

令和2年度「EBPMをはじめとした
統計改革を推進するための調査研究」
(教育政策の特性を踏まえた根拠に基づく
政策形成のあり方についての研究業務)
報告書

令和3年3月



METRICS WORK CONSULTANTS INC.

本報告書は、文部科学省の教育政策推進事業委託費による委託事業として、株式会社メトリクスワークコンサルタンツが実施した令和2年度「EBPMをはじめとした統計改革を推進するための調査研究事業」の成果物です。

目次

I. 本調査研究の目的・方針.....	1
1. 背景・目的	1
2. 検討体制.....	2
II. 執務便覧（案）	3

Ⅰ. 本調査研究の目的・方針

1. 背景・目的

政府の統計改革推進の動きと相まって、EBPM（Evidence-Based Policy Making：根拠に基づく政策形成）に関する議論が政府全体で本格化している。こうした環境変化の中で、教育分野においても「客観的な根拠を重視した教育政策の推進」を志向する動きが始まっている。例えば、第3期教育振興基本計画（平成30年6月15日閣議決定）においては、「より効果的・効率的な教育政策の企画・立案等を行う観点や、国民への説明責任を果たす観点から、客観的な根拠を重視した行政運営に取り組んでいくことが重要である」との言明があり、EBPMの重要性が確認されている。

EBPMの重要性は論を俟たないが、教育分野におけるEBPMの推進には固有の課題や懸念が指摘されている。同基本計画では、「他の政策分野と比較して、成果が判明するまでに長い時間を要するものが多いこと、成果に対して家庭環境等他の要因が強く影響している場合が多く、政策と成果との因果関係の証明が難しいものが多い」といった記載がなされており、「数値化できるデータ・調査結果のみならず、数値化が難しい側面（幼児、児童、生徒及び学生等の課題、保護者・地域の意向、事例分析、過去の実績等）についても可能な限り情報を収集・分析し、あるべき教育政策を総合的に判断して取り組む必要がある」といった注記が示されている。

文部科学省において上記のような教育分野の特性も踏まえて、政策形成過程の改善に向けた取組を一層進めるため、EBPMに関する知見や意欲のある職員による「教育分野におけるEBPM推進チーム（以後、「教育EBPMチーム」と呼ぶ）が設置された。

本研究業務は、教育EBPMチームとともに教育分野の特性を踏まえながら、EBPMの考え方に沿った政策過程の実践を検討し、文部科学省におけるEBPM推進に役立つ執務便覧を作成するものである。

2. 検討体制

本調査研究は、公募を経て、株式会社メトリクスワークコンサルタンツが受託し、実施した。執務便覧の作成に向けて、「教育 EBPM チーム」とともに、文部科学省職員が感じる疑問や困難点を把握しつつ、EBPM の考え方に沿った政策過程（PDCA サイクル）のあるべき姿を開催する検討会を組織した。本業務内で実施した検討会第 2 回から 8 回までの概要を下表に示す¹。

	実施日	内容
第 2 回	2020 年 7 月 27 日	・ EBPM 推進にかかる省内の課題と本業務の方向性
第 3 回	2020 年 9 月 9 日	・ 学校給食・食育総合推進事業を素材とした文部科学省におけるモデル事業の試行的事例分析（前編）
第 4 回	2020 年 9 月 24 日	・ 学校給食・食育総合推進事業を素材とした文部科学省におけるモデル事業前編の論点議論
第 5 回	2020 年 10 月 16 日	・ 学校給食・食育総合推進事業を素材とした文部科学省におけるモデル事業の試行的事例分析（後編）
第 6 回	2020 年 10 月 29 日	・ 指標設定 ・ 学校給食・食育総合推進事業を素材とした文部科学省におけるモデル事業後編の論点議論
第 7 回	2020 年 11 月 20 日	・ 学校給食・食育総合推進事業を素材とした文部科学省におけるモデル事業（RCT 以外のリサーチデザイン、指標の設定） ・ 教育の特性についての検討方針
第 8 回	2021 年 1 月 28 日	・ 教育分野の指標設定・開発にかかる論点 ・ アウトカムの多元化に係る論点

一連の検討会においては、「学校給食・食育総合推進事業を素材とした文部科学省におけるモデル事業」を仮想施策として設定し、政策立案から効果検証に至るまでの各段階において職員が検討・実施すべき事項をなぞりながら執務便覧作成のための論点を抽出していった。また、そこで明らかとなった疑問点に対して下表に示す有識者を招いた意見交換会（第 8 回）を開催した。

有識者	所属及び肩書
岡崎 善弘	岡山大学大学院 教育学研究科 助教
原 祐一	岡山大学大学院 教育学研究科 講師

¹ 第 1 回検討会は本業務開始前の教育 EBPM チームのキックオフの位置付け。

II. 執務便覧（案）

本研究業務にて、教育 EBPM チームとともに教育分野の特性を踏まえながら、EBPM の考え方に沿った政策過程の実践を検討し、作成した執務便覧案を次頁から示す。

執務便覽

(案)

目次

1. 本マニュアルの位置付け	1
2. EBPMの中心メッセージ	4
3. 知っておいて欲しいTips集	9
(1) アウトカムの設定方法	10
(2) 既存エビデンスの探し方	14
(3) 実証デザインの検討方法	21
(4) 既存アウトカム指標の参照方法	28
(5) 新規心理尺度の開発方法	32
(6) 実施規模の決定方法（サンプルサイズ）	35

本マニュアルの 位置付け

- 2017年8月の「EBPM推進委員会」発足を皮切りに、証拠に基づく政策立案（Evidence-Based Policy Making: EBPM）に関する議論が政府全体で本格化しました。EBPMを推進するための各種取組が進展する中、これまでに各所でEBPMの解説書やマニュアルが整備されてきました。また、EBPMの考え方を伝える研修等も頻繁に実施されています。このような取組を通じて、全省でEBPMの理解が着実に浸透してきています。
- しかしながら、EBPMの考え方に沿って具体的な政策・施策に取り組もうとすると、政策マネジメントの各工程において、具体的にどのような作業をどのようにこなしていけばよいのか戸惑いを感じる場面も少なくないでしょう。
- さらに、EBPMの一般論ではなく、教育領域におけるEBPMということを見ると、他の政策領域には見られない特有の難題に直面します。例えば、第3期教育振興基本計画においては、以下のように記されています。

教育政策は、幼児、児童、生徒及び学生の成長や可能性の伸長等を目指して行われるものであり、一人一人の様々な教育ニーズを踏まえて、教育活動が行われる。このため、成果は多様であり、その評価は多角的な分析に基づくべきものであることに留意する必要がある。

また、他の政策分野と比較して、成果が判明するまでに長い時間を要するものが多いこと、成果に対して家庭環境など他の要因が強く影響している場合が多く、政策と成果との因果関係の証明が難しいものが多いことなどの特性があることにも留意し、研究者や大学、研究機関など、多様な主体と連携・協力しながら、数値化できるデータ・調査結果のみならず、数値化が難しい側面（幼児、児童、生徒及び学生等の課題、保護者・地域の意向、事例分析、過去の実績等）についても可能な限り情報を収集・分析し、あるべき教育政策を総合的に判断して取り組むことが求められる。

成果の多様性という特徴によって、個々の教育ニーズを包含する単一のアウトカム指標の設定が難しくなることは想像に難くありません。また、アウトカム発現までに要する時間の長さや政策実施以外にアウトカムに影響を与える要因の多さは、政策効果の把握を行う際に用いることのできる効果検証デザインの選択肢を狭めたり、結果の解釈が他領域以上に容易ではなくなるといった問題につながってきます。第3期教育振興基本計画で触れられたこと以外にも、様々な困難が想起できます。例えば、非認知能力の向上を目指す政策が執り行われた場合を考えてみましょう。非認知能力を測定する指標は様々なものが考案されているため、政策効果を把握するためにはどの指標が最適なのか判断が難しいと感じたことがある職員もいるのではないのでしょうか。こうした教育領域の固有事情を念頭に置いて細部にまで配慮されたマニュアルは十分に整備されているとは言い難い状況になっています。

- 本マニュアルは、**教育領域を担当する行政官には十分とはいえない既存資料の行間を埋めることを目的**として作成されました。EBPMの考え方に基いて政策マネジメントを行っていくという教育行政官向けに、企画立案段階で行うべき作業項目について紹介しています。特に教育領域の特殊性が色濃く出る項目については、踏み外してはいけないポイントについて詳述してあります。

本執務便覧の読み方

- 本執務便覧は、①EBPMの基本メッセージの確認と、②PDCA過程で行う具体的な作業の手助けとなるような解説の2つのパートから構成されています。
- 読者はまず①のEBPMの基本メッセージに目を通し、EBPM的思考法に基づいたPDCAサイクルの姿がどのようなものになっているのかを俯瞰することをお勧めします。特に要求年度に発生する立案過程（Planning）で行うべき作業事項を確認して下さい。
- 立案を進める中で難しさを感じる場面が出てきたら、2つ目のパートに当該場面が扱われていないか覗いて見て下さい。僅か6つの項目しか取り上げていませんが、文部科学省の職員の方々が腹落ち感なく、疑問を感じながら取り組まれているポイントを取り上げてあります。悩みを解決してくれるヒントが隠されているかもしれません。

EBPMの 中心メッセージ

ポイント

- EBPMとは政策課題を的確に把握した上で、「その政策課題に対する打ち手を検討する際に、有効性を示す情報（エビデンス）を参照する」という行政官が取るべき行動様式を意味します。**既に有効性が示されている打ち手を知る**ことによって、政策手段の妥当性を最大限高めることができるようになるでしょう。
- もしエビデンスに裏打ちされていない政策介入を採用する場合は、**政策実施前に効果検証の方法も併せて検討し**、政策を終えた暁には今後参照可能となるエビデンスを産出できるようにしましょう。それによって、自身の政策の見直しはもとより、同様の政策課題に直面した第三者も妥当な政策手段を選択することができるようになります。
- 政策効果を正確に把握するには、周到に準備された政策対象者・非対象者のデータが必要になります。当面、その実践は容易ではないでしょうから、まずはできることとして**政策実施前後のアウトカムデータを収集する**というところから始めることを意識してみましょう。

- **証拠に基づく政策立案（Evidence-Based Policy Making: EBPM）**とは政策マネジメントの考え方・実践方法と考えることができます。「EBPMという考え方を持って政策マネジメントを行っていくべきである」、或いは「政策マネジメントはEBPMが提唱する実践方法に沿って行うべきだ」という意見の背後には、EBPMではない政策マネジメントの考え方・実践方法があるということになります。従来型の政策マネジメントver.1.0にEvidence-Basedという考え方が導入されるとVer.2.0にアップデートされるというイメージを持つと、EBPMがもたらす付加価値を理解する際の助けになるでしょう。
- 一般に、**政策マネジメント**は「企画立案（Plan）」、「実施（Do）」、「評価（Check）」及び「次の企画立案への反映（Action）」からなる4要素（4段階）によって構成されると考えられています。最後に位置するActionは次の意思決定（Plan）を通じて具体的な実行につながっていくため、政策マネジメントとは本質的にサイクルとして描かれたものであり、これによって政策パフォーマンスが継続的に向上していくことが企図されています。この**PDCAサイクル**と呼ばれる政策マネジメントの実践方法がVer1.0から2.0にアップデートされるとは、1巡目の企画立案段階で用いる情報に**エビデンス**（この介入手段は有効なのか？（What works?）という問いの答え）が明示的に加わることを意味します。
- 政策の論理構造を見える化したものにロジックモデルと呼ばれる表現方法があります。上記のような理解を持つと、EBPMとは、立案時に構築するロジックモデルをエビデンスによって裏付けられたものとするすることで、政策手段に関するアカウンタビリティを高めるための考え方だという捉え方ができるでしょう。

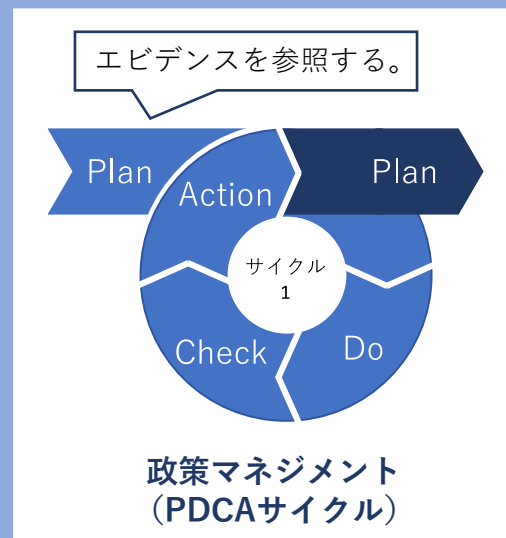
政策マネジメントのバージョンアップ

EBPMの考え方が導入される以前の企画立案の方法（政策マネジメントver.1.0）

- 教育政策課題を特定し、その課題の発生原因（問題構造）を特定するために、**統計データ等を活用してファクトベースの検討**を行う。
- 問題構造を踏まえ、最終アウトカムを引き出すためにはどのような途中アウトカムが発生する必要があるかを**論理的に検討**し、初期アウトカムを引き出すための活動を考案する。
- ファクトと論理的思考により、**仮説としてのロジックモデル**が描かれる。

EBPMの考え方を踏まえた企画立案の方法（政策マネジメントver.2.0）

- 問題構造の把握後、**エビデンスを参照**することで、その課題に対して効果的であることが立証されている打ち手がないかを吟味する。
- もしくは、打ち手を考案し、その打ち手の有効性を裏付けるエビデンスがないかを確認する。
- その結果、**エビデンスに裏打ちされたロジックモデル**が描かれる。



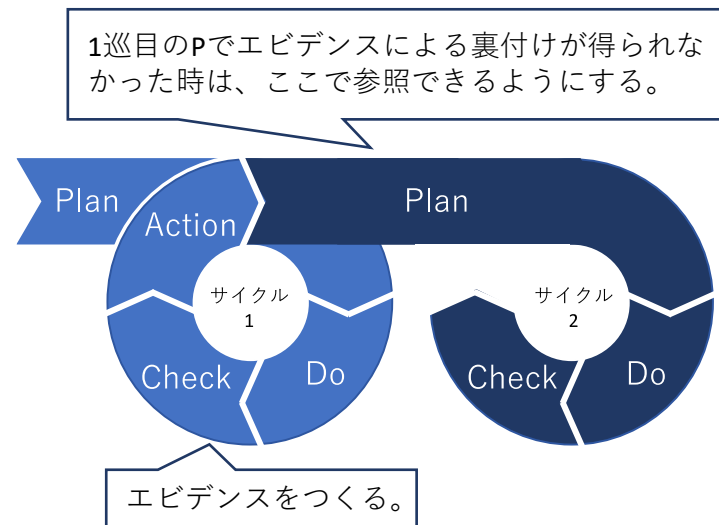
■ EBPMとはエビデンスによって裏付けられたロジックモデルを構築することであると述べました。しかし、**有用なエビデンスが見つからない**ということがあるかもしれません。政策課題は多様であり、また時代とともに状況は大きく変化していくことから、むしろ有用なエビデンスが見つかることの方が稀かもしれません。

■ 有用なエビデンスがないと、描かれるロジックモデルは論理的に十分検討された仮説に過ぎなくなります。これはエビデンスによって裏付けられたロジックモデルではありません。この場合は、仮説的ロジックモデルに基づくPDCAサイクルが一巡した際に、その有効性を判断できる情報（自前のエビデンス）を手にすることができるようにしておくことが望ましいと言えるでしょう。そうすることで2巡目のPlanの段階で、拡大、継続、修正、廃止といったような最適な意思決定を行うことが可能となります。なお、こうしたエビデンスを率先してつくれば、今後他所で同じような政策課題が生じた時に、このエビデンスが参照されることにもなるでしょう。

■ 2巡目のPlanにエビデンスを利用できるようにするには、**1巡目のPlan時に効果検証（Check）を行うために必要となる「仕込み」を計画**し、その内容に沿って政策介入を行い、必要に応じて1巡目の実施過程を通じてデータを収集していくことが不可欠となります。

■ PDCAサイクルの1巡目のPである政策立案時、ないしは2巡目以降のPである政策見直し時において、政策の有効性を示す情報（What works?の答え）を参照するという意味は以上のとおりです。EBPMという効果検証の方法論や、そのためのデータ収集・整備に目が向きがちです。しかし、EBPMの本質は、これから行う政策手段の妥当性を可能な限り高めていこうという点にあるといえます。これまではロジックを詰めることで手段の妥当性を説明しようとしていました。それに対して、EBPMでは、「論より証拠」という考え方で、**その手段を取ったことで問題を解決することができたのか？という実証分析の結果を重視**しています。

■ もちろん問題解決のための手段はエビデンスだけによって決まるものではありません。実際には、政治、世論、その他様々な要素を総合的に加味して決定することになるでしょう。しかし、その検討要素の一つに、これまで然程意識されなかった「政策の有効性を検証した過去の実証分析の結果」というWhat worksを議論した情報（エビデンス）を加えることが重要です。



エビデンスがない場合の政策マネジメント

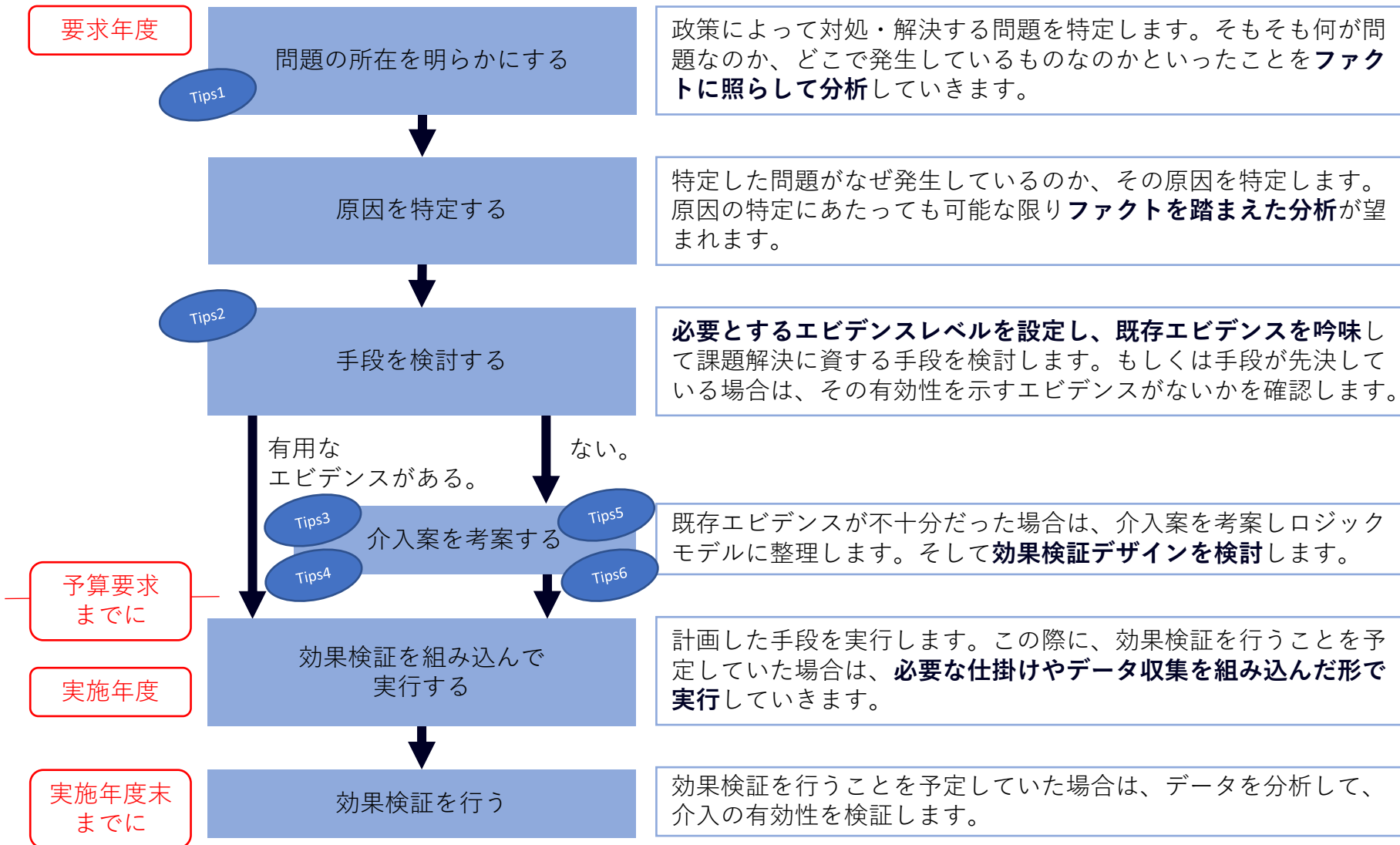
- EBPMが政策マネジメントにもたらした付加価値について、最後に1点補足をしておきます。エビデンスにはレベルがあるということが認識されています。**エビデンスのレベル**とは、現在述べられている政策効果の確証度合いを意味します。これから行う効果検証に関しては、得られる結果の確証度合いを意味すると理解して差し支えないでしょう。政策マネジメントにエビデンスという考え方を導入するにあたっては、エビデンスレベルを意識しながら上手にエビデンスと付き合っていくことが肝要です。

エビデンスレベル

高い	今後さらなる研究を実施しても、効果推定への確信性は変わりそうにない。
中	今後さらなる研究が実施された場合、効果推定への確信性に重要な影響を与える可能性があり、その推定が変わるかもしれない。
低い	今後さらなる研究が実施された場合、効果推定への確信性に重要な影響を与える可能性が非常に高く、その推定が変わる可能性がある。
非常に低い	効果推定が不確実である。

- あらゆる政策において高いレベルのエビデンスを求めることは現実的ではありません。**政策課題の性質に照らして必要となるエビデンスレベルを十分に吟味することが重要**になります。その上で、入手できたエビデンスが必要レベルを満たしているのであれば、それ以上のエビデンスを追求する必要はないといえるでしょう。つまり、1巡目の立案時に適切なレベルのエビデンスでロジックモデルの有効性を裏付けることができたのであれば、その後の厳密な効果検証の必要性は乏しく、政策実施に関するモニタリングを適切に行なっておくことで十分であるといえます。
- 政策課題の状況によってはエビデンスを必要とせず、論理的な検討がなされていれば十分と判断することもあります。
- 既存エビデンスのレベルが不十分な場合は、PDCAサイクルの1巡目を通じて自前のエビデンスをつくっていくこととなりますが、その際も常に過度に高いエビデンスレベルを求める必要はありません。「得られた検証結果が間違っているかもしれない」というリスクをどこまで受け入れるか吟味し、それに応じた適切な効果検証デザインを組み込んでおくことが重要となります。
- そうはいつても、高いエビデンスレベルの需要に応えることは容易ではないでしょう。**当面は、エビデンスレベルは低いものとなりますが、政策介入の事前事後におけるアウトカム比較を行っていくことも一案**です。ただし事前事後比較の結果解釈には注意が必要です（事前事後比較についてはp27の解説を参照）。

- これまで制作過程を「企画立案（Plan）」、「実施（Do）」、「評価（Check）」及び「次の企画立案への反映（Action）」という4段階で捉えてきましたが、次のように企画立案の段階を細分化することで、EBPMの考え方を取り入れた場合に具体的にいつ何を検討する必要があるのかをより明確に把握することができるでしょう。



知っておいて欲しい Tips集

ポイント

- EBPMの思考を取り入れた政策マネジメントを行う際に、教育領域の特性を踏まえると以下の点は特に慎重に作業を進めていく必要があります。
 1. アウトカムの設定方法（合意方法）
 2. 既存エビデンスの探し方
 3. 実証デザインの選択方法
 4. 既存アウトカム指標の参照方法
 5. 新規心理指標の開発方法
 6. 実施規模（サンプルサイズ）の決定方法

アウトカムの設定方法

アウトカム設定の重要性

- 政策立案を始める際のきっかけとなる情報は、曖昧な表現で行政官に提供されることもあるでしょう。示された一つの表現が多くの意味を持っていることも少なくありません。その場合、「問題」の受け手である行政官が、その表現が暗に焦点を当てている主たる領域について共通認識を持つ必要があります。一旦領域が正確に定義されれば、問題分析がなされ、アウトカムも自ずと決まっていきます。ただし、アウトカムの設定は重要である一方、合意形成を要する価値判断の問題になるため、確立された決定方法はありません。
- 問題領域が曖昧になっていると、何を問題と捉えればよいのか判然とせず、問題分析の方向性も見えず、アウトカムも一意に定まらないことを意味します。そのような状況でエビデンスを意識した政策過程を踏んでいこうとすると、様々な不具合が生じてしまいます。
- 特にエビデンスをつくるという過程で有効性が認められなかった場合、この政策のアウトカムは本来違うものだったといった議論が起きがちです。第三者が、本来主眼としていないアウトカムを取り上げて政策効果を論じてしまうということもあるでしょう。政策のアウトカムを明確に表明しておけば、このような事態は回避することができます。

① 問題の洗い出し

政策課題として行政官に伝えられたアジェンダが、具体的にどの領域に焦点を当てている／当てるべきものなのか、考える領域を洗い出します。

例：「食生活の乱れ」の多義性

「子供たちの食生活が乱れている」といった問題提議がなされたとします。しかし、これだけでは「食生活の乱れ」とはどのような領域を問題視しているのか判然としません。人によって食生活の乱れから連想するものは右記のように異なるかもしれません。毎日、朝昼晩決まった時間に食事をしていても摂取している栄養に偏りがあることを問題視する人もいるでしょう。この場合は、「食生活の乱れ」は回数や時間ではなく、栄養の偏りを意味します。他方で、家族揃って食事が取れていない家庭が増えていることを問題だと捉えている人にとっては、「食生活の乱れ」は孤食を指していることになります。

- 食事の回数
- 食事の量
- 食事の内容
- 食事の時間
- 食事の場所
- 食事を一緒に食べる人
- 食事中の姿勢、etc.

例：「心の豊かな子供」の構成要素

教育現場では「心の豊かな子供を育てる」といったことが目標にされることがあります。昨今の研究では「心の豊かさ」という概念を構成する要素を抽出する試みがなされており、一例として右記のような7要素が提案されています。このようにある概念が複数の構成要素から成り立っている場合も、注意が必要です。「心の豊かさ」という表現で、全体概念のことを指し示している場合はよいですが、構成要素の一部のみを指している場合は分解して問題視している要素を理解しておく必要があります。共感・共鳴の向上を目指した政策に対して、自己理解・受容が生じていないといった批判は回避する必要があります。

- 自己向上・増強
- 自己理解・受容
- 安穩的感性の獲得
- 高揚的感性の獲得
- 自己表出・評価獲得
- 協働
- 共感・共鳴

【補足・解説】

- 政策による対応が必要となる「問題」とは、理想と現実の間にギャップが生じている状態のことをいいます。アウトカムはそのギャップが改善された／解消された状態を描いたものになります。ここでは、例に示したような「食の乱れ」が意味するものを漏れなくリストアップするということではなく、「食の乱れ」というテーマの中で、問題が発生していそうな領域に限定して検討候補を羅列することで十分です。最終目的は、政策が目指すアウトカムを定めることです。
- 当該領域においてどのような**問題**が生じているのかあたりをつけるためには、有識者へのヒアリングや、ブレインストーミングが有用となるでしょう。

② 問題の特定

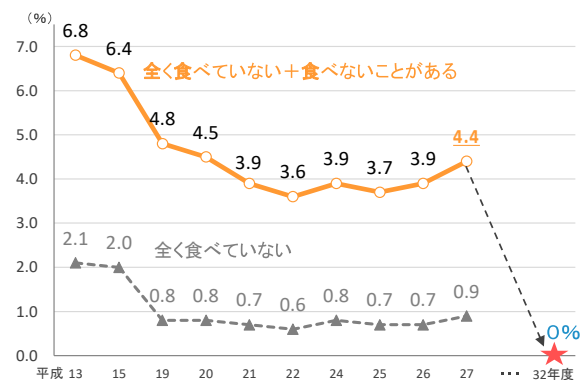
検討すべき問題候補があがったら、データ・ファクトを確認することで、理想的状態に対してギャップが生じているのかを詳細に検証します。この過程で、実際には問題は生じていないと分かるものもあるでしょう。

例：朝食欠食

食生活の乱れのうち、子供たちの朝食欠食を問題として取り上げるべきではないかという検討が始まったとします。こうした問題は全国規模で誰にでも発生している現象とは限りません。朝食欠食はどこで生じているのかをファクトをつぶさに見ていくことで把握する必要があります。また、その実態が理想とする水準から乖離していることを確認することで、現状が問題であると認識されます。

例えば左図のように現状の全体像を把握するだけでは、問題の認識としては不完全です。性別や年齢、地域といった観点から、対処しなくてはならない問題（ギャップ）が生じている場所を特定することが不可欠です。

朝食を欠食する子供の割合



資料：文部科学省「全国学力・学習状況調査」(平成19年度～平成27年度)
国立教育政策研究所教育課程研究センター「教育課程実施状況調査」
(平成13年度、平成15年度)

【補足・解説】

- 問題解決術を扱う書物では、問題の発生場所を特定することの重要性が説かれているものが少なくありません。上記の食生活の乱れの例で言えば、「子供たち」といっても、ある特徴や傾向を持った特定の家庭・地域の子供たちの中で問題が深刻化していることもあるでしょう。朝食欠食といったように問題領域を定めるだけでなく、対象を「漏れなく・ダブリなく」（MECE: Mutually Exclusive, Collectively Exhaustive）分け、「朝食欠食であればどこで（誰に）発生している現象なのか」を明らかにしていくことが大切です。
- この過程である程度アウトカム指標が意識されるはずですが、「朝食欠食が問題化しているか？」をデータで問うということは、朝食欠食という現象を何らかの形で測定し、その現状値と理想値を比較することが必要となるためです。しかし、この段階では、明確なアウトカム指標に踏み込むことなく、様々な観点からデータを観察することに徹しましょう。
- 理想的なデータが常に利用可能なわけではないので、データから読み取れることを適切に理解してデータに向き合しましょう。そもそものデータに偏りがあったり（女の子の状況しか分からない等）、本来の定義とずれていたり（データから分かる朝食欠食とみなす状態が、知りたい情報と異なっている等）する場合は注意が必要です。

③ アウトカムの決定

確認できた問題群の中から政策が解決を目指す領域を絞り込みます。そしてその問題が改善された／解消された状態を本政策のアウトカムとして決定します。

【補足・解説】

- 現実と理想のギャップは客観的なデータ・ファクトから特定されますが、たとえギャップがあったとしても政策介入が必要となるような深刻な問題なのかどうかは価値判断に委ねられます。さらに、どの問題を政策の対象として取り上げるかも判断に関わるものです。したがって、政策の目標を一意に定め、合意を得ることは容易な作業ではありません。
この決定プロセスに機械的な手順は存在しません。
- アウトカム（政策が取り組む問題）は必ずしも一つに絞る必要はありません。ただし、同時に複数の問題に取り組む時は、この後の介入内容を決定する際に取組とアウトカムの対応関係をロジカルに整理しておくことが不可欠です。ある一つの取組が複数のアウトカムを同時達成するのか、それぞれのアウトカムには別々の取組がなされるのかは意味が大きく異なります。この整理が不完全だとエビデンスの活用や、政策効果の判断に支障が生じてしまいます。

用語の確認：アウトカム・アウトカム指標

アウトカムとアウトカム指標は異なります。アウトカムを考える上で指標が自ずと意識されたり、その表現に含まれることもあります。必ずしも1対1対応するものではありません。アウトカムの設定は明確であるにも関わらず、指標がアウトカムを的確に反映するものではないといったことは起こりえます。アウトカム指標の設定方法についてはTips4も参照して下さい。

アウトカム

政策介入の結果もたらされる何らかのよい**変化のこと**。

例

- 食事の回数が改善する
- 食事の時間が適切になる
- 心の豊かさが増す
- 自己効力感が向上する

アウトカム指標

アウトカムが生じたか否かを把握するための指標のこと。

例（食事の回数）

- 過去1ヶ月の1日の食事の平均回数
- 過去1週間の1日の食事の平均回数
- 過去1ヶ月に朝昼晩の三食を毎回摂取しなかった日の割合

既存エビデンスの探し方

👉 既存エビデンスの重要性

- 焦点を当てている社会的課題の中には、**既に他の地域や組織で類似の課題が発生していて様々な解決策が試みられているもの**も少なくありません。その中からうまく機能した事例を見出すことができれば、より効率的に眼前の問題を解決することができます。
- 多くの情報があふれる中で、目についた既存エビデンスだけを参照していると真に有用な解決策を見落としてしまうリスクがあります。**既存エビデンスは可能な限り網羅的に、しかし効率的に収集していく必要**があります。既存エビデンスを集約したデータベースなどが一定程度整備されているので、そのようなインフラを駆使することが望まれます。
- 場合によっては、政策課題と解決策が同時に提議されていることもあります。その場合は、検討している解決策の有効性を既存エビデンスで裏付けることができるかという観点から、検索をしていきます。
- したがって検索には、問題を出発点として解決策を検索する場合と、解決策を出発点として検索していく場合の2パターンがあることになります。いずれにしても、政策立案時に既存エビデンスを適切に検索することで、政策手段の妥当性を高めていくことができます。

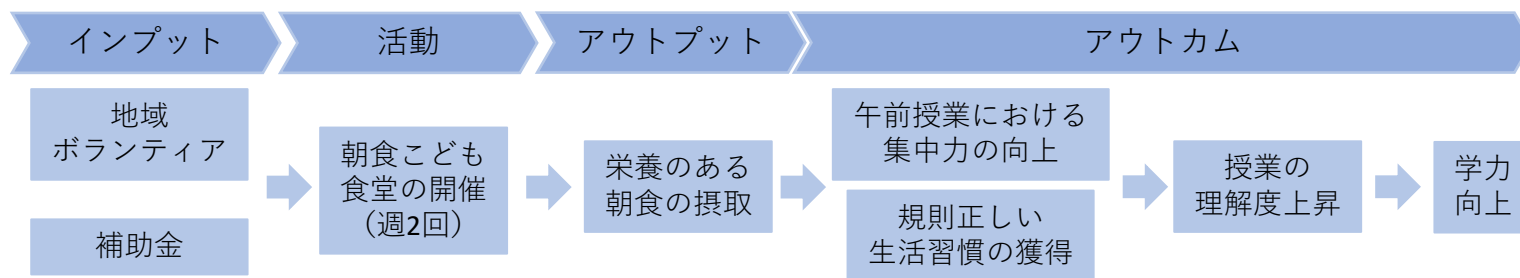
① 収集すべき情報の整理

集めるべきエビデンスを見定めるために、必要としている情報をPICOという以下の4つの要素に整理します。

P 対象	誰に対して	対象者の属性、抱える問題、地理的要件など
I 取組	何をすると	検討している取組
C 比較	何と比べて	他の取組の選択肢（何もしないことも含む）
O アウトカム	何が変化するか	期待される変化

例：朝食欠食（架空事例）

- 全国学力・学習状況調査（小学校6年生）の結果が毎年低迷していることを問題視したとしましょう。学力分布を詳細に検討したところ、就学援助の対象となっている児童が顕著に低い学力を示していました。
- 就学援助対象児童の家庭環境のデータを見たところ、多くの家庭で朝食が用意されていないことが分かりました。午前中の授業態度を注意深く観察すると、そのような家庭の子供たちは十分に集中できていない様子です。
- そこで小学校の敷地を地域に開放し、地域のボランティア団体による小学生を対象とした「朝食こども食堂」への補助金事業を計画しました。この事業のロジックモデルは下記のようなものになります。




- この政策を裏付ける既存エビデンスの検索を行います。そのために必要となるPICOを整理すると、右のようになります。

P	就学援助対象の小学生に対して
I	小学校区の地域ボランティアによって朝食こども食堂を週2回実施すると
C	何もしない場合と比べ
O	学力試験のスコアは向上するか？

【補足・解説】


- 一般に既存エビデンスはデータベース等から入手しますが、明確な指針なくエビデンスを闇雲に探してしまつては、必要とするものに中々辿り着けません。そこで、介入効果の有無という問いを、**対象 (Population)、取組 (Intervention)、比較 (Comparison)、アウトカム (Outcome)の4つの要素 (PICO)**に分解して整理しておくことで検索を効率的に行うことができます。具体的にはPICOという形で整理した情報がデータベースを用いてエビデンスをキーワード検索する際のヒントになります。また、本マニュアルでは触れていませんが、検索で得られたエビデンスが果たして自分が必要としているものか否かを判断するチェック項目としても活用できます。
- PICOへの整理が曖昧だと、無関係な情報を収集してしまうことにつながります。一方で、問題を細かく整理しすぎると、関連のある検索結果を十分に得ることができない可能性があります。また、直接的には関連しなくても参考になる情報を見落としてしまうリスクもあります。PICOを整理する際には、以下の点を意識しながら作業を進めるとよいでしょう。

 各要素に示された事項は問題の所在・詳細、取組内容を的確に表現しているか？
明確性

- ✓ 対象者が絞れているか？
- ✓ アウトカムは課題と対応しているか？
- ✓ 介入内容は曖昧ではないか？

学力向上施策の例

- | | | |
|---|---------------|---------------------|
| P | 中学1年生に対し、 | ⇒ 学力に不安を抱える中学1年生に対し |
| I | 補習授業の機会を提供すると | ⇒ 習熟度別補習授業の機会を提供すると |
| C | 何もしない場合と比べ | |
| O | 学力は向上するか | |
- 中学1年生一般を対象としているエビデンスなのか？ (P)
 - どのような補習授業を想定しているのか？ (I)

 情報は過度に具体化されておらず、適度な加減となっているか？
適当性

- ✓ 取組の対象が極端に限定されていないか？
- ✓ 特殊すぎるアウトカムに注目していないか？
- ✓ 取組の内容は適度に抽象化できているか？

食育施策の例

- | | | |
|---|------------------------------|-------------------|
| P | 朝食欠食が週3回以上ある中3女子に対し | ⇒ 朝食欠食がある中学生女子に対し |
| I | 月一の栄養バランスの整った朝食メニューに関する授業提供は | |
| C | 月一の保護者に対する朝食の重要性を訴える手紙配布に比べ | |
| O | 朝食欠食の回数が減少するか | |
- 上記のPICOは明確だが余りに限定されすぎていて、このような問いに答えたエビデンスは存在しない可能性が高い。
 - 例えば朝食欠食の定義と対象学年を緩めることも検討する。

- PICOを整理してみると、政策の立案状況に応じて埋めることができない要素がでてくることに気づくかもしれません。アウトカムを定め、それに対して有効な打ち手にはどのようなものがあるのかという関心からエビデンスを探している場合（取組からのアプローチ）は、当然取組（I）は空白となります。他方で、既に何らかの施策のアイデアがある場合や、既存施策の見直しを含む政策立案が行われている場合（課題からのアプローチ）は、取組（I）は特定されることとなります。しかしながらアウトカム（O）が空白になることはありません。アウトカムが未定だったり、一意に明確に定まっていない場合は、エビデンスを収集する前に、アウトカムの設定に戻る必要があります。

課題からのアプローチ

- 「対処すべき問題の解決に対して、どのような取組が有効と考えられるか？」と問う。
- PICOのうち、I（取組）は埋まらない。
- 問題から取組の検討を行うため、問題解決のアプローチとしては一般的な形となる。特に新規の政策を立案する際に適したアプローチである。
- ロジックモデル作成の工程においては、問題やアウトカムを定めた後、もしくは原因分析を行い打ち手の検討を始める段階でエビデンスを参照することになる。

取組からのアプローチ

- 「対処すべき問題の解決に対して、検討している取組は有効と考えられるか？」と問う。
- PICOの4要素が定まる。
- 取組から考えることは問題解決のアプローチとしてはあまり望ましくはない。しかし既に何らかの政策アイデアがある場合や既存政策の見直しの際にはこうしたアプローチが取られる。
- ロジックモデルの構築プロセスにおいては、論理的な思考に基づいてロジックモデルを組み上げた段階でエビデンスを参照することになる。

② 検索の実行

PICOに整理した情報を基に、既存エビデンスが収録されたエビデンスポータル、エビデンスデータベース、論文検索サイトなどでフィルター機能やキーワード検索を駆使して既存エビデンスを探します。

【補足・解説】

- エビデンスは学術論文や報告書といった形で発信されています。しかし、個々の報告書や論文を探し出し、それを読み解いていくことは、多忙な行政官にとっては簡単なことではありません。そこでエビデンスを収集する際には、以下のようなサービスを活用するとよいでしょう。分野によって充実度は異なりますが、エビデンスへのアクセスが容易になります。ただし、日本国内の取組を対象とした質の高い効果検証は少なく、エビデンスの蓄積が十分ではありません。そのため、国外の情報ソースから英語で書かれたエビデンスを収集することが避けられません。また、海外のエビデンスは特に、日本国内への適用可能性を吟味する必要があります。

① エビデンスポータル

既存のエビデンスを整理し、概要や質、実施コストなどを一覧できる形で公表しているプラットフォーム。ただし、ポータルが作成されているトピックはまだまだ少なく、カバーされている取組も限られているというデメリットもある。

② エビデンスデータベース

介入効果の検証を扱った個別の報告書や学術論文をデータベースとして集約し、キーワード検索やフィルタリングサービスを提供しているもの。エビデンスポータルのように、内容や質などが直観的にわかるようには整理されていないため、個々のエビデンスの中身については個別に読み解いていく必要がある。ただし、こうしたデータベースは報告書や論文の質に関する掲載基準を設定しているため、そこに含まれる報告書や論文は一定程度の質が担保されていると考えることができる。

③ 論文検索サイト

その名の通り学術論文の検索に特化した検索サイト。記述研究や理論研究などあらゆる内容の学術論文を対象としているため、広範な論文検索ができる一方で、自ら効果検証型の研究を選り分ける必要がある。また、学術論文といえども、質の高いものとそうでないものが混在しているため、質の吟味も重要になる。

主要なエビデンスポータル

分野	名称	運営組織/機関
教育（5-16歳）	Teaching and Learning Toolkit	Education Endowment Foundation（英国）
早期教育	Early Years Toolkit	Education Endowment Foundation（英国）
教育	Find What Works	Institutes of Education Sciences（米国）
教育	Evidence For ESSA	Center for Research and Reform in Education（米国）
子供の教育・福祉	Early Intervention Foundation Guidebook	Early Intervention Foundation（英国）
犯罪防止	Crime Reduction Toolkit	What Works Centre for Crime Reduction（英国）
福祉	Intervention Tool	Centre for Homelessness Impact（英国）
子供の福祉	Evidence Store	What Works for Children's Social Care（英国）

主要なエビデンスデータベース

キャンベル共同計画（Campbell Collaboration）

ビジネス・経営、刑事司法、障害、教育、国際開発、社会福祉などの分野における系統的レビューに関するデータベース。

ジャミール貧困アクションラボ（J-PAL）

ランダム化比較試験によって得られたエビデンスに特化したデータベース。もともとは開発途上国における貧困削減に関するエビデンスの産出を行っていたが、近年は北米やヨーロッパなどの先進国における取組に関するエビデンスも増えている。

インパクト評価に関する国際イニシアティブ（3ie）

国際開発分野におけるエビデンスを取りまとめているデータベース。個別の取組に関するエビデンスがまとめられているデータベース（Impact Evaluation Repository）と系統的レビューがまとめられているデータベース（Systematic Review Repository）がそれぞれある。

例：エビデンスポータル：Teaching and Learning Toolkit

具体例として、教育分野のTeaching and Learning Toolkit（教えと学びのツールキット）を見てみます。Webサイトにアクセスすると、以下のようなリストが表示されます。

各行に一つの取組について、その取組の費用、エビデンスの質（エビデンスレベル）、効果の大きさ（通常の学習月数で表現）が要約されています。例えば、一番上のBlock scheduling（授業数を少なくし、1授業当たりの時間を増やす取組）は効果(④)が0なので、効果が認められていない取組であることがわかります。

他方、3つ目のEarly years intervention（早期教育）は効果(④)が5なので、通常の学習の5か月分という比較的大きい効果があったことがわかります。また、エビデンスの強さ(③)も4と信頼性の高い情報であることが示されています。一方で、費用(②)も大きいため、費用対効果では必ずしも最適ではないこともわかります。

The screenshot displays the 'Teaching and Learning Toolkit' interface. At the top, it reads 'An accessible summary of the international evidence on teaching 5-16 year-olds'. Below this, there are four main filter categories: ①取組 (Intervention), ②費用 (Cost), ③質 (Evidence Strength), and ④効果の大きさ (Impact). The 'Filter Toolkit' section on the left allows filtering by keywords and includes sliders for Cost (set to £), Evidence (set to +1), and Months Impact (set to +1). The main content area shows three intervention rows:

①取組	②費用	③質	④効果の大きさ
Block scheduling Very low or no impact for very low cost, based on limited evidence.	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	0
Digital technology Moderate impact for moderate cost, based on extensive evidence.	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+4
Early years interventions Moderate impact for very high cost, based on extensive evidence.	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5

実証デザインの 検討方法

実証デザインの重要性

- 効果検証の詳細が描かれたものを**実証デザイン**と呼びます。政策介入が一通り終わってから（PDCAのCheckの段階を向かえてから）、実証デザインのことを考え始めたのでは、必要十分な結論が得られない可能性があります。Checkの段階ではデータを用いた分析が始まるのであり、その時にどのような分析を行うのか、こういったデータが利用可能になっていけばいいのかといったことは介入が始まる前の**Planning**の段階で計画されていなくてはなりません。その計画次第で、介入のあり方やデータ収集の対象や規模、タイミングが規定されます。
- したがって、外部有識者やコンサルタント等に効果検証を委託する際にも、**Planningの段階から相談を始める**ことが重要です。
- **Planning**時に実証デザインを検討する際には、求めるエビデンスレベルを見据え、制約条件の中で最善を尽くすことが求められます。例えば、高いレベルのエビデンスを求める場合は、政策介入の対象とならない比較群（統制群）の設定が不可欠となります。また一定規模のサンプルサイズが必要となります。求めるエビデンスレベルと取得可能なエビデンスレベルの間に大きな差があることが判明した時は、この状況下で実証分析を行うことの価値を問うことも大事でしょう。

① 求めるエビデンスの質の決定

効果検証を行うことで、どのようなレベルのエビデンスを必要としているのかを決定しましょう。

【補足・解説】

- エビデンスレベルとは、**介入の有効性を示す情報の確証度**を意味します（p.7も参照のこと）。実証分析を行えば、須らく信用できる結果が得られるわけではありません。実証分析の結果は間違えているかもしれないのです。
 - エビデンスレベルは実証デザインによって規定される（手法に応じてランク付けされる）性質のものではありません。どの実証デザインであっても、そのデザインが要請する適用条件を完全に満たしていれば、得られる結果は妥当なものとなります。ただし、適用条件が満たされているかどうかは、分析者の信念に過ぎず客観的に立証できない性質のものであります。一方で、実証デザインによって適用条件が満たされていると期待できる度合いには濃淡があることが経験的に知られています。この点から、**実証デザインが確証度合いの目安を与えてくれる**とはいえるでしょう。一例として、『平成30年度内閣府本府E B P M取組方針』に示されている整理を記載します。
- | | | |
|--------------------|-------|--------------------------|
| ↑
↑
質が
高い | レベル1 | ランダム化比較実験 |
| | レベル2a | 差の差分分析、傾向スコアマッチング、操作変数法等 |
| | レベル2b | 重回帰分析、コーホート分析 |
| | レベル3 | 比較検証、記述的な研究調査 |
| | レベル4 | 専門家等の意見の参照 |
- 次の②で述べるように、**利用可能な実証デザインは政策実施の環境が生み出す様々な制約条件によって影響を受けます**。それに応じて取得できるエビデンスレベルにも限界がでてきます。
 - エビデンスレベルの観点からは、大は小を兼ねるといえるので、与えられた制約条件の中で最もレベルが高いエビデンスが得られる実証分析デザインを選択するというのとは一つのコストの観点からは、常にハイレベルのエビデンスが優越するとは言えません。場合によっては低いレベルのエビデンスで十分なため、あえて簡便なアプローチを採用するということもあるでしょう。実証デザインの詳細を検討する前に、どのようなレベルのエビデンスを必要としているのか、すなわち本効果検証の位置付けを明確にしておくことが有益となります。
 - ただし、**必要とするエビデンスレベルは「決め」の問題なので具体的な決定方法はありません**。一つの判断材料は、エビデンスを用いて行う将来的な意思決定の重要性の大きさです。例えば、介入の実施コストが巨額となる場合は、介入の有効性を高いレベルのエビデンスによって慎重に見極めてから実施することが好まれるかもしれません。或いは、扱っている問題が広く社会の関心を集めており、不適切な介入を行ってしまったときの反響が大きいという時も、厳密な効果検証を経てから本格実施に至るといった工程を踏むほうが望ましいかもしれません。

② 制約条件の確認

実証デザインに影響を与える各種事項について、自由度がどこまで残されているかを確認しましょう。以下は代表的な検討事項です。

- 介入対象の操作可能性
- 利用可能（収集予定）データ時点
- 時間的制約

【補足・解説】

- 実証デザインを政策立案の初期段階から検討していれば、①で確認した必要なレベルのエビデンスを作り出すためのデザインを考えることもできるでしょう。しかし現実には、実証デザインを検討し始めるタイミングや様々な意味でのリソースといった点で、デザインを検討するにあたって**制約条件となってしまう決定事項がある**ことが多いでしょう。そのため制約条件を十分に洗い出し、その下で定めたエビデンスレベルに限りなく近づけることができるデザインを特定していくことになります。

介入対象の操作可能性

効果検証の肝は適切な比較対象群を設定することです。政策の介入対象がまだ決まっていなければ、対象者決定プロセスに様々な工夫をこらすことによって、作為的に適切な比較対象を**作り出す**余地が残されます。RCTで用いられるランダム割付という介入対象者の決定方法は具体的な一つの例です。他方で、既に何らかの基準や理由によって対象者が決定してしまっている場合は、非対象者の中から効果検証に用いることができる適切な比較対象が存在するのかを**探し出す**ことしかできなくなります。なお、未介入者が存在しない場合は、原則事前事後の比較（分割時系列分析を含む）を行うことしかできません。

利用可能（収集予定）データ

実証デザインによって必要となるデータの時点数、構造等が異なります。例えば、予算やその他の何らかの事情で、介入実行前のデータが取れないことが前もって分かっている場合は、パネルデータ（複数の主体に対して2時点以上の観察記録を持つデータ）を前提とした差の差法のような効果検証デザインは選択肢から外れることになります。

時間的制約

効果検証の結果をいつの時点までに出すのかによって、設定するアウトカム指標やデータ収集のタイミングに影響が出てきます。

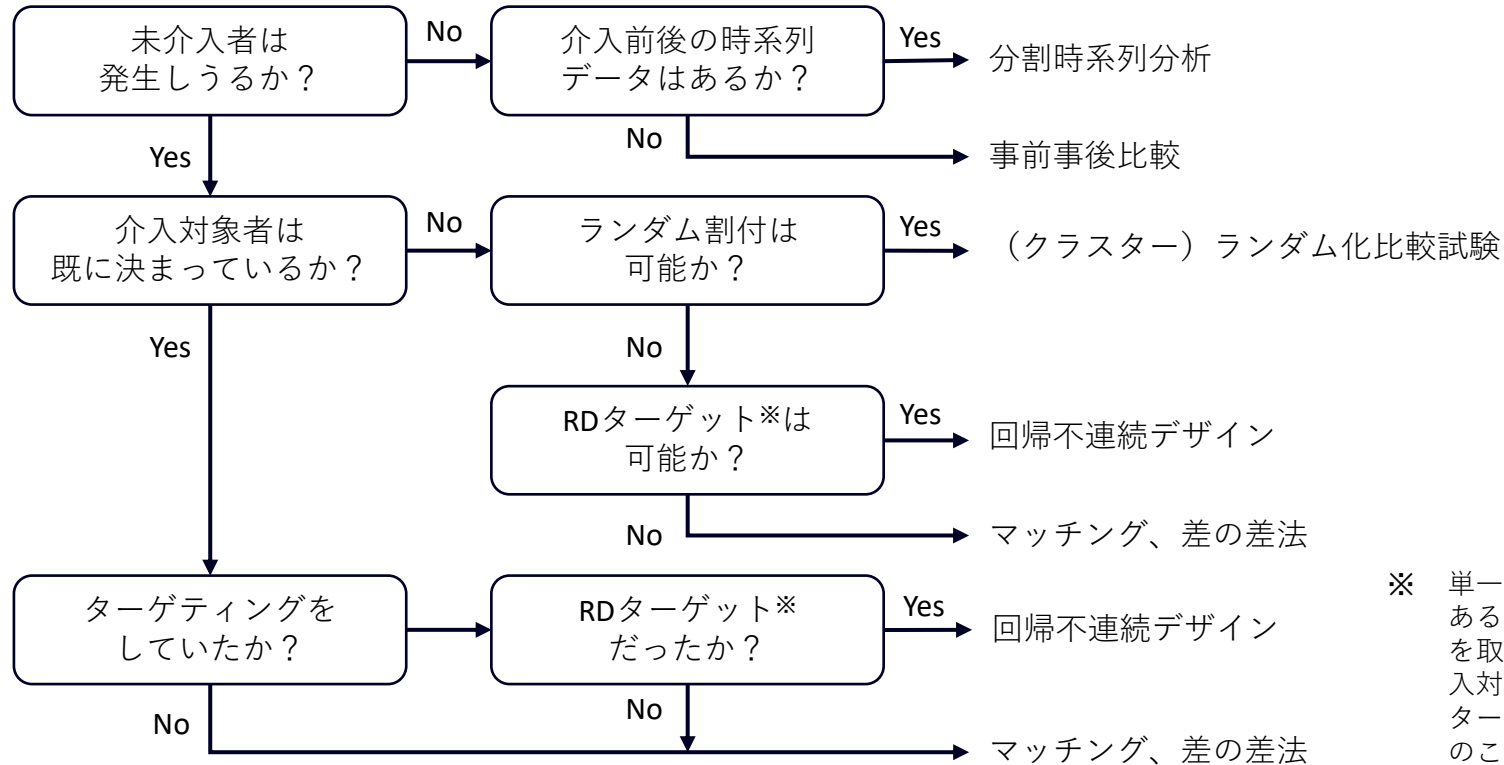
③ 実証デザインの決定

①と②を踏まえて、最適な実証デザインを決定しましょう。

【補足・解説】

手始めに選択可能な実証デザインを判断するには、以下のフローチャートが参考になるでしょう。マッチングや差の差法の利用可能性は、更に丁寧に適用可能条件が成り立つかを確認していく必要があります。また、このフローチャートは便宜的なものであり、機械的に実証デザインを提示してくれるものではありません。

なお、②に挙げた検討項目は実証デザインを左右する要素の一部に過ぎません。更に、具体的なアウトカム指標やサンプルサイズを検討する過程で実証デザインの見直しが必要になることもあるでしょう。この段階では仮決定という理解をしておくといよいでしょう。



※ 単一指標を用いて、ある値以上（以下）を取った場合には介入対象とするようなターゲティング方法のこと。

いなそうでいる未介入者

「政策は全国で実施される」ことが決まっていると、一見して未介入者は存在しない印象を与えます。しかしながら、段階的に対象範囲を拡大し最終的に全国がカバーされるといった過程を取る場合、拡大期においては一時的に未介入者が存在することになります。

或いは、全国の学校を介してサービスデリバリーがなされるものの、その対象となる個人は一部に過ぎないということもあるでしょう。この場合も、政策は全国で実施されていますが、対象となる個人とならない個人が存在しています。

政策実施の実態を仔細に調べてみることで、より好ましい実証デザインを考案する道が開けることもあります。

能動的に仕込む回帰不連続デザイン

実証デザインを扱うテキストの中には、回帰不連続デザインは自然実験的状況の一つとして位置付けられているものもあります。自然実験的状況とは、比較対象者が効果検証のために理想とする形で意図的に用意されたのではなく、偶然結果的に理想的な比較対象群が出来上がっていた状況のことをいいます。例えば、学力テストである点数以下を取った場合に補習への参加を促すという形で学力向上支援が設計されているようなものです。これは効果検証を念頭に考案された設計というわけではありません。自然実験として回帰不連続デザインを位置づけると、このような好都合な状況が見つかった時にのみ回帰不連続デザインは適用できるという誤った理解につながってしまいます。

政策立案時に取りうる実証デザインを検討している行政官としては、効果検証のために回帰不連続デザインが適用できるようにターゲティングを意図的に行うことができないかという発想を持つことが大切です。先の例で言えば、学力向上支援策の有効性を検証したいと考えるのであれば、回帰不連続デザインが使える状況を作為的に作り出すことを意図して、学力テストで一定の成績以下となった生徒に補習授業を提供するというターゲティング方法を仕込むという思考です。

ランダム割付の実行可能性を高める工夫

公平性を重視する教育行政の現場において、ランダムに選定された一部の対象だけに行政サービスである政策介入を実施する（逆の見方をすれば、残された一部の対象を行政サービスの受益者から排除する）ことに強い抵抗感を感じる人も少なくありません。

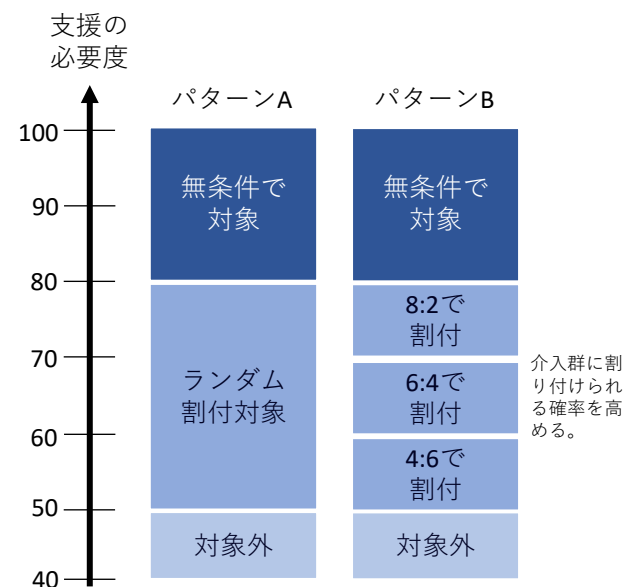
効果検証は結果の説明責任を果たすという目的もありますが、主たる狙いは有効な取組の開発を行うという点にあります。したがって、検証対象になっている介入は、まだ有効性が確立していないものであり（有効性が確立しているのであれば、わざわざ効果検証を行う必要はありません）、必ずしも有益な成果を生むとは限りません。そのような取組を大規模に行ってしまうことの方が、むしろ倫理観を問われるという考え方もあります。

しかし、そうは言ってもやはり抵抗感を払拭することは容易ではないでしょう。その抵抗感を多少は軽減できるようなランダム割付の工夫が多数考案されています。ここでは代表的な工夫を1つ紹介します。実証分析の専門家はこうした現場の抵抗感を和らげる知恵を有しているので、厳密な効果検証をしたいが難しさを感じている場合は、相談してみることも一案です。

例：境界周辺でのランダム割付

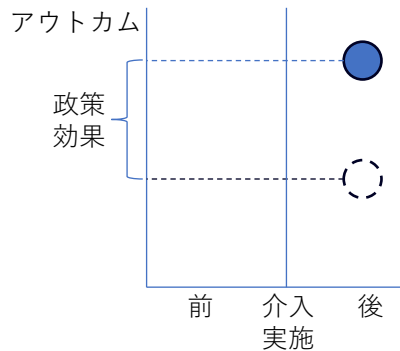
政策現場では、支援の必要性が高い人に対しては確実に介入する方が好ましいという考えからランダム割付に抵抗が生じることが少なくありません。支援の必要性が高いにも関わらず、偶然によって介入対象から外れてしまうことは認め難いという思いです。その場合は、右図（パターンA）に示すように非対象者になることが許されない層は無理にランダム割付の対象とせず、非対象者が出て許容できる層に限定してランダム割付を行うといった工夫を施すことが考えられます。つまり介入対象者の全てがランダム割付によって決まるのではなく、「ランダム割付の下で対象者に選ばれた人」+「無条件で対象者とする人」に介入が行われることとなります。ただし、分析に用いるのはランダム割付対象者のみなので、介入対象者数という点では非効率になります。

ランダム割付の対象となる層の中で、より必要度が高い人ほど介入を受けやすいようにするといった更なる工夫を組み込むことも考えられます（パターンB）。

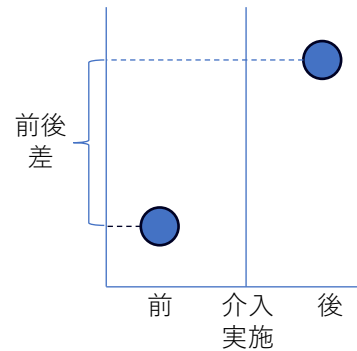


事前事後比較による政策効果把握

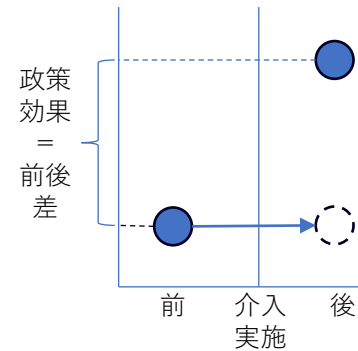
- 政策効果とは、同時点における政策介入が行われた時のアウトカム水準と、政策が行われなかった時のアウトカム水準の差として定義されます（下図①）。いずれか一方は実際には生じ得ない反事実的状況になります。
- 一方で、事前事後比較とは文字通り政策介入実施の前後時点のアウトカム水準を比較するものです（下図②）。この差と政策介入の事前事後のアウトカム水準の差は注目しているものがそもそも違うため、基本的には両者は一致しません。
- 両者が一致する唯一の状況は、「政策介入前のアウトカム水準」と「政策が行われなかった時のアウトカム水準」が等しい時です。つまり、政策が行われなかったとしたら、アウトカム水準は変化することなく推移するという状況です。この場合、事前事後比較は「政策介入が行われた時のアウトカム水準と、政策が行われなかった時のアウトカム水準の比較」と実質的に同じになっているので政策効果を正しく把握できることとなります（下図③）。
- では、この唯一の状況は頻繁に見られることなのでしょうか？ 私達は、このような状況になることを期待してもよいのでしょうか？ 殊、教育領域に関していえば、第3期教育振興基本計画において「成果に対して家庭環境など他の要因が強く影響している場合が多」と記載されていたように、アウトカム水準は政策以外にも様々な要因に影響されると一般的に考えられているでしょう。政策が行われていなかった時に、他の要因が変化することによってアウトカムが変化することは稀な事ではありません。
- もし政策が本質的に効果を生むものであったとしても、他要因が変化したことの影響を受けてアウトカム指標が改善していれば、事前事後比較によって把握される政策効果は過大評価となります。逆に、他要因が負の影響を及ぼした場合は政策効果は過小評価となります（下図④）。もし真の政策効果が負であっても、他要因が生み出す正の影響が非常に大きければ、事前事後比較の結果は過大評価どころか符号すら逆になってしまいます。事前事後比較は、このように常に政策効果を正確に捉えてくれないため、エビデンスレベルは比較的低いものとなります。



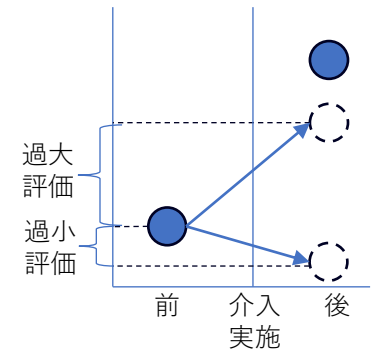
① 政策効果の定義



② 事前事後比較



③ 事前事後比較が正しい状況



④ 事前事後比較が誤る状況

既存アウトカム指標の参照方法

既存アウトカム指標活用の重要性

- アウトカム指標の設定は実証デザインの重要な一角をなします。政策立案の冒頭でアウトカムを合意できていたとしても、その水準を妥当な指標によって計測することができなくては、アウトカムの達成・未達成を判断することができません。
- アウトカムを的確に測定している指標を独自に考え出すことは容易ではないでしょう。特に独りよがりのアウトカム指標を考案し設定してしまった場合には、せっかく労力をかけて効果検証を行っても、指標が妥当ではないということで検証結果が軽んじられてしまうリスクも高まります。
- まずは**確立された指標を用いることができないかを十分に検討する**ことが望まれます。その際には、既存エビデンスの参照時と同様に、データベースなどを用いて効率的に検索を行うとよいでしょう。

① アウトカムの意味内容の確認

測定しようとしているアウトカムの意味内容を確認しましょう。

【補足・解説】

- アウトカムは指標とほぼ一体化した形で表現されることもありますが、一般的には例えば「学力が上がる」「非認知能力が高まる」といったように抽象的な表現となります。実際にアウトカムが達成されたかどうかを判断するためには、アウトカムを何らかの指標で捉え、その測定を行う必要があります。これが**アウトカム指標**と呼ばれるものです（P14も参照のこと）。アウトカム指標を検討する際には、まずアウトカムの意味内容を確認しておくことが不可欠です。
- 政策立案過程の冒頭でアウトカムは一意に定まっているはずですので、基本的にはその内容を確認することで十分です。もしこの段階でアウトカムが一意に定まっていない事に気づいたら、指標の検討を始める前にアウトカムの設定に戻る必要があります。

② アウトカムデータベースの参照

①で確認したアウトカムの意味内容を念頭に、データベース等を用いて指標や学術論文を検索しましょう。

【補足・解説】

- アウトカム指標を検討するに際しては、**標準化された指標**がないかを確認することが重要です。確立している指標は構成概念（指標で捉えようとしている内容）さえ正しければ、それを適切に測定できていることが担保されていると考えて差し支えないでしょう。また、その指標を用いた既存の効果検証結果との比較可能性を確保することができるといったメリットがあります。
- そのような既存指標を探すにあたっては、まずは指標が網羅されているデータベースを当たることが近道となるでしょう。**心理測定尺度であれば『心理測定尺度集Ⅰ～Ⅵ』が有用**です。収録されている心理測定尺度は以下のウェブサイトに掲載されています。

https://rnavi.ndl.go.jp/research_guide/entry/theme-honbun-400301.php

- もし尺度集のようなデータベースから適当な指標が見つからなかった時は、個別の報告書や学術論文に当たってみるとよいでしょう。この時も、**PICO**を踏まえて**Google scholar**などを駆使することで効率的に検索を行うことができます。学術論文や報告書の中には、効果検証が主題ではなくても注目しているアウトカムの指標化を行っているものがあるかもしれません。効果検証型の文献に限定せず、広く検索を行うことで最適な既存資料を見つけることが期待できます。



生理指標・行動指標・尺度指標

- 人々の状態や変化を捉えようとする時、指標が必要になります。数値化された指標を通じて現在の状態がどの水準にあるのか、ある期間にどのように変化したのかといったことを把握することができるようになります。
- 指標は大まかに生理指標、行動指標、尺度指標の3種類に分類することができます。
- **生理指標とは生体の生理反応を捉えた指標**であり、脳波や血圧、脈拍、呼吸、皮膚電位などといえば想像がつくでしょう。例えば人がストレスを感じている時には、コルチゾールが副腎皮質から分泌することが知られています。そこで唾液中コルチゾール量を指標とすることでストレスの有無を把握するといったことができます。
- **行動指標とは、観察することのできる人々の行動を何らかの形で指標化したもの**です。教育関係者の間で有名となったマシュマロテストでは、子供の衝動・感情をコントロールする能力を把握することを試みています。この直接的には目に見えない能力を把握するために、目の前に置かれたマシュマロを食べてしまうか否かという行動を観察し、食べた場合はコントロール力なし、食べなかった場合はコントロール力ありとみなしています。この例のように特異な環境下（実験的環境）で人々が見せる行動に限らず、朝食を食べた・食べない、挨拶をした・しない、地域活動に参加した・しないといった行動も行動指標に含まれます。
- **尺度指標とは、質問紙を通じて行動や、観察することのできない人々の内面（ココロや考え・意見）を観測し、必要に応じてその回答に統計的処理を施し得点化したもの**です。性格、感情状態、適応状態などを数値化したもので、目にする機会も多いのではないのでしょうか。
- 各指標には利点と欠点があります。いずれかの指標が常に他を優越するということはありません。一方で、教育行政の現場で活用することを考えると、教育現場で観察することのできる行動指標、及び質問紙を通じて大規模かつ短時間で収集できる尺度指標を活用することが一般的です。

推奨	種類	欠点	利点
X	生理指標	<ul style="list-style-type: none"> ・ 時間を要する ・ 高額な測定器具 	<ul style="list-style-type: none"> ・ 心理的要因で生起する ・ 客観性が高い
◎	行動指標 (学校での観察)	<ul style="list-style-type: none"> ・ 記録管理者が必要 ・ 形式が統一されてない 	<ul style="list-style-type: none"> ・ 過去データも利用可能 ・ 教師と変化を共有可能
△	行動指標（実験）	<ul style="list-style-type: none"> ・ 考案に時間を要する ・ 実験環境下のデータ 	<ul style="list-style-type: none"> ・ 条件を操作できる ・ 各条件のデータが揃う
○	尺度指標	<ul style="list-style-type: none"> ・ 妥当性・信頼性等が他指標より劣る ・ 質問項目が多い 	<ul style="list-style-type: none"> ・ 短時間で収集できる ・ 選択肢が豊富にある

③ 指標の妥当性（適用可能性）検討

検索の結果、入手できた指標が当該政策の文脈において適用できる指標かどうかを検討しましょう。適用可能性を判断する際には以下の視点が有用になるでしょう。

- アウトカムの意味内容に合致しているか
- 文脈を踏まえて適切な事象を測定しているか
- 指標を測定することができるか

例：アウトカムの意味内容

「食生活の乱れの改善」というアウトカム*を検証した先行事例があり、ここでは「一週間の内、朝食欠食であった日数」というアウトカム指標が用いられていたとします。しかし、政策立案の過程でアウトカムが「食事の回数が改善する」という形で合意されていた場合、朝食欠食の日数は「食事の回数が改善」を部分的に捉えているかもしれませんが、適切な指標とは言い難いでしょう（食事の回数は朝食欠食のことを指しているのでしょうか？）。

*「食生活の乱れ」はアウトカムとしては曖昧過ぎるものでした。はじめから、このような曖昧なキーワードで指標の検索を行わないことも重要です。

例：文脈からの妥当性

小学校において教師が子供に向き合うための努力量の増加をアウトカムとする政策を考えてみます。既存の文献を検索した結果、多くの発展途上国で同種のアウトカムを目指した政策が行われており、具体的なアウトカム指標として「教師の欠勤率」が用いられている事がわかりました。

発展途上国においては、低賃金や学校への交通費といった問題から教師がそもそも学校に来ないという事態が広範に見受けられます。このような状況下においては「教師が子供に向き合うための努力量の増加」を欠勤率で把握するというのは妥当な方法だと一般に考えられています。しかし、日本の文脈で考えれば、アウトカムを捉える指標としてはいささかポイントがずれたものになることは明らかでしょう。

例：測定可能性

「栄養の偏りの改善」というアウトカムに対して、数日間に渡る自記式食事記録による食品摂取多様性スコアという確立した指標が見つかったとします。アウトカムの意味内容に照らして適切だと判断できたとしても、調査予算に照らして十分な精度や規模で自記式食事記録という方法でデータ収集を行うことはできないと思われる場面もあるでしょう。

新規心理尺度の開発方法

心理尺度開発手順の重要性

- アウトカムを的確に把握できる既存指標がない場合には、不完全な指標であっても妥協して使用することも重要です。しかし、新規の指標を開発したいという場面もあるでしょう。
- 質問紙によってアウトカムを把握することも少なくありませんが、**質問紙は十分な開発プロセスを踏まえて完成させる必要があります**。適当に作成した質問紙から作られるアウトカム指標では、測定したいアウトカムを適切に測れないでしょう。特に行動ではなく人々の内面を捉えようとしている場合は、新規心理尺度の開発手順を踏まえる必要があります。
- 新規心理尺度を作成する場合は、暫定的な質問項目を用いて予備調査を実施し、その回答を踏まえて改訂・最終化を図るという工程を伴います。作成される指標は、あくまでこの予備調査の対象が代表する母集団において機能する指標になります。最終的な効果検証の対象となる母集団を意識して、指標作成工程を踏んでいくことが不可欠ですが、既存の教育行政制度の中で、どのように調査を含む各工程を実施していくかは慎重な検討が必要になります。
- 新規心理尺度の作成難易度から、第一に既存指標を用いることを検討し、新規作成は最後の選択肢とすることが賢明でしょう。

① アウトカムの意味内容の確認

測定しようとしているアウトカムの意味内容を確認しましょう。

【補足・解説】

- このステップは、既存アウトカム指標の探し方の①と同じになります。

② 質問項目の作成

どのような質問をすればアウトカムが意味する概念を測定することができるか、質問項目を考えましょう。

【補足・解説】

- 様々な角度、観点からアウトカムを捉えることができるような質問項目を絞り出すことが求められます。アイデアを得るための標準的な方法として、①一部の介入対象者に対して聞き取りを行ったり、自由記述による意見をもらう、②関連概念を扱った既存の心理尺度の項目を参考にする、③専門家にヒアリングをする、④ブレインストーミングをするといったことがあります。
- 質問項目の素材から具体的な質問項目（質問文・質問紙の構成）を作成する際には、一般的な社会調査法で指摘される質問紙における禁じ手に留意しましょう。

③ 予備調査の実施

②で作成した質問紙ドラフトを用いて、予備調査を実施する。

【補足・解説】

- 予備調査を行う際は、最終的に介入対象となっている母集団（完成した心理尺度を用いてアウトカム指標を収集していく対象）と同一の母集団を対象に行うことが望まれます。極端な例をあげるならば、最終的には日本の大学生を対象とする介入効果を検証することを想定しているにも関わらず、予備調査をアメリカの大学生を対象に行ったのでは、日本の大学生にとって妥当な質問紙とはならないことは直感的にもうなずけるでしょう。

④ 予備調査で得られた回答結果の分析

質問項目が満たすべき性質を満たしていたか、得られた回答を分析して確認しましょう。

【補足・解説】

- この工程では、得られた回答結果に対して様々な検討を加えていきます。はじめに項目分析を行い、不適切な項目を排除します。さらに、妥当性、信頼性、内的整合性といった観点から統計的な検討を行うことが一般的です。
- 各分析の方法や妥当性や信頼性等を検証する具体的な作業方法については専門的となるため、本マニュアルで説明することはしません。詳細については、専門書を参照したり、専門家の助言を得てください。

⑤ 質問項目の修正

③④の結果を踏まえて、質問項目の修正や削除を行い、質問紙の改訂を行います。

【補足・解説】

- 必要に応じて2回目の予備調査を行い、④⑤の工程を繰り返します。
-
- このように、尺度指標の開発は高い専門性が必要となり、非常に手間のかかる作業となります。したがって、安易に新規尺度を開発するという判断をしてしまうのは賢明とはいえません。入念な検討をした上で、やはり新規尺度を開発しない限りアウトカムを捉える適切な指標が手に入らないということであれば、尺度開発の専門家を交えた検討が不可欠です。

実施規模の決定方法 (サンプルサイズ)

サンプルサイズ設計の重要性

- 統計分析によって介入効果の検証を行うことを想定している場合は、分析時に扱うデータの規模（サンプルサイズ）を十分に検討しておくことが不可欠です。サンプルサイズは効果検証の事業規模によって規定されるため、サンプルサイズを検討することは介入の規模を検討することに繋がります。
- サンプルサイズが不足している場合は、**統計分析を行っても介入が意図した効果を生んでいるのか否か明確に判断できない**という深刻な問題を引き起こします。
- 本来、サンプルサイズは予算や協力者の意向から自ずと決まってしまうものではありません。目的と実施環境を加味して、**計算公式を通じて求めるものです。必要サンプルサイズに基づいて、介入の実施規模を決定するという思考順序を持つことが重要**になります。
- 何らかの事情で介入の実施規模が確定してしまっている場合は、その状況で得られるサンプルサイズから分かることは何かを検討し、目的に叶った効果検証ができるのかどうかを十分に吟味しておく必要があります。

① 受容可能最小政策効果量の設定

政策効果として許容しうる最小限の大きさを決めましょう。具体的には、①政策介入があった時に期待しているアウトカム指標の水準、②なかった時に想定できるアウトカム指標の水準、そして③アウトカム指標の標準偏差の3つの値を定めま

【補足・解説】

- 一般に、統計分析においてはサンプルサイズが大きくなればなるほど、小さな介入効果しか生じていなくても、つまり介入群と統制群/比較群のアウトカム水準の差が小さくても、統計的に有意な差があるという結論が得られます。逆の見方をすれば、サンプルサイズが十分でない、小さな介入効果は統計的に有意であるという結果にはならないことになります。
- 小さな介入効果であっても政策的に重要な成果であると考えられるのであれば、「介入が全く効果を生まないものだとしても偶然観察できるような誤差とみなせる。したがって、介入効果はあるとはいえないだろう」という結論に至るのではなく、「介入は小さいながらも確かに効果を生んでいると考えてよい」という結論が得られなくては分析を行う意義が薄れてしまいます。サンプルサイズが小さすぎるために前者のような結論になることは避けなくてはなりません。
- サンプルサイズの設計とは、効果ありと見なさなくてはいけない最小効果の大きさを適切に扱えるサイズを見極めることです。この最小効果の大きさを**受容可能最小政策効果量**と呼びます。次ページに示すように、受容可能最小政策効果量を述べるには「政策介入を行った場合と、行わなかった場合の平均アウトカム水準の差」と「アウトカムの標準偏差」の2つの値を決めなくてはなりません。この内、前者は施策実施者が、「この政策は最低限どのくらいのアウトカムを産まなくていけないと考えているか」「この政策の目標値はいくつと考えているのか」といった**意見**になります。他方で後者はアウトカムの標準偏差なので**客観的事実**として存在する値になります。
- 受容可能最小政策効果量を提示するにあたっては前者は意見なので必ず決めることができます。他方で、後者の情報を得ることは難しいことが多いでしょう。先行事例や関連統計から当て推量を得ることを目指しますが、手がかりがまったくない時は、アウトカム指標の分布を知るための小規模な調査を行い、その結果を参照するというのがもっとも安全な方法になります。ただし、このような小規模調査を実施することは時間的にも予算的にも現実的ではない事が多く、実際には専門家の意見なども参考にしつつ、標準偏差が大きい場合、小さい場合というようにいくつかのシナリオごとに受容可能最小政策効果量を計算しておくことが一般的です。なお、意見であるアウトカム水準の差についても、大胆な値、控えめな値といったようにいくつかのシナリオを想定しておいてもよいでしょう。
- なお、サンプルサイズを設計する際には、アウトカムを一つに絞る必要があります。アウトカムが複数ある場合は、それぞれのアウトカムごとに同様のプロセスを踏んで、各アウトカムに対応するサンプルサイズを求めます。

用語の確認；受容可能最小政策効果量

- 政策介入を行った場合と、行わなかった場合の平均アウトカム水準の差のことで、一般的に効果量と表記されます。
- ただし、測定単位に影響されない値とするために、標準偏差の値（客観的事実）で除した形で表現されます。これは偏差値の算出と同様の考え方です。
- サンプルサイズの検討においては、政策実施者が期待している効果の大きさの下限值（裁量値）という理解になります。

例. 学力向上を狙った教育政策

介入を行わなかった場合は50であるアウトカム値が、介入によって55まで上がると**期待している**とします。アウトカムの標準偏差は10であることが分かりました。この時、効果量は、以下の値になります。

$$\text{効果量} = \frac{\text{アウトカムの差}}{\text{標準偏差}} = \frac{55 - 50}{10} = 0.5$$

効果量の目安

前述の通り、標準偏差の妥当な値が分かることは極めて稀なため、受容可能最小政策効果量の見積もりはサンプルサイズ設計の最大の難関です。

アウトカムによっては、政策介入がおおよそどの程度の介入効果を持ちうるかという目安が同種の効果検証の経験蓄積から提示されている場合があります。例えば、教育分野におけるテストスコアに対する効果量については、**0.1**以下は「**Small effect**」、**0.3**以上は「**Large effect**」、**0.5**であれば「**Very large effect**」とする見解があります。当然アウトカムが違ったり、アウトカムを生むための困難さは文脈に大きく依存するため一概に**0.3**が**Large effect**であるとみなせませんが、情報が限られている場面においては一定の参考値となるでしょう。

② その他のパラメータの値の決定

サンプルサイズを計算するにあたって必要となる受容可能最小政策効果量以外のパラメータの値を決定します。

【補足・解説】

- 受容可能最小政策効果量以外のパラメータは、実証デザインによって異なります。例えば、RCTを用いる場合は、①有意水準 (α)、②検出力 ($1-\beta$) の値を指定する必要があります。クラスターRCTの場合は、これらに加えて③クラスター数、④クラスターサイズ、⑤級内相関係数 (Intra-Cluster Correlation: ICC) を決定する必要があります。このうち、有意水準は5%、検出力は80%に設定することが一般的であり、それ以外の値を積極的に用いる理由がない限りは、これらの値を用いればよいでしょう。
- 教育領域の政策では、例えば教育委員会や学校単位で政策の対象者を決め、生徒や教員からアウトカムを計測するといったように、介入はクラスター単位となり、アウトカムはクラスターの構成員から収集するという構造を持つことが多いと思われます。その場合は③-⑤の値を決定する必要があります。これらの値はいずれも客観的な事実となりますが、中でも⑤級内相関係数の値は正確な情報が得られないことが多いでしょう。級内相関係数の値はサンプルサイズに非常に大きな影響力を持つため、この見積もりを失敗するとサンプルサイズが足りなかったということになりかねません。前ステップで検討した標準偏差同様、正確な値は小規模調査から得られますが、実施できることは稀なので、様々な情報ソースを駆使して、妥当な値にたどり着けるよう最大限の努力をする必要があります。
- ③-⑤の値は非常に専門的な知識が必要になってくるため、専門家に判断を仰ぐことも一案です。ここではこれ以上の解説は控えます。

有意水準

- 統計的仮説検定において、帰無仮説（政策には効果がない）が真であるにもかかわらず、帰無仮説を偽として棄却してしまう確率のこと。

検出力 (Power)

- 帰無仮説が偽であるときに正しく帰無仮説を棄却する（介入効果はゼロではないと判断する）確率のこと。

級内相関係数 (Intra-cluster correlation coefficient: ICC)

- 個々のアウトカム値が、クラスターごとにどの程度のまとまりを持っているかを示す値で、0~1の間の値を取ります。
- クラスターの環境が大きく異なる時、同一のクラスターに属する個体のアウトカムは似た値を取る傾向があります。
- クラスターランダム化比較試験のようにクラスター介入を行い、アウトカムはその構成主体から測定するという場合に限り、ICCの値を加味してサンプルサイズを設計する必要があります。

③ サンプルサイズの計算

①②で決定したパラメータの値を基にサンプルサイズを計算します。実際の計算にあたっては専用のソフトウェアを用いることが多いでしょう。

【補足・解説】

- サンプルサイズを求める計算は手計算で行うことはほとんどありません。RやStataといった統計分析専門のソフトウェアや、ウェブ上で実行できるアプリなどが多数提供されています。操作方法が簡便で、使いやすいものを用いるとよいでしょう。基本的にどのソフトウェア、サービスもこれまで検討してきたパラメータの値を入力することを求められるので、それらを指定すると自動的に必要サンプルサイズを表示してくれます。
- 受容可能最小政策効果量をはじめとする裁量的パラメータ、及び様々な当て推量の結果得られている客観的パラメータの値を変えることで、必要サンプルサイズがどの程度変化するか確認しておくことも有用です。

数値例：ランダム化比較試験

海外留学への奨学金付与のように生徒単位で行われる政策介入をランダム化比較試験によって検証する場合を考えてみましょう。ここでは、介入対象者と非対象者の比率を1：1とする場合を扱います。受容可能最小政策効果量と必要サンプルサイズの関係を示したものが下表になります（有意水準は5%、検出力は80%と標準的な値を用いるものとします）。

効果量が0.1の場合は、3130名（介入群1565名、統制群1565名）というサンプルサイズが必要となります。他方で、効果量を0.8と大きくした場合は、僅か52名（介入群26名、統制群26名）のサンプルサイズで十分となります。

効果量	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
サンプルサイズ	3130	800	350	200	130	90	70	52

注：全体のサンプルサイズ（介入群と統制群の合計）を示しているため、各群に必要なサンプルサイズは半分となる。

数値例：ランダム化比較試験②

海外留学への奨学金付与のような介入の場合、アウトカムデータは独自調査によって把握することになるかもしれません。そのため、検出力（通常は80%）を確保できる最小のサンプルサイズを選択することが理に適っているといえるでしょう。その場合の介入群・統制群の配分比率は1:1になります。

他方で、全国学力・学習状況調査のように効果検証とは無関係に予定されている大規模な調査の結果をアウトカム指標として用いる場合は、全てのデータを活用し必要な検出力を確保できるような介入群・統制群の配分比率を検討することも考えられます。

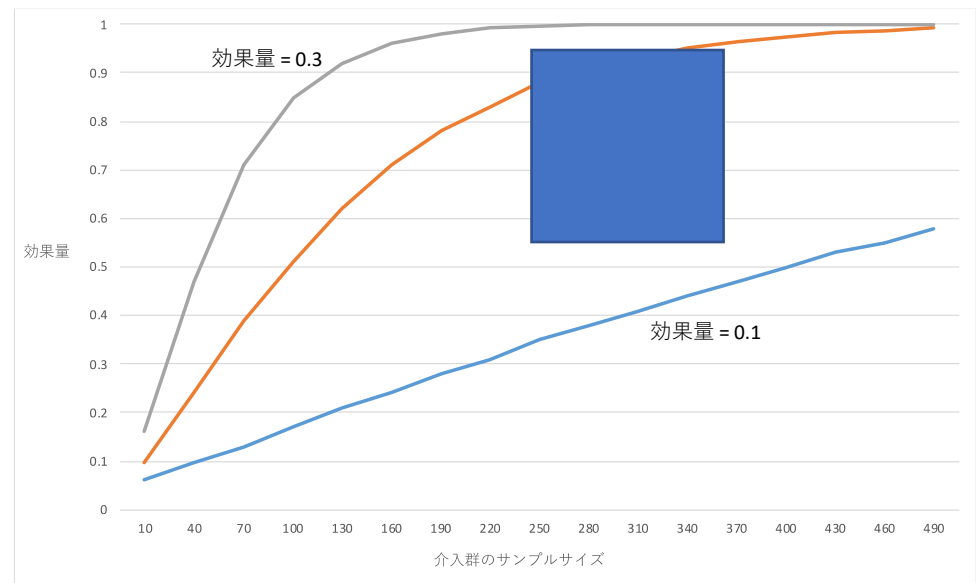
例えば、問題が発生している対象が10,000人いて、彼らのアウトカムデータはコストをかけることなく利用可能であるとします。この時、受容可能最小政策効果量ごとに、検出力と介入群への配分比率（人数）の関係を示したものが下図になります。

受容可能最小政策効果量が0.2の場合、10,000人のうちランダムに選定された202人に対して介入を行えば、アウトカムデータを利用することができる10,000人を全て用いた分析を行うことで検出力80%を確保することができます。

先の1:1の配分比率の場合と介入規模の比較をしてみましょう。効果量が0.2だとすると400名に介入する必要がありました（サンプルサイズは800人）。しかしサンプルサイズを10,000人にすることで、介入対象数は201人で済みます。

図には記載されていませんが、効果量が0.1の場合は、10,000人のうち860名に介入すれば十分となります。1:1の配分比率の時は介入群に1,565人が必要だったことを考えると、かなり介入コストを削減することができます。

このように全体のサンプルサイズを最小化するのか、介入対象者を最小化するのかによって必要となるサンプルサイズ設計の考え方も変わってきます。



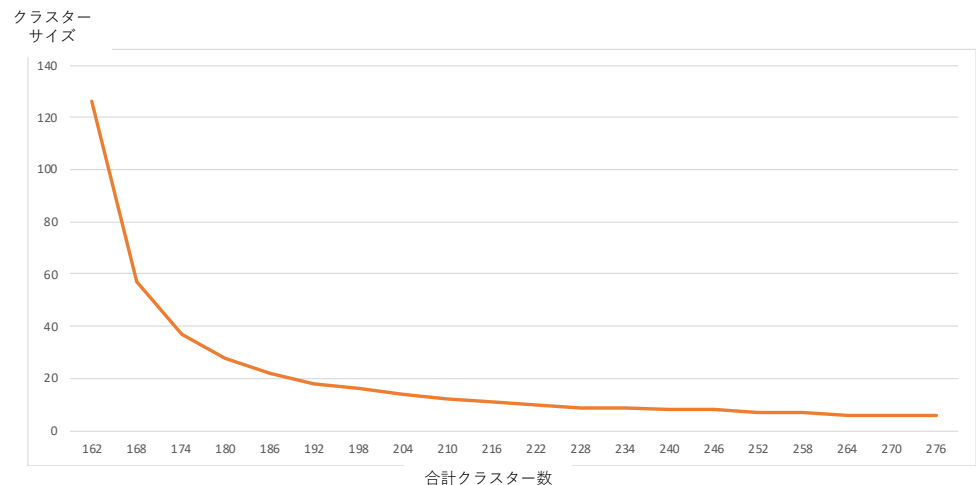
数値例：クラスターランダム化比較試験

学校単位で食育プログラムを実施し、個々の生徒の食生活をアウトカム指標とするような政策介入の効果をクラスターランダム化比較試験によって検証する場合を考えてみましょう。この場合、まず受容可能最小政策効果量と客観的パラメータであるICCの値を固定します。次に、裁量的パラメータのうち、クラスター数とクラスターサイズを検討します。クラスター数×クラスターサイズでサンプルサイズが決まることになりますが、クラスター数を増やしクラスターサイズを小さくした方が、一定の検出力を小さなサンプルサイズで確保することができます。換言すると、一定のサンプルサイズを所与とした時には、クラスター数を大きく取った方が検出力が大きくなることを意味します。

クラスターサイズの上限（例えば学校の生徒数）に気をつけながら、定めた受容可能最小政策効果量とICCの下でクラスター数とクラスターサイズの組み合わせを吟味し、実施可能な規模を探っていきます（有意水準は5%、検出力は80%と標準的な値を用いる）。例えば効果量が0.2、ICCが0.1のケースを考えてみます。各クラスターの上限が100人であった時、それらすべてからデータを取得すると必要クラスター数は86校（介入クラスター43校、未介入クラスター43校）となります。この時、合計クラスター数（学校数）が最小化されています。アウトカムデータは86校×100人＝8,600人から収集することになります。

他方で、クラスター数をそれよりも増やし、各クラスター内でのデータ収集対象を減らすと、より少ないサンプルサイズで済ますことが可能となります。クラスター数とクラスターサイズの関係を右図に示してあります。例えば、クラスター数を180校（介入クラスター90校、非介入クラスター90校）にするとクラスターサイズは28人となります。サンプルサイズは180校×28人＝5,040人で済みます。

最適な組み合わせを費用最小化という観点から選択するのであれば、クラスター数を増やすことのコスト増とサンプルサイズが増えることのコスト増を比較していくことになります。



なお、ランダム化比較試験の数値例②のように、介入クラスターと統制クラスターの配分比率を1:1にする必要は必ずしもありません。配分比率を変えることで、より効率的なサンプルサイズの設計（介入クラスター、データ収集対象の設計）が可能になります。ただし、パラメータが増えることで複雑度が増していくため、ここではこれ以上の説明は行いません。

- サンプルサイズを計算する際に用いられる代表的なソフトウェアやウェブサービスとして以下があります。それぞれ操作方法は異なりますが、基本的には実証デザインごとに必要なパラメータの値を入力していくという形になります。
- 個々の操作方法は各種マニュアルやウェブサイトを参照して下さい。

汎用的統計分析ソフトウェア

- R
- Stata

専門ソフトウェア

- Optimal Design

ウェブサービス

- PowerUp!

```

. power twomeans 30 33, sd(5)
Performing iteration ...
Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1 versus Ha: m2 != m1

Study parameters:
alpha = 0.0500
power = 0.8000
delta = 3.0000
m1 = 30.0000
m2 = 33.0000
sd = 5.0000

Estimated sample sizes:
N = 90
N per group = 45

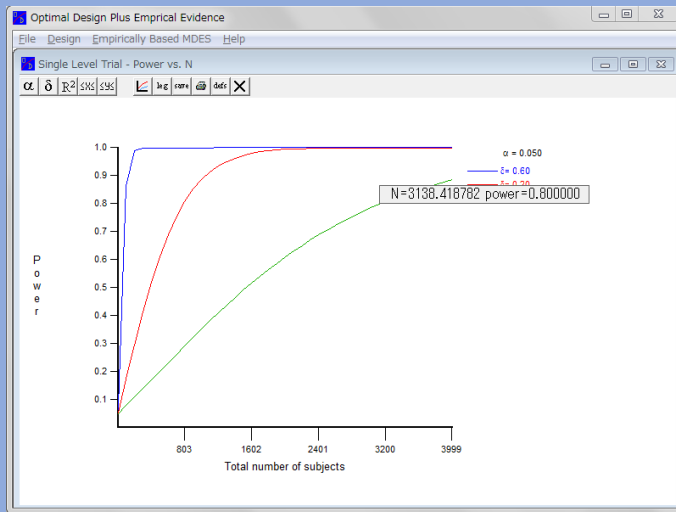
. power twomeans 30 33, sd(5)
Performing iteration ...
Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1 versus Ha: m2 != m1

Study parameters:
alpha = 0.0500
power = 0.8000
delta = 3.0000
m1 = 30.0000
m2 = 33.0000
sd = 5.0000

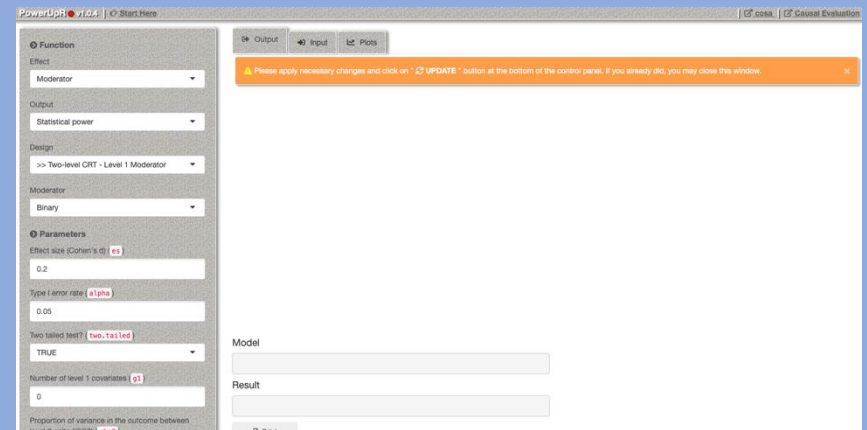
Estimated sample sizes:
N = 90
N per group = 45

```

Stata



Optimal Design



PowerUp!

④ サンプルサイズの確定

以下のような点を加味して最終的なサンプルサイズを決定します。

- データの不備といったような不測の事態が発生する見通し
- 複数のアウトカムを見ている場合

【補足・解説】

- ③で求めたサンプルサイズは最終的に分析に用いることのできるサンプルサイズを意味しています。現実には、収集したデータのうち、不備があり使用できないケースがあったり、そもそもアウトカムが入手できないケースがあったりします。不測の事態があっても必要サイズを確保できるように、ある程度水増しをして最終サンプルサイズとします。
- また、複数のアウトカムを想定している場合は、大は小を兼ねるという考えから、最大となるサンプルサイズ（水増し調整済み）を最終サンプルサイズとして選択することを検討します。或いは、主要アウトカムのサンプルサイズを優先させるといった判断をします。
- 最終的に、ここで確定したサンプルサイズを確保できるような規模で政策介入を行っていきます。なお、ここで求まるサンプルサイズは介入規模ではない点に注意して下さい。非介入者も含めた比較分析に用いるサイズになります。介入群と統制群/比較群が1:1となる場合は、実際の介入規模は求まったサンプルサイズの半分になります。

- 前述の通り、サンプルサイズは事業規模に先決して検討を行うべきものです。しかし、現実的には様々な事情から事業規模を所与としなくてはならない場面も少なくはないでしょう。
- そのような場合には、与えられたサンプルサイズの下で**検知できる効果量の下限値（Minimum Detectable Effect: MDE）**を求めることになります。
- MDEが受容可能最小政策効果量を上回ってしまっている時は、たとえ政策が価値があると考えられる効果を生んでいたとしても、「効果があるのかないのか分からない」という結果になる可能性が高いことを意味します。
- そのような状況であることを十分に理解した上で、まずは、所与の事業規模の中で少しでも検出力を大きくする工夫を施さないかを考えることが重要です。代表的な工夫の方法としては分析に用いる説明変数を増やしたり、データ収集の時点数を増やすといったことが考えられます。またRCTのようにランダム割付を伴うデザインを検討しているのであれば、層別ランダム割付を行うといった工夫もあるでしょう。
- それでも状況が改善しきらないのであれば、はっきりとした結果が得られないかも知れないというリスクを受け入れて効果検証を組み込むか、抜本的に事業規模を大きくすることができないか再考し直すという判断をしなくてはなりません。

数値例：クラスターランダム化比較試験

海外留学への奨学金付与の効果を検証する場合を考えてみます。本事業には全国から1,000人の応募があり、そのうち300名が要件を満たしていたとします。また、奨学金を付与できる上限は150名だったとします。公平性を期すために、奨学金付与者150人は応募要件を満たしている300人から抽選で決めることにしました。抽選で介入対象者を決定するので、ランダム割付を行っていることと同じ状況になります。

この300名からアウトカムデータを取ることができる場合、介入群、統制群の最大サイズは各150名となります。この時、MDEは0.325となります。つまり、もし受容可能最小政策効果量が0.325よりも小さいのであれば、十分な検出力を確保できないことを意味します。想定していた受容可能最小政策効果量が0.2だった場合、このサンプルサイズにおける検出力は約41%になります。

このような小さな検出力の場合、有意義な効果検証にはならないでしょう。そこで検出力を上げるために、介入前のアウトカムの値（ベースライン値）や、その他の説明変数を計測し分析に組み込むことができないかといった検討をしていくことになります。