Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-0959
https://www.castledown.com/journals/lea/

Language Education & Assessment, 5(1), 18–33 (2022)
https://doi.org/10.29140/lea.v5n1.769

The Role of Expert Judgement in Language Test Validation

DAVID CONIAM[a]

TONY LEE[b]

MICHAEL MILANOVIC[b]

NIGEL PIKE[b]

WEN ZHAO[c]

[a] *PeopleCert, UK*
David.Coniam@PeopleCert.org;

[b] *LanguageCert, UK*
Tony.Lee@PeopleCert.org;
Michael.Milanovic@PeopleCert.org;
Nigel.Pike@PeopleCert.org

[c] *School of Foreign Studies, Jinan University, China*
dianawen@hotmail.com

Abstract

The calibration of test materials generally involves the interaction between empirical analysis and expert judgement. This paper explores the extent to which scale familiarity might affect expert judgement as a component of test validation in the calibration process. It forms part of a larger study that investigates the alignment of the LanguageCert suite of tests, Common European Framework of Reference (CEFR), the China Standards of English (CSE) and China's College English Test (CET).

In the larger study, Year 1 students at a prestigious university in China were administered two tests—one with items based on China's College English Test (CET), and the other a CEFR-aligned test developed by LanguageCert—the LanguageCert Test of English (LTE). Comparable sections of the CET and the LTE involved sets of discrete items targeting lexico-grammatical competence.

Data Availability Statement: *All relevant data are within this paper.*

In order to ascertain whether expert judges were equally comfortable placing test items on either scale (CET or CEFR), a group of professors from the university in China who set the CET-based test, were asked to expert judge the CET items against the nine CSE levels with which they were very familiar. They were then asked to judge the LTE items against the six CEFR levels, with which they were less familiar. Both sets of expert ratings and the test taker responses on both tests were then calibrated within a single frame of reference and located on the LanguageCert scale

In the analysis of the expert ratings, the CSE-familiar raters exhibited higher levels of agreement with the empirically-derived score levels for the CET items than they did with the equivalent LTE items. This supports the proposition that expert judgement may be used in the calibration process where the experts in question have a strong knowledge of both the test material and the standards against which the test material is to be judged.

Keywords: expert judgement, test validation, reading and usage, CEFR, CSE

The CEFR and the CSE

For the past two decades, the CEFR has come to be accepted as illustrating standards of language ability by many stakeholders: policy makers, exam bodies and test developers (Deygers *et al.,* 2018). Not only in Europe, but in many countries around the world (Little, 2007), the CEFR has become the common currency for specifying levels of language ability (Figueras, 2012).

The CSE reflects an overarching notion of language ability, with which language knowledge and strategies co-function in performing a language activity. Its development attempts to pull together all the different English language curriculums and assessment instruments into one overarching framework.

With the aim of achieving one overarching framework, scholars in China proposed that a unified Asian framework of English language proficiency be established to promote the development of language teaching, learning and assessment in Asia (see Yang & Gui, 2007). Jin et al. (2017) describe the development of the *Common Chinese Framework of Reference for English (CCFR-E): Teaching, Learning, Assessment,* which began to be developed in 2014. In line with the CCFR-E, also in 2014, the Ministry of Education launched the development of China's Standards of English (CSE), a unified framework of English language ability (Liu, 2017). The CSE, released in 2018, has three major levels, each subdivided into three sublevels (Liu, 2019). Figure 1 illustrates the two frameworks, giving an indication of how they appear to be aligned.

Common European Framework of Reference (CEFR)		**China Standards of English (CSE)**	
Label	Level	Level	Label
Proficient User	C2	Level 9	Advanced Stage
	C1	Level 8	
		Level 7	
Independent	B2	Level 6	Intermediate Stage
	B1	Level 5	
		Level 4	
Basic User	A2	Level 3	Elementary Stage
	A1	Level 2	
		Level 1	

Figure 1 *CEFR and CSE levels.*

The Use of Expert Judgement in Language Assessment

'Expert judgement' in language assessment has been a long-accepted practice in test development both in item writing and in the estimation of item difficulty—which in turn impacts level setting and cut scores. In the case of test setting, the use of the 'expert' is critical. In a study of minimally-trained item writers Coniam (2009) reported such personnel as achieving a quality setting rate of only approximately 20%; i.e., items that may be defined as having good item statistics (see Falvey *et al.,* 1994). A number of ground rules for the setting of good items was proposed by Haladyna & Downing (1989); many of these also appear in Alderson et al.'s (1995) discussion of the qualities needed of an "expert item writer". To be able to efficiently produce good tests—with good items and an accurate reflection of a given proficiency level—it is therefore clear that test item writers need to be both familiar and experienced with the test they are engaging with, as well as being well-trained.

There has been considerable discussion of the use of expert judgement in standard setting, with research on one side supporting the use of experts (e.g., Shiotsu, 2010), with some dissenting voices in other quarters (see e.g., Mehrens, 1995; Alderson & Kremmel; 2013).

Studies comparing expert judge ratings against expected difficulty or empirical scores have reported mixed results. Studies which required independent predictions of judges have reported correlations in the 0.3 range (see e.g., Melican *et al.,* 1989; Hambleton *et al.,* 2003). In contrast, studies which provided raters with a clear framework and with training have reported higher correlations—in the 0.7 range (see e.g., Attali *et al.,* 2014). Lu & Read (2021), whose study compared two groups of experts' judgements on reading task item content, reported a general convergence of about 53% of the items.

Generally, the use of expert judgement has been widely employed in the field of language assessment for test validation and standard setting (e.g., Bachman *et al.,* 1995). In recent expert judgement validation studies, judges have reportedly reached comparatively high levels of agreement (e.g., Gao & Rogers, 2011; van Steensel *et al.,* 2013).

The Use of Expert Judgement within the CEFR

The use of expert judgement and opinion has been quite extensive in establishing correspondences with the CEFR. Utilising the framework laid out in the Manual for linking examinations to the CEFR (Council of Europe, 2009), a number of studies have explored the correspondences between the CEFR and the CSE. In Dunlea et al.'s (2019) study, for example, the methodology involved the expert judgement of items against CSE and CEFR levels and the assignment of CSE descriptors against tasks, followed by field testing the proposed levels with Chinese teachers of English. Peng and associates have undertaken a number of studies investigating correspondences between CEFR and CSE levels. In one study, Peng (2021) investigated level alignments between the CSE and CEFR writing descriptors, while Peng & Liu (2021) investigated the alignment of CSE listening skill levels with those of the CEFR. Zhao et al. (2017) illustrated how College English vocabulary levels might be linked with the CEFR.

Other key uses of expert judgement in the Asian context may be seen in the studies exploring correspondences between the CEFR and Taiwan's General English Proficiency Test (GEPT). Brunfaut and Harding (2014) explored the linking of the GEPT listening test to the CEFR; Knoch and Frost (2016) explored the linking of the GEPT Writing test to the CEFR; Green and Inoue (2017) explored how GEPT Speaking Tests might be compared to the CEFR.

Positioning the Current Study

Against the above backdrop, the current study presents a comparative picture. In the first instance, a single group of experts rate items from a test with which they are familiar against standards with which they are also familiar. In the second instance, the same group of experts rate items from a test with which they are not familiar against standards with which they are much less familiar.

Within this context, four sets of data regarding the College English Test (CET) and LanguageCert Test of English (LTE) discrete item subtests constitute the current dataset [Note 1]:

- Test taker results on a CET test
- Test taker results on an LTE test
- Expert judgement of CET discrete test items against the CSE
- Expert judgement of LTE discrete test items against the CEFR

Both sets of test taker responses to, and expert ratings of, test items were then calibrated together within a single scale of reference; following which the expert ratings were analysed on a single scale—the scale developed and used by LanguageCert.

Current Study: Assessment Instruments, Test Taker Sample, Hypotheses

This section briefly outlines the background and make-up of the tests and the self-assessment ratings which test takers completed.

In late 2020, approximately 2,500 Year 1 university CET students took a 65-item multiple-choice reading and language use test prepared by experts from the China university. Three months later, this same set of students took a 53-item multiple-choice reading and language use CET test adapted from previously-validated LanguageCert Test of English (LTE) material (Coniam *et al.,* 2021). The items in the LTE test used in the study were selected on the basis of having been calibrated to represent the spectrum of difficulty across the six CEFR levels. As long as students have not been given results nor had additional practice, a three-month period has been shown to have no effect on test results (Cheng, 1993). Since the students had had no preparation or coaching for the LTE test, student ability may be taken as similar for both tests.

Item difficulty in LTE tests is predicated on the overarching LanguageCert Item Difficulty (LID) scale; see Table 1. This scale lays out item difficulty levels generally adopted in LanguageCert assessments (Coniam *et al.,* 2021).

For analysis and calibration purposes, 100 is taken as the mid-point of the scale. To this end, Rasch logit values are rescaled to a mean of 100 and a standard deviation (SD) of 20 (see Coniam *et al.,* 2021).

Table 1 *LID Scale Values*

CEFR level	LID scale range
C2	151–170
C1	131–150
B2	111–130
B1	91–110
A2	71–90
A1	51–70

As Zhao & Coniam (2022) describe, in a comparative analysis of the make-up of the two reading and usage tests, despite some differences, the content of the two tests, and even the order in which the different sections of the test appeared to test takers, exhibited a great deal of similarity.

The data in the current study involved *discrete items* which assessed grammar, syntax, vocabulary and usage. There were 30 such items in the CET test and 23 items in the LTE test. Appendix 1 provides detail on the two tests; Section 2 are the items in focus in the current study. There were eight expert judges, professors with higher degrees in their late 40s or 50s from the Foreign Studies Department. All of them had been at the university for over ten years, had been involved in the teaching and assessment of English, and had been setting CET items and other assessment test types for the university's students. As confirmed by senior staff at the university, the eight professors—the expert judges—had a clear picture of standards in the CSE. Also, given the fact that they were all English language professionals, most had knowledge of, albeit not in-depth familiarity with, the CEFR.

Before rating took place, training and standardisation sessions were conducted for the expert judges participating in the study. The purpose of these sessions was to increase rater reliability and familiarity with the two frameworks. In an initial orientation session, judges were given the CSE and CEFR frameworks and key issues relevant to the two frameworks were outlined. Following this orientation, judges trial-rated sample CET items against the nine CSE levels, after which they shared scores, with a discussion of discrepancies leading to further standardisation. This process was then repeated for the LTE items. Judges trial-rated sample LTE items against the six CEFR levels. Subsequent to the training, the expert raters were then given the 30 CET items to rate against the nine CSE levels and the 23 LTE items to rate against the six CEFR Levels.

The overarching hypothesis in the study was that levels of agreement achieved by expert judges rating the CET items against the CSE—with which they were very familiar—would be better than levels of agreement achieved rating LTE items against the CEFR—with which they were less familiar.

Against this backdrop, the hypotheses in the current study are defined as levels of agreement as expressed by the statistic Cohen's Kappa (for which a definition is provided below). In the discussion below, the term "assessed" is used as shorthand for "test takers' mean scores on the items"; "rated" then refers to "expert judges' ratings of item difficulty".

Hypothesis 1: On the CET items, a high level of agreement between assessed and expert-rated values will be obtained. Such agreement will be exhibited via a Kappa value of 0.8 ('strong agreement').

Hypothesis 2: On the LTE items, a low level of agreement between assessed and expert-rated values will be obtained. Such agreement will be exhibited via a Kappa value of 0.4 ('fair agreement').

Statistical Analysis

This section briefly outlines the statistics used in the current study.

Rasch Measurement

The manner for gauging test fitness-for-purpose in the current study, and for linking the data—the two different tests and self-assessments—involves the use of Rasch measurement, which will now be briefly outlined.

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. Once a common metric is established for measuring different phenomena (test takers and test items in the current instance), test taker ability may be estimated independently of the items used, with item difficulty estimates also estimated independently from the sample (Bond *et al.,* 2020).

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.

Key analytics usually reported when doing Rasch measurement, and which have been reported on in previous LanguageCert studies (e.g., Coniam *et al.,* 2021), involve the 'fit' of the data to the Rasch model. This refers, in essence, to how well obtained values match expected values. While a perfect fit of 1.0 indicates that obtained values match expected values exactly, acceptable ranges of tolerance for fit are generally taken as ranging from 0.5 to 1.5 (Lunz & Stahl, 1990). Key statistics usually reported are infit and outfit mean squares, and reliability.

Kappa

Cohen's Kappa is a statistical measure for examining the agreement between two rated categories. It aids in determining the implementation of a given coding system.

In the current study, Kappa determines levels of agreement between the two variables—assessed and rated item values—against the nine CSE and six CEFR levels. Following recoding of the CEFR levels as 1–6 (A1 = 1, C2 = 6 etc), and 1–9 for the nine CSE levels, Kappa values were calculated using the software SPSS. According to Landis & Koch (1977), a level of 0.21–0.40 for Kappa indicates "fair agreement", 0.41–0.6 "moderate agreement", 0.61–0.8 "substantial agreement", and 0.81 and above "strong agreement". These are the values referred to in the two hypotheses above.

Data and Frame of Reference

To recap, there are four sets of assessment data in the current study (see Appendix 1):

- 30 CET items expert rated against the nine CSE levels
- 23 LTE items expert rated against the six CEFR levels
- test taker mean scores on the 53-item LTE
- test taker mean scores on the 65-item CET

Since all four datasets were collected from the same test takers, the data configuration may be taken as a unified collection, in that all data are referenced to the same candidates and to their English language

ability. The *person links* (Boone, 2016) in the four datasets embrace a coherent *frame of reference* (FOR), defined by Humphry (2006) as "compris[ing] a class of persons responding to a class of items in a well-defined assessment context." While the expert judges only rated the discrete items in each test (23 in the LTE, and 30 in the CET), all items in both tests (53 LTE items, and 65 CET items) were included in the calibration. The reason for this is that the assessed locations—the assessed values—of the expert-judged items need to be expressed in the context of the whole FOR, that is within the FOR of the total number of items (118) in both CET and LTE tests.

In the one frame of reference calibration, scores for all four elements were converted to the same measurement scale—LID scale values, as laid out in Table 1 above.

Results

Since the current study involves rater judgement of item difficulty, a baseline in the analysis involves rater consistency concerning the judgements made. A summary of the analysis of the eight judges' consistency is presented below. As mentioned above, acceptable tolerance for fit is generally taken as ranging from 0.5 to 1.5 (Lunz & Stahl, 1990).

Table 2 reports the judges' rating of the CET items, and Table 3 their rating of the LTE items.

Reliability for both tests was high at 0.9, an important baseline. Fit statistics were generally good. With the experts judging the CET items, infit and outfit figures for all eight judges were good, indicating that they all fit the model. With the experts judging the LTE items, Judge 1's infit and outfit statistics were below the 0.5 threshold, indicating misfit. The general picture, however, was that rater consistency was good.

Equating Test Taker and Expert Judge Assessment Component Scores

Table 4 presents the results for test takers' means scores on the items ("assessed" items in Column 2) and for judges' mean ratings of item difficulties ("rated" items in Column 2). All assessment components, it should be noted, were anchored at 100—the mid-point of the LanguageCert scale, and the point at which all LanguageCert tests are anchored (Lee *et al.,* 2022).

Table 2 *CET Items Rated by Expert Judges*

Fair		Model	Infit	Outfit	Judges
Average	Measure	S.E.	MnSq	MnSq	
4.25	–2.44	0.29	0.64	0.67	1
5.37	0.05	0.27	1.14	1.13	3
5.84	1.03	0.26	0.87	0.89	4
4.28	–2.36	0.29	1.50	1.42	5
6.17	1.71	0.26	0.89	0.88	6
5.16	–0.38	0.27	0.67	0.65	7
6.53	2.40	0.27	1.20	1.17	8
5.37	0.00	0.27	0.99	0.97	Mean
0.89	1.89	0.01	0.31	0.28	SD Sample

RMSE .27 Adj (True) S.D. 1.87 Separation 6.89 Strata 9.52 Reliability .98 Fixed (all same) chi-square: 281.8 d.f.: 6 significance (probability): .00

Table 3 *LTE Items Rated by Expert Judges*

Fair-M		Model	Infit	Outfit	Judges
Average	Measure	S.E.	MnSq	MnSq	
4.01	1.46	0.27	0.33	0.34	1
4.84	0.16	0.26	0.77	0.78	2
5.86	–1.28	0.25	1.14	1.18	3
3.74	1.91	0.28	0.83	0.81	4
5.13	–0.29	0.25	1.27	1.33	5
6.41	–1.96	0.25	1.00	1.01	6
3.78	1.83	0.28	0.70	0.81	7
6.30	–1.83	0.25	1.36	1.23	8
5.01	0.00	0.26	0.92	0.94	Mean
1.10	1.61	0.01	0.34	0.32	SD Sample

RMSE .25 Adj (True) S.D. .94 Separation 3.72 Strata 5.30 Reliability .93 Fixed (all same) chi-square: 44.4 d.f.: 3 significance (probability): .00

Table 4 *Test Takers 'Mean Scores and Judges' Mean Ratings of Items*

Test	Assessment component	Items	Mean	SD	Reliability
LTE	Assessed items	23	102.96	32.10	0.97
LTE	Rated items	23	104.89	32.77	1.00
CET	Assessed items	30	104.28	22.43	1.00
CET	Rated items	30	96.25	20.22	0.84

As a baseline, reliabilities for all four elements of the dataset were high, above 0.8.

On the 23 LTE items, the "assessed mean" (the test taker mean score) was 102.96. The "rated mean" (the expert judge mean rating) was 104.89, a difference of 1.93.

On the 30 CET items, the "assessed mean" was 104.28. The "rated mean" was 96.25, a difference of 8.03.

While the standard deviations (SD) are broadly comparable within each assessment component pair, they differ considerably between tests. There is also a degree of difference between the mean scores. In light of this, the means and SDs need to be aligned into a single frame of reference (Humphry, 2006; Linacre, 2006). In Rasch terms, this means that the mean orientations of the scales (the zero logit points), and the logit widths (the SDs) need to be expressed in terms of similar values, and aligned to the LID scale.

For differences between means and SDs to be smoothed out, and for the four dataset components to be aligned, two parameters need to be applied where appropriate (Linacre, 2006). The two parameters relate to:

(1) the raw differences between the assessed means and the expert-rated item means (i.e., assessed means *minus* rated item means)

Table 5 *Assessed and Rated LTE and CET Item Differences*

1	2	3	4	5	6	7
Test / mode	**Items**	**N**	**LID values**	**Assessed *minus* Rated**	**SD**	**Assessed *divided by* Rated**
CET assessed	30	2,328	104.28	+8.03	22.43	1.11
CET rated	30	8	96.25		20.22	
LTE assessed	23	4,218	102.96	–1.93	32.10	0.98
LTE rated	23	8	104.89		32.77	

(2) the proportional differences between assessed and expert-rated standard deviations (i.e., assessed SDs *divided by* rated item SDs)

Table 5 below is an expansion of Table 4. It includes the results for parameter (1) [in Column 5] and for parameter (2) [in Column 7] for assessed and rated CET and LTE items.

Regarding parameter (2), if the proportional difference between the SDs for two scales is close to 1.0, the width of the two scales may be taken as being the same (Linacre, 2006). As Column 7 in Table 5 shows, both SDs are very close to 1.0 (1.11 for the CET and 0.98 for the LTE). Against this backdrop, no action needs to be taken for parameter (2). It is only parameter (1) which needs to be brought to bear.

In order to map the expert-rated CET items onto the assessed CET items, the mean for each item needs to be raised by 8.03 points (104.28–96.25), as in Column 5 of Table 5. Conversely, with the expert-rated LTE items, the mean should be lowered by 1.93 points (102.96–104.89).

Mapping Assessed and Rated Values to CEFR Levels

With the parameters above in place, it is now possible to map assessed and rated values to CEFR levels (or LID values). Refer back to Table 1 above.

CET items

Table 6 below presents the results—as LID scale values—for the CET items. For each item, Column 2 provides test taker scores as LID values, with Column 3 expressing these values as CEFR levels. Column 4 provides the original expert-rated score. Column 5 provides the adjustments (via parameter (1)) to the original scores, adding 8.03 to each. Column 6 expresses the Column 5 scores as CEFR levels. Column 7 presents the difference between Columns 6 and 3—the rater level compared with the assessed level.

Table 7 presents a summary of the fit between CET assessed levels and expert-judged levels.

As can be seen, 27/30 (90%) of the expert ratings of the CET items matched CEFR level values which emerged from test taker scores. Three items were under-rated by one level.

With CEFR levels recoded as 1–6 (A1 = 1 through to C2 = 6), Kappa was then calculated between assessed CEFR level scores (Column 3 in Table 6) and adjusted expert CEFR level ratings (Column 6 in Table 6). With the CET items, a Kappa of 0.92 ($p < .001$) emerged—a "strong" agreement between the two variables.

Table 6 *CET Items: Assessed and Expert Rating Values (Sorted by CEFR Level)*

[1]	[2]	[3] (from [2])	[4]	[5] ([4] + 8.03)	[6] (from [5])	7 ([6] vs [3])
CET item number	**Assessed score**	**Assessed score, as CEFR level**	**Original expert rating**	**Adjusted expert rating (+8.03)**	**Adjusted expert rating, as CEFR level**	**Rated level vs assessed level**
C329	62.5	A1	51.3	59.3	A1	=
C341	75.4	A2	64.2	72.2	A2	=
C343	77.8	A2	66.6	74.6	A2	=
C317	78.92	A2	67.72	75.72	A2	=
C334	79.66	A2	68.46	76.46	A2	=
C345	83.61	A2	72.41	80.41	A2	=
C339	96.54	B1	85.34	93.34	B1	=
C326	98.92	B1	87.72	95.72	B1	=
C331	103.44	B1	92.24	100.24	B1	=
C318	104.3	B1	93.1	101.1	B1	=
C344	104.58	B1	93.38	101.38	B1	=
C327	107.71	B1	96.51	104.51	B1	=
C335	108.19	B1	96.99	104.99	B1	=
C322	109.47	B1	98.27	106.27	B1	=
C328	110.8	B2	99.6	107.6	B1	–1
C321	111.33	B2	100.13	108.13	B1	–1
C342	115.88	B2	104.68	112.68	B2	=
C336	116.02	B2	104.82	112.82	B2	=
C320	118.71	B2	107.51	115.51	B2	=
C325	118.99	B2	107.79	115.79	B2	=
C338	120.28	B2	109.08	117.08	B2	=
C337	127.62	B2	116.42	124.42	B2	=
C333	128.64	B2	117.44	125.44	B2	=
C319	132.06	C1	120.86	128.86	B2	–1
C332	136.65	C1	125.45	133.45	C1	=
C330	141.77	C1	130.57	138.57	C1	=
C316	142.23	C1	131.03	139.03	C1	=
C323	144.93	C1	133.73	141.73	C1	=
C340	149.96	C1	138.76	146.76	C1	=
C324	158.32	C2	147.12	155.12	C2	=

LTE items

Table 8 now presents the picture for the LTE items. For each item, as before, Column 2 provides test taker scores as LID values, expressed as CEFR levels in Column 3. Column 4 provides the original expert-rated score. Column 5 provides the adjustments to the original scores, subtracting 1.93 from each. Column 6 expresses Column 5 scores as CEFR levels. Column 7 presents the difference between Columns 6 and 3—the rater level compared with the assessed level.

Table 7 *Fit of Assessed Levels and Expert-judged Levels: CET Items*

Fit	N = 30
Over-rated by one level	0
Exact fit	27 (90.0%)
Under-rated by one level	3 (10.0%)

Table 9 presents a summary of the fit between LTE assessed levels and expert-judged levels.

The expert ratings of the LTE items matched test takers' LTE scores much less closely than was the case with the CET items. Only 5/23 (21.7%) of the expert ratings on the LTE items matched test takers' scores on the LTE items. 18 items were under-rated by one level.

Table 8 *LTE Items: Assessed and Expert Rating Values (Sorted by CEFR Level)*

[1]	[2]	[3] (from [2])	[4]	[5] ([4] –1.93)	[6] (from [5])	7 ([6] vs [3])
LTE item number	**Assessed score**	**Assessed score, as CEFR level**	**Original expert rating**	**Adjusted expert rating (+8.03)**	**Adjusted expert rating, as CEFR level**	**Rated level vs assessed level**
L82	69.56	A1	55.75	53.82	A1	=
L75	80.19	A2	66.38	64.45	A1	–1
L76	84.96	A2	71.15	69.22	A1	–1
L74	94.71	B1	80.9	78.97	A2	–1
L77	93.96	B1	80.15	78.22	A2	–1
L80	105.23	B1	91.42	89.49	A2	–1
L81	109.89	B1	96.08	94.15	B1	=
L83	90.06	B1	76.25	74.32	A2	–1
L85	103.21	B1	89.4	87.47	A2	–1
L87	103.45	B1	89.64	87.71	A2	–1
L94	108.82	B1	95.01	93.08	B1	=
L73	115.85	B2	102.04	100.11	B1	–1
L78	110.55	B2	96.74	94.81	B1	–1
L84	110.55	B2	96.74	94.81	B1	–1
L89	111.51	B2	97.7	95.77	B1	–1
L90	115.97	B2	102.16	100.23	B1	–1
L91	124.62	B2	110.81	108.88	B1	–1
L92	114.29	B2	100.48	98.55	B1	–1
L93	116.52	B2	102.71	100.78	B1	–1
L95	114.82	B2	101.01	99.08	B1	–1
L79	146.13	C1	132.32	130.39	C1	=
L86	149.55	C1	135.74	133.81	C1	=
L88	132.33	C1	118.52	116.59	B2	–1

Table 9 *Fit of Assessed Levels and Expert-judged Levels: LTE Items*

Fit	N = 23
Over-rated by one level	–
Exact fit	5 (21.7%)
Under-rated by one level	18 (78.3%)

With assessed CEFR level scores (Column 3 in Table 8) and adjusted expert CEFR level ratings (Column 6 in Table 8) again recoded as before, a Kappa of 0.40 ($p < .001$) emerged between the two variables—a "fair" agreement.

Discussion and Conclusion

In the expert ratings of the eight judges, a much better correspondence was observed between raters judging CET item difficulty than was the case with the raters judging LTE item difficulty.

With the CET items, a 90.0% exact fit was recorded against the 30 CET items. In contrast, with the LTE items, the exact fit was considerably lower, at only 21.7%.

Discrepancies between items in both tests differed, it must be stated, by only one CEFR level: there were no instances of two-level discrepancies. This suggests that raters were generally within acceptable ranges, even if their ratings did not all exhibit an expert, i.e., exact, match. This was particularly the case with the LTE items. The judges were less familiar with the CEFR than they were with the CSE; their ratings were, nonetheless, within generally tolerable ranges with a discrepancy of one level in six being within the range of acceptability for marking consistency purposes (see Attali & Burstein, 2005).

The purpose of the current study has been to make a meaningful contribution to the discussion surrounding the viability of expert judgement. The current study has explored expert judgement from two perspectives, namely with a single set of expert judges who rated two sets of items. These experts were very well acquainted with one set of items (the CET items) and the scale (the CSE) against which to assess them. As language teaching and assessment professionals, they were familiar (although less so than with the CSE) with the other set of items (the LTE items) and the scale (the CEFR) against which the LTE items were to be assessed.

The study pursued two hypotheses.

Hypothesis 1 stated that, with the CET items, a "strong" level of agreement of 0.8 (or 64% 'shared variance') would be achieved between assessed and expert-rated values. Of the 30 CET items, a Kappa value of 0.92 emerged and 27, or 90.0%, of the items recorded an exact fit between the expert-rater scores and the test taker scores. Hypothesis 1 was therefore accepted.

Hypothesis 2 set out rather lower expectations: that, with the LTE items, only a "fair" agreement' of 0.4 (or 16% 'shared variance') would be achieved between assessed and expert-rated values. A Kappa value of 0.40 actually emerged, with an exact fit only recorded between five, or 21.7%, of the 23 LTE items. Hypothesis 2 was also accepted.

The conclusion that emerges from the current study is that judges who are very familiar with their own assessment situation in terms of test material, test constructs, assessment levels etc., are able to make more accurate assessments than are judges who are less familiar with the material they are assessing, and the levels at which test items should be assessed. While the current results might appear to be somewhat self-evident, the results lend support to the argument that, with adequate training and standardisation, and a strong background in the material to be judged, expert judgement is a methodology that may be reliably utilised in test validation.

The current expert judgement study forms part of a larger study, whose overarching purpose involves exploring potential alignment between the CEFR-based LanguageCert tests and the CSE in the context of the reading and usage components. What the current study reveals in the context of experts rating within their own assessment domains is that China experts may be reliably used to rate China CSE-linked test items, and experts who rate (and set) items within the context of the CEFR may be used to rate LanguageCert's CEFR-linked test items (Zhao & Coniam, 2022).

The current study has only involved one set of raters—Chinese raters rating CSE- and CEFR-linked test items. A parallel study currently being planned involves redoing the current study from the opposite perspective: that is, having UK-based expert setters of CEFR-linked items rate the two sets of items used in the current study.

Note

1. The College English Test (CET) is China's ESL test which examines the English proficiency of undergraduate and postgraduate students in China. It is intended to ensure that Chinese tertiary students reach English language levels specified in the National College English Teaching Requirements (see Mini, 2018).

The LanguageCert Test of English (LTE) is an English 'for work' exam intended for people over the age of 18 in or about to enter the workplace, as well as those in higher or further education. The level-agnostic qualification is offered in two paper-based versions measuring CEFR levels A1-B1 or A1-C2, and as an adaptive test measuring CEFR levels A1-C2 (see Coniam *et al.,* 2021).

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*(4), 535–556. https://doi.org/10.1177/0265532213489568

Attali, Y., & Burstein, J. (2005). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*(3).

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

Cheng, M.Y. (1993). *Testing and re-testing in Hong Kong F5 and F5 English Secondary Classes*. Unpublished M. Ed. Dissertation. Hong Kong: University of Hong Kong.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the impact of training in test development principles on improving test quality. *System*, *37*(2), 226–242. https://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*. London, UK: LanguageCert.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for Languages: Relating Language Examinations to the 'Common European Framework of Reference for Languages: Learning, Teaching, Assessment' (CEFR). A Manual*. Strasbourg: Council of Europe.

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. *Language Assessment Quarterly, 15*(1), 44–58. https://doi.org/10.1080/15434303.2017.1421955

Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students*. Hong Kong: Longman.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477–485. https://doi.org/10.1093/elt/ccs037

Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-based methods for judgementally estimating item difficulty parameters (LSAC Research Report 98-05).* Newtown, PA: Law School Admission Council.

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

Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China Standards of English: challenges at macropolitical and micropolitical levels. *Language Testing in Asia, 7*(1), 1–19. https://doi.org/10.1186/s40468-017-0032-5

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

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. https://doi.org/10.46451/ijts.2022.01.12

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

Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645–655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Liu, J. D. (2017). China's Standards of English and its applications in English learning. *China Foreign Languages, 14*(6), 4–11.

Liu, J. D. (2019). China's Standards of English Language Ability. *China Foreign Languages, 16*(3), 1+11–12.

Liu, J.D. (2021). Validating China's Standards of English Language Ability. *Modern Foreign Languages, 44*(1), 86–100.

Liu, X., & Read, J. (2021). Investigating the skills involved in reading test tasks through expert judgement and verbal protocol analysis: Convergence and divergence between the two methods. *Language Assessment Quarterly, 18*(4), 1–25. https://doi.org/10.1080/15434303.2021.1881964

Mehrens, W. A. (1995). *Methodological issues in standard setting for educational exams. In Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments* (Vol. 2, pp. 221–263). Washington DC: National Assessment Governing Board and National Center for Education Statistics.

Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement, 49*, 467–478. https://doi.org/10.1177/0013164489492020

Mini, G. (2018). *An Introduction to China's College English Test (CET). World Education News + Reviews*. Retrieved from https://wenr.wes.org/2018/08/an-introduction-to-chinas-college-english-test-cet

Ministry of Education of the People's Republic of China. (2018). *China's Standards of English Language Ability*. Beijing: Ministry of Education.

Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. Cambridge: Cambridge University Press.

Yang, H. Z., & Gui, S. C. (2007). On establishing a unified Asian level framework of English language proficiency. *Foreign Languages in China, 2*, 34–37.

Zhao, W., & Coniam, D. (2022). Using self-assessments to investigate comparability of the CEFR and CSE: An exploratory study using the LanguageCert Test of English. *International Journal of TESOL Studies, 4*(1), 169–186. https://doi.org/10.46451/ijts.2022.01.11

Appendix 1: CET and LTE Test Components

Section	CET	LTE
1	*Cloze:* 15 items One cloze passage Assessing grammar, syntax, discourse, vocabulary	*Cloze:* 15 items Three cloze passages Assessing grammar, syntax, discourse, vocabulary
2	*Discrete items:* 30 items Assessing grammar, syntax, vocabulary, usage	*Discrete items:* 23 items Assessing grammar, syntax, vocabulary, usage
3	*Reading comprehension:* 20 items Four reading comprehension passages, each with 5 items Assessing a range of reading comprehension skills	*Reading comprehension:* 15 items Three reading comprehension passages, each with 5 items Assessing a range of reading comprehension skills
	65 items	53 items

Author Biodata

David Coniam [corresponding author] is Head of Research at LanguageCert. He has been working and researching in English language teaching, education and assessment for almost 50 years. His main publication and research interests are in language assessment, language teaching methodology and academic writing and publishing.

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.

Wen Zhao is Dean of the School of Foreign Studies at Jinan University, Guangzhou. Her main publication and research interests are in corpus linguistics, English curriculum and instruction, and EFL writing. She has been working and researching in English language teaching and learning, and has been involved in national English curriculum development for senior secondary vocational education and College English education.