教育心理学研究, 2006, 54, 265—279

〔展　望〕

テストへの適応

——教育実践上の問題点と解決のための視点——

村　山　　航*

学習者は，テストを受ける中で，“テスト作成者はこういったことを評価したいのだ”とその評価基準・意図を推察し，それにあわせて自らの学習行動を変化させることがある。本稿では，そのような現象を“テストへの適応”と呼び，関連領域を包括的に概観しながら，1）実証的にどのような形で支持されているのか，2）どのような教育実践上の問題点を持っているのか，3）その問題を解決するための視点として何が考えられるか，の3点に関して検討を行った。実証的な支持に関しては，テスト期待効果研究と学習方略研究を取り上げ，それらを統合的に捉える仮説を提出した。問題点としては“学習行動の危機”と“妥当性の危機”という2点を指摘した。最後にこれらの問題を解決するために，テストワイズネス・テストスキルの個人差の排除，新しい評価（alternative assessment）の導入，インフォームドアセスメント（informed assessment），妥当性概念の拡張，表面的妥当性への意識，という5つの視点を提出した。

キーワード：テストへの適応，テスト期待効果，インフォームドアセスメント，妥当性，新しい評価

問題と目的

“テスト”とは，学習者の学力を診断・把握するためのツールである。そのため，テストには，客観的で正確に学習者の能力を測定することが求められる。従って，“客観的で正確な測定をするテスト”が“よいテスト”であるといえるだろう。しかし，テストは，学力を測定するためのものさしとしてのみ存在するわけではない。テストは，教師と学習者との関係の中で，社会システムの一つとして実施される。このような社会的文脈の中で，テストは単なる“ものさし”を越えて，さまざまな意味に価値づけられ，多くの学習者に影響を与えると考えられる。テストが社会的文脈の中に置かれたとき，よいテストのあり方とは，テストの測定の客観性・正確さだけに依拠するのではない。“テストがどのような影響を学習者に与えるのか”ということまでも含めて考えるべきなのである（Frederiksen & Collins, 1989；Messick, 1989；村山, 印刷中）。

テストが学習者に与える影響に関しては，さまざまなものがある（包括的なレビューとしてCrooks, 1988）。1つは，動機づけに対する影響である。テストを事前に予告したり，そのフィードバックを返したりすることは，学習者の動機づけに影響を与えると考えられる。テストの予告やフィードバックの返し方の違いによって，学習者の動機づけがどのように変化するかに関しては，多くの実証研究がある（レビューとしてHarkins, 2001；鹿毛, 1996）。2つは，テストによる学習の定着（consolidation）効果である。テストを受けること自体が，その学習内容の再確認・再活性化に繋がり，学習の定着を促進すると考えられる。このことは，テスト効果（test effect）と呼ばれ，Jones（1923-24）による古典的な研究を先駆けとして，研究が行われてきた（e.g., Halpin & Halpin, 1982；Nungester & Duchastel, 1982）。3つめとして，学習者の自己改善が挙げられる。テストによる診断結果をもとに，学習者が自分の学習を改善することは，自己制御学習（self-regulated learning；Zimmerman, 2001）の重要な要素の1つであり（Butler & Winne, 1995），テストの正統的な効用だと考えることができる（Bloom, 1981；Bloom, Hastings, & Madaus, 1971）。

このように，テストは学習者にさまざまな形で影響を与える。そして，教師自身もテストのこのような影響を念頭に置き，利用することが多い（Mullis, Martin, Gonzalez, & Chrostowski, 2004）。しかし，学習者はテストからの影響をただ教師の思惑通りに受けるわけではない。テストを繰り返し受ける中で，学習者はテストと戦略的に（テスト開発者の意図を積極的に読み取りながら）相

* 日本学術振興会・東京工業大学大学院社会理工学研究科
murakou@orion.ocn.ne.jp

互交渉し，テストに“適応”を起こすことが考えられる。すなわち，学習者は，テストを受ける中で，“テスト作成者はこういったことを評価したいのだ”とその評価基準・意図を推察し，それにあわせて自らの学習行動を変化（変容）させることがあると思われる。例えば，数学のテストで計算問題ばかりが出題されると，学習者は計算問題の練習を中心に行い，文章題などに関する勉強はあまりしないようになるだろう。テストが学習者に与える影響を考えるとき，こういった学習者の戦略的な“テストへの適応”も考慮に入れる必要があるだろう。

学習者がテストに適応して自らの学習行動を変容させることは，古くは百年以上も前から指摘があった（Latham, 1877）。その後も，認知心理学の実験室実験から，教育政策に関する事例研究まで，さまざまなレベルの研究で示唆されてきた。また，関係が深いと思われるテーマも数多く存在する。しかし，これらを“テストへの適応”という観点で，包括的にまとめた研究は見当たらない。本稿の目的は，“テストへの適応”という筆者独自の観点から，関係する知見を幅広く概観し，“テストへの適応”が1）実証的にどのような形で支持されているのか，2）どのような教育実践上の問題点を持っているのか，3）その問題を解決するための視点として何が考えられるか，について議論を行うことである。以下では，この3つの目的に沿った形で，3つの節を構成し，議論を進めていくこととする。この議論の中で，“テストへの適応”という枠組の妥当性も明らかになっていくだろう。

なお，本稿では“テスト”を，学習者の学力を測定するために，何らかの課題や項目を与えて，その反応を調べる自記式の検査ツールと定義する[1]。近年の教育評価研究では，テストという概念自体の限界が指摘され，テストを超えた，パフォーマンス評価やポートフォリオ評価といった評価の形態・モデルが模索されつつある（Gipps, 1994；鹿毛, 2004；National Research Council, 2001；田中, 2002a, 2002b）。本稿では，実証的研究の蓄積が多いテストに主として焦点を当てるが，この近年台頭してきた評価形態を“新しい評価”（alternative assessment）と総称し，後半部分で簡単な検討を加えることとする。また，ただ単に“評価”と表記したときには，テストや新しい評価に限定しない，学習評価一般のことを指すこととする。

テストへの適応に関する実証的証拠

テストへの適応を実証的に検討していると考えられる研究として，テスト期待効果（test-expectancy effect）研究と，ヨーロッパ・オーストラリアの学習方略研究が挙げられる。しかし，この両者の結果にはやや不一致が見受けられる。本節では，それぞれの研究を概観してその不整合を指摘した上で，この両者を繋ぐために，“学習方略による媒介”という新しい仮説を提出したい。

テスト期待効果研究

学習者のテストへの適応を，最も直接的な形で検討しているのが，テスト期待効果の研究である。テスト期待効果とは，どのような形式のテストを予期するかによって，学習者のテストに対するパフォーマンスが変わってくる効果のことを指す（Lundeberg & Fox, 1991）。テスト期待効果研究では，テスト形式の予告を行ったり同じ形式のテストを繰り返し実施したりすることによって，学習者に課題後に実施されるテスト形式を予期させる。そして，課題終了後，予期と合致した形式のテストと，予期と合致しなかった形式のテストのパフォーマンスを比較する。実験は実験室での記憶実験の場合もあれば（e.g., Neely & Balota, 1981），実際の授業を用いた授業実験の場合もある（e.g., Gay, 1980）。多くの場合，テスト形式として，再認テストと再生テストが使用される。

ここで，もしテストへの適応が実際に生じているのならば，再生テストを予期した群は，再認テストを予期した群に比べ，再生テストの記憶成績（パフォーマンス）が高く，再認テストの記憶成績が低いことが予測される。しかし，この予測は実証的にはそれほど支持されていない（Marton & Säljö, 1976b）。例えば，再認テストに関して，予測どおりの結果が得られた研究もあれば（Sax & Collet, 1968；Tversky, 1973），予測と正反対（再生テスト予期群の方が高い記憶成績を示す）の結果を報告している研究も存在する（Meyer, 1934；Neely & Balota, 1981；Thiede, 1996；Breen & Wilding, 1984）。両者の間に違いを見出せなかった研究も多い（Gay, 1980；May &

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999)ではテストを，標準化された方法を用いて受験者のある領域における行動を収集・評価・得点化するための装置・手続きとしている。このように広義に定義した場合，テストには学力以外の能力・特性を調べるための検査（心理検査など）も含まれる。また，次に記す“新しい評価”も，部分的にはテストに含まれることになる。しかし，本稿ではテストをより一般的な意味で捉えるため，本文で述べたような定義で議論を進めることとする。ただし心理検査などに対する示唆も，“終わりに”の節で少し議論することとする。

Thompson, 1989 ; Oakhill & Davies, 1991)。同じように，再生テストに関しても，予測を支持する知見が多い一方で (d'Ydewalle, Swerts, & Corte, 1983 ; Meyer, 1934 ; Sanjivamurthy & Kumar, 1983 ; Schmidt, 1983)，そのような効果を見出していない研究もある (McDaniel, Blischak, & Challis, 1994 ; Rickards & Friedman, 1978)。

さらに，テスト期待効果は，いくつかの変数によって，その結果が変わることも明らかになっている。例えば，Oakhill & Davies (1989) では，実験を実施する時間が午前の場合には，再生テスト・再認テストともに再認テストを予期した群の記憶成績がよいが，時間が午後になると，再生テストを予期した群の記憶成績が全般的に高くなることを見出した。同じように，記銘単語の親近性 (Balota & Neely, 1980)・イメージ性 (Wnek & Read, 1980)・記銘リストのカテゴリ化の程度 (Connor, 1977) など，記銘材料の性質がテスト期待効果に影響を与えていることが示されている。また，ノート取りの指導 (Carrier & Titus, 1981)・記銘方略の指導 (清水, 1990) など，指導内容の違いも，テスト期待効果を調整することが示されている。Lundeberg & Fox (1991) のメタ分析では，実験状況の違い (授業場面か実験室場面か) が，テスト期待効果に影響を与えていることが明らかになっている。

以上のように，これまでのテスト期待効果研究を総括する限り，その結果は非常に不安定であるといえよう。テスト期待効果研究においては，テストへの適応という現象が，それほど安定していないように思われる。

ヨーロッパ・オーストラリアの学習方略研究

学習者のテストへの適応を，別の形で検討しているのが，ヨーロッパ・オーストラリアを中心とした，大学生を対象にした学習方略研究である。この研究の嚆矢となった Marton & Säljö (1976a) は，大学生の文章読解方略をインタビューで検討し，深い処理(deep-level processing) の方略と浅い処理 (surface-level processing) の方略という質的に違う２種類の方略があることを見出した。そして，この２つの方略使用を規定する要因について，特性と状況の相互作用という観点から，多くの研究が実施された (レビューとして Ramsden, 1988)。ここで，特に学習方略の状況的な規定因として注目されたのがテストであった (Laurillard, 1979)。

このとき，もしテストへの適応が実際に生じているのならば，評価方法に応じて，学習者はそれに適した学習方略を使うようになることが予測される (cf. 村山, 2003a)。この予測は，これまでの先行研究において，概ね支持されているように思われる。例えば Scouller (1998) は，学期中に実施された多肢選択式テストと記述式のアサインメントを学習者に想起させ，それらに対してどのような学習方略を使用したかを尋ねた。その結果，学習者は，記述式のアサインメントに比べ，多肢選択式テストに対して，浅い処理の方略を多く使い，深い処理の方略をあまり使わなかったことを報告した。このような研究に代表されるように，学習方略研究では，多肢選択式テストや単答式テストに対して，学習者が浅い処理の方略を多く使用し，深い処理の方略をあまり使用しないことが明らかになっている (Marton & Säljö, 1976b ; Newble & Jaeger, 1983 ; Scouller & Prosser, 1994 ; Thomas & Bain, 1984)。

これら一連の研究には方法論上の欠陥が多いため，結果の妥当性に疑問を投げかけることもできる。例えば Scouller (1998) や，Scouller & Prosser (1994)，Newble & Jaeger (1983) では，テストを実施してからかなり経った状態で回顧報告を求めているため，学習者のテストに対する知識 (テスト形式スキーマ; 村山, 2005, 2006) によるバイアスを受けている可能性がある。Thomas (1982) や Thomas & Bain (1984) では，同じ被験者に異なる形式のテストを実施したうえで，その方略使用を調査しているため，被験者が過度にテスト間の違いを対比させて結果を報告した可能性がある。しかし，後の村山 (2003b, 2004) では実験的統制を厳密にした場面で，テストの違いが学習方略使用に影響を与えることを見出しており，学習者がテストに適応して方略を変容させることは，確実な知見だと考えられる。

テスト期待効果研究と学習方略研究の統合―学習方略による媒介仮説―

これまで見てきたように，テストへの適応に関して，学習方略研究では支持する知見が多い一方，テスト期待効果研究ではそれほど明確な傾向が得られていない。この不一致を統合的に捉えることはできないだろうか。

ここで有力な仮説として，“学習方略による媒介”を挙げることができる。この仮説では，テストの違いは，学習方略には直接の影響を与えるのに対し，記憶成績 (パフォーマンス) には学習方略を媒介した，間接の効果しか与えないと考える。そして，テスト期待効果研究で結果が安定していないのは，記憶成績という，テストへの適応を間接的にしか反映していない測度を用いているからであり，学習方略研究で結果が安定しているのは，学習方略という，テストへの適応を直接反映する測度を用いているからだと推察する。確かに，テ

スト形式の予期の違いが直接記憶成績に影響を与えるとは考えにくく，媒介するプロセスとして学習方略の変容を仮定するのは妥当のように思われる。

Murayama (2005) は，テスト期待効果の実験パラダイムで，学習方略と記憶成績の両方を測定し，この仮説を直接検討した。具体的には，テストの予期の違いが，異なった学習方略使用を媒介して，課題成績に与える媒介モデルを検討した。交互作用が交互作用を媒介するモデルになるため，潜在曲線モデルを応用したモデルを用いた (cf. DeSteno, Petty, Wegener, & Rucker, 2000)。結果が FIGURE 1 である。テスト形式の予期の違いが方略使用に影響し，方略使用の違いがパフォーマンスに影響を与えていることが示された。一方，テスト形式の予期がパフォーマンスに直接の影響を与えているという証拠は見出されなかった ($\chi^2(1)=0.51, ns$)。この結果は学習方略による媒介仮説を支持するものである。

テスト期待効果研究と学習方略研究には，研究パラダイムの違いも大きく，上記の研究だけで仮説が検証されたとは言い難い。だがいずれにせよ，学習方略による媒介仮説の重要な点は，これまで別個に行われてきた“テスト期待効果研究”と“学習方略研究”を，統合的に把握しようとした点である。学習者の“テストへの適応”を捉えるためにも，今後こういった統合的な視点をもとにした検討が大切になるだろう。

テストへの適応が生み出す教育実践上の問題点

前節ではテストへの適応に関する実証的証左として，テストにあわせた学習方略の変化を指摘した。本節では，この学習方略の変化というメカニズムを足掛かりに，テストへの適応が生み出す教育実践上の問題点を検討する。その検討を通して，テストへの適応という現象に対するより詳細な理解を深めたい。具体的には，“学習行動の危機”と“妥当性の危機”の2つを取り上げて，検討を行うこととする。

学習行動の危機

高次の能力を反映していないようなテストを実施したとき，学習者がそのテストに適応してしまい，深い処理の学習方略などを阻害してしまうことがある (村山, 2003a, 2003b, 2004)。すなわちテストへの適応は，学習方略といった学習行動をネガティブな方向に変化させてしまう可能性を持っている。このことは，これまで本稿で見てきた実証研究でも示唆されているが，実際の教育・評価制度に焦点を当てた研究でも明らかになっている。

例えば，1970-1980年代のアメリカで，テストへの適応による負の影響が1つの社会問題となった。当時のアメリカでは，教育のアカウンタビリティの議論が高まり，高等学校などで，標準化テストや最低基礎能力テスト (minimum-competency test) を実施する州が多かった (Jaeger, 1989 ; Madaus, 1983)。このテストの多くは多肢選択式の問題より構成されていた。Madaus (1988) は，このような標準化テストが教師の教授内容や学習者に与える影響に関する調査のレビューを行い，テストがハイ・ステイクス (high-stakes) であるほど，学習者の学習内容を狭めてしまったり，そのテストに特化した低次の認知的スキルを増大させてしまったりすることを主張した。ハイ・ステイクスなテストとは，その結果が人の将来や意思決定に，大きな影響を及ぼすようなテストのことである。Cannell (1988) は，アメリカのどの州においても，標準化テストの平均点が，アメリカ全州の平均点 (そのテストが標準化されたときの平均

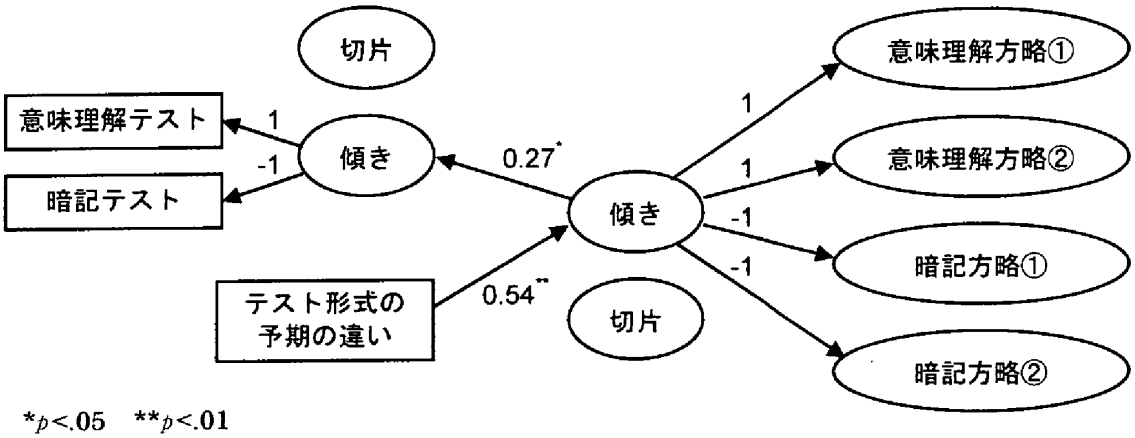

FIGURE 1 テスト形式の予期の違いが，方略を媒介してパフォーマンスに与える影響

注) Murayama (2005) より。方略を構成する観測変数，切片に関するパス，誤差変数は省略してある。

点)より高いというパラドキシカルな現象を見出し，その背後にはテストのための教授・学習（テストへの適応）があると推察した。また，Frederiksen (1984)，Haladyna, Nolen, & Haas (1991) や Mehrens & Kaminski (1989)，Smith(1991)でもこのような標準化テストの悪影響に関して，多くの観点から議論がされている。

日本国内でもテストへの適応が学習者に悪影響を与えることが指摘されている。古くは学制が頒布されて間もない明治20年(1887年)，福島私立教育会雑誌十七号には“生徒平常の受業皆試験の為に備るが如く一々記憶に収め置きて所謂復生想像力のみを養ひて構成想像力を使用せず受動的の能力に富みて発動的の能力に乏しきに至る且之を施すか為往々疾病を醸ことあり”といった，テストが低次の学習を増大させてしまうことの悪弊が記されている（福島県教育委員会, 1972)。近年では藤澤(2002)が，テストにあわせて自らの学習内容・学習方法を矮小化させてしまうことを“ごまかし勉強”と呼び，その問題点を指摘している。

以上のように，テストにあわせた学習行動の変化は，低次の能力でも解決できるようなテストを実施した場合に，ネガティブな結果を生むと考えられる。テストへの適応は，学習行動の危機を生じさせる可能性を内包しているのである。

妥当性の危機

こういった学習行動の危機は，それ自体が問題であるだけでなく，テストの妥当性にも負の影響を与えると考えられる。すなわち，学習者にあわせて学習行動を変化させることにより，そのテストで測定したいもの以外の要素が系統的に混入し，テストの妥当性が低下する可能性が考えられる。例えば，テストへの適応によって学習内容を限定することは，“ヤマの張り方の上手さ”といった要素をテスト得点に混入させることになる (e.g., Broekkamp, Van den Bergh, Van Hout-Wolters, & Rijlaarsdam, 2002 ; Broekkamp, Van Hout-Wolters, Van den Bergh, & Rijlaarsdam, 2004)。また，テストにあわせて学習行動を暗記などの低次の方略に限定することは，テスト作成者が本当に測定したい学習プロセスを遮蔽することになる。

テストへの適応において，テストの妥当性が危機に陥る要因はそれだけではない。ここではさらに，テストワイズネスとテストスキルという要因を指摘したい。

テストワイズネス　テストワイズネス(test-wiseness)とは，テスト特有の手掛かりを利用して，テストの点を高めることができる能力のことである (Millman, Bishop, & Ebel, 1965 ; Sarnacki, 1979 ; Towns & Robinson, 1993)。特に狭義のテストワイズネスは，正答するための知識がまったくなくても，正答の確率を上げることができる能力を指す。例えば，多肢選択式の問題で，選択肢の長い文章を選ぼうとすることや，極端な副詞が混じった選択肢を避けようとすることは，一種のテストワイズネスである。テストワイズネスは，テストを受ける経験の中で獲得されるものだということが明らかになっている (Crehan, Koehler, & Slakter, 1974 ; Evans, 1984 ; Sarnacki, 1979 ; Wahlstrom & Boersma, 1968)。すなわち，テストへの適応とテストワイズネス能力には密接な関係があると考えられる。

もしテストワイズネス能力に個人差があれば，同じような能力を持った学習者も，テストで異なった点数を取ってしまうことになる (e.g., Dolly & Williams, 1986 ; Evans, 1984)。従って，テストワイズネス能力の個人差によって生じた分散が，テストの妥当性を変化させていると考えることができる (Diamond & Evans, 1972 ; ただし McMorris, Brown, Snyder, & Pruzek, 1972)。すなわち，テストへの適応は，テストワイズネス能力を介して，テストの妥当性に影響を与えていると言うことができよう。

テストスキル　テストスキル (test-taking skill) とは，テストで学習者が自分の能力を十分に発揮できるようにするためのスキルであり (Scruggs, White, & Bennion, 1986)，テストにおける設問形式への慣れ (Vernon, 1962) や効率的な時間配分のスキル (Scruggs et al., 1986) などが挙げられる。このテストスキルも，テストを繰り返し受ける中で育成されるものであり，テストへの適応との関係が強いと思われる。

やはりもしこのテストスキルに個人差があれば，同じような能力を持った学習者も，テストで異なった点数を取ってしまうことになる (e.g., Bangert-Drowns, Kulik, & Kulik, 1983 ; Deaton, Halpin, & Alford, 1987 ; Samson, 1985)。従って，テストスキルの有無によって生じた分散が，テストの妥当性を変化させていると考えることができる (Vernon, 1956)。すなわち，テストへの適応は，テストスキルも介して，テストの妥当性に影響を与えているといえよう (Bond, 1989)。

まとめ　以上見てきたように，テストへの適応は，いくつもの要因を媒介して，テストの妥当性に影響を与えていると考えることができる。ただし本稿ではこれらの要因を分けて記述したが，厳密には区別するのが難しい (cf. Millman et al., 1965)。いずれにせよ，テストへの適応は学習行動の危機を生じさせるだけでなく，テストの妥当性にも脅威を与えるのである。

問題解決のための視点

これまで，テストへの適応に関して“学習行動の危機”と“妥当性の危機”という2つの問題点があることを指摘してきた。本節では，これらの問題を解決するための5つの視点を提出し，その意義や限界などについて議論する。これらの視点は相互に独立したものではなく，相互に関連しあっているものである。

テストワイズネス・テストスキルの個人差の排除

先述したように，テストワイズネス・テストスキルの個人差はテストの妥当性に影響を与えていると考えられる。テストで妥当性の高い測定をするためには，こういった個人差の影響を排除してやる必要がある。

テストワイズネスにおける個人差の影響を減らす方法は2つ考えられる。1つは，すべての学習者にテストワイズネスを教示することである。これまでの研究で，テストワイズネスの能力を教示によって伸ばせることが示されている（e.g., Scruggs et al., 1986 ; Slakter, Koehler, & Hampton, 1970 ; Wahlstrom & Boersma, 1968）。しかし一方で，テストワイズネス能力は領域固有的で複数の異質なスキルの集合体だという見解も強い（Diamond & Evans, 1972 ; Evans, 1984 ; Rogers & Bateson, 1991）。Dolly & Williams (1986) では，テストワイズネスのトレーニングによって，教えたテストワイズネス能力は伸びたが，他の問題への転移は起こらなかったことを見出している。従って，すべての学習者にテストワイズネスの教示を行うことが，どの程度有効かは疑わしい点がある。2つは，テストワイズネスへの脆弱性がないテストを開発することである。テストワイズネスにどのような種類のものがあるかについて，これまでの研究には多くの蓄積がある（e.g., Millman et al., 1965 ; Towns & Robinson, 1993）。教育場面では，授業時間には制限がある。そのことを考えると，限りある授業時間を使用してテストワイズネスを教示するよりは，テスト作成者が注意してテストを作成しようとする後者のアプローチの方が，より有効であろう。

テストスキルの個人差は，テストの改善によってある程度克服できるが，学習者に固有の問題もある（テスト形式への慣れなど）。従って，明確な教授によって，テストスキルの個人差を減らす必要がある。テストスキルトレーニングの有効性は多くの研究で示されている（Bangert-Drowns et al., 1983 ; Bond, 1989 ; Deaton et al., 1987 ; Samson, 1985 ; ただし Scruggs et al., 1986）。また，そのようなトレーニングがテストの妥当性を向上させることも示されている（Powers, 1985）。

以上のように，テストへの適応によって生じる妥当性の問題に対処するためには，テストワイズネス・テストスキルといった要因を意識し，その個人差ができるだけ生じないようにテストの作成・実施をする必要がある。しかし，このようなことを意識したとしても，テストへの適応に関する問題がすべて解決されるわけではない。例えば，過剰なテストスキル能力は，学習内容や行動の限定に繋がり，学習行動の危機を生じさせる可能性がある（Mehrens & Kaminski, 1989）。このことは，テスト作成者が測定したい認知プロセスを測定できないという意味で妥当性の危機を再燃させてしまう。テストへの適応によって生じる問題を解消するためには，ここで述べた視点だけでなく，次に述べる“新しい評価の導入”といった視点なども取り入れていく必要があるだろう。

新しい評価の導入

テストへの適応によって，学習行動が低次のものに限定されてしまうことの1つの原因として，そのテスト自体が，低次のスキルによって解答できてしまうからだということが挙げられる。裏を返せば，もしテストが適切なものにさえ変われば，テストへの適応は逆に学習者に高次の認知スキルを促すことになるだろう。

この考え方は新しいものではない。到達度評価の主導者である Popham は，1980年代に測定主導の学習指導（Measurement Driven Instruction, MDI）という考え方を提出した。これは，適切な到達度評価が用いられれば，教師の指導も学習者の学習も改善されるはずだという主張である（Airasian, 1988 ; Popham, 1987 ; Popham, Cruse, Rankin, Sandifer, & Williams, 1985）。また，Wiggins (1989) は“*A true test*”と題された論文の中で，オーセンティック評価（authentic assessment）と呼ばれる評価概念を提出した。オーセンティック評価とは，教育評価が，大人が現実世界で直面するような問題解決場面で（オーセンティックな文脈で）なされるべきだという主張である（遠藤, 2003）。そのような評価のもとでは，評価への適応が必然的に日常的な問題解決スキルや高次の思考能力を促進することになる。また，評価への適応では，学習者が学習内容を限定してしまうことも問題になっていたが，このことも“日常的な問題解決のなかで何が重要かを学習者が取捨選択する”という肯定的な意味を持つこととなる。

Wiggins の主張は従来のテスト概念自体に疑問を投げかけるものであり，教育評価の考え方に大きなインパクトを持った。それ以降，パフォーマンス評価（performance assessment）[2]やポートフォリオ評価（port-

folio assessment) といった新しい評価の具体的な実践に繋がった。新しい評価に関しては，理論・実践的検討 (e.g., Wiggins, 1998) や，信頼性・妥当性などに関する計量心理学的検討 (Ruiz-Primo & Shavelson, 1996 ; Yen & Ferrara, 1997)など，多くの研究が1990年代以降なされている。日本においても，古くは橋本 (1981, 1983) の到達度評価に関する著書をはじめに，ポートフォリオ評価 (西岡, 2003 ; 田中・西岡, 1999)，パフォーマンス評価 (お茶の水女子大学21世紀COEプログラム, 2004) など，いくつもの理論的・実践的な研究がなされている (総括的なものとして田中, 2002a, 2002b)。

それでは，こういった新しい評価の導入は，学習者にポジティブな影響をもたらしうるのだろうか。何らかのポジティブな影響を持つことを期待して導入したテストや新しい評価法が，教師や学習者に与える影響のことを総称して波及効果 (washback effect, backwash effect ; Alderson & Wall, 1993) と呼ぶ[3]。近年の波及効果の研究をまとめたCheng, Watanabe, & Curtis (2004) は，波及効果の影響過程は複雑であり，必ずしも意図した通りの結果が得られるとは限らないということを主張している。実際，新しい語学テストの導入に関する波及効果研究では，テストによって教師や学習者が影響を受けた事例は散見されるが，その影響も限定的であることが見出されている (Alderson & Hamp-Lyons, 1996 ; Wall & Alderson, 1993; Watanabe, 1996)。短期的な効果は見られたが，長期的になると効果にばらつきが生じるという知見もある (Shohamy, Donitsa-Schmidt, & Ferman, 1996)。

パフォーマンス評価導入の効果を見た研究でも，そのような評価法の導入が実際に学習者の問題解決力を高めたり，高次の認知能力を高めたりしているかに関しては，まだ十分な証拠が得られていない。例えば，Firestone, Mayrowetz, & Fairman(1998)は，パフォーマンス評価を導入したアメリカのある州の事例を調査した。その結果，パフォーマンス評価の考え方と旧来のテストの考え方との間に大きなギャップがあるため，教師がパフォーマンス評価の理念を十分に理解できず，教師や学習者に与える影響が非常に限定的になっていることを見出した (同様の報告としてGuskey, 1994)。Torrance (1993) は，イギリスの標準評価課題 (standard assessment task, SATs) が教師や学習者にどのような影響を与えているかに関して調査を行った。標準評価課題にはパフォーマンス評価が含まれている。調査の結果，標準評価課題の導入は，教師にかなりの負担を与えるため，なかなか思った効果を得ることができないことが報告された。また，慣れない評価であるために，教師が結局はマニュアル・パッケージに頼ってしまい，評価が形骸化してしまうことも示された。

新しい評価の導入は，確かにテストへの適応という現象に対する1つの解答となりうるだろう。しかし，ここで見てきたように，その評価をただ漫然と導入するだけでは，実施者が意図した効果を持ち得ないと思われる。新しい評価を導入する際には，次に述べるようなインフォームドアセスメントの視点を織り交ぜるなど，多角的な配慮が大切だろう。

インフォームドアセスメント

新しい評価の導入が，それほどポジティブな影響を持っていない原因の1つとして，実施者や学習者の“評価意図や評価基準に関する知識”の不足を挙げることができる。オーセンティック評価をその理念とする新しい評価は，確かに学習者の高次の認知を促進する可能性がある。その一方で，評価の形態がやや新奇であるため，評価の意図や基準が実施者や学習者に分かりにくいという問題があるように見受けられる。この点が不明確であれば，学習者もどのように適応していいかが分からず，結果として学習の促進は生じないだろう。

そこで本稿ではインフォームドアセスメント (informed assessment) という考えの重要性を主張したい。インフォームドアセスメントとは，評価の目的や基準に関して，実施者と受け手との間にしっかりとした知識の伝達・合意がなされているような評価のあり

---

[2] オーセンティック評価は具体的な評価方法ではなく，評価の理念である。このようなオーセンティック評価を具体化する方法の一つがパフォーマンス評価である。パフォーマンス評価とは，学習者に特定の活動 (パフォーマンス) を要求し，そのパフォーマンスから学習者の問題解決プロセスを評価する評価方法である。ただし，その定義には地域・研究者によって大きな違いがある。単なる記述式のテストをパフォーマンス評価に含める場合もあれば，より具体的な活動 (理科のある問題に対して自分で仮説を立てて実験を行う活動や，人の前でプレゼンテーションする活動など) の評価を意味する場合もある。実際に作品 (国語の教科書の朗読テープや歴史新聞など) を作ってもらい，それをパフォーマンスとみなしてアセスメントする場合もある (村山, 印刷中)。

[3] 波及効果研究は主として，語学教育の分野で研究がなされてきた(e.g., Cheng et al., 2004)。しかしここでは，“ポジティブな影響を意図して実施したテストや新しい評価法が，教師や学習者に与える影響”を扱った研究をすべて波及効果の研究として捉え，領域を限定しない。波及効果研究では，オーセンティック評価といった新しい評価の重要性を主張しているが(Bailey, 1996)，オーセンティック評価の効果だけを波及効果と呼ぶわけではない。

方を指す。こうしたインフォームドアセスメントを意識し，評価に関する知識を共有することが，テストであれ新しい評価であれ，学習者にポジティブな影響をもたらす鍵になってくると思われる（Spratt, 2005）[4]。

実際，いくつかの研究で，テストや新しい評価に関する知識の重要性は実証的に示されている。例えば，村山（2005）は，学習者のテスト形式に対する知識をテスト形式スキーマと呼び，この知識がテストへの適応に大きな役割を果たしていることを示した。村山（2006）は，同じテストでも，テスト形式スキーマに介入するだけで，学習者の行動に違いが見られたことを報告している。また，Firestone et al.（1998）の事例では，新しい評価に対する教師の無理解が問題になっていた。さらに，Fuchs, Fuchs, Karns, Hamlett, Dutka, & Katzaroff（2000）は，パフォーマンス評価の評価基準を明示的に伝えることで，学習者の学習が改善されることを報告している。こうした研究は，インフォームドアセスメントという評価のあり方の有効性を示唆するものであろう。

インフォームドアセスメントにおける評価の知識の共有は，さまざまな形でなされるものであるが，特に有効な形態として，グループモデレーション（group moderation）を考えることができる。グループモデレーションとは，具体的な答案・作品の事例と評価基準をつき合わせながら，評価基準を改善していく集団での検討会である（村山，印刷中）。Fuchs, Fuchs, Karns, Hamlett, & Katzaroff（1999）は，グループモデレーションを通して教師がパフォーマンス評価に対する理解を深め，パフォーマンス評価の実施が学習者の学習を促進したことを報告している。このように，モデレーションは評価に対する知識を深め，評価のポジティブな影響力を引き出す可能性を持っている。また，通常の教育活動では得られない，さまざまな“気づき”を促してくれるという意味で，教育実践的な意義も大きい。

ただしこうした作業に伴うコストも，教育実践では重要な要因となっている点に注意が必要である。Torrance（1993）の事例からも明らかなように，新しい評価が教育現場でなかなか根付かない背後には，こうしたコストの問題がある（Black, 1994；Kleinert, Kennedy, & Kearns, 1999）。ただコストが高まるだけで効果が可視的にならなければ，教育現場に浸透するのは難しいだろう（Guskey, 1994）。効果をアナウンスできるだけの，新しい評価に関する実証的な検討も，今後精力的に行う必要があるだろう。

妥当性概念の拡張

これまで，テストへの適応が妥当性の危機を生じさせると考えてきた。言い換えるなら，テストへの適応によって，妥当性が変化をするとしてきた。しかし，そのように考えるのではなく，テストへの適応ということを視野に入れた上で，妥当性概念自体を拡張することも可能だろう。すなわち，妥当性とは“測定したいものが測定できているのか”というのが古典的な定義であるが，“学習者がテストに適応することも考慮に入れた上で，測定したいものが測定できているのか”とする視点である。

またさらに，“学習行動への危機”までも視野に入れて，妥当性概念を拡張することも可能だと思われる。すなわち，上記のような考え方に加えて，“その評価が学習者にポジティブな影響を与えること”も妥当性の要件の一つとする視点である。このような妥当性概念のもとでは，その評価が高次認知を測定しているかどうかだけが問題にならない。その上で，学習者の高次の認知能力や問題解決能力を促進するような影響を与えるのか，ということまで問われる。

こうした妥当性概念の拡張は，近年の妥当性研究の流れにも合致している（レビューとしてMoss, 1992）。Messick（1984, 1989, 1996）は，“解釈・使用の適切性”という観点から妥当性を定義した上で，何らかの評価を実施したことによって生じる社会的な結果や副次効果も，妥当性の一つの側面とした。これを結果妥当性（consequential validity）という。Frederiksen & Collins（1989）は，教師の教授方法や学習者の学習方法をより発展させるような評価をシステム妥当性（systemic validity）の

[4] もちろん新しい評価の導入がそれほどポジティブな影響を持っていない理由はそれだけではない。後述するように，評価の実施に伴うコストの問題もある。また，実際の教育場面では，テストがさまざまなレベルで存在していることも要因の一つとして指摘できる。中学校や高校における定期考査の内容は，制度レベルのテスト，すなわち入学試験に大きな制約を受ける。教室レベルでテストを改善したとしても，入学試験の内容が変わらなければ，学習者への影響は限定的になると思われる。逆に，入学試験レベルのテストが変わったとしても，それが教室レベルのテストや教師の教え方にうまく反映されなければ，やはりその効果は弱いものになるだろう。波及効果研究はこのような複雑なシステムとしての関係がある中で，単一のテストの効果を見出そうとしたために，効果を検出できなかったのかもしれない。実際，Firestone et al.（1998）では，制度レベルの要因が，パフォーマンス評価の影響を弱めたことを考察している。現実の教育現場でテストの効果を考える場合，さまざまなレベルの要因が，システムとして複合的に働いていることを意識する必要があるだろう（Frederiksen & Collins, 1989）。

ある評価と呼んだ。Morrow (1986) は波及効果研究の中で波及妥当性 (washback validity) という概念を提出し，ポジティブな波及効果を持つかということも，妥当性の要件だとした。評価が与える影響も含めて妥当性を概念化すべきだという主張は，他にも Linn, Baker, & Dunbar (1991) などでなされている[5]。

妥当性概念を拡張することで，“いい評価法を一度開発すればそれで終わりである”といった静的な妥当性の考え方から脱却できる。妥当性は，学習者のテストへの適応によってダイナミックに変化するのである。評価開発者にとって，信頼性と妥当性を保障することは，大きな課題である。そのとき，従来の静的な妥当性の考え方ではなく，テストへの適応を視野に入れた動的な妥当性の捉え方をすることが大切だろう。

表面的妥当性への意識

学習者にポジティブな影響を与えるような評価を考える場合に，表面的妥当性という視点も大切になってくると思われる。村山 (2005) は，たとえ意味理解が有効な空所補充型テストを実施しても，学習者は空所補充型テストという表面的な形式に囚われてしまい，意味理解型の学習方略を使用するようにならないことを示した。このことは，学習者によい影響を与える評価を開発するためには，内容だけでなく，その表面的な見え方（すなわち表面的妥当性）にも気を配る必要があることを示している (e.g., Thiede, 1996)。これまで，表面的妥当性はにせものの妥当性として，ほとんど重視されてこなかった。しかし，テストへの適応という視点から，テストや新しい評価と学習者との相互作用を考えたとき，その評価が学習者にどのように“見える”のかは，大きな問題になってくると思われる(村山, 印刷中 ; Watanabe, 2001 ; Wiggins, 1989)。

終わりに

本稿では，“テストへの適応”という概念を提出し，その実証的基盤を明らかにすると同時に，その概念がもたらす教育実践上の問題点について検討した。そして最後に，その問題点を解決するためのいくつかの視点を提出した。

ここで扱ったのは主として，認知的領域の教育評価に関しての議論であった。しかし，テストへの適応による学習行動の変化という議論はそこだけに留まるものではないと思われる。第1に，情意的領域の評価にも適用可能だと考えられる。情意的領域の測定は，どうしても学習者の表面的な行動や自己報告に頼ってしまうだけに，テストへの適応の問題はより大きいものになるだろう (cf. McClelland, Koestner, & Weinberger, 1989)。現時点ではこの点に関する研究はほとんど見受けられないが，今後検討していく必要があると思われる。第2に，教育評価だけでなく，心理検査といった心理測定にも関係が深い問題だと考えられる。例えば，態度測定の研究では，態度を測定すること自体が，被調査者の行動を促進することが示されている (Chandon, Morwitz, & Reinartz, 2005 ; Feldman & Lynch, 1988)。このような現象の結果，妥当性係数が増大してしまうことを，自己生成的な妥当性 (self-generated validity) と呼ぶ。この自己生成行動は，測定への適応というわけではないが，測定が被調査者に影響を与えているという意味で，本稿での主張と合致する知見であると言えよう。

“テストへの適応”は，さまざまな領域の問題が複合している，多面的な現象である。それがどのような認知的メカニズムで生じているのかという認知心理学的な問題も関与していれば，それが測定の妥当性にどのような影響を与えているかという心理測定的な問題もある。さらに，新しい評価の方法を模索している教育評価研究に関係もあれば，テストをどのように実践の中で生かすかという教育実践的な問題にも示唆がある。それだけ意義が大きいと同時に，研究が難しい領域でもある。本稿で示した包括的な視野を足掛かりに，今後より一層の研究がなされる必要があるだろう。

[5] 最近では，このような妥当性概念の拡張に，否定的な立場も出てきている。Borsboom, Mellenbergh, & Van Heerden (2004) は，“*The concept of validity*”というタイトルの論文の中で，近年の妥当性の定義が複雑になりすぎているとして，Messick らの概念化を鋭く批判した。その上で，従来の“測定したいものが測定できているのか”という定義を再認識すべきだと主張し，潜在変数と観測変数の因果的関係という立場から妥当性を考えた。

引用文献

Airasian, P. W. 1988 Measurement driven instruction : A closer look. *Educational Measurement : Issues and Practice*, **7**(4), 6-11.

Alderson, J. C., & Hamp-Lyons, L. 1996 TOEFL preparation courses : A study of washback. *Language Testing*, **13**, 280-297.

Alderson, J. C., & Wall, D. 1993 Does washback exist ? *Applied Linguistics*, **14**, 115-129.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. 1999

Standards for educational and psychological testing. Washington, DC : American Educational Research Association.

Bailey, K. M. 1996 Working for washback : A review of the washback concept in language testing. *Language Testing*, **13**, 257-279.

Balota, D. A., & Neely, J. H. 1980 Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology : Human, Learning, and Memory*, **6**, 576-587.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. 1983 Effects of coaching programs on achievement test performance. *Review of Educational Research*, **53**, 571-585.

Black, P. J. 1994 Performance assessment and accountability : The experience in England and Wales. *Educational Evaluation and Policy Analysis*, **16**, 191-203.

Bloom, B. S. 1981 *All our children learning*. New York : McGraw-Hill. (稲葉宏雄・大西匡哉(監訳) 1986 すべての子どもにたしかな学力を 明治図書)

Bloom, B. S., Hastings, T. H., & Madaus, G. F. 1971 *Handbook on formative and summative evaluation of student learning*. New York : McGraw-Hill. (梶田叡一・渋谷憲一・藤田恵璽(訳) 1973 教育評価法ハンドブック 第一法規)

Bond, L. 1989 The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. Washington, DC : American Council on Education/Macmillan. Pp.429-444. (池田 央・柳井晴夫・藤田恵璽・繁桝算男 (監訳) 1992 教育測定学 (下巻) みくに出版 Pp.137-159.)

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. 2004 The concept of validity. *Psychological Review*, **111**, 1061-1071.

Breen, L. K., & Wilding, J. 1984 Noise, time of day and test expectations in recall and recognition. *British Journal of Psychology*, **75**, 51-63.

Broekkamp, H., Van den Bergh, H., Van Hout-Wolters, B. H. A. M., & Rijlaarsdam, G. 2002 Will that be on the test? Perceived task demands and test performance in a classroom context. *European Journal of Psychology of Education*, **17**, 75-92.

Broekkamp, H., Van Hout-Wolters, B. H. A. M., Van den Bergh, H., & Rijlaarsdam, G. 2004 Teachers' task demands, students' test expectations, and actual test content. *British Journal of Educational Psychology*, **74**, 205-220.

Butler, D. L., & Winne, P. H. 1995 Feedback and self-regulated learning : A theoretical synthesis. *Review of Educational Research*, **65**, 245-281.

Cannell, J. J. 1988 Nationally normed elementary achievement testing in America's public schools : How all 50 states are above the national average. *Educational Measurement : Issues and Practice*, **7**, 5-9.

Carrier, C. A., & Titus, A. 1981 Effects of note-taking pretraining and test mode expectations on learning from lectures. *American Educational Research Journal*, **18**, 385-397.

Chandon, P., Morwitz, V. G., & Reinartz, W. J. 2005 Do intentions really predict behavior? : Self-generated validity effects in survey research. *Journal of Marketing*, **69**(2), 1-14.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.) 2004 *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Connor, J. M. 1977 Effects of organization and expectancy on recall and recognition. *Memory and Cognition*, **5**, 315-318.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. 1974 Longitudinal studies of test-wiseness. *Journal of Educational Measurement*, **11**, 209-211.

Crooks, T. J. 1988 The impact of classroom evaluation practices on students. *Review of Educational Research*, **58**, 438-481.

d'Ydewalle, G., Swerts, A., & Corte, E. D. 1983 Study time and test performance as a function of test expectations. *Contemporary Educational Psychology*, **8**, 55-67.

Deaton, W. L., Halpin, G., & Alford, T. 1987 Coaching effects on California Achievement Test scores in elementary grades. *Journal of*

Educational Research, **80**, 149-155.

DeSteno, D., Petty, R. E., Wegener, D. T., & Rucker, D. D. 2000 Beyond valence in the perception of likelihood : The role of emotion specificity. *Journal of Personality and Social Psychology*, **78**, 397-416.

Diamond, J. J., & Evans, W. J. 1972 An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement*, **9**, 145-150.

Dolly, J. P., & Williams, K. S. 1986 Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement*, **46**, 619-625.

遠藤貴広 2003 G. ウィギンズの教育評価論における「真正性」概念—「真正の評価」論に対する批判を踏まえて 教育目標・評価学会紀要, **13**, 34-43.

Evans, W. 1984 Test wiseness : An examination of cue-using strategies. *Journal of Experimental Education*, **52**, 141-144.

Feldman, J. M., & Lynch, J. G., Jr. 1988 Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, **73**, 421-435.

Firestone, W. A., Mayrowetz, D., & Fairman, J. 1998 Performance-based assessment and instructional change : The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, **20**, 95-113.

Frederiksen, N. 1984 The real test bias : Influences of testing on teaching and learning. *American Psychologist*, **39**, 193-202.

Frederiksen, J. R., & Collins, A. 1989 A systems approach to educational testing. *Educational Researcher*, **18**(9), 27-32.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. 2000 The importance of providing background information on the structure and scoring of performance assessments. *Applied Measurement in Education*, **13**, 1-34.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., & Katzaroff, M. 1999 Mathematics performance assessment in the classroom : Effects on teacher planning and student problem solving. *American Educational Research Journal*, **36**, 609-646.

藤澤伸介 2002 ごまかし勉強（上）学力低下を助長するシステム 新曜社

福島県教育委員会 1972 福島県教育史 第一巻（先近代近代前期編） 岩瀬書店

Gay, L. R. 1980 The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, **17**, 45-50.

Gipps, C. V. 1994 *Beyond testing : Towards a theory of educational assessment*. London : Falmer Press. (鈴木秀幸(訳) 2001 新しい評価を求めて—テスト教育の終焉— 論創社)

Guskey, T. 1994 What you assess may not be what you get. *Educational Leadership*, **51(6)**, 51-54.

Haladyna, T. M., Nolen, S. B., & Haas, N. 1991 Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher*, **20**, 2-7.

Halpin, G., & Halpin, G. 1982 Experimental investigations of the effects of study and testing on student learning, retention, and ratings of instruction. *Journal of Educational Psychology*, **74**, 32-38.

Harkins, S. G. (Ed.) 2001 *Multiple perspectives on the effects of evaluation on performance : Toward an integration*. New York, NY : Kluwer Publishers.

橋本重治 1981 到達度評価の研究—その方法と技術— 図書文化

橋本重治 1983 続・到達度評価の研究—到達基準の設定の方法— 図書文化

Jaeger, R. M. 1989 Certification of student competence. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. Washington, DC : American Council on Education/Macmillan. Pp. 485-514. (池田 央・柳井晴夫・藤田恵璽・繁桝算男 (監訳) 1992 教育測定学（下巻） みくに出版 Pp.215-257.)

Jones, H. E. 1923-1924 Experimental studies of college teaching : The effect of examination on permanence of learning. *Archives of Psychology*, **10**, 5-70.

鹿毛雅治 1996 内発的動機づけと教育評価 風間書房

鹿毛雅治 2004 教育評価再考—実践的視座からの展望— 心理学評論, **47**, 300-317. (Kage, M. 2004 Educational evaluation reconsidered: From the perspective of educational practice. *Japanese Psychological Review*, **47**, 300-317.)

Kleinert, H. L., Kennedy, S., & Kearns, J. F., 1999 The impact of alternate assessments : A statewide teacher survey. *Journal of Special Education*, **33**, 93-102.

Latham, H. 1877 *On the action of examinations considered as a means of selection.* Cambridge, England : Deighton, Bell & Co.

Laurillard, D. 1979 The processes of student learning. *Higher Education*, 8, 395-409.

Linn, R. L., Baker, E. L., & Dunbar, S. B. 1991 Complex, performance-based assessment : Expectations and validation criteria. *Educational Researcher*, **20**(8), 15-21.

Lundeberg, M. A., & Fox, P. W. 1991 Do laboratory findings on test expectancy generalize to classroom outcomes ? *Review of Educational Research*, **61**, 94-106.

Madaus, G. F. 1983 Minimum competency testing for certification : The evolution of test validity. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing*. Boston, MA : Kluwer-Nijhoff Publishing. Pp.21-61.

Madaus, G. F. 1988 The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum*. Chicago, IL : University of Chicago Press. Pp. 83-121.

Marton, F., & Säljö, R. 1976 a On qualitative differences in learning I : Outcome and process. *British Journal of Educational Psychology*, **46**, 4-11.

Marton, F., & Säljö, R. 1976 b On qualitative differences in learning II : Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology*, **46**, 115-127.

May, R. B., & Thompson, J. M. 1989 Test expectancy and question answering in prose processing. *Applied Cognitive Psychology*, **3**, 261-269.

McClelland, D. C., Koestner, R., & Weinberger, J. 1989 How do self-attributed and implicit motives differ? *Psychological Review*, **96**, 690-702.

McDaniel, M. A., Blischak, D. M., & Challis, B. 1994 The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology*, **19**, 230-248.

McMorris, R. F., Brown, J. A., Snyder, G. W., & Pruzek, R. M. 1972 Effects of violating item construction principles. *Journal of Educational Measurement*, **9**, 287-295.

Mehrens, W. A., & Kaminski, J. 1989 Methods for improving standardized test scores : Fruitful, fruitless, or fraudulent ? *Educational Measurement : Issues and Practice*, 8, 14-22.

Messick, S. 1984 The psychology of educational measurement. *Journal of Educational Measurement*, **21**, 215-237.

Messick, S. 1989 *Validity*. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. Washington, DC : American Council on Education/ Macmillan. Pp.13-103. (池田 央・柳井晴夫・藤田恵璽・繁桝算男(監訳) 1992 教育測定学(上巻) みくに出版 Pp. 19-145.)

Messick, S. 1996 Validity and washback in language testing. *Language Testing*, **13**, 241-256.

Meyer, G. 1934 An experimental study of the old and new types of examination : I. The effect of the examination set on memory. *Journal of Educational Psychology*, **15**, 641-661.

Millman, J., Bishop, C. H., & Ebel, R. 1965 An analysis of test-wiseness. *Educational and Psychological Measurement*, **25**, 707-726.

Morrow, K. 1986 The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing*. London : NFER/Nelson. Pp.1-13.

Moss, P. A. 1992 Shifting conceptions of validity in educational measurement : Implications for performance assessment. *Review of Educational Research*, **62**, 229-258.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. 2004 *Findings from IEA's trends in international mathematics and science*

study at the fourth and eighth grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murayama, K. 2005 *Test format and memory : A mediational analysis*. Paper presented at the 6th Tsukuba International Conference on Memory, Tsukuba.

村山 航 2003 a 学習方略の使用と短期的・長期的な有効性の認知との関係 教育心理学研究, **51**, 130-140. (Murayama, K. 2003a Learning strategy use and short- and long-term perceived utility. *Japanese Journal of Educational Psychology*, **51**, 130-140.)

村山 航 2003 b テスト形式が学習方略に与える影響 教育心理学研究, **51**, 1-12. (Murayama, K. 2003 b Test format and learning strategy use. *Japanese Journal of Educational Psychology*, **51**, 1-12.)

村山 航 2004 テスト形式の違いによる学習方略と有効性の認知の変容 心理学研究, **75**, 262-268. (Murayama, K. 2004 Effects of test format on learning strategy and perceived utility. *Japanese Journal of Psychology*, **75**, 262-268.)

村山 航 2005 テスト形式の予期による方略変容メカニズムの検討 教育心理学研究, **53**, 172-184. (Murayama, K. 2005 Exploring the mechanism of test-expectancy effects on strategy change. *Japanese Journal of Educational Psychology*, **53**, 172-184.)

村山 航 2006 テスト形式スキーマへの介入が空所補充型テストと学習方略との関係に及ぼす影響 教育心理学研究, **54**, 63-74. (Murayama, K. in press Effects of test format scheme on the relationships between objective tests and learning strategies. *Japanese Journal of Educational Psychology*, **54**, 63-74.)

村山 航 印刷中 教育評価 鹿毛雅治(編) 講座教育心理学 朝倉書店

National Research Council 2001 *Knowing what students know*. Washington, DC : National Academy Press.

Neely, J. H., & Balota, D. A. 1981 Test-expectancy and semantic-organization effects in recall and recognition. *Memory and Cognition*, **9**, 283- 300.

Newble, D. I., & Jaeger, K. 1983 The effect of assessments and examinations on the learning of medical students. *Medical Education*, **17**, 165-171.

西岡加名恵 2003 教科と総合に活かすポートフォリオ評価法—新たな評価基準の創出に向けて— 図書文化

Nungester, R. J., & Duchastel, P. C. 1982 Testing versus review : Effects on retention. *Journal of Educational Psychology*, **74**, 18-22.

Oakhill, J., & Davies, A. M. 1989 The effects of time of day and subjects' test expectations on recall and recognition of prose materials. *Acta Psychologica*, **72**, 145-157.

Oakhill, J., & Davies, A. M. 1991 The effects of test expectancy on quality of note taking and recall of text at different times of day. *British Journal of Psychology*, **82**, 179-189.

お茶の水女子大学21世紀COEプログラム 2004 JELS第3集 算数・数学学力調査報告

Popham, W. J. 1987 The merits of measurement-driven instruction. *Phi Delta Kappan*, **68**, 679-682.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. 1985 Measurement-driven instruction : It's on the road. *Phi Delta Kappan*, **66**, 628-634.

Powers, D. E. 1985 Effects of test preparation on the validity of a graduate admissions test. *Applied Psychological Measurement*, **9**, 179-190.

Ramsden, P. 1988 Context and strategy : Situational influences on learning. In R. R. Schmeck (Ed.), *Learning strategies and learning styles*. New York : Plenum Press. Pp.159-184.

Rickards, J. P., & Friedman, F. 1978 The encoding versus the external storage hypothesis in note taking. *Contemporary Educational Psychology*, **3**, 136-143.

Rogers, W. T., & Bateson, D. J. 1991 Verification of a model of test-taking behavior of high school seniors. *Journal of Experimental Education*, **59**, 331-350.

Ruiz-Primo, M. A., & Shavelson, R. J. 1996 パフォーマンスアセスメントにおけるレトリックと現実：最新の情報 学習評価研究, **27**, 10-31.

Samson, G. E. 1985 Effects of training in test-taking skills on achievement test performance : A quantitative synthesis. *Journal of Educational Research*, **78**, 261-266.

Sanjivamurthy, P., & Kumar, V. K. 1983 Test mode anticipation and performance : A classroom experiment. *Contemporary Educational Psychology*, 8, 355-365.

Sarnacki, R. E. 1979 An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, **49**, 252-279.

Sax, G., & Collet, L. 1968 An empirical comparison of the effects of recall and multiple-choice tests on student achievement. *Journal of Educational Measurement*, **5**, 169-173.

Schmidt, S. R. 1983 The effects of recall and recognition test expectancies on the retention of prose. *Memory & Cognition*, **11**, 172-180.

Scouller, K. 1998 The influence of assessment method on students' learning approaches : Multiple choice question examination versus assignment essay. *Higher Education*, **35**, 453-472.

Scouller, K. M., & Prosser, M. 1994 Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, **19**, 267-279.

Scruggs, T. E., White, K. R., & Bennion, K. 1986 Teaching test-taking skills to elementary-grade students : A meta-analysis. *Elementary School Journal*, **87**, 69-82.

清水寛之 1990 再生に及ぼす検査予期とリハーサル方略の効果：同時提示事態での検討 心理学研究, **61**, 268-272. (Shimizu, H. 1990 The effects of test expectancy and rehearsal strategies on recall in the situation of simultaneous presentation. *Japanese Journal of Psychology*, **61**, 268-272.)

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. 1996 Test impact revisited : Washback effect over time. *Language Testing*, **13**, 298-317.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. 1970 Learning test-wiseness by programmed texts. *Journal of Educational Measurement*, **7**, 247-254.

Smith, M. L. 1991 Put to the test : The effects of external testing on teachers. *Educational Researcher*, **20**, 8-11.

Spratt, M. 2005 Washback and the classroom : The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, **9**, 5-29.

田中耕治 2002 a 新しい教育評価の理論と方法 第Ⅰ巻 理論編 日本標準

田中耕治 2002 b 新しい教育評価の理論と方法 第Ⅱ巻 教科・総合学習編 日本標準

田中耕治・西岡加名恵 1999 総合学習とポートフォリオ評価法 入門編 日本標準

Thiede, K. W. 1996 The relative importance of anticipated test format and anticipated test difficulty on performance. *Quarterly Journal of Experimental Psychology*, **49A**, 901-918.

Thomas, P. R. 1982 Consistency in learning strategies. *Higher Education*, **11**, 249-259.

Thomas, P. R., & Bain, J. D. 1984 Contextual dependence of learning approaches : The effects of assessments. *Human Learning*, **3**, 227-240.

Torrance, H. 1993 Combining measurement-driven instruction with authentic assessment : Some initial observations of National Assessment in England and Wales. *Educational Evaluation and Policy Analysis*, **15**, 81-90.

Towns, M. H., & Robinson, W. R. 1993 Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching*, **30**, 709-722.

Tversky, B. 1973 Encoding processes in recognition and recall. *Cognitive Psychology*, **5**, 275-287.

Vernon, P. E. 1956 *The measurement of abilities. (2nd ed.)*. London : University of London Press.

Vernon, P. E. 1962 The determinants of reading comprehension. *Educational and Psychological Measurement*, **22**, 269-286.

Wahlstrom, M., & Boersma, F. J. 1968 The influence of test-wiseness upon achievement. *Educational and Psychological Measurement*, **28**, 413-420.

Wall, D., & Alderson, J. C. 1993 Examining wash-

back: The Sri Lankan impact study. *Language Testing*, **10**, 41-69.

Watanabe, Y. 1996 Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, **13**, 318-333.

Watanabe, Y. 2001 Does the university entrance examination motivate learners? A case study of learner interviews. 秋田英語英文学イングラム先生来秋記念特別号抜刷, 100-110.

Wiggins, G. 1989 A true test : Toward more authentic and equitable assessment. *Phi Delta Kappan*, **70**, 703-713.

Wiggins, G. 1998 *Educative assessment : Designing assessments to inform and improve student performance*. San Francisco, CA : Jossey-Bass.

Wnek, I., & Read, J. D. 1980 Recall and recognition encoding differences for low- and high-imagery words. *Perceptual and Motor Skills*, **50**, 391-394.

Yen, W. M., & Ferrara, S. 1997 The Maryland School Performance Assessment Program : Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement*, **57**, 60-84.

Zimmerman, B. J. 2001 Attaining self-regulation : A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation*. San Diego, CA : Academic Press. Pp.13-39.

謝　辞

本論文をまとめるにあたりご指導・ご助言いただきました東京大学の市川伸一教授，秋田大学の渡部良典助教授に感謝いたします。

(2005.3.10 受稿, 10.5 受理)

"Adaptation to the Test" : A Review of Problems and Perspectives

KOU MURAYAMA (RESEARCH FELLOW OF THE JAPAN SOCIETY FOR THE PROMOTION OF SCIENCE/DEPARTMENT OF HUMAN SYSTEM SCIENCE, GRADUATE SCHOOL OF DECISION SCIENCE AND TECHNOLOGY, TOKYO INSTITUTE OF TECHNOLOGY)

JAPANESE JOURNAL OF EDUCATIONAL PSYCHOLOGY, 2006, 54, 265—279

When students expect to be tested, they often accommodate their method of learning to the test demands or to their teacher's evaluation criteria. This phenomenon is, in the present article, called "adaptation to the test." In a review of the relevant literature, 3 issues are addressed. First, the critical role of learning strategy was confirmed from a review of empirical findings on test expectancy effects and studies of learning strategies. Second, the present review points out that adaptation to the test leads to 2 crises : problems with students' learning behavior, and problems with test validity. Third, the following were presented as possible ways to overcome these difficulties : (1) the elimination of individual differences in test-wiseness and test-taking skills, (2) the introduction of alternative assessment methods, (3) informed assessment, (4) an expansion of the concept of validity, and (5) a concern about face validity.

Key Words : adaptation to the test, test expectancy effect, informed assessment, test validity, alternative assessment