福岡教育大学紀要，第65号，第４分冊，1―13（2016）

テストをめぐる学際的検討
―教育方法学・教育心理学・教育社会学の対話

An interdisciplinary discussion on assessment

川口俊明　松尾剛　樋口裕介

Toshiaki KAWAGUCHI　Go MATSUO　Yusuke HIGUCHI

学校教育講座　教育心理学講座　学校教育講座

（平成27年９月30日受理）

1. はじめに

本稿の目的は，とくに小・中学校教育における学力テストに焦点をあて，どのように子どもたちの能力を評価・測定すればよいのかについて，教育方法学・教育心理学・教育社会学の知見を駆使しながら論じることにある。

近年，日本では，全国学力・学習状況調査を始めとし，さまざまな学力調査が実施されている。しかし一方で，実際に実施されている学力調査にはさまざまな問題点があり，教育の在り方を歪めているという批判も繰り返しなされている。それでは，どのような評価の在り方が適切なのか。教育評価に関わる領域が，あまりに多岐に渡るため，この議論は非常にむずかしい。本稿では，評価に関わる教育学諸領域のうち，教育方法学・教育心理学・教育社会学の知見をもとに，適切な教育評価の在り方を探ろうと思う。

後に２節で扱うように，教育評価には「メジャメント（measurement）からエバリュエーション（evaluation）へ」という流れが存在する。そこで本稿では，前半部分である２節・３節・４節において，おもにエバリュエーションに関わる内容を扱う。具体的には，２節で教育評価の考え方とその展開，３節で評価を行う際に重要な事項となる妥当性・信頼性という概念，４節で現在の教育評価の論点について整理する。その上で後半である５節・６節では，メジャメントに関わる近年の研究動向に焦点を当てる。５節で新しいテスト理論であるIRT（項目反応理論）を扱い，６節では「測定された学力」を利用した社会的不平等の解明に関する研究について述べる。最後に７節で，これまでの議論を踏まえ，教育評価の課題と今後の展望を述べることにしたい。

2. 教育評価の目的と立場

2.1. 教育評価の目的―「メジャメント（measurement）」から「エバリュエーション（evaluation）」へ―

戦後の教育評価研究の展開をまとめた田中耕治は，「「教育評価」という用語は，第二次世界大戦後にアメリカから移入された「エバリュエーション（evaluation）」を邦訳したものである。」と説明したうえで，教師の指導と子どもたちの学習活動の改善をめざす行為として教育評価が位置づけられたことを指摘している（田中 2014，p.182）。田中自身も「「エバリュエーション」としての教育評価の目的とは，子どもたちをネブミして，序列・選別することではなく，教育実践それ自体に反省を加えて，修正・改善することである」（田中，2008，p.83）と述べているように，教育評価は（教師あるいは子どもの）実践の反省と改善のためのものだという考えが定着していると言えよう。

このように教育評価の目的が規定される背景には，教育評価における「メジャメント（measurement）」から「エバリュエーション（evaluation）」へ，というパラダイム転換がある。「メジャメント」運動が栄え

た背景を田中耕治は以下のように説明している。

> この時期のアメリカは，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。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，「メジャメント」運動が提案した「テスト」法に期待が集まっていた（田中 2008，p.16）

このように，教育評価において，「メジャメント」は公平な人材配分のためのものであり，子どもを区分し，序列化することに主眼が置かれていると理解されてきた。それ故，「メジャメント」運動に浸透する社会ダーウィニズムや「テストや測定の宿命論的解釈（知能テストで測定されるのは教育課程の差であるのに生得的知能の差と考えること）」（田中 2008，p.22）に批判が向けられてきたのである。

そこにあるのは，「「個人差」は生まれながらの恒常的なものではなくて，それはとりわけ教育活動との関数関係として生じ，したがって可変的なものである」（田中 2008，p.23）という考えであり，だからこそ「教育活動との関数関係」を測定することが重点になる。こうした問題意識を背景として登場するのが，「エバリュエーション」概念である。「エバリュエーション」概念の特徴は以下のように整理される。

> ①評価の規準は，教育目標である，②教育目標は，高次の精神活動を含む重要な目標群を含むべきである，③教育目標は，生徒に期待される行動で記述すべきである，④目標実現の度合いを知るための多様な評価方法が工夫されるべきである，⑤もし，目標に未到達の子どもがいた場合には，治療的授業が実施されるべきである，⑥以上のことは，カリキュラムや授業実践の改善につながる，⑦以上のことは，実践家と研究者の協力によって行われることが望ましい（田中 2008，pp.26-27）

このように，「エバリュエーション」は，子どもたちの序列化を目的とするのではなく，カリキュラムや授業の目標が子どもに達成されたかどうかを明らかにすることを目的とするのである。このパラダイム転換は，「相対評価」から「目標に準拠した評価」へ，という教育評価の立場の転換につながる。

もう一つ，教育評価の基本的な考え方をおさえるうえで重要な概念として「指導と評価の一体化」がある。これには，大きく二つのとらえ方がある。「第一は，評価の結果を次の指導に生かすという意味での一体化であり，第二のそれは，評価すること自体を指導にしていくという意味での一体化である」（諸岡 1987，p.546）。前者は，教育活動の改善のための評価という意味である。後者では，評価すること自体が子どもにどういった作用をもたらすのかということを重視する。前者の場合，日本の教育課程行政を鑑みれば，文部科学省の「告示」である学習指導要領が教育実践に対して強い法的拘束力を持つ中で，「評価行為の結果に照らしてカリキュラムの改善を志向するという条件が弱まる」（田中 2008，p.86）。そうしたなかでは，国レベルや学校レベルのカリキュラムの改善というよりは，短期的な一つ一つの授業に目が向きがちである。また，後者の場合であればなおさら，子どもに対する即時のフィードバックが求められることになるだろう。

2.2. 教育評価の立場の展開―「相対評価」から「到達度評価」「目標に準拠した評価」へ―

それでは，戦後の教育評価の立場はどのように展開したであろうか。「戦後半世紀にわたる公的な教育評価の構造は，「相対評価」を基軸とする多層構造であった」と田中が述べるように，まず，戦前の「考査」の時代がもたらした教師の主観的・恣意的な絶対評価を克服するものとして，「客観性」[(1)] をもつとされる相対評価が展開する（田中 2014，p.183）。しかし，相対評価の非教育性が指摘されるに連れ，「到達度評価」「目標に準拠した評価」への要求が高まっていく。田中（2008）によれば，相対評価の問題点は次の 4 点にまとめられている。

> 第一の問題点は，必ずできない子がいるということを前提とする非教育的な評価論であること。・・・（中略）・・・第二の問題点は排他的な競争を常態化させて，「勉強とは勝ち負け」とする学習観を生み出すこと。・・・（中略）・・・そして，第三の問題点は学力の実態を映し出す評価ではないこと。たとえ「5」をもらったとしても，その意味するところは集団における相対的な位置が上位であるというこ

> とであって，そこで獲得した学力が教師のめざす教育目標に達していたかどうかを証明してはいないということである。・・・（中略）・・・第四の問題点として，したがって「相対評価」のもとで学業不振が起こったとして，その責任は子どもたちの努力不足，才能不足に帰せられてしまうことになる（田中2008，pp.47-48）

このように，教育評価論にとって相対評価は「メジャメント」を前提とする立場であり，子どもたちを序列化することに主眼があると理解されてきた。相対評価では，当該の集団内での序列は明らかになるものの，目標の達成度合いは明らかにできない。そこからは，カリキュラム・授業，あるいは子どもの学習への改善指針は導き出されない。

こうして，「到達度評価」「目標に準拠した評価」が求められることとなる。これらは，前述した「エバリュエーション」概念を前提としつつ，国家・学校による子どもの学習権保障・学力保障という意識にもとづくものである。

3. 教育評価における妥当性と信頼性

エバリュエーションとしての教育評価においても，測定の質を高めることで学力や態度などを適切に反映したデータを得ることは重要な営みである。なぜなら，教育評価が測定の対象とする学力や態度といったものは直接には観察不可能な構成概念であり，その測定方法で適切に対象を数値化できているという保証はないからである。「相対評価を行なうべきか，目標に準拠した評価を行なうべきか」といった点をどれだけ慎重に議論しても，そもそも議論の対象となっているデータが適切な測定方法によって得られたものでなければ，その評価によって教育を改善していくことは困難になってしまう。そこで，以下では教育評価における測定の質をいかに高めるかという点について「妥当性」と「信頼性」の観点から論じていく。

3.1. 測定の「妥当性」

測定の「妥当性」とは，その測定によって目的となる対象が適切に測定できているか，という概念である。例えば，足から頭までの長さを cm という数値に置き換えたもの，これが身長という性質を測定するために適切な指標であるということに異論は少ないだろう。

では，測定対象が構成概念である場合，どのような点に配慮が必要だろうか。まず，測定しようとする構成概念の定義を明確にしなければならない。例えば，学力を「短時間でより多くの単純な情報を処理する能力」と定義した場合，2 桁の足し算を 3 分間に何問解く事ができるかといった問題によってテストを実施することで，その学力の違いを適切に点数に反映できる，といったことが考えられよう。

ただし，上記の問題はあくまでも学力を「大量の情報を素早く処理する能力」と定義した場合に妥当性の高い測定方法（問題）ということであり，学力の概念定義が異なれば妥当性の高い測定方法も異なる。学力という構成概念は，時代，社会・文化的背景，実践の文脈などによって異なるものであるし，しばしば，一元的な概念ではなく，多元的な内容から定義されることもある。したがって，まずは，評価の対象となる構成概念を明確に定義した上で，その概念を適切に測定できる方法（問題）を計画することが必要となる。

また，相対評価においては，どのような構成概念を評価対象とした場合であっても，最終的な評価結果が示していることは所属集団内の順位であり，目標となる能力の直接的な習得状況を示した指標ではない。エバリュエーションとしての教育評価においては，しばしばこのような妥当性の低さという点で相対評価が批判の対象となる。

心理尺度やテストの作成過程では，様々な段階において妥当性を高めるための手続きが実施される。例えば，専門家が協議することで測定対象となる構成概念の定義を明確化する。これは，測定の結果として得られたデータが測定したい構成概念と対応しているという「構成概念妥当性」を高めるための手続きである。また，多様な対象に予備調査を実施するなどして幅広く項目を収集し，体系的に整理することで，偏りのない質問項目や問題のセットを作成しようとする。これは，目的となる概念が偏りなく測定されているという「内容的妥当性」を高めるための手続きである。

上記の内容は，心理尺度の質問項目やテストの問題項目の作成段階での手続きに関するものであるが，実際に測定を実施した後も，得られた結果と外的基準との関連性を統計的に分析する作業が必要である。例え

ば，作成した尺度やテストの得点と，関連する別の構成概念を測定する既存の尺度やテストの得点との間にどのくらいの相関があるかを確認しなければならない。作成した尺度やテストが目的となる構成概念を適切に測定しているのであれば，関連する構成概念を測定している尺度やテストの得点との間には比較的高い相関が見られるはずである。このような方法は「併存的妥当性」とよばれ，外的基準となる既存の尺度やテストの妥当性が保証されているという条件付きではあるが，作成している尺度やテストの妥当性を保証する根拠となる。

また，測定対象としている構成概念を反映した未来の行動を外的基準とする「予測的妥当性」という観点もある。例えば，入社試験の成績が入社後のパフォーマンスと関連しているか，という観点から適切な入社試験のあり方について検討を行なうといった方法である。

3.2. 測定の「信頼性」

測定の「信頼性」とは，その尺度やテストが目的となる対象をどのくらい精度よく測定できているか，という概念である。例えば，ある学生に対してあるテストを行って，ある点数が得られたとする。このとき，テストの点数（＝測定値）が，その学生の実際の学力の程度（＝真値）を正しく反映しているという保証はない。なぜなら，あるテストによって得られる測定値は，真値と誤差の成分の影響を受けた点数だからである。例えば，学力が高い学生であっても，正しい選択肢を選ぶための根拠が全くないような質の悪いテスト問題に回答しなければならない場合には，勘や運だのみで回答しなければならないだろう。その場合，テストの結果として得られる点数というのは，「学力」の真値ではなく「まぐれ」という偶然の誤差の影響を大きく受けた点数となる。この場合，どんなに妥当性の高い問題を作成していたとしても，その点数をもって学生の能力を適切に測定できていると判断することは不適切である。

学力が構成概念である以上，真値そのものを直接に確認することは不可能である。しかし，測定値が真値を反映している精度を確認したり，その精度を高めたりする工夫は可能である。例えば，同じ尺度を用いて異なる二時点での測定を行い，その得点間の相関を計算する。仮に，測定値が真値の影響ではなく，それ以外の何らかの要因（誤差）による影響を大きく受けたものであるならば，前半の測定値と後半の測定値の間の一貫性は低くなるため，相関係数も低くなるはずである。また，問題や質問の項目数なども信頼性に大きく影響する。できるだけ多数の項目を用いて測定することは，問題選択の偶然性による測定値の変動を低く抑えるための１つの手段である。

妥当性と信頼性の両方を担保したテスト作成は非常に困難なことでもある。例えば「論理的な思考力」といった構成概念を測定しようとする場合，妥当性を高めるために，しばしば，ある程度まとまった量の論述を学生に求めるような問題を含むテストが用いられる。そのような問題の場合，限られた時間で回答できる分量には限界があるため，問題数を少なくすることになる。そうすると，たまたま，その問題の内容に関する予備知識があった（なかった）から，点数がとれた（とれなかった）という偶然の影響が大きくなってしまうため，測定の信頼性に乏しくなる。かといって，このような論述形式の問題を多く出題してしまうと，後半の特定には「疲労」の効果が影響することになり，極端な表現をすれば学力よりも体力を測定しているような妥当性の低いテストにもなりかねない。

4. 教育評価の方法の展開

エバリュエーションとしての教育評価の方法は，上記のような「妥当性」や「信頼性」といった観点に基づきながら，以下のような展開を見せている。

4.1.「真正の評価」論からの示唆

「目標に準拠した評価」が重視される今日，その展開に示唆を与えているのが「真正の評価」論である。「真正の評価」論の登場の背景には，日常の文脈からは断絶した学力を測定していたにすぎないという「標準テスト」への批判がある。すなわち，「「標準テスト」で良い成績をおさめたとしても，それは学校の中でしか通用しない特殊な能力を評価したにすぎず，生きて働く学力を形成したという保証にはならないのではないかという疑問や批判」（田中 2008，p.71）である。例えば Saxe（1988）は，ブラジルの路上でキャンディーを売る子どもたちの観察を通じて，彼らが仕入れと売買のために非常に高度な計算を行っているこ

と，その能力が学校に行った経験の長さと関連しないこと，同様の計算を抽象的な形式の問題にした場合には成績が下がってしまうことなどを示している。また，Arimoto（1991）は「4つのリンゴと7つのオレンジがあります。かけるといくつでしょう」といった意味のない計算問題に小学5年生の8割～9割がためらいなく回答してしまうことなどを示している。こうして，「実社会」「生活」「リアルな課題」といったことが意識され，子どもたちに実社会や生活のなかで生きて働く学力の形成がめざされるとともに，その評価のあり方が追究されたのである。

田中耕治は，「真正の評価」論をふまえた「目標に準拠した評価」の展開を以下の6点に整理している。すなわち，①評価の文脈と目標が「真正性」を持っていること，②構成主義的な学力観を前提としていること，③評価は学習の結果だけでなくプロセスを重視すること，④学習した成果を評価する方法を開発し，さらには子どもたちも評価方法の選択ができること，⑤評価は自己評価を促すものでなくてはならないこと，⑥評価は教師と子どもとの，さらには保護者や地域住民も含む参加と共同の作業であること，である（田中2008，pp.76-78）。

ここでは教育評価の具体的な例として，ポートフェリオ評価とパフォーマンス評価の二つを取り上げてみたい。はじめに，ポートフォリオ評価である。ポートフォリオとは，「学習において，自分はどのようなことに努力しているか，どこがどのように成長したか，何を達成したかなどについての証拠となるものを，目的，目標，規準と基準に照らして，系統的・継続的に収集したもの」（岸本2010，p.106）のことである。ポートフォリオに収められるのは，「①学習の成果としての作品や学習のプロセスを示す作業メモ，②子どもの自己評価，③教師による指導と評価の記録など」（岸本2010，p.106）である。すなわち，ポートフォリオ評価とは，こうして収集されたものへと注目して，子どもの学びのプロセスを把握・評価の対象とするものである。

ポートフォリオ評価の意義は，学びのプロセスに注目した評価法であるということだけではない。「ポートフォリオ検討会」によって，①「教師と子供たちの共同作業としての教育評価」になること，②「保護者や地域住民も「ステイクホルダー」」として評価に巻き込むこと，③「「自己評価」をはげます「検討会」」となること，といったさらなる意義をもつものになる（田中2008，pp.161-164）。指導と評価の一体化という考え方に示されているように，このような活動は，教師にとっては評価であるが，子どもにとっては学びのプロセスでもある。すなわち，自らの学習の履歴を振り返り，価値づけ，次の学習についての見通しをもつという営みを通じて「自己調整学習」（Pintrich & De Groot 1990）のための能力の育成が期待されるのである。

次に，パフォーマンス評価について見てみよう。パフォーマンス評価とは，「ある特定の文脈のもとで，様々な知識や技能などを用いて行われる人のふるまいや作品を，直接的に評価する方法」（松下2007，p.6）である。具体的には，「「パフォーマンス課題」を与えて解決・遂行させ，それを複数の評価者が，「ルーブリック」と呼ばれる評価基準表を用いながら，評価」する（松下2007，p.7）。

パフォーマンス評価のように，子どもの自由な表現を引き出す評価課題では，子どもの反応に多様性と幅が生じるため，質的な判断が求められることになる。そこで，「ルーブリック」と呼ばれる質的な採点指針を用いることが有効となる（石井2010，p.48）。ルーブリックの開発手順は，例えば以下の通りである。「①試行としての課題を実行し多数の児童生徒の作品を集める。②あらかじめ数個の観点を用いて作品を採点することを同意しておく。③それぞれの観点について一つの作品を少なくとも3人が読み6点満点で採点する。④次の採点者にわからぬよう付箋に点数を記して作品の裏に貼り付ける。⑤全部の作品を検討し終わった後で全員が同じ点数をつけたものを選び出す。⑥その作品を吟味しそれぞれの点数に見られる特徴を記述する。」（石井2010，p.49）さらに，モデレーション（moderation: 調整）を通して，ルーブリックはたえず再検討されることによって妥当性と信頼性を高められる。モデレーションの進め方は以下のようになる[(2)]。すなわち，明確な評価基準の策定と作品例の提供，評価者への訓練，統計的手法，査察，被評価者によるアピール，機関レベルの認定，本質的なモデレーション，である。

ポートフォリオ評価やパフォーマンス評価は，いわゆる低次の教育目標から高次の教育目標まで対応して妥当性を高めることを追求して多様に展開されてきた。そこには，西岡（2010a）が「「真正の評価」は，信頼性を多少不問にしても，妥当性を確保することが重要だという考えにもとづいています。」（西岡2010a，p.69）と述べるように，信頼性よりも妥当性を重視する傾向が見られる。

一方，教育測定の立場からは，「一般化可能性理論」の考え方を用いてパフォーマンス評価の信頼性を検

討する試みも提案されている。一般化可能性理論においては，評価結果の点数のばらつきを受験者の能力の影響，課題の難易度による影響，評定者の違いによる影響，それぞれの組み合わせの影響，といったように分解して検討を行うことができる。例えば，あるパフォーマンス評価の基準では評定者間での採点のばらつきはあまりないが，項目と受験者と課題の組み合わせによる影響が大きい（項目によって受験者の順位が入れ変わっている）といった情報を得ることができる。この場合，基準自体を精緻化することは信頼性の向上にはあまり寄与せず，項目数の数を増やすことが信頼性を高める有効な手立てとなりうる，といった示唆をえることができ，信頼性と妥当性のバランスを検討することが可能になる。

4.2. 目標にとらわれない評価

目標に準拠した評価への転換が図られる一方で，それを全面的に採り入れることへの批判意識もある。それは，目標にとらわれない評価を重視する立場からの指摘である。カリキュラム開発の二つのモデルとして「工学的アプローチ」と「羅生門的アプローチ」が挙げられるが，この場合，前者が「目標に準拠した評価」を，後者が「目標にとらわれない評価」を必要とする。

「目標にとらわれない評価」に注目が集まるのは，「「目標に準拠した評価」は，教師による「目標」が規準となることから，それに回収されない子どもたちの活動を見落とす危惧があり，まずは子どもたちの，ひいては保護者や地域住民の教育評価への「参加」が保障される必要があり，そのことによって，多面的に多層的に教育活動を把握することが可能になってくるのではないか」（田中 2014，p.184）というように，一つには評価行為への多様なステイクホルダーの参加の要求という形で位置づけられている。

また，ナラティブ・アプローチの観点から，学力形成および評価をとらえ直そうとする試みもあり，これもまた「目標にとらわれない評価」を授業あるいはカリキュラムづくりに位置づけ直す動向と言えるだろう。例えば，黒谷和志は，「知識や技能を獲得させることに囚われるのではなく，学習者にとっての生活現実がいかに新たな意味をもってつかみ直されるのかを問う」（黒谷 2012，p.25）ことの重要性を主張している。「学力テストが広がりを見せる時代においては，学力を数量的に捉える思考様式や学力・リテラシーを効率的に向上させる手法が広がりをみせ，子どもたちから語り出される物語は，余分なものと位置づけられがちとなる（黒谷 2012，pp.32-33）」と指摘し，子どもたちの物語を授業に位置づけなおすことを提起している[(3)]。

こうした「目標にとらわれない評価」の試みは測定における妥当性や信頼性を高めることを否定するものではない。あえて言えば，そうした手続きを経て，より「確かに」数量的に捉えられるもので教育実践への評価を完結あるいは収斂することに対する危機意識であろう。「現在の PISA では，＜機能的（適応的）側面―批判的（創造的）側面＞，＜経済的観点―政治的観点＞，＜労働力形成―市民形成＞といった対立軸のうち前者のみが肥大化している。後者の側面・観点・目的を取り戻していくということが対抗的な教育実践の大きな方針になるだろう」（松下 2013，p.20）という指摘にもあるように，数量的に捉えられたものは，それぞれの前者と親和性の高いものであるが，教育という営みにおいて後者を忘れてはならないはずである。

「目標にとらわれない評価」という発想は，学力を矮小化せずに子どもを評価しようとするという点において，妥当性の高い評価方法といえるだろう。また，事前の固定された枠組みに子どもを無理にあてはめるのではなく，実際の学びの営みを通じてダイナミックに評価規準や基準を生成していこうとする試みであるということもできるかもしれない。

ただ，教育測定の立場から言えば，こうした利点を積極的に主張するのであれば，やはり評価しようとしている対象は何であり，どのような事実を，どのように解釈した結果，どのような評価になったのか，という過程を明確化し，他者と共有しうるものにしようとする営みが不可欠である。その営みが欠けてしまえば，妥当性という観点での議論の俎上に乗せることすら不可能であろう。これは，信頼性の観点からも同様であり，その尺度やテストが，何を，どのように数量化しているかという点が明確にされていなければ，その数値が意味することを共有することは不可能である。「目標にとらわれない評価」の展開においては，どのような方法（定量的か定性的か）といった点だけでなく，教育実践を多様に解釈，言語化する高度な実践的鑑識眼をいかに評価者に育んでいくかという議論が不可欠であろう。

5. テスト理論の展開

ここまでは，エバリュエーションという概念を重視した教育評価の展開について論じてきた。一方でメジャメントの側面を重視した教育評価においては，テスト理論の枠組みから，受験者集団に依存せず，複数のテストの結果について統一的な見解を得るための方法として「項目反応理論」に関する知見が蓄積されている。以下では，古典的テスト理論の考え方と対比しながら，項目反応理論を用いたテスト開発の特徴について概説する。

5.1. 古典的テスト理論

一般的にテストを作成する場合には，実施前に問題のセットを作成する段階が中心だと考えられているのではないだろうか。確かに，どのような構成概念について，どの程度の，どういった種類の理解を測定したいのか，といった点を十分に検討し，その概念に合致した，偏りのない，かつ現実的な制約の中で実施可能な範囲で，できるだけ多くの問題を準備するといった作業にコストをかけることは，妥当性や信頼性を高める上で非常に重要な作業である。

一方で，そのようにして作成したテストが，実際に想定通りに適切に機能していたのかという観点においては，テストの結果をもとにしながら，個々の問題の性質に関する様々な情報を引き出すことが重要となる。このような作業が項目分析である（日本テスト学会 2010)。具体的には，各問題の正答率（通過率）を計算することで，各問題の難易度に関する情報を得る。また，テストの総得点と分析対象となる問題の得点との相関係数（IT 相関）や，分析対象となる問題以外の問題の点数の総和と分析対象となる項目の得点との相関係数（IR 相関）を指標とすることで，各項目がテスト全体で測定しようとしている何らかの構成概念を一貫して測定しており，また，受験者の能力を識別する性能（識別力）をどの程度有しているか，という指標を得る事ができる。例えば，難易度が高すぎるために受験者の大半が不正解になってしまうような問題や，逆に，難易度が低すぎるために受験者の大半が正解する問題，また，正しい選択肢が存在しなかったり，複数存在したりするなど内容に不備がある問題，テスト問題全体として測定している概念とは明らかに異なる概念を測定しているような問題，などが含まれていると一般に識別力は低くなる。

受験者を成績順に複数のグループに分けて，各グループの受験者の正答率の変化や選択肢の選択率の変化などをグラフ化した，項目特性図を用いることで，各問題がどのような受験者層に対して有効に機能するのかという情報を得る事なども可能である。例えば，難易度の低い問題は全体としての識別力は低いが，特に学力が低い受験者を識別するためには有効な問題項目として機能している可能性も考えられる。項目分析では，テスト結果に関する様々な情報を用いてテスト問題の検討を行なう。

5.2. 項目反応理論

上記の項目分析の方法は古典的テスト理論と呼ばれる枠組みから示されてきたものである。このような項目分析を用いることで，テスト問題に関する豊かな情報を引き出すことができる。ただし，古典的テスト理論の考え方では，項目分析から得られた知見について，テストを受験した受験者集団の特性とテスト問題の性質を切り離して検討することが困難だという限界がある。仮に，A 学校でテスト α を実施し，B 学校でテスト β を実施したとする。α と β は数学の同じ領域に関するテストである。もし，A 学校の平均点が 70 点で B 学校の平均点が 60 点であったとした場合に，学校 A の生徒は学校 B の生徒よりも学力が高いと言えるだろうか。もしかすると，実際には学校 B の生徒の方が学校 A の生徒よりも学力が高いにも関わらず，たまたまテストに含まれていた問題の難易度が高かったために点数が低くなってしまったのかもしれないし，本当に学校 B の生徒の学力が低かったのかもしれない。仮に点数を標準化して偏差値に換算したとしても状況は同じである。A 学校における偏差値 50 と B 学校における偏差値 50 が同じ能力であるという明確な根拠はなく，偏差値 50 という得点がもつ意味は各学校の受験者集団の水準に依存して異なってくる。

もちろん，A 学校と B 学校の学生が共通した試験問題を受験すれば直接的な比較が可能である。しかし，しばしば，現実的な制約のためにそういった処遇が不可能なケースもある。また，2014 年度の学生と 2015 年の学生の能力について比較したいとか，同じ学生集団について 2014 年時点の学力と 2015 年時点の学力を比較して経年変化を調べたいといった調査を行なう場合，2 時点間で同じ問題を使用することは不可能であり，それぞれの時点で異なる問題を使用することになる。こういった場合に，受験者集団の能力とテストの

性質を切り離して考えることができない古典的テスト理論の枠組みでは，テスト間で結果を比較することが困難となる。

このような古典的テスト理論の限界性を克服するために，因子分析のように，各個人の潜在的な特性を想定し，その潜在的な特性を推測した尺度値を用いて分析を行なうことで，個人の能力と問題の特性とを切り離した形で分析可能にしているのが，項目反応理論と呼ばれる新しいテスト理論の考え方である（豊田2002）。例えば，1母数のモデルでは，個人の能力特性値$\theta = 0$の受験者が50%の確率で正答できる問題，$\theta = 1$の受験者が50%の確率で正答できる問題，といった形式で，個人の能力推定値に基づいた問題の性質が導かれる事になる。また，2母数のモデルでは，問題項目がどの能力値の受験者に対して，どのような正答確率の変化を示すか，といった識別力に該当する指標等を得る事もできる。また，複数の異なるテスト間で受験者の一部を共通させたり，問題の一部を共通させたりすることによって，その情報を基準としながら難易度や識別力を等化することで，テスト間での比較を行なうことも可能となる。

このように項目反応理論を用いて能力特性値で問題の特徴を把握することにより，異なるテスト間で解答者の能力水準を比較する事が可能になる。例えば，ある児童の小学校3年生の時の算数の能力と4年生の時の算数の能力について，それぞれ異なるテスト問題で測定したとしても，能力特性値という共通した尺度上での比較が可能になるということである。酒井・野口（2015）は，成人を対象とした精神的健康のスクリーニングテストである，GHQ-30，UPI，K10という3つのテストを共通尺度化している。これらのテストはそれぞれ異なる項目，異なる採点方法で精神的健康の度合いを判断するものであり，GHQ-30は7点以上，UPIは30点以上，K10は15点以上をそれぞれカットオフポイントとしている。大学生548名がこの3種類のテストに回答した結果をもとに，項目反応理論を用いた共通尺度化を行なった結果，各テストはほぼ同程度の健康度（$\theta = 1.60 \sim 1.79$）をカットオフポイントとしていることが示された。このように，項目反応理論を用いることで，受験者や項目の異なる複数のテスト間での比較なども可能となる。

また，様々なテスト問題が共通した尺度上での難易度と識別力という観点で特徴づけられることによって，どのような受験者にはどのような問題を用いると最も効果的に能力を識別する事が可能か，といった点についての検討も可能となる。このようなメリットはTOEICやTOEFLなど様々なCBT（Computer Based Testing）において活用されている。

6. メジャメントと学力格差問題

2節で述べたように，教育評価ではエバリュエーションが重視される一方，メジャメントの問題は後方に退いていった。しかし，正規分布を利用したり，子どもたちの成績を測定したりすることが，教育の改善のために必要な場面も存在する。その典型的な例が，近年の学力格差をめぐる研究の進展である。ここでは，学力格差をめぐる研究の進展と，そこでどのようにメジャメントの理論が利用されているかという点について論じてみたい。

6.1. 学力格差をめぐる議論

2000年以降，日本でも学力格差をめぐる議論が進展してきた。その中心となったのは，おもに教育社会学者たちである。かれらの議論の特徴は，「学力とは何であるか」という議論はいったん置いておいて，「測定された学力（≒成績）」に焦点を当てたことにある。教育社会学の学力研究は，議論をいわゆる成績に限定することで，家庭の経済力や文化，両親の学歴，本人の性別などの社会的要因により，子どもたちの成績に差が生じていること，その差が次第に拡大していること，さらにこうした格差を克服する実践の在り方を明らかにしてきた（苅谷他2002，苅谷・志水編2004，耳塚2007，志水編2009）。

こうした研究は，子どもたちの家庭環境（社会経済的背景／Socio Economic Status: SES）がきわめて成績に強い影響を持っていることを指摘するのみならず，なぜ，皆が平等に学校に通っているにもかかわらず，こうした関連が持続するのか，という問題を追及してきた。教育評価の文脈に照らして重要なのは，こうした議論が生み出した「学校は社会の平等化を推進するどころか，社会的不平等の再生産とその正当化に役立っているに過ぎない（志水1990，p.18）」という視座に立つ研究の知見である。これらの研究のうち，教師に関わる部分を要約すると，次のようになる。教師たちは中産階級の出身であり，社会経済的に恵まれない子どもたちと文化的な距離がある。それが，社会経済的に恵まれない子どもたちの問題行動が，教師た

ちの目に理解できないものと認識される大きな原因であり，結果として，かれらに「できない子」「問題行動を起こす子」といった近視眼的なレッテルを貼ることを許してしまう。そして，振る舞いを正そうと厳しく統制したり，補充的な学習ばかりを行ったりすることで，かれらのやる気を損なわせ，不平等の再生産に荷担してしまうのである（西田 2012，pp.183-186）。これらの研究は主として欧米の研究の知見を元にしたものだが，日本でも格差・貧困をめぐって同様の指摘がなされている（長尾・池田編 1990，盛満 2011）。さらに，こうした研究は，成績だけでない多様な指標で子どもを評価しようという試みが，コミュニケーション能力等のあいまいな指標を持ち込むことで，社会経済的に不利な立場に置かれやすい子どもたちを排除し，けっきょく中産階級の子どもたちを利することになるというプロセスを描いている[(4)]。

教育測定の文脈においては，こうした文化的偏りを検知する方法も存在している。たとえば，特異項目機能（differential item functioning: DIF）という技法を使えば，文化・民族性・性別などの要因によって，正答確率が変動する項目を発見し，取り除くことが可能である（鈴川 2013，岩間 2013）。しかし，こうした偏りはテスト場面のみならず，日々の授業や学校生活を通して，ときに無自覚に行われるものであり，統計的技法で検出できるのは，その一部に過ぎない。とはいえ，議論を「測定された学力」というメジャメントの世界に限定することで，学力格差という新たな問題が日本の教育の争点として浮上することができたという点は，強調しておく必要があるだろう。

6.2. 大規模調査をどのように設計するか

教育評価においてメジャメントの議論が不十分であったことは，大規模な学力調査の信頼性を失わせることにもつながっている。たとえば，2015 年度の全国学力・学習状況調査において，大阪府の中学校の順位は大きく向上した。これは，全国学力・学習状況調査を内申点に反映させるという教育委員会の方針を受け，関係者が成績を向上させようと躍起になったためであると思われる。この事例は，評価の結果的妥当性（西岡 2010c）を歪めた典型例であると言えるだろう。こうした動きを，学力調査の意図（一人一人の子どもの達成状況を測るために行っている）を踏まえない行為であると批判することもできる。しかし，これはむしろ，調査設計のミスと考えた方がよい。

一般に，社会調査は調査対象に影響を与えないように行うことが好ましい。今回の大阪の事例のように，被調査者の一部だけに調査に積極的に答えるインセンティブが与えられると，調査データに分析者が統制できない偏りが生じ，適切な分析を施すことができなくなってしまうからである。他にも，悉皆で実施されている全国学力・学習状況調査にはさまざまな欠点がある。たとえば，すべての子どもに同じ問題を出題するため，出題範囲が限られてしまい，けっきょく全体のことがよくわからなくなっている。また，すべての問題を公開することを原則としているため，毎年問題をすべて変更する必要がある。これでは，子どもたちの成績が変化したときに，それがテスト問題が違うためなのか，何らかの要因によるものなのか，区別することが不可能になる。

こうした問題を回避するための統計技法が，PISA や TIMSS という大規模学力調査には，いくつも採用されている[(5)]。代表的な技法は，標本抽出法である。一般に，教育政策の影響を知りたいのであれば，すべての学校・子どもを調査対象に含める必要はない。全体を調査するとなると，時間も予算もかかりすぎるし，データセットが大きくなりすぎて整理することもむずかしくなる。さらに，今回の大阪府の事例のように，テスト結果を当初の目的と異なる形で運用しようとする人々が出てくる可能性も高くなる。加えて，5 節で述べた IRT の技法を使い，複数の年度の調査を同一尺度上に等化することも必要である。

その他，大規模学力調査で重要な統計技法として，ここで重複分冊法と推算値法（Plausible Values）を取り上げておきたい。重複分冊法とは，出題したい問題を複数の冊子に分解し，配布する方法である。冊子間には重複する箇所があり，その重複を利用して IRT による等化を行うことで冊子間の成績を比較することが可能になる。このように，重複分冊法には，個々の回答者が答える問題数を抑えることができる一方で，全体としては出題数を多く保つことができるという利点がある。

推算値法は，重複分冊法に伴う個々の推定値の荒さをカバーするための統計技法である。重複分冊法では，個々の回答者は，すべての問題に答えるわけではないため，回答は欠損した状態になっている。推算値法では，この欠損を回答者の情報（配布された質問紙に対する回答のみならず，生活実態調査や属性情報なども含む）を用いて補完することで，より正確な推定値を得ようとする。これは，データの欠損をどう処理するかという統計学の近年の研究を応用した分析法である。

重複分冊法にせよ，推算値法にせよ，いずれの技法も悉皆調査を前提とできない状況で，標本抽出によって全体をよりよく推定しようという発想に立っている。近年の大規模学力調査には，こうしたメジャメントに関わる技術がいくつも採用されているのである。

この点，全国学力・学習状況調査が悉皆で実施されること・調査問題をすべて公開することを後押しする理論として，梶田や田中らの教育評価の発想が使用されたという点も指摘しておかねばならない。大規模学力調査では，田中（2008）が批判した「成績が正規分布すること」こそが重要である。たとえば，少人数指導の影響力を測定することを考えてみよう。「目標に準拠した評価」の場合，テストは全員満点が好ましい。達成基準に全員が到達することを目指すのだから，これは当然である。ところが，全員が満点を取ってしまうと，小員数指導をした学校もしていない学校も全員満点で，いったい少人数指導の影響はどうなのか，という肝心の問題はわからなくなってしまう。

木村（2006）や鳶島（2010）らは，教育評価の必要性を訴える梶田叡一が「悉皆調査でなければわからないこともいっぱいあります」と述べ，指導に活かすことを全国学力・学習状況調査の目的の一つと捉える教育評価の論理を使って，全国学力・学習状況調査の悉皆調査を後押ししたことを指摘している。これを妥当性・信頼性という枠組みから捉えるなら，妥当性を重視するあまり信頼性を失ってしまったということが，現行の学力調査の問題点の一つだと言えるのである。

6.3. メジャメントによる議論の限界

ここまで，教育評価におけるメジャメントの重要性を指摘してきた。一方で，メジャメントを重視する学力研究に，一定の限界があることもまた認めなければならない。PISA はその典型であり，各種の調査・統計技術を駆使したにもかかわらず，日本を含め，各国の教育政策に強力なインパクトを与えている。また，教育測定が盛んなアメリカでも，NCLB 法の成立以後，各学校の成績は厳しく管理されるようになり，テスト主義による弊害が次々と生まれている。

とくに根本的な問題は，測定しようとする何らかの尺度は，それを測定する人々が，現在の社会において必要だと判断したものであり，はじめから社会の中心に近い人々にとって有利な尺度である可能性が高いという点である。近年のアメリカの右派による新自由主義・新保守主義的な教育改革に批判的な立場をとる Apple は，こうした改革を支える集団として，教育測定にも批判的な目を向けている。

> 右派同盟の四つ目のグループは，専門職に従事したり，マネージメントの専門知識を持つ新中産階級である。この派に属するものは，経営管理や効率に関する専門知識を駆使し，新自由主義的な市場化と新保守主義的な知の支配が要求するアカウンタビリティ，評価，情報の生産，測定制度などを実施する役割を担う。教育においては，この派は，狭義のアカウンタビリティに基づいた卒業・進級テストや統一テストなどの政策を支えると同時に，彼ら自身こうした諸政策から恩恵を受けている。というのも，彼らこそがこうしたシステムを稼働させるうえで不可欠な技術的ノウハウを持っているからである。（Apple & Wayne 2009, p.26）

こうした Apple の指摘には，日本の現状を批判する上で，必ずしも適切とは言えない部分もある。日本の場合，IRT に代表される教育管理の技術はそれほど高度化されておらず，木村（2006）が指摘するように，学力調査にテストの専門家が関わっているわけでもない。さらに，先にも述べたように，教育社会学者たちが「測定された学力」に議論を限定したことが，学力格差という新しい問題を，日本の教育の争点として浮上させたことも事実である。こうしたことを踏まえるなら，Apple らが言うような教育測定の危険性は認めつつも，教育測定の議論を適切に利用しながら現状を改善する道を模索するべきであろう。

ここで鍵になるのは，これまで日本で培われてきた教育評価の再検討であるように思う。そもそも，何のため，誰のために学力テストを実施するのか。学力テストの結果をどう理解し，何のために使うのか。教育評価の考え方を，たとえば「指導と評価の一体化」という言葉で教室内に閉じ込めてしまうのではなく，より広い文脈から位置づけ直す作業が必要とされているように思う。その意味では，「メジャメントからエバリュエーションへ」ではなく，「メジャメントもエバリュエーションも」が，おそらくあるべき解なのではないだろうか。

7. まとめ

本稿では，教育評価をめぐる，教育方法学・教育心理学・教育社会学の知見を整理してきた。前半では，おもに教育評価におけるエバリュエーションの問題を扱い，相対評価批判から目標に準拠した評価へいたる流れと，妥当性・信頼性概念の再検討，新しい教育評価の流れについてまとめた。後半では，「メジャメント（measurement）」に関わる新しい潮流として，近年の IRT や学力格差といった研究を紹介してきた。

本稿の議論は多岐にわたっているが，多くの節で論点となっているのは，妥当性と信頼性の両立をどのように目指すかという問題意識である。教育評価では「メジャメントからエバリュエーションへ」という言葉からもわかるように，妥当性に傾倒する傾向が強い。しかし，本稿の各節で指摘したように，信頼性を欠いた測定は，「一体何を測っているのか？」という重要な問題に答えないまま，教育実践を行うことにつながりかねない。あるいは，意図しないままに教育格差を拡大したり，不適切な学力調査を後押ししたりして教育現場の混乱を招く可能性もある。

2 節で触れたように，メジャメント批判は相対評価への批判として成立してきた。しかし，新しいテスト理論である IRT は，正規分布を利用しているとは言え，単純に相対評価であると切って捨てることのできない豊かな理論的含意を有している。その意味では，単純なメジャメント批判は，今日すでにその前提が崩れているのである。メジャメントにおける新しい潮流を取り入れながら，あらためて教育評価の在り方を再構築することが求められていると言えよう。

残念ながら，こうした取り組みは本稿の枠を超える。おそらく，「目標にとらわれない評価」とは何なのか，あるいはそもそも「評価」とは何なのかといった，より根源的な問いを扱わなければならないのだろう。こうした問題は，今後の課題としたい。

最後になるが，本稿は 3 名の教育学に関わる研究者の共同執筆によるものである。分担は川口が 1・6・7 節を，松尾が 3・5 節を，樋口が 2・4 節をそれぞれ主に担当している。研究会を開始した当初は，同じ教育学領域と気楽に構えていたのだが，教育現象に対する見方が異なる 3 名による執筆は，なかなか論点の整理がむずかしく，負担の大きいものであった。他方で，それぞれのこだわりや，分析視覚の違いから，新たな発見や課題が見えることは少なくなく，多分に刺激的な集まりでもあった。

自身の例で恐縮だが，教育社会学領域の研究者は，社会調査の精度にはこだわるが，教育評価の妥当性にはまったくといっていいほど無頓着である。ことに日本の場合，近年の学力研究の前提であるはずの IRT が言及されることすら稀である。IRT は教育心理学領域を中心に発展した技法であり，教育社会学者には馴染みが薄い。これを学問の専門分化といえば聞こえはいいが，IRT を使っていない学力調査を前提とする学力研究では，諸外国の水準にはとうてい及ばない。やむなく他領域に踏み込んで IRT や教育評価を学んだものの，他領域は他領域で閉じているように思えることが常々不満であった。こうした状況をどうにかしたいという思いが，本稿執筆の動機の一つである。学問分野のタコツボ化が批判されて久しいが，こうした試みが，現状を変えていくことを願ってやまない。

【引用文献】

Apple, M. W. & Wayne A, 2009,「批判的教育学の政治，理論，現実」Apple, M. W, Whitty, G & 長尾彰夫編著『批判的教育学と公教育の再生―格差を広げる新自由主義を問い直す―』明石書店，pp.9-38.

Arimoto, N, 1991, “A computer tool designed to change children’s concept of school math”, *Educational Technology Research*, Vol. 14, pp.11-16.

石井英真，2010,「ルーブリック」田中耕治編『よくわかる教育評価　第 2 版』ミネルヴァ書房，pp.48-49.

岩間徳兼，2013,「特異項目機能（DIF）― IRT に基づかない方法―」豊田秀樹編『項目反応理論【中級編】』朝倉書店，pp.100-118.

苅谷剛彦・志水宏吉編，2004,『学力の社会学』。

苅谷剛彦・清水睦美・志水宏吉・諸田裕子，2002,『調査報告「学力低下」の実態』岩波ブックレット。

岸本実，2010,「ポートフォリオ評価法」田中耕治編『よくわかる教育評価　第 2 版』ミネルヴァ書房，pp.106-107.

木村拓也，2006,「戦後日本において「テストの専門家」とは一体誰であったのか？―戦後日本における学力調査一覧と「大規模学力テスト」の関係者一覧―」『教育情報学研究』第 4 号，pp.67-100.

黒谷和志，2012,「これからの学力・リテラシー形成と教育方法」山下政俊・湯浅恭正編著『新しい時代の教育の方法』ミネルヴァ書房，pp.20-34.
松下佳代，2006,「評価の妥当性・信頼性・客観性」辰野千壽・石田恒好・北尾倫彦監修『教育評価事典』図書文化社，p.66.
松下佳代，2007,『パフォーマンス評価―子どもの思考と表現を評価する』日本標準。
松下佳代，2013,「PISA の影響の下で，対抗的な教育実践をどう構想するのか」日本教育方法学会編『教育方法 42　教師の専門的力量と教育実践の課題』図書文化社，pp.10-24.
耳塚寛明，2007,「小学校学力格差に挑む だれが学力を獲得するのか」『教育社会学研究』第 80 集，pp.23-39.
盛満弥生，2011,「学校における貧困の表れとその不可視化―生活保護世帯出身生徒の学校生活を事例に―」『教育社会学研究』第 88 巻，pp.273-294.
諸岡康哉，1987,「指導と評価の一体化」吉本均編『現代授業研究大事典』明治図書，pp.545-547.
長尾彰夫・池田寛編，1990,『学校文化―深層へのパースペクティブ―』東信堂。
日本テスト学会，2010,『見直そう，テストを支える基本の技術と教育』金子書房。
西岡加名恵，2010a,「妥当性と信頼性」田中耕治編『よくわかる教育評価　第 2 版』ミネルヴァ書房，pp.68-69.
西岡加名恵，2010b,「比較可能性とモデレーション」田中耕治編『よくわかる教育評価　第 2 版』ミネルヴァ書房，pp.72-73.
西岡加名恵，2010c,「公正性と実行可能性」田中耕治編『よくわかる教育評価　第 2 版』ミネルヴァ書房，pp.74-75.
西田芳正，2012,『排除する社会・排除に抗する学校』大阪大学出版会。
OECD, 2014, PISA 2012 *Technical Report*, OECD Publishing.
Olson, J. F, Martin, M.O, & Mullis, I.V.S. eds, 2008, TIMSS 2007 *Technical Report*, TIMSS & PIRLS International Study Center.
Pintrich, P. R. & De Groot, E. V, 1990, “Motivational and Self-Regulated Learning Components of Classroom Academic Performance”, *Journal of Educational Psychology*, 82, pp.33-40.
酒井渉・野口裕之，2015,「大学生を対象とした精神的健康度調査の共通尺度化による比較検討」『教育心理学研究』第 63 巻，pp.111-119.
Saxe, G. B. (1988). Candy selling and math learning *Educational Researcher* 17, 14-21.
鈴川由美，2013,「特異項目機能（DIF）― IRT に基づく方法―」豊田秀樹編『項目反応理論【中級編】』朝倉書店，pp.78-99.
鈴木和夫，2005,『子どもとつくる対話の教育―生活指導と授業』山吹書店。
志水宏吉，1990,「学校文化論のパースペクティブ」長尾彰夫・池田寛編『学校文化―深層へのパースペクティブ―』東信堂，pp.11-42.
志水宏吉編，2009,『「力のある学校」の探求』大阪大学出版会。
田中耕治，2008,『教育評価』岩波書店。
田中耕治，2013,『教育評価と教育実践の課題―「評価の時代」を拓く』三学出版。
田中耕治，2014,「教育評価」日本教育方法学会編『教育方法学研究ハンドブック』学文社，pp.182-185.
鳶島修治，2010,「全国学力テストの悉皆実施はいかに正当化されたか―教育評価と＜学力保障＞のポリティクス―」『社会学年報』第 39 巻，pp.75-86.
豊田秀樹，2002,『項目反応理論［入門編］』朝倉書店。
山田哲也，2014,「カリキュラムと学力」耳塚寛明編『教育格差の社会学』有斐閣アルマ，pp.25-52.

【注】

(1) 評価の客観性について松下佳代は次のように説明している。

> 妥当性と信頼性が対になって使われる測定論の専門用語であるのに対し，客観性はより一般的な用語であり，大きく 2 つの文脈で使われる。1 つは，測定論の文脈で使われる場合である。この場合は，

だれが評価しても常に同じ結果が得られるという意味であり，採点の信頼性とほぼ同義である。つまり，客観性は信頼性に含まれる。もう１つは，最近の教育評価の理論・実践の文脈で使われる場合である。この場合には，評価は本来，当事者の主観を含んでなされるものだとされ，客観性は異質な複数の評価者による評価結果の突き合わせ・交渉という形で実現されることになる。」（松下，2006, p.66）

相対評価は，評価用具ではないため，松下の説明をそのまま当てはめることはできないだろう。また，田中も，その後の「目標に準拠した評価」の登場に対して客観性を問う指摘があったことの背景として，「「相対評価」の「客観性」に対する幻影が，未だ根強く息づいているということ」（田中，2013，p.27）と指摘しているように，相対評価がもっているとされる「客観性」については，正確に検討する必要があろう。

(2) モデレーションの手続きの詳細については，西岡（2010b）を参照されたい。
(3) ここで取り上げられている実践については，鈴木（2005）を参照されたい。
(4) 学力との関わりで，こうした問題を整理したものとしては，山田（2012）を参照されたい。
(5) 本稿で紹介した大規模調査に関わる技法としては，PISA や TIMSS の Technical Report を参照されたい（OECD2014, Olson et al. 2008）