

Part 7 Assessment and Testing

33 The Practice of Language Assessment

Glenn Fulcher

<発表の構成>

- I. 著者情報
- II. 本論文の構成
- III. 本論文の要約
- IV. 考察点
- V. 参考文献

【おねがい】

このレジュメは、東京学芸大学大学院国語教育専攻日本語教育コースの授業「日本語教育方法論演習」（授業担当者：南浦涼介）での大学院生で取り扱った、Hinkel, E. (Ed.) 2017. *Handbook of Research in Second Language Teaching and Learning*. の発表のレジュメです。
 教育的価値、資料的価値としてウェブでの掲載を行っておりますが、いわゆる「論文」ではありませんので、論文等への引用や掲載は固くお断りいたします。また、分析対象の著作権は著作者、資料文書の著作権は発表者に記しますので、無断転載はご遠慮ください。質問については、東京学芸大学南浦研究室 (<http://minamiura-lab.com/>) までお願いいたします。

I. 著者情報

氏名 Glenn Fulcher

現職 レスター大学（イギリス） 英文学科 応用言語学/言語評価学教授

著作 Sawaki, Y. (2012). Technology in language assessment. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 426-437). New York: Routledge.

Fulcher, G. (2015). *Re-examining Language Testing. A Philosophical and Social Inquiry*. London and New York: Routledge.

II. 本論文の構成

1 Introduction	1 導入
2 The Language Test Design Cycle	2 言語テストのデザインサイクル
Activity1: From Test Purpose to Test Tasks	活動1：テストの目的からタスクまで
Activity2: Specification Writing and Iteration	活動2：仕様の詳述と反復
Activity3: Prototyping and Piloting	活動3：プロトタイプとパイロット
Activity4: The Field Test and the Go/No-go Decision	活動4：現場テストと実施の決定
3 Trends and Controversies	3 傾向と論争
4 Perspectives	4 概観
Validation Processes	検証プロセス
Substantive Validation	本質的検証
Social Responsibility and Accountability	社会の反応と責任
5 Conclusion and Future Directions	5 結論と今後の展望
6 Acknowledgements	6 謝辞
7 References	7 参考文献

- ☞ (1) 言語を評価するとはどういうことか。
- (2) 言語能力を測るテストとは何か。どのような形式で提示されるべきか。
- (3) テストの汎用性をどう考えるべきか。
- (4) テストや評価において、それぞれの文化的背景や個人の知識量をどう捉えるか。

『言語評価の実践』

◎本論文の解釈にあたって 〈 Assessment について 〉

Assessment の目的 (Cohen 1994 参考)

General purpose	Specific purpose
Administrative 学校運営に関わる機能	全体の評価
	クラス／レベル分け
	履修の免除
	卒業／修了
	昇進
Instructional 教育に関わる機能	(各学習者における言語力の) 診断
	上達の度合い／伸び
	学習者へのフィードバック
	教え方／カリキュラムの評価
Research 研究に関わる機能	評価・検証
	実験
	言語学習や言語使用に関する知識

Ⅲ. 要約

1 Introduction 【導入 テストと評価というものの位置づけ】

言語学者と応用言語学者は、言語テストと評価に常に従事している。非公式であろうと公式であろうと、評価手順は（テスト受験者の）意思決定のための証拠となるように設計されている。

- ①low-stakes の教室では、学習者に FB を与え、足場学習、学習の質とペースを評価し、教育哲学を学び、カリキュラム設計に取り入れる。
- ②high-stakes では、識別可能なテストイベントで学習者の将来の人生を左右する決定を下すために使用されるスコアが生成される。
- ③研究では、SLA 習得に関わる問題提示のために使われるデータの生成を目的としてテストが使われる。

これらの文脈でそれぞれのテストや評価を使用する際には、いくつかの一般的な仮定をしている。

1. 実施されるタスクは、行われる決定または研究課題に関連する。
2. タスクによって誘発される反応は、決定を下すことまたは質問に対処するのに有用である。
3. スコアは応答を適切に要約している。
4. スコアに応じて必要性を持つが、必ずしも十分な決定の証拠となるとは限らない。

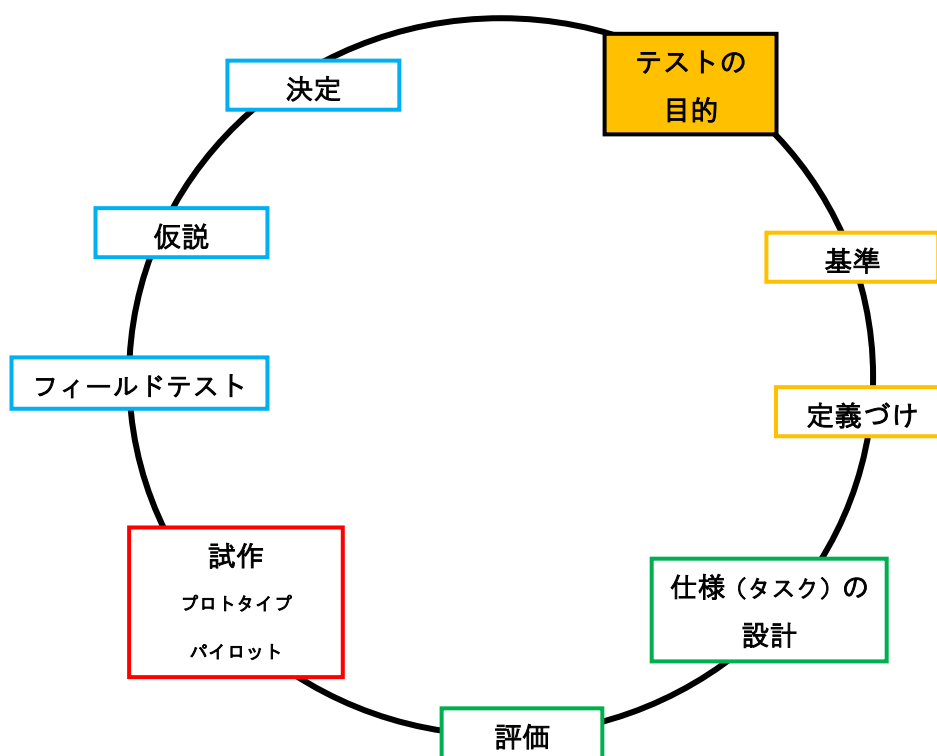
すべての文脈で仮定していることは、**テストと評価の慣行を慎重にする必要があるという主張につながる**。慎重な考え、計画、文書化が必要である。これは学習のためのインフォーマルな教室評価¹であっても、学習者が現在できることと次のゴールとのギャップを認識する手助けとなる FB を形成することが第一の目的である。

¹ informal とは教師の記述する評価についてのメモやノートのことをいう

➡学習評価の計画は、学習者間の相互作用を含むタスクの設計に焦点を当て、教室管理は学習者が FB を反映し、それに応答するための時間を提供するべきである。意図的に計画を立てるには、実践を導く原則と概念だけでなく評価構築における実践スキルを含む評価リテラシーのレベルが必要である。

2 The Language Test Design Cycle 【言語テストのデザインサイクル】

設計者が他の場所で広く採用されているメタファーに従って、テストを構築するプロセスを用いる。建築家が行なうように初めに建物の目的を検討し、目的がわかれば、その用途を考慮することができる。



■Activity1 (テストの目的からタスクまで)

①「このテスト／評価は何のためか」

「生成された証拠に基づいてどのような決定を下すべきなのか」

すべてのテストは目的のためのもの (Ingram1968, p.70) であり、それが何であるかの明確なアイデアなしに構成されたテストはうまくいかないということを説明している。つまり、サイクル上の「決定」によって「目的」を導くことができる。

②「テスト基準は何か」

たとえば、英語学習者の教育環境内の共通の基準は、L2 言語設定での数学についての学習を指す。この場合、基準は学校の学術コンテンツにアクセスするために必要な数学的な英語の領域になる。

③「学習者が目標を達成するために、何の知識、技能、能力 (KSAs) を目指すか」

言語テストでは、KSA はしばしば要素と呼ばれる。構造図を評価するために、使用される可能性のあるタスクタイプの指示には図が頻繁に提供される。アメリカにおける英語学習者の共通基準は、「知識とアイデアの統合」を一つの構成要素に挙げている。

一般的な記述レベルで推奨されるタスクタイプの一つについての説明

テキスト内の言葉で表現された量的または技術的情報を視覚的または数学的に表現された情報に変換する。学習者がテキストと表／図／方程式との間で情報を翻訳できれば、彼らは「知識とアイデアを統合する」という証拠を提供していることを主張できる。

p466【タスク】

学習者に、学校旅行のオプションに関するデータが数値と表で提示されている。問題は「どの旅行がクラスにとって最も適しているか」の判断のために、学習者がデータを操作する必要がある。その判断と理由が書面で説明されることによって、回答者が「この領域内の知識とアイデアを統合できるか」（そしてどの程度まで）推論するかを推測する証拠となる。テストプロセスは、推論が妥当で正しい判断であるかを決定づけるのに有効である。

p467【テストの目的を明示することについて】

1) 2) は評価の目的を明確に示し、テスト基準を指定するためのコンテンツの種類も示している。現実世界の活動には、1) 仕事の読み込み、社会的な情報のリーフレットの閲覧、2) ビジネスミーティングへの参加、契約の交渉などがある。それぞれの目的に必要な言語は非常に異なる可能性が高く、ある目的のために作成されたタスクを正当化するために作成された検証の説明は、容易に別の目的に翻訳される可能性は低い。

■Activity 2 (仕様の詳述と反復)

◎建築における青写真＝テストと評価においては test spec（テスト仕様）がそれにあたる

明示的なテスト仕様を描くことは、設計と開発プロセスの慎重さを強制する。テスト仕様が定まれば、フォームが作成でき、そのフォームは別のテストに応用できる。つまり、同じことを、同じ難易度・異なる内容で作成することが可能となる。試験受験者が異なる時間に試験を受ける場合、または試験前や試験後の研究設定で学習を評価するために試験が実施されている場合にも必要となる。

1)Evidence specification どんない印象にするか

closed response か opened response かなど、問題の問い方をどのようにしてどう答えさせるか。また、そのスコアの意味は何かを考慮するということ。（→見出しの Assumptions 3、4 に関連）

※課題あり…評価には個々の判断に差が出る。現状では評価基準のルールブックを提示する。

2)Assembly specification どんな問題をどのくらい扱うか

テスト作成者は、完全なフォームを構成するために必要な各タイプのアイテムの数を示す。テストの各形式で出題数を量的に均質することを保証するような仕様にするということ。

3)Delivery specification どう実践するか

テストの媒体（紙／PC）、人数、指示の方法、補助の有無などを明確にする。

4)Presentation specification 視覚的な適切さ

フォントサイズ²、字体、アイコン、色合いなどについてテストの現物に関する検討。障害のある受験者（視覚障害）への影響を考慮しなければならない。アイコンやナビゲーションを不適切に使用すると、タスクへの応答よりもテストの時間を要する。

仕様プロセスは、テストの目的が確立されるとすぐに開始される。目的、基準、関心領域、および要素の記述はすべて仕様の一部であり、各タスクタイプは繰り返し実行される。テストデザイナーが構造の理解を洗練し、テスト手順を通じて有益な証拠を得るための最良の方法として、タスクタイプの変更と改定が求められる。これはプロトタイプの課程で最も顕著に起こる。

■Activity 3 (プロトタイプとパイロット)

〈プロトタイプ〉

作成にあたり、通常は少数のテスト受験者（約 5 人）に新しいタスクを試す。彼らはタスク実行中または完了直後に、自らがどのように結果を導いたかを口頭で、直面した問題を説明するよう求められる。プロトタイピングは分析のために 2 つのタイプのデータを生成する。①タスクへの応答、②参加者がどのようにタスクに応答したかの言語化である。データ調査により、①作成者の意図と合っているか、②出題項目・形式が対象の学習者の難易度にあっているかを検証する。(Nissan & Schedl, 2012, p.283)

タスクの大部分の問題は、このプロセス 2～3 回の反復だけで発見され、識別された問題を取り除くために各反復でタスク（および仕様）が変更される。障害のある学習者を考慮する場合は、障害の影響を緩和するための調整が同時に設計され、プロトタイプ化されることが不可欠である。また、少数の参加者でプロトタイピングを行うことができるという事実は、正式なテストだけでなく、学校での評価タスクの使用のための予行としても実行可能であることを意味する。

プロトタイピングは基本的な研究であり、テストデザイナーまたは教師は、タスクの使用と応答の意味について批判的な妥当性の検証ができる問いをすべきである。教師が協力して課題の仕様を洗練する文脈で、プロトタイプが学校で実施されるとき、この研究は学習目標の共通理解と構築物の実用的意味の作成と進化を可能にする。

〈パイロット〉

テストのための統計データを生成する必要がある場合にのみ実行される。つまり、テストの運営が成立するかどうかの検証である。このデータにより研究者は、あまりにも困難または容易すぎるタスクタイプを特定することができる。

■Activity 4 (フィールドテストと実施の決定)

high-stakes テストでは、設計プロセスのすべての反復が完了したときに、少なくとも 2 つの完全な形式のテストが、多数の参加者サンプルに与えられる。この段階でテストの基準が確立し、管理手順とセキ

² テストがコンピュータによって提供された場合にフォントサイズが変更されるかどうかも含む

ユリティ機能がすべて正常に動作することを確認するためにテストされる。フィールドテストが終了した時点で、テスト開発者は意図された目的のために現実の評価にテストを使用できるかどうかの判断をし、情報に基づく決定ができる。(これは、テスト開始はプロセスの最後ではないことを表している。)

➡ 試験集団は時間とともに変化する可能性があるため、試験結果の絶え間ない監視が不可欠である。適用される言語研究は構築物の定義およびコミュニケーションの理解を向上させるので、新しいタスクタイプにつながる構築物の新しい操作を生成することが可能になる。時間が経つにつれて、いくつかのタスクタイプが廃止されたり新しいタイプが導入されたりすることで、テストが向上していく。

3 Trends and Controversies 【傾向と論争】

近年、High-stakes テストの最も重要な傾向に「再利用」が挙げられる。もともとは設計されていなかった第二の目的のためにテストが使用された。政策の文脈で言語テストを使用する傾向は、教育の商業化と移民政策の成長に支えられて制御されていない再利用をもたらした。従来と違う効果を測る場合には、測定法が検証されていないため正当化が必要である(AERA et al.,1999, p.18)。

移民政策のための言語テストは、現在世界中で普及している(Kunnan,2012)。例えば IELTS は「新しい目的のための適切性を確立するために適切に再検証されるべきだった」が、NZ では政府の費用の欠如によって試験設計ができず、実施されたという例がある。

さらに CEFR の使用が問題を複雑化させている。既存のテストを CEFR レベルにマッピングすることによって、どのテストも有効であると主張することができる。同様に、健康専門職への入学で AE のテストを使用したり、一般的なビジネス英語のテストを使用して軍事要員の準備導入に関する決定を下すなどのような例もある。

これは、現代の言語テストの実践において深刻な論争をもたらしている。IELTS は特定の分野での移住や就労を目的としたものではなく、英語圏の学界や訓練の分野で勉強しようとする候補者の英語力を測定するためのものであるし、TOEIC は一般的な職場英語のテストとして設計されており、軍のコミュニケーションの準備についての決定をしない。

これに対し、Ahern(2009)はテストと教育の商品化、Shohamy,McNamara(2009)はテスト使用の政治化であると批判している。Fulcher(2016)は、「検証は政治機関による認知として再解釈された」として、検証理論の崩壊を述べている。一方で、テスト所有者が資金を提供する研究においては、市場シェアを維持するためには政策立案者のニーズに敏感であることを単に示唆している。

4 Perspectives 【概観／観点】

Validation Processes (検証プロセス)

この章では、スコアの意味と本来の目的のために設計された試験手順に対する推論を制限する慣習から派生された議論について触れる。この視点に求められるのは、目的の理由とデザインサイクルを取り巻く検証の語りの解釈である。検証の議論において、点数の定説というものを正当化することができる。例えば TOEFL ibt などでは明確にこれが提示されており、二つの構成要素を含んでいる。

①確認調査（Confirmatory research：仮説検証）

テストサイクルの初期段階で、決定の設計が試験目的に有用であることを確認する。

②非確認調査（Non-confirmatory research：非仮説検証）

パイロティングやフィールドテストが実施されたのちに、ライブテストを行うが、これにより、**試験の目的と意図された評価構成に関係のないあらゆる要素によって、スコアが影響を受ける**という弱点が主張される³。例えば、受験者が想定している言語使用の領域とはほとんど関係性がない試験内容が疑われている場合、これらの基準を元に試験課題、トピック、言語使用、手順の比較研究が行われる。仮に、テスト仕様に列挙されたコンテンツが基準を満たしていない場合、試験の有効性に対して疑問を投げかける根拠があるということである。

Substantive Validation（本質的検証）

試験作成者が試験問題を作成する際、受験者が適切な答えを選択することと特定の方法で実行することを前提としている。プロトタイプの合理性は、これが本物の受験者に当てはまるかを検証することにある。本質的な検証の根底には「受験者が試験問題に対して試験作成者の意図したとおりに答えるか」である。方法論として、同時プロトコル解析または遡及プロトコル解析⁴（これらを略して PA 研究）が挙げられる。

作成手順時のこれらの研究の目的は、試験作成者の想定に挑戦する非仮説検証を構築することである。先行文献では、最も教育的な PA 研究の一つに Buck（1991）があり、一連のリスニング用テキストと問題を学習者にどのように彼らが答えたかを提示した。発話プロトコルを分析しているとき、Buck は**文化的背景や個人の知識量によって問題の解釈の仕方に根本的な違い**を発見した。構成と無関係な要素によるテスト分散が道理的に「間違った」答えへ導き得ることが分かったのだ。

西アフリカに住んでいた私の友達のスーザンは、そこに住んでいる間、物取りに悩まされていた。2 か月の間、毎週日曜の夜に誰かがスーザンの寝室の窓から家に押し入り、家からとても小さいものを盗んでいたのだ。彼女は窓に棒を仕掛けたり、警備を雇ったりしてこれらが起きることを防ごうとした。そして、それでもまだ、毎週日曜の夜に誰かがどうやってか、窓から忍び込んで何かを盗んだ。

質問）スーザンは問題に対して何をしましたか。

☞ ある学生は、「彼女は警備を雇い、つけかけ棒を使って窓をふさぎました。」と回答した。

Social Responsibility and Accountability（社会的責任と信頼）

パイロティングにおける統計的根拠は、研究者にとって、予備集団に対して差別的な設問がないかを確認することも可能にする。すべてのテスト作成者、受験者は次の点に留意すべきである。仮に、得点が年齢、社会的性差、障害、人種、宗教、その他の保護された特徴などに敏感な場合、有効でない得点の使用は訴訟にもつながる。実践的な教訓は、利害関係が大きくなるにつれて、慎重な設計、調査、検証に費やす努力がより煩雑となる。

³ これは構成概念と無関係なテスト分散（construct-irrelevant variance）という用語を使用し要約されている

⁴ 遡及とは以前に遡ることをいい、問題解決のための原因分析に使用する <http://consul-navi.tmg.co.jp/tie/detail3/detail21.html>

5 Conclusion and Future Directions 【結論と今後の展望】

これらの実践は 20 世紀にかけてゆっくりと進化してきて、それぞれの世代ごとの指針や教科書の中で書面化されてきており、この章で議論されたテストデザインサイクルを反映している。現在においては特に 2 つの重要な課題があり、研究者が結論を出すまでにしばしの時間を要することだろう。

課題 1) 試験実施によりコミュニティに対して与えるインパクト

- ・「テストが実施対象の地域において、被教育者にとって馴染みのあるものとなっているか」
- ・「実施対象地域で全く知見のない巨大試験機関によって実施されるのか」を考慮すべきである。

政治的問題を含むが、実施対象地域の当局または外部委託にすべきなのか、評価を適切に運用するには特定のコミュニティや、利害関係にある当事者の視点も考慮することで、大規模な試験実施に対する実施対象地域の参与による波及効果や将来の試験の再考に影響を与えるかもしれない。

また、特定の意思決定という文脈に使われる国際試験と比較して、地域で設計された試験がどれほど機関・地域・国にとって有益であるか。この問いはそれぞれの使用文脈において吟味される必要があるが、調査の一般化モデルは現時点では存在していない。

課題 2) 教員の継続的専門性の向上におけるテストデザインの役割

Davidson and Lynch はテスト要件の解釈を意思決定の過程に関連づける部分に着目し、テストデザインの文脈において、教員の集団力学を初めて考察した。教員同士の共通理解に着目するために、各々がそれぞれの解釈の理解と学習目的を認識するための spec development ワークショップを開いた。特定の機関における言語プログラム内では、専門性の向上は地域密着型のテストデザイン（解釈への同意を通して、教育と評価に共有できる設計タスク、共通評価基準の構築）に多大な貢献をするかもしれない。その分野においては現在、どのように試験設計者が決定に至るのか、どのような手順で発展した教育哲学を支持しているのか、などのテストデザインの活動力に関する研究が十分でない。

実践的テストデザインは、経験研究と論理的理解による知見を用いた人的決定を必要とする。研究と合理性によって支持された決定の語りはデザインによる有効性の主張の構成要素となり、スコアの意味とその使用目的を促進させるという主張を明白にする。

テストデザインの手順は一つの発見である。研究業界において、グループの結合力と相互作用を必要とし、専門性の向上とチームビルディングにとってきわめて有益である。理想的には、【学習者間の言語使用を想定する領域に関連し、意思決定に有用な試験】を最終製品としたい。

IV. 考察点

- ・テスト作成時にその目的や評価基準を明確化することと、評価リテラシーをどう考えるべきか。
- ・テスト内容の精査へのプロセスはテストの規模によっても異なるが、相対的な評価は明確でなければならないと思う。言語テストの評価のあり方とは何なのか。
- ・言語テストが本来の目的とは異なる文脈で使用されている現状をどう捉えるべきか。
- ・テストの「妥当性」を求めれば求めるほど構成プロセスは複雑になる。一方で「汎用性」や「有効性」を保つには、どんな点に考慮しなければならないのか。

V. 参考 URL

<https://translate.google.co.jp/?hl=ja>

<https://www.jrc.sophia.ac.jp/uploads/2016/06/c368164e4dd0bcbbaa0422d09663544a9.pdf>

○追加（授業でのディスカッションのまとめ）

言語評価（assessment）は学習者の学習の到達度や理解度を可視化するために行われるものであり、点数評価（evaluation）とは区別して認識する必要がある。

大規模テストの再利用の問題は、学習者や学習目的の多様化との関連が大きく、その一つとして移民・難民問題が考えられる。テストの対象者が多様化している一方で、各々の言語能力を測るテストが一様化しているという現象を見ると、テストの有効性に疑問を投げかけざるを得ない。個々人の学習目的、学習段階にあった能力が身に付いているかどうかを測るためにはローカルな視点が必要であるにも関わらず、大規模テストを代用するような動きがある。

また、対象者の多様化は、テストに織り込まれる情報面にも大きく影響を与える。私たちが日ごろ同じ言語を使用していても理解の齟齬が起きる場合があるように、テスト上でも文化背景（国家間の異文化のみならず個々人の生活環境から生まれる文化差）や既存の知識量によって、テストの結果に影響が出てしまう場合があることを、テスト作成の段階で配慮することが必要である。

発表者としては、この論文を通して「評価リテラシー」のあり方について深く考えさせられた。特に学習者の多くが共通して目指す大規模テストから個々の能力を判断することの意味や、ローカルなテストで個々の能力を評価するには何に照準を定めるべきなのかという点について、テストや評価に関わる教育者一人ひとりが意識し、向き合っていく必要があるのではないかと思う。