Language Cert

Validating the LanguageCert Test of English Scale: The Adaptive Test

David Coniam
Tony Lee
Michael Milanovic
and
Nigel Pike

Abstract

This paper reports on the calibration to a common scale of the LanguageCert Test of English (LTE) in its computer adaptive mode which in turn builds on the calibration of LTE paper-based tests reported in Coniam et al. (2021). The work described here now ensures that the LanguageCert Item Difficulty (LID) scale can be used for all LTE modes of delivery and that item difficulties in LTE align with all LanguageCert tests that make reference to the LID scale. The calibrated LTE item bank is therefore a robust source of materials for both the paper-based and computer-based adaptive tests.

Introduction

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 current study builds on Coniam et al. (2021), which documented the first phase of measurement scale development for the LanguageCert Test of English (LTE). That study described the validation of the LID scale via the LTE paper-based tests. The LID scale was created between 2017-2019 on the basis of classical test statistics and expert judgement. The LID scale is the empirical basis for the alignment of current and future LanguageCert assessment products to the same measurement scale that is itself aligned to the CEFR.

The Coniam et al. (2021) study focused on the LTE paper-based tests, which, after being calibrated, were placed on a common scale. The current study extends the LTE calibration process by demonstrating how the LTE adaptive test is calibrated to the same common LID scale as the paper-based tests. It demonstrates how candidates taking either a paper-based or an adaptive LTE test will be placed at more or less the same point on the LID scale regardless of which form of the test they take.

Current Study: Background

The LanguageCert Test of English (LTE) comprises three products, as in Table 1 below.

Table 1: Three LanguageCert test 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 candidates at all CEFR levels
(3) an adaptive test measuring CEFR A1-C2	Test for candidates at all CEFR levels

The Coniam et al. (2021) study reported on the validation, linking, and establishing of a common scale for paper-based variants (1) and (2). The current study reports on the alignment of variant (3), the adaptive test, to the LID scale. The purpose of the current study, as mentioned, is to ensure that candidates taking any variant (paper-based or adaptive) will be consistently placed at the same point on the LID scale. Given that the scores are interchangeable, consistency of measurement across modes of delivery and different versions of the same test is essential.

Initial development and calibration of the LID scale had its origins in a compilation of the LTE paper-based tests (Coniam et al., 2021), with the latter study showing the four paper-based tests to be robust and the calibrated scale which emerged to be consistent with the data. The initial scale provided an acceptable basis for the development of the full LanguageCert

scale on the basis of the adaptive test data – the focus of the study reported in the current paper.

While the current study reports on the LTE adaptive test, it must be restated that it is the LID scale, not the adaptive test, that is the focus of this study. For detail on adaptive testing and an overview of the LTE adaptive test, its algorithm and operation, the reader is referred to Pike & Coniam (2021).

Since the calibration for both the paper-based tests and the adaptive test have made use of Rasch measurement, via Winsteps (Linacre, 2006), a brief description of Rasch will now be provided.

Rasch Measurement

The use of the Rasch model enables different facets to be modelled together. First, in the standard Rasch model, the aim is to obtain a unified and interval metric for measurement. The Rasch model converts ordinal raw data into interval measures which have a constant interval meaning and provide objective and linear measurement from ordered category responses (Wright, 1997). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred as 'logits') evenly spaced along the ruler. Rasch measurement achieves its goal by estimating the theoretical probability of success of candidates answering items. Such theoretical probabilities are derived from the sample assessed, yet independent from it due to the use of the statistical modelling techniques. Therefore, the measurement results based on Rasch analysis, can be interpreted in a general way (like a ruler) for different candidate samples assessed using the same test or different tests aligned to the same scale. Second, once a common metric is established for measuring different phenomena (candidates and test items being the most obvious), person ability estimates are independent of the items used, with item difficulty estimates being independent of the sample because the estimates are calibrated against a common metric rather than against a single test situation (for person ability estimates) or a particular sample of candidates (for item difficulty estimates). Third, Rasch analysis prevails over Classical Test Analysis statistics by calibrating persons and items onto a single unidimensional latent trait scale (Bond et al., 2020).

In Rasch analysis, person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties may be conducted. Consequently, results can be interpreted with a more general meaning. One of these more general meanings involves the transferring of values from one test to another via anchor items. Once a test, or scale, has been calibrated (see e.g., Coniam et al., 2021), the established values can be used to equate different test forms.

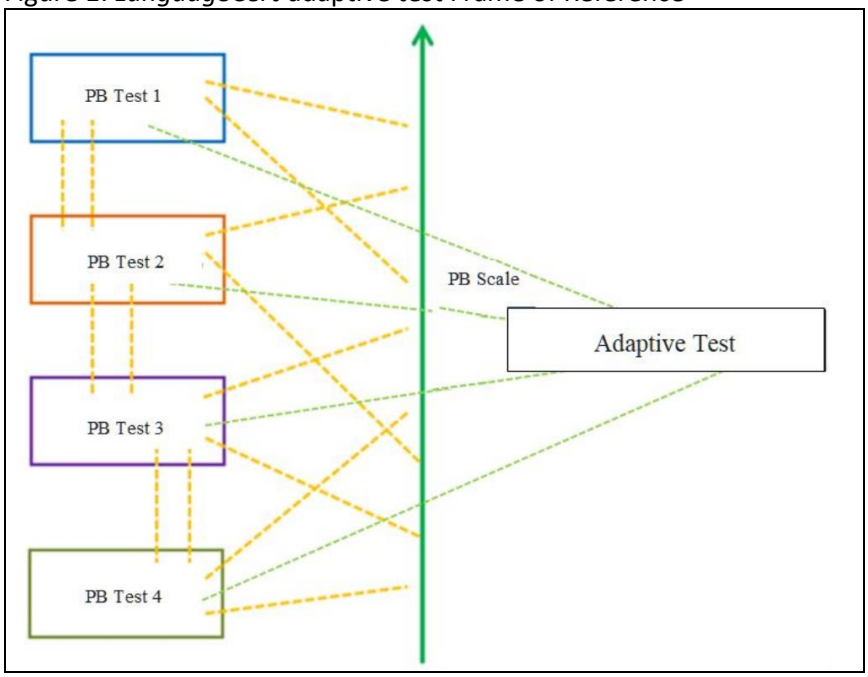
Frame of Reference

To put the operation of an adaptive test vis-a-vis the paper-based tests into perspective, reference needs to be made to the concept of the frame of reference (FOR) for measurement, and the parameters under which different tests may subsequently operate. Humphry (2006) defines a frame of reference as "compris[ing] a class of persons responding to a class of items in a well-defined assessment context." The relevance for this in the current context is that while the LTE adaptive test may be drawn from the same item bank as the paper-based tests, the "well-defined assessment context" for each test – which contributes to the computation of the overall total raw score of a test – are not necessarily congruent. The tests have, in Rasch terms, their own "internal logic" (Goodman, 1990). This internal logic refers to the starting point for any Rasch calibration and is the maximum possible raw total score of the test, computed from a particular set of items, from which the general probability of the particular test may be extrapolated (Goodman, 1990). This is the essence of the frame of reference, which Figure 1 illustrates in the context of the calibration of the paper-based LanguageCert Test of English (LTE) (Coniam et al., 2021). In that validation study, a common scale was constructed for the LTE via four paper-based (PB) – referred to as Tests 1-4 in Figure 1 below. The green arrow separating the two sets of tests is the calibrated LID scale.

In operational terms, two yardsticks indicate whether an item may be accepted within the FoR of two tests:

1. That item difficulty in both tests is comparable: there is less than 0.5 of a logit between item measures.
2. That item values occur in roughly similar positions in both tests; i.e., both items are, say, within the top 25th percentile.

Figure 1: *LanguageCert* adaptive test Frame of Reference



Key: PB = paper-based

The LTE adaptive test includes items which may also appear or have appeared in the paper-based tests. As a combined data matrix, however, the adaptive test constitutes a distinct and separate FOR from the paper-based tests. This is somewhat different from an adaptive test in the usual sense of the term. There, candidates are presented with items one item at a time from the interim paper-based scale and thus remain within the same FOR of the paper-based scale. Anchoring the items in the LTE adaptive test with values from previously-calibrated paper-based tests may not necessarily fit the new FOR: each test is an individual entity, and as such, values cannot simply be transferred from one to another. The two conditions laid out above first need to be satisfied.

In terms of analysis, the corollary is therefore that an FOR should be established for the paper-based tests, through linking via anchor items. A similar procedure is then conducted for the adaptive test – internal linking via anchor items. Once robust scales have been determined for both FORs, a merging of the two scales – of the two FORs, that is – may then be attempted. How this is achieved practically in the context of the LTE adaptive test in relation to the already existing calibration of the paper-based tests – with both being eventually merged onto a common scale – is discussed below.

The LanguageCert CAT

The LTE adaptive test assesses listening and reading from CEFR levels A1 to C2. Development began in 2019, with an initial item bank of approximately 400 items consisting of a range of listening and reading items and testlets (mini tasks of 2-5 connected items) which assessed different listening and reading constructs. The item bank furnishes test materials for both the paper-based and the computer adaptive tests. The adaptive test was trialled in late 2019 and went live in April 2020. The trial adaptive bank had approximately 900 items and the first live adaptive bank approximately 800 items. Items are continually being added to the core LTE item bank, and by early 2021, the bank comprised over 1,500 items. As is necessary with all item banks, the item bank will continue to be refreshed and grow in the future.

Analysis

The analysis of the adaptive test was conducted in early 2021, at which point the dataset consisted of 827 items and 5,870 candidates. The analysis of this dataset via the Rasch analysis software Winsteps (Linacre, 2006) is described below. Table 2 below details the summary statistics for the calibration.

Table 2: Adaptive Test Calibration Details

PERSON	5870 INPUT	5870 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	38.1	56.9	134.92	6.66	1.00	.0	1.00	.0
P.SD	5.3	2.4	32.07	1.00	.12	.9	.38	.9
REAL RMSE	6.74	TRUE SD	31.35	SEPARATION	4.65	PERSON RELIABILITY		.96
ITEM	827 INPUT	827 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	270.3	403.8	99.25	3.95	.99	-.2	1.00	-.2
P.SD	195.6	266.6	39.93	4.97	.11	2.3	.27	2.3
REAL RMSE	6.35	TRUE SD	39.42	SEPARATION	6.21	ITEM RELIABILITY		.97

A total of 5,870 candidates and 827 items were included in the calibration. Candidates took on average 57 items, from which a mean raw score of 38.1 emerged. Item reliability is high at 0.97, as is person reliability at 0.96, the latter being the equivalent of classical test theory reliability (Anselmi et al., 2019). Person infit mean-square (1.00) and outfit mean-square (1.00) fit statistics are both within the acceptable range of 0.5 to 1.5, suggesting that the calibration of persons may be taken as acceptable. By the same token, item infit mean-square (0.99) and item outfit mean-square (1.00) fit statistics are also acceptable. The overall summary calibration statistics point, therefore, to a test that may be viewed as sound.

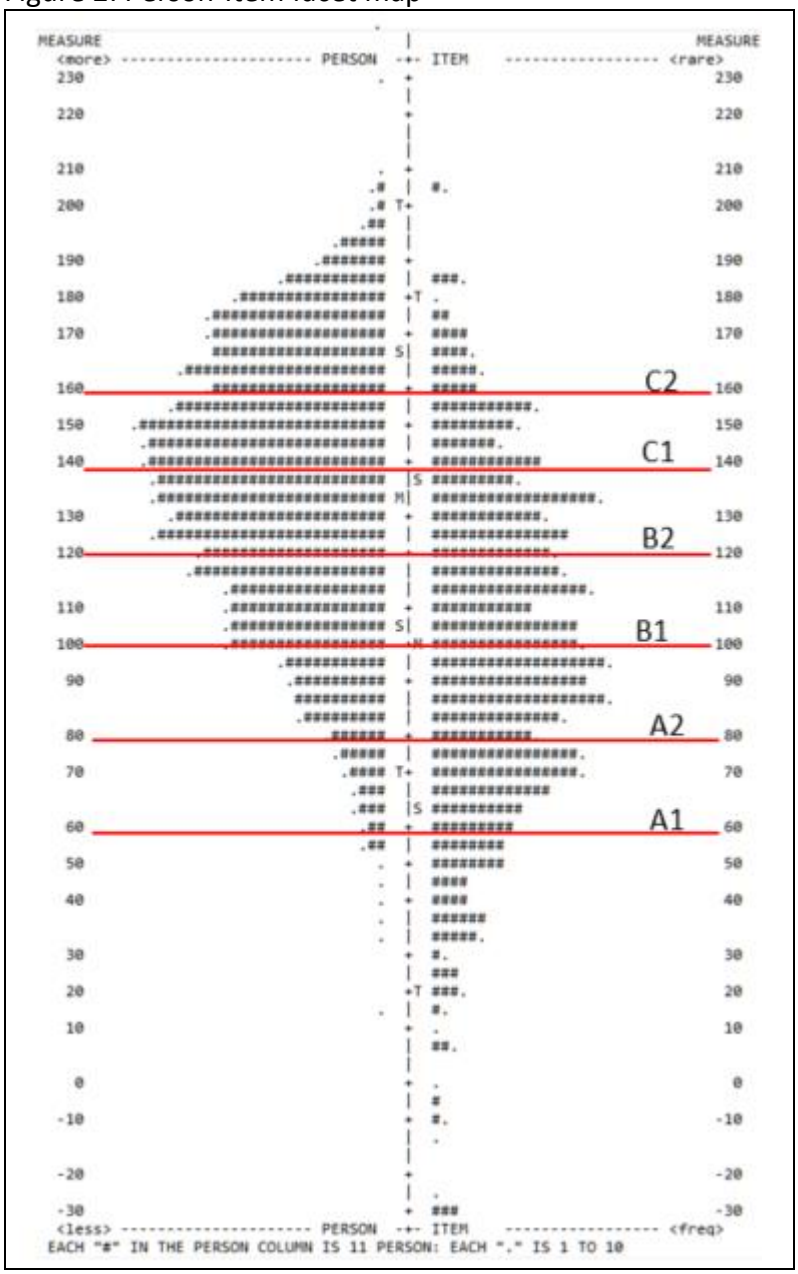
The overarching LanguageCert Item Difficulty (LID) scale lays out item difficulty levels generally adopted in LanguageCert assessments (Coniam et al., 2021). These are presented in Table 3.

Table 3: LID scale

CEFR level	LID scale range	Mid point
C2	151-170	160
C1	131-150	140
B2	111-130	120
B1	91-110	100
A2	71-90	80
A1	51-70	60

To give a visual overview of the measurement, the vertical ruler (the 'facet map') produced in the Winsteps output is presented below in Figure 2. This is a visual representation of where facets (items and candidates) are located on the scale. In Figure 2 below item/person maps are laid out such that the person spread (in logits) appears to the left-hand side of the ruler while the item spread (in logits) appears to the right-hand side of the ruler. Higher level persons (candidates) appear towards the upper left side of the map while lower level persons appear towards the lower left side of the map. Similarly, more difficult items appear towards the upper right side of the map while easier items appear towards the lower right side of the map.

Figure 2: Person-Item facet map



As Figure 2 illustrates, person and item distributions are quite wide and comparatively even in spread. Both extend approximately 120 points, or six logits – the rule-of-thumb operational range (Bond et al., 2020). Persons (on the left-hand side) extend from 60 to 200 while Items (on the right-hand side) extend from 30 to 170.

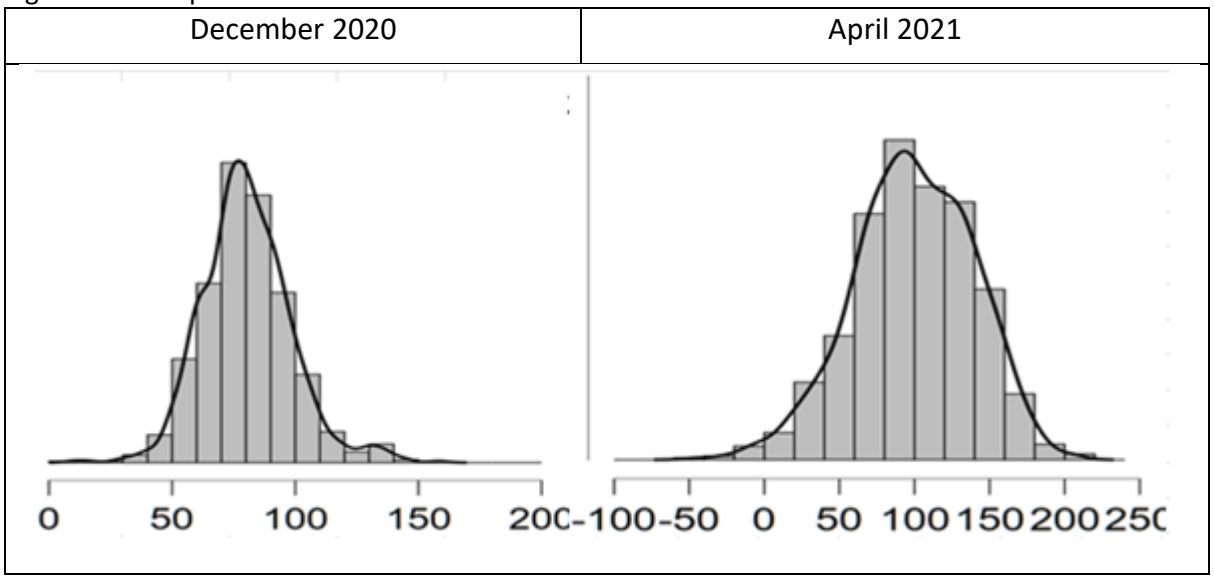
Candidates generally match with items. The midpoint of the item curve may be seen to be around B2; with persons, the midpoint of the curve may also be seen to be around B2. The Person distribution is, however, dependent upon the nature of the test population in this sample. It is known, for example, that there were a considerable number of high ability candidates in the sample.

Brief comparison with earlier calibration

The analysis described in this paper was conducted in early 2021 when the adaptive test consisted of 827 items and 5,870 candidates. A previous, exploratory analysis had been conducted in late 2020, at which point the dataset consisted of 820 items and 1,575 candidates. The section below presents some comparative analyses of the two datasets, in order to give a sense of how, with the increase in size, dataset robustness has, unsurprisingly, improved.

Figure 3 presents a virtual anchoring of items in each dataset, with the curve indicating the peak, the mid-point of each dataset.

Figure 3: Item spread in datasets



While the Pearson correlation between the two sets of data was 0.96, the mid-point has shifted slightly upwards – from items centring around 80 (B1) to around 100 (B2)

Table 4 presents an elaboration of the April 2021 dataset, with the percentiles indicating CEFR levels, and LID scale values.

Table 4: Percentiles indicating CEFR LID scale position (as of April 2021)

		Level	LID scale value
No. of items	816		
Mean	100.76		
Std. Deviation	37.98		
Maximum	204.47		
		C2	150
		C1	130
75th percentile	129.42		
		B2	110
50th percentile	99.92		
		B1	90
25th percentile	73.64		
		A2	70
		A1	50
Minimum	6.95		

The expanded calibration of the adaptive test, in terms of both item and candidate numbers, has shown improvement in the rigour of the LID scale from two key aspects:

1. The scale mid-point (the 50th percentile) is now 100 (99.92), which closely matches the item distribution mean (100.76).
2. Levels A1 and A2 now occur in the bottom 25th percentile, levels B1 and B2 in the central 50th percentile, and C1 and C2 in the top 25th percentile. Such a distribution might possibly be expected of any large candidate sample size. Everything else being equal, the mid-range ability group would be expected to occupy the major central region of the distribution while the higher and lower ability groups would be expected to occupy the upper and lower narrower range of ability.

Conclusion

As outlined in Coniam et al. (2021), the LanguageCert LID scale for all LanguageCert tests, was developed and calibrated initially against a set of paper-based tests. The initial calibrated scale that emerged provided a validation of the paper-based tests, showing them to be robust and consistent with the data. The initial scale therefore provided an acceptable basis for the further development of the LanguageCert Item Difficulty scale and the integration of LTE on to the overarching LID scale on the basis of the adaptive test data.

The focus of the current study has been to calibrate the expanded set of items in the item bank against the cohort of candidates who have thus far taken adaptive tests from the LTE item bank.

With the extension and expansion of the scale and the item bank, measurement statistic configurations necessary to achieve the goal of a robust calibrated scale have had to be taken account of. Specifically, the concept of the frame of reference for measurement has been instructive in setting parameters for co-configuring the paper-based tests as one entity, and subsequently incorporating the expanded item bank dataset and adaptive test

into a single frame. It is now possible to see, post hoc, after anchoring, that the different tests match up. All the tests may now be viewed – and may operate – within the same frame of reference.

The calibrated LanguageCert Item Difficulty (LID) scale may now be considered to be a comprehensive scale, linked to an item bank which provides both anchoring from individual tests with different FORs and individual item-based adaptive tests.

With a coherent LID LTE scale having been developed, two further projects are now being undertaken. The first of these involves a comparative study of both versions of the LTE. This involves administering – to a representative sample of candidates – versions of both the LTE paper-based test and the adaptive test. The second project, which is ongoing, involves an expansion in the size of the item bank, with concomitant confirmatory re-analysis. Currently, as reported in this paper, the item bank comprises 827 items. Once the candidate cohort reaches 10,000, further analysis will be conducted and the robustness of the calibration revisited.

References

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. Frontiers in Psychology, 10, 2714.

Bond, T., Yan, Z., & Heene, M. (2020). Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.

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

Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. Sociological Methodology, 20, 249-294.

Humphry, S. (2006). The impact of differential discrimination on vertical equating. *ARC report*.

Humphry, S. M., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement*, *9*(3), 249-264.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Pike N., & Coniam, D. (2021 in press). Adaptive testing and the LanguageCert Test of English adaptive test. ELTNEWS.

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

Language Cert

languagecert.org