Linguaskill

Quality Assurance for Linguaskill

Reading & Listening

Version 1.0

CAMBRIDGE English

Introduction

The purpose of this document is to outline the quality assurance processes in place to ensure the validity of the Reading and Listening Linguaskill tests. This document is divided into two main sections: pretest analysis and post-test analysis. Together, these sections provide a comprehensive overview of the methods and rationale 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. Furthermore, this document includes references for the statistical limits used in these analyses.

Quality Assurance Framework

Pretest and Post-Test Analysis

Our approach to quality assurance begins with detailed pretest analysis, where each Reading and Listening item in our question bank is calibrated to determine its difficulty. These values are established during pretesting and may be further refined in a post-test, full bank recalibration. We utilize a version of Item Response Theory (IRT), specifically the Rasch model, to analyse our data from both pretest and live testing scenarios. This analysis allows us to accurately place each item on a continuous difficulty scale.

Linking to the CEFR

As part of our broader research activities, we also ensure that our test items are appropriately linked to the Common European Framework of Reference for Languages (CEFR). During the standard setting process, our research team uses the calibrated difficulty scale to select cutscores representing transitions between CEFR levels. The methods employed for standard setting include the binary or 'yes/no Angoff method' (Cizek & Bunch, 2007; Impara & Plake, 1997) and the Bookmark method (Cizek & Bunch, 2007; Lewis, Mitzel, & Green, 1996; Hambleton & Pitoniak, 2006). This ensures accurate and consistent representation of CEFR levels across our tests.

By integrating these detailed quality assurance processes, we ensure that the Reading and Listening Linguaskill tests are robust, valid, and aligned with international standards. This document focuses specifically on the pretest and post-test analyses, which are integral to maintaining the validity of our assessments.

Pretest analysis.

To ensure that the items deployed into the live bank are of sufficient quality, all items are first pretested within the live test. This process involves embedding uncalibrated items into a reserved slot, a few questions into the test. Embedding pretesting provides several benefits:

- High Candidate Motivation: Candidates remain highly engaged, as they are unaware which items are being pretested.
- Appropriate Targeting: Items are exposed to candidates suitable for their difficulty level.
- Population Match: The pretest population mirrors the target population.
- Increased Pretest Data: Higher pretest numbers are achieved compared to standard pretesting methods.

The response data from these pretests are analysed using Item Response Theory (IRT). We utilize the Winsteps software, version 3.66.0, for this analysis. More details about Winsteps can be found on their official website. Each time a pretest analysis is conducted the performance of the calibrated items is checked against that item's historical performance. This group of items are known as anchors, and are used to ensure that the uncalibrated, or pretest items, are calibrated with respect to items which are performing consistently. In this way, each pretest analysis ties the newly calibrated difficulty values to the same scale. Data for the pretest analysis must comply with various quality control tolerances. These limits are defined here:

- Sample size – More than 250 observations required per item, suitable for high-stakes purposes (Linacre, 1994).
- Anchor selection (outlier removal) – Studentised residual >3 (Montgomery and Vining 2003), and Cook's d > 1 (Cook & Weisberg, 1982)
- L1 proportion – Item observations are sampled to ensure no more than 30% of the candidates which make up the observations are collected from a single first language, i.e. no more than 30% Spanish. This ensures that the items are less likely to exhibit bias with respect to L1.

Once the analysis is complete, items are flagged using the following item-level statistics. These statistics are:

- Item infit – 0.8 to 1.2. Linacre states that 0.5 to 1.5 are value which can be used to ensure items are productive for measurement (Linacre, 2002), however we have further constrained these to ensure only the most quality items pass through the pretest process.
- Standard error – 0.2. Suitable for high-stakes and ensures high confidence in calibrated values (Linacre, 2024).
- The range of item difficulties for multi-interaction-items – Set at approximately one CEFR level to ensure items are never inappropriate for a given test-taker.

Pretest data are reviewed during Pre test Review (PTR) meetings, prior to being made live and added to the live test bank. Items which fall outside of the above tolerances are rejected. If an item is flagged as inadequate by the statistical metrics detailed above, but the Assessment experts (internal and external) are confident that an edited version would fix the issues, the item may be re-pretested. Items are not to be re-pretested more than twice unless there are valid reasons to do so.

Post-test analysis

Item recalibration

As items are pretested within the live bank with high numbers of observations (typically over 250), they are considered fully calibrated once pretested. However, to ensure that Linguaskill items remain of high quality and reliability, the item bank is periodically recalibrated using live test data.

To validate the recalibrated values and ensure they surpass the original calibrated values, it is required that each item meets at least the initial pretest requirements in terms of sample size. However, many items accumulate thousands of observations by this stage, making their recalibrated difficulty values much more precise.

To guarantee the quality of the data used for recalibration, the data undergoes a rigorous cleaning process. This includes removing any incomplete tests, excluding candidates suspected of malpractice, and filtering out responses where the time taken for each task is significantly outside the expected range (using z-scores of ±3).

The statistical analysis employed to evaluate response data is known as Item Response Theory (IRT). Once this analysis is complete, 'anchor items' are selected from the item bank based on their stability, or consistency in difficulty over time. These anchor items play a crucial role in the pretesting process described earlier.

Currently, the certificated Linguaskill test is not yet live, and therefore, a live bank recalibration using live data has not occurred. However, the plan is to perform recalibrations periodically, approximately once or twice per year, depending on the testing volume.

Item diff analysis.

As part of the future live bank recalibration and review process for the Linguaskill test, each item will be scrutinized to determine if it exhibits differential item functioning (DIF) across various demographics, such as age groups, gender, first language (L1), and country. The Mantel–Haenszel test will be employed to identify significant biases in items toward any specific subset of test-takers. Items that demonstrate significant bias will be flagged in the Live Test Review Report and subsequently forwarded to the assessment team for further evaluation after each bank recalibration. Since Linguaskill is a new test, this process has not yet been applied to live data, as we have not accumulated sufficient live testing data to date. However, these measures will be integral to maintaining the fairness and reliability of the test as we move forward.

Test reliability.

As Linguaskill is an adaptive test, it is challenging to report test reliability using a standard statistic such as Cronbach's alpha. Instead, the reliability statistic reported is the scale-separation reliability (SSR), calculated using the following formula:

$$SSR = \frac{Var(Estimated\ abilities) - mean(SEM^2)}{Var(Estimated\ abilities)}$$

In this formula, the Standard Error of Measurement (SEM) represents the measurement error for candidates who take the test over a specified period. Since the SEM is calculated during the test delivery process and serves as a criterion for stopping the test, it is generally stable and unlikely to vary significantly from month to month.

However, each month's calculated SSR is compared against data from the previous 12 months to detect any significant deviations. This comparison helps identify potential changes in test reliability. The data is analysed across different facets such as age, country, gender, and test

purpose to determine if any shifts in reliability are due to substantial changes in the candidate demographics.

It is important to note that these are planned quality assurance processes, which will be thoroughly reviewed as the live test is implemented. Currently, there is no live data for the Certificated version of Linguaskill, and therefore, no monthly statistical reports have been generated yet.

Test score distribution.

A monthly product performance report is utilized to monitor significant changes in the distribution of test results. Each month's data is compared against a rolling 12-month period to identify any substantial deviations. The analysis is segmented by various factors including age, country, gender, and test purpose to determine whether changes in results are associated with shifts in the candidate demographics.

To measure changes in the proportions within each category, a chi-squared test of proportions is employed. Significant changes, identified by a p-value of less than 0.01, are reported to the Delivery team. If a suspicious increase in candidate performance is observed, it is thoroughly investigated. These changes are cross-referenced with information on the test sessions created. For example, if a large group of English teachers takes the Linguaskill test, this might result in a notable increase in C1 and above-level results. Such an increase would be flagged as significant but could be expected given the proficiency of this particular cohort.

These monitoring processes are crucial for maintaining the integrity and reliability of the Linguaskill test results as they help identify and explain any unusual patterns in the data.

Test use review

The monthly product performance report monitors that the test is being used for its intended purpose. Biometric 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 Owner for review.

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.

Content demand and commissioning planning

As Linguaskill demand increases, items are used with more frequency across the world. Having a sufficiently large item bank makes item harvesting, where test uses illegally copy test content, an ineffective form of cheating. A large bank also means that repeat candidates are unlikely to see the same items again, which supports the test validity.

To target the item task types most urgently required by the bank, the Senior Assessment Manager subject leads in collaboration with the Assessment Group Managers, use a demand management tool to focus commissioning on the areas which are required. This ensured that the bank continues to grow in an adaptive and intelligent fashion.

Malpractice

To ensure the integrity and reliability of the Reading and Listening components of the Linguaskill test, three pre-release malpractice checks are conducted. These checks utilize data from past test administrations to identify potential irregularities or suspicious behaviour in candidate performance. The checks are as follows:

- Jagged Profiles: Using historical live data, a multiple linear regression is performed on candidates who have taken all four components of the current Linguaskill test. By using three (or fewer) components as predictors for the fourth (or most recent) component, a likely score is calculated for each candidate. The actual score is then compared to this predicted score. If the difference between the predicted and actual scores exceeds a set tolerance of ±3 z-scores, the candidate is flagged for further investigation.
- Response Time Analysis: Historical data provides median response times for each task in the Reading and Listening bank. Candidates' response times are compared to these median values, and those who complete tasks significantly faster than the median time are flagged as suspicious. If a candidate has more than 80% of their responses marked as suspiciously fast, they are flagged for further investigation.
- Resitter Improvement Analysis: Using past live data, analysis was conducted on all resitting candidates to determine the expected improvements for candidates across the ability range at 5 CES score intervals. ‘Reasonable’ improvements were calculated using the 75th percentile as a placeholder-tolerance. Once live, this can be revisited to adjust the efficacy of the statistical check. Candidates who received a score on a second sitting which is outside of the 75th percentile of historic resitters (in terms of improvement) are flagged for further investigation.

These pre-release checks are crucial for maintaining the credibility of the Linguaskill test and ensuring that all candidates' scores are a true reflection of their abilities.

References

Bramley, T. (2015). ***Investigating the Reliability of Adaptive Comparative Judgment***. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available at: www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf (Accessed: 11 June 2024).

Cizek, G.J. and Bunch, M.B. (2007). ***Standard Setting: A Guide to Establishing and Evaluating*** Performance Standards on Tests. Thousand Oaks, CA: SAGE Publications, pp. 88-92.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. New York, NY: Chapman & Hall.

Hambleton, R.K. and Pitoniak, M.J. (2006). "Setting Performance Standards." In: R.L. Brennan (ed.) Educational Measurement. 4th edn. Westport, CT: Praeger, pp. 433-470.

Impara, J.C. and Plake, B.S. (1997). "Standard Setting: An Alternative Approach." Journal of ***Educational Measurement***, 34(4), pp. 353-367.

Lewis, D.M., Mitzel, H.C. and Green, D.R. (1996). "Standard Setting: A Bookmark Approach." In: D.R. Green (ed.) ***Setting Performance Standards: Concepts, Methods, and Perspectives***. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 219-248.

Linacre, J.M. (1994). "Sample Size and Item Calibration Stability." Rasch Measurement Transactions, 7(4), p. 328. Available at: www. rasch.org/rmt/rmt74m.htm (Accessed: 11 June 2024).

Linacre, J.M. (2002). "What do Infit and Outfit, Mean-square and Standardized mean?" Rasch Measurement Transactions, 16(2), pp. 878-879. Available at: www. rasch.org/rmt/rmt162f.htm (Accessed: 11 June 2024).

Linacre, J.M. (2024). Standard Error and Item Calibration. Winsteps User Manual. Available at: www.winsteps.com/winman/index.htm?standarderror.htm (Accessed: 11 June 2024).

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2003). Introduction to Linear Regression Analysis. 3rd edn. New York: John Wiley & Sons.