International Journal of TESOL Studies (2022)
Vol. 4 (1) 187-202 https://doi.org/10.46451/ijts.2022.01.12

Article

Equating Rasch Values and Expert Judgement Through Externally-Referenced Anchoring

Tony Lee*
LanguageCert, UK

Michael Milanovic
LanguageCert, UK

Nigel Pike
LanguageCert, UK

Received: 15 September 2021/Accepted: 31 January 2022/Published: 30 March 2022

Abstract

This paper reports on the use of externally-referenced anchoring by LanguageCert as a methodology for calibrating language test materials and aligning test forms. The datasets used are taken from tests at each of the six levels of LanguageCert IESOL suite, all of which have been aligned to the CEFR through expert judgement. We illustrate in this paper the extent to which externally-referenced anchoring, using Item Response Theory (IRT) but based on expert judgement, can be used as an effective, reliable and valid methodology. The approach is based on the premise that successful anchoring may be achieved by reference to well-targeted, expertly-written test forms aligned to the underlying traits of a particular CEFR level by expert judgement and verified through the use of IRT.

This study focuses on the analysis of 18 LanguageCert test forms, three at each CEFR level. The LanguageCert Item Difficulty (LID) scale, which underlies all LanguageCert test materials, is linked empirically to the CEFR, and each test was placed on the LID scale based at the midpoint of its distribution. This midpoint setting was then set as the externally-referenced anchor for a given CEFR level.

The findings of this study indicate that, while the match between the distribution of items in the selected LanguageCert IESOL tests and the LID scale was not perfect, in general, a relatively close match between the items in the tests and the LID scale was found and, as a consequence, the corresponding CEFR level. For each test, most of the items fell between the 25th and 75th percentile of any given level: this range representing the lower and upper bounds of LID scale values for each CEFR level. These results demonstrate that LanguageCert IESOL test items are well set and appropriately positioned at respective CEFR levels on the basis of expert judgement. The study illustrates that externally-referenced anchoring based on expert judgement may be used as a methodology for aligning test forms to an external frame of reference, in this case the CEFR.

*Corresponding author. Email: Tony.Lee@PeopleCert.org

Keywords

Externally-referenced anchoring, calibrating test materials, aligning test forms, Rasch

1 Expert Judgement and Test Setting

'Expert judgement' in language assessment is 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. Rodriguez (1997) refers to item writing as an art, while Bristol (2015) describes the creation of examination questions as both an art and a science. Haladyna and Downing (1989) provide a set of seven ground rules originally selected for good item setting, some of which are echoed in Alderson et al. (1995) where the qualities of an expert item writer are cogently discussed. What is clear however, is that training and experience are necessary characteristics of successful item writing. Coniam (1997) suggests that well-trained and competent item writers may be expected to achieve a 'quality setting' rate of around 70% and above; that is, 70% of the items such writers produce make their way into a live test or examination. In a follow up article, Coniam (2009) observes that barely-trained item writers are unlikely to achieve a quality setting rate of more than 20%. These findings lead us to the conclusion that good tests – with good items and an accurate reflection of a given proficiency level – can be produced efficiently by well-trained and experienced writers.

There has been considerable discussion of the use of expert judgement in standard setting, with differences of opinion in some quarters – Alderson and Kremmel (2013), for example. Generally, however, the use of expert judgement has been widely employed in the field of language assessment for test validation and standard setting (see Lumley, 1993; Bachman et al., 1995; Gable & Wolf,1993). Recent validation studies involving expert judgement include VanderVeen et al. (2007), Song (2008), Gao and Rogers (2011), and van Steensel et al. (2013), studies in which judges were reported to have reached high levels of agreement.

1.1 LanguageCert and the CEFR

There are six examinations in the LanguageCert International ESOL suite, all aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The examination specifications reflect the requirements of the CEFR; and test materials writers represent the highest international standards, having extensive expertise in, and knowledge and understanding of, the CEFR.

All LanguageCert test setters meet minimum requirements in terms of professional qualifications and experience in order to be eligible for consideration as an item writer. For guidance, there is an extensive item writing manual which lays out in detail how to write items and how to achieve appropriate quality standards.

Each IESOL 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 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 is presented in Table 1.

Table 1
LID Scale

CEFR level	LID scale range
C2	170-150
C1	150-130
B2	130-110
B1	110-90
A2	90-70
A1	70-50

Studies by Coniam et al. (2021) 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.

1.2 The Rasch model

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

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

1.3 Frame of reference (FOR)

To further put Rasch measurement into perspective, it is also important to understand the concept of the frame of reference (FOR) for measurement, and the parameters under which different tests may operate. Humphry (2006) defines a frame of reference as "compris[ing] a class of persons responding to a class of items in a well-defined assessment context." (p. 3) The relevance for this in the current context is that each test has, in Rasch terms, its own "internal logic" (Goodman, 1990). This internal logic refers to the starting point for Rasch measurement models: the basis for Rasch measurement is the total score of the test, computed from a particular set of items, from which the measurement based on the theoretical

probability of the particular test is extrapolated (Goodman, 1990). The theoretical probability estimated from a particular test is independent of the test (items, persons, and any other relevant facets) but not separated from it. The theoretical measurement estimated is, therefore, an objective measurement albeit specific to the test measured. Rasch calls this "specific objectivity", and is the case, for example, when we measure a rectangle and a circle with the metric. The two objects may be equal in reference to the metric system (the theoretical and objective measurement) yet different in reference to one being the measurement of four straight lines and the other that of a circumference. Thus, the Rasch measurement of a test has to be interpreted within a particular FOR.

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 candidate 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.

2 LanguageCert, the CEFR and Externally-Referenced Anchoring

Coniam and Lampropoulou (2020) in their analysis of 62 LanguageCert IESOL Listening and Reading tests using classical test statistics showed that LanguageCert IESOL tests are well constructed and robust. However, despite being robust and comparable, these IESOL tests had not been calibrated using IRT and anchor items to a single scale.

The most frequent manner of calibrating tests onto a single scale generally involves using common items between the different tests and cross-calibrating them via the Rasch scale, or via persons found in both tests and the Rasch scale. At times, however, the construction of the tests is such that there are no common elements – test items, person, or even examiners, through which linking via Rasch scale locations may be established. An alternative approach, which is investigated in this study may be referred to as 'virtual', or 'externally-referenced' linking. Linacre (2018) outlines situations where no common (or identical) items exist although items do exist that might be defined as measuring the same trait.

Boone and Staver (2020) exemplify the concept of virtual linking – or 'virtual anchoring' – in the context of mathematics where two simple addition items are presented as being construed to share the same underlying trait. While there has been some research reported on the use of virtual anchoring, this has only been in the context of test equating: Longford (2015); Boone and Staver (2020); Luppescu (1996, 2005). Further, in the latter two studies, the focus has been on the tracking of persons, with the methodology essentially being that of regression onto a latent variable from raw scores. In contrast, the current research presents externally-referenced anchoring in the context of test items. Following the use of fit statistics to first explore the robustness of the measurement, the focus in the current study is on revealing the latent trait.

A similar use of externally-referenced anchoring to that used in the current study was conducted by Humphry et al. (2014). In the context of standard setting, and the use of a modified Angoff approach, Humphry et al. used a form of externally-referenced anchoring to explore how, via use of Rasch measurement, the expert rater scale might be aligned with the test taker scale.

In the current context, externally-referenced anchoring may, therefore, be seen through the lens of expert setters. Test forms have no common items but comprise items which have been set at predefined and well-accepted CEFR levels. The fact that the levels have been internalised by expert setters through many years of experience is akin to, or rather one step up from, considering two content-related items (as with the two maths addition items referred to above).

As mentioned above, in line with Rasch principles, a test should ideally be anchored at the mid-point of the item distribution of a given scale. The mid-points of the LID scale for the six CEFR levels are presented in Table 2.

Table 2
LID Scale

CEFR level	LID scale range	Mid point
C2	170-150	160
C1	150-130	140
B2	130-110	120
B1	110-90	100
A2	90-70	80
A1	70-50	60

While there are many IESOL test forms at each CEFR level, typically there are no linking items or candidates by which cross-calibrating may be conducted. Externally-referenced anchoring using the calibrated mid-point of a given CEFR scale is therefore the method used in the current study in order to anchor the different IESOL tests onto the LID scale. The frame of reference in this case does not constitute the items but rather the CEFR scale locations calibrated through the items involved. The critical anchoring parameters in this instance are therefore the expert-rated CEFR levels of the items in a given test and the calibrated CEFR locations on the LID scale.

In order to investigate the extent to which such expertly-written yet uncalibrated test forms were indeed equivalent in terms of difficulty and level, the externally-referenced anchoring approach was applied whereby each test's midpoint was taken as an accurate representation of the level in question. The midpoint of each test in this context would then:

1. enable an effective calibration of the items in each of the IESOL tests given that no other restrictions are imposed on the items.
2. reveal the goodness of fit between the calibrated item distributions and the expertly assigned CEFR levels. The fit is determined by whether a broadly bell-shaped distribution of item measures emerges where the majority of item measures are clustered around the mean and fall between the 25th to 75th percentile and so largely within a given level.

When test development takes place, the mid-point of an individual test is intended by the test developers to represent a given CEFR ability level. It was decided to anchor the tests to the LID scale level via the mid-point for each test, which, it is argued, in turn anchors each test to the CEFR. The goodness of match of the anchoring is evaluated by the extent to which the mid-range of the items in the tests coincides with the CEFR levels on the LID scale and the extent to which the mid-range of the test item distribution includes most of the items in each test.

In this study, three IESOL tests randomly selected for each CEFR level – 18 test forms in total – are anchored by external referencing following the procedure described.

3 Current Study: Key Analytics

A number of key analytics are usually conducted when doing Rasch measurement – and have been reported on in previous LanguageCert studies (see e.g., Coniam et al., 2021). The first of these involves 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 then item outfit mean squares, item infit mean squares, and reliability.

A summary of the analysis of the 18 tests – three at each CEFR level, with each test comprising approximately 50 items – is presented below.

3.1 Item infit and outfit

The majority of the items in all tests fell 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, and not more than 5% (2-3 out of 50) items on any one test. Appendix 1 presents fuller details.

At A2, B1, C1 and C2 levels, all test item infit and outfit mean-square values were within the 0.7 and 1.3 range, indicating that the items performed well.

With A1, all infit and outfit mean-square values were within the 0.7 and 1.3 range, except for a marginally higher outfit figure on Test A1-T1, indicating a slight outlier effect.

With B2, all infit and outfit mean-square values were within the 0.7 and 1.3 range except for an outfit of 2.26 on Test B2-T2, and 2.01 on Test B2-T3 – although these relatively high outfits only occurred at the 90th percentile.

3.2 Reliability

Test reliability, for a 50-item test, is proposed to be at 0.7 or above (Ebel, 1965). For an 80-item test, 0.8 or better is the projected figure, and it is this which is taken as the baseline in the current study. For the 62 tests reported on in the Coniam and Lampropoulou (2020) study, almost all test reliabilities – via the KR20 statistic – were above 0.8.

The equivalent of classical test measures of reliability in Rasch is person reliability (Anselmi et al., 2019); this statistic is currently reported for all 18 tests in the current study. As Appendix 1 illustrates, the target of 0.8 or better was achieved by externally-referencing all tests for all levels apart from one A1 test with a reliability of 0.75, and one A2 test with a reliability of 0.77.

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

A fuller picture of the data is available in Appendices 1, 2 and 3. Appendix 2 illustrates test C1 T1, for which the midpoint is 140 on the LID scale. As can be seen the item distribution is quite regular and bell-shaped. Appendix 3, which illustrates test A1 T1 for which the midpoint is 60, is not quite as regular, being somewhat bimodal with a set of more demanding items towards the upper end of the scale. In general, however, as discussed below, the results reflect more the picture presented by the C1 than the A1 test.

4 Externally-Referenced Anchoring: The Study

Table 3

Item Distributions in A1 Externally-referenced Anchored IESOL Tests

A1 ERA midpoint = 60	A1 T1	A1 T2	A1 T3
No. of items	52	52	50
Mean	60	60	61.3
Std. Deviation	37.68	24.48	24.45
25th percentile	32.1	38.83	50.43
50th percentile	53.05	63.93	62.35
75th percentile	71.51	76.21	75.4

An analysis of the 18 tests from two perspectives is presented below. First, tables are presented with test means and measures that emerged after externally-referenced anchoring, in particular at the means recorded at the 25th, 50th and 75th percentiles. Second, graphs are presented which provide a more visual representation of the outcome of the externally-referenced anchoring (ERA).

The externally-referenced anchoring midpoint for A1 was 60. For Test A1 T1, the mean measure at the 50th percentile was 53.05, a third of a logit (i.e., 6 points) below the midpoint; for Test A1 T2 and T3, the mean measure at the 50th percentile was very close to the midpoint of 60. 70 is the top end of the A1 cut score; figures recorded at the 75th percentiles for all three tests were very close to this figure of 70. This confirms the fact that the majority of items at this level are in the level.

Table 4

Item Distributions in A2 Externally-referenced Anchored IESOL Tests

A2 ERA midpoint = 80	A2 T1	A2 T2	A2 T3
No. of items	52	52	52
Mean	80	80	80
Std. Deviation	20.38	21.21	21.75
25th percentile	68.36	69.94	64.79
50th percentile	78.08	82.78	78.32
75th percentile	97.63	90.89	92.8

The externally-referenced anchoring midpoint for A2 was 80. At the 50th percentile, all three tests were very close to this figure. With 90 as the top end of the A2 cut score; the 75th percentiles of A2 T2 and T3 had means very close to this figure; A2 T1 had some rather more demanding items, with a slightly higher mean measure of 97.63.

Table 5

Item Distributions in B1 Externally-referenced Anchored IESOL Tests

B1 ERA midpoint = 100	B1 T1	B1 T2	B1 T3
No. of items	45	47	50
Mean	98.88	91.29	98.18
Std. Deviation	22.18	17.72	19.04
25th percentile	81.48	81.84	81.25
50th percentile	106.2	91.87	100.09
75th percentile	116.18	101.63	111.41

The externally-referenced anchoring midpoint for B1 was 100. Tests B1 T1 and T3 were very close to this figure at the 50th percentile; items in B1 T2 were slightly easier. With 110 as the top end of the B1 cut score, a similar picture emerged: B1 T1 and T3 were very close to the 75th percentile, while B1 T2 had items of slightly easier values at 101.63.

The externally-referenced anchoring midpoint for B2 was 120. Tests B2 T1 and T3 were very close to this figure at the 50th percentile; items in B2 T2 were slightly easier. With 130 as the top end of the B2 cut score, a similar picture emerged. At the 75th percentile, the B2 T1 and T3 mean measures were very close to this cut score, while B2 T2 had items which were rather more demanding at 151.88.

Table 6

Item Distributions in B2 Externally-referenced Anchored IESOL Tests

B2 ERA midpoint = 120	B2 T1	B2 T2	B2 T3
No. of items	48	44	47
Mean	120	117.9	120
Std. Deviation	20.35	37.72	29.42
25th percentile	106.09	91.98	101.13
50th percentile	119.11	113.78	118.11
75th percentile	133.67	151.88	137.63

Table 7

Item Distributions in C1 Externally-referenced Anchored IESOL Tests

C1 ERA midpoint = 140	C1 T1	C1 T2	C1 T3
No. of items	51	52	52
Mean	139.06	140	140
Std. Deviation	16.2	17.41	19.09
25th percentile	128.7	128.24	122.89
50th percentile	141	141.59	140.78
75th percentile	149	149.55	149.44

The externally-referenced anchoring midpoint for C1 was 140. All three were almost exactly at this figure, showing an extremely close fit. Similar pictures were recorded at the 25th and 75th percentiles. With 150 being the top end of the C1 cut score, a very similar picture emerged, with all three tests having mean measures almost exactly at this figure.

Table 8

Item Distributions in C2 Externally-referenced Anchored IESOL Tests

C2 ERA midpoint = 160	C2 T1	C2 T2	C2 T3
No. of items	50	50	50
Mean	158.22	158.59	158.43
Std. Deviation	18.71	14.68	16.04
25th percentile	143.51	147.5	146.75
50th percentile	160.73	157.6	158.64
75th percentile	172	171.2	169.13

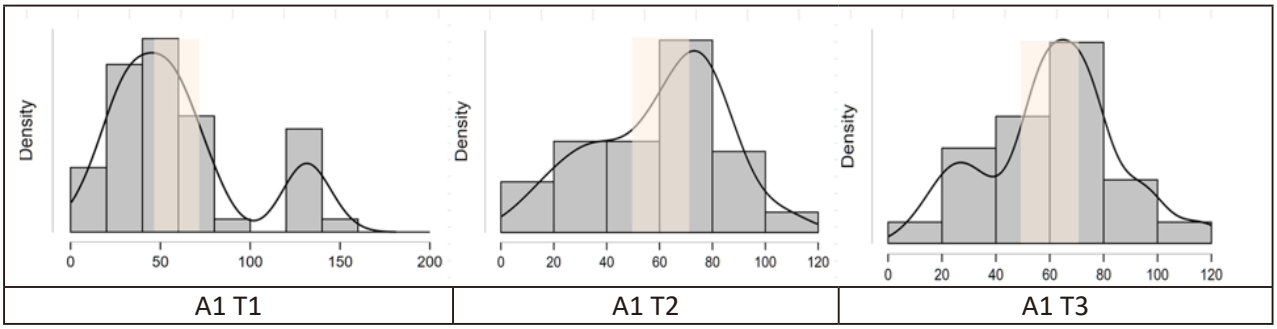
Figures recorded for C2 were very similar to those returned for C1. With the externally-referenced anchoring midpoint for C2 being 160, all three C2 were almost exactly at this figure. Similar pictures were recorded at the 75th percentiles. With 170 the top end of the C2 cut score, a similar picture to C1 again emerged, with all three tests having mean measures almost exactly at the 170-point C2 top end cut score figure.

As a parallel view, and a reframing of the data presented in the tables above, the charts in Figures 1 - 6 below contain the results of anchoring as matched visually against the LID CEFR levels.

The grey bars and the trend graphs represent the IESOL item distributions; the shaded areas are the LID CEFR ranges. The density represents the frequency of items at a given LID scale range.

Figure 1

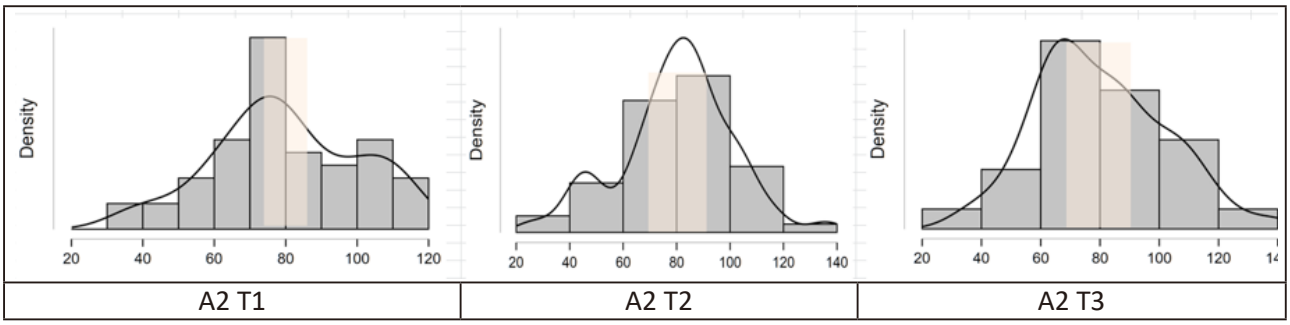
Externally-referenced Anchoring of A1 Level Tests



60 is the externally-referenced anchoring midpoint for A1, repeated by the orange shading. Tests A1 T2 and T3 show quite a normal distribution, in particular with test A1 T3. Test A1 T1 is less regular – being somewhat bimodal with a number of items which are more demanding than might be expected at A1 level.

Figure 2

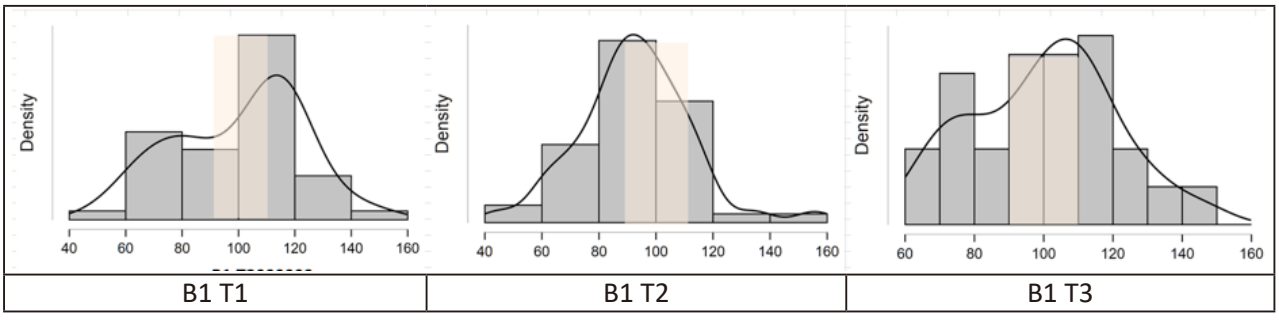
Externally-referenced Anchoring of A2 Level Tests



80 is the externally-referenced anchoring midpoint for A2. A2 T2 fits a normal distribution well, as does A2 T1 although this test has quite a large number of items exactly around the midpoint of the A2 scale.

Figure 3

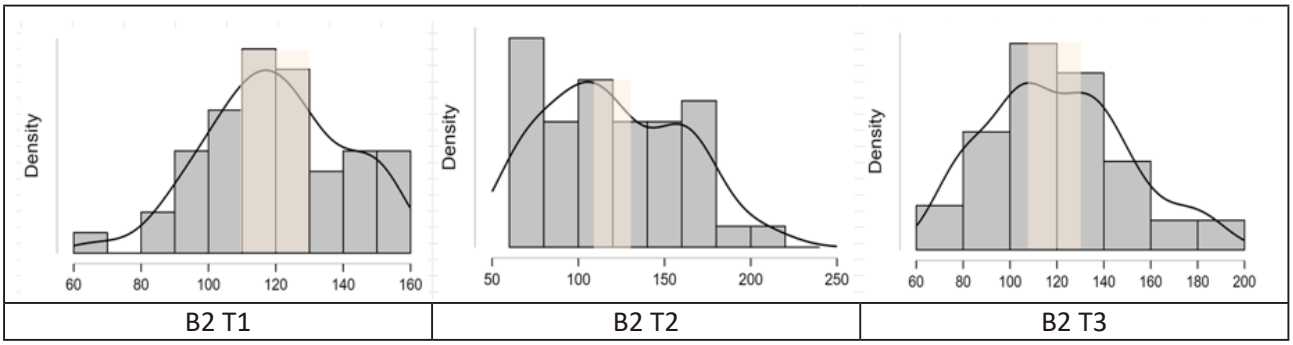
Externally-referenced Anchoring of B1 Level Tests



100 is the externally-referenced anchoring midpoint for B1, repeated by the orange shading. B1 T2 shows quite a normal distribution. The B1 T1 and T3 tests are slightly negatively skewed towards more demanding items.

Figure 4

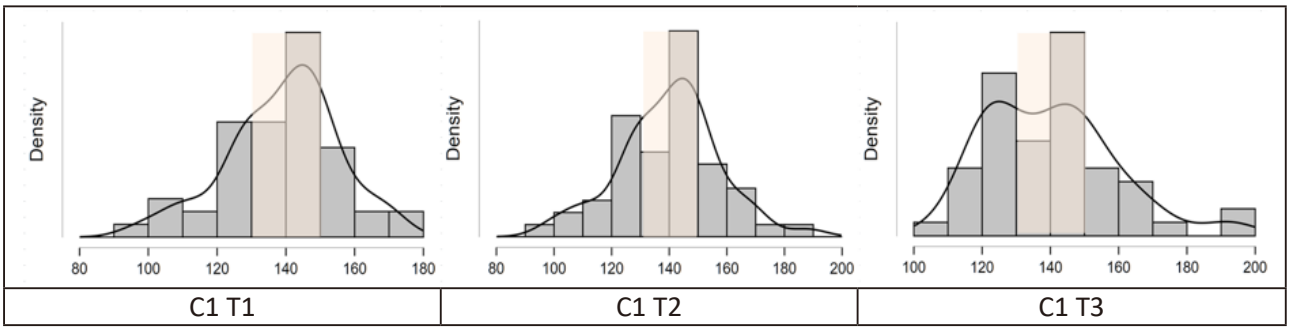
Externally-referenced Anchoring of B2 Level Tests



120 is the externally-referenced anchoring midpoint for B2. B2 T1 and T3 show quite normal distributions; B2 T2 items are distributed in a slightly narrower range.

Figure 5

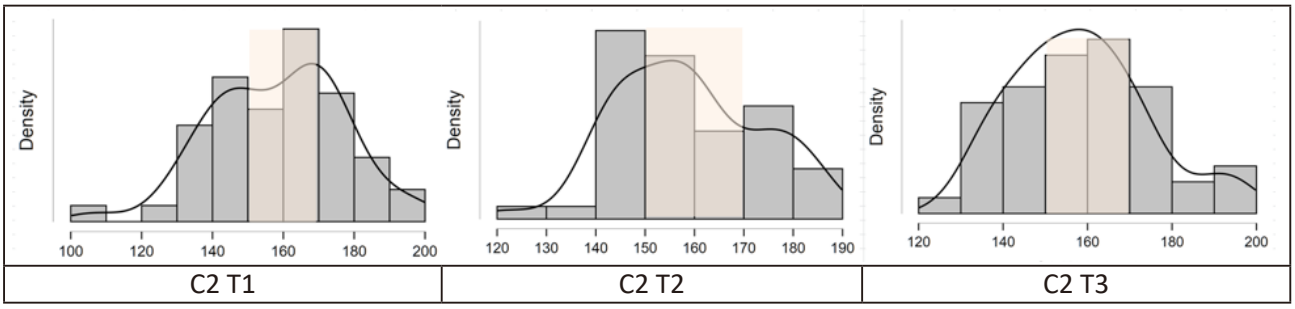
Externally-referenced Anchoring of C1 Level Tests



140 is the externally-referenced anchored midpoint for C1. All three tests show generally normal distributions.

Figure 6

Externally-referenced Anchoring of C2 Level Tests



160 is the externally-referenced anchoring midpoint for C2. All three tests again show generally normal distributions.

It can be seen that the LID CEFR zones in general occupy the centre of IESOL item distribution, with this distribution including a substantial number of the items in a given test. The expert-rated CEFR levels for the IESOL tests match well with the calibrated LID scale CEFR levels. The IESOL tests may therefore be considered to be acceptably well anchored onto the LID scale.

5 Discussion and Conclusion

This paper has reported on the externally-referenced anchoring of LanguageCert IESOL tests against the LanguageCert LID scale CEFR levels. Calibrating tests onto a single scale generally involves using common items between different tests and cross-calibrating them using Rasch measurement. When there are no linking items available, other methods, however, need to be used. One of these, proposed by Linacre (2018), involves the use of items that measure the same trait, i.e., externally-referenced anchoring. In the current context, externally-referenced anchoring is illustrated through the lens of expert setters who have been producing quality items (see Coniam & Lampropoulou, 2020) at predefined and well-understood CEFR levels for many years.

Two related hypotheses regarding the validity of externally-referenced anchoring are investigated. The first is that good Rasch infit and outfit statistics from the externally-referenced anchoring process are achieved. At each of the six CEFR levels, three different test forms were selected at random for analysis and good Rasch infit and outfit statistics are indeed found for each test. The first hypothesis is therefore confirmed.

The second is that broadly bell-shaped item measure distributions would emerge from the analysis. All analyses generally recorded a good match between IESOL-assigned CEFR levels and the LID scale CEFR levels, with sets of items, for the most part, showing generally balanced distributions. The majority of items in almost all tests fell within the 25th to 75th percentiles: the points at which these percentiles broadly match the upper and lower end of the cut scores determined for a given CEFR level. Hypothesis two is also confirmed.

As may be seen in Appendix 1, not all matches between the items distributions and the LID scale are perfect; in general, however, a close match is reported, with the majority of items falling between the 25th and 75th percentiles – the lower and upper bounds of LID scale values for a given CEFR level. Consequently, two findings emerge: the results indicate that LanguageCert IESOL test items are generally appropriate for the respective CEFR level; and the concept of externally-referenced anchoring as a methodology is also validated.

The match in the current study between externally-referenced anchored levels and LID scale CEFR levels reinforces the argument that LanguageCert IESOL tests have been well set, and statistically verify expert judgements. The fact that the majority of the items fall within the 25th to 75th percentiles confirms the contention that the items in the IESOL tests are well-targeted at the appropriate CEFR level by expert setters. The present study lends further support to the use of expert ratings in assessment.

While the externally-referenced anchoring outcomes obtained from the current study confirm the robustness of LanguageCert tests reported elsewhere (Coniam & Lampropoulou, 2020), only three tests were analysed at each CEFR level. Using externally-referenced anchoring principles, a study is therefore currently underway to analyse a single dataset containing 15 tests at any given CEFR level. Results from the new study will supplement the picture of the current study and will be reported on in due course.

Appendix 1
Fit Statistics and Person Reliabilities

A1	A1-T1 Items	A1-T1 Infit	A1-T1 Outfit	A1-T2 Items	A1-T2 Infit	A1-T2 Outfit	A1-T3 Items	A1-T3 Infit	A1-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	60	0.96	1.25	60	1	0.88	57.85	0.99	0.96
S.D.	37.68	0.11	0.7	24.48	0.15	0.39	29.97	0.13	0.29

25th pc'tile	32.1	0.89	0.84	38.83	0.89	0.67	44.85	0.91	0.74
50th pc'tile	53.05	0.96	1.05	63.93	0.97	0.81	60.43	0.97	0.95
75th pc'tile	71.51	1.04	1.43	76.21	1.05	1.03	73.83	1.05	1.08
Reliability	0.84			0.75			0.82		

A2	A2-T1 items	A2-T1 Infit	A2-T1 Outfit	A2-T2 items	A2-T2 Infit	A2-T2 Outfit	A2-T3 items	A2-T3 Infit	A2-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	80	0.99	1.04	80	0.99	0.96	80	1	0.98
S.D.	20.35	0.17	0.42	21.24	0.17	0.41	21.76	0.13	0.33
25th pc'tile	68.37	0.87	0.74	69.98	0.89	0.72	64.79	0.93	0.75
50th pc'tile	78.08	0.97	0.93	82.81	0.97	0.86	78.31	1.02	0.96
75th pc'tile	97.6	1.1	1.32	90.93	1.04	1.09	92.8	1.09	1.17
Reliability	0.88			0.88			0.77		

B1	B1-T1 items	B1-T1 Infit	B1-T1 Outfit	B1-T2 items	B1-T2 Infit	B1-T2 Outfit	B1-T3 items	B1-T3 Infit	B1-T3 Outfit
Valid	46	46	46	52	52	52	52	52	52
Missing	6	6	6	0	0	0	0	0	0
Mean	100	1	1.04	100	1	1.28	100	1	1.01
S.D.	23.21	0.21	0.39	32.1	0.14	1.23	20.81	0.14	0.43
25th pc'tile	81.55	0.8	0.7	83.33	0.9	0.8	83.27	0.92	0.76
50th pc'tile	106.72	0.99	0.93	94.31	1.02	0.97	101.38	0.99	0.92
75th pc'tile	116.78	1.12	1.27	108.11	1.1	1.23	112.67	1.08	1.07
Reliability	0.88			0.85			0.84		

B2	B2-T1 items	B2-T1 Infit	B2-T1 Outfit	B2-T2 items	B2-T2 Infit	B2-T2 Outfit	B2-T3 items	B2-T3 Infit	B2-T3 Outfit
Valid	48	48	48	48	48	48	47	47	47
Missing	0	0	0	0	0	0	1	1	1
Mean	120	0.99	1	737.44	0.97	1.22	120	0.99	1.13
S.D.	20.35	0.18	0.27	2416.94	0.24	1.63	29.42	0.17	0.7
25th pc'tile	106.09	0.84	0.76	93.48	0.85	0.53	101.13	0.86	0.75
50th pc'tile	119.11	0.98	1	117.83	1	0.71	118.11	0.97	0.88
75th pc'tile	133.67	1.12	1.19	161.12	1.1	1.11	137.63	1.1	1.19
Reliability	0.86			0.88			0.87		

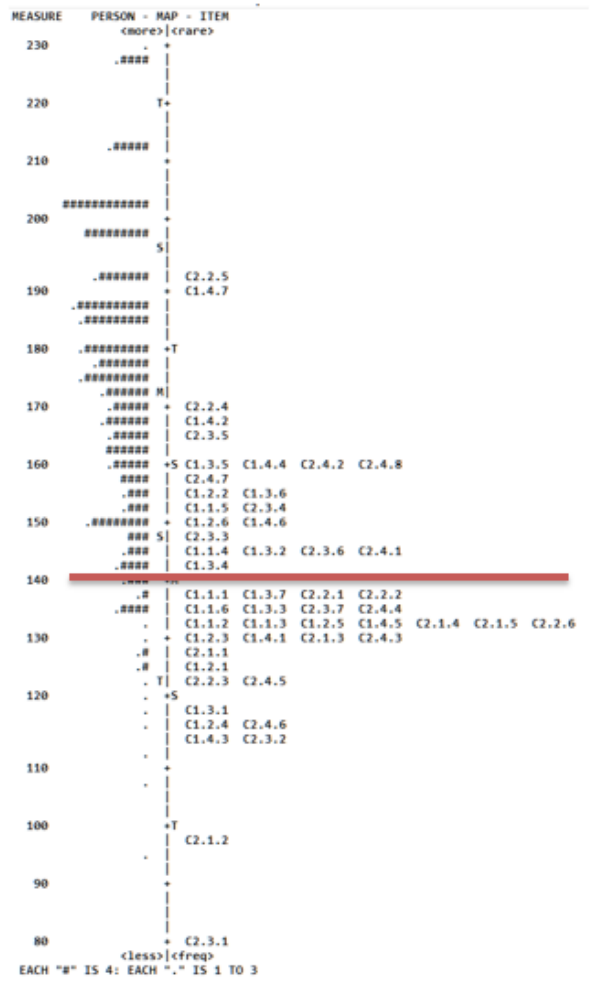
C1	C1-T1 items	C1-T1 Infit	C1-T1 Outfit	C1-T2 items	C1-T2 Infit	C1-T2 Outfit	C1-T3 items	C1-T3 Infit	C1-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	140	1	0.98	140	0.99	0.93	140	0.99	0.98
S.D.	19.99	0.1	0.3	17.41	0.1	0.25	19.09	0.14	0.3
25th pc'tile	131.01	0.93	0.79	128.24	0.92	0.77	122.89	0.91	0.76
50th pc'tile	137.29	1	0.93	141.59	0.97	0.91	140.78	0.97	0.91
75th pc'tile	153.25	1.05	1.06	149.55	1.05	1.06	149.44	1.02	1.06
Reliability	0.85			0.81			0.88		

C2	C2-T1 items	C2-T1 Infit	C2-T1 Outfit	C2-T2 items	C2-T2 Infit	C2-T2 Outfit	C2-T3 items	C2-T3 Infit	C2-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	160	1	0.92	160	0.99	0.93	160	0.99	0.95
S.D.	20.46	0.13	0.27	16.06	0.12	0.24	17.6	0.14	0.27
25th pc'tile	144.8	0.91	0.75	148.12	0.92	0.82	147.18	0.88	0.76
50th pc'tile	161.54	0.97	0.92	157.96	0.96	0.92	159.58	0.99	0.92
75th pc'tile	172.22	1.08	1.05	172.2	1.05	1.05	170.69	1.05	1.06
Reliability	0.83			0.81			0.84		

Appendix 2
Sample C1 Test Outputs from IESOL Calibrations

Test C1 T1 (midpoint = 140)

```
-------------------------------------------------------------------------------
| PERSON     670 INPUT     670 MEASURED                INFIT         OUTFIT   |
|          TOTAL      COUNT      MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      39.4       53.7       172.55    8.83      1.00     .1    .98     .1|
| P.SD       8.8        2.4        24.41    3.58       .13     .7    .35     .7|
| REAL RMSE   9.53 TRUE SD   22.48  SEPARATION  2.36  PERSON RELIABILITY  .85|
|-----------------------------------------------------------------------------|
| ITEM      54 INPUT      52 MEASURED                INFIT         OUTFIT   |
|          TOTAL      COUNT      MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN     507.3      666.6       140.00    2.30      1.00     .0    .98    -.1|
| P.SD      90.4        2.9        19.80     .65       .10    2.0    .30    2.3|
| REAL RMSE   2.39 TRUE SD   19.65  SEPARATION  8.24  ITEM   RELIABILITY  .99|
-------------------------------------------------------------------------------
```



Appendix 3
Sample A1 Test Outputs from IESOL Calibrations

Test A1 T1 (midpoint = 60)

PERSON	518 INPUT	518 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	35.3	51.9	82.98	9.14	.98	-.2	1.24	-.1
P.SD	7.2	.5	23.48	2.63	.56	1.7	1.46	1.5
REAL RMSE	9.51	TRUE SD	21.46	SEPARATION	2.26	PERSON RELIABILITY		.84
ITEM	54 INPUT	52 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	351.7	516.7	60.00	2.80	.96	-.2	1.25	1.2
P.SD	143.7	1.2	37.32	.65	.10	1.5	.69	3.1
REAL RMSE	2.88	TRUE SD	37.21	SEPARATION	12.93	ITEM	RELIABILITY	.99

```
MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
  159             .  +
                     |
                     |
                ###  |
  149                +
                     |  1.2.1
                  #  |
  139                +
                     |  1.1.4
                  .  |T 2.1.3
                     |  1.1.5  1.1.6
  129              T+  1.1.1  1.2.2
                  .  |  1.1.2
                     |
                ###  |  2.1.5
  119                +
                .##  |
                     |
  109     #########  +
                    S|
       .###########  |
       .##########   |
   99            #   +
         .#########  |S
                  .  |
     ############## |
   89    ##########  +
     ############## |
         .######## M|  1.3.3
   79          .#   +
         #########  |  1.2.4  1.3.1  2.3.4
       .##########  |
            ######  |  2.2.2  2.3.2
   69      .#######  +
              ####  |  1.1.3
           #########|  1.2.3  1.2.5
   59            ### S+M 1.3.2  2.2.3
               .###  |  1.2.6
               .##   |  2.1.2  2.2.4  2.2.5  2.3.7  2.4.7
                ##   |  2.1.1
   49         #####  +  1.3.5  2.4.8
               .##   |  1.4.1
                .#   |  1.3.6  1.4.4
                 .   |  2.3.6  2.4.6
   39            .   +  1.2.7  2.4.3
               .### T|  2.3.1
                  #  |  1.4.3  2.3.5  2.4.5
   29             .  +  1.4.5  2.1.6
                  .  |  1.4.2  2.3.3  2.4.4
                  .  |  1.3.4
                  .  |S 2.1.4
   19                +  2.4.1  2.4.2
                     |  2.2.1
                     |
                     |
    9                +  1.1.7
                  .  |
                     |  1.4.6
    0                +
              <less>|<freq>
EACH "#" IS 3: EACH "." IS 1 TO 2
```

References

Alderson, C. J., Alderson, J. C., Clapham, C., Wall, D., & Swan, M. (1995). *Language test construction and evaluation*. Cambridge University Press.

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. http://doi.org/10.1177/0265532213489568.

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, *10*, 2714. http://doi.org/10.3389/fpsyg.2019.02714.

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 University Press.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*, 4th ed. Routledge. http://doi.org/10.4324/9780429030499.

Boone, W. J., & Staver, J. R. (2020). *Externally-referenced equating of test forms. In advances in Rasch analyses in the human sciences*. Springer. http://doi.org/10.1007/978-3-030-43420-5_12.

Bristol, T. (2015). Test item writing: 3Cs for successful tests. *Teaching and Learning in Nursing*, *10*(2), 100-103. http://doi.org/10.1016/j.teln.2015.01.004.

Coniam, D., & Lampropoulou, L. (2020). *A review of LanguageCert IESOL listening and reading test reliabilities 2018-2020*. LanguageCert.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System*, *37*(2), 226-242. http://doi.org/10.1016/j.system.2008.11.008.

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

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, *14*(2), 5-22. http://doi.org/10.1558/cj.v14i2-4.15-33.

Ebel, R. L. (1965). *Measuring educational achievement*. 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.). 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. http://doi.org/10.1177/0265532210364380.

Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. *Sociological Methodology*, *20*, 249-294. http://doi.org/10.2307/271088.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*(1), 37-50. http://doi.org/10.1207/s15324818ame0201_3.

Humphry, S. (2006). *The impact of differential discrimination on vertical equating*. ARC report. Western Australia Department of Education & Training.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education*, *27*(1), 1-18. http://doi.org/10.1080/08957347.2014.859492.

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

Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*, *40*(3), 227-253. http://doi.org/10.3102/1076998615574773.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*(3), 211-234. http://doi.org/10.1177/026553229301000302.

Luppescu, S. (1996). *Virtual equating: An approach to reading test equating by concept matching of items* (Doctoral dissertation. University of Chicago, Department of Education.

Luppescu, S. (2005). Externally-referenced equating. *Rasch Measurement Transactions*, *19*(3), 1025.

Rodriguez, M. C. (1997). *The art & science of item writing: A meta-analysis of multiple-choice item format effects*. In annual meeting of the American educational research association, Chicago, IL.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435-464. http://doi.org/10.1177/0265532208094272.

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. http://doi.org/10.1177/0265532212440950.

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*. Lawrence Erlbaum Associates.

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

Tony Lee is Senior Psychometrician at LanguageCert. He has been involved in language assessment statistical analysis work since 1980 in universities in Hong Kong and Australia. His major language assessment work includes the assessment management of the Australian Federal Government's migrant English assessment system ACCESS as well as the Hong Kong Government's English Language Ability scale.

Michael Milanovic is Chairman of LanguageCert and a member of its Advisory Council. Previously CEO of Cambridge Assessment English, he has been working extensively with PeopleCert since 2015. He worked closely with the Council of Europe on its Common European Framework of Reference, has held, and still holds a number of key external roles.

Nigel Pike was previously Director of Assessment at Cambridge Assessment English, directing the delivery of all Cambridge English examinations. Nigel holds an MBA, and has extensive experience with national and local ministries of education around the globe, delivering consultancy, customised examinations and developing language policy for governments.