Language Cert

Externally-Referenced Anchoring of LanguageCert SELT Tests

Michael Milanovic
Tony Lee
David Coniam
and
Yiannis Papargyris

Abstract

This paper reports on the use of externally-referenced anchoring by LanguageCert as a methodology for vertically aligning test forms: i.e., aligning test forms to a calibrated midpoint.

An analysis is presented of a sample of the Listening and Reading test forms which comprise the LanguageCert SELT tests which assess at CEFR levels B1–C1. Using Rasch measurement to vertically align tests on the basis of prior expert judgement (Lee et al., 2022), the robustness of the LanguageCert SELT B1–C1 tests is illustrated. An analysis of the test forms reveals three findings of close matches:
1) between the items in the different test forms;
2) between the test forms and the LanguageCert Item Difficulty (LID) scale; and as a consequence;
3) between the test forms and the respective CEFR levels.

The results provide support for the claim that LanguageCert SELT tests are well set, with each test appropriately positioned at its respective CEFR level.

Introduction

This report extends LanguageCert's exploration of quality in its examinations (see e.g., Coniam et al., 2021a; 2021b). Considerable importance is now attached to English language qualifications for work and study; this is reflected by the UK Visas & Immigration (UKVI) establishing Secure English Language Tests (SELT) tests for movement and work to the UK. LanguageCert was approved in 2020 as a provider of UK Home Office approved SELT tests and offers LanguageCert SELT (LST) four-skills tests at a range of levels, mapped to the Common European Framework of Reference (CEFR) for UK Visas & Immigration (UKVI) worldwide, covering all visa type requirements to live, work or study in the UK.

In line with the type of visa being applied for to the UKVI, a language test exhibiting proof of competency in English at a particular level needs to be passed. Against this backdrop, this paper examines the statistical quality of the LST B1–C1 Listening and Reading Tests, approved for UKVI language certification purposes, and which were produced over the period 2020-2021. All test forms comprise 52 items.

Against the key test qualities of validity and reliability (Bachman & Palmer, 2010), central validity issues include how well the different parts of a test illustrate what a test taker can do – i.e., communicate – in English, and how well test scores provide an indication of test taker ability in relation to communicative language competence (Messick, 1989; Bachman & Palmer, 2010). The LST tests assess the communicative skills that test takers will be expected to control at particular levels of ability (i.e., in relation to the CEFR). Test content matches target test takers – in terms of grammar, functions, vocabulary, topics etc., and the tasks have correspondingly relevant 'communicative' contexts.

If tests are to be of high validity and reliability, they need to be well constructed (Hughes, 2003). In this regard, LanguageCert test item writers are of the highest international standard and have extensive expertise in, and knowledge and understanding of, the different CEFR levels (see Papargyris & Yan, 2022). Test items are linked to the CEFR by expert judgement, a methodology which has proven – as long as adequate training and standardisation are in in place – to be robust (Coniam et al., 2022).

The LST B1-C1 test forms analysed constitute a sample of the test forms delivered by LanguageCert in the 18-month period from mid 2020 to late 2021. For security purposes, all LST Listening and Reading tests are currently constructed as standalone tests. Since test forms are separate from one another, there are no linking items or test takers by which direct cross-calibrating may be conducted. Nonetheless, the externally-referenced anchoring methodology pioneered by Lee et al. (2022) permits tests which have no common linking items to be vertically linked against the test's midpoint using previously-established item values by expert judgement. It is therefore this methodology – externally-referenced anchoring – which is used in the current study to explore how accurately the different LST B1–C1 test forms are anchored onto the LanguageCert Item Difficulty (LID) scale, and hence to the CEFR.

The key to establishing the appropriate points on the LID scale involves the use of expert setters and their concomitant expert judgement. Such 'expert judgement' in language assessment is therefore a key factor in test development both in the area of item writing and test setting as well as in the estimation of item difficulty, which in turn impacts level setting and cut scores.

In the case of test setting, the use of experts is a critical requirement. While there has been debate over the use of expert judgement in standard setting (e.g., Alderson & Kremmel, 2013), generally, the use of expert judgement has been accepted as having a valid role in the field of language assessment for test validation and standard setting – see Lumley, 1993; Gable & Wolf,1993; Bachman et al, 1995. Relatively recent validation studies involving expert judgement include VanderVeen et al. (2007), Song (2008), Gao and Rogers (2011), and van Steensel et al. (2013). In these studies, judges were reported to have reached high levels of agreement. The positive use of expert judgement is reflected in Lee et al.'s (2021) study utilising externally-referenced anchoring with other LanguageCert CEFR-related tests – the IESOL suite of tests (see also Coniam et al., 2022).

The LanguageCert SELT Tests

The LST suite comprises tests at CEFR levels B1 to C2. Examination specifications reflect the requirements of the CEFR; with test materials writers having extensive expertise in, and knowledge and understanding of, the CEFR.

Each LST test has a designated CEFR level, with, as mentioned, all test forms carefully set using expert judgment and reviewed by other expert staff in the LanguageCert Assessment Team. The LanguageCert Item Difficulty (LID) scale referred to above is the metric against which items are linked to the CEFR on the basis of item difficulty. The LID scale was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement by a group of assessment and item writing experts who are highly experienced in writing test materials and aligning them to the CEFR. The LID scale may be found in Table 2 below.

Studies by Coniam et al. (2021a; 2021b) have validated and extended the LID scale beyond its original CTS origins to a Rasch-based calibration where all levels are statistically validated and linked.

The four-skills LST tests are located on the LanguageCert Global Scale [Note 1] along with other LanguageCert test products: the LanguageCert Test of English, and the International IESOL suite of English language tests.

Since the methodology surrounding externally-referenced anchoring relates to the use of Rasch measurement, a brief overview of Rasch will now be presented.

Rasch Measurement

The use of the Rasch model enables different facets to be modelled together, converting raw data into measures which have a constant interval meaning (Wright, 1997). 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. In Rasch measurement, test takers' theoretical probability of success in answering items is gauged; scores are not derived solely from raw scores. While such 'theoretical probabilities' are derived from the sample assessed, they are able to be interpreted independently from the sample due to the statistical modelling techniques used. Measurement results based on Rasch analysis may therefore be interpreted in a general way (like a ruler) for other test taker samples assessed using the same test. In recent decades, Rasch analysis, it should be noted, has complemented and in some cases replaced classical test statistics in enabling stakeholders to appreciate better what is being measured and how it is being measured with greater sophistication than before.

In Rasch analysis, test taker measures and item difficulties are placed on an ordered trait continuum. Direct comparisons between test taker abilities and item difficulties, as mentioned, may then be conducted, with results able to 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. Anchor items are a number of items that are common to both tests; they are invaluable aids for comparing students on different tests. Once a test, or scale, has been calibrated (Coniam et al., 2021), the established values can be used to equate different test forms.

To achieve meaningful test anchoring, it is important to consider a fundamental tenet: that the starting point of a Rasch calibration is the mid-point of the calibration. This is the estimation of the point in a test at which a test taker has a 50/50 chance of answering the item/s correctly. A test, if specified to measure at a particular level of ability, should have the mid-point of the item distribution of the test in question anchored at a position in a scale representing that level of ability.

There are a number of key analytics usually conducted when doing Rasch measurement – and which have been reported on in previous LanguageCert studies (see e.g., Coniam et al., 2021a; 2021b). At the forefront, is the 'fit' of the data to the Rasch model, referring, in essence, to how well obtained values match expected values. Fit itself is 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 (Bond et al., 2020). Key statistics usually reported on are item infit and outfit mean squares and reliability.

Externally-Referenced Anchoring, CEFR levels and test forms

The methodology used in the current study is based on, as mentioned, externally-referenced anchoring (ERA) (Lee et al., 2022). In ERA, test forms which have no common items but comprise items which have been set at predefined and well-accepted CEFR levels are anchored using the calibrated midpoints of a test form against the LID scale and against the CEFR. For each test level, the frame of reference (see Humphry, 2006) constitutes the respective CEFR scale locations calibrated through the test forms and items for that level.

Table 1 below first provides detail on the number of test forms and their candidatures analysed.

Table 1: LST test forms and candidatures

CEFR level	Test forms	Candidates
B1	9	10,808
B2	6	2,732
C1	6	581

The focus in the current study is B1 to C1. Due to a comparatively small candidature, the C2 test forms do not form part of the current analysis.

The analysis in the study examines nine test forms at LST B1 level, six at B2 and six at C1. There are, as mentioned, for reasons of security, no linking items or test takers by which cross-calibrating may be conducted within or across test forms or levels. In the current study, ERA uses the calibrated midpoints of B1–C1 on the LID scale to explore the anchoring of these LST levels on the LID scale, and against CEFR levels. LID scale ranges and midpoints for the three CEFR levels explored are presented in Table 2.

Table 2: LID scale

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

On the basis of vertical midpoint anchoring, ERA:

- enables an effective calibration of the items in each test form – given that no other restrictions are imposed on the items.
- reveals the items' goodness of fit between expertly-assigned values and calibrated item distributions.

The anchoring goodness of fit is then evaluated by two metrics:

1) The extent to which a test's midpoint corresponds to the LID scale level.
2) The fit in terms of the extent to which the item distribution around a test's midpoint includes most of the items in a given test. Such fit is determined by a broadly bell-shaped distribution of item measures with the majority of item measures being clustered around the mean and falling between the 25th to 75th percentiles.

Research Questions

The research questions pursued in the current study may be summarised thus:

1. Do good Rasch infit and outfit statistics emerge from the externally-referenced anchoring of the LST B1–C1 test forms?
2. Do broadly bell-shaped item measure distributions emerge on the LST B1–C1 test forms?

Background Statistical Analysis

Item Infit and Outfit

Analysis in the current study has been conducted via the Rasch analysis software Winsteps (Linacre, 2018). Appendices 1, 2 and 3 provide details of fit statistics. The majority of the items in all LST B1–C1 test forms had infit and outfit fit statistics within the acceptable fit range of 0.7-1.3, indicating good fit to the Rasch model. Misfit, where it occurred, was only in a small percentage of items, less than 5% of the items on any one test.

Reliability

Test reliability, for a 50-item test, is proposed as being 0.7 or above (Ebel, 1965). The equivalent of classical test reliability in Rasch is person reliability (Anselmi et al., 2019). As Appendices 1–3 illustrate, 0.8 or better was achieved by all LST B1–C1 test forms. This indicates that satisfactory test reliability has occurred in the data available for this study.

These two sets of background statistics are indicative of a set of robust, well-constructed tests. This means that the picture of test robustness confirms that the externally-referenced anchoring is being conducted against a backdrop of reliable tests.

Externally-referenced Anchoring Results

Test means and measures that emerged after externally-referenced anchoring are now examined, in particular means recorded at the 25th and 75th percentiles. Ideally, the 25th percentile will be located half a logit (10 LID scale points) below and the 75th percentile half a logit above the test midpoint (Lee et al., 2022).

Two sets of linked analyses are presented below. The first set provides a summary of percentile distribution values; the second provides a more visual impression in the form of item difficulty distribution graphs.

Percentile Distribution Values

Summary analyses of the LST B1–C1 test forms in table form are presented in Tables 3–5 below. Acceptable values are in green font; values which are greater than five LID scale points (a quarter of a logit) away from the established range are in red font.

Table 3 provides the relevant detail for the B1 level test forms.

Table 3: Percentile distributions in LST B1 test forms
(LID scale range: 91-110; midpoint: 100)

	T206	T207	T208	T209	T384	T409	T414	T446	T593
Mean	100.00	100.01	100.00	100.00	99.99	100.00	100.00	100.00	100.00
SD	20.72	19.95	20.14	19.59	20.57	25.26	24.64	20.88	21.03
Maximum	159.34	145.40	139.98	141.43	150.09	157.75	175.02	138.25	158.75
75th percentile	117.08	111.89	116.59	113.48	116.51	115.78	118.29	115.14	112.91
50th percentile	98.92	101.33	100.66	99.60	97.07	103.78	97.17	97.71	100.54
25th percentile	87.72	90.97	83.65	85.17	86.95	82.36	82.51	86.32	85.99
Minimum	56.24	54.60	62.72	48.86	63.40	40.67	48.48	47.06	41.20

As can be seen, at the 25th percentile, all nine test forms are acceptably close to the lower scale range of 91. At the 75th percentile, there is some divergence, with six test forms showing a diverge of more than 5 LID scale points above the top of the LID scale range of 110 – in particular Tests T206 and T414. Nonetheless, the divergence seen is within half a logit (10 LID scale points) (Zwick et al., 1999), which means that the divergence is within acceptable bounds.

Table 4 provides the relevant detail for the B2 level test forms.

Table 4: Percentile distributions in LST B2 test forms
(LID scale range: 111-130; midpoint: 120)

	T211	T219	T220	T363	T385	T421
Mean	120.00	120.00	120.00	120.00	120.00	120.00
SD	23.13	23.60	20.91	19.94	20.21	17.53
Maximum	183.97	172.19	186.28	189.18	156.26	153.73
75th percentile	**134.75**	**134.11**	**130.88**	**131.22**	**138.34**	**132.54**
50th percentile	**118.92**	**120.46**	**117.59**	**118.83**	**120.15**	**117.87**
25th percentile	103.95	102.19	**109.34**	**107.21**	102.35	**107.80**
Minimum	84.77	69.00	82.48	78.75	80.70	84.38

At the 75th percentile, all six test forms are close to the upper scale range of 130. At the 25th percentile, there is more divergence, with three test forms showing a diverge of more than 5 LID scale points – in particular Tests T219 and T385. Such divergence is, however, within half a logit of difference, despite some items being slightly easier than intended in three of the tests.

Table 5 provides the detail on C1 level test forms.

Table 5: Percentile distributions in LST C1 test forms
(LID scale range: 131-150; midpoint: 140)

	T210	T222	T356	T364	T386	T588
Mean	140.00	140.00	140.00	140.00	140.00	140.00
SD	16.26	21.97	19.59	18.35	18.78	21.29
Maximum	175.56	196.41	190.32	179.01	186.88	190.73
75th percentile	**152.56**	**151.16**	**152.73**	**152.08**	**155.40**	**148.38**
50th percentile	**140.40**	**140.04**	**136.16**	**142.24**	**140.20**	**140.71**
25th percentile	**127.75**	**127.75**	**125.85**	**125.16**	**126.98**	**126.79**
Minimum	106.72	73.50	104.07	102.35	102.05	100.32

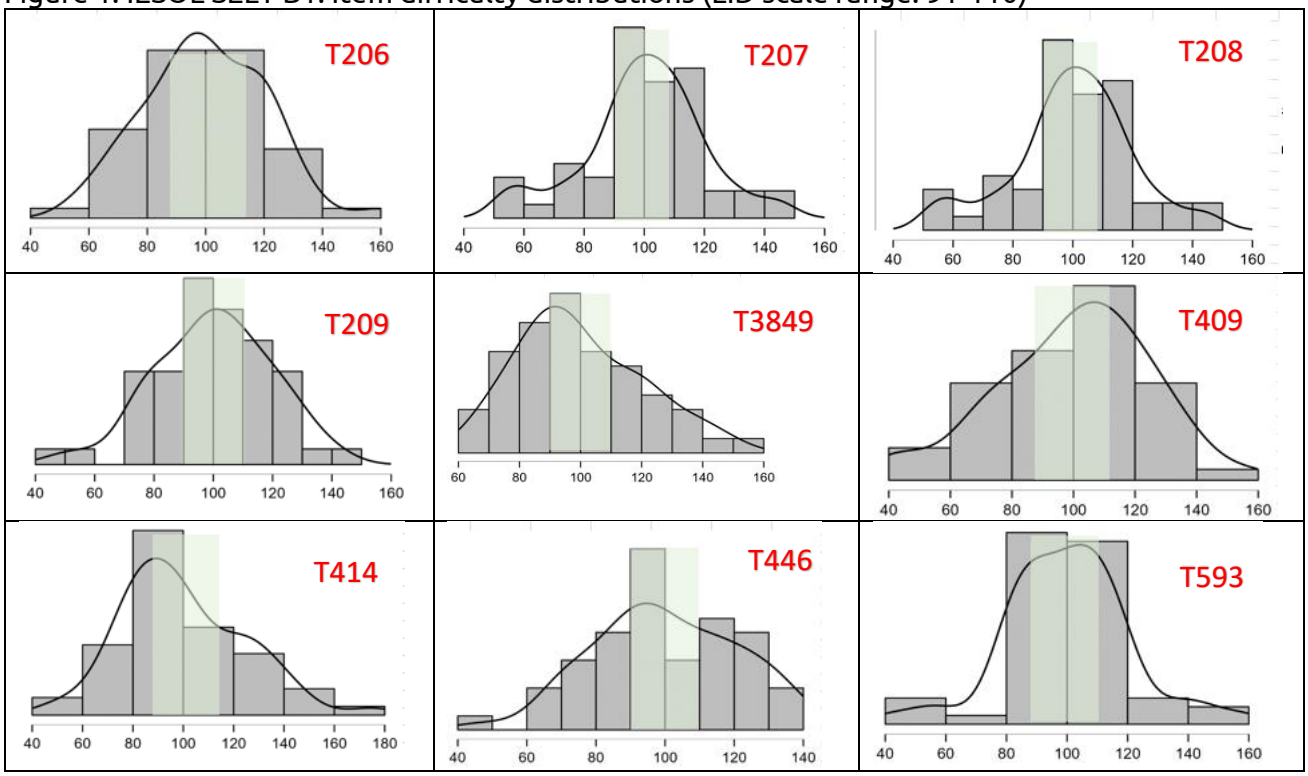
The C1 test forms show a close match with their LID scale ranges. At both 25th and 75th percentiles, all six test forms are close to the upper and lower scale ranges of 150 and 131. This means that all six tests have been well targeted at the C1 level.

Item Difficulty Distribution Graphs

To provide an accessible visual impression, item difficulty distributions are now presented in graph form in Figures 1–3. The green shading denotes the LID scale range for each test form. Frequency trend lines included across the scale for each test form provide a visual indication of the general shape of the distributions.

Figure 1 presents the item difficulty distributions for LST B1.

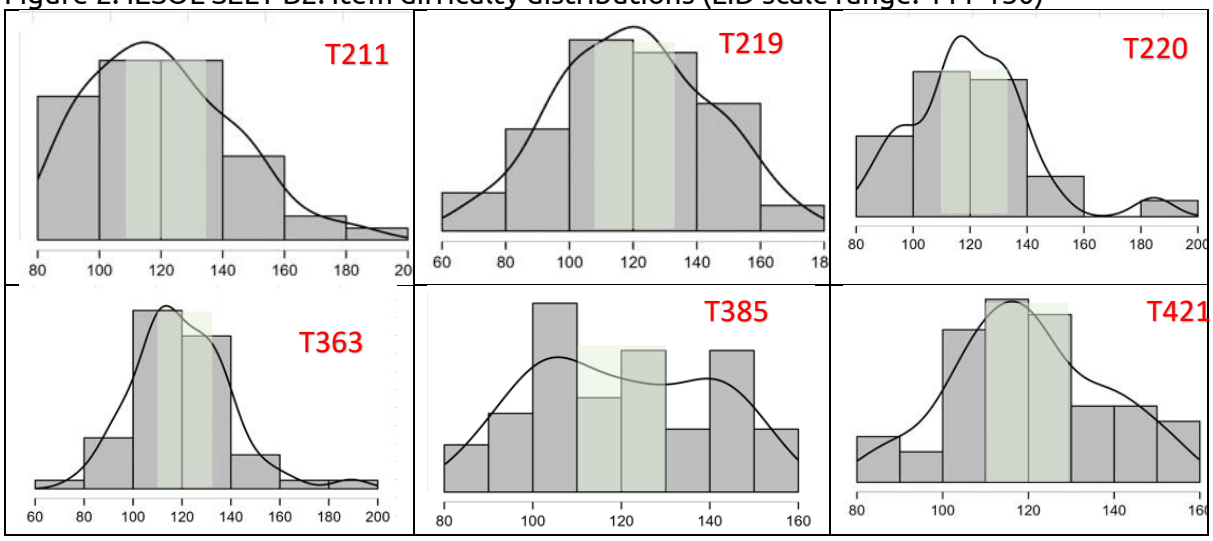
Figure 1: IESOL SELT B1: Item difficulty distributions (LID scale range: 91-110)



With the B1 test forms, there is a range of distributions. T414 is skewed slightly to the easy side; T446 has a comparatively wide distribution; T593 bulges around the midpoint. Nonetheless, in general, the green zones (the LID scale range) in the centre of the item distributions include a substantial number of the items in the B1 test forms. While not uniformly bell-shaped, the frequency trend lines do nonetheless indicate a regularity of shape.

Figure 2 presents the item difficulty distributions for LST B2.

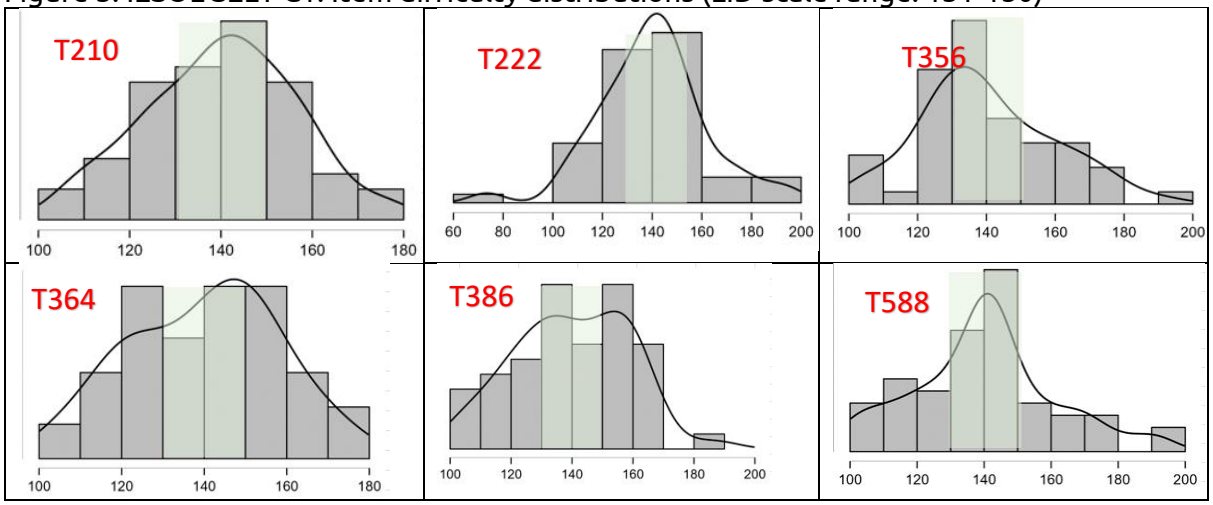
Figure 2: IESOL SELT B2: Item difficulty distributions (LID scale range: 111-130)



With the B2 test forms, distributions again show some divergence in their patterning. T211 is skewed slightly to the easy side; T220 has some outlying difficult items at the top end; T385 has a fairly flat distribution. Nonetheless, in general, the green zones (the supposed LID scale range) in the centre of the item distributions include a substantial number of the items in the B2 test forms. The frequency trend lines indicate a general regularity of shape, however, in general approaching a bell shape.

Figure 3 presents the item difficulty distributions for LST C1.

Figure 3: IESOL SELT C1: Item difficulty distributions (LID scale range: 131-150)



The C1 test form item distributions can be seen to be slightly more regular and bell-shaped than those for B2. T386 and T588 have some outlying difficult items at the top end of the scale, but the LID scale range (the green zones) again occupy a key section of the curve. The frequency trend lines again indicate a regularity of shape, approaching a bell shape.

In summary then, it can be seen that the expert-set items for the LST B1–C1 test forms match well with calibrated LID scale CEFR levels. This lends support to the claim that the LST B1–C1 test forms may be seen to be acceptably anchored on the LID scale.

Conclusion

This paper has reported on the externally-referenced anchoring of LanguageCert SELT tests (LST) at levels B1–C1. The study was pursuing two related research questions.

The first research question explored the extent to which good Rasch infit and outfit statistics would emerge from the externally-referenced anchoring of B1–C1 test forms. As has been described, the majority of B1, B2 and C1 test forms exhibited good Rasch infit and outfit statistics. This may be interpreted as a baseline of test quality.

The second research question explored the extent to which broadly bell-shaped item measure distributions would emerge from the analysis. The analyses generally exhibited a good match between CEFR levels B1–C1 and LID scale levels. Items on all test forms showed generally balanced distributions, with the majority of items in the majority test forms falling within the 25th to 75th percentiles -- the percentiles point which broadly match the upper and lower end of the cut scores determined for respective B1–C1 CEFR levels.

The match in the current study between the externally-referenced LST B1–C1 anchored levels and LID scale CEFR B1–C1 levels supports the argument that LanguageCert LST B1–C1 tests have been well set, with the results of the study statistically verifying expert judgements. The fact that the majority of items on the B1–C1 test forms fell within the 25th to 75th percentiles confirms the claim that LST B1–C1 tests are well targeted at the appropriate CEFR levels.

The test forms and items have been shown to be located acceptably on the LID scale – and against CEFR levels. Against this backdrop, vertical anchoring can now be brought to bear to place composite tests for each CEFR level on to the LID and hence LanguageCert Global scales. This research will be reported upon in a subsequent paper.

Notes

1. The **LanguageCert System** reports scores on the LanguageCert Global Scale of 0-100 that is derived directly from the 180-point LID scale. It provides candidates, employers, education institutions and government agencies an easy-to-understand results system. It applies across all the tests in the LanguageCert System. The Global Scale defines specific levels of attainment needed to fulfil certain requirements. For example, entrance into a university or for migration and employment purposes. The levels of attainment can relate to overall performance in an examination, performance by skill (e.g., speaking), or both these parameters.

References

Alderson, J. C., Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, 30, 535–556.

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

Bachman, L. F., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language*. Cambridge: Cambridge University Press.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge. https://doi.org/10.4324/9780429030499.

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

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

Coniam, D., Zhao, W., Lee, T., Milanovic, M., & Pike, N. (2022). The role of expert judgement in language test validation. *Language Education & Assessment*.

Ebel, R. L. (1965). *Measuring educational achievement*. Prentice-Hall, NJ: Englewood Cliffs.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings* (2nd ed.). New York, NY: Springer Science & Business Media.

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28(1), 77–104.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

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

Lee, T., Milanovic, M., & Pike, N. (2022). Equating Rasch values and expert judgement through externally-referenced anchoring. *International Journal of TESOL Studies*, 4(1), 187-202. doi.org/10.46451/ijts.2022.01.12.

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

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10(3), 211–234.

Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LanguageCert Writing Tests. International Journal of TESOL Studies, 4(1), 203-212. doi.org/10.46451/ijts.2022.01.13.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435–464.

van Steensel, R., Oostdam, R., & van Gelderen, A. (2013). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing*, 30(1), 3–21.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT reasoning test. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Zwick, Rebecca, Dorothy T. Thayer & Charles Lewis. 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement,* 36(1). 1-28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x.

Appendix 1: LST B1: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary
T206	SELT B1 T206 | PERSON 10810 INPUT 1314 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 44.0 51.7 154.41 13.89 .99 .1 .95 .1| | P.SD 8.9 1.9 33.64 8.86 .14 .6 .57 .8| | REAL RMSE 16.47 TRUE SD 29.33 SEPARATION 1.78 PERSON RELIABILITY .76| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1112.2 1307.6 100.00 2.08 .99 .0 .94 -.4| | P.SD 123.7 5.6 20.52 .53 .14 2.8 .43 3.1| | REAL RMSE 2.15 TRUE SD 20.41 SEPARATION 9.51 ITEM RELIABILITY .99|
T207	SELT B1 T207 | PERSON 10810 INPUT 1295 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 40.8 52.0 141.01 11.14 1.00 .1 .97 .1| | P.SD 10.5 .0 33.05 7.00 .11 .7 .41 .8| | REAL RMSE 13.16 TRUE SD 30.31 SEPARATION 2.30 PERSON RELIABILITY .84| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1014.8 1295.0 100.00 1.79 .99 -.1 .97 -.3| | P.SD 140.8 .0 19.76 .40 .14 3.4 .39 3.4| | REAL RMSE 1.83 TRUE SD 19.68 SEPARATION 10.74 ITEM RELIABILITY .99|
T208	SELT B1 T208 | PERSON 10810 INPUT 1384 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 41.0 51.7 141.70 10.99 1.00 .1 .96 .1| | P.SD 10.2 2.0 31.95 6.47 .11 .6 .48 .8| | REAL RMSE 12.75 TRUE SD 29.29 SEPARATION 2.30 PERSON RELIABILITY .84| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1091.4 1375.1 100.00 1.76 .99 -.2 .96 -.4| | P.SD 152.6 8.5 19.95 .36 .15 3.3 .37 3.4| | REAL RMSE 1.79 TRUE SD 19.86 SEPARATION 11.08 ITEM RELIABILITY .99|
T209	SELT B1 T209 | PERSON 10810 INPUT 1411 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 42.3 51.8 146.37 12.11 1.00 .2 .93 .1| | P.SD 9.7 1.5 32.77 7.78 .10 .6 .46 .7| | REAL RMSE 14.40 TRUE SD 29.44 SEPARATION 2.05 PERSON RELIABILITY .81| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1146.5 1404.9 100.00 1.82 .99 .1 .93 -.4| | P.SD 138.9 6.3 19.40 .43 .13 3.1 .33 3.1| | REAL RMSE 1.87 TRUE SD 19.31 SEPARATION 10.34 ITEM RELIABILITY .99|
T384	SELT B1 T384 | PERSON 10810 INPUT 1365 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 41.5 51.8 143.43 11.20 .99 .1 .95 .1| | P.SD 9.8 1.6 31.91 6.48 .14 .7 .57 .9| | REAL RMSE 12.94 TRUE SD 29.17 SEPARATION 2.25 PERSON RELIABILITY .84| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1090.0 1358.8 100.00 1.79 .99 .1 .95 -.3| | P.SD 162.8 6.5 20.37 .33 .13 3.2 .36 3.3| | REAL RMSE 1.82 TRUE SD 20.28 SEPARATION 11.12 ITEM RELIABILITY .99|
T409	SELT B1 T409 | PERSON 10810 INPUT 1344 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 43.3 51.7 151.86 12.38 1.00 .2 .92 .1| | P.SD 8.8 2.5 31.74 7.22 .13 .6 .70 .7| | REAL RMSE 14.33 TRUE SD 28.32 SEPARATION 1.98 PERSON RELIABILITY .80| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1120.1 1335.4 100.00 2.14 .99 .1 .92 -.6| | P.SD 146.9 6.8 25.02 .87 .14 3.1 .35 3.1| | REAL RMSE 2.31 TRUE SD 24.91 SEPARATION 10.77 ITEM RELIABILITY .99|

T414

SELT B1 T414

PERSON	10810 INPUT		1401 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	41.8	51.8	145.66	11.19	.97	.1	1.00	.2
P.SD	9.3	1.2	31.86	5.60	.19	.7	.85	1.1
REAL RMSE	12.52	TRUE SD	29.30	SEPARATION	2.34	PERSON RELIABILITY		.85
ITEM	52 INPUT		52 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	1126.1	1396.3	100.00	1.86	.99	-.1	1.00	-.5
P.SD	196.0	5.5	24.41	.47	.14	3.1	.62	3.2
REAL RMSE	1.92	TRUE SD	24.33	SEPARATION	12.67	ITEM RELIABILITY		.99

T446

SELT B1 T446

PERSON	10810 INPUT		655 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	41.0	51.6	141.82	11.18	1.00	.1	.94	.1
P.SD	9.8	2.7	32.46	7.13	.12	.7	.49	.8
REAL RMSE	13.26	TRUE SD	29.63	SEPARATION	2.23	PERSON RELIABILITY		.83
ITEM	52 INPUT		52 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	516.1	650.3	100.00	2.56	.99	.1	.94	-.2
P.SD	76.6	4.2	20.68	.62	.13	2.4	.34	2.5
REAL RMSE	2.63	TRUE SD	20.51	SEPARATION	7.80	ITEM RELIABILITY		.98

T593

SELT B1 T593

PERSON	10810 INPUT		641 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	41.7	51.7	145.88	12.24	.99	.1	.96	.1
P.SD	10.0	2.2	34.51	7.93	.14	.7	.63	.8
REAL RMSE	14.58	TRUE SD	31.27	SEPARATION	2.14	PERSON RELIABILITY		.82
ITEM	52 INPUT		52 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	514.6	637.2	100.00	2.69	.99	.0	.96	-.2
P.SD	71.6	3.6	20.83	.71	.15	2.5	.43	2.6
REAL RMSE	2.78	TRUE SD	20.64	SEPARATION	7.43	ITEM RELIABILITY		.98

Appendix 2: LST B2: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary
T211	SELT B2 211 | PERSON 2732 INPUT 528 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 31.0 51.9 132.15 7.56 1.00 .0 1.00 .0| | P.SD 10.2 .7 25.77 2.64 .15 .9 .41 1.0| | REAL RMSE 8.00 TRUE SD 24.50 SEPARATION 3.06 PERSON RELIABILITY .90| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 314.5 527.1 120.00 2.26 1.00 -.1 1.00 .0| | P.SD 101.5 1.2 22.91 .25 .14 2.8 .28 2.6| | REAL RMSE 2.27 TRUE SD 22.79 SEPARATION 10.05 ITEM RELIABILITY .99|
T219	SELT B2 219 | PERSON 2732 INPUT 569 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 31.7 51.8 135.44 8.07 1.00 .0 1.00 .0| | P.SD 11.4 1.5 29.79 3.63 .15 .9 .40 .9| | REAL RMSE 8.85 TRUE SD 28.45 SEPARATION 3.21 PERSON RELIABILITY .91| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 347.2 567.0 120.00 2.25 .99 -.1 1.00 -.1| | P.SD 103.4 2.1 23.37 .30 .15 2.8 .38 2.8| | REAL RMSE 2.27 TRUE SD 23.26 SEPARATION 10.24 ITEM RELIABILITY .99|
T220	SELT B2 220 | PERSON 2732 INPUT 547 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 33.5 51.8 138.19 8.07 1.00 .1 .98 .0| | P.SD 11.1 2.2 28.30 3.41 .14 .8 .32 .9| | REAL RMSE 8.76 TRUE SD 26.91 SEPARATION 3.07 PERSON RELIABILITY .90| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 352.2 544.8 120.00 2.27 1.00 -.1 .98 -.1| | P.SD 87.5 1.8 20.71 .26 .14 2.7 .28 2.6| | REAL RMSE 2.28 TRUE SD 20.58 SEPARATION 9.02 ITEM RELIABILITY .99|
T363	SELT B2 363 | PERSON 2732 INPUT 573 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 37.7 51.8 149.97 9.42 1.00 .1 .97 .0| | P.SD 10.6 1.9 30.75 5.38 .15 .7 .36 .8| | REAL RMSE 10.85 TRUE SD 28.78 SEPARATION 2.65 PERSON RELIABILITY .88| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 415.4 571.1 120.00 2.40 .99 -.1 .96 -.2| | P.SD 80.1 1.7 19.74 .34 .15 2.7 .29 2.4| | REAL RMSE 2.42 TRUE SD 19.60 SEPARATION 8.09 ITEM RELIABILITY .98|
T385	SELT B2 385 | PERSON 2732 INPUT 280 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 35.4 51.9 144.96 8.86 1.00 .1 1.00 .1| | P.SD 10.9 .9 31.00 4.99 .12 .8 .44 1.0| | REAL RMSE 10.17 TRUE SD 29.29 SEPARATION 2.88 PERSON RELIABILITY .89| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 190.8 279.5 120.00 3.29 1.00 -.1 1.00 .0| | P.SD 42.6 .8 20.02 .42 .15 2.1 .39 2.1| | REAL RMSE 3.32 TRUE SD 19.74 SEPARATION 5.95 ITEM RELIABILITY .97|
T421	SELT B2 421 | PERSON 2732 INPUT 235 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 28.9 51.7 126.60 7.12 1.00 .0 1.01 .0| | P.SD 10.2 3.2 23.14 3.13 .10 .8 .23 .9| | REAL RMSE 7.78 TRUE SD 21.79 SEPARATION 2.80 PERSON RELIABILITY .89| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 130.5 233.4 120.00 3.17 1.00 -.1 1.01 .0| | P.SD 37.7 1.1 17.36 .26 .14 2.3 .22 2.1| | REAL RMSE 3.18 TRUE SD 17.06 SEPARATION 5.36 ITEM RELIABILITY .97|

Appendix 3: LST C1: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary
T210	SELT C1 T210 PERSON 581 INPUT 135 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 30.6 51.9 150.01 6.85 1.00 .0 1.00 .0 P.SD 10.2 .6 21.06 1.08 .11 .9 .21 .9 REAL RMSE 6.94 TRUE SD 19.88 SEPARATION 2.87 PERSON RELIABILITY .89 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 79.6 134.7 140.00 4.22 1.00 -.1 1.00 .0 P.SD 19.8 .5 16.10 .34 .17 2.0 .25 1.9 REAL RMSE 4.23 TRUE SD 15.53 SEPARATION 3.67 ITEM RELIABILITY .93
T222	SELT C1 T222 PERSON 581 INPUT 100 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 33.4 52.0 158.73 8.18 1.01 .0 1.01 .0 P.SD 11.2 .0 29.10 3.82 .14 .8 .31 .8 REAL RMSE 9.03 TRUE SD 27.66 SEPARATION 3.06 PERSON RELIABILITY .90 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 64.2 100.0 140.00 5.40 .99 .0 1.01 .1 P.SD 16.3 .0 21.76 1.06 .15 1.3 .39 1.3 REAL RMSE 5.50 TRUE SD 21.05 SEPARATION 3.83 ITEM RELIABILITY .94
T356	SELT C1 T356 PERSON 581 INPUT 115 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 36.1 52.0 163.88 8.06 1.00 .1 .99 .0 P.SD 9.3 .0 24.82 3.49 .12 .8 .28 .8 REAL RMSE 8.78 TRUE SD 23.22 SEPARATION 2.65 PERSON RELIABILITY .87 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 79.9 115.0 140.00 5.03 1.00 .0 .99 .0 P.SD 18.1 .0 19.40 .74 .15 1.4 .33 1.4 REAL RMSE 5.08 TRUE SD 18.73 SEPARATION 3.68 ITEM RELIABILITY .93
T364	SELT C1 T364 PERSON 581 INPUT 120 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 32.2 52.0 155.07 7.54 1.00 .0 1.08 .1 P.SD 10.9 .0 26.02 2.59 .11 .8 .55 1.0 REAL RMSE 7.97 TRUE SD 24.76 SEPARATION 3.11 PERSON RELIABILITY .91 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 74.3 120.0 140.00 4.66 .99 -.1 1.08 .1 P.SD 18.4 .0 18.18 .48 .15 1.5 .54 1.5 REAL RMSE 4.68 TRUE SD 17.56 SEPARATION 3.75 ITEM RELIABILITY .93
T386	SELT C1 T386 PERSON 581 INPUT 55 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 29.2 51.9 147.57 6.90 1.00 .0 1.01 .0 P.SD 9.7 .8 21.35 1.39 .12 .8 .26 .9 REAL RMSE 7.04 TRUE SD 20.15 SEPARATION 2.86 PERSON RELIABILITY .89 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 30.9 54.9 140.00 6.67 .99 -.1 1.01 .0 P.SD 9.4 .3 18.59 .72 .22 1.5 .40 1.5 REAL RMSE 6.71 TRUE SD 17.34 SEPARATION 2.58 ITEM RELIABILITY .87
T588	SELT C1 T588 PERSON 581 INPUT 56 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 30.6 52.0 150.40 7.02 .99 .0 1.03 .1 P.SD 9.7 .2 21.46 1.12 .12 .8 .31 1.0 REAL RMSE 7.11 TRUE SD 20.25 SEPARATION 2.85 PERSON RELIABILITY .89 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 32.9 56.0 140.00 6.77 1.00 -.1 1.03 .0 P.SD 10.2 .2 21.08 .87 .16 1.2 .34 1.2 REAL RMSE 6.82 TRUE SD 19.95 SEPARATION 2.92 ITEM RELIABILITY .90

Language Cert
languagecert.org