広島大学大学院教育学研究科紀要　第三部　第57号　2008　99－104

教育評価における妥当性・信頼性に関する一考察

北　川　剛　司

（2008年10月２日受理）

A Consideration on the Validity and Reliability of Educational Evaluation

Takeshi Kitagawa

Abstract: In the research of the educational evaluation, we have been discussing not only the evaluation method but also the criterion of the evaluation to judge the adequacy of one. In the research of the measurement, we have developed some criteria, “validity” and “reliability”. Validity is defined as the extent to which the information tests supplies suits the purpose of the measurement. Reliability is defined as the consistency of the measurement. Though these criteria are terms of the “measurement”, we have applied them to the argument of the “evaluation”. And, doing so limits the development of new methods of the evaluation. The purpose of this paper is to consider the possibility of new methods and new criteria of the evaluation.

Key words: evaluation, validity, reliability, “Forth Generation Evaluation”
キーワード：評価，妥当性，信頼性，「第四世代評価」

Ⅰ．問題の所在

教育評価研究においては，評価方法の開発だけでなく，評価方法が適切かどうかということが議論されてきた。そして，評価の適切性（adequacy）を問う（判断する）ために，伝統的に妥当性・信頼性が規準とされてきた[1)]。評価の適切性を判断する規準というと，新たな評価方法の開発に従属して議論されるものというイメージを持つが，歴史においては，その逆に作用してきた。すなわち，評価の適切性を判断する規準としての妥当性・信頼性ありきで，その枠組みのなかで評価方法の開発が行われてきたのである。このような中にあっては，従来の妥当性・信頼性の枠組みにとらわれない全く新しい評価方法が議論されにくい。

妥当性・信頼性の枠組みにとらわれることなく評価方法が議論されることで，教育評価の方法は豊かになり，そのことが子どもをよりよく評価すること，ひいては教育評価の主目的である「指導と学習を改善すること」へとつながっていくと考える。

本研究の目的は，妥当性・信頼性にとらわれない，新たな評価方法およびその適切性を判断する新たな規準の可能性について考察することである。

そのために，まず，これまでの教育評価における方法の展開とともに妥当性・信頼性の展開を整理する。その上で，従来の評価論に対して，近年まき起こされた，構成主義的な評価を主張する立場からの批判を概観する。そして最後に，構成主義的な評価の方法とその適切性を判断するための規準について検討する。

Ⅱ．教育評価の方法および妥当性・信頼性の展開

教育評価は，時代背景や社会の影響によって，評価の方法を変化させてきた。この点に注目して，本稿では，方法の変化にしたがって，三つの時代区分で，教育評価の展開を整理することとする。

１．測定の時代

近代の評価は，測定（measurement）により幕を開けたとされる[2)]。アメリカにおいて測定論が勃興し

たのは，20世紀を前後する頃であった。当時のアメリカは，「法人資本主義の進行と階級・階層の移動と分裂，東欧系を中心とする新移民の激増，学歴社会の展開とりわけ高校進学率の上昇のなかで，誰もが納得のいく人材配分の公開の装置を必要としていた」[3]。そこでは，血統，財産，門地，年功に左右されない，「誰がいつやっても同じ結果が出る」という測定法（テスト法）が期待されたのである[4]。

測定論は，自らの測定の適切性を判断するための規準として，妥当性や信頼性という概念を作り出した。南風原朝和によると，妥当性とは，「意図されている測定目的に対して，テスト得点の与える情報が適切で有用である程度」[5]，信頼性とは，「同一の集団に対して，同様な条件のもとでテスト実施を繰り返すとき，一貫したテスト得点が得られる程度」[6]と定義される。

妥当性・信頼性が高いほど，その測定は適切であったということを示すため，適切な測定を求めようとする限りにおいては，妥当性・信頼性は，本来，同時に追及されるべきものである。しかし，「誰がいつやっても同じ結果が出る」ということを測定で重視する必要があった社会においては，当時の教育測定家たちの関心を，テストの信頼性を高めることに集中させた。当時開発された標準テストや知能検査などの精神測定法は，明らかに，信頼性を重視して開発されたものであった。

このように，この時代の教育評価は，自らが作り出した妥当性・信頼性のうち，信頼性を軸として発展していった。

２．記述の時代

第一次大戦終了後まもなく，アメリカの高校に，小学校レベルの教育しか受けていない学生が大量に入学してきた[7]。すると，これらの学生たちは，高校側に教育内容の変革を訴えた[8]。というのも，当時の高校が実施していた，大学進学の準備のためのカリキュラムは，彼らの要求を適切に満たせなかったからである。彼らの要求とは，高校で親の社会的・経済的地位を越えるために必要な能力を身につけさせてほしいといった，実社会で有用なスキルを求めるものであった[9]。そこで，政府は高校の履修単位や条件を変えるなどして，学生の要求に対応しようとした。しかし，この変更には障害があった。それは，大学の標準的なカリキュラムについていけない高等学校卒業生を受け入れざるをえなくなることに対する大学側の懸念であった[10]。

このような大学側の懸念に対し，その懸念が妥当であるかどうかを調べるプロジェクト（８年研究）が1933年にスタートした。調査の目的は，正統でない（unorthodox）カリキュラムで学んだ高校生たちが大学に入学した場合，彼らが大学の授業についていけるかどうか，を調査するためであった[11]。

この調査の中心にいた人物がタイラーであった。タイラーが最初に直面した問題は，新カリキュラムが果たして意図したように機能しているかどうか，を確かめるための評価法を，まず考案することであった。評価法を考案するにあたって，タイラーは「当時の教育測定運動に生じていた，測定条件に統計学的手法を導入することによって測定行為の自己目的化をもたらす傾向を批判し」[12]た。そして，そうした批判を乗り越える新しい評価方法として，あらかじめ定められた「教育目標」を評価規準として，その目標に対する強みと弱みを記述するという方法を採用した。ここにおいて，記述志向の評価（evaluation）が登場した。

ここで注目すべきは，従来の測定志向の評価において議論されてきた妥当性・信頼性が，記述志向の評価においても議論されていることである[13]。すなわち，記述志向の評価においても，測定論の概念である妥当性・信頼性が，評価の適切性を判断する規準として採用されたのである。

３．判定の時代

1960年代後半になると，評価者の役割には目標に準拠して記述することだけでなく判定が含まれるべきだという主張が，ステイク（Stake, R. E.）によってなされた[14]。こうした主張がされはじめた当初は，「従来の評価のように，科学的で，一般に価値観から独立しているとみなされてきた業務に，価値観を伴う判定基準を導入することは，当時の多くの評価者にとってとても受け入れがたいこと」[15]と考えられていた。しかし，それにも関わらず，評価者が判定者としての役割を受け入れざるを得なかったのは，当時の社会的な要請[16]の影響が大きかったためである。このように，社会的な要請に後押しされるようにして，判定志向の評価は登場した。

判定志向の評価については，さまざまな論者によって，さまざまな判定モデルが提唱されたことが特徴的である。例えば，スクリヴァン（Scriven, M. S.）はゴール・フリーモデル（goal free model）を提唱し，評価者は目標を評価規準とはせず，目標から独立した独自のチェックリストを用いて価値の判定を行うべきだと主張した[17]。アイズナー（Eisner, E. W.）は鑑識眼（connoisseurship）にもとづく評価方法の理論を確立し，芸術批評で用いられるような主観的で，質的な方法で価値の判定を行うことを提唱した[18]。

このように多様な判定モデルが提唱されたが，そのなかでもグーバ（Guba, E. G.）とリンカーン（Lincoln,

Y. S.）の判定モデルは，単なる方法の議論にとどまらず，評価をパラダイムで整理するなど，マクロな視点で議論を展開している点が興味深い。グーバとリンカーンによると，先にとり挙げた，測定志向の評価，記述志向の評価，そしてスクリヴァンやアイズナーの判定モデルはいずれも科学的前提[19]にもとづいているとして批判される[20]。

そこで，本稿では，今後の教育評価の発展をめぐる重要な提言として，グーバとリンカーンの判定モデルに着目し，彼らが展開する批判と，その批判のうえに提唱される評価方法およびその適切性を判断する規準を以下で検討する。

Ⅲ．グーバとリンカーンによる科学的評価に対する批判

グーバとリンカーンは従来の科学的前提に立つ評価（科学的評価）に対して，社会科学的な知見と構成主義的な評価論の立場から以下のような批判を展開している。

第一の批判は，従来の評価が「管理主義（managerialism）へ傾倒」[21]しているということに対してなされる。一般に，評価者が評価を引き受け，指示を受けて，そのアウトプットである報告書を提出する先は管理者である。ここでいう管理者とはすなわち，評価レポートに基づいて実行する責任をもつエージェントのことである。教育評価における管理者とは，例えば，教育長や学校長などがそれにあたる。従来の評価のように，評価が管理主義へ傾倒していると，管理者と評価者との関係は，次のように，幾つかの望ましくない結果を生み出す。

望ましくない結果の一つ目は，管理者が無害なように保護されることである[22]。すなわち，管理者は評価結果に対して責任を負わなくてすむということである。管理者が評価の外にいる限り，管理の資質や技量は問われないし，評価の結果，何が生み出されようとも，管理者はその責任を問われない[23]。望ましくない結果の二つ目は，管理者と評価者の典型的な関係が，不公平な関係になってしまうことである[24]。すなわち，両者の関係が，権限を持つ者と権限を持たない者の関係となることである。このような関係においては，管理者と別の解釈をする可能性のあるステイクホルダー（評価者自身を含む）は，事実上弱い立場におかれることになる。つまり，管理者が評価に大きな権限を有している以上，評価に関心を持つステイクホルダーの意見を反映することは不可能ではないが，一般に難しいということである[25]。望ましくない結果の三つ目は，管理者のみが評価結果を公表する権限を持ってしまうことである[26]。こうなると，管理者にとって都合のいい評価結果のみが公表されることになり，結果として，ステイクホルダーは本当の評価結果にアクセスすることができなくなる。

第二の批判は，従来の評価が「多元的価値の問題に対応できていないこと」[27]に対してなされる。従来の評価においては，「客観的」とされる評価手法の開発にも価値判断が関わりあっていることや，明確な目標設定の際にも価値の合意を前提にしていること，などの事実が見過ごされてきた。このように，従来から，価値の問題が存在していたにも関わらず，その評価結果が信用されてきたのは，評価手法が科学的であれば，価値の問題に無関係である，という議論が行われてきたためである[28]。

第三の批判は，従来の評価が「科学的パラダイムに依存しすぎてきたこと」[29]に対してなされる。社会科学における調査に関わってきた人々は，これまで自然科学における手法を積極的に取り入れてきた。しかし，科学的手法への極端な依存は，いくつかの望ましくない結果をもたらした。第一に，文脈剥離の問題である[30]。文脈剥離とは，評価対象があたかも現実の文脈のなかではなく，人口的に注意深く統制された環境下だけに存在して，対象を評価することである。第二に，正確に定量化できるデータのみを収集したことである[31]。そして第三に，科学的手法に基づいた評価結果は，ある種の権威を持ったことである[32]。すなわち，評価結果が権威をもったがゆえに，被評価者は評価の結果について議論することや拒否することが難しくなった。第四に，科学的手法を使うことによって，その結果を「真実」として受け入れるように強制力が働いたことである[33]。このような強制力が働いたことで，評価対象を異なった視点から考察する機会は奪われてしまった。第五に，科学は価値観と無関係であるとされることにより，評価者は科学的パラダイムを信奉しさえすれば，評価活動に対する道義上の責任を軽減されるようになるかのような錯覚をもたらしたことである[34]。

以上のような，従来の科学的評価が持っていた問題点を克服するために代替案がもたらされた。その代替案の一つが，グーバとリンカーンの提唱する「応答的構成主義的評価（responsive constructivist evaluation）」であり，彼らはこの種の評価が次なる世代を形成するという意味で，これを「第四世代評価」と称した。第四世代評価は，応答的評価の焦点の当て方と構成主義的パラダイムにもとづく方法論を合わせ持った評価である[35]。すなわち，「ステイクホルダーが評価対象に

持っている評価の焦点，要求，争点は，相互作用の進行とともに変化していくので，事前に決めておくことはできない」[36]という反事前決定型（anti-preordinate）の焦点と，「真実は客観的に存在しているのではなく，人々の相互作用をとおして社会的に構成される」という前提のもとで分析・批判を繰り返して調査者と回答者との間から洗練された構成物を作り出すという方法論とを用いて行う評価が「第四世代評価」である。

以下では，「第四世代評価」の評価方法およびその適切性を判断する規準について述べる。

Ⅳ．「第四世代評価」の方法およびその適切性を判断するための規準

1．「第四世代評価」の方法

グーバとリンカーンによると，「第四世代評価」の基本的な方法は，次の12のステップから成るという[37]。

［1］クライアント／スポンサーとの契約に着手する。
［2］評価の実施にむけて手配する。
［3］ステイクホルダーを特定する。
［4］ステイクホルダー集団内の共同解釈を発展させる。
［5］ステイクホルダーの共同解釈をテストし拡大する。
［6］解決した要求（claims），懸念（concern），争点（issues）を分類する。
［7］未解決の項目に優先順位をつける。
［8］未解決の項目を解決するための情報を収集する。
［9］交渉のための議題を準備する。
［10］交渉を実行する。
［11］報告する。
［12］再循環させる。

ここに見出せるのは，ステイクホルダー同士の共同解釈の産出→合意にいたらなかった事項に関するステイクホルダー同士の交渉（negotiation）→合意としての共同解釈の産出，という「第四世代評価」の基本的な流れである。「第四世代評価」においては，評価方法の定式化に関して，これ以上のことは議論されない。

「第四世代評価」が提案される以前の評価においては，評価は評価者が行うということが前提とされてきた。しかし，「第四世代評価」においては，評価はステイクホルダーが行うということが前提となることが特徴的である。すなわち，評価者はコーディネーターのようにふるまい，ステイクホルダー同士の交渉を組織することを要求される。

2．「第四世代評価」の適切性を判断するための規準

評価とは，統制のとれた探究の一つの形式であると認めるならば，探究の質を判断するための規準が重要になる[38]。このことは，「第四世代評価」においても例外ではない。

グーバとリンカーンによると，「第四世代評価」の適切性を判断するには，妥当性・信頼性といった伝統的な規準はふさわしくないという[39]。その理由は，「『第四世代評価』は，構成主義的な評価であり，そのような評価の適切性を，妥当性・信頼性といった科学的で実証主義的な（positivistic）規準によって判断するのは適当でない」[40]と述べられる。

それでは，「第四世代評価」においては，どのような規準を用いるのが適当なのであろうか。グーバとリンカーンが「第四世代評価」にふさわしい規準として挙げるのが，真正性規準（authenticity criteria）といわれる諸規準で，この規準は「構成主義自体の基本的前提から直接的に導き出されるものである」[41]という。グーバとリンカーンが真正性規準として挙げるのは次の五つである。以下では，それぞれの規準がどのようなものであるかを述べる。

(1) 公平性（fairness）

公平性とは，「評価のプロセス内で，異なる構成物とそれら構成物の根底にある価値構造が，受け入れられ，尊敬され」[42]ているかをみる規準である。探究は，価値をおびていて，価値に状況付けられているものなので，探究によってもたらされた構成物もそれぞれが価値をもっている。このことが受け入れられていることが公平（fair）な状態である。

(2) 存在論的真正性（ontological authenticity）

存在論的真正性とは，「個々の応答者自身がつくった構成物が，改善され，熟成され，拡張され，精緻化され，その結果，より洗練された構成物にな」[43]っているかをみる規準である。個々の構成物は，完全ではなく，完全に向けて常に再構成されつづけなければならないという前提に立ち，より洗練された構成物となることを目指して再構成し続けることが，存在論的真正性を保証する。

(3) 教育的真正性（educative authenticity）

教育的真正性とは，「ステイクホールディング・グループの外側にいる人々の構成物に対して，個々の応答者がそれを理解し評価し」[44]ているかをみる規準である。「ステイクホルダーは少なくとも，自身とは全く異なった人々の構成物に向き合う機会を持つべき」[45]とされる。というのも，そうすることで，評価者をとりまく争点に対する全く異なった解決策を学びとることができるからである。すなわち，自身の構成物だけでなく，他の構成物に目をむけ，そこから学ぶことで教育的真正性は高められる。

(4) 触媒的真正性（catalytic authenticity）

触媒的真正性とは，「評価プロセスによって，人々の行動は刺激され，促進させられ」[46]たかをみる規準である。評価の目的は，評価結果にもとづき，人々が行動するようになったり，意思決定するようになることである。評価は，人々を次の行動に向かわせる実質的な機能をもつものでなければならない。

(5) 戦術的真正性（tactical authenticity）

戦術的真正性とは，「ステイクホルダーが行動する権限を付与されている」[47]かをみる規準である。評価は，触媒的真正性をもつだけでは不十分である。すなわち，評価は，人々を行動に駆り立てるだけでは不十分である。というのも，評価がたとえ，人々の行動を駆り立てるものとなっていたとしても，人々が行動する権限をもたなければ，その行動が実施されえないからである。

以上のような規準が，「第四世代評価」の適切性を判断するにふさわしい規準として提示されている。これらはいずれも，妥当性・信頼性がみようとしている評価の側面とは違う側面に焦点化している。すなわち，妥当性・信頼性が，評価の方法に焦点化するものであるのに対して，上記の真正性規準は，評価を行う者の意識や評価を行う者の権限などにも焦点化する。この意味で，これらの規準は，これまでにない，新たな規準の可能性を示唆しているといえる。

Ⅴ．おわりに

本研究をとおして，評価の妥当性・信頼性に関して，次の二つのことが明らかになった。

第一に，評価方法の適切性を判断する規準に関する議論は，測定志向の評価，記述志向の評価においては，妥当性・信頼性に関する議論に集中しており，その他の規準について論じたものは見られないが，判定志向の評価においては，その他の規準に関する論議が一部でなされるようになったということ。本稿でとりあげたグーバとリンカーンの議論は，妥当性・信頼性以外の規準に言及した数少ない議論である。

第二に，妥当性・信頼性は，評価の適切性を判断する規準として必ずしもふさわしいとはいえないということである。すなわち，妥当性・信頼性以外の規準が想定され，評価によっては妥当性・信頼性以外の規準が使用されるべきであるということである。すでに述べたが，妥当性・信頼性は実証主義的な規準であり，例えば「第四世代評価」のような構成主義的な評価のための規準としてはふさわしくない。新たな評価方法について議論する際には，妥当性・信頼性以外の規準の可能性も視野に入れておく必要がある。

本稿でとりあげた「第四世代評価」のような評価方法は，妥当性・信頼性の枠組みにとらわれた評価論においては，議論されえない。したがって，教育評価を今後さらに発展させるにあたっては，妥当性・信頼性のような規準を無条件に受け入れるのではなく，本稿でみたような規準の導入についても検討していかなくてはならないだろう。

【註】

1）Cf., Lincoln, Y. S. & Guba, E. G. (1986). “But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation”. In Williams, D. D. (Ed.), *Naturalistic Evaluation*. Jossey-Bass, p.74.

2）グーバとリンカーンは，測定を評価の第一世代と呼んでいる。(Cf., Guba, E. G. & Lincoln, Y. S. (1989). *Forth Generation Evaluation*, Sage Publications, pp.22-26.)

3）田中耕治（2008）『教育評価』岩波書店，16頁。

4）同上書，参照。

5）南風原朝和（1988）「妥当性」東洋ほか『現代教育評価事典』金子書房，402頁。

6）南風原朝和「信頼性」同上書，343-344頁。

7）Guba, E. G. & Lincoln, Y. S. (1989). op. cit., p.27.

8）Ibid.

9）Cf., ibid.

10）Ibid.

11）Ibid.

12）田中耕治（2008），前掲書，24頁。

13）例えば，Baron, D & Bernard, H. W. (1958). *Evaluation Techniques for Classroom Teachers*. McGRAW-HILL.

14）Cf., Stake, R. E. (1967). “The countenance of educational evaluation”. *Teachers College Record*, **68**, p.524.

15）Guba, E. G. & Lincoln, Y. S. (1989). op. cit., p.30.

16）スプートニク・ショック以来，アメリカでは科学教育の必要性が自覚され，科学教育に関わる複数のプログラムが実施された。それらのプログラム改善を一刻も早く行うために，即座に効果を判定することが求められた。

17）スクリヴァンのゴール・フリーモデルの方法については，次の文献に詳しい。根津朋実（2006）『カリキュラム評価の方法－ゴール・フリー評価論の応用－』多賀出版。

18）アイズナーの鑑識眼にもとづく評価方法について

は，次の文献に詳しい。Eisner, E. W. (2003). “Educational Connoisseurship and Educational Criticism: An Arts-Based Approach to Educational Evaluation”, In Kellaghan, T., Stufflebeam, D. L., & Wingate, L. A., *International Handbook of Educational Evaluation*. Part 1 Perspective, Kluwer Academic Publishers.／森谷宏幸「E. W. アイスナーの質的教育研究方法論の検討」『福岡教育大学紀要』第43号，第２分冊，1994年。

19）科学的前提とは，「真実は客観的に存在している」と考えることである。

20）Guba, E. G. & Lincoln, Y. S. (1985). *Naturalistic Inquiry*. Sage Publications. において，科学的パラダイムに対する批判が明確に述べられている。続く，Guba, E. G. & Lincoln, Y. S. (1989). *Forth Generation Evaluation*, Sage Publications. では，前著の批判にもとづいて，構成主義的パラダイムにもとづく評価（判定）モデルが提唱されている。

21）Guba, E. G. & Lincoln, Y. S. (1989). op. cit., p.32.

22）Cf., ibid., p.32.

23）Ibid.

24）Cf., ibid.

25）Cf., ibid., pp.32–33.

26）Cf., ibid., p.33.

27）Ibid., p.34.

28）Cf., ibid.

29）Ibid., p.35.

30）Cf., ibid., p.36.

31）Cf., ibid., p.37.

32）Cf., ibid.

33）Cf., ibid.

34）Cf., ibid., p.38.

35）Cf., ibid., p.11.

36）Ibid.

37）Cf., ibid., pp.188–226.

38）Cf., ibid., p.228.

39）Cf., ibid., p.245.

40）Ibid.

41）Ibid.

42）Ibid., pp.245–246.

43）Ibid., p.248.

44）Ibid.

45）Ibid.

46）Ibid., p.249.

47）Ibid., p.250.

（主任指導教員　深澤広明）