Linguaskill

Quality Assurance for Linguaskill

Speaking & Writing

Version 1.0

CAMBRIDGE
English

The purpose of this document is to outline the quality assurance processes in place to ensure the validity of the Speaking and Writing Linguaskill tests. This document provides a comprehensive overview of the methods and rationales behind our quality assurance efforts. They support and evidence various aspects of validity, including context, cognitive, criterion-referenced, scoring, and consequential validity of the test.

Contents

1 Item Trial and Pre-test

To ensure that the tasks designed to measure the target skills are appropriate for the target test-taker, all the task types in Writing and Speaking are trialled as part of test development.

To trial different task types for the New Linguaskill, the Delivery Team recruited a large number of candidates (2330) with 54 first languages to sit the tests. Different versions of task types which differ in task format, rubrics and topics were trialled in three rounds. For each round of trial, appropriateness of item format was evaluated, as well as the target CEFR levels assessed and item/test psychometric properties. The trial results from previous rounds informed the Assessment Team to make corresponding changes for the following round. Through the trial process, the task types are finalised as a template. The new tasks/items are generated following a standardised item writing and production process to ensure the comparability of tasks.

Due to the nature of the tasks in Writing and Speaking (e.g., smaller item bank), it was deemed necessary to balance the spread of pre-testing of new items and so reduce the risk of item exposure. The MLT team is now exploring existing accredited pretesting procedures used by IELTS Speaking and Writing. Plans are to implement and align our process with ALTE accredited products. At the same time, the MLT team have followed a standardised quality check process to minimise item bias. In addition, alternative ways to pre-test the items are being piloted, for example, some new Writing items are being pre-tested in Writing & Improve, a Cambridge online platform to help learners to practice and improve their English writing ability.

2 Post-test Psychometric Analysis

2.1Item analysis

The item analysis is based on the Classical Test Theory (CTT). The following statistics are reported in the assessment PowerBI dashboard:

- *item difficulty*: the ratio of the mean item score to the maximum-possible item score. The higher the value, the easier the item.
- *item discrimination (item_total correlation)*: Pearson correlation coefficient between the item score and the candidate's total score in which the item appears. It indicates the degree to which responses on one item are related to responses on other items within a test. Higher item-test correlation is desired.
- *Number of candidates taking the item, item mean and SD*
- *Item difficulty curve*: a line chart denoting the item difficulty values at each quartile. This curve indicates how item difficulty varies along the score scale. It is desirable that the item is more difficult for candidates with lower total scores and easier for the candidates with higher total scores.
- *Item score distribution*: score frequency distribution on an item. It shows the number of candidates awarded at each score category, which is used to examine the score spread of the item and if a full scale of scores have been applied by examiners.
- *item mean, SD and z test (DIF) across subgroups including age, gender, first language (L1) and country*. This is to examine if any item has bias toward a certain subgroup. Items which show significant bias are flagged in the live test review report and removed as well as reviewed by the Assessment Team.
- *Average word count and range (for Writing):* to check similar length of responses can be elicited across the items.

2.2Test analysis

Since the Writing test only consists of one item, the test analysis is the same as the item analysis. The following statistics for each test version are reported for Speaking:

- *Number of candidates, mean and SD for each test version*
- *Test reliability (Cronbach's alpha)*
- *Test score distribution*: score frequency distribution for each test version.
- *Test mean and SD across subgroups including age, gender and L1 and country.*

These statistics are checked to ensure no significant changes from one test version to another.

2.3Test use review

The monthly product performance report monitors that the test is being used for its intended purpose. Demographic data (age, gender, country and L1) for the candidates are collected along with the candidate reported 'Test purpose'. Large proportions of candidate taking the test for any other reason than those specified in the test specifications are reported to the product management group for review and relevant action if it is suspected that the test is being used for reasons such as immigration.

Each month's data is compared to a rolling 12-month period and significant differences flagged. The data is faceted by age, country, gender, and test purpose to ascertain whether changes to the reported test purpose may have been caused by large changes to the candidature.

3 Test Scoring

3.1 Test response allocation

The responses derived from Speaking and Writing tests are marked in Metrica by Linguaskill examiners. The marked responses are in text format in Writing and in audio format for Speaking. All the responses are randomly ordered to be accessed by examiners, which is to avoid any examiners mark a particular group of candidates. For the Speaking test which consists of five tasks, the task responses from a single test (i.e., one candidate) are marked among a minimum of 3 examiners. This process will prevent a candidate's test score from being affected by any systematic errors specific to an examiner as well as halo effect (McCaffrey et al., 2022).

3.2 Examiner marking

Speaking and Writing components elicit extended responses that are marked by examiners. The reliable marking of the responses serves as the basis for accurate estimation about candidates' target language abilities. Therefore, a rigorous examiner quality assurance process is implemented to reduce the risk of errors due to examiner scoring. This section summarises the steps taken to ensure the examiners marking quality. The detailed procedures on how we implement the QA process can be found in the Examiners QA document.

Examiner recruitment

The Linguaskill examiners are recruited by following the Cambridge examiner recruitment process. The minimum requirements for speaking & writing examiners are

- education to first degree level or equivalent
- a recognised language teaching qualification
- proof of substantial, relevant, recent teaching experience ideally equivalent to at least 1800 hours.

Please refer to Recruiting and Managing Speaking Examiners for detail.

Examiner training and certification

Before the examiners can do any live marking, they are invited to complete Induction, Training, and Certification of Assessment. **The Certification of Assessment** is an online standardisation course which must be completed by all new and existing examiners once a year before they can mark live responses. There are three stages for this exercise:

- **Core training:** looking at marked responses and commentaries on why each response received the mark level.
- **Review and practice quiz:** practicing marking to unmarked responses and reviewing pre-awarded marks and commentaries.
- **Certification Assessment:** assessing if examiners have awarded the marks within a reasonable tolerance to be accepted for live marking.

To help examiners understand why the responses have been awarded certain marks, commentaries are produced which explain the marks using the language of the Mark Scheme as well as other analytical language. For further detailed information regarding Certification of Assessment process, please refer to the Examiners QA document.

Examiner monitoring

The examiner monitoring is to check if examiners maintain accuracy and consistency in their marking. A variety of metrics are examined for this purpose on monthly basis. Some example metrics include

- Descriptive statistics for each examiner
  - o Responses marked for each part and in total
  - o Mean and SD on each part marked
  - o Mark frequency distribution
  - o Marking time for each part
  - o Part scores comparison at the candidate level
- Many-Facet Rasch Measurement analysis (e.g., examiner facet measures, infit mean square, etc.)

Detailed explanations and threshold values for all metrics can be found in the Examiners QA document.

Examiner Intervention and Feedback

An examiner review meeting is held every month. An examiner who is flagged by the monitoring process will be reviewed and follow-up intervention will be discussed. For detailed intervention process, please refer to Examiners QA document.

Reporting

Examiners certification results and monitoring metrics are all reported in PowerBI. Please refer to the PowerBI reports Linguaskill Speaking Examiner QA and Linguaskill Writing Examiner QA

4 Grading

Grading refers to the methodology used to define the CEFR levels and cut scores of the test corresponding to the CEFR levels. A grading process is run periodically to ensure the accuracy of the grading. This process is managed by the Research department. Standard setting reports detailing the process used and results attained for prior standard settings.

References

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). **Best practices for constructed-response scoring** (ETS RR–22-17). *ETS Research Report Series*, 2022(1), 1-58. https://doi.org/10.1002/ets2.12358

We help people **learn English** and **prove their skills** to the world

Discover more:
cambridgeenglish.org/linguaskill

Find out more at
cambridge.org/english

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

Where your world grows

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

All details are correct at the time of going to print in February 2025.

ENG/9672/V1/JAN24