第 1 章（評価のパラダイムの転換）～第 3 章（テストのもたらす弊害）までの簡単なまとめ

・精神測定学的視点の問題点を指摘
　ノルム準拠テスト（偏差値、得点による順位付け、相対評価）
　信頼性を重視しすぎてテストの形式や内容を制限してしまっている
　　⇓　パラダイムの転換
　構成主義的な学習モデル
　クライテリオン準拠評価（到達度評価、絶対評価）
　パフォーマンス評価の導入：学習指導や学習活動につながる評価の開発

第 4 章 妥当性と信頼性

妥当性（validity）　（p.81）

妥当性の伝統的な定義：テストによる測定が、どの程度正確であるかというもの
初期の研究書による 4 つの種類の妥当性
- 予測妥当性：テストが将来の学習の状況を正確に予測できるか
  （例：共通テストの得点が大学入学後の学習状況を予測できているか）
- 併存的妥当性：テストが同じような技能を測定する別のテストと相関しているか、または実質的に同じ結果をもたらすか
- 構成妥当性：テストが構成概念、すなわち評価されるものの基礎にある（説明となる）技能を適切に測定しているか
- 内容妥当性：適切に必要な内容を把握しているか。
  （例：定期テストをある一定の範囲とした場合、テスト問題がその範囲を反映した内容となっているか）

➡予測妥当性と併存的妥当性は基準妥当性と言われ、ある基準と比較したパフォーマンスの予想

➡近年の妥当性研究書は、構成妥当性を統一的なテーマとする統合概念であることを強調している（メシック, 1989a；クロンバック, 1988）

統合概念としての妥当性　（p.83）

妥当性に関する認識の変化：メシック（1989a）

> 妥当性は統合的概念であり、その意味は、構成妥当性に体現される得点の意味が、得点に基づくすべての推測の基礎となっていることを意味する。しかし、妥当性について十全で統合された見方をするためには、得点に基く推測の適切性や有意味性、有効性がテストのもたらす社会的な結果にも大きく依存することを認識しなければならない。そのため、妥当性の考察において社会的な価値を無視することはできない。

[ここに入力]

➡テストの解釈やテストの使用を支持する証拠がどれだけあるかに結び付けて妥当性を考えている。

妥当性に対する 2 つの強迫観念を軽減する証拠や議論を集める必要がある

①構成概念非代表性（construct underrepresentation）

テストで測定している内容が、対象となる構成概念（知識、スキル、能力など）の代表となっていないこと。

②構成概念不適合性（construct irrelevent variance）

テストで測定している内容が、対象となる構成概念に適合していないこと。

例）ある教科の知識についてテストで文章を読ませて理解することを求めるとき、読むことの不得意な生徒にとっては、（構成）妥当性は低くなる。

➡対象となるすべての構成概念を具体的に示すことをメシックは提唱

➡メシック（1989a, 1992）とクロンバック（1980,1988）は妥当性に関する議論をテストの機能的な価値を超えた問題とみている。つまり、構成妥当性はテストの解釈を正当化するだけではなく、テストの使用をも正当化するために必要であるとする。

妥当性とテストの使用の帰結　（p.87）

結果妥当性：テストの使用とその解釈によってもたらされる結果の考察

例）多肢選択式の共通テストがもたらす高校での指導やカリキュラムへの影響

メシック（1989a）による結果妥当性に関する 2 つの問題提起

・テストは私たちが評価しようとしている特徴を測定する手段として適切か

・テストの結果を提案された目的に使うべきであるかどうか

➡テストそのものだけではなく、テストのそれぞれの使用法の妥当性。

➡私たちはテストの解釈や使用によって起こりうる結果や、実際に生じた結果が意図した目的に役立つかどうかだけでなく、他の社会的価値と両立するかどうかについても問わなければならない。

妥当性の側面（メシックによる表）

	テストの解釈	テストの使用
証拠の基準	構成妥当性	構成妥当性＋関連性／有用性
結果の基準	価値的な意味	社会的な結果

例）テストの女子の点数が低い（テストの社会的な結果が有害な例）

・テストが（構成）妥当性を欠如している

・評価される構成概念を正しく反映している

[ここに入力]

メシックとクロンバックによるテストの結果と妥当性の関係に関する異なった視点

メシック：構成妥当性が確認されるならば、有害な影響それ自身がテストを妥当性のないものとするとは見ていない。

クロンバック：有害な社会的な結果それ自身が、テストの使用の妥当性に対して疑問をなげかけるものである。

➡結果妥当性はどちらかといえば得点の解釈よりも、その使用のほうに直接的に関係している。

➡アメリカで現在パフォーマンス評価が強調されているのは、学習指導に特定の結果をもたらそうとの願いからである。つまり、実際的な活動や問題解決学習、高次の技能の育成を目指した学習指導を促進しようとしているのである。

➡パフォーマンス評価の開発にあたって、指導や学習に対する意図した結果や、意図せざる結果についての証拠を収集する必要があるだけでなく、指導や学習に対してどのような影響を与えようとしているのかをもっと明確にしておかなければならない。

体系的妥当性（結果妥当性の特別な形式）

「体系的に妥当なテストとは、テストが測定しようとしている認知技能の発達を引き起こすようなカリキュラムや学習指導の変化を、教育システムに引き起こすものである」（フレダーリクセンとコリンズ, 1989）

個人の結果―クラス、学校、地区での結果の相違に関するもう一つの妥当性の問題（リン）

・私たちはテストの点数の解釈や使用を個々の生徒について適切だと結論づけても、それらを合計したレベルではその妥当性を再考しなければならない。

・テストを使う目的がレベルの異なる使用者によって大きく異なる。

➡テストの結果の使用をコントロールするのは不可能であるし、現実的でもない。

ティトル（1989）による妥当性に関する検討

テストの推測や使用の妥当性　⇔ テストの得点（教師の指導や地域の状況が影響）

「このように妥当性は 2 方向へ相反する理論である。一方では普遍性に基盤を置こうとする。また一方で、私たちが妥当であってほしいと願うテストの得点は、普遍性を持ちえないのである。」

➡妥当性に関する研究をさらに進めるには、指導や学習のための教育評価において、評価する構成概念の選択基準と同様に、クラスのモデルや教師の教育目標が具体的に示されていなければならない。（ティトル）

[ここに入力]

信頼性（reliability）　(p.94)

評価全体の妥当性を高めるために、パフォーマンス評価やオーセンティック評価が考案された。一方で、これはしばしば妥当性と対立する信頼性において問題を引き起こす

信頼性：テストが測定しようとしている技能や達成事項をどの程度正確に測定しているかを問うことである。「生徒のパフォーマンスの一貫性」(再現可能性) と「その評価の一貫性」（統一性）に関係する。

テストの信頼性を調べる方法

・テストー再テスト法：同じテストを数日おいて再び実施
・並立テスト法：同じテストを異なった形式にして同じ母集団に実施し比較
・テスト折半法：テストを無作為に 2 つに分割し、半分どうしの得点が一致する程度を調査（内的一貫性）
・テストの可能なすべての分割を考える方法：折半法を拡張し、すべての可能な相関関係の平均を求める統計的手法を用いて、内的一貫性の係数を求める

テストの誤差の原因

①測定手続きに内在する変動、②特定の課題の選択にともなう変化、③日によって異なる各人の状態、④課題をやり遂げる各人の速度の違い

クライテリオン準拠評価は個人間の差異を際立たせるために作られていないため、伝統的な信頼性の概念が不適当。測定の一貫性を評価する異なった方法を必要とする。

パフォーマンス評価では、「採点のスタンダードの一貫性」とともに、「評価の課題に対する取り組みの一貫性」を考慮する必要がある。「採点のスタンダードの一貫性」とは、異なった採点者の評価基準の解釈を同じにすること、「評価の課題に対する取り組みの一貫性」とは、課題の実施に関する。

採点の信頼性　(p.96)

採点の信頼性を調べる方法

・評価者間信頼性：異なった採点者が同じ課題の解答を採点する
・評価者内信頼性：同じ評価者が同じ課題の解答を異なった時に採点する

➡択一式や穴埋め式など採点が単純ならどちらも一致の程度は高い。作文方式やパフォーマンスに基づく評価の課題では、採点計画の複合性と判断の主観性のために採点の違いが大きくなる。

➡採点者は生徒の名前からジェンダーや人種など推測でき、採点に影響することがある。

[ここに入力]

伝統的な信頼性に関する主だった 3 つの批判
①達成したことを正確に測定することが可能であるという仮定をしていること。
②再現可能性を確保するために、精神測定のテストに固有の条件として、テストを非常に標準化された条件で実施する必要がある。
③内的な一貫性を高めるために、テストの問題は同質であることを必要とし、そのため単一の技能や属性を評価することになる。
➡精神測定とは違った信頼性の検証方法を見ておかなければならない：評価過程の統一と評価結果の統一

評価の一貫性の確保　（p.99）
評価過程の統一：評価の実施過程の標準化や一貫性を確保する方法
評価結果の統一：評価した結果が統一されるよう比較調整し、評価の一貫性を確保することに焦点を当てている。
➡評価の一貫性を確保するための方法：モデレーション

モデレーションの方法　（p.100）
＜参照テストやスケーリングの方法を用いた統計的モデレーション＞
共通の参照テストの得点を用いることで、教師の判断の全体的な偏りを是正する
例）オーストラリア学習適性テスト
スケーリング：科目間の評価の統一性を調べるために用いられる統計的な方法
➡【問題点】スケーリングの過程で、一定の選択パターンが男の生徒に有利に働く。参照テスト自体が一定のグループに対して有利で偏っている可能性をもつ。
＜査察に基づくモデレーション＞
グレードや資格認定の付与に責任を持つ機関が、私見の答案や評価されたレポートをサンプリング調査する。課題の内容が求められた条件を満たしているか、また、指示された通り採点され、グレードを与えられているかを確かめるため。
＜再調査委員会＞
グレードの一致の程度や、特定の課題の等級づけに焦点を当てて再調査が行われる：評価結果の統一。
＜コンセンサス・モデレーション＞
教師の評価に対して、教師や専門家、モデレーション担当者などのグループや委員会でコンセンサスを取る。
＜グループ・モデレーション＞
評価基準について共通の理解に達するために教師や指導者が生徒の課題の事例を使って討議する：評価過程の統一と評価結果の統一両方を検討。
＜機関レベルの承認＞

特定の資格などの付与の責任を有する組織が、ある機関や学校に適切な教育課程を提供し、かつ関連する評価を実施しているとして承認するもの。

＜本質的モデレーション＞

経験を積んだ教師が試験官として、シラバスの共通目標を用いて、試験を作成し、チーム方式で生徒の答案を採点する：評価過程の統一。

結論　(p.105)

・伝統的には妥当性は信頼性よりも重要だと考えられている。しかしながら、テスト開発の場面では信頼性を強調しがちで、高い正確性と再現可能性をもったテストを作るために、妥当性はないがしろにされてきた。
・パフォーマンス評価や学校内での教師の評価への動きは、信頼性と妥当性の均衡を回復する試みのひとつである。
・妥当性と信頼性という相反する関係に対処する方法として、高い水準の妥当性（内容、構成妥当性）を確保しながら、テストの目的に応じた最大限の信頼性をもつことがテストの質である。(ハーレン, 1994)

一般化可能性（generalizability）:妥当性と信頼性を結合させる概念

評価とは行動の一部として取り出したサンプルをもとに、行動の全体を一般化して推測すること

例）読みのテストの結果は、一般的に読みの能力を示すものである

➡行動の全体が注意深く規定されていること（構成妥当性)、評価そのものが信頼性を有すること

一般化可能性に対する主要な難問

・コンテクスト

得点が特定のコンテクストに限られるものか、一般化できるものか問う必要がある。

評価が現実的なもの、パフォーマンスに基づくようになると、他のコンテクストへの一般化可能性が正当化できなくなる。

➡テストの適用可能なコンテクストの範囲を特定しておくこと、そしてテストの使用や得点の解釈をこの範囲に限定すべきこと

➡パフォーマンス評価においては課題の数を増やすことが一般化可能性を高める最も効果的な方法である（リン, 1993a）

➡抽出の概念や一般化可能性そのものを放棄し、意味のあり、よく説明されたコンテクストでの最もすぐれたパフォーマンスを引き出す。

信頼性の概念そのものも拡張されるべきである（リンたち,1991）

[ここに入力]

参考資料　鈴木秀幸(2021)『これだけはおさえたい学習評価入門　「深い学び」をどう評価するか』図書文化社