LanguageCert

Exploring Item Bank Stability in the Creation of Multiple Test Forms

David Coniam
Tony Lee
and
Michael Milanovic

Exploring Item Bank Stability in the Creation of Multiple Test Forms

David Coniam, Tony Lee and Michael Milanovic

LanguageCert, UK

Abstract

The engine facilitating the construction of LanguageCert tests is a complex item banking system. These item banks contain large amounts of test material covering a wide range of content and construct characteristics. They are calibrated on the basis of Rasch difficulty estimates and fit statistics, and classical test statistics analysis.

When effectively constructed and managed, item banks allow for the creation of test forms which are consistent and comparable both in terms of content and difficulty. This is relevant not only when creating tests intended to measure at a particular level such as CEFR level B1 but also when developing tests which measure across multiple levels from A1 to C2.

The current paper is the second in a set of linked studies investigating one of the item banks developed by LanguageCert UK. The first study (Lee et al., 2022) involved exploring item bank stability in terms of model fit and regression line statistics in both the live dataset (13,000 test takers, each doing 60 items) and in a simulated 'full' dataset generated via model-based imputation (i.e., 13,000 test takers, each having done all 820 items). The purpose of the current study involved submitting the item bank to a real-world test in order to examine the quality of actual tests derived from the item bank. To achieve this, three tests were compiled from the calibrated item bank, and subsequently administered to a sample of test takers. In the analysis of the three tests, good fit statistics emerged, with high correlations between each test – an indicator of robust joint calibration and further evidence as to the stability of the item bank.

The paper concludes with the claim that the items that comprise the item bank have been well set, with strong support for the robustness of the item bank as a clearing house from which many different tests may be constructed.

Introduction

This paper reports on a study investigating the stability and robustness of one of the item banks developed by LanguageCert UK [Note 1]. Given that both paper-based and adaptive high-stakes tests are produced from these item banks, key issues are item bank stability and item quality in terms of tests generated from the item bank (Mills & Steffen, 2000). Indeed, test quality is of the utmost importance for any organisation administering high-stakes examination.

In the previous study (Coniam et al., 2022), an overview was first presented of issues relevant to item bank size (Choppin, 1968; Ree, 1981; Derner et al., 2008; Rudner, 2009) and item bank stability (Gao & Chen, 2005; Weiss & von Minden, 2012; Sahin & Weiss, 2015). Following this, the makeup of the LanguageCert adaptive test item bank and test taker dataset was outlined. The adaptive item bank contains 820 items, with subsets of approximately 60 items administered (as adaptive tests) to approximately 13,000 test takers.

By using imputed values (Peugh & Enders, 2004), via the software Winsteps (Linacre, 2018), a simulated 'full' dataset was constructed on the basis of responses for each test taker being imputed for all 820 items based on test takers' actual responses. From the existing 0.78 million data points (13,000 x 60), a full dataset containing 10.66 million data points (13,000 x 820) was therefore generated.

Two hypotheses were pursued in the study involving simulations.

The first was that the regression line (R^2) value of the simulation would be a minimum of 0.75 [the rule of thumb for 'substantial' R^2 values – Ringle & Sinkovics (2009)]. The R^2 values for the simulation were, in fact, 0.99, and the hypothesis was thus accepted.

The second was that Rasch infit and outfit statistics would be within acceptable ranges at the key 25th and 75th percentiles. For both live dataset values and simulated dataset values at the both percentiles, fit statistics were well within acceptable ranges. This hypothesis was also accepted.

The conclusion drawn from the comparison of the 'full' and live (comparatively sparser) dataset was that as the live dataset expands in terms of data points (i.e., items and test takers), stability is likely to improve further. It was recommended that to provide corroborating evidence to support this claim, the item bank needed to be submitted to a real-world test whereby actual tests were generated, run and analysed. It is this procedure which the current study is now pursuing.

Assessment Context

The exploration reported in the current paper relates to the LanguageCert Test of English (LTE). The LTE, which is accredited by the UK's Office of Qualifications and Examinations Regulation (Ofqual), is an English 'for work' exam intended for people over the age of 18 in or about to enter the workplace, as well as those in higher or further education.

The LTE comprises three products, three level-agnostic tests, as in Table 1 below.

Table 1. LanguageCert Test of English (LTE) products

Test product	**CEFR levels aimed at**
(1) a paper-based test measuring A1-B1	Test aimed at beginner to intermediate cohorts.
(2) a paper-based test measuring A1-C2	Test for test takers at all CEFR levels
(3) an adaptive test measuring CEFR A1-C2	Test for test takers at all CEFR levels

Reference is made in this paper to "one" item bank. There are, however, a number of LTE item banks, with different banks used at different times to produce both paper-based and adaptive tests. Test taker results are reported against CEFR (the Common European Framework of Reference for Languages) levels, which have been defined on the basis of LID scale scores; these are laid out in Table 2 below.

Table 2. LID scale

CEFR level	Mid point
C2	160
C1	140
B2	120
B1	100
A2	80
A1	60

Rasch Measurement

The current study, as mentioned, is predicated on the use of the Rasch model, a brief overview of which is provided below.

The use of the Rasch model enables different facets (e.g., person ability and item difficulty) to be modelled together, converting raw data into measures which have a constant interval meaning (Wright, 1997) and which provide objective and linear measurement from ordered category responses (Linacre, 2012). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred to as ‘logits’) evenly spaced along the ruler. Once a common metric has been established for measuring different phenomena (test takers and test items, for example), person ability estimates may then be calculated independently from the items used, with item difficulty estimates also being calculated independently from the sample recruited.

In this manner, the Rasch model enables persons and items to be calibrated onto a single unidimensional latent trait scale – also known as the one-parameter IRT (Item Response Theory) model (Bond & Fox, 2007). Person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties can be easily conducted, and results subsequently interpreted with a more general meaning.

Rasch Model Fit

Broad criteria in assessing model fit are the *Infit* and *Outfit* mean square statistics (i.e., estimates of population variance, or standard error) and the *Standardised Infit* and *Outfit* (i.e., Z-score) statistic. These statistics are outlined briefly below.

Infit may be seen as the ‘big picture’ in that it scrutinises the internal structure of an item or person. High infit mean square values indicate rather scattered information within the item or person, providing a confused picture about the placement of the item or person. Very small infit values indicate only very small variation and, provide therefore, little information to articulate clear and meaningful judgments about an item or person.

Outfit gives a picture of ‘outliers’, that is responses from persons or items that appear to be considerably out of line with where a person or item would expect to be placed.

For both Infit and Outfit, a perfect fit of 1.0 indicates that obtained values match expected values 100%. While acceptable ranges of tolerance for fit vary, acceptable ranges are generally taken as from 0.5 for the lower limit to 1.5 for the upper limit (Lunz & Stahl, 1990).

The Study: Real-World Validation

While the results from the previous study suggest an a priori robust item bank, test taker results obtained from the item bank need be seen to be stable irrespective of what tests are constructed or derived – paper-based or adaptive – from the item bank. Consequently, in the current study, three live tests produced in mid 2021 and administered to actual test taker groups were analysed against calibrated values.

Three level-agnostic A1-C2 tests, each consisting of 110 items, were compiled from the recalibrated item bank. The three tests were constructed with a number of overlapping items from the item bank, thereby permitting anchoring and direct comparisons to be made. The three tests were administered to a sample of European university students (average age 23 years), whose English language proficiency was estimated by their professors to range from B1 to C1.

Hypotheses

The current study is pursuing the following hypotheses.

1. Rasch infit and outfit statistics will be within acceptable ranges: between 0.5-1.5 at the 25th and 75th percentiles.

2. LID values at the 50th percentile for all three tests will be close to the usual calibration mid point of 100 (i.e., within half a logit, or 10 LID points).

3. Pearson correlations between LID values across the three tests will be over 0.8.

Results

Table 3 below provides details of the number of different items in each test (i.e., overlapping items have been removed) and the number of test takers for each test.

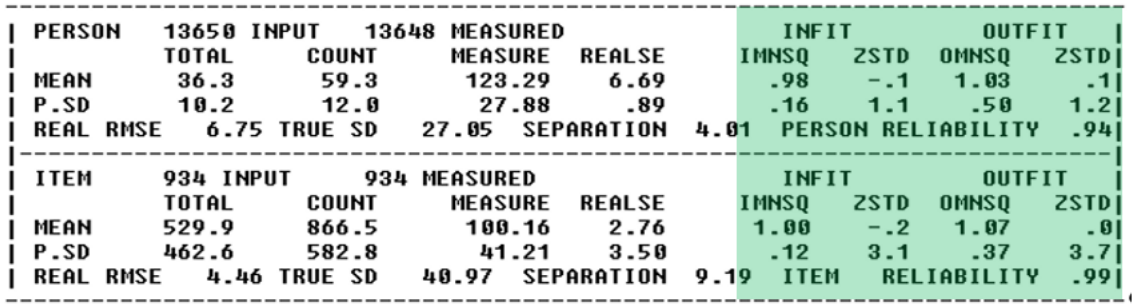
Table 3. Test items and test takers

	Test 1	Test 2	Test 3	Totals
Discrete items	110	77	25	212
Test takers	108	241	228	577

The dataset comprised 212 discrete items administered to 577 test takers.

As mentioned, item difficulty in LTE tests is predicated on the overarching LanguageCert Item Difficulty (LID) scale (Table 2 above). For calibration purposes, the mid-point of the scale is set at 100 (B1 in LID value terms), with a standard deviation (SD) of 20 (see Coniam et al., 2021). This was the starting point for the analysis, with the Test 1 items first anchored with these values against the entire item bank. Table 4 presents a summary of the analysis of the whole item bank of 820 items with the 110 items in Test 1.

Table 4. Combined analysis of Test 1 with whole item bank



```
-------------------------------------------------------------------------------
| PERSON   13650 INPUT   13648 MEASURED                    INFIT        OUTFIT  |
|          TOTAL      COUNT      MEASURE  REALSE       IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      36.3       59.3       123.29    6.69         .98    -.1   1.03     .1|
| P.SD      10.2       12.0        27.88     .89         .16    1.1    .50    1.2|
| REAL RMSE    6.75 TRUE SD    27.05  SEPARATION  4.01  PERSON RELIABILITY  .94|
|-------------------------------------------------------------------------------|
| ITEM      934 INPUT      934 MEASURED                    INFIT        OUTFIT  |
|          TOTAL      COUNT      MEASURE  REALSE       IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN     529.9      866.5       100.16    2.76        1.00    -.2   1.07     .0|
| P.SD     462.6      582.8        41.21    3.50         .12    3.1    .37    3.7|
| REAL RMSE    4.46 TRUE SD    40.97  SEPARATION  9.19  ITEM   RELIABILITY  .99|
-------------------------------------------------------------------------------
```

As may be seen in the section highlighted in green in Table 4 above, infit and outfit figures are good – very close to 1.0. Reliability figures are high for both persons and items, being in the high 0.9 range.

Following the calibration of Test 1 to the whole item bank, Tests 2 and 3 were then calibrated against Test 1. Table 5 presents the comparative results for the three tests.

Table 5. Test analysis details – item spread

Statistic	**Test 1**	**Test 2**	**Test 3**
Valid	110	77	25
Mean	100.00	97.26	101.55
Std. Deviation	28.22	38.73	26.34
Maximum (99th percentile)	150.34	176.18	135.28
75th percentile	121.26	126.93	125.75
50th percentile	104.44	99.76	105.20
25th percentile	83.49	69.63	82.67
Minimum (1st percentile)	16.11	-8.41	53.19

As can be seen, at the 50th percentile, LID values for all three tests are close to the mean of 100 (B1). Means are likewise quite comparable at the 75th percentile of 120, where values are one logit (20 points) higher, at B2. There is some divergence at the lower end of the scale at the 25th percentile: Tests 1 and 3 are one logit lower around 80 – CEFR level A2. Test 2 exhibits a wider range of difficulty, and is another half logit lower at 69.63. Despite the divergences (some of which may be attributed to guessing), it can be seen that the three tests are broadly aligned.

Table 6 presents the Pearson correlations between LID values across the three tests.

Table 6. Correlations between LID values

		Test 1	Test 2
Test 2	Correlation	0.923	—
	p	< .001	—
Test 3	Correlation	0.917	0.964
	p	< .001	< .001

The scores from all three tests correlate very highly – above 0.9, with p values < .001. This is an indicator that the joint calibration of the three tests is robust.

Conclusion

The current study has explored how a dataset such as the LanguageCert Test of English (LTE) adaptive test may be assessed in terms of robustness. In the study, the item bank was submitted to a real-world test whereby three tests were compiled from the calibrated item bank, and administered to a representative sample of test takers. Three hypotheses were pursued in this study.

Hypothesis 1 was that Rasch infit and outfit statistics would be within acceptable ranges: between 0.5-1.5 at the 25th and 75th percentiles. Infit and outfit statistics were within acceptable ranges, and the hypothesis was therefore accepted.

Hypothesis 2 was that LID values at the 50th percentile for all three tests would be within half a logit of the usual calibration mid point of 100. At the 50th percentile, LID values for the three tests within five LID scale points (a quarter of a logit) of the mean of 100; the hypothesis was therefore accepted.

Hypothesis 3 was that Pearson correlations between LID values across the three tests would be above 0.8. Test scores from all three tests correlated at above 0.9, and this hypothesis was also accepted.

The previous – background simulation – study (Lee at al., 2022) was indicative of current, and future, stability. The current study's real-world testing of tests compiled from the item bank and administered to a representative sample of test takers provides corroborating evidence for the above claim. In the current study, good fit statistics, comparable LID score levels at major percentile levels, and high inter-test correlations emerged on the three tests: all of which underscore the stability of the item bank.

The conclusion that may therefore be drawn from the current study is that the items that comprise the item bank used in the construction of LTE tests are of good quality and have been well set. This in turn supports the claim regarding the robustness of the LTE as an assessment instrument.

Notes

1. Reference is made in this paper to "one" item bank. It should be noted that LanguageCert tests access multiple parallel item banks.

References

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: fundamental measurement in the human sciences (2nd ed.). Mahwah, N.J.: Erlbaum.

Choppin, B. (1968). Item Bank using sample-free calibration. Nature, 219, 870-872. https://doi.org/10.1038/219870a0.

Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). Validating the LanguageCert Test of English scale: The adaptive test. LanguageCert: London, UK.

Derner, S., Klein, S., & Hilber, D. (2008). Assessing the Feasibility of a Test Item Bank and Assessment Clearinghouse: Strategies to Measure Technical Skill Attainment of Career and Technical Education Participants. MPR Associates, Inc.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. Applied Measurement in Education, 18(4), 351-380. https://doi.org/10.1207/s15324818ame1804_2.

Lee, T., Coniam, D., & Milanovic, M. (2022). Exploring Item Bank Stability Through Live and Simulated Datasets. Journal of Language Testing & Assessment (2022) Vol. 5: 13-21. https://doi.org: 10.23977/langta.2022.050102

Linacre, J. M. (2012). A user's guide to WINSTEPS. Chicago, IL: Winsteps.com.

Linacre, J. M. (2018). Winsteps Rasch measurement computer program user's guide. Beaverton, OR.

Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. Evaluation and the Health Profession, 13, 425-444. https://doi.org/10.1177/016327879001300405.

Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In Computerized adaptive testing: Theory and practice (pp. 75-99). Springer, Dordrecht. https://doi.org/10.1007/0-306-47531-6_4.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74, 525-556. https://doi.org/10.3102/00346543074004525.

Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. Advances in International Marketing, 20, 277-319. https://doi.org/10.1108/S1474-7979(2009)0000020014.

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In Elements of adaptive testing (pp. 151-165). Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-8_8.

Sahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. Educational Sciences: Theory & Practice, 15(6), 1585-1595.

Weiss, D. J., & von Minden, S. V. (2012). A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG. St. Paul, MN: Assessment Systems Corporation.

Wright, B. D. (1997). A history of social science measurement. Educational Measurement: Issues and Practice, 16(4), 33-45. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x.

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926.

Copyright © 2022 LanguageCert

All rights reserved. No part of this publication may be reproduced or transmitted in any form and by any means (electronic, photocopying, recording or otherwise) except as permitted in writing by LanguageCert. Enquiries for permission to reproduce, transmit or use for any purpose this material should be directed to LanguageCert.

DISCLAIMER
This publication is designed to provide helpful information to the reader. Although care has been taken by LanguageCert in the preparation of this publication, no representation or warranty (express or implied) is given by LanguageCert with respect as to the completeness, accuracy, reliability, suitability or availability of the information contained within it and neither shall LanguageCert be responsible or liable for any loss or damage whatsoever (including but not limited to, special, indirect, consequential) arising or resulting from information, instructions or advice contained within this publication.

Language
Cert
languagecert.org