査読付き論文

評価の質*－評価書の事後的分析の試み

益田　直子**

（立教大学法学部特任准教授）

1. はじめに

評価の質が重要であることに異論はないだろう。日本においては「政策評価に関する基本方針」（2001年12月28日閣議決定）[1]や「政策評価分科会の当面の活動の重点～政策評価制度の発展に向けて～」（2004年4月30日）が，評価の質の向上を求めている。しかし，評価の質の向上について特別に採り上げ，議論し，具体的取組を定め，実行しているようには見えない[2]。また，日本における評価研究においても，評価の質の重要性の指摘は見られるものの[3]，質とは何か，実際はどのようになっているか，といった研究は見受けられない。本論文は，実務及び研究上のこれらの欠如を一部ではあるが埋めることを目的としている。

これらの欠如に何か理由はあるのだろうか。評価の質の保証に関する各国及び国際機関の取組をまとめたシュワルツとメインは，最高会計検査機関を例外として[4]，政府機関の多くは評価の質を比較的気にかけていないという（Schwartz and Mayne, 2005b, p.315）。そして彼らは，評価の質を保証するシステムの構築を妨げていると考えられる要因として，バイアスのある評価書を作成するように仕向ける政治的組織的圧力，質を自主管理するための人材の不足，基準の設定と実施の難しさ，評価情報をつくる経験の少なさ，評価で使うデータの収集に必要な訓練の不足，そして評価情報を組織活動に役立てることに高い価値を置いていない組織においては評価情報の質や信頼性に関心が払われにくい点を挙げている（Mayne and Schwartz, 2005, pp.10-13& Schwartz and Mayne, 2005b, pp.314-316）。このように，評価の質は，政治，評価能

* 2014年3月17日受付，5月28日受理。2名の匿名の査読者の方々より思慮深く鋭いコメントを頂戴し，論文の内容を改善させることができた。また，立教大学社会情報教育研究センター助教の廣瀬毅士先生より統計分析に関する助言を頂いた。記して感謝の意を表したい。もとより本文への反映は筆者の責任に帰する。

** 2005年4月日本学術振興会特別研究員（～2007年3月），2005年8月アメリカ・ペンシルベニア州立大学客員研究員（～2006年8月），2008年2月総務省行政評価局一般職任期付職員（～2010年1月），2008年3月東京大学大学院法学政治学研究科博士課程修了（博士（法学）），2012年4月立教大学法学部特任准教授（現職）。日本行政学会，日本政治学会，アメリカ評価学会に所属。主な著書は，『アメリカ行政活動検査院（The U.S. Government Accountability Office）―統治機構における評価機能の誕生』（2010年2月，木鐸社）など。

1) 「I-3-ア」，「III-1」を参照。

2) 但し，『政策評価の実施に関するガイドライン』（2005年12月16日政策評価各府省連絡会議了承）は，その目的を「評価の質の向上」とは明記していないが，「法に基づく政策評価の円滑かつ効率的な実施のための標準的な指針を示したものである」としている。

3) 山本清（2004），山谷清志（2007）。

4) シュワルツとメインによれば，最高会計検査機関（又は会計検査院）は，通常，組織内部において報告書の質を保証するための仕組みをもつという。また，その考えられる理由として，最高会計検査機関にとって報告書が組織の存在理由（レゾンデートル）であるため，報告書の信頼性が組織の成功にとって不可欠であることを挙げている（Schwartz and Mayne, 2005b, p.316）。

力，評価の利用など他の要素と分かちがたく結びついていることがわかる。日本の現状もこれらの要素と関わっていることが考えられる。

例えば，日本においても評価の利用は主要な課題となっている。2001 年に政策評価制度が導入されて以降，多くの評価情報が各府省によってつくられてきた。法律に基づき政策評価の実施が始まった 2002 年度から 2012 年度までの各府省における政策評価実施件数の合計は，67,014 件である[5]。しかし，評価結果が行政活動に活かされないことから，「評価疲れ」という言葉が聞かれるようになって久しい[6]。評価研究においては，評価の利用を促す要因の代表的なものの 1 つとして「評価の質」を捉えている[7]。一方，「評価の質」の基準はさまざまであるが，シュワルツとメインは，評価の質の良さに関する基準を，①評価書の質，②評価書を作成する過程の質，③評価書の有用性の 3 つの種類に分けている（Schwartz and Mayne, 2005b, pp.304-305）。つまり，「評価の質」と「評価の利用」は相互に関連している。だからこそ難問であるため[8]，日本において実務及び研究上扱いにくかったことも考えられる。

本論文は，評価の質に関する現状分析の第一歩として，「評価の質への具体的取組の不足は，評価書間の質の不均一性をもたらしている」という仮定に基づき分析をする。共通の基準や指針も持たずに，又はそれらに近いものがあったとしてもその解釈と実際の評価活動への適用について各評価担当者間で異なっていたり，又は必要な訓練も受けずにそれぞれの評価書がつくられていれば，評価書間の質は不均一になるはずである。このことは問題を孕んでいる。質の不均一性は評価情報の信頼性を喪失させ利用への障害となりうる。他方，質が不均一かどうかも分からず，質の低い情報が利用された場合には，意思決定に悪影響を与える危険性を高めることになる。よって，本論文は，質の不均一性（ばらつき）の存在を確認し，質に影響を与えると考えられる要因を挙げることにより，質の向上についての議論を喚起することを目指すものである[9]。

2. 評価の質に関する先行研究

2. 1. 研究アプローチ

評価の質に関する研究は多様な視点から質について論じてきた。トロキムとビスコは評価の質に関する研究を，①良き手法に関する研究（サンプリング，測定，調査設計，データ分析），②評価書の質を判断する研究，③評価の質の向上のための基準の開発に関する研究の 3 つに分けた（Trochim and Visco, 1985, pp.93-94）。シュワルツとメインは，評価情報の質を向上するために使われてきた多くの異なるアプローチを，次の 4 つに分類している（Mayne and Schwartz, 2005, pp.8-10）。①構造的アプローチとは，基準や指針の開発と普及，及びそれらの実施に関するアドバイスの提供を行うことを指す。②形成的アプローチとは，評価情報をつくる過程で評価情報の質を管理することを指す。③総括的アプローチとは，つくられた評価情報を事後的に評価することを指す。④制度的アプローチとは，評価情報をつくり出している制度や手続きを調査することを通じて評価情報の質を向上させようとすることを指す。

---

[5] 総務省『政策評価等の実施状況及びこれらの結果の政策への反映状況に関する報告』における評価実施件数の合計値を 2002 年度から 2012 年度（2014 年 2 月現在，最新の報告）まで累計した数字。

[6] 2005 年 12 月 7 日総務省主催の政策評価フォーラムにおける新村保子委員の発言を参照。

[7] Cousins and Leithwood (1986) &GAO (2013)

[8] 評価研究の代表的教科書の 1 つである Wholey et al. (eds.) (2010) は，評価活動における 4 つの課題を示しており（pp.668-673），「評価の質の管理」と「評価結果の利用」はそれに含まれている。

[9] 「一定水準以上の質が確保されていれば，質が多少均一でなくても問題ではない」，という考えもありうるだろう。しかし，本論文では「一定水準」の決定は行わない。それは，本文「2.5.」にある通り，関係者間の合意が必要になると考えるためである。

2. 2. 定義

評価の質に関するさまざまな定義が示されてきた。クックシーとカラチェリ（Cooksy and Caracelli, 2005, p.35）による手際よいまとめによると，有用性，実行可能性，適切性，正確性を使って定義したり，他に論者によっては，透明性，バランス，関連性，信頼性，妥当性，正統性，文化能力，体系性といった用語をそれぞれ用いて評価の質を説明している。評価を行う目的や使用すべき評価方法に関する多様な考え方に基づき評価活動は行われており，評価の質に関する異なる見方は，こうした評価活動の多様性を反映している。

2. 3. 共通の基準

以上の通り研究アプローチも定義も多様であるが，共通の基準や指針を示す努力もなされてきた。アメリカの評価活動に従事する人々に広く知られている評価活動の指針として，『プログラム評価基準（The Program Evaluation Standards）』がある。第1版（1981年）と第2版（1994年）においては有用性，実行可能性，適切性，正確性の4つの特性に基準を分けていた。最新版の第3版（Yarbrough et al., 2010）では「評価のアカウンタビリティ」を加え，5 つの特性に変更している。この変更は，体系的な評価の評価（メタ評価）を通じた評価自身の説明責任の確保の必要性が強調されていることを反映している（Yarbrough et al., 2010, p. xiv&Part V）。この特性は評価の質の改善と強く関係している。

作成された評価書の質に関する基準に焦点を当てると，シュワルツとメインは，次の5つの要素が良質な評価情報をもたらすとしている。①よく定義づけられた評価範囲（評価目的や範囲の明確性），②評価基準の有効性，③正確なデータ（収集データの妥当性と信頼性），④理論的に確かな分析（確かな方法論に基づいたデータ分析），⑤実証的かつ偏見のない結果（評価結果が収集された証拠から説明でき，また，偏見なしに客観的な方法で公表されること）である（Schwartz and Mayne, 2005b, pp.304-305）。他に，実践的な基準として，「連邦政府の評価者たち（Federal Evaluators）」と呼ばれる，アメリカ連邦政府で評価活動に従事する職員が構成員となる非公式なネットワークが公表しているチェックリストがある。このネットワークは，1999年にアメリカ行政活動検査院（又は政府監査院）の評価手法担当官が中心となって作ったものであり，評価の方法論・政策・実践に関する情報の共有を目的としている。2013年現在，全省等から約900人が参加している。このネットワークが2006年に示した「プログラム評価の質と有用性を評価するためのチェックリスト（Checklist of Questions for Assessing the Quality and Usefulness of a Program Evaluation）[10]」の詳細は，「3.2.」で説明するが，先に紹介したシュワルツとメインによる，良質な評価情報をもたらす5つの要素と同種の項目はこのチェックリストにおいても含まれている[11]ことから，評価書の質の基準としては学問的にも妥当な内容であるとみなすことができる。

[10] Federal Evaluators (2006) 内に示されている。

[11] 5つの要素の①はチェックリストの「評価目的の明確さ」と，②はチェックリストの「評価設計の適切さ」と，③はチェックリストの「データ収集者の選定と訓練の適切性」・「データ収集者間の信頼性確保の手続きの確保」・「データ収集過程における不足や課題の存否」と，④はチェックリストの「統計手続きの明示と適切性」・「評価設計の明確さ」・「評価設計の適切さ」と，⑤はチェックリストの「結論の確かさ」と主に関連している。

2. 4. 近年の研究

近年にも新たな動きがあり，古くて新しいテーマであることが分かる。2010 年のアメリカ評価学会年次大会のテーマが「評価の質」であった。同大会時の学会長であったクックシーは，2012 年に共著で「評価の質に及ぼす影響」という論文を発表している。同論文では，評価の質を定めるいかなる基準を使うにしても，その基準に合うのは難しいことが多いという。そして，質を達成することは，少なくとも部分的には，評価者の能力，評価活動の背景（評価政策や政治的背景等），そして評価者が専門家集団に参加をすることを通じて得られる支援が合わさった結果である，と論じている[12)]。

日本においては，総務省行政評価局による委託調査として 2009 年 3 月に『諸外国における政策評価のチェックシステムに関する研究』が公表されている。評価の品質確保のための評価基準についてメタ評価[13)]の事例を紹介している。

2. 5. 小括

研究アプローチ，定義，共通の基準に関する先行研究からは多様性が示されている。しかし，本論文では日本における評価の質に関する現状分析の第一歩として，完成された評価書を対象に事後的に分析（総括的アプローチ）をする。言い換えれば，評価書を作成する過程の質を分析するものではない。また，評価書の利用について分析するものでもない[14)]。事後的な分析には評価活動に従事するアメリカ政府職員のネットワークが公表するチェックリストを，評価の質を測る基準として利用する。その理由は，先行研究の「2.2.」の説明にもある通り，評価の質は評価活動の背景にある考え方を踏まえて定義されるものであり，その意味では評価活動に従事する政府職員等との合意の上で評価の質の定義と質を測る基準を設定することが望ましい。しかし，日本においてはそうしたものが見当たらないため，また「2.3.」の説明にもある通り同チェックリストは学問的にも妥当な内容を備えているとみなしうることから，他国ではあるが日本の政策評価制度との類似性も見られる制度を持つアメリカの政府職員によるネットワークが公表したチェックリストを，分析の意図と実効性を考慮し修正したものを基準として使うこととする。

3. 方法

本論文は，完成した評価書の質を，設定した基準に基づき事後的に測るものである。それによって，評価書間の質の均一性又は不均一性について確認をし，その要因を探ることを目的としている。

3. 1. 分析対象

分析対象となる評価書は，総務省行政評価局が作成した統一性・総合性確保評価である。統一性・総合

---

12) Cooksy and Mark, 2012, p. 80

13) メタ評価については，拙著（2006）「メタ評価」（「I.2 政策評価の概念とメタ評価」内），大山達雄（編）『公共政策評価の理論と実際』，現代図書，25-27 頁も参照されたい。

14) 前述の通り，評価の質は評価の利用等他の要素と関連しており，またそれらの各要素の中には評価研究における理論的柱を成す大きさのものも含まれている。各要素をほぐし，その上で相互関係を慎重に分析する必要があると考える。よって，本論文はその第一の試みとして評価書の質の事後的分析を行うものである。なお，評価の利用に関する理論的・実証的分析は別稿で行う予定である。日本の政策評価制度初期の段階での評価の利用については，益田（2004）が論じている。一方，評価書を作成する過程の質については，評価組織の内部を対象とした調査（インタビュー調査や内部資料等の調査）が必要になる。本論文が依拠する外部からの調査（評価書の調査）という方法では困難である。また，評価書を作成する過程の質に関する分析結果が今後出てきたとしても，それが持つ意味を解釈する段階において，本論文の総括的アプローチによる結果が必要になると考える。

性確保評価[15]とは，「政策を所掌する各府省とは異なる立場から，各府省では行うことのできない又は十分に達成できない評価として，複数府省にまたがる政策について，政府全体の統一性または総合性を確保するための評価」（総務省行政評価局『政策評価Q&A』）である。本論文では，2014年2月現在までに公表された全23本の評価書を対象とする。

同評価を分析対象とする理由は，第一に，政策評価制度上，同局は評価専担組織であり，他の各府省が評価活動以外を主たる業務としているのと異なり，評価活動そのものが中心的業務となっている点にある。この点では，先ほど最高会計検査機関に言及したが（脚注4），同局も類似の性質を持つと考えることができる。よって，政策評価制度の中で作成される評価書としては，比較的質への配慮が高く，質が均一であることが期待できるためである。言い換えれば，統一性・総合性確保評価を最も質が高く均一性が保たれている評価と仮定して分析対象とすることにより，行政機関全体の水準を予測しようとするものである。第二に，同局は，客観性担保評価活動とよばれる，各府省が作成した評価書の点検を行っている点にある。その目的は，評価の質の向上とそれを通じた政策の見直し・改善にあるとしている。よって，質への高い関心と質に関する知見の蓄積があると考えることができる。第三に，統一性・総合性確保評価はあまり調査対象になってこなかった点にある。制度上，各府省の評価書の質は同局が点検することになっているが，統一性・総合性確保評価に対しては同様の仕組みはない。また，研究上も同評価はほとんど対象となってこなかった。数少ない研究として，南島（2008）は，統一性・総合性確保評価のうち1つの評価書を検証している。また，拙著（2006）は，評価設計から予測される評価情報の性質と，実際に産出された評価情報の性質を比較分析している。しかし，包括的に同評価の質を分析しようとした研究はまだない。

なお，シュワルツとメインによれば，評価書の事後的な評価の主な機能は将来の評価活動への教訓を得ることにある，と指摘している（Schwartz and Mayne, 2005b, p.309）。本論文の目的も同様である。便宜上測定値を示すが，個々の評価書の良し悪しを示すことは目的ではない。評価書間の質の均一性又は不均一性を確かめ，その要因を探ることによって，評価の質の向上に向けた将来の取組に資することを目的としている。

分析対象となる統一性・総合性確保評価の名称は次の通りである。

(1)　地域輸入促進に関する政策評価書
(2)　容器包装リサイクルの促進に関する政策評価書
(3)　リゾート地域の開発・整備に関する政策評価書
(4)　障害者の就業等に関する政策評価書
(5)　政府金融機関等による公的資金の供給に関する政策評価書
(6)　特別会計制度の活用状況に関する政策評価書－歳入歳出決算における表示内容を中心として－
(7)　経済協力（政府開発援助）に関する政策評価書
(8)　検査検定制度に関する政策評価書
(9)　少子化対策に関する政策評価書－新エンゼルプランを対象として－
(10)　湖沼の水環境の保全に関する政策評価書
(11)　留学生の受け入れ推進施策に関する政策評価書
(12)　大都市地域における大気環境の保全に関する政策評価書

[15] 「行政機関が行う政策の評価に関する法律」（2001年6月制定）の第12条に基づく。

（13）　少年の非行対策に関する政策評価書
（14）　リサイクル対策に関する政策評価
（15）　PFI 事業に関する政策評価
（16）　自然再生の推進に関する政策評価
（17）　外国人が快適に観光できる環境の整備に関する政策評価
（18）　配偶者からの暴力の防止等に関する政策評価
（19）　世界最先端の「低公害車」社会の構築に関する政策評価
（20）　バイオマスの利活用に関する政策評価
（21）　児童虐待の防止等に関する政策評価
（22）　法曹人口の拡大及び法曹養成制度の改革に関する政策評価
（23）　ワーク・ライフ・バランスの推進に関する政策評価

なお，（5）政府金融，（6）特別会計，（8）検査検定に関する評価書の 3 つは，統一性確保評価である。統一性確保評価とは，「各行政機関の政策それぞれに共通する側面について統一した観点により横断的に評価」（「政策評価に関する基本方針」。以下，「基本方針」）するものである。その他の評価書は総合性確保評価である。それは，「複数の行政機関の所掌に関係する政策について，その総合的な推進を図る見地から，全体として評価」（「基本方針」）するものである。

3. 2. 評定基準

分析で用いる基準は，「連邦政府の評価者たち」が公表している「プログラム評価の質と有用性を評価するためのチェックリスト」に基づき設定しているが，分析の意図と実効性を考慮し修正している。具体的な相違点の 1 つ目は，「サンプリングの過程と対象の明示と適切性，サンプリング過程の政策決定者による一般化の可能性，分析計画の明示と適切性，データ収集者の選定と訓練の適切性，データ収集者間の信頼性確保の手続きの確保，データ収集過程における不足や課題の存否，統計手続きの明示と適切性，評価結果の解釈とは異なる解釈の可能性」を基準として採用していない点にある。つまり，サンプリングの過程，分析計画，データ収集過程に関する事項といった，評価書を作成する過程を分析対象としていない。完成した評価書の事後的な評価からは，これらの過程に関する評定に必要な情報を得ることができないためである。同様の理由で，データ収集者の能力も扱わない。また，統計手続きの明示と適切性，評価結果の解釈とは異なる解釈の可能性については，評価書が扱う個別政策領域の専門家と共同で判断しなければ，評定は困難である。他の手続きや解釈の可能性について幅広く検討しながら，当該評価書が適切な対応をしているのかを判断する必要があるためである。しかし，本論文では，評価書を事後的に評価する上で，評価書の構成要素として最低限の事項については押さえていると考える。つまり，「評価目的の明確さ，評価目的の適切さ，評価設計の明確さ，評価設計の適切さ，評価設計の実行実績，測定変数の適切さ，結論の確かさ，評価の限界の特定」について評定を行う。

また，相違点の 2 つ目は，上記の採用した基準における修正内容であり，詳細は表 1 の（*2）～（*6）に示したのでそちらに委ねるが，大きく分類すると，上記チェックリストに評定可能な程度の詳細が明示されていないため，評価研究の代表的教科書からそれを補足したもの（*2），次に，統一性・総合性確保評価の評価書の特徴を考慮し，より具体的な判断基準を置いたもの（*3）～（*5），そして同様の理由から内容の一部を除いたもの（*6）となる。

評定は次の 5 段階で行う。各項目の特徴に合わせた用語を使用している。

A1：不明確，A2：やや不明確，A3：どちらでもない，A4：やや明確，A5：明確

B1：不適切，B2：やや不適切，B3：どちらでもない，B4：やや適切，B5：適切

C1：不実行，C2：あまり実行されず，C3：どちらでもない，C4：やや実行，C5：完全に実行

表 1：「評価書の質」の評定基準

番号	項目名	内容	記号 (*1)
1	評価目的の明確さ	評価目的は明確か。	A
2	評価目的の適切さ	○ 評価目的は，プログラム[16]の発達段階の観点から考えて適切か。 ○ ロッシ 他（2005，39 頁）の“EXHIBIT 2-E”（Pancer & Westhues（1989）の引用）を基準にしている。(*2) ○ 「政策/施策の実施開始から評価開始までの期間」，「採用している評価の観点」の 2 つを主な判断基準としている。(*3) ○ なお，判定については本来評価目的（評価書では「評価の観点（又は視点）」の項目）からのみ判断することが求められている項目であるが，統一性・総合性確保評価の評価目的が，ほとんどの評価書において同様の記述となっており，また，その記述も「一括して」「全体として」「総体として」「総合的観点から」というように漠然とした表現を使用しているため，そのほとんどが B3 の評定がされかねない。ここではこうした実情を踏まえて，評価目的の項目のみならず評価書の内容も考慮することにより，評価書間の特徴を捉えようとした。このことから，全体に評定結果が目的の記述のみで判断した時よりも高めになっている可能性がある。(*4)	B
3	評価設計の明確さ	○ 評価設計は明確か。 ○ 評価の問い又は観点，手法又は測定指標のそれぞれの選択理由，及び関係性を明確に説明しているかを，主な判断基準としている。(*5)	A
4	評価設計の適切さ	評価設計は，調査目的を考慮に入れると，適切か。	B
5	評価設計の実行実績	提示された評価設計は実行されたか。	C
6	測定変数の適切さ	測定変数は，評価目的と関連および適切に翻訳しているか，また，評価目的に適切であるか。(*6)	B
7	結論の確かさ	結論は，データと分析によって支えられているか。	C
8	評価の限界の特定	評価の限界は特定されているか。	C

(*1) 「記号」は，5 段階評定の A，B，C のどれを各項目で採用するかを示している。

(*6) 参考にしている「プログラム評価の質と有用性を評価するためのチェックリスト」では，本項目の質問に「クライアントの質問への回答として適切か」という内容も含まれているが，統一性・総合性確保評価のクライアントは特定されていないため，本調査の項目からは除いた。

3. 3. 手順

- まず初めに，各評価書を評価書の質の評定基準に従って項目ごとに評定する。
- 評定結果に基づき標準偏差及び変動係数を算出することにより，質の不均一性（ばらつき）が生じているかを確認する。その結果，質の不均一性（ばらつき）が明らかであることが分かった場合には，どのようなパターンがあるのかをクラスター分析により検証する。
- 以上の分析結果を踏まえ，評価書の質を高めるために本分析から考えうる方策を探る。

[16] プログラムとは，目的，又は目標の組合せを持つ，あらゆる活動・プロジェクト・機能・政策を指す（GAO, 2011, *Performance Measurement and Evaluation: Definitions and Relationships*）。本論文では，政策又は施策を指すとしている。

4. 結果と考察

4. 1. 評定結果

23 の統一性・総合性確保評価を，評価書の質の評定基準に従って評定した結果は，表 2 の通りである。評定の際に，評価書が持つ特徴を踏まえ判断した事項を，＜留意事項＞に示している。

4. 2. 標準偏差と変動係数

質の不均一性（ばらつき）が生じているかを確認するために，評定結果に基づき標準偏差及び変動係数を算出した結果は，表 3 の通りである。標準偏差がゼロ（0）であれば，各項目において評価書間に質のばらつきがなく，均一であることになる。しかし，分析結果を見てみると，評定結果の合計値の標準偏差は 4.23 であり全体にばらつきが見られる。ここでは評定段階 1～5 を計算に使っており，標準偏差の数は評定段階の数字と同じ意味を持つことになる。よって，評価の限界の特定（1.56），評価設計の適切さ（1.06）の 2 つの項目は，標準偏差が「＞1」となっており，意味のない小さなばらつきとは言い難い。また，どの項目が全体のばらつきに寄与しているのかを探ると，各項目間の相対的なばらつきの統計量である変動係数は，評価の限界の特定（61.73），評価設計の適切さ（30.41）となっており，その他の項目と比して大きい。そのため，この 2 つの項目について詳しく見ていく。

この 2 つの項目のばらつきが大きい理由は何であろうか。「評価の限界の特定」については，評定 5 が 3 評価書，評定 4 が 5 評価書，評定 3 が 4 評価書，評定 2 がなし，評定 1 が 11 評価書となっている。本項目は，例えば，評価書で採用するデータや手法の妥当性や信頼性などに係る制約を明らかにし，それが当該評価書の分析に与える影響，又はそれに対する対応策の適切さなどを記入しているか否かを確認する項目である。言い換えれば，一般に評価に制約はつきものだが，各評価書が評価の問いに答えるために採用した方法の妥当性や信頼性の範囲について明示する項目である。評定 5 の評価書では，指標又はデータに係る制約，採用した手法に係る限界の記載がある。一方，既存データが不在であるため採用された代替的な手法については制約の記述がない評価書があり，それらの評定結果はそうではない場合よりも低くなっている。

次に，「評価設計の適切さ」については，評定 5 が 5 評価書，評定 4 が 6 評価書，評定 3 が 7 評価書，評定 2 が 5 評価書，評定 1 がなし，となっている。本項目は，調査目的を考慮してその設計の適切さを判断するものである。評定 5 の評価書では，評価目的と評価設計に一貫性があることを確認できる。一方，①評価目的と実際の分析での対象の範囲が顕著に異なるもの，②政策/施策の内容とそれ以外の内容の両方を混在させたまま説明しようとするもの，③政策介入の時点ではなくそれ以前の時点を基準として変化を説明しようとするもの，又は政策介入がもたらす変化の測定に使用するデータの取得期間の理由が不明なもの，④調査対象数の設定理由が不明なもの，⑤政策/施策が着手されたばかり又は改定されたばかりであるにもかかわらずその効果を測ろうとするために，評価目的と評価設計の間の一貫性を維持できなくなっているもの，⑥政策/施策の体系や指標設定が不十分であることの影響を受けて，評価目的と評価設計の関係が不明確になっているもの，等はそうではない場合よりも評定結果は低くなっている。

表 2 :「評価書の質」の評定結果

番号	1	2	3	4	5	6	7	8
項目名 / 評価書名	評価目的の明確さ	評価目的の適切さ	評価設計の明確さ	評価設計の適切さ	評価設計の実行実績	測定変数の適切さ	結論の確かさ	評価の限界の特定
FAZ	3	5	4	3	3	3	3	1
容器包装	3	5	5	4	5	5	5	4
リゾート	3	5	5	4	5	4	5	1
障害者	4	4	5	2	5	4	5	1
政府金融	5	5	5	5	5	5	5	5
特別会計	5	3	5	4	5	3	5	1
経済協力	4	5	3	3	5	3	4	1
検査検定	5	4	5	5	5	5	4	5
少子化	3	4	5	2	4	3	3	1
湖沼	3	5	5	5	5	4	5	5
留学生	3	5	3	3	5	3	5	1
大気環境	3	5	5	5	5	5	5	1
少年非行	3	3	3	2	5	3	5	4
リサイクル	3	5	5	5	5	5	5	4
PFI 事業	3	5	3	3	5	3	5	4
自然再生	3	2	3	2	5	3	5	3
外国人観光	3	5	5	4	5	4	4	1
配偶者暴力	3	5	5	3	5	5	5	3
低公害車	5	5	5	4	5	4	4	1
バイオマス	3	5	5	4	5	5	5	3
児童虐待	3	5	3	2	3	3	5	3
法曹人口	3	5	3	3	5	3	5	1
ﾜｰｸ･ﾗｲﾌ･ﾊﾞﾗﾝｽ	3	3	5	3	5	5	5	4

＜留意事項＞

(1) 「1. 評価目的の明確さ」について：例えば評価書において「各種施策が総体としてどの程度効果を上げているかなどの総合的な観点から」という表現が繰り返し使われているが，「総合性確保評価である」という以上の情報を得られないので，この表現と類似した表現を使う評価書も含めて「A3」と評定した。目的が特定され明確である程，評定結果が高くなっている。

(2) 「2. 評価目的の適切さ」について：プログラムの発達段階（政策/施策の実施開始から評価開始までの期間）から考えて採用している評価の観点（又は方法）が，適しているかという点で判断している。例えば，政策/施策の開始後年数が経っているにもかかわらず（言い換えれば，プログラムの発達段階が進んでいるにもかかわらず）初期段階に適している評価方法（ニーズアセスメントなど形成的評価）を採用している評価書や，逆に，政策/施策開始後年数が経っていないにもかかわらず（言い換えれば，プログラムの発達が初期段階であるにもかかわらず），発達段階が進み運営や機能が安定していると考えられるプログラムに適した評価方法（プロセス評価，アウトカム評価，費用便益分析・費用効果分析）を採用している評価書は，そうではない場合よりも評定結果が低くなっている。また，政策/施策の発達段階から考えて適していると考えられる観点や方法が理由もなく採用されていない場合も，そうではない場合よりも評定結果が低くなっている。

(3) 「5. 評価設計の実行実績」について：「3. 評価設計の明確さ」の評定結果が低いにもかかわらず，本項目の評定結果が高い評価書がある。論理的には，明確さの程度が低ければ，実行実績も判別しにくく評定結果が低くなるはずであるが，本調査では，評価設計の大枠しか説明されていない場合でも，その大枠が実行されていれば，「評価設計が実行された」と判断しており，評定結果が高めに出ている。

(4) 「6. 測定変数の適切さ」に関する「経済協力」の判定を「B3」としているが，手法の性質ゆえ点数化の為に便宜上置いた要素が強い。

(5) 「7. 結論の確かさ」について：「把握の結果」（個々の分析内容）と「評価の結果」（分析内容を踏まえた評価結果）がほぼ同一内容（例えば，「評価の結果」が「把握の結果」の要約）となっている場合にも，「結論がデータと分析によって支えられている」と判断されてしまう。結論が収斂するのではなく，さまざまな比較的小さな事項の集合体のようになっている場合にも同様の判断となる可能性がある。よって，判断が「5」となっていても，上に該当する場合が考えられるため，必ずしも高い評定結果に見合う評価書とはいいきれない可能性が残る。

(6) 5 段階評定について：上記の通り評価書の持つ特徴や評価項目の特徴への配慮を行うと同時に，各段階の判断の一貫性への配慮も行っている。例えば，「評価設計の適切さ」において，判断が難しいと考えられる「やや適切」，「どちらでもない」，「やや不適切」を採り上げる。基本的には，「やや適切」は，「適切」よりも 1 つ程度不足事項を確認した場合とし，次の「どちらでもない」は，2 つ程度不足事項を確認した場合とし，そして「やや不適切」は，3 つ程度以上，又は 2 つ程度でもその内容が評価設計の中心部分と関係している場合としている。なお，不足事項の数え方やその重要性の程度の判断は，評定者によって異なる可能性があるため，評定者が複数になる場合にも各段階の判断の一貫性に配慮をして評定を行うことが必要になる。

表３：標準偏差と変動係数

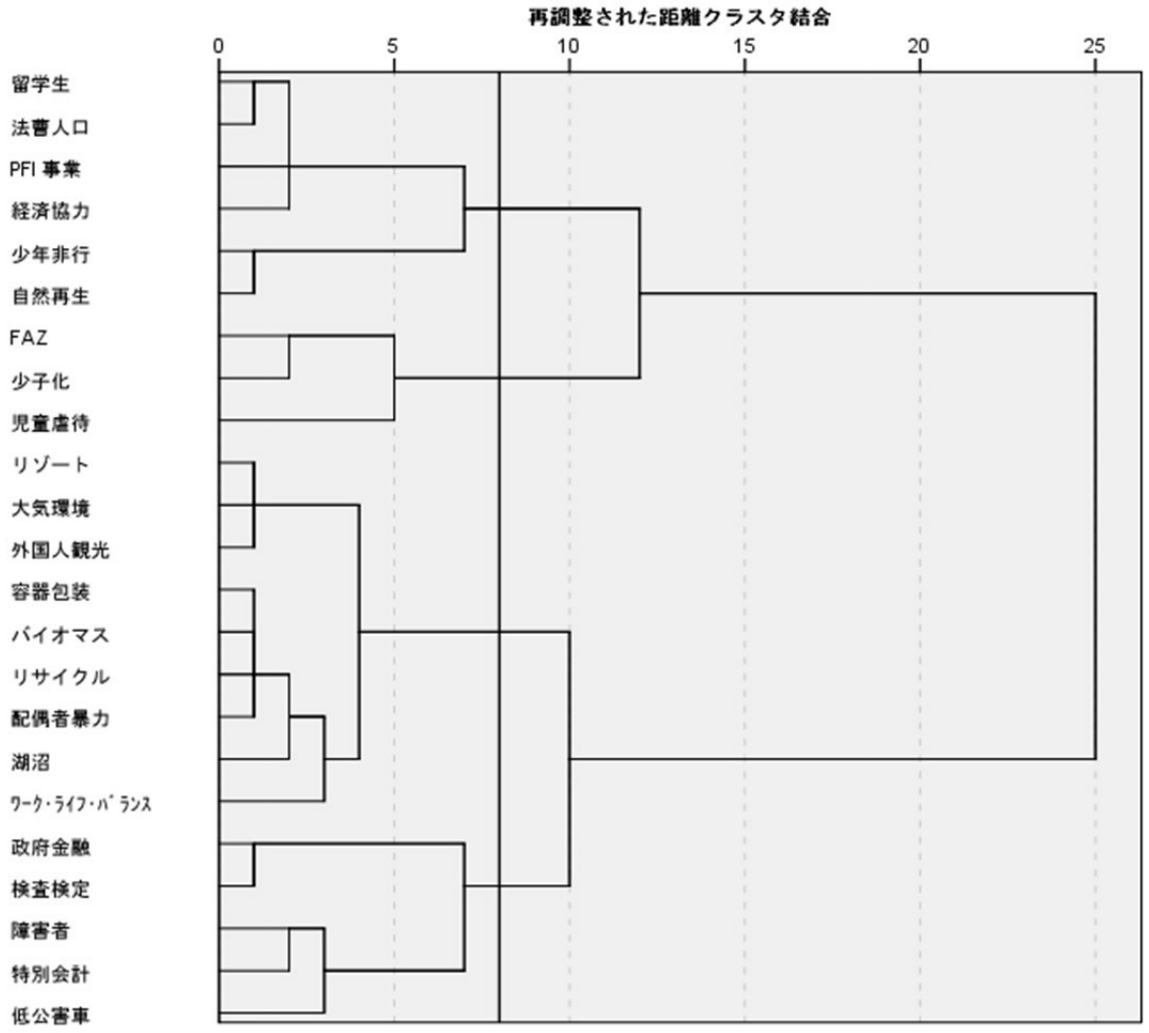
	1	2	3	4	5	6	7	8	
	評価目的の明確さ	評価目的の適切さ	評価設計の明確さ	評価設計の適切さ	評価設計の実行実績	測定変数の適切さ	結論の確かさ	評価の限界の特定	合計(*)
標準偏差	0.77	0.88	0.91	1.06	0.59	0.88	0.63	1.56	4.23
変動係数（%）	22.43	19.61	21.02	30.41	12.26	22.50	13.61	61.73	13.38

(*) 合計とは，すべての評価書の評定結果の合計値を使って算出した数値。表２においては，本論文が個々の評価書の良し悪しを示すことを目的としていないため，各評価書の合計値を示していない。

4．3．クラスター分析

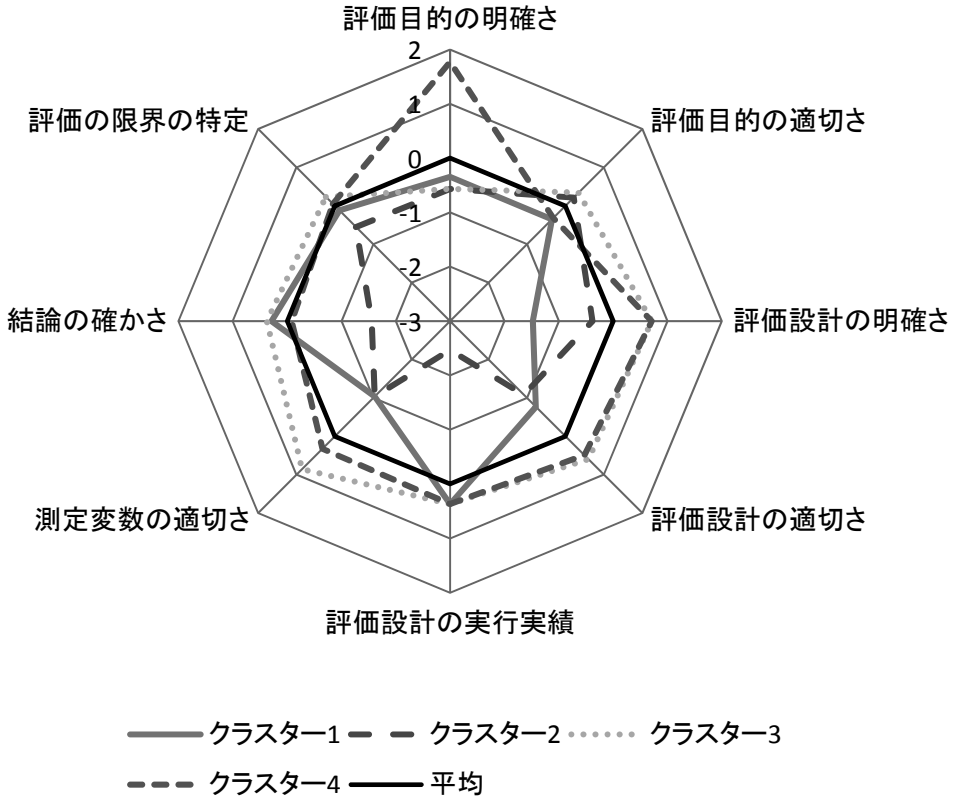
評価書の質のばらつきにどのようなパターンがあるのかを，評定結果を用い，階層的クラスター分析（Ward 法）を行った。各評定結果の標準化を行い，個体間の距離は平方ユークリッド距離で測定した。分析の結果は，図１のデンドログラムの通りである。評価書の質は４つのクラスターに区分された。クラスター数の決定に当たっては，図中のクラスター間の距離と説明の容易さの点から総合的に判断をした。

図１　評価書の質に関する評定結果のクラスター分析の結果



各評定項目の平均値[17]をクラスター毎にレーダーチャートで示すことにより，各クラスターの特徴を示したものが，図2である。全23本の評価書の平均（0）と比べながら相対的な高低に言及をして特徴を説明すると，次の通りとなる。クラスター1 は，設計関連項目（評価目的の明確さ・適切さ，評価設計の明確さ・適切さ，測定変数の適切さ，評価の限界の特定）の評定は相対的に低いが，実効性（評価設計の実行実績）や「結論の確かさ」は相対的に高い傾向を示している。クラスター2 は全体に相対的に低い傾向を示している。クラスター3 と 4 は共に全体に相対的に高い傾向を示しているが，両者を明らかに分けるのは，「評価目的の明確さ」の評定結果であり，クラスター4 の方が相対的に高い傾向を示している。

図2　各クラスターのプロファイル（各評定項目のクラスター平均値）



次に，担当班名，公表年，調査時期，頁数の4点を採り上げ説明を試みる。その理由は，前述の「2.4.」において評価の質に及ぼす影響として考えられる事項として紹介をした，評価者の能力，評価活動の背景（評価政策や政治的背景等），評価者が専門家集団に参加をすることを通じて得られる支援との関連にある。各事項について直接説明できる情報を統一性・総合性確保評価の評価書から得ることはできな

[17] クラスター分析に当たっては，各評価書の評定結果を標準化（平均0，標準偏差1）をしているので，図中のスコアも同様に標準化した数値となっている。

い。例えば，評価者の能力とは，クックシーの説明[18]によれば多岐にわたるが，その多くが社会科学的手法の理解と運用の能力に関連している。評価者個人が持つこうした能力に関する情報を評価書からは得られない。そこで，評価書に記載されている「担当班名」を使って間接的な説明を試みる。「担当班名」でクラスター毎に何らかの特徴が現れた場合には，評価組織全体で共有する，評価書の質に関する指針等が確認できない現状においては，各担当班の過去の評価経験の蓄積が評価書の質に影響を与えている可能性を考えることができる。また同様に，評価活動の背景のうち評価政策や政治的背景に関する情報を評価書からは得られない。例えば，評価政策とは，クックシーの説明[19]によれば，評価と関係するあらゆる決定を含むものである。よって，評価政策は明示的・非明示的な内容を含む概念である。どのような評価政策が評価書の質に直接影響しているのかは評価書からは判然としない。そこで，評価書に記載されている「調査・公表時期」と「頁数」を使って間接的な説明を試みる。「調査・公表時期」によってクラスター毎に何らかの特徴が現れた場合には，評価と関係する何らかの決定が評価書の質に影響を与えている可能性を考えることができる。また，「頁数」は，各評価書が対象とする政策を構成する施策数及び事業数，そしてそれらの対象となる組織・集団・個人の数といった評価対象の詳細を正確に数えることができない場合において，評価範囲のおおよその大きさを示すと考える。この「頁数」によってクラスター毎に何らかの特徴が現れた場合には，評価範囲に関わる評価政策からの影響の可能性を考えることができる。他方，評価者が専門家集団に参加をすることを通じて得られる支援の具体的内容については，評価書からは直接的にも間接的にも説明を試みることが困難である。以上の考えにより，「担当班名」，「調査・公表時期」，「頁数」[20]の情報をクラスター毎に示した結果が表４である。

[18] Cooksy and Mark, 2012, pp. 80-81
[19] Cooksy and Mark, 2012, p. 81
[20] なお，「頁数」（評価の結果及び意見，又は勧告までのもの。関係資料，参考資料等は数に含まれない）を公表年の順に並べると次の表の通りとなる。2007 年以降，150 頁を超える評価書が現れている。それ以前と比べて評価書が大部となる傾向についての明確な要因は，本論文の範囲からは説明できない。しかし，評価書が評価結果の体系性・総合性・緻密性を以前に増して重視した結果とも考えられる一方で，利用者の視点から考えれば，利用のしやすさから頁数の増加を問題視し指摘する程の利用者がなかったと考えることもできる。

評価書名	公表年	頁数
FAZ	2003年	82
容器包装	2003年	79
リゾート	2003年	92
障害者	2003年	34
政府金融	2003年	123
特別会計	2003年	112
経済協力	2004年	31
検査検定	2004年	65
少子化	2004年	37
湖沼	2004年	60
留学生	2005年	63
大気環境	2006年	96

評価書名	公表年	頁数
少年非行	2007年	88
リサイクル	2007年	222
PFI 事業	2008年	105
自然再生	2008年	141
外国人観光	2009年	125
配偶者暴力	2009年	151
低公害車	2009年	43
バイオマス	2011年	285
児童虐待	2012年	166
法曹人口	2012年	359
ﾜｰｸ･ﾗｲﾌ･ﾊﾞﾗﾝｽ	2013年	161

表 4：「担当班名」・「調査・公表時期」・「頁数」とクラスター

クラスター	評価書名	担当班名	公表年	調査開始	調査終了	頁数
1	経済協力	外務・文部科学担当	2004 年	2002 年 5 月	2004 年 4 月	31
1	留学生	国土交通担当	2005 年	2003 年 8 月	2005 年 1 月	63
1	PFI 事業	国土交通担当	2008 年	2005 年 12 月	2008 年 1 月	105
1	自然再生	国土交通担当	2008 年	2006 年 12 月	2007 年 3 月	141
1	少年非行	法務・外務・文部科学担当	2007 年	2005 年 4 月	2007 年 1 月	88
1	法曹人口	法務・外務・文部科学等担当	2012 年	2011 年 1 月	2012 年 4 月	359
2	少子化	厚生労働担当	2004 年	2003 年 8 月	2004 年 7 月	37
2	児童虐待	内閣・総務・厚生労働・防衛担当	2012 年	2009 年 12 月	2012 年 1 月	166
2	FAZ	内閣・総務・法務担当	2003 年	2001 年 1 月	2002 年 12 月	82
3	リゾート	規制改革等担当	2003 年	2001 年 1 月	2003 年 4 月	92
3	外国人観光	国土交通担当	2009 年	2007 年 8 月	2009 年 3 月	125
3	容器包装	農林水産・環境担当	2003 年	2001 年 1 月	2002 年 12 月	79
3	湖沼	農林水産・環境担当	2004 年	2002 年 12 月	2004 年 8 月	60
3	大気環境	農林水産・環境担当	2006 年	2004 年 12 月	2006 年 3 月	96
3	リサイクル	農林水産・環境担当	2007 年	2005 年 12 月	2007 年 8 月	222
3	バイオマス	農林水産・環境担当	2011 年	2008 年 12 月	2011 年 2 月	285
3	ﾜｰｸ･ﾗｲﾌ･ﾊﾞﾗﾝｽ	復興・総務・国土交通担当	2013 年	2011 年 12 月	2013 年 6 月	161
3	配偶者暴力	法務・外務・文部科学担当	2009 年	2007 年 3 月	2009 年 5 月	151
4	検査検定	規制改革等担当	2004 年	2002 年 8 月	2004 年 4 月	65
4	障害者	厚生労働担当	2003 年	2001 年 8 月	2003 年 4 月	34
4	特別会計	財務・経済産業等担当	2003 年	2002 年 12 月	2003 年 10 月	112
4	低公害車	財務・経済産業等担当	2009 年	2006 年 12 月	2009 年 6 月	43
4	政府金融	特殊法人等担当	2003 年	2002 年 1 月	2003 年 6 月	123

*頁数は，評価の結果及び意見，又は勧告までのもの。関係資料，参考資料等は数に含まれない。

この表から分かることは，「担当班名」が少なくとも 4 つのクラスターを分ける特徴とはまったく関係がないとは言い難い点である。言い換えれば，担当班内において過去の評価の経験をその後の評価活動においても継承し，それが各評価書の質を決めている可能性を考えることができる[21]。但し，この過去の評価の経験の具体的内容はさまざまな事項が考えられるが（例えば，評価対象となる政策の選定や評価手法の選定の考え方等），この表からは特定することはできない。他方，「調査・公表時期」と「頁数」は 4 つのクラスターを分ける特徴とは言うことができない。しかし，「調査・公表時期」は第 3 クラスターの，「頁数」は第 4 クラスターの個別の特徴の説明を補足している。詳細は次に説明をする。

クラスター別に見ていくと，第 1 クラスターは，国土交通担当と法務・外務・文部科学等の担当による評価書の分類となる。第 2 クラスターは，内閣・総務・厚生労働・法務・防衛等の担当による評価書の分類となる。第 3 クラスターは，農林水産・環境担当による評価書を中心とし，規制改革等担当（総合性確保評価（リゾート）の場合），復興・総務・国土交通担当の分類となる。規制改革等担当の評価書は，第 4 クラスターにもあるが，検査検定は統一性確保評価でありこちらの要素が効いていることが考えられる。なぜなら，第 4 クラスターにすべての統一性確保評価が集まっているためである。次に，復興・総務・国土交通担当の評価書（ワーク・ライフ・バランス）については，現段階では他にないため，今後同担当班の評価書が公表された際に改めて検証が必要である。この点についてはリゾートも同様である。

[21] 他にも，評価対象の政策を担当する省庁による評価活動への協力度合いの違いや，評価対象の政策の測定のしやすさといった性質の違い等，担当班が必ずしも予見できるとは限らない事項の影響を受けた結果である可能性も考えられる。

他方，外国人観光（国土交通担当）及び配偶者暴力（法務・外務・文部科学担当）については，第1クラスターではなく，第3クラスターに分類されている理由を説明することが難しい。但し，両方の評価書が調査時期を2007年から2009年までとし，公表が2009年であるという共通点はある。総務省年金記録確認中央第三者委員会の報告書（2011年6月）における事務局体制の変遷を見ると（p. 38），459人（2007年7月12日現在）→896人（2008年2月1日現在）→約2,000人（2008年7月時点）→約2,200人（2009年4月1日現在）→約1,900人（2011年4月1日現在）となっており，年金記録確認に要する活動量が2007年7月から2009年4月の間に急増していることが伺える。よって，年金記録確認に要する活動量の増加という決定が評価活動に関わる決定に変化を与え，担当班の評価活動に，そうした決定のない時期とは異なる特徴が現れた可能性はある。但し，低公害車も同時期を共有しているが第4クラスターに分類されている理由を，次に説明する。

第4クラスターは，評価の範囲の限定性（「頁数」）が分類に影響していると考える。また，「担当班名」による影響の可能性も残されている。統一性確保評価（検査検定，特別会計，政府金融）では，各ツールが実際にもたらした作用を統一的な方法で分析をするため，評価対象の種類も評価の方法も限定的である。他方，障害者と低公害車は総合性確保評価ではあるが，「頁数」を見ると，前者は34，後者は43となっており他の評価書に比して少ないことから，評価の範囲が限定的であることが伺える。それぞれ内容を確認すると，前者については，評価が対象とする政策は，「盲・聾・養護学校の高等部在学中から卒業後の職場への適応・定着に至る段階における障害者の就業等に関する政策」であるとしている。しかし，実際の評価は，「知的障害者を教育する養護学校の高等部」のみを「例にとる」（対象）としており（評価書，p.4&「表1」p.5），対象を限定的に捉えている。後者については，低公害車（CNG自動車/電気自動車/ハイブリッド自動車/メタノール自動車/低燃費かつ低排出ガス認定車）と燃料電池自動車という限定的な対象となっている。また，低公害車については，財務・経済産業等担当という担当班の要素も影響している可能性がある。

5. 制約

本論文には主に2つの制約がある。1つ目の制約は，評定者が1名である点である。可能であれば複数の評定者による評定が望ましい[22)]。この制約への対応として，本論文では，評定基準の内容と評定時の留意事項を明確かつ詳細に示した。

2つ目の制約は，各クラスターを特徴づける担当班の組合せ（特に第1クラスター）の理由を，評価書を事後的に分析する方法のみからでは説明できない点である。「行政評価局調査の流れ図」によれば，テーマが決定した後に実施段階に入り，「事前準備→実地調査の実施→調査結果のとりまとめ」を行う。その後，勧告等結果公表へと続く。また，各評価書によれば，政策評価計画と調査の状況（政策評価の方向性）の段階で政策評価分科会の審議に付されている。こうした評価活動の過程において質を保証する取組の存否や内容を確認することが必要になる。

[22)] 複数の評定者による評定が可能となった場合にも，次の点に留意が必要である。評価研究や社会科学的調査に関する知識を持つこと，及び評定が一貫性を保つためのルールの明確化と共有化を図ることである。本論文は試論であり，より多くの評価関係者等が評価の質の向上及び保証の為の具体的取組を始めるための，評価の質に関する現状分析の第一歩である。

6. おわりに

本論文では，評価書の質を評定基準に基づき事後的に測り，評価書間の質の不均一性を確認し，そのパターンを分析した。これらの結果に基づき，評価書の質を高めるための方策について 2 点を示し，本論文を締めくくりたい。

1 つ目は，政府内の評価専担組織である行政評価局内に，担当班単位ではなく全体で評価の質を保証する仕組みを機能させ，その取組を広く示すことは，自己評価を基本とする各府省にとってもよい参考となり，政府全体の評価の質が向上することに繋がると考えることである[23)]。まずは局内で評価の質とは何かについて確認及び共有するところから始め，徐々に仕組みをつくっていくことが，逆機能を引き起こさないためにも重要であると考える[24)]。

2 つ目は，評価対象の設定（又は評価範囲の設定）に関する制度上の定めについてである。統一性・総合性確保評価については，「基本方針」（2005 年 12 月 16 日閣議決定）の「III - 2 - (3) - (ア) 統一性又は総合性を確保するための評価活動」に「関係施策が極めて多岐にわたっている政策については，評価の結果を政策に適切に反映するために合理的と認められる単位により評価するものとする」とある。しかし，実際の評価書を見ると，例えば直近のワーク・ライフ・バランスに係る評価書においては，評価の対象が，6 府省が担当する主な 11 政策となっている。複数の評価書の束で 1 つの評価書が作られていると言えるような大きさである。これは本評価書以外にも見られる傾向である。合理的と認められる単位についての判断がどのようなものであったのかについては，評価書の事後的分析からは分からないが，評価の質と関わる内容であり検討が必要だろう。

[23)] 評価の質を保証する仕組みを整える際には，それが，評価の質を向上させる結果をもたらす場合もあれば，有害な副作用をもたらす場合（逆機能）もあることに注意が必要である。シュワルツとメインは，逆機能として，資源の無駄遣い（wasted resources. 質確保活動に資源を費やしたにもかかわらず質の向上が顕在化しなかった等），威力が弱い（decoupling. 質確保活動が，基準が高すぎる又は主要業務として重要と思われないゆえに無視される，又は儀礼的に扱われる），支配（colonizing. 質確保に執着しすぎて，視野狭窄や誤説を招き，評価情報をゆがめうる）の 3 つを挙げている（Schwartz and Mayne, 2005a, p.10& Schwartz and Mayne, 2005b, pp.312-314）。

[24)] Schwartz and Mayne, 2005, p.321

参考文献

総務省（2001）「政策評価に関する基本方針」http://www.soumu.go.jp/main_sosiki/hyouka/hyoka_hosinhonbun.html（2014 年 2 月 28 日参照）。

総務省（2004）「政策評価分科会の当面の活動の重点～政策評価制度の発展に向けて～」http://www.soumu.go.jp/menu_news/s-news/daijinkanbou/040430_1.pdf（2014 年 2 月 28 日参照）。

総務省（2005）「政策評価の実施に関するガイドライン」http://www.soumu.go.jp/main_content/000152600.pdf（2014 年 2 月 28 日参照）。

総務省（2011）「年金記録確認第三者委員会報告書－信頼回復へ向けたこれまでの活動と今後の課題－」http://www.soumu.go.jp/main_content/000117966.pdf（2014 年 2 月 28 日参照）。

総務省（2012）「政策評価等の実施状況及びこれらの結果の政策への反映状況に関する報告」http://www.soumu.go.jp/menu_news/s-news/72647.html（2014 年 2 月 28 日参照）。

総務省行政評価局（2009）「諸外国における政策評価のチェックシステムに関する研究」http://www.soumu.go.jp/main_sosiki/hyouka/seisaku_n/chousakenkyu/houkoku_2103.html（2014 年 2 月 28 日参照）。

総務省行政評価局（2011）「政策評価 Q&A（政策評価に関する問答集）」http://www.soumu.go.jp/main_sosiki/hyouka/seisaku_n/q_and_a.html（2014 年 2 月 28 日参照）。

南島和久（2008）「PFI 政策の評価をめぐる考察－総務省『PFI 事業に関する政策評価』（平成 20 年 1 月）の概要とそのレビュー」『評価クォータリー』第 5 号，2-12 頁。

益田直子（2004）「政策評価制度は何を解決するのか－制度運用の実態と理論の考察を通じて－」『本郷法政紀要』第 13 号，東京大学大学院法学政治学研究科，251-294 頁。

益田直子（2006）「評価情報の性質－総合評価を事例に－」，大山達雄（編）『公共政策評価の理論と実際』現代図書，28-37 頁。

山本清（2004）「政策評価の質的向上に向けて－課題の克服と対策（1）～（5）」『会計と監査』55（7）～（11）。

山谷清志（2007）「政策評価の質とその改善－実践と研究の交錯」『評価クォータリー』創刊号，2-13 頁。

ロッシ，ピーター・H，マーク・W・リプセイ，ハワード・E・フリーマン（監訳：大島巌，平岡公一，森俊夫，元永拓郎）（2005）『プログラム評価の理論と方法 - システマティックな対人サービス・政策評価の実践ガイド』日本評論社。（Rossi, Peter H, Mark W. Lipsey and Howard E. Freeman (2004) *Evaluation: A Systematic Approach Seventh Edition,* SAGE Publications.）

Cooksy, Leslie J. and Melvin M. Mark (2012) “Influences on Evaluation Quality,” *American Journal of Evlauation*, Vol.33(1), pp.79-87.

Cooksy, Leslie J. and Valerie J. Caracelli (2005) “Quality, Context, and Use: Issues in Achieving the Goals of Metaevaluation,” *American Journal of Evaluation*, Vol.26, pp.31-42.

Cousins, J. Bradley and Kenneth A. Leithwood (1986) “Current Empirical Research on Evaluation Utilization,” *Review of Educational Research*, Vol.56, pp.331-364.

Federal Evaluators (2006) *Evaluation Dialogue between OMB Staff and Federal Evaluators: Digging a Bit Deeper into Evaluation Science*.

Mayne, John and Robert Schwartz (2005) “Assuring the Quality of Evaluative Information,” Schwartz, Robert and John Mayne (Eds.), *Quality Matters: Seeking Confidence in Evaluating, Auditing, and Performance Reporting*, New Brunswick and London, Transaction Publishers, pp.1-17.

Pancer, S. Mark and Anne Westhues (1989) “A Developmental Stage Approach to Program Planning and Evaluation,” *Evaluation Review*, Vol.13 (1), pp.56-77.

Schwartz, Robert and John Mayne (2005a) “Assuring the Quality of Evaluative Information: Theory and Practice,” *Evaluation and Program Planning*, Vol.28(1), pp. 1-14.

Schwartz, Robert and John Mayne (2005b) “Does Quality Matter? Who Cares about the Quality of Evaluation Information?,” Schwartz, Robert and John Mayne (Eds.), *Quality Matters: Seeking Confidence in Evaluating, Auditing, and Performance Reporting*, New Brunswick and London, Transaction Publishers, pp.301-322.

Trochim, William M. K. and Ronald J. Visco (1985) “Quality Control in Evaluation,” *New Directions for Program Evaluation*, No. 27, pp.93-106.

U.S. Government Accountability Office (2013) *Program Evaluation: Strategies to Facilitate Agencies' Use of Evaluation in Program Management and Policy Making*, GAO-13-570.

U.S. Government Accountability Office (2011) Performance Measurement and Evaluation: Definitions and Relationships, GAO-11-646SP.

Wholey, Joseph S., Harry P. Hatry, and Kathryn E. Newcomer (Eds.) (2010) *Handbook of Practical Program Evaluation (3rd ed.)*, San Francisco, Jossey-Bass.

Yarbrough, Donald B. , Lyn M. Shulha, Rodney K. Hopson and Flora A. Caruthers (2010) *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users (3rd ed.)*, Thousand Oaks, CA, SAGE Publications.