Linguaskill

ライティングトライアルレポート　2017 年 6 月

ケンブリッジ大学英語検定機構

調査チーム

Kevin Cheung

Jing Xu

Gad Lim

目次

リンガスキルとは？

～マルチレベルのオンライン英語能力テスト～

リンガスキルは、個人やグループの受検者の英語力レベルをさまざまな機関が迅速にチェックできる、便利なオンラインテストです。1 種類のテストで、あらゆるレベルの受検者の英語力を評価することができます。

試験は迅速・簡単に、かつ費用効率よく実施できるように設計されており、しかも信頼性の高いテスト結果を提供いたします。テスト結果レポートには Cambridge English スケールや、言語運用能力を示す国際基準である CEFR（Common European Framework of Reference）[1]の該当レベルが表示されます。

リンガスキルのライティング

ライティングテストは 45 分間で、2 パートで構成されています。

・パート 1：電子メール作成（50 words 以上）。与えられた情報を使って、既知の読み手に向けた電子メールを作成します。
・パート 2：長文作成（180 words 以上）。未知の読者に向けて、記事や報告書などの長文を作成します。

テストはコンピュータにより自動採点されます。これは実質的に、専門の採点者（人間）によって採点された学習者のテスト解答を収集した膨大なデータから採点方法を学習した、一連のコンピュータ・アルゴリズムです。

リンガスキルのトライアル

リンガスキルのライティングテストのトライアルは、23 カ国から総勢 3,918 名の英語学習者に参加いただき、2016 年 12 月から 2017 年 2 月にかけて実施されました。

トライアルの目的は、以下のとおりです。

・リンガスキルによって、あらゆるレベルの受検者の言語熟達度を正しく示しているかの評価
・コンピュータ処理されたテストスコアの信頼性調査
・基準設定による、リンガスキルのライティングスコアと CEFR レベルの整合性確保

リンガスキルのライティングテストは、オンラインテストプラットフォーム「Metrica」で実施されます。Metrica により下記のデータが収集されました。

・受検者のテスト解答
・受検者が取り組んだ問題
・コンピュータが自動採点したテストスコア

[1] 詳しくは www.cambridgeenglish.or/cefr を参照のこと。

コンピュータ処理されたテストスコアの信頼性調査のため、テスト解答の一部を選び、人間の採点官[2]による検証を実施しました。
さらに、コンピュータによる自動採点の信頼性についての洞察を深め、リンガスキルのライティングスコアと CEFR レベルの整合性を図るため、上記とは別の一部解答について、ライティング評価の専門家[3]による検証を行いました。

ご存知ですか？

大半の受検者（73%）は、リンガスキルのライティングテストを受けてよかったと感じています。よくなかったと答えた受検者はごくわずか（6%）でした。
受検者の大部分（84%）は、このテスト結果が自身の英語ライティング能力[4]を反映していると感じています。

トライアル結果

リンガスキルは、あらゆるレベルの受検者について言語熟達度を正しく示しているか？

主な所見

リンガスキルの全てのタスクやプロンプトは、CEFR が規定する言語熟達度全範囲におけるライティング・パフォーマンスを有効的に引き出している。
リンガスキルはマルチレベル・テストです。したがって、達成可能なスコア水準に応じて異なるライティング解答を、テストタスクにより引き出せることが重要です。
言語熟達度の低い受検者が取り組むことのできないタスクは、マルチレベル・テストに含まれるべきではありません。同様に、言語熟達度の高い受検者がその能力を立証できないタスクもまた、マルチレベル・テストの妥当性を損ないます。

方法：トライアルで提出された 3,918 名全員のライティング解答を、コンピュータの自動採点で評価しました。また、任意の一部のテストスクリプトをケンブリッジ大学英語検定機構の採点官が審査し、コンピュータの自動採点で最低点となった答案が CEFR のライティングレベル A1 以下に相当し、最高点となった答案が CEFR のライティングレベル C1 以上に相当することを確認しました。トライアルでは、リンガスキルのタスク 1 とタスク 2 で異なるプロンプトが実施されました。
所見：コンピュータで自動採点されたテストスコアによると、トライアルで使用された全てのテストプロンプトやタスクが、全対象範囲のパフォーマンスから解答を引き出していました。採点官（人間）に

[2] 評価スケールの活用方法、ならびに採点がケンブリッジ大学英語検定機構の英語能力テストの基準に合致していることを保証するための基準化や定期的な見直しに関する訓練を受けたライティングの専門評価者です。
[3] 専門家委員会は、CEFR レベルと整合する能力記述を活用したライティング解答の評価に関して、経験豊富な 6 人のケンブリッジ大学英語検定機構の研究マネージャーと評価マネージャー、ならびに 5 人の熟達したライティング採点官およびスピーキング試験官で構成されました。テストと CEFR レベルの整合性を図るための基準設定手続きは概して、テストスコアと外的基準（2009 年欧州評議会）との関連付け方法を決定する専門家委員会によって実施されます。
[4] ライティングテストの終了時、受検者に Metrica によるオンライン調査への回答協力を依頼しました。その結果、3,026 名（77%）の受検者から回答が得られました。

よる検証はこの調査結果を裏付けており、リンガスキルがテストの対象となる全範囲（A1～C1 以上[5]）からライティング・パフォーマンスを有効的に引き出していることを確認しました。
また、自動採点されたスコアは正規分布しており、リンガスキルのタスクが言語熟達度の全範囲から解答を引き出すよう適切に目標設定していることを示しています。トライアルで引き出された解答の平均水準を調査するため、自動採点で最も頻度の高かったスコアを付けられた一部のスクリプトを採点官が検証しました。これにより、リンガスキルのライティングタスクが最も多く引き出した答案は B1 レベルであることを確認しました。

コンピュータの自動採点によるテストのランク付けやスコアには信頼性はあるのか？

主な所見

コンピュータの自動採点によるテストスコアの信頼性を調査するためのトライアルの一環として、2 つの研究が実施された。得られた結果は以下のとおり。

- コンピュータの自動採点によるテストスコアと採点官（人間）の採点したテストスコアの平均との間には、強い正の相関関係がある。
- コンピュータの自動採点と採点官（人間）の採点とでは、テストの解答を似た順序（最高品質の解答から最低品質の解答へ）でランク付けする。

信頼性の研究 1：コンピュータの自動採点によるテストスコアは信頼性が高いか？

本研究では、コンピュータの自動採点によるテストスコアと採点官（人間）の採点によるテストスコアにどのような違いが生じるかを評価しました。

方法：5 人の経験豊富なケンブリッジ大学英語検定機構の採点官が、トライアル参加者の答案をそれぞれ 60 名分採点しました。これらの答案はコンピュータの自動採点で全範囲のスコアを網羅しています。
採点官は 6 段階レベルの能力記述による包括的な採点スキームを活用しました。Task Achievement（タスクの達成度）、Language Resource（言語資源）、Text Organization（文章構成）を基準に、リンガスキルのタスクごとにスコアが付けられました。テスト解答に 2 つのレベルの特徴をほぼ同水準で満たしている場合は、採点官は 6 段階の各レベルの中間スコアを付けるよう指示されました。これにより、有効なテスト解答に対して 11 段階のスケールが形成されました。さらに、採点官は意味のない解答の場合、または内容が主題からずれている場合に、そのテスト解答のスコアを 0 とすることができました。
テストスコアには、5 人の採点官全員が出したスコアの平均値を用いました。集計された採点官（人間）のスコアは、採点官の採点のばらつきによる影響が低減されているので、1 人の採点官が付けたスコアよりも信頼性の高い指標です。

[5] タスク 1 の最高スコアの解答で、採点官の判定によって C2 レベルのライティングであることを立証するものは比較的少なかったのに対して、タスク 2 の最高スコアの解答は C2 レベルでした。受検者が達成可能な最高レベルは「C1 以上」として報告されており、2 つのタスクで引き出されたライティングレベルはリンガスキルにとって適切であると言えます。

所見：スピアマンの順位相関係数計算は、コンピュータの自動採点によるテストスコアは採点官グループ（人間）の集計スコアと強い正の相関があることを示しています。

- **リンガスキル・タスク 1：**コンピュータが自動採点したテストスコアと 5 人の採点官（人間）全員の集計スコアの相関は ρ=0.82 である。
- **リンガスキル・タスク 2：**コンピュータが自動採点したテストスコアと 5 人の採点官（人間）全員の集計スコアの相関は ρ=0.88 である。

次に、個々の参加者のテストスコアを総計し、両タスクにおけるパフォーマンスを反映させました。コンピュータの自動採点によるテストスコアの総計は、個々のテストタスクが単独で評価された場合に比べて、採点官（人間）が付けた総合スコアとの間にさらに強い正の相関がありました。

- **リンガスキルのライティング総合スコア：**両タスクにおいてコンピュータが自動採点したテストスコアと 5 人の採点官（人間）全員の総合スコアの相関は ρ=0.90 である。

比較のため、採点官（人間）間の一貫性もスピアマンの相関分析により検証しました。
テストスコア全体を見ると、個々の採点官（人間）のスコアの相関は 0.84～0.95 となっており、平均値は ρ=0.91 でした。これは、コンピュータの自動採点で計算された相関と酷似していました。
これらの結果から、コンピュータの自動採点は、採点官（人間）と同様のパフォーマンスを示しており、同じ採点スキームを使用した一部の採点官（人間）のパフォーマンスよりも優れていることがわかりました。したがって、自動採点によるスコアが正確かつ信頼できるものであると確信しています。

表 1：両タスクにおいて採点官（人間）が付けた平均スコアの相互相関

	採点官 1	採点官 2	採点官 3	採点官 4	採点官 5
採点官 1	–	0.92	0.91	0.84	0.91
採点官 2	–	–	0.94	0.88	0.91
採点官 3	–	–	–	0.9	0.95
採点官 4	–	–	–	–	0.91
他の採点官の平均	0.90	0.91	0.93	0.88	0.92

信頼性の研究 2：コンピュータの自動採点によるテスト解答の得点ランキングは信頼性が高いか？
本研究では、コンピュータの自動採点および専門家（人間）の採点が、テスト解答の品質（最高から最低まで）に関して整合性ある判定をしているのか調査することにより、コンピュータの自動採点の信頼性についての洞察を深めました。

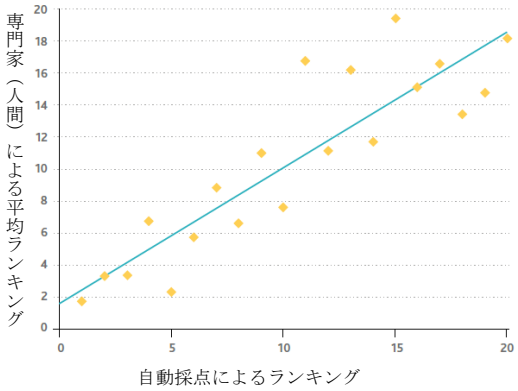
また、コンピュータによる自動採点の信頼性に関する根拠を提供するのに加えて、リンガスキルと CEFR との整合性を図る基準設定演習も支援しました。詳細は 8 ページに記載しています。

方法：10 人のライティング評価の専門家による委員会[6]は、リンガスキル・タスク 1 の 20 件の解答とリンガスキル・タスク 2 の 20 件の解答の品質を（最高から最低まで）ランク付けしました。これらのテスト解答は、コンピュータの自動採点による全スコア範囲を網羅しています。委員会は解答品質のランク付けに困難はなかったと報告しています。
前述同様に、ランキングは個々の委員会メンバーの採点のばらつきの影響を低減し、各解答の「真のランキング」を正確に予測するために、全専門家（人間）間で平均化されました。

所見：コンピュータの自動採点がさまざまな水準のライティング解答を有効的に識別しており、テスト解答の品質は望ましい範囲内に分布したことについて、委員会全員の認識が一致しました。
スピアマンの相関計算は、コンピュータの自動採点によるテスト解答のランキングは専門家委員会によるテスト解答の総合ランキングと強い正の相関があることを示しています。

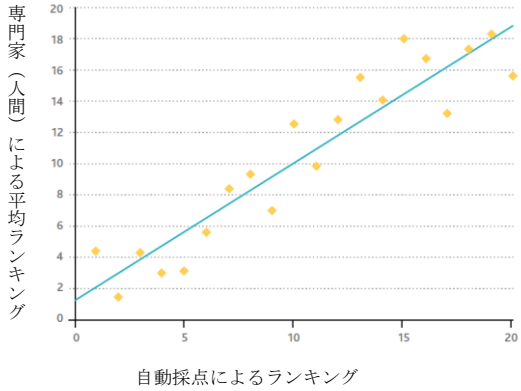
- **リンガスキル・タスク 1：**コンピュータの自動採点によるテスト解答のランキングと専門家委員会の総合ランキングとの相関は 0.88 であった。
- **リンガスキル・タスク 2：**コンピュータの自動採点によるテスト解答のランキングと専門家委員会の総合ランキングとの相関は 0.92 であった。[7]

図 1：リンガスキル・タスク 1 におけるコンピュータの自動採点によるテスト解答のランキングと専門家委員会による平均ランキングの相関散布図



[6] ワークショップのファシリテーター（進行役）が、この演習に参加する前にコンピュータで自動採点されたスコアを見直していたため、自動採点スコアのランキングは今回の報告に含めていません。
[7] タスク 1 でランク付けされた 20 のスクリプトは、タスク 2 でランク付けされた 20 のスクリプトとは別の受検者によって作成されたものです。したがって、総合ランキング作成上これらをリンクさせることはできません。

図 2 : リンガスキル・タスク 2 におけるコンピュータの自動採点によるテスト解答のランキングと専門家委員会による平均ランキングの相関散布図



比較のため、専門家（人間）間の一貫性がここでもスピアマンの相関により検証されました。多くの演習が判断に基づいたものであるのと同様に、個々の委員会メンバーによるランキング順もさまざまでした。専門家間において、完全にランキングの一致したペアはありませんでした。

・リンガスキル・タスク 1: 個々の委員会メンバーによるランキングの相関は 0.83〜0.96 の範囲であった。

・リンガスキル・タスク 2: 個々の委員会メンバーによるランキングの相関は 0.75〜0.97 の範囲であった。

前回の研究と同様、コンピュータの自動採点の相関の程度は、専門家（人間）同士を比較した場合に見られる範囲内でした。つまり、コンピュータの自動採点と専門家（人間）の採点の整合性は、専門家（人間）同士に見られる整合性と同程度であると言えます。

結論

これら 2 つの研究から得られた結果により、コンピュータの自動採点による答案評価の信頼性は高く、採点官（人間）と同様にテストスコアを算出し、テスト解答のランク付けを行うと確信するに至りました。

リンガスキルのライティングの CEFR レベルとの整合性

次に、リンガスキルのライティング項目を CEFR レベルに関連付け、リンガスキルのスコアを CEFR レベルに照らしてレポートできるよう、体系的なプロセスを使用しました。基準設定のためのワークショップやその準備を通じ、CEFR レベルに整合したリンガスキルのライティング・パフォーマンスのレベル決定に取り組みました。このプロセスは、語学能力試験を CEFR に関連付けるためのマニュアル（欧州評議会、2009 年）により通知されました。

方法：11 人のライティング評価専門家による委員会（ワークショップ・ファシリテーターを含む）[8]が、以下の体系的プロセスに参画しました。

1	**ワークショップ事前演習**
	基準設定には周知プロセスを含むべきであるとの欧州評議会勧告に従って、委員会メンバーにはワークショップ前に教材と演習課題が配布されました。 リンガスキルのタスクおよび受検者の解答を確実に理解するために、ワークショップ事前演習が実施されました。委員会のメンバーは、リンガスキルのタスクごとに個別に 20 名分の解答のランク付け（最高品質から最低品質まで）[9]を求められました。
2	**基準設定のワークショップ：ワークショップ事前演習の検証**
	ワークショップの冒頭、委員会メンバーの CEFR レベルに関する見識が最新のものであることを確認するために、ワークショップ事前演習の結果に関する協議が行われました。 委員会メンバーは互いのランキングならびにコンピュータの自動採点によるランキングを検証しました。委員会メンバーは、一部のテスト解答を詳細に分析し、スクリプトの品質比較の判定理由を考察し、これがいかに CEFR の能力記述と関連付くかを検証しました。
3	**基準設定のワークショップ：基準の設定**
	委員会では、解答事例を選び、リンガスキルのライティングのスコアを CEFR レベルとして確信をもってレポートできるように、コンピュータの自動採点のカットスコア（分割点）を設定するためにブックマーク・メソッドを活用しました。 スコアと CEFR レベルとの整合を図るため、それぞれのタスクの判定を 2 ラウンド行うことにより、委員会メンバーが各 CEFR レベルの境界について合意することができました。
	3.1　第 1 ラウンドの判定 ・委員会メンバーは、テスト解答の自身のランキングを検証しました。そして、各 CEFR レベルの言語熟達度に相当する最初の解答を特定するように求められました。A1 未満のレベルと A1 レベルの境界にあると判定された解答の選定から始まり、B2 レベルと C1 レベル以上の境界に達するまで進められました。

[8] 全委員会メンバーはケンブリッジ大学英語検定ならびにテスト開発活動の検証を通じて、CEFR に精通しています。
[9] 演習の結果は、本レポートの前節（6 ページ）に掲載されています。

	・判定は個別に匿名で実施されました。次に、11 人の委員会メンバー全員の判定の取りまとめを行いました。委員会メンバーの中で意見が相違する部分については、今後の協議事項として確認しました。
	3.2　第 2 ラウンドの判定 ・委員会メンバーは、コンピュータの自動採点によるテスト解答のランキングを検証しました。これらの解答は 1 から 20 までの順位で並べられ一冊にまとめられました。委員会メンバーは、その冊子の中で、各 CEFR レベルの言語熟達度に相当する最初の解答を特定するように求められました。 ・CEFR レベルごとに、委員会メンバーは、解答が順位で並べられている冊子の最初のページから始め、全てのテスト解答を順位に沿って検証するように指示されました。これにより、委員会メンバーが同一の答案を異なる CEFR レベルの境界に相当する解答として選ぶ可能性が生じました。 ・判定はここでも個別に匿名で実施されました。各 CEFR レベルの言語熟達度に相当する最初の解答を特定するため、CEFR レベルごとに、11 人の委員会メンバー全員の平均カットスコアが算出されました。 ・これは、レベルごとの CEFR 能力記述と並行して、委員会全体で検証されました。レベルごとに確実に合意がなされ、最終的な基準として採用することになる自動採点スコアを特定するために、協議を行いました。

基準設定演習の結果

第 1 ラウンドでの委員会メンバーの判定は、特に A2 レベルと B1 レベルの境界ならびに B2 レベルと C1 レベルの境界においてばらつきがありました。第 2 ラウンドでは、ランキングの標準偏差の縮小に見られるように、合意が大幅に進展しました。

表 2 : 解答が順番に並べられている冊子の中で各 CEFR レベルの言語熟達度に相当する最初の解答についての委員会メンバーの判定の標準偏差[10]

リンガスキル・タスク 1	**A1**	**A2**	**B1**	**B2**	**C1**
ラウンド 1 標準偏差	1.8	1.82	2.76	2.38	2.41
ラウンド 2 標準偏差	1.15	1.28	2.02	1.4	2.64
リンガスキル・タスク 2	**A1**	**A2**	**B1**	**B2**	**C1**
ラウンド 1 標準偏差	-	1.4	2.95	2.84	3.39
ラウンド 2 標準偏差	0.45	1.21	1.75	1.21	1.97

プロセスを通じてコンセンサスが形成され、またこの体系的なアプローチを活用することにより、目標としていた 5 つの CEFR レベル（A1、A2、B1、B2、C1 以上）に対応するカットスコアを特定しました。

[10] この標準偏差は、委員会によって検証されたスクリプトのランキングによるもので、この単位は 20 のスクリプトの中から順位付けられたスクリプトの数を示しています。

委員会メンバーがコンピュータの自動採点による順位と異なる順位の判定を下した例が、わずかながらありました。基準設定では個々の専門家の独立した判定を重視したため、自動採点と異なる順位については正当な判断として扱われました。とはいえ、それぞれのタスクにおいて順位付けに食い違いが生じた件数は少なく、コンピュータの自動採点によるスコアの頑健性への信頼はさらに高まりました。

結論

概して、本レポートから得られた所見は、リンガスキルのライティングスコアが自信を持って使用できる、CEFR レベルと有意に関連付けられた指標であることを示しています。

ご連絡先

ケンブリッジ大学英語検定機構

1 Hills Road

Cambridge

CB1 2EU

United Kingdom

本書は、本書印刷時点である 2018 年 2 月における情報をもとに作成しています