作成：中野茂

新しい評価を求めて テスト教育の終焉

キャロライン・V・ギップス著

鈴木秀幸訳

論創社

2001 年

（第 1 章 評価のパラダイムの転換）

（第 2 章 評価と学習の関係）

（第 3 章 テストのもたらす弊害）

（第 4 章 妥当性と信頼性）

第 5 章　クライテリオン準拠評価

ノルム準拠評価とクライテリオン準拠評価の相違点

「私がクライテリオン準拠測定と呼ぶものは、質についても絶対的な水準に準拠するものであり、一方ノルム準拠測定と呼ぶものは、相対的な水準に準拠するものである」（グレイザー　1963 年）

「このように、生徒の到達事項をクライテリオン準拠によって評価する測定方法は、特定の生徒の達成した能力の度合いについて、他の生徒のパフォーマンスを参照することなく、情報を提供するものである」（グレイザー　1963 年）

教育評価は「個人の成長（個人間の違いではなく）に関心を持って」おり、「テストは学習した内容に関連している」ので、「クライテリオン準拠テストが発する問いは「X は Z で Y よりも高い得点をあげたか」ではなく「X と Y は Z ができるか」なのである」

ノルム準拠評価とクライテリオン準拠評価の相互作用

ドイツのグレード分けの仕組みであるノーテンスカラの例

→「教育課程の各段階で指導され修得しなければならない知識や技能は国によって定められている。」（クライテリオン準拠）

→「所属する学校の種類によって「よい」という評価基準が違っている」（ノルム準拠）

1970 年代クライテリオン準拠評価の開発が後押しされた背景と現状

ノルム準拠テストの限界：ノルム準拠試験においては、「ほとんどの生徒ができるような問題は削除される」

→学習してきたことを問う試験ではなく、生徒の違いを見つけるための試験

「教育評価に携わる者から見れば、何よりもまずうまく作成されたクライテリオン準拠テストの恩恵は、何が測定されているかについて、明確な表現を得られるところにある。このことは次の段階で、テストの得点の意味が大切にされているため、より正確な測定を可能とすることになる。そして、最終的に明確さの増すにつれて、教育評価携わるものは、政策決定者に対してより説得力のある解釈可能なデータを提供できる」（ポファム 1987 年）

「うまく作成されたクライテリオン準拠テストでの明確な表現は、教師が生徒の十分にできなかった分野の指導を可能にする。一方、ノルム準拠テストでは、指導すべき分野を知るには、問題の一つ一つの出来具合を見なくてはならない」

→テストができなかった生徒への指導の効率

「クライテリオン準拠テストへの関心の高まりとともに、アメリカのテストを編集している出版社は、ノルム準拠テストにクライテリオン準拠テストの解釈を付け加えるようになった」（ポファム 1992 年）

「クライテリオン準拠テストはアメリカで広く採用されているが、多くの場合、最低限習得度テストや完全習得学習の教育計画の枠内においてである。」

「修正された形式として、英国では等級別評価があり、さらに、1980 年代中ごろには GCSE 試験とナショナル・カリキュラムの評価をクライテリオン準拠で行うことが決定された」

現在のクライテリオン準拠テストの際立った特徴

「現在のクライテリオン準拠テストの際立った特徴は、グレイザーが言っているように得点を基準となるパフォーマンスと比較することではなく、評価される内容やドメイン（領域）を詳しく規定することであると考えられている（シェパード　1991）。（ここでドメインというのは、教科の知識の分野の意味であり、たとえば数学のドメインの測定というように）ドメイン準拠テストとは、クライテリオン準拠テストの一種である。このモデルでは、領域が明確に定義され、このドメインでテストされる可能性のあるすべての内容（全内容領域という）が規定されることになる。テストの問題は、この全内容領域から定められたサンプリングの規則に則って選び出されることになる。このテストでの受験者の得点は、そのドメイン全体について受験者が習得したかの指標となる。」

ドメイン準拠テストとクライテリオン準拠テストの違い

「どちらの評価も、学習内容に焦点を当てているところは同じである。しかし純粋な形のドメイン準拠テストでは、ドメインに含まれる内容を規定する明確な規則と、テストの対象とする項目を選び出す抽出手続きを規定することにより、テストとして選ばれた項目の問題の得点から、領域全体を推定できるようにしている。そのような詳細な規定（「見せかけの正確性」とリンは言っている）は、極めて狭い領域でのみ可能とされる。よく使われる事例は、0 から 9 までの数字を 2 つたす 100 通りの計算である。（…）リンも指摘しているように、これはクライテリオン準拠テストに対する批判の一つであるが、クライテリオン準拠テストは詳細で限定された課題に焦点を当てるというわけではない。クライテリオン準拠テストの実施される多くの領域で、すべての要素を特定することは不可能であり、もっと広い規定の仕方に注目すべきである。」

- **技術的な問題(p. 114)**

テスト問題への影響

ノルム準拠テストとクライテリオン準拠テストを規定する概念の違いはテストの作成に影響を及ぼす。

「（クライテリオン準拠テストは）生徒ができる課題とそうでないものを見つけることを目的としている。したがって、あまりに易しかったり、難しかったりして受験者

の点数が開かないような問題であっても、もしそれが学習分野で重要な要素であれば、テスト問題として採用されることになる。」

- **クライテリオン準拠評価の信頼性(p. 115)**

正確性の点からみたクライテリオン準拠評価とドメイン準拠評価の違い

「クライテリオン準拠評価では、一般に修得したかどうかを判断して採点される。これはドメイン準拠評価の採点に比べて正確性に欠ける」

「クライテリオン準拠評価では、評価の課題について特定の割合以上にできたり、基準となるパフォーマンスを示した生徒は修得したと分類され、そうでない生徒は未修得とされるのである」

クライテリオン準拠評価の信頼性の測定

「ショーガンとハッチソン（1991）は、英国のナショナル・カリキュラムの評価と関連する、クライテリオン準拠評価の開発に用いられた信頼性の測定方法を詳しく説明している。」

「Key Stage 1 のナショナル・カリキュラムの評価を検証した研究チームは、信任性という用語を使った。」

＊信任性：「信頼性と妥当性を統合した概念が信任性である」

→「この短絡がちな議論の示すように、クライテリオン準拠評価の信頼性を検証する適切な方法がどのようなものであろうかについて、コンセンサスを得るにいたっていないのである。また、ナショナル・カリキュラムの評価の信頼性を検証する方法も、荒っぽいものであることがわかる。」

→「成功/失敗を分ける点数という概念をもたず、一定の広がりを持った得点を導き出す(それによって生徒を区別して、グレードを振り分けることになる)GCSE 試験のようなクライテリオン準拠評価では、伝統的な信頼性の検証手段が適切なものとなっている。ウッドによれば信頼性の向上はテストをより個別的、具体的なものに限定して、グレード記述に合わせた採点を心がけることによって可能になるという。」

ノルム準拠テストとクライテリオン準拠テストの違いを示すもう一つの分野である一般化可能性

＊「一般化可能性とは、特定の評価で得た結果を、評価の問題や課題を越えた広い範囲の能力や技能の学習状況を示すものと考えてよいかというものです（…）このように、一部から全体の状況を推測してよいかということを一般化可能性と言います。」（p.52）

「ノルム準拠テストでは、テストの結果を学習内容の全体についての推定に用いることはできない。そのかわり、標準となる集団に対しての位置を推定することはできるのである。」

「対照的に、クライテリオン準拠評価の提唱者によると、上手に作成されたクライテリオン準拠テストのテスト問題は、学習内容の領域に見合ったものとなるため、テスト問題からより広い領域の学習内容にまで一般化できるという。このように、クライテリオン準拠テストの質を決める決定的な要因は、内容領域の規定が十分に行われ、テストに含まれる問題が内容領域を分析して得られた適切な課題となっていることである（ピィリナー、1979）

- **まとめること(p.118)**

クライテリオン準拠評価の大きな問題としての「まとめること」

「個々人の詳細な学習状況を示す情報を、単一の数字やグレードで報告して台無しにすること」が問題

「また、点数配分やまとめのための複雑なシステムによって最終的な点数が歪められ、習得状況が見えなくなってしまう」

- **妥当性(p.120)**

クライテリオン準拠評価に関連する妥当性の３つの要素

「領域は適切に規定されているか、内容（抽出された問題）は領域や求める構成概念を適切に代表しているか、評価の課題は構成概念を適切に反映しているか（すなわち、課題が私たちの評価しようとしているものをきちんととらえているか）」

クライテリオン準拠評価の妥当性の検証

「ウッド（1991）の指摘するように、領域または構成概念を注意深く規定し、そこから課題を抽出することにより、クライテリオン準拠評価は推定手続きを低く見積もっている。推定する部分が少ないということは、評価しようとする技能や目的を厳格に

規定しなくてはならない。これがクライテリオン準拠評価での抜き差しならない不合理な状況を作り出している。すなわち、あまりにも狭く厳格に規定された評価基準は断片が分離した評価課題へ過度に依存しがちになる。一方で、あまりにも一般的で緩やかに規定された評価基準は評価の妥当性と信任性を低くする。」

「ハンブルトンやロジャースによれば（1991）、専門家の判定がクライテリオン準拠テストの内容妥当性を検証する主な形式であり、彼らは内容妥当性をテストの使用がどうあれ重要だとみなしている。クライテリオン準拠評価の妥当性の検証方法についての詳しい議論は、ハンブルトンとロジャースを参照してほしい。(…) クライテリオン準拠評価の妥当性についての問題は、1980 年代に適切な枠組みによって研究されて、優れた研究が登場したとみている。」

(ハンブルトンやロジャースは)妥当性の検証に関して気を付ける点として、検証の労力の総計はテストの重要性に関連させるべきだと考えている。

クライテリオン準拠評価に関する技術的な問題としてのコンテクスト

「テストの問題に対してのコンテクストの影響を除く事は不可能であり、非常に精密な規定をして同質の問題を作ることも不可能である。代わりの方法は、評価の規定を特定のグループを念頭に置いて作り、その中で共通理解を基礎として評価を作っていこうとするものである。」

- **GCSE(p. 123)**

GCSE 試験でのクライテリオン準拠評価の開発の目的 1

各試験委員会が単一で一貫性を持ち、明確に定義されたグレード基準の策定

GCSE 試験でのクライテリオン準拠評価の開発の目的 2

全体の水準の引き上げ

「ノルム準拠テストでは、どの生徒も平均水準やそれ以上に到達させると言っても意味を持たない。（…）クライテリオン準拠テストでは、理論的にはすべての生徒が最高のグレードを得ることができる」

クライテリオン準拠評価を設計することの難しさ

クライテリオン準拠評価は簡単に規定できるような実行能力（50 メートル泳ぐことができる）には理想的なものであるが、課題がより複雑なものになると、そうでなくな

ってくる。つまり、評価がより複雑なものにならざるを得ないか（例えば、車の免許のテストでは、1 対 1 の徹底的な評価を必要とする）、評価基準をより一般的なものにせざるをえない。もし、基準がより一般的なものになっていくと、解釈の問題が起こりがちになるため、信頼性は低くなっていく。」

グレード別評価基準

GCSE にはすでにおおまかなグレード記述があったが、「パフォーマンスについてのもっと具体的な表現」が求められていた。

各教科ごとの作成委員会が作られ、領域、能力規定、パフォーマンスの定義や基準を作成した。

→グレード基準が複雑すぎて使えなかった

→「グレード基準の原案に基づいた指導や評価の方法が、教科を孤立した課題に分断」

→「グレード別評価基準の原案は、全体として機能しないと結論づけられたのである」（キングダンとストバート、1987）。

パフォーマンス・マトリックス

パフォーマンス・マトリックス：「GCSE 試験の各グレードを得るために必要な、知識、理解、技能等の能力を一覧表にして作成したもの」

1988 年にパフォーマンス・マトリックスは凍結される。

クライテリオン準拠評価の実現可能性

→「達成目標と達成基準の不正確さ、まとめの問題により、せいぜいまとまりに欠けたクライテリオン準拠システムしか作り出せないように思われる。」

複雑なデータを合計する難しさ

GCSE 試験のシラバスにおいて「生徒の各教科のレベルは、各達成目標の点数を合計した点数によって決定される」

→レベル 8 の能力がなくても基本問題でよい点数を取ることで、レベル 8 を得ることが可能

→厳密なクライテリオン準拠評価は、成績の平均を出すのではなく、パイロットの試験のようにすべての基準をクリヤーすることが求められている。

- **結論（p. 129）**

細分化された目標から幅広い目標へ

「厳密な形のクライテリオン準拠評価の条件を満たすためには、評価基準を細かい点まで具体的に示す必要がある。しかしこれは、あまりにも細部まで決めたり、非常に矮小化され厳格な目的を規定することにつながるのである。アメリカでのクライテリオン準拠評価の強力な推進者であるポファムは、以前の細分化された目標の推奨は改めて、いくつかの幅広い目標だけを述べる方向を支持している（ポファム、1987b と 1993a）。」

→幅広い目標は「評価が信任されるかという問題を生み出す」

サドラーズのクライテリオン準拠評価に対する批判

「比較的高度な統計的、技術的な方法に依存」

「生徒の学習の質を直接観察できる人間の判断によってしか優れた評価は不可能」

スタンダード準拠評価

「スタンダード準拠評価は、熟練した教師が日常の学習指導の中で行っている専門的な能力を用いて、質的な判断をしようというものである。スタンダード準拠評価のもう一つの特徴は、最終的なグレードやレベルを判断するために合格点を用いるのではなく「重視されるのは一連のテクストの内容や評価の課題にわたる生徒の様子やパフォーマンスの傾向である」(サンドラー、1987)」

「サドラーはスタンダード準拠評価は実行可能で、信頼できる評価方法であると言う。（…）基準となるスタンダードを規定する方法は、言語表現とこれを説明する事例の組み合わせによるとサドラーは言っている。」

オーストラリアのスタンダード準拠評価

「オーストラリアのスタンダード準拠評価では、シラバスに目標が述べられている。教師は生徒のパフォーマンスを評価する明確な評価基準を与えられ、特定の段階で生徒が典型的にできることを述べた説明書がある(例えば事例集)。」

「このシステムは 1993 年の時点でまだ開発中であり、効果のほどは確かめられていない。しかしそれはノルム準拠評価でもなく、クライテリオン準拠評価でもない評価のモデルの一つのあり方を示している。」

まとめ：クライテリオン準拠評価の限界と可能性

「厳密なクライテリオン準拠評価は明らかに実施不可能であり、望ましいものでもない。特に教育評価の枠内ではそうである。クライテリオン準拠評価が教育的にも技術的にも無理があることが証明されたのであるから、スタンダード準拠評価のような評価の開発は、それ自身として確固としたモデルとして見るべきであり、クライテリオン準拠評価からの妥協の産物ではなく、そこからの発展の方向を示すものである。」

「別の方法として、クライテリオン準拠評価と対照的なものとして、クライテリオン準拠またはスタンダード準拠による報告を考えることである。言い換えるとこれは標準的なパフォーマンスを規定し、この標準に従って生徒のパフォーマンスを報告することを条件として、各学校や各学区が生徒のパフォーマンスをそれぞれのやり方で評価できるようにするものである。これは、アメリカやオーストラリアで現在試みられている方法である。しかし当然これは各評価の統一のてんで大きな問題を抱えている。」

「興味深いことにリン(1993b)は最近、クライテリオン準拠評価がパフォーマンス評価やオーセンティック評価を考えるための概念的枠組みを提供すると言っている。」

「クライテリオン準拠評価の価値は、指導の結果として生じるパフォーマンスに注目することであるとリンは主張する。（…）このような見方は、オーセンティック評価の動向と一致するものである。両方ともその目的は、テストと学習指導の目的の乖離をなくそうというところにある。」

「クライテリオン準拠評価の概念は普段考えられているよりももっと豊かであり、リンドクゥィストの達成テストの第 1 の目的と相応する。すなわち、テストの問題は受験者にどのように問題の内容が複雑であろうと、同じようにできることを要求することである(リンドクゥィスト、1951、リンによる引用、1993b)。これが教育評価の様々な形を開発しようと言う現在の運動の基本をなす哲学である。」