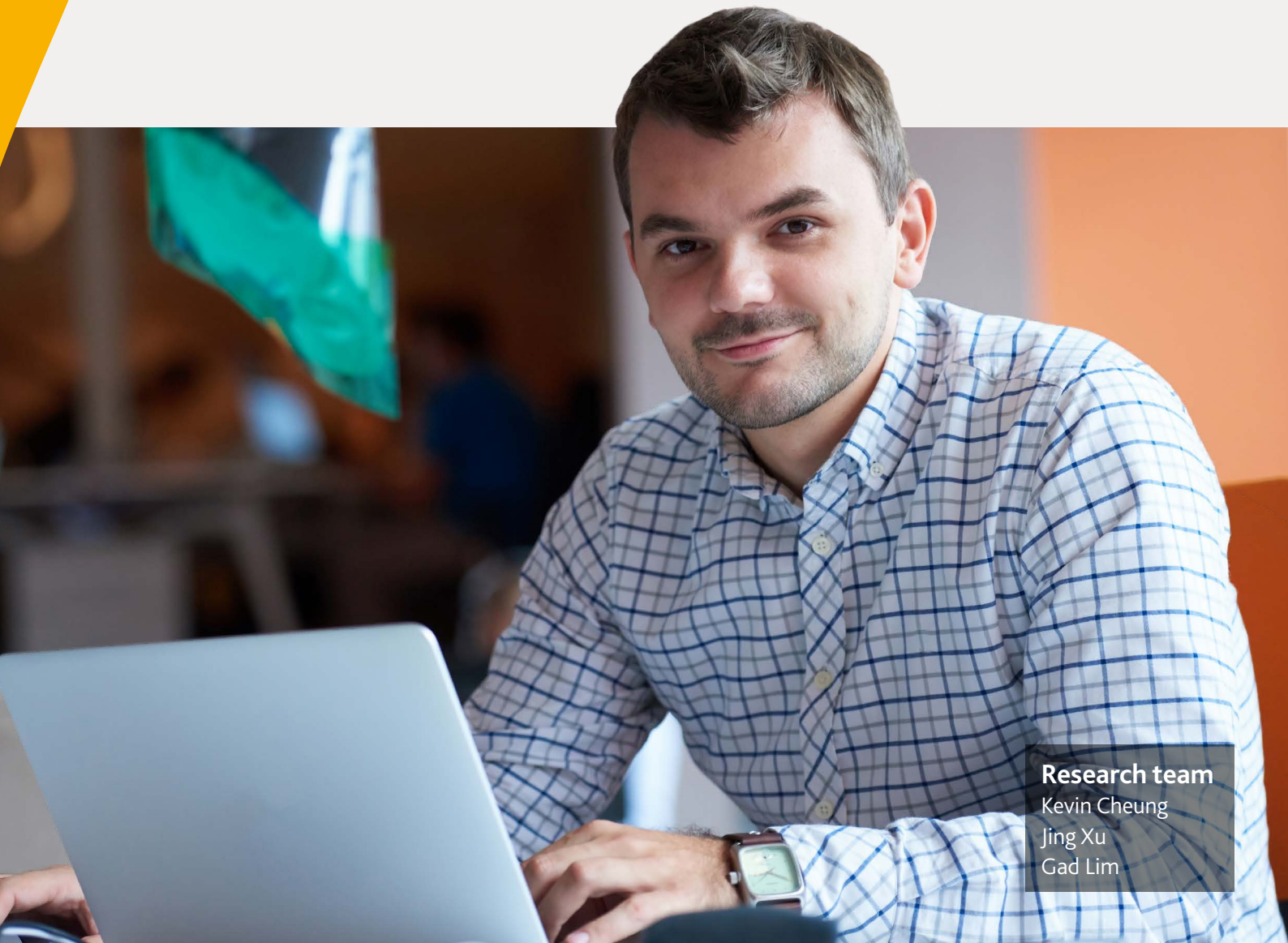
Cambridge Assessment English

Linguaskill

## Writing

# Trial report

June 2017



**Research team**
Kevin Cheung
Jing Xu
Gad Lim

Linguaskill

## Contents


What is Linguaskill? ..... 3

The Linguaskill trial ..... 4

Trial results ..... 5

Aligning Linguaskill Writing to the CEFR ..... 9


# Online, multilevel English testing

## What is Linguaskill?

Linguaskill is a quick and convenient online test to help organisations check the English levels of individuals and groups of candidates. Candidates of all abilities can be assessed with just one test.

The testing experience is designed to be quick, easy and cost effective, with robust results that you can trust. Test results are reported using the Cambridge English Scale and the Common European Framework of Reference (CEFR), the international standard for describing language ability.[1]

### Linguaskill Writing

The Writing test is 45 minutes long and consists of two parts:

- Part 1: Email (at least 50 words). Write an email to a known reader using given information.
- Part 2: Longer writing (at least 180 words). Write a longer piece e.g. article or report to an unknown reader.

Tests are scored by a computer-automarker. This is essentially a series of computer algorithms that has learned how to mark test responses from a large collection of learner responses marked by expert human markers.

1 **www.cambridgeenglish.org/cefr**

## The Linguaskill trial

The Linguaskill Writing test was trialled from December 2016 to February 2017. A total of 3,918 English language learners located in 23 different countries participated in the trial.

The aims of the trial were to:

- evaluate whether Linguaskill allows candidates of all abilities to demonstrate their proficiency
- investigate the reliability of computer-generated test scores
- align Linguaskill Writing scores to the CEFR through standard setting.

The Linguaskill Writing test is delivered through an online testing platform called Metrica. The following data was collected through Metrica:

- participants' test responses
- the questions that participants attempted
- computer-automarker test scores.

A subset of the test responses was reviewed by human examiners[2] in a reliability study of computer-generated test scores.

A different subset of the responses was then reviewed by writing assessment experts[3], to provide further insight into the reliability of the computer-automarker and to align Linguaskill Writing scores to the CEFR.

### Did you know?

Most participants (73%) felt positive about their Linguaskill Writing test. Very few participants (6%) had a negative experience of the test.

The vast majority of participants (84%) felt that the test allowed them to show their English writing ability.[4]

2 Professional assessors of writing, who had received training on how to use the rating scale, as well as standardisation and regular moderation to ensure their marking is up to the standard of Cambridge Assessment English's language tests.

3 The panel consisted of six Cambridge Assessment English research managers and assessment managers with substantial experience of assessing written responses using descriptors aligned to the CEFR, and five experienced writing and speaking examiners. Standard-setting procedures to align tests with the CEFR typically use a panel of experts to determine how test scores link with external criteria (Council of Europe 2009).

4 At the end of their Writing test, participants were invited to complete an online survey authored into Metrica. The survey was completed by 3,026 participants (77%).



## Trial results

## Does Linguaskill allow candidates of all abilities to demonstrate their proficiency?

### Key finding

All Linguaskill tasks and prompts successfully elicit writing performances across the entire range of language proficiency, as defined by the CEFR.

Linguaskill is a multilevel test and it is therefore important that test tasks elicit written responses that vary across the range of achievable scores.

Tasks that low proficiency test takers cannot attempt should not be included in a multilevel test. Similarly, tasks that do not encourage highly proficient language users to demonstrate their ability also compromise the validity of a multilevel test.

**The methodology:** All 3,918 written responses submitted for the trial were evaluated by the computer-automarker. A subset of test scripts was also reviewed by Cambridge Assessment English examiners, to check that the responses awarded the lowest scores by the computer-automarker represented CEFR A1 level writing or below, and that those awarded the highest scores represented CEFR C1 or above level writing. Different prompts for Linguaskill Task 1 and Task 2 were trialled.

**Findings:** The test scores provided by the computer-automarker show that all the trialled test prompts and tasks elicit responses across the entire targeted range of performances. The review by human examiners supports this finding, confirming that Linguaskill successfully elicits writing performances across the range targeted by the test (A1 to C1 or above[5]).

The scores awarded by the automarker are also normally distributed, indicating that Linguaskill tasks are appropriately targeted to elicit responses across the entire range of language proficiency. A subset of the scripts receiving the most commonly awarded scores from the automarker was reviewed by examiners, to investigate the average standard of responses elicited during the trial. This confirmed that Linguaskill Writing tasks most commonly elicit writing at B1 level.

5 Relatively few of the highest scoring responses to Task 1 demonstrated C2 writing, as judged by examiners, whereas the highest scoring responses to Task 2 were at C2 level. The level of writing elicited by the two tasks is acceptable for Linguaskill, as the highest result that test users can achieve is reported as 'C1 or above'.

# Does the computer-automarker rank and score test responses reliably?

## Key findings

Two studies were conducted as part of the trial to investigate the reliability of computer-automarker test scores. These studies showed that:

- there is a strong positive correlation between the test scores provided by the computer-automarker and the averaged test scores provided by human examiners
- the computer-automarker and human experts rank test responses in a similar order (from highest quality response to lowest quality response).

## Reliability study 1 – does the computer-automarker provide reliable test scores?

This study evaluated how test scores awarded by the computer-automarker compare to test scores awarded by human examiners.

**Methodology:** Five experienced Cambridge Assessment English examiners each marked the test responses of 60 trial participants. These responses covered the full range of scores awarded by the computer-automarker.

The examiners used a generic mark scheme with descriptors for six levels. A score was awarded for each Linguaskill task, based on Task Achievement, Language Resource and Text Organisation. Examiners were instructed to award scores between the six levels where test responses included features of two levels in approximately equal measure. This formed an 11-point scale for valid test responses. Additionally, examiners could score a test response as 0 if there was no meaningful response or if the writing was off-topic.

Test scores were averaged to provide a score across all five examiners. An aggregated human examiner score is a more reliable indicator than scores awarded by a single examiner, because the impact of examiner variability is reduced.

**Findings:** Spearman's correlation calculations show that the test scores awarded by the computer-automarker correlate strongly and positively with the aggregate scores awarded by the human examiners.

- **Linguaskill Task 1:** The correlation between computer-automarker test scores and the aggregate scores awarded across all five human examiners is Rho = .82.
- **Linguaskill Task 2:** The correlation between computer-automarker test scores and aggregate scores awarded across all five human examiners is Rho = .88.

Then an overall test score was calculated for each individual participant, reflecting their performance across both tasks. The overall test scores awarded by the computer-automarker correlate even more strongly with aggregate human examiner scores, than when individual test tasks are reviewed in isolation.

- **Overall Linguaskill Writing score:** The correlation between computer-automarker test scores for both tasks and the aggregate scores awarded across all five human examiners is Rho = .90.

As a comparison point, the consistency amongst human examiners was also reviewed using Spearman's correlations.

Looking at overall test scores, the correlations between different human examiners ranged from .84 to .95. The average was Rho = .91. This is very similar to the strength of relationship calculated with the computer-automarker.

These findings indicate that the computer-automarker performs similarly to a human examiner, and even outperforms some of the human examiners using the same mark scheme. This allows us to be confident that the automarker is awarding scores accurately and reliably.

Table 1: Inter-correlations between average scores awarded across both tasks by human examiners

| | Examiner 1 | Examiner 2 | Examiner 3 | Examiner 4 | Examiner 5 |
|---|---|---|---|---|---|
| Examiner 1 | - | 0.92 | 0.91 | 0.84 | 0.91 |
| Examiner 2 | - | - | 0.94 | 0.88 | 0.91 |
| Examiner 3 | - | - | - | 0.90 | 0.95 |
| Examiner 4 | - | - | - | - | 0.91 |
| Average across other examiners | 0.90 | 0.91 | 0.93 | 0.88 | 0.92 |



## Reliability study 2 – does the computer-automarker rank test responses reliably?

This study provided further insight into the reliability of the computer-automarker by investigating whether the computer-automarker and human experts make consistent judgements about the quality of test responses (from highest quality to lowest quality).

In addition to providing evidence about the reliability of the computer-automarker, this study also supported the standard-setting exercise aligning Linguaskill to the CEFR, explained in more detail on page 9.

**Methodology:** A panel of 10 writing assessment experts[6] ranked 20 Linguaskill Task 1 responses and 20 Linguaskill Task 2 responses (from highest quality to lowest quality). These test responses covered the full range of scores awarded by the computer-automarker. The panel reported no difficulties in ranking the quality of the responses.

As before, rankings were averaged across all the human experts to reduce the impact of variation between individual panel members and provide a more accurate estimate of each response's 'true ranking'.

**Findings:** All the panel members agreed that the computer-automarker had successfully identified writing responses of varying standards, and the test responses varied in quality across the desired range of scores.

Spearman's correlation calculations show that the ranking of test responses by the computer-automarker correlates strongly and positively with the aggregate ranking of test responses by the panel of experts.

- **Linguaskill Task 1:** The correlation between the ranking of test responses by the computer-automarker and the aggregate ranking provided by the panel of experts was .88.
- **Linguaskill Task 2:** The correlation between the ranking of test responses by the computer-automarker and the aggregate ranking provided by the panel of experts was .92.[7]

6 The workshop facilitator reviewed the computer-automarker scores before taking part in this exercise, so their rankings are not reported here.

7 The 20 scripts ranked for Task 1 were written by different candidates from the 20 scripts ranked for Task 2. Therefore, they cannot be linked to provide an overall ranking.

Figure 1: Linguaskill Task 1 scatterplot of the correlation between the ranking of test responses provided by the computer-automarker and the average rankings provided by the panel of experts

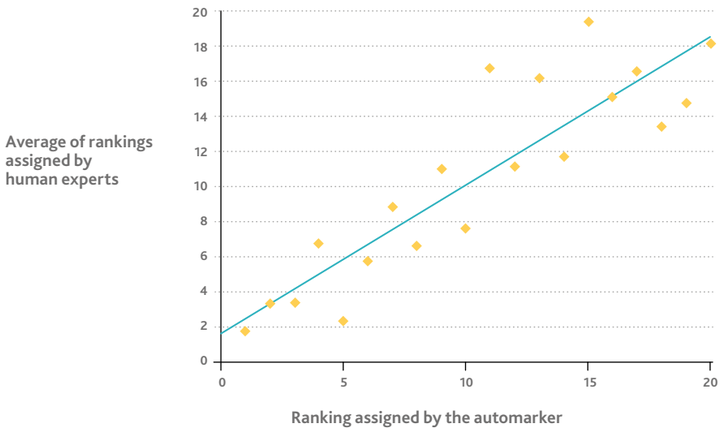
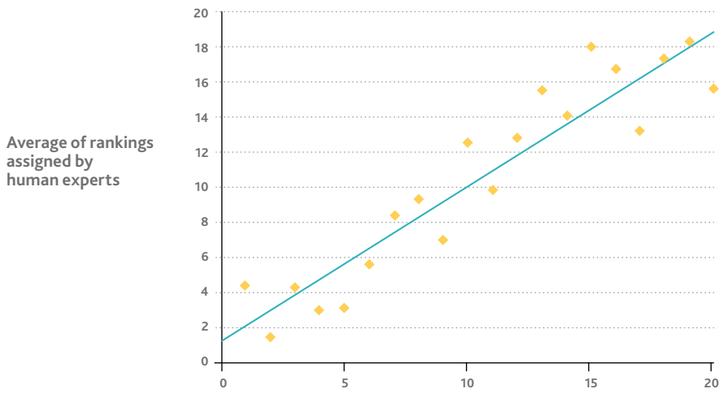



Figure 2: Linguaskill Task 2 scatterplot of the correlation between the ranking of test responses provided by the computer-automarker and the average rankings provided by the panel of experts




As a comparison point, the consistency amongst the human experts was again reviewed using Spearman's correlations. As with most exercises that rely on judgement, the ranking orders provided by different panel members varied. There was not an exact agreement between any pair of the experts.

- **Linguaskill Task 1:** The correlation between the ranking provided by different panel members ranged from .83 to .96.
- **Linguaskill Task 2:** The correlation between the ranking provided by different panel members ranged from .75 to .97.

As in the previous study, the correlation strengths for the computer-automarker fall within the ranges observed when comparing humans with each other. In other words, the computer-automarker's consistency with human experts is comparable to the consistency observed between different human experts.

## Conclusion

The findings across these two studies allow us to be confident that the computer-automarker evaluates test responses reliably – providing test scores and ranking test responses in a similar way to human markers.



## Aligning Linguaskill Writing to the CEFR

A systematic process was then used to align the Writing component of Linguaskill to the CEFR and facilitate the reporting of Linguaskill scores as levels of the CEFR. A standard setting workshop and pre-workshop preparation were used to determine the levels of performance on Linguaskill Writing that align with CEFR levels. This process was informed by the Council of Europe's (2009) manual for relating language examinations to the CEFR.[8]

8 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.

**Methodology:** A panel of 11 writing assessment experts (including the workshop facilitator)[9] took part in the following systematic process.

**1 Pre-workshop exercises**

In line with Council of Europe recommendations to include familiarisation when standard setting, panel members were provided with materials and exercises before the workshop.

A pre-workshop exercise was used to ensure familiarisation with Linguaskill tasks and candidate responses. Members of the panel were asked to independently rank 20 responses for each Linguaskill task (from highest quality to lowest quality).[10]

**2 Standard-setting workshop: review of the pre-workshop exercises**

In the first part of the workshop, the results of the pre-workshop exercise were discussed to confirm that panellists' knowledge of the CEFR was up to date.

Panellists reviewed each other's rank orders, and the ranking provided by the computer-automarker. Panellists examined a subset of the test responses in detail to consider the reasons for judging one script higher than another and how this linked to CEFR descriptors.

**3 Standard-setting workshop: setting the standard**

The panel used the bookmark method to select example responses and set cut-scores for the computer-automarker that allow scores from Linguaskill Writing to be confidently reported as CEFR levels.

Alignment of scores to the CEFR was completed using two rounds of judgements for each task, which allowed panellists to agree on a boundary for each CEFR level.

**3.1 First round of judgements**

- Panellists reviewed their own rank order of test responses. Panel members were asked to identify the first test response that demonstrated language proficiency at each CEFR level, by selecting a response judged as being at the below A1/A1 boundary and progressing upwards until they reached the B2/C1 and above boundary.
- Judgements were made independently and submitted anonymously. Judgements were then aggregated together across the 11 panellists. Areas of disagreement between panellists were identified for discussion.

**3.2 Second round of judgements**

- Panellists reviewed the computer-automarker's rank order of test responses. These responses were put together, from one to 20, in an ordered response booklet. Panel members were asked to identify the first test response in their booklet that demonstrated language proficiency at each CEFR level.
- For each CEFR level, panellists were instructed to start from the beginning of their ordered response booklet and review all the test responses in order. This meant it was possible for panellists to select the same test responses for different CEFR boundaries.
- Judgements were again made independently and submitted anonymously. For each CEFR level, the average cut-score across all 11 panellists was calculated to identify the first test response that had been identified demonstrating each level.
- This was reviewed by the panel as a group, alongside the CEFR descriptors for each level. Discussion was used to ensure agreement for each level and identify an automarker score which was then adopted as the final standard.

9 All the panel members were familiar with the CEFR through examining Cambridge English Qualifications and test development activities.

10 The results of this exercise were presented in the preceding section of this report, on page 7.



## Results of the standard-setting exercise

In the first round, panel members' judgements showed variation, particularly for the A2/B1 and B2/C1 boundaries. In the second round, there was much greater agreement as can be seen in the standard deviation of rank orders becoming smaller.

**Table 2: Standard deviation of panellists' judgements of the first test response in their ordered response booklet that demonstrated language proficiency at each CEFR level.[11]**

| **Linguaskill Task 1** | **A1** | **A2** | **B1** | **B2** | **C1** |
|---|---|---|---|---|---|
| Round 1 standard deviation | 1.80 | 1.82 | 2.76 | 2.38 | 2.41 |
| Round 2 standard deviation | 1.15 | 1.28 | 2.02 | 1.40 | 2.64 |
| **Linguaskill Task 2** | **A1** | **A2** | **B1** | **B2** | **C1** |
| Round 1 standard deviation | - | 1.40 | 2.95 | 2.84 | 3.39 |
| Round 2 standard deviation | 0.45 | 1.21 | 1.75 | 1.21 | 1.97 |

Consensus was developed throughout the process, and using this systematic approach, cut-scores were identified for the five targeted CEFR levels (A1, A2, B1, B2, and C1 and above).

In a small number of cases, a panellist provided their judgements in a different order from that provided by the computer-automarker. As standard setting is focused on the independent judgement of individual experts, these were treated as valid judgements. However, the number of disordered selections made for each task was small, providing further confidence in the robustness of scores assigned by the computer-automarker.

# Conclusion

Overall, the findings in this report demonstrate that Linguaskill Writing scores can be used with confidence, and that they link meaningfully to levels of the CEFR.

11 As this standard deviation is derived from the rank order of scripts reviewed by the panel, the unit here is number of ordered scripts, out of 20.

# Contact us

Cambridge Assessment English
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

All details are correct at the time of going to print in February 2018.

cambridgeenglish.org

/cambridgeenglish

/cambridgeenglishtv

/cambridgeeng

/cambridgeenglish



*4307359366*