Cambridge Assessment English

Linguaskill

スピーキングテストの妥当性に関する考察

Jing Xu、Mark Brenchley、Edmund Jones、
Annabelle Pinnington、Trevor Benjamin、
Kate Knill、Gaelle Seale-Coon、Martin Robinson、
Ardeshir Geranpayeh

Linguaskill

目次

概要

Linguaskill スピーキングテストは、自動採点技術により強化されたコンピュータ方式の英語オーラルテストです。このテストでは、人工知能と経験豊かな採点者（人間）による判定のそれぞれの長所やメリットを組み合わせたハイブリッドな方法を採点に取り入れています。短期間でたくさんの英語学習者を評価する必要がある法人ユーザーにとって魅力的なテストです。また、個々の学習者にとっては、ブラウザベースのスピーキングテストはとても利便性が良く、高速インターネットに接続された Windows パソコンがあれば、自宅でも受検することができます。Linguaskill のテスト結果は、自動採点技術により 48 時間以内に報告されます。

本書では、テストスコアの意図された解釈および使用法を裏付けるために収集された研究データに関する説明を織り交ぜながら、Linguaskill スピーキングテストの妥当性についてのエビデンスを示します。まず、テストの目的、目標言語使用 (TLU) 領域および形式について説明します。次に、自動採点の設計、人間の試験官向けトレーニングプログラムおよびテストで使われるハイブリッド採点モデルを明らかにします。そして、Linguaskill の妥当性に関する論拠の構成を整理して述べた後、その中の各要素がテストの妥当性を論証するために不可欠である理由を説明します。最後に、妥当性の論拠における 3 つの要素、すなわち、*テストの出題内容*、*返答の採点*、*テスト結果の解釈*に関連する研究のエビデンスを提示します。テストの妥当性検証は累積的なプロセスであることは注目に値します。妥当性の論拠を裏付ける新たなエビデンスを集めるために更なる研究が行われています。

本書で示されたエビデンスは、テストに関する以下の妥当性の主張を後押しします。

- **テスト内容：**日常生活、社会活動での対話、職場でのやりとり、通信手段を使ったコミュニケーションなどのトピックは、受検者が実社会で遭遇するであろうコミュニケーションの状況を総じてよく描写しています。
- **テスト内容：**扱うトピックは興味を持たれる内容で、簡単すぎたり、難しすぎることはありません。
- **テスト内容：**受検者はコンピュータに向かって話すことに一般的には抵抗感がありません。
- **返答の採点：**人手採点の信頼性は十分あります。
- **返答の採点：**ハイブリッド採点を実施した結果、自動採点は人間の試験官との間で CEFR グレードに関して 95.6% の完全一致と 100% の隣接一致を達成しました。
- **返答の採点：**自動採点は、英語ではないと疑われる発話を検出し、人間の試験官に確認して検証を求めます。
- **テスト結果の解釈：**テストと CEFR との関連性を確証するために、定期的に標準化の予行演習を行います。このため、CEFR に基づく高い信頼度でテスト結果を解釈することができます。
- **テスト結果の解釈：**確認的因子分析は、単一の包括的なスピーキング構成がテストによって評価されていることを示唆しています。

はじめに

Linguaskill スピーキングテストは、自動採点技術により強化されたコンピュータ方式の英語オーラルテストです。他の自動スピーキング評価とは対照的に、Linguaskill スピーキングは人間の試験官と自動採点技術を組み合わせたハイブリッドな方法で採点を行います。コンピュータが返答の採点への確信度が低いと判断した場合、人手採点に回すことになります。このハイブリッドモデルは、最新の自動採点技術と経験豊富な人間の採点者の判定を結びつけることで、完全自動化の評価へ向けた課題に対処することを目的としています。

自然言語処理、機械学習、音声認識技術の進歩に伴い、自動スピーキング評価の注目度が急上昇しています。従来の対面式のスピーキング試験と比較して、コンピュータやモバイル端末に配信される自動スピーキング評価は、スコアレポートの迅速な提供、テスト管理の簡素化、試験日の自由設定などのメリットがあります。

スピーキング評価の自動化は、対面式のスピーキング試験を大規模に実施することが難しいと考えている法人ユーザーには特に訴求力があります。個々の学習者は、遠隔試験監視システムの導入により、自宅でもスピーキングテストを受けることができるため、自動評価の利便性が向上すると考えるかもしれません。ただし、自動評価には、測定範囲に対する懐疑的な見方や受検者の不正行為の可能性など、人手採点には通常ありえない課題や問題も発生します（Chun 2006, Fan 2014, Xi 2010, Xu 2015）。

本書では、Linguaskill　スピーキングテストの妥当性の論拠を示します。妥当性の論拠は、研究のエビデンスとなる様々な要素に基づいて、提示された解釈や使用に賛成または反対のいずれかの立場で首尾一貫した分析を行うことで、テストスコアの意図した解釈や使用についての総合的な評価を与えます（Cronbach 1988, Kane 2013）。私たちはまず、意図されたテストの目的、目標言語使用領域、テスト形式を含むテスト仕様を説明することから始めます。次に、自動採点と人間の試験官が採点上の課題を共有するオーラル評価に応用するハイブリッド採点モデルを紹介します。以降では、テスト作成サイクルの各段階での批判的な妥当性の考慮すべき事項を整理して、明確な妥当性確認のフレームワークを提示します。本書の最後では、このフレームワークに基づいて集められた妥当性の根拠を提示します。

1. テストの目的

Linguaskill スピーキングテストは、受検者の日常的なコミュニケーションに必要なオーラルな英語力を評価します。スピーキング単体でも、他の Linguaskill モジュールのリーディングとリスニング、ライティングを一緒に受検することもできます。Linguaskill は、言語学習および習得の進捗状況を記述するための基準として広く認知されている CEFR（Council of Europe 2001, 2018）、および Cambridge English スケールに基づくより細かなスコアに基づいて、迅速かつ信頼性が高く、分かりやすく判断できる結果の提供を目的としています。Linguaskill の利用目的は、a) 教育機関における進学、進級、卒業のための英語レベルの測定、b) 企業における採用やキャリアアップの機会のための英語レベルの測定が挙げられます。Linguaskill の対象となる受検者は、16 歳以上の英語学習者となっています。

2. 目標言語使用領域

目標言語使用（TLU）領域とは、受検者がテスト以外でその言語を使用できる状況や文脈を仮定して記述したものです。この領域の範囲を定義し、その領域での言語使用の主な特徴を特定することで、テスト開発者は、これらの言語使用活動を模倣したタスクを設計することができます。そして、受検者のテスト時の行動は、TLU 領域において予測される言語パフォーマンスのサンプルと見なすことができます。

Linguaskill は、複数のテスト目的に対応するように設計されているため（セクション 1 参照）、その TLU 領域は日常生活と職場の両方での英語を使用のさまざまな状況やタスクに広範に対応したものでなければなりません。この領域でのコミュニケーションの文脈は多様であるため、テスト開発者は、いくつかの重要な TLU タスクを特定して記述し、それらがテスト内容に含まれていることを確認する必要があります（詳細はセクション 5.1 参照）。

Linguaskill で選択された重要な TLU タスクは、通例では 4 つのカテゴリーに分類されます。それは、日常生活（例. 余暇の過ごし方について話し合う、好みを伝える）、社会活動での対話（例. 状況や問題を説明する、ニュースや個人的な経験を語る）、職場でのやり取り（例. 問題または課題を提起する、データに関する報告を行う）、テレコミュニケーション（例. 情報を要求する、留守電を残す）です。

3. テストの説明

Linguaskill スピーキングテストは、Web ブラウザで動作しますので、試験監督者が監視する中で高速インターネットに接続された Windows パソコン[1] を使って受検します。自宅での受検を希望する場合は、遠隔での試験監視が行われます。出題はパソコン画面とヘッドフォンを通して受検者に提示され、受検者の返答が録音されてコンピュータアルゴリズムまたは試験官によって遠隔で評価されます（セクション 4 を参照）。このテストはマルチレベルの測定が可能で、A1 未満、A1、A2、B1、B2、C1 以上といった CEFR に基づいた習熟度レベルのオーラル能力を導き出して評価する設計になっています。テスト結果は 48 時間以内に報告されます。

Linguaskill Speaking テストは 5 つのパート、インタビュー、音読、プレゼンテーション、視覚情報ありのプレゼンテーション、コミュニケーション・アクティビティで構成されています。すべてのパートは均等に重み付けされ、スピーキング能力の様々な側面に焦点が当てられています。5 つのパートの形式、テストの目的、評価基準は以下の通りで、表 1 に要約しています。

3.1 インタビュー

a. 形式

インタビュー・タスクでは、受検者は自分自身に関する 8 つの質問に答えます。最初の 4 つの質問は全テスト共通の質問で受検者には各質問に答える時間が 10 秒与えられます。質問 5 ～ 8 は各テストのバージョンによって異なり、習慣、経験、嗜好などに関するシンプルで個人的な質問が出題されます。これらの各質問に答えるために 20 秒の時間が与えられます。

表 1.Linguaskill スピーキングテストのタスク概要

パート	タスク	説明	応答時間	準備時間	配点
1	インタビュー	自分自身に関する 8 つの質問に答えます。	4 問 x 10 秒および 4 問 x 20 秒	なし	20%
2	音読	8 つの文を声に出して読みます。	8 問 x 10 秒	なし	20%
3	プレゼンテーション	与えられたトピックについて話します。	1 分	40 秒	20%
4	視覚情報を用いたプレゼンテーション	与えられた図表の情報に基づいてプレゼンテーションします。	1 分	1 分	20%
5	コミュニケーション・アクティビティー	シナリオに関連した 5 つの質問について意見を述べます。	5 問 x 20 秒	40 秒	20%

[1] 2020 年 6 月の時点では Mac での Linguaskill テストはサポートされていませんでしたが、現時点ではサポートされています。パソコンでの受検は、Google Chrome または Mozilla Firefox の使用を推奨します。

b. テストの目的
受検者をコンピュータ形式のテストに慣れさせるだけでなく、このテストパートでは個人的な質問に返答する能力を評価し、習熟度の低い受検者に対しては、より身近で達成感のあるタスクを与えることに重点を置いています。

c. 評価基準
受検者は、発音や流暢さ、言語資源の観点から言語的なアウトプットが評価されます。

3.2 音読

a. 形式
音読タスクでは、8 つの文を読み上げることが求められます。各センテンスを読み上げる時間は 10 秒です。文は、受検者が現実の場面で読み上げなければならないような内容で、音韻の特徴や統語的構造などを幅広くカバーし、段々と難易度が高くなります。

b. テストの目的
このテストパートでは、書かれた英文を音声に変換する能力と、文レベルでの発音の要素を扱う能力を評価します。

c. 評価基準
受検者は、全体的な明瞭さ、各人の個別の音を発声する能力や、強勢、リズム、イントネーションといった音韻的基準に基づいて評価されます。

3.3 プレゼンテーション

a. 形式
プレゼンテーションのタスクでは、与えられたトピックについて 1 分間話すことが求められます。トピックは選べません。受検者のプレゼンテーションが録音開始されるまでに準備時間は 40 秒あります。

b. テストの目的
このテストパートでは、受検者の長く話す能力を評価します。状況や論点の説明だけでなく、受検者は箇条書きの項目に沿って意見を述べ、正当性を説明することが奨励されています。

c. 評価基準
受検者は、発音や流暢さ、言語資源と談話管理の観点から、言語的なアウトプットを評価されます。加えて、採点ではルーブリックや指示に従ってタスクを適切に完了する能力も考慮されます。

3.4 視覚情報ありのプレゼンテーション

a. 形式
視覚情報を用いたプレゼンテーションタスクでは、受検者は視覚的な形で提示された情報について 1 分間話すことが求められます。受検者のプレゼンテーションが録音開始されるまでに準備時間は 1 分あります。受検者は、友人にボイスメールを残したり、クラスでプレゼンテーションを行うなど、特定の文脈の中で情報を描写するように求められます。

b. テストの目的
このテストパートでは、とてもシンプルな視覚情報を解釈し、勧める、説明する、または提案することを含む長く話す能力を評価します。

c. 評価基準
受検者は、発音や流暢さ、言語資源と談話管理の観点から、言語的なアウトプットを評価されます。加えて、採点ではルーブリックや指示に従ってタスクを適切に完了する能力も考慮されます。

3.5 コミュニケーション・アクティビティー

a. 形式
コミュニケーション・アクティビティーのタスクでは、あるシナリオに関連した 5 つの質問に答えることが求められます。最初の質問を聴く前に 40 秒の準備時間が与えられます。各質問ごとに 20 秒の返答時間で、受検者は意見を述べたり、仮説を推測したり、評価することが求められます。

b. テストの目的
このテストパートでは、音声による指示に従って、与えられたトピックについての意見や考えを表現する能力を評価します。これは、より高いレベルの受検者が、さらに高いレベルのスキルを発揮する機会となります。

c. 評価基準
受検者は、発音や流暢さ、言語資源と談話管理の観点から、言語的なアウトプットを評価されます。加えて、採点ではルーブリックや指示に従ってタスクを適切に完了する能力も考慮されます。

4. スピーキング・パフォーマンスの採点評価

Linguaskill スピーキングテストでは、ハイブリッド、つまり*人間が採点*の過程に含まれるモデルを採用しています。ライブ評価で自動採点技術が使われますが、採点に人間の試験官が関与するモデルです。このセクションでは、自動採点の設計、試験官のトレーニングと認定およびハイブリッド採点の適用方法について説明します。

4.1 AI 自動採点技術

Cambridge English の AI 自動採点技術とは、スピーキングやライティングのように構成されたテストの応答を採点するために設計されたコンピュータアルゴリズムのセットです。ケンブリッジ大学英語検定機構（以下、Cambridge English）は、ケンブリッジ大学の研究グループとの共同研究により、2012 年に非ネイティブ英語の自然発話の自動採点の開発に着手しました。Linguaskill スピーキングテストで使用される自動採点は、CASE（Custom Automated Speech Engine）と呼ばれ、ケンブリッジ大学の学際的な研究センターである ALTA（Institute for Automated Language Teaching and Assessment）から機械学習技術を用いて移転された技術を基に、Enhanced Speech Technology 社が開発しました。

CASE は、図 1 に示すように、音声認識、特徴抽出モジュール、グレーダーの 3 つの主要コンポーネントで構成され

図 1. 自動採点の構成（Knill 他 2018）

ています（Knill 他 2018, Wang 他 2018）。音声認識では、発話の音声信号を基礎となる単語変換から、構造化された表現にテキスト変換する自動音声認識（ASR）を実行します。この音声認識は、Cambridge English が提供した学習者の音声を使用し、クラウドソースのテキスト変換を組み合わせてディープニューラルネットワークモデルに基づいて訓練されました（Lu 他 2019 ASR2 システムを参照）。特徴抽出は、音声信号と構造化されたテキスト変換の両方から、発話構成に関連する特徴（例えば、流暢さ、発音の正確さ、語彙の多様性）をグレード評価の基準として引き出します。これらの特徴に基づいて、グレーダーは最先端の機械学習モデルを使用してスコアの分布を返し、そこから CEFR グレードなどの受検者へのフィードバックを導き出します。グレーダーの訓練サンプルには、さまざまな第一言語および CEFR 全レベルの学習者によって生成された Linguaskill スピーキングテストの応答の大規模なセットと、試験官がこれらの応答を評価した採点も含まれています。さらに、CASE で使用されている機械学習グレーダーモデルは、入力データと訓練データの類似性に基づいて不確実性の尺度を返すように設定されています（van Dalen, Knill and Gales 2015, Malinin, Ragni, Knill and Gales 2017）。この不確実性の尺度は、自動採点スコアの信頼性を示す重要な指標であり、人手採点を必要とするテストの応答を識別するのに役立ちます。

4.2 試験官

Linguaskill スピーキングテストの試験官は全員、資格を得るために厳格なトレーニングプログラムを受けています（図　2）。試験官の候補者は、最低限の専門的な要件を満たしている必要があります。具体的には、大学卒または同等レベルの教育を受けていること、言語教授の認定資格を有していること、直近 2 年以内に実質的かつ関連性のある教育実務経験を証明すること、および適切な英語能力を有していることです。

要件を満たし承認された志願者は、オンライン・ポータルサイトを通じてトレーニング教材が提供されます。採点手順に関する豊富な資料と、採点済みのスピーキング解答例、パフォーマンスと採点根拠に関する詳細なコメントが用意されています。志願者は、この資料をもとに指導されます。

アクセスしてから 30 日以内に、資格認定テストを受ける必要があります。資格認定テストには、経験豊富で信頼できる試験官によって事前に採点されたスピーキング項目が、最終的に承認された採点として統計的に調整された平均スコア[2]とともに含まれています。志願者は、適正な採点の 80% 以上 (承認済み採点の 0.5% 以内) を取らなければなりません。異なるバージョンのテストが 2 回実施されます。両方のテストに合格しなかった志願者は、ポータルサイトへのアクセス権が自動的に取り消されます。

志願者が資格認定テストに見事合格すると、認定試験官のシステム ID が付与され、採点プールに追加されて受検者の採点を開始することができます。試験官が採点業務をスタートさせると、その後の業務は統計的に監視されます。採点行動分析は、バイアスの可能性、一貫性および非準拠行動を識別するために実施されます。統計的にフラグが立てられた試験官は、その行動が不十分であると確認された場合には、調査されて採点プールから除外されます。再認定は 2 年ごとに新しいトレーニングとテスト教材が用意されます。

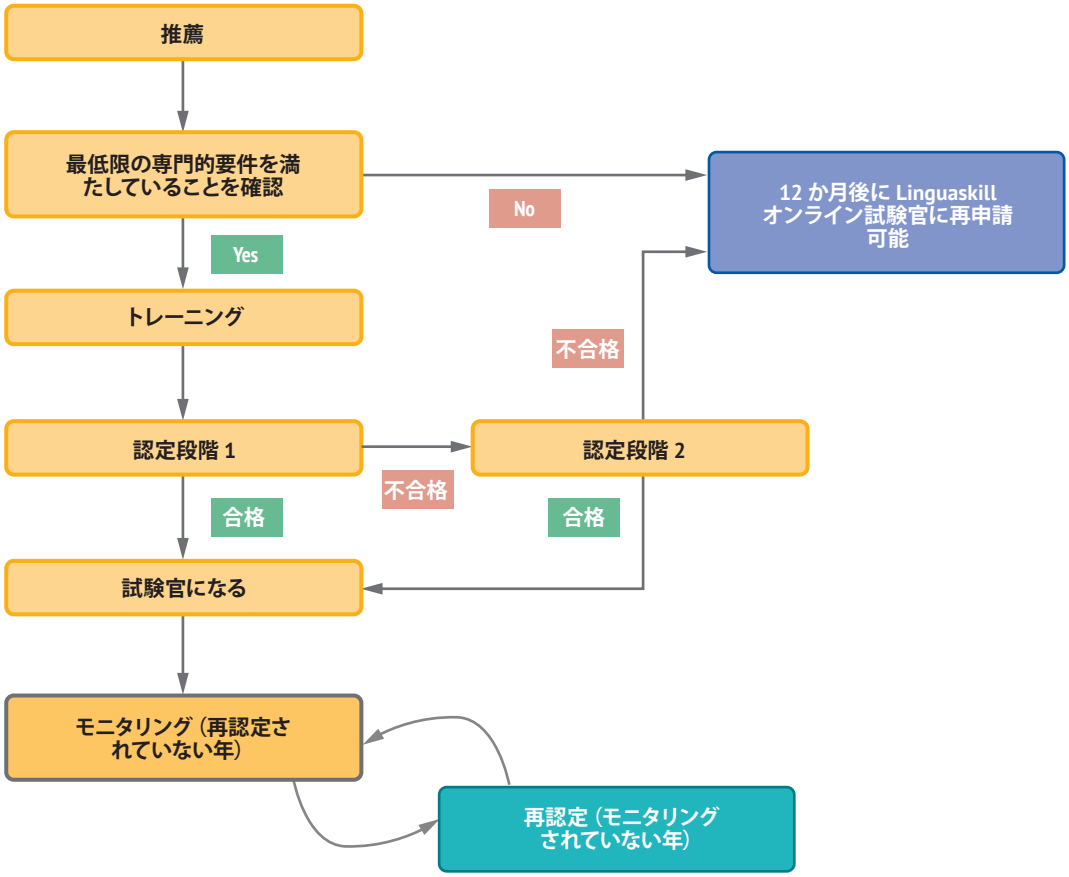
4.3 ハイブリッド採点

ハイブリッド採点は、人工知能（AI）と人間の試験官のそれぞれの長所とメリットを組み合わせます。コンピュータによる採点はスピーディーで費用対効果に優れていますが、採点される応答は AI システムの訓練サンプルに近いものでないと、信頼性が低下します。自動採点の精度の低下は、音質の悪さ、異常な発話行動、訓練サンプルの不足など複数の要因があります。音質が悪いと、ASR の精度が著しく低下し、自動採点のパフォーマンスに影響を与える可能性があります。また、学習者は採点システムを騙してより高い評価を得ようとする戦略を立てようとします（Xi, Schmidgall and Wang 2016）。このように、人間の試験官は、信頼性の低い自動採点のスコアが受検者に公開されるのを防ぐためのゲートキーパーとして重要な役割を果たしていると言えます。

ハイブリッド採点モデルは、人間の試験官の専門知識を活用して AI 自動採点技術をサポートし、さらに発展させようというものです。また、コンピュータがスコア予測の確信度を示す情報を提供することを前提としています。この確信度が低いと、テストの対応にフラグが立ち、人手採点に回すことになります。Cambridge English ハイブリッド採点モデル（図 3）では、人間の試験官への確認は、予測されたスコアに加えて、自動採点によって生成された 3 つの特徴に閾値を設定することによって決定されます。それは、評価品質スコア、言語品質スコア、音声品質スコアです。*評価品質*スコアは、グレーダーによって生成される不確実性の尺度であり、グレーダーがスコア

[2] 多相ラッシュ分析 (Linacre 1989) により、採点を厳格に調整した公正な平均スコアを活用。

図 2.Linguaskill スピーキング試験官の認定手順



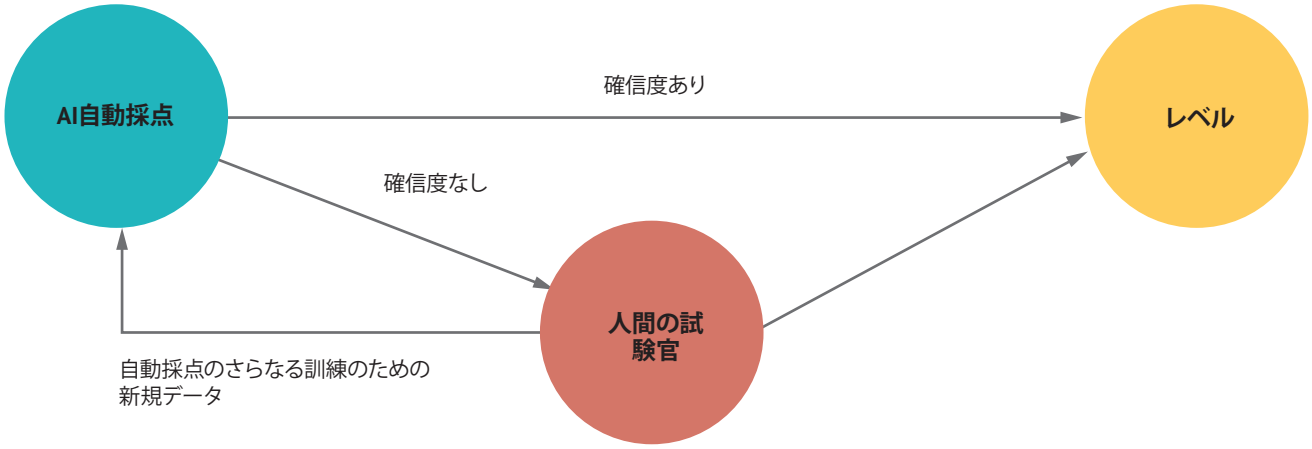
予測に対して持っている確信度の量を示します (セクション 4.1 参照)。*言語品質*スコアは、音声認識によって返される ASR 確信度スコアです。これは、システムの転写の正確さに対する確信度を表し、テスト中に実際に英語を話していない受検者を識別するのに役立つ指標となります（Lu　他 2019　を参照）。*音声品質スコア*は、音声録音の明瞭度を示し、3 つの別々の測定から導き出されます。ダイナミックレシオ（音声の大きい部分と小さい部分の間の振幅の差）、クリッピング（可能な最大値 / 最小値に達した結果、歪みが発生したオーディオフレーム）、およびノイズです。また、音質や音声からテキストへの変換中の中間処理に関連した他の様々な ASR エラーも含まれています。さらに、自動採点スコアが特定のカットオフ値を下回ったり上回ったりしたテスト応答は、試験官の採点のためにフラグが立てられます。これは、自動採点の評価によると、自動採点のスコアは採点スケールの下限と上限では信頼性が低い傾向があることを示唆しています。現在のハイブリッド採点モデルでは、テスト応答の大部分が人間の試験官によって採点され、採点の質を保証して自動採点をさらに訓練するために採点データを提供しています。自動採点の向上に伴い、人手採点の割合は徐々に減少していくと考えられます。自動採点とハイブリッド採点モデルの評価については、セクション 6.2 で詳しく説明します。

5. テスト検証のフレームワーク

妥当性は、評価の最も基本的で重要な点です。妥当性とは、意図されたテスト使用のため、結果の解釈をエビデンスや理論によって裏付けられる度合いのことです (AERA、APA、NCME 2014)。言語テストの妥当性検証に使用される 2 つの一般的なフレームワークは、論拠ベースのフレームワーク（Bachman and Palmer 2010, Kane 2013）と社会認知的フレームワーク（Weir 2005）です。前者は、実践的な議論を中心とした妥当性の探求を構造化することで、Messick（1989）の複雑な妥当性理論を分解することに焦点を当てています。後者は、Messick（1989）の妥当性理論を言語評価に適用し、Messick と同様にエビデンス収集に累積的なアプローチをとります。図 4 に示すように Linguaskill の妥当性の論拠は、妥当性の主張をすること、裏付けとなるエビデンスを評価すること、この 2 つがまとめられて構成されています。

Linguaskill の妥当性の論拠は、次の 6 つの部分で構成されています。*「テスト内容」*、*「応答プロセス」*、*「返答の採点」*、*「テスト結果の解釈」*、*「テストの用途」*、*「テストのインパクト」* です。これらは、テストの構成からテストが利害関係者に与える影響まで、典型的なテストサイクルが構成する一連の流れを表しています。このセクションでは、これらの概念が何を意味するのか、Linguaskill の妥当性の検証研究にどのように役立つのかについて説明します。

図 3.Cambridge English ハイブリッド採点モデル



5.1 テスト内容

テストの妥当性に関する共通理解は、テスト自体または能力を測定するために構築された仕組みに関係しています。つまり、テスト設計は高品質で目的に合ったものであるかどうか？この理解は誤りではありませんが、妥当性の 1 つの側面に過ぎません。

伝統的に、テスト内容に関する妥当性の側面は、*内容妥当性*（APA, AERA, NCME 1974）または*文脈妥当性*（Weir 2005）と呼ばれています。この考え方は、テストの出題が意図されたテスト目的に関連しており、その目的に関連した重要な知識やスキルを網羅しているべきであるというものです。言語評価では、テスト内容の妥当性の根拠は、通常、タスクと TLU 領域 (受検者がテスト外で遭遇する可能性の高い状況や文脈を説明) との関連性についての専門家によるレビューによって収集されます。テスト内容のレビューの目的は、タスクの特性が TLU 領域の言語使用活動の特性を反映しているか、または適切に表現していることを確認することです。この概念は、一部の言語テスト研究者 (例えば、Bachman and Palmer 1996) によって*真正性*とも呼ばれ、妥当性の論拠の基礎として考えられています (Bachman and Palmer 2010, Chapelle, Enright and Jamieson 2010)。

5.2 応答プロセス

言語テストを受けると、受検者の言語使用が誘発されて観察されます。応答プロセスに関する妥当性は、目標言語能力に起因する行動の誘発に関係しています。このことは、テストによって評価される特性や能力に関する*構成概念妥当性*の独自な概念と結びつきます (Cronbach and Meehl 1955)。応答プロセスの妥当性の根拠には、受検行動 / 戦略と言語能力構成理論の間の一貫性、適切なタスクの運用、明快なテスト指示、特別な配慮が必要な受検者への対応などが挙げられます。このようなエビデンスは、テストでのパフォーマンスに影響を与えた目標言語能力以外の要因が、テストスコアの選択的解釈を排除するのに役立ちます (AERA、APA、NCME 2014)。

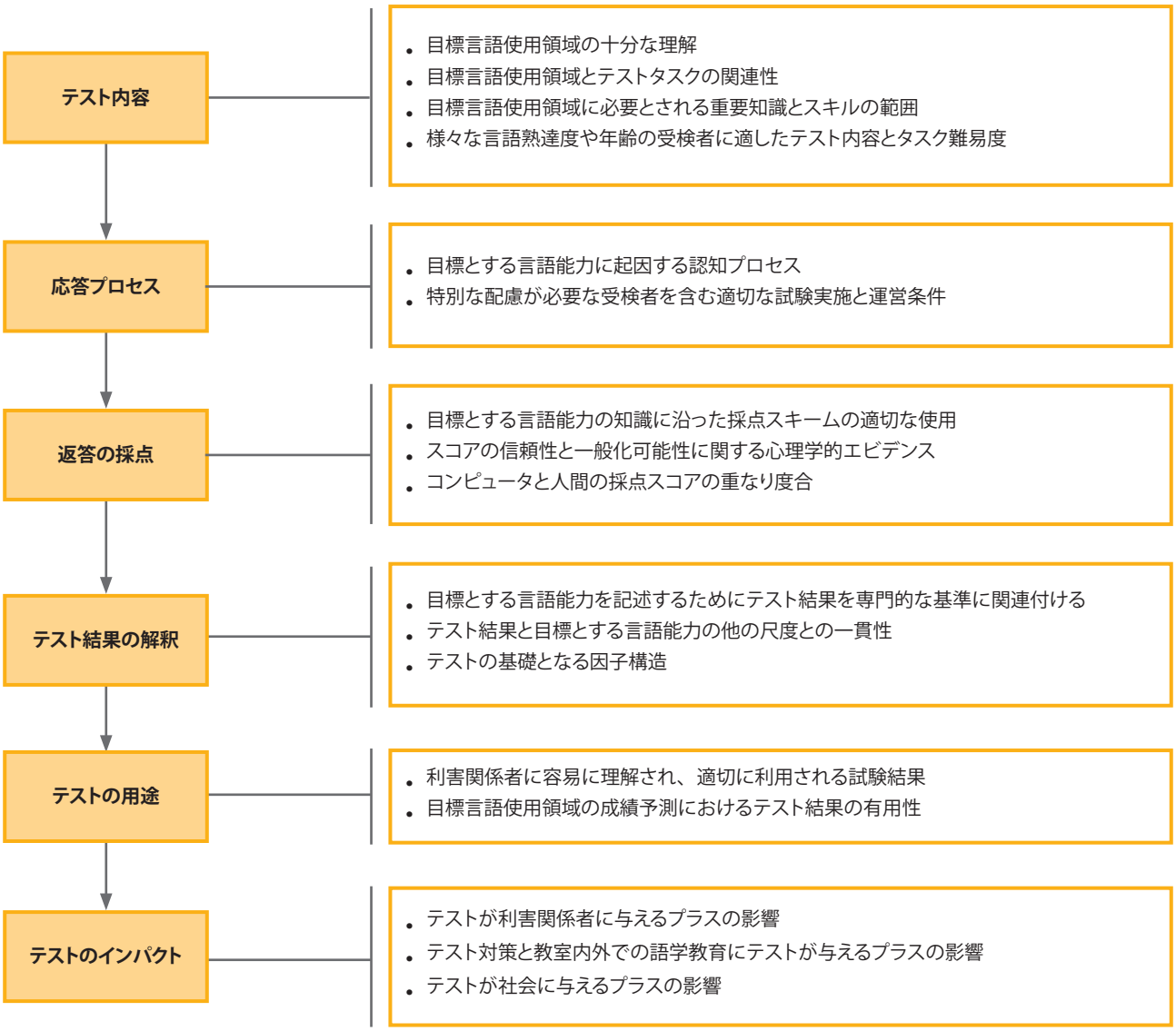
5.3 返答の採点

テストの採点は、試験官またはコンピュータアルゴリズムのいずれかによって行われます。テスト方式があらかじめ用意した多肢選択から正解を選ぶのではなく、構成された返答を引き出す場合、テストのパフォーマンスを評価するための採点スキーム (採点基準とも呼ばれる) が必要になります。採点に関連する妥当性の根拠は、採点スキームの検証、人手採点プロセスや理論的根拠の分析、採点の一貫性の調査から得られることがあります。例えば、採点スキームの開発は、目標とする言語能力に関する理論や研究に基づいており、試験官はそのような能力に起因する言語行動の重要な側面に信用を与えることが期待されています。さらに、スコアの重み付けや組み合わせなどの採点プロセスは、正当化され、目標とする言語能力を推定するための最善のアプローチを反映している必要があります。また、受検者が受け取るテストスコアは、同じテストの並列形式や採点プールから無作為に選ばれた受検者が得たであろうスコアをほぼ推定することが期待されています。この概念は一般的に*信頼性* (Haertel 2006) または*スコアの妥当性* (Weir 2005) と言われます。機械が人間のスコアを予測するために使用される場合、自動化されたスコアの信頼性は、人間の試験官と機械の一致、または人間のスコアから機械のスコアの偏差という観点から示す必要があります。さらに、構成された返答がコンピュータによって採点される場合、コンピュータと人間の採点基準の重なりの度合いについてのエビデンスを探す必要があります。そうしなければ、コンピュータの採点と人間の採点を同じように解釈することが困難になります。（Xi 2010, Xu 2015）。

5.4 テスト結果の解釈

テストのスコアは単なる数値なので、さまざまな目的で使えるように意味を持たせなければなりません。これがスコア解釈の核心です。ほとんどの場合、テスト開発者は、スコアに関連する能力の Can Do ステートメントという形で、テストユーザーに推奨される解釈を提供しますが、この解釈は、認知プロセス、言語発達、または第二言語習得の理論に裏付けされていなければなりません（Weir 2005）。スコア解釈の妥当性の根拠は、CEFR（Council of

図 4.Linguaskill の妥当性に関する論拠の概要



Europe 2001, 2018）のような言語能力を記述するための理論および / または研究に基づく基準にテストスコアを関連付けることを目的とした標準化の予行演習から得ることができます。さらに、このエビデンスは、テスト結果と目標言語能力の他の測定値との関係を調べる同時並行の調査から収集されることがあります。これは慣例では*併存的妥当性*と呼ばれています (APA, AERA, NCME 1974)。スコア解釈の妥当性の根拠は、テストの基礎となる因子構造を調査する潜在因子分析から収集されることもあります (例えば、Sawaki, Stricker and Oranje 2009)。このエビデンスは、統合言語評価に特に関連性が高く、2 つ以上の言語スキル (リスニングとスピーキングなど) が同時に評価されます。

5.5 テストの用途

テスト結果によって、受検者、教師、雇用主、入試担当者などの利害関係者の多くは、次のような行動を取ります。例えば、受検者は特定のスキルを向上させるためにさらに努力する意思を固める、教師は生徒の学習ニーズに合わせてレッスンプランを微調整する、雇用主は海外市場を拡大するためにスコアの高い受検者からチームを選抜する、学校の入試担当者は志願者の合否を決定する、などが考えられます。テストの用途に関連する妥当性の根拠とは、テスト結果が、利害関係者が十分な情報に基づいた意思決定を行ったり、正しい行動をとったりする場合に、どの程度役立っているかに大きく関わっています。このエビデンスは、次の 2 つの分野で探し求められます。1 つ目は、意図的でないスコアの解釈や使用を避けるために、提案されたテスト活用やテストスコアの意味をテストユーザーに十分理解されている必要があります。2 つ目には、テスト結果は、言語の使用に関連する仕事の成果や学業成績など、今後の重要な行動を予測するのに役立つものである必要があります。このような予測的な意味での妥当性の使用は*予測的妥当性*と呼ばれ、1950 年代に*構成概念妥当性*が登場する以前は主流でした (APA, AERA, NCME 1974)。

5.6 テストのインパクト

テストスコアの使用は、さまざまな教育や学習、社会的な文脈の中で利害関係者に影響を与えます。例えば、ハイステークスな言語テストがどのように設計されているかは、学習者がどのように言語を学ぶか、教師がどのように言語を教えるか、さらには言語能力と公平性に関する社会的価値観にまでインパクトを与える可能性があります。このインパクトは、*社会的重要性*、*波及効果*、*結果的妥当性*とも呼ばれ (Cheng 2014, Messick 1996, Weir 2005)、妥当性の概念に不可欠な側面です。テストのインパクトは、テストがどのように使用されるかということに密接に関係しています。もし Linguaskill が意図しない目的のために誤って使用された場合、テストはマイナスに影響することがあります。テストのインパクトをポジティブにする責任は、テスト提供者とテストユーザーの双方にあります。

6.Linguaskill スピーキングテストの妥当性の根拠

このセクションでは、Linguaskill スピーキングテストをその意図した目的に合わせて使用するための研究エビデンスを紹介します。テストの検証は、累積的かつ継続的なプロセスです (Messick 1989)。妥当性の根拠は、より多くのデータが収集されるにつれて時間の経過とともに洗練されていきますが、Linguaskill テストは比較的新しいテストであるため、妥当性の論拠のすべての側面がまだ完全に文書化されているわけではありません。*テストの内容*、*返答の採点*、*テスト結果の解釈*については、広範なエビデンスが得られています。テストが使用されている特定の国や地域において、*応答プロセス*、*テストの用途*、*テストのインパクト*に関する新たなエビデンスを収集するためにさらなる調査が行われています。

6.1 テスト内容の妥当性の根拠

TLU 領域へのテストタスクの関連性を証明するには専門家の判断が不可欠です（Messick 1989, p.39）。Linguaskill の場合、内容の適切性と構成要素の範囲 (すなわち、テストで評価されるスキル) に関する判断は、作問者、上級試験官、言語テスト研究者からなる専門家グループが、テスト開発段階のテスト検討会議で行います (図 5)。スピーキングの設問審査では、設問の難易度、指示や説明の明確さ、トピックや状況の本物らしさ、返答に必要な背景知識、評価されるスキルに重点が置かれます。審査で不合格となった設問は、破棄されるか、改訂されてからもう一度審査されます。専門家による審査に合格した設問は、本番テストに採用される前に、対象となるテストの母集団から選ばれた学習者の大規模グループで試行されます。試行テストによって見つかった問題ある設問は、開発段階に戻されます。

Xu and Gallacher（2017）は、Linguaskill スピーキングテストのグローバルトライアルとして、23 カ国の成人英語学習者 3,601 人を対象に調査を実施しました。受検参加者の多くが、タスクは現実世界での英語の使い方に似ている (65.3%)、トピックは自分の生活に密接に関連している (57.9%) と答えました。さらに、参加者の約 70% が、Linguaskill スピーキングテストは英語を話す能力を証明することを可能にする、に同意または強く同意しました。

試行テストの質に関するアンケート調査の分析では、スピーキングのトピックは興味深く、簡単すぎず難しすぎず、日常生活や仕事に関連することが示されました。例えば、ある参加者は、テストの質問を自分の日常的な言語使用状況に結び付けています。

> 「トピックは適切であり、簡単すぎず、難しすぎないと思います。これらのトピックは、日常生活で普通に起こることに関連していると思います。これらのトピックは、現実の社会で起こることなので、英語を学ぶほとんどの人がマスターすべきトピックだと思います。」（参加者 ID 1742853）

多くの参加者は、スピーキングテストは予想したほどストレスを感じなかったと報告しています。ある参加者は、次のように報告しています。

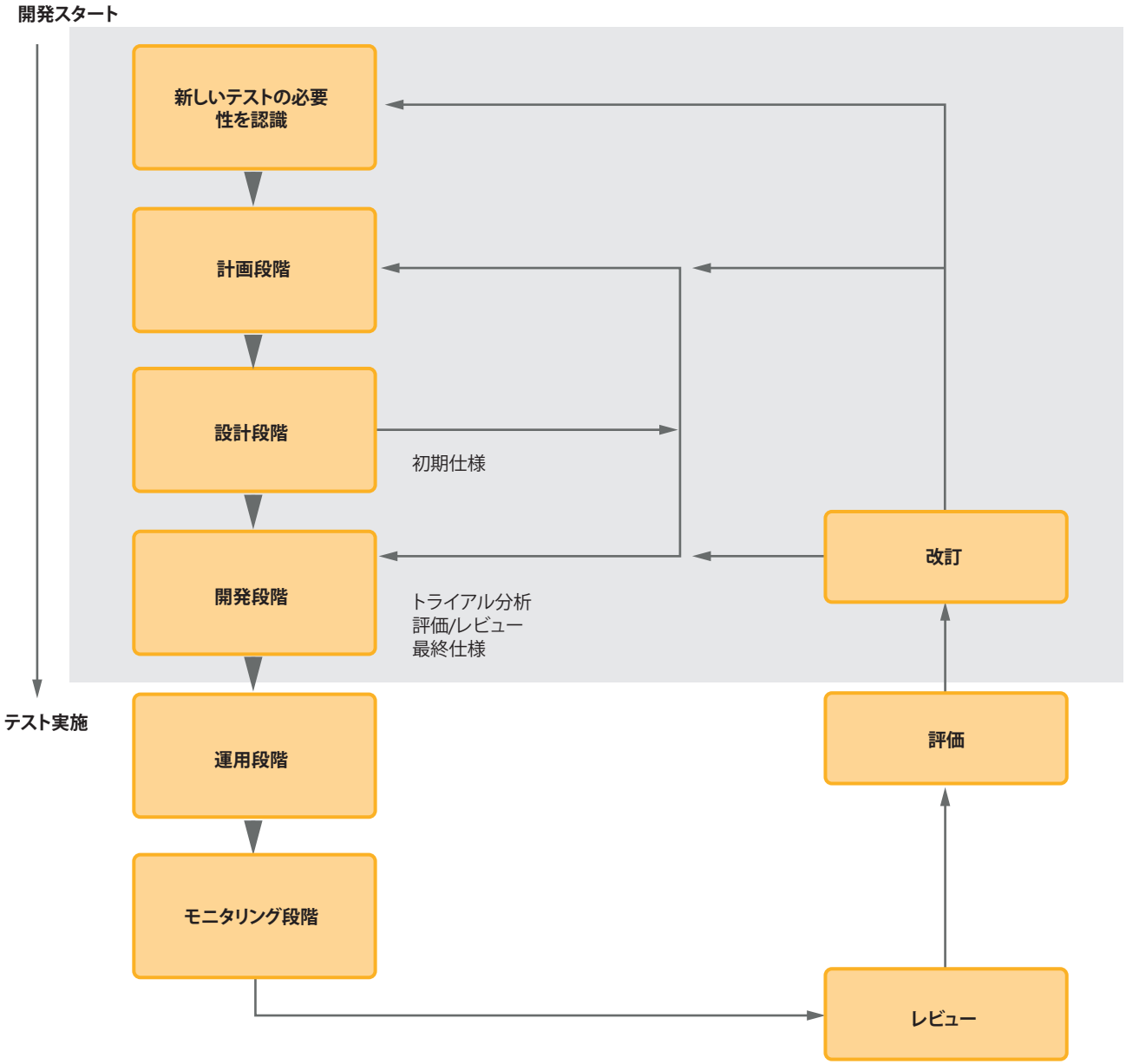
> 「試験ではいつも不安を感じていましたが、質問を聞いているうちに気持ちが楽になり、リラックスすることができました。だから私にとっては意外に簡単でした。」（参加者 ID 1721614）

参加者はコンピュータと話すことについて、さまざまな感情を持っていました。コンピュータと話すことは「本物の人と話すのと同じ」（参加者 ID 1750497）、またはスピーキング試験官と話すよりもストレスが少ない（参加者 ID 1741200）と感じている人もいました。また、「最近では携帯電話を使ったやりとりが当たり前になっている」（参加者 ID 1756618）ので、デジタル機器でのやりとりには慣れているとの回答もありました。参加者のごく一部（19.3%）は、現実の日常生活におけるオーラルコミュニケーションでは、情報交換（参加者 ID 1646082）や人の顔を見ること（参加者 ID 1714910）を求めるなど、依然として人と対話することを望む声もありました。

つまり、Linguaskill のコンテンツ制作を支える品質保証プロセスと、この大規模なトライアル調査から得られた知見は、Linguaskill スピーキングテストの内容が、受検者が現実世界で遭遇するであろうスピーキングタスクを総じて適切に表現したものであるという結論にいたりました。

対話相手との対話能力 (Brown 2003, Galaczi and Taylor 2018) は、通常では対面式のインタビューで評価されますが、Linguaskill スピーキングテストはそれができません。したがって、Linguaskill スピーキングテストのスコアを相互作用能力の直接的な尺度として説明することはできません。とはいえ、一人でのスピーキングパフォーマンスは、ある程度は相互作用型スピーキングのパフォーマンスを予測する可能性があり (Bernstein, Van Moere and Cheng 2010)、Linguaskill スピーキングテストは、さまざまなコミュニカティブ・スピーキング機能をカバーするように設計されていると主張できます。

図 5.Linguaskill のテスト開発サイクル（Cambridge English 2016）



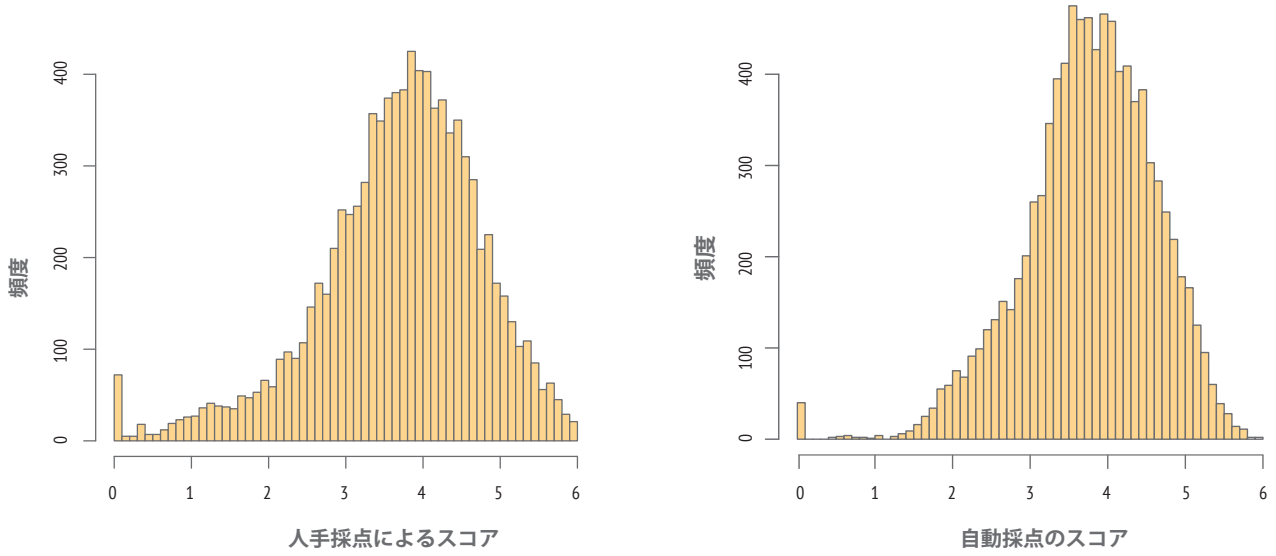
6.2 返答採点の妥当性の根拠

信頼性の高い採点は、受検者の目標とする言語能力を正確に推定するための基礎となります。自動スピーキング評価に関連する注意事項は、決まった解答のない発話の採点の信頼性に関連しています (Xu 2015)。Linguaskill で使用されている自動スピーキング採点の CASE の信頼性について説明する前に、まず、試験官の採点の信頼性について報告します。これは、a) Linguaskill スピーキングテストの大部分が今でも試験官によって採点されている、b) 自動採点は人間が採点した音声データで訓練されているため、訓練サンプルの中で最高の試験官を上回ることができない、という理由からです。

6.2.1 試験官の採点の信頼性

Xu and Gallacher（2017）は、Linguaskill スピーキングテストにおける人手採点の信頼性を調査する研究を行いました。大勢いる資格認定者登録リストから無作為に選ばれた 5 人の Linguaskill スピーキング試験官に、様々な習熟度レベルの 60 人の受検者によるテスト応答からなる共通データセットの採点を依頼しました。言い換えれば、テストの各パートは同じ 5 人の試験官によって採点が行われたということです。テストの各パートおよびテスト全体 (各パートの平均値) における人間による採点の信頼性を、級内相関係数 (ICC) を用いて推定しました。この係数は、1 つの返答に対する一人の採点が、同じ返答に対して他者が同じ採点をする度合いを示しています (Shrout and Fleiss 1979)。一般的に、ICC 値が 0.75 ～ 0.90 の間の値は信頼性が良いとされ、0.90 以上の値は信頼性が優れていると見なされます (Cicchetti 1994)。テストの各パートとテスト全体の ICC 値を表 2 に示します。同一試験官による採点の信頼性は、5 つのテストパートで

図 6. ヒストグラム　ハイブリッド採点導入以前の試験官と自動採点の素スコア



0.84 から 0.91 まで変化し、テスト全体では 0.91 であることから、タスクレベルでの人手採点の十分な信頼性と、テストレベルでの優れた信頼性を示すことが分かります。Brenchley（2020）は、3 人の試験官によって個別に採点された 204 セットの Linguaskill スピーキングテストの大規模なデータセットを用いて、試験官の間の信頼性を再調査しました。この調査研究では、テスト全体で単独採点 ICC が 0.90 であったことが報告されています。

6.2.2 自動採点の信頼性

自動採点に対する評価は、多くの場合、コンピュータ採点と人手採点との間の相関または一致を計算することによって行われます (例えば、Bernstein 他 2010、Wang 他 2018)。Jones, Brenchley and Benjamin (2020) は、自動採点の現行バージョンの評価研究を実施し、最初に自動採点単体の性能、つまりハイブリッド採点システムに組み込まれていない性能に焦点を当てました (次セクション参照)。評価は、実際の受検者である 9,286 人分の Linguaskill スピーキングテストのデータセットに基づいています。データセットの人手採点による CEFR グレードの分布は、Below A1 が約 1%、A1 が 5%、A2 が 13%、B1 が 36%、B2 が 35%、C1 以上が 10% でした。データセットには、スペイン語 (29%)、アラビア語 (26%)、ポルトガル語 (15%) の母語話者が最も多く含まれていました。

この調査研究では、自動採点と人手採点の素スコアに同じ CEFR カットオフ値を適用した場合、56.8% のテストで自動採点が試験官と同じ CEFR グレードを与えたことがわかりました。テストの 96.6% では、自動採点と人手採点の差は、同等の CEFR レベルか 1 レベル以内でした。テストの 3.4% では、自動採点と人間の採点者の差は CEFR の 1 レベル以上異なっていました (表 3)。運用テストでは、これらの誤った自動採点のスコアは、次のセクションで説明するように、試験官のスコアによって無効にされます。

この研究ではまた、自動採点と人間による素スコアの分布はほぼ重なっているものの (図 6)、自動採点はスコアの高い方では比較的厳しく、低い方では比較的甘くなっていることも明らかになりました (図 7)。繰り返しますが、これらの不正確さは、ハイブリッド採点の採用および自動採点の継続的な訓練と改善によって解決されます。自動採点の素スコアの二乗平均平方根誤差 (RMSE) は 0.64 で、CEFR バンドの半分程度でした。RMSE は残差 (自動採点予測誤差) の標準偏差であり、人間と機械の正確な一致が達成される対角の回帰直線の周りにデータポイントがどれだけ集中しているかを示す指標です (図 7)。

試験官と自動採点の一致に加えて、この研究では、応答の中にある英語以外の発話やチンプンカンプンな言葉を識別するために、音声認識の確信度の指標である言語品質スコア (セクション 4.3 参照) の有用性も評価しました。異常な発話態度を含むデータのサブセット（n = 284）に基づいて、正常な英語の発話は英語以外の発話よりも言語品質スコアが有意に高いことがわかりました（図 8 参照）。このスコアにカットオフ値を適用することで、データセットに含まれる 19 件の非英語話者の応答をすべて識別することができました。この研究によると、言語品質スコアは英語以外の発話に感度が良く、自動採点をゲームのように試そうとする受検者を認識するのに役立ちます。

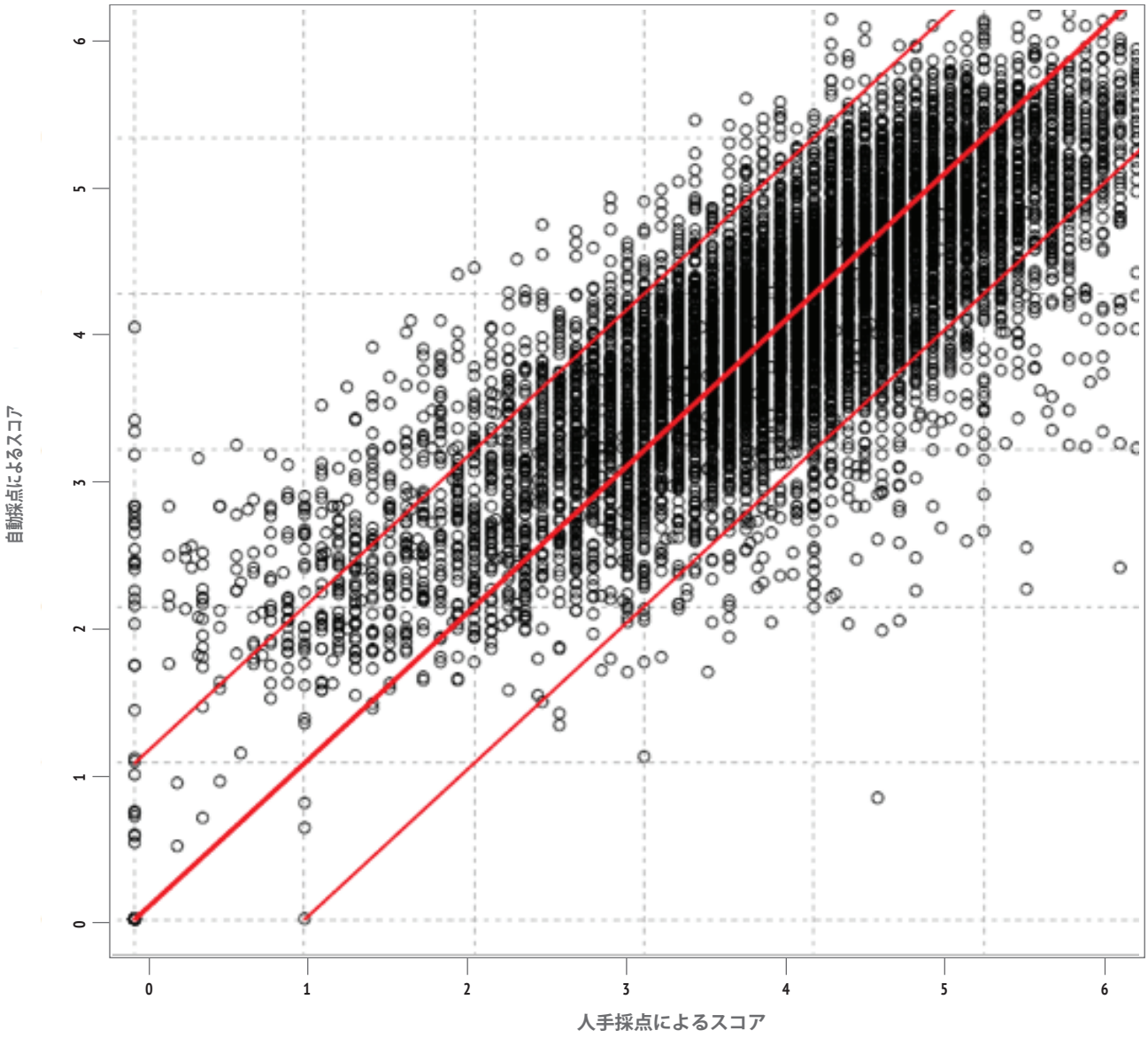
表 2.1 人の試験官の採点におけるクラス内相関係数（Xu and Gallacher 2017）

パート 1	パート 2	パート 3	パート 4	パート 5	テスト全体
0.84	0.87	0.90	0.88	0.91	0.91

表 3. 自動採点と人手採点による CEFR レベルの一致率（n=9,286）

人間と機械の一致	%
完全一致 (または差異なし)	56.8%
隣接一致 (または CEFR の差が 1 レベル以内)	96.6%
採点ミス (または CEFR が 1 レベル超えて異なる)	3.4%

図 7. 散布図　ハイブリッド採点前の試験官と自動採点の素スコア



6.2.3 ハイブリッド採点の信頼性

セクション 4.3 で説明したように、ハイブリッド採点では、自動採点が*採点ミスした*可能性のある応答は人間の試験官において確認され、自動採点と人間のスコアがスコアスケールで 1 CEFR レベル以上離れている可能性が高い場合、と定義されています。Linguaskill ハイブリッド採点モデルでは、評価品質、言語品質、音声品質、自動採点スコア（下限値）、自動採点スコア（上限値）など、自動採点によって生成された多くの特徴に規定が適用されます。各規定は、*言語品質スコアが 0.9 未満である場合*といった不等式で表されます。いずれかの規定に達した応答は、人間の試験官に引き渡されます。規定で使用される閾値（例 0.9）は、最適化のプロセスによって決定されます。

この制約付き最適化は、brute-force search　として知られる、しらみつぶし探索を使って行われました。各変数に対して、可能な閾値のセットが作成されました。例えば、言語品質スコアは　0　から　1　までの範囲なので、閾値のセットは　0, 0.01, 0.02, ... , 1　となります。5 つの閾値のすべての可能な組み合わせについて、9,286 件の Linguaskill スピーキングテストのデータセットに基づいて、最適な (最高の) 再現率が算出されました (Jones et al 2020)。パーセントで報告される再現率は、フラグの完全性を示しています。例えば、再現率 0.90 は、自動採点によって採点ミスされたすべてのテスト応答のうち、90% が規定の適用によって正常にフラグが立てられたことを意味します。

再現率とともに報告されることが多い統計用語は適合率で、これはフラグの精度を示す指標です。たとえば、適合率 0.90 は、フラグを立てられたすべてのテスト応答のうち、90% が自動採点によって実際に採点ミスされたことを意味します。適合率と再現率の間には常にトレードオフがあります。再現率の値が高いと適合率の値が低くなり、その逆も同様です。Linguaskill の仕様を設計する

図 8. 英語スピーキングテストと英語以外のスピーキングテストの言語品質スコア比較

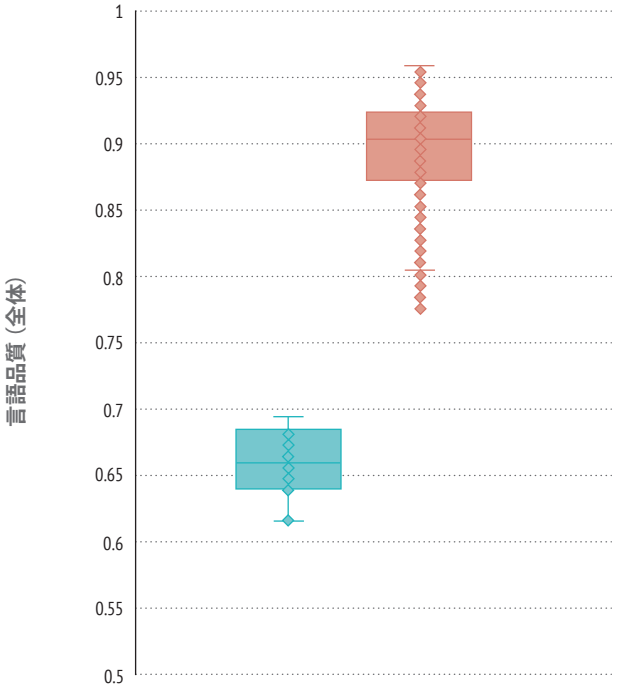
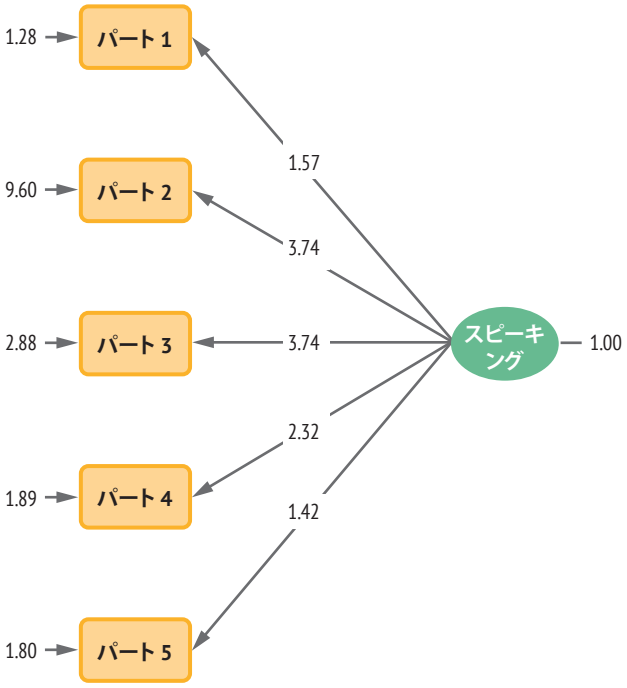


図 9. 一因子モデル（Xu and Seed 2017）



際に、信頼性の低い自動採点スコアが受検者に提供されるのを防ぐために、高い再現率の値を追求しました。

採点の高い信頼性を重視しているため、当初は、人間の試験官に対するテスト応答の大部分を確認させる代価として、再現率 0.96 という閾値を選択しました。その結果、再現率が高くなり、自動採点の RMSE は 0.16 と小さく、人間と機械の一致率は CEFR　グレードで 95.6% の完全一致、隣接するグレードでは 100% 一致で非常に優れていました。 しかしながら、自動採点を継続的に改良し、閾値を評価することで、試験官が採点するテスト応答の割合を減らすようにしています。

6.3 テスト結果解釈の妥当性の根拠

言語テストの結果の解釈は、目標言語能力に関する構成理論によって支持される必要があります。一方では、言語に関する構成理論は、テスト設計に情報を与え、テストスコアに意味を与えて、テストスコアの差異を説明するために、テスト開発者によって選択されます。
他方で、テストの検証は、観察されたテストデータが、スコア解釈のために選択された理論を確認したり、反論したりするという意味で、理論検証のプロセスでもあります（Cronbach and Meehl 1955)。

構成理論は、CEFR のように言語発達の過程を詳細に記述した言語能力記述子の集合であっても構いません。あるいは、言語能力の構成に関する憶測である場合もあります。Linguaskill スピーキングテストの提案されたスコア解釈をサポートするための妥当性の根拠は、標準化と因子分析によって収集されてきました。前者はテストの成績をスピーキング能力の向上に関する理論と結びつけているのに対し、後者はテストの目標となるスピーキングの構成要素の基礎構造を検証しています。

6.3.1 標準化

Linguaskill スピーキングテストは CEFR に基づいたテスト結果をレポートしているため、定期的に標準化の予行演習を行い、テスト結果を CEFR フレームワークに関連付けるようにしました。この調整により、試験利用者は CEFR が提供する言語能力記述子を参照することで、より広い文脈で試験結果を解釈することができます。

標準化とは、試験において *1 つ以上のカットスコアを設定するプロセス*を指します (Cizek and Bunch 2007, p.13)。Linguaskill の場合、カットスコアは、受検者を CEFR の習熟度レベルに沿って 6 つのグループ A1 未満、A1、A2、B1、B2、C1 以上に分けて使用されます。Linguaskill スピーキングテストに関する直近の標準化の予行演習は、言語テストを CEFR に関連付ける手引書 (Council of Europe 2009) の中で勧められた改訂 Bookmark 法に従い、Lopes and Cheung (2020) によって実施されました。

6.3.2 因子構造

標準化に加えて、Linguaskill スピーキングテストの基礎構造を調べるために因子分析が行われました。5 つのテストパートで評価される能力は一次元的であり、単一の包括的なスピーキングの構成がテストで評価されることを意味しているという仮説が立てられました。しかし、テストのパート 2 にある音読は、他の 4 つのパートの自発的スピーキングタスクとは少し異なる構成要素を評価するように見えます。

上記の仮説を検証するために、Xu and Seed (2017) は、試験官のみが採点した 3,250 のスピーキングテストを対象に、設問レベルの検証的因子分析を実施しました。この調査では、一因子モデル (図 9）がデータに適合し、その結果、CFI（Comparative Fit Index）値が 0.99、NNFI（Non-Normed Fit Index）値が 0.98、RMSEA（Root Mean

Square Error of Approximation）値が 0.08 であったことがわかりました。通常、CFI や NNFI の値が 0.90 以上、または RMSEA の値が 0.80 以下であれば、十分なモデル適合度があることを示しています (Sawaki 他 2009)。この調査結果は、1 つのスピーキング構成要素が 5 つのパートすべてのテスト成績を説明することができたことを示唆します。これにより、5 つのパートの平均を算出し、全体のテストスコアを生み出す手法の裏付けが得られました。

しかし、パート 2 の音読に関連した残差 (誤差) 用語は他のパートに関連したものよりも相対的に大きくなっていることも指摘されています。研究者らは、これをスピーキング評価において音読と自発的発話を区別するエビデンスの一つと考え、コミュニケーション能力を測定するために制約のある発話タスクだけを単独で使用することに対して注意を促しました。

References

AERA, APA and NCME (2014) *Standards for educational and psychological testing*, Washington, DC: AERA.

APA, AERA and NCME (1974) *Standards for educational and psychological tests*, Washington, DC: APA.

Bachman, L F and Palmer, A S (1996) *Language testing in practice*, Oxford: Oxford University Press.

Bachman, L F and Palmer A S (2010) *Language assessment in practice*, Oxford: Oxford University Press.

Bernstein, J, Van Moere, A and Cheng, J (2010) Validating automated speaking tests, *Language Testing* 27 (3), 355–377.

Brenchley, M (2020) *Re-examining the reliability of human marking in the Linguaskill Speaking test*, Cambridge Assessment English internal research report.

Brown, A (2003) Interviewer variation and the co-construction of speaking proficiency, *Language Testing* 20 (1), 1–25.

Cambridge Assessment English (2016) *Principles of good practice: Research and innovation in language learning and assessment*, Cambridge, UK: Cambridge Assessment.

Chapelle, C A, Enright, M K and Jamieson, J M (2010) Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice* 29 (1), 3–13.

Cheng, L (2014) Consequences, impact, and washback, in Kunnan, A J (Ed.) *The Companion To Language Assessment* (Vol. III) Chichester, West Sussex: John Wiley and Sons, 1,130–1,146.

Chun, C W (2006) An analysis of a language test for employment: The authenticity of the PhonePass test, *Language Assessment Quarterly* 3 (3), 295–306.

Cicchetti, D V (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology, *Psychological Assessment* 6 (4), 284–290.

Cizek, G J and Bunch, M B (2007) *Standard setting: A guide to establishing and evaluating performance standards on tests*, Thousand Oaks, CA: Sage.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Strasbourg: Council of Europe.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual*, Strasbourg: Council of Europe.

Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, teaching, assessment (Companion volume with new descriptors)*, Strasbourg: Council of Europe.

Cronbach, L J (1988) Five perspectives on validity argument, in Wainer, H and Braun, H I (Eds) *Test validity*, Hillsdale, NJ: Lawrence Erlbaum, 3–17.

Cronbach, L J and Meehl, P E (1955) Construct validity in psychological tests, *Psychological Bulletin* 52 (4), 281–302.

Fan, J (2014) Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach, *Language Testing in Asia* 4 (6), 1–17.

Galaczi, E and Taylor, L (2018) Interactional competence: Conceptualisations, operationalisations, and outstanding questions, *Language Assessment Quarterly* 15 (3), 219–236.

Haertel, E H (2006) Reliability, in Brennan, R L (Ed.) *Educational Measurement* (4th edn), Westport, CT: Praeger, 65–110.

Jones, E, Brenchley, M and Benjamin, T (2020) *An investigation into the hybrid marking model for the Linguaskill Speaking test*, Cambridge Assessment English internal research report.

Kane, M T (2013) Validating the interpretations and uses of test scores, *Journal of Educational Measurement* 50 (1), 1–73.

Knill, K, Gales, M, Kyriakopoulos, K, Malinin, A, Ragni, A, Wang, Y and Caines, A (2018) Impact of ASR Performance on Free Speaking Language Assessment, *Proc. Interspeech* 2018, 1,641–1,645. https://doi.org/10.21437/Interspeech.2018-1312

Linacre, J M (1989) *Many-facet Rasch measurement*, Chicago: MESA Press.

Lopes, S and Cheung, K (2020) *Final report on the December 2018 standard setting of the Linguaskill General papers to the CEFR*, Cambridge Assessment English internal research report.

Lu, Y, Gales, M, Knill, K, Manakul, P, Wang, L and Wang, Y (2019) Impact of ASR performance on spoken grammatical error detection, *Proceedings of the Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, September 2019, 1,876–1,880. https://doi.org/10.21437/Interspeech.2019-1706

Malinin, A, Ragni A, Knill, K and Gales, M (2017) Incorporating Uncertainty into Deep Learning for spoken language assessment. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* 2, 45–50. https://doi.org/10.18653/v1/P17-2008

Messick, S (1989) Validity, in Linn, R L (Ed.), Educational measurement (3rd edn), New York: Macmillan, 13–103.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13 (3), 241–256.

Sawaki, Y, Stricker, L J and Oranje, A H (2009) Factor structure of the TOEFL Internet-based test, *Language Testing* 26 (1), 5–30.

Shrout, P E and Fleiss, J L (1979) Intraclass correlations: Uses in assessing rater reliability, *Psychological Bulletin* 86 (2), 420–428.

van Dalen, R C, Knill, K and Gales, M (2015) Automatically grading learners' English using a Gaussian process. *SLaTE 2015: Workshop on Speech and Language Technology in Education*, 7–12. https://www.isca-speech.org/archive/slate_2015/sl15_007.html

Wang, Y, Gales, M J F, Knill, K M, Kyriakopoulos, K, Malinin, van Dalen, R C and Rashid, M (2018) Towards automatic assessment of spontaneous spoken English, *Speech Communication* 104, 47–56. https://doi.org/10.1016/j.specom.2018.09.002

Weir, C J (2005) *Language testing and validation: An evidence-based approach*, Basingstoke: Palgrave Macmillan.

Xi, X (2010) Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing* 27 (3), 291–300.

Xi, X, Schmidgall, J and Wang, Y (2016) Chinese users' perceptions of the use of automated scoring for a speaking practice test, in Yu, G and Jin, Y (Eds) *Assessing Chinese learners of English: Language constructs, consequences and conundrums*, Basingstoke, Hampshire: Palgrave Macmillan, 150–175.

Xu, J (2015) *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech*, unpublished doctoral dissertation, Iowa State University, Ames, IA.

Xu, J and Gallacher, T (2017) *Linguaskill Speaking trial report*, Cambridge Assessment English internal research report.

Xu, J and Seed, G (2017) *Automated speaking tests: Merging technology, assessment and customer needs*, paper presented at the Language Testing Forum 2017, Huddersfield, UK.

お問い合わせ

私たちケンブリッジ大学英語検定機構 (Cambridge Assessment English) は、ケンブリッジ大学の一部門である非営利機関です。英語を学び、そのスキルを世界に向けて証明することをお手伝いします。

私たちにとって、英語を学ぶことは、試験や成績だけが目的ではなく、コミュニケーションに自信をもち、生涯にわたり、豊かな経験と機会にアクセスすることだと考えます。

適切なサポートがあれば、言語の習得はワクワクする旅になります。私たちはあなたのすべてのステップをサポートします。

Cambridge Assessment English
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

cambridgeenglish.org
/cambridgeenglish
/cambridgeenglishtv
/cambridgeeng
/cambridgeenglish
/cambridge-assessment-english

本書の掲載内容は、2020 年 6 月発行の英語版を翻訳したものです。