Language Cert

Validating the LanguageCert Test of English Scale: The Paper-based Tests

David Coniam
Tony Lee
Michael Milanovic
and
Nigel Pike

Abstract

An appropriately validated measurement scale is a necessary prerequisite for any examination/assessment system. Such scales can be developed on the basis of expert judgement through the use of statistical techniques or through a combination of both approaches. However it is likely that effective scale construction will use a combination of expert judgement and statistical techniques.

This report documents the first phase of measurement scale development for the LanguageCert Test of English (LTE). The study describes the validation of the initial LanguageCert Item Difficulty (LID) scale which was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. The study builds on the original LanguageCert Item Difficulty scale through the use of Item Response Theory (IRT) and Rasch analysis in addition to expert judgement and CTS. This enhanced LID scale will form the empirical basis for the alignment of all current and future assessment products to the same measurement scale that is itself aligned to the CEFR.

LTE tests are produced from the LTE item bank. At the time of analysis (early 2021) the bank contained a total of approximately 1000 items from which four paper-based tests were produced to form the basis for the current study.

The report details how the test with the largest candidature (Test 3) was used as the starting point for the analysis required to establish a baseline measurement scale. Following this the other three tests were calibrated in their own right in order to provide an initial view of the distribution of persons (candidates) and items. Linking items were then anchored against Test 3 logit values after which the three tests were recalibrated.

Having calibrated the four tests onto a single scale using IRT this scale was aligned to the original LID scale. Rescaling the calibrated scale from standard logit values to a mid-point of 80 with a spacing factor of 20 resulted in a scale with was comparable to the original LID/CEFR level scale.

The fact that the calibrated Rasch scale produced from the LTE paper-based tests has emerged as well aligned to the original LID scale provides support for further integration of LanguageCert products onto the common scale and validates the use of expert judgement and CTS in the original LID scale creation. The whole process and the successful outcome support the view that that expert judgment a time proven human element in assessment and rigorous statistical modelling can and should work hand in hand for the benefit of both.

Introduction

This report documents a study based on data gathered for the LanguageCert Test of English (LTE) in order to validate the LanguageCert Item Difficulty (LID) scale created in 2017. The *LanguageCert Test of English* (LTE) 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 has been accredited by the UK's Office of Qualifications and Examinations Regulation (Ofqual) and can therefore be regarded as a high-stakes exam.

The report provides the basis for a variant of the original LID scale based on IRT that will form the basis for the alignment of all current and future assessment products to the same scale that is itself aligned to the Common European Framework of Reference (CEFR).

Current Study: Purpose

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 PB test measuring A1-B1	Test aimed at beginner to intermediate cohorts.
(2) a PB 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 purpose of the current study is to validate link and establish a common scale for paper-based variants (1) and (2). A follow up study will align this scale to the adaptive LTE test scale ensuring that candidates taking any variant (PB 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.

Test Development and Test Administration

The LTE item bank contains a total of approximately 1000 items calibrated in line with the *LanguageCert Item Difficulty* (LID) scale as laid out in Table 2.

Table 2. LanguageCert Item Difficulty (LID) scale

CEFR Level	**LID cut score**
C2	160 +
C1	140- 159
B2	120 - 139
B1	100 - 119
A2	80 - 99
A1	60 - 79
Below A1	0 - 59

The LID scale was developed on the basis of the expert judgement of a group of assessment and item writing experts who are highly experienced in writing test materials and aligning them to the CEFR. It is aligned to the six CEFR levels measuring item difficulty in a 0-200 scale where 60 is the cut score level for A1 80 for A2 and moving up by 20 points per CEFR level arriving at 160 at CEFR level C2. Items with a difficulty below 60 are included in the tests as these items measure at the Pre-A1 level. Items above 160 have a ceiling difficulty of 180.

Two PB tests measuring the range A1-B1 and two measuring A1-C2 were assembled as shown in Table 1 above. Table 3 below presents an overview of the four tests constructed and the number of candidates taking the tests in this study.

Table 3. Four paper-based tests

Study test name	Items	Number of candidates	Target CEFR levels
Test 1	72	721	A1-B1
Test 2	72	93	A1-B1
Test 3	110	1161	A1-C2
Test 4	110	137	A1-C2
Total	364	2112	

Tests 1 and 3 have considerably larger sample sizes thus making the analysis for these tests more generalisable.

Common items were included across the four tests and it is on this basis that scale development and calibration were conducted. Table 4 shows the location of common items across the four tests.

Table 4. Common items across tests

	Test 1	Test 2	Test 3	Test 4	Total
Test 1			19	21	40
Test 2			22	20	42
Test 3	19	22			40
Test 4	21	20			42

While there were 364 items in total across the four tests 82 of these were common items. This meant that there were 282 discrete items in the four-test database.

Brief Overview of Rasch Analysis

This section presents a brief overview of the Rasch statistical procedures used in the study.

In contrast to Classical Test Statistics (CTS) the use of the Rasch model enables different factors or facets (e.g. person ability and item difficulty) to be modelled together.

Firstly 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 (Linacre 2006). 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.

Secondly once a common metric is established for measuring different phenomena (test takers and test items being the two most obvious) person ability estimates can be considered independent of the items used with item difficulty estimates being independent from any sample used 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 test takers (for item difficulty estimates).

Thirdly Rasch analysis improves on CTS by calibrating persons and items onto a single unidimensional latent trait scale – also known as the one-parameter IRT (Item Response Theory) model (Bond and Fox 2007; Wright 1992). 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. Consequently results can be interpreted with a more general meaning; that is extrapolating reliably beyond the current test. Further as the Rasch model provides a great deal of information about each item in a scale its use enables a better evaluation of individual items and how these items function in a scale (Törmäkangas 2011).

A brief description of key Rasch terms and statistical procedures is provided in Appendix 1.

Calibrating the LTE tests

Introduction

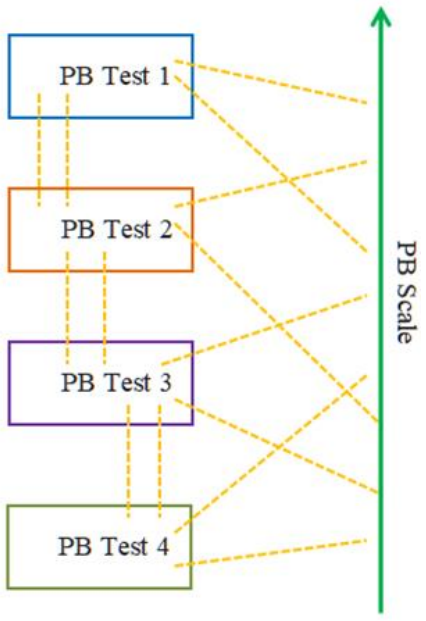
This section describes the calibration of the four paper-based (PB) tests. The calibration exercise consisted of two key stages. The first involved establishing the internal consistency of the items in all four PB tests and linking these to a unified metric with a view to establishing internal consistency reliability The second involved pulling all items together to form an initial PB test measurement scale.

Frame of Reference

In the context of Rasch assessment it is important to bear in mind the concept of "frame of reference" (FOR). Humphry (2006) discusses how in the context of assessment a frame of reference "comprises a class of persons responding to a class of items in a well-defined assessment context." Given that the output of a Rasch calibration are sample-independent item difficulties and test-free person ability estimates there may be a temptation to believe that the tests themselves are unimportant. However this is not the case. Because the basis of Rasch estimates is the total score on a test when working with multiple tests as is the case in this study the overall frame of reference must be taken into account.

In Figure 1 below the four PB tests analysed in the current study are linked to the common PB scale by equating via common items. In the first instance while they retain distinct FORs and item locations and are hence legitimately placed on the PB scale they have to be interpreted within their own respective FOR.

Figure 1. AT Frame of Reference



Scale Construction Strategy

The principal statistical tool used for calibrating the Rasch scale is the unidimensional Rasch measurement model using Winsteps (Linacre 2020). The common or linking items across the PB tests provide anchoring points where all items from the four PB tests form elements of the scale. In addition the existing linkage to the CEFR levels in the original LID scale underlying the four PB tests and whether and to what extent the scale is aligned to CEFR levels is investigated.

The scale construction included the steps laid out below.

Step 1

Given that Test 3 had the largest candidature (N=1161) and number of items (N=110) Test 3 was taken as the starting point for the analysis to establish a baseline measurement scale. The larger sample size in Test 3 enables a higher degree of precision and stability for this baseline than is the case with smaller sample tests. (Following this initial results were investigated to establish that the goodness-of-fit for Test 3 was adequate to provide the baseline and starting point of the scale construction.)

Step 2

Test 1 was first calibrated on its own so as to provide an initial view of the distribution of persons (candidates) and items. Linking items were then anchored at Test 3 logits after which Test 1 was recalibrated. The results of the two calibrations were then compared for any significant distortions that may have emerged in the anchored results. A large discrepancies between the two would indicate 'disturbance' – that is anchored item values being either under- or over-estimated in the recalibration of either items and/or persons.

Step 3

The same method as in the Step 2 process was followed with Test 2.

Step 4

The same calibration approach was then used for Test 4. Taking anchor items from both Tests 1 and 2 enabled Test 4 to be linked to Test 3 despite the lack of any direct links between the two tests. In a similar fashion Tests 1 and 2 were linked via linking items obtained from Test 3.

Background to Analysis

The key Rasch analysis elements that form the basis for the analysis and discussion in this report are:

Overall Calibration Tables

Here reference is made to Infit Mean Squares (IMNSQ) and Outfit Mean Squares (OMNSQ).

Variable (Item / Person) Maps

In the Figures 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.

In standard Rasch output logits are presented so that zero is the mid-point with an SD or spacing factor of 1 between logits. Under such an output above zero (a positive value) means a person of higher level and a more demanding item; below zero (a negative value) means a person of lower level and an easier less demanding item. To make the interpretation of logit values more user-friendly logits may be rescaled – often with the intention of all values being positive. In the analyses of the initial calibrations of the four tests presented below 100 was set as the initial mid-point of the scale (zero logits) with one SD rescaled as 20. The red lines in the Figures below indicate these calibration mid-points.

It should be noted that of the four tests, Tests 1 and 2 aimed at A1-B1 candidates. Candidates sitting these tests have thus only been able to be graded up to B1.

Data Analysis and Interpretation

This section presents the calibration analyses for each test. Test 3 is calibrated first to produce anchor items against which the other three tests may be subsequently linked. The other three tests (Tests 1 2 and 4) are then calibrated twice: firstly in their own right and secondly against Test 3.

Test 3: Initial Calibration

The overall calibration results are presented in Table 5.

Table 5: Test 3 – overall calibration results

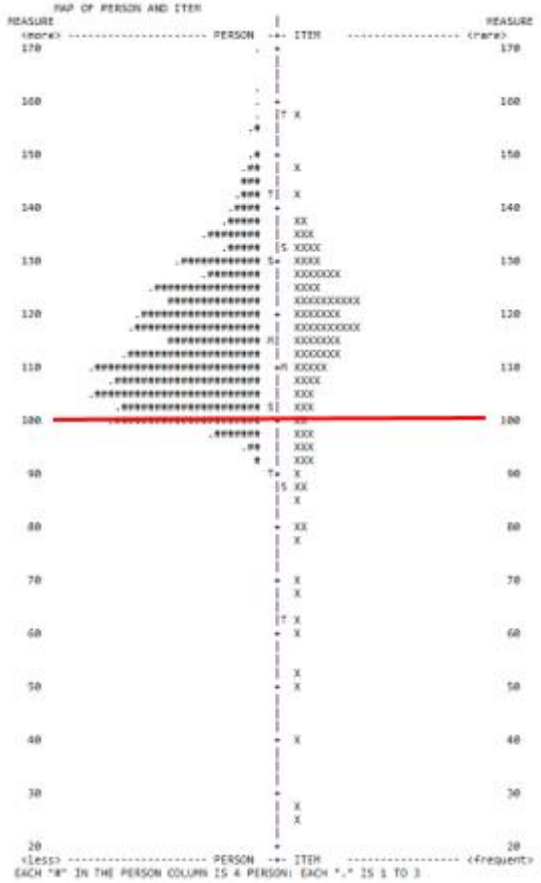
PERSON	1161 INPUT	1161 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	58.6	108.9	115.87	4.47	1.00	.0	.99	.0
S.D.	14.0	4.1	13.08	.36	.11	1.3	.23	1.1
REAL RMSE	4.48	TRUE SD	12.29	SEPARATION	2.74	PERSON RELIABILITY		.88
ITEM	110 INPUT	110 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	618.9	1149.2	110.00	1.48	1.00	-.1	.99	-.2
S.D.	221.1	23.8	23.23	.61	.10	4.7	.15	4.7
REAL RMSE	1.61	TRUE SD	23.17	SEPARATION	14.43	ITEM RELIABILITY		1.00

The key indices to be noted in Table 5 are:

- Reliability (overall) which for the test items at 1.0 is very good.
- The Separation index of 14.43 indicates that the True SD (the amount of variance among items) is more than 15 times the error indicating that there is a large separation and a small standard error in the item calibration.
- The Item Outfit Mean Square (OMNSQ) is the measure (in standard errors [SE]) of how items are grouped around the calibrated measure. In Table 6 item outfit at 0.99 is less than one SE indicating there are no clear outliers among the items. This confirms that the items form a relatively coherent assessment.
- Item Infit Mean Square (IMNSQ) measures the SEs within an item. Table 6 shows item infit to be 1.0 SE indicating good information (neither too wide nor too narrow) from the options in the items. This suggests that the items have been well constructed.

Figure 1 below lays out the Person/Item calibration map for Test 3. In the analysis of Test 3 logits have been rescaled to a mean of 100 and an SD of 20. The red line indicates the position of the mid-point of calibration.

Figure 1. Person / Item calibration map for Test 3



Persons: Left Side

Items: Right Side

Scale: Mid-point 100

One logit = 20

'S' = 1st SD

'T' = 2nd SD

'M' indicates the Person and Item mean

- From Figure 1 we see that both Person and Item distributions are quite wide and comparatively even in spread. Persons (on the left-hand side) extend from 90 to 150 (3 logits) while Items (on the right-hand side) extend from 90 to 140 (2.5 logits).
- Candidates are generally well matched with items except for the most able candidates (to the top left of the figure) where there are very few items which match the person abilities.

- Items below 1st SD (85) are too easy as all or nearly all candidates in this sample were able to answer these items correctly.

Calibration of Tests 1 2 and 4

Following the initial calibration of Test 3 the three remaining tests were analysed. In this procedure each test was first calibrated in its own right to examine its initial goodness of fit. Tests 1 2 and 4 were then recalibrated with items anchored to Test 3 through linked items. A number of items which appear in either Test 1 or Test 2 also appear in Test 3. When recalibration of Test 1 and Test 2 was carried out the items common with Test 3 were anchored at the values set in Test 3. Similarly re-calibration of Test 4 was anchored at the values of common items between Test 1 or Test 2 with Test 4. Pre- and post-anchoring results were then compared to explore whether any noticeable mismatch occurred between the two calibration exercises. Finally anchored results were aligned to produce a common scale.

For the sake of efficiency only the procedure adopted for Test 1 is discussed below.

Test 1: Analysis and Calibration

Test 1 when calibrated in its own right

The overall calibration results for Test 1 are presented in Table 6.

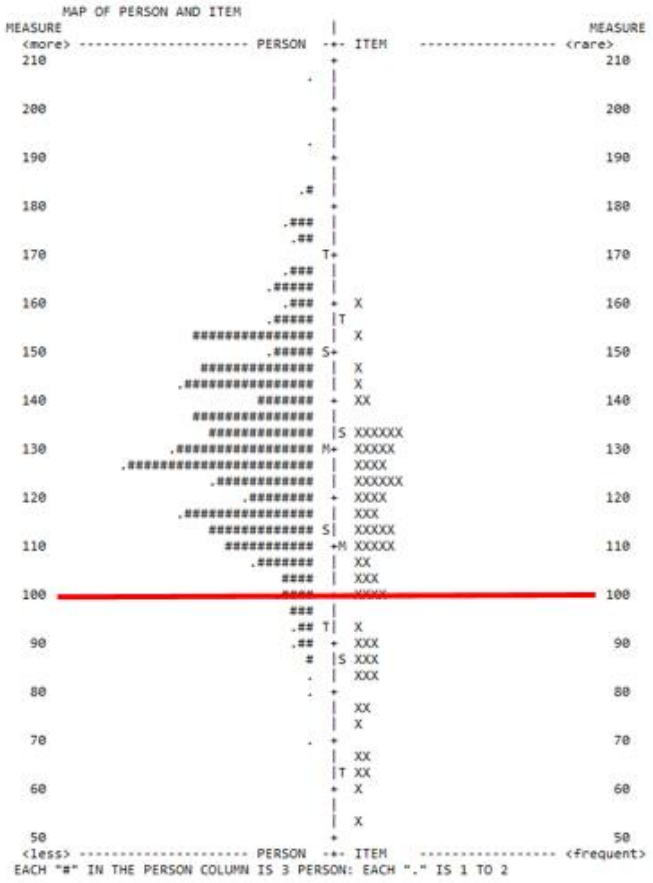
Table 6. Test 1 – overall calibration results.

PERSON	721 INPUT	718 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	48.3	71.2	131.61	6.37	1.00	.1	.95	.0
S.D.	11.6	4.2	19.91	1.33	.18	1.3	.37	1.2
REAL RMSE	6.50	TRUE SD	18.82	SEPARATION	2.89	PERSON RELIABILITY		.89
ITEM	72 INPUT	72 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	481.5	709.5	110.00	2.08	.99	.1	.94	-.2
S.D.	128.3	7.3	23.59	.59	.14	3.5	.30	3.7
REAL RMSE	2.17	TRUE SD	23.49	SEPARATION	10.84	ITEM RELIABILITY		.99

- Overall item reliability of 0.99 is very high as is item separation at 10.84.
- Item Outfit Mean Square (OMNSQ) is 0.94 is less than one SE and indicates there are no clear outliers among the items. Item Infit Mean Square (IMNSQ) is 0.99 indicating good information was provided from the options in the items. This confirms that the items have been constructed well and the items form a coherent assessment.

Figure 2 presents the Person/Item calibration map for Test 1. Logits have been rescaled to a mean of 100 and an SD of 20.

Figure 2. Person / Item calibration map for Test 1



The Rasch analysis suggests the following:

- Both Person and Item distributions are quite wide and even in spread. Persons extend from 85 to 180 (4 logits) while Items extend from 60 to 150 – 4.5 logits.
- Candidates are located higher on the scale than items indicating that the test is relatively easy for this group of candidates. Items below the 1st SD (85) and especially the 2nd SD (65) are too easy as all or nearly all candidates were able to answer them correctly.

As mentioned above Test 1 (and Test 2) was intended only for A1-B1 candidates. Candidates scoring above B1 are graded as B1 by default. This in part may help to account for the discrepancies in the two sets of analyses presented.

Some items may actually be at B2 level (120 in Figure 2 above) given that a few items appeared very difficult for the cohort

Test 1 calibrated with Test 3 item anchors

The analysis presented below is a reanalysis of Test 1 anchoring it to Test 3 using the 16 common items in the two tests. Table 7 presents these results.

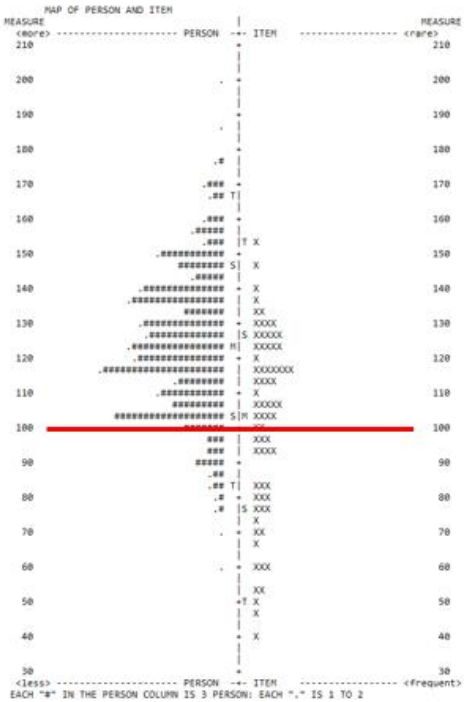
Table 7. Test 1 Calibration with Test 3 item anchors

PERSON	721	INPUT	718	MEASURED		INFIT		OUTFIT
	TOTAL	COUNT		MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	48.3	71.2		124.49	6.56	1.07	.6	1.16	.6
S.D.	11.6	4.2		20.55	1.29	.19	1.3	.46	1.2
REAL RMSE	6.68	TRUE SD	19.43	SEPARATION	2.91	PERSON RELIABILITY			.89
ITEM	72	INPUT	72	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT		MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	481.5	709.5		102.08	2.31	1.13	1.1	1.16	.7
S.D.	128.3	7.3		26.05	1.17	.55	4.0	.98	4.0
REAL RMSE	2.59	TRUE SD	25.92	SEPARATION	10.01	ITEM	RELIABILITY		.99

- Item overall reliability at 0.99 and separation at 10.01 are both high.
- Item Outfit Mean Square (OMNSQ) is 1.16 – less than one SE; Item Infit Mean Square (IMNSQ) is 1.13 indicating good information being provided from the options in the items.

Figure 3 presents the Person/Item calibration map for Test 1 after anchoring. The logits have been rescaled to a mean of 100 and an SD of 20. The red line indicates the mid-point.

Figure 3. Person / Item calibration map for Test 1 after anchoring



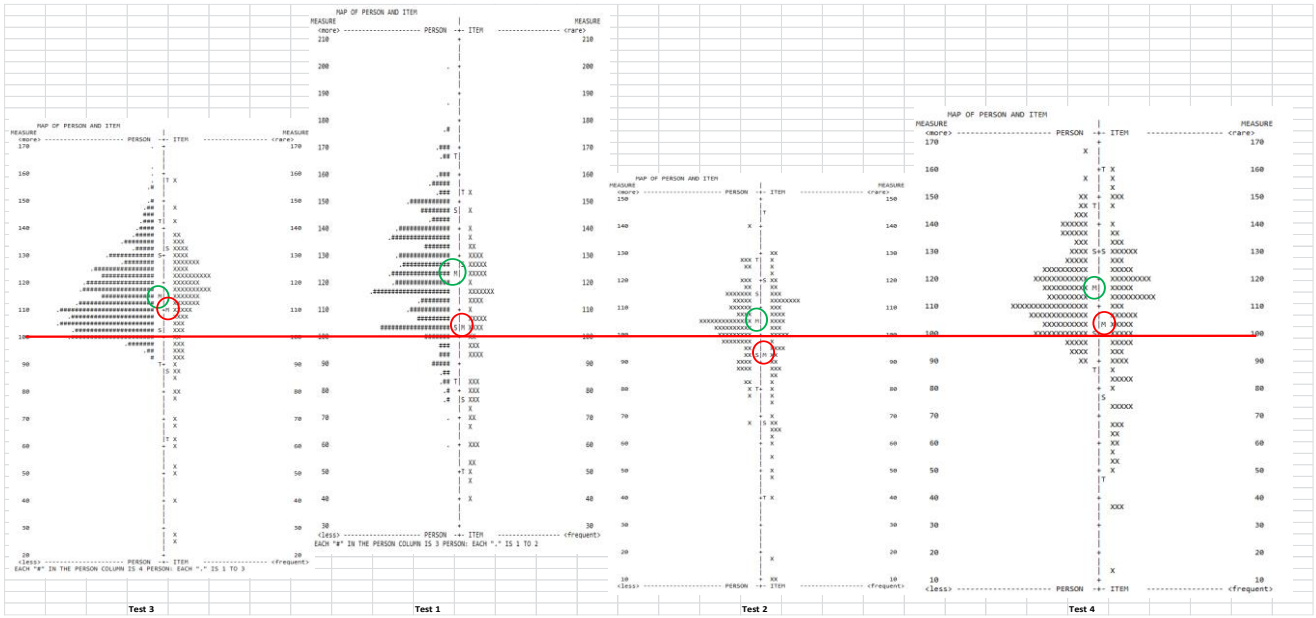
- After anchoring the Person distribution remains relatively unchanged if slightly lower ranging from 80 to 170 (4.5 logits).
- The item distribution also remains almost unchanged after anchoring perhaps shifting slightly lower with a range of 50 to 150 (5 logits). Quite a number of items to the bottom right of the scale are below the level of the candidates although this is to be expected with a slightly truncated sample.

As mentioned above the procedure conducted with Test 1 regarding the initial calibration and then recalibration with Test 3 items anchors was also conducted with Tests 2 and 4.

Recalibrating

Figure 4 now presents a composite picture of the Person/Item maps of the four anchored test calibrations. The mid-points for both Persons and Items are indicated by the circled 'M' – green for Persons and red for Items. The red horizontal line indicates the mid-point the origin of the Rasch scale of 100 or zero logits. The order of presentation of the tests in Figure 4 follows the order in which the tests were calibrated; namely Test 3 first followed by Test 1 Test 2 and Test 4.

Figure 4. Candidate and Item distributions across the four tests after anchoring

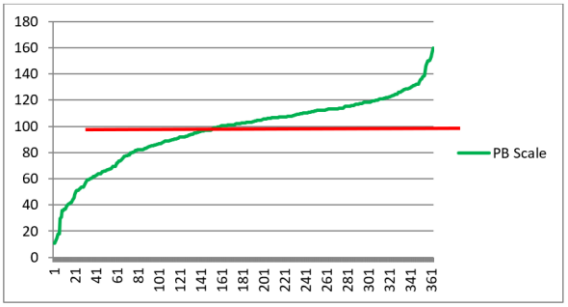


Legend: Red circled 'M' = Item mean; green circled 'M's = Person mean

Figure 4 illustrates from the positioning of the mid-point of 100 that Persons and Items in the four tests were generally above the scale mid-point. With the exception of items in Test 2 candidates were generally slightly more able and items slightly more demanding.

Figure 5 presents the relative difficulty of all 364 calibrated items in the four tests after anchoring.

Figure 5. TCC (Test Characteristic Curve) of 364 calibrated items



The vertical axis of the TCC in Figure 5 represents item difficulty levels and the horizontal axis represent the number of items. We can see that there is a steep progression of difficulty up to 60 indicating that there are relatively few items (about 40) between 10 to 60. Difficulty progression then moves upward steadily until it reaches 130 when the slope becomes steeper with about 15 items in this section. There are therefore about 55 items at the two ends of the item difficulty spectrum or about 15% of the total and covering the range A1-B2+. The majority of items (about 85%) fall between 60 and 130 across the mid-range of the scale.

Recalibrating the Scale

Having calibrated Tests 1-4 onto a single scale – taking Test 3 as the baseline – the next step involved examining the alignment of the newly-calibrated scale with the original LID scale.

To establish a baseline for Test 3 logit values had initially been rescaled to a mid-point of 100 with a spacing factor of 20. An advantage of Rasch is that as long as the recalibration with the new mid-point does not alter the original calibration results different mid-points may be used to suit specific calibration exercises (see https://www.winsteps.com/winman/rescaling.htm for an elaboration). Such a procedure may be viewed as being similar to changing individual tests' mid-points via anchoring.

Rescaling was subsequently conducted following discussion with the test development team such that a new mid-point of 80 was applied to match initial LID scale with the 20-point spacing factor maintained. Following this realignment the whole test calibration process with anchoring was performed again with Test 3 as the starting point and the other three tests calibrated to Test 3 values. It must be pointed out here that like all statistical procedures Rasch calibration is content free. The interpretation of calibration results are guided by considerations beyond the statistical procedure as long as the principles underlying the statistical procedures are not violated. The anchored calibration of the paper-based tests based on Test 3 is the best amongst equals. The initial anchored calibration will gradually be refined as the item bank develops.

The final mapping of the four PB tests onto a single scale with the mid-point of 80 is shown in Table 8.

Table 8. Test 3 – Candidate distributions via LID and Rasch-calibrated (mid-point 80) scales

CEFR level	LID level cut scores	Candidates achieving grade via LID scale	Candidates achieving grade with Rasch (80) scale
C2	160	0%	0%
C1	140	2%	1%
B2	120	10%	10%
B1	100	21%	35%
A2	80	36%	48%
A1	60	28%	5%
pre-A1	40	2%	1%

With the mid-point of 80 the two scales are more closely aligned. Apart from some misalignment at the A level the two scales are quite comparable.

With the mid-point of 80 Figure 6 now presents candidate distributions across all four tests after anchoring. The order of presentation follows the order of calibration with Test 3 first.

Figure 6. Candidate and Item distributions across all four tests after anchoring

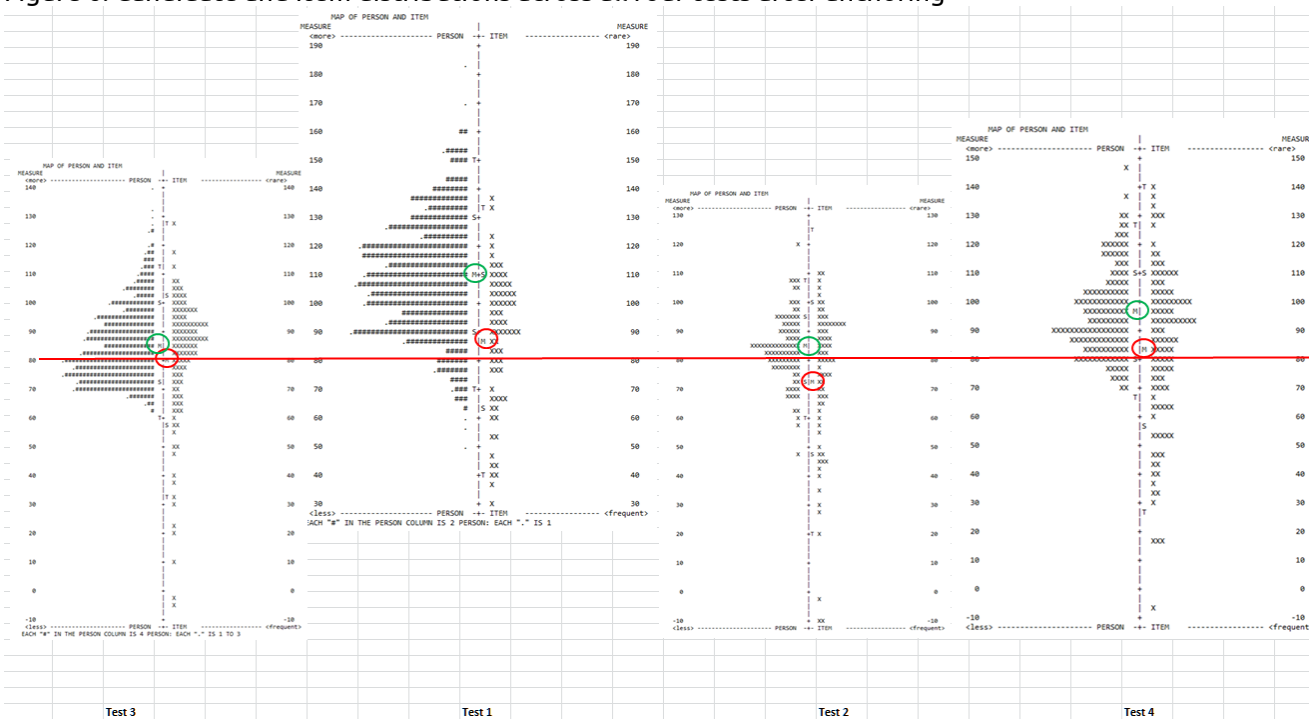


Figure 6 shows that the comparative standing of the four tests vis-a-vis one another has not changed with the use of the new scale mid-point of 80.

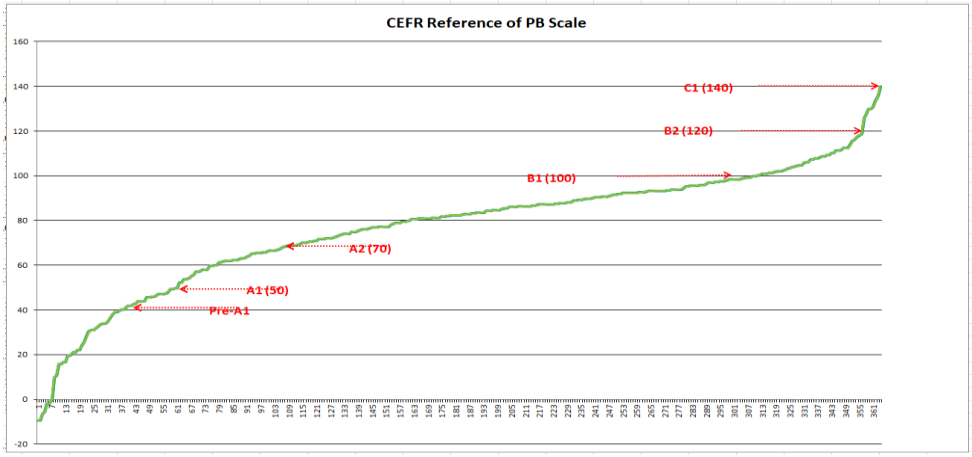
Table 9 and Figure 7 below present the finalised Rasch-calibrated PB scale illustrating how the recalibrated scale matches CEFR levels and the original LID scale.

Table 9. LID and Rasch-calibrated PB scale cut score match

CEFR level	LID cut scores	Recalibrated Rasch scale cut scores
C2	160	
C1	140	140
B2	120	120
B1	100	100
A2	80	70
A1	60	50
Below A1		

In Figure 7 below the red arrows and values represent the cut score points for the different levels.

Figure 7. Finalised Rasch-calibrated PB scale



The LID scale was initially developed as a linear scale with the cut scores for each CEFR level as in Table 9 above. As can be seen there is a very close fit between the original LID scale and the Rasch-calibrated scale. C1 B2 and B1 match exactly with 20 points (or one logit) between each level. Between B1 and A2 the Rasch analysis suggests a slightly wider gap – 1.5 rather than one logit. Between A2 and A1 there is again a 20-point difference. Between A1 and pre-A1 the Rasch analysis suggests the gap should be only half a logit or 10 points.

In sum then the Rasch-calibrated scale from pre-A1 up to C1 extends 100 points or five logits with the Rasch rescaling corresponding very closely – with the exception of A1 and A2 – to the original LID scale. The weaker alignment here needs to be investigated further.

Conclusions

The study reported above had two major objectives. The first was to calibrate using Rasch measurement the existing paper-based version of LanguageCert Test of English (LTE) onto a common scale; the second was to examine the subsequent alignment of the common scale produced with the existing LanguageCert Item Difficulty (LID) scale developed on the basis of Classical Test Statistics (CTS) and expert judgement in order to lay the foundations for a single unified measurement scale aligned to the CEFR that would underlie all LanguageCert assessment products.

The report details how Test 3 with the largest candidature (N=1161) and number of items (N=110) was taken as the starting point in terms of establishing a baseline measurement scale. The other three tests after having been first calibrated in their own right were then anchored to Test 3 via linking items drawn from Test 3 after which the three tests were then recalibrated. This process resulted in all four tests eventually being calibrated onto a single scale.

With all four tests on a single scale the calibrated scale was rescaled to a mid-point of 80 with a spacing factor of 20 in order to align the calibrated Rasch scale and the original LID scale. The rescaling of the Rasch scale in this manner produced a comparable alignment between the two scales although there were some differences at the A1 A2 and B1 levels that need to be further explored.

The next step is to calibrate the LTE adaptive test also generated from the LTE item bank to the common Rasch scale produced in the current study. This will also entail a revisiting of the Frame of Reference concept.

References

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

Fisher Jr W. P. (1992). Reliability statistics. Rasch Measurement Transactions. Chicago USA: MESA Press.

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

Linacre J. M. (2006). A user's guide to WINSTEPS: Rasch-model computer programs. Chicago IL: Winsteps.com.

Linacre J. M. (2006) Winsteps. Rasch measurement computer program. Chicago: MESA Press.

McNamara T. (1996). Measuring second language performance. New York: Longman.

Weigle S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263-287.

Weir C. J. (2005). Language testing and validation: An evidence-based approach. Houndmills UK: Palgrave Macmillan.

Wright B. D. (1992). IRT in the 1990s: Which models work best? Rasch Measurement Transactions, 6, 196-200.

Wright B. D. & Masters G. N. (1982). Rating scale analysis: Rasch measurement. Chicago Illinois: MESA Press.

Appendix 1: Rasch Statistics

The unit of measurement in Rasch analysis is the 'logit'. These are evenly spaced along the Rasch 'ruler' or scale.

In the default Rasch output logits are presented such that zero is the 'origin'. This origin is labeled the 'mid-point' (used hereafter in the report) with a Standard Deviation (SD) or spacing factor of 1 between logits. The zero origin represents the situation where item difficulty equals person ability in the measurement – similar to when a scale is balanced in weighing where the object being weighed is equal to the weight applied. Under such a system of presentation above zero (a positive value) means a more able person (or candidate) and a more demanding item; below zero (a negative value) means a less able person or an easier item. If person ability is greater than item difficulty the measurement result will be positive; in the case of person ability being lower than item difficulty the measurement result will be negative.

Rescaling Logits

To make the measurement generally intelligible the measurement results can be rescaled to suit specific situations with all values appearing positive. For example by making 50 the mid-point and 10 the Rasch scale unit (the logit) the resulting measurement results would be on a scale of about 0 to 100 making it popular in many assessment contexts – see Wright and Stone (1979) for a discussion of this. Such rescaling changes do not affect the measurement results.

In the analyses of the initial calibrations of the four tests presented below 100 was set as the initial mid-point of the scale (zero logits) with one SD rescaled as 20. The red lines in the Figures below indicate these calibration mid-points. This was implemented in order to align the results with the LID scale.

Fit

Fit in Rasch may be seen as the 'fit' of the data to the Rasch model. In essence this refers to how well obtained values match expected values. Fit itself is then divisible into a number of related if slightly different categories. A perfect fit of 1.0 indicates that obtained values match expected values 100%. Acceptable ranges of tolerance for fit range from 0.7 to 1.3 (see e.g. Weigle 1998).

Infit

Infit Mean Square measures the standard error (SE) within an item or a person. A large SE indicates a random or scattered distribution within an item; in contrast if Infit Mean Square is small this indicates too little variation in the items indicating limited useful information.

An Infit of 2 or more would mean rather scattered information within the item or person providing a confused picture about the placement of the item or person. An infit index which was very small (e.g. <0.5) would indicate that there is only very small variation and therefore very little information to place the item or person. A low Infit figure indicates too small a variation to articulate clear and meaningful judgments and measure across a limited range.

Infit may therefore be seen as the 'big picture' in that it scrutinises the internal structure of an item or person.

In the Winsteps output Item Infit Mean Square is IMNSQ.

Outfit

Outfit is the measure of how items are grouped around the calibrated measure and is measured in terms of standard errors. Outfit refers to an item or person as an element within the pool of items or persons being calibrated. An Outfit larger than 2 would flag an item or person as being out of line with the rest in the pool an outlier. A small degree of Outfit is of little consequence within the bigger picture.

In the Winsteps output Item Outfit Mean Square is OMNSQ.

Reliability

Reliability in Rasch-calibrated terms is comparable to reliability (KR-20) in CTS. The major difference is that Rasch statistics are based on calibrated totals with higher scores reflecting a higher ability. This is in contrast to classical reliability which is based on un-weighted and un-calibrated totals. Classical reliability and Rasch reliability are therefore the same thing except that the former is based on un-weighted and un-calibrated totals while the latter is based on calibrated totals. A minimum reliability of 0.8 is taken as the benchmark for classical reliability with a similar threshold recommended for Rasch although achieving reliability estimates as close to 1.0 as possible is desirable in most assessment contexts.

Rasch calibration reports item and person reliability in respect of the measure rather than KR20. As Rasch reliability is based on calibrated and assessment-measure-aligned total scores the metric is more precise than the KR20.

Separation

Separation is the ratio of True SD / Real RMSE (Residual Mean Square). To elaborate separation indicates how much True SD (i.e. variance) there is in each unit of error with the greater the separation figure the better. In general Rasch tolerates a minimum of 3 and an optimum of 9 in separation. With reliability recommended as 0.8 or better under CTS three distinct strata would be preferred in a Rasch analysis.

Language Cert

languagecert.org