訳　小　林　麻衣子（常磐大学）

NONEXPERIMENTAL REPLICATIONS OF SOCIAL EXPERIMENTS IN EDUCATION, TRAINING, AND EMPLOYMENT SERVICES [REVISED PROTOCOL]
教育、訓練、雇用事業に関する社会実験の非実験的手法による再現（改訂版プロトコル）

2002 年 12 月

レビューワ：Steven Glazerman, David Myers, Dan Levy

主任レビューワの連絡先
Mathematica Policy Research, Inc.
600 Maryland Avenue SW, Suite 500
Washington, DC 20024
202-484-4834phone;202-863-1763fax
sglazerman@mathematica-mpr.com

本レビューは、スミスリチャードソン財団、ウィリアム・フローラヒューレット財団の助成による支援を受けている。

Ⅰ 背景と目的

統制実験では、対象が無作為に介入を受けるように割り当てられるが、これは望ましいものである反面、とりわけ社会的背景の中では実行不可能であり、またあまりにも負担が大きいことがしばしばある。したがって、非実験的（また、準実験的とも呼ばれる）方法が、代用としてよく利用される。準実験的方法は、概して押しつけが少なく、またコストも時折低くなるが[1]、よりはっきりした仮説を必要とする。異なる準実験的方法は、異なる仮定を要する。そうした仮定は重要であるが、通常仮定をチェックすることができない。予定されているレビューでは、綿密に計画・実行された実験研究から得られる結果に、準実験的方法が最も近づく条件を調べる。これは、便宜的標本と 1 つかそれ以上の準実験的方法を用いた社会実験から得た調査結果の反復を試みた一連の研究を収集し、要約することで行われる。

A. 本研究の指針となる論題

レビュー計画では、以下の論題に取り組む：

・ 準実験的な方法は、綿密に計画・実行された無作為化試験から得られるような結果に近似することができるか？

・非実験的な方法がバイアスのない効果推定値を生み出す条件とは何か？

・単一の非実験的効果推定値のバイアスは、複数の実験的効果推定値の集積を通して相殺できるだろうか、あるいは補正できるだろうか（例えば、研究、研究場所、あるいは方法に渡って）。

我々は、一般的な調査方法論に関する以下の実際問題に取り組むため、これらの実証的論題に答える必要がある。

・ システマティックレビューではどのように準実験的証拠を扱わなければならないか？

・ どのような研究デザインと方法が将来の研究では用いられるべきか？

B. レビューされている分野の状況

[1] 例えば、準実験的研究は時として、行政記録を使って後ろ向きに行うことができるが、実験的研究はたとえ行政記録を使うとしても、前向きで行われなければならない。

レビューされているトピックについては、十分な理論的議論がなされている。多くの研究報告で、1 つのアプローチまたは別のアプローチの利点と不利点や、いつ実験的あるいは非実験的方法が要求されるかについて議論されている。例えば、Burtless（1995）、Heckman と Smith（1995）らによる Journal of Economic Perspectives での討論会を参照してもらいたい。他は、Cook と Cambell（1979）にさかのぼり、推定量や研究デザインごとの統計的特性、仮説または「妥当性への脅威」の目録を作成している。ここでは、推定量は分析の重要な単位である。なぜならば、各推定量は観察値を用いてプログラムの効果の推定値を構築するのに異なる原則に拠っているためである。本システマティックレビューの目的は、さまざまな状況下におけるさまざまな推定量の性質について知ることである。

実証的根拠は豊富ではないが、システマティックレビューが大いに価値を付加するであろうレベルにまで集積し始めている。それは現状の知識を評価し、そして、結果で得た新たな差異を調整し、またこの分野の今後の活動に対する共通的枠組みを提供することによってなし得るものである。レビューの詳細な論理的根拠と動機は、Shadish（2000）により提供された。レビューされている分野でよく引き合いに出される最初の試みは、Lalonde（1986）の政府後援職業訓練の実証研究である。Lalonde のデータは、Dehejia と Wahba（1999）、そして、Smith と Todd(2002)によって、ここ数年の間に 2 回再分析されている。さらに多くの研究では互いに異なる非実験的推定値を比較しているが（例：Heckman と Hotz 1989）、しかし、実験的な推定値を含んで比較を行っているものはまれである。ここ数年は勢いを持ち直しているものの、それ以後は、Friedlander と Robins（1995）、Bell ら（1995）による実験的デザイン反復研究が 2，3 あるのみである。（参照：Heckman et al. 1998; Olsen and Decker 2000; Agodini and Dynarski 2001）追加研究が進行中であり、本レビューに含めるのに間に合う終了が期待される。

我々が知る限りでは、この分野のシステマティックレビューは現在まで行われていない。これらの研究の歴史を考察する試みもあるが（Shadish 2000）、しかし我々の目的である、教訓を得るために研究を統合しているものはない。本レビューの構成研究の殆どすべては、先行文献についての要約を含んでいる。我々が知る限りでは、最も包括的なサマリーは 1 番最近のもので(Bloom et al. 2002)、本文献の重要箇所、すなわち義務福祉制度を扱っている部分について検討を行っている。

我々がレビューを行った研究は研究内比較として、Shadish によって参照されている。というのも、そういった研究は非実験的方法を使って、実験的結果を再現しようと試みているためである。いくつかの*研究間*比較では、大量の一次的実証研究文献を要約するのにメタアナリシスが使われ、またエフェクトサイズ推定値の差異を説明する要因として（実験的、半実験的な）デザインを用いている。しかし、それぞれの構成要素となる研究は、単一のデザインを使っている。例については Cooper ほか（2000）を参照のこと。また、こうしたメタアナリシスに対するメタアナリシスが、我々がここで提示したものと類似した論点に答えるものとして行われている。（実験的あるいは非実験的な）推定量は違いを生むだろうか？例については、Shadish と Ragsdale（1996）を参照のこと。これらのサイト間研究（別のデータでの研究）では、方法間の結果の差異が、研究者や介入、あるいは研究

内容によるものではなく方法それ自体によるものであるということを、レビュー案に含まれるサイト内研究（同じデータに対する研究）ほど強く証明できない。

II. 本レビューに含む研究の検討基準

A. 研究の種類

本レビューの結果変数は、単一の非実験的推定量のバイアス推定値である。レビューに採用する主な基準として、各研究はプログラム効果の実験推定値と、同様に少なくとも 1 つの非実験効果推定値を含んでいなければならない。あるいは同様に、少なくとも 1 つの非実験効果推定量のバイアス推定値を含まなければならない。バイアスは、2 つの効果間の差として、または無作為化対照群と何かの代替となる比較群間の修正された平均成果における差として概算することが可能である（第 4 節のフレームワークを参照のこと）。

いろいろな設定での実験的・準実験的な効果推定値を比較するメタアナリシスも、それが同じ設定（介入、期間、地理的位置を含む）で異なるアプローチ法を用いて研究比較をしていなければ、本レビューの範囲外となる。また、異なる非実験的推定値を互いに比較していても、実験による基準をもたない研究は含まれない。

B. 介入のタイプ

レビューに焦点を若干与えるため、我々は教育や雇用に関連した結果改善に向けた介入効果に関する研究に注意を限定した。これにより、興味深いと思われる研究、つまり医療結果についての介入効果を測定している研究が除外され、その中には無作為化・作為化臨床試験が含まれる。そういった研究が多くあるという点で、それらは別のシステマティックレビュー対象になるべきである。健康介入に焦点を当てたこのジャンルの研究を除外する 2 つの追加的理由は、(1)サンプル選択手順（非実験的方法がモデル化しようとするもの）が、教育・訓練介入に比べ医療介入ではかなり異なること。そして、(2)二分野の証拠（エビデンス）の基準が必ずしも同じものである必要がない。なぜなら損失関数がかなり異なると思われるためである。換言すれば、健康介入に関する証拠は、より頻繁に有害な治療結果を発見し、回避しようとしているが、対して職業訓練プログラムの成功についての根拠は、帰無仮説の間違った棄却に対してペナルティが低い。そうした場合の消極的結果は、政府が無効なプログラムに資金を使っているということであり、そしておそらく、訓練を受けるものが時間を無駄にしているということだろう。[2]

[2] 健康政策評価と労働政策評価の違いの例としては、医療介入の臨床試験では研究者が介入群のメンバーによって告訴されるのを恐れるという報告がある。反対に、社会福祉事業プログラムの無作為化実証研究では、評価者は対照群のメンバーによって告訴されるのを恐れるだろう。

C. 効果指標のタイプ

こうした研究群では、5 つの広義的な効果指標が見つけられよう。

1. 試験得点

2. 所得

3. 職業

4. 学校参加（出席、中退、学業成績）

5. 生活保護受給

最初の 2 つ（試験得点と所得）は、概して連続変数である。後者の 2 つ（仕事と学校出席や成績）は、概して 2 項変数である。場合によっては、所得と職業の結果は、いくつかの時間点で計測される。

III. 研究特定のための検索方法

この主題に該当する研究を特定するのは挑戦的な仕事である。我々が称するように、デザイン再現研究を記述する共通語は存在しない。我々は見つかる望みのないものをみつけようとしているー大量の研究の中で、基準を満たすようなものは殆どないと思われ、また表題あるいは抄録が、その研究が本レビューに適合するかどうか必ずしも示唆するとは限らない。

それにもかかわらず、電子検索は ERIC、Econlit、CIS、C2-SPECTR、policyfile.com や、統計資料あるいは他の政策関連資料を索引した追加データベースを通して行われる予定である。手動検索は、評価調査会社と政府機関の刊行リストを使って行われる。追加検索は NBER 監査調書と他の特定経済・公共政策機関（Mathematica Policy Research、Rand、Abt Associates、MDRC、Urban Institute、the World Bank）や政府機関(DoL, ED, HHS)の監査調書を利用して行われる予定である。進行中の研究をもりこむため、主な社会科学と公共政策研究学会（APPAM、ASSA、AERA、AEA、AEFA、ASA、APSA、PAA）[3]の

---

[3] AEA = American Evaluation Association; AEFA = American Education Finance Association; ASSA = Allied Social Science Association; ASA = American Sociological Association; APSA = American Political

最新プログラムを盛り込む。これら全ての検索戦略には、熟練した研究者による検索を通して慎重な組み合わせが必要である。我々はまた、研究総合事業の説明と我々が既に確認した幾つかの研究リストと併せて、特定の研究者へ手紙（または E メール）を郵送する予定である。

我々は、公表結果と進行中の評価をともに含む既知の研究リストから取りかかる。また、これらの論文の参照リストやデータベースに含まれるいかなるものも運用する予定である。

A. 研究方法

各情報源内の検索戦略は、以下に挙げる 3 つの各方法に従うものとする：

1. 介入による検索　　実験的評価の対象となっている介入を特定すること。それから、非実験的評価が同様に行われているかどうか調べること。

介入による検索では、1 つ以上のデザイン戦略を用いた下記のものが研究されていることが分かっている：

・政府後援職業訓練計画

・職業生活保護

・職業団体

・中途退学実証研究

・学級規模の減少

2. メタアナリシス間の検索　　関心領域にある介入タイプに制限し、同じ介入が 1 つ以上の評価法で調査されているようなメタアナリシスは全て検索する。*同一設定内での再現法に着目を制限する。*

3. 方法論的文献の検索　　評価戦略（推定量）を比較するのに実証的根拠を用いている方法論的文献については、経済学、心理学、社会学、教育学、統計学雑誌を参考とする。政府機関に対して実験的評価を行っている企業の公表データベースと、政府機関そのもののデータベース（可能であればウェブサイト上にあるもの）を検索する。

Science Association; PAA = Population Association of America.

B. 検索語

検索語の定義は、特に挑戦的な仕事である。本リストは、以下の用語を含む：準実験的;非実験的;無作為割付;デザイン AND 比較;代替法 AND デザイン;評価方法;デザイン AND 再現;実験的 AND 再現

IV. レビュー方法

A. フレームワーク

このレビューでの目的と方法は、形式的記号を使って明確に示すことができる。θ は関心のあるパラメータ（たとえば処遇群への介入の真の効果）を表すこととする。システマティックレビューの目的は、異なる非実験的推定量 θ に関連するバイアスを調べ、異なる状況下でバイアスがどのように変化するかを理解することである。真の効果が未知のため、バイアスは決して直接観察することができない、しかし、本レビューでは、実験的にバイアスを推定できる 2 種の研究を包含している。第 1 種の研究群には、K 個までの非実験推定値、つまり関心パラメータである $\theta_{hat\,k}$（k＝1,...K）と、$E[\theta_{hat0}]=\theta$ となるような 1 実験推定値 $\theta_{hat\,0}$ まで提示しているものが含まれる。第 2 種の研究は、対照群の平均成果 $\bar{Y}_0$ を、非実験的方法 k に基づいたいくつかの総合対照群の平均成果 $\bar{Y}_k$ と比較している。$\bar{Y}_k$ はマッチングされた比較群であるか、または処遇を受けなかった対象の便宜的標本に対する回帰修正平均結果であることが多い。これら変数間の関係は方程式 (1.1) と (1.2) で示される。ここで $\bar{Y}_T$ は治療群の平均結果を示し、$B(\theta_{hat\,k})$ はバイアスである。

$$\bar{Y}_T - \bar{Y}_k = \theta_{hat\,k} \qquad (1.1)$$

$$\bar{Y}_T - \bar{Y}_0 = \theta_{hat\,0} \qquad (1.2)$$

方程式 (1.2) を方程式 (1.1) から差し引くと、2 形態のバイアス推定値が生じる。そして、上記で議論された 2 種の報告形式と一致する：

$$\bar{Y}_0 - \bar{Y}_k = \theta_{hat\,k} - \theta_{hat\,0} = B\,hat(\theta_{hat\,k}) \qquad (1.3)$$

このように、後者のタイプは処遇群の情報を利用していないが、2 種の研究は同等である。これらの推定値を利用して、$B(\theta_{hat\,k})=E[\theta_{hat\,k}-\theta]$ として定義される各 k 推定値に関連し

たバイアスを推定することが可能である。この公式から、推定バイアスは、真のパラメータが既知であることを必要としているのがわかる。その代わりに、我々は非実験的推定量と実験的推定量間の差としてバイアスを推定する。実験が良好に実行されれば、推定バイアスは、式（1.4）で証明されるようにそれ自体不偏である。

$$E[B_{hat}(\theta_{hat\,k})]=E[\theta_{hat\,k}]-E[\theta_{hat0}]=E[\theta_{hat\,k}-\theta]=B(\theta_{hat\,k}) \qquad (1.4)$$

1. 多重レベルモデルの指定

本レビューでの分析の目的は、$B(\theta_{hat\,k})$をモデル化することである。$B(\theta_{hat\,k})$は、Z ベクトルで表現される研究とその介入の特徴と内容、そしベクトル変数 W 上における推定量自体の特徴の関数とされる。これによって、我々はこの質問に答えることができる。“バイアスは、用いる推定量タイプ、設定、そして設定と推定量タイプ間の相互作用によってどのように異なるか？”この分析自体が多重モデルの役に立つ。ここで、j は研究を指数化し、そして、k は各研究内の推定量を指数化している。

$$B(\theta_{hat\,jk}) = f\ (Z_j, W_k, Z_jW_k) \qquad (1.5)$$

Heckman その他（1998）は、異種の処遇効果を考慮した‘厳密な’バイアス定義を使用している。彼らの定式化では、バイアスは個の特徴である X の関数である。彼らの設定は、マッチングに依存する非実験的研究によく共通する”common support”（注：定訳なし。マッチングする相手がとれないということ）問題を理解するのに役立つ（Lechner 2000）。従って理想としては(1.5)は３つのレベルによる：1 つは個人、1 つは推定量、そして 1 つは研究となる。ここでは、我々は 2 つのレベル、推定量と研究のみを使用する予定である。個々の特質をレビュー案に加えることに付随する実際問題は、地域 X による治療効果の違い（共通支援の有り、無し）や、研究間の共通法一式についての X 値など、情報源となる研究が十分な情報を報告していないことである。しかしながら、我々は Z ベクトルの一部として研究母集団の集約的指標と、各比較群を生成するのに用いられた平均的な背景特徴の測定をコード化し包含することを提案する。マルチレベル（注：定訳）モデルとして式（1.5）の設定例があげられよう。

$$B(\theta_{hat\,jk})=\alpha_{0j}+\alpha_1 W_{jk} +\alpha_{2j}W_{jk}Z_j+\varepsilon_{jk} \qquad (1.6)$$

$$\alpha_{0j}=B_0+B_1Z_j+\omega_j \qquad (1.7)$$

ここではαとβは評価されるパラメータベクトルであり、ε とωは確率的誤差項である。

方程式（1.6）は、推定値レベルの方程式である。方程式（1.7）は、推定量が導かれた研究の特徴の関数として、バイアスへの各推定値の効果をモデル化する。

2. 推定

方程式（1.6）（1.7）にあるようなモデル推定では、いくつかの問題が浮上してくる。すなわち、研究群全体で意味ある従属変数を正確な単位で表すこと、変数内のサンプリングと推定エラー、そしてモデル内のパラメータ推定値を説明すること、分析単位を選択して、推定量レベルの観察値の非独立性を説明すること、適切なソフトウェアを選択すること、そして関連したパラメータを識別することである。

第一に、従属変数（統合案で結合されている効果）はバイアス推定値である。各バイアス推定値は、表から直接導かれるか、あるいは非実験的推定値から実験推定値を引き算することによって研究から抽出される。これらは、研究全体を通して共通単位に変換されなければならない。効果またはバイアス推定値は、自然な単位（例えば所得結果の年間ドル）で、各著者から報告される事が多い。大部分の研究が結果として所得を用いているため、我々は全ての効果を 2 タイプの単位に変換する予定である：(1)関連がある場合は、年間修正ドル、そして、(2)標準化効果サイズ。これは結果が同じ単位である標準偏差で分割された結果の自然な単位として形式的に表示される。第 1 タイプの単位を用いる分析は、所得結果のある研究に限定される。標準化効果サイズを用いる分析は、全ての研究を利用する予定である。

第 2 の推定問題は、モデル変数とモデルパラメータ推定値における推定とサンプリングエラーの説明である。従属変数は、バイアス推定値であるために、エラーと共に推定されることがわかっている値である。このバイアス推定値は、実験推定量の分散と非実験推定量の分散、そしてこの 2 つの共分散によって決定される独自のサンプリング分散をもつ。各推定量のタイプに関連する変動をとらえる分散要素を推定可能にする過程で、モデル（1.6）はさらに設定が可能である。これを行う 1 つの方策は、それぞれ有用な解釈をもつ平均と分散のある、ランダムな係数として α_{1k} を推定することである。α_{1k} の平均がゼロと異なれば、それはその非実験的アプローチと関連するシステマティックバイアスを示唆していることになる。分散が高い場合、それは非実験的アプローチが他のアプローチより効率的でないことを意味している。

第 3 の問題は、分析単位を選択し、単位間の非独立性を説明することである。マルチレベルモデルによって、我々は研究群内の集団を説明することができる。同じ研究から得られた推定値によって、研究特有の分散要素あるいは固定効果を共有することができるだろう。もう一つの非独立性源は、本質的に同じデータや非常に類似した技術を用いた同一研究範囲内にある推定量から導かれる。たとえば、異なる 10 種の傾向傾向スコアマッチング推定量は、異なる 10 の傾向スコア対応デザインの反復を表している訳ではない。レビュー筆者は、統計専門家と協力し、必要な場合は統合し、これらの依存性の説明が可能な場合は明確にモデル化を行う。

第 4 の問題は、データ・コレクションと分析に使用されるソフトウェアである。データは、単純なスプレッドシートに入力される。研究レベルのデータ用に 1 つと推定量レベルデータ用に 1 つとなる。その後これらのファイルは SAS または Stata フォーマットに変換される。そして、各バイアス推定に対して関連した研究レベルデータを反復する。これにより、最も一般的に利用可能な統計ソフトウェアパックを使って分析できる、単純矩形のデータセットができあがるだろう。

最後に、難しい問題は、研究の関心パラメータが、文献において現在利用可能な推定値サンプルを使って識別することが果たして可能であるかということである。事実、我々は、既知の変動要素の数に対して、非実験的推定値が少数であるためにモデル（1.6）が過剰適合されるのではという懸念を抱いている（下記 B 項を参照のこと）。それにもかかわらず、たとえメタアナリシスの統計力が低いとしても、それは少なくとも 2 つの理由により有用である：(1) メタアナリシスは、異なる要因、これらは異なる非実験的推定量に関連するバイアスのばらつきを説明するものであるが、そうした要因の考察に形式的枠組みを提供する;そして、(2)メタアナリシスは将来の研究が結合される際にこうして問題を設定し、そしてレビュー更新は容易で実りあるものになるだろう。この分析計画は、すでに適切なものとなっていると思われる。

B. 研究コード化分類

付録Aには、レビューにおいてコード化される項目のリストが載せてある。このリストは、我々の主な分析で使用される予定のものよりも長い。コーディング計画は、研究レベルと推定量レベルで特徴を測る。いくつかの項目は、現在の研究サンプルにおいて殆どあるいは全くばらつきがなく、分析段階で分解され、統合され、あるいは除外される必要がでる可能性もある。また、我々がすでに識別したおよそ 12 の独立実験的な研究には、単純な結論に達するにはあまりに多くのばらつき源があるかもしれない。しかしながら、コードはデータベースを将来更新するために維持されるであろう。

コーディングフォームの重要な側面は、各推定量と関連した研究デザインタイプを識別できる一連の項目である。最も重要なのは、フォームによって、比較群の出所（文献で比較されるデザインの全てが何らかの不等価比較群デザインであるため)、比較群間の差を修正するのに用いられる統計手法と処遇集団、そして、差異修正に用いられる背景データの質などを記述することが可能になることである。「データの質」についての批判的側面は、研究者が結果指標について事前介入指標をもっているかどうかである。

C　質的調査の取扱い

計画されているレビューでは、社会的介入の評価に対する非実験的アプローチと関連したバイアスの量的推定値を検討する。しかし、ほとんど全ての非実験的研究において、こうした同じ問題を質的に扱う試みが少なくともいくつかある。多くの例で、これらの質的

な論議は、非常に厳密で形式にかなっており、それ自体サンプル選択バイアスの推定値として考慮されうる。そのような質的な推定値のレビューは、検索戦略（すでに量的研究で困難なものである）がより挑戦的なものになるため、かなりの労力を必要とする。その上、これらの研究から情報を抽出し、コード化して、それらを共通単位に入れることで、さらなる挑戦が加わる。したがって、その労力は現行レビューの範囲外であるが、我々は将来野心的な調査統合者によってこれが遂行されるよう推奨する。

V.　時間的枠組みと更新計画

以下のスケジュールはプロトコル草案と共に提出され、完了期日予定は 2002 年 7 月となっている。

作業	実際の/予想される完了
出版物の検索と包括の最終決定	2002 年 3 月
研究に関する予備知識データのコード化	2002 年 4 月
結果データの標準化	2002 年 5 月
統計分析	2002 年 6 月
報告書準備	2002 年　7 月

我々は主にこの予定を順守し、ボルチモア（MD）でのキャンベル共同計画法会議で発表するために、7 月に中間報告、そして 9 月に最終報告を出した。我々は、2003 年 2 月のキャンベル共同研究会に間に合うよう、最終報告用の分析を延長・更新する予定である。

レビューされている研究分野は、いまだ大変活気づいており、新しい根拠が定期的に作り出され発行されている。加えて、我々は本レビューの普及によって、多くの評価研究者が刺激を受けあと少しのステップを取って、現在彼らが持つ根拠を本文献にさらに貢献するようなデザイン反復研究へと再形成することを期待する。したがって、我々はレビューが将来更新されうると考える。我々は、より多くの研究の追加が非常に簡単にできるような方法で、全てのデータ記憶と分析手順を設計している。レビューを更新するのに適当な時期は、入手可能となる予定の新印刷物の進行具合に多少なりとも左右される。データベースとプログラムは、その後第三者がレビューを更新するのに利用できるようにされる予定である。

VI. 潜在的な利害対立

なし

付録A

コード体系

コーダー氏名: _____
日付: _____
コード化される研究の著者/出版年: _____

1. 研究水準の特徴

研究概要と規模

研究ID No. _____ (空白でも可、後から割当てる)

査読水準/出版状況

a. 研究報告書
b. 政府出版または著者出版報告書, 論文審査のある学術専門誌には載せられていない
c. 論文審査のある学術専門誌に掲載予定、または掲載済み
d. 既刊書あるいは書籍内の章

研究から得られたバイアス推定値の数

a. 個別に報告された推定量の数: ____
b. 個別に報告された部分集合の数: ____　全サンプル　はい/いいえ
c. 個別に報告された場所の数: ____　全場所の平均　はい/いいえ
d. 個別に報告された期間の数: ____　全期間　はい/いいえ

合計: _____

正式な決定解析が研究に含まれているか？ はい/いいえ

コード研究の逐語的結果： _____
_____
_____
_____

治療のタイプと抽出

介入タイプ (複数回答可)

a. 正規の学級教育（学術的教科）
b. 非公式教育　(例：生活技能訓練)
c. 職業訓練

d. 雇用サービス
e. その他:_____

プログラム参加規則
a. 義務参加
b. 自発的参加、強い動機
c. 自発的参加、弱い動機

プログラム資格規定
主観的基準を使用　はい/いいえ/わからない
客観的基準を使用　はい/いいえ/わからない

成果の指標 (複数回答可)
a. 試験得点
b. 持続性 (例：出席、脱落、プログラム修了、学位修了)
c. 雇用
d. 収入
e. その他: _____

対象集団

全サンプルサイズ (これは、全体的な研究規模を測るため、記述統計用に用いられる)
治療群: _____
対照群: _____
潜在的比較集団: _____

研究集団の年齢・性別
a. 児童 (高校入学前)
b. 青年 (高校～30歳)
c. 成人男子 (身障者?)
d. 成人女子 (身障者?)
e. 成人男女（男 + 女）
f. 成人, 特定集団　(明記すること):_____

研究に含まれた場所の地理的位置
a. 単一地域
b. 一国内の複数地域、便宜的地域標本

c. 複数国、便宜的国家標本
d. 複数国、全標本または代表標本

実験の属性

データ収集の形態（複数回答可）
a. 自己記入調査データ、郵送アンケート
b. 自己記入調査データ、ウェブアンケート
c. 電話インタビュー
d. 対面インタビュー
e. 行政記録、学校が行ったテスト得点を含む
f. 研究者による直接観察、研究者が行ったテストやアセスメントを含む
g. 情報提供者による観察・アンケート（例：両親、教師）
h. その他: _____

対象集団の選択は無作為割当デザインによって制約されたか （例：チーム学習教室は無作為割当ができないため除外された）
a. はい (記述): _____
b. いいえ
c. ディスカッションされなかった、または不明

研究参加率 (=100% - 非定率)
a. 報告された:_____
b. 報告されなかった

順守率
対照群の交差または混在
a. 報告率: ____
b. 得られた質的記述: _____
c. 情報なし

処遇受給率(これは現行の研究にはあまり重要でなく、全ケースで入手不可能であるか、適用されない可能性がある)
a. 報告率:_____
b. 得られた質的記述: ____
c. 情報なし

研究脱落

脱落の差の評価に関する情報

a. 対照群で報告された: ___ 治療群: ____

b. 報告されなかった

c. 該当しない

行政データあるいは他のデータがバイアス評価に利用された。はい/いいえ

著者に尋ねたい質問事項、または要求したい項目リスト：

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

2.　推定量レベルの特徴

同じ研究が複数の推定量を有する場合がある。というのもそうした推定量は、異なる研究地、期間、部分集団、アウトカム、比較標本、あるいは推定方法に関連しているからである。こうした疑問は、下に挙げた全ての回答が書き入れられるようなグリッドによって置き換えが可能であろう。各ID設問は列ラベルに示される。各記述式設問は別個の行に示される。

研究: _____ (著者と年度、あるいは研究ID番号)
バイアス推定値: _____
単位: _____

記述子

研究地ID: _____
研究地名: _____

期間: _____
期間の単位: _____ (例、無作為割当からの経過月数、暦年)

部分集団ID: _____
部分集団タイプ

a. 全研究サンプル
b. 男性のみ
c. 女性のみ
d. 青少年
e. 成人
f. その他: _____

比較群のデータの質(各比較群に対して1つだけ回答すること)

比較サンプル:_____

比較群の出所

a. 無作為対照群の非無作為部分標本
b. 事前介入コホート
c. 地理、管轄区域により不適当
d. 他の理由により不適当： (明記すること)_____

e. 全国データセットより抽出
f. 他の評価より抽出: 例：複合地域研究のために他の地域から得た対照群
g. 同じ評価から抽出：プログラム欠席者あるいは治療群における非参加者
h. 同じ評価抽出：落選者（審査により除外）
i. 比較群なし（例：一群デザイン）

比較群は同じ労働市場、学校（区域）、あるいは他の関連する地理的地域から抽出されたか。

a. はい
b. いいえ
c. 必ずしもそうではない、マッチは市場・区域間と市場・区域内に含まれている

実験が可能でなかった場合は、比較群は評価できる対象を代表しているか（はい/いいえ）

処遇（対照）群と同じ計測手段を用いたか（はい/いいえ/わからない）

処遇（対照）群と同じデータ収集形態（自己報告、行政データ）用いたか（はい/いいえ/わからない）

比較群のデータ有用性のために分析変数集合が制限されているか（はい/いいえ/わからない）

処遇（対照）群と同じ回答率だったか（はい/いいえ/わからない）

結果は、SSAから得た収入のような集合データを用いて測られたか（はい/いいえ）

回帰に用いられた推定方法とデータ（各推定量につき１回答）

方法ID: _____
方法の記述: _____
効果推定で共変量を用いたか（共変量は回帰調整されたか）　はい/いいえ

回帰調整にどれくらい共変量データが用いられたか（複数回答可）

a. なし
b. 段階 I: 行政データから得られる基本人口学的情報（例：年齢、性、人種/民族）
c. 段階 II: 総分類で計測されたが、労働市場結果に対する結果に影響を与えると思われる基本情報。これには教育成績と受理学位が含まれる。教育成果には、無料昼食や減額資格が含まれているため。

d. 段階 III: 収入、職歴 (詳細程度は?)
e. 段階 IV: 個人情報：結婚歴、世帯構成、障害程度、労働組合への参加（労働市場結果用）、両親、教育、学歴（学校教育結果用）など詳細な調査によってのみ入手可能なもの

マッチングはサンプル抽出において用いられたか？はい/いいえ

マッチングは効果推定で用いられたか？はい/いいえ　（いいえの場合は、マッチングに関する設問は飛ばして進むこと）

マッチング、あるいは傾向スコア推定に用いられたデータ量（該当する場合
a. なし/適用せずまたは該当なし
b. 段階 I: 行政データから得られる基本的な人口学的情報（例：年齢、性別、人種/民族）
c. 段階 II: 総分類で計測されるが、労働市場結果に対する結果に影響を与えると思われる基本情報。これには、学業成績、学位が含まれる。教育成果には、無料昼食や減額資格が含まれているため。
d. 段階 III: 収入、職歴（詳細程度は？ Before Ashenfelter dip?）
e. 段階 IV: 個人情報：結婚歴、世帯構成、障害程度、労働組合への参加（労働市場結果用）、両親、教育、学歴（学校教育結果用）など詳細な調査によってのみ入手可能なもの

サンプル調整：マッチング手続きによって除外されたサンプルメンバーがいるか？
a. 調整なし：治療群と比較群全部を用いた
b. X共通指示体以外から得た全サンプルを除外
c. 傾向スコア分布における観測上下__%を除外
d. 傾向スコア分布における比較観測上下__%を除外
e. 該当なし

マッチング手続：マッチは復元抽出されたか
a. はい、復元あり　　比較群は複数回利用されうる
b. いいえ、復元なし　　比較群は１治療群メンバーとのみマッチされた

推定量の種類とタイプ（複数回答可）
a. マッチング
a. セルマッチング
b. 傾向スコア：1対1
c. 傾向スコア：1対複数（例：カーネル密度、重み付け、層内平均）

b. 不観測値に関する抽出
   a.　パラメトリック（例：ヘックマンの補正）
   b.　セミパラメトリック・ノンパラメトリック
   c.　操作変数
c. 差異
   a. 単純プレポスト（事前事後）
   b. 個別固定効果
   c. 成長曲線モデル
   d. その他: _____
   e. なし　（平均あるいは共変量調整における単純相違のみ)

標準誤差方法、概要：
   a.　ブートストラッピングあるいは途中経過を明らかにする他の手順を使用
   b.　傾向スコアが正確であると仮定
   c.　標準誤差計算方法について記述なし
   d.　標準誤差の報告なし

バイアスの標準誤差の算出法
   a.　直接報告されていない、効果推定値についてのみ単独で報告あり。共分散項なし。
   b.　バイアスについて報告あり（例：ブートストラップ方法を用いて）
   c.　報告なし

結果報告と解説

実験をしなかった場合、こうした方法（あるいはサンプル）は信頼性のある効果推定値をするのに用いられたと思いますか。

a. その可能性は非常に高い
b. 多少は可能性がある
c. おそらくない
d. わからない/どれともいえない

非実験的方法に関して特定された全仮説は報告書の文中で明確に述べられているか。
（はい/いいえ）

特定された仮定について正当化はなされたか？

a. 仮説検証
b. 一応の/逸話的正当化
c. 文献におけるアプローチ使用について言及
d. 正式な質的根拠
e. 正当化おこなわれず

傾向スコアによるマッチングの際、診断報告がなされているか。（複数回答可）

a. 治療群と対照群間の差異に関してF検定
b. 傾向スコア層内における処遇群と対照群間の差異に関してF検定
c. 処遇状況に合わせた傾向スコアのヒストグラム
d. 傾向スコアモデルに対する適合度検定
e. その他 (明記すること):_____

訳　小　林　麻衣子（常磐大学）

付録B

第1回目の批評に対する回答

訳　小　林　麻衣子（常磐大学）

レビューワＡ

1. ***筆者によると、準実験も含めるとのことだが、まさにそれが何を意味しているのか包括基準が明確ではない。 不等価比較群デザイン、時系列、ケースコントロールあるいは後ろ向き研究を含むのだろうか？この質問に答える何か基準があるのか？***

レビューでは、我々の包括基準を満たす全てのデザイン反復研究において実施された非実験的デザイン・方法の全てを調べる予定である。我々が特定した研究における殆ど全てのデザインが不等価比較群を用いている。デザインは主として、比較群の出所、差異調整するのに用いられた背景的データ、そして差異調整に用いられた統計的手法の点で異なっている。よってそれらが、我々がコード化した属性である。

2. ***筆者は、真の効果の近似値を求めるため、無作為実験から得られた結果を用いたバイアス推定を目的としている。これは、彼らが同様にして全無作為実験を扱うことを除けば妥当である。しかし、無作為実験が全て同じように優れているとはいえないではないか？筆者は優れた準実験と劣った準実験について何らかの説明方法を検討すべきである。例：高い脱落のあるものと脱落が低い、あるいはないもの、サンプルサイズが大きいもの小さいものなど***

我々は、レビューワが“優れた実験と劣った実験”を意味したものと仮定している。この記述には賛同するので、基準として用いられた実験の質についてのディスカッションに叙述的セクションを充てたい。このディスカッションでは、非無作為脱落、溢出効果や置換バイアス、治療割当への不順守などに関する問題を考慮に入れる予定である。 我々の全般的な認識では、レビューする研究で用いられた実験の殆ど全ては、いくつかの特定可能なパラメーターに関連して効果推定値を不偏なものとして扱うのに十分高い質である。脆弱性を確認するまで、我々は、劣った研究に依存したバイアス推定値を包括・除外することで自分達の結論がどのように影響を受けるか調査する予定である。

3. ***著者は、標準化効果サイズを推定するために対照の標準偏差を用いるとしているが、私は合併標準偏差が良く好まれるというコンセンサスがメタ分析者間にあると認識している。合併標準偏差を使わない理由が何かあるのか？***

我々は、利用可能な最良母分散推定値を用いる計画である。合併分散が報告されるか、または報告された情報から計測が可能であれば、それを用いるだろう。そうでなければ、無作為対照群を使って母分散を推定する予定である。我々のレビューにおける殆どの研究には、標準化有効サイズを算出できるような情報は全くなく、よって我々は共通の有意な測定基準を作成するため代替案を探索している。

4. 筆者は、モデル特定のためにあまりに多くの予測値をモデル化してまう可能性について賢明に論じている。たとえ完全な多変量モデルが特定されずとも、筆者は教育的である可能性をもつバイアスについて、いくつかの簡単な記述統計を示すことが可能であると認識している筈である。

まさに我々が意図しているところである。既にいくつかの記述的統計について算出を行っており、さらに継続の予定である。未解決の問題は、巨大化しそうな記述的情報をどのように要約するかについてである。結果は、選択した図で集計するかあるいは説明するかもしれないが、文中で全てを報告する予定である。

5. すでに2002年5月であることを考えると、筆者の予定表はきわめて野心的であるように思う。おそらく、本当は作業の殆どを終えていると思われるが、そうでなければスケジュールに間に合わせることは不可能であろう。

我々はおおむね予定に従っている。6月17日には論文の草稿を終えた。7月にプロトコルを更新し、7月・8月で追加コーディングと分析を行った。そして9月末までにレビューの改訂版を提出しようとしている。C2方法協議会で9月に最終研究成果中間報告を提出し、2月のC2研究会に向けて、また別の分析を完成させ詳述する予定である。

6. コーディングについて評定者内の信頼性の論議が行われていないが、これは不可欠であり取り組まれるべきである。.

批評者は、重要な懸念をあげているが、つまりもし別のコーダーによっていくつかのコーディング決定が再び行われれば、それらは異なってなされたかもしれないということである。2つの懸念事項とは誤差と主観性である。誤差あるいは主観的変動に対する各コード化項目の感受性は、コード化されている各項目で異なり、情報が引き出された出版物の明確さによって異なることがある。我々のアプローチでは、まず1人のレビューワ（Glazerman）に研究半分を主として担当してもらい、別のレビューワが（Levy）残りの半分を受け持つようにした。しかしながら、何よりもまず我々は論文を選択し、両レビューワにそれぞれその論文を読んで、別々に完全なコード化をしてもらった。それから我々は項目ごとにコーディング結果を比較し、全ての差異調整にじっくり取り組んだ。両レビューワは全論文を読み、我々が関与したものについてコード化を行った。次に、我々はそれぞれのコードをチェックし、最も困難である項目または最重要項目にはとりわけ注意を払った。いくつかの事例では、特定の論文やデータ項目について論議するため協議を行った。誤差を少なくするため、リサーチアシスタントに選択論文を精査してもらい、第１コーダーについてもう別の確認を行うための情報を得た。主観性を少なくするためには、より経験のある研究者(Myers)に選択項目のレビューを同様にしてもらった。

7.　筆者は、C2研究デザイン要約においてすでに作成されたいくつかのコードもまた活用できるのではないだろうか。

この提案に従う予定だが、研究に包含されている情報そのものによって多くの決定が導かれた。例えば、もし我々がコード計画(1)「デザイン種」を採用するとなると、レビューにおける殆ど全ての非実験的推定量は、同じデザイン、つまり「不等価比較群」を有することになるだろう。これについて証明するデータはないが、我々の経験から言えるのは、社会プログラムを評価するのに最もよく用いられる無作為実験の代替法は、たいていこの分類に入るということである。我々のコード体系は、以下の3局面に基づいてこのカテゴリー内での区別をしようというものである：比較群の出所、処遇群と非処遇群間の差異調整に使用できるデータの質、差異調整に用いられた統計的手法。.

レビューワB

しかしながら、以下の最初の点が考慮されなければ合意はできかねる。他の点はこれに比べると重要度は低い。

1. *準実験などというものは存在しない。ただし準実験的デザインというものはある。どの準実験的デザインが他より優れており、ゆえに無作為化実験の代替法として相応しいかが既知である場合、いくつかの低級の準実験に対して実験を対照させるのは意味がないことである。よって、もしここでの対照事例が“準実験”あるいは“非実験”であれば、我々は時間を無駄にすることになる。準実験の類型構成、その類型を指数化するのに信頼できるコード体系の考案、そして実験に関連した異種の準実験効果を推測できる統計的検出力のある分析の作成、これらについて計画はあるのだろうか？実験はバイアスがより少ない、という結論付けはあまり役には立たない。*

上述の批評は、我々は各ケースで最良の利用可能な準実験的デザインに集中すべきだということだと理解している。批評者は、我々は先験的に低級準実験と優れた代替法間の差異について既知であると提案している。また、彼/彼女は 入念なデザイン種のコーディングにより、我々はその区別ができるとも述べている。

この領域で我々が行った文献読解によって、どの準実験デザイン[5]が（データを使ってどのデザインが最良に機能するかを把握する）無作為化実験を最も反復する可能性があるかどうかについてはかなりの不確実性があることがわかっている。例えば、一般的に単純すぎるという理由で却下された方法が、より洗練された、あるいはデータハングリーな方法と同じくらい優れている、あるいはさらに優れているのかを知ることは有用な発見であろう。

我々は、実際に多くの準実験的デザイン/方法が使われていることについて確実に認識している。 事実、我々の目的は、非実験的根拠が社会介入の効果について適切に設計され実行された無作為化実験と同じ情報を提供できる条件について、よりよい理解をすることである。本キャンベルレビューおいて我々ができる最善のことは、このトピックについて何が文献で分かっているのか、可能な限り完全かつ系統的に総合することである。この過程で、いかにして異なる準実験的方法とアプローチをお互い比較するかについて我々は学ぶことになろう。

我々は、研究デザインアプローチの類型と、その類型を指数化する信頼性高いコード化体系を作成している。 我々は複雑な類型から取りかかり、それが実行不可能であるとわかった。 結局、7つの非相互的な限定指示変数から成る単純な類型は、レビューした文献で見られた一連のデザインを把握するのに十分記述的であった。 最後の段階である、異種の

[5] 我々は、準実験的と非実験的という用語を相互に使用し、無作為割付によって統制されていない全てのアプローチを意味することとする。名詞の場合は、準実験という用語を使う。

準実験の効果推定に対して統計検出力のある分析法の作成はより難しいものであった。なぜなら、統計的検出力はデータによって制限されてしまうからである。他のキャンベルレビューと同様、我々は最も優れ、最も適切である利用可能な分析手段を用いなければならない。しかしデータによって強固な結論が立証されなければ、それが、レビュー結果であると報告する。

2. ***実証的研究を、教育と労働力分野に限定することには全く異存はない。しかし、筆者は教育を極めて狭義に思い描いているように感じる。例えば、ある種の音声学あるいはテストスコアについての音素論効果を評価した研究は、過去15年で1000以上公表されているが、そのうち56が実験、あるいは準実験である。すべての教育調査研究で、今までテストスコアについてなされた教育介入の全コーディングを行おうとしたら一生かかってしまうだろう。筆者は、経済学者であり、その専門分野における研究者によって主として影響された研究を展望しているようにみえる。しかし、教育専門家による教育研究は果てしなく存在する。提案書の筆者が考えている領域は、もっと限定されているように思う。特定することはできるのか？さもなければ、著者はデータに溺れてしまうか、学問的に短絡的であると非難されてしまうかどちらかであろう。***

我々は、多くの学問分野を検索し、視点を経済学だけに制限はしなかった。批評者が引用した文献は、同様に我々が行った検索の一部分である。この文献の中で、真の研究内デザイン反復はたった1つしか見つからなかった (Aiken et al. 1998, who studied college remedial writing)。文献では、多くの無作為化試験があったが、非無作為化比較群を用いて無作為化対照群の結果は近似されうるかどうかについて方法論的研究は皆無であった。

3. 筆者は、実験と準実験結果を直接的に比較するメタ分析文献にもっと留意すべきであると思う。特に、LipseyとWilson(1993, Psych Bulletin)の研究を参照されたい。そこでは、実験と準実験では約50のトピックにわたり同じ回答が得られたが、特定のトピックに関する実験では標準誤差が準実験よりもかなり小さかったという議論がなされた。言い換えれば、実験だと同じ解答に到達するのがより早いということである。これは、バイアス軽減におけるその優位性とは無関係であるという実験の価値（とりわけ新分野における）に対して重要な主張を成す。もちろん、準実験はデザインの質に関しては不均一な集まりであるので、タイプによっては他よりもより優れた無作為化試験結果の近似を導けることもあるだろう。これはLipsyとWilsonによって探索されなかったものの、私が上記1で強調していることである。

批評者は、研究間デザイン反復について言及している。これらの比較は、大変興味深く、我々が提示しているものと同じ研究論題に取り組んでいるが、異なる手法を用いて行われている。研究間アプローチは、広域な研究探索を用いているという利点があるが、研究デザインを研究それ自体と潜在的に混同してしまうという不利点もある。よって、我々は、システマティックレビューにおいて、背景として以外は研究間反復を含まない。将来キャ

ンベルレビューで、プログラムの効果推定値について研究デザイン効果の研究間エビデンスが要約されることを望む。

4. ***本論文の目的は、キャンベル共同計画の中核を成すものであり、他の分野（刑事司法、健康、幼児教育、本PIが扱わないタイプの教育など）で同様の事が行われるのを期待している。こうした分野のより成熟したレビューがすでにあるとすれば（例えば健康分野）、それらは私のように、より社会科学的な典型に知られるべきであり、筆者によって引用される必要があろう。***

我々は、多分野で少数のデザイン反復研究（ほとんどが健康関連）を特定したが、これらはキャンベルレビューに値する段階まで蓄積されていない。しかしながら、これらの分野の個別レビューを行う他のレビューワのいかなる試みも歓迎し支持する予定である。

レビューワ C

1.　このプロトコルの理解は困難であった。統計モデル作成については長く、提示された論点及び統合可能な研究の性質に関する批評的事項については短い。

我々は、こうした批評で論議されている詳細を加えるため、プロトコル記述を改訂した。セクションIV.Aの統計的詳細は、連続性を失うことなく省略可能である。

2.　しかし、本プロトコルはC２レビューだけでなく幅広い分野の介入研究にも概して影響をもつ、大変重要なトピックについて取り組んでいる。私は、実験研究に対する準実験的近似におけるバイアス問題を啓発するようないかなる研究も支持したい。こうした好意的気持ちが私にはあるのだが、本プロトコル草案内での論議から、レビュー案が意味ある、または有用な結果を提供するであろうどんな方法でも実行可能であると確信するにはいたらない。

実行可能性についての疑問は、レビューの形式的メタアナリシス要素に最も当てはまりやすいと考えている。よってこの意味で批評者の懸念に感謝する。システマティックレビューの実行可能性に関する限りは、本文献は誰かが統合しなければならないところまで蓄積しており、よってトピックに関心のある人は、16の大変複雑な研究を新たにやり遂げる必要はない。我々は、説話的レビューと、メタアナリシスに向けられた協調努力から得られた（少なくとも）要約の両方を提示しようと計画している。我々はまた、個々の出典研究の山よりも、読者がよりよくデータを理解できるよう、記述的情報を盛り込む予定である。

3.　調査の第一の論題は、準実験的方法が、無作為割付研究から得られる研究成果の近似値を求めるのに用いられるだろうか、ということであるが、私にとってはあまりにも広範すぎ、レビュー案が実際取り組む範囲を超えているように思う。準実験的方法には、不等価群比較、不連続回帰デザイン、多重時系列、そして他のいくつかのばらつきに関して多くの差異がある。無作為割付研究結果を推定するこれらの能力に関して、これら全てに渡ってどんな一般化も可能だと信じることはできない。各々に異なる長所・短所があるだけでなく、それぞれ適用される特定の状況次第で、多かれ少なかれ成功する可能性がある。

レビューから得た主要な洞察は、社会科学では多くの巧妙な計画が因果関係特定に用いられているが、プログラム評価は一般的に不等価比較群デザインに最も依存しているということである。この点で、我々が行うのは大変広範な準実験的デザイン類のレビューである。この部類内で、我々は出所の異なる比較群、背景データタイプ、差異調整に用いた統計的手法について検討を行う予定である。研究論題を明確にするため、異なる手法と技

術が、より優れたまたはより劣った働きをする条件について我々は知りたいと思っている。これは、批評者が提言していることと、我々がこの重要点を明らかにするためプロトコル改訂したことと一致しているであろう。

4.　プロトコル草案のディスカッションから推測されるに、特に焦点が向けられているのは、不等価比較群デザインである。この不等価比較群デザインは、実験群と非無作為対照群間の初期差異から生じた全てのバイアスを少なくする統計的調整、あるいはマッチングを通した試みが有るか無いものである。そうしたデザインは広範であり問題があるが、もっと直接的かつ全面的に提示される必要性を考えると、この焦点は適切である。ちなみに、コード化体系には１群事前事後デザインに関するカテゴリーも提示されている。このデザインは、より一層問題をはらんでおり、不等価比較よりも異なるバイアス源にさらされやすいという、かなり異種のものである。そうしたデザインの個別分析に対応する十分なデザイン反復研究が見つからない限り、これを含むのは賢明ではないと思う。

　コーディングフォームは更新されている。具体的には、1群事前事後デザインに対してはもはや欄を設けていない。実際は違っていたものの、遭遇すると思われたデザインに関しするコードを初めに作成した。

5. 焦点が非無作為化群にあると仮定すると、背景ディスカッションと適格性基準の重要な点は、このタイプのアプローチ範囲についての記述であろう。コード化体系は観測されない変数について、マッチングと抽出をそれぞれのいくつかのサブカテゴリーと共に提示している。もし、こうしたカテゴリーとサブカテゴリーがより完全に定義されたならば、統合案の範囲と焦点はより明確なものになっただろう。また、マッチングあるいは統計的補正を用いていない不等価比較が含まれるのかどうかについて明らかでない。コード化体系にはそうした比較について適切な箇所が見当たらない。

　デザイン代替の範囲について我々の理解を反映するため、コード化体系は更新されている。特に、カテゴリーはより全面的に定義されており、コード化体系ではマッチングあるいは統計的調整を用いていない比較も考慮に入れられている。

6.　また、私は包括されるべき研究の種類と、適格な研究を抽出する基準があいまいであると思う。この件に関する簡単な記述から、介入群、無作為割付対照群、そして非無作為化対照群を使った研究は、基本的に該当していると私は解釈している。他の基準（例えばこれら全３群が同じ場所の同じ集団から抽出される場合など）があるかどうかについては明確でない。さらに、プロトコルでは、もし同じ介入、時間など同じ背景において行われた研究を含むならば、メタアナリシスが適格であろうと示唆している。こうした目的に照らした極めて限定的な基準を満たすメタアナリシスがあるか疑わしいと思うが、もしあるならば、適格性を自ら持つ原始研究について情報を提供する以外に、どのようにして統合に

それらが取り入れられるのかが明確でない。

研究内比較として包含されるためには、処遇群と統制群は必ず同じ場所のものでなければならない。比較群は必ずしも同じ場所から抽出されなくともよいが、無作為化対照群の代用としてふさわしいものでなければならない。全3群は、同じ処遇個体群に関連している。(処遇群は、自らの成果を事実に基づいた状態で評価する。統制群と比較群はその成果を事実に反した状態で評価を試みる)これらは異なる処遇群の比較を含むため、たとえ各処遇群が同じ治療を受けると意図されていても、メタアナリシスは望ましくない。

7. 統合案の成功への重要要素は、分析に利用可能な適格性のある研究の数である。プロトコルは、とりわけこの点に関する様々な側面について疑わしい。第一に、教育または雇用成果の改善を目的とした介入に制限されている。これら2分野だけに着目する論理的根拠が明確でない。より多くの介入領域を対象とすることで、さらに多くの適格な研究がもたらされるだろう。しかし、非実験的デザインの妥当性は、比較群設定に関する介入の質やサンプル、実践と関連して異なる介入によって変化する可能性がある。そうしたばらつきを分析するには、多様な介入領域の1つ1つから得た多くの研究が必要となるだろう。あるいは、十分適格性のあるデザイン反復研究を含む充実した介入領域について、これを分類する試みも可能と思われる。雇用関連プログラムは、単一領域として有効であろう。2領域を選択することは、交差領域の変動を十分に検討することもなければ、その領域特有の結論を支持するのにどちらか1つに十分な研究があると必ずしも意味しているわけでもない。プロトコルのある時点で、12の適格研究がすでに入手されたと述べられている。どのようにこうした教育と雇用間の分類がなされたのか明示されていない。しかし、たとえこれらが2領域の片方における全てだとしても、不等価統制群を扱う過程（マッチング、傾向スコア、ANCOVAなど）で、研究が多くの変動を含む場合、その数が分析に十分であるかどうかは疑わしい。

本レビューでどの介入について考察を行うかを決定した際、我々は成果の類似性を維持する一方で十分に広範な領域を持つことと、プログラム参加を左右するのに適当なプロセス（すなわち選択バイアス手順）の間のバランスを取るよう努めた。

非実験的デザインの妥当性は異なる介入領域で変化するという点については批評者に同意する。理想的には、我々のシステマティックレビューによって、介入領域を含め、どの異なる非実験的方法が多かれ少なかれ実りあるものかが分かるだろう。可能な程度まで、我々は異なる非実験的方法がどの介入領域でより良好に働くのかを探求するつもりだが、データには制限がある。

8. 別の場所では、“失敗した”メタアナリシスでさえ有用であると主張されている。理由は、それはディスカッションに形式的枠組みを提供し、今後の研究が利用可能になるにつれ有効になるからである。もし、形式的枠組みの適用性を有意義に評価するには研究が少

なすぎるとしても、多くが得られたことについては全く明らかでない。

キャンベルレビューの目的は、利用可能な最善の根拠を統合し、分野における知識状況について我々が何を知っていて何を知らないのかを示すことである。文献に格差があるとすれば、我々はそうした格差を実証する。12か16の研究では、明確なメタアナリシスを行うには不十分ということがありえる。もちろん、そうした研究の豊かさに留意しておくことが重要である．これらの多くは、複合的なアプローチを用いてバイアス推定値を出している。我々はすでに1000を越えるバイアス推定値をたった11の研究から抽出している。我々はそうした推定値間の非独立性について明らかにするが、有効に使える追加的変動があることも認識する。

9. *本プロトコルで提示された枠組みと、添付されたコード体系内では、脱落の重要な役割について、これが受けるに値するほど着目がされていない。不偏介入効果を推定するのに実験的比較を用いるためには、割付後に実験群、統制群を問わず最小限の脱落もあってはならない。この問題については、プロトコル内の脚注＃３で間接的に言及しており、コード体系で最低限の脱落をいくつか得ている。論議されていないのは、介入研究において脱落がどれだけ広範囲に渡っているのか、仮にあったとしても、脱落のある実験比較はどのようにして介入効果の不偏推定値を得ることができるかについてである。非実験的比較群のさらなる複合的脱落は、初期不等価性に関連した一番上のバイアス効果推定に独立的にバイアスをかけるかもしれないが、これもまた論議を要する。脱落が皆無あるいはわずかな研究を見つけられると著者が見込んでいるのか、または分析案では比較基準とされる実験効果サイズにバイアスをかける可能性があるにもかかわらず、彼らが何とかして脱落を分析に組み込もうと計画しているのか、プロトコルでは明白でない。*

非無作為脱落は、実験にしろそうでないにしろ、どんな縦断的研究においても重要問題となる。レビュワーAによる指摘（＃２）を参照のこと。

10. *要約するに、本プロトコルは重要なトピックに取り組んでいるものの、2つの主要点において疑問がある。第一に、包括される準実験デザインについての明確な詳細、そうしたデザインが変動すると予測される基本寸法、そして統合に盛り込むことに関連した“デザイン反復”研究の基準を与えられていない。*
項目３から６の回答を参照のこと。

11. *第二に、提案された方向に沿った非実験的デザインバイアス問題の分析は、デザインタイプ、対象サンプル、そして（少なくとも）介入タイプに渡って変動の検討を必要とするだろう。変動分析は、重要変数によって十分に変動する比較的多数の研究を必要とする。この種の有意な分析に対応する関連研究が十分に存在するのか疑わしい。代わりに、２，３のデザイン変動のみを比較し、限定分析に十分対応する１介入領域における少数の研究*

があると思われる。雇用プログラムとはこうした領域であると思われるが、プロトコルはそうした主張もしていなければ、利用可能なデザイン反復研究の数やその変動範囲について具体的に説明もしていない。

我々が何を見つけるかついて、批評者の推測は正しかったことが分かった。参加者の収入についてプログラム効果に関するバイアス推定値を使った研究は11である。これらは完全なデータセット間よりも、より多くの共通性があることに対して核心を成している。残りの研究の殆ど全てについては、標準化効果サイズについて結果をコード化し、追加的だがより限定された比較を研究間で行うことができた。有用なシステマティックレビューを行うことはより困難だが、それでも研究数が比較的少数の副次的領域において実行可能であることが分かった。.

付録C

第２回目の批評に対する回答

訳　小　林　麻衣子（常磐大学）

レビューワＡ

1.　“我々がレビューした研究で使われた全ての実験は、特定可能ないくつかのパラメーターに関する効果推定値を不偏として扱うのに十分高い質を備えている“、と言うのは、我々が“確認可能“というのを深刻に受け止め、またこうした研究からの脱落が事実上ない場合を除けば信じ難いように思える。後者の場合は、この回答に対して何ら反論はない。しかし、そうした無作為割付されたものから脱落が少なからずある場合（例：10%またはそれ以上）、研究されているパラメーターが“確認可能”であるかどうかについては議論の余地が極めてあるが、ディスカッションだけでなく分析においてもこのことは考慮されていない。

実験効果推定の不偏性には多くの潜在的脅威がある。それらのうちの１つは研究脱落に関連しており、本研究で我々が測定しているものである（コーディングフォームの４ページを参照）。特定の脱落率（たとえば10%）が懸念を生じさせるか否かは、脱落の理由、脱落者対継続者で異なる効果の尤度、そして脱落率が処遇群と対照群間で異なるかどうかに左右される。こうした情報の全ては、体系的に抽出できるような方法では滅多に公表されない。代わりに、我々は、実験の全議論を検証し、そして条件を満たしているかどうかについて推断するという研究者のスキルに頼らなければならない。妥当性や不偏性に対する他の多くの潜在的脅威と共に、この決定は脱落を考慮することになるだろう。どんなデザイン反復結果も、実験効果推定が不偏であるという仮定次第である。もしバイアスがあると我々が確信する場合、バイアスのサイズと方向を予想し、問題になっている実験－非実験比較が省略されるならば、我々の結論がどのように変わるかを検討なければならない。これは、データサブセットに対して実験の質表示を用いることにより、メタアナリシスでは容易に行える。

2.　内部評価者の信頼性について、もし私が出版用にこれをレビューしていたならば、筆者が提供するよりもう少し多くのデータを見たいと思うだろう。

方法論的レビューとして、我々が行ったものは、典型的なシステマティックレビューより技術的に厳しいものであるため、十分に訓練されたより少ないレビューワ（リサーチアシスタントではない）が、原研究に多くの時間を費やすことになる。また、メタアナリシスは、システマティックレビューのごく一部にすぎない。よって、我々は全データポイントのコーディングに形式的な重複を入れるのは費用効果的ではないと決断した。 それにもかかわらず、我々は明らかな重複を結合し（第１回目の批評に対する回答で記述）、選択項目についてかなりのインフォーマルディスカッションを行った。このより知的な段階的手続きは、どのデータポイントがレビューを必要としているかを決定するものだが、メタアナリシスで従来用いられてきたであろう機械論的なものよりも効率的であることがわかった。.読者はこの情報に照らして研究結果の判断が可能である。

レビューワB

[回答レターについて]

1. *回答のいくつか、とくに、非実験的対照のタイプを先験的な立場で区別するのは不可能であるという主張に関しては不安を起こさせるものであった。準対照実験の１つの形としてまさにマッチング領域内において、著者は一卵性双生児デザインを二卵性双生児デザインと同じくらい優れているとみなすのであろうか？；後者は兄弟対照デザインと同じか：後者は結果同変数に対するｎ年の事前価値についてマッチされた非家族比較デザインと同じか：これは１年マッチングをしたデザインと同じくらい続くだろうか：人口学的データのマッチングはあるが、結果については行っていないデザインと同じだろうか：介入が極めて局所的な場合に、全国データセットから得られた傾向スコアや個別レポートなどに基づいた外部比較と同じか？もしＣ２において全種の非実験が、どのデザインが他よりも優れているかについて事前の意識なしにホッパーに投げ込まれたのだとしたら、私には大変な懸念である。もしくは、お互いに相殺し合う２つと共にいくつかの準実験は過小評価され、他は過大評価されるのを期待している。Ｃ２の公的信頼性については、まさに緊急の課題である。*

我々は、先験的に非実験デザインタイプを区別する、あるいはその優劣を判断することが不可能だとは考えていないことを明確にしなければならない。事実、我々は研究の多くの側面とそこで用いられた非実験的推定量を入念にコード化した。(付録Ａを参照)
先験的にデザインの強弱を区別することに関して、問題はそれが客観的に行うことができないことである。少数の明らかな場合（双生児の例のように）を除き、与えられた状況においてどれが優れたデザインであるか、あるいはどのデザインが、結論が内的に妥当であると確信するのに十分なほど実行されたかについては相当な人が異論を唱えることが多い。また、基本データをもっと収集するというような、デザインの改善がコストに見合うだけの十分な差異を生み出したかについても意見が異なる。

現在の我々の知識だと、そうした人々の間で行われる理論的議論は、更なる理論によって解決されそうにはない。我々の論文の核心は、一部のデザインが他よりも優れた働きをするかどうかを*実験的に*調べることである。もし有力で注目に値する人物が全てのコード化を自発的にしてくれるのならば、我々は、メタアナリシスにおける仲介者として研究の質の主観的コードを喜んで用いる。事実、我々はそうした情報を限られた基準で有効に生かしている。なぜならば、1事例でHeckman and Hotz (1989) は怪しい推定量を除外するために、政府後援職業計画の非実験的反復に対して形式的な規格試験を行ったためである。しかしながら、我々のアプローチは、客観的に測定可能なデザインの特徴を用いて、データ自体に語らせようというものである。

[プロトコルについて]

2. 無作為化実験が不偏の結果をもたらすことについて、以下の2つの点においていかにして確信できるのか？(a)現実世界では、完全な無作為化の必要性と実行された無作為化の実用性には必ずある程度のずれがある。(b)無作為化実験とその準実験的比較間で時間、セッティング、あるいは集団の変化がある場合、どうやって実験と準実験の差が処遇バイアスであると分かるのか？

(a) 無作為化実験によってプログラム効果の不偏推定量が与えられるとは、我々は決して確信できない。不偏性の仮定というのは、ただ主張されただけの仮定である。それにもかかわらず、社会プログラムの無作為化試験は、非実験的推定量の機能測定に対して最高の基準を示していると我々は確信する。我々が調査した実験における他の質に関する問題については、レビューワＡへの回答を参照されたい。

(b) レビューに包括されるため、デザイン反復研究は異なる方法を用いて、(時間・設定・集団について定義された)*同じ*パラメーターの推定値を含まなければならないという点で、時間・設定・集団はすべて一定にされている。多くの場合、準実験デザイン*/*方法は、与えられた処遇群について反事実的結果を概算するために、異なる時間・設定・集団からの情報に依存していた。そうした依存はそのデザインの必須部分であり、まさに我々が評価しようとしているものである。我々は、反事実的条件文を推測するために、データが異なる時間・設定、または時間から収集できる条件について理解しようと努めている。

3. 準実験種間の差異を調査する可能性が存在することがもっと明らかになればいいと思う。これらは概念的に平等に作り出されていない。

様々な準実験種間における差の調査は重要であるという批評者と同意見である。事実、これは我々のレビューの核心である。この批評者が様々な実験種ということで何を意味しているのが確かではないが。おそらく、我々が調査したＮＸデザインの範囲（すなわち、デザイン反復の著者によって検証されたデザイン範囲）は、批評者が好むものよりも限られているのだろう。これは我々の手の負うところではない。そして我々の包括基準を満たしており、より広範囲の評価デザインを対象としたデザイン反復を行った全ての新しい研究を喜んで包括したい。教育評価に対する非実験的研究デザインの現実社会的適用はきわめて狭いというのが我々の認識である。我々がレビューした研究に反映されているように、その殆どがマッチングや他の統計的調整と共に比較群法を用いている。