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６章　パフォーマンス評価

【はじめに】

● パフォーマンス評価の定義づけ

ここでは、**パフォーマンスに基礎を置いた課題を用いて実施される評価**を指す。パフォーマンス評価は伝統的な択一式の標準テストへの依存を改めたいと願う人々によって広く使用されている用語であるが、**定義が大雑把**である（アメリカでは択一式テスト以外の全ての評価を指す）。

パフォーマンス評価の目的は、評価によって学習指導を歪めないこと。

「パフォーマンス評価は、話したり書いたりコミュニケーションする技能、問題解決技能など、私たちが生徒に取り組んで欲しいと願っている現実の学習活動をモデルにして評価しようとするもの」（p.135）。イギリスでは標準評価課題とGCSE 試験のコースワークが該当する。

● パフォーマンス評価とオーセンティック評価の違い

オーセンティック評価：

- 通常の授業活動の中で行われるものであって、評価のために特別に設定されたものではない（パフォーマンス評価でないオーセンティック評価は考えにくい）。ポートフォリオがその一例。
- 評価者がどの面がオーセンティックであるかについて具体的に述べられるべき。（メイヤー、1992）

パフォーマンス評価：

- 以前に獲得した知識によって、新しい問題の解決や特定の課題を遂行する能力を測定する体系的な試み（ステイギンズ）。
- 受験者に何かを作ったり、活動に従事することにより、彼らの能力を直接示すように求めるものである。（ハートル、1992）
- 反応の評価については、観察や専門家の判断に大きく依存する。教師が

採点者としてコースワークやエッセイを採点することも含まれる。

- パフォーマンス評価は、アカウンタビリティー目的に使う場合は問題を引き起こす。（フレッチリング、1991）

アカウンタビリティーのためのテスト：
テスト範囲が広い、採用している学校が多いため安価。短時間で実施可能。広い範囲を浅く調べる。採点が簡単で信頼性が高くなければならない。
パフォーマンス評価：
時間がかかる。特定の技能や分野について詳しく多面的な情報を提供することが多い。採点は複合的な観点からなされるため、担任の参加が不可欠。パフォーマンスの標準化はできないため、伝統的な意味での信頼性は高くない。

教科の学習についての研究から、より相互交流的な形式のパフォーマンス評価が登場。テストの形式が学習指導や学習に与える影響を考えれば、パフォーマンス評価をどのような条件下で大規模なテストに使うかを検討する必要がある。
パフォーマンス評価の技術的側面は興味深いが、ハイ・ステイクスな目的に使う場合は技術的側面に注意を払う必要がある。

【パフォーマンス評価の妥当性】

「教育評価のあるべき姿からして、パフォーマンス評価は重要」
妥当性を向上させ、高次の技能を評価する機会を提供している。
実践的な数学や科学的探究といった種類を評価しているように見える→ 表面的な妥当性は高い傾向がある。→ 一般の人々の理解が得られる。

<u>問題となるのは、構成妥当性と領域に関する対応と定義</u>
構成妥当性：領域を厳しく規定して、評価する構成概念を考えて作成されなくてはならない。

- 正しい領域が設定されているか？
- 領域は十分に規定されているか？
- 課題の抽出を正しく行なっているか、正しく抽出したとしても、課題から領域についての推定を行えるか？

領域の適用範囲や一般化可能性を無視して課題（作品やパフォーマンス）を見るが、得点の意味を推定するときには構成概念（領域を構成しているものの理解）

を用いる。（メシック、1992）

構成妥当性が危うい！
作品やパフォーマンスにテストで問うべき構成概念が欠けている場合（構成概念非代表性）や、関係ない知識や技能に依存する場合（構成概念不適合性）。

構成妥当性の検証方法：
生徒が課題を遂行する過程を調べる（課題について生徒と話す、課題終了後、課題の遂行に使われた技能を調べるために生徒と面接する）。

結果妥当性：パフォーマンス評価の評価結果を統一する。
パフォーマンス課題は、学習の結果を正確に反映するものであり、高次の技能やそのプロセスの学習指導を促進する効果を持っている。そのためには、評価の課題で求める基礎的な認知能力を明確にしておかなければならない。

批判：「課題が単にオーセンティックであるからといって、複合的な技能を修得したことを推定できるわけではない」（ウィギンズ、1992）
生徒に高次の思考技能を用いさせようとするが、学習について明確な概念的枠組みの作成を怠っている。場合によっては、
①実際にやっているうちに課題をこなすだけになってしまう。
②機械的に繰り返すだけになっている。
③パフォーマンス評価の課題（課題の遂行方法）を教えてしまう。
↓
教師は、評価される領域や準領域に向けた指導をしなくてはならない。

体系妥当性：カリキュラムや学習指導の改革を引き起こす。
直接、認知能力を評価すること、得点を決める場合に要求される判断の程度が指摘されている。
テストの点数を上げようとする学習指導が、より発展した課題でのパフォーマンスの向上をもたらしたり、課題のコンテクストの中での認知的技能の表現の向上にもなる。（フレダーリクセン、コリンズ）

評価者は採点の範疇を理解し、どのようにそれを用いるのか指導されなければならない。
→ **研修**による専門的な能力の向上と、優れた問題解決や歴史分析の重要な特徴について、メタ認知的な知識を発達させる基盤となる。

パフォーマンス評価の主な特徴は採点における主観性である。パフォーマンス評価の概念や技術的な問題に必ず関係するので、定義に含めるべきもの。

表面的妥当性では不十分。評価の結果をもっと考えるべき。学習指導に与えた有形無形の影響についての証拠を集めて検証すべきである。（リン、ベイカー、ダンバー、1991）

- パフォーマンス評価における公平の問題についても考えなくてはならない。
- パフォーマンス評価は受け入れられているからといって、必ずしも公平であると考えてはいけない。
- パフォーマンス評価で用いられる課題の数はこれまでの評価より少ない。
- 採点の判断のなかに採点者が持っている先入観や偏見が入ってはいけない。

採点者の訓練と結果のモデレーションが重要！
表面的妥当性を超えた分析が必要！

- パフォーマンス評価では１つの領域について少数の課題を使うだけなので、課題そのものが生徒の時間や労力を求めるだけの価値があることを確認しないといけない。
- テストの開発にあたり教科の専門家を加えることも大切。

パフォーマンス評価の弱点

- 評価する内容の範囲がずれることで、学習指導とのずれが生じることがあってはならない。
- パフォーマンス評価は領域から選択的に深く掘り下げて調べるものであり、その課題を実行するために多くの時間を要するため、対象領域を広く調べることを深く調べるために犠牲としなければならない（対象領域が広い方が、評価の結果から広い領域への一般化ができる）。

【パフォーマンス評価の信頼性】

信頼性：テストが意図した技能や達成事項をどの程度正確に測定しているか。
生徒のパフォーマンスの一貫性（再現可能性）と**パフォーマンスの評価の一貫性**（評価の統一性）に分けられる。

伝統的なテスト：妥当性を犠牲にして信頼性を過度の重視

パフォーマンス評価：信頼性を犠牲にして妥当性を過度に重視

パフォーマンス評価とは、標準化しすぎた評価の手続制約から逃れようとする動向の一部である。もし、アカウンタビリティーや資格認定の目的で用いる場合は信頼性の問題は無視できない。

前提：目的への適合性
どんな評価であっても、その特定の目的に応じて、一般に受け入れられるレベルの信頼性と妥当性を必要とする。

パフォーマンス評価の基準は主観的になるから、採点者間の評価の一貫性が注目される。
→　**綿密な訓練と採点説明書があれば、採点者間の一貫性の精度は高められる。**
標準化されたパフォーマンス評価を用いるためには、明確な採点説明書、採点者の訓練、いくつかのレベルやグレードでのパフォーマンスの事例の提供があれば、採点者間の信頼性は高くなる。

パフォーマンスの一貫性
パフォーマンス評価でのパフォーマンスは事例ごとに独自であるがゆえに、同じ領域に関する異なった課題のパフォーマンスや、同じように見える課題のパフォーマンスでも一貫性はそこそこにとどまる。

評価方法によってもパフォーマンスの測定結果は異なる。
評価に用いる課題の数を増やすことが、評価者の数を増やす依も得点の信頼性を高めることとなる。一般化可能性も高めることとなる。（リン、1993）

【一般化可能性】

- ひとつの課題から別の課題への一般化ができない（パフォーマンスは個別の課題によってかなり左右される）。
- パフォーマンス評価が形成的な目的やクラス内での使用以外に利用されるときは深刻な問題になる。

各課題が独自であること、領域から課題をサンプリングするのに限界があるから、パフォーマンスの結果を領域全体に一般化することは困難。一般化可能性を高めるには、課題の数を増やして領域を網羅するようにすべき。

パフォーマンス評価は、アカウンタビリティーや学校改革ではなく、学習の向上のための評価を強調している。（シャベルソン、1992）
→ 試みるなら、新たな枠組みが必要となる。

- 自国語以外の言語の測定に関して、正確性や一貫性といった概念そのものの再興を求める。（スウェイン、1990）
- 全体としての評価の仕組みは、構造化された課題と、構造化されていない自由な解答を許す課題の組み合わせを用いるべき。（メシック、1992）

パフォーマンス評価での課題全般の一般化可能性は低い。思考や学習がコンテクストに強く影響されるという結果と一致するものである。

一般化可能性を高めようとすると評価に必要な時間、評価のための負担が増大する。一般の人々にも受け入れられない。

内的な一貫性による信頼性は一般化可能性の指標としては十分ではない。特定の評価の課題から、これを越える広い範囲での到達状況を一般化することは、相応の根拠を必要とする（これは全ての評価方法にも言えること）。

領域全体を扱う方法としては、格子状計画手法やマトリックス・サンプリングがある。しかし、一般化可能性を求めるならば、多くの時間を要することを認めるべき。（ベイカー、1991）

一般化可能性の程度に影響する２要因：**①課題の類似性の程度②指導の種類**

ハートル（1993）によるパフォーマンス評価での一般化可能性のレベル
① 一つの事例に対して、採点者の間に一貫性のある採点ができるか。
② 同じ課題が、時間と場所を変えてもその意味が変わらないか。
③ 同じ種類の課題で、一般化可能性があるか。
④ 異なる種類の課題の間に、一般化可能性があるか。

採点の明確な説明書と採点者の訓練があれば、①は大した問題じゃない。
②は 3 つの要因に依存する。

- 課題の実施方法（制限時間を設けているか、生徒と教師が話すことを認めているか、生徒間の協働を認めているか、何が課題のやり方を教えることになるかの規則の有無）

- 課題に成功するために必要な補助的な能力の役割
- 評価に先行する指導

③は補助的な能力が課題によって異なるかどうか、また先行する指導にも依存する（ただし、課題の類似性を維持したとしても、２つの課題を同じように機能させることは困難である）。また、パフォーマンス課題のために準備した場合、この課題に対するパフォーマンスを同じ領域の他の課題にまで一般化することはできない。
④はあり得ない。活動内容の相違によってパフォーマンスは異なる。

新しい評価方法：その目的は、ある種の複雑な達成事項を認めること。
解釈論的な方法を採用→ 主体や使用者の一般化可能性に関する解釈を優先。
パフォーマンス評価では、参加者自身の経験の概念化や再構成のあり方、成果の見方に注目する（モス、1992）。大事な点は、

① パフォーマンス評価の特徴（妥当性や高次の技能を重視すること）を後退させてまで、パフォーマンス評価を標準化しないこと
② 課題の評価していることが重要であることをはっきりさせること

【パフォーマンス評価の使用に関する問題】

アメリカ
パフォーマンス課題の多くは診断的な目的や指導のための教室での評価として開発されたが、現在では大規模なパフォーマンス評価が多く行われている。

概念的で全体的な（ホリスティックな）指導や学習を促進する方向へ、指導を変えていこうとする明確な意図に基づいている。
例）職業適正ポートフォリオ、新スタンダード計画、全国試験システム

パフォーマンス評価をアカウンタビリティーの目的で使うと、すぐにハイ・ステイクスなものになり、技術的な内容がかわってくる。また、行政的に実行可能で、専門的立場から信頼され、一般にも受け入れられ、法的に容認でき、経済的に可能であることが必要。

ハイ・ステイクスな目的で使うなら評価の統一性が必要になるので、以下の多くのことを行うべき。（ベイカー、オニール、リン、1991）p.151

- 具体的な規定を要するもの
- 評価の調整

- モデレーション
- 研修
- 確認手続や監視手続

課題を採点するために、人間を使うために必要な余分な労力と費用は歓迎されるべき。妥当性の高いデータが得られるとともに、教師を関与させることで専門的な技能の発達の機会ともなる。←この波及効果は強調したい。

ただし、パフォーマンス評価を可能とするためには、すべての学年で、すべての生徒を、すべての科目にわたって実施するという考えは放棄すべき。（シェパード、1992）

イギリスの事例

イングランド、ウエールズ、北アイルランドでは、16 才の生徒に対して実施される中等教育修了資格試験（GCSE 試験）とアカウンタビリティーのための試験では、**全国的なレベルでパフォーマンス評価を導入している。**

（GCSE 試験：数学や科学での探究的な活動、英語と外国語での口述試験、ポートフォリオの作成、歴史や地理、経済、ビジネス科目などでのプロジェクト研究。ペーパー試験は択一式の問題はほとんど使っておらず、解凍はエッセイ形式か、短文方式（アメリカ的な用語法ではこれもパフォーマンス評価）である）

一定の領域をカバーしたり、評価や採点の一貫性を確保するために、生徒、教師、試験官がかなりの時間を掛ける必要がある（英国で可能なのは 1 つの年齢だけで実施されていること。ナショナル・カリキュラムでは困難であることがわかっている）。

コースワークは教師や学校から評判がいい：実際的な活動、口頭表現の重視、学習の深化など、学習指導に対して良い影響がある。（ギップス、ストバート、1990）

一般からの反応に苦しめられる GCSE 試験

- 教師の関与は不適当である。
- 教育格差がもたらす家庭学習の不公平さ

→　多くのコースワークは学校内で一定の条件下で行い、成果物が生徒自身によって作成されたことを教師が保障する。

パフォーマンスに基づく評価を実施している GCSE 試験は、学習指導や生徒の

学習のあり方に望ましい影響を与えている。しかも運営可能で、現場で価値も認められている。しかし、反対派の政治家と衝突し、コースワークの比重が下げられた。

【７才でのナショナル・カリキュラムの試験】

- イングランドとウエールズのナショナル･カリキュラムにおける７才児にチアする評価。コア科目（英数理）について４段階で各生徒の達成レベルを評価する。
- 教師は自分が望ましいと思われる方法で評価して良いが、観察、日常的な評価、作品の事例を保存することが奨励されている。
- 夏学期前半と春学期後半に、達成評価課題（SATs）から生徒の達成状況を調べる。
- SATs はオーセンティックであること、直接性、認知活動の複合性、主観的な採点などから、パフォーマンス評価と言える。

- SATs が抱える問題

時間がかかる→もっと短い標準化されたペーパーテストに移行→活発な活動が減少し、別の方法が用意され、追加の課題が増える。→評価時間が増え、教師の不安が大きくなり、ほとんど使われない。

問題は、現実的な学校の状況にある。

実施の労力が大きすぎるうえに、評価全体の実施は実行不可能。かなり大きな学校全体の組織改革が求められ、混乱が生じる。教師は困難でストレスが多く、時間がかかる仕事に忙殺される。

さらに評価への関心の高さが、ストレスの原因となる。

- レッテル貼りの評価を実施する危惧
- 低学年の子どもを評価、類別する不適当さ（診断的な目的で行うべき）
- （４月生まれと３月生まれの差のような）教育期間の相違
- 学校入学前の教育の内容
- 家族の違い
- 社会的な背景

1992 年の２回目の SATs は、１回目の経験からストレスは軽減されたが、信頼性と実施可能性は問題として残っている。

【評価の実施から得られた信頼性、妥当性に関する教訓】

標準化されていないことにより、実施のあり方にばらつきがあった。

背景：SATs で重視していたことは、何が求められているかを生徒が理解すること。そのためには大人たちの手助けや説明に関する制約もなかったが、生徒間で課題を説明することは認められていなかったし、非英語母語話者の子どもたちへの母語の説明も許されなかった。教師間だけでなく、同じ教師による実施でも対応にばらつきが出る。

ばらつきの原因

- 小グループの編成方法（生徒のパフォーマンスに影響）
- 課題で使用する物の組み合わせの選択（パフォーマンスの統一性に影響）
- 実践課題かワークシートを用いた課題か（課題の性質、評価方法に影響）
- 不明瞭な評価基準（評価方法とスタンダートとなるパフォーマンスを議論れば評価の統一は可能）

２年目には教師の評価と SATs や別のテストの結果が一致する度合いが極めて大きくなる→ これは人為的に作り出された部分がある。

- 教師の評価が SATs に影響された
- SATs の結果を教師の評価とした教師がいた。

教師の評価と SATs の結果の不一致：

「２つの形式の評価を同じものとみなしてはならない」（SEAC、1991）

教師は長い期間にわたる、より広い達成事項を対象としている。SATs より信頼性は低いかもしれないが、教師の評価は全般的にわたっていて、達成事項についての全体的な状況を示すものである。

レベル到達の決定の判断について規則の有無に違いがあった（SATs は全てを満たす必要があったが、教師の評価にはそういう規則がなかった）

観察者が訓練をうけ、採点の説明書が与えられれば、パフォーマンスに基づく課題についての採点者間の統一性は高い。（シャベルソン他、1992）

パフォーマンスのスタンダードについて議論する機会があれば、評価基準の共通理解が深まったというデータがある。（ジェイムズ、コナー、1993、NFER/BGC、

1992）
教師たちが集まり、モデレーターや外部の専門家が加わり、生徒の作品を材料としてパフォーマンスの評価について議論する。→ 評価の統一性に貢献
実際、ペーパー試験形式よりも、採点者間の評価の統一性が高いことが報告されている。（ブラウン、1992）

SATs の形式は、課題の内容妥当性と構成妥当性を高めること、学習指導に有害な結果をもたらさないことを期待していた。評価の結果が学校や地域のレベルでアカウンタビリティーの目的に用いられ、個人のレベルでも用いられる（クラス編成、選抜）ことを考えれば、SATs の意図は歓迎された。

- 批判：
  - 評価されるべき内容と構成概念を教師が妥当性をもって評価していない場合がある。（ジェイムスとコナーの事例研究、1993、作文の例：ピリオドや大文字の使用が重視されるが、書かれた文章の思考力についての判断が求められていない）。← これはテスト開発者の問題
  - 生徒に与えられた資料の質を高めるように。

評価に関わる問題点について教師の理解が進む。
ペーパー上の課題を用いずに実践的な課題を使った教師の多くは、ペーパー上の課題の妥当性に疑問を持っていたためである。

- SATs によるパフォーマンス評価を利用して妥当性を高める試みは、信頼性については妥協的なものになった。

妥当性を高めようとすると信頼性が損なわれる。
パフォーマンス評価は妥当性は高いが信頼性は低く、一方標準化されたテストは信頼性は高いが、妥当性は低いという評価に至る。
  - 読解力のテストの場合、子どもたちは選択できる本のうち、馴染みのある本を選ぶ。
  - 教師の観察による評価の場合、課題が同じでなく、難易度も異なるので、すべての教師が同じような方法で、同じようにレベルを判断していると確信できない。

- 結果妥当性についてはいくつかのレベルで検討が必要

子ども個人レベル：一組の評価基準でパフォーマンスを綿密に観察することは困難である。

教師レベル：綿密な観察と詳しい評価によって実践に有益な影響を与えた。テスト以外でも、カリキュラムに対して注意深くなり、子どもが何ができるかを注意深く観察するようになり、子どもの可能性に目を開くようになる。
指導レベル：基本の強調と実践的な数学や科学の活動へ学習指導を広げる効果を上げた。グループ活動や自主的な活動の導入をもたらした。

当初は学校のレベルを把握して、ランキング表を作ることを構想していた。ただし、道徳的、技術的な観点から批判をうけることとなる。厳密性にも欠ける。パフォーマンス評価についてはまだまだ慎重に扱わないといけない。

【結論】

アカウンタビリティー目的にパフォーマンス評価を用いることには問題がある。
複合的で時間がかかるパフォーマンス評価は、これまで用いてきた標準テストに比べて劣っていることとなる。これまでとは異なる妥当性の基準が必要。

パフォーマンス評価においてモデレーション（チェック機能）は重要。
→ 採点者間の評価の一貫性を向上させる
→ 評価過程の統一が可能となる。

標準化せずに、評価から導き出される結論を正当化しようとするならば、**基準となるパフォーマンスと最も優れたパフォーマンスを引き出す状況と活動について、教師が共通の理解を持つようにしなければならない。**
再現可能性や一般化可能性ではなく、**パフォーマンスの質や採点の公平さを考慮すべきである。**

評価の議論から、適切な課題の考案の問題、そして学習指導に対しても影響を与える。
モデレーションの欠点：多くの時間がかかるため、実行可能性を損なう。
モデレーションの利点：教師の学習指導の向上に効果がある。

低学年の子どもを対象とする場合は事情が異なる。子どもに何が適切かを考慮して、異なった評価形式が必要となる。標準化されたパフォーマンス評価を実施することは不可能。→ 一般化可能性の問題
パフォーマンス評価の有利な点
広い範囲の技能を様々な方法で評価するための強力な手段である。