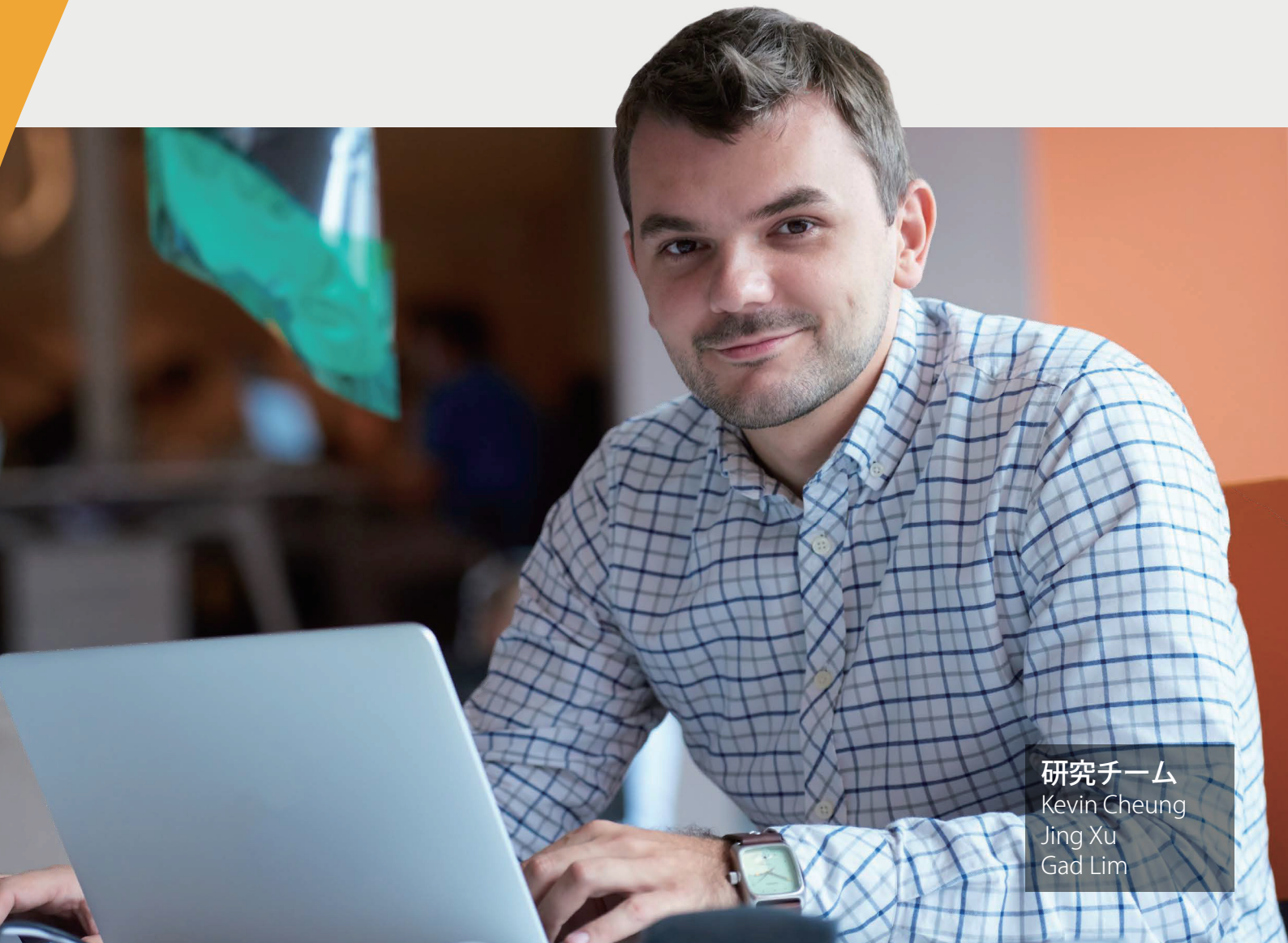
Cambridge Assessment English

ケンブリッジ英語検定4技能CBT
（Linguaskill　リンガスキル）

ライティングのトライアルに関する報告書

2017 年 6 月



研究チーム
Kevin Cheung
Jing Xu
Gad Lim

Linguaskill リンガスキル
ケンブリッジ英語検定4技能CBT

目次

Linguaskill リンガスキル とは ……3
リンガスキルのトライアル……4
トライアルの結果……5
リンガスキルのライティングテストで
CEFR のレベルを判定 ……9

複数レベル対応のオンライン英語テスト

ケンブリッジ英語検定4技能CBT（Linguaskill リンガスキル）とは

リンガスキルとは、迅速で便利なオンラインテストです。個人だけでなく、団体受検にも対応しているため、学校や大学、企業などの組織が生徒や従業員などの英語レベルを確認するのに役立ちます。このテストだけで、受検者のさまざまな英語能力を評価できます。

迅速、簡単に実施でき、コスト効果に優れるこのテストでは、信頼性の高い結果が得られます。テストの結果は、Cambridge English スケールおよび、言語能力を説明する国際的基準であるヨーロッパ言語共通参照枠 ）を使用して報告されています。[1]

リンガスキルのライティング

ライティングテストの所要時間は 45 分で、次の 2 つのパートで構成されています。

- パート 1：E メール（50 単語以上）。与えられた情報を使って、知っている人に E メールを書きます。
- パート 2：長文のライティング（180 単語以上）。記事や未知の相手への報告書など、長文を書きます。

テストの採点は、コンピューターの自動採点ツールで行われます。このツールは、経験豊かな採点者が採点した多量の学習者の解答から採点方法を学んだ、コンピューターの一連のアルゴリズムと言えます。

1 www.cambridgeenglish.org/cefr

ケンブリッジ英語検定4技能CBT（Linguaskill リンガスキル）のトライアル

2016 年 12 月から 2017 年 2 月にかけて、リンガスキルのライティング分野のトライアルテストが実施されました。世界 23 か国から、合計 3,918 名の英語学習者がこのトライアルテストを受検しました。

トライアルの目的：

- 受検者が自分の総合的な英語能力をリンガスキルで実証できるかどうかを判断する
- コンピューターによって生成されるテストスコアの信頼性を見極める
- リンガスキルのライティングのスコアを標準的な設定で CEFR のレベルと一致させる

リンガスキルのライティングテストは、「Metrica」と呼ばれるオンラインのテストプラットフォームを通じて配信されます。Metrica を通じて次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
- コンピューターで自動採点されたテストのスコア

コンピューターが算出したテストスコアの信頼性調査では、採点官[2]が一部のテスト解答を見直しました。

次に、コンピューターによる自動採点の信頼性をさらに検証し、リンガスキルのライティングスコアを CEFR のレベルと一致させるために、ライティング評価の専門家[3]がテスト解答の別の部分を見直しました。

豆知識

参加者の大半（73%）は、リンガスキルのライティングテストに肯定的な印象を抱きました。受検体験に「否定的」だった参加者はごくわずか（6%）にとどまりました。
ほとんどの参加者（84%）は、テストで英語のライティング能力を実証できたと感じました[4]。

2 ライティング分野の熟練評価者。ケンブリッジ大学英語検定機構の言語テスト基準を満たす採点ができるように、評価尺度（スケール）の使用方法、および標準化と通常の調整手順に関するトレーニングを受講済みです。
3 CEFR に準拠する能力記述子（descriptor）を用いて記述解答を評価した経験豊富な 6 名のケンブリッジ大学英語検定機構の調査マネージャーと評価マネージャー、および 5 名のライティングとスピーキングの経験豊富な採点官から成る審議会。テストを CEFR に適合させる基準の設定手順では、通常、専門家からなる審議会を利用して、テストスコアを外部基準（欧州評議会 2009）と関連付ける方法を決定します。
4 ライティングテストの終了時に、Metrica へ組み込まれるオンラインアンケートへの参加を受検者にお願いしました。3,026 名（77%）の受検者に、アンケートにご協力いただきました。

トライアルの結果

受検者はケンブリッジ英語検定4技能CBT（Linguaskill リンガスキル）で自分の英語能力を総合的に実証できますか？

主な調査結果

リンガスキルのすべてのタスクとプロンプトを通じて、英語能力をヨーロッパ言語共通参照枠（CEFR）のレベルで総合的に判断できます。

リンガスキルは複数のレベルに対応するテストであるため、達成可能なスコアの広いレンジにおいて異なる記述解答が引き出されるように、タスクを工夫する必要があります。

初級レベルの英語学習者にとって難しすぎるタスクは、複数レベルに対応するテストに含めるべきではありません。同様に、英語上級者にとって挑戦しがいのないタスクを入れると、こうしたテストの有効性が問われることになります。

テスト方法：トライアルでは 3,918 名の受検者がライティング分野のテストに参加し、解答はコンピューターで自動採点されました。ケンブリッジ大学英語検定機構の採点官もテストの解答文の一部を見直し、コンピューターの自動採点で最下位レベルに分類された解答が CEFR A1 以下のレベルのライティングであるかどうか、最上位レベルに分類された解答が CEFR C1 以上のレベルのライティングであるかどうかを確認しました。リンガスキルのタスク 1 とタスク 2 では、異なるプロンプトが試行されました。

調査結果：コンピューターで自動採点されたテストスコアによると、テストで試されたすべてのプロンプトとタスクの解答は、ターゲットとなる成績レンジ全体に分散していることが判明しました。この調査結果は採点官による見直しで裏付けられ、リンガスキルのテストによって、ターゲットレンジ（A1 ～ C1 以上[5]）全体を網羅するライティング能力を判定できることが確認されました。

自動採点スコアの分布状態が標準的である点も、すべてのレンジの言語能力で解答を引き出せるようにリンガスキルのタスクが適切に構成されていることを示しています。自動採点のスコア分布が最も集中した解答文について、一部の解答が採点官によって見直され、トライアルでの平均的な解答が考察されました。この考察で、リンガスキルのライティングタスクで最も平均的な成績は B1 レベルであることが確認されました。

5　タスク 1 で最上位レベルのスコアがライティングで C2 レベルと評価されるケースは比較的少なかったのに対し、タスク 2 で最上位に評価された解答のスコアは C2 レベルでした。この 2 つのタスクの解答レベルでは受検者が達成可能な最高成績は「C1 以上」と報告されるため、この解答水準は、リンガスキルにとって許容できるものです。

コンピューターの自動採点ツールによるランクと解答のスコアは信頼できますか？

主な調査結果

トライアルでは、コンピューターが自動採点したテストスコアの信頼性を調査するために、2 つの分析が行われ、以下の結果が得られました。

- コンピューターで自動採点されたテストスコアと採点官が採点したテストスコアの平均値には、強い正相関が見られる。
- コンピューターの自動採点と専門家による評価とでは、類似の順番（最上位の解答から最下位の解答に至る）で解答がランク付けされている。

信頼性分析その 1 – コンピューターの自動採点ツールは、信頼性の高いテストスコアを提供するか？

この分析では、コンピューターで自動採点されたテストスコアには、採点官が採点したテストスコアと比べてどのような特徴があるかが考察されました。

分析方法：ケンブリッジ大学英語検定機構の経験豊富な採点官 5 名が、それぞれ 60 名のトライアル参加者の解答を採点しました。コンピューターで自動採点されたこれら解答のスコアは、幅広いレンジにわたっています。

採点官は、6 つのレベルの能力記述子を用いる一般的な採点方法を採用しました。リンガスキルの各タスクについて、タスクの達成度、言語リソース、および文章構成を基準にスコアが付けられました。採点官は、スコアを 6 つのレベルに分類するよう指示されました。解答の中には、2 つのレベルの特徴の測定値がほぼ等しいものもありました。そのため、有効なテスト解答は 11 ポイントのスケールに分けられました。

さらに、有意な解答がない場合やライティングがトピックとずれている場合には、採点官はその解答にゼロ（0）というスコアを付けることもできました。

5 名の採点官全員の採点が反映されるように、合計スコアの平均が計算されました。各採点官のスコアを集計すると、採点官の評価のばらつきによる影響が軽減されるため、1 名の採点官の付けたスコアより信頼性の高い指標となります。

調査結果：スピアマンの順位相関係数の計算では、コンピューターで自動採点されたテストスコアと各採点官の付けた合計スコアには、強い正相関が示されています。

- **リンガスキルのタスク 1：**コンピューターで自動採点されたテストスコアと 5 名の採点官が付けたスコアの合計との相関関係は、Rho（スピアマンの順位相関係数）= 0.82 と算定されます。
- **リンガスキルのタスク 2：**コンピューターで自動採点されたテストスコアと 5 名の採点官が付けたスコアの合計との相関関係は、Rho（スピアマンの順位相関係数）= 0.88 と算定されます。

次に、両方のタスクの成績を反映するテストの総合スコアが、参加者ごとに計算されました。コンピューターで自動採点されたテストの総合スコアと採点官が付けた合計スコアとの相関性は、各タスクを別々に分析した場合よりも強くなっています。

- **リンガスキルのライティングテストの総合スコア：**コンピューターで自動採点されたテストスコアと 5 名の採点官が付けたスコアの合計との相関関係は、Rho（スピアマンの順位相関係数）= 0.90 と算定されます。

比較点として、各採点官の一貫性もスピアマンの順位相関係数を使って検証されました。

総合テストのスコアを見ると、各採点官の相関係数は 0.84 ～ 0.95 のレンジにあり、平均は Rho = 0.91 です。この値は、コンピューターの自動採点スコアで計算された相関性の強度にきわめて近似しています。

上記の調査結果からは、コンピューターによる自動採点が採点官の採点と似通っていること、同じ採点方法を採用する一部の採点官より優れた成果さえあげていることがわかります。以上より、自動採点ツールが正確で信頼性の高いスコアを付けていると結論付けることができます。

表 1：採点官による両方のタスクの平均スコア間の内部相関性

	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	-	-	-	0.90	0.95
採点官4	-	-	-	-	0.91
ほかの採点官の平均	0.90	0.91	0.93	0.88	0.92

信頼性分析その 2 – コンピューターの自動採点ツールは、信頼性の高い試験解答のランク付けを提供するか？

この分析では、コンピューターの自動採点と人間の専門家は、テスト解答の質について、（最上位の質も最下位の質も）一貫性のある判断を下しているかどうかを調査することで、コンピューターの自動採点の信頼性をさらに考察しました。

この分析では、コンピューターの自動採点の信頼性を実証するだけでなく、リンガスキルと CEFR のレベルを一致させる基準設定の演習にも対応しています。詳しくは、9 ページの説明をご覧ください。

分析方法：ライティング評価の専門家[6]10 名からなる審査団は、リンガスキルのタスク 1 の解答 20 件、およびタスク 2 の解答 20 件の質を（最上位から最下位へ）ランク付けしました。コンピューターで自動採点されたこれらのテスト解答は、幅広いレンジにわたります。また、審査団は、解答の質をランク付けする際、困難は感じなかったと報告しました。

分析その 1 同様、専門家間のばらつきの影響を軽減し、各回答者の「真のランキング」をより正確に推察するため、専門家全員のランキングの平均値を出しました。

調査結果：審査団のメンバーは、コンピューターの自動採点ツールがさまざまな基準のライティングの解答を問題なく特定したこと、およびテスト解答の質は、スコアの目標レンジ全体で多岐にわたったことに全員賛同しました。

スピアマンの順位相関係数の計算では、コンピューターの自動採点ツールによるテスト解答のランキングと専門家グループによるテスト解答の合計ランキングとの間には、強い正相関が示されています。

- **リンガスキルのタスク 1：**コンピューターの自動採点によるテスト解答のランキングと専門家グループによる累計ランキングとの相関性は、0.88 でした。
- **リンガスキルのタスク 2：**コンピューターの自動採点によるテスト解答のランキングと専門家グループによる累計ランキングとの相関性は 0.92 でした[7]。

6 ワークショップの進行役は、演習に参加する前に、コンピューターの自動採点スコアを確認したため、進行役のランキングはこの文書では報告していません。
7 タスク 1 でランク付けされた 20 の解答文は、タスク 2 でランク付けされた 20 の解答文とは異なる受検者が記述したものです。したがって、これらの解答を組み合わせて総合ランキングを出すことはできません。

図1：リンガスキルのタスク1において、コンピューターの自動採点ツールによるテスト解答のランキングと専門家グループによる平均ランキングとの相関性散布図

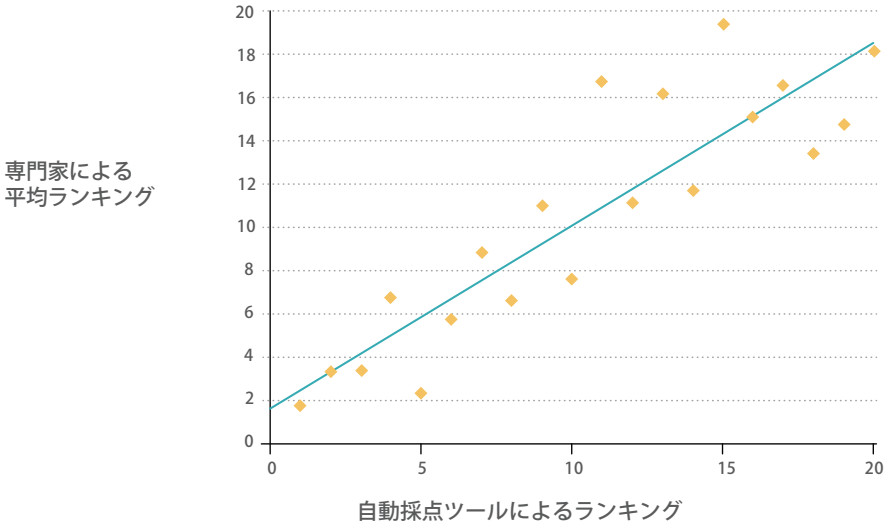
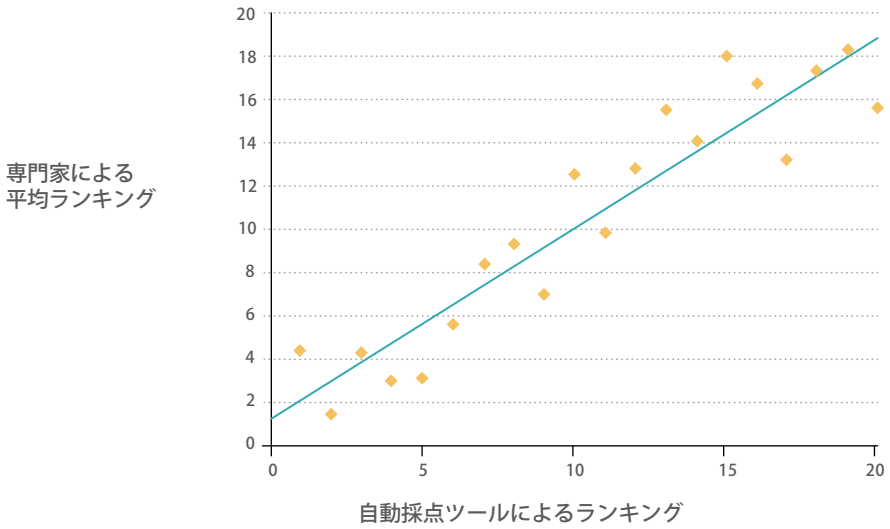


図2：リンガスキルのタスク2において、コンピューターの自動採点ツールによるテスト解答のランキングと専門家グループが付けた平均ランキングとの相関性散布図



比較点として、各専門家の一貫性についてもスピアマンの順位相関係数を使って検証されました。判断力に依存する大半の演習同様、審査団の各メンバーが付けたランキング順位には一貫性がなく、専門家の意見が完全に一致した例は 1 組もありませんでした。

- **リンガスキルのタスク 1：**審査団の各メンバーが設定したランキングの相関性は、0.83 ～ 0.96 のレンジに分散しました。

リンガスキルのタスク 2：審査団の各メンバーが設定したランキングの相関性は、0.75 ～ 0.97 のレンジに分散しました。

先ほどの分析と同じく、コンピューターの自動採点の相関性強度は、専門家同士を比較した場合と同じレンジ内に収まっています。つまり、コンピューターの自動採点と専門家の判断を比べた場合と、各専門家の判断を比べた場合とでは、一貫性はほとんど変わりません。

結論

2 つの分析結果により、コンピューターの自動採点において人が採点した場合と同じようにランク付けされるため、信頼性の高い評価が得られるとの確信を強めました。

ケンブリッジ英語検定4技能CBT（Linguaskill リンガスキル）のライティングテストで CEFR のレベルを判定

次に、体系的なプロセスを用いてリンガスキルのライティング要素を CEFR のレベルに一致させ、リンガスキルのスコアを CEFR のレベルで説明するのを促進できるようにしました。基準を設定するワークショップとワークショップ前の準備活動を利用して、リンガスキルのライティングテストの成績を CEFR の各レベルに照らして判定しました。このプロセスでは、言語試験を CEFR に関連付けるために欧州評議会（2009）が作成したマニュアルを参考にしました。[8]

8 欧州評議会（2009）Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), A Manual, Strasbourg: 欧州評議会。

分析方法：11 名の評価分野の専門家から成る審査団（ワークショップの進行役を含みます）[9] は、次の体系的なプロセスに参加しました。

1	**ワークショップ前の演習**
	基準設定時に、基準に慣れることを勧める欧州評議会の勧告に従い、ワークショップの前に、審査団のメンバーに資料が配布され、演習が実施されました。 ワークショップ前の演習は、リンガスキルのタスクと受検者の解答に慣れるために、利用されました。審査団のメンバーはそれぞれ、リンガスキルの各タスクの解答 20 件の質をランク付け（最上位から最下位へ）するように指示されました。[10]
2	**基準設定ワークショップ：ワークショップ前の演習**
	ワークショップの最初の部分では、ワークショップ前の演習について説明を受け、CEFR について審査団の知識が最新であることが確認されました。 審査団は、お互いのランキングとコンピューターの自動採点によるランキングとを確認しました。審査団は一部のテスト解答を詳細に考察し、解答文の優劣を判断した理由、および CEFR の能力記述子とその理由の関係を検討しました。
3	**基準設定ワークショップ：基準の設定**
	審査団は、ブックマーク方法を使って、解答例を選択し、リンガスキルのライティングのスコアを CEFR レベルで自信を持って報告できるように、コンピューターの自動採点の足きりスコアを設定しました。 各タスクにつき判断を 2 回にすることで、審査団は CEFR の各レベルの境界に関して合意に達し、スコアの CEFR への適合が完了しました。
	3.1 1 巡目の判断 • 審査団は、テスト解答の自分のランキングを見直してから、言語能力を CEFR の各レベルで実証する最初のテスト解答を特定するよう指示されました。具体的には、Below A1/A1 の境界と判断される解答を選択し、 B2/C1以上の境界に達するまで、レベルを上げながら解答の選択を続けました。 • 判断は各自が行い、匿名で提出されました。次に、11 名の審査団全員の判断が累計されました。審査団のメンバー間で意見が食い違う部分が特定され、協議されました。
	3.2 2 巡目の判断 • 審査団は、コンピューターの自動採点によるテスト解答のランキングを見直しました。これらの解答は、順番に並べられた解答冊子に 1 から 20 までまとめられました。審査団のメンバーは、自分の冊子（言語能力を各 CEFR レベルで実証）で最初のテスト解答を特定するよう指示されました。 • CEFR の各レベルについて、順番に並べられた解答冊子の最初から開始してすべてのテスト解答を順番に見直すよう、審査団に指示が出されました。つまり、審査団はレベルが異なる CEFR の境界で、同じテスト解答を選択する可能性がありました。 • 今回も判断は各自が行い、匿名で提出されました。CEFR の各レベルについて、11 名の審査団メンバー全員の平均足きりスコアが計算され、各レベルを表す最初のテスト解答が特定されました。 • 審査団はグループとして、この判断および各レベルの CEFR 能力記述子を見直しました。各レベルについて合意し、自動採点スコアを特定するために話し合いが持たれ、このスコアが最終的な基準として採用されました。

9 審査団メンバーには全員、ケンブリッジ英語検定とテスト開発活動の考察を通じて、CEFR に関する知識がありました。
10 この演習の結果は、当報告書の 7 ページに掲載されています。

基準設定演習の結果

1 巡目には、審査団のメンバーの判断にはばらつきがあり、A2/B1 および B2/C1 の境界で特に顕著でした。2 巡目には、ランキングの順番の標準偏差が低下していることからわかるように、判断のばらつきは小さくなっています。

表 2：順番に並べられた解答冊子（言語能力を各CEFRレベルで実証）の最初のテスト解答に対する審査団の判断の標準偏差。

リンガスキルのタスク1	A1	A2	B1	B2	C1
1巡目の標準偏差	1.80	1.82	2.76	2.38	2.41
2巡目の標準偏差	1.15	1.28	2.02	1.40	2.64
リンガスキルのタスク2	**A1**	**A2**	**B1**	**B2**	**C1**
1巡目の標準偏差	-	1.40	2.95	2.84	3.39
2巡目の標準偏差	0.45	1.21	1.75	1.21	1.97

プロセス全体を通じて総意が形成され、この体系的なアプローチを用いて、5つの目標CEFRレベル（A1、A2、B1、B2、およびC1以上）について足きりスコアが特定されました。

少数のケースではありますが、コンピューターの自動採点によって提供されたものとは異なる順序で判断を下した審査員がいました。基準設定は個々の専門家の独立した判断に焦点を当てているため、これらは有効な判断として扱われました。しかし、各タスクで異なる順序で選択された数は少なく、コンピューターの自動採点によって付与されたスコアの信頼性がさらに向上しました。

結論

この報告書の調査結果は全般的に、リンガスキルのライティングのスコアが自信を持って利用可能であること、およびこれらのスコアは CEFR の各レベルと高い関連性があることを実証するものとなっています。

11 標準偏差は審査団が評価する解答文のランキングから算出されているため、ここでの単位は 20 件中順番に並べられた解答文の数です。

お問い合わせ先

ケンブリッジ大学英語検定機構　試験開発部門　日本統括
Email: InfoJapan@cambridgeenglishreps.org
〒101-0054　東京都千代田区神田錦町1-21-1
ヒューリック神田橋ビル9階 ケンブリッジ大学出版株式会社
M: 080-5545-8969、T: 03-3518-8276、F: 03-3518-8274
www.cambridgeenglish.org (English)、
www.cambridgeenglish.org/jp（日本語）

上述の内容はすべて、2018年2月時点での正確な情報に基づいています。

cambridgeenglish.org
/cambridgeenglish
/cambridgeenglishtv
/cambridgeeng
/cambridgeenglish

Copyright © UCLES 2018 | CER/6080/8Y02

4307359366