【研究論文】

学力調査と教育評価

田　中　耕　治

皆様、こんにちは。只今、ご紹介のありました京都大学の田中耕治です。本日は、教育研究に関する国際会議にお招き頂き、大変に光栄に思っています。京都大学は、千年間も都が置かれた京都市にありまして、日本ではノーベル賞を多く受賞する大学として知られています。

私の専門は、おもに学校におけるカリキュラムや授業、さらには教育評価のあり方について研究をしております。最近では、私は心理学の立場ではなく、教育学の立場から、教育評価の理論や実践について研究を進めております。

さて、振り返ってみますと、第二次世界大戦後の日本におけるカリキュラム改革の歴史

は、常にそれぞれの時代の学力問題と学力調査が契機となって、展開してきました。現代の日本で進行しているカリキュラム改革も、1999年から始まった「学力低下論争」を契機としています。そして、2008年にはナショナル・カリキュラムである「学習指導要領」が改訂され、文部科学省は「ゆとり教育」政策から「学力向上」政策へと大きく方針を転換しました。この改革論議の最中に、2007年から実施されるようになった「全国学力・学習状況調査」は、今日のカリキュラム改革とともに教育評価研究に大きな影響を及ぼすことになりました。

本日は、「全国学力・学習状況調査」を分析することを通して、現代日本で行われている教育改革の特徴とともに、今後の課題を明らかにしたいと思います。本日の発表は、次の三点に即して、話をいたします。

1．「全国学力・学習状況調査」の背景
2．「全国学力・学習状況調査」の内容分析
3．「全国学力・学習状況調査」の社会的文脈

1．「全国学力・学習状況調査」の背景

まず最初に、「全国学力・学習状況調査」がなぜ実施されるようになったのかを説明します。

日本では、1999年から「学力低下論争」が開始されました。その論争を象徴する本が、1996年6月に出版されました。その本の名前は、『分数ができない大学生』というものでした。大学・短期大学の進学率がほぼ50%に達した現代の日本において、「小学校で教える分数」が解けない大学生が存在するという事実は、大きな衝撃を与えました。

大学生の学力低下問題は、その原因として、学習指導要領における「ゆとり教育」政策にあるとの批判へと発展していきます。教育現場ではいわゆる「読み・書き・算」の基礎学力向上をめざすドリル学習が盛んとなりました。このような「ゆとり教育」政策への批判的な状況に呼応するように、文部科学省から2002年には「学びのすすめ」というアッピールが出され、教育課程行政の転換が始まりました。

ところで、この度の「学力低下論争」においては、過去にはなかった新しいインパクトが加わりました。それがPISAの調査結果であり、PISAが採用した学力の「活用力」を意味する「リテラシー」概念が注目を受けることになりました。ご承知のように、PISAが提起した「リテラシー」は今やグローバル・スタンダードとして、各国の教育政策に影響を与えています。日本においては、とくに2003年に実施された第2回PISAの結果に注目が集まりました。日本の生徒は科学や数学の分野では国際的にトップ・クラスの成績でしたが、「読解力」では成績が良くありませんでした。そこで、文部科学省は「読解力」重視を強調するようになります。

以上のように、この度の学力問題においては、その

当初いわゆる「読み・書き・算」の基礎学力が低下しているという指摘から出発しました。その後、PISAにおける「リテラシー」概念の影響を受けて、学力の発展性を示す「活用力」を重視するという展開となっていったのです。

2007年から実施されるようになった「全国学力・学習状況調査」は、全国の公立小学校のすべての6年生と、同じく公立中学校のすべての3年生を対象とする、きわめて大規模な「悉皆調査」です（私立学校については、調査を希望する学校が参加しています)。調査項目としては、国語と算数・数学の学力テストとともに、児童生徒の学習意欲、学習環境、生活習慣、さらには学校での指導方法や設備の整備状況も調査されています。とくに注目しておきたいのは、学力テストの構成が、A問題(「知識」に関する問題－基礎的・基本的な内容）とB問題（「活用」に関する問題－実生活を想定した場面で知識を活用して答えを出す内容）となっていることです。この間の学力問題が明らかに反映したものになっています。ちなみに、一回の「全国学力・学習状況調査」には、約60億円が費やされています。

文部科学省は、このような全国規模の調査を行うことによって、児童生徒や各学校の学力実態を把握するとともに、教育改善の資料を得ようとしています。しかしながら、国が実施するようになった、「悉皆調査」としての「全国学力・学習状況調査」は、日本のカリキュラム改革に多くの問題や課題を提起することになりました。

ここでは、まず「全国学力・学習状況調査」の内容について分析をします。次に、「全国学力・学習状況調査」を社会的な文脈から分析を加えたいと思います。

2.「全国学力・学習状況調査」の内容分析

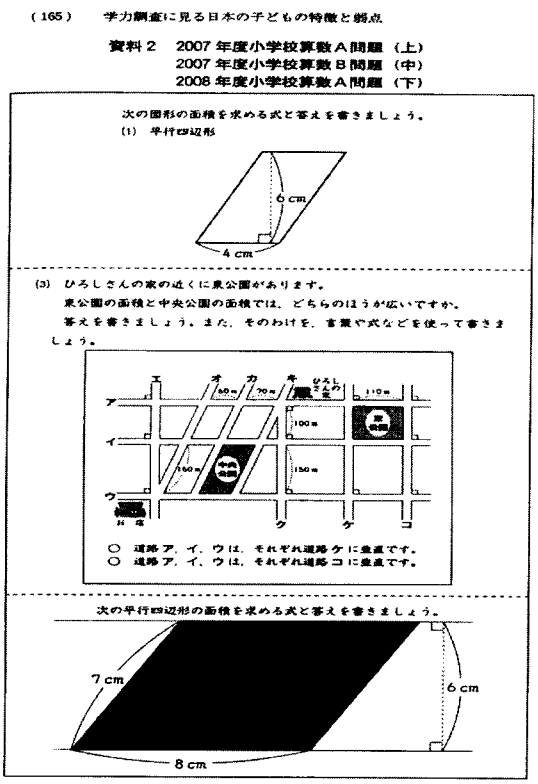
一般に、このような大規模な学力調査では、「信頼性」を確保するために、「標準テスト」や「客観テスト」を使用してきました。すなわち、「標準テスト」や「客観テスト」においては、いつどこで誰が実施しても、評価結果の一貫性が保証されるからである。しかし、「標準テスト」や「客観テスト」で測れる学力の質には明らかに限界があって、高次の学力を測ることができないと指摘されてきました。他方、大規模な学力調査において、「妥当性」を求めて、高次の学力を測る評価方法を採用しても、「信頼性」は確保できないと指摘されてきました。つまり、「信頼性」と「妥当性」の関係は、相対立する関係であると理解されてきたのです。

しかしながら、1980年代の後半から、アメリカで主張されるようになった「真正の評価」論と、その代表的な評価方法である「パフォーマンス評価」は、この「信頼性」と「妥当性」の関係に新たな光を投げかけることになりました。この「真正の評価」論と「パフォーマンス評価」は、PISAの問題作成に大きな影響を与えました。そして、PISAが日本における「全国学力・学習状況調査」のB問題に影響を与えたのです。

「真正の評価」論とは、評価を行う文脈が実際の生活を反映したものでなければならないという主張です。または、実際の生活で起こるリアルな課題に子どもたちを取り組ませることによって、子どもたちの学力を評価しようとする考え方です。「標準テスト」や「客観テスト」では、作為的な問題が出されることが多く、このような問題を解けたからといって、子どもたちには「生きる力」に結びつく学力は形成されないと考えるのです。

ところで、「真正の評価」論は、子どもたちの生活の文脈に即して問題作成を行うことから、子どもたちにとっては親しみやすい問題場面になります。しかし、問題場面が親しみやすいということは、その問題が簡単に解けるということではありません。むしろ、かなり高度な学力が要求されるのです。なぜならば、実際の生活は、問題を解くために必要な情報だけが存在しているわけではないからです。そこでは、不必要な情報が多くあったり、必要とする情報が不足している場合もあります。子どもたちには、必要とする情報を選び出して、どのような知識や法則を使って解けるのかを判断するという、高次な学力が要求されるのです。

日本において、2007年に実施された「全国学力・学習状況調査」で出題された小学校の算数の例をあげてみましょう。


（165）　学力調査に見る日本の子どもの特徴と弱点

資料２　2007年度小学校算数Ａ問題（上）
2007年度小学校算数Ｂ問題（中）
2008年度小学校算数Ａ問題（下）

次の図形の面積を求める式と答えを書きましょう。
(1)　平行四辺形

(3)　ひろしさんの家の近くに東公園があります。
東公園の面積と中央公園の面積では，どちらのほうが広いですか。
答えを書きましょう。また，そのわけを，言葉や式などを使って書きましょう。

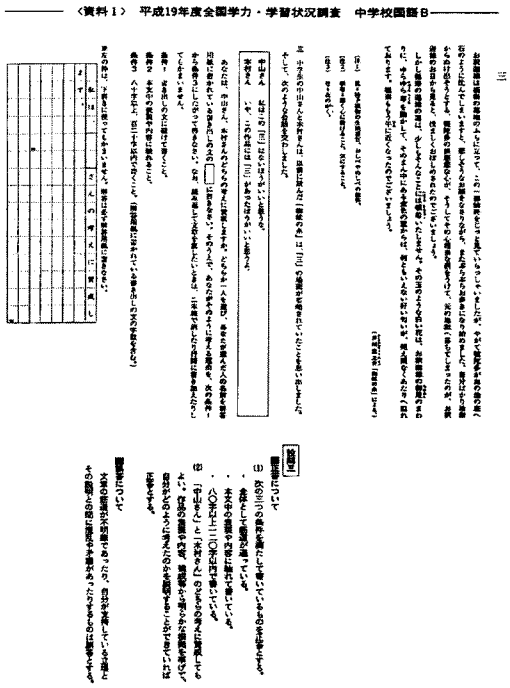
○　道路ア，イ，ウは，それぞれ道路ケに垂直です。
○　道路ア，イ，ウは，それぞれ道路コに垂直です。

次の平行四辺形の面積を求める式と答えを書きましょう。

— 21 —

ここでは「平行四辺形の面積を求める」という学力が、評価の対象となっています。A 問題では、「底辺」と「高さ」の長さが書かれた「平行四辺形」が示されて、面積を求めることになります。この問題は大変に平易で、正答率が 96%でした。しかし、B 問題では、さまざまな数値が書き込まれた「地図」が示されて、そこに位置する「平行四辺形」の公園の面積と「長方形」の公園の面積を比較して、大きい方の公園はどちらかを答え、その理由を書くという問題が出されました。その結果は、18.2%という大変に低い正答率でした。その大きな原因のひとつは、「地図」に「平行四辺形」の「斜辺」の長さが示されていて、34.4%の子どもたちが「底辺」と「斜辺」を掛けて面積を出していました。つまり、現実の世界は、本来は情報が過多であって、その図形を「平行四辺形」と判断してよいのか、「平行四辺形」であれば「底辺」と「高さ」はどれなのかといった、高次の判断力や思考力が連続して要求されるのです。この問題は、「真正の評価」論の分かり易い例と言えるでしょう。

ここで、もうひとつの例をあげてみます。この問題は、2007 年に実施された「全国学力・学習状況調査」で有名になった中学校の国語の B 問題です。この問題は、黒沢明監督の映画「羅生門」の原作者として有名な芥川龍之介の小品（「蜘蛛の糸」）を全文読ませて、その作品の最終場面の描写が必要なのか、必要でないのかを選択させて、その理由を書かせるものでした。


〈資料１〉　平成19年度全国学力・学習状況調査　中学校国語Ｂ

ご承知のように、PISA に類似の問題があり、それに影響されたものです。日本における従来の国語教育では、テキストの主題を解釈させたり、登場人物に共感させるような指導が行われていました。このような国語教育に対して、著名な芥川龍之介の作品構成を批判的に読ませるという、高度な学力を要求する問題は、日本の教師たちに大きな衝撃を与えました。たしかに、現実の世界ではテキストの真偽や質を吟味する、高次な学力が要求されるのです。

以上の二つの問題は、「パフォーマンス評価」の例でもあります。「パフォーマンス評価」とは、子どもたちが知識を実際の世界にどの程度うまく活用させて

いるのかを測るものです。その際、「客観テスト」がテスト用紙に書き込まれた既存の解答を選択させるという方式を採るのに対して、「パフォーマンス評価」は学び得たことをさまざまなメディアを使って表現させるという方式を採ります。そのメディアとは、文字による表現だけでなく、図やグラフや絵という表現もあり、さらには実際に演出するという表現もあります。このようなメディアを通じての「表現」によって、いわゆる「生きる力」としての学力を測ろうとするのです。「平行四辺形」の面積を求めるB 問題では、大きい面積であると「判断した理由」を書くことを求めており、芥川龍之介の作品構成を批判的に読ませる問題でも「判断した理由」を書くことが求められています。この「判断した理由」を子どもたちが書けるためには、算数・数学や国語の教科内容を深く理解しているとともに、その理解したことを伝えるためには「メタ認知能力」が形成されていることが必要でしょう。つまり、高次な学力が求められるのです。

このように、日本における「全国学力・学習状況調査」は、PISA の影響を受けて、「真正の評価」論や「パフォーマンス評価」を採用しました。しかしながら、ここから二つの課題が明らかになってきました。

そのひとつは、大規模な学力調査において、高次な学力を測るために、「妥当性」の高い「パフォーマンス評価」を採用した場合に、「信頼性」をどのように保証するのかという課題です。実のところ、日本における「全国学力・学習状況調査」では、この「信頼性」の問題が十分に検討されないままに、実施されたのではないかと思います。たとえば、芥川龍之介の作品構成を批判的に読ませる問題では、75.5%という高い正答率を示しました。先に指摘しましたように、従来の日本の国語教育では、批判的に読むという指導が行われてこなかったことから、この正答率は余りにも高いと判断できるでしょう。その原因は、この問題の採点基準が、「文章として筋が通っていれば良い」という程度の、かなり甘いものであったことにあるでしょう。このように、「パフォーマンス評価」を採用しても、「信頼性」が確保されない場合には、その意義は半減してしまいます。

この「信頼性」を保証するために、「真正の評価」論や「パフォーマンス評価」では、採点基準として、「ルーブリック」の設定を求めています。「ルーブリック」とは、「尺度」、評価基準を示す記述語と、それぞれの「尺度」には具体的で典型的なサンプルが付されていて、「信頼性」を確保しようとしています。芥川龍之介の作品構成を批判的に読ませる問題を例にすると、3 段階か4 段階の「尺度」を設定して、それぞれの「尺度」には、子どもたちが「判断した理由」を書いた文章の質的な特徴と、それを典型的に示している子どもの作文例が付されていることが必要です。このような「ルーブリック」が周到に準備されてこそ、「パフォーマンス評価」は「妥当性」とともに「信頼性」を確保することになるのです。

もうひとつの課題は、「パフォーマンス評価」は「全国学力・学習状況調査」のような「紙と鉛筆」で回答を求めるテスト形式に限定されるのではなく、学び得たことをさまざまなメディアを使って表現させることに特徴があることを確認することです。たとえば、「あなたは博物館の学芸員になって、ある中学校からやって来た見学者たちに、日本の江戸時代の人々の暮らしを、ビジュアルなパネルを作成して、わかりやすく説明して下さい」という「パフォーマンス課題」が提起されます。この場合には、「学芸員」になった子どもたちが、どのようなパネルを作成するのか、どのようにプレゼンテーションするのかが、評価の対象となります。「パフォーマンス評価」は、子どもたちの生き生きとした学習活動を総体として捉えようとする、きわめてダイナミックな評価方法なのです。

このことを強調するのは、PISA や「全国学力・学習状況調査」で出題された問題が解けさえれば、高次な学力の目標である、「活用力」としての「リテラシー」が完全に形成されたとは言えないということです。子どもたちが高次の学力を本当に身につけたかどうかを捉えるためには、「紙と鉛筆」で回答を求めるテスト形式だけでなく、教育実践と統合されたダイナミックな「パフォーマンス評価」を実施する必要があるのです。

3.「全国学力・学習状況調査」の社会的文脈

さて、「全国学力・学習状況調査」の実施が公表される中で、その実施に対して、厳しい批判が行われるようになりました。とくに、「全国学力・学習状況調査」が「悉皆調査」である点に批判が集中しました。従来、

文部科学省は「教育課程実施状況調査」という「抽出調査」を行っていました。それがなぜ今になって「悉皆調査」を実施する必要があるのか。文部科学省は、「悉皆調査」を行うことによって、国家の統制力を強化しようとしているのではないか。文部科学省は、「悉皆調査」の結果を公表して、各学校を競争させることによって、学力を向上させようとしているのではないか。そのようになれば、各学校は点取り競争に走って、教育現場が混乱するのではないか。このような批判がなされたのです。

このような批判が行われた大きな理由として、日本では、次のような苦い歴史的な経験があるからです。文部科学省（当時は文部省と呼称）は、1960年代の初めに、中学生を対象とした学力に関する「悉皆調査」を実施したことがあります。その結果、学力調査における「日本一」をめざして、教育現場は過激な競争を行いました。その中で、事前にテスト問題を生徒に漏らしたり、成績の悪い生徒をテスト当日に休むように強制する事件が発生しました。このため、この「悉皆調査」に対して強い反対運動が起こり、「悉皆調査」は中止に追い込まれました。

このような苦い歴史的な経験を踏まえて、文部科学省は、今回の「全国学力・学習状況調査」について、競争することが目的ではなく、学校教育を改善することが目的であるという点を強調しました。そのために、市町村単位や学校別の成績は公表していません。しかしながら、都道府県単位の成績を公表したことから、成績が悪かった都道府県では、市町村単位の成績を公表して競争を煽ったり、テストの点数を上げるための即席の方法を奨励するなどの問題が発生しています。あらためて、「悉皆調査」の問題点が問われるとともに、その「悉皆調査」で採用された「パフォーマンス評価」に対しても、考えなくてはならない課題を提起しました。

確かに、「パフォーマンス評価」は、「標準テスト」や「客観テスト」に対して、質の高い学力を測ることができるという点で、優れた評価方法です。しかし、国家が実施する「悉皆調査」に「パフォーマンス評価」が採用されると、そして調査結果がハイ・ステイクスな性格を強めるようになると、教育現場では勢い「パフォーマンス評価」というテスト形式への対処療法（いわゆる試験対策）が優先されることになります。その結果として、高次の学力を形成するための地道な「授業研究」が軽視されます。また、教育現場では「紙と鉛筆」という形式に縛られた「パフォーマンス評価」しか視界に入らなくなります。このような「パフォーマンス評価」が置かれた状況では、「生きる力」に結びつく「リテラシー」は形成されないでしょう。

ここで、「真正の評価」論が提起する「stakeholder アプローチ」に注目しておきたいと思います。このアプローチは、評価の結果に利害を有する「stakeholder 」たちは、評価行為に「参加」する権利があるという主張です。最近では、このようなアプローチは「参加型評価」というネーミングで知られるようになっています。

従来の評価場面では、「評価する人」と「評価される人」は不平等な関係であって、「評価される人」は「評価する人」の規準や評価方法に従属しなくてはなりませんでした。ましてや、国家が主催する学力調査となると、この不平等な関係は明白でした。この不平等な関係を組み換えて、評価場面に民主主義を貫こうとしているのが、「stakeholder アプローチ」なのです。

このアプローチを支えているのが、「モデレーション」という方法論です。「モデレーション」とは、評価場面でのさまざまな意見の相違や対立を調整して、合意形成を促そうとする行為です。「モデレーション」は、まずはスタンダードとしてどのような教育目標を設定するのかについて、教育研究者や教師や保護者たちの意見が反映すべきことを求めます。そして、どのような評価方法を採用するのかについても、「モデレーション」が必要です。さらには、「パフォーマンス評価」の「信頼性」を確保するために必要な「ルーブリック」を設定するためには、子どもたちの「パフォーマンス」をどのように段階づけて「尺度」にするのかについて、教師間の「モデレーション」が実施されます。

また、「真正の評価」論のもうひとつの有名な評価方法である「ポートフォリオ評価」においては、教師と子どもたちによる「カンファレンス」が重視されます。この「カンファレンス」こそ、「ポートフォリオ」を素材にして、教師と子どもたちが「モデレーション」を行っている場面なのです。

このように評価場面で「モデレーション」が行われ

るようになると、評価行為が固定化したり、画一化したり、形骸化することを防ぐとともに、真に教育実践を改善するための評価行為のあり方を追求することができます。また、「モデレーション」は、それに参加した人々の専門性（つまり「評価リテラシー」）を高めることにもなります。「全国学力・学習状況調査」が実施されたことによって、「stakeholder アプローチ」とそれを支える「モデレーション」の役割がますます強く自覚されています。

以上、日本で実施されるようになった「全国学力・学習状況調査」が、教育評価研究に

投げかけた課題を考察してきました。そこでは、高次の学力を測る「パフォーマンス評価」の意義とともに、それが本来の役割を果すための条件としての「モデレーション」の重要性を指摘しました。この日本での経験が、アジアの教育改革に少しでも参考になれば幸いです。ご静聴ありがとうございました。

参考文献

Gipps, C. V.(1994). *Beyond testing: towards a theory of educational assessment.* London: Falmer Press.

Torrance,H (ed).(1995). *Evaluating Authentic Assessment: Problems and Possibilities in New Approaches to Assessment.* Buckingham: Open University Press.

Janesick, V.(2006). *Authentic Assessment.* New York, NY: Peter Lang.

田中耕治. (2008). 教育評価. 岩波書店.

田中耕治(2008). 新しい学力テストを読み解く. 日本標準.

Wiggins, G. (1998).*Educative Assessment: Designing Assessments to Inform and Improve Student Performance.* San Francisco: Jossey-Bass Publishers.

Wiggins G. and McTighe, J.(2005). *Understanding by Design. (2nd Edition).* Alexandria, VA: ASCD.

（本稿は、2009年10月29-30日に開催された The 10th International Conference on Educational Research by Education Research Institute of Seoul National University において発表したものである）。

（教授）