LanguageCert

Aligning LanguageCert SELT Tests to the LanguageCert Item Difficulty (LID) Scale

Tony Lee
Yiannis Papargyris
Michael Milanovic
Nigel Pike
and
David Coniam

Abstract

This paper reports on the alignment of LanguageCert SELT tests to the LanguageCert Item Difficulty (LID) Scale. The paper builds on a previous study which established that the LanguageCert SELT B1–C1 tests are robust though the use of externally-referenced anchoring.

The paper explores the alignment of LanguageCert SELT tests in relation to the two objectively marked components of Listening and Reading. The use of externally-referenced anchoring enabled the robustness of the four CEFR test levels B1–C2 to be demonstrated.

As the paper illustrates, the LanguageCert SELT tests in general assess at their designated CEFR level but also contain items which allow them to assess across levels. At the C1 level, there are items which assess above C1 and, at the other end, below C1. Likewise, at the B2 level, there are items which assess both above and below B2.

Introduction

LanguageCert has been an approved provider, delivering Secure English Language Tests (SELT) tests to the UK Home Office for UK visas & immigration purposes, for movement and work to the UK, since 2020.

LanguageCert SELT Test (LST) four-skills tests are offered at a range of levels (B1 to C2), mapped to the Common European Framework of Reference (CEFR). The previous study (Milanovic et al., 2022) illustrated how LanguageCert calibrates test material and aligns test forms to the respective CEFR levels. Building on the previous study, the current study demonstrates the alignment of all four LST levels (B1–C2) incorporating all B1 to C2 test forms produced since 2020.

The LST tests used in the current study constitute a number of the test forms for the respective CEFR levels delivered by LanguageCert in the 18-month period from mid 2020 to late 2021.

The LanguageCert SELT tests

The LanguageCert SELT Test (LST) suite of tests form an integral part of the LanguageCert System [Note 1]. The suite comprises four tests from B1 to C2, each aligned to its respective CEFR level as well as three 2-skill tests ranging from A1-B1. Examination specifications reflect the requirements of the CEFR; test materials writers represent the highest international standards and have extensive expertise in, and knowledge and understanding of, the CEFR, the latter being crucial in ensuring validity and reliability (Hughes, 2003). Test items are linked to the CEFR by expert judgement, a methodology which has been shown to be robust (Coniam et al., 2022).

The B1-C1 tests comprise 52 items: 26 Listening and 26 Reading items; the C2 tests comprise 56 items: 30 Listening and 26 Reading items. In line with the key test qualities of validity and reliability (Bachman & Palmer, 2010), the LST tests assess the communicative skills that test takers will be expected to control at particular levels of ability. Test content matches target test takers – in terms of grammar, functions, vocabulary, topics etc., and the tasks have correspondingly relevant ‘communicative’ contexts.

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. 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 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.

Test data

Table 1 below provides detail on the number of test forms at each level and candidates.

Table 1: SELT IESOL test forms and candidatures

CEFR level	Test forms	Candidates
C2	3	111
C1	6	581
B2	6	2,732
B1	9	10,808

Via externally-referenced, or vertical, anchoring (see detail below), test forms are anchored at the midpoint of the item distribution of a given scale. The C2 sample is small, as can be seen from Table 1. As Lee et al. (2022) illustrate, externally-referenced anchoring is nonetheless a methodology that works even with small samples. On this basis, C2 is included in the current analysis.

The midpoints of the LID scale for the six CEFR levels are presented in Table 2. In line with the LanguageCert Global Scale, Table 2 includes correspondences between the LID scale and the Global Scale.

Table 2: LID scale

CEFR level	LID scale range	LID scale midpoint	Global scale range	Global scale midpoint
C2	151-170	160	90-100	95
C1	131-150	140	75-89	82
B2	111-130	120	60-74	67
B1	91-110	100	40-59	50
A2	71-90	80	20-39	30
A1	51-70	60	10-19	15

Externally-Referenced Anchoring

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. 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 (Lee et al., 2022).

Research Questions

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

> **Can the four SELT tests (B1-C2) be accurately placed on the LID scale and hence against the CEFR?**

Background Statistical Analysis

Item Infit and Outfit

Accuracy mentioned in the research question above will be measured through good Rasch infit and outfit statistics emerging from the analysis at each of the four test levels. Analysis in the current study has been conducted via the Rasch analysis software Winsteps (Linacre, 2018). Appendix 1 provides detail on fit statistics. Most of items in tests at all four LanguageCert SELT Test levels had infit and outfit fit statistics within the acceptable fit range of 0.7-1.3, indicating good fit to the Rasch model.

Reliability

Test reliability, for a 50-item test, is proposed at 0.7 or above (Ebel, 1965). The equivalent of classical test reliability in Rasch is person reliability (Anselmi et al., 2019). As Appendix 1 illustrates, 0.8 or better was achieved on all four levels of test.

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

Externally-referenced Anchoring Results

Test means and measures that emerged after the introduction of externally-referenced anchoring are now examined, in particular means recorded at the 25th, 50th and 75th percentiles. As mentioned, the 25th percentile will ideally 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).

Summary analyses of the LST B1–C2 test forms are presented 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.

Two sets of linked analyses for the composite LST tests 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.

Table 3 provides the relevant detail for the composite LST tests. Each level has two sets of entries: the LID scale level range (in blue font) to the left-hand side and the distributions which emerged (in green font) to the right-hand side.

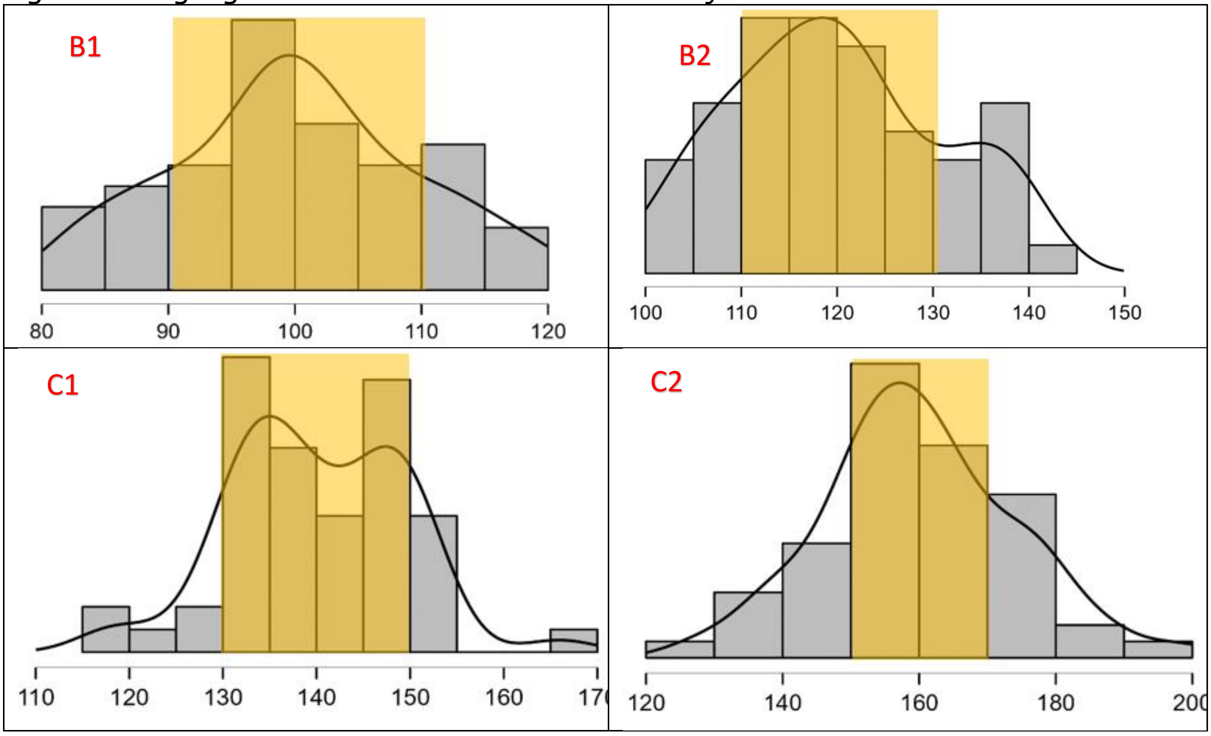
Table 3: Percentile distributions in composite LanguageCert SELT Test tests

	B1		B2		C1		C2	
No. of items		52		52		52		56
Mean		100		120.00		140.00		160
SD		9.59		10.83		9.28		14.09
Maximum		119.55		141.02		165.98		198.53
75th percentile	**110**	**105.64**	**130**	**126.43**	**150**	**147.69**	**170**	**167.96**
50th percentile		**99.45**		**119.29**		**139.50**		**159.15**
25th percentile	**91**	**94.04**	**111**	**112.78**	**131**	**133.45**	**151**	**150.72**
Minimum		82.05		100.28		117.51		127.34

As can be seen, at the 25th percentile, all test levels are acceptably close to the lower LID scale range. Similarly, at the 75th percentile, all test levels are acceptably close to the upper LID scale range. There is a degree of divergence, although this is within the accepted half a logit (10 LID scale points) of difference (Zwick et al., 1999) which means that tests have been generally well targetted at their intended level.

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

Figure 1: LanguageCert SELT Test tests: Test difficulty distributions



As can be seen, each level shows a broadly bell-shaped distribution, as confirmed by the best fit lines that wrap around the columns. The distributions are not perfect – C1 shows a somewhat irregular pattern in the centre of the graph. In general, however, the distributions are comparatively regular, indicating that the tests are performing as expected.

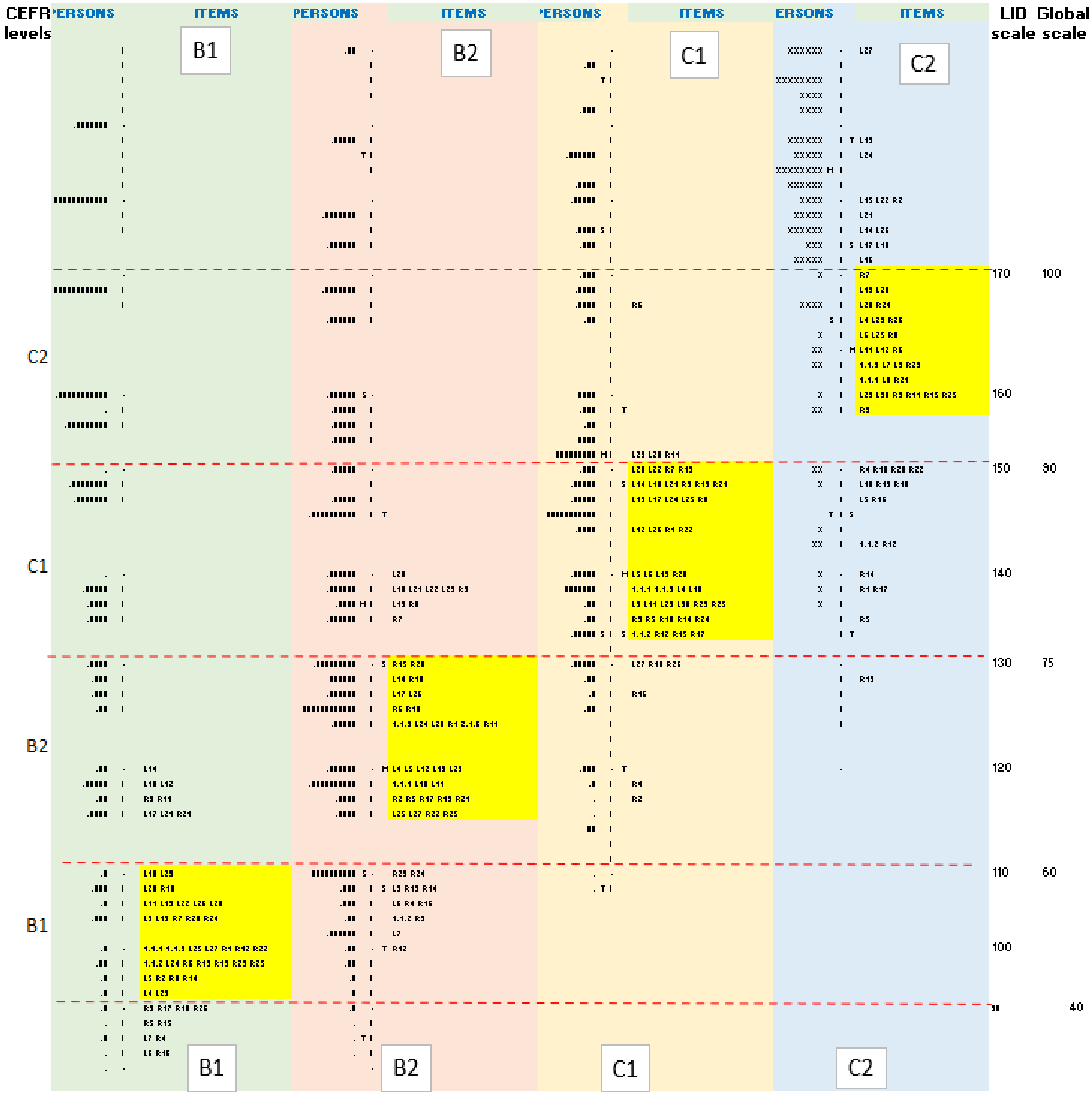
Placing LanguageCert SELT on the LID Scale

It has been established that the test forms have been well set and are robust in terms of fit statistics and reliability. The tests are located at appropriate points across the ranges of the LID scale, and hence at appropriate points against the CEFR.

Figure 2 below presents the Rasch person and item distributions on the LID and Global scales. The B1 test is green; the B2 salmon; the C1 beige; the C2 blue. LID scale values are to the right-hand side of the maps; CEFR levels to the left-hand side. The red tram lines indicate the LID scale cuts for each level. The highlighted yellow sections are the CEFR / test item match.

The maps should be read such that candidates (persons) are located to the left-hand side of a particular map, items to the right-hand side. More able candidates are situated towards the upper left end of the map, and less able candidates towards the lower left end. More demanding items are situated towards the upper right end of the map while easier items are situated towards the lower right end.

Figure 3: LanguageCert SELT Test Common Scale



As can be seen from Figure 3, for each LST test, the majority of the items (the highlighted yellow sections) fall within the CEFR level for which they are intended. This is an indicator of validity, indicating that the LST tests are generally well set, and are being targetted at the appropriate level.

It is also clear from Figure 3 that while tests assess in general at a particular CEFR level, the tests also assess across levels. Taking the beige C1 test as an example and reading up from the bottom of the C1 row, it can be seen that the bulk of the items assess at C1 level, as intended. There are, however, a number of items which assess at B2 below C1 and another set which assess at C2 above C1.

Likewise, with the salmon B2 test, the majority of items assess at B2 level, but substantial numbers assess at B1 and at C1 levels. This is the value and utility of a common scale: the reach across levels. While tests in principle assess at a given level, with appropriate calibration, tests can also be used across levels.

Conclusion

This paper has explored the alignment of LanguageCert SELT tests to the LID Scale. The use of externally-referenced anchoring has enabled the robustness of the four CEFR test levels B1–C2 to be demonstrated.

As the Rasch item/person maps illustrate, while the LST tests principally assess at their designated CEFR level, tests also contain items which assess across levels. At the C1 level, there are items which assess above and below C1. Likewise, at the B2 level, there are items which assess both above and below B2.

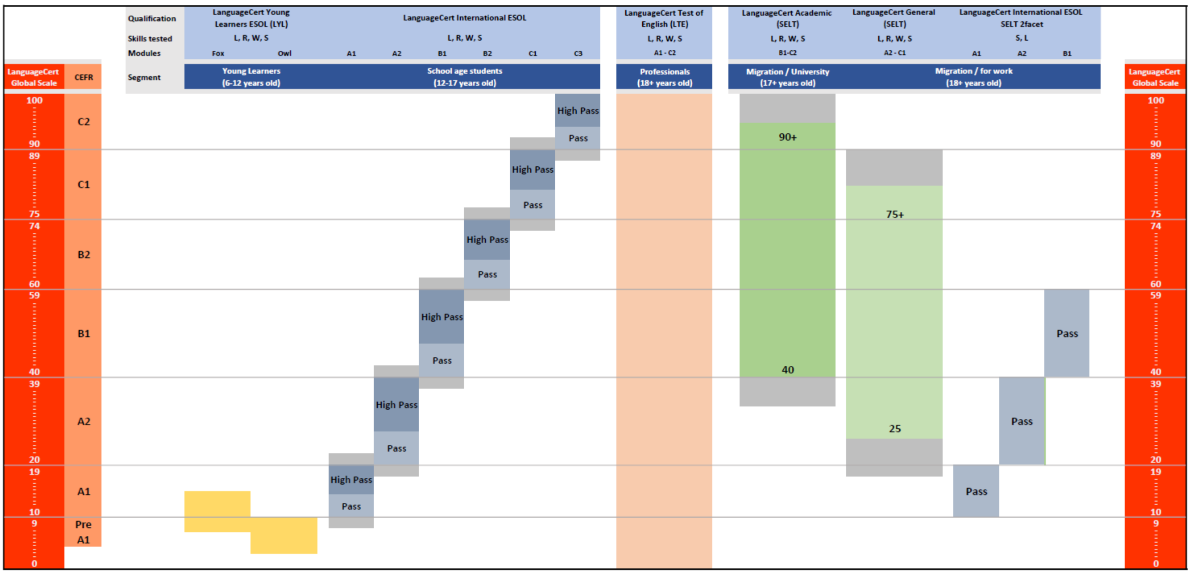
The research question pursued in the study was that LanguageCert SELT tests could be accurately placed on the LID scale and hence the CEFR, accuracy being defined as good Rasch infit and outfit statistics being obtained in the analysis at each of the four test levels. Rasch levels were indeed within acceptable levels, supporting the claim that the tests are accurately placed.

This exercise forms part of the overall research drive that is being undertaken at LanguageCert to locate its various test products on the LID and hence LanguageCert Global Scale. The extensive research and calibration undertaken with the LanguageCert Test of English (Coniam et al., 2021a; b) is now being extended to other LanguageCert products. The research conducted with the SELT tests in the current study forms part of that endeavour.

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 (see below). 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.

The LanguageCert Global Scale



References

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.

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). Validating 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.

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.

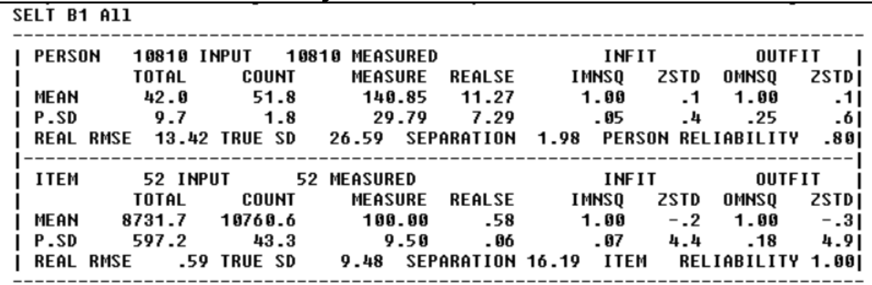
Linacre, J. M. (2018). *Winsteps Rasch measurement computer program user's guide*. Winsteps.com: Beaverton, OR.

Milanovic, M., Lee, T., Coniam, D., & Papargyris, Y. (2022). Externally-Referenced Anchoring of SELT tests. London: LanguageCert.

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.

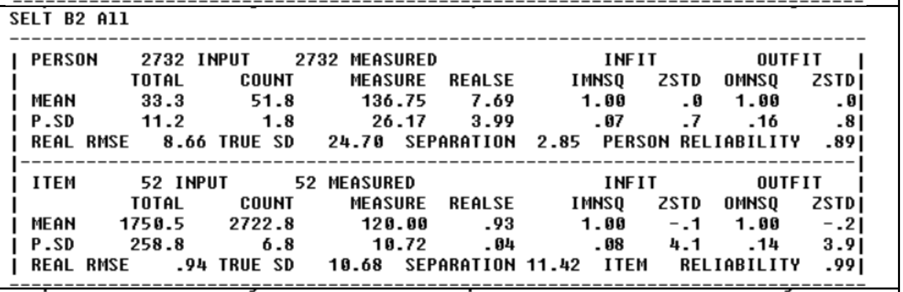
Appendix 1: LanguageCert SELT Test: Fit Statistics and Person Reliabilities

<table>
<tr><th>Test level</th><th>Rasch statistics summary</th></tr>
<tr><td>B1</td><td>



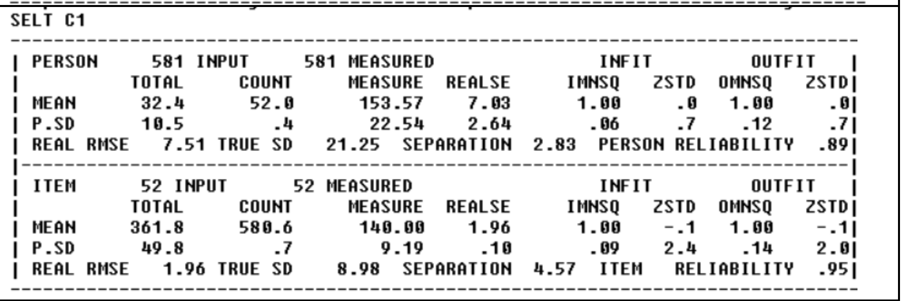
```
SELT B1 All
-------------------------------------------------------------------------------
| PERSON    10810 INPUT    10810 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN       42.0       51.8      140.85   11.27     1.00     .1   1.00     .1|
| P.SD        9.7        1.8       29.79    7.29      .05     .4    .25     .6|
| REAL RMSE  13.42 TRUE SD    26.59  SEPARATION  1.98  PERSON RELIABILITY  .80|
|------------------------------------------------------------------------------|
| ITEM        52 INPUT        52 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN     8731.7    10760.6      100.00     .58     1.00    -.2   1.00    -.3|
| P.SD      597.2       43.3        9.50     .06      .07    4.4    .18    4.9|
| REAL RMSE    .59 TRUE SD     9.48  SEPARATION 16.19  ITEM   RELIABILITY 1.00|
-------------------------------------------------------------------------------
```

</td></tr>
<tr><td>B2</td><td>



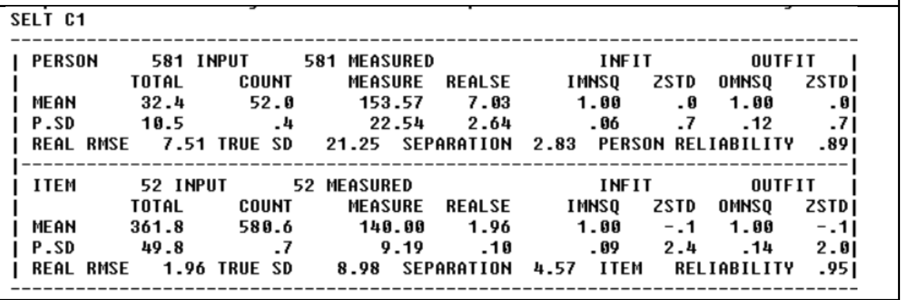
```
SELT B2 All
-------------------------------------------------------------------------------
| PERSON     2732 INPUT     2732 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN       33.3       51.8      136.75    7.69     1.00     .0   1.00     .0|
| P.SD       11.2        1.8       26.17    3.99      .07     .7    .16     .8|
| REAL RMSE   8.66 TRUE SD    24.70  SEPARATION  2.85  PERSON RELIABILITY  .89|
|------------------------------------------------------------------------------|
| ITEM        52 INPUT        52 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN     1750.5     2722.8      120.00     .93     1.00    -.1   1.00    -.2|
| P.SD      258.8        6.8       10.72     .04      .08    4.1    .14    3.9|
| REAL RMSE    .94 TRUE SD    10.68  SEPARATION 11.42  ITEM   RELIABILITY  .99|
-------------------------------------------------------------------------------
```

</td></tr>
<tr><td>C1</td><td>



```
SELT C1
-------------------------------------------------------------------------------
| PERSON      581 INPUT      581 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN       32.4       52.0      153.57    7.03     1.00     .0   1.00     .0|
| P.SD       10.5         .4       22.54    2.64      .06     .7    .12     .7|
| REAL RMSE   7.51 TRUE SD    21.25  SEPARATION  2.83  PERSON RELIABILITY  .89|
|------------------------------------------------------------------------------|
| ITEM        52 INPUT        52 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      361.8      580.6      140.00    1.96     1.00    -.1   1.00    -.1|
| P.SD       49.8         .7        9.19     .10      .09    2.4    .14    2.0|
| REAL RMSE   1.96 TRUE SD     8.98  SEPARATION  4.57  ITEM   RELIABILITY  .95|
-------------------------------------------------------------------------------
```

</td></tr>
<tr><td>C2</td><td>



```
SELT C1
-------------------------------------------------------------------------------
| PERSON      581 INPUT      581 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN       32.4       52.0      153.57    7.03     1.00     .0   1.00     .0|
| P.SD       10.5         .4       22.54    2.64      .06     .7    .12     .7|
| REAL RMSE   7.51 TRUE SD    21.25  SEPARATION  2.83  PERSON RELIABILITY  .89|
|------------------------------------------------------------------------------|
| ITEM        52 INPUT        52 MEASURED              INFIT          OUTFIT   |
|           TOTAL      COUNT      MEASURE  REALSE    IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      361.8      580.6      140.00    1.96     1.00    -.1   1.00    -.1|
| P.SD       49.8         .7        9.19     .10      .09    2.4    .14    2.0|
| REAL RMSE   1.96 TRUE SD     8.98  SEPARATION  4.57  ITEM   RELIABILITY  .95|
-------------------------------------------------------------------------------
```

</td></tr>
</table>

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

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

Language
Cert
languagecert.org