Institute for
Effective Education
Empowering educators with evidence

評価ハンドブック

ー新しい教育の効果を評価する方法ー

（Evaluation handbook）

日本語版（Ver.1.0）

2021 年 8 月

Institute for
Effective Education
Empowering educators with evidence

・原書は Institute for Effective Education (IEE)が作成しました。
・執筆者：アリシア・ショー(Alicia Shaw)
・原書を引用する際は以下のように記載してください。
・Institute for Effective Education (2020) Evaluation handbook. York: Institute for Effective Education

・翻訳者：岡崎善弘（岡山大学学術研究院教育学域）
・連絡先：okazakiys@okayama-u.ac.jp

Institute for
Effective Education
Empowering educators with evidence

日本の先生方へ

学校や教師は新しい教育方法を試すことがよくあります。教師がひとりで開発した教育方法、同僚と協力して開発した教育方法、教師以外の人が作った（効果は検証されていない）教育方法、効果が実証された教育方法を生徒のニーズに合わせて修正した方法等が挙げられます。

どうすれば「新しい教育方法の効果」を知ることができるでしょうか？多くの場合、教師や学校は簡単な評価方法に頼る傾向があります。例えば、「新しい教育方法についてどのように思いますか？」「全体的に改善したと思いますか？」等の質問をします。これらの質問で得た回答は確かに役立つかもしれません。しかし、信頼できる評価方法ではありません。

Institute Effective Education（IEE）は、教師や学校が自分たちだけで新しい教育方法を評価するプロジェクト（27 件）を支援しました。IEE は、学校がよりロバストな評価を実施することができるかどうかを知りたかったのです。そして、その過程の中で、評価方法ついてたくさんのことを学びました。

本書では、上記の評価プロジェクトで得たことを集約していますので、新しい教育方法を評価したいと考えている教師や学校の役に立つかもしれません。本書が日本の学校の皆様のお役に立てるようでしたら幸いです。

Jonathan Haslam
Alicia Shaw

Institute for
Effective Education
Empowering educators with evidence

目次	ページ
イントロダクション	01
本書の使い方	02
エビデンスを参照して学校を改善する	03
01. 取り組みたい問題・課題を特定する	04
02. 問題の理解と解決策を決定するためにエビデンスをレビューする	05
03. リサーチクエッションを書く	06
04. インパクト評価を計画する	09
a) 参加者と募集方法	09
b) 条件に割り当てる	11
c) 新しい方法	14
d) アウトカムの測定	16
e) プロセス評価	21
f) 分析	23
g) タイムライン	25
h) 予算	25
05. 計画した通りに評価する	26
06. データを分析する	27
a) 測定したアウトカム	27
b) プロセス評価の分析	28
07. 結果を書く	29
08. 評価の限界を書く	30
09. 結論を書く	31
10. 次のステップを決める	33
a) 実践から示唆を得る	33
b) 評価から示唆されること	33
c) わかったことを発信する	33
d) リフレクション	34
リファレンス	35
エビデンスが掲載されているサイト	37
用語集	39
Appendix A	46
Appendix B	49

イントロダクション

教師が「新しい方法」を取り入れることは日常的です。教師1人で考えた方法や学校の同僚と一緒に開発した方法を試すこともあるでしょう。また、他校の教師が作った（効果を検証していない）方法や、効果が実証された方法をアレンジして実施することもあるでしょう。しかし、方法を新しく開発しても、公平に正しく評価していなければ、「新しい方法」が以前の方法よりも優れているかどうかを判断することはできません。

公平に正しく評価することができれば、以下のことが理解できるようになります。

・「新しい方法」は生徒にどのような効果を与えるのか。
・「新しい方法」を実施するために必要な支援は何か。
・「新しい方法」の実施に関わる潜在的な課題は何か。

「新しい方法」を広める前に「評価」することができれば、教師やスクールリーダーは質の高い情報に基づいてリソース配分等の意思決定ができるようになります。少人数の生徒やクラスを対象にして事前に評価することができれば、学校全体に広く導入する前にその方法に関する障害を取り除くことができるでしょう。

研究の種類は様々です。研究は教師のさまざまな質問・疑問に答えてくれるでしょう。「研究の種類」と「教師が知りたいこと」の対応関係は、「Engaging with Evidence Guide」で詳しく説明しています。ある方法が生徒のアウトカムに与える効果を知るためには、インパクト評価（impact evaluation）を実施することが最も良いでしょう。インパクト評価では、「新しい方法」で教育を受けた生徒と受けていない生徒を比較するアウトカム評価（outcome evaluation）と、「新しい方法」がどのように実施されたのかを調べるプロセス評価（process evaluation）を組み合わせます。2 つの評価から、「アウトカムの改善に貢献したのか」や「なぜそのような結果になったのか」を理解することができます。

本書の使い方

本書の目的は、学校を改善する実践の「評価の計画・実施」を支援することです。評価を実施する前に、新しい評価方法を学びましょう。本書では、最初に、「問題・課題を特定する方法」について説明します。次に、インパクト評価（impact evaluation）の計画と実施方法を説明します。そして、結果を分析して結論を出すまでの枠組みを紹介します。

エビデンスを参照して学校を改善するプロセスとして、次ページに示した 10 のステップをお勧めします。現在進行中のプロジェクトで本書を利用する場合は、関連する部分だけを詳しく読むという使い方でも良いでしょう。

本書の中に見慣れない用語が含まれているかもしれませんので、本書の最後に用語集を載せています。教育研究に関する用語を包括的に集めた用語集ではありませんが、本書を理解する上で参考になれば幸いです。

評価を計画する時は Appendix A をぜひご利用して下さい。プロジェクトが完了したら、「Writing up your innovation evaluation report」を参考にして、評価結果を執筆して、共有すると良いでしょう。

本書では、「学校」という言葉は、幼児教育、特別学校、オルタナティブ教育、大学等、あらゆる教育機関を指す言葉として使用しています。また、イノベーション（Innovation）という言葉は、評価したい新しい授業方法や介入方法を指す言葉として使用しています。
（注：日本語版では、イノベーション（Innovation）を「新しい教育」と翻訳しています）

エビデンスを参照して学校を改善する

「エビデンスを参照した学校の改善」では、直面している問題を理解するために学校のデータを利用します。そして、対応方法のエビデンスを調べて、実施した方法の効果を評価します。

評価は以下の 10 のステップに従って進めると良いでしょう。

01. 取り組みたい問題・課題を特定する
02. 問題の理解と解決策を決定するためにエビデンスをレビューする
03. リサーチクエッションを書く
04. インパクト評価を計画する
05. 計画した通りに評価する
06. データを分析する
07. 結果を書く
08. 評価の限界を書く
09. 結論を書く
10. 次のステップを決める

あなたが評価したい「新しい方法」をすでに用意している場合でも、ステップ 01 から始めることをお勧めします。データとエビデンスに基づいて「新しい方法」を再検討しましょう。

1．取り組みたい問題・課題を特定する

学校でどのような問題が起きているでしょうか。生徒の学習、教師の定着率、保護者の関与、生徒の問題行動、不登校、メンタルヘルス等、問題は学校生活の様々な側面と関連している可能性があります。広い問題よりも、特定の問題の方が理解しやすく、対処もしやすいので、問題は明確に特定しましょう（例えば、「児童は行動に問題がある」ではなく、「言葉やコミュニケーションの支援が必要な児童は長期間の停学を受ける傾向がある」のように、解決したい問題を明確にしましょう）。そして、特定した問題に介入することができるか確認してください。

問題に対処するためには、どのようなことが問題なのか、誰が影響を与えているのか、考えられる本質的な原因をよく理解することが重要です。例えば、あなたの学校の小学１年生の児童の読解力が期待されていた水準よりも低かったとしましょう。この問題に対処するためには、読解の要素（フォニックス、流暢な読解力、単語の認識力、理解力、語彙力等）に問題はないか、偏った教育を受けている児童のグループがあるのか、また、このような望ましくない結果の原因となっている要因が他にもあるのかどうかを知ることが大切です（例えば、出席率が低い、授業中の児童の行動が良くない等）。

問題の原因を十分に把握することができない場合、仮説検証の方法を用いると良いでしょう。仮説を検証するためには、問題を引き起こしている可能性が最も高い原因を仮説（hypotheses）にしてリストアップします。エビデンスを参照して可能性が最も高い仮説を特定したり、問題に関連する理論を同僚に聞いたりすることも有効です。次に、仮説を検証するためにデータを集めます。教師の評価、日常的に収集されている学校のデータ、潜在的に関連する標準化された指標（standardized measures）の測定値、スタッフの認識、生徒の意見、保護者の意見等、さまざまな情報を使うことができます。すべてのデータの収集を集めることができれば、問題の原因の理解ができるかもしれません。複数の要因が問題の原因になっていると考えられる場合は、これらの要因のうち、**比較的容易に変更できる要因や介入が可能な要因**を１つ選ぶことをお勧めします。

2. 問題の理解と解決策を決定するためにエビデンスをレビューする

この時点で問題の原因は何か見当がついていなければなりません。問題と原因について、研究からわかっていることを確認しましょう。学校に関する問題は広範囲にわたって研究されており、問題の対処方法や改善方法を決定したい時に研究の知見が役立ちます。

解決策を見つける時に役立つ「研究の要約」が多くの組織から提供されています。これらの組織は「エビデンスが掲載されているサイト」（37 ページ）に記載しました。「Engaging with Evidence Guide」では、エビデンスの種類、それぞれのエビデンスからわかること、エビデンスの質や信頼性（reliability）の評価方法を紹介しています。

次に、解決策を決定するためにエビデンスを探しましょう。見つけたら、効果が期待できるかどうか検討して下さい。変化の理論（theory of change）を作成して、その解決策が効果的に機能するメカニズムを説明してみましょう。

解決策を選択したら、一貫して正しく実施される可能性を最大限に高める方法を検討します。EEF の「Putting evidence to work: a school's guide to implementation」は、実施計画を立てる際に役立つでしょう。

Box1：変化の理論

変化の理論とは、選択した解決策の効果を説明するモデルです。変化の理論では、「解決策」と「望んでいる結果」をつなぐ道筋を仮説的に示します。変化の理論の作成を通して、「どのように対処するのか」や「なぜ対処できると考えているのか」を深く考えることができるでしょう。変化の理論を作成するためには十分な時間の投資が必要です。このプロセスを怠ってしまうと、流行っている方法や、自分の好みで解決策を選択してしまう危険性が高まります。

変化の理論に正解はありません。期待する変化を視覚的に表現するために、フローチャート、コンセプトマップ、ウェブ、表等を利用すると良いでしょう。変化の理論を作成する方法を紹介している資料は数多くあります。「Community Tool Box」や「The Center for Theory of Change」をぜひ参照して下さい。

3. リサーチクエッションを書く

評価したいリサーチクエッションを具体的に書いてください。一貫性のある評価を計画・実施する上で、明確なリサーチクエッションを持つことはとても大切です。評価対象の「新しい方法」、実施する期間、測定するアウトカム、対象者は必ず記述して下さい。エビデンスを参照して解決策を決めることができたとしても、アウトカムの設定において、ステップ 1 で指摘したような問題が生じるかもしれません。リサーチクエッションは、以下のフレームを利用すると良いでしょう。

[どれくらいの期間] で [どのような方法] を提供したら
[誰] の [どのようなアウトカム] にどのような効果を与えますか？

リサーチクエッションの各要素は明確に定義しましょう。例えば・・・

[1 日 10 分, 週 4 回, 4 か月] で [Toddler Talk プログラム] を提供したら
[英語が母国語ではない 2 歳児] の [表現力] にどのような効果を与えますか？

効果を複数のアウトカムで調べたいと思うかもしれません。複数のアウトカムで調べたい場合、1 つのアウトカムに対して 1 つのリサーチクエッションを書いて下さい。しかし、リサーチクエッションの数が多過ぎると評価が難しくなります。一般的に、リサーチクエッションは多くても３つです。3 つ以上のリサーチクエッションを 1 回で答えようとする評価はお勧めできません。

何が起きるのかを具体的に予測する（仮説を立てる）のも良いでしょう（例：4 ヶ月間の Toddler Talk プログラムに参加した 2 歳児の表現力は、対照群の 2 歳児の表現力よりも高くなるだろう）。仮説を検証する際は、対照群を設けて下さい。そして、設定しているアウトカムが似ている研究を探して参照して下さい。

Box2：良くないリサーチクエッション（1）

良いリサーチクエッションの作成は、とても難しい作業です。誰かにリサーチクエッションを読んでもらって、すべての要素が具体的かどうか確認してもらうと良いでしょう。ピットフォールはいくつかありますが、時間をかけて考えたり、クリティカルな意見を言ってくれる友人の助けを借りたりすれば、いずれも解決できます。

明確にアウトカムを定義していないケース

[6カ月間] で [高校生の生徒と毎週メンタリング] は
[中学3年生の生徒] の [到達度] にどのような効果を与えますか？

上記の例では、「到達度」は良く定義されたアウトカムではありません。これ以上は詳しく説明することができない段階まで「到達度」とは何かを問い続けて下さい。

・「到達度」とは何ですか？
　「年度終わりの到達度」です
・「年度終わりの到達度」とは何ですか？」
　「年度末のテスト成績」です
・「年度末のテスト成績」とは何ですか？
　「主要科目の年度末のテスト成績」です
・「主要科目の年度末のテスト成績」とは何ですか？
　「年度末の英語，数学，科学のテスト成績」です

方法、期間、参加者の記述が不十分なケースもあります。これらの要素を明確にするために、上記のような質問（「・・・とは何ですか？」）を繰り返しましょう。

アウトカムが複数あるケース

[1年間] の [スターのバッジを与えるシステム] は
[key stage2（小学3年生～小学6年生）の生徒] の
[破壊行動・クラスから退出・停学] にどのような効果を与えますか？

上記の例では、リサーチクエッションに3つのアウトカムが含まれています。3つのリサーチクエッションに分けておくと、結論が書きやすくなるでしょう。

Box2：良くないリサーチクエッション（2）

リサーチクエッションが広過ぎるケース

［2 学期分］ の ［TA の授業補助］ は［担当している生徒］ の
［学習のしやすさ・進捗度・潜在能力の発揮］ にどのような効果を与えますか？

関心の範囲が広過ぎると、何が知りたいのか不明になることがあります。上記の例では、方法が明確ではありません。TA の配置を改善する方法は 1 つではないので、どの方法の効果を知りたいと考えているのかわかりません。そして、3 つのアウトカムが明確に定義されていません。さらに、アウトカムの 1 つは測定することができません（生徒が潜在能力を発揮したかどうかを判断する客観的な方法はありません）。

「向上させたい具体的なアウトカム」と「アウトカムを向上させる対象者」をリサーチクエッションで示す必要があります。アウトカムや対象者を明確に定めることができない場合は、ステップ 1 に戻りましょう。問題に対する新たな理解が得られた場合は、ステップ 2 も作り直しましょう。そして、解決策と、その解決策を提供する期間を明記して下さい。解決策の最適な提供期間がわからない場合は、ステップ 2 を繰り返してください。

解決策のインパクトに対する予想を含んでいるケース

［1 コマの授業時間］ の
［ナレッジオーガナイザーを利用して「二都物語」の理解を深めること］ は
［高校 3 年生の生徒］ の ［「二都物語」のテスト成績］ にどのような効果を与えますか？

上記の例では、「ナレッジオーガナイザーを利用すれば二都物語の理解が深まる」という想定が含まれています。ナレッジオーガナイザーを利用すれば二都物語の理解が深まるかどうかはわかりません。大切なことは、中立的な質問を書くことです。期待していた結果にならない可能性を受け入れることも大切です。

4. インパクト評価を計画する

リサーチクエッションに答える評価方法を計画しましょう。科学的に妥当な評価をするためには、参加者、参加者の募集、デザイン、測定法、評価したい方法、プロセス評価、データ分析を事前に決定しておくことが重要です。考える順番は特に決まっていません。計画する際は、評価計画（Appendix A 参照）を利用すると良いでしょう。

意思決定する際は、ステークホルダーの協力を得ましょう。関係者全員から理解を得ておくことは大切です。計画を明確に書いて、評価を実施する前に関係者全員に読んでもらいましょう。計画書をコピーして、関係者全員から同意のサインを得ておくと良いでしょう。評価に参加するすべての学校関係者に同意してもらうことはとても大切です。

評価を開始する前に計画を公表することはとても良いことです。「何をしようとしているのか」を明確に説明することができれば、実施中に仕方なく変更・妥協した部分がわかります。したがって、取り組みの透明性が保証されるため、十分な情報を得た上で結論を出すことができます。学校のウェブサイトや評価を一緒に実施している組織のウェブサイトで計画を公開することも検討してみましょう。

a） 参加者と募集方法

評価の対象になる標本（sample）を決めて下さい。参加者（participants）は、リサーチクエッションで書いたグループを代表している必要があるので、参加者として必要な特性（学年、年齢、現在の到達度、母国語、性別等）を決めましょう。参加者として必要な特性は、採用基準（inclusion criteria）と呼ばれています。

次に、参加者の募集方法を決めましょう。所属している学校外の生徒を募集する場合は、いつ、どのように他校に伝えるかを決めて下さい。そして、関係する各学校のスタッフとミーティングを実施して、「何をしようとしているのか」を全員が明確に理解しているか確認して下さい。関係者から賛同を得る方法も計画しておきましょう。

生徒の参加に同意が必要かどうか、誰に同意をしてもらうのか、どのような同意を求めるべきかを決めて下さい。16 歳未満の生徒は、評価を開始する前に親の同意を得る必要があります。また、必要に応じて生徒の同意も得た方が良いでしょう。少なくとも、参加者が 16 歳以上であるなら、生徒の同意を必ず得て下さい。インタビューは、年齢に関係なく、必ず参加者の同意を得て下さい。参加の同意を求めない場合は、生徒のデータを分析に含めることに同意を求めるだけでよいでしょう（つまり、生徒全員を参加させて、自分の子どものデータを分析に含めるかどうかを保護者に決定してもらいます）。「新しい方法」が日常的に実施している小さな実

践の変更であれば、オプトインの同意ではなく、オプトアウトの同意（「新しい方法」で教育を受けた後、保護者が自分の子どもを評価に参加させたくないと回答すること）の方が生徒の参加率は高くなります。どのような形で同意を得るにしても、評価が有害にならないように配慮することはあなたの責任です。「British Education Research Association ethical guidelines」は、教育研究の倫理の理解に役立つガイドです。

チェックリスト

・参加者の採用基準は決まりましたか？
・参加者を募集する方法は決まりましたか？
・どのような同意が必要ですか？
・同意を得る方法は決まりましたか？

b) 条件に割り当てる

生徒は常に成長し続けています。成長は「新しい方法」の効果ではなく、他の要因の影響かもしれません。「新しい方法」が現在の方法よりも優れているかどうかを調べたい時、どのようにして「新しい方法」の効果を評価すると良いでしょうか。「新しい方法」の効果を調べるためには、「新しい方法」で教育を受けた生徒と受けていない生徒のアウトカムを比較する必要があります。正しく比較しなければ、間違った結論を導く可能性があります。例えば、典型的な進歩（例：中学 2 年生頃に代数の理解が向上する）や時間の経過（例：中学 2 年生の数学の理解度が 1 年間で上がっている）が「新しい方法」の効果として解釈されるかもしれません。生徒の学業成績以外の能力も時間とともに変化するため（例えば、学校の関与は年齢とともに減少する）、「新しい方法」を経験していない生徒と比較することは、学業成績以外の測定であっても重要です。

条件（condition）とは、参加者が所属する群のことです。介入群（intervention group）と対照群（control group）があります。介入群は、あなたが評価したいと思っている「新しい方法」で教育を受けます。対照群は、これまで通りの方法で教育を受けます。開始する時点で、介入群と対照群は、できるだけ類似している必要があります。そして、どちらの群も「新しい方法」で教育を受けていないことが必須です。さらに、両群の生徒は採用基準（inclusion criteria）を満たしている必要があります。

条件割り当て（assignment to condition）とは、参加者を各条件に配置する方法のことです。「新しい方法」の規模に合わせて参加者を異なる条件のどちらかに割り当てる必要があります。例えば、新しい方法が個人で提供される場合（例：1 対 1 の個人指導プログラム）は、各参加者を対照群と介入群に割り当てます。「新しい方法」がクラスで提供される場合（例：数学の新カリキュラム）は、各クラスを対照群と介入群に割り当てます。「新しい方法」が学校全体で提供される場合（例：健康的な食事プログラム）は、各学校を対照群と介入群に割り当てます。

対照群や介入群に参加者をランダムに割り当てる方法（random assignment）は、関係者が恣意的に参加者を各群に割り当てる方法よりも、バイアスを減らすことができます。ランダム化比較試験（randomized control trials）は、参加者を各群にランダムに割り当てます。介入群と対照群に参加者を無作為に割り当てる方法として、乱数発生の利用（excel の機能やオンライン上で利用できます）、コインをはじく、袋の中に入れた生徒の名前を 1 つずつ取り出す等があります。

参加者をランダムに各群に割り当てたら、介入群と対照群の特性がよく類似している（well-matched）かどうかを確認して下さい。例えば、「メンタルヘルスプログラムが主観的幸福感に

与える効果」を評価したい時、(a) 無料で学校給食を受けていない高校 3 年生の生徒と(b)無料で学校給食を受けている高校 3 年生の生徒を介入群と対照群にランダムに割り当てたとします。そして、プログラムを開始する前に、高校 3 年生の生徒全員が同程度の幸福度を持っていたかどうか、また、各群に無料で学校給食を受けている生徒が同程度の割合で含まれているかどうかを確認しなければなりません。介入群と対照群の特性がよく類似していない場合は、満足できるまでグループを再ランダム化しましょう。

現実的な理由から、参加者をランダムに各群に割り当てることができない場合もあります（例えば、時間割の制約により、参加するすべてのクラスのうち「新しい方法」で教育できるクラスが 1 つしかない）。参加者をランダムに割り当てない実験は準実験（quasi-experiments）と呼ばれています。各群の割り当てがランダムではない場合は、評価を開始する前に、測定するアウトカムや人口統計学的特性の観点から、各群がよく類似しているか確認することが特に重要です。

チェックリスト

・生徒を介入群と対照群に割り当てる方法は決まりましたか？
・各群が十分に類似しているか確認する方法は決まりましたか？

Box3：対照群

対照群を設けることに反対する人もいます。対照群が「新しい方法」で教育を受けないことは不公平だと感じるからです。しかし、現時点では、その「新しい方法」が現在の方法よりも優れているかどうかはわかりません。だからこそ、評価が必要なのです。「新しい方法」はプラスの効果を与えることがすでにわかっている場合には、「新しい方法」の利用を保留することは非倫理的です。EEF が実施した評価の約 80%は効果を示すことができていないため、「新しい方法」が本当に優れた介入であることはとても稀です。

対照群を設置することに対して不快である場合は以下を参照して下さい。

・対照群は、いつも通りに過ごす対照群（business as usual control group）です。つまり、あなたの学校で現在使用している方法で教えることができます。「新しい方法」で提供されるコンテンツを対照群に教えることが大切です。対照群に教えないことは、非倫理的であるだけでなく、「新しい方法」が現在の方法よりも効果的であるかどうかを理解する助けにもなりません。

・「新しい方法」が効果的であるとわかった場合、評価が完了した後に対照群に「新しい方法」を使用することができます。対照群を待機群（waiting list control group）として扱うことができるので、対照群は両群の利点を得ることができます。「新しい方法」が現在の方法よりも効果的であれば、対照群に参加した生徒はその恩恵を受けることができます。現在の方法が効果的であるとわかれば、対照群の生徒は、効果が低い方法で教育を受けることはありません。

c) 新しい方法

評価したい「新しい方法」を明確に説明しましょう。関係者全員が「何を期待して、何をするのか」を正確に理解しておく必要があります。評価のレポートを読んだ人々がよく理解できるように、明確に記述してください。以下のアウトラインに沿って「新しい方法」を明確にしましょう。

- 「新しい方法」で何を教えるのか（科目・トピック・スキル等）。
- 参加者は「何を」「どれくらい」経験することになるのか。
  - 「新しい方法」を提供する時間の長さ。
  - 「新しい方法」を提供する回数、1回あたりの時間、提供する頻度、提供する時期。
  - 参加者が取り組む内容。
  - 「新しい方法」に間接的に影響するコンテンツの提供（例：宿題）。
- 「新しい方法」を提供する規模（例：個人、グループ、クラス、学校全体）。
- 「新しい方法」を生徒に提供する人は誰なのか。
- 「新しい方法」を提供する人に与えられるトレーニングは何か。
- 「新しい方法」が提供されている間に支援は何かあるのか。
- 「新しい方法」を実践している途中でスタッフが退職した場合はどうするのか。
- 新しいスタッフが途中で参加した場合はどのようにトレーニングするのか。
- 手引き書や授業計画等、「新しい方法」を支援する様々なリソースはあるのか（共有された資料から明確な方向性を知ることができれば、一貫性を高めることができます）。
- 「新しい方法」に関連するその他の情報はあるのか。

評価期間中に対照群に提供する内容も決める必要があります。「いつも通りに過ごす（例：教師や学校がいつも通りの方法で教育を提供する）」、「代替の方法で教育を提供する」、「一部を変更した「新しい方法」で教育を提供する」等が挙げられます。進行の途中で対照群が「新しい方法」の要素を経験した時に起こる汚染（contamination of control group）をどのようにして回避するかも検討しましょう。

「新しい方法」を評価する時、評価者は関わり方を慎重に検討する必要があります。あなたが「新しい方法」を開発・提案している場合は特に注意が必要です。介入群と対照群のどちらにも教師として参加しない方が理想的です。介入群を別の人に任せることができれば、あなたが気づかなかった良いアイデアを得ることができるかもしれません。また、対照群も別の人に任せることができれば、コンタミネーション（対照群に「新しい方法」の要素を誤って使用してしまうこと）のリスクを減らすことができます。

チェックリスト

・「新しい方法」を明確に記述していますか？
・「新しい方法」を提供する人のトレーニングの内容は決まりましたか？
・「新しい方法」を提供する人はどのようなトレーニングを受けますか？
・「新しい方法」を提供する人はどのようなサポートを受けますか？
・必要なリソースをすべて特定しましたか？また、それらをいつ購入、収集、作成しますか？
・対照群は何をするか把握していますか？
・対照群に参加している人が「新しい方法」を学ぶことができないようにしていますか？
・対照群が「新しい方法」に関する情報にアクセスできないようにしていますか？

d) アウトカムの測定

「新しい方法」の効果を確認するために、客観的なアウトカム指標（outcome measures）を利用しましょう。多大な時間と労力を投資しているため、プラスの効果を与えていないデータが示されたとしても、「プラスの効果を与えていたはずだ」と信じたがる傾向があります。データはプラスの効果を示さなかったとしても、参加者の意見だけに頼っていては、「新しい方法」の効果を正しく把握することはできません。したがって、リサーチクエッションで示した「問い」のアウトカムは客観的な指標を用いて測定するべきです。可能であれば、アウトカムの測定は学校の外で作成された標準化された指標（standardized measures）を利用して下さい（「新しい方法」の開発・提案や評価に関わる人がテストを作成してはいけません）。標準化された指標は、自作した指標よりも信頼性と妥当性が高く、偏っていません。しかし、自由に利用できる標準化された指標は多くありません。つまり、評価に使用するアウトカム指標を独自に作成する必要があるかもしれません。

「新しい方法」を経験した生徒（介入群）と経験していない生徒（対照群）のアウトカムを比較するために、介入群と対照群は同じアウトカム指標で測定する必要があります。「新しい方法」を始める前にアウトカムを測定して（事前テスト）（pre-test）、生徒のスタートポイントを把握します。そして、対照群と介入群の類似度を判断して下さい。また、「新しい方法」の提供が終了した後もアウトカムを測定します（事後テスト）（post-test）。介入群と対照群の生徒がどれくらい伸びたのかを確認してみましょう。また、効果が時間の経過とともに持続するかどうかを確認するために、「新しい方法」の提供が終了してから数週間後や数ヶ月後にアウトカムを測定することもできます（遅延テスト）（delayed post-test）。

事前・事後テストが生徒や教師の大きな負担にならないように配慮して下さい。可能であれば、生徒が必ず受ける一般的な測定（例：典型的な学年末のテスト）を使用してください。そして、公平な測定になるようにして下さい。介入群に教えたことだけ（対照群には教えていないこと）を評価してはいけません。介入群の生徒だけが見慣れている（対照群の生徒は見慣れていない）教材等に基づいてテストを作成してはならないということです。可能であれば、事前テスト、事後テスト、遅延テストは互いに比較できるように設計しましょう。例えば、標準化されたテストの等価版（equivalent versions of standardized measures）があります。

Education Endowment Foundation の「DIY Evaluation Guide」には、さまざまなタイプの指標（全国的な評価、標準化されたテスト、自作の指標）の情報があります。EEF の「SPECTRUM database」には、学業以外のスキルの測定方法が掲載されています。

アウトカムの測定について、以下のことを決定して下さい。

・どのような指標を使用するのか。
・いつ測定するのか。
・誰がどのように測定するのか（測定はすべての参加者に対して同じ条件で行うべきです。可能であれば、評価に関与していない人が測定して下さい）。
・テストの採点方法と採点者（可能であれば、採点者は個々の生徒を特定できないようにして下さい。各生徒がどのような状態にあるのかもわからないようにして下さい）。
・テストの採点の信頼性をどのように保証するか（例：採点基準を作ってから採点する）。

生徒や教師の行動がリサーチクエッションのアウトカムになっている場合は、テストの前後に構造化された観察（structured observation）を実施することが適切です。構造化された観察では、あらかじめ定義されたグループ（1 人、参加者の中から選ばれたグループ、参加者全員）の一定の時間内に生じた行動を、観察者が記録するという方法です。対象となる行動を明確に定義して、観察者が確実にその行動を識別できるように訓練することがとても大切です。観察計画（observation schedule）が観察時に役立つかもしれません。

Box 4：構造化された観察（1）

構造化された観察（structured observation）は、行動を記録する体系的な方法です。「新しい方法」が意図された通りに実施されているかどうかをチェックする際に役立ちます。行動の変化に関心がある場合は、行動をアウトカム指標として利用することができます。構造化された観察をする前に、あなたが気になっている行動を明確に定義する必要があります。観察をアウトカム指標として使用する場合は、何を行動としてカウントするのか、何をカウントしないかを明確に書きましょう。また、プロセス評価の一部として観察を行っている場合は、生じることが予想される「新しい方法」の特徴をすべて列挙して下さい。他の人があなたと同じように行動や「新しい方法」の特徴を特定することができるように、詳しく具体的に書いて下さい。観察する行動をリスト化したら、観察計画（observation schedule）を作成しましょう。

例：1：検索練習授業の忠実度チェック

日時：　　時間：　　クラス：　　担当教員：　　観察された生徒：

検索練習の授業の特徴	チェック
・子どもたちは教室に入ったらすぐに検索練習を始めている。	
・生徒たちは、学習したことを思い出す活動している。フラッシュカード、検索練習ソフト、仲間たちと問題を出し合う、マインドマップの作成等が含まれる。	
・検索練習は 5 分から 10 分ほどの時間で実施されていた。	

次の項目は「新しい方法」の中に存在しないはずです。
観察された行動があれば〇をつけてください。

・生徒がノートを読み直している。
・生徒が教科書や過去に書いたノートを使ってフラッシュカード等を作り直している。
・生徒が質問の回答を見つけるために、教科書、ノート、インターネットを使用している。
・活動が 5 分未満で終わっていた。
・活動が 10 分以上続いていた。

Box 4：構造化された観察（2）

例 2：教師の質問に答えるために手を挙げた生徒の行動観察

日時：　　時間：　　クラス：　　担当教員：　　観察された生徒：

生徒が手を挙げていると判断できる行動

・手が顎の高さより上にある。
・手が頭に触れていない状態。
・教師が質問をしている時に手を上げる。

生徒が手を挙げていると判断できない行動

・手が顎の高さより下にある。
・手が顔や髪に触れている。
・教師が質問をしていない時に手を挙げた場合。
・1 つの質問に対して生徒が何回も手を上げ下げした場合は、カウントは 1 回。

	集計（教師の質問に対して各生徒が挙手する度に 1 カウントする）
生徒 A	
生徒 B	

行動のカウントは、離散的な行動を知りたい時に有効です。長い時間にわたって続く行動に関心がある場合は、観察を小さな時間のブロックに分割して、それぞれの時間帯に起きた行動を記録して下さい。その行動が観察された比率を報告することができます。

Box 4：構造化された観察（3）

例 3：課題に関する行動の観察・課題に関係ない行動の観察

日時：　　時間：　　クラス：　　担当教員：　　観察された生徒：

授業と関係がある行動

・教師が話をしている時、教師の方を見ている。
・ディスカッションに参加している。
・ライティングに取り組んでいる。
・指示された通りに取り組んでいる。

授業と関係がない行動

・授業と関係ない移動をしていた。
・授業と関係ない話題を教師や他の生徒に投げかけていた。
・授業と関係ない話題を他の生徒と話していた。
・30 秒以上、教室の周りや窓の外を見ていた。
・一方向を 30 秒以上見つめる（教師や話している人ではなく）。

時間	1	2	3	4	5	6	7	8
課題中								
課題以外								

生徒全員の行動を正確に記録することは不可能なので、観察する必要がある場合は、クラスの中から代表的な生徒を選んで観察して下さい。リサーチクエッションで特定の特徴を持つ生徒に言及している場合は、該当する生徒を観察する必要があります。すべての生徒の行動に興味がある場合や、対象基準を満たす生徒が多過ぎて正確に記録できない場合は、観察するクラスの中から代表的な生徒を選んで下さい。対照群のクラスで観察される生徒は、介入群のクラスで観察される生徒と同じ基準で選ばれていることを確認してください。クラスのすべての生徒を観察したい場合は、授業を撮影しましょう。後で観察記録を作成することができます。

誰を観察するにしても、どのように観察するとしても、観察をアウトカム指標として用いるのであれば、事前テストと事後テストで同じ生徒を観察しなければなりません。

e) プロセス評価

プロセス評価（process evaluation）は、新しい教育方法が「どのように実施されたか」を示す情報です。「新しい方法」が意図された通りに実施されたかどうか（実装の忠実度）（implementation fidelity）を知りたい時に役立ちます。また、「新しい方法」に対する教師と生徒の意見、成功したことや改善できたこと等の認識を知ることができるため、プロセス評価は、アウトカム評価の結果の理解する時も役立ちます。

忠実度（Fidelity）

「新しい方法」が計画した通りに実施されているかどうかを判断するために、「新しい方法」が実施されている様子を観察する必要があります。構造化された観察（structured observation）は、構造化されていない観察よりも要約や分析が容易で、客観的である可能性が高いため、プロセス評価をする時に役立ちます。「新しい方法」の主な特徴を特定した上で、目立たないように観察する計画を作成します。Box4 に忠実度の観察計画（observation schedule）の例を示しました。対照群が「新しい方法」の要素を 1 つも受け取っていないことを確認するために、対照群も観察した方が良いでしょう。観察は、「新しい方法」で教育を始めてからある程度の時間が経過した後に実施することが一般的です。

計画書や生徒のノートをチェックして、「新しい方法」が計画した通りに実施されているかどうかを確認することもできます。「新しい方法」で教育を提供する人に記録やチェックリストの記入を求める場合は、どれくらいの負担が増えるのか、慎重に検討してください。

チェックリスト

・計画した通りに「新しい方法」が提供されているか確認する方法は決まりましたか？
・対照群が「新しい方法」の要素を受けていないことも確認しますか？
・観察する場合、誰が、いつ、観察するかを決めましたか？
・観察計画は作成しましたか？
・忠実度を確認する場合、誰が、いつ、どのような基準で確認するか決めましたか？

参加者の声・視点

「新しい方法」を提供した人や経験した人に感想を尋ねることは有益です。関係者全員に「新しい方法」に対する意見を聞くだけでなく、「新しい方法」で教育を実施した人に、経験してわかった課題や、実施の成功を支えた要因を尋ねることも有益です。もしあなたがこの方法を再び利用することを選択した場合、トレーニング方法や支援方法の改善を考える時に彼らの意見がとても参考になります。

主要な関係者に調査（survey）を行い、「新しい方法」に対する彼らの見解を確認する必要があります。インタビュー（interviews）、質問紙（questionnaires）、フォーカスグループ（focus groups）等が一般的な調査方法です。調査の目的をよく考えて、目的に合った質問を書く時間を十分に確保してください（質問紙の作り方は Appendix B で詳しく説明しています）。素早く簡単に分析できるクローズドクエッションと、参加者の意見をより深く理解するオープンクエッション（ただし、分析に時間がかかる）の両方を含めると良いでしょう。インタビューやフォーカスグループは、少数の人々から詳細なフィードバックを収集する柔軟な方法です。フォーカスグループはとても難しいことに注意してください。

専門知識のない人がフォーカスグループを実施した場合、グループの力関係や強い個性が回答に影響を与えることがよくあります。

質問紙やインタビューは、新規性効果（新しいという理由だけで何かを好きになったり嫌いになったりすること）や学習曲線効果（参加者が時間をかけて理解を深めてゆくこと）を避けるために、通常は評価の最後に実施します。「新しい方法」の開発・提案や評価に関与していない人に対面式のインタビューを依頼することを検討してください。

チェックリスト

・誰の意見を収集するか決まりましたか？
・意見を収集する方法（インタビュー、質問紙、フォーカスグループ）は決まりましたか？
・質問内容は決まりましたか？
・質問紙の実施方法、インタビューする人、それらを実施する予定は決まりましたか？

f) 分析

アウトカム（outcome）

アウトカムの測定から得られた定量的（数値的）（quantitative）データをどのように分析するのか、分析方法を決めて下さい。事前にデータ分析の計画を立てることは、行き過ぎた計画のように思えるかもしれません。しかし、分析の信頼性、頑健性、偏りのなさを確保するためには、とても大切なステップです。計画していない分析の危険性は、こちらのブログ記事をご覧ください。データを得たら、介入群と対照群の事前テスト、事後テストの平均値を算出してみましょう（計画していた場合は遅延テストも平均値を算出する）。また、両群の平均的な進歩（事前と事後のスコアの差）を計算するのも良いでしょう。使用するデータを要約する方法は以下の通りです。

- 平均値（mean）とは、データを足し合わせて数値の個数で割って求められる値のことです。スコア間の距離が一定であるデータの平均を求める時に使用できます。テストの点数、テストやアクティビティにかかった時間、ある行動が観察された回数等に平均値を使うことができます。

- 中央値（median）は、データの中央に位置する値です。評価スケールの回答等、スコア間の距離が一定ではないデータの代表値を求めたい時に使用します。社会的・情緒的な能力や経験の自己報告等は評価尺度を使用しているので、中央値はこれらの計算に適した代表値です。

効果量（effect size）も計算してみましょう。効果量とは、介入群と対照群のアウトカムの差を示す指標です。事後テストのスコア（事後テストのスコアが高い群はどちらなのか）、または、進歩（事前と事後のスコアの差）を用いて効果量を計算することができます。ただし、事後テストのデータと進歩のどちらを使用して効果量を計算するのか事前に決めておきましょう。事後テストと遅延テストの両方を使用している場合、効果量を計算してどのように比較するのかも決めておく必要があります。

高度な統計を用いる必要はありません。しかし、使うかどうかは自由です。あなたの評価は小規模で実施される可能性が高いため、統計的有意性の検定（tests of statistical significance）によって結果は偶然であることが示される可能性が高いです。参加者の数が少なければ少ないほど、統計的有意性の検定で、結果は「新しい方法」の効果ではなく、偶然による結果になる可能性が高くなります。

生徒をサブグループ（例：到達度が異なる生徒、性別、経済的困窮の指標等）に分けて効果を分析する場合は、どのような分析をするか、事前に分析方法を決めておきましょう。分析の数が増えれば増えるほど、強いインパクトを示唆する（ように見える）結果に出会う可能性が高くなるため、分析方法を事前に決めておくことがとても大切です（多数の分析からポジティブな結果のみを報告することは、「チェリーピッキング」と呼ばれています）。

・グループをどのように定義するのか？（例えば、高い到達度の生徒のデータを分析したい場合、どのような指標を用いて生徒の到達度を測定するのか？指標で高い到達度を定義する場合、カットオフ値はどれくらいなのか？）

・予定しているデータ分析：各サブグループのどのような指標を分析するのか？平均値や中央値は算出するのか？効果の大きさは計算するのか？効果量を計算する場合、テスト後のスコアと進歩のどちらを比較するのか？

チェックリスト

・どのような代表値を算出するか決めましたか？
・比較する代表値は決まりましたか？
・効果量は計算しますか？
・特定のサブグループの生徒のデータを分析する場合、サブグループの定義や分析方法は決めていますか？
・プロセス評価の定量的・定性的データを分析する方法は決まりましたか？

プロセス評価（process evaluation）

データを分析する方法を決めて下さい。定量的なデータ（quantitative data）（例えば、観察計画のデータ、クローズドクエッションや質問紙の回答）がある場合、回答の平均値（mean）、中央値（median）、最頻値（mode）のいずれを利用することが最も適切なのかを判断する必要があります（最頻値とは、データの中で最もよく出現する値です。生徒が最も楽しんだと答えた「新しい方法」の要素等、数値以外のカテゴリーデータの代表値として使用されます）。

シンプルなオープンクエッションの質的な回答は、カウントすることができます。例えば、「どの本が最も面白かったですか」という質問の回答を分析したい場合、各本の名前が出るたびにカウントすることができるので、どの本が最も人気があるのか（どの本が最も人気がないのか）がわかります。複雑な質問（例：生徒が放課後のセッションに参加しないことを決めた理由は何だと思いますか）の回答等、より複雑な質的データは、主題分析（thematic analysis）を利用します。主題分析では、参加者の回答パターンを調べて、重要なテーマを特定します。

g）タイムライン

プロジェクトの詳細なタイムラインやガントチャート（Gantt chart）を作成して、マイルストーンや責任者等を記載しましょう。チーム、教師、評価に関わる人たちと合意が必要な事項も含めて下さい。

h）予算

計画段階で必要な予算を計算しておきましょう。評価を完了するために必要な資金の総額を把握することができるでしょう。予算計画があれば、適切な時期に資金の調達を検討することができます。

チェックリスト

・プロセス評価の定量的・定性的データを分析する方法は決まりましたか？

5．計画した通りに評価する

計画した通りに評価を実施します。アウトカムの測定や「新しい方法」の実践だけでなく、授業観察等のプロセス評価も忘れないで下さい。

評価期間中は、関係者全員と一貫性のあるコミュニケーションをとって、問題が発生したら直ちに対応してください。また、評価期間中は、予算の変化を注意深くモニターしましょう。

計画に沿ってあらゆる努力をするべきですが、さまざまな理由で意図した通りには進まないかもしれません。それは仕方のないことです。当初の計画から外れたことは忘れずにメモしておきましょう。結果を報告する際、何が変わったのか、なぜ変わったのかを明確に説明してください。

6. データを分析する

データの収集を終えたら、データを分析しましょう。アウトカムやプロセスを分析するために収集したすべての情報を分析します。

a) 測定したアウトカム

計画した通りにデータを分析しましょう。機密保持のため、各生徒の氏名の代わりに ID 番号を割り当てて下さい。氏名と ID 番号の記録は別の場所に安全に保管して下さい。

すべての参加者の完全なデータセットは得られない可能性があります（例えば、テストの日に欠席した参加者がいるかもしれません）。各分析では、すべての参加者のデータセットを対象として下さい。ただし、データが欠落している参加者はデータ分析から除外して下さい。例えば、体育の補習授業が、生徒の「体重」と「800m 走のタイム」に与える効果を評価するケースで考えてみましょう。事前テストで両方を測定して、事後テストでも体重を測定しました。しかし、事後テストで 800m 走のタイムを測定することができませんでした。このような場合、生徒の体重はデータ分析に含めるべきですが、その生徒の 800m 走のタイムはデータ分析に含めないで下さい。各計算に参加者数は必ず報告しましょう。一般的な結果の記述では、参加者数は「n」として報告されます。

「新しい方法」で学習していないと判断した参加者のデータを持っているかもしれません。意図していない場面で「新しい方法」の要素が利用されていることがあるかもしれません。このような場合でも、最初に割り当てられた条件に基づいて、すべての参加者のデータを分析に含める必要があります。脱落させた参加者は、完了した参加者とは異なる特性を持っていた可能性があります。開始前に介入群と対照群が同等であった場合、対照群にも同じような特性を持った参加者がいた可能性があります。しかし、彼らを特定することはできないので、対照群から外されることはありません。したがって、より公平な比較をするために、ITT 分析（intention to treat analysis）に基づいて比較することをお勧めします。

効果の大きさを計算する計算式は以下の通りです。

$$\text{効果量} = \frac{\text{介入群の平均値} - \text{対照群の平均値}}{\text{全体の標準偏差}}$$

事後テストの結果に基づいて効果量（effect size）を計算する場合、（1）介入群の事後テストの結果の平均値、（2）対照群の事後テストの結果の平均値、（3）サンプル全体の事後テストの標準偏差を使用して下さい。進歩（事前と事後のスコアの差）に基づいて効果量を計算する場合は、介入群の進歩の平均値（事後テストのスコアから事前テストのスコアを引いた平均値）、対照群の進歩の平均値、群全体の進歩の標準偏差を使用して下さい。

「新しい教育」が特定のグループの生徒に大きな影響を与えているかどうかを判断するために、計画段階で決めたサブグループ（性別、到達度、無料学校給食の資格等）の結果を分析する場合、計画した通りにデータを分析する必要があります。

b) プロセス評価の分析

プロセス評価で得たデータを計画した通りに分析しましょう。

Box 5：エクセルを使って効果量を計算する方法

1. 各参加者の結果データをスプレッドシートに入力しましょう。介入群と対照群参加者のデータを別々の列に入れます。わかりやすくするために、効果量を計算するデータ以外はシート内に置かないようにしましょう。次に、データを持っていない参加者の行を削除します。
2. 介入群の平均を計算します（Excel では「平均（A）」と表記されています。「数式」の「オート SUM」から選択して下さい）。
3. 対照群の平均を計算します。
4. サンプル全体の標準偏差（standard deviation）を計算します（Excel の「数式」の「その他の関数」にある「統計」から「STDEV.S」を選択して下さい。介入群と対照群の生徒全員の結果データを選択して「OK」をクリックして下さい）。
5. 介入群の平均値から対照群の平均値を引いて、その数値をグループ全体の標準偏差で割ります。算出された数値が効果量（effect size）です。

7. 結果を書く

得られた結果を記述して下さい。このセクションでは、分析の結果を解釈する必要はありません。分析して得られた結果のみ記述しましょう。結果を記述する際には、データの分析結果とプロセスの分析結果の両方を書いて下さい。表、グラフ、その他のインフォグラフィックを使用するとわかりやすく記載することができます。

8. 評価の限界を書く

評価の限界について記述しましょう。結論を出す際には、制約や方法論的な限界を考慮することが重要です。「新しい方法」以外の要因が影響を与えていたかもしれません。これらは内的妥当性（internal validity）の脅威と呼ばれています。例えば・・・

・参加した学校・クラスの一部または全部において、「新しい方法」以外の介入があった。
・「新しい方法」が始まる前の介入群と対照群の類似度が低かった。
・介入群と対照群に提供された教育内容に類似性がなかった。
・一方または両方の群に脱落した参加者がいた。
・対照群が「新しい方法」の要素を受けていた。
・対照群が「新しい方法」で教育を受けていないことに対する違和感や「介入群ではない」という認識を持ったことによって、対照群の行動や対応に変化が生じている。

評価にバイアスを与える要因を考えてみましょう。「新しい方法」を開発・提案した人が、その方法の提供や評価に強く関与していた場合、バイアスが生じやすくなります。バイアスがあると考えられる場合、得られた結果に対して自信が持てなくなります。

調査結果の一般化（外的妥当性）（external validity）についても考慮する必要があります。調査の規模や対象となる学校の特徴を踏まえて、どのような学校で同様の結果が得られるかを考えてみましょう。

9. 結論を書く

評価の限界を理解した上で分析結果から結論を導き出して下さい。結論は、あなたが最初に作ったリサーチクエッションに対する回答です。結果は先行研究と似ているのか、または違うのか、なぜこのような結果になったのかを考えてみましょう。

Box 6：結果の解釈と結論の出し方

効果量（effect size）は、介入群と対照群のアウトカムを比較する方法の 1 つです。アウトカムは、事後のスコアや進歩（事前と事後のスコアの差）が利用されます。2 つの群の結果の差を群全体の標準偏差と関連づけて、差の大きさを計算します。プラスの効果量は、介入群が対照群よりも良い結果をもたらしたことを示しており、マイナスの効果量は、対照群が介入群よりも良い結果をもたらしたことを示します。効果量が大きいほど、2 つの群の結果の差は大きいことを意味します。

効果量の「良さ」を解釈する共通のルールはありません。効果量を評価するカテゴリーはいくつか存在します（例えば、0.2 は小さい、0.5 は中程度、0.8 以上は大きい等）。しかし、効果量の解釈は単純ではありません。効果量は、サンプルの均質性、測定されるアウトカム、テストの信頼性等、評価に関する特性の影響を受けます。また、効果の大きさが実際にどのような意味を持つのかは、コストや潜在的なスケーラビリティ等、「新しい方法」の特性によって異なります。

結論を出す際には、アウトカム評価とプロセス評価から収集したすべてのデータを使用して下さい。両群が達成した進歩（事前と事後のスコアの差）を理解するために、効果の大きさ、事前テストと事後テストの平均値、進歩（事前と事後のスコアの差）の平均を見てください。

表やグラフを使うと、結果を簡単に把握することができます。研究の限界、評価が行われた状況の理解、プロセス評価の一部として収集された参加者の意見と合わせて結論を検討しましょう。そして、同じような環境で実施した場合の効果の可能性について結論を出してください。

効果量について詳しく知りたい方は、教育介入の効果量の解釈に関する論文（Kraft, 2019）を最初に読むと良いでしょう。

10. 次のステップを決める

評価してわかったことを利用する方法を考えましょう。他の人たち（他校の教師や研究者を含む）が知っておくべきことは何でしょうか。どのように他の人たちと結果を共有するかを決めてください。これは、評価の実施から学んだことを振り返る良い機会です。

a) 実践から示唆を得る

・今回の結果を受けて、あなたの教室では何を変えますか？
・今回の結果を受けて、あなたの学校では何を変えますか？
・今回の結果を受けて、他校の教師やスクールリーダーに提言したいことは何ですか？

アウトカム評価とプロセス評価の結果から、再び実施する前に取り組み方を変更した方が良いと結論づけるかもしれません。大幅な変更を予定している場合は、更新された教育方法が効果的に機能すると確信する前に、改めて評価する必要があります。

b) 評価から示唆されること

結果を分析している途中、リサーチクエッションの範囲を超えた多くの疑問が浮ぶかもしれません。疑問は、学校関係者のコメントや、参加者の観察データ、データの中で気づいたパターンに関連しているかもしれません。また、異なる状況、対象、年齢層において、「新しい方法」がどのように機能するかという疑問も出てくるでしょう。これらは、別の研究プロジェクトで調査する予定にしましょう。現在は調査することができない深いリサーチクエッションかもしれません。以下の 2 つを考えてみましょう。

・新規に実施してみたい評価内容
・この分野の研究者に対して提言したいこと

c) わかったことを発信する

他の学校や教師も恩恵を受けることができるように、評価の結果は共有しましょう。想定した結果ではなかった場合や、評価のすべての要素を完了していない場合でも、結果を共有することは重要です。ネガティブな結果の共有は他の学校や教師を助けることにつながります。例えば、成功しない方法を試そうとしたり、よく似た方法（すでに評価されている方法）を評価して時間を無駄にしたりする可能性を減らすことができます。プロセス評価の結果、教師が「新しい方法」を好まなかったために実施を中止していた場合、その方法に関心を持つ人たちは、継続の可能性を高めるために「新しい方法」の内容や計画を工夫して改善するかもしれません。また、「新しい方法」が参加者に予期せぬ悪影響を与えたことがわかって評価を中止していた場合、他の学校や教師が同じ様に悪影響を与えなくて済むので、結果を共有することはとても大切です。

報告書を書いて公開することをお勧めします。報告書は、学校のウェブサイトや、評価に協力してくれたパートナーのウェブサイトで公開できるかもしれません。TES のような教育関係の出版社は、教師が行った実験の結果に興味・関心を持っていますので、あなたの調査結果を出版したいかどうか、アプローチしてみましょう。評価報告書の書き方は、「wiring up your evaluation report」を参照してください。また、会議やイベント、地域のネットワークミーティングで評価結果を発表・共有することもできます。

d) リフレクション

評価を終えた後は、プロセスを振り返ってみましょう。何が成功に役立ったのか、何を学んだのか、今後の評価では何を変更して実践すると良いのか等を考えてみてください。次のような質問を考えてみてください。

評価を実施する

・評価でうまくいったことは何ですか？
・どのような課題に直面しましたか？
・当初の予算は現実的でしたか？
・想定していたよりも多い（少ない）費用でしたか？
・当初のスケジュールは守られましたか？
・評価に予想以上の時間がかかりましたか？（かかりませんでしたか？）
・今後の評価で、より多くの時間を割く必要がある部分はありますか？
・今後の評価では何を変えると良いですか？
・どのようなサポートが役に立ちましたか？
・今後も同じサポートを受けることはできますか？

実践する

・あなたは「新しい方法」を今後も教室で使いますか？使うとしたら、誰と一緒に使いますか？
・今後、「新しい方法」の影響をどのようにモニターしますか？

次のステップ

・今回の評価の結果を受けて、今後は何をする予定ですか？
・教育実践を新規に評価するために必要な資金を申請しますか？
・「新しい方法」を深く評価することに興味がありますか？
・「新しい方法」の評価を実施した経験は、あなたの実践に何か変化をもたらしましたか？

リファレンス

References British Educational Research Association, Ethical Guidelines for Educational Research
https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018

Community Tool Box, Logic Model resource
https://ctb.ku.edu/en/table-of-contents/overview/models-for-community-health-and-development/logic-model-development/main

Education Endowment Foundation, DIY Evaluation Guide
https://educationendowmentfoundation.org.uk/tools/diy-guide.Education

Endowment Foundation, Putting evidence to work: a school's guide to implementation
https://educationendowmentfoundation.org.uk/public/files/Publications/Implementation/EEF_Implementation_Guidance_Report_2019.pdf.Education

Endowment Foundation, Spectrum database
https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/measuring-essential-skills/spectrum-database

Institute for Effective Education, Engaging with Evidence Guide
https://the-iee.org.uk/what-we-do/engaging-with-evidence

Institute for Effective Education, Unplanned Analyses, Data Dredging and Cherry Picking
https://the-iee.org.uk/2020/10/15/unplanned-analyses-data-dredging-and-cherry-picking

Institute for Effective Education, Writing up your innovation evaluation report guidance
https://the-iee.org.uk/wp-content/uploads/2017/03/Writing-up-your-innovation-evaluation-report.pdf

Kraft, Interpreting Effect Sizes in Educational Interventions
https://scholar.harvard.edu/mkraft/publications/interpreting-effect-sizes-education-interventions

Research Methods Knowledge Base, External Validity
https://conjointly.com/kb/external-validity

Research Methods Knowledge Base, Internal Validity
https://conjointly.com/kb/internal-validity

The Center for Theory of Change
https://www.theoryofchange.org/what-is-theory-of-change

エビデンスが掲載されているサイト

Association for Science Education
https://www.ase.org.uk

Campbell Collaboration
https://campbellcollaboration.org

Chartered College of Teaching
https://chartered.college

Deans for Impact
https://deansforimpact.org

Early Intervention Foundation
www.eif.org.uk

Education Endowment Foundation Early Years Toolkit
https://educationendowmentfoundation.org.uk/evidence-summaries/early-years-toolkit

Education Endowment Foundation Guidance Reports
https://educationendowmentfoundation.org.uk/resources/guidance-reports

Education Endowment Foundation Teaching and Learning Toolkit
https://educationendowmentfoundation.org.uk/resources/teaching-learning-toolkit

EPPI-Centre
https://eppi.ioe.ac.uk/cms

Institute for Education Sciences
https://ies.ed.gov

Institute for Effective Education Best Evidence Encyclopaedia
www.bestevidence.org.uk

Institute for Effective Education Best Evidence in Brief
www.beib.org.uk

Institute for Effective Education Evidence 4 Impact
www.evidence4impact.org.uk

National Centre for Excellence in the Teaching of Mathematics
https://www.ncetm.org.uk

National Foundation for Educational Research
https://www.nfer.ac.uk

The Nuffield Foundation
https://www.nuffieldfoundation.org

Research Schools Network
https://researchschool.org.uk

Teacher Development Trust
https://tdtrust.org

The Learning Scientists
www.learningscientists.org

Wellcome
https://wellcome.org

What Works Clearinghouse
https://ies.ed.gov/ncee/wwc

用語集

条件に割り当てる（Assignment to condition）
参加者を異なる条件に割り当てる方法のこと。

条件（Condition）
参加者のグループのこと（例：介入群または対照群）。

対照群の汚染（Contamination of the control group）（diffusion of treatment）
対照群が「新しい方法」の一部またはすべての要素を受けること。一般的に、「新しい方法」の効果について結論を出すことができなくなることを意味しています。

対照群（Control group）（Comparison group）
「新しい方法」で教育を受けない参加者のグループ。

アクティブコントロール群（Active control group）
介入群と異なる方法で介入を受ける対照群のグループ。

いつも通りに過ごす対照群（Business as usual control group）
学校や教師がいつも通りに教育・支援するグループ。学校教育に関するインパクト評価で使用される典型的な対照群です（何も教育を受けない対照群やプラセボ対照群は、教育評価では稀です）。

待機群（Waiting list control group）
採用基準を満たしている参加者全員に「新しい方法」を同時に提供することができないケースがあります（例えば、スタッフが少なかったり、学校に十分な部屋がなかったりします）。同時に「新しい方法」を提供することができない場合は、参加者を 2 つの群に分けて実施します。2 番目の介入群に選ばれた参加者は、「新しい方法」の提供を待っている対照群です。1 番目の介入群で効果が認められたら、待機群の参加者にも「新しい方法」で教育が提供されます。

効果量（effect size）

「新しい方法」を経験した参加者（介入群）と経験しなかった参加者（対照群）の結果の違いを示す指標です。一般的な教育研究では、「−1.0」から「＋1.0」の間の数値が示されます。効果量の値がマイナスを示した場合、介入群よりも対照群の方が良い結果だったことを意味しています。効果量の値がプラスを示した場合、対照群よりも介入群の方が良い結果だったことを意味しています。効果量が大きいほど、2 つの群間の差は大きくなります。

外的妥当性（External validity）

研究結果を他の状況や集団に適用できる範囲のこと。外的妥当性の脅威や外的妥当性を改善する方法の詳細はコチラをご覧ください。

フォーカスグループ（Focus group）

インタビューを一緒に受ける参加者のグループを意味しています。フォーカスグループの参加者は、インタビュアーの質問に答えるために参加者間で交流します。一緒にインタビューを受けて個別に回答するグループインタビューの方法とは異なる方法です。

ガントチャート（Gantt chart）

特定の期間に完了する課題を横向きの棒グラフで表した視覚的なタイムライン。

仮説（Hypothesis）

既にわかっていることに基づいて立てる予測。正しいかどうかを検証することができる予測。

インパクト評価（Impact evaluation）

アウトカムとプロセスの評価を同時に実施して「新しい方法」の効果を検討する実験の 1 つ。

実施の忠実度（Implementation fidelity）

「新しい方法」が意図した通りに実施されたかどうかを意味しています。

採用基準（Inclusion criteria）

評価対象として参加者が備えておく必要がある特性。到達度、人口統計学的特性、過去の経験等が含まれます。特定のグループ（例：事前の到達度が低い生徒や出席率の低い生徒等）を想定して「新しい方法」が評価される場合、想定しているグループに該当する項目が参加者の採用基準に含まれます。

除外基準（Exclusion criteria）

参加希望者が持っていてはいけない特性。到達度、人口統計学的特性、行動特性だけでなく、過去に「新しい方法」で教育を受けた経験や、評価期間中に別の群に参加したこと等が含まれます。

ITT 分析（Intention to treat analysis）

参加者のデータは、すべての活動に参加したかどうかにかかわらず、最初に割り当てられた条件で分析されます。

内的妥当性（Internal validity）

「新しい方法」の効果に対する確信の程度。評価の内的妥当性には多くの脅威があります。詳細はコチラをご覧ください。

介入群（Intervention group）

「新しい方法」で教育を受ける参加者のグループ。

インタビュー（Interview）

情報を引き出すために、インタビュアーが質問をすること。

平均値（Mean）

データを足し合わせて数値の個数で割って求められる値。スコア間の距離が一定のデータ（間隔尺度）の代表値を求める場合に使用されます（例：身長、スピード、テストのスコア）。

中央値（Median）

データを大きい順（小さい順）に並べた時に中央に位置する値。中央値は、スコア間の距離が一定ではないデータの平均を求める場合に使用されます（例：1～10 のスケールで評価、生徒のテスト成績の順位）。

最頻値（Mode）

データの中で最もよく出現する値。数値以外のカテゴリーデータ（例：お気に入りの本）の代表値として使用される。

アウトカム評価（Outcome evaluation）

「新しい方法」が生徒の到達度やその他のアウトカムに影響を与えているかどうかを検討する研究の１つ。

アウトカム指標（Outcome measure）

パフォーマンスを測定するために使用するテスト。

参加者（Participant）

評価に参加する人。

母集団（Population）

標本（Sample）を構成できる、すべての人（採用基準を満たしているすべての人）。

事後テスト（Post-test）

参加者が「新しい方法」を完了した後に実施するアウトカムの測定。

事前テスト（Pre-test）

参加者が「新しい方法」を始める前に実施するアウトカムの測定。

遅延テスト（Delayed post-test）

参加者が「新しい方法」を完了してから数週間後（数ヶ月後）に実施するアウトカムの測定。

プロセス評価（Process evaluation）

「新しい方法」が計画していた通りに実施されたかどうか（実施の忠実度）を評価すること。「新しい方法」の実施状況の観察、インタビュー、質問紙等があります。

質的データ（Qualitative data）

非数値のデータ。一般的に質的データは言葉が利用されます（例：オープンクエッションに対する回答、観察のナラティブなレポート等）。

質的研究（Qualitative research）

ナラティブな解釈で結果が表現される研究。

準実験（Quasi-experiment）

参加者が介入群と対照群にランダムに割り当てられていない実験。

量的データ（Quantitative data）

統計的手法を用いて分析される数値形式のデータ。

量的研究（Quantitative research）

数値形式のデータを用いて調査結果を報告する研究。

質問紙（Questionnaire）

参加者に回答を求める質問項目をまとめて記したもの。

無作為割り当て（Random assignment）

参加者をランダムに介入群または対照群に割り当てる方法。

ランダム比較試験（Randomized controlled trial）

研究参加者を介入群と対照群にランダムに割り当てる研究デザイン。2 つのグループが最初から同等である可能性が高いため、群間の結果の違いは「新しい方法」の効果である可能性が高くなる。

信頼性（Reliability）

結果（各測定値）は安定しているか（一貫性があるか）どうかを調べるために、同じ標本から異なる機会にデータを採取しても同じ結果が得られるか検討する。

標本（Sample）

評価に参加する人たちのグループ。標本の各個人は参加者（Participant）と呼ばれる。

標準化された指標（Standardized measures）

標準化された方法で実施・採点する測定法。膨大な人々のデータと比較することができる。標準化されたテストは、様々なチェックを経て、各出版社から販売されている。

標準化されたテストの等価版（Equivalent versions of standardized measures）

標準化されたテストの中には、同じスキルを異なる問題で測定できるテストがあります。異なる時期に実施することができるので、事前テストと事後テストに同等のテストとして実施することができます。

標準偏差（Standard deviation）

スコアのばらつきを示す指標です。数値が小さいほど、スコア間のばらつきが少ないことを意味しています。

統計的有意性（Statistical significance）

統計的有意性の検定に基づいて、結果が偶然に発生した可能性を見極める考え方の総称。

構造化された観察（Structured observation）

事前に決めたカテゴリーに基づいて行動や出来事を観察すること。

観察計画（Observation schedule）

構造化された観察をする際に観察者が記入する記録用紙。

調査（Survey）

質問して参加者の意見・経験を調べる方法。インタビューや質問紙に答えてもらう方法等がある。

統計的有意性の検定（Test of statistical significance）

結果が偶然に得られた可能性が高いかどうかを判断する。統計的検定では p 値で結果が示される。p 値は 0 から 1 の間で示される。数値が小さいほど、偶然に生じた可能性が低いことを示す。p 値が 0.05 より小さい場合、結果は統計的に有意であるとして、偶然に起因する可能性は低いと判断されます。

主題分析（Thematic analysis）

質的データの分析方法の 1 つ。質的データのパターンを特定・分析・記録します。

変化の理論（Theory of Change）

「新しい方法」がアウトカムに与える効果の因果関係のアウトライン。「新しい方法」がアウトカムの変化につながるプロセスを説明するモデル。

よく類似している（Well-matched）

「新しい方法」の評価を始める前に、介入群と対照群がどれくらい類似しているかを検討する時に用いられる用語。一般的に、2 つの群の標準偏差がどちらも 0.25 以下であれば、2 群は類似していると判断されます。群がよく類似している場合、同等（equivalent）という用語が用いられることもあります。

Appendix A: 評価計画（Evaluation planning）

1．問題・課題：解決したい問題・課題を記述してください
・あなたの学校では、どのような問題・課題を解決する必要がありますか？ ・1 つの問題・課題を明確に特定して、その問題・課題を引き起こしている原因や問題を持続させている要因について仮説を立てましょう。

2.既存のエビデンス：現在どのようなエビデンスがありますか？
・先行研究は、学校が抱えている問題・課題や解決策について何を言っていますか？ ・問題・課題に対処するために、どのような解決策を使用しますか？ ・変化の理論（Theory of change）を概説してください。

3．研究課題・仮説
[どれくらいの期間] で [どのような方法] を提供したら [誰] の [どのようなアウトカム] にどのような効果を与えますか？

4．方法：評価する方法を具体的に記述しましょう。
（a）標本
・誰が評価に参加しますか？ ・満たさなければならない採用基準は何ですか？ ・どのような方法で参加者や他校の参加を募りますか？ ・参加者数はどれくらいになりそうですか？ ・参加者からどのような同意を得る必要がありますか？

（b）条件の割り付け
・参加者を介入群と対照群にどのように割り当てますか？ ・「新しい方法」の評価を始める前に、介入群と対照群が類似していることをどのようにして確認しますか？

(c)「新しい方法」
・「新しい方法」を詳しく説明してください。 ・「新しい方法」で教育する人たちにどのようなトレーニングと支援が与えられますか？ ・どのようなリソースを作成する必要がありますか？ ・トレーニングや支援はいつ実施されますか？ ・対照群は何をしますか？ ・対照群が「新しい方法」の要素を取り入れていないことをどのように確認しますか？

(d) アウトカムの測定
・どのような方法でアウトカムを測定しますか？ ・いつ、どのようにアウトカムを測定・採点しますか？

(e) プロセス分析
・実装の忠実度はいつ・どのように確認しますか？ ・誰が忠実度を調べますか？ ・対照群が「新しい方法」の要素を使用していないこともチェックしますか？ ・どのような参加者の意見を集めますか？ ・意見の収集はいつ・どのように実施しますか？ ・参加者の意見を収集する責任者は誰ですか？

(f) 分析
・アウトカムのデータをどのように分析しますか？ ・どのような分析を行いますか？ ・参加者のサブグループのデータを分析しますか？ ・サブグループのデータをどのように分析しますか？ ・サブグループをどのように特定・定義しますか？ ・データを分析する方法を明記してください。 ・プロセス評価の量的・質的データをどのように分析しますか？

(g) タイムライン

以下のような表を用いて、マイルストーンや責任者等、プロジェクト計画の詳細なタイムラインを作って下さい。また、ガントチャート等の別のフォーマットでタイムラインを作成することもできます。

活動	開始日と最終日	活動の責任者

(h) 予算

以下のような表（予算と予算が必要な時期を記載した表）を作成してください。

予算項目	予算が必要な日	金額
		合計：

Appendix B: 調査作成のガイドライン（Guidelines on writing a survey）

計画書に書く必要がある事項を紹介します。

目的（Purpose）

・今回の評価で知りたいこと、知りたい理由を明確に記述してください。次に、評価がリサーチクエッションの回答に役立つかどうかを確認して下さい。プロセス評価をするために質問紙を作成する場合は、質問は具体的に書いて下さい。
・データ分析の方法を検討して、分析方法に合う適切なデータが収集できるようにしましょう。
・リサーチクエッションを立てた理由と、今回の評価が何に貢献するのかを明確に書いて下さい。

構成（Structure）

・質問紙の最初に簡潔で明確な説明を書いて下さい。質問の種類が変わる時は、必要に応じて改めて説明して下さい。混乱を招く可能性がある場合は、例を示すと良いでしょう。
・質問の順序
  –つながりのある順番で質問を整理して、類似した質問をまとめて下さい。
  –簡単で難しくない質問から始めると良いでしょう。
  –長い質問の最後に最も重要な質問を置かないでください。
  –前の質問が後の質問の回答に影響する可能性を検討して下さい（質問順序効果）。
・どのようなタイプの質問をするのか決定しましょう（例：オープンな質問、クローズドな質問、またはその 2 つを組み合わせた質問）。オープンな質問はより詳細な情報を提供しますが、クローズドな質問は早く回答が得られるため、コード化や分析が容易です。
・複数の回答を選択するクローズドクエッションを使用する場合は、以下の方法が有効です。
  –妥当性が高い回答を得るために、最初に試験グループ（テスト回答してくれる人たち）に対してオープンクエスチョン形式の質問をしてみましょう。試験グループは、評価に参加していないが、参加者の特性が類似している人たちである必要があります（例：同じ学年の生徒、同じ科目の教師）。
  –選択式の質問において、回答の選択肢を提示する順番が回答に影響を与える可能性を検討して下さい(回答順効果)。

質問を書く（Writing questions）

・言葉の選び方・使い方（Wording）

–回答者の意図した通りに解釈されるようにしましょう。人によって異なる意味を持つ可能性がある用語は明確に定義して下さい。

–専門用語の使用は避けて下さい。

–頭字語は、最初に使う時に書き出すか、完全な形で言いましょう。

–できるだけ短い質問になるように工夫しましょう。

–否定的な質問は避けて下さい。

–1 つの質問で 1 つのことを尋ねて下さい。

・対象者の認識（Awareness of audience）

–質問紙を使用する場合は、回答する生徒の読解力を考慮した質問を書きましょう。

–回答者の知識の量を認識した上で質問を書いて下さい(特に学校関係者ではない場合、例えば保護者の意見を収集する場合)。

・行動について質問する場合は、質問が曖昧になるので、「いつもの」という言葉は使わないようにしましょう。行動に関する質問には、期間を含めると便利です（例:「昨年度…を教える時にどのようなアプローチをしましたか」）。

・尺度を使用する場合、尺度は 5 件法以下にすると良いでしょう。5 件法を超える尺度を使用する必要がある場合は、選択肢が提示された視覚的な情報を添えてください。6 件法を超える尺度を用いた場合、生徒は回答に困る傾向があります）。

・可能な限り、参加者が回答に困る質問は避けて下さい。

–質問紙の目的を明確にしましょう。

–質問紙ではオープンな質問ができるので、インタビューでは回答しにくいことを尋ねたい時に有効です。

事前調査（Pilot the survey）

・可能であれば、評価に参加しない数名の生徒に試験的に回答してもらいましょう。

・不明確だと感じた項目、追加すべきだと感じた質問、複数の選択肢の中に欠けているものがないか、複数の選択肢が重なっていないか、紛らわしい質問になっていないか、立ち入り過ぎている質問や不適切な質問等がないか、テスト回答してくれた人からフィードバックを求めてください。

・参加者を代表するグループでパイロットテストを行うのが現実的でない場合は、同僚や家族、友人に質問紙を見てもらって、言葉遣いや意味が明確かどうかを確認してもらいましょう。

リファレンス

Mertens M D (2005). Research and evaluation in education and psychology. London: Sage.

Robson C and McCarthan K (2016). Real world research. Chichester: John Wiley & Sons.