第２章　アメリカにおけるメタ評価の現状

佐々木　亮

１　メタ評価の種類

メタ評価は、現在は次の二つの意味で使われているとされる（Bustelo, 2002）。

（ア）評価の質的管理（評価デザイン批評）

（Evaluation quality control (Evaluation design critique)）

（イ）評価結果の統合

（Evaluation Synthesis）

さらにスクリヴェンは、評価の結果が客観的証拠（いわゆる Evidence）に基づいて、論理的な結論が導出されているかを検証する意味でもメタ評価が使われていると理解しており、次のように名づけている（Mathison 2008: 250）

（ウ）独立的立場からの評価結果の再検証

（Independent confirmation of evaluative conclusion）

上記のうち、（ア）は、何らかのチェックリストを使用して、評価報告書を構成する一般的な項目がカバーされているかとか、評価活動において行うべき活動項目や遵守すべき注意事項が遵守されたかどうかなどをチェックすることを具体的内容とする。

上記のうち、（イ）は、いわゆるメタ分析（Meta-analysis）に近いものである。ただしメタ分析が、類似の複数のインパクト評価で明らかになった効果サイズ（Effect size）を、合成して平均的な効果量を計算することを目的とするのに対して、メタ評価は類似の複数の評価結果をレビューして、総合的な評価的結論（「やって良かった／悪かった」「やるだけの価値があった／なかった」等といった価値的結論）を導出することを目的とする。つまり、メタ評価も評価の一種であるから、価値に触れる用語（Value-laden words）を使って明快に評価的結論（Evaluative conclusion）を出すことが求められる。

上記のうち、（ウ）は、厳格な方法・デザイン（科学的あるいは客観的な方法・デザインとも言う）によってエビデンスが得られたかどうか、そして得られたエビデンスによって論理的・合理的な結論が導出されたかどうかを検証する作業である。

２　メタ評価のためのチェックリスト

前節で解説したメタ評価の３種類のうち、（ア）評価の質的管理としてのメタ評価のための基準はどのようなものであるべきだろうか。この点に関してスクリヴェンは次のように述べている（Scriven 2009）。「（各種の）評価の体系的なアプローチは、自動的にメタ評価のための体系的な基礎を提供するものだ。明らかな例は、CIPP モデルチェックリスト（CIPP Model Checklist）（Stufflebeam 2002）、プログラム評価基準（Program Evaluation Standards (PES)）（Sanders & Joint Committee 1994）、基幹評価チェックリスト（Key Evaluation Checklist）（Scriven 2006）などのチェックリスト群である。ただし、それらの本来的な有効性にも関わらず、単純に形式的に適用され重要な点を見失う結果になることもある。」[1]　なお、評価の質的管理の作業の一部として、（ウ）独立的立場からの評価結果の再検証も含まれる場合が多いのが現実のメタ評価の運用である。

ここでは、「基幹評価チェックリスト」（Scriven 2008）を掲載する。また、「評価者の基準原則によるメタ評価チェックリスト」（Guiding Principles Checklist for Evaluating Evaluations）の概略も述べる。また、「プログラム評価チェックリストのメタ評価チェックリスト」（Stufflebeam 1999）については、本報告書の第１章を参照されたい。

[1] “A systematic approach to evaluation automatically provides a systematic basis for metaevaluation. Obvious examples are the checklist approaches, e.g., CIPP (Stufflebeam 2007), the Program Evaluation Standards (PES) (Joint Committee, 1994), and the Key Evaluation Checklist (KEC) (Scriven, 2007). … However, you’re then pinning your effort to the validity of these checklists and even if they are intrinsically valid, they may be applied ritualistically, and thus miss important points.”

BOX 1 「基幹評価チェックリスト」（Key Evaluation Checklist）

スクリヴェンが 1980 年に作成して提案して以来（Scriven 1980. The Logic of Evaluation)、広く用いられている基幹評価チェックリストであり、メタ評価にそのまま利用可能である。

A：準備（Preliminaries）	
I．要約	通常は項目 11～15 の要約を書く。
II．序文	評価を依頼された目的と背景を書く。(i)統制評価、形成評価、帰属的評価のいずれの目的で実施するのか。(ii)等級、順位、側面描写等のどれを総合評価として求められているのか。(iii)評価依頼者、利害関係者（ステークホルダー)、その他で報告書を読むのは誰か、を記述する。
III．方法	どのような評価方法や評価モデルを用いて評価を実施するのかを記述する。例）実験デザイン、マッチングデザイン、フォーカスグループ、インタビュー等。
B：基礎（Foundations）	
1．背景と状況	政策／施策／事業の歴史や背景を書くとともに、次の項目を説明する。(i)川上のステークホルダー（政策／施策／事業の資金負担者等)、(ii)関係する立法や政策変更、(iii)公式のプログラム・セオリー等、(iv)類似の介入の結果、(v)以前の評価結果その利用結果（もしあれば）
2．政策／施策／事業の概要	政策／施策／事業の正確で完全な記述を行う。主要な目標と予想される成果、活動内容、活動を取り巻く環境や状況を説明する。それらは、公式のプログラム・セオリーとかなり違うことがあることに留意する。また、当該評価で使用することが避けられない専門用語があれば、ここで解説する。
3．受益者（Consumers or impactees）	受益者は、(i)サービス／生産物の利用者（川下の直接的‘被影響者’）と(ii)サービス／生産物の利用から間接的に影響を受ける人々（川下の間接的‘被影響者’）に分けて説明する。政策／施策／事業の実施に携わる職員等は受益者ではないので記載する必要はないことに留意する。
4．投入資源（Resources）	政策／施策／事業実施のための投入資源を説明する。投入資源は、財務的、物理的、人的、そして社会能力的な資源に分けて記述する。社会能力的資源とは、知識、技術、職員・ボランティア・地域住民・支援者等の‘良い意志’のことである。
5．評価基準（Values =Evaluative criteria）	政策／施策／事業を評価するために用いる価値基準である「評価基準」を特定して記載する。なお、「評価基準」は変化しないものではなく、その時代やその場所によって変化することに留意する。「評価基準」は以下の情報源から得られた情報を統合して特定する。 「評価基準」は、「価値側面」(Criteria of Merit）と「価値水準」(Standard of Merit)で構成される（「価値側面」は、後述の‘過程、成果、効率性、比較優位’などのことで、「価値水準」とは、それぞれの価値側面における‘優、良、可、不良、不可’のそれぞれの水準の状態の具体的な記述で

	ある)。 (i)受益者のニーズ査定結果、(ii)定義的情報（評価対象の政策／施策／事業の上位目標および目標が設定されている場合で、十分に論理的な場合はそれを援用する)、(iii)論理的情報、(iv)法律的情報、(v)倫理的情報、(vi)個人的・組織的目標、(vii)誠実さ（コンプラアンス）に関する基準、(viii)準法規的基準、(ix)専門的基準、(x)熟練者による基準、(xi)歴史的／伝統的／文化的基準、(xii)科学的情報、(xiii)技術的情報、(xiv)市場的情報、(xv)政治的情報、等。 得られた情報にしたがって、評価基準を整理して記載する（縦横のマトリックスにするとわかりやすい)。さらに、それぞれの価値側面に加重点(Weights)を決めて併せて記載しておくことが勧められる。さらに、必要に応じて、基準点切り(Bar)として、特定の価値側面に関して受け入れ可能な最低基準の設定をしておく（いわゆる「足切り」)。そのほか、より洗練された方法として、段階付け(Stepping)も用いることができる。
C：部分評価（Subevaluations）	
それぞれの部分評価では、(i)事実を特定する、(ii)特定された事実に評価水準を適用して価値判断を下す、という二つの作業を行う。	
6．過程 （Process）	政策／施策／事業の成果（アウトカム）が得られるまでに起こった全ての重要な物事を明らかにし、それらの意義や重要性を評価して記述する。具体的には、(i)目標（ゴール)、(ii)計画（デザイン)、(iii)実施過程、(iv)管理（マネジメント)、(v)活動・手続き、(vi)知識習得、(vii)態度変容、(viii)他に分けて評価するとよい。
7．成果 (Outcome / Impact)	受益者が被った成果（アウトカム）を明らかにして、それらの意義や重要性を評価して記述する。成果には、(i)直接・間接、(ii)意図したもの・意図しなかったもの、(iii)直後・短期・長期の別があることに留意しつつ、(i)通常の成果、(ii)社会的成果、(iii)環境的成果に分けて評価するとよい。また、重要な成果のいくつかは、政策／施策／事業の実施中にも発生することに留意する。 なお、非常に重要な成果は予期していなかったものであることが多いので、いわゆる仮説検証の手法は適さない。したがって、あらかじめ設定された目標にとらわれずに、成果を特定する洗練された作業が要求される。
8．効率性 (Cost Efficiency)	金銭的コストと、非金銭的コストの双方を明らかにして、それらの効率性を評価して記述する。開始／維持／アップグレード／終了・閉鎖などの発展段階に分けて特定して評価するとよい。
9．比較優位 (Comparative analysis)	より少ない投入資源で類似の便益*をあげられる他の政策／施策／事業（”Critical Competitors”）と比較する。５．～７．は、政策／施策／事業と受益者の価値観等に基づく「評価基準」の対比で評価が行われたが、８．では当該政策／施策／事業と、他の政策／施策／事業の比較で評価が行われることになる。比較により、当該政策／施策／事業の意義・値打・重要性を明らかにする。 *ここで言う「便益」とは、過程、成果、効率性の各段階で得られた積極的な（プラスの）価値判断のことである。
10．持続性 （一般化可能性）	政策／施策／事業が、今後とも自立的に持続するかどうかを明らかにする（持続性)。また、他の時間／場所／職員／受益者／環境でも類似の便

(Sustainability / Generalizability)	益をあげられるかどうかを明らかにする（一般化可能性）。後者は、キャンベル（Campbell, 1969）の「外部妥当性」とほぼ同じ概念である。(i)財務面、(ii)（人的な）能力面、(iii)組織・制度面、(iv)社会面、(v)環境面、(vi)政治面、(vii)その他の条件に分けて評価するとよい。
D：結論と示唆（Recommendations and Implications）	
11. 総合化と結論	５つの部分評価を総合して総合評価を記述する。加重・合計、基準点切り、段階付け等を施して、単一の等級や順位を決定して記述する。総合化が難しい場合には、少なくとも「側面描写」（５つの部分評価を並列的に記述する）を行う必要がある。 通常は、受益者のニーズに対する政策／施策／事業の成果（現時点の実績と将来時点の予想）にもっとも焦点を当てて記述する。また、評価委託者やその他のステークホルダーの情報ニーズ（委託者の組織目標はどの程度達成された等）に応える結論も記述する。
12. 教訓と提言（可能であれば）	提言の記述は、通常考えられているほど簡単なことではない。提言を書かなくとも、評価結果を記載することが要求されている評価報告書としては何ら問題はないとされる。 提言は、ミクロレベルの提言（Micro-recommendations）とマクロレベルの提言（Macro-recommendations）に分けられる。ミクロレベルの提言は、管理運営や資器材選択に関するマイナーな改善提言であり、通常は非常に有用である。 一方、マクロレベルの提言は、政策／施策／事業の拡大・中断・廃止などに関する提言であるが、このレベルの提言を行うには、(i)当該分野に関する豊富な知識、(ii)追加資源を要するような多大な作業努力、(iii)政策レベルの意思決定者にアクセスできる位置にいること等の高度な条件が要求される。 また、今後の政策／施策／事業の計画立案に参考になる「教訓」もここで記載する。ただし、無理に書かない。
13. 結果責任の所在（可能であれば）	この記述は、万が一求められた場合で万が一それを特定するだけの能力がある場合に実施する。評価者は気軽にこの作業を引き受けてはならない。
14. 報告と今後の支援	政策／施策／事業の実施者に対する場合と、一般市民に対する場合と、議会関係者に対する場合等では、評価報告の体裁や用語のレベルはまったく違うものになるだろう。それぞれの対象に対してどのような報告を行うべきかを記述する。 また、評価が終了したあとに、受益者、利用者、職員、資金提供者等に追加で行う支援サービスについても記述する。
15. メタ評価	メタ評価とは「評価の評価」のことである。当該評価結果に関して、その利点と制約を明らかにして記述する。通常は、当該評価を実施した評価者とは別個の独立した品質管理部署がメタ評価を実施する。ただし、投入資源がない場合に評価実施者自身が行ったとしても、実施する価値は高い。メタ評価の視点は、有効性（Validity）、有用性(Utility)、信頼性(Credibility)である。

（出典）Scriven (2004) *Key Evaluation Checklist*（一部、筆者（佐々木）が変更）。

BOX 2 「評価者の基準原則によるメタ評価チェックリスト」

（Guiding Principles Checklist for Evaluating Evaluations）

全米評価学会で承認されている「評価者のための基準原則」を利用して作成したチェックリストである。以下の５項目で構成される。この基準原則は、各分野のメタ評価の基準を含んでいないので、プログラム評価チェックリスト（PES）などと併用することが勧められる。

- ◆ 系統的調査スタンダード（Systematic Inquiry）
- ◆ 作業能力（Competence）
- ◆ 誠実な態度（Integrity / Honesty）
- ◆ 人々に対する敬意（Respect for people）
- ◆ 一般・公共福祉に対する責任（Responsibilities for General and Public Welfare）

（出典） http://www.wmich.edu/evalctr/checklists/guidingprinciples2005.pdf

３　アメリカ会計検査院（ＧＡＯ）によるメタ評価の実践状況

（１）ＧＰＲＡ、ＧＡＯ、ＯＭＢの関係

政府業績成果法（GPRA）、およびその普及を推進してきたアメリカ会計検査院（GAO）、そして GPRA と予算審査を連動させようと試みるアメリカ議会予算局（OMB）の関係は複雑である。廣瀬（2006）、伊藤（2006）がこれらの関係をうまくまとめているので、別途参照されたい。

1993 年に成立した「政府業績結果法」（Government Performance and Results Act of 1993、通常 GPRA と呼称される）は、アメリカ連邦政府の各省庁に、戦略計画（Strategic Plan）と業績測定（Performance Measurement）およびプログラム評価（Program Evaluation）を義務付けた法律である。各省庁は、連邦議会との協議を経て、３年間をターゲットとする戦略計画を策定する。戦略計画を踏まえて「年次計画」（Annual plan）を策定するが、その中で、業績指標（Performance Indicators）を選定して業績目標（Performance Target）を設定することが義務付けられている（龍・佐々木 2004）。なお、年次計画は OMB と連邦議会に提出されて承認を得る必要がある。そして、各省庁は、業績目標の達成度合を測定して、「年次業績報告書」（Annual Performance Report）を策定し、大統領と連邦議会に提出することになっている（廣瀬 2006）。GPRA に基づく戦略計

画と業績測定は、1994 年度に試行期間が始まり 1998 年度から本格実施された。

GPRA の導入を OMB とともに推進したのは GAO であり、GAO は、戦略計画および各省庁が業績測定を実施するための各種指針等の策定、連邦議会の要請に基づく GPRA の実施状況に関する研究報告書の策定、GPRA そのものの有効性に関する評価の実施等を行っている。

さらに最近の動きとしては、OMB は、GPRA によって策定される戦略およびその業績と、連邦政府の予算を連動させるために、PART（Program Assessment Rating Tool）を導入したことが注目されている（伊藤 2006）。

（２）ＧＡＯにおけるプログラム評価の実際

GAO は、1921 年の設置以来、伝統的な会計検査を実施してきたが、1960 年代後半から徐々に活動範囲を拡大し、1970 年の立法府改革法や 1974 年議会予算法によりプログラム評価が正式に GAO の活動として認められるに至った。現在では、GAO の活動の大半（約 90%）が、議会の要請によるプログラム評価の実施で占められている（廣瀬 2006）。残りの 10%は GAO が独自の判断で実施するプログラム評価であるが、議会の要請によるプログラム評価同様、報告書は議会に提案して承認を得るため、報告書の記述からは GAO 独自の判断で実施したものがどれかかは判断できない。プログラム評価報告書は、通常の評価項目（プロセス評価、インパクト評価、費用便益分析等）で構成されていることが多い。

現在、GAO が行うプログラム評価は、「取り決めのための電子支援ガイド」（Electronic Assistance Guide for Leading Engagements (EAGLE)）というシステムにより報告書作成手続きが電子的に管理されている。同支援ガイドは、議会要請の受付から報告書公表までの業務を 7 段階に分けて、それぞれの段階の手続きを詳細に定めている。 7 つの段階とは、(i)業務受付、(ii)業務開始、(iii)任務分担の設計、(iv)メッセージ合意、(v)第一担当者の承認（報告書原案の作成後）、(vi)各省庁のコメント拝受、(vii)報告書の発行、である（黒田 2003）。以下にその概要を示すが、詳細については、黒田（2003）によくまとめられているので参照されたい。また参考までに GAO の組織図およびプログラム評価担当部署も示す。

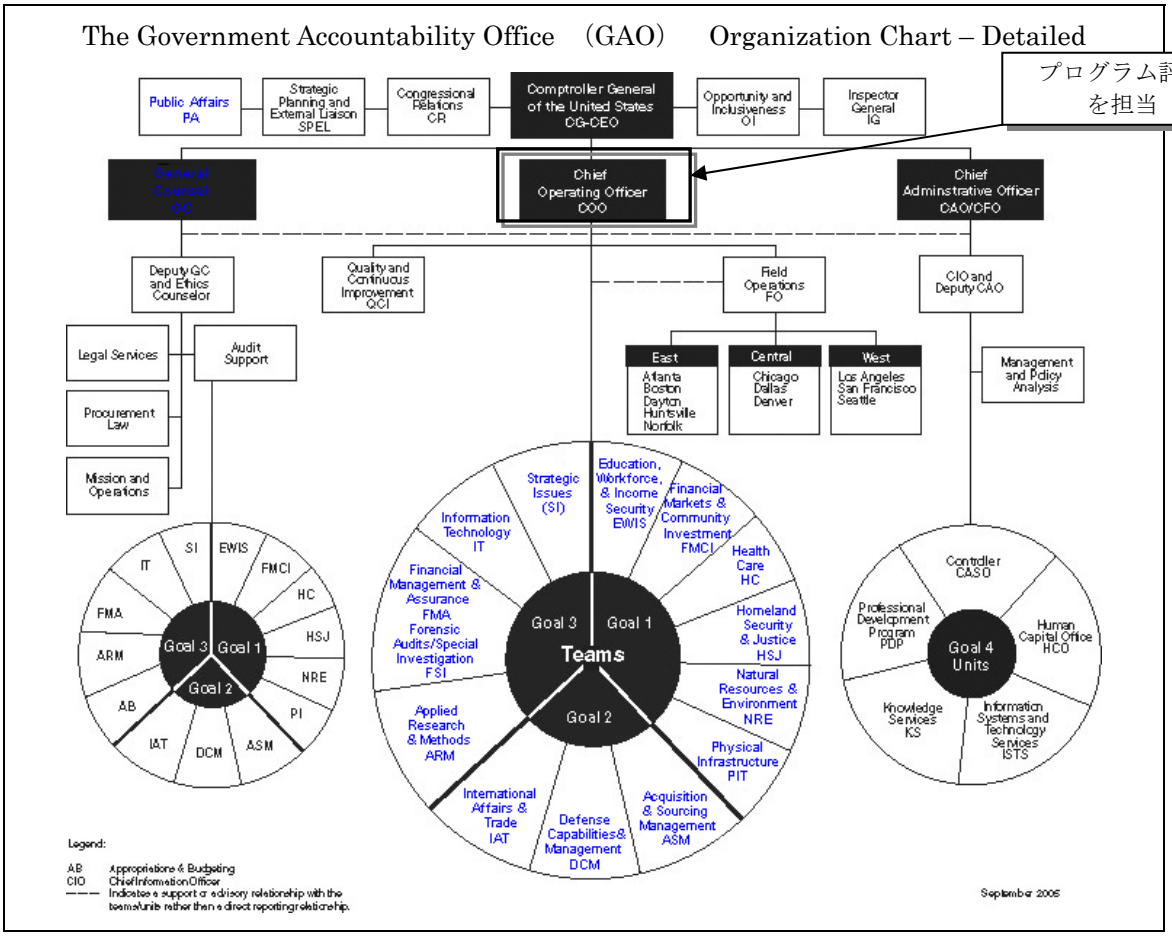
表１　ＧＡＯ「取り決めのための電子支援ガイド」の概要

（Electronic Assistance Guide for Leading Engagements (EAGLE)）

段階	具体的作業項目
(i) 業務受付	1-1　要求の受付 1-2　業務受入会議(EAM: Engagement Acceptance Meetings)への議題登録 1-3　受入にあたっての要望の評価とほかの主要な決定 1-4　業務受入会議の決定の内部伝達 1-5　業務受入会議の決定についての要望者との話し合い
(ii) 業務開始	2-1 業務の開始、2-2 要求者との会合、2-3 業務設計の開始の承認 2-4 関係省庁への通知
(iii)任務分担の設計	3-1 業務設計、3-2 設計についての合意（Design Matrix の策定を含む） 3-3 評価方法、報告書、データについての合意を得るための要求者との会合 3-4 報告書発行日の承認
(iv)メッセージ合意	4-1 データ収集と分析、4-2 メッセージの作成、4-3 メッセージについての合意、 4-4 文書形式の合意、4-5 要求者との意思疎通
(v) 第一担当者の承認（報告書原案の作成後）	5-1 報告書原案の作成、5-2 報告書原案の更新、 5-3 報告書原案の承認（第一担当者による報告書原案の審査と承認）
(vi) 各省庁のコメント拝受	6-1 第二担当者の同意 6-2 各省庁への送付 6-3 業務審査会議に対して各省庁への送付を周知
(vii) 報告書の発行	7-1 各省庁のコメントの反映 7-2 各省庁のコメントの扱いの審査 7-3 最終承認、7-4 要求者への事前通知、7-5 最終プロセス（印刷、インターネットへの掲載等）、7-6 報告書発行後の作業（少なくとも年一回、報告書が実現した便益や改善効果についてレビューする等の作業）

（出典）1. 黒田（2003）資料１より転載（一部筆者が修正）。
2. Electronic Assistance Guide for Leading Engagements (EAGLE)

図１　アメリカ会計検査院（ＧＡＯ）の組織図



（出典）http://www.gao.gov/about/workforce/orgchartdet.html

（３）ＧＡＯ内部におけるメタ評価（報告書の質的管理）の実際

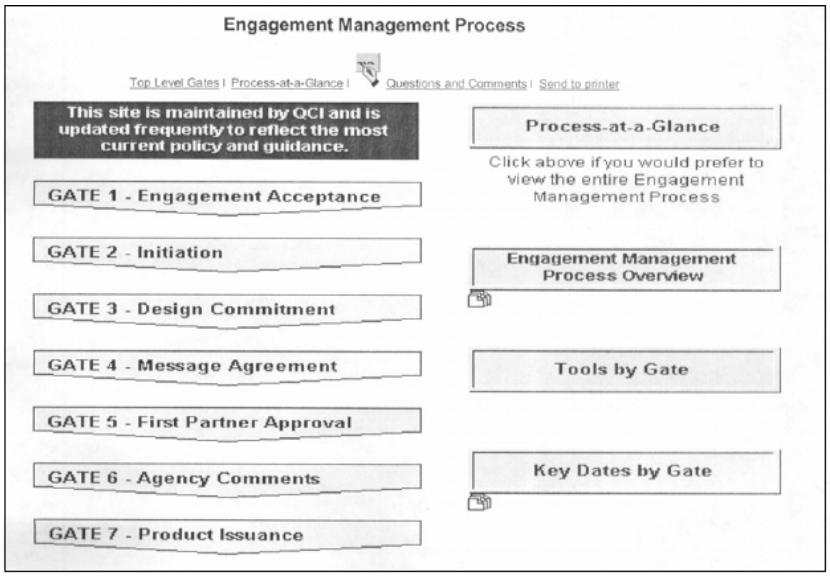
GAO は毎年、1,000 冊におよぶ報告書（Performance Audit Report を含む）を公表するが(Smith 2005)、報告書の質的管理はどのように行われているのだろうか。少し古くなるが、金本（1990）が次のように記載している。「報告書の内容の正確さを確保するために、GAO はきびしい内部チェックのシステムを持っている。この内部チェックは何層にもわたって行われるが、最も重要 なのは GAO で referencing と呼ばれている手続きである。すべての報告書は他の調査官（evaluator）の検査を受けるが、この検査は報告書の一行一行についてデータの裏付けと論理性をチェックするものである。また、この検査では調査に使ったすべての内部資料や作業メモが検査者に渡され、それらを用いて検査が行われる。この referencing 以外にも、管理職によるチェックや経済的問題に関しては内部のエコノミストによるチェックが行われ、通常は、10～12 ヶ月の調査に 1～3 ヶ月を費やし

て内部検査が行われる。」

この内部チェックの手続きを具体的に定めたのが、既出の「取り決めのための電子支援ガイド」（EAGLE）であると言える。一連の作業の中で、メタ評価（報告書の質的管理）に関係する作業は２回行われる。第１回目は、「5-3 報告書原案の承認（第一担当者による報告書原案の審査と承認）である。第２回目は、「6-1 第二担当者の同意（第二担当者の報告原案の審査と第一担当者への同意)」である。さらに、メタ評価ではないが、関係する各省庁へ報告書原稿を送付してコメントを拝受するとしている（ただし自動的に反映されるわけではなく、別途コメントの取り扱いを審査するとしている)。

では２回行われるメタ評価ではどのような基準が用いられているのであろうか。EAGLE は GAO 内部向けのイントラネットであり、外部からアクセスすることはできない（図２参照)。ただし EAGLE は、これに関して公表された二人の担当者が用いる共通の基準は、「GAO の主要価値」「専門的基準」であり、「GAO の主要価値」として、(i) 説明責任（Accountability)、(ii) 誠実（Integrity)、(iii) 信頼（Reliability）があげられている（黒田 2003)。

図２　「取り決めのための電子支援ガイド」（ＥＡＧＬＥ）のトップページ



（出典）Patrick (2006). Figure 3.4

「専門的基準」として何が用いられているかは明らかではないが、GAO が行うすべての評価は有名な「GAO イエローブック」（GAO Yellow Book. 正式名称 Generally

Acceptable Government Accountability Standards (GAGAS)）に従うことになっており(Sherman 2009)、同イエローブックが定める基準がその主要な部分になっていることは想像に難くない。ただし今回は現地調査を実施しなかったため、GAO 内部で利用されている EAGLE を直接参照することはできなかったことから、今後の現地調査で確かめられる必要がある。

表２　ＧＡＯの Yellow book で定められている基準

(Chapter 2 Auditors' ethical responsibility からの抜粋)

- 一般的基準
  - 独立性(Independence)
  - 専門的判断(Professional Judgment)
  - 高い能力(Competence)
  - 質的管理及び質的保証(Quality Control and Assurance)
- 倫理的基準
  - 公共の福祉（The public interest)
  - 専門家としての振る舞い（Professional behavior)
  - 誠実（Integrity）
  - 客観性（Objectivity)
  - 政府の情報・資源・職位の適切な利用（Proper use of governmental information, resources, and position)

（出典）GAO (2006)

（４）ＧＡＯにおけるメタ評価（評価統合）の実際

GAO が手がけたプログラム評価のいくつかは Synthesis report である。つまり、いくつかの評価報告書をレビューして統合した報告書である。著名な評価研究者であるエレノア・チェリムスキーが、GAO のプログラム評価・手法局の局長（Director）だった時期に、GAO のプログラム評価・手法課（Program Evaluation and Methodology Division）によって、The Evaluation Synthesis と題する報告書が作成されている。その中では「評価統合」の意義を次のように述べている。

「評価統合（Evaluation synthesis）は、特別のプログラムに関するクライアント（Clients）（議会等の評価委託者）の情報ニーズを満たすために行われ、いくつかの別々の報告書から、確認された事実を統合する体系的な手続きである。それは、いくつかの共

通の評価質問をあらかじめ設定して、違う時間、違い場所で、違う人々によって作成された評価報告書から評価結果を収集して統合する作業である」（GAO 1992: 1 & 6）

そして、GAO は、今までどのようにプログラムが実施されどんな効果が実現したのかを知りたいという議会の要求に対して回答を出すために、評価統合を用いてきたとも述べている。なお、評価統合の対象となるもともとの報告書は、GAO が作成した評価報告書、各省庁が作成した評価報告書であり、「可能な限り多数の評価報告書を収集する」としている（GAO 1999）。評価統合により、ある特定の介入行為（プログラム）がもたらす平均的なインパクト幅を推定することができる。また、個別の介入行為（プログラム）はそれぞれの場合でいろいろな外部要因から影響を受けるが、複数の評価結果を統合することによりそれら外部要因による影響が緩和されて、より一般的なインパクト幅を推定することが可能になるのだ。

複数の評価報告書を収集して、比較したり総合的に分析することには、ひとつの評価報告書をメタ評価するのとは別の意味がある。複数の報告書を統合分析することで初めて可能となる利点があると指摘し、主に以下の利点を列挙している。

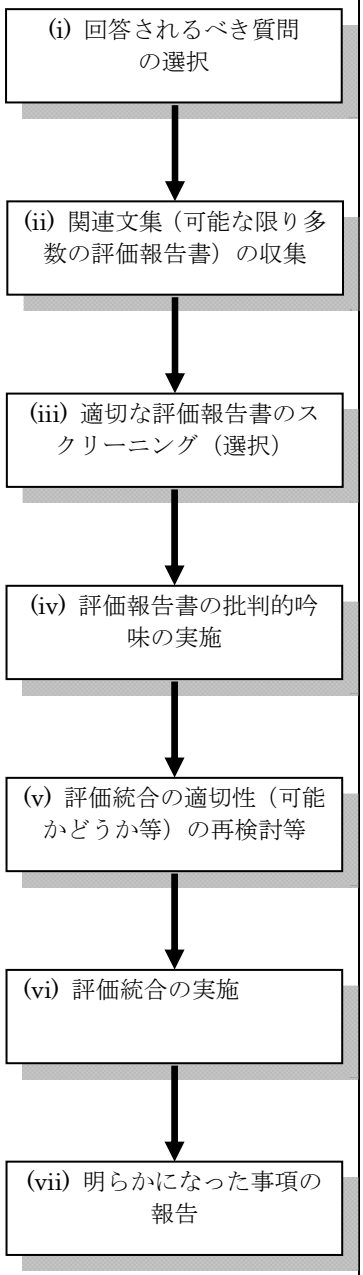
表３　メタ評価（評価統合）の利点

（ア） 受益者（Impactees）のタイプと介入のタイプを一致させて効果を説明する
（イ） 特に重要な介入行為を峻別して効果を説明する
（ウ） 矛盾する結果を説明する
（エ） 重要なアウトカムに関して「絶対的効果」のみならず「相対的効果」も説明する
（オ） 介入行為による効果の安定性を査定する
（カ） 適用したリサーチデザインについて査定する

（出典）GAO (1999)

次のページに、評価統合として適用される作業段階について図示した。また、必要な評価設問および定量的・非定量的統合手法についても列挙した。なお、評価統合でも「価値判断が行われるので、評価者は伝統的な口述によるレビューに傾きがちになるだろうがそれは間違っている」と指摘している。評価統合により、（表４で列挙したような）統合手法を駆使して、可能な限りパワフルな回答を得るべきだと述べている。

表４　ＧＡＯによる評価統合（Evaluation Synthesis）の手順



●評価統合の第一段階（(i)回答されるべき質問を選択）では通常、次の質問が選択される。

（１） 平均的な効果は何か？

（２） プログラムは、どこで誰にとって効果があったか、あるいは効果がなかったか？

（３） 当該プログラムはここで機能するのか？別の言い方をすれば、特定の場所で、プログラムを実施する場合の実践的なガイドラインはないのか？

具体的には、「実施に関する質問群」（Program operations questions）と「プログラム効果に関する質問群」（Program effects questions）が設定されることが多い。

● 評価統合の実施（iv）で用いられる統合の方法は次のとおりである。

＜定量的手法＞

―平均効果量（Average effect size）を計算する（Glass, 1981）

―統合有意検定（Combined Significance Test）を実施する（Rosenthal, 1978）

＜定性的手法（非定量的手法）＞

以下の５種類がある。

（１） シングル・ケーススタディ(Single case studies)

（２） 非定量的な集合研究（Nonquantitative aggregate studies）

（３） 定量的研究における非定量的な情報分析(Nonquantitative information in quantitative studies)

（４） 専門家判断(Expert judgment)

（５） 口述的レビュー(Narrative review of collections of research studies)

（出典）GAO (1999)

４　アメリカ行政管理予算庁（ＯＭＢ）によるメタ評価の実践状況

（１）ＧＰＲＡによるプログラム評価の義務付け

1993 年の GPRA の導入により、各省庁には、戦略計画の策定・実施・モニタリングだけではなく、プログラム評価の実施も義務付けられた。

（２）ＯＭＢによるメタ評価：ＰＡＲＴの導入

一方、OMB では、「施策の査定と格付けツール」（Program Assessment and Rating Tool (PART)）という仕組みを用いて、GPRA で義務付けられた戦略策定やプログラム評価の実施が適切に行われているかどうかを評価するとしている。各省庁で実施されたプログラム評価の結果を評価するということは、すなわちメタ評価であり、その基準がどのようになっているかを把握することが本研究にとって重要である。OMB では、各省庁を担当する職員が、メタ評価の作業にあたる。下表に PART で定められている設問と配点についての概略を述べる。

表５　ＰＡＲＴの共通設問と配点基準

セクション１：プログラムの目的とデザイン（設問数 5 問、ウェイト 20%）
プログラムの目的とデザインが、明確かつ確固たるものかを評価する。
セクション２：戦略計画（設問数 8 問、ウェイト 10%）
プログラムが長期および年次に関する妥当な目標と指標を持っているかを評価する。
セクション３：プログラム・マネジメント（設問数 7 問、ウェイト 20%）
財政面や改善努力を含め、省庁のプログラム・マネジメントを格付けする。
セクション４：プログラムの成果とアカウンタビリティ（設問数 5 問、ウェイト 50%）
目標・指標に照らして、または他の評価を通じて、プログラムの実績を格付けする。

（出典）左近（2008）

「施策の査定と格付けツール」（PART）の導入経緯とその運用については、左近（2008）が適切に解説しているので別途参照されたい。以下に概略を述べる。

導入以来 10 年余りが経過した GPRA に対する主な批判は、多数の指標と膨大なデータ（指標値の経年変化等）を産出した一方で、それらが予算要求や予算査定と必ずしも関連しているわけではないため、「政策コストと成果の関係（費用対効果）が見えにくく、結果

として政策資源の配分や決定や日常のマネジメントにおいて十分に情報が活用されていない」という点だったとされる（左近 2008）。こうした状況を改善するために導入されたのが「施策の査定と格付けツール」（PART）である。毎年、各省庁が 5 分の 1 のプログラムを評価対象として選定して、まずは省庁自身が PART を適用して評価を行うことにより評価報告書を作成して OMB に提出する。OMB では、予算審議官（Examiner）が審査および格付けを行い、その結果に基づいて OMB 内部で予算調整を進め、最終的に議会に提出する大統領予算にもその結果を添付することになっている（左近 2008）。

つまり、PART は、プログラムの目的・デザイン・戦略計画・予算の出来具合および成果が出ているかが評価されるので、プログラム評価のツールであると言える。さらに、プログラムの成果に関して行われた評価結果を、再び OMB がチェックする作業があり（後述の質問 2.6 と質問 4.5）、その点に注目すればメタ評価の手法であるとも言える。本解説では、後者のメタ評価に重点を置いて解説する。

なお、PART は 4 つのセクションに分かれており、ウェイトもそれぞれ違う。たとえば、「セクション IV:プログラムの成果とアカウンタビリティ」に対するウェイトが 50%と非常に高くなっている一方で、同セクションの平均点は 4 セクションの中でもっとも低くなっていることが注目される（他のセクションが 76-87 点であるのに対して、セクション IV は 49 点（FY2009））。

次のページに「予算編成に用いた PART の共通設問と配点基準」を掲載した。2005 年度用のものであるが、現在の 2008 年度用のものと違いはない。この中で、プログラム評価に関する記載があるのは、質問 2.6 と質問 4.5 である。どちらも、各省庁自身によってプログラム評価がなされているかどうかに関して査定することになっている。したがって、これはメタ評価（評価の質的保証）にあたる作業だと言える。

表６　予算編成に用いたＰＡＲＴの共通設問と配点基準（2005 年用）

セクション I:プログラムの目的とデザイン　20%

（プログラムの目的とデザインが、明確でかつ確固たるものかを評価）

1.1　プログラムの目的は明確か。

1.2　プログラムは、現存する問題・利害・ニーズに対応するものか。

1.3　プログラムは、連邦政府、州政府、地方政府や民間における他の取組と重複しないようにデザインされているか。

1.4　プログラムのデザインにおいて、プログラムの効果や効率性を損なうような欠点はないか。

1.5　プログラムには、その資源が意図した受益者に到達するように、あるいは、プログラムの目的に直接的に対応するように、効果的な目標が設定されているか。

セクション II:戦略的計画（Strategic plan）　10%

（プログラムが、長期および年次に関する妥当な目標と指標を持っているかを評価）

2.1　プログラムには、限られた数の具体的な長期的業績指標（プログラムの成果に関するものであり、かつプログラムの目的を意義ある形で反映するもの）が設定されているか。

2.2　プログラムには、長期の業績達成に向けて、野心的な目標やタイムフレームがあるか。

2.3　プログラムには、限られた数の具体的な年次業績指標（プログラムの長期目的の達成に向けた進展状況を示しうるもの）が設定されているか。

2.4　プログラムには、年次の業績に向けて、ベースラインと野心的な目標があるか。

2.5　すべてのパートナー（交付金の受け手、委託先、費用分担者、政府における他のパートナー等）は、プログラムの長期的および年次の目的の達成に向けて、業務を実施しているか。

2.6　適切な範囲・質を有する独立した評価が、定期的にもしくは必要に応じて、当該プログラムの改善や、有効性、問題・利害・ニーズに照らした妥当性等の評価のために、実施される予定であるか。

2.7　予算要求はプログラムの長期および年次の業績目標の達成状況に応じたものとなっているか。プログラムに必要な資源は、その予算において、完全かつ透明性のある形で提示されているか。

2.8　プログラムは、その戦略的計画における課題に対して、意義ある対策を講じてきたか。

セクション III:プログラムのマネジメント　20%

（財政面や改善努力を含め、省庁によるプログラムのマネジメントを格付け）

3.1　省庁は、時宜を得たかつ信頼性のある業績情報を、定期的に収集しているか。また、省庁はこれらの情報を、プログラムを管理して業績を改善することに活用しているか。

3.2　連邦政府のマネージャーやプログラム実施のパートナーは、プログラムの費用、スケジュール、業績等に関して責任を果たしているか。

3.3　資金は時宜を得て出されたか。また、意図した目的のために費やされたか。
3.4　プログラムは、その実施に際して効率性（費用対効果）を測定しかつ実現するための手続き（例えば、市場化テスト、コスト比較、IT 関連の改善、適切なインセンティブ等）を持っているか。
3.5　プログラムは、関連する他のプログラムの調整や協力が効果的になされているか。
3.6　プログラムは、財政面でのマネジメントを実施しているか。
3.7　プログラムは、マネジメント上の欠陥に対して、意義ある対策を講じてきたか。
セクション IV:プログラムの成果とアカウンタビリティ　50%
（目標・指標に照らして、または他の評価を通じて、プログラムの実績を格付け）
4.1　プログラムは、長期の業績目標の達成に向けて、適切な進展を示しているか。
4.2　プログラム（およびそのパートナー）は、年次の業績目標を達成したか。
4.3　プログラム実施の効率性（費用対効果）は、毎年の業績目標の追及に際して、改善していると言えるか。
4.4　プログラムの業績は、政府および民間等による他の同様の目的を持つプログラムと比べて好ましい状況と言えるか。
4.5　適切な範囲・質を有する独立した評価は、当該プログラムが有効であり、その目的を達成していると評価しているか。

注：セクション内のウェイトは OMB との協議により省庁ごとに設定される。
（出典）左近（2008）の訳を採用（一部、筆者（佐々木）が変更））。また、OMB（2008）も参照した。

（３）ＰＡＲＴにおけるメタ評価の詳細

次に、設問 2.6 と設問 4.5 をより詳細に見てみる。まず、質問 2.6 の具体的な質問は以下のようになっている（OMB 2008: 30-32）。とくに注目される点としては、評価の「質」を確保するために、厳格な手法が用いられることが推奨されていることと、独立性が重視されていることである。

表 7　設問 2.6 の詳細

2.6 適切な範囲・質を有する独立した評価が、定期的にもしくは必要に応じて、当該プログラムの改善や、有効性、問題・利害・ニーズに照らした妥当性等の評価のために、実施される予定であるか。

目的：パフォーマンス情報に関するギャップを埋めるために、バイアスのない評価を定期的にあるいはニーズがあるときに参加できることを保証する。

YES のための要素：YES のためには以下について証拠を示し明確に説明できることが必要である。

✓　プログラムは、以下に記載された基準を満たすような評価を実施すべきである。

―高い質

―適切な範囲

―バイアスがない（中立である、独立である）

―プログラムの改善を支援するために定期的に実施される

(1) 質（Quality）：評価は、プログラム効果に関する情報を提供するために、十分に厳格でなければならない。省庁は、（選択した評価手法によって）最も厳格な証拠を産出する評価手法を選択したことを説明できねばならない。たとえば、ランダム化比較試験などは、インパクトを特定することに特に向いている。他のタイプの評価アプローチ、たとえばよく計画された擬似実験デザインなどは、プログラムのインパクトに関して有用な情報を提供するだろう。

(2) 範囲（Scope）：評価は、プログラムの一側面ではなく、プログラム全体の効果を総合的に評価していなければならない。

(3) 独立性：独立であるためには、利益相反（Conflict of interest）がなく、かつバイアスがない（＝不偏不党の）者によって評価が行われなければならない。もし省庁やプログラムが第三者に評価委託する場合には、その第三者は十分に独立的でなければならない。評価が、省庁の検査官（Inspector General）あるいはプログラム評価担当部署によって行われる場合には、それらは独立であるとみなされる。

(4) 頻度：プログラム評価の情報の鮮度を保つために、定期的に更新されねばならない。

エビデンス／データ：エビデンスとして、プログラム評価計画（実施日程）、および評価概要（評価のタイプ、範囲、質、独立の評価者を選択する基準等）が提出されねばならない。

（出典）OMB（2004）

次に、設問 4.5 の具体的な質問は以下のようになっている（OMB 2008: 55-60）。基本的に質問 2.6 を踏襲しているが、事後評価に限定しており、実際に厳格かつ客観的な手法に

よって評価が実施されたかどうかに注目している。

表８　設問 4.5 の詳細

4.5 適切な範囲・質を有する独立した評価は、当該プログラムが有効であり、その目的を達成していると評価しているか。

目的：独立かつ総合的な評価が、プログラムは効果的かどうかを判断していること。この質問は、特に、定量的な実績測定方法を決定することが本質的に難しいプログラムにとって重要である。

YES のための要素：YES のためには以下について証拠を示し明確に説明できることが必要である。

✓ 独立かつ総合的な評価が、プログラムは効果的かどうかを判断していること。評価は、質問 2.6 で定義されている質的基準、範囲基準、独立基準の３つの基準に従うことが求められる。

この質問では「Not applicable」は取り得ない。全てのプログラムは、質、範囲、独立の各要素を満たすように評価を実施せねばならない。詳しくは 2.6 を見よ。

エビデンス／データ：エビデンスとして、学術機関、リサーチ機関、政府契約、その他の独立機関、GAO、あるいは査察官（Inspectors General）が関係した評価の結論の議論の要約が提出されるべきだ。なお、質問 2.6 とちがって今後実施されるであろう評価計画はエビデンスにはなりえず、実際に実施された評価結果だけがここではエビデンスに該当する。

（出典）OMB（2004）

（４）ＰＡＲＴにおけるメタ評価の詳細

さて、質問 2.6 で言及された「厳格な評価」とは何を指すのであろうか。OMB の考えるプログラム評価は、「何がプログラムの効果に関する強力なエビデンスを構成するか？（What Constitutes Strong Evidence of a Program's Effectiveness?）」で詳細に解説されている（OMB 2004）。それよると次の５種類があげられており、エビデンスが強力である順番（＝厳格である順番）に並べられていると考えられる。

表９　ＯＭＢ「何がプログラムの効果に関する強力なエビデンスを構成するか？」の概略

●ランダム化比較試験（Randomized Controlled Trials（RCT）） ランダムアサインメント（無作為割当）によってプログラムを適用するグループと適用しないグループに分けることにより、純粋にプログラムの効果を測定する評価デザイン。 ●Direct Controlled Trials 効果に影響を及ぼすと考えられる外部要因を評価者が直接管理することができる場合に、直接管理して影響が及ばないようにしてから、プログラムの効果を測定する評価デザイン。宇宙開発や兵器開発で用いられる。 ●Quasi-Experimental ランダム化を用いないけれども近似的に比較グループを構築して、プログラムの効果を測定する評価デザイン ●Non-Experimental Direct Analysis プログラムを適用したグループだけを観察する評価デザイン。代表的なものは「事前・事後比較デザイン」 ●Non-Experimental Indirect Analysis 独立した専門家パネルを設置して専門家が直接観察により評価を行う。

（出典）OMB（2004）

上記の図にあるように OMB のガイドラインは、ランダム化比較試験をベストな方法として推奨していることが分かる（GAO 2005: 27）。

なお、OMB のメタ評価者（OMB examiner）と省庁側の評価担当者（Federal evaluation officials）の間で、評価の質をどのように査定するかという一連の議論が行われている。まず、2004 年夏に OMB は Interagency Program Evaluation Working Group を立ち上げて一連の議論を行った。省庁側の評価担当者は OMB が「厳格な評価」を過度に狭く定義していることに懸念を表明している（GAO 2005: 28-29）。

その後、2005 年春に再び集まり、彼らがもっとも適切だと思う評価アプローチについてまとめた。それが次の表 10 である。

表 10　省庁側の評価担当者の見方：プログラム効果の評価のためのデザイン

● プロセス／アウトカムのモニタリング・評価 事前に設定された目標（値）と比べる。 ● 擬似実験モデル：単一グループ 実施グループのみで、事前事後比較を行う。 ● 擬似実験モデル：比較グループ 近似した比較グループを設定して、比較を行う。 ● ランダム化比較試験 ランダム化により介入効果を測定する。

しかし、OMB は 2007 年の段階でも、ガイダンスの大幅な見直しは行っておらず、プログラム効果を評価するためのベストなデザインとしてランダム化比較試験を推奨している。

５　アメリカ教育省（ＤＯＥ）によるメタ評価の実践状況

アメリカにおけるメタ評価の利用として最も注目される動向の一つがアメリカ教育省による「エビデンス（科学的根拠）に基づく研究」である。以下に概略を述べる。なお、この動向については、田辺（2006）がよくまとめているので、別途参照されたい。

（１）ＮＣＬＢ法と「エビデンス（科学的根拠）に基づく研究」

ブッシュ政権はさまざまな改革に取り組んだが中でも重視したのが教育改革である。アメリカの教育レベルの長期的凋落傾向に危機感を抱いた同政権は、限られた予算を有効に活用するために、確かに効果があると認められた教育施策を明らかにしてそうした施策に重点的に予算配分することを目指した。

こうした発想を具体化したのが 2002 年に超党派の支持を受けて成立した「子供を一人も落ちこぼれにしない法案」（No Child Left Behind Act of 2001：NCLB 法）である。アメリカの教育水準を引き上げることを目的とし、学区や学校に大幅な裁量権を認めることと引き換えに、子供の学力向上という「成果」を厳しく問う内容であった。そして「成果」を挙げた学区や学校には予算を増額するとともに、「成果」をあげた取り組みを全米の学区や学校に普及することを目指していた。

ではどうやって「成果」を明らかにするのであろうか？ここに NCLB 法の特徴がある。

NCLB 法は、「エビデンスに基づく研究」（Evidence-based Research）を重視するとしている。そして、「エビデンスに基づく研究」で有効性（つまり「成果」）が確認された施策のみに連邦予算を配分するように求めている。

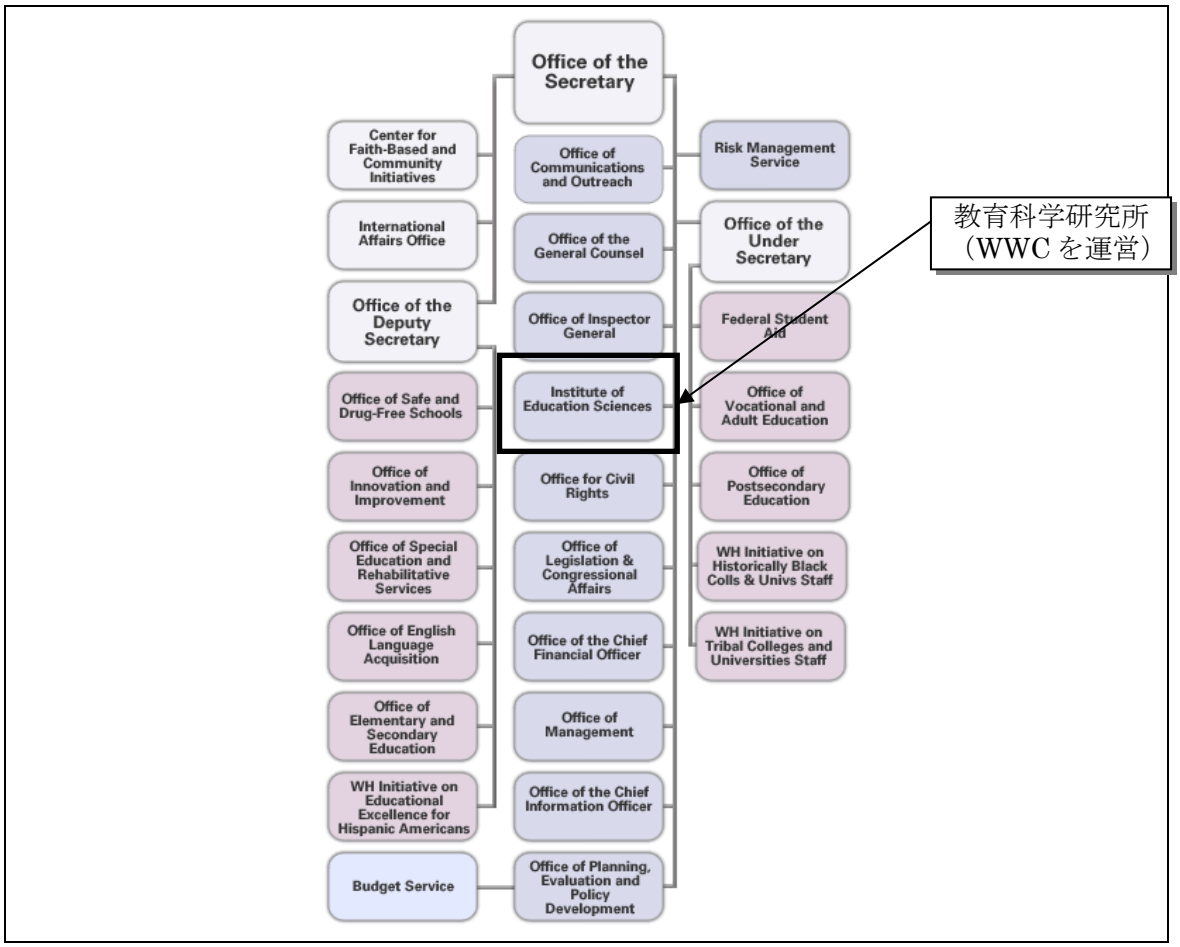
それでは「エビデンス」とは何を指すのであろうか。NCLB 法 9101 条では、「教育活動や施策に関する妥当で信頼できる知識を得るための厳格・体系的かつ客観的な手続きの適用を含む研究」であって、具体的には、観察や実験に基づく実証研究、仮説検証を伴うデータ分析、実験・疑似実験デザインを用いた研究などが含まれるとされた（田辺 2006）。ただし、同条項の定義は必ずしも明確ではない一方で、連邦予算は「エビデンスに基づく研究」のみに配分されることが求められたことから、教育現場で疑念と混乱が広がったとされる。

この状況を受けて、アメリカ教育省により、「厳密なエビデンスによって裏付けられた教育実践の識別と実施：ユーザーフレンドリーガイド」（Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide）が策定されて公表された。同ガイドでは、エビデンスを産出するための方法や制約について詳細に解説されているが、そのなかで最も厳格なエビデンスが得られるのは、ランダム化比較試験（いわゆる RCT）だと明記された。

（２）アメリカ教育省によるメタ評価の仕組み：What Works Clearing House

エビデンスに基づく教育を普及させるためには、蓄積された情報を体系的に整理して、教育関係者に提供する仕組みが必要である。その必要を満たすために、教育省が創設したのが、What Works Clearing House（WWC）という仕組みである。邦訳すれば「何が機能するのかの情報拠点」ということになろう。WWC は、教育省傘下の教育研究所（Institute of Education Science）がキャンベル共同計画（Campbell Collaboration）と共同で開発・運営を行っている（教育省の組織図（図３）を参照）。WWC は、教育に関わる政策、施策、活動、製品の有効性について、最新で質の高いレビューを提供し、エビデンスに関する信頼できる情報源となることを目的としている（田辺 2006）。

図３　教育省の組織図



（出典）U.S Department of Education のウェブサイト（http://www.ed.gov）

（３）ＷＷＣのハンドブックに記載されたメタ評価の手順

WWC では、ランダム化比較試験などを利用して作成・提出される膨大な評価報告書のなかから、特定の施策に関する報告書を選定して収集し、それぞれの評価報告書に関してメタ評価を行う。そして基準を満たした複数の報告書に記載された有効性の度合いの組み合わせにより、当該施策の一般的効果を「好ましい効果」（Positive effect）から「否定的な効果」（Negative effect）まで６段階で結論するとしている。そして選定から最終結論を得るまでのメタ評価の過程を公開し、教育関係者に対して、何が機能し何が機能しないのか、の情報を提供している。教育関係者はそれを参照して、機能するとされた施策を、自分の学区・学校で適用するための判断材料にすることになる。

WWCの「手続き・基準ハンドブック」では、メタ評価のプロトコルを次のように定めている。なお、同ハンドブックでは「メタ評価」という用語を使用しているわけではなく、

レビューという一般的な用語を使用していることに注意する必要がある。

表 11　「手続き・基準ハンドブック」に記載されたメタ評価の手順

1．レビュー対象分野の特定
(1) レビュー対象分野の特定
(2) 調査報告書の検索
(3) （レビュー対象分野としての）的確性の評価
2．調査報告書レビュー作業
(1) レビュープロセス（The review process）
(2) エビデンスの基準（Evidence standards）
(i) 研究のデザイン（Study Design）
(ii) サンプル集団からの脱落度合（Attrition）
(iii) RCT による同質性の確保（Establishing Equivalence in RCTs）
(iv) 対抗要因（Confounding Factor）
(v) 基準を満たさなかった理由（Reasons for Not Meeting Standards）
(vi) 訂正・修正（Correction and Adjustments）
3．レビュー結果のとりまとめ
(1) ドラフト報告書の作成
(2) クオリティ保障レビュー（Quality Assurance Review）の実施
(3) 独立評価スタッフと外部専門家によるレビュー
(4) 最終報告書の作成と公表

（出典）Institute of Education Sciences, Department of Education (2008)

以上の手続きに基づいて、個別の評価報告書に関してメタ評価を行うわけであるが、複数の報告書に記載された有効性の度合いの組み合わせにより、当該施策の一般的効果を以下の通りの 6 段階のいずれかを用いて結論するとしている。

表 12　WWCの効果レーティング

好ましい効果（=正の効果）(Positive Effects)	好ましい効果を示す強力なエビデンスがあるとともに、それを上回るような反対のエビデンスがない。 ・ 二つ以上の研究が統計学的に有意な正の効果を示しており、少なくともそのうちの一つが、WWCエビデンス基準を満たしている。 ・ 統計的に有意な負の効果を示す研究がない。
部分的に好ましい効果（=部分的に正の効果）(Potentially Positive Effects)	好ましい効果を示すエビデンスがあるとともに、それを上回るような反対のエビデンスがない。 ・ 少なくとも一つの研究が統計学的に有意な正の効果、あるいは本質的に重要な正の効果を示している。 ・ 統計学的に有意な負の効果を示す研究がないとともに、本質的に重要な正の効果を示す研究と同数あるいはそれよりも少ない数の研究が中間的な効果（Intermediate effects）を示している。
混合的な効果（Mixed Effects）	一貫しない効果のエビデンスがある。 ・ 少なくとも一つの研究が統計学的に有意な正の効果あるいは本質的な重要な正の効果を示すと同時に、少なくとも一つの研究が統計学的に有意な負の効果あるいは本質的に重要な負の効果を示している。ただしそうした負の効果を示す研究の数が、そうした正の効果を示す研究の数を超えていない。 ・ 少なくとも一つの研究が統計学的に有意な効果あるいは本質的な重要な効果を示すと同時に、それらの研究の数よりも、中間的な効果を示している研究の数の方が多い。
効果は明確ではない (No Discernible Effects)	・ 正の効果に関する肯定的なエビデンスがない（No affirmative evidence of effects）。 ・ 統計学的に有意あるいは本質的に重要な効果を示す研究がない（正の影響でも負の影響でも）。
部分的に否定的な効果（=部分的に負の効果）(Potentially Negative Effects)	否定的な効果に関するエビデンスがあるとともに、それを上回るような反対のエビデンスがない。 ・ 少なくとも一つの研究が統計学的に有意な負の効果、あるいは本質的に重要な正の効果を示している。 ・ 統計学的に有意な負の効果を示す研究がない。あるいは統計学に有意な負の効果あるいは本質的に重要な負の効果を示す研究の数が、統計学的に有意な正の効果あるいは本質的に重要な正の効果を示す研究の数よりも多い。
否定的な効果（=負の効果）(Negative Effects)	否定的な効果に関して強力なエビデンスがあるとともに、それを上回るような反対のエビデンスがない。 ・ 二つ以上の研究が統計学的に有意な負の効果を示しており、少なくともそのうちの一つが、WWCエビデンス基準を満たしている。 ・ 統計的に有意な正の効果を示す研究がない。

（出典）Institute of Education Sciences, Department of Education (2008)

６　結論と考察

アメリカで提案されて普及した「メタ評価」の本来の意味は、厳格な方法・デザインによって質の高いエビデンスが得られたかどうか、そして得られたエビデンスによって論理的・合理的な結論が導出されたかどうかを検証する作業ということである。また、その意味が拡大し、メタ分析に近い「評価統合」という意味合いで用いられる場合でも、まずは、複数のもともとの評価報告書の質がチェックされることになる。本章では、アメリカにおけるこうしたメタ評価の実践を、GAO、OMB、そして最近最も注目されているアメリカ教育省の事例を用いて分析した。日本における政策評価の実践に対しては以下の示唆を得ることができる。

（１）「日本版メタ評価チェックリスト」の検討・策定

本章で詳細に解説したいくつかのチェックリストを踏まえて、日本の状況に合うメタ評価チェックリストを総務省が検討・策定して公表することが勧められる。総務省自身が使用するほか、他の省庁でもそれらを利用することにより、自身の評価報告書の質をみずからチェックすることができるようになる。また、公益法人や学校法人などでも広く利用されることが見込める。このことを通じて、日本における政策評価の質が高まると考えられる。

（２）エビデンスの質に関する検討

一方で、何が質の高いエビデンスなのかの議論は、アメリカでも未だに論争が続いており、エビデンスの質やそれを得るための評価デザインの厳格さの順位を示すようなリストを作成して公表することには慎重な議論を要する。OMBやアメリカ教育省の例に見られるようにランダム化比較試験（実験デザイン）が最も望ましいという見解が強くなっている一方で、それに反対する評価研究者も多数おり(Donaldson & Christie 2005)、今後の議論の展開を見守る必要がある。

（参考）　アメリカにおける評価人材育成

GAOにおける職員の採用基準は変化してきたとされており、以前（1960年代まで）は、会計士の資格を持つものが多かったが、1970年代以降は、公共行政学、公共政策学、経済学、心理学、社会学等を専攻した者に採用対象を拡大してきた（黒田2003）とされる。

アメリカでは大学院教育によって評価人材が供給されている。とくに、公共政策学大学院が主な人材供給源となっている。公共政策学修士課程（M.P.P: Master of Public Policy）で教えられる科目は、どの大学院でもほぼ同一である。通常はコア科目と選択科目の２段構えになっている。コア科目としては、以下の５つの科目で構成されていることが普通である。なお、一部の大学ではこれに加えて、費用便益分析が必修である場合がある。

―ミクロ経済
―財務管理
―統計学（基礎編）
―マネジメント
―公共政策（基礎編）

上記科目は、通常、公共行政学修士（M.P.A. Master of Public Administration）と公共政策学修士（M.P.P. Master of Public Policy）の２つの専攻で共通のコア科目であり、公共政策学修士課程ではさらに以下のような選択科目を履修することになっている場合が多い（ただしこの限りではない）。

―公共経済・財政
―統計データ分析（重回帰分析、多変量解析）
―プログラム評価
―政策分析
―特定分野の選択科目（教育、保健、行政一般、国際開発等）

さらに、National Association of Schools of Public Affairs and Administration (NASPAA)という団体が、一定の基準を満たした公共政策・公共行政大学院に対して「認証」（Accreditation）を行っており、その認証制度が質的保証の役割を果たしている。主要な公共政策大学院はすべてNASPAAの認証を獲得しているが、認証を得るためには一定の基準が設定されている。そのなかで、カリキュラムに関する基準があり、それを満たすためには上記のようなスタンダードな科目を揃えることが必須になっている。こ

うした事情を受けて、主要な公共政策大学院の科目構成はほとんど同一となっているのだ。また、それぞれの科目で使用される教科書もほぼ定まったスタンダードなものがある。たとえばプログラム評価であれば、Rossi, Lipsey & Freeman (2006). *Evaluation: A Systematic Approach*が幅広く用いられているし、別の教科書を使う場合（例えばWeiss, C. (1999). *Evaluation 2nd edition*）を使う場合でも内容はほぼ同じである。したがって、どの大学院を卒業しても同じ技術と同じ知識を身につけた人材が輩出されることになっている。その知識の例が、ロジックモデリングや実験デザイン・疑似実験デザインであり、公共政策大学院を卒業して修士号を保持する者で、これらの概念と技法を知らないものはいないと言っても過言ではない。そのように同じ技術・同じ知識を身につけた者が入省してくることが、GAOによるプログラム評価の質を保つ大きな要因になっている。

（参考文献）National Association of Schools of Public Affairs and Administration (NASPAA) (2008)、および主要な公共政策大学院のホームページ