

II. スピーキングの測定・評価 (pp. 188- 204)

2. スピーキング能力のテストとアセスメント

スピーキング技能を測るテストは、何をどのように測ってどのように評価するのか？

■ Fulcher(2003)

L2 としてのスピーキングのテストは、他の技能などのテストよりも相当難しい。

スピーキングテストでは、実際に何かを話させる必要がある。言語についての知識と運用能力は違う。

■ Underhill (2003)

オーラルテストは質的にも他の種類のテストとは異なり、実際的なオーラルテストを仮定した場合、対面で互いに話し合う必要がある。→**human approach** 方式でデザインしなければならない。

■ Louma (2004)

話し言葉としての文法特性や語彙特性があることを認識することが重要。

■ 構成概念妥当性のための基盤として、以下の4観点の考慮が必要

- ・ 特定の状況下におけるテストしたいスピーキングの種類
- ・ 上記のテストのためのタスクと評価基準の開発
- ・ テスト受験者に対する情報の告知
- ・ テスティングと評価方法についての実施計画の確認

2.1 スピーキングテストの測定と評価

■ Louma (2004) タスクを大きく2分類した。

制約を設けないタスク(open-ended speaking tasks)→自由に述べさせる

構造的タスク(structured speaking tasks)→多肢選択的なので判定の正当化につながる

その他にロールプレイなどのセミ構造的タスクがある。

■ タスクは受験者が個人、ペア、グループで構成されるか、教育的タスクか実社会タスクかという違いで、必要な情報が異なる。また、録音の有無で評価法に影響が与えられる。

■ タスクにはタスク用の題材が用意されることが重要。

表 1. スピーキングタスクのタイプ (教科書 p. 190) 参照

■ スピーキングテストにおける評価法

全体的評価(holistic rating)

分析的評価(analytic rating)

これらの評価法をテストの目的によって使い分ける必要がある。

(1) Flucher (2003)

- L2 スピーキングテストの分野が注目されたのは WWII 以降のことであり、それまでは採点者に信頼を置けないことが議論されていたため関心が低かった模様。
- 1950 年台に開発された FSI (Foreign Service Institute) において、軍用外国語コミュニケーションを対象とした FSI 評価尺度がスピーキング評価の原点。

(2) OPI

- ILR (Interagency Language Roundtable)尺度は FSI を元に作成されたものだが、教育現場に適さないことから、それを発展させた評価尺度が OPI による評定尺度である。
- ACTFL (the American Council on the Teaching of Foreign Languages)と ETS が共同で、Proficiency Guidelines を開発し、その評価尺度を適用したのが ACTFL OPI である。直接対話方式の口頭能力面接試験。評定は 10 段階設けられている。(p. 192 参照)

■ Flucher(2003)

Skehan(2001)を拡大し、スピーキングテストのパフォーマンスとの関係を p. 193 に示される図に示した。

(3) Bachman & Parmer (1996)

評価上の問題として、評価者間の評定尺度の解釈の違いや、基準の厳格さの違い、評定尺度に関連していない要素への反応があげられるが、評定者を適切に選び、訓練することで避けられる。

(4) Underhill (2003)

オーラルテストは常に高い信頼性を得られるとは限らない。テストは受験者と対面で行うべきで、評定者の人数を増やす、または録音することが有効である。Inter-rater reliability も考慮し、少数の評定者がすべてのテストに関わって評定を下すべき。

(5) Louma (2004)

TSE の評定尺度について。全体的尺度と分析的尺度を適用し、20～60 の 5 段階評定になっている。以下の 4 つの評定尺度が重要であると Louma (2004)は述べている。

- ①評定者重視の評定尺度
- ②受験者重視の評定尺度
- ③テスト管理者重視の評定尺度
- ④CEFR の can-do による評定尺度
→range, accuracy, fluency, interaction, coherence

(6) Flucher & Davidson (2007)

- 授業でのアセスメントについて
 - ・インタラクションとコミュニケーションを促し、目標の達成の分かち合い感、教授者と学習者間・学習者同士のフィードバックが目的になる。
 - ・テストではこれらの行為が cheating につながるが、グループで行うテストには認められる。
 - ・スピーキングテストの管理と運営について、テスト環境を整え、対話者と評定者を用意し、事前訓練を行い、評定者間信頼性と評定者内信頼性に留意する必要。

(7) Hughes, R. (2002)

スピーキングの評定基準の比較

1. IELTS

流暢さと一貫性、語彙の使用範囲と量、文法の使用範囲と正確さ

2. CPE

談話管理、語彙の試用、文法の試用、発音、相互のやり取り

3. CELS

談話管理、文法と語彙、発音、相互のやり取り

→言語能力とスピーキング能力のどちらを評価するのか疑問を呈している。

2.2 スピーキングテストの実証的研究

■ McNamara (2000)のパフォーマンステスト

テストの役割を考えて、テストの設計図を描く前に、テスト業務遂行への制約があることを念頭に置くべき。テスト内容と採点方式を含めたテスト方法を考え、受験者のテスト問題への応答形式から応答の真正性も考慮すべき。

■ テスト実施のための管理・運営については、テスト問題文の指示(prompt)、受験者への指示(rubric)などを決め、評価基準と採点手順が開発されるべき。

■ 訓練を受けた評価者が、合意された評定手順によって発話サンプルを評定する。

■ タスクは実社会タスクであるべき。

■ パフォーマンスのどの点を判定するか考慮が必要。尺度の他に level descriptors を用意しておくといよい。

■ 全体的評定を行うよりは、パフォーマンスのそれぞれの面を分析的に評価するべき。

(2) Hughes, A. (2003)の口頭能力テスト

①受験者へのタスクは、タスクのすべてを達成できるような代表的なものを設定

②受験者の能力を真に反映できる言語行動を引き出すべき

③引き出された言語行動のサンプルは、妥当性、信頼性を持って採点される

タスクの設定には代表的なタスクを設定し、すべての内容を明確に記す必要がある(e.g. p.199)。

■ テストの計画と構成

①実行可能な限り時間を長くする、②テストを注意深く立案する、③受験者にはできるだけ多くの「フレキシブルなスタート」を与える、④面接には 2 番目の面接官が控える、⑤出題のタスクとトピックは母語でなら容易に取り組める、⑥試験室は面接には静で声がよく聞こえるところ、⑦受験者の実力が出せるように気楽にさせる、⑧関連する情報を十分に集める、⑨しゃべりすぎない、⑩面接担当者の選考には注意深く、そして選んだ面接担当者をトレーニングする

■ 面接担当者のトレーニング

①背景と概要の説明、②受験者のレベル分け、③インタビューの練習、④トレーニングの考査、⑤採点の妥当性と信頼性の確保。

(3) 英語スピーキングテスト SST

- ACTFL と (株) アルクの共同開発。9段階の判定が可能。日本人の学習者に多い中級レベルを詳細に測定でき、母語話者以外も試験官になる。
- 15分間受験者と1対1の対面インタビューを行い、レベルに応じた質問を行う(質問は固定されていない)。試験官とは別の「評価官」は録音されたサンプルを聞き、言語機能、内容、発話の形、正確さ(文法、語法、発音、流暢さ、社会言語学的適切さ)を総合的に判断する。(SSTの詳細は pp. 200-201 参照)
- 最低2名の評価官が独立してインタビューを評価し、初級、中級、上級の3レベルにパフォーマンスが評価される。評価基準は以下の4観点を包括的に評価する。
 1. 総合的タスク・機能
 2. 話題・状況
 3. テキストの型
 4. 正確さ

(4) 今井裕之・吉田達弘編著(2007)のHOPE 中高生のための英語スピーキングテスト

中高生のスピーキング能力を評価するためのインタビューテスト「HOPE (High school Oral Proficiency Examination)」。

(5) コンピュータを利用したテスト

例 1. Versant™

電話による試験(コンピュータを利用しての受験も可)全体で5パートある

A: Read aloud(8問)

B: Repeat sentence (16問)

C: Short-answer questions(24問)

D: Build sentence (10問)

E: Open questions (3問)

(現在では Open question の前に Summary が3題用意されている)

- 80点満点で①総合スコア、②文章構文、③語彙、④流暢さ、⑤発音が示される。5分以内に結果が出るので使い勝手が良い。

例 2. 英語スピーキングテスト試用版 (English Recognition and Production Test: ERPTTest)

- 大学生初級レベルのスピーキングテスト。5~6分後に採点結果を閲覧できる。

(現在は利用できないもよう)

2.3 スピーキングテストの評定の信頼性・妥当性

(1) Underhill (2003)

テストの妥当性検討のための6種の要素

- ①表面妥当性 (face validity)、②内容妥当性 (content validity)、③構成概念妥当性 (construct validity)、④信頼性 (reliability)、⑤併存妥当性 (concurrent validity)、⑥予測妥当性 (predictive validity)

(2) Hughes, A. (2003)

採点の妥当性と信頼性の確保のため、

- ①適切な評定尺度の作成、②使用する評定尺度の目盛付、③評定者、および面接担当者の訓練、④受入可能な採点手順を踏む

ことが必要であるとし、訓練も含めて時間と努力が必要。

- 内容妥当性、基準関連妥当性(criterion-related validity)、基準関連妥当性のその他の証拠、採点における妥当性、表面妥当性は構成概念妥当性によってすべてまとめられる。
- 妥当性検証はテスト開発者の義務であり、それが無理な場合、
 - ①図るべき構成概念についてテスト細目規定を実際に書く
 - ②可能な限り直接テストィングを使う
 - ③テストの信頼性を高めるために、可能なあらゆることをする。

(3) Louma (2004)

タスクを評価の基盤とし、評定基準を考える。スコアの妥当性は評定基準とタスクの関係に依存する。スコアの詳細情報を受験者にフィードバックできるようにし、評定報告書をデザインしておくべき。

評定者内信頼性、評定者間信頼性、並行テスト信頼性を得るために、評定者訓練を行うことを主張している。

2.4 スピーキングテストの波及効果

- スピーキングテストによって、次の学習指導に結果を生かしていく効果が期待できる。波及効果の現れはスコア報告書によって学習者と教授者にも期待できる。
- Hughes, A. (2003)

テストによる波及効果の研究が更に進むことを予測しながらも、有益な波及効果を達成するための言語テストに関する基本方針を上げている

 - ①伸ばそうとする能力のテストを行う、②広く、予測不可能な範囲から出題する、③直接テストィングを用いること、④テストィングは基準参照的なものにする、⑤到達度テストは目的に基づくものとする、⑥テストは内容・形式を生徒・教師に周知すること、⑦必要ならば教師を支援すること、⑧コストを計算すること

ACTFL の改訂とレベルの導入について

- ACTFL 初版は 1986 年発行（1982 年にひな形となる ACTFL Provisional Proficiency Guidelines が発表されている）。1999 年の改訂でそれまで上級が 2 レベルだったものが上級—上・中・下という 3 レベルに変更されたことが最大の変更点。
（1986 年版では上級レベルが **Advanced** と **Advanced-Plus** となっている）
- もう一点の大きな変更点として、1986 年版では **Novice**→**Intermediate**→**Advance**→**Superior** と下級レベルから順にソートされていたのに対し、1999 年の改訂版では **Superior**→**Advanced** …と逆になっている。
- これは例えば、～**High** レベルのパフォーマンスがそのレベルよりも上位レベルの示す言語機能により関連が強い（**Intermediate-high** の場合は **Intermediate-mid** よりも **Advanced-low** の基準により近いものと見なす）という考えに基づくものである。
- ～**High** レベルの話者は、一つ上のレベルの言語機能を継続的に使えないということを強調し、**descriptor** の中で否定的な表現や、繰り返しになる表現を避けることも目的している。

参考

- American Council on the Teaching of Foreign Languages (1999). ACTFL PROFICIENCY GUIDELINES -Speaking. [ONLINE] Available at:
<http://www.actfl.org/sites/default/files/pdfs/public/Guidelinesspeak.pdf>. [Last Accessed 12 December 2012].
- Fulcher, G. (2003). Testing Second Language Speaking. Great Britain: Pearson Education.