、探求心・受動的なきっかけ→EBP的思考の共有化→チーム形成→実践（教育データ会議・実践）・EBPメンター→リフレクション→EBP的思考の定着となっていた。

他方、上記の変容はあくまで学校組織全体としての年間の変容であるため、それぞれの立場においては、様相が異なっていた。例えば、リフレクションについて多くの語りが見られたのは、実践を行っている担任教諭が多かった。教務主任や校長は組織マネジメントについての語りが多かった。そして、教務主任のEBPメンターとしての中核的な働きかけによるEBP的思考の広がり・定着をみることができた。

5. 総合考察

本研究では、EBP的思考によって、教師や学校がどう変容するのかを事例的解明を試みた。EBP的思考に着目することで、実証研究の乏しさを乗り越え、EBPが学校現場でどのような変容を辿るのかについて実践的示唆を行うことを目指した。その結果、調査対象となったZ学校では以下の変容が確認された。「探求心・受動的なきっかけ→EBP的思考の共有化→チーム形成→実践（教育データ会議・実践）・EBPメンター→リフレクション→EBP的思考の定着」である。では、なぜこのような変容をたどったのだろうか。先行研究により以下4点を考察した。

(1) EBP的思考の広がり

本事例の中で脈々と流れるのは、EBP的思考の醸成である。A教諭や教務主任・校長の「実践上での探求心」や「平均点のみで捉える指導法への懸念」といったEBP的思考が、組織の中でB教諭、C教諭に広がっていく様相が確認された。

Z学校の校長は「教育データの利活用をうちの学校でやるとしたらどういうふうにできるだろうなって考えてやってきた。」と語った。このような語りは、植田（2022：32）の「エビデンスを活用する学校づくりを学校管理職が中心となって行っていくことが重要である」という指摘が具現化されたものであると捉えることができる。

本事例では、校長や教務主任が、EBPの重要性を認識し、組織文化づくり役となっていたことが確認された。学校管理職が学校研究の文化に与える影響については、これまで多くの研究で指摘されている（木原ら 2015：173）。このため、学校管理職のエビデンス受容認識や管理職のEBP的思考の形成と学校組織の影響については、今後さらに校長個人やその他の事例を通して検証していくことが必要である。

(2) EBP的思考のリフレクション

本事例で確認できたリフレクションは、EBP的思考のリフレクションだと考えられる。教師のリフレクションについては、これまで教師の成長過程の中で重要な要素として指摘されてきた（ドナルド・ショーン 2001）。とくに、同僚教諭との会話は、自身のメタ認知を促し、実践の改善につながる。これまでのリフレクション研究では、同僚教諭との会話が主な手立てとして指摘されていたが、本事例は、EBP的思考のリフレクションが確認された。言うなれば、先行研究では、「同僚教諭との会話」によるリフレクションに対して、本事例では、「教育データとの会話」によるリフレクションである。両者の違いや発生要因等、詳細を明らかにすることはできていないが、本事例で確認されたこと自体が、今後の研究の新たな視点を提示することができる。

(3) EBPメンターの存在

EBPメンターの存在は本事例を語る上で、欠かすことができない。Melnykら（2022）は、EBP推進においてチームアプローチを組み込むことの重要性を強調している。本事例では、EBPチームが意図的に組織されており、Melnykらが示唆したチームでのEBPに取り組む重要性を実践的に示した例と言える。このEBPメンターの重要性を認識すれば、EBP的思考を有するチーム作り、EBPメンターの働きかけの有無によって、学校におけるエビデンス活用の功罪（内田 2015）の大きな分岐点になる可能性がある。

(4) 各立場に応じたEBP推進の役割

校長、教務主任、各教諭によって、EBP的思考

の発展は異なっていた。組織のEBP発展については、これまで推進要因や阻害要因について明らかにされてきた（Langerら 2016）が、本事例からは、役職によって、その要因が異なる可能性が推測される。教務主任やA教諭のように、実践上の課題意識を強く抱いていた教諭とそうでなかった教諭でも、EBP育成や支援の在り方が異なる可能性が高い。今後は、これらの点を踏まえた上で、教師のEBP的思考を高めるプログラムや支援ツールの在り方を開発していくことが求められる。

6. 今後の展望

最後に、本研究の限界と今後の課題について述べる。本研究は、EBP的思考によって、教師や学校組織はどうなるのかについて事例の解明を試みた。その結果、組織内の変容、リフレクションやEBPメンターの存在を明らかにすることができた。一方、本研究は、先進的な学校を対象にした探索的な事例研究である。今後は、教諭・教務主任・管理職といったそれぞれの役職の違いによる影響、キャリア形成における指導観の変容等、教師個人や学校組織に対して、更なる介入の効果検証を計量的に解明することが期待される。

また、本研究では、教育データやエビデンスそのものの質や多義性（米原 2024：3）についての議論もあるが、本稿では、扱っていない。このため、インタビュー調査から得られた語りの中には、データとエビデンスの混同がみられた。しかし、このような語りを一つ一つ精査していくと、教師や学校の変容を事例的にでも明らかにしようとする本研究の趣旨と異なる。ただし、教育データやエビデンスそのものの質に着目することで、EBPの発展の様相が異なる可能性を否定するものでもない。また、本事例で主に語られた学力データは、あくまでZ学校において得られた学力テストによるデータであるため、測定しやすい学力のみを取り扱ってしまう危険性も孕んでいる。

これらの点を踏まえると、目指すべき学力像や教育的価値、エビデンスの質や、その多義性を踏まえたEBPによる教師や学校の変容について、継続的な実践研究が必要であろう。その際、看護分

野におけるエビデンス活用プロセスを示したARCCモデル（Melnykら 2022）等、一定の理論モデルを活用して事例を考察することで実証性を高めることができると考えられる。

それでも、学校という総合的な実践の場から、EBP的思考に着目し、事例研究を蓄積していくこと自体は、教育分野におけるEBPの実践的示唆を深化させることになる。今後、教育分野におけるEBPの実践的研究が一層取り組まれることを願っている。

謝辞

本研究についてご指導頂きました内田良先生（名古屋大学）、日本評価学会社会実験分科会の皆様、インタビュー調査にご協力いただきました先生方に心より感謝申し上げます。

参考文献

- 植田みどり（2022）「学校経営におけるエビデンス活用の限界と可能性」、『日本教育経営学会紀要』、64：21-37
- 内田良（2015）「教育実践におけるエビデンスの功と罪」、『教育学研究』、82（2）：277-86
- 木原俊行・島田希・寺嶋浩介（2015）「学校における実践研究の発展要因の構造に関するモデルの開発―「専門的な学習共同体」の発展に関する知見を参照して―」、『日本教育工学会論文誌』、39（3）：167-179
- 佐藤浩一（2016）「学校における効果的いじめ防止要素の考察―包括的いじめ防止プログラムの開発に向けて―」、『子育て研究』、6：9-20
- 杉田浩崇・熊井将太（2019）『「エビデンスに基づく教育」の関を探る：教育学における規範と事実をめぐる』、春風社
- 杉谷和哉（2022）『政策にエビデンスは必要なのか：EBPMと政治のあいだ』、ミネルヴァ書房
- 鈴木久米男・佐藤進（2018）「学校経営における教育データ活用に関する一考察：A中学校における学校評価の分析結果を踏まえた教育活動への活用とは」、『岩手大学教育学部附属教育実践総合センター研究紀要』、17：73-85
- 津富宏（2008）「少年非行対策におけるエビデンスの活用」、『少年非行の行動科学』、北大路書房
- ドナルド・ショーン（佐藤学・秋田喜代美訳）（2001）『専門家の知恵』、ゆみる出版
- 松下良平（2015）「エビデンスに基づく教育の逆説―教育の失調から教育学の廃棄へ―」、『教育学研究』、82（2）：202-215
- 森俊郎（2019）「EBEを実践で語ろう」、『エビデンスに基づく教育』の関を探る：教育学における規範と事実をめぐる』、春風社、130-167
- 文部科学省（2024）「教育データ利活用の実現に向けた実効的な方策について（議論のまとめ）」https://www.mext.go.jp/content/20240328-mext_syoto01-000034992_1.pdf（2024年12月3日閲覧）
- 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
- 米原あき（2024）「評価研究及び評価実践におけるエビデンス理解の多様性と多義性」、『日本評価研究』、24（1）：3-5
- American Psychological Association, Presidential Task Force on Evidence-Based Practice (2006). *Evidence-based practice in psychology*. *American Psychologist*, 61 (4), 271-285.
- Andersen, I.G. (2020). What went wrong: Examining teachers' data use and instructional decision making through a bottom-up data intervention in Denmark. *International Journal of Educational Research*, 102.
- Laurenz Langer, Janice Tripney, David Gough (2016). The Science of Using Science Researching the Use of Research Evidence in Decision-Making, *Technical report*, April.
- Kayama, M. Asahara, K. Miura, Y. Oyamada, K. Sakaki, M. Nishimura, Y. Yamamoto, N. Gureggu, M. & Kinoshita, Y. (2022). 「日本語版 Standards for Reporting Qualitative Research (SR QR)」『IMPACT: New Frontiers in Medical Technology』、23、16-18、<https://quarin.jp/guideline/srqr>（2024年4月2日閲覧）.
- Melnyk, B.M. & Fineout-Overholt, E. (2022). *Evidence-Based Practice in Nursing & Healthcare. A Guide to Best Practice*. Philadelphia: Lippincott Williams & Wilkins.
- Ploeg, J. Markle-Reid, M. Davies, B. Higuchi, K. Gifford, W. Bajnok, I. McConnell, H. Plenderleith, J. Foster, S. & Bookey-Bassett, S. (2014). Spreading and sustaining best

practices for home care of older adults: A grounded theory study. *Implementation Science*, 9, 162.

Sackett, D.L. (1997). Evidence-Based Medicine. *Seminars in Perinatology*, 21 (1), 3-5.

(2024.12.18受理)

What happens to teachers and schools with ‘EBP thinking’?

Toshiro Mori

Nagoya University • Yoro Tobu Junior High School, Gifu.

m072132@gmail.com

Abstract

In recent years, the use of educational data has been promoted in schools. The merits and demerits of Evidence Based Practice, which aims to link evidence derived from educational data to improvements in practice, have been discussed. However, practical suggestions are lacking in school settings, where various factors are intertwined. Therefore, in this study, a school where ‘EBP Thinking’ was developed was surveyed, and a case study was conducted to elucidate what triggered the change, what happened and how the school organisation and individual teachers were transformed. As a result of interviews with several staff members, a series of transformations such as ‘exploration and passive triggers → shared EBP thinking → team formation → practice (educational data meetings and classroom practice) , EBP mentors → reflection → consolidation of EBP thinking’ were identified. Through this study, the significance of reflection on EBP thinking, the existence of EBP mentors and the fostering of EBP thinking was suggested.

Keywords

Education date, Evidence Based Practice, EBP mentor, Teacher

【企画記事】

インパクト評価の発展の可能性： 望ましい社会的価値の取り込み

佐々木 亮

国際開発センター

sasaki.ryo@idcj.or.jp

インパクト評価（Impact Evaluation）のデザインには、実験デザインとそれ以外のデザインがある。「実験デザイン」（Experimental Design）は近年は、RCT（Randomized Controlled Trial）と一般に呼ばれているし、ICT分野ではA/Bテスト（A群（介入群）とB群（統制群）の2群のテストの意味）とも呼ばれている。「それ以外のデザイン」としては、回帰・分断デザイン、マッチングデザイン、一般指標デザインなどがあるとされる（Rossi et al. 1991, p.261）。さらに、たいへん単純だが広く使われているデザインに事前－事後デザインがある。

こうした定量的なインパクト評価のデザインで使用される分析のツールの一つとして回帰分析（Regression Analysis）がある。現場で収集したデータに適用して、どんな要因（変数）が影響しているのかを分析するツールである。日本では中学校で学ぶ「 $Y=aX+b$ 」という直線の式をイメージすると分かりやすく、その式の中の aX が複数あるということである。もちろんその複数の X の一つが、私たちの介入適用の有無（通常は1/0で表すことが多い）である。

そして、実験デザイン（RCT）とともに回帰分析が併用されることがある。少なくとも国際開発分野では、実験デザイン(RCT)で得られたデータにシンプルな2群の平均値差検定（t検定）を適用して効果の有無を判定したあとに、同じデータに回帰分析を適用して、介入変数のほかに、複数の属性変数（例：性別や年齢など通常は自分の意思で変えられない要素）や環境変数（例：地域の経済状況など）を導入して、それら属性変数・環境変数の影響を推定するとともに、属性変数・環境変数の影響をコントロールしたあとでも介入変数がインパクト（効果）を発揮したかどうかを判定するという手続きである。インパクト評価の論文に掲載された表で、介入行為の X のみの単回帰分析の結果（2群の平均値差検定（t検定）の結果と同一）が第1列目に掲載され、介入行為の X の他に複数の X を導入した重回帰分析の結果が第2列目に掲載された表を見ることがある（実例としてKozuka et al. 2016）。2群の平均値差検定（t検定）の結果と重回帰分析の結果を別々の2枚の表にもできるが、スペースの節約のために一つの表で表示していると理解される。

ICT技術の発展により、ICT分野でも回帰分析が広く応用して使用されており、SNSの広告表示や商品のレコメンド（推奨機能）でも回帰分析で得られた式が（もちろんより高度の分析手法もあるが）一般に使われている。

こうした状況に関連して、MITメディアラボ所長兼教授（2011-2019）であった伊藤穰一は次のように指摘している（伊藤 2019）。なお以下の引用文のなかのアルゴリズムとは、ICT分野でも日常的に使用される回帰分析を含む統計分析一般のことを指していると理解している。

過去のデータ（中略）に頼るだけでは、歴史的な不公正を固定化してしまうだろう。既存のアルゴリズムとそれ

に基づいたシステムは、理想の未来ではなく過去の規範に従っている。これでは社会の進歩にはつながらず、むしろそれにブレーキをかけることになるはずだ。

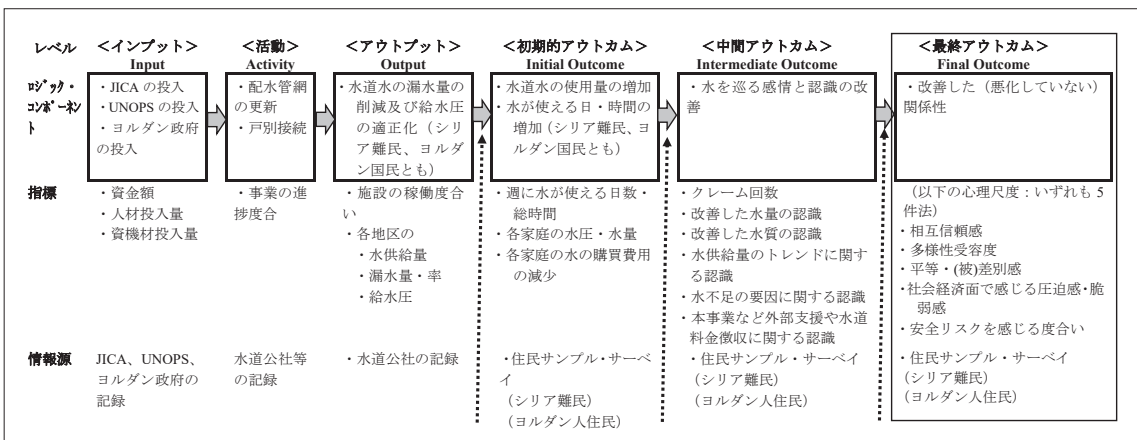
つまり、介入以外の属性や環境の指標値が差別を含んでいる場合には、その差別まで再生産しかねないということである。これはBlack Lives Matter (BLM) 運動が提起した構造的な差別に関する伊藤の考察でもあったが、結論として伊藤（2019）は、倫理判断や価値基準といったものを排除して純粋に数学的な手法で「公正さ」を定義しようとする傾向を批判している。

この状況を改善する方法はあるだろうか。評価は「ものごとの価値を明らかにすること」（Evaluation is determination of value of things）（Scriven 1991）と定義され、評価＝事実特定＋価値判断と公式のように説明されることがある。これからわかることは評価は価値中立ではないということで、インパクト評価も「評価」と言うからには、積極的に社会的な価値の判断基準を導入していく方法を考えるべきということである。

具体例を一つ上げる。世界では難民の発生と流入による多民族化が頻繁に見られるし、日本社会でも多民族化は必然となる可能性がある。そして「多様性」は今後の好ましい社会的な価値である。この多様性の価値を取り込んだインパクト評価の例である。図1のロジックモデルの右端が最終アウトカムであるが、その最終アウトカムの定義に多様性の概念が取り込まれている。これに基づいて多様性の受容に関する被質問を設定してその回答をインパクト推定の説明変数として用いている。UNICEFのサーベイ用紙（Valli et al. 2018）とWorld Food Programme（2017）のサーベイ用紙を参考にして慎重に作成した設問文とリッカートスケール型回答選択肢が開発されて用いられている事例である。

もう一つの具体例として、属性変数であるジェンダーの自己認識を、従来の男性／女性という2値の選択肢(binomial)から、男性／女性／ノンバイナリーという3値の選択肢とすることが考えられる。これはジェンダーの「多様性」という新しい社会的な価値を反映させる提案である。これにより3群のサブグループ分析が可能になるし、回帰分析でも説明変数の一つとして2値項（1/0）ではなく、図2のようにジェンダー・コード・マトリックスによる複数項（3つに分解されたX）として入れることになる（ただし実際の分析作業が複雑になるわけではない）。なお、ジェンダー研究のサーベイではこの複数のジェンダー分類の選択

図1 ロジックモデルの例（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）



（出所）国際協力機構（2023）『ヨルダン国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』（委託先：国際開発センター）（https://www.jica.go.jp/Resource/activities/evaluation/ku57pq0000198n75-att/Water_Supply_Improvement_Report_ja.pdf 最終アクセス日：2024年12月11日）

図2 新しいジェンダー・コード・マトリックスの例

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

(出所) 筆者作成

肢をすでに取り入れていると認識している。また、その他の自己認識の項目でも同様に多様性の取り込みを検討できる可能性がある。

以上、多様性という望ましい社会的な価値の取り込みの例を二つ紹介した。

私たちは、インパクト評価の主要な分析ツールの一つである回帰分析について、価値中立であるべきという態度を超えて、望ましい社会的な価値の判断基準を積極的に導入していくべきである。それが統計学の発展に関する評価学の貢献となり、社会の改善に貢献することになると言える。

なお、他の研究者が類似の主張をすでにしている場合もあるが、この小論で述べているのはそれを踏まえたうえで、望ましい社会的価値の取り込みを今まさに実践しようということである。

参考文献

- 伊藤穰一 (2019) 「アルゴリズムがつくる「公正さ」には、差別を助長する危険性が潜んでいる」、『WIRED Opinion』 (<https://wired.jp/2019/05/18/ideas-joi-ito-insurance-algorithms> 最終アクセス日：2024年11月16日)
- Kozuka, E., Sawada, Y. and Todo, Y. (2016). How Can Community Participation Improve Educational Outcomes? Experimental Evidence from a School-Based Management Project in Burkina Faso. *JICA-RI Working Paper*, No.112. JICA Research Institute. (https://www.jica.go.jp/Resource/jica-ri/publication/workingpaper/jrft3q000000273e-att/JICA-RI_WP_No.112.pdf Access date: 16 Nov.2024.)
- Rossi, P.H. Freeman, H. & Lipsay, M.W. (1999). *Evaluation: A Systematic Approach (6th ed.)*. Sage Publications, Inc.
- Scriven, M. (1991). *Evaluation Thesaurus (4th ed.)*. Sage Publications Inc.
- Valli, E., Peterman, A., & Hidrobo, M. (2018). *Economic Transfers and Social Cohesion in a Refugee-hosting Setting*. UNICEF Office of Research.
- World Food Programme. (2017). *Social Cohesion in Turkey: Refugee and Host Community Online Survey. Round 1*. WFP.

【特集に寄せて】

日本の業績測定はハトリから何を学べるか

田中 啓

静岡文化芸術大学

ハリー・ハトリ（Harry P. Hatry）の名は、公共部門で評価に関わる者の間で広く知られている。ハトリは「業績管理の父（the father of performance management）」と形容されるように、業績管理や業績測定の研究と実務に先駆的に取り組んできたパイオニアであり、今日世界の公共部門で広く採用されている業績測定の手法面の原則の確立と実務への応用において、多大なる貢献があった。ハトリの貢献がなければ、恐らく業績測定はこれほど広く普及していなかったであろうし、もしかしたら、業績測定自体も今日とはかなり異なるものになっていたかもしれない。

そのハトリが2023年2月20日に亡くなった。享年92歳であった。最晩年までアーバン・インスティテュートの研究者として、業績測定と業績管理の研究・実務に第一線で関わることを貫いた生涯であった。

ここでハトリの経歴を簡単に紹介しておく。ハトリはロングアイランド（ニューヨーク州）出身で、イェール大学（エンジニアリング専攻）を卒業後、ゼネラル・エレクトリック社に入社した。その後の足取りには不明な点も多いが、国防省のロバート・マクナマラ長官の知遇を得て、同長官の下で国防省におけるPPBSの実務に携わったとみられる。1968年にマクナマラが国防長官を辞任し、ワシントンD.C.にアーバン・インスティテュートというシンクタンクを設立すると、間もなくハトリもそのメンバーに加わった。

アーバン・インスティテュートでは、連邦政府機関をはじめとする公的機関の依頼に応じて、プログラム評価や業績測定に関連する調査分析プロジェクトに従事した。さらに、アメリカ国内を飛び回るのはもちろんのこと、時には外国に出向いて、政府機関や州・地方政府、非営利組織による業績管理の実践を支援した。また彼は、自身の研究成果や公的機関に対する実務的支援から得た知見を読みやすい著作物としてまとめて積極的に公表してきた。

ハトリの業績はアメリカの国内外で高く評価されており、生涯に多数の賞を受賞している。特筆すべきは、1999年に米国行政学会（ASPA）がハリー・ハトリの名を冠した賞（the Harry Hatry Award for Distinguished Practice in Performance Measurement）を設けたことである。存命中に自身の名を冠する賞が作られるということは、彼の功績の大きさを端的に物語っている。また2014年には、ジョセフ・ホウリー（Joseph S. Wholey）との共著論文“The Case for Performance Monitoring”（Wholey and Hatry 1992）が1940年以降 *Public Administration Review* に掲載された論文の中で、最も影響力のある75篇のうちの1篇に選ばれた。ちなみに、ホウリーはプログラム評価の分野で大きな功績のあった人物であり、ハトリにとって重要な共同研究者でもあったが、彼もまた2023年8月に亡くなっている。

本特集は、ハトリの逝去を受けて構想された、いわば追悼企画である。『日本評価研究』誌上において、これまで故人を追悼する特集が組まれたことはなく、同誌としても新しい試みになる。偉大な故人とはいえ、アメリカ人であるハトリの追悼企画を『日本評価研究』で組むことについては、異論の向きがある

かもしれない。

本特集を企画したのは、主に以下の2つの理由による。

第1に、ハトリの日本における影響力の大きさに鑑みてである。ハトリが日本の業績測定に与えた影響の大きさを客観的に測ることは難しいものの、本稿の読者の中で、ハトリの名を聞いたことがない人は、恐らく極めて少ないであろう。むしろ読者の多くは、ハトリの著作に何らかの形で触れてきたのではないだろうか。

もちろん本特集の寄稿者（田中、左近会員、小野会員）も例外ではない。いずれも業績測定の分野を主要なフィールドとして研究や実務に関わる中で、ハトリが書いたことや述べたことから多くを学んできた。

第2に、ハトリを追悼することは、単にハトリの功績を振り返ることを超えて、日本における業績測定を見直す好機を提供することにつながると考えたためである。

今や日本においては、政府、自治体、非営利組織の多くが業績測定を実践しており、業績測定は日本の公共分野にすっかり定着した感がある。その一方で、本特集の小野論文が指摘するように、業績測定の実践が「マンネリズム」に陥っていることがしばしば指摘されている。

こうした状況において、ハトリの主張に改めて光を当てることは、業績測定の原点に立ち戻って、日本の業績測定の現在地を確認することにつながる。ただし、それは日本の業績測定を過去に巻き戻すことを意味しない。ハトリが再三強調してきたような、業績測定における重要な原則を守りつつも、現在の組織を取り巻く最新の状況に即した業績測定のあり方の再考を促すことにつながることを期待される。

このような意図の下、本特集は3本の論文から構成されている。

田中論文は、ハトリの主要な著作物を幅広く収集・分析してハトリの業績の全体像を把握した上で、著作物を絞ってその内容を詳細に検討し、業績測定に関してハトリが一貫して重視してきた主張を抽出している。

左近論文は、自身がハトリを含むアメリカの業績測定の専門家に対して数多くのインタビュー調査を実施してきた経験に基づき、業績測定の実践に関してハトリらが強調した論点の中から、8つの要点を抜き出している。さらに、その8つの要点のそれぞれについて、日本の業績測定の現状を照らし合わせて検討をおこなっている。

最後に小野論文は、ハトリを中心とする専門家の意見を参照しながら、日本の業績測定型評価の問題点を明らかにしている。その上で、日本においてもEBPMへの要請が高まりつつあることを踏まえて、エビデンスの提供という観点から、日本の業績測定型評価の現状を検証し、エビデンスの提供において業績測定型評価が果たすべき役割を論じている。

3本の論文の役割をおおまかに整理すれば、田中論文はハトリの示した基準を確認する作業、左近論文は、その基準に照らして日本の業績測定の現在地を確認する作業、そして小野論文は、発展的な視点を導入して業績測定の向かうべき方向を示す作業をそれぞれおこなったといえる。

本特集の意図が成功しているかどうかは、読者の判断に委ねたい。いずれにしても、本特集を契機に、さまざまな分野でハトリの業績や主張が改めて参照され、業績測定のよりよい実践につながっていくことを期待したい。

参考文献

Wholey, J. S., and Hatry, H. P. (1992). The Case for Performance Monitoring. *Public Administration Review*, 52(6), 604-610.

【Research Note】

Harry P. Hatry's Contribution to Performance Measurement: A Review of 55 Years of His Work

Hiraki P. Tanaka

Professor, Faculty of Cultural Policy and Management,
Shizuoka University of Art and Culture
hiraki@suac.ac.jp

Abstract

This study aims to clarify the contributions of Harry P. Hatry (1930-2023), a pioneer in the field of performance measurement. To achieve this goal, it collects and examines the works he published over his 55-year career. Following a comprehensive search, 106 works were identified, of which 28 were selected for closer examination. An analysis of these works revealed general trends in the topics and sectors they addressed, as well as Hatry's ideas and arguments regarding performance measurement. Despite some of the limitations, this study can provide valuable insights that deepen our understanding of Hatry's arguments, which could lead to better implementation of performance measurement.

Keywords

Harry P. Hatry, performance measurement, performance management, program evaluation

1. Introduction

1.1 Introduction to Hatry

Harry P. Hatry (1930-2023) has been a leader in developing performance management and program evaluation procedures for public agencies since 1970s. Due to his significant contributions, he has been referred to as the “father of performance management” (The Urban Institute n. d.).

Hatry was affiliated with the Urban Institute from its establishment in 1968 and, until his death in 2023, led research and practice in the field of performance measurement and program evaluation at the institute. Considering the period during which he was affiliated with the Urban Institute, he was engaged in the field of performance measurement and program evaluation for around 55 years.

He traveled extensively across the United States and even abroad to support the implementation of performance measurement practices in government agencies, local governments, and nonprofit organizations. He also actively published accessible documents that summarized his research findings and insights gained from his practical support for public organizations.

1.2 Purpose and Methodology of This Study

While Hatry's name and individual works are widely known, it cannot be said that his overall achievements have been fully understood. This study seeks to review Hatry's achievements in performance measurement and deepen our understanding of his work. In particular, it focuses on the following question: what were Hatry's views on performance measurement and what did he consider particularly important? Understanding what the "father of performance management" emphasized in terms of performance measurement is considered meaningful for those interested in this field.

In order to answer this question, the following two objectives are set. The first objective is to grasp a comprehensive view of Hatry's achievements. Understanding the fields and themes he engaged with at each stage of his 55-year career is essential to comprehending the significance of his achievements. The second objective is to clarify Hatry's ideas on performance measurement and performance management. This study seeks to reveal Hatry's perspectives on the principles and methodologies of performance measurement.

To achieve these objectives, this study adopts a method of reviewing Hatry's works. Information on his works was collected extensively through the internet, and the bibliographic details and content were investigated and analyzed. The following sections present the findings based on this analysis.

The structure of this paper is as follows. In Chapter two, after examining the relationship between the development of performance measurement and Hatry's activities, an overall picture of his achievements is presented. Chapter three extracts and examines some of Hatry's works in detail, revealing his views and arguments on performance measurement. Chapter four summarizes the findings from Chapters two and three, discusses the significance and limitations of this study, and concludes the paper.

In this paper, the terms "performance measurement" and "performance management" frequently appear. According to Hatry (2014), performance measurement and performance management are defined as follows:

"Performance measurement" is a process in which a governmental or non-governmental public service organization undertakes regular collection of outcome and/or output data (preferably both) throughout the year (not only at the end of the year) for at least many of its programs and services.

"Performance management" refers to the practice of public service managers using performance data to help them make decisions aimed at continually improving services to their customers. (Hatry 2014, p.1)

However, the term "performance measurement" is often used to include the concept of performance management. In this paper, unless a clear distinction is required, "performance measurement" will be used as a concept encompassing performance management.

2. Overview of Hatry's Works on Performance Measurement

2.1 Performance Measurement and Hatry's Activities

According to Bouckaert and Halligan (2008, pp.11-15, 45-49), the development of performance measurement can be broadly divided into three periods.

The first period extends from the early 20th century to the end of World War II. During this period, there was a strong focus on improving the efficiency of government services in the United States. Various federal, state, and local government agencies worked diligently to measure the efficiency of government activities.

The second period spans from the post-World War II era to the 1980s. This period can be defined as the time

when performance measurement expanded its scope from activities that focused on “measurement” to the “management” aspects of budget and program operation.

During this period, the U.S. federal government, driven by the significant expansion of federal expenditures on urban and social policies, sought more rational budgeting methods. Examples include the PPBS (Planning, Programming, and Budgeting System), MBO (Management by Objectives), and ZBB (Zero-Based Budgeting). All of these systems aimed to improve program management in federal agencies, with performance measurement as an integral component. Although the application of PPBS in the federal government was unsuccessful, its principles and components became widely adopted at the local government level. Many state and local governments were inspired by the ideas behind PPBS and later began implementing performance measurement.

The third period, which began in the 1990s and continues to the present day, is notable for the widespread adoption of performance measurement across various levels and sectors of the public domain, and even globally.

The New Public Management (NPM) movement had a significant influence on this trend. One of the primary implications of NPM is the orientation toward the management of public services based on results and outcomes (Hood 1991). Consequently, in performance measurement, there was an increasing emphasis on outcome measurement and the use of these results in managing public services.

Although there have been some new developments since the 2010s, the focus of performance measurement has still remained on measuring outcomes and using these results for managing public services.

Comparing the development of performance measurement with Hatry's career, he was involved in performance measurement research and practice for around 55 years, from the mid-second period (late 1960s) through the third period.

Incidentally, Hatry was involved in PPBS introduced at the Department of Defense under Secretary Robert McNamara in the 1960s, as suggested by several sources¹. Although the details are unclear, it is assumed that Hatry was involved in systems analysis and program evaluation during the implementation of PPBS. It is important to note that He was engaged in systems analysis and program evaluation at the beginning of his career and later became involved in issues related to performance measurement.

2.2 Hatry's Works on Performance Measurement

2.2.1 Classification of Hatry's Works

In order to grasp the overall picture of Hatry's achievements to performance measurement and to clarify his ideas and arguments on the subject, this study adopted a method of reviewing documents authored by Hatry. The documents examined included articles, reports, guidebooks, books, book chapters, newsletters, and blog posts. Searches were conducted using digital libraries and search engines on the Internet to locate writings in which Hatry was the author or co-author².

Using the method described above, a total of 106 documents³, either authored solely by Hatry or co-authored, were identified. While these 106 documents may not cover his entire works, they are believed to encompass the majority of documents currently accessible. Table 1 summarizes the results of classifying these 106 works by year of publication, document type, whether they were single- or co-authored, and the sector they address.

First, with respect to the years of publication, the oldest works were two documents published in 1967. After that, there were 13 works from the 1970s, 14 from the 1980s, 22 from the 1990s, 28 from the 2000s, and 27 from 2010 onward. A relatively larger number of works from the 1990s onward reflects the fact that Hatry's research activities became more vigorous during this period.

When classified by type, reports made up the largest category (40 works), followed by articles (22), book chapters

(20), and books (17). Many of the reports and books were published by the Urban Institute, with a substantial portion of the reports being co-authored with Hatry's colleagues at the institute.

Of the 106 works, 31 were authored solely by Hatry, and 75 were co-authored with one or more individuals. Among the 75 co-authored works, Hatry was the first author in 38.

Hatry's works can be broadly categorized by the sectors they address, including works that target the public sector in general, the federal government or foreign central governments, state and local governments, nonprofit organizations, and specific policy fields.

The largest category was one that addressed the public sector in general (40 works), focusing on the current state of performance measurement and program evaluation. These documents were written for a wide range of readers. The second most common category was works targeting specific fields⁴ (28 documents). Works addressing state and local governments totaled 22. There were eight documents targeting the federal government and nonprofit organizations respectively, which were less than the other categories.

Table 1 Classification of Hatry's Works

Item	Category	# of documents	Item	Category	# of documents
Year of Publication	1960s	2	Document Type	Articles	22
	1970s	13		Books	17
	1980s	14		Book Chapters	20
	1990s	22		Reports	40
	2000s	28		Guidebooks	3
	2010s & Beyond	27		Newsletters	2
Targeted Sectors or Fields	Public Sector in General	40		Blog Posts	2
	Federal Government, Central Government	8	Authorship	Single-author (Hatry only)	31
	State Governments, Local Government	22		Co-authors	75(38)
	Nonprofit Organizations	8		2 authors	25(15)
	Specific Policy Fields	28		3 or more authors	50(23)
			Total		106

Note: The figures in brackets () indicate the number of documents in which Hatry is the first author.

Source: Author

2.2.2 Subjects of The Works

As a result of examining the titles and contents of the 106 works, the subjects were broadly classified into seven categories: Program Evaluation, Program Analysis, Policy Study⁵, Productivity Measurement, Performance Measurement, Performance Management, and Citizen Survey. According to Table 2, the most common subject for works from the 1960s to the 1970s was productivity measurement, with eight works on this theme, followed by four works on citizen survey and three works on program analysis. The focus on citizen surveys reflects the need to understand the effectiveness of public services from the perspective of citizens and clients.

In the 1980s, the predominant subject of Hatry's works was policy study, with five works in this category. However, each of these works addressed different policies and programs, and thus there was little coherence or commonality in their contents. In terms of terminology, while one article published in 1980 (Hatry 1980) used the term "performance measurement" in its title⁶, "productivity measurement" remained the dominant term in most of the works throughout the 1980s.

In the 1990s, the number of works on performance measurement increased to 10, reflecting the fact that the practice of performance measurement in the public sector gained momentum during this period. The three works

classified under the program evaluation category are due to the publication of the first edition of the *Handbook of Practical Program Evaluation* in 1994 (Wholey, Hatry, and Newcomer 1994), which included three of Hatry's works.

The 2000s are characterized by a remarkable increase in works categorized under performance management, with 10 works in this category. This trend seems to reflect the actual expansion of outcome management practices that use outcome measurement results.

In the 2010s and beyond, the trend remained largely the same as in the 2000s. Ten works classified under program evaluation are a result of the *Handbook of Practical Program Evaluation* being revised in 2010 and 2015, with several of Hatry's works included in each edition (Wholey, Hatry, and Newcomer 2010; Newcomer, Hatry, and Wholey 2015).

To summarize, the subjects of Hatry's works shifted over time: from a focus on productivity measurement in the 1970s and 1980s, to performance measurement in the 1990s, and finally to performance management from the 2000s onwards.

Table 2 Subjects Addressed in Hatry's Works

Subject Category	Examples	1960s-1970s	1980s	1990s	2000s	2010s-present	TOTAL
Program Evaluation	evaluation, program evaluation, mini-RCTs	0	2	3	4	10	19
Program Analysis	PPBS, program analysis, policy analysis, program planning	3	2	1	0	4	10
Policy Study	study/examination of specific policy/program/practice	0	5	6	3	1	15
Measurement		8	4	12	9	3	36
Productivity Measurement	measurement, productivity measurement, measurement of effectiveness	8	3	2	0	0	13
Performance Measurement	performance measurement, outcome measurement, performance monitoring, service efforts & accomplishments (SEA)	0	1	10	9	3	23
Performance Management	outcome management, results-based government, managing for results, data-driven performance reviews	0	1	0	10	8	19
Citizen Survey	citizen survey, client survey, customer survey, sample survey	4	0	1	1	1	7
TOTAL		15	14	23	27	27	106

Source: Author

3. Analysis of Hatry's Arguments

3.1 Method of Selecting the Works for Close Examination

The second objective of this study is to clarify Hatry's ideas and arguments regarding performance measurement. To achieve this objective, this study selected works from Hatry's above-mentioned 106 documents that were considered to clearly reflect his ideas and arguments, and examined their contents in detail. The works selected for examination were mainly single-authored works by Hatry. Several co-authored works, those with a small number of co-authors and in which he was the first author, were selected⁷. The reason for selecting such works is that they are considered to more clearly reflect Hatry's individual perspectives⁸. Although the focus was on works related to performance measurement, the ones addressing program evaluation and program analysis were also included, in order to verify whether Hatry's ideas and arguments were consistent across different subjects.

A total of 28 works were selected and the list of the works is shown in Table 3. The results of a detailed examination

Table 3 Selected Documents from Hatry's Works for Close Examination

No	Authors	Published Year	Title	Document Type	Subject	# of citation (as of Sep.14, 2024)
1	Hatry	1967	Criteria for Evaluation in Planning State and Local Programs	report	program analysis	43
2	Hatry	1970	Measuring the Effectiveness of Nondefense Public Programs	article	productivity measurement	23
3	Hatry	1971	Status of PPBS in Local and State Governments in the United States	article	program analysis	10
4	Hatry	1972a	Issues in Productivity Measurement for Local Governments	article	productivity measurement	63
5	Hatry	1972b	Reflecting The Consumer Viewpoint in State and Local Government Fiscal and Expenditure Decisions	article	citizen survey	7
6	Hatry	1976	Approaches to Productivity Measurement and Program Evaluation	article	productivity measurement	N.A.
7	Hatry	1978	The Status of Productivity Measurement in the Public Sector	article	productivity measurement	212
8	Hatry	1980a	Performance Measurement Principles and Techniques: An Overview for Local Government	article	performance measurement	196
9	Hatry	1980b	10. Pitfalls of Evaluation	book chapter	evaluation	N.A.
10	Hatry	1984	The Status of Productivity Measurement in the Public Sector: an Update	report	productivity measurement	N.A.
11	Wholey, Hatry	1992	The Case for Performance Monitoring	article	performance measurement	574
12	Hatry	1997	Where the Rubber Meets the Road: Performance Measurement for State and Local Public Agencies	article	performance measurement	76
13	Hatry	1999	Performance Measurement: Getting Results [1st Ed.]	book	performance measurement	(1800)
14	Hatry	2002	Performance Measurement: Fashions and Fallacies	article	performance measurement	254
15	Hatry, Cowan, Hendricks	2004a	Analyzing Outcome Information: Getting the Most from Data	report	performance measurement	N.A.
16	Hatry, Newcomer	2004b	19. Pitfalls of Evaluation	book chapter	evaluation	N.A.
17	Hatry	2004c	Public and Private Agencies Need to Manage for Results, Not Just Measure Them	newsletter	performance management	N.A.
18	Hatry	2006	Performance Measurement: Getting Results [2nd Ed.]	book	performance measurement	1800
19	Hatry	2008	GOVERNING FOR RESULTS: IMPROVING FEDERAL GOVERNMENT PERFORMANCE AND ACCOUNTABILITY: SUGGESTIONS FOR THE NEW FEDERAL ADMINISTRATION	newsletter	performance management	4
20	Hatry	2010	Looking into the Crystal Ball: Performance Management over the Next Decade	article	performance management	57
21	Hatry, Davies	2011a	A Guide to Data-Driven Performance Reviews	report	performance management	53
22	Hatry	2011b	Measuring government performance is important but not easy	blog post	performance measurement	N.A.
23	Hatry	2013	Sorting the Relationships Among Performance Measurement, Program Evaluation, and Performance Management	article	performance management	93
24	Hatry	2014	Transforming Performance Measurement for the 21st Century	report	performance measurement	65
25	Hatry	2016	Using Performance Data: Regular Data-Initiated Reviews	blog post	performance management	N.A.
26	Hatry	2020	Evaluation Guide for Public Service Program Managers	report	evaluation	5
27	Hatry, Katz, Barrett, Greene	2022a	Dos and Don'ts: Tips for Strengthening Your Performance Management Systems	report	performance management	4
28	Hatry, Katz	2022b	Mini-RCTs: A Tool for Program Managers to Help Identify What Works	report	evaluation	1

Note: The figures in the far-right column are the number of citations by google scholar (as of September 14, 2024).

Source: Author

of these 28 works are discussed below.

3.2 Hatry's Arguments on Performance Measurement

A detailed examination was conducted on 28 selected works. For each work, methodological principles and advice presented by Hatry concerning performance measurement and related issues were extracted as comprehensively as possible. Among these, those that appeared repeatedly and were emphasized, regardless of the publication year of the work, were selected. The fact that these views are expressed consistently across different periods suggests that Hatry regarded them as particularly important.

As a result of above examination, several ideas and arguments have been presented throughout his long career. The following six key arguments were identified as highly consistent across his works.

(1) Emphasis on the need to set a variety of indicators

Hatry consistently argues for the need to set and track a variety of indicators in performance measurement. This argument is already evident in his early works. In Hatry (1967), which discusses PPBS, he states that all relevant criteria related to the program's objectives should be used in program analysis (Hatry 1967, p.6). In his works published in the 1970s and the 1980s, the need to use a variety of measures is repeatedly emphasized. In Hatry (1999), he cites "comprehensiveness" as one of the criteria for setting outcome indicators, and notes, "The set of indicators should include outcomes that identify possible negative or detrimental effects" (Hatry 1999, p.58). In more recent works, he argues that even if an outcome indicator is one over which the government has little influence, it should still be tracked if the government is responsible for the results (Hatry 2014).

The reasons Hatry emphasizes the use of diverse indicators are summarized as follows. First, it is difficult to fully capture a program's effects using only a single indicator. The effects of a program manifest in various aspects and unfold in stages. Additionally, since it is often challenging to quantitatively measure the outcomes of a program, he believes that relying on a single indicator would be insufficient for understanding the full scope of the program's results.

Second, an insufficient group of indicators may induce perverse actions. Perverse actions refer to when the established indicators incentivize undesirable behavior among public officials. This issue was already highlighted in an early article discussing productivity measurement (Hatry 1972a). He also states that if there is a possibility of negative side effects occurring as a result of implementing a program, it is necessary to establish indicators to monitor those effects (Hatry 1972b).

Hatry, when advocating for using diverse indicators, primarily refers to those that focus on aspects of program outcomes. However, this does not imply that he believes solely outcome indicators should be established. Since the government's ability to influence outcomes is limited, Hatry insists that both outcome and output indicators should be tracked (Hatry 2022a).

(2) Emphasis on indicators that measure effects on citizens and clients

Hatry also stresses the need to set indicators that measure the effects on citizens and clients in performance measurement. As discussed in (1), he emphasizes the necessity of setting diverse indicators, but among those indicators, he places particular importance on those that measure the effects from the citizens' and clients' viewpoints.

The effects on citizens and clients refer to how the conditions of public service recipients and program participants have changed, as well as how they have evaluated the services and programs provided. To grasp these aspects, it is necessary to conduct citizen or client surveys, for which Hatry offers practical advice on their implementation (Hatry 1972b, 1976, 1978, 1997, 2002, 2014).

In relation to citizen surveys, Hatry points out several important considerations. First, he is not keen on measuring the “satisfaction” of citizens and clients. Since the goal of public services and programs is not to increase satisfaction, he believes it is more important to collect information on improvements in the “conditions” of citizens and clients through citizen surveys (Hatry 2002).

Second, Hatry argues that it is not necessary to seek a high level of statistical reliability in citizen surveys. Surveys conducted on large sample sizes yield more reliable results, but they are costly, and it is not feasible to conduct such surveys frequently. He suggests that it is preferable to conduct more frequent citizen surveys with modest sample sizes, even if the reliability of the results is somewhat compromised.

Third, Hatry emphasizes the need to include not only current recipients of services or participants in programs but also those who have previously used the services or participated in the programs, as survey targets. He sees the tracking of these individuals and conducting follow-up surveys after they have left the program as part of “aftercare” (Hatry 2014).

(3) Emphasis on using “breakouts”

Hatry insists that outcome indicators should not only be aggregated but also disaggregated by target group attributes or service type for analysis. He refers to this as “breakouts” or “disaggregation.” This idea has been repeatedly expressed in his works, from the 1960s to his most recent writings.

In Hatry (1967), for example, he proposes focusing on specific “target groups” to understand the program’s effects on those groups. In Hatry (1970), he uses the term “subgroup” to discuss the need for analysis at the subgroup level. The term “breakouts” first appears in Hatry (1997). In Hatry (1999), he devotes an entire chapter⁹ to discussing the importance and usefulness of “breakouts.”

Hatry emphasizes the use of “breakouts” according to the following reasons. Firstly, in order to check whether services are being provided equitably among the citizens and clients concerned, it is necessary to understand the status of service provision and the realization of effects for each target group, not just to look at the aggregate numbers. He refers to this consideration as an “equity concern” (Hatry 2014).

Secondly, using “breakouts” is one of the primary methods for analyzing performance information. For Hatry, analysis in performance measurement essentially means making some form of “comparison.” The comparison methods he suggests include time-series comparisons, comparisons between actual and target values, ones between departments or with other organizations (benchmarking), ones by client characteristics, and ones by service characteristics (Hatry 1972b, 1999, 2006, 2014). To carry out comparisons by client characteristics or by service types, it is necessary to use performance information that has been disaggregated into smaller categories.

(4) Emphasis on improving services rather than fulfilling accountability

Hatry places greater emphasis on improving services and programs, rather than using performance measurement to fulfill the accountability of government agencies and nonprofit organizations. This argument has become more clearly articulated in Hatry’s works since the 1990s.

From the 1960s to the 1980s, before the U.S. public sector was exposed to the influence of NPM, many of the Hatry’s works focused on themes such as PPBS and productivity measurement. For this reason, references to accountability rarely appear in his writings from this period.

In the 1990s, the focus of measurement shifted to outcome measurement, and performance results started being used by government agencies to fulfill their accountability obligations. In response to this development, Hatry expressed the view that performance measurement should be used not only for external accountability, but also for program improvement (Hatry 1992).

In his works from the 2000s onwards, Hatry repeatedly emphasizes that outcome information should be used to improve services. For example, he argues that using outcome information to improve services represents a shift from “outcome measurement” to “outcome management” (Hatry 2014). Furthermore, he recommends its use in a variety of practical applications, which include performance-based pay, employee recognition, performance contracts, and data-driven performance reviews (Hatry 2008, 2014).

(5) The need to estimate future effects

As pointed out in (4), Hatry consistently holds the view that performance measurement should be used for service improvement. In many of his works, however, he also claims that while performance measurement and program evaluation are “about the past,” decision-making regarding services and programs is “about the future.”

Therefore, he maintains that it is necessary not only to carefully analyze past measurement and program evaluation results but also to estimate the future effects¹⁰ of programs. This argument is first presented in Hatry (1970), where he states that program effects should be considered not only for the next fiscal year but also for subsequent fiscal years. In Hatry (1999), published much later, he also states that if the effects of a program proposed in the budget formulation process do not appear within the next fiscal year, the effects for future years should be considered in the process. Similar points are made frequently in his works from the 2000s onwards (Hatry 2002, 2008, 2010).

Hatry emphasizes the need to estimate the future effects of programs because he recognizes that the results of performance measurement or program evaluation are used as a basis for formulating budget request for the following fiscal year. The effects of programs do not always come out in the following year, and even if they do, the effects may continue to remain in subsequent years. Thus, when deciding whether to adopt a program and how much to allocate for it in the following fiscal year, Hatry believes it is necessary to make decisions not only based on past information but also by estimating the effects expected in the coming years and taking their magnitude into account.

Measuring and analyzing the past performance of a program and predicting the future effectiveness of a program require different methods. However, Hatry does not provide specific methodologies for estimating future effects in his works examined in this study. At least, the performance information obtained as a result of performance measurement can be the basis for estimating the future effectiveness of the program. Hatry expects that more sophisticated methods for doing so will become available due to advances in information technology in the near future (Hatry 2010).

(6) Complementary relationship between performance measurement and program evaluation

Hatry's recognition that the relationship between performance measurement and program evaluation is complementary, remained consistent throughout his career.

In his early works, Hatry argues that “to determine the extent to which an agency is responsible for observed changes in performance measures, more in-depth program evaluation and analysis are needed” (Hatry 1980a, p.338). In this sense, performance measurement is complemented by program evaluation.

In later works, he also suggests that “Outcome measurement can be a substantial aid to in-depth program evaluation” (Hatry 1997, p.42), indicating that performance measurement can complement program evaluation. In the same document, he asserts, “The position of this author is that program evaluation and performance measurement are complementary” (Hatry 1997, p.42).

This kind of discussion leads to the following question: Which did Hatry think was more important: performance measurement or program evaluation? On this point, he stated that “Performance measurement can be considered part of program evaluation” (Hatry 1997, p.41), suggesting that he viewed performance measurement as a component of program evaluation. However, this is merely an indication of the taxonomic relationship between the methods, and it is not appropriate to understand that he considered program evaluation to be the primary focus.

He argues that program evaluation is expensive to carry out and requires specialized skills, limiting its implementation, whereas performance measurement can be implemented consistently and cover a wide range of government activities. Based on this practical understanding, he insists that more focus should be placed on performance measurement in the management of public organizations (Hatry 1997). In fact, he suggests conducting performance measurement as the primary approach, with program evaluation being implemented as a supplementary measure. He further notes that it is more practical to implement quasi-controlled approaches rather than controlled experiments (Hatry 1976)¹¹.

From 2014 onwards, mini-RCTs (or mini-experiments) began to appear in his works. In these writings, Hatry proposes to implement mini-RCTs while primarily conducting performance measurement (Hatry 2014, 2020, 2022b). He recommends conducting mini-RCTs not to clarify the causal relationship of program effects, but as a convenient method for obtaining insights into the effectiveness of alternative service proposals.

4. Summary and Conclusion

4.1 Summary of Results

4.1.1 Overview of Hatry's Contributions

This study aimed to gain a comprehensive view of Hatry's achievements by collecting and categorizing a wide range of his works. After reviewing 106 of his publications, the following points have become clear.

First, the publication years of the works span from the late 1960s to the early 2020s, and a significant increase in the number of publications was observed after the 1990s. Second, as for the type of works, reports constitute the largest portion of his works, followed by book chapters, articles, and books. Regardless of the type, the majority of the works deal with the practical considerations of performance measurement and program evaluation for public sector practitioners. Third, although performance measurement and performance management are the primary subjects of his works, there are also a certain number of works focused on program evaluation, specific policies, and program analysis.

4.1.2 Hatry's Arguments on Performance Measurement

This study closely examined 28 of Hatry's works, focusing primarily on his single-authored publications, to analyze his arguments on performance measurement. Six key points consistently emphasized by Hatry throughout his career were identified. These arguments were repeatedly presented from the early stages of Hatry's career to his more recent works, with little substantial change in content.

4.2 Conclusion

Before concluding this paper, some of the limitations of this study should be mentioned. First, because this study focused solely on Hatry's published works, it was not possible to fully assess his achievements that were not reflected in these publications. The advice he provided directly to government officials in practical settings, as well as the content of his lectures and training sessions, were not considered¹². Hatry's expertise may have been demonstrated in such activities, but it is expected that many of the issues he considered particularly important are reflected in his major works which this study examined.

Second, this study was not able to thoroughly examine most of his co-authored works. As a result, many of the reports co-authored with his colleagues at the Urban Institute were excluded from the analysis. Given the objectives

of this study, however, it seems appropriate to focus primarily on his single-authored works.

Despite these limitations, this study can provide valuable insights that deepen our understandings of Hatry's arguments, which could lead to better implementation of performance measurement¹³.

Notes

- 1 In the preface of Hatry (1999), Hatry does not explicitly state his involvement, but he explains in detail the circumstances under which PPBS was introduced to the Department of Defense in the 1960s. An obituary published online (Barrett and Greene Inc. 2023) to report Hatry's passing mentions that Hatry was involved in the management reform led by Secretary McNamara at the Department of Defense in the 1960s as one of "McNamara's Whiz Kids."
- 2 Searches were conducted using JSTOR, the Urban Institute's website, and Google Scholar.
- 3 Different editions of a document were counted as separate entries.
- 4 Specific fields include education, police departments, healthcare services, economic development, and museum grants.
- 5 Some works focused on specific policies, programs, or practices, conducting detailed investigations and summarizing the results. These were grouped under the category of "Policy Studies."
- 6 The earliest use of the term "performance measurement" in Hatry's works appeared in Hatry (1972).
- 7 Although Wholey and Hatry (1992) was not authored by Hatry as the lead author, it was included for analysis due to its significant influence (it has 574 citations according to Google Scholar as of September 14, 2024).
- 8 In addition to the above conditions, the number of citations of the work and the need to secure a certain number of works for each decade were also taken into consideration.
- 9 Chapter 8: Making Outcome Information Useful: Providing Indicator Breakouts.
- 10 Hatry does not provide a clear definition of "effect" in his work, but based on the context, it can be inferred that he uses this term to refer to the direct changes expected to occur as a result of implementing a program.
- 11 Hatry uses the term "controlled experiments" to refer to randomized controlled trials (RCTs), and calls the method of comparing groups without randomization the "quasi-controlled approach".
- 12 Such activities are usually based on contracts and therefore their details are not usually made public. Attempts were made in this study to obtain information on such activities, but sufficient information could not be obtained.
- 13 The arguments on performance measurement that this study has extracted from Hatry's work can also serve as a valuable reference for many people involved in performance measurement in Japan. However, as Hatry's assertions may not be fully applicable to the Japanese context, it will be necessary to reinterpret them in the light of Japan's current circumstances.

References

- Barrett & Greene Inc. (2023). *Hatry, H. P., Leader and Performance Management Pioneer, dies at 92*. Blog Post. Retrieved December 1, 2023, from <https://www.greenebarrett.com/post/harry-hatry-leader-performance-management-pioneer-dies-at-92>
- Bouckaert, G. (2002). Review: Performance Measurement: Getting Results. *Public Performance & Management Review*, 25(3), 329-335.
- Bouckaert, G. and Halligan, J. (2008). *Managing Performance: International Comparisons*. New York: Routledge.
- Hatry, H.P. (1967). Criteria for Evaluation in Planning State and Local Programs. In Committee on Government Operations, United States, *Criteria for Evaluation in Planning State and Local Programs*. Washington, D.C.: U.S. Government Printing Office, 1-42.
- Hatry, H. P. (1970). Measuring the Effectiveness of Nondefense Public Programs. *Operations Research*, 18(5), 772-784.
- Hatry, H. P. (1971). Status of PPBS in Local and State Governments in the United States. *Policy Sciences*, 2(2), 177-189.

- Hatry, H. P. (1972a). Issues in Productivity Measurement for Local Governments. *Public Administration Review*, 32(6), 776-784.
- Hatry, H. P. (1972b). Reflecting The Consumer Viewpoint in State and Local Government Fiscal and Expenditure Decisions. *National Tax Journal*, 25(3), 363-368.
- Hatry, H. P. (1976). Approaches to Productivity Measurement and Program Evaluation. *Public Productivity Review*, 1(3), 21-28.
- Hatry, H. P. (1978). The Status of Productivity Measurement in the Public Sector. *Public Administration Review*, 38(1), 28-33.
- Hatry, H. P. (1980a). Performance Measurement Principles and Techniques: An Overview for Local Government. *Public Productivity Review*, 4(4), 312-339.
- Hatry, H. P. (1980b). 10. Pitfalls of Evaluation. In Majone, G., & Quade, E. S. (Eds). *Pitfalls of Analysis*, New York: John Wiley & Sons, 159-178.
- Hatry, H. P. (1984). The Status of Productivity Measurement in the Public Sector: an Update. In Windle, C. (Ed). *Program Performance Measurement: Demands, Technology, and Dangers*. Washington, D.C.: U.S. Government Printing Office, 37-50.
- Hatry, H. P. (1997). Where the Rubber Meets the Road: Performance Measurement for State and Local Public Agencies. *New Directions for Evaluation*, 1997(75), 1-98.
- Hatry, H. P. (1999). *Performance Measurement: Getting Results*. Washington, D.C.: the Urban Institute.
- Hatry, H. P. (2002). Performance Measurement: Fashions and Fallacies. *Public Performance & Management Review*, 25(4), 352-358.
- Hatry, H. P., Cowan, J., and Hendricks, M. (2004a). *Analyzing Outcome Information: Getting the Most from Data*. Washington, D.C.: the Urban Institute.
- Hatry, H. P., and Newcomer, K. E. (2004b). 19. Pitfalls of Evaluation. In Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds). *Handbook of Practical Program Evaluation 2nd edition*. San Francisco: CA Jossey-Bass, 547-570.
- Hatry, H. P. (2004c). *Public and Private Agencies Need to Manage for Results, Not Just Measure Them*. Washington, D.C.: the Urban Institute. Retrieved September 16, 2024, from <https://www.urban.org/research/publication/public-and-private-agencies-need-manage-results-not-just-measure-them>
- Hatry, H. P. (2006). *Performance Measurement: Getting Results 2nd edition*. Washington, D.C.: the Urban Institute.
- Hatry, H. P. (2008). *Governing for Results: Improving Federal Government Performance and Accountability: Suggestions for The New Federal Administration*. A Paper prepared for the Government Performance Workshop held Oct.15, 2008.
- Hatry, H. P. (2010). Looking into the Crystal Ball: Performance Management over the Next Decade. *Public Administration Review*, 70(Supplement), 208-211.
- Hatry, H. P., and Davies, E. (2011a). *A Guide to Data-Driven Performance Reviews*. IBM Center for The Business of Government, Improving Performance Series, IBM Center for The Business of Government.
- Hatry, H. P. (2011b). *Measuring government performance is important but not easy*. Urban Wire. Retrieved September 16, 2024, from <https://www.urban.org/urban-wire/measuring-government-performance-important-not-easy>
- Hatry, H. P. (2013). Sorting the Relationships Among Performance Measurement, Program Evaluation, and Performance Management. *New Directions for Evaluation*, 2013(137), 19-32.
- Hatry, H. P. (2014). *Transforming Performance Measurement for the 21st Century*. Washington, D.C.: the Urban Institute.
- Hatry, H. P. (2016). *Using Performance Data: Regular Data-Initiated Reviews*. Blog Post, ICMA. Retrieved September 16, 2024, from <https://icma.org/blog-posts/using-performance-data-regular-data-initiated-reviews>
- Hatry, H. P. (2020). *Evaluation Guide for Public Service Program Managers*. Washington, D.C.: the Urban Institute.
- Hatry, H. P., Katz, B., Barrett, K., and Greene, R. (2022a). *Dos and Don'ts: Tips for Strengthening Your Performance Management Systems*. Washington, D.C.: the Urban Institute.
- Hatry, H. P., and Katz, B. (2022b). *Mini-RCTs: A Tool for Program Managers to Help Identify What Works*. Washington, D.C.: the Urban Institute.
- Hood, C. (1991). A Public Management For All Seasons? *Public Administration*, 69, 3-19.

- The Urban Institute. (n.d.). *Harry P. Hatry*. Retrieves September 16, 2024, from <https://www.urban.org/author/harry-p-hatry>
- Van Dooren, W. (2011). Better Performance Management: Some Single- and Double-Loop Strategies. *Public Performance & Management Review*, 34(3), 420-433.
- Wholey, J. S., and Hatry, H. P. (1992). The Case for Performance Monitoring. *Public Administration Review*, 52(6), 604-610.
- Wholey, J. S. (1999). Performance-Based Management: Responding to the Challenges. *Public Productivity & Management Review*, 22(3), 288-307.
- Wholey, J. S., Hatry, H. P., and Newcomer, K. E. (Eds). (1994). *Handbook of Practical Program Evaluation*. San Francisco: CA Jossey-Bass.
- Wholey, J. S., Hatry, H. P., and Newcomer, K. E. (Eds). (2004). *Handbook of Practical Program Evaluation 2nd edition*. San Francisco: CA Jossey-Bass.
- Wholey, J. S., Hatry, H. P., and Newcomer, K. E. (Eds). (2010). *Handbook of Practical Program Evaluation 3rd edition*. San Francisco: CA Jossey-Bass.
- Newcomer, K. E., Hatry, H. P., and Wholey, J. S. (Eds). (2015). *Handbook of Practical Program Evaluation 4th edition*. San Francisco: CA Jossey-Bass.

(Accepted 27th December, 2024)

【研究ノート】

米国の業績測定の実践から我が国取組が学ぶべきこと —日本の行政評価の取組を振り返る—

左近 靖博

株式会社浜銀総合研究所（地域戦略研究部）

y-sakon@yokohama-ri.co.jp

要 約

米国における業績測定は、成果志向の経営（Managing for Results）の考え方が根幹にあり、測定した実績値の分析と、改善・見直しの意思決定の仕組みが一体化して運用されている。この米国の取組は、行政に対する透明性確保やアカウンタビリティ（説明責任）の実現を意図して導入が進んだ我が国の行政評価の取組の今後の見直しの参考になる。本稿では、米国の実践から日本の業績測定の見直しの方向性をハリー・P・ハトリ氏へのインタビューより整理する。

キーワード

業績測定、ドリルダウン分析、戦略的な計画、スタットモデル、成果志向の経営（MGR）

1. はじめに

日本の行政評価の起点は、三重県において1995年度に始められた「さわやか運動」の根幹として導入された「事務事業評価システム」と知られている。以降、その影響を受け公共部門に広く普及し、また国においては2002年の「行政機関が行う政策の評価に関する法律（以下、「政策評価法」と言う）」施行により制度化されている。この三重県の「事務事業評価システム」の導入に理論的影響を与えたのは、オズボーン（David Osborne）とゲーブラー（Ted Gaebler）の共著「行政革命」と言われており、米国自治体においての成果志向の経営（Managing for Results）の考え方が背景にある（窪田 1998、p.18）。また、当時の日本では、予算の不適切使用（カラ出張）や官官接待のような行政機関の不祥事、無駄で多額の公共事業投資、

財政赤字などの問題が注目され、公金の無駄遣いや行政運営の非効率性が批判されており、行政内部や政策立案、運営の抜本的見直しが喫緊の課題とされた時代であった。そのような中で注目されたのが、行政に対する透明性確保やアカウンタビリティ（説明責任）実現を意図した運営見直しの仕組みとしての行政評価であった。

本稿では、このように米国の取組に端を発し、当時の日本の状況を背景に独自進化をした日本の行政評価、その中心である業績測定（業績達成度評価）について、日本の取組にも多大な影響を与えたハリー・P・ハトリ氏や、在米の有識者、実務担当者への筆者らによるヒアリング調査の結果を整理し、米国での実践と比較して、日本の取組において学ぶべきことを主に実務担当者向けの視点から示す。

2. ハリー・P・ハトリー氏へのインタビューを振り返る

筆者は民間シンクタンクの研究員として、主として2000年代以降、我が国の行政評価導入支援を行うために国、自治体のプロジェクトに数多く参画して、米国、カナダ、英国、仏国、独国などの主要先進国の取組を調査し、その過程で現地調査にて実務担当者や研究者へのインタビューを多数実施した。その際にハリー・P・ハトリー氏とは、2005年、2008年、二度、ワシントンD.C.のアーバンインスティテュートにて、主に実務の観点から業績測定の実状と在り方についての話を聞く貴重な機会を得た（いずれも国土交通省受託調査による。なお、報告書は公開されていない。その他に総務省、会計検査院などでの米国連邦政府、地方政府・自治体への調査実績がある）。この時のインタビューを振り返り改めて整理すると、業績測定の実務・実践において以下のような8つの要点を得た。（『』に発言要旨として記載）

（1）詳細分析（ドリルダウン分析）

『業績測定は、目標達成度の評価が基本であるが、それを基にして達成要因を分析し、見直しを進めることが肝要』である。そのためには、集計したデータの階層レベルを掘り下げていき集計項目を参照し詳細分析を行う、「ドリルダウン分析」が求められる。目標達成状況は、特定の地域や属性、対象に偏っていないか、その理由は何かを検討することで、データをより正確に分析でき、適切な見直し方策の検討に活用できる。『実績値は必ずドリルダウンによって詳細に分析する必要』がある。

（2）未達成部分の課題の分析と特定

『目標達成度は業績測定の基本的な視点であるが、分析においては未達成の部分にも着目すること』が求められる。未達成の部分には、プログラムの問題点や課題が内在している他、100%の達成目標を掲げるような場合には、未達があること自体がより大きな課題である場合もあるからである。なお、それらの課題は業績測定から結論や問題が自ずと導き出されるものではなく、仮説に基

づいて別の方法による検討が必要となる。

（3）戦略的な計画と成果目標

『公共部門の業績測定は、アウトカム目標の設定と、その前提となる戦略計画が基本的な構成』となる。計画は、期限を設定して、解決しなければならない優先的な目標に限定した、体系的かつロジカルな戦略的な構成が、成果志向の経営の観点からは望ましい。『プログラムが成果を実現するストーリーの仮説であるロジックモデルがその中核』である。『業績測定は、地域や社会の課題を解決するための成果実現に向けた進捗を計測するための経営のツール』である。

（4）意思決定の仕組みとリーダーシップ

『業績測定は、データのモニタリング、分析・評価により目標達成度と改善と見直しの仮説を導出することが役割であるが、公共部門ではそれをどのように意思決定に反映するのか、という仕組み作りも併せて重要』である。ハリー・P・ハトリー氏は、ボルティモア市のスタットモデルを例示した。定期的に業績測定を行い、組織の長が自らのリーダーシップにより、その結果を基にして意思決定をして、改善見直しを継続して取組む仕組みがボルティモア市では構築されている。

業績測定をどのように実務において活用するのかの議論は、米国でも2000年以降のことである。ボルティモア市の例は、業績測定の結果活用において、「定期的な会議体設置」と「リーダーシップ」という、業績測定の結果活用に求められる2つの前提要素の重要性を示唆している（連邦運輸省、住宅都市開発省職員らも同様に指摘）。

（5）プログラム評価による詳細分析

—業績測定との役割分担—

『業績測定は設定した目標の達成度を示すことはできるが、成果が真にプログラムの効果なのか、どのような要因が影響して、達成されたのかを示すことは困難』である。業績測定は実務的には簡易なツールであり、成果の現状を客観的に示すことは可能であるが、より詳細にプログラムの効果の貢献度合いを分析・検証するには、プログラム評価を一体として実施することが望ましい。

米国連邦政府の政府業績成果法（Government Performance and Results Act: GPRA法）は、成果志向の政府を構築するための基盤となる法律で、各省においての戦略計画策定と業績測定を中核にしているが、関連して実施したプログラム評価について記載することになっており、相互に補完的な役割を果たすことが求められている。

(6) 第三者による評価

『業績測定は公共部門の自己評価を基本とするが、米国では施策などの効果を客観的に検証するプログラム評価の多くは、シンクタンクや研究機関、NPOなどの第三者による評価実施を基本』としている。インタビューを踏まえると、それは専門性と客観性を担保するという考え方であり、また、ここワシントンD.C.には評価実務を担える数多くの機関がある、とのことである。米国では、評価実務は市場化されている（ジョージワシントン大学Kathryn Newcomer教授らも同様に指摘）。

(7) 最低限度の統計的な知見

『分析するデータの取り扱いには、最低限度の統計的な知見が必要』である。ここで言う知見とは、高度の分析手法の習得ではなく、数字を分析する際の最低限度の知識である。母集団とサンプル数、複数視点、経年分析、類似対象との比較、平均の罨、相関と因果についての理解、また公共政策の効果分析の場合には必ず外部要因の影響について踏まえる必要があり、データ単体のみからの厳密な効果検証には限界がある。データは客観的状況を示すが、読み方の誤りは結論導出に間違いを引き起こす可能性もある（ウィリアム・アンド・メアリー大学 John B. Gilmour教授も同様に指摘）。

(8) 市民への積極的な説明

『業績測定は成果志向の経営を実現するツールであるとともに、納税者に対する説明にも積極的に活用』できる。『米国では多くの地方政府において、成果情報と併せて次年度予算について市民向けの説明会を開催しており、多くの団体で制度化』されている。それは行政機関の担当者による市民向けの成果と予算の説明の場であり、市民か

らの意見や要望の収集機会でもある。なお、このBudget Public Hearingは、コロナ渦でもオンラインで開催されていた。カンサスシティ市、フェニックス市など多数。

3. 米国業績測定の実践からみた日本の現状

(1) 日米の業績測定の実践の対比

以下では、ハリー・P・ハトリ氏らが指摘する上記の業績測定の実践における8つの要点を参考にしながら、米国の取組との対比で日本の状況を整理し、業績測定の見直しに向けての方向性や在り方を示す。なお、インタビュー時期からやや時間は経過しているが、各要点については、現状もなお、我が国取組の改善において参考となり得る。

①詳細分析（ドリルダウン分析）

我が国においての業績測定の実践は、多くの団体で実施されている業績測定をベースにした事務事業評価である（総務省 2017, p.1）。一般的には定型的な共通の評価シートにおいて、設定した一つまたは複数の指標の目標値と実績値、達成度分析と見直しなどにより整理されている。フォーマット化されていることで、担当者はそれを記載することにのみ注力する傾向がある。この達成状況の測定からの見直しにおいて、詳細分析（ドリルダウン分析）を行っている例は、総務省（2017, p.5）などの調査では現状確認はできないが、少なくとも評価書においては、あまり多くはないのではないか。そのため、データ分析に基づく見直しの意思決定の水準が不十分である可能性がある。多くの実務担当者は、指標の適切性やデータ収集の問題を指摘するが、実績値をどのように分析するのかという点には十分に配慮されていないと思われる。

例えば就業支援プログラムにおいて「修了者の就職率」を達成度を測る単一の指標とした分析の場合、ドリルダウン分析により、高卒者と大卒者、性別などにより偏りがあった場合にはプログラムの内容自体に問題があり、必要に応じて対象別のカリキュラム見直しが求められたのではないかと推定できる。この場合「事業の業績指標の達成度

は全体で85%であり、経年的に改善しており順調」と結論してしまうと、プログラムが各ターゲット層に対して真に有効ではない可能性があると思定される、などである。以下、ドリルダウン分析の視点の例を示す。

表1 ドリルダウン分析の視点

基本属性	年齢層、性別、職業、学歴
地域	居住地区、地域特性
その他	経験・経歴、志向

(出所) 筆者作成

②未達成部分の課題の分析と特定

未達成の要因や背景分析に着目している例も、同様に総務省（2017、p.5）などの調査では現状確認はできないが、あまり多くはないのではないか。上記の就業支援プログラムの他、例えば乳幼児訪問や乳児健診の受診率は93%以上を経年的に維持しているような場合、それを数値のみに着目して前向きに「順調である」と評価した場合、一方で残る7%の層に、まったく健診を受診しない家庭や子どもがいることも想定される。その背景には育児放棄や家庭環境に問題があるなど、より積極的な公的支援が必要にも関わらず、その状況を放置することになる。

表2 未達成の要因に着目すべき業績指標例

受診率、利用率・参加率（未利用）、定員充足率、理解度・満足度、普及率・カバー率、到達度、申請率（未申請者）

(出所) 筆者作成

上記に未達成に着目すべき業績指標の例を示す。業績測定の基本的な視点は達成度であることを踏まえると、多くの指標が未達部分の分析にも着目すべきであると言える。

③戦略的な計画と成果目標

我が国の自治体では、政策や施策を評価対象にするなど、総合計画を基点として政策体系を構築して、目標設定して業績測定を行う場合が多い（総務省 2017、p.5；日本生産性本部 2024、p.5）。総合計画は、基本構想策定義務付けの規定が1969年

に地方自治法に設けられた際、自治体がめざす将来像と将来の目標を明らかにする基本構想を頂点に、基本計画、実施計画を要素とする総合計画の標準的な構成・内容が示されて以降、総合的かつ計画的な行政の運営をすすめる中核的柱として位置付けられ、その結果として多くの団体に総花的で似たような計画になっている。

平成23年8月1日に地方自治法の一部を改正する法律により、基本構想の策定を義務付けていた規定が廃止されたが、現在も多くの団体において継続して計画が策定され、また構成もあまり変化していない（日本生産性本部 2024、p.4）。そのため、制度の義務付けが廃止された現在でも、限定された目標に注力した戦略的な計画はあまり多くはない。また、総合計画の期間と首長任期を連動させて、首長の政策を反映した戦略的な総合計画とする動きも一部にはあったが、現状「総合計画と首長の任期の年限が一致している」のは、8.8%と、それらは必ずしも取組の主流ではない（日本生産性本部 2024、p.17）。背景には、「自治体のめざす将来像と目標を示す基本構想を頂点とするため全市的な視点となり、分野・課題が網羅的な体系になりやすい」「その結果、計画が政策分野別の分類や整理になっている」や、「企画部門策定の計画とはいえ、総合計画に担当課の施策の記載がないと、全市的な計画のため事業がすすめられない、予算が付かないという行政担当者の意識」などと想定される。

ハリー・P・ハトリ氏は、『実務において業績測定は成果を実現する経営のツールである』と言う。公共部門の成果志向の経営という視点から、中期もしくは短期に解決する課題や成果目標を前提とした、仮説をベースにしたロジカルで戦略的な計画が前提として必要となる。地方自治体において、各団体の地域、社会の状況、財政状況は多様であり、似たような構成、内容の計画ではなく、より地域課題に注力した個性ある計画が期待される。そのような計画を前提にして、民間経営のように目標必達で施策運営や予算配分、改善・見直しに活用するツールとして業績測定が位置付けられることで、より機能する。

なお、米国連邦政府では、2011年に改正された政府業績結果法近代化法（GPRA Modernization

Act of 2010) では、各省庁の戦略計画と目標体系に合わせて、中期を想定した省庁優先目標 (Agency Priority Goals: APGs)、そして政府横断的な取組については、優先順位が高い業績目標 (High Priority Performance Goals: HPPGs) の設定と定期的なレビューが義務化されているなど、より戦略的な制度運用に見直されている。

以下、本稿では参考として、米国ワシントン州政府のガイドラインより、米国においての戦略計画の基本的構成要素を提示する。

表3 ワシントン州の戦略計画の要素

■ ミッション

組織が何をするのか、何故それをしているのか、何を目指し、そして誰のためにしているのかを説明

■ ビジョン

目指すべき目標と成果を達成したときの姿を簡潔に力強く説明する声明

■ ヴァリュー (価値)

組織の信念や行動を示す声明、戦略計画の価値観の声明・表明は、職員が業務遂行する指針に位置付け

■ 外的課題と機会の分析

組織が目標や業績目標を達成する能力に影響を与える重要な＋の外部要因を説明

■ 組織の能力と財務

組織の能力 (人、資本、技術)、財務の状況を整理

■ 戦略分析

組織が現在の状態と将来の望ましい状態との間のギャップをどのように埋めようとしているかを分析

■ 目標、目的、戦略

達成しうる上位目標、組織の取組の成果を示す計測可能な目的、これらを達成する戦略

■ 業績指標

業績測定は、機関の業務と成果を数値で表現したもの。これらはデータに基づいており、機関や活動が望ましい目標を達成しているか、政策や組織の目標に向かって進展があるかどうかについてのストーリーを可視化し監視

(出所) Office of Financial Management Budget Division, pp.5-8 より筆者要約

要点を整理すると、戦略的な計画は以下のような構成要素を踏まえている。上位の目標や理念と施策体系、業績目標が予算も含めて関連付けられ

ている、すなわち計画・目標、戦略、予算が一体化している。

- ・ 組織の使命・役割は何か
- ・ どこににいるのか (状況分析、SWOT)
- ・ どうなりたいか (ビジョンとミッション)
- ・ どのように到達するのか (目標、目的)
- ・ 財政的にどうするのか
- ・ どのような行動をとるべきか
- ・ どのように評価、調整、見直するのか

なお、日本の自治体では基本構想や基本計画にて予算額を示しているのは合わせて12.6%で、多くの団体 (95.1%) は実施計画に示されている (日本生産性本部 2024, p.10)。

④意思決定の仕組みとリーダーシップ

我が国においても業績測定の結果を活用するため、庁内で意思決定する幹部により構成される会議体の仕組みを取り入れている団体数は総務省 (2017, p.3) などの調査では現状確認はできないが、事例は確認できる。岩手県大船渡市、茨城県桜川市、愛知県みよし市、兵庫県川西市、佐賀県神埼市などであるが、団体数としては多くはない。事務事業や施策の評価による見直しの決定は、担当者レベルでの対応では限界があり、組織的な判断や意思決定を組み込まないと、評価結果による制度的あるいは予算措置に関わる見直しは実現しない。担当者による評価書作成のみであれば、対外的な説明としての資料作成の役割にのみに陥る。

米国においては、意思決定の仕組みとリーダーシップを一体化した取組として、業績スタットモデルがある。ボルティモア市ではマーティン・ジョセフ・オマリー市長の時の創設期のスタット会議は、二週間に一度、また、連邦政府の政府業績結果法近代化法では、四半期ごとに省庁優先目標の達成度を分析する幹部統括の会議体の設置と開催が義務付けられている。スタットモデルは、組織トップのリーダーシップにより目標達成に向けての改善と見直しの意思決定をオンゴーイングで進める役割を果たす。ハリー・P・ハトリ氏も、業績スタットモデルをテーマにした論文 (A Guide to Data-Driven Performance Reviews) を共著で2011年に執筆している。ハリー・P・ハトリ氏を含む複数者へのヒアリングでは、公的部門の業績情

報の活用において、業績スタットモデルが機能するのは、「リーダーの関与（意思決定との連動）」、「定期的な開催」、「データに基づく迅速な意思決定」の3点と確認した。

表4 業績スタットモデルの特徴

- ・データ活用を基本として定期的な会議開催
- ・オンゴーイングな改善と見直しの意思決定
- ・客観的・科学的な業績データを基に、どうすれば課題解決・改善できるか具体策（改善アクション）をその場で議論して決定
- ・改善アクションに、双方が合意（コミット）し、次回以降の会議で徹底的にフォロー

（出所）筆者作成

既述のように我が国においても、業績測定において、庁内で意思決定する幹部により構成される会議体の仕組みの事例はある。しかし、米国のようなトップの関与とリーダーシップを前提にした迅速な意思決定による改善・見直し、それを実現するための定例開催の会議体の設置を参考に、我が国においても業績情報に基づく意思決定をより実効的にし得る可能性はあるのではないだろうか。

⑤ プログラム評価による詳細分析

—業績測定との役割分担—

自治体においてプログラム評価を実施している例もあまり多くはないのではないかと。自治体において独自施策はあるものの、多くは法定のものであることや、評価・分析に知見がある担当者が多くない、また評価のための労力や予算が限定的という状況も推察される。また、中央政府においても、政策評価法第5条に基づく「政策評価に関する基本方針」にて総合評価として定義されているが、法施行後、積極的に実践しているのは国土交通省、農林水産省、厚生労働省、公正取引委員会、外務省の国別、分野別、スキーム別評価など、限定的である。

業績測定は成果志向の経営のためのパフォーマンスを測定するものであり、米国では施策などの効果検証は、業績測定とは別にプログラム評価により行われるなど、役割分担をしている。我が国では前掲の政策評価法（政策評価に関する基本方

針）において、業績測定とプログラム評価は評価制度として同じ枠組みに位置付けられているが、米国では業績測定は政府業績成果法に規定されている一方、プログラム評価全般を規定する法律はない。我が国実務においては、業績測定とプログラム評価を同じものとして捉えられていることが想定され、業績測定を行うことが、すなわち効果検証する評価であると認識されているのではないかと。なお、その他の取組ではカナダ連邦政府により導入されたRMAF（Results-based Management and Accountability Frameworks）は、対象プログラムについて業績目標設定と実績モニタリング計画、そしてプログラム評価を一体化する仕組みで、プログラム評価が業績測定を補完する役割と捉えて制度化された例もある。

我が国と米国の業績測定とプログラム評価について論じている、田辺（2014、p.13）によると、我が国では「指標設定と分析の欠如が相まって、業績測定が意思決定やマネジメントに利用されないという現状につながっているものと考えられる」と指摘し、「業績測定における指標設定、分析、意思決定への利用という面で、プログラム評価に期待できる効果は大きい。たとえ時間がかかっても、米国の事例を参考に、プログラム評価の実施を進めるべきではないだろうか。」と提言している。そのうえで米国の事例からの示唆として、まず「業績測定を「使える」ものにしていくためには、ある程度のコストがかかることを覚悟する必要がある」ということ、次に「プログラム評価の併用を目指すためには、プログラム評価の実施を後押しする仕組みが必要である」などと指摘している。このうち、「評価には時間とコストを要する」という点は、評価を業務負担でしかないと捉えられている現状の仕組みと意識を見直すための改善や工夫の他、後述する⑥の第三者による（プログラム）評価の実施による対応などが想定される。また、後押しの仕組みについての指摘は、例えば自治体の総合計画や単独施策などにおいて、限られた財源の有効活用と説明責任を意図した評価計画の記載などでの対応は可能である。これら田辺論文の指摘は、現在もなお日本での業績測定とプログラム評価の関係と課題、在り方を示している。

表5 米国における業績測定とプログラム評価
(定義、目的、データ、一般的な評価者)

■業績測定 ・事前に選定された指標に関するデータの収集、監視、レビュー、および報告を継続的に実施 ・目標達成の進捗、達成度、運用ミスを確認 ・データは定量的 ・組織内部者が日常業務の一部として実施
■プログラム評価 ・政府介入がどの程度効果的であるかを検討するための個別の体系的な研究で、政府介入の効果を、存在しない場合や、複数の代替アプローチと比較して評価を実施 ・活動と、その影響を受けた人々への成果との因果関係を分析、併せて異なる環境や集団における有効性のばらつき、要因を検証し疑問を検証 ・データと分析手法は評価の問いによる。一般的には定量的データと定性的データの両方、因果関係を調べる研究の場合、他からの影響を分離するために複雑・高度な分析方法を採用 ・組織から独立した外部研究者により実施され、高度な方法論に関する技術的専門知識が必要

(出所) Evaluation.gov'Performance Measurement and Evaluation'より筆者要約

⑥第三者による評価

多くの公共部門では、第三者の評価は評価制度運用全般を助言する組織や、評価結果のチェック者としての第三者委員会設置などの例は多いが、独立した評価者として第三者の立場で評価するものは、外務省などのODA評価以外の例はあまり多くはないのではないかと。また、総務省(2017、p.3)によれば、自治体において、外部評価の体制について「内部評価から独立している」としたのは、523団体のうち83団体(15.9%)である。米国では、業績測定を補完するプログラム評価は、成果検証を目的としており、統計的な数値はないがインタビューによれば、専門性がある外部者、第三者による客観的な視点、立場での実施が基本である。なお、「政策評価の手法等に関する研究会(平成11年8月27日第一回開催)」の検討過程においては、外部の評価の知見活用も議論されたが(第19回会

議録より)、最終的には、各府省の「自己評価」を基本とし、一方で、いわゆる「お手盛り」評価にならないように評価の客観性を確保することや、各府省にまたがる政策について評価専任組織として総務省が評価を行うことになった経緯がある(新井2013、pp.10-11)。

このような専門家による第三者評価の多くは、米国においても委託業務として実施されると想定されるが、その際には発注者である行政機関は、評価者に対して事実誤認などを除いて、自らの有利な結果や説明を指示、誘導するような過度な助言や介入を行うことを排除するなど、独立性を確保することが前提として求められる。

⑦最低限度の統計的な知見

多くの公共部門では、評価書作成マニュアルはあるが、データ収集の方法や分析の指南になるようなものは、あまり多く策定されていないのではないかと。そのため、行政職員において基本的な統計分析や数字の見方についての知見がない状況で、業績測定と分析が行われていることが想定される。業績測定に必要な統計的な知見であれば、例えば高度に数学的な素養がなくても、また数日を要しての研修をしない範囲でのレクチャー程度での習得は可能である。指標設定マニュアルなどは数多いが、データ分析・基礎的統計マニュアルなどの策定と運用が各公共部門において求められる。なお、統計に関する行政職員向けの研修もある。例えば、総務省統計研究研修所の「初めて学ぶ統計」は4時間、「統計担当者向け入門」は7時間30分で、これらを担当職員に参加を促すことでもよい。

統計的な知識の習得は、小野(2013、p.32)が指摘しているように、「評価における数字の誤用・濫用」、すなわち、数値により間違っている結論をもたらすような使い方の防止にもつながるなど、重要な論点である。

⑧市民への積極的な説明

行政評価は、従来はブラックボックスであった市民、議会などに対する施策や事業の内容や予算、成果の可視化において一定の役割を果たしたと言える。しかし、多くの市民は終わったものを説明されても、一般的にはあまり関心は高まらないと想定される。他方で、次年度や将来のことになる

と、地域課題の解決や実生活との関係から特定分野においては自身に影響があり、興味を引くことが想定される。我が国においても、予算編成過程を公表する取組はある。札幌市、名古屋市、善通寺市、北九州市などで、市民からの質問に対する書面フィードバックも行われている。

しかし米国での取組との比較では、「成果と予算との関係づけが明確には示されていない」、「想定シナリオがない市民との直接対話の機会がない」、という点で相違があるのではないかと。連邦制の米国と我が国自治体においては、組織の立ち位置や事情が異なり、財源構成や予算編成の方法も違いがあり、成果と予算を関連付けるという点において、実務的な対応や対市民への説明に対する必要性、必然性の事情もやや異なる。とは言え、納税者であるという市民目線に立てば、業績や成果と予算を関連させて対外的に説明する機会は、広く関心事であり、むしろ自治体からの積極的な説明、開示の対応により市民の関心を喚起するという取組が期待される。

米国自治体では、予算編成過程での対市民向けの対話の機会や公聴会出席が制度化されている。例えば、4月上旬、フェニックス市は、2024-25会計年度（FY）予算案に対する意見を収集するため、11のコミュニティで予算公聴会を開催している。過年度のコロナ禍においてもオンラインなどで実施されていた。また、カンサスシティ市は、公開イベント、オンラインやソーシャル・メディアなどを通じて、事業計画や予算サイクルのプロセスに市民が参加して、影響を与えることができる。市は、事業計画と予算策定への市民の参加は、多様な地域課題・ニーズを踏まえた事業選択を正当化し、それらの選択が利害関係者の優先事項やサービスレベルに対する意向を反映しているかどうかを確認するために極めて重要であるとしている。そして、市民とビジネスコミュニティは、公共サービスの顧客であり、納税と投票によって所有者となり、市の目標達成に取り組むパートナーであり、カンサスシティでは、市民の参加を積極的にすすめている、としている。

表6 カンサスシティの市民参加の形態

■住民ワークセッション

住民ワークセッションは、全市事業計画の提出後と年次予算の提出後に複数回実施される。その都度、市内各所でフォーラムが開催される。フォーラムの目的は、全市事業計画について議論し、次年度の年間予算編成に活用できる市民の意見を得ることである。利益団体を招いての公開セッションも行われる。

■予算公聴会

市民は、予定されている予算公聴会に出席し、提出された予算案に意見を述べるができる。市議会議員は、予算公聴会のため3つの地域社会を訪問し、また予算審議の際には市民意見を受け付ける。

■オープン予算KC

カンサスシティは、内部データ提供と市民フィードバック活用を約束する。このコミットメントの一環として、住民の参加を促進し、市のデータにアクセスできる「オープン予算KC」は、住民らが、市議会が優先する項目に市がどのように資金を配分しているかを調べることができるツールである。

（出所）カンサスシティ HPより筆者要約

4. 日本と米国の異同から 日本が学べること

本稿では、主として2000年代半ばに筆者が実施したハリー・P・ハトリ氏らへのインタビューにおける「業績測定の実務での在り方についての指摘」を振り返りながら、我が国における取組を比較し、現状をレビューした。米国、日本ともに行政評価の制度は統一されていないものの、理論的背景や取組の基本要素や構成は共通しており、それを前提にした上で、全ての要点について、米国の取組を導入すべきという示唆は、本稿での現況分析が十分ではないため、やや拙速であるものの、我が国における業績測定の高度化や活用の進化・進展に向けての検討において参考となる論点になると思われる。

表7 米国の取組を踏まえた我が国の業績測定において参考にすべき点（要点）

- ① 業績測定においては、詳細分析（ドリルダウン分析）の視点を取り入れる。
- ② 目標の達成と併せて未達成にも着目して分析する。
- ③ 地域や社会の課題解決に対応したロジカルで戦略的な計画を策定、運用する、もしくは包括的な計画では優先目標と指標を設定して、限定した目標と取組に対して、より注力する仕組みを取り入れる。
- ④ トップ・幹部が参画する定例開催の会議体を組成し、業績情報に基づく改善実行と迅速な意思決定の仕組みを構築する。
- ⑤ 業績測定は継続的な改善と見直し、評価は客観的で総括的な分析と区分・整理して、相互に役割を補完する。
- ⑥ 第三者による運営や結果のチェックと併せて、客観的・専門的な立場からの独立した第三者による評価を実施する。
- ⑦ 評価者、評価担当は、最低限度の統計的な知見を身に付ける。
- ⑧ 成果と次年度予算とを関連させて、市民とのシナリオのない、オープンで積極的な説明対応・対話とニーズ把握を行う。

（出所）筆者作成

5. まとめ

我が国における業績測定は、行政評価の枠組みにおいて実践され、1990年以降の取組を通じて、国、自治体において一定の定着を得ている。しかしながら他方で、行政運営に対する説明責任やP-D-C-Aを取り入れる仕組みを意図して制度化されて導入されたものの、時間の経過によって担当者にとっての定型的な作業となり、行政の取組内容や予算、決算、結果を年度末に公表することが目的化し、いわゆる作業に陥り、実務的な負担、評価疲れになっており、結果として担当者は実施目的を見失っているとの指摘もある（総務省2017、p.7；東京市町村自治調査会2023など）。端的に言えば、評価書作成を通じて事務負担が増え説明責任の役割は一定程度果たしているものの、改善・見直しを行う経営ツールとしては十分には機能していないのではないだろうか。米国の取組

と対比すれば、日本はより「説明責任志向型」の行政評価という特徴である。

一方で米国では、業績測定は評価そのものではなく、「成果志向の経営型」のマネジメント・ツールであり、評価（Evaluation）とは異なるものと位置付けられている。このことが業績測定について本稿で整理したように、業績測定の取組について両国では、改善・見直しのためのドリルダウン分析、戦略的な計画、会議体設置による定期的な業績レビューに基づく意思決定とリーダーシップ、市民への積極的説明において、違いが生じている要因の一つではないかと考えられる。

行政評価の起源を米国として発展してきた日本との制度や運用面で、このように仮説的に「説明責任型」「成果志向の経営型」と整理した違いについての要因分析は、本稿では詳細には取り扱わない。この点についての背景要因として、各自治体の役割が地域によって多様異なる米国では、より市民志向の経営的な志向が求められていることがあるのではないかと。一方で日本での行政評価導入の直接的な背景は、行政に対する透明性確保やアカウンタビリティ（説明責任）実現にあったこと、またとりわけ自治体は共通、統一的な行政サービスが求められ、他団体の取組を参考にする場面が多いこと、行政運営において継続性を志向する考え方やボトムアップ的な運営を志向する考え方があるのではないかと。それらの複層的な影響によって、行政評価導入が意図したP-D-C-Aのマネジメント・サイクルを構築するという、当初に意図した経営的な側面での導入効果が十分に実現できなかったのではないかと考えられる。

中央政府であれ自治体であれ、行政機関として法令に従い誠実に施策を執行することが責務であるという前提であるものの、行政評価が導入された1990年代以降においては、より複雑化・高度化する地域や市民の課題を解決するために、民間経営のように成果志向で、柔軟で弾力的な対応を行うことが期待されるようになってきている。そのような中、庁内や関係機関との調整や協議、また対外的な説明において、客観的なデータ活用は必然であるということは、当時も今も変わらない。

筆者は、シンクタンク研究員として行財政改革をテーマに主要先進国を調査し、数多くの実務担

当者や有識者へのヒアリング調査を実施した。その際にインタビュー対象者から、「我々も常により良い解決策を模索している、日本の経験も共有して欲しい」という発言を多く得た。1990年後半から2000年にかけて行政評価の取組が進展した時期においては、先行事例、先進事例に注目が集まり、学会やフォーラムなどで実践的な取組の研究や意見交換も盛んで、実務者も研究者も他より学ぶという姿勢や傾向があったように記憶している。業績測定は、適切な指標の設定と定期的な実績モニタリングを中核としているが、実績の詳細な分析と、それを活用したマネジメントを実現する仕組みを一体としている一方、高度な評価（プログラム評価）は専門家、第三者が実施するのが米国の基本的スタンスである。我が国公共部門の業績測定の取組を、米国の取組を参考にして改めて経営の視点から、行政運営を成果志向に変革するためのツール、仕組みとして見直すことで、業績測定の取組も変革し、より進化する可能性があるのではないかと考え、またそれを期待したい。経営の仕組みは課題や時代に応じて変化し、進化する、その過程では課題解決に向けて戦略的思考であること、そして他者より学ぶことの意義が重要なのではないだろうか。

参考文献

- 新井誠一（2013）「政策評価制度10年の軌跡～制度導入以降の省察と今後の展望～」、『日本評価研究』、13（2）：3-19
- 荒川潤、片山泰輔（1998）「行政のアカウンタビリティ」待望論の盲点～前提としての政策評価の重要性」、『経済月報』、733：3-8
- 小野達也（2013）「政策評価と実績測定～府省の実績測定における計量・計数の現状～」、『日本評価研究』、13（2）：21-36
- 会計検査院受託：新日本監査法人「アメリカの政府業績成果現代化法（GPRAMA）等の運用から見た我が国の政策評価の実施及び会計検査に関する調査研究」https://www.jbaudit.go.jp/koryu/study/pdf/itaku_h27_1.pdf（2024年11月15日閲覧）
- 北川正恭（1998）「三重県における行政改革」、『日本公共政策学会年報1998』
- 窪田好男（1998）「三重県の事務事業評価システム～評価を通じた政策改善、あるいは評価システムによる「行政革命」」、『日本公共政策学会年報1998』
- 左近靖博（2015）「政策評価の結果の予算への反映—実現に向けて求められる要素と実際—」、『季刊評価クウォータリー』、34：32-39
- 総務省受託：三菱UFJ リサーチ&コンサルティング株式会社（2008）「カナダにおける複数府省にまたがる政策の評価に関する調査研究—報告書—」https://www.soumu.go.jp/main_sosiki/hyouka/seisaku_n/pdf/0803_canada.pdf（2024年11月15日閲覧）
- 総務省「政策評価の手法等に関する研究会」https://www.soumu.go.jp/main_sosiki/hyouka/81303.htm（2024年11月15日閲覧）
- 総務省（2017）「地方公共団体における行政評価の取組状況」https://www.soumu.go.jp/iken/02gyosei04_04000062.html（2024年11月15日閲覧）
- 田中啓（2009）「日本の自治体の行政評価」、『分野別自治制度及びその運用に関する説明資料No.14』
- 田中啓（2014）『自治体評価の戦略—有効に機能させるための16の原則』、東洋経済新報社
- 田辺智子（2014）「業績測定を補完するプログラム評価の役割—米国GPRAMAの事例をもとに—」、『日本評価研究』、14（2）：1-16
- 玉村雅敏（2014）『総合計画の新潮流—自治体経営を支えるトータル・システムの構築—」、公人の友社
- 東京市町村自治調査会（2023）「基礎自治体における行政評価の効果的、効率的な実施に関する調査研究報告書」https://www.tama-100.or.jp/contents_detail.php?co=new&frmId=1215（2024年11月15日閲覧）
- 日本生産性本部（2024）「自治体総合計画に関するアンケート調査」https://www.jpc-net.jp/research/assets/pdf/20240828_report_v2.pdf（2024年11月15日閲覧）
- City of Phoenix‘About Budget Hearings’ <https://www.phoenix.gov/budget/hearings>（2024年11月15日閲覧）
- Evaluation.gov‘Performance Measurement and Evaluation’<https://www.evaluation.gov/assets/resources/Performance-Measurement-and-Evaluation.pdf>（2024年11月15日閲覧）
- Kansas City‘Budget Process – Public Process’ <https://www.kcmo.gov/city-hall/departments/finance/office-of-management-and-budget-omb/public-process>（2024年11月15日閲覧）
- Office of Financial Management Budget Division（2019）

‘2019-21 STRATEGIC PLAN’ <https://ofm.wa.gov/sites/default/files/public/budget/instructions/other/2019-21StrategicPlanGuide.pdf> (2024年11月15日閲覧)
(2024.12.9受理)

What Japan Can Learn from U.S. Performance Measurement Practices —Review of Japan’s performance measurement and evaluation efforts—

Yasuhiro Sakon

Hamagin Research Institute
y-sakon@yokohama-ri.co.jp

Abstract

Performance measurement in the U.S. is based on the concept of “Managing for Results,” which integrates the analysis of measured performance and the decision-making mechanism for improvement and review. This will serve as a reference for future improvement of evaluation and performance measurement practices in Japan, where the introduction of such practices has progressed with the intention of ensuring transparency and achieving accountability. This paper presents the main points for improving Japan’s performance measurement efforts with reference to U.S. practice.

Keywords

Performance measurement, Drill-down analysis, Strategic Planning, Stat models, Managing for Results (MGR)

【総説】

業績測定に何を求めるか —EBPMとマンネリズムの間隙を巡って—

小野 達也

追手門学院大学

t-ono@otemon.ac.jp

要 約

業績測定による評価は、この約30年間であらゆる公共部門に行き渡った感がある。近年は評価作業のルーティン化や評価疲れが取り沙汰されるが、一方でEBPMの標榜を受けて業績測定にも重要な任務が課された格好である。本論文は、日本で多く行われている業績測定による評価の定量的機能について、ハリー・ハトリーらの文献をあらためて参照しながら、様々な側面をエビデンスという観点から点検し、改善が求められる事項を再整理するとともに今後への期待を述べる。これからの日本の政策にエビデンスは必要であり、そこで業績測定による評価が果たすべき役割は広く、重い。

キーワード

業績測定、エビデンス、評価指標、ハリー・ハトリー

1. はじめに

2023年、ハリー・ハトリー、そしてジョセフ・ホーリーが他界した。日本の政策評価、とくに業績測定の研究と実践は多くを両氏に負うだろう。この機会に、業績測定の機能をもつ評価制度・システム（本稿ではこれを「業績測定型評価」と呼ぶ）の来し方行く末について、近年重視されるエビデンスという観点から、思いを巡らすこととしたい。

一般に、業績測定型評価はパフォーマンスの定期的な測定・評価・対策提言・公表という過程である（上野・上野2007）。日本における実際の評価制度・システムにおいては定性的な情報や判断が一定の役割を担う場合が少なくないが、本稿の考察対象は業績測定型評価の中核である定量的機能¹である。具体的には、評価指標の設定とモニタ

リングのほか、目標値との比較やプログラム間の比較、他地域との比較などを含む。

本稿の目的は2つある。第1に、業績測定について数多くの著作を残したハトリーらの議論を踏まえつつ、日本で行われてきた業績測定の段階ごとにあらためて点検し、改善すべき点を明らかにすることである。第2は、近年のEBPM（Evidence-Based Policy Making：エビデンスに基づく政策立案）の要請を踏まえ、上記の点検において業績測定はどのようなエビデンスをもたらし得るのかという観点を加味して行い、今後業績測定に期待すべき事柄を明らかにすることである。

以下ではまず、日本における業績測定型評価のこれまでを振り返った上で、本稿においてエビデンスという観点から考察することの意味を確認する。続いて、評価指標の設定・測定段階、実績値

の比較段階の順に検討を加え、最後に業績測定型評価の今後について述べる。

2. エビデンスという観点からみた業績測定型評価

(1) マンネリズムとEBPM

1996年に三重県の事務事業評価システムが導入されたのを皮切りに、全国の自治体において業績測定型評価が急速に普及し、導入から四半世紀を越えた団体も多い。同一の評価システムが長期間続くことはほとんどないが、変遷を経ながらも評価の実践が各所で定着したことは間違いない。

政府でも2001年に導入された政策評価制度において、施策レベルで業績測定を行う実績評価方式及び目標管理型政策評価が重要な役割を担ってきた。事務事業単位の業績測定機能をもつ行政事業レビューも2010年の導入決定以来、制度の改定を経て今日に至る。

このように普及し定着したとはいえ、業績測定型評価の実践においてルーティン化の傾向が見られることも否定できない。マンネリズムである。評価疲れの末、評価制度を廃止した自治体も出てきている。

一方、数年前からは日本でもEBPMへの取り組みが具体化している。政府の見解を総合すれば、すべての政策のあらゆる段階がEBPMの対象である（小野2024）。

EBPM推進の方針のもと政府の政策評価もリニューアルが進んでいるが、少なくとも当面は、政策効果の証拠という広義のエビデンスをあらためて重視することが中心となる。その舞台は政策評価制度と行政事業レビューにおける業績測定である（小野2023）。

EBPM志向は自治体にも浸透しつつある。様々な取り組みがあるが、大半の団体が取り組む業績測定型評価において、当然エビデンスという観点が意識されることになろう。

(2) 業績測定型評価とエビデンス

現行の業績測定型評価はエビデンスを提供しているだろうか。また今後、より質の高いエビデ

ンスを提供できるだろうか。政府がEBPM推進を標榜する現在、政府・自治体を問わず日本中の業績測定型評価に突き付けられた問いである。本稿では、業績測定型評価に期待されるエビデンスという観点から、エビデンスを「政策の効果の（広い意味で）証拠となる数量データ」とする²。

日本の業績測定型評価における評価指標の取り扱い、即ち指標設定から指標値の各種比較までの過程については、これまで多くの課題が指摘されてきた³。それらの多くについてハトリーらが早くから繰り返し強調してきたことが実行されていないといえる一方、独自の発展を遂げた観のある日本の業績測定型評価に特有のものもあるだろう。

一方、定量的に評価を行うことの目的として多く掲げられ、求められて来たのは客観性の担保である。また社会科学における測定が充たすべき条件として多く言及されるのが妥当性と信頼性であり、評価指標の設定・測定においてもしばしば求められる。しかし、これらは政策効果の証拠となるための必要条件であっても十分条件ではないだろう。業績測定型評価の課題はエビデンスという観点からは必ずしも検討されてこなかったきらいがある。

本稿の目的をあらためて述べれば、ハトリーや他の論者による見解と日本の現状を照らし合わせ、さらにエビデンスという観点を交えて吟味し、日本の業績測定型評価の課題を整理の上、今後を展望することにある。

3. 評価指標の妥当性と信頼性

(1) ロジックと妥当性

評価対象プログラムのロジックモデルを描き、その上で評価指標を設定することが業績測定型評価のいわば標準的な手順となった感がある。翻訳もされたハトリーのテキスト（Hatry 1999）が果たした役割が大きいと思われるが、1990年代後半にはいくつかの文献がその手順を推奨している（Swiss 1995; Hatry and Kopczynski 1997; Poister 1999）。

そのメリットは、ロジックを踏まえて、測定すべきものを測定する指標を設定できること、すな

わち評価指標の妥当性validity（本稿ではこれをロジック上の妥当性と呼ぶ）が得られることである。

ところで、測定対象の事象とデータが自然科学のように密着していない社会科学の測定全般で、妥当性は最も基本的な条件とされるが（小野2013）、業績測定でも少なからぬ文献が、指標の設定基準として妥当性を筆頭に掲げる⁴。しかし妥当性が具体的に何を指すか、見解は様々である。

ロジック上の妥当性は、Hatry（2006）やCAP（1998）がvalidityという語を用いずに、指標の選択・設定に係る最重要な基準として挙げたrelevance（プログラムの目的との関係が論理的かつ直接的である）やimportance（アウトカムの重要な側面を測定している）、results oriented（アウトカムに焦点をあてている）などに該当する。

なお、因果推論や統計学的な有意性の検証を志向するインパクト評価（プログラム評価）では、指標というより推論の妥当性を様々な観点から問うが、そこでの議論は業績測定型評価には直接該当しない⁵。

さて、多くの場合、ロジックモデルは少なからぬ数のブロックで構成されるだろう。精密な把握と分析には多数の指標が必要であり、各種リソースの制約を考えれば少数の指標が得策である。両者は明らかにトレードオフの関係にある（Hatry 2006; Swiss 1995; Wholey 2006; Van Dooren et al. 2010）。

その上で、どのような観点から評価指標の集合を構成すべきか、即ち指標群の妥当性とは、ハトリーらが評価指標群に求める包括性（comprehensiveまたはbalanced and comprehensive、プログラムの重要な側面が漏らさずカバーされていること）に尽きるだろう（Hatry et al. 1994; Ammons 1995; Hatry 2006, 2014; Poister 2003）。その内容は、以下の4点に整理できる。

第一に、最終アウトカムに近い指標を含むこと。効果を把握するためにはこれが不可欠だが、しばしば単純なことではない。例えば国の施策で自治体の取り組みを促す場合、自治体の取り組みをアウトカムとすることがあるが、国民から見ればアウトプットである。

第二に、必要に応じてアウトプットや直接的・中間的アウトカムの進捗を把握すること。最終段

階のアウトカムの発現に至った場合も至らなかった場合も、その因果関係を示唆するエビデンスとなりうる。

第三は、外部要因の影響が大きい指標と小さい指標を使い分けること。重要なアウトカムは、外部要因の有無と程度を定性的にでも認識の上で指標を設定し、明確な議論をすべきである。外部要因の有無を曖昧にしたままの指標は、エビデンスの観点では全くの逆機能となる。また、直接的アウトカムやアウトプットの指標は外部要因の影響が相対的に小さいだろう。

最後の第四は、重要な経路をカバーすること。ロジックモデルが複数の経路で構成される場合、重点事項や予算規模など何らかの観点で重要な経路はカバーすべきである。

ロジックモデルに基づいて評価指標を設定する取り組みが広がりつつある状況は喜ばしい。しかし、評価指標群の実績値がエビデンスとなるためにはこの4点が求められる。

（2）もう1つの妥当性と測定の信頼性

ロジックモデルのあるブロックについて評価指標を設定したとする。この指標の測定値がエビデンスとなるためには、当該ブロックに係る事象の変化を的確に反映しなければならない。

この概念は文献では応答性responsive（CAP 1998）、感度sensitive（Poister 1999; Van Dooren et al. 2010）などと呼ばれるが、本稿ではこれを指標定義の妥当性とした。即ち、業績測定における指標の妥当性は、ロジック上の妥当性と指標定義の妥当性という2段階に分けて考えることができる⁶。

現実の場面で2段階の妥当性を十分に有する指標（群）を設定できるとは限らないが、ハトリーらが述べるように、指標のニーズを先に考えることが重要である（Hatry 2006; Swiss 1995）。設定できなくても可能であれば代理指標を設定すべきであり、それも難しければ必要な指標（値）がないことを明示すべきであろう。そして、指標群がロジックモデルのうちカバーしていない部分を明確にする必要がある。このことを抜きにして、いかなる指標群も政策効果のエビデンスとはなりえないだろう。

さて、社会科学の測定において、しばしば妥当性と並ぶ条件とされるのが信頼性である。業績測定においても当然、同一の事象であれば、誰がいつ測定しても同じ結果が得られなければならない。

指標データを得るために記録・観測・調査を行う場合はもとより、既存の統計データを使う場合にも信頼性の吟味が必要であり、その説明が求められる。エビデンスであるためには、データは単なる数字ではなく「数字+（データの履歴などを説明する）補助情報」でなければならないが、信頼性に関する情報は、もっとも重要な補助情報の1つである。

一例を挙げれば、各種の標本調査の結果数字が評価指標に使われるが、その誤差情報を目にするのが日本では減多にない。標本調査の結果数字は、誤差の大きさを知らずに解釈することはできないはずである⁷。

以上により、評価指標（群）とその測定値が満たすべき条件はロジック上の妥当性、指標定義の妥当性、測定値の信頼性という3条件となるが⁸、妥当性を欠く指標についていくらか信頼性が高い測定結果があってもエビデンスにならないことはいうまでもない。

(3) 効率を測る指標

効率も業績測定型評価における重要概念である。効率を最も一般化した定義は「産出÷投入」であろうが、この投入と産出はシステムへの何らかの投入・産出であり、例えば投入は行政資源に限らない。すなわちロジックモデルを左から右に向かう連鎖図として描くとすれば、相対的に左側のブロックと右側のブロックの比が広義の効率に相当する。

業績測定において効率をどのように測るべきか、多くの論者が様々な分子・分母の組合せを示しているが⁹、それらは次の3種類に集約できるー①インプットとアウトプットの比、②インプットとアウトカムの比、③アウトプットとアウトカムの比。さらにインプットは費用、スタッフの人数、所要時間という3通りがある。何れも、効率向上の要因あるいは取り組みとして、分子の変化と分母の変化という2つがあることになる。効果のエ

ビデンスという観点からは、ハトリも主張する通り、可能な限りアウトカムを産む効率に焦点をあてるべきであろう（Hatry 2014）。

日本の業績測定型評価でもほぼ例外なく効率が導入目的のキーワードの1つとなっているが、米国とは異なり、効率を上のような種類の指標で直接測ることは少ない。その背景にはプログラムの費用を人件費を含めて具体的に把握するのが容易でないこと、1つの数字で効率を評価することへの抵抗感などがあるだろうか。効率に関して多いのは、定性的な判断に基づく3段階などの評定である。評定の根拠が説明されていればよい方だが、それでもそのような評定結果の記号や数字をエビデンスと考えるのは難しい。やはり客観データを用いて効率を算出する工夫が求められる。

もう一点の懸念は、インプット（予算や職員数）そのものを効率の指標とし、その削減をもって効率化とみなす風潮である。アウトカムが不変であることが暗黙の前提となっている。ハトリも警告するように、見かけ上効率が向上していても（あるいはそうみなしていても）サービスの質などを測る必要がある（Hatry 2006）。アウトカムとインプット双方のエビデンスを組み合わせない限り効率のエビデンスは得られない。

なお、ハトリは米国等でもアウトカムを向上させるのに必要な費用がまだまだ不確かで更なる取り組みが必要である（Hatry 2008）と述べる一方、今後はコストのデータをアウトカムと結び付けて効率を明らかにすることが広く当たり前のように行われるだろうとも予測していた（Hatry 2010）。

4. 実績値の比較

(1) 実績値の比較対象

一般に指標の値は何かと比較することで初めて判断に資する意味をもつ。業績測定においても、実績値が測定されれば、何らかの比較によって解釈や分析を行うことになる。比較の対象として多く挙げられるのは、過去の実績値、何らかの標準や基準値、予め掲げた目標値、他の特定の類似機関や地域、機関内・地域内・対象者等の様々な区分、民間セクターなどである¹⁰。ハトリをはじめ、

これら比較対象をベンチマーク、その比較をベンチマーキングと呼ぶことが多い。

掘り下げた分析を行わないことが前提の業績測定において、ベンチマーキングの中でも、何らかの基準値（多くは目標値）との比較はプログラムの進捗が順調か否かを判断する有力な手段となる。米国でも業績測定の普及に貢献した先行事例のオレゴン州等のベンチマーキングがそうであったように目標値との比較が重視されるようになった（Greiner 1996）¹¹。

米国に多くを学んだ格好の日本ではその傾向がさらに強い。三重県の事務事業評価は必ずしもそうではなかったが、今や他の比較はほとんどせずに目標値との比較のみ行うのが1つの典型となった感さえある。

さて、評価制度としての業績測定の目的は個々のプログラムに止まらず、プログラムの集合のPDCAサイクルに資する情報提供でもある。指標ごとの測定結果や進捗状況は、プログラム単位への集約と、さらにそれらの比較ないし全体の集約が何らかの形で求められることとなる。日本では、各指標の目標達成状況を踏まえてプログラム単位の状況把握を行い、さらに組織全体や予算全体、計画全体というレベルで比較と集約を行うことが多い。指標の目標達成状況の全体での集計を直接行うことも少なくない。

以下の（2）～（4）では、日本で特徴的な目標値と実績値の比較及びその横断的比較という2段階の比較の組合せについて、エビデンスの観点を交えながら3つの課題を取り上げる。続く（5）（6）ではエビデンス提供につながる別の方向の2種類の比較について述べる。

（2）目標値の設定

実績値を比べる対象である目標値の設定方法については多くの解説があるが¹²、要は各種データの参照・分析と関係者間の議論である。その際、結果的に重要な意味をもつことになるのは目標値を達成することの難しさである。達成したか否か、あるいはあと少しまで近づいた場合の解釈は、その難しさ次第で変わるはずだからである。

業績測定型評価の有名先行事例であったオレゴン・ベンチマークは、目標値設定の姿勢と基本的

手順を一定程度明確にしていた¹³。それらを大いに参照した日本の業績測定であったが、目標値の設定手順を明確にした例、達成の難しさを説明した例、達成の難しさを明示的に統一した例は非常に少ないのが現状である。

達成の難しさが不明なままでは、個々の指標の達成状況はもとより、指標間・プログラム間での達成状況の直接・間接の比較の意味はさらに曖昧なものになる¹⁴。なお、政策の企画・立案の初期段階で、目標値が象徴的な指標とともに予め決まっている場合もあるが、ここでの議論はそのまま当てはまる。

政策の目標値の性格について、かつて西尾（1976）は①理想状態を示す期待値、②一応の水準として満足できる充足値、③絶対達成すべき限界値、という3分類を提示した。さらに、政策評価制度導入のタイミングで田辺（2001）はこれを再整理し、①達成可能か否かは別として尽力しようという理想値、②十分達成可能な期待値、③最低限達成すべき限界値、としている。

本稿では、これにいわゆるストレッチ・ターゲット（現状では難しいが改善があれば可能な水準）（Poister 2003; 田中2014）という考え方を加味し、①理想的な水準、②挑戦的だが可能な水準、③現実的な水準、④最低限の水準という4分類を提示する。

ところで、目標値が客観的・科学的に設定できる基準値でない場合、つまり様々な立場の人間による協議や判断の産物である限り、目標値と実績値の比較自体はエビデンスにはならない。エビデンスである（と期待される）指標値に関する有力な説明情報である。

目標値の設定には、様々な政治の介入やgamingが起ころう（Hatry 2006; NAPA 1994; Radnor 2008; Markiewicz and Patrick 2016）。その防止策として、目標値の設定手順の客観性や透明性、目標値の性格の明示はもとより、様々な制約の下では暫定的な目標値、複数レベルの目標値、範囲の形の目標値、必要に応じた目標値の変更など（Hatry 2006; NAPA 1994）も有効だろう。

（3）フロー/ストックの区別¹⁵

1つのプログラム内での指標同士の比較やプロ

グラムの枠を越えた指標同士の比較において指標の比較可能性は前提となる。目標達成状況の比較はもとより、実績値の変化を比較する場合に懸念されることの1つがフロー指標とストック指標の区別である¹⁶。

経済統計で定義される通り、フローとは一定期間に発生する量であり、ストックとはある時点に存在する量である。一定量の行政の注力が続き、そのまま指標値に反映すればフロー指標は横這い、ストック指標は一定のペースでの増加が自然である。即ち両者の推移はベースラインが異なり、目標値の達成状況はもとより推移を直接比べることはできない。1つのプログラムの中でフロー指標とストック指標を比べる際にも注意が必要だが、プログラムAのフロー指標とプログラムBのストック指標を素朴に比べてもその結果はエビデンスとは無縁である。

日本では、両者が全く区別されないまま混在している場合が非常に多い。推移を直接比べるには、ストック指標の差分（＝期末値－期首値）を取ってフロー指標に変換する必要がある。なお、一定期間のフローを合計した累計は、ストックとはやや異なる概念となる。

業績測定で設定される指標には、その定義がフローとストックのどちらでもないものも少なくない。しかし、フローとストックの定義を拡張して、同じの量の政策等を継続することが指標値の維持をもたらすのが自然であるものを「フロー的」、指標値の上昇をもたらすのが自然であるものを「ストック的」とすれば、ほとんどの指標はそのどちらかに該当するだろう（小野2008、2011b）。

（4）目標達成状況の数値化と横断的比較

目標が数値で設定され、実績も数値で測定されれば、目標達成状況を計算すること（その計算結果を以下では目標達成度と呼ぶ）は自然な発想である。実際、日本の業績測定型評価の多くは、何らかの方式で目標達成度を計算しているが、ここでは、個々の指標の目標達成度の計算、及び目標達成度の指標間・プログラム間の比較において懸念がある¹⁷。

目標達成度の計算式は、①ある時点の比率「実績値÷目標値」（目標年度または各年度の目標値

に対する各年度の実績値の比率）、②計画期間（評価対象期間）の変化量の比率「（終了時または途中の実績値－開始時の実績値）÷（終了時の目標値－開始時の実績値）」、③計画期間の総量の比率「計画期間の実績値の合計÷計画期間の目標値の合計」という3種類に大別される（小野2024）¹⁸。

①～③は目標達成状況を異なる角度から捉えており、使い分ける必要がある。例えばストック指標の場合、①の計算結果の解釈は難しい。また理想的な高い水準を目標値としている場合は①や③の計算結果よりも②の計算結果が重要だろう。数量の増加を目指す指標で目標未達の場合、①の達成度は②よりも必ず高くなることにも注意が必要である。

ところが、現行の業績測定においてこれらの計算式を明示的に選択あるいは使い分けた例はほとんど見られない。数値の低下を目指す目標値の達成度として①の逆数を計算してしまうなど、意味不明の処理も見られる。

実績値と目標値の比較は多くの業績測定型評価において中核的な手順となっており、それは政策効果のエビデンスである実績値について、目標達成状況という観点からいわば補助線を引くように解釈を提供する最重要段階といえる。

さて、一般に業績測定において、測定結果をプログラム間で比較することは分析の基本でもあるが、日本で特徴的なのは、目標達成度の比較が広く行われることである。指標ごとの目標達成度を指標間・プログラム間で比較し、しばしば計画や予算全体の集約も行われる。自治体では達成度をカテゴライズして集計した表をしばしば作成し、政府の目標管理型政策評価では目標達成状況を5段階で評定し集計した。欧米の文献では、計算した目標達成度を比較し集計する例を筆者は目にすることがない¹⁹。

ただし、このような比較はどのような指標も直接比較できる合理的な手段である一方²⁰、指標の進捗状況の比較を単純化した計算で済ませてしまっているともいえる。

その背景として、日本では計画の進行管理としての業績測定の第一の目的が計画全体の目標達成状況の管理・報告となっていること、あるいは事務事業評価の第一の目的が行財政改革の観点から

の予算全体のスリム化であったというような事情も指摘できる。

ここでも問題となるのは比較可能性である。達成度の数値を直接比べる場合、目標値達成の難しさが異なったり、フローとストックが混在したりすれば、カテゴライズして集計する意味は曖昧である。

また、政策体系のある階層で1つ下の階層の複数の指標の目標達成度を総合するための評定も行われる。担当部局の判断による場合はもとより、達成度の水準に基づくルール（一定の達成水準以上の指標を数えるなど）による場合も、目標達成度の比較可能性が十分でなければ、評定結果の数字や符号の質は危うい。

目標達成度の計算とその比較・集計・総合を比較可能性が不十分なまま行った結果、例えばある10年間の計画の中間時点で75%のアウトカム指標の目標達成度が50%を超えており、計画は順調に進捗しているという説明がなされたとしよう。これでは計画の進行管理とアカウンタビリティの何れにも資するとは言い難く、指標の実績値が政策効果のエビデンスであったとしてもEBPMの機運に水を差すことになってしまう。

(5) アウトカムの内訳の比較

ハトリは共著や編書を含め少なくとも16本の著作において、アウトカム指標を何らかの観点で分割した内訳（breakouts、disaggregationsなどと呼ばれる）に注目することの重要性を繰り返し説いている²¹。ハトリ以外の文献では多くはないが、Roy and John（1996）、Poister（2003）が概ね同じ趣旨の見解を詳細に述べ、Smith（2006）はアウトカム測定値の統計分析という観点から言及、Van Dooren et al.（2010）は分割と統合（総合指標化など）を合わせて1つの分析戦略として論じている²²。

国であれ自治体であれ、政策のアウトカムをまず総数や平均等の代表値で捉えることは自然であるが、その水準が目標値等の一定の水準を上回ったことをもって事足りりとする訳にいかないのは明らかである。

Hatry（2006）は総数等の単一の値のみのアウトカム指標値を、湖の深さの平均値に例える。平均

が浅いからといってこの湖は安全に泳げるといえないことは明らかなが、そもそもいかなるデータの集合であっても、代表値だけに注目してはいけないのは統計学の初歩である。

ところが、業績測定型評価において、例えば国の政策が自治体を經由して実施されたアウトカムの発現状況を地域別の値で見たり、自治体のサービスのアウトカムを市民の属性別に確認したりすることがほとんど行われない。1つの指標値が全体の状況に関するエビデンスであっても、一定規模の政策のPDCAサイクルを回すのであれば、それだけでは政策効果のエビデンスとして著しく不足であろう。Hatry（2008）は、世界中の業績測定においてアウトカムの内訳が広く無視されてきたとさえ述べているが、日本でもハトリの懸念がそのまま当てはまるだろう。

プログラム評価であればアウトカムの分割は当然のことであるとしても（Hatry 2014）、また評価制度とは別に政策立案等の現場で内訳に注目することがあるとしても、これは業績測定型評価の課題である。評価フォーマットのレイアウトやスペースの都合で評価指標の値が代表値1つであっても、その解釈や対応策の検討において内訳に言及しないのでは、業績測定型評価が書類作成のルーティンワークとなりかねない。

(6) 地域間比較

指標値を他の地域及び機関と比べることも、アウトカム指標の分析における常套手段の1つと位置づけられる。多数の地域の間での比較をハトリらはcomparative performance measurement（CPM）と呼び、地域差の分析の有用性を説く（Morley et al. 2001）。日本では業績測定型評価の枠内で明示的に他地域との比較を行う例は少なく、その方法も自地域と周辺地域の指標値の推移を折れ線グラフで示すなどに止まることが多い。

他地域との比較は実績の解釈において重要な情報を提供するが、その比較がエビデンスとしての性格をもつためには、やはり比較可能性が前提となる。

例えば同一の定義のアウトカム指標であっても、外部要因が影響を及ぼす場合、地域によって外部要因の影響度合いが異なれば比較は成り立た

ない。米国におけるCPMの取り組みで行われるように、回帰分析によって指標値を調整したadjusted performance measures (APMs) を用いる必要がある (Rubenstein et al. 2003)²³。

また、業績測定型評価に限らず、人口等地域ごとの規模の違いを調整する目的で、人口千人当たりなどの関係比率に加工して指標値を比較することも多い。しかし、実は関係比率の分子と分母が概ね比例関係にない関係比率として機能しない。精密な比較のためには、やはり回帰分析が必要となる場合がある (小野 2011a)。

各種指標の地域間比較においては、総合指標が作成される場合がある。Morley et al. (2001) はCPMとして扱っているが、方法論としてこれは長い歴史をもつ社会指標の領域である。日本でも様々な地域総合指標がこれまで作成されており、元の基礎指標をアウトカム指標とみなせばCPMとみることもでき、業績測定ともみなせる。ただし、複数の指標をどのようなウェイトで総合するかという重要課題を始め、地域総合指標の比較可能性の確保には、様々な統計的吟味が求められる (小野 2014)。

本章の冒頭で述べたように、業績測定型評価の指標はその実績値を比較することで判断に資する。目標値との比較とその結果の横断的比較は幅広く行われているが比較可能性の点で明らかな課題を抱えており、折角のエビデンスを台無しにしかねない。一方、内訳の比較と地域間比較は行われない事自体が課題であるが、比較可能性を担保して行えば有用なエビデンスが得られるはずである。

5. 業績測定型評価の今後に向けて

(1) インパクト評価・因果推論への接続

業績測定とプログラム評価 (特にインパクト評価) の関係については諸説あれど、両者は補完関係にあるとするのが今日の着地点であろう²⁴。

ハトリは単著・共著を通じ、補完関係として3点を挙げている²⁵。第一に、アウトカムが変化しても業績測定はその原因を明らかにせず、外部要因の影響もわからない。特に結果が目立って高い

か低い場合など、プログラム評価等の深い分析による説明が求められる。第二に、業績測定は実験や準実験デザインの評価に詳細なデータやその収集手段を提供できる。その際、4 (5) で述べたアウトカムを分割した内訳は有用である。第三に、プログラム評価の実行によって業績測定の評価指標やデータ収集法の改善をもらすことができる。第三の項目には、業績測定のために描いたロジックモデルの裏付けが得られること (Poister et al. 2015) も加えてよいだろう。

また、業績測定自体も因果推論と無関係な訳ではない。Hatry (2006) は、業績測定の枠内で事前事後比較や実験群・対照群の比較等の準実験、小規模な無作為配分による実験を行うことによって因果推論が可能な場合を掲げている。また、4 (6) で取り上げた回帰分析による指標値の地域間比較は、準実験の差の差分析と重なる部分がある。さらに3 (1) でも言及した通り、ロジックモデル上で最終的アウトカムに至る途中のブロックの測定指標値の変化は、因果効果としての最終アウトカム発現の必要条件でもある。

日本では、国・自治体を問わず業績測定型評価がすでに幅広く実行されている。エビデンス重視の機運のもと、その品質改善がなされれば、業績測定型評価と互いに補完する形でのインパクト評価の実行を促すことにもつながると期待したい。

(2) エビデンスを提供する業績測定へ

業績測定では、例えば①指標設定、②指標計算式や測定方法の決定、③目標値の設定、④実績値の測定、⑤実績値と目標値の比較、⑥目標達成状況の集約、という段階があるとして、ある段階の品質は論理的に次の段階の品質の必要条件となるが、エビデンス及びその解釈の提供という観点ではこの論理的関係は一層厳格なものとなろう。つまりある段階で問題があれば、それ以降の議論が意味を失う。

本稿は日本の業績測定の問題点をいちいち論じた観もあるが、それらの問題群は誤った意思決定をもたらしてきたであろうか。仮に不都合な事態が出来していなければ、業績測定の結果が本格的には活用されていないことの皮肉なエビデンスになってしまう。国・自治体において画期をなす改

革として導入されて以来四半世紀を経てマンネリズムに陥っている日本の業績測定型評価にとって、近年のEBPMはその使命を再確認する千載一遇の好機である。

なお、業績測定が提供する広義のエビデンスは、いわば因果推論にもとづく狭義のエビデンスを支える層である。一方、広義のエビデンスを支える層として欠かせないのが、「統計数字のウソがないこと」である。立て続けに露呈した統計不正の出鱈目ぶりをみるにつけ、評価の基礎データとなる統計数字には何らかの監査が必要というべきだろう（小野2022a、2024）²⁶。

これからの日本の政策にエビデンスは必要である。そこで業績測定型評価が果たすべき役割は広く、重い。Hatry and Newcomer（2015）曰く、評価において欠けているものの認識は評価の弱点ではなく、むしろよい評価の強みである。けだし金言である。

謝辞

本論文の初稿に対して匿名の査読者から、論旨を明確にする上で有益な指摘をいただいた。深く感謝の意を表する。

注記

- 1 後述するように定量的なエビデンスの提供とその解釈に係る段階からなるといえる。なお、これは業績測定型評価の「測定」部分といってもよい。英語の performance measurement もそうだが、業績測定という語には量を測るという測定そのものに加えて、しばしば測定した指標値を何らかの比較などにより解釈・分析する過程を含む。広く読まれたハトリの1999年のテキスト Performance Measurement も5分の3は指標値の測定後の話である。
- 2 政府・自治体における政策評価の現場のエビデンスの解釈も概ねこのようなものであろう。実際に、政府が具体的にEBPMに取り組む端緒となった統計改革推進会議による最終取りまとめは、EBPMの基本的な考え方について、「統計等データなどの客観的な証拠」を用いて政策課題の把握、政策効果の予測・測定・評価を行うものと説明する（統計改革推進会

議 2017）。このような広義のエビデンスには、因果関係を統計学的に実証した根拠などと定義される狭義のエビデンスも当然部分集合として含まれる。なお、業績測定も、本稿の後段で触れるように狭義エビデンスの把握に資することが期待され、狭義エビデンスと無縁のものではない。

Urban Institute（1980）には、業績測定は定量的なエビデンスを提供するものとの記述がある。その後のEBPMに即した表現ではないが、本稿の定義と重なるものといえる。

- 3 例として、典型的な課題を列挙した小野（2011b）、府省の実績測定を点検して課題を指摘した小野（2013）、都道府県についての小野（2008、2016）、中核市・施行時特例市についての小野（2022b）を挙げる。
- 4 例えば、Hatry（1980、2008）、Wholey（2006）、Van Dooren et al.（2010）など。
- 5 プログラム評価論の妥当性の概念を詳細に論じたものとして、評価デザインと統計解析法を統合して体系的に述べたLangbein（2012）、プログラム評価の一段階としての業績モニタリング・評価における妥当性についても併せて論じたMcDavid et al.（2013）、Stufflebeam and Coryn（2014）などを挙げる。
- 6 妥当性概念をこのように2段階に分ける考え方は、より一般に測定理論（measurement theory）とも合致する（Brewer 2006）。
- 7 業績測定型評価において、常に統計的検定が求められる訳ではない。判断に必要な情報が求められる（Hatry et al. 1997）。
- 8 Poister（1999）及びBrewer（2006）は本稿と概ね同趣旨の3条件を掲げている。
- 9 今回、効率を測る指標について参照した文献（Swiss 1995; CAP 1998; Adamaschek 1998; Hatry 2006; Poister 2003; Van Dooren et al. 2010; Pidd 2012; McDavid et al. 2013）では、驚くべきことにその分類や用語（efficiency、productivity、cost-effectiveness等）の定義、効率を「出力÷入力」と「入力÷出力」のどちらで計算するかまで全く区々であった。
- 10 Hatry（1980、2006、2014）、Hatry et al.（1990）、Poister（2003）を参照した。
- 11 米国の会計検査院（GAO）は、かつて業績測定を「インプット・作業（活動）・アウトプット・アウトカム・効率等の一連のデータを定期的に収集し報告するこ

- と」(U.S. GAO 1992)と定義したが、その後「プログラムの実績、特に事前に設定した目標や基準の達成度合いについての体系的で継続的なモニタリングと報告」(U.S. GAO 2012)としている。
- 12 方法を具体的に解説している文献として、例えば Hatry (2006)、Poister (1995, 2003)、Bovaird (1996)、Van Dooren et al. (2010)。
- 13 オレゴン・ベンチマークでは第Ⅱ期への改定にあたり、目標値設定の姿勢を「積極的だが現実的」とし、調整や協議のベースとする水準について、上昇基調か下降基調か、標準型か積極型という観点から4通りの計算法を設定していた(小野・田淵2001)。
- 14 オレゴン州のように目標値設定の共通の基本手順を明らかにした事例はあるが、今回参照した欧米の文献では、目標達成の難易度を個別または統一的に明示すべきだという議論は見つからなかった。合意形成の重要性は多く指摘され、目標値は「合理的」(Poister 1995)で「やや野心的であると同時に実際に達成可能」(Poister 2003)であるべきなどの議論はある。なお4(5)で取り上げるアウトカムの内訳に関する議論において、Hatry (2006)及びHatry et al. (1990)は、1つのプログラムの中の何らかの区分ごとの業務遂行の難易度を考慮せずに達成状況を見ることの問題や難易度を踏まえた目標設定の必要性を指摘している。
- 15 フロー指標とストック指標は、本文で述べる通り両者を区別せずに比較することが問題となる。無造作な比較を避けるためには指標設定時の区別が有効であり、フロー指標に統一するという方法もあることから指標設定時の課題と位置づけることもできるが、指標間の比較を行わない業績測定(必ずしも両者を区別する必要がない)があること、比較時にストック指標をフロー指標に変換する等の対応法があることから、本稿では「実績値の比較」の課題として位置づけている。
- 16 今回参照した欧米の文献では、フローとストックの区別が必要であるという議論は見当たらなかった。少なくとも重要な課題とされていない模様である。その理由としては、個々の指標の目標達成状況の分析が重視されていること、指標同士の比較は1つのプログラム内が多いこと、プログラム間の比較は同一の指標や類似のプログラムを対象にすることが多いこと、目標達成状況の一覧表はあってもそれを集計・集約することがないこと、などが考えられる。
- 17 欧米の文献には、目標達成状況を含む測定結果の解析に関する記述にページ数を割くものが少なくないが、目標達成度の計算については、今回参照した範囲でHatry (2006)及びHatry et al. (2004)が「実績値－目標値」という引き算を、Liner et al. (2001)及びRosen (1995)がそれぞれ「実績値÷予測値」「実績値÷期待値」という割り算を示した例にとどまる。その背景にあるものとして、目標達成状況の吟味において、計算よりも数字そのものの解釈や議論に重点が置かれていること、現場の行政職員や意思決定者に数学的な思考を求めないこと(一方、後述する指標値自体のクロスセクション比較を行うとなれば回帰分析等の統計解析が必要となる)などが考えられる。
- 18 ③はフロー指標を想定した計算式である。ストック指標の場合は②の計算がこれに相当し、その場合フロー指標の②の計算に相当するのはストック指標をフロー指標に変換してからの②の計算となる。
- 19 目標値と実績値を並置した棒グラフを並べたものや注17で述べた素朴な計算結果を並べたものはある。
- 20 目標達成度ではなく、評価指標の測定値自体を比較し総合する作業は、同一ないし類似の指標同士でない限り、業績測定の枠内では定性的なものとならざるを得ない。ハトリーらも、第一義的に主観的なものとする(Hatry et al. 2004)。
- 21 16本の著作を年代順に挙げると、Hatry et al. (1981)、Hatry et al. (1990)、Hatry et al. (1992)、NAPA (1994)、Hatry et al. (1994)、Hatry (1996, 1997)、Hatry and Kopczynski (1997)、Hatry (1999)、Liner et al. (2001)、Hatry (2002)、Hatry et al. (2003)、Hatry et al. (2004)、Hatry (2006, 2008, 2014)。
- 22 日本でアウトカムを分割した内訳に着目する必要性や重要性を指摘したものとして、小野 (1999, 2004)、田中 (2014)など。
- 23 外部要因等の変数で実績値を回帰した残差を指標(APMs)とする。なお、回帰式に処置群・対照群の別や事前・事後の時期を表すダミー変数を導入すれば、厳密なエビデンスを指向する準実験の1つとして多く行われる「差の差分析」となる。
- 24 両者の関係を周到に論じたものとして、Hatry (2013)、Hunter and Nielsen (2013)、Nielsen and Hunter (2013)など。

- 25 第一の点は最初期のUrban Institute (1980)、Wholey and Hatry (1992)等10本以上の著作で、Hatry (2006)及びHatry et al. (2015)は第一・第二の点、Hatry (2013)は3点すべてについて述べている。
- 26 ハトリとホーリーも、指標の妥当性・信頼性ととともに、数字の操作や過誤の可能性を最小化するために、財務データと同様に定期的な監査をすべきであると述べている (Wholey and Hatry 1992; Wholey 2006)。

参考文献

- 上野宏・上野真城子 (2007) 「パフォーマンス・メジャーメント」、三好皓一編『評価論を学ぶ人のために』、世界思想社、61-78
- 小野達也 (1999) 「行政評価の実践テクニック」、島田晴雄・三菱総合研究所政策研究部『行政評価：スマート・ローカル・ガバメント』、197-243
- 小野達也 (2004) 「評価指標のかんどころ」、梅田次郎・小野達也・中泉拓也編『行政評価と統計』、日本統計協会、173-213
- 小野達也 (2008) 「都道府県の評価の10年ー定量評価のこれまでとこれから」、『日本評価研究』、8 (1) : 19-38
- 小野達也 (2011a) 「自治体ベンチマーキングと指標の比較可能性」、『日本評価研究』、11 (2) : 13-30
- 小野達也 (2011b) 「業績測定型評価における目標設定と達成度評価の妥当性ー行政評価の形骸化を避けるための条件」、『地域学論集 (鳥取大学地域学部紀要)』、8 (2) : 1-25
- 小野達也 (2013) 「政策評価と実績測定ー府省の実績測定における計量・計数を巡ってー」、『日本評価研究』、13 (2) : 21-36
- 小野達也 (2014) 「都市ランキングの虚と実ー地域総合指標による比較を巡ってー」、『地域学論集 (鳥取大学地域学部紀要)』、11 (2) : 1-27
- 小野達也 (2016) 「自治体における業績測定型評価の現状と課題ー20年を経過した都道府県の取り組みの点検結果からー」、『日本評価研究』、16 (1) : 3-16
- 小野達也 (2022a) 「EBPMの傍らで統計不正とは」、『評価クォーターリー』、61 : 1
- 小野達也 (2022b) 「業績測定による評価における活用ー測定結果は活用するに足りるのか」、『日本評価学会第23回全国大会発表要旨集録』、209-212
- 小野達也 (2023) 「国の府省の政策評価のパラダイム転換に向けてー評価制度とEBPMの関係を巡ってー」、『日本評価研究』、23 (2) : 3-4
- 小野達也 (2024) 「統計が政策のエビデンスとなる条件：遥かなるEBPMの呼び声」、『統計』、75 (1) : 4-12
- 小野達也・田淵雪子 (2001) 『行政評価ハンドブック』、東洋経済新報社
- 田中啓 (2014) 『自治体評価の戦略：有効に機能させるための16の原則』、東洋経済新報社
- 田辺国昭 (2001) 「実績評価の課題」、『都市問題研究』、53 (10) : 100-114
- 統計改革推進会議 (2017) 『統計改革推進会議 最終取りまとめ』 https://www.kantei.go.jp/jp/singi/toukeikaikaku/pdf/saishu_honbun.pdf (2024年9月1日閲覧)
- 西尾勝 (1976) 「効率と能率」、辻清明他編『行政の過程』 (行政学講座3)、東京大学出版会
- Adamaschek, B. (1998). *Intermunicipal Comparative Performance Measurement: Innovation through Competition*. Bertelsmann Foundation Publishers.
- Ammons, D. N. (1995). Performance Measurement in Local Government. In D. N. Ammons, (Ed.), *Accountability for Performance: Measurement and Monitoring in Local Government*. International City/County Management Association, 15-32.
- Bovaird, T. (1996). The Political Economy of Performance Measurement. In A. Halachmi & G. Bouckaert (Eds.), *Organizational Performance and Measurement in the Public Sector: Toward Service, Effort and Accomplishment Reporting*. Conn: Quorum Books, 145-166.
- Brewer, G. A. (2006). All Measures of Performance are Subjective: More Evidence on US Federal Agencies. In G. A. Boyne et al. (Eds.), *Public service performance: Perspectives on measurement and management*. New York: Cambridge University Press, 35-54.
- Center for Accountability and Performance (1998). *Performance Measurement Concepts & Techniques*. American Society for Public Administration.
- Greiner, J. M. (1996). Positioning Performance Measurement for the Twenty-First Century. In A. Halachmi & G. Bouckaert (Eds.), *Organizational Performance and*

- Measurement in the Public Sector: Toward Service, Effort and Accomplishment Reporting*. Conn: Quorum Books, 11-50.
- Hatry, H. P. (1980). Performance Measurement Principles and Techniques: An Overview for Local Government. *Public Productivity Review*, 4(4), 312-339.
- Hatry, H. P. (1996). Foreword. In A. Halachmi & G. Bouckaert (Eds.), *Organizational Performance and Measurement in the Public Sector: Toward Service, Effort and Accomplishment Reporting*. Conn: Quorum Books, ix-xii.
- Hatry, H. P. (1997). Where the Rubber Meets the Road: Performance Measurement for State and Local Public Agencies. In K. E. Newcomer (Ed.), *Using Performance Measurement to Improve Public and Nonprofit Programs. New Directions for Evaluation*, 75, 31-44.
- Hatry, H. P. (1999). *Performance Measurement: Getting Results*. Washington, D.C.: The Urban Institute Press. (上野宏・上野真城子訳 (2004) 『政策評価入門－結果重視の業績測定』、東洋経済新報社)
- Hatry, H. P. (2002). Performance Measurement: Fashions and Fallacies. *Public Performance & Management Review*, 25(4), 352-358.
- Hatry, H. P. (2006). *Performance Measurement: Getting Results* (2nd ed). Washington, D.C.: The Urban Institute Press.
- Hatry, H. P. (2008). Epilogue: The Many Faces of Use. In W. Van Dooren & S. Van de Walle (Eds.), *Performance Information in the Public Sector: How It Is Used*. London: Palgrave Macmillan, 227-240.
- Hatry, H. P. (2010). Looking Into the Crystal Ball: Performance Management Over the Next Decade. *Public Administration Review*, 70, s208-211.
- Hatry, H. P. (2013). Sorting the Relationships Among Performance Measurement, Program Evaluation, and Performance Management. In S. B. Nielsen & D. E. K. Hunter (Eds.), *Performance Management and Evaluation. New Directions for Evaluation*, 137, 19-32.
- Hatry, H. P. (2014). *Transforming Performance Measurement for the 21st Century*. The Urban Institute, <http://urban.org>. (2024年8月23日閲覧)
- Hatry, H. P., Blair, L. H., Fisk, D. M., Greiner, J. M., Hall, J. R., and Schaenman, P. S. (1992). *How Effective Are Your Community Services? Procedures for Measuring Their Quality* (2nd ed). Washington, D.C.: The Urban Institute and International City/County Management Association.
- Hatry, H. P., Cowan, J., and Hendricks, M. (2004). *Analyzing Outcome Information: Getting the Most from Data*. The Urban Institute, <http://urban.org>. (2024年8月26日閲覧)
- Hatry, H. P., Fountain, J.R., and Sullivan, J. M. (1990). Overview. In H. P. Hatry et al. (Eds.) *Service efforts and accomplishments reporting: Its time has come: An Overview*. Conn: Governmental Accounting Standards Board.
- Hatry, H. P., Gerhart, C., and Marshall, M. (1994). Eleven Ways to Make Performance Measurement More Useful to Public Managers. *Public Management*, 76, s15-18.
- Hatry, H. P., and Kopczynski, M. (1997). *Guide to Program Outcome Measurement for the U.S. Department of Education*. Washington, D.C.: U.S. Department of Education.
- Hatry, H. P., Marcotte, J. E., van Houten, T., and Weiss, C. H. (1997). *Customer Surveys for Agency Managers: What Managers Need to Know*. Washington, D.C.: The Urban Institute Press.
- Hatry, H. P., Morley, E., Rossman, S. B., and Wholey, J. S. (2003). *How Federal Programs Use Outcome Information: Opportunities for Federal Managers*. Washington, D.C.: IBM Endowment for The Business of Government.
- Hatry, H. P., and Newcomer, K. E. (2015). Pitfalls in Evaluations. In K. E. Newcomer, H. P. Hatry & J. S. Wholey (Eds.), *Handbook of Practical Program Evaluation* (4th ed). San Francisco, Calif.: Jossey-Bass Publishers, 701-723.
- Hatry, H. P., Newcomer, K. E. and Wholey, J. S. (2015). Evaluation Challenges, Issues, and Trends. In K. E. Newcomer, H. P. Hatry & J. S. Wholey (Eds.), *Handbook of Practical Program Evaluation* (4th ed). San Francisco, Calif: Jossey-Bass Publishers, 816-832.
- Hatry, H. P., Winnie, R. E., and Fisk, D. M. (1981). *Practical Program Evaluation for State and Local Governments* (2nd ed). Washington, D.C.: The Urban Institute Press.
- Hunter, D. E. K. and Nielsen, S. B. (2013). Performance Management and Evaluation: Exploring Complementarities. In S. B. Nielsen & D. E. K. Hunter

- (Eds.), *Performance management and evaluation. New Directions for Evaluation*, 137, 7-17.
- Langbein, L. (2012). *Public Program Evaluation: A Statistical Guide* (2nd ed). New York: M.E.Sharpe.
- Liner, B. et al. (2001). *Making Results-Based State Government Work*. Urban Institute, <http://urban.org>. (2024年8月8日閲覧)
- Markiewicz, A. and Patrick, I. (2016). *Developing Monitoring and Evaluation Frameworks*. Thousand Oaks, Calif: Sage.
- McDavid, J. C., Huse, I., and Hawthorn, L. R. L. (2013). *Program Evaluation and Performance Measurement: An Introduction to Practice* (2nd ed.). Thousand Oaks, Calif: Sage.
- Morley, E., Bryant, S. P., and Hatry, H. P. (2001). *Comparative Performance Measurement*. Washington, D.C.: The Urban Institute Press.
- NAPA Advisory Panel On Improving Government Performance (1994). *Toward Useful Performance Measurement: Lessons Learned from Initial Pilot Performance Plans Prepared Under the Government Performance and Results Act*. National Academy of Public Administration.
- Nielsen, S. B. and Hunter, D. E. K. (2013). Challenges to and Forms of Complementarity Between Performance Management and Evaluation. In S. B. Nielsen & D. E. K. Hunter (Eds.), *Performance Management and Evaluation. New Directions for Evaluation*, 137, 115-123.
- Pidd, M. (2012). *Measuring the Performance of Public Services: Principles and Practice*. New York: Cambridge University Press.
- Poister, T. H. (1995). Productivity Monitoring: Systems, Indicators, and Analysis. In D. N. Ammons (Ed.), *Accountability for Performance: Measurement and Monitoring in Local Government*. International City/County Management Association, 98-123.
- Poister, T. H. (1999). Developing Performance Indicators for the Pennsylvania Department of Transportation. In R. C. Kearney & E. M. Berman (Eds.), *Public Sector Performance: Management, Motivation, and Measurement*. Oxford: Westview Press, 341-362.
- Poister, T. H. (2003). *Measuring Performance in Public and Nonprofit Organizations*. San Francisco, CA: Jossey-Bass.
- Poister, T. H., Aristigueta, M. P., and Hall, J. L. (2015). *Managing and Measuring Performance in Public and Nonprofit Organizations: An Integrated Approach* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Radnor Z. (2008). Hitting the Target and Missing the Point? Developing an Understanding of Organizational Gaming. In W. Van Dooren & S. Van de Walle (Eds.), *Performance Information in the Public Sector: How It Is Used*. London: Palgrave Macmillan, 11-23.
- Rosen, E. D. (1995). Measuring Productivity. In Ammons, D. N. (Ed.), *Accountability for Performance: Measurement and Monitoring in Local Government*. International City/County Management Association, 44-65.
- Roy, C.-H. and John, L. (1996). Societal Outcome Measurement: the European Dimension. In P. Smith (Ed.), *Measuring Outcome in the Public Sector*. London: Taylor and Francis.
- Rubenstein, R., Schwartz, A. E., and Stiefel, L. (2003). Better than Raw: A Guide To Measuring Organizational Performance with Adjusted Performance Measures. *Public Administration Review*, 63(5), 607-615.
- Smith, P. C. (2006). Quantitative Approaches Towards Organizational Performance. In G. A. Boyne et al. (Eds.), *Public Service Performance: Perspectives on Measurement and Management*, New York: Cambridge University Press, 75-91.
- Stufflebeam, D. L., and Coryn, C. L. S. (2014). *Evaluation Theory, Models, and Applications* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Swiss, J. E. (1995). Performance Monitoring Systems. In D. N. Ammons (ed.), *Accountability for Performance: Measurement and Monitoring in Local Government*. International City/County Management Association, 67-97.
- Urban Institute (1980). *Performance Measurement: A Guide for Local Elected Officials*. Washington, D.C.: The Urban Institute Press.
- U.S. General Accounting Office (1992). *Program Performance Measures: Federal Agency Collection and Use of Performance Data*. General Accounting Office Report GAO/GGD-92-65.
- U.S. Government Accountability Office (2012). *Designing Evaluations: 2012 Revision*. Government Accountability

Office Report GAO-12-208G.

Van Dooren, W., Bouckaert, G., and Halligan, J. (2010).

Performance Management in the Public Sector. London: Routledge.

Wholey, J. S. (2006). Quality Control: Assessing the Accuracy and Usefulness of Performance Measurement Systems. In H. P. Hatry, *Performance Measurement: Getting Results* (2nd ed). Washington, D.C.: The Urban Institute Press, 267-286.

Wholey, J. S., and Hatry, H. P. (1992). The Case for Performance Monitoring. *Public Administration Review*, 52(6), 604-610.

(2025.1.13 受理)

Bridging the Gap Between the Practice of Evidence-Based Policy Making and Mannerism in Performance Measurement

Tatsuya Ono

Otemon Gakuin University
t-ono@otemon.ac.jp

Abstract

Management by performance measurement seems to have spread throughout all public sectors over the past 30 years. In recent years, however, the routinization and fatigue of its management processes have been topics of discussion. On the other hand, performance measurement also has been assigned an important role by the EBPM-oriented government. This paper examined the process of measurement in performance management practices in Japan, with reference to the literature by Harry P. Hatry and others. Then, dos and don'ts in measurement process were newly presented from the perspective of evidence. Future policies in Japan have to be evidence-based. Performance measurement should play wide and important role in EBPM initiatives.

Keywords

Performance Measurement, Evidence, Performance Indicator, Harry P. Hatry

【Article】

Examining the Appropriateness of Using Year-Over-Year Accident Numbers to Evaluation the Effectiveness of Product Safety Administration

Shunsuke Saito

Nagoya Institute of Technology,
Department of Engineering
s.saito.278@stn.nitech.ac.jp

Hironobu Kawamura

Nagoya Institute of Technology,
Department of Architecture
kawamura@nitech.ac.jp

Abstract

Predictive indicators are crucial for budget planning and outcome-based performance measurement. In Japan's product safety administration, the year-over-year change in the number of product accidents is often used to evaluate government agency activities. However, predicting product accidents is inherently variable. This study examines the predictive accuracy of using the previous year's data as forecasts and compares it with the actual target value of a "3% reduction." The results show that such forecasts have an average prediction error of approximately 8%, which exceeds the target value. This finding suggests the irrationality of past performance evaluation methods. Goal setting often includes reduction rates from predictive values, requiring careful consideration of prediction errors. However, achieving goals significantly different from prediction errors may be unrealistic. In outcome-based evaluations, goals must be distinguishable from prediction errors and attainable. If these criteria are not met, a reassessment of outcome-based evaluation methods may be necessary.

Keywords

Performance measurement, Product Safety, Accident Prediction, Impact Assessment

1. Introduction

1.1 Accidents Caused by Consumer Products

Consumer products, such as bicycles, refrigerators, and toys, are indispensable in modern life, but accidents do occur involving these products. Some of these accidents are relatively minor, such as minor burns caused by overheating product parts, whereas others are more serious, such as fires caused by internal product components and bone fractures caused by bicycle parts detaching and causing falls. Approximately 2,000 accidents involving consumer products occur in Japan each year, resulting in 40 to 50 deaths each year (National Institute of Technology and Evaluation [NITE], 2022, 2023).

Although the beginning of the recognition of product accidents as a social issue in Japan is unclear, the relevant department of the Ministry of Economy, Trade, and Industry (METI) traces its roots to the enactment of laws related to this issue and the evolution of these laws (Product Safety Division, Commerce and Information Policy Bureau, METI, 2001). The Consumer Product Safety Law, a comprehensive law on product safety, was enacted in 1973. During this period, accidents involving consumer products, such as glass bottle explosions and climbing lifeline failures, were being widely reported by the mass media. This was also a period of economic growth in Japan, when people's incomes were rising and technological advances were increasing the complexity and sophistication of consumer products. As products became increasingly complex, making judgments about product safety became more difficult for consumers, and product-caused accidents and complaints started rising. The law was enacted against this background.

At the time of its enactment, the Consumer Product Safety Act established a government certification system with which the government verified whether a product conformed to statutory technical standards. However, the system was changed in 1999 so that, in principle, manufacturers must voluntarily confirm conformity to technical standards. This was partly because verifying product safety had become inefficient for the government due to improvements in product safety and quality and enhancements in the technological capabilities of inspection agencies. Moreover, the increased number of products in circulation had made such verification impractical for the government.

Another reason for this change was liability in the event of an accident. The use of the government certification system required the government to check the safety of a product before distribution and give it a passing or failing result. This meant that the government was responsible for the safety of the product. This also muddled the scope of responsibility of the manufacturer such that the government could not issue an order to the manufacturer to take measures to prevent accident recurrence. Subsequently, a number of precedents related to product accident trial accumulated, and rulings that manufacturers should be responsible for the safety of their products led to a reconsideration of the scope of their responsibilities. Thus, the system was changed as described above, and product safety is now the primary responsibility of product manufacturers or importers.

The Japanese government deals with product accidents mainly through the Consumer Affairs Agency and the METI. The NITE, an independent administrative agency under the jurisdiction of the METI, cooperates with METI in various activities. The main roles of government agencies in product safety are the formulation of rules for ensuring safety and the execution of appropriate postmarketing measures, such as recall orders. Therefore, the government is implementing various measures, the most important of which are the collection of accident information, development of safety standards, and educational activities.

As for information collection, when a product accident occurs, the government collects information about the product accident from the product manufacturers and importers and local authorities, including the fire and police departments. The purpose is to share accident information with society as a whole to enable the government, manufacturers, and consumers to play their appropriate roles in creating a safe society. The information gathered is also used by the government to consider the issuance of orders to prevent recurrence and establish safety standards. Reporting to the government has been mandatory since 2007 for serious product accidents, namely, those that result in death, serious injury (requiring treatment for more than 30 days), permanent injury, carbon monoxide poisoning, or fires. This was made mandatory because accident information was previously rarely provided by manufacturers to consumers and the government, leading to delays in government response (Consumer Affairs Agency & METI, 2018a). In addition, the NITE investigates the causes of product accidents, whether serious or nonserious. This allows the government to understand the actual status of accidents in the market, provide accurate guidance to manufacturers, and take measures to prevent recurrence.

As for the development of safety standards, the following laws have been established in Japan regarding technical

standards for consumer products:

- Consumer Product Safety Act
- Electrical Appliances and Materials Safety Act
- Gas Business Act
- Act on the Securing of Safety and the Optimization of Transaction of Liquefied Petroleum Gas

These laws establish technical standards for safety for certain products and prohibit the distribution of products that do not conform to these standards. For example, a gas stove with two or more burners must be equipped with an overheat protection device to prevent fire accidents caused by forgetting to turn off the burner, and lighters must have child-resistant structures to prevent accidental operation by children. Manufacturers are required to have their products inspected for compliance with these standards, either by themselves or by a certification body, and to keep documentary evidence of such inspection for a certain period of time. The government verifies the compliance of manufacturers with the law through on-site inspections. In addition, the government may establish new technical standards or modify existing standards in light of accidents that occur in the market. These technical standards play an important role in product safety (Magat & Moore, 1995) and have been proven effective in many cases (Hanway & Rodgers, 2020; Moore & Magat, 1996; Rodgers, 2002, 2022; Rodgers & Adair, 2009; Rodgers & Topping, 2012; Smith et al., 2002; Viscusi & Cavallo, 1996; Viscusi & Cavallo, 1994; Viscusi & Dalafave, 2022). However, if the set criteria are inappropriate, they may be ineffective or may even be counterproductive (Magat & Moore, 1995; Morrongiello et al., 2007; Rascoff & Revesz, 2002; Rodgers, 1988, 1996; Viscusi, 1984).

The government conducts educational activities, which are mainly categorized into consumer- or manufacturer-oriented ones, for product safety. Consumer-oriented educational activities mainly include publicity to alert consumers. These alerts are about accidents that are confirmed to be highly dangerous and likely to occur frequently or accidents that are expected to occur frequently in the coming season. The government provides publicity to alert consumers, encouraging them to change their behavior and prevent accidents. The NITE's press releases play a particularly large role in this activity. The institute holds press conferences to alert the public, and each company disseminates the given information through its own media. The information therefore reaches a large number of citizens. The NITE's press releases are generally held once a month, providing the media with statistics related to accidents, materials summarizing the accident mechanisms, and accident reenactment videos, which are then reported on TV, newspapers, and the Internet.

The government prepares and publishes guidelines for manufacturers to ensure that they properly fulfill their responsibilities regarding product safety. These include handbooks that describe the procedures and basic responsibilities required in the event of a product accident, recommendations for manufacturers to work on product safety, and the basic procedures for conducting risk assessments. The METI also fosters a corporate culture of product safety, such as by awarding companies that implement outstanding product safety efforts.

1.2 Difficulty and Current Status of Product Accident Prediction

The government aims to prevent product accidents through the aforementioned activities. However, these initiatives require personnel and financial resources, whose budget is accordingly planned by the government. Budget planning is ideally based on predictions of product accidents for the next fiscal year. However, the number of product accidents (number of accidents) is difficult to predict for many reasons, with the main ones being the stochastic nature of product accidents and the diversity of product content.

Product accidents are stochastic by nature. For example, some people will have accidents while others will not, even when using the same product. A person may not have an accident one day but may experience one on another day. This stochastic nature makes it difficult to predict the occurrence of product accidents or explain past trends.

Product accidents are diverse. In Japan, products that can be purchased and generally used by consumers are typically classified as consumer products; the few exceptions include foodstuffs, pharmaceuticals, and automobiles. Accidents vary widely, including injuries, fires, and carbon monoxide poisoning. Furthermore, accidents have wide-ranging causes, such as the overloading of internal parts due to poor product design, ignition due to part defects, and mishandling due to improper interfaces. Nonetheless, some accidents are not caused by products, such as those due to user carelessness, lack of common sense, and natural disasters.

Because of these various types of product accidents, the occurrence trends of product accidents have a highly complex background. Predicting the number of accidents entails constructing an actuarial model that considers these backgrounds. However, a large amount of resources is needed to investigate explanatory variables covering these backgrounds and construct a mathematical model. Researchers have rarely attempted to do so; as far as we have ascertained, only Visucci (1985) has published a related study, where increases in product accidents were explained by an increase in household disposable income.

These characteristics hinder the prediction of the number of accidents and proper budget allocation. As mentioned earlier, NITE conducts accident investigations, which constitute one of the major costs. The costs associated with these investigations, particularly labor costs, are budgeted based on forecasts. Overinvesting may consume budget that should be used in other areas, whereas underinvesting may result in dysfunctional product safety administration and threaten public safety. For example, after a continuous decrease in the number of accidents for a certain period of time, if the government naively assumes that accidents will continue to decrease in the future and does not allocate the minimum necessary budget, then product safety administration may become dysfunctional.

Predictions can also be the basis of performance measurement. For example, if the number of accidents decreased in a given year, then the interpretation of the obtained data will completely differ depending on whether one considers that the decrease occurred despite a predicted increase or that the decrease matched the prediction. In the former case, the decrease may be assumed that government activities such as NITE's press releases had some influence on the decrease; however, this is difficult to assume in the latter case. However, as discussed later, studies on the prediction of the number of accidents remain limited. The US Consumer Product Safety Commission (CPSC), the public agency responsible for product safety in the US, does not be evaluated performance based on an increase or decrease in the number of accidents, at least since 2014, when documents are currently available on the CPSC website. This may be due to the aforementioned circumstances (Agency Reports, n.d.).

However, the performance measurements for NITE often sets a target for the year-on-year comparison of indicators related to product accidents. For example, in 2019, a target of reducing the number of serious product accidents based on the Consumer Products Safety Act by 3% from the previous year was set in performance measurements for NITE (NITE, 2020). The rationale for setting such a goal may be the growing trend of evaluating government agencies on an outcome basis. However, even if a 3% decrease occurred, it may fall within the error range due to the stochastic nature of product accidents. If the target of 3% were based on the stochastic nature of the number of accidents, then there is no problem. Otherwise, then this indicator is based on insufficient consideration. In addition, targets were set for 2019 and beyond based on year-on-year comparisons of various indicators related to the number of accidents. In light of this, studies should be conducted to validate the prediction based on the actual results of the previous year. This will optimize the investment of resources in administrative activities and ensure the appropriateness of performance measurement. However, as far as the author can ascertain, no study has been found that examines performance measurement from this perspective.

The results of performance measurements are not independent of resource allocation. Regarding the evaluation of incorporated administrative agencies and resource allocation, the Ministry of Internal Affairs and Communications outlines the following two points in the Guidelines for the Evaluation of Incorporated Administrative Agencies: 1)

agencies are expected to use the evaluation results to revise internal budget allocations, operational methods, and personnel treatment; and 2) the competent minister is required to appropriately incorporate the evaluation results into budget requests, including the calculation of operational subsidies (Ministry of Internal Affairs and Communications, 2022). In other words, the accuracy of goal-setting serves as the foundation for appropriate resource allocation. In the field of social sciences, causal relationships are often not clearly established. Additionally, administrative organizations may face situations where they must set goals without sufficient validation due to limited resources available for goal-setting processes. As a result, it might be considered inevitable that some goals are established without fully considering causal relationships or statistical properties. However, in areas related to the quality of life of citizens, particularly safety, using inadequately validated goals as a basis for resource allocation poses significant risks. Therefore, this study examines the prediction accuracy of the number of accidents forecasted by the previous year's values. In addition, through a comparison of the analysis results and the 3% reduction targeted in performance measurement for the NITE in 2019, the validity of the target value at that time is verified, and performance measurement using the predicted values is discussed. It should be noted that this study focuses on the "performance measurement of administrative organization", rather than the "evaluation of individual policies". While these areas share many similarities, they are based on different fundamental guidelines and often differ (Ministry of Internal Affairs and Communications, 2022, 2023).

2. Method

2.1 Data

In this study, data provided by the NITE are analyzed. The NITE collects information about product accidents that occur in Japan and has one of the largest product accident databases in the country. In Japan, manufacturers and importers of products are required to report product accidents classified as serious (referred to as "serious product accidents") if they result in death, severe injury requiring 30 or more days of treatment, permanent disability, carbon monoxide poisoning, or fire. NITE also obtains this information and compiles it into a database. The NITE's data for each accident contain product information, such as the model and manufacturer of the product that caused the accident, and the results of the accident investigation, such as the nature and cause of the accident.

This study analyzes the annual number of accidents from 2008. Because, the current accident information collection rules were established in April 2007. Serious accidents had occurred before then, and the data had been accumulated at NITE. Nevertheless, data collected before April 2007 are not used due to the considerable change in the ability to collect accident information. Before the 2007 introduction of the mandatory reporting of serious product accidents, whether or not to report a product accident to the NITE was left to the discretion of the manufacturer. However, because reporting was voluntary, accidents were almost never reported to the NITE (Consumer Affairs Agency & Ministry of Economy, Trade and Industry, 2018b). Therefore, data preceding April 2007 were excluded from this study because their basic properties may differ from those of data from the covered period.

In addition, data from January to March are missing for 2007. Furthermore, the number of serious accidents was uniquely low around April 2007. This may be due to the fact that it was immediately after the mandatory reporting of serious product accidents, and may be due to the perception of the operators who are obliged to report, or due to the system on the part of the administrative agencies. Therefore, 2007 was also excluded from the analysis in this study.

This study aims to determine the basic properties of product accident prediction. Conducting analysis over a period of time that includes an event with high social impact, such as the COVID-19 pandemic, the results may be strongly influenced by this event. Therefore, the analysis was conducted using monthly data on the number of serious

product accidents through December 2019, just before the impact of COVID-19 appeared.

This study also aims to verify the prediction accuracy of the forecasts based on the previous year's numbers and provide insights into the goal setting of performance measurement. The discussion includes goal of the performance measurement for NITE in 2019: reducing the number of serious product accidents by 3% relative to the previous year. Therefore, nonserious product accidents are excluded from the data in this study, and modeling is only performed for the number of serious product accidents. Therefore, this study included the annual number of serious accidents from 2008 to 2019 in the analysis. METI also discloses the number of reported serious product accidents in various contexts. However, the aggregation rules differ, which may result in slight discrepancies from the data used in this study. For instance, some serious product accidents are erroneously reported and accepted. These are included when aggregated as "reported cases" but excluded when aggregated as "occurrence cases." In this study, to ensure an accurate count of "occurrence cases," the data were independently aggregated from NITE's database.

2.2 Modeling and Accuracy Examination Current Forecasting Approach

As noted in Section 1.2, performance measurement for the NITE in 2019 set a goal of a 3% reduction in the number of accidents from the previous year. This goal was established to evaluate the accident prevention effects brought about by NITE. When evaluating the effect of an intervention, it is crucial to compare the actual observed outcomes after the intervention with the outcomes that would have been observed had the intervention not been implemented. Since the latter cannot be directly observed, it is typically simulated under various assumptions. Applying this concept, this goal implicitly assumes that if the activities in a given year are the same as those in the previous year, then the number of accidents will be the same as that in the previous year. This can be expressed as the following prediction model:

$$\hat{Y}_t = Y_{t-1} \quad (1)$$

Where Y_t denote the annual number of accidents and $\hat{Y}_t$ denote the annual predicted number of accidents. In this study, we considered this as a kind of forecasting model and verified its forecasting accuracy. However, it is important to note that NITE has not explicitly stated that such an assumption was made when setting the goal. This interpretation applies the framework of statistical impact evaluation methods to the performance measurement method that is the focus of this study.

Various indicators are used to evaluate forecast accuracy (Ramos et al., 2015). Three representative ones are used in this study: the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Each indicator is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{11} \sum_{T=2009}^{2019} (Y_T - \hat{Y}_T)^2} \quad (2)$$

$$MAE = \frac{1}{11} \sum_{T=2009}^{2019} |Y_T - \hat{Y}_T| \quad (3)$$

$$MAPE = \frac{1}{11} \sum_{T=2009}^{2019} \left| \frac{Y_T - \hat{Y}_T}{Y_T} \right| \times 100 \quad (4)$$

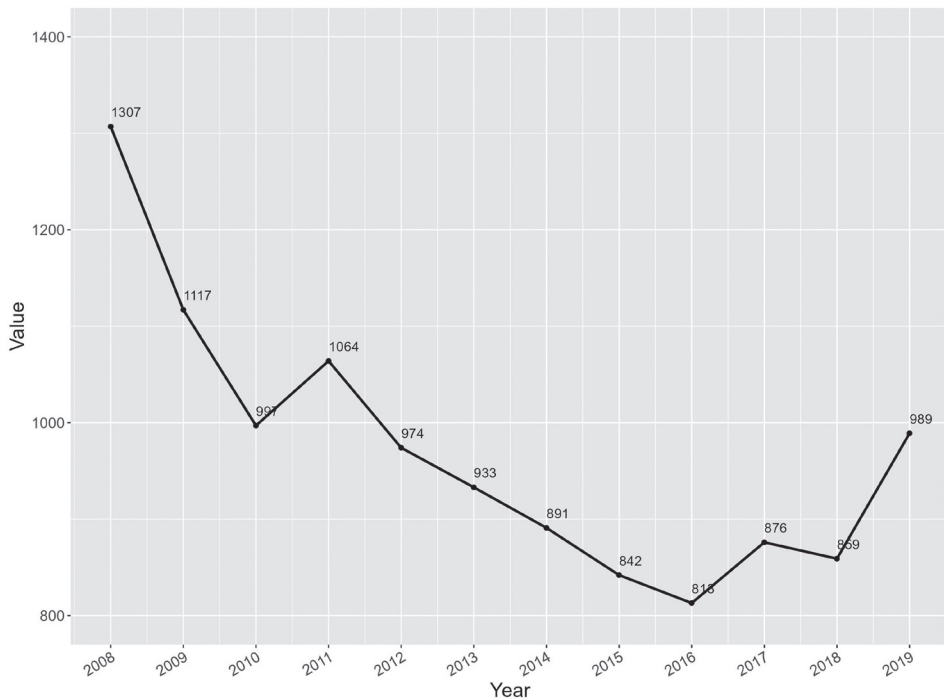
These indicators all represent the average magnitude of errors. Both RMSE and MAE calculate the absolute size of average errors, but RMSE squares the errors, making it more sensitive to outliers. In contrast, MAPE expresses

errors as percentages by dividing the errors by the observed values. For example, if the predicted value is 120 and the observed value is 100, the error is 20%. MAPE represents the average of these percentage-based errors. As for the MAPE, the error is calculated as a percentage, which improves the ease of comparison of values with different forecast value scales. The MAPE, which indicates prediction accuracy in terms of the percentage, is also important considering that the target reduction in serious product accidents is also expressed as a percentage. Therefore, the discussion focuses on the MAPE. The RMSE and MAE are included for reference as a concise indicator of the absolute value of the error.

3. Result

Figure 1 shows the number of accidents. The number of accidents declines from 2008 to 2016. As noted in the Introduction, this may be due to the increased concern about product accidents around 2007. Product manufacturers may have designed their products with greater attention to safety. Consumers may also have become more concerned about accidents. In addition, this was a period when government regulations were strengthened. The RMSE, MAE, and MAPE were calculated as follows: RMSE was 90.92, MAE was 76.18, and MAPE was 7.76. These results indicate that, on average, there were about 80, or about 8% fluctuations in the number and percentage of incidents, each year.

Figure 1. Trends in the number of serious product accidents



Source: Created by the author

4. Discussion

The analysis shows that model which reflects the current approach to target setting, has a forecast error of 7.76%. From these findings, we discuss the target set for 2019 “reducing the number of serious product accidents by 3% relative to 2018”. This target calls for activities to be conducted based on the model approach to forecasting to reduce the number of accidents by 3% relative to the forecast value. However, according to the results of this study, using the number of accidents in the previous year as the forecast value results in a forecast error of approximately 7.76% with respect to the actual data obtained. Under such circumstances, an actual decrease of 3% will not be sufficiently distinguishable from the forecast error. Therefore, assessing the activities of the administrative agency based on whether this target is achieved or not is irrational.

Thus, in the setting of the target value as a percentage of the decrease relative to the forecast value, the target value should be set to a value that is sufficiently distinguishable from the forecast error. However, setting targets based solely on this concept may be dangerous. In this study, for example, if the target is set to decrease by 20% from the forecast value, it may be distinguishable from the forecast error. However, from a practical standpoint, such a goal is difficult to achieve. The evaluation of an administrative agency based on such a difficult-to-achieve goal can be unfair to the agency being assessed. Furthermore, determining resource allocation based on the results of such an evaluation may compromise administrative services. In target setting, the magnitude of forecast errors should be considered and the set targets should be sufficiently distinguishable from them. Moreover, sufficient attention should be given to whether these targets are realistically achievable. If the two are difficult to balance, then decision-makers should not adhere to outcome-based evaluation alone. For example, for an indicator of a business process that can be controlled to some extent by the efforts of the evaluated agency, such as the percentage of reported accidents that are handled within a certain period of time, a numerical target can be set considering the evaluated agency’s ability to handle the process and society’s expectations.

The arguments presented in this study may seem self-evident at first glance. However, the study re-examined actual evaluation metrics through the framework of statistical impact evaluation and specifically identified their shortcomings. This process has clarified how goal-setting should be approached in the future. Furthermore, it has provided valuable insights into whether it is appropriate to use such evaluation methods in the first place. While this research merely highlights issues using a single case study, it raises the possibility that outcome-based evaluations are being indiscriminately overused. Although outcome-based evaluation may be an ideal approach in theory, if irrational evaluations disguised as outcome-based become pervasive due to prevailing trends, they could ultimately result in negative consequences rather than benefits.

5. Conclusion

In Japan, approximately 2,000 product accidents occur each year. To prevent these accidents, the Japanese government undertakes various activities, with NITE playing a major role in these efforts. Recently, the evaluation of these activities has become outcome-based and quantitative, though the methods employed remain questionable. In this study, the validity of the performance measurement was tested using an evaluation method based on the percentage reduction in accidents from the previous year. This evaluation method was previously applied to NITE.

The accuracy of the model’s predictions, which used the previous year’s values as a basis, was tested and found to have an average prediction error of about 7.76%. Therefore, the goal, namely, a 3% reduction in the number of serious product accidents relative to the previous year, is not sufficiently distinguishable from the forecast error range.

Therefore, the setting of such a goal was unsuitable evaluation indicator for administrative agencies. In addition, a goal that calls for reducing the number of accidents by a percentage that is sufficiently distinguishable from this forecast error is likely to be unrealistic and inappropriate. In the assessment of administrative agencies, both predictability and feasibility in setting goals should be considered. Determining resource allocation through irrational evaluation approaches that do not adequately consider this may disadvantage the public. If both goals are difficult to achieve, then decision-makers should not adhere to evaluation via quantitative outcome measures alone.

Funding statement

No outside funding was received to conduct this study.

Data Availability Statement

This study uses externally provided, non-public data. Data cannot be made public.

References

- Agency Reports. (n.d.). *U.S. consumer product safety commission*. Retrieved October 17, 2023, from <https://www.cpsc.gov/About-CPSC/Agency-Reports>
- Consumer Affairs Agency, & Ministry of Economy, Trade and Industry. (2018a). Consumer product safety act (serious product accident information reporting and publication system) handbook for businesses. https://www.meti.go.jp/product_safety/producer/guideline/download.html
- Consumer Affairs Agency, & Ministry of Economy, Trade and Industry. (2018b). Consumer product safety act (serious product accident information reporting and publication system) handbook for businesses. https://www.meti.go.jp/product_safety/producer/guideline/download.html
- Hanway, S. J., & Rodgers, G. B. (2020). Impact of the voluntary safety standard for liquid laundry packets on child injuries treated in US hospital emergency departments, 2012–2018. *American Journal of Public Health*, 110(8), 1242–1247. <https://doi.org/10.2105/AJPH.2020.305650>
- Magat, W. A., & Moore, M. J. (1995). *Consumer product safety regulation in the United States and the United Kingdom: The case of bicycles*. National Bureau of Economic Research Cambridge, Mass., USA.
- Ministry of Internal Affairs and Communications. (2022). *Guidelines for the Evaluation of Incorporated Administrative Agencies*. https://www.soumu.go.jp/main_content/000797278.pdf
- Ministry of Internal Affairs and Communications. (2023). *Guidelines for the Evaluation of Policy*. https://www.soumu.go.jp/main_content/000871192.pdf
- Moore, M. J., & Magat, W. A. (1996). Labeling and performance standards for product safety: The case of CPSC's lawn mower standards. *Managerial and Decision Economics*, 17(5), 509–516. [https://doi.org/10.1002/\(SICI\)1099-1468\(199609\)17:5<509::AID-MDE792>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1099-1468(199609)17:5<509::AID-MDE792>3.0.CO;2-W)
- Morrongiello, B. A., Walpole, B., & Lasenby, J. (2007). Understanding children's injury-risk behavior: Wearing safety gear can lead to increased risk taking. *Accident; Analysis and Prevention*, 39(3), 618–623. <https://doi.org/10.1016/j.aap.2006.10.006>
- National Institute of Technology and Evaluation. (2020). *Report on Business Results for Fiscal Year 2019*. <https://www.nite.go.jp/data/000111533.pdf>
- National Institute of Technology and Evaluation. (2022). *Accident Information Analysis Report 2021*. <https://www.nite.go.jp/data/000141124.pdf>
- National Institute of Technology and Evaluation. (2023). *Accident Information Analysis Report 2022*. <https://www.nite.go.jp/jiko/report/annual/2022fy/2022.html>

- Product Safety Division, Commerce and Information Policy Bureau, & Ministry of Economy, Trade and Industry. (2001). *Explanation of the consumer product safety law*. Economic and Industrial Research Institute, Economic and Industrial Information Division.
- Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, 34, 151–163. <https://doi.org/10.1016/j.rcim.2014.12.015>
- Rascoff, S. J., & Revesz, R. L. (2002). The biases of risk tradeoff analysis: Towards parity in environmental and health-and-safety regulation. *The University of Chicago Law Review*, 69(4), 1763. <https://doi.org/10.2307/1600618>
- Rodgers, G. B. (1988). Reducing bicycle accidents: A reevaluation of the impacts of the CPSC bicycle standard and helmet use. *Journal of Products Liability*, 11(4), 307–317.
- Rodgers, G. B. (1996). Bicyclist risks and helmet usage patterns: An analysis of compensatory behavior in a risky recreational activity. *Managerial and Decision Economics*, 17(5), 493–507. [https://doi.org/10.1002/\(SICI\)1099-1468\(199609\)17:5<493::AID-MDE791>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1099-1468(199609)17:5<493::AID-MDE791>3.0.CO;2-A)
- Rodgers, G. B. (2002). Effects of state helmet laws on bicycle helmet use by children and adolescents. *Injury Prevention*, 8(1), 42–46. <https://doi.org/10.1136/ip.8.1.42>
- Rodgers, G. B. (2022). Evaluation of the impact of the voluntary safety standard for liquid laundry packets on the rate of child exposures reported to Poison Control Centers in The US. *Injury Prevention*, 28(1), 43–48. <https://doi.org/10.1136/injuryprev-2020-044115>
- Rodgers, G. B., & Adair, P. K. (2009). Exemptions to the children's sleepwear flammability standards: A description of policy considerations and an evaluation of the effects on burn injuries to children. *Journal of Consumer Policy*, 32(1), 59–71. <https://doi.org/10.1007/s10603-009-9092-y>
- Rodgers, G. B., & Topping, J. C. (2012). Safety effects of drawstring requirements for children's upper outerwear garments. *Archives of Pediatrics and Adolescent Medicine*, 166(7), 651–655. <https://doi.org/10.1001/archpediatrics.2011.1269>
- Smith, L. E., Greene, M. A., & Singh, H. A. (2002). Study of the effectiveness of the US safety standard for child resistant cigarette lighters. *Injury Prevention*, 8(3), 192–196. <https://doi.org/10.1136/ip.8.3.192>
- Viscusi, W. W. (1984). The lulling effect: The impact of child-resistant packaging on aspirin and analgesic ingestions. *The American Economic Review*, 74(2), 324–327.
- Viscusi, W. K. (1985). Consumer behavior and the safety effects of product safety regulation. *The Journal of Law and Economics*, 28(3), 527–553. <https://doi.org/10.1086/467100>
- Viscusi, W. K., & Cavallo, G. (1996). Safety behavior and consumer responses to cigarette lighter safety mechanisms. *Managerial and Decision Economics*, 17(5), 441–457. [https://doi.org/10.1002/\(SICI\)1099-1468\(199609\)17:5<441::AID-MDE787>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1099-1468(199609)17:5<441::AID-MDE787>3.0.CO;2-4)
- Viscusi, W. K., & Cavallo, G. O. (1994). The effect of product safety regulation on safety precautions. *Risk Analysis*, 14(6), 917–930. <https://doi.org/10.1111/j.1539-6924.1994.tb00061.x>
- Viscusi, W. K., & Dalafave, R. (2022). The broad impacts of disposable lighter safety regulations. *Journal of Benefit-Cost Analysis*, 13(2), 149–165. <https://doi.org/10.1017/bca.2022.6>

(Accepted 26th December, 2024)

【研究論文】

総額型研究開発税制が研究開発費の促進に与える影響

河瀬 豊

神戸学院大学

kawase@ba.kobegakuin.ac.jp

要 約

試験研究費に係る法人税の税額控除制度は、研究開発投資の促進を目的としている。各府省の政策評価書では、この制度が研究開発支出を押し上げていると評価している。本研究では、総額型の研究開発税制が本当に研究開発費を促進しているか否かについて検証した。その結果、総額型制度が研究開発費を増加させている明確な証拠は確認できなかった。ただし、負債比率の高い企業では研究開発費の増加が観察された。また、繰越控除の上限規制が研究開発費に影響を与えたという証拠も得られなかった。これらの結果から、制度に存在意義がないと断定することはできない。重要なのは、①立法趣旨と政策目的に整合性を持たせる、②有効性を検証するモデルを確立させることである。

キーワード

研究開発費、税額控除、租税政策、租税特別措置

1. はじめに

企業が研究開発投資を実施すると、一定の条件を満たせば、税額控除が適用される「試験研究を行った場合の法人税額の特別控除」（租税特別措置法42条の4、以下、「研究開発税制」）制度がある。この制度の現在の目的は、研究開発投資を促進するために存在するものと考えられる。経済産業省をはじめとする各府省の政策評価書には、当該制度により、研究開発投資及びGDPの押し上げ効果が認められる旨の記述がある。果たして、本当にそのような効果があるのだろうか。本稿では、①研究開発税制に関する政策評価の問題点を指摘し、②研究開発費が増加しているか否かについて実証分析を行い、③政策評価書を改善するための考え方を示す。

研究開発税制は1967年に創設され、当初は、過去の試験研究費を超える部分が優遇の対象となっていた。2003年度の改正により、企業が支出した試験研究費の全額が優遇の対象となった。本稿では、後者の制度を導入したことによる有効性の確認を中心的なテーマとする。

河瀬（2023）では、繊維産業に限定して総額型の研究開発税制の有効性を検証し、当該税制が研究開発費を増加させているとは言えない結果を得ている。本稿では、サンプルの範囲について金融機関を除く一般事業会社全般に拡大し、制度導入当初だけでなく、長期的な効果も検証するなど、検証モデルに様々な工夫を取り入れた。さらに、政策評価書との関連を検討し、公共政策に応用する可能性についても考察した。

2章では、研究開発税制の有効性に関する先行

研究を簡単にレビューする。3章では、我が国における研究開発税制の概要と変遷について、分析対象期間を中心に説明する。4章では、研究開発税制が政策評価書において、どのように評価されているかを概観する。5章では、実証分析のリサーチ・デザインと結果を示す。6章ではむすびを述べる。

なお、研究開発費は会計上の科目名であり、試験研究費は税法上の用語である。これらの定義は異なる。試験研究費は税額控除の対象となるため、範囲が厳格に定められており、通常、研究開発費より狭い範囲を指す。本稿では、研究開発費の額に税制が影響を与えているか否かを分析している。制度の目的も、試験研究費を増加させることではなく、より広く研究開発投資を増加させることに重点を置いていと考えられる。試験研究費に焦点を当てるよりも、むしろ研究開発費に着目する方が目的に適合していると考えられる。

2. 先行研究

これまで研究開発に関する税制が研究開発投資に与える影響について、国内外で多くの研究が行われてきた。

(1) 国外の先行研究

Hall and Reenen (2000) は、米国の10研究及び世界10研究をメタ分析し、研究開発費税制の有効性を検証している。税価格弾力性は概ね-1であり、税額控除が1%増加するとR&D支出も約1%増加することが示された。

Bloom et al. (2002) は、9か国を対象に1979年から1997年までのデータを用いて研究し、税価格弾力性が概ね-0.1であることを明らかにした。短期的な効果は小さいものの、長期的には大きな効果になると主張している。

(2) 日本の先行研究

①増加型についての先行研究

Koga (2003) が日本の製造業904社について、1989年から1998年のパネルデータを用いて実証分析を行っている。その結果、税価格弾力性は-0.68

と算出され、諸外国と比較して低い値であることが示された。また、規模の大きな企業においては、税価格弾力性は概ね-1となっており、増加型制度は主に大企業の研究開発投資を促進すると結論付けている。

②総額型についての先行研究

大西・永田 (2010) は、Difference in Differences (DID) の手法を用いて平成15 (2003) 年度税制改正の影響を検証している。サンプルには、文部科学省「民間企業の研究活動に関する調査」、総務省「科学技術研究調査」(2003年度、2004年度調査に回答している企業を採用) をもとに、経済産業省「企業活動基本調査」のデータを加えた485社を採用している。研究開発支出の代理変数には、「科学技術研究調査」の「研究開発費総額」を使用している。社内の研究費に限定するために、ここから社外研究費を除外する調整を行っている。総額型制度利用企業と未利用企業の研究開発費増加分の平均値の差の検定を行っている。分析の結果、総額型の税額控除を利用した企業群と利用していない企業群の研究開発支出の増加分の平均値に有意な差が認められないことを明らかにした。これにより、総額型の税制改正は研究開発支出を増加させているとは言えないと結論付けている。また、研究開発投資が増加しない理由の一つとして、税額控除の上限が低い(当時の税額控除の上限は、法人税額の20%である。) ことを指摘している。

米谷・松浦 (2010) は総額型の導入が企業の研究開発行動に及ぼす影響を分析している。分析対象は1999年度から2006年度までの全上場企業のうち、金融業を除く一般事業会社から、一定の条件を満たす3,989企業-年である。総額型導入後は税額控除の限度額の影響を受ける企業はそうでない企業に比べて研究開発投資を増加させていることが示された。これは、企業が当該税制改正とは無関係に研究開発投資を行っていることを示唆している。

Kasahara et al. (2014) も、2003年度の税制改正の影響を検証している。サンプルは、経済産業省の2001年から2003年までのデータを使用し、業種を製造業に絞った6,156を採用している。Bloom et al. (2002) のモデルを参考に一階差分を取った次

式等をGMM（Generalized Method of Moments）により推定している。

$$\Delta \ln RD_{it} = \beta \Delta \tau_{it} + \gamma \Delta \ln Y_{it} + \delta \Delta \left(\frac{b_{it}}{A_{it}} \right) + \theta \Delta \left(\tau_{it} \frac{b_{it}}{A_{it}} \right) + \Delta Z'_{it} \alpha + \Delta \eta_{it} + \Delta \epsilon_{it}$$

ここで、RDは研究開発支出（添え字*i*は年、*it*は企業）、 τ は試験研究費税額控除の率、*Y*は売上高、*b*は負債、*A*は固定資産、*Z*は他のコントロール変数、 η は時間効果である。

負債の比率の高い会社においては、税額控除の効果が認められると結論付けた。先行研究の結論は一致していない。総額型の導入によって研究開発支出が増加したとは言えない。ただし、特定の条件を満たす企業においては研究開発支出を増加させる効果が認められる。特に、負債の比率の高い企業にはその促進効果が顕著である。

3. 研究開発税制の概要と変遷

ここでは研究開発税制について、主要な変遷を説明する。本稿の分析期間に関連する重要な改正については、表1にまとめた。検証期間について重点的に説明するため、最新の制度についての制度説明は割愛する。

研究開発税制は昭和42（1967）年度に創設された制度で、過去の特定の事業年度に支出した試験

研究費を超えて支出した場合に、超えた部分の一定割合（当初は10%）を税額控除できるものである（旧租税特別措置法42条の6）。この制度は過去の試験研究費から増加して支出した場合にのみ適用があるので、以下「増加型」と呼ぶ。研究開発税制は、昭和45（1970）年3月31日までの時限立法であったが、制度に変更を加えながら延長を繰り返し、現在も存続している。この制度の導入当初は、試験研究費を増加させるというよりも、試験研究活動を税制面から助成することに主眼が置かれている¹。立法趣旨と政策目的に齟齬がある可能性があるので、この趣旨は政策評価においても重要な意味を持つ。

昭和60（1985）年度税制改正では、中小企業技術基盤強化税制が導入された。中小企業者等が支出する試験研究費の額の全額を税額控除の対象（税額控除の額は試験研究費の6%）とする制度である。中小企業者等については、このときに総額型制度が導入されている。本稿では、上場企業を分析の対象としているため、この制度改正は直接の影響はない。しかし、改正内容は重要であるため言及した。

平成15（2003）年度税制改正では、試験研究費の総額に一定率を乗じて税額控除できる制度が創設された（以下、「総額型」）。同時に、税額控除が使いきれなかった場合、翌年に繰り越すことができる制度も導入された（以下、「繰越控除制度」）。これまで過去の特定期間の試験研究費の額を超えなければ、当該制度による税額控除は適用できな

表1 研究開発税制の主な変遷

改正年度	改正内容
1967	試験研究費税制創設（増加型）
1985	中小企業技術基盤強化税制の導入
2003	総額型の創設 税額控除の上限を法人税額の20%に引き上げ（改正前15%） 繰越控除制度の創設
2013	税額控除の上限を法人税額の30%に引き上げ
2015	繰越控除制度が廃止 税額控除の上限を法人税額の25%に引き下げ
2017	増加インセンティブ制度の導入

（出所）武田（1979）、旧租税特別措置法を参考に筆者作成

かったので、試験研究費を増額しなくても優遇税制を適用できるという意味で総額型は抜本的な改正である。総額型は、制度の変更を伴いながら、実質的に現在も存続している。

平成25（2013）年度税制改正では、税額控除の上限を法人税額の20%から30%に引き上げた。平成27（2015）年度税制改正で、繰越控除制度は廃止され、税額控除の上限も25%に引き下げられた。繰越控除が廃止されたため、この改正以後は、課税所得がプラスの法人のみ総額型を適用できる状況になっている。これは総額型の制度を適用できる会社とできない会社を区分する上で重要な改正である。

平成29（2017）年度税制改正では、企業が支出する試験研究費の増減に応じて、税額控除率が6%から14%まで変動する仕組みが導入された。これは、試験研究費が増加するほど控除率も上昇するため、企業が研究開発費を増やすインセンティブが強化されている（以下、「増加インセンティブ制度」）。

令和元（2019）年度税制改正では、増加インセンティブ制度を強化した。以前の制度では、増加試験研究費について増加率の基準を5%としていたが、これを8%とした。これは政策目標を達成するために必要な増加率を元に算出したと考えられる²。

4. 研究開発税制の政策評価

研究開発税制は、政策評価上、研究開発政策ではなく、租税特別措置に分類されている。平成22（2010）年に租税特別措置の適用状況の透明化等に関する法律（租特透明化法）が施行された。それに伴い、租税特別措置に関する政策評価が導入された。

導入された理由について、総務省の「政策評価Q&A」に次のように説明されている。

租税特別措置等は、その多くが特定の者の税負担を軽減することなどにより産業政策等の特定の政策目的の実現に向けて経済活動を誘導する手段となっています。こうした租税特別措置等は、税負担の公

平の原則の例外であり、これが正当化されるためには、その適用の実態や効果が透明で分かりやすく、納税者が納得できるものである必要があります。

これを踏まえると、租税特別措置等を新設したり、拡充又は延長を行ったりする際には、租税特別措置等の必要性や有効性等について、可能な限り客観的なデータを明らかにして事前・事後の政策評価を実施し、その結果を政策決定の判断材料として活用することは極めて重要です。（総務省行政評価局 2017、p.29）

この制度の導入背景には、民主党の政策が大きく関わっている。民主党に政権が交代したときに各政策を実現するために歳出などの「ムダづかい」にメスを入れようとした。租税特別措置はその多くが時限立法であるが、その効果や存在意義が検証されないまま存続しているものが多くある。そこで、延長する必要のない租税特別措置を廃止することを試みた³。

政策評価の実施内容は、必要性、有効性、及び相当性であり、これらについて説明しなければならない（租税特別措置等に係る政策評価の実施に関するガイドラインⅡ4（1）・（2））。

試験研究を行った場合の法人税額等の特別控除については、経済産業省が主管となり、平成22（2010）年から各府省が政策評価を実施している。この税制の延長や新制度の創設の前年に事前評価を行っている。

本来、事前評価の結果は、事後評価によって必ず検証されなければならないはずである（同ガイドラインⅡ4）。ただし、租税特別措置の事後評価については、既存制度の拡張又は延長の要望について事前評価を実施した場合は、事後評価の要素を含んでいるとして、事後評価を実施することを要しない（同ガイドラインⅡ5（2））。つまり、事後評価を省略した場合には、事前評価において政策を実施した結果を検証する必要がある。

研究開発税制は租税特別措置の政策評価が導入される以前から存在している。したがって、この制度についての要望は、通常、拡充又は延長に該当するので、制度上、事後評価を実施する必要はない。経済産業省は事前評価のみを行っており、事後評価は行っていない。以下、特に断りのない

表2 試験研究を行った場合の法人税額等の特別控除についての政策評価書の概要

実施年	主要望の内容	政策目的	根拠など	有効性等
2010	総額型の控除上限を10%に引上げ	研究開発投資（官・民併せて）をGDP比4%以上	新成長戦略	経済産業省の試算
2011	増加型・高水準型の恒久化	GDP比4%以上	新成長戦略	経済産業省の試算
2012	総額型の控除上限を30%に引上げ	GDP比4%以上	日本再生戦略	経済産業省の試算
2013	増加型上乘せ措置の控除率を30%へ引上げなど	民間研究開発投資を3年以内にGDP比3%以上	日本再興戦略	経済産業省の試算
2014	オープンイノベーション型制度の導入など	民間研究開発投資をGDP比で世界1位に復活	日本再興戦略、日本再興戦略改訂2014	kasahara et al. (2011) (ママ)によると、総額型は控除額に対し民間研究開発投資を約2.33倍押上げる効果がある。
2016	試験研究費にサービス開発の追加、試験研究費の増減に応じた控除率の設定など	民間研究開発投資を2020年度までに対GDP比3%以上	第5期科学技術基本計画、成長戦略（日本再興戦略改訂2016）、科学技術イノベーション総合戦略2016、経済財政運営と改革の基本の方針2016	kasahara et al. (2011) (ママ)、経済産業省のアンケート調査
2018	控除上限引上げなど多数	民間研究開発投資を2020年度頃までに対GDP比3%	第5期科学技術基本計画、未来投資戦略2017・2018、統合イノベーション戦略2018、経済財政運営と改革の基本方針2018	kasahara et al. (2011) (ママ)
2019	オープンイノベーション型の対象機関の拡大	民間研究開発投資を2020年度頃までに対GDP比3%	第5期科学技術基本計画、未来投資戦略2017・2018、経済財政運営と改革の基本方針2019、科学技術・イノベーション創出の活性化に関する法律49条、統合イノベーション戦略2019	目標を達成するためには、民間研究開発投資を毎年7.8%増加させる必要がある。
2020	控除上限引上げなど多数	民間企業の研究開発投資を維持・拡大	統合イノベーション戦略2020、経済財政運営と改革の基本方針2020、成長戦略フォローアップ、Beyond 5G推進戦略	Kasahara et al. (2014)

(出所) 経済産業省 (2010; 2011; 2012; 2013; 2014; 2016; 2018; 2019; 2020) をもとに、筆者作成

限り、研究開発税制に関連する政策評価書は、経済産業省が作成したものを指す。

他の府省は、経済産業省の政策評価書とほぼ同じ内容のものが多い。一部、各府省独自の政策評価を行うこともある。例えば、農林水産省は独自に事後評価を実施している（農林水産省 2012; 2017）。

表2に経済産業省の政策評価書の内容をまとめた。

総額型制度は2003年に導入されているが、租税特別措置の政策評価制度が当時は存在していなかった。そのため、平成22（2010）年からの政策評価書でこの制度を評価している。

試験研究を行った場合の法人税額等の特別控除の政策目的として、平成22（2010）年の政策評価書には、「2020年度までに、官民合わせた研究開発投資をGDP比4%以上にする。」（経済産業省 2010、p.41）と記載されている。その根拠は、新成長戦略（内閣府 2010、p.48）である。新成長戦略では、「2020年度までに官民合わせた研究開発投資をGDP比の4%以上にする。そのため、政府の関与する研究開発投資を第4期科学技術基本計画に沿って拡充することとし、効果的、効率的な技術開発を促進するための規制改革や支援体制の見直し、官民連携の強化、民間研究開発投資への税制優遇措置など研究開発投資の促進に向けた各種施策を検討・実施する」（内閣府 2010、pp.48-49）と記載されている。その翌年に第4期科学技術基本計画（内閣府 2011）が策定された。その計画によると、将来にわたる持続的な成長と社会の発展の実現、我が国が直面する重要課題への対応、及び基礎研究と人材育成の強化の政策を遂行するために、官民合わせた研究開発投資をGDP比4%（政府1%、民間3%）以上及び総額約25兆円を目標としている。研究開発費の数値目標は存在するが、なぜ研究開発投資がGDPの4%必要なのかについての説明はない。総額型による研究開発投資の押し上げ効果については、次のように試算している。財務省が試算した平成22（2010）年度減収額1,712億円に、当該税制による押し上げ効果を1.38倍とし、2,362億円（ $\div 1,712 \text{ 億円} \times 1.38$ ）の押し上げ効果があると試算している。平成26（2014）年の事前評価書（経済産業省 2014）から、

Kasahara et al. (2011) のディスカッションペーパー⁴を引用し、「控除額に対し民間研究開発投資を約2.33倍押し上げる効果がある」としている。令和2（2020）年の事前評価書（経済産業省 2020）では、Kasahara et al. (2014) を引用し、「研究開発税制（総額型）の導入により、研究開発投資が3.0～3.4%増加した。」（経済産業省 2014、p.3）としている。政策評価書では、実際に適用された税額控除額をもとに効果が推計されている。しかし、多くの先行研究では、税額控除を適用した企業全体に対して促進効果が確認されていない場合が多い。有効性を認める研究もあるが、それらは負債比率が高いなど、限定された条件下でのみ効果を認めている。したがって、有効性の検証には、促進効果が発現する射程範囲についても十分に検討する必要がある。また、Kasahara et al. (2014) までは、大西・永田（2010）が公表されているにもかかわらず、他のディスカッションペーパーを引用している点にも注目すべきだろう。

5. リサーチ・デザイン

(1) 研究開発費の予測モデル

総額型制度の導入による研究開発費の促進効果を検証するため、次に示す式に固定効果モデルを適用して、パネルデータ分析を行った。研究開発費は会計上の勘定科目であり、会計学のモデルを利用するのが適切である。小嶋（2008）は、財務会計分野において研究開発費の説明モデルを開発した⁵。本稿では、小嶋（2008）のモデルをもとに、税制改正の効果を検証するものに拡張する。

$$RD_{i,t} = \beta_{0,i} + \beta_1 RD_{i,t-1} + \beta_2 OCF_{i,t} + \beta_3 FA_{i,t} + \beta_4 IRD_t + \beta_5 IFA_t + \beta_6 \ln GDP_t + \beta_7 LEV_{i,t} + \beta_8 Black_{i,t} + \beta_9 Gross_t + \beta_{10} GNC_t + \sum \beta_k \cdot Year + \varepsilon_{i,t}.$$

ここで、 $RD_{i,t}$ は企業*i*の*t*年の研究開発費を売上高で除したもの（以下、財務数値は売上高で除している。 $\ln GDP$ 、 LEV 、ダミー変数を除く。）、 OCF は営業キャッシュ・フロー、 FA は固定資産の取得による支出、 IRD は業種別の研究開発費の平均値、 IFA は業種別の固定資産の取得による支出

の平均値、 $\ln GDP$ は国内総生産（名目）を対数変換（自然対数）したもの、 LEV は負債を総資産で除したもの、 $Black$ は税金等調整前当期純利益（税引前当期純利益）が正の値のときは1、そうでないときは0、 $Gross$ は総額型が適用される年は1をとり、そうでないときは0、 GNC は総額型制度の年で、かつ、税額控除の繰越制度がない年のときは1、そうでないときは0、 $Year$ は年次ダミー変数（添え字の年に該当したときは1、そうでないときは0）、 ε は誤差項である。

$RD_{i,t}$ から $\ln GDP$ までは、 RD を予測する変数で、本稿ではコントロール変数としての役割を果たす⁶。事前分析として、 RD 予測モデルの妥当性を確認する。そのモデルがmodel0である。結果は表5に示したとおりであるが、企業の研究開発費は前年のそれや業種の影響を大きく受けることが確認できる。このモデルに適宜、交差項を追加して、DIDによる検証を行う。

(2) サンプル

サンプルは、日経NEEDS Financial Questを使用した。使用したデータの期間は2001年から2020年⁷で、3月決算の一般事業会社（金融業、保険業が除かれている）を対象とした。

全ての年に研究開発費があるなどモデルに必要なデータが揃っているなどの条件を満たすサンプルサイズ18,000企業-年を得た⁸。

(3) 仮説

総額型を導入したことにより、研究開発費が増加しているか否かを検証する。他にも現在までに議論の対象となっている税制改正についての有効性も同時に検証する。

先行研究では、税制改正当初の効果に焦点を当てているが、総額型は改正後も継続している制度である。 $Gross$ 変数を利用して、総額型の長期的な効果も検証する（model10、11）。

仮説1：平成15（2003）年度税制改正により、研究開発費が増加した⁹。

この仮説を検証するために、model1、model2を使用し、総額型の導入が試験研究費の増加に寄与したか否かを分析する。

仮説2：研究開発税制の税額控除限度額は、研究

開発費の増減に影響を与える。つまり、①平成25（2013）年度税制改正により、研究開発費が増加した。②平成27（2015）年度税制改正により、研究開発費が減少した。

この仮説を検証するために、model1、model3、model4を使用する。

仮説3：増加インセンティブ制度は、研究開発費を増加させる。つまり、①平成29（2017）年度税制改正により、研究開発費が増加した。②令和元（2019）年度税制改正により、研究開発費が増加した。

この仮説を検証するために、model1、model6、model7、model8、model9を使用する。

仮説4：総額型の税額控除制度が適用できる法人は、適用できない法人に比べて研究開発費を増加させる。

この仮説を検証するために、model5、model10、model11、model12を使用する。

(4) 基本統計

記述統計及び変数間の相関係数を表3・4に示す。

研究開発費は1期前との相関が約0.967であり、毎年ほぼ同額の研究開発予算が計上されていることが推測される。また、研究開発費とその業種平均との相関が約0.624であることから、業種の影響を受けやすい費目であることがわかる。

次に記述統計を確認する。 RD の平均値は約0.024で、中央値は約0.015である。実証研究によっては、研究開発費／売上高が数パーセント以下の企業はサンプルから除外する研究もあるが、本稿では研究開発費が規模に比して少額の企業も含めて研究開発費を増額することが制度の健全な目的にも合致していると考える。したがって、研究開発費が少額な企業もサンプルから除外しない。

(5) 結果

分析結果を表5に示した。

$Year_{2004}$ などの変数は、年次ダミーである。2003年度税制改正が反映されるのは、この分析では2004年からのなので、 $Year_{2004}$ の係数を確認する。つまり、総額型導入については、 $Year_{2004}$ （係数の予想符号は正、以下同じ）、同様に、税額控除上限の拡大は $Year_{2014}$ （予想：正）、縮小は $Year_{2016}$ （予想：

表3 記述統計

	RD	OCF	FA	IRD	IFA	lnGDP	LEV	Black
サンプルサイズ	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000
平均値	0.024	0.069	-0.047	0.024	-0.047	13.177	0.490	0.874
標準偏差	0.031	0.066	0.045	0.019	0.022	0.035	0.197	0.332
最小値	0.000	-0.803	-0.898	0.000	-0.265	13.117	0.021	0.000
第一四分位	0.006	0.034	-0.062	0.011	-0.056	13.153	0.340	1.000
中央値	0.015	0.066	-0.036	0.022	-0.046	13.178	0.493	1.000
第三四分位	0.033	0.102	-0.018	0.031	-0.037	13.198	0.635	1.000
最大値	0.394	0.562	0.000	0.124	-0.004	13.230	1.473	1.000

(出所) 筆者作成

表4 変数間の相関係数

	RD	RD _{t-1}	OCF	FA	IRD	IFA	lnGDP	LEV	Black
RD	1								
RD _{t-1}	0.9672	1							
OCF	0.1754	0.1823	1						
FA	-0.1192	-0.1021	-0.2917	1					
IRD	0.6243	0.6209	0.1393	-0.0906	1				
IFA	-0.1166	-0.1091	-0.2755	0.4850	-0.1867	1			
lnGDP	-0.0144	-0.0070	0.0428	-0.0277	-0.0231	-0.0571	1		
LEV	-0.2678	-0.2620	-0.2597	0.0833	-0.2141	-0.0355	-0.0757	1	
Black	-0.0293	-0.0126	0.2395	0.0009	-0.0065	0.0033	0.1502	-0.1860	1

(出所) 筆者作成

負)、増加インセンティブ制度の導入は $Year_{2018}$ (予想：正)、増加インセンティブ制度の拡張は $Year_{2020}$ (予想：正)を確認する。

単純に改正年度の翌年を検証したものがmodel1である。 $Year_{2004}$ 、 $Year_{2016}$ 、 $Year_{2020}$ は有意ではない。 $Year_{2014}$ と $Year_{2018}$ は、5%水準以上で有意であるが、符号が予測と異なり説明が困難である。研究開発費を促進する制度が導入されたにもかかわらず、その支出が有意に減少しているからである。

負債比率の高い企業に対して、研究開発費の促進効果については、DIDの手法を使用した。これに該当するのはmodel2、model3、model4、model6、model8、model11である。model2の $LEV \times Year_{2004}$ の係数は正で10%水準ではあるが有意である。5%水準で有意ではないものの、負債比率の高い企業

の研究開発費を増加させる傾向が確認できる。同様に、model3の $LEV \times Year_{2014}$ は、正で有意(1%水準)である。このことから、負債比率の高い企業に対しては、税額控除上限の引き上げ政策が研究開発費を促進した可能性がある。他のモデルでは、負債比率の高い企業に対する政策効果は認められなかった。

次に、税額控除を適用できる企業と適用できない企業に研究開発費に差があるかを検証する。3章で確認したとおり、平成15(2003)年度税制改正では、総額型と繰越控除の制度が導入されたため、課税所得が発生しているか否かで税額控除適用可能企業とそうでない企業を分けることができない。課税所得が発生せず税務上の欠損が生じている場合でも、翌期に繰り越されるであろう税額

表5 分析結果

RD	model0	model1	model2	model3	model4	model5	model6	model7	model8	model9	model10	model11	model12
RD _{t-1}	0.653*** (0.006)	0.651*** (-0.005)	0.652*** (-0.006)	0.654*** (-0.006)	0.652*** (-0.006)	0.651*** (-0.006)	0.653*** (-0.006)	0.651*** (-0.006)	0.653*** (-0.006)	0.651*** (-0.006)	0.653*** (-0.006)	0.653*** (-0.006)	0.651*** (-0.006)
OCF	-0.012*** (0.001)	-0.011*** (-0.001)	-0.013*** (-0.001)	-0.013*** (-0.001)	-0.013*** (-0.001)	-0.01*** (-0.001)	-0.013*** (-0.001)	-0.01*** (-0.001)	-0.013*** (-0.001)	-0.01*** (-0.001)	-0.013*** (-0.001)	-0.013*** (-0.001)	-0.01*** (-0.001)
FA	-0.018*** (0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)	-0.018*** (-0.002)
IRD	0.574*** (0.026)	0.525*** (-0.026)	0.555*** (-0.026)	0.55*** (-0.026)	0.557*** (-0.026)	0.552*** (-0.026)	0.56*** (-0.026)	0.556*** (-0.026)	0.555*** (-0.026)	0.549*** (-0.026)	0.559*** (-0.026)	0.559*** (-0.026)	0.544*** (-0.027)
IFA	-0.021*** (0.006)	-0.013** (-0.006)	-0.019*** (-0.006)	-0.017*** (-0.006)	-0.018*** (-0.006)	-0.019*** (-0.006)	-0.018*** (-0.006)	-0.019*** (-0.006)	-0.017*** (-0.006)	-0.017*** (-0.006)	-0.016*** (-0.006)	-0.016*** (-0.006)	-0.017*** (-0.006)
lnGDP	-0.002 (0.002)	-0.001 (-0.002)	-0.004*** (-0.002)	-0.005*** (-0.002)	-0.004** (-0.002)	-0.0001 (-0.002)	-0.003* (-0.002)	0.001 (-0.002)	-0.004*** (-0.002)	-0.0001 (-0.002)	-0.004** (-0.002)	-0.004** (-0.002)	-0.002 (-0.002)
LEV		-0.007*** (-0.001)	-0.005*** (-0.001)	-0.005*** (-0.001)	-0.005*** (-0.001)		-0.005*** (-0.001)		-0.005*** (-0.001)		-0.006*** (-0.001)	-0.006*** (-0.001)	
Black		-0.002*** (-0.0002)				-0.002*** (-0.0002)		-0.002*** (-0.0002)		-0.002*** (-0.0002)			-0.002*** (-0.0002)
Year ₂₀₀₄		0.0002 (-0.0002)	-0.001 (-0.001)										
Year ₂₀₁₄		-0.001*** (-0.0002)		-0.003*** (-0.001)									
Year ₂₀₁₆		-0.00004 (-0.0003)			-0.0001 (-0.001)	-0.0005 (-0.001)							
Year ₂₀₁₈		-0.001** (-0.0003)					-0.001 (-0.001)	-0.001 (-0.001)					
Year ₂₀₂₀		0.0001 (-0.0003)							0.0001 (-0.001)	0.001 (-0.001)			
LEV × Year ₂₀₀₄			0.002* (-0.001)										
LEV × Year ₂₀₁₄				0.004*** (-0.001)									
LEV × Year ₂₀₁₆					0.0005 (-0.001)								
Black × Year ₂₀₁₆						0.001 (-0.001)							
LEV × Year ₂₀₁₈							0.001 (-0.001)						
LEV × Year ₂₀₂₀								0.0002 (-0.001)					
Black × Year ₂₀₁₈								0.001 (-0.001)					
Black × Year ₂₀₂₀									-0.001 (-0.001)				
Gross											-0.001*** (-0.0002)	-0.001 (-0.0004)	
LEV × Gross												0.00004 (-0.001)	
GNC													-0.00005 (-0.0005)
Black × GNC													0.0003 (-0.0005)
Observations	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000	18,000
R ²	0.487	0.494	0.489	0.49	0.489	0.49	0.489	0.49	0.489	0.49	0.49	0.49	0.49
Adjusted R ²	0.460	0.467	0.462	0.463	0.462	0.463	0.462	0.463	0.462	0.463	0.463	0.463	0.463
F Statistic	2,708.89***	1,280.90***	1,819.12***	1,823.55***	1,818.32***	1,825.82***	1,819.44***	1,826.03***	1,818.41***	1,826.42***	2,049.86***	1,821.99***	1,826.09***

(注) *p<0.1、**p<0.05、***p<0.01 () 内は標準誤差

(出所) 筆者作成

控除を考慮して、研究開発費を増加させる可能性もある。平成27（2015）年度税制改正により繰越控除制度が廃止されるため、この改正後は、課税所得が発生している企業のみが税額控除制度を利用できることになる。このことを検証したのは、model5、model7、model9である。着目すべき変数は、*Black*と年次ダミーとの交差項である。例えば、model5では $Black \times Year_{2016}$ が該当する。係数は正であるが、有意ではない。したがって、税額控除を適用可能な企業の研究開発費を促進したとは言えない結果となっている。増加インセンティブ制度の導入についても、model7及びmodel9の交差項を確認することにより、効果が認められないことが確認できる。

最後に、総額型の長期的な効果を検証する。これまでのモデルでは、税制改正直後の企業行動を検証しているが、改正された制度はその後にも継続されることが多い。したがって、税制改正直後にはそれほど促進効果が表れなかったとしても、徐々に制度が浸透し、長期的に研究開発投資を促進している可能性もある。これについて、model10、model12を用いて検証する。model10の*Gross*の係数は負で、1%水準で有意である。この結果の解釈は難しいが、少なくとも総額型が研究開発費を促進したとは言えない。model12の $Black \times GNC$ の係数は正であるが、有意ではない。これは増加型の税額控除を適用できる状況にある企業は制度によって研究開発費の支出が促進されているとは言えないことを意味する。この結果は重要な意味を持つ。つまり、企業は税制とは関係なく、自社に必要な研究開発支出を行っている可能性が高い。

6. むすび

租税特別措置に関する政策評価について、これまでほとんど研究が行われてこなかった。本稿では、特に研究開発税制の有効性に焦点を当てて実証分析を行った。この視点での研究は他に例がない。また、会計学における研究開発費の説明モデルを用いて分析を行っており、これは先行研究にはないアプローチである。この方法を採用するこ

とによって、各年度の税制改正に対する有効性を多面的に検証することができる。

研究開発税制の有効性について、総額型制度を中心にDifference in Differencesの手法を用いて検証した。その結果、総額型制度は研究開発費を増加させているとは言えない。ただし、負債比率が高い企業には、効果が認められることが明らかになった。この点においては、先行研究と本稿は整合的である。さらに、税額控除の限度額規制による影響は認められなかった。また、繰越控除廃止後の状況を分析し、税額控除が適用できる企業とできない企業とを比較検証した。この分析は、本稿における重要な貢献の一つである。また、総額型は導入後も制度は継続しているため、長期的な効果についても検証したが、効果は認められなかった。負債比率の高い企業についても同様に効果は認められなかった。繰越控除の上限についても研究開発費への影響は認められなかった。増加インセンティブ制度についても、簡単に分析を行ったが、有効性は確認できなかった。

本稿では、研究開発費の増分を被説明変数とした検証はできなかった。研究開発費の増減を説明するモデルが検証期間では十分に機能しなかったためである。研究開発費の増減を説明するモデルの開発とそれを利用した税制の有効性の検証は今後の課題としたい。

この結果をもって、当該制度の存在意義がないと主張しているのではない。重要なことは以下の2点である。

第一は、立法趣旨と政策目的を明確にし、矛盾なく整理することである。前述のとおり、制度導入時は、試験研究を助成するものであった。現在においてもこの考え方が完全に否定されているとは言えないだろう。そうであるならば、研究開発支出が増加しなくても、趣旨どおりの結果が得られている（つまり、試験研究費を支出した企業に税制面で助成している）と解釈することも可能である。

我が国では、法律制定時に各条文の立法趣旨が明確に公表されることはない。さらに、法律の趣旨は時代と共に変化する。重要なことは、現在における法律の趣旨を明確にし、政策目的と合致させることであろう。趣旨や目的が明確でなければ、

政策の有効性を検証することは困難である。

第二に、政策の有効性を検証するモデルの開発が重要である。本稿のように財務数値を使用したパネルデータ分析を行えば、DIDなど政策評価で通常用いられる手法の適用も容易であり、公表データを利用して外部からの検証も可能である。また、2003年の総額型導入当初の有効性を評価の中心とするだけでなく、長期的な視点、現在における効果など多面的な分析が必要であろう。

以上の点を改革・改善すれば、政策評価がより実質的かつ有効な役割を果たすことが期待される。

謝辞

本研究の一部は、京都大学生存圏研究所の生存圏科学共同研究からの助成を受けて実施しました。また、同研究所の西村耕司先生には有益な助言をいただきました。この場を借りて厚く御礼申し上げます。

付記

1. 年の表記について

年号については、視認性を高めるためにも西暦を使用した方が良い。しかし、各年度の税制改正は元号で表記すべきである。政策評価書も元号表記されている。本稿では、視認性と正確性を両立させるために、西暦と元号を並記した箇所もある。

2. 政策評価書の文献表記について

政策評価書などのリンクについての記載は省略した。政策評価書は「政策評価ポータルサイト」に掲載されている。府省ごとに取り扱いは異なるが、一定期間経過後に削除され、国会図書館アーカイブから入手することになる。

注記

- 1 「民間企業は一般に規模も小さく内部留保も少ないので試験研究への投資が大きな負担となり、試験研究のための支出は景気の好不況により著しく変動しているのが実情であります。」(国税庁編 1967, p.97) このような実情を踏まえ、「民間企業の試験研究活

動に対して税制面からも積極的な助成措置を講ずることが緊要であるとされたのであります。」(国税庁編 1967, p.97) とある。

- 2 「総額型、中小企業技術基盤強化税制の適用金額については、『2020年頃に名目GDP600兆円達成』及び『民間研究開発投資を対GDP比3%以上』という目標を達成するためには、民間研究開発投資を毎年7.8%増加させる必要があることから、平成31年度税制改正要望内容に伴う適用金額に107.8%に相当する額を平成32年度推計値に、平成32年度推計値の107.8%に相当する額を平成33年度推計値とした。」(経済産業省 2018, pp.10-11) と政策評価書に記述がある。
- 3 民主党のマニフェストに「不透明な租税特別措置を全て見直して、効果の乏しいもの、役割を終えたものを廃止する。」(民主党 2009, p.4) と記載されている。
- 4 政策評価書には、「kasahara et al. (2011)」(ママ) と記載があるだけで、文献を特定する記載はない。評価書に記載されている他の引用文献も同様である。
- 5 小嶋 (2008) のモデルは次のとおりである。

$$RD_t = \alpha + \beta_1 \cdot FUNDS_t + \beta_2 GNP_t + \beta_3 RD_{t-1} + \beta_4 IRD_t + \beta_5 CAP_t + \beta_6 ICAP_t + \varepsilon_t$$

RD は研究開発支出を売上高で除したもの(添え字 t は事業年度を表す)、 $FUNDS$ は(経常利益+研究開発支出)/売上高、 GNP は実質GNPを売上高で除したものを1万で除したもの、 IRD は、 RD の産業平均、 CAP は設備投資、 $ICAP$ は CAP の産業平均、 ε は誤差項である。

本稿では、このモデルを修正した上で、税制改正関連の説明変数を加えてモデルを拡張した。主な修正は、① $FUNDS$ を経常利益ベースから営業キャッシュ・フローに変更したこと、② GNP を対数変換した GDP としたことである。①については、国際会計基準の普及に伴い、経常利益が存在しない企業が増えたこと、会計基準の選択適用の幅が増えたことから、客観的に企業の稼得能力を測定するために営業キャッシュ・フローを適用した。②については、現在では生産力を表す指数としては GDP の方が適切であること、1万で除するだけでは係数(β)がすべて0になってしまう結果の意味を読み取るのが困難になることを避けるためである。

- 6 OCF について、経常利益や営業利益を採用すると、検証期間においては、どの会計基準を採用するかに

よってデータが存在しない、あるいは、利益の内容が異なることがあるため、比較的客観的に測定されるであろう営業活動によるキャッシュ・フローを採用した。FAについては、設備投資額を採用することも検討したが、データの欠損を最小限に抑えるために、これを採用した。業種は日経業種分類の中分類による。本稿では、LEV（負債／総資産）を負債比率とする。

- 7 差分データを使用しているため、データの収集期間は2000年から2020年までである。
- 8 企業グループで研究開発費が増加しているかを検証することが目的であるため、連結優先（連結会計を実施していれば、連結データ、そうでなければ単体データ）のデータを使用している。また、証券番号により企業を管理しているので、データベースにアクセスしたとき（2024年7月13日）に上場したものに限られる。生存バイアスなどは生じるが、通常、上場廃止になる企業の多くは研究開発を積極的に拡大するとは考えにくいので、このバイアスが税制を無意味化する結論に導くことはないと思われる。
- 9 帰無仮説は、平成15（2003）年度税制改正は法人の研究開発費の増減に影響を与えないとなる。他の仮説についての帰無仮説は省略するが、同様に考えれば自明である。

参考文献

- 大西宏一郎・永田晃也（2010）「研究開発優遇税制は企業の研究開発投資を増加させるのか—試験研究費の総額に係る税額控除制度の導入効果分析—」、『研究技術計画』、24（4）：400-412
- 河瀬豊（2023）「試験研究費税制の有効性に関する実証研究—繊維産業における試験研究費税制が研究開発投資に与える影響」、『神戸学院大学経営学論集』、神戸学院大学経営学会、20（1）：35-43
- 経済産業省（2010）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2011）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2012）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2013）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2014）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2016）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2018）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2019）「租税特別措置等に係る政策の事前評価書」
- 経済産業省（2020）「租税特別措置等に係る政策の事前評価書」
- 国税庁編（1967）『改正税法のすべて 昭和42年度版』（大蔵財務協会（2003）『改正税法のすべて：法人税関係（平成15年更新版、CD-ROM版）』に収録されたものを確認）
- 小嶋宏文（2008）「期待外利益の回避と研究開発支出の裁量的調整」、『會計』、森山書店、174（1）：89-100
- 米谷健司・松浦良行（2010）「税額控除が研究開発投資に与える影響」、『會計』、森山書店、178（4）：124-139
- 総務省行政評価局（2017）「政策評価Q&A（政策評価に関する問答集）」（https://www.soumu.go.jp/main_content/000359598.pdf、2024年9月3日最終閲覧）
- 武田昌輔編（1979）『DHCコンメンタール法人税法』、第一法規、〔最終加除：追録2248号（2024年3月15日発行）〕
- 内閣府（2010）「新成長戦略」〔平成22年6月18日閣議決定〕
- 内閣府（2011）「科学技術基本計画」〔平成23年8月19日閣議決定〕
- 農林水産省（2012）「租税特別措置等に係る政策の事後評価書」
- 農林水産省（2017）「租税特別措置等に係る政策の事後評価書」
- 民主党（2009）「民主党政権政策Manifesto」（http://archive.dpj.or.jp/special/manifesto2009/pdf/manifesto_2009.pdf、2024年9月3日最終閲覧）
- Bloom, N., Griffith, R. and Reenen, J. V. (2002). Do R&D tax credits work? Evidence from a panel of countries 1979-1997, *Journal of Public Economics*, 85, 1-31.
- Hall, B. and Reenen, J. V. (2000). How effective are fiscal incentives for R&D? A review of the evidence, *Research Policy*, 29, 449-469.
- Kasahara, H., Shimotsu, K. and Suzuki, M. (2011). How

Much Do R&D Tax Credits Affect R&D Expenditures?
Japanese tax credit reform in 2003, *RIETI Discussion
Paper Series*, 11-E-072.

Kasahara, H., Shimotsu, K. and Suzuki, M. (2014). Does an
R&D tax credit affect R&D expenditure? The Japanese
R&D tax credit reform in 2003, *Journal of The Japanese
and International Economies* 31, 72-97.

Koga, T. (2003). Firm size and R&D tax incentives,
Technovation, 23, 643-648.

(2024.12.26 受理)

The Impact of a Volume-Based R&D Tax Credit on R&D Expenditures in Japan

Yutaka Kawase

Kobe Gakuin University
kawase@ba.kobegakuin.ac.jp

Abstract

Japan's corporate income tax credit system for R&D system is designed to promote investments in research and development (R&D). Government ministry and agency policy evaluation reports indicate that this system has increased R&D spending. The findings of this study indicate increased R&D spending in firms with high debt-to-asset ratios but no clear evidence that the gross system increased R&D spending. In addition, the study did not find evidence that limitations on carryforward deductions impacted R&D spending. Based on these results, Japan's corporate income tax credit system's effectiveness remains inconclusive. We suggest developing a model to test the system's effectiveness to obtain a comprehensive understanding of the current status.

Keywords

Research and Development Expenses, R&D Tax Credit, Tax Policy, Special Taxation Measures

【実践報告・調査報告】

地方自治体におけるEBPMの実態と課題 —福島県市町村調査からの示唆—

加藤 穂高

福島大学

r429@ipc.fukushima-u.ac.jp

鈴木 あい

福島大学

r430@ipc.fukushima-u.ac.jp

高森 智嗣

福島大学

h047@ipc.fukushima-u.ac.jp

要 約

本研究では、地方自治体におけるEBPMの実態と課題を明らかにする端緒をつかむため、福島県内の59市町村を対象としたアンケート調査、および特に取組が進んでいる事業等があると回答した4団体へのヒアリング調査を実施した。

調査の結果、①EBPMの推進は、大規模自治体を中心として一部にとどまっていること、②「エビデンス」や「EBPM」等の言葉がきわめて高い多義性を持っている可能性があること、③市町村の政策形成において実際に行われているのは、いわゆる「広義のEBPM」が中心であり、「独自の効果検証」や「学術論文を対象とした文献調査」といった「狭義のEBPM」を行っている団体は一部にとどまること、④EBPM推進にあたっての課題は主に手法やノウハウの不足、人手不足、庁内の理解不足であること、⑤EBPM推進の地域的な広がりについて、周辺自治体へEBPMが連鎖・波及している傾向は見られなかったことなどが確認された。本研究より、地方自治体において様々な取組が行われる一方、改めて地方自治体におけるEBPM実践には困難さがあることが示唆された。

キーワード

地方自治体、EBPM、アンケート調査、ヒアリング調査、福島県

1. はじめに

近年、EBPM（Evidence Based Policy Making）が注目を集めている。一般的に、EBPMは「エビデンスに基づく政策立案」とされるが、本研究では、内閣官房行政改革推進本部事務局（2019）の「(1)政策目的を明確化させ、(2)その目的のため本当に効果が上がる行政手段は何かなど、当該政策の拠って立つ理論を明確にし、それに即してデータ等の証拠を可能な限り求め、「政策の基本的な枠組み」を明確にする取組」という定義を用いる。

なお、後述するアンケート調査結果やヒアリング調査の分析に際しては、小林（2022）を参考に、政策立案時に行われることが想定される10の取組を①他自治体の事例調査のような「エピソードに基づく政策形成」、②公的統計等既存データの調査・分析、研究者へのヒアリング調査、独自のアンケート調査、政策目的の明確化、ロジックモデルの作成、KPIの設定、アウトカム指標の設定といった「広義のEBPM」、③学術論文を対象とした文献調査、独自の効果検証のような「狭義のEBPM」に分け、この3つの観点から各団体の取組

状況を判断することとした。

EBPMの起源は、医療行為の意思決定において、経験的知識ではなくエビデンス（科学的根拠）を利用するというEBM（Evidence Based Medicine：証拠に基づく医療）にあるとされている（田辺2020）。また、EBPMという用語が普及する以前、評価研究の文脈では、保健・福祉等の対人サービス領域におけるEBP（Evidence Based Practice：科学的根拠に基づく実践）が報告されている（大島2010）。

このような、「エビデンスに基づく」という意思決定のあり方は、政策立案のあり方にも影響していく。英国では、1992年から「コクラン共同計画」が開始されており、世界中の医療研究成果の収集と統合が行われていた。これを受けて、社会科学分野においても2000年に「キャンベル共同計画」が実施され、「エビデンスの重視」は医療分野以外の政策分野にも拡大していく（杉谷2022）。英国ブレア政権や米国オバマ政権において、EBPMに関する様々な取組が実行に移されたことは広く知られている。

日本においても「経済財政運営と改革の基本方針2017について」（平成29年6月9日閣議決定）において「統計改革推進会議最終取りまとめ」等に基づき、証拠に基づく政策立案（EBPM）と統計の改革を車の両輪として、一体的に推進する」とされ、EBPM推進の方針が示された。このようなEBPMの主たる目的は、その名称が表すとおり、政策立案や意思決定をエビデンスに基づいて行うことにあると言えるだろう。

他方、「エビデンスとは何か」ということについては未だ議論の余地があると思われるが、少なくとも「根拠に基づいた」取組は、EBPM推進以前から行われてきている。例えば、日本においては、EBPMより以前に「政策評価」が行われてきており、一定の蓄積がある。政策評価による評価の対象は、①理論（セオリー）、②実施過程（プロセス）、③改善効果（インパクト）、④効率性（コスト・パフォーマンス）であり、評価の目的は①政策に関する意思決定の改善、②資源配分の最適化・効率化、③納税者への説明責任の向上といった課題に資する情報を提供することにあるとされる（龍・佐々木2004）。

日本において、このような「政策評価」は地方自治体で普及しているが、ここで重要なのは「評価」の特性である。佐々木（2010）は、評価の本質は「価値判断」にあるとしつつ、評価が「事実特定」との組合せによって成り立つとしている。すなわち、政策評価を含む評価活動においては、何らかの根拠資料・データ等の収集や分析（事実特定）が行われているはずであり、EBPMの普及以前は「エビデンス」不在だったというわけではない。また、前述のとおり、政策評価の主たる目的は、前例主義から戦略的意思決定への転換、諸資源の配分の最適化・効率化としての資源配分、アカウンタビリティの履行にある。このように、政策評価の結果がプログラムの修正・改善、継続・停止、拡大・縮小を左右するものであるならば、政策評価やそれを取り巻くロジックモデル等のツールはEBPMと類似している、または親和性が高いと考えられる。

他方、2022年に策定された「デジタル田園都市国家構想総合戦略」（令和4年12月23日閣議決定）において、地方版総合戦略の見直しにあたってEBPMを行うことが明記されるなど、近年、地方自治体においてもEBPMの推進が求められている状況にある。

地方自治体におけるEBPMに関する研究には、広島県の取組に着目したもの（西畑2020、石田2022）等がある。また、「令和5年度 自治体経営改革に関する実態調査報告」（土方他2024）でも自治体におけるEBPMに関する調査を行っている。しかしながら、先進事例に関する研究や一般市レベルまでの調査・報告はあるものの、町村レベルまでカバーした研究については、管見の限り見当たらなかった。

政策評価での経験の蓄積は、EBPM推進においても活用可能であるものの、現在、地方自治体において、どのようなEBPMに関する取組が行われているか、どのようなことが課題となっているかなど、その実態と課題は十分に明らかであるとは言い難い。

そこで、本研究では、地方自治体におけるEBPMの実態と課題を明らかにする端緒をつかむため福島県を事例に、県内に所在する自治体（市町村）に対し、EBPMに関するアンケート調査と

ヒアリング調査を行った。福島県をフィールドとしたのは、第一に筆者らが所属する大学と各市町村との間に各種連携関係があり、アンケート調査で高い回答率が期待できる。第二に福島県は町村数では北海道、長野県に次いで全国で3番目に多いうえに、沿岸部から山間部まで多様な自治体が存在することから、特に小規模自治体の動向を把握するという本稿の目的を達成するうえで望ましい条件を備えている。第三に福島県は地域性が比較的明瞭であり、単一の県の調査であっても地域性による差を分析できる。以上、3つの理由からである。

2. 福島県におけるEBPM取組状況調査

(1) 調査目的・方法

本研究では、福島県内市町村のEBPM取組状況の定量的把握およびEBPM推進のための方策を検討する基礎資料とすることを目的に福島県内の59市町村を対象とする郵送調査を実施した。調査票の作成にあたっては、土方他（2023）を参考にしつつ、想定される県内市町村の実態に合わせた設問の設定を行った。

調査期間は2024年6月3日から21日までの約3週間とし、締切から3営業日が経過してもなお調査票の返送がなかった団体に対しては電話による督促を行った。調査票は、角2封筒に依頼状・返信用封筒と共に各市町村の企画政策担当部署に郵送し、回答は同封した返信用封筒を利用して返送、またはメールによって回答された。また、調査期間中はメールおよび電話による問い合わせを受け付けるとともに調査者の所属組織のウェブページに調査を実施している旨を掲載した。

なお、各団体には、当該団体がどのような回答を行ったのかを特定できないような形で公表する旨を伝えている。また、本調査は県内の市町村（団体）を対象とした調査であるため「福島大学におけるヒトを対象とする実験及び調査研究等に関する指針（平成18年1月17日）」に基づき倫理審査の対象外となる。

調査の結果、最終的に59市町村のうち93.22%にあたる55市町村から回答を得た。回答団体の規模

としては市が13団体、町村が42団体であった。また、福島県は県土が広域なため一般的に「浜通り」（太平洋に接する地域）・「中通り」（東北新幹線・東北道沿いの地域）・「会津」（奥羽山脈と越後山脈に挟まれた地域）の3地域に区分されるが、回答団体を地域別にみると浜通り11団体、中通り29団体、会津15団体であった（表1）。

なお、福島県内には政令市は存在しないが、中核市が3団体存在する。また、町村は半分以上が人口1万人未満の小規模団体である。

(2) 福島県内の市町村におけるEBPM取組状況

調査では、のべ21個（対象者を限る設問を含む）におよぶ事項について調査を行っているが本稿では、このうち「総合計画等へのEBPM記載状況」、「取組の実施状況」、「EBPMの推進状況」、「EBPM推進上の課題」について調査結果を概観する。なお、統計分析にあたってはLight Stone社のStata/MP 18.0を使用した。

①総合計画等へのEBPM記載状況

各団体の総合計画（またはそれに類する行政計画）にEBPMが明記されているかを尋ねたところ、「明記されている」が14.5%、「EBPMという言葉は入っていないが、同じ趣旨は記載されている」が18.2%、「いいえ」が63.6%となった（表2）。

これを団体規模別にみると、市では「明記されている」が23.1%、「EBPMという言葉は入っていないが、同じ趣旨は記載されている」が15.4%であったのに対し、町村では11.9%、19.0%と、町村に比べ市で総合計画等にEBPMが記載されている傾向が見られた。

また、被説明変数を総合計画等へのEBPM記載状況（「明記されている」=3、「EBPMという言葉は入っていないが、同じ趣旨は記載されている」=2、「いいえ」=1）、説明変数を住基人口、地域ダミー（浜通り=0とし、その他をそれぞれ1にとる、以下同じ）とする順序ロジット・モデルによる分析を行ったところ、10%水準で住基人口が有意になった。

②取組の実施状況

各団体に、施策・事業の立案または改善に際して、他自治体の事例調査、公的統計等の既存データの調査・分析、学術論文を対象とした文献調査、

表1 回答団体の規模・種別

	市	人口10万人以上	人口10万人未満	町村	人口1万人以上	人口1万人未満	計
浜通り	3/3 (100.0%)	1/1 (100.0%)	2/2 (100.0%)	8/10 (80.0%)	3/3 (100.0%)	5/7 (71.4%)	11/13 (84.6%)
中通り	8/8 (100.0%)	2/2 (100.0%)	6/6 (100.0%)	21/21 (100.0%)	8/8 (100.0%)	13/13 (100.0%)	29/29 (100.0%)
会津	2/2 (100.0%)	1/1 (100.0%)	1/1 (100.0%)	13/15 (86.7%)	3/4 (75.0%)	10/11 (90.9%)	15/17 (88.2%)
計	13/13 (100.0%)	4/4 (100.0%)	9/9 (100.0%)	42/46 (91.3%)	14/15 (93.3%)	28/31 (90.3%)	55/59 (93.2%)

(注) 分母は対象数、括弧内は回答割合。人口は2023年1月1日現在の登録人口に基づく。

(出所) 筆者作成。

表2 総合計画等へのEBPM記載状況

	明記されている	EBPMという言葉は入っていないが、同じ趣旨は記載されている	いいえ	未回答
全体 (n = 55)	14.5%	18.2%	63.6%	3.6%

【規模別】

市 (n = 13)	23.1%	15.4%	61.5%	0.0%
町村 (n = 42)	11.9%	19.0%	64.3%	4.8%

【地域別】

浜通り (n = 11)	0.0%	27.3%	72.7%	0.0%
中通り (n = 29)	27.6%	13.8%	55.2%	3.4%
会津 (n = 15)	0.0%	20.0%	73.3%	6.7%

(出所) 筆者作成。

表3 取組の実施状況

	①ほとんどの施策・事業で行っている	②一部の施策・事業で行っている	③行っていない	④把握していない	未回答	【再掲】 ① + ②
KPIの設定	38.2%	60.0%	1.8%	0.0%	0.0%	98.2%
自治体の事例調査	40.0%	54.5%	1.8%	1.8%	1.8%	94.5%
独自のアンケート調査	10.9%	81.8%	3.6%	3.6%	0.0%	92.7%
公的統計等の既存データの調査・分析	12.7%	76.4%	3.6%	5.5%	1.8%	89.1%
政策目的の明確化	49.1%	38.2%	0.0%	10.9%	1.8%	87.3%
アウトカム指標の設定	18.2%	52.7%	14.5%	12.7%	1.8%	70.9%
研究者へのヒアリング調査	5.5%	45.5%	14.5%	32.7%	1.8%	50.9%
ロジックモデルの作成	3.6%	30.9%	34.5%	29.1%	1.8%	34.5%
学術論文を対象とした文献調査	3.6%	18.2%	30.9%	45.5%	1.8%	21.8%
独自の効果検証	0.0%	20.0%	40.0%	38.2%	1.8%	20.0%

(出所) 筆者作成。

研究者へのヒアリング調査（委員会への参加を含む）、独自のアンケート調査、独自の効果検証（RCTやDIDなどの手法を用いた分析）、政策目的の明確化、ロジックモデルの作成、KPI（重要業績評価指標）の設定、アウトカム指標の設定、といった取組を行っているかを「ほとんどの施策・事業で行っている」、「一部の施策・事業で行っている」、「行っていない」、「把握していない」の4択で尋ねた。

その結果、「ほとんどの施策・事業で行っている」と「一部の施策・事業で行っている」の合計が最も多かった取組は、「KPIの設定」で98.2%、次いで「他自治体の事例調査」で94.5%、「独自のアンケート調査」で92.7%となった。一方、最も少なかった取組は、「独自の効果検証」で20.0%、次いで「学術論文を対象とした文献調査」で21.8%、「ロジックモデルの作成」で34.5%となった（表3）。

③EBPMの推進状況

各団体に、団体としてEBPMを推進しているかを尋ねたところ、「推進している」は10.9%、「推進に向けて具体的な検討を進めている」は5.5%、「推進に対して関心はあるが具体的な検討に至っていない」は72.7%、「推進の必要性を感じていない」は9.1%となった（表4）。

なお、総合計画等にEBPMが明記または同趣旨が記載されているにもかかわらず、EBPMの推進状況として「推進に対して関心はあるが具体的な検討に至っていない」または「推進の必要性を感じていない」と回答した団体が見られた。これについては、調査をとおして回答者にEBPMが適切に理解してもらえず、このような一見矛盾した回答が見られた可能性が考えられる。この点については、調査票設計上の今後の課題としたい。

また、これを団体規模別にみると、市では「推進している」が30.8%、町村では4.8%となっており、市と町村で大きな差がある。また、市では「推進の必要性を感じていない」の回答は見られないが、町村では11.9%に上っている。また、地域別にみると、中通りで「推進している」が13.8%で最も多い。ただし、「推進している」と「推進に向けて具体的な検討を進めている」の合計でみると、会津が26.7%で最も多くなる。

ただし、団体規模と地域で交絡している可能性

があるため、被説明変数をEBPMの推進状況（「推進している」=4、「推進に向けて具体的な検討を進めている」=3、「推進に対して関心はあるが具体的な検討に至っていない」=2、「推進の必要性を感じていない」=1）、説明変数を住基人口、地域ダミーとする順序ロジット・モデルによる分析を行ったところ、疑似決定係数が低く本モデルの説明力が低い点に留意する必要があるが、住基人口（5%水準）、会津（10%水準）ともに有意な関連が見られた。

では、EBPMの推進に向けた取組を進めている団体は、そうでない団体と比べてどのような取組を行っているのか。これを確認するため、被説明変数に各取組の実施状況（「ほとんどの施策・事業で行っている」=4、「一部の施策・事業で行っている」=3、「行っていない」=2、「把握していない」=1）、説明変数にEBPMの推進状況とコントロール変数として人口、地域を加えた順序ロジット・モデルによる分析を行った。その結果、推進中とした団体は、本稿で広義のEBPMとしている「公的統計等の既存データの調査・分析」（1%水準）、「ロジックモデルの作成」（10%水準）が実際に取組として行われている傾向が見られた。一方で、狭義のEBPMとしている「学術論文を対象とした文献調査」や「独自の効果検証」ではEBPM推進状況による取組実施状況の差は見られなかった（表5）。

④EBPM推進上の課題

EBPMについて「推進している」、「推進に向けて具体的な検討を進めている」、「推進に対して関心はあるが具体的な検討に至っていない」と回答した49団体に対して、EBPM推進に際して課題になっていることを複数回答で尋ねた。

その結果、最も回答が多かった課題は「手法に関するノウハウ・知識が足りない」で83.7%、次いで「人手が足りない」で57.1%、「庁内の理解が足りない」で51.0%となった（表6）。

また、EBPM推進状況によって課題が異なるかを確認するため、被説明変数にEBPM推進上の課題（課題として回答した=1、非回答=0）、説明変数にEBPM推進状況ダミー（未着手=0とし、その他をそれぞれ1にとる）とコントロール変数として人口、地域ダミーを加えたロジット・モデル

表4 EBPMの推進状況

	推進している	推進に向けて具体的な検討を進めている	推進に対して関心はあるが具体的な検討に至っていない	推進の必要性を感じていない	未回答
全体 (n=55)	10.9%	5.5%	72.7%	9.1%	1.8%
【規模別】					
市 (n=13)	30.8%	0.0%	69.2%	0.0%	0.0%
町村 (n=42)	4.8%	7.1%	73.8%	11.9%	2.4%
【地域別】					
浜通り (n=11)	9.1%	0.0%	72.7%	18.2%	0.0%
中通り (n=29)	13.8%	0.0%	79.3%	3.4%	3.4%
会津 (n=15)	6.7%	20.0%	60.0%	13.3%	0.0%

(出所) 筆者作成。

表5 EBPM推進状況別の実際の取組 (順序ロジット・モデルによる分析結果)

	(A) 他自治 体の事 例調査	(B) 公的統計等 の既存デー タの調査・ 分析	(C) 学術論文 を対象と した文献 調査	(D) 研究者へ のヒアリ ング調査	(E) 独自のア ンケート 調査	(F) 独自の 効果検 証	(G) 政策目的 の明確化	(H) ロジッ クモデ ルの作 成	(I) KPIの 設定	(J) アウト カム指 標の設 定
EBPMの推進状況 (reference = 必要なし)										
未着手	0.407 (1.034)	0.954 (1.130)	-1.028 (0.859)	0.157 (0.879)	-1.338 (1.163)	-0.552 (0.876)	0.014 (0.889)	0.159 (0.884)	-1.213 (1.079)	0.192 (0.948)
検討中	1.627 (1.599)	1.777 (1.741)	0.750 (1.379)	-0.696 (1.289)	-2.936 (2.334)	-0.074 (1.402)	-0.017 (1.379)	1.087 (1.323)	-0.962 (1.596)	-0.469 (1.381)
推進中	1.012 (1.630)	4.114** (1.844)	0.079 (1.508)	0.202 (1.473)	-0.057 (1.837)	-1.898 (1.685)	16.125 (1491.292)	2.577* (1.476)	0.057 (1.764)	0.079 (1.433)
住基人口 (千人)	-0.004 (0.007)	-0.002 (0.006)	-0.006 (0.006)	-0.005 (0.006)	-0.001 (0.007)	0.005 (0.006)	0.045* (0.023)	-0.004 (0.006)	0.038** (0.019)	0.008 (0.006)
地域ダミー (reference = 浜通り)										
中通り	1.333 (0.832)	0.945 (0.912)	-0.209 (0.693)	-0.983 (0.708)	2.769** (1.243)	0.213 (0.724)	0.161 (0.742)	-1.426** (0.716)	1.276 (0.971)	0.455 (0.712)
会津	0.879 (0.926)	2.313* (1.229)	-0.220 (0.778)	0.243 (0.858)	3.840*** (1.426)	0.225 (0.787)	0.822 (0.871)	-1.348* (0.798)	1.759 (1.080)	0.888 (0.816)
cut1	-2.847 (1.410)	-1.086 (1.150)	-1.281 (0.944)	-1.248 (0.962)	-2.988 (1.337)	-0.776 (0.921)	-1.091 (0.928)	-1.737 (0.972)	-3.560 (1.422)	-1.106 (1.002)
cut2	-2.124 (1.219)	-0.512 (1.098)	0.230 (0.920)	-0.569 (0.947)	-2.140 (1.242)	1.102 (0.927)	1.288 (0.920)	-0.049 (0.937)	1.575 (1.155)	-0.134 (0.982)
cut3	1.786 (1.142)	4.747 (1.449)	2.311 (1.101)	2.469 (1.070)	3.938 (1.584)			2.842 (1.158)		2.375 (1.042)
疑似R ²	0.057	0.166	0.042	0.047	0.169	0.014	0.147	0.065	0.197	0.032
N	53	53	53	53	54	53	53	53	54	53

(注) 括弧内は標準誤差。***、**、*はそれぞれ1%、5%、10%水準で有意であることを示す。

(出所) 筆者作成。

表6 EBPM推進上の課題（複数回答）

	全体（n=49）
手法に関するノウハウ・知識が足りない	83.7%
人手が足りない	57.1%
庁内の理解が足りない	51.0%
国による指針など実施のよりどころになるものがない	34.7%
解説書や事例集など参考になる情報が足りない	24.5%
研究者等の専門家とのネットワークが足りない	22.4%
予算が足りない	22.4%
議会の理解が足りない	10.2%
その他	8.2%
特に課題はない	2.0%

（出所）筆者作成。

表7 EBPM推進上の課題（ロジット・モデルによる分析結果）

	(1) 手法に関するノウハウ・知識が足りない	(2) 解説書や事例集など参考になる情報が足りない	(3) 研究者等の専門家とのネットワークが足りない	(4) 国による指針など実施のよりどころになるものがない	(5) 予算が足りない	(6) 人手が足りない	(7) 庁内の理解が足りない	(8) 議会の理解が足りない	(9) その他
EBPMの推進状況 (reference = 未着手)									
検討中	-1.024 (1.517)	0.000 (0.000)	1.625 (1.641)	0.695 (1.476)	0.000 (0.000)	-1.365 (1.404)	0.000 (0.000)	18.999 (5484.745)	0.000 (0.000)
推進中	-3.661** (1.743)	-1.882 (2.465)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-1.500 (1.371)	0.763 (1.300)	0.000 (0.000)	1.867 (1.819)
住基人口（千人）	0.009 (0.008)	0.007 (0.010)	0.024 (0.019)	0.015 (0.018)	-0.031 (0.026)	-0.007 (0.009)	-0.004 (0.006)	-0.629 (0.408)	0.006 (0.008)
地域ダミー (reference = 浜通り)									
中通り	0.147 (1.059)	-0.488 (0.850)	1.293 (1.174)	-0.311 (0.832)	-1.027 (0.878)	-0.163 (0.921)	-0.070 (0.784)	-0.029 (1.453)	-1.799 (1.332)
会津	0.000 (0.000)	-0.017 (1.015)	0.036 (1.530)	-1.178 (1.078)	-2.299* (1.307)	-0.289 (1.063)	0.525 (0.945)	-18.327 (5484.745)	0.000 (0.000)
切片	1.688* (0.896)	-0.817 (0.763)	-2.427** (1.173)	-0.256 (0.775)	0.529 (0.841)	0.983 (0.842)	-0.105 (0.706)	2.240 (2.457)	-1.831** (0.891)
疑似R ²	0.1741	0.0265	0.0937	0.0428	0.108	0.1097	0.0197	0.478	0.2376
N	40	46	43	43	40	48	46	43	37

（注）括弧内は標準誤差。***、**、*はそれぞれ1%、5%、10%水準で有意であることを示す。

（出所）筆者作成。

による分析を行った。その結果、疑似決定係数が低く本モデルの説明力が低い点に留意する必要があるが、推進中とした団体では5%水準で「手法に関するノウハウ・知識が足りない」を課題として挙げない傾向が見られた（表7）。

（3）アンケート調査結果からの考察

土方他（2024）と同じく、福島県内の市町村においても団体規模が大きいのほどEBPMを「推進している」と回答する割合が高くなる傾向が見られた。特に、町村においては全体の7割以上で「推

進に対して関心はあるが具体的な検討に至っていない」と回答しているだけでなく、1割強で「推進の必要性を感じていない」と回答しており、大規模団体とのEBPMに対する温度差が大きいことが伺える。

また、取組の実施状況をみると、「KPIの設定」や「他自治体の事例調査」は多くの団体で実施されている一方で、狭義のEBPMである「独自の効果検証」や「学術論文を対象とした文献調査」は広く普及しているとは言い難い状況である。

加えて、広義のEBPMである「ロジックモデルの作成」も3割強であるほか、「公的統計等の既存データの調査・分析」も「ほとんどの施策・事業で行っている」との回答は12.7%にとどまっており、広義のEBPMも基礎自治体においては広がりや欠いていると言わざるを得ない。

ただし、現実問題として小規模団体において（狭義・広義を問わず）EBPMを推進することは技術的な問題が大きい点を考慮する必要があるだろう。EBPM推進上の課題として、「手法に関するノウハウ・知識が足りない」との回答が最も多くなっている。一方、EBPM推進中とした団体では、ノウハウ・知識が課題としてあまり挙がっていない点を考えると、技術的ハードルが最大のネックとなっていると考えられる。

もちろん、この点については大規模団体、さらには中央省庁においても同様だが、特に地方の小規模団体の場合は技術的な助言ができる人材が団体の内外を問わず十分にいないと考えられる。「研究者へのヒアリング調査」を「ほとんどの施策・事業で行っている」「一部の施策・事業で行っている」と回答した団体は50.9%に上っているものの、他の取組と比べるとやや低い。一方で、課題として「研究者等の専門家とのネットワークが足りない」を挙げた割合は全体でも22.4%にとどまっている。この点については、そもそも自治体が外部に技術的助言を求めるという意識が高くない可能性がある。

以上を踏まえると、基礎自治体においては、以前より指摘されているように他自治体の事例調査に代表される「エピソードに基づく政策形成」が多い一方、学術論文の参照や効果検証の実施といった「狭義のEBPM」だけでなく、ロジックモ

デルの作成や既存データの調査・分析といった「広義のEBPM」も広まっているとは言えない状況にある。ただし、一部の団体においてはEBPMの推進に向けた取組が進んでおり、団体によって濃淡がある。

3. EBPM推進団体に対するヒアリング調査

(1) 調査目的・方法

上述したEBPM取組状況調査では、「EBPMの推進状況」について「推進している」または「推進に対して具体的な検討を進めている」と回答した団体に対して、特に取組が進んでいる事業等の有無、そして、取組の詳細について尋ねている。本研究では、特に取組が進んでいる事業等があると回答し、取組についてのヒアリングに応諾した4団体（A～D）に対し、取組の詳細について尋ねるため、ヒアリング調査を実施した。

具体的には、2024年7月30日（団体A、B）、8月1日（団体C）、8月8日（団体D）に各団体を訪問し、対面で実施した。ヒアリング調査の所要時間は各団体1時間半程度だった。質問事項は、各団体に事前に送付した。質問事項は各団体によって異なるが、主な質問事項は、EBPMの推進を始めたきっかけ、EBPMの推進状況の詳細、施策・事業の立案または改善に際して行っている取組の詳細、EBPM推進上の課題の詳細である。

(2) ヒアリング結果

①団体A

団体Aは、福島県内の市である。団体Aでは近年、EBPMやデータ利活用に関する様々な取組が行われているがその背景として、現首長のデータ重視の姿勢が大きい。例えば、2016年度には全国の自治体で構成される「オープンガバメント推進協議会」に団体Aとして加入したほか、2017年度には県内大学の教員による統計に関する庁内研修等も実施している。

特に大きな動きとしては組織改編がある。2018年度には、これまで統計を所管していた係を、主に庁内の情報システムを所管する部署から総合計画を推進する部署に移管し、さらに、2022年11月

には係から課への昇格を行った。ただし、部署の人員は従来の5名から変更はない。そのため、国の基幹統計調査等の通常業務に追われており、データ分析・全庁的なEBPM推進は今後の課題になっている。

団体AではEBPMに関する取組として、2023年度の「EBPM推進統括リーダー」および「EBPM推進リーダー」の設置が挙げられる。EBPM推進リーダーは、所属ごとに選任され、EBPMに関する取組を進める役割を担い、2024年4月現在で116名がいる。また、EBPM推進統括リーダーはEBPM推進リーダーからその統括として選任される。2024年度にはEBPM推進リーダー向けにEBPMの重要性やe-Statの使い方（総務省統計局主催の研修）に関する研修を企画し実施する予定である。

ただし、各職員のデータ利活用に関するスキルセットや資格保有状況等は把握していない。また、現時点でEBPM推進リーダー等によるEBPMに関する具体的な取組が行われているわけではない。さらに、通常業務に加えデジタルリーダー等他の職務を兼務しているほか、選任者の人事異動もある。現在は、団体Aにおける人口や各産業に関するデータの「見える化」に主眼を置き、各所属のEBPM推進リーダーと協力しながら、市政のデータ集を作成・公表している。

団体Aにおいては、庁内全体として、EBPMの必要性に対する認識は高まっているという。その要因として、上述したような首長の姿勢の他に、各所属に担当者が置かれていることや課の新設によって窓口が明確化し問い合わせがしやすくなったことなどが考えられるという。

②団体B

団体Bは、福島県内の市である。団体Bでは、行政評価（事務事業評価）においてロジックモデルの試行的導入を行っている。

団体Bでは長年、総合計画をもとに行政評価を行ってきた。現行の総合計画は2017年度から10年間で計画期間となっており2022年度は中間年であった。このとき、行政評価の実効性が課題となっていた。また、同時期に少子化・人口減少対策が市政の重要課題となっており、政策の有効性の検証も課題となっていた。その中で市幹部から、ロ

ジックモデルを取り入れてはどうかという意見が挙がり、行政評価の所管課である企画調整課内での検討を経て、2023年度行政評価から一部事業でのロジックモデルの導入が始まった。

行政評価（前年度事業の総括と来年度新規事業の事前評価）は例年5月に始まり、11月に評価結果の公表を行う。2023年度のロジックモデル導入は新規事業と少子化・人口減少対策事業を対象とすることになり、最初に企画調整課ではロジックモデル等を加えた行政評価フォーマットを作成した。そのうえで、ロジックモデルの浸透を企図して、外部講師（民間コンサルティング会社社員）による庁内研修（幹部職員向け・一般職員向けの二部立て）を行政評価が始まる5月に実施した。その後、原課においてフォーマットの記入を行い、8月ごろの「新規事務事業ヒアリング」（原課・企画調整課・財政課・人事課で構成）、「経営戦略会議」（市長、副市長等で構成）を経て、事業実施可否の判定を行っている。

ロジックモデルの導入に対しては、幹部級からの受け止めは好意的なものが多いとのことだった。特に、総合計画が自治基本条例に基づくものであり、議会から総合計画と新規事務事業の対応関係を言及されることも多いため、ロジックモデルによりその説明がしやすくなるという側面が評価されている。2024年度の行政評価では、ロジックモデル導入対象事業を新規事業とそれと関連する他事業に変更し、作成されたロジックモデルのチェックのため「新規事務事業ヒアリング」に外部有識者（前述の民間コンサルティング会社社員）も同席することとなった。

一方で、ロジックモデル作成の対象事業ながらロジックモデルの導入がそぐわない事業（法令で実施が義務付けられている事業等）やフォーマットに記載されているKPI（現行総合計画に記載のKPIを記載するルールとなっている）が適当でない事業も存在している。この点について、企画調整課としては総合計画に基づいた事務事業が求められるうえ、予算査定や人事配置の庁内ルールに則った運用が求められるため一定の齟齬が生じる点はやむを得ない面があると考えている。ただし、この点については次期総合計画策定にあたって考慮する（総合計画に記載するKPIはアウトカム指

標に限定する、進行管理に使用するアウトプットは個別計画に記載するなど）ことで課題の解消を図る予定とのことだった。

ロジックモデルに限らず団体Bでは行政評価のフォーマットの簡略化等行政評価の改善を続けており、幹部級では行政評価の必要性の認識は共通のものとしてある。また、行政評価と予算編成が紐づいていたり、外部評価が行われたりする点も大きい。加えて、原課レベルでは行政評価が事業廃止の説明に使えるという面もあるという。

団体BにおいてEBPMの取組が進んでいる他の事例として「総合戦略」における地域幸福度（Well-Being）指標の導入がある（地域幸福度指標については団体Cで詳述）。団体BではWell-Being指標を進捗管理に用いることは困難であると考えているため、あくまで参考指標として位置付けているが、政策の必要性を説明する際には一定の有用性はあるとも認識しているとのことだった。

③団体C

団体Cは、福島県内の町村である。団体CはEBPMに関する取組が進んでいる事例として地域幸福度（Well-Being）指標（以下、単に「Well-Being指標」）を用いた総合計画策定を挙げた。

Well-Being指標とは、一般社団法人Xが開発・普及促進を行っている指標で、住民のWell-Being等を主観指標と客観指標から数値化・可視化するものである。団体CでWell-Being指標の活用を検討し始めたきっかけは、2023年度に同団体のCDO（Chief Digital Officer、最高デジタル責任者）から、団体Bで開催されたWell-Being指標についてのセミナーを紹介され、受講したことだったという。その後、一般社団法人Xと関係ができたことに加え、小規模団体でのWell-Being指標の実証を行いたかったXの意向とも合致したことから、団体CにおいてWell-Being指標の活用に向けた本格的な検討が開始されるに至った。

具体的には、2024年度から始まった総合計画策定作業にあたって、団体Cの主観指標を住民アンケート調査により調査・集計を行い、その結果を総合計画検討会議（住民委員25名から構成）・策定プロジェクトチーム（町職員25名から構成）での重点政策選定において参照する資料の一つとして用いることを想定している。今後、Well-Being

指標とその他の指標の関連性をロジックモデルで明確化できれば、KPIとしてWell-Being指標を用いることも想定しているという。

また、団体CにおけるEBPM推進の特徴として、文献調査や研究者へのヒアリングの実施も挙げられていた。具体的には、新規事業の立案・検討に際しては、他自治体（近隣の町村）への情報照会や公的統計の参照は必ず行っているという。また、文献調査等は日ごろから関係を構築している研究者等に情報提供をお願いすることが多いという。このような域外の有識者とのネットワークは、元首長が研究者に熱心に協力を仰いでいたことが大きく、その後、関係を築いた方からさらに他の方に繋げてもらうなどの循環ができているという。

④団体D

団体Dは、福島県内の町村である。団体Dでは外部団体と連携した観光データ分析が先進事例として挙げた。

取組が始まった経緯は比較的偶発的な側面が大きい。そもそも、団体Dでは首長から、域内の若者の意見を集約する枠組みが必要ではないかという話が出され、その枠組みのファシリテーターを、他県にある大学の教員（Y氏）に委託した。また、Y氏はEBPMに取り組んでいる一般社団法人Zの立ち上げメンバーとかねてから面識があった。一般社団法人Zは、EBPMを自治体で実践したいと考えていたことから、その話をX氏経由で団体Dに打診したことで取組が始まったという。Y氏から一般社団法人Zの話を最初に受けた際には、団体DのCDO（団体DのCDOは団体CのCDOを兼務している）からデータ活用の重要性の話も聞いていたこともあり、企画政策担当課として対応することとしたという。

団体Dと一般社団法人Zは、2022年7月からオンラインベースで打合せを開始した。議論の結果、域内にある道の駅を主な対象とした観光客のデータ分析を行うこととなった。具体的には、民間事業者が提供する位置情報データや道の駅の売上データを使用して、道の駅への来訪状況の分析や周辺宿泊施設の利用状況を重回帰分析等で分析を行った。データ分析は一般社団法人Zが行い、データは団体Dが提供している。なお、民間事業者が提供する位置情報データは有料サービスだが、団

表8 ヒアリング調査対象団体の特徴

団体	取組概要	発案者	取組の広まり
団体A（市）	各課室におけるEBPM推進担当者の任命	首長	全庁的な展開
団体B（市）	行政評価へのロジックモデル導入	自治体幹部	全庁的な展開
団体C（町村）	総合計画策定でのWell-Being指標の活用	外部団体	—
団体D（町村）	外部団体と連携した観光データ分析	外部団体	一部での展開

（出所）筆者作成。

団体Dがこれまで行っていた利用者数調査を廃止し、その財源を充当した。一般社団法人Zへの委託料等は生じていない。

さらに現在は、観光担当部署と連携して町のプロモーション（web広告を想定）の費用対効果分析を試みている。企画政策担当課では、これらで得られた成果を団体Dが策定する観光戦略に盛り込むことが当面の目標となると認識しているという。

また、これらの動きを首長や一部の議員はおおむね好意的に捉えているが庁内全体の理解についてはまだまだ課題があるという。例えば、得られたデータ分析結果は、これまでも庁内で感覚的には認識されていたことでもあり、「わざわざデータ分析を行う必要があるのか」という反応がされることもあるという。将来的に、職員自身にデータ分析のスキルを身につけてほしいとは考えていないが、データを使って推移を見て、考えるという姿勢は身につけてほしいと企画政策担当課としては考えているという。なお、担当課としては、団体Dにおける現在の取組は一部にとどまっており、全庁的な展開とは言い難いため「EBPMを推進している」状況ではないと考えている。ただ、業務効率化を考えると、データ利活用等をおして不要な事業の廃止等にもつなげていきたいという思いもあるとのことだった。

（3）ヒアリング調査結果からの考察

本章では、EBPM推進に関して「特に取組が進んでいる事業等」があると回答した福島県内の4団体に対して行ったヒアリング調査の結果をまとめた。いずれの団体も取組を開始した経緯やその内容は異なっており、共通点を見出すことや一

般化することは難しい。そこで、4団体が推進する「EBPM」の特徴を2点挙げておきたい（表8）。

①2種類のEBPM推進者の存在

調査時点において、地方自治体にはEBPM実施を義務付ける法令は存在していない。そのため、団体としてEBPMに取り組むには法的拘束力以外の推進力が必要となる。その役割を団体Aや団体Bでは、首長や自治体幹部といった内部のトップ（級）が担った。一方、団体Cでは外部団体が開催した研修会に参加した政策担当者がそのアイデアを庁内に持ち込み、団体Dでも外部団体からのアプローチを受けて取組が開始されており、外部団体がEBPMの推進役として機能している。

また、庁内のEBPMの広まりをみると、内部のトップ（級）が推進役となった団体A・Bでは全庁的な取組として展開されている。一方、外部団体が推進役となった団体Dでは、一部の部署への波及は見られるもののその広がりには現状として限定的である。団体Cにおいては総合計画策定プロセスに「EBPM」を取り入れる取組のため全庁的な展開という側面もあるものの、現在進行形で行っているため現段階での評価は難しい。ただし、団体C・DともにEBPMが全庁的に取り組まれているとは言い難い。

今回のヒアリング調査の対象となった4団体のうちトップ級がEBPMを推進している団体では①比較的規模の大きい自治体であること、②職員の意識変化がみられること、③「広義のEBPM」ととどまっていること、という特徴が見られる。また、外部団体がEBPMを推進している団体では①比較的規模の小さい自治体であること、②庁内理解に課題があること、という特徴が見られる。ただ、繰り返しになるが、いずれもどこまで一般化

が可能な特徴であるかは今後の検討課題である。

②「EBPM」の多義性

もう一つの特徴として「EBPM」の多義性を指摘しておきたい。今回の調査では、内閣官房行政改革推進本部のEBPMに関する説明を定義として用いたが、実際の現場では多種多様な数字や事柄が「エビデンス」や「EBPM」として扱われていた。

例えば、団体Aでは組織改組や役職の設置、研修の実施等がEBPMとして扱われていた。また、団体Bではロジックの裏付けとなる因果関係を示すエビデンスの収集には至っていなかったが、行政評価におけるロジックモデルの導入がEBPMとして扱われていた。導入されたロジックモデルは事業実績・短期成果・中期成果・最終成果と成果の推移（KPI）から構成される比較的シンプルなものであり、現行の庁内の枠組みに合わせたものを作成していた。このように各団体で実現可能な取組から着手しており、それを「EBPM」と呼称しているのが現時点での状況と言えるのかもしれない。

また、参照する「エビデンス」の捉え方も団体によって様々であった。顕著な例がWell-Being指標の捉え方である。団体Cでは、Well-Being指標の活用そのものをEBPMと捉え、同指標の有用性に大きな期待を寄せていた。一方で、団体BでもWell-Being指標を取り入れているが、一定の有用性を評価しながらも実務面での活用の難しさを認識しており、活用には比較的慎重な姿勢であった。すなわち、ある分野に関する主観的な指標の変動を見ただけでは、その変動が事業の効果によるものか否かの判断がつかないという認識である。このように同じWell-Being指標をEBPMの文脈で使用しているが、捉え方は全く異なっていた。

以上のようにEBPMを推進する地方自治体の現場においてもEBPMの捉え方に多義性が存在していた。もちろん、このような現象は政府におけるEBPM推進議論でも指摘（三輪 2020）されていたことであり、真新しさはない。ただ、地方自治体の現場においても同様の事象が生じており、実現可能な取組を試行錯誤しながら実施する中で「エビデンス」や「EBPM」の捉え方が揺らぎ、結果としてこれらの言葉がきわめて高い多義性を持っている可能性がある。

4. おわりに

本稿では、福島県内の市町村を対象にEBPMの取組状況をアンケート調査とヒアリング調査を行い、地方自治体におけるEBPMの実態と課題を調査した。調査で明らかになった点としては以下のとおり。

①EBPMを推進していると回答した市町村は大規模自治体を中心として一部にとどまっている。特に町村においては推進していると回答した団体がさらに限られるだけでなく、そもそもEBPMの必要性を感じていないという団体も一定数存在しており、市町村の間にも温度差があることが明らかになった。

②ただし、EBPMを推進していると回答した団体においても、そうでない団体と比較したときに具体的な取組の実施状況に統計的に有意な差が確認できたのは、ロジックモデルの作成等、一部の取組に限られていた。また、ヒアリング調査においても具体的な取組につながっているケースは限られていた。EBPMを推進していると回答した市町村の多くでは、現時点では実現可能性が高い取組をEBPMと呼称していた。

③市町村の政策形成において実際に行われているのは「KPIの設定」や「他自治体の事例調査」といったいわゆる「広義のEBPM」が中心であり「狭義のEBPM」とされる「独自の効果検証」や「学術論文を対象とした文献調査」を行っている団体は一部にとどまった。

④市町村におけるEBPM推進にあたっての課題は主に手法やノウハウの不足、人手不足、庁内の理解不足であった。これらの課題に対して、EBPMに関する具体の取組が進んでいる団体においては、手法やノウハウ不足を主に外部有識者との連携によって解決を図っており、団体トップ級がEBPMを推進している団体においては庁内理解が比較的進んでいく傾向が見られた。

⑤EBPM推進の地域的な広がりとしては、アンケート調査では県内の一部地域で集中している傾向が見られたものの、ヒアリング調査からは積極的に周辺自治体へEBPMが連鎖・波及している傾向は見られなかった。

以上の結果からは、改めて地方自治体における

EBPM実践の困難さが示されたとも言える。もちろん、限られたリソースの中でEBPMが政策の有効性の向上やアカウンタビリティの履行に資するのであれば、福祉政策等を最前線で担いながらも財源・人員に限りのある地方自治体においてEBPMを推進していく必要性は高いはずである。一方で、RCTに代表される厳密な効果測定や海外文献を含む学術論文の調査等を実施するのは地方自治体の現状のリソースからすれば、技術的ハードルの高さは否めない。必要性和高い難易度のはざま（幸いにして技術的ハードルが克服できた一部の団体を除いて）様々な取組がEBPMと呼ばれたり、そもそもEBPM推進をあきらめたりする団体が出てくることは、やむを得ないと言っても良いだろう。もし、この状況を打開したいのであれば、地方自治体に対する何らかの技術的なサポートを充実させるか、地方自治体の実情に合わせたEBPMの形を模索していく他ないはずである。

最後に、繰り返しになるが、本稿は福島県という限られた地域を対象とした調査であり、サンプルサイズが限られるうえ、ここで見出だされた傾向がどこまで一般化できるか不明確な点で限界がある。もちろん、地域が限定されとはいえ、地方自治体における状況を定量・定性の両面から明らかにした意義はあるが、全国の自治体においても同様の状況であるという保証はない。今後、他の自治体においても、同様の調査が実施されることで、地方自治体におけるEBPMの現状の理解が深まることが期待される。

謝辞

本稿の執筆にあたり「福島県におけるEBPM取組状況調査」およびヒアリングにご協力いただいた市町村ご担当者の皆様、ならびに実査にあたって事務処理を担っていただいた福島大学「地域×データ」実践教育推進室の新妻佳奈氏に感謝の意を表します。

参考文献

石田直人（2022）「広島県におけるEBPM推進の歩み」、大竹文雄・内山融・小林庸平編著『EBPM エビデ

ンスに基づく政策形成の導入と実践』、307-347、日本経済新聞出版社

大島巖（2010）「精神保健福祉領域における科学的根拠にもとづく実践（EBP）の発展からみたプログラム評価方法論への貢献～プログラムモデル構築とフィデリティ評価を中心に～」、『日本評価研究』、10（1）：31-41

小林庸平（2022）「政策評価とEBPMをどう連動させるか？—基本的な考え方と実践例—」、『評価フォータリー』、2022年4月号：27-47

佐々木亮（2010）『評価論理【評価学の基礎】』、多賀出版

杉谷和哉（2022）『政策にエビデンスは必要なのか—EBPMと政治のあいだ—』、ミネルヴァ書房

田辺智子（2020）「エビデンスに基づく政策立案（EBPM）の推進に向けて—医療の経験からの示唆—」、『日本評価研究』、20（2）：19-31

内閣官房行政改革推進本部事務局（2019）「資料1 EBPMの推進について」、第4回EBPM推進委員会、<https://warp.ndl.go.jp/info:ndljp/pid/11987457/www.kantei.go.jp/jp/singi/it2/ebpm/dai4/gijisidai.html>（最終閲覧2024年9月13日）

西畑宏治（2020）「地方自治体におけるEBPMの組織への定着に向けた課題の考察」、『日本評価研究』、20（2）：65-76

土方孝将・鈴木淳・大塚敬・沼田壮人（2023）「令和4年度自治体経営改革に関する実態調査報告」、https://www.murc.jp/library/report/seiken_230728/（最終閲覧2024年7月19日）

土方孝将・鈴木淳・大塚敬・沼田壮人・片平春樹（2024）「令和5年度自治体経営改革に関する実態調査」、https://www.murc.jp/library/report/seiken_240910/（最終閲覧2024年9月19日）

三輪芳朗（2020）「日本政府のEBPM推進の取組み、その開始から現状に至る過程—大橋弘編[2020]『EBPMの経済学—エビデンスを重視した政策立案』（東京大学出版会）の刊行を契機に」、『経済学論集』、83（1）：55-122

龍慶昭・佐々木亮（2004）『「政策評価」の理論と技法』、多賀出版

（2024.12.26 受理）

The Actual Situation and Challenges of EBPM in Local Governments —Implications from a Study in Fukushima Prefecture—

Hodaka Kato

Fukushima University
r429@ipc.fukushima-u.ac.jp

Ai Suzuki

Fukushima University
r430@ipc.fukushima-u.ac.jp

Tomotsugu Takamori

Fukushima University
h047@ipc.fukushima-u.ac.jp

Abstract

The current study performed a questionnaire survey on evidence-based policy making (EBPM) of 59 municipalities and interviews with four municipalities with advanced EBPM in Fukushima Prefecture to understand the situations of EBPM among small municipalities in countryside. The results of the current study revealed that 1) the situation of promoting EBPM varies from municipality to municipality; the larger the size of the municipalities, the higher the percentage of respondents who answered that they have promoted EBPM, 2) the words “evidence” and “EBPM” may have the ambiguity, 3) policy making in local municipalities is mainly based on so-called “EBPM in a broad sense”, 4) the challenges in implementing and promoting EBPM in municipalities are mainly lack of methods, manpower, and understanding within the municipalities, and 5) there was no trend of EBPM actively cascading or spreading to surrounding municipalities. These findings suggested that while various efforts have been made, there are still difficulties in implementing and promoting EBPM in local municipalities.

Keywords

Local Government, EBPM, Questionnaire Survey, Interview Survey, Fukushima Prefecture

日本評価学会第25回全国大会
「評価のインパクト～Impact of Evaluation～」
開催の報告とお礼

2024年12月21日（土）、22日（日）開催（於：早稲田大学）の日本評価学会第25回全国大会には、198名の方々にご出席いただきました。誠にありがとうございました。会員各位の日頃の研究や実践活動の報告をもとに、評価研究者、実務者の間の経験、情報、知識を共有化する場として、皆様にとって有益な機会となりましたら幸甚に存じます。多くの方のご尽力により本大会を開催することができましたこと、心より感謝申し上げます。今後とも当学会の活動にご高配賜りますようお願い申し上げます。

第25回全国大会実行委員長 田辺智子（早稲田大学）
企画委員長 南島和久（龍谷大学）

第1日目 2024年12月21日（土） ※カッコ内は座長名

09:30-11:30	(井深大記念ホール) 基調講演「業績測定の影響データはどのように使われるのか」 Donald Moynihan, Ph.D. (ミシガン大学公共政策大学院 教授) (同時通訳付き)		
11:40-	理事会 (会場3)		
13:00-15:00	共通1 (会場1) 「JICA事業評価におけるDXの推進について」 (佐藤真司)	RTD1 (会場2) 「日本の政策評価のこれまでとこれから —山谷清志を囲んで—」 (湯浅孝康)	自由1 (会場3) 「評価技法」 (新藤健太)
15:10-17:10	共通2 (会場1) 「学校評価マップの活用 —学校評価学と実践の往還—」 (小澤伊久美)	共通3 (会場2) 「政策形成・評価に関する改革の 取組～総務省行政評価局に おける最近の事例～」 (尾原淳之)	自由2 (会場3) 「規制・社会実装と評価」 (佐藤由利子)
17:10-17:50	総会、学会賞授与式 (会場1)		

第2日目 2024年12月22日（日）

09:30-11:30	共通4（会場1） 「中国の対外経済協力の現状とDAC 評価による国際協調の可能性」 （河野 撰）	共通5（会場2） 「エビデンスと政治」 （杉谷和哉）	自由3（会場3） 「ODA評価」 （石田洋子）	RTD2（会場4） 「社会的インパクト評価に関する現状整理と今後」 （今田克司）
11:40-	各委員会（会場1・2・3・4）			
13:00-15:00	共通6（会場1） 「発展的評価（DE）の実践報告から見えること：DEの守備範囲と効果」 （中谷美南子）	共通7（会場2） 「国立研究開発法人の評価制度と研究評価：機関業績とイノベーション創出と科学的インパクト」 （白川展之）	自由4（会場3） 「インパクト評価」 （齊藤貴浩）	RTD3（会場4） 「実務者と研究者の交流に基づく学術論文執筆ラウンドテーブル・その2:「投稿」から「査読」へ」 （米原あき）
15:10-17:10	RTD4（会場1） 「評価能力向上のための国際的ネットワーク構築」 （石田洋子）		共通8（会場3） 「評価の利用を促進する評価結果の報告方法」 （窪田好男）	

日本評価学会第25回全国大会 「評価のインパクト～Impact of Evaluation～」 プログラム詳細

○印は主発表者

1日目（2024年12月21日（土））	
09:30 ～ 11:30	
井深大記念ホール	
基調講演「業績測定のインパクト—業績データはどう使われるのか—」 講師：Donald Moynihan, Ph.D.（ミシガン大学公共政策大学院教授）（同時通訳付き） 司会：田辺智子（早稲田大学）	
13:00 ～ 15:00	
会場1	
共通論題1：JICA事業評価におけるDXの推進について 座長・司会：佐藤真司（JICA） 討論者：正木朋也（JICA） - 趣旨説明 - 報告1：○阿部俊哉・西野恭子（JICA）「JICAの事業評価の概要」 - 報告2：長野悠志（JICA）「JICA事業におけるデータ活用枠組と事業評価への示唆」 - 報告3：○三原智子・丸山真司（JICA）「JICA事業評価における衛星データの利活用の課題と展望」 - 総合討論：質疑応答・フロアとの意見交換	
会場2	
ラウンドテーブルディスカッション1：日本の政策評価のこれまでとこれから—山谷清志を囲んで— 座長・司会：湯浅孝康（大阪国際大学） - 趣旨説明 - 話題提供1：山谷清志（同志社大学） - 話題提供2：南島和久（龍谷大学） - 話題提供3：山谷清秀（大阪経済大学） - 話題提供4：鏡圭佑（日本大学） - 話題提供5：三上真嗣（長野県立大学） - 全体のまとめ	
会場3	
自由論題1：評価技法 座長・司会：新藤健太（日本社会事業大学） - 津富宏（立教大学）「対人支援の成果をどう測るか：批判的検討」 - 西村君平（東北大学）「不確実性に適応的なシステムティックレビュー—RAMESSESプロジェクトを中心に—」 - 陳凌雲（同志社大学大学院生）「地域福祉実践における多様な主体が評価プロセスに参加する意義」	
15:10 ～ 17:10	
会場1	
共通論題2：学校評価マップの活用—学校評価学と実践の往還— 座長・司会：小澤伊久美（国際基督教大学） 討論者：石田洋子（広島大学） - 趣旨説明 - 報告1：橋本昭彦（日本女子大学）「学評マップの概要と外国のマップから考える活用案」 - 報告2：佐々木保孝（天理大学）「社会教育人材にとっての学評マップ—学校評価士に期待されること—」 - 報告3：石田健一（計画と評価専門家）「評価者と社会運動—協力と変革の道筋」 - 報告4：小澤伊久美（国際基督教大学）「学会運営の評価から学評マップへの示唆—組織評価と『評価の場作り』—」 - 指定討論 - 全体ディスカッション	

会場2
共通論題3：政策形成・評価に関する改革の取組～総務省行政評価局における最近の事例～ 座長・司会：尾原淳之（総務省） 討論者：南島和久（龍谷大学） ・趣旨説明 ・報告1：木村智宏（総務省）「租税特別措置等の効果検証手法について」 ・報告2：高崎浩一（総務省）「浄化槽行政に関する調査」 ・報告3：堤雅也（総務省）「太陽光発電設備等の導入に関する調査」 ・討論者と各発表者のディスカッション ・質疑応答・意見交換
会場3
自由論題2：規制・社会実装と評価 座長・司会：佐藤由利子（日本学生支援機構） ・村上裕一（北海道大学）「EU における政策評価の進展：「より良い規制」に向けた事後評価の活用」 ・〇渡邊直之・宜保友理子（慶應義塾大学）「研究開発型スタートアップへのインパクト投資の可能性と課題」 ・野呂高樹（政策研究大学院大学）「欧州連合（EU）における研究＆イノベーション（R&I）施策の評価に関する一考察—Horizon 2020 の事後評価および Horizon Europe の中間評価を事例に一」 ・齊藤貴浩（大阪大学）「大学の社会的インパクトと大学評価の社会的インパクト～大学が自らの価値を社会と利害関係者に示す評価とするために～」
17:10 ～ 17:50
会場1
総会、学会賞授与式
2日目（2024年12月22日（日））
09:30 ～ 11:30
会場1
共通論題4：中国の対外経済協力の現状と DAC 評価による国際協調の可能性 座長・司会：河野摂（外務省） ・趣旨説明 ・報告1：徐一睿（専修大学）・佐藤一光（東京経済大学）・〇河野摂（外務省）「中国の対外経済協力の現状と『開発協力』と『投資』が混在する中での DAC評価の役割～霞山会研究会成果報告を元に」 ・報告2：村岡敬一（OECD/DAC開発評価ネットワーク前副議長）「DAC評価基準改定と中国対外援助へのインプリケーション」 ・ディスカッションおよびフロアとの意見交換 ・質疑応答
会場2
共通論題5：エビデンスと政治 座長：杉谷和哉（岩手県立大学） 司会：鏡圭佑（日本大学） 討論者：松村一志（成城大学） ・報告1：杉谷和哉（岩手県立大学）「エビデンスの政治性について」 ・報告2：清水右郷（京都大学）「方法論的進歩と望ましい結果を政策で結ぶ」 ・報告3：清水習（宮崎公立大学）「『科学的エビデンス』の何が政治的なのか？」
会場3
自由論題3：ODA評価 座長・司会：石田洋子（広島大学） ・牟田博光（国際開発センター／大妻女子大学）「ミャンマー連邦共和国におけるコロナ禍とクーデターによる人的資源蓄積減少の評価」 ・〇倉田正充（上智大学）・青柳恵太郎（メトリクスワークコンサルタンツ）「災害時の緊急現金給付が貧困層のフード・セキュリティに与える影響—バングラデシュにおける自然実験分析—」 ・高橋真美（元早稲田大学大学院博士課程）「持続性の評価基準と日本の政府開発援助」

会場4
ラウンドテーブルディスカッション2：社会的インパクト評価に関する現状整理と今後 座長：今田克司（CSO ネットワーク／ブルー・マーブル・ジャパン） 司会：米原あき（東洋大学） ・趣旨説明 ・発表1：今田克司（CSO ネットワーク）「『社会的インパクト評価に関する現状整理』概説について」 ・発表2：「『社会的インパクト評価に関する現状整理』キーポイント」 - 津富宏（立教大学）（民間非営利活動の分野から見て） - 桑原亜希子（日立製作所）（社会イノベーション事業の分野から見て） - 中村美亜（九州大学）（文化芸術の分野から見て） 「社会的インパクト評価の現状紹介」 - 野呂高樹（未来工学研究所）（科学技術イノベーション（STI）分野） - 齊藤貴浩（大阪大学）（大学評価分野） ・発表3：大澤望（インパクト・マネジメント・ラボ）「『社会的インパクト評価に関する現状整理』に対する会員フィードバック（アンケート結果の報告）」 ・グループディスカッション＆発表 ・まとめと今後について
11:40 ～
会場1・2・3・4
各委員会
13:00 ～ 15:00
会場1
共通論題6：発展的評価（DE）の実践報告から見えること：DEの守備範囲と効果 座長：中谷美南子（チームやまびこ） 司会 津崎たから（ウェスタン・ミシガン大学大学院博士課程） 討論者：津富宏（立教大学）／今田克司（ブルー・マーブル・ジャパン） ・発表者・討論者の紹介 ・報告1：千葉直紀（ブルー・マーブル・ジャパン）「効果的な発展的評価（DE）を実施するためのレディネスと評価プロセスの活用」 ・報告2：中村美亜（九州大学）「アウトカム・ハーベスティングを用いた評価アプローチの開発—文化団体における発展的評価の事例—」 ・報告3：中谷美南子（チームやまびこ）・清水潤子（武蔵野大学）「発展的評価（DE）の評価プロセスの活用とその効果検証モデルの提案」 ・指定討論者からのコメント ・セッション参加者からの質疑応答
会場2
共通論題7：国立研究開発法人の評価制度と研究評価：機関業績とイノベーション創出と科学的インパクト 座長・司会：白川展之（新潟大学） 討論者：三上真嗣（長野県立大学） ・趣旨説明 ・報告1：西山慶司（山口大学）「独立行政法人の評価制度：再考」 ・報告2：○柳瀬恵一・宮崎英治（JAXA）「宇宙航空研究開発機構（JAXA）における研究開発の運用と評価」 ・報告3：山本佳子・○植松郁哉（NEDO）「新エネルギー・産業技術総合開発機構（NEDO）の研究開発プロジェクトのマネジメントと評価」 ・発表4：白川展之（新潟大学）「政策評価・法人評価・研究評価：多層的・重層的評価制度の発散と共通のアウトカムとインパクト」 ・総合討論
会場3
自由論題4：インパクト評価 座長：齊藤貴浩（大阪大学） ・中野友絵・畑島宏之（THE WORLD BANK GROUP）「開発の新時代におけるインパクトの評価システム構築の課題：世界銀行グループ MIGAの取り組みから」

会場4
ラウンドテーブルディスカッション3：実務者と研究者の交流に基づく学術論文執筆ラウンドテーブル・その2：「投稿」から「査読」へ 座長・司会：米原あき（東洋大学） ・趣旨説明 ・話題提供1：窪田好男（京都府立大学）・新藤健太（日本社会事業大学）「『日本評価研究』の実践報告・調査報告に求められるもの」 ・話題提供2：村上裕一（北海道大学）「憎まず、腐らず、諦めず：査読というコミュニケーション・プロセスについて」 ・話題提供3：西野桂子（関西学院大学）「査読をお願いされたら？査読結果（あなたの意見）の伝え方」 ・フロア討論
15:10 ～ 17:10
会場1
ラウンドテーブルディスカッション4：評価能力向上のための国際的ネットワーク構築 座長・司会：石田洋子（広島大学） 討論者：佐藤由利子（日本学生支援機構）／新井和久（外務省） ・趣旨説明：石田洋子（広島大学）「評価能力向上のための国際ネットワーク構築」 ・話題提供1：津崎たから（ウェスタン・ミシガン大学大学院博士課程）「スリランカにおけるアカデミックシンポジウムに参加して」 ・話題提供2：西野桂子（関西学院大学）「国際的ネットワーク強化の動向」 ・話題提供3：石田洋子（広島大学）「APEA とJES のパートナーシップ」 ・討論者コメント ・フロア討論
会場3
共通論題8：評価の利用を促進する評価結果の報告方法 座長：窪田好男（京都府立大学） 司会：池田葉月（東京通信大学） 討論者：小島卓（NTT データ経営研究所） ・趣旨説明 ・報告1：池田葉月（東京通信大学）「評価結果の報告に関する先行研究の現状と課題」 ・報告2：窪田好男（京都府立大学）「自治体評価における評価結果の報告の方法」 ・討論者と各発表者のディスカッション ・質疑応答・意見交換 ・まとめ

日本評価学会第25回全国大会セッション報告

基調講演

「業績測定のインパクト—業績データはどう使われるのか—」

講師 ドナルド・モイニハン（ミシガン大学）

司会 田辺 智子（早稲田大学）

日本の公的機関では業績測定が普及しているが、得られた業績データがどのように利用されているか、また測定が業績向上につながっているかについては検証が進んでいない。一方、海外では近年、業績測定が政府のアウトカムを向上させるかについて活発な実証研究が展開されている。本基調講演では、当該分野の第一人者である米国ミシガン大学のドナルド・モイニハン教授をお迎えし、これまでの自身の研究成果を基にご講演いただいた。

まず、業績データの利用には4種類の類型があり、本来の目的である組織の業績改善に利用する「Purposeful」、法令に基づき業績を測定するが積極的に利用しない「Passive」、組織を擁護するためにデータを利用する「Political」、公共目的に反して業績を操作する「Perverse」があることが紹介された。また、業績データの利用に当たっては人々に認知の歪みがあり、統計データより逸話のほうが印象に残るためデータを逸話に組み込むと効果的であること、業績が高い場合より低い場合のほうがより関心を引きやすいこと、自らの信念に引き付けてデータが利用されやすいこと等が指摘された。さらに、米国の経験からは、政府業績成果法近代化法（GPRAMA）で採用された四半期ごとの業績レビューが学習フォーラムとして機能し、業績データ利用を促進したことが紹介された。

質疑応答ではフロアから活発な発言が寄せられた。業績マネジメントと業績測定の関係、ナラティブ情報の役割、業績マネジメントと民主主義の関係、業績測定とプログラム評価の従事者、政府業績成果法（GPRA）存続の理由、アウトカムとインパクトの違い、業績データの分析の必要性、目標の設定方法等について質問やコメントが寄せられ、講師との間で有意義な議論が行われた。

共通論題セッション

共通論題1「JICA事業評価におけるDXの推進について」

座長・司会 佐藤 真司（国際協力機構）

討論者 正木 朋也（国際協力機構／北里大学大学院）

本セッションは、JICAの事業評価における最新の取り組みについて紹介し、セッション参加者との議論から、今後のJICA事業評価の改善につなげることを目的に開催した。

報告1では、「誰一人取り残さない」や「人々のウェルビーイング」の視点を事後評価に導入したことや、外部要因の捉え方など紛争影響国の事業評価の難しさを紹介。加えて、OECD-DAC EvalNetにて2023年に初めてAIが取り上げられ、世界銀行、UNDPから、評価でのAI活用ははまだ文書の要約などに限定されていることが報告されたことを紹介。

報告2では、デリーメトロのEnterprise DXという、JICA事業DXの取組みを紹介。JICAの顧客である相手国側自身がデータを作成・収集・連携する能力を強化することで、JICAがそれらのデータへのアクセスが可能となる点を説明。そのことが、事業計画（より詳細なLogic Model・Theory of Changeの構築）、事業監理（LM・ToCの精緻化）、事業完了後フォローアップ（インパクトエンゲージメント）、という事業運営全体をデータ駆動に変容させる可能性に言及。

報告3では、衛星データの主な活用目的について、①インパクト指標の代替、②定量的効果指標の補完、③実施機関提供データの補完、④現地調査踏査地の選定の4つがあることを紹介。その上でこれまでに得られた課題として、①衛星データの特徴やメリットを最大化できていない点、②衛星データ活用を念頭においた指標が設定されていない点、③衛星データのスキルを持つ人材が不足している点を、今後の展望と共に説明。

フロアからは、DXや衛星データは主にインパクトを定量的に捉える議論と理解したが、現場のナラティブな情報なくして事業効果の真の原因がなにかは分からないのではないかとの問題提起あり。JICAからは、ストーリー性とそれらを記録に残す重要性は認識しており、事後評価でも定量的な分析だけでなく、受益者へのインタビューを可能な限り実施するなど、数値に現れない行動変容などを捉える努力をしていることを紹介。

共通論題2「学校評価マップの活用—学校評価学と実践の往還—」

座長・司会 小澤 伊久美（国際基督教大学）

討論者 石田 洋子（広島大学）

本共通論題では、まず、橋本昭彦会員（日本女子大学）が、学校評価分科会有志で作成中の学校評価マップ（以下、学評マップ）について、その構成の概況を紹介した。その上で、米国などの外国のマップから参考になる情報を整理して示し、我が国の学校評価の現状に適した独自のマップの構成方法とその活用を論じた。

次に、佐々木保孝会員（天理大学）は、地域と学校の協働活動を支える「地域コーディネーター」の協力を得て学評マップのブラッシュアップの可能性を検討し、マップによる整理が今後の学校評価の実践に

において持ちうる意味や有用性を指摘した。

続いて石田健一会員（計画と評価専門家）は、ノンフォーマル教育を通して社会運動に取り組む団体に外部評価者として関わっている経験に基づき、評価の実践において団体関係者らと重ねた対話の意味を問うた。具体的には当該評価実践の経験を共有した上で、評価と事業形成、評価と人的育成、評価を通じたケア、評価を通じた気づきの促し、評価を通じた代替案の協働提起などを考える重要性を提起した。

一方、小澤伊久美会員（国際基督教大学）は、学会運営改善の事例を報告し、組織評価の結果ではなく、評価に取り組むこと自体が評価の場作りにつながることで、運営改善を通して評価に好意的な内部者をつくるのが評価人材育成の第一歩になる可能性を論じた。

最後に石田洋子会員（広島大学）から、まず、個々の報告についてのコメントがあり、続けて、報告全体を踏まえて、学校評価士は、どのような立場から、どのような役割で、どのように評価に責任を負うのか、学校評価士にとって学評マップがガイドラインとなるのかという問いが投げかけられた。また、報告者らが指摘した対話やナラティブといったキーワードが学評マップにないと指摘し、これらを取り入れる改善案が指摘された。

共通論題3「政策形成・評価に関する改革の取組 ～総務省行政評価局における最近の事例～」

座長・司会者 尾原 淳之（総務省）

討論者 南島 和久（龍谷大学）

尾原淳之会員（総務省）から当局における政策評価の推進と行政運営改善調査の概要について説明した。その後、木村智宏会員（総務省）より「租税特別措置等の効果検証手法について」、高崎浩一会員（総務省）より「浄化槽行政に関する調査」、堤雅也会員（総務省）より「太陽光発電設備等の導入に関する調査」について、それぞれ報告した。

木村会員からは、EBPMの実践を後押しするための政策効果の把握・分析手法の実証的共同研究の一環として、租税特別措置の効果検証を①ロジックモデルの作成、②リサーチデザインの整理、③データの入手、④分析の実施の4つの手順で実施し、得られた示唆と今後に向けた方向性について報告した。

高崎会員からは、単独処理浄化槽のうち生活環境の保全等に重大な支障が生じるおそれのあるものが、都道府県等において「特定既存単独処理浄化槽」として的確に判定・措置されるよう、都道府県等における取組状況などを調査し、①判定の考え方の見直し・定量的基準の設定、②清掃業者や保守点検業者からの情報収集の仕組みを有効に機能させるための措置、③維持管理の向上に向けた浄化槽台帳の整備・活用方法の提示、デジタル化の検討等を環境省に求めた旨を報告した。

堤会員からは、太陽光発電設備等の適正な導入を推進するための改善策を検討するため、市区町村及び経済産業局の対応状況を調査し、令和5年8月に住民説明のポイントを経済産業省に情報提供し、令和6年3月に現地調査の強化や法令違反等の状態が改善されない場合に交付金の留保等の必要な措置を適確に実施すること等を経済産業省に求めた旨を報告した。

討論者の南島和久会員（龍谷大学）や会場からは、勧告した後の改善状況について等の質問や必要なデータの取得に必要な手続や期間についての意見交換等があった。

共通論題4「中国の対外経済協力の現状とDAC評価による国際協調の可能性」

座長・司会 河野 撰（外務省）

本セッションは、2024年霞山会上半期研究会「グローバルパワー移行期における中国の対外経済協力—インフラ投資を中心に」の研究成果を踏まえた中国の対外経済協力の現状についての報告と、DAC評価改訂プロセスや改訂後の当評価を通じた国際協調の可能性についての報告の2本で構成され、中国をはじめとした新興国台頭下でのDAC評価による国際協調の可能性について検討が行われた。

第1報告では徐一睿氏（専修大学）と佐藤一光氏（東京経済大学）が財政学の専門的知見から中国の対外協力の現状とその背景について報告を行った。徐氏は開発学で論じられる特に開発途上国でのインフラ分野における中国のプレゼンスの拡大は、資金的ボトルネックに直面している開発途上国が中国企業とPPPを通じた資金調達をせざるをえない状況も一因であることを報告した。佐藤氏は中国のプレゼンス拡大の背景として、世界の貿易・産業・エネルギー構造の転換から中国が「世界の工場」から「世界のインフラ投資国」へと変貌を遂げていることがPPPを通じた対外経済協力の拡大に繋がっていると分析をした。これを受けて、河野撰会員（外務省）は、中国をはじめとした新興国の台頭により開発途上国の現場では「開発協力・援助」とPPPのような「投資」が混在し、従来のDAC評価では実態を捉えにくくなっている点を報告した。かかる現状において、DAC評価の包摂的な改定プロセスや、新興国等による新DAC評価の適用は混沌とした現場において国際協調の機能がある点を報告した。

第2報告は村岡敬一会員（OECD/DAC開発評価ネットワーク前副議長）による「DAC評価基準改定と中国対外援助へのインプリケーション」であった。本報告ではSDGsへの適合を目的に改訂されたDAC評価基準が世界的に普及しつつある中、新評価基準が中国対外援助にもたらす影響と、基準の共有を通じた伝統的援助国（機関）との援助協調の可能性を考察した。

共通論題5「エビデンスと政治」

座長 杉谷 和哉（岩手県立大学）
司会 鏡 圭佑（日本大学）
討論者 松村 一志（成城大学）

エビデンスの活用によって政策の合理化を企てる、EBPM（Evidence Based Policy Making）の動きが世界的に活発化し、様々な議論が交わされている。本学会においても、特集号や学会のセッション等で様々な角度から議論が交わされてきた。しかしながら、政策過程が本来的に政治と無関係ではありえない以上、EBPMが活用を企図するエビデンスもまた、政治的なものから自由ではありえない。

本共通論題では、まず司会が全体の趣旨説明を行った。次に座長が、「エビデンスの政治性について」と題した報告で、EBPMにおける政治性をめぐる議論を概括し、多義的なエビデンスを踏まえないければ、かえって政策決定が非合理的になることを示唆した。続けて、清水右郷氏（京都大学）が「方法論的進歩と望ましい結果を政策で結ぶ」の報告において、医学・疫学における臨床試験の歴史とそこから得られるエビデンス解釈をめぐる議論を緻密に解説した。そのうえで、科学技術政策の方向性を再検討する重要性が提示された。清水習会員（宮崎公立大学）は、『科学的エビデンス』の何が政治的なのか？で、EBPMのプロジェクトが最終的に「エビデンスを通じた統治」に帰結すると喝破し、政策言説においてエビデンスが覇権性を有している現状を直視し、それを支えているメカニズムや構造に目を向けなければな

らないと指摘した。

討論者の松村一志氏（成城大学）からは、それぞれの報告に対してコメントが加えられ、「政治性」という言葉で何を指しているのか、エビデンスに潜むバイアスの内実、「社会」をどのように捉えるかといった幅広い論点が提起され、全体でこれを討議した。

共通論題6「発展的評価（DE）の実践報告から見えること： DEの守備範囲と効果」

座長 中谷 美南子（チームやまびこ）
司会 津崎 たから（ウェスタン・ミシガン大学大学院博士課程）
討論者 津富 宏（立教大学）
今田 克司（（株）ブルー・マーブル・ジャパン）

発展的評価（DE）は、従来型のプログラム評価では対応しきれない状況や評価対象を評価するために開発されたアプローチである。本セッションにおいては実践事例を題材に進められた分科会の議論を踏まえ、DEの守備範囲を再確認するとともに、期待される効果やその効果検証について報告された。

冒頭、司会の津崎たから会員（ウェスタン・ミシガン大学大学院博士課程）よりセッションの趣旨説明等がされた後、千葉直紀会員（（株）ブルー・マーブル・ジャパン）より、報告1「効果的な発展的評価（DE）を実施するためのレディネスと評価プロセスの活用」が行われ、DEの守備範囲といわれている5つの状況やDEの強みが発揮される条件である関係機関のレディネスについて発表された。

次に、中村美亜会員（九州大学）からDEの事例紹介として「アウトカム・ハーベスティングを用いた評価アプローチの開発：文化団体における発展的評価の事例」が報告された。本事例においてはDEプロセスにおいて評価者が担った役割や、その影響を受けた評価関係者の変化等についての具体的な報告があった。

最後に、中谷美南子会員（チームやまびこ）と清水潤子会員（武蔵野大学）から報告3「発展的評価（DE）の評価プロセスの活用とその効果検証モデルの提案」が発表され「評価プロセスの活用（Evaluation Process Use）」に焦点をおいたDEの効果検証の具体的方法論について、事例をもって提案された。

指定討論者の津富宏会員（立教大学）からは、DEの守備範囲やプリンシプルの適用等に関する詳細な質問が寄せられ、それに対しての回答が行われた。また今田克司会員（（株）ブルー・マーブル・ジャパン）からは、報告3で提示されたDEの効果検証を新しい試みとして期待を示しつつも、そのモデルが実際に効果検証の枠組みとして成立し得るかについて懸念が示された。中谷・清水会員からは、指摘を受けた点を踏まえ、今後事例の設計を調整しながら検証を進めていく旨の見解が表明された。

共通論題7「国立研究開発法人の評価制度と研究評価： 機関業績とイノベーション創出と科学的インパクト」

座長・司会 白川 展之（新潟大学）
討論者 三上 真嗣（長野県立大学）

科学技術評価分科会の共通論題は、科学技術イノベーション政策における評価制度の複雑な構造と実践的課題、そして今後の改善の可能性について、評価制度の運用実態を俯瞰して示す問題意識から実施され

た。国立研究開発法人や独立行政法人の勤務経験を持つ登壇者からの報告により、実務経験に基づく事例と論点が共有された。2020年から準備をはじめ科学技術評価分科会が設立されて4回目となる本共通論題は、省庁縦割りや政策評価や研究評価など評価の種別を超え、縦割りを越えた日本の評価学の発展を示す機会となった。

座長からの冒頭趣旨説明の後、西山慶司会員（山口大学）から「独立行政法人の評価制度：再考」と題して、制度の基本原則の変容、運用実態、総合科学技術・イノベーション会議との関係性について考察が示された。

2つの国立研究開発法人からの発表があり、柳瀬恵一会員・宮崎英治会員（共に宇宙航空研究開発機構）から「宇宙航空研究開発機構（JAXA）における研究開発の運用と評価」として、大規模プロジェクトと小規模研究開発の評価の実態と課題が提示された。加えて、山本佳子氏・植松郁哉氏（新エネルギー・産業技術総合開発機構）から「NEDOの研究開発プロジェクトのマネジメントと評価」と題して、研究評価制度の概要と、評価結果の活用による好影響事例が紹介された。

最後に、分科会長・座長を兼ねる白川展之会員（新潟大学）から「政策評価・法人評価・研究評価：多層的・重層的評価制度の発散と共通のアウトカムとインパクト」として、多層的で重層的評価制度のなかで、目指すアウトカムとインパクトが入り組む構造についての分析が示された。

これら発表を受け、指定討論者の三上真嗣会員（長野県立大学）が共通論題と本分科会の4年間の活動を総括した。通常の行政研究者などが複雑で扱いにくい科学技術と関連政策の評価について、独立行政法人等で実務経験を持つ実務者・研究者が一堂に会し評価の制度運用実態と課題を検討できたことは、何よりも貴重な機会であるとした。

共通論題8「評価の利用を促進する評価結果の報告方法」

座長 窪田 好男（京都府立大学）
司会 池田 葉月（東京通信大学）
討論者 小島 卓弥（（株）NTTデータ経営研究所）

本セッションは評価結果の報告方法の研究を活性化させることを目的に開催した。評価結果の報告方法の研究が進展すれば、評価結果の利用が促進され、公共政策や社会の改善が実現するという展望が背景にある。約50名の参加があり盛況であった。

インフルエンザ等の大流行もあり、報告予定であった青柳恵太郎会員（（株）メトリクスワークコンサルティング）と討論者の予定であった佐藤徹会員（高崎経済大学）は止むを得ず欠席した。

本セッションではまず司会の池田葉月会員（東京通信大学）と座長の窪田好男会員（京都府立大学）がそれぞれ趣旨説明を行った。続いて報告があり、まず池田葉月会員が「評価結果の報告に関する先行研究の現状と課題」と題する報告を行った。続いて窪田好男会員が「自治体評価における評価結果の報告の方法」と題する報告を行った。

池田報告は、近年の理論や実践では、評価結果の報告に続く想定利用者や社会とのコミュニケーションが重視されていることを指摘し、評価結果の効果的な報告のあり方もそれに対応した双方向性のあるものへ転換することが意図されていることを明らかにしようとした。

窪田報告は、自身が重要な役割を果たした三重県事務事業評価システムと京都市政策評価制度を事例に、自治体評価で従来どのように評価結果の報告が行われてきたか、自治体評価において評価結果の報告を制約する条件は何かを明らかにした。

討論では、評価結果の利用や方向のようなキーワードについての定義の重要性、分野別評価に横串を指

すような理論が成立する可能性をより丁寧に議論する必要性などが議論され、今後の全国大会でのセッションの継続的な開催や将来の分科会の設置についても言及があった。

ラウンドテーブルディスカッション1

「日本の政策評価のこれまでとこれから—山谷清志を囲んで—」

座長・司会 湯浅 孝康（大阪国際大学）

日本の政策評価は、中央府省では2001年1月6日の中央省庁等改革によって制度化された。そこから四半世紀近くが経過し、最近では評価制度の大幅な見直しも見られた。ひるがえって、評価制度の導入にあたって何が期待され、それは現時点でどの程度達成できているのか。これまで評価制度はどうとらえられ、どう運用されてきたのか。評価制度の導入によって何が変わり、何が変わらなかったのか。そして、評価制度は日本の政策にどのような帰結をもたらしたのか。これらの論点を含めた日本の政策評価のこれまでについて、本ラウンドテーブルディスカッション（RTD）では、中央府省での制度導入の際から研究会メンバーとして携わってきた山谷清志のオーラルヒストリーを実施し、これからの日本の政策評価について討論者を交えて議論した。

当日はまず司会者から企画の趣旨を説明した後、山谷清志会員（同志社大学）から『日本の政策評価』のこれまでとこれから」をテーマに、1時間程度の概説があった。おおむね時系列順に、「政策評価前史」「改革の時代」「省庁再編と政策評価」「『地方分権改革』と政策評価」「政策評価の現状分析と将来」の5つの話題が提示・説明され、残された課題として「改革と評価」「評価とアカウンタビリティ、レスポンスビリティ」「中央政府と地方政府」「官と民」「評価の評価」があげられた。これらを踏まえ、討論者の三上真嗣会員（長野県立大学）、鏡圭佑会員（日本大学）、山谷清秀会員（大阪経済大学）、南島和久会員（龍谷大学）から山谷清志会員への質問応答がなされた。フロアの参加者からの質疑応答の時間も取ることができ、1時間弱、活発な議論がなされた。

本RTDには60名前後に参加いただいた。本RTDを通じて日本の政策評価のこれまでについてさらなる理解が深まったこと、また日本の政策評価のこれからについての興味・関心が喚起されたことを期待する。

ラウンドテーブルディスカッション2

「社会的インパクト評価に関する現状整理と今後」

座長 今田 克司（（一財）CSOネットワーク／
（株）ブルー・マーブル・ジャパン）
司会 米原 あき（東洋大学）

本セッションは、社会的インパクト評価分科会による「社会的インパクト評価に関する現状整理（案）」の内容を検討し、会員相互の意見交換を通じて、分科会としての今後の取り組み検討への一助とする目的で企画されたラウンドテーブル・セッションである。同分科会では、2024年9月に分科会としての「現状整理（案）」をまとめ、これを公表するとともに、会員等からアンケート形式でのフィードバックを収集してきた。本セッションでは、グループディスカッションを行い、セッション参加者からのさらなるフィードバックを収集した。

冒頭、座長より、「現状整理（案）」の説明があり、続けて、分科会メンバーより分野ごとの解説があっ

た。津富宏会員（立教大学）は民間非営利活動の分野から見て、桑原亜希子会員（（株）日立製作所）は社会イノベーション事業の分野から見て、中村美亜（九州大学）は文化芸術の分野から見て、いかに社会的インパクト評価が理解・実践されているか、いかなる課題が存在しているかについての説明があった。なお、分野ごとの解説では、当初4件の報告が予定されていたが、1件は報告者が不参加となったため、結果として3件のみの報告となった。加えて、野呂高樹会員（（公財）未来工学研究所）より科学技術イノベーション（STI）分野における、齊藤貴浩会員（大阪大学）より大学評価における社会的インパクト評価の現状の簡単な紹介があった。また、大澤望会員（（一社）インパクト・マネジメント・ラボ）より、アンケート結果の報告があった。

これらを受けて、本セッションでは、1グループあたり5～6名のグループに分かれ、分科会メンバーがファシリテーター役として加わったグループディスカッションを行った。ここでは、プログラム評価との違いを明確にするとよい、「社会的インパクト」といっても立場によって見ているものが違うため様々な分野における発展を丁寧に整理するとよい、など多様な意見が寄せられた。

今回のセッションの結果も含め、分科会として「現状整理（案）」のまとめや今後の方向性について議論を行う予定である。

ラウンドテーブルディスカッション3

「実務者と研究者の交流に基づく学術論文執筆ラウンドテーブル

その2：『投稿』から『査読』へ」

座長・司会 米原 あき（東洋大学）

本ラウンドテーブルは、昨年に続き、実務者による論文投稿を促進・支援する目的で、編集委員会によって企画されたものである。今年も昨年に続き「実践報告・調査報告に求められるものとは何か」というテーマを設定し、そして今年は新たに「投稿者と査読者の建設的なコミュニケーションとはどのようなものか」というテーマを加え、投稿者・査読者の両方の立場から、活発な議論が交わされた。

フロアディスカッションに先立って3つの話題提供が行われた。

まず、窪田好男会員（京都府立大学）と新藤健太会員（日本社会事業大学）による共同報告「『日本評価研究』の実践報告・調査報告に求められるもの」においては、大会報告や会報記事とは異なる、学術誌である『日本評価研究』に掲載されるべき実践報告・調査報告とはいかなるものなのか、そこには何が求められているのかという問題提起がなされた。窪田会員、新藤会員自らの実践報告論文を事例としたアクティブな報告が展開された。

続いて村上裕一会員（北海道大学）による報告「憎まず、腐らず、諦めず：査読というコミュニケーション・プロセスについて」においては、『日本評価研究』における論文査読が、研究者と実務家の間で知識や経験を交換するコミュニケーションの場としてどのように機能しているか／機能し得るかについて、村上会員の経験を踏まえた議論が展開された。

最後に、西野桂子会員（関西学院大学）による「査読をお願いされたら？査読結果（あなたの意見）の伝え方」においては、西野会員の長きにわたる編集委員の知見に基づくメッセージが、投稿者・査読者・担当編集委員それぞれの（潜在的）立場の参加者に対して発信された。『日本評価研究』が、評価学と評価実践の架橋に貢献する学術誌として発展していくために、投稿者・査読者・担当編集委員というそれぞれのステイクホルダーが協働すべき方向性が示された。

ラウンドテーブルディスカッション4 「評価能力向上のための国際ネットワーク構築」

座長・司会 石田 洋子（広島大学）

本ラウンドテーブルの開始に当たって、座長・司会の石田洋子会員（広島大学）が日本評価学会（JES）による国際交流や国際協力の活動を振り返り、今後の国際的ネットワークのあり方についての議論を目的とするという趣旨説明を行った。

話題提供として、まず、津崎たから会員（ウェスタン・ミシガン大学大学院学際的評価学博士課程）が、「スリランカ国スリ・ジャヤワルダナプラ大学評価センターとアジア太平洋評価学会（APEA）共催の『モニタリング評価に係る学術的教育』に参加して」と題し、2024年12月2日～4日にコロンボで開催された同シンポジウムでのセッション発表報告を行った。

さらに、西野桂子会員（関西学院大学）が「国際的ネットワーク強化の動向」と題し、SDGsの達成を目指し、評価専門家の自発的組織の国際的ネットワークである国際評価協力機構（IOCE）やEvalPartner、Eval4Action等の評価推進のパートナーシップ、若手評価者支援のためのEvalYouthネットワーク等の活動状況を報告した。

最後に、石田洋子会員（広島大学）が「APEAの活動とJESとのパートナーシップ」と題してAPEAによる評価人材の育成、若手評価者を中心としたウェビナー実施などの様々な活動を紹介した。同会員は2024年1月からAPEA会長を務めており、2025年11月11日～14日に、第5回APEA大会が日本評価学会との共催で開催されることを発表した。

続いて、佐藤由利子会員（日本学生支援機構）、並びに新井和久氏（外務省）の討論者2名から、JESにおける国際的ネットワーク強化、国内外へ向けての日本の評価経験や成果に関する効果的発信の必要性などが指摘された。さらに、これらの課題について、約30名の参加者とともに議論が行われ、こうした課題解決のために第5回APEA大会の機会を有効活用することが確認された。

自由論題セッション

自由論題1「評価技法」

座長・司会 新藤 健太（日本社会事業大学）

本セッション最初の発表は、津富宏会員（立教大学）による「対人支援の成果をどう測るか：批判的検討」で、対人支援という「介入」のもつ権力性について吟味し、その「成果」を測定するための指標の在り方について報告された。特に、3つの支援機関を対象にしたインタビューの結果も引用して、介入によって被支援者の変化を促し、これを測定・評価することを「加工モデル」と表現し、そのあり方に疑問を投げかけた。

次の発表は、西村君平会員（東北大学）による「不確実性に適応的なシステマティックレビュー：RAMESESプロジェクトを中心に」で、教育や福祉のような不確実性の専門的対人サービス領域ではRCTを補完あるいは代替する研究デザイン／評価デザインが提案され、エビデンスの多様化・多義化が進んでいるものの、こうした知見を統合するためのシステマティックレビューの在り方に関する検討は十分ではないという問題意識のもと、RAMESES（Realist And Meta-narrative Evidence Syntheses: Evolving Standards）というシステマティックレビューの特徴や意義、限界、日本のEBPMへの示唆が報告された。

最後の発表は、陳凌雲会員（同志社大学大学院生）による「地域福祉実践における多様な主体が評価プロセスに参加する意義」で、中国江蘇省塩城市の「Green Living, Blooming Communityプログラム」を事例にプログラム計画策定段階の評価ワークショップの効果等について報告された。具体的には、個人レベルの変化として「評価能力の形成」、「評価に対する意識・行動の変化」があり、組織レベルの変化として「協働的なプログラム運営体制の構築」、「評価的思考の浸透」があったとのことだった。

なお、本セッションは、当初4つの報告が予定されていたが、1件は報告者がインフルエンザに罹患して発表できなかったため、結果として3件のみの報告となった。

いずれの報告に対しても参加者からは多くの質問が寄せられ、積極的な質疑応答が交わされたことで有意義な時間となった。

自由論題2「規制・社会実装と評価」

座長・司会 佐藤 由利子（日本学生支援機構）

村上裕一会員（北海道大学）による「EUにおける政策評価の進展：「より良い規制」に向けた事後評価の活用」では、EUの政策評価において、政策サイクルを意識し、政策立案や立法の教訓を得るための事後評価を重要視するようになっており、政策の経済性・効率性・有効性を検証する欧州会計検査院や、法政策の対社会パフォーマンスを評価する欧州議会との連携が図られていることを報告し、日本への示唆を示した。

渡邊直之会員・宜保友理子会員（共に慶應義塾大学、第一発表者を渡邊直之会員に変更）による「研究開発型スタートアップへのインパクト投資の可能性」では、社会課題の解決に資する先端技術（ディープテック）が注目される中、研究開発型スタートアップの投融資におけるインパクト・アセスメントについて、海外のインパクト投資ファンドや国内外の大学発のベンチャービルダーへのヒアリングに基づき、重

視されている視点や課題、得られた示唆を示した。

野呂高樹会員（政策研究大学院大学）による「欧州連合（EU）における研究&イノベーション（R&I）施策の評価に関する一考察：Horizon 2020の事後評価およびHorizon Europeの中間評価を事例に」では、Horizon 2020の事後評価やHorizon Europeの中間評価を概観し、「主要インパクト経路（Key Impact Pathways：KIPs）」への着目が特徴的であること、「まず評価する」という原則を適用し、ミッション志向やトランスフォーマティブなアプローチの評価を試行していることを紹介し、我が国への含意を示した。

齊藤貴浩会員（大阪大学）による「大学の社会的インパクトと大学評価の社会的インパクト：大学が自らの価値を社会と利害関係者に示す評価とするために」では、大学評価は教育研究等の水準の向上と社会への説明責任を目的として行われてきたものの、何をして何ができたかを形式的に示すだけになってしまっている現状を指摘し、大学の社会的インパクトを示す評価の在り方について提案を行った。

自由論題3「ODA評価」

座長・司会 石田 洋子（広島大学）

本セッションでは、ODA評価に関連する3つの報告があった。

第1報告では、牟田博光会員（国際開発センター／大妻女子大学）から「ミャンマー連邦共和国におけるコロナ禍とクーデターによる人的資源蓄積減少の評価」と題した報告が行われた。ミャンマーでは、コロナ禍及び軍部クーデターにより初等・中等教育就学率が下がった。同研究では、過去の就学状況の推移傾向からコロナ禍・クーデターがなかった場合のカウンター・ファクチュアルの児童生徒数推計値を求め、現実の在籍者数と比較して、今後の傾向値を推計した。2023年では幼稚園から9年生までの就学率は66%と計算され、政変がなかった場合に比較して21%ポイント低くなることが推計された。

第2報告は、倉田正充会員（上智大学）他による「災害時の緊急現金給付が貧困層のフード・セキュリティに与える影響－バングラデシュにおける自然実験分析－」と題する報告であった。日本のODA事業「金融包摂強化プロジェクト」による災害時の緊急現金給付の影響について、全顧客約15,000名から層化抽出法により最終的に無作為抽出した顧客2,900人を対象に調査分析を行った。同プロジェクトによって提供された金融現金給付には、フード・セキュリティに関するサイクロンの負の影響を緩和する効果があることが確認された。

第3報告では、高橋眞美会員（元早稲田大学大学院社会科学科博士後期課程）から「持続性の評価基準と日本の政府開発援助」と題した報告が行われた。SDGs制定に伴い、OECD-DACは評価6基準を定め、評価基準の定義も改訂した。同研究は持続性評価に注目する。他の援助機関は持続性評価の実施が限定的であるが、JICAでは事後評価において持続性にも継続的に注力している。現状では持続性の計測方法や学びの活用は明示的でない。持続可能な社会実現のために、持続性に関するさらなる分析が求められる。

自由論題4「インパクト評価」

座長・司会 齊藤 貴浩（大阪大学）

本セッションでは、インパクト評価に関する1つの報告が行われた。中野友絵会員・畑島宏之会員（共にTHE WORLD BANK GROUP）より、報告「開発の新時代におけるインパクトの評価システム構築の課題：世界銀行グループMIGAの取り組みから」があった。世界銀行グループがグループ全体の改革を進める中

で、新たなインパクト評価システムの一環として、多数国間投資保証機関（MIGA）のワンストップショップ型保証プラットフォームと、世界銀行グループの新しいコーポレート・スコアカードの導入が紹介された。世界銀行の新たな政策だけに他の国際開発金融機関への影響、さらには各プロジェクトのモニタリング・評価を含むマネジメントや、開発援助の在り方にも影響を与えうる取組と考えられ、今後の改革の進展を見守ることは当然として、その改革に関与する者にはより良い事業とすることが期待される。

なお、本セッションは、当初は社会的インパクト評価に係る3つの報告が予定されていたが、1件は発表取消、1件は報告者がインフルエンザに罹患して発表できず、結果として1件のみの報告となったことは残念であった。しかしながら多くの参加者があり、データベースとインパクト評価、AIの活用、実施体制に至るまで、活発な意見交換とともにインパクト評価の理解を深めることができたことを報告する。

～投稿案内～

日本評価学会では、「日本評価研究」掲載のための投稿原稿を募集しております。
投稿の締め切りは9月末日（翌年3月刊行）及び3月末日（9月刊行）です。
ご興味をお持ちの方は投稿規定・執筆要領・査読要領、並びに原稿見本をご参照のうえ、
奮ってご投稿ください。<http://evaluationjp.org/activity/journal>

投稿先：特定非営利活動法人日本評価学会 編集委員会事務局
E-mail: journal@evaluationjp.org

『日本評価研究』第25巻第1号
2025年 3 月31日

編集・発行 特定非営利活動法人日本評価学会 編集委員会
〒113-0034 東京都文京区湯島3-31-1
中川ビル5階
一般財団法人行政管理研究センター内

印 刷 株式会社 研恒社

© 日本評価学会

本誌に掲載されたすべての内容は、日本評価学会の許可なく転載・複写はできません。

Japanese Journal of Evaluation Studies

Vol. 25, No. 1, March 2025

CONTENTS

Preface

Evaluation of the Japan Evaluation Society

Takahiro Saito

Mini-Special Issue: Global Trend of Evidence-based Practice (EBP)

Editor's Note: Global Trend of Evidence-based Practice (EBP)

Ryo Sasaki, Satoko Tanabe

The Role of Qualitative Research
in Evidence-Based Policy Making (EBPM)

Satoko Tanabe

Why Evidence-Based Policymaking Falls Short in Japan: A Comparative Study
with the U.S. and Insights from the Evidence-Support System

Yohei Kobayashi

What happens to teachers and schools with 'EBP thinking'?

Toshiro Mori

Possibility for Progress of Impact Evaluation:
Introducing Desirable Social Values

Ryo Sasaki

Mini-Special Issue: The Development of Performance Measurement in Japan and a Reassessment of Its Significance: In Connection with The Work of Harry P. Hatry

Editor's Note: What We Can Learn from Hatry's Arguments

Hiraki P. Tanaka

Harry P. Hatry's Contribution to Performance Measurement:
A Review of 55 Years of His Work

Hiraki P. Tanaka

What Japan Can Learn from U.S. Performance Measurement Practices

- Review of Japan's performance measurement and evaluation efforts - Yasuhiro Sakon

Bridging the Gap Between the Practice of Evidence-Based Policy Making
and Mannerism in Performance Measurement

Tatsuya Ono

Article

Examining the Appropriateness of Using Year-Over-Year Accident Numbers to Evaluation
the Effectiveness of Product Safety Administration

Shunsuke Saito, Hironobu Kawamura

The Impact of a Volume-Based R&D Tax Credit on R&D Expenditures in Japan

Yutaka Kawase

Report

The Actual Situation and Challenges of EBPM in Local Governments

- Implications from a Study in Fukushima Prefecture -

Hodaka Kato, Ai Suzuki, Tomotsugu Takamori

Report of the 25th Annual Conference

Japan Evaluation Society