CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 5

CEFR and CSE comparability study: An exploration using the Chinese College English Test and the LanguageCert Test of English

David Coniam, LanguageCert
Michael Milanovic, LanguageCert
Wen Zhao, Jinan University, Guangzhou

https://doi.org/10.37546/JALTSIG.CEFR5-3

This paper outlines how different studies can be brought together to reveal how two separate examinations, based on different assessment frameworks, may be compared. The paper reports on data obtained from a cohort of a comparatively large sample of Chinese university test-takers who took two separate tests – the Chinese College English Test (CET), which is linked to the descriptive scales of the CSE (China Standards of English) and the LanguageCert Test of English (LTE) linked to the descriptive assessment framework of the CEFR (Common European Framework of Reference for Languages). In addition to the test-taking, two further studies were conducted.

Analysis indicated that it was possible to make reasonably definitive pronouncements about the comparability of the two frameworks in reading and language use. The findings contribute to the assessment research literature in that they provide relevant stakeholders with a means of comparing performances on either the LTE (linked to the CEFR scale) or the Chinese CET (linked to the CSE framework).These findings are particularly valuable for western institutions of higher education who, when considering the admission of Chinese students, postgraduate or other, are presented with CET results based on the CSE framework and LTE results based on the CEFR framework.

Keywords: test validation, reading and language use, CEFR, CSE, self-assessment, expert judgement, Rasch measurement

1 Introduction: Exploring comparability between the CEFR and the CSE

The overarching purpose of the research reported in the current paper was to explore how comparability of reading and language use tests such as the LanguageCert Test of English (LTE) and the Chinese College English Test (CET) could be established between the CEFR (the *Common European Framework of Reference for Languages*) and the CSE (the *China Standards of English*) assessment frameworks by using a combination of data from two tests.

The research – which comprised two separate studies – was undertaken because stakeholders needed to be able to compare results using the two assessment frameworks as growing numbers of Chinese students were beginning to be candidates for the LTE. Students in China who have taken the CET, based on the CSE, are nowadays taking the LTE, based on the CEFR, in much greater numbers. Institutions in Europe, including the UK which recognises the LTE for visa and migration purposes, will find the comparisons made in this article valuable in establishing the language proficiency of potential applicants from China, often seeking postgraduate courses, when they apply for admission to courses.

The first study used expert judgement to determine the 'fit' between the CET-type items and the CSE framework, with which the raters were familiar, and the LTE items and the CEFR framework, with which they were less familiar. While there was already existing evidence of the LTE items linkage with the CEFR, the study was expected to establish a reliable linkage between the CET-type items and the CSE framework. The second study was an exercise in which test-takers reported their self-assessments on each of the two tests. This study enabled for a comparison of the accuracy of the test-takers' self-assessments by matching them against their actual test results.

Data for the studies comprised a large cohort (N=2,500) of Year 1 students at a prestigious university in China. These students took two tests and also completed a set of Can Do self-assessments derived from both the CEFR and the CSE. In the first study (Zhao and Coniam 2022), expert judges were asked to place all the CET (College English Test) items on the CSE scale, with which they were very familiar. In the second study (Coniam et al. 2022), students self-assessed their English language ability against CEFR and CSE levels using Can Do statements. The intention of this study was to explore which framework levels students were best able to judge themselves against.

The two studies were originally published separately because there was too much to include in a single paper. The current paper pulls them both together, extending their individual reach, with an attempt to illustrate the issue of comparability from a larger perspective. Descriptions of both studies are consequently essential to provide readers with an adequate understanding of how the findings of the two studies were built upon in a further process of bringing together the analysis of a variety of data to provide a useful and valid picture of how the two scales compare.

2 Background to the CEFR and the CSE frameworks

For the past two decades, the CEFR has come to be accepted as illustrating standard descriptors of language ability by many stakeholders: e.g., 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 language activities. The CSE development attempts to pull together —in the context of China—a wide range of different English language curriculums and assessment instruments into one overarching framework. The development of the CSE began with the "Common Chinese Framework of Reference for English (CCFR-E): Teaching, Learning, Assessment", which began in 2014 (Jin et al. 2017). This development then became known as the *China Standards of English* (CSE) which was finally released in 2018 and consists of three major levels, each subdivided into three sublevels as illustrated in Figure 1.

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

Figure 1. CEFR and CSE Frameworks

As mentioned above, the test aligned to the CEFR in the current study is derived from the LanguageCert Test of English (LTE). The LTE is a 'level agnostic' test (one that is independent of levels and modes), existing in two parallel formats, both drawn from the same item bank: a paper-based version and an online adaptive test. The validation of the paper-based version of the LTE was reported in . (2021a), and the validation of the adaptive test in Coniam et al. (2021b).

The overarching LanguageCert Item Difficulty (LID) scale is aligned to the CEFR (Coniam 2021a) as illustrated in Table 1. The LID scale has been developed by LanguageCert. It is a validated measurement scale, a necessary prerequisite for any examination/assessment system.

Table 1. ***LID scale***

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

The next section reviews details of previous comparability studies.

3 Relevant comparability studies

One of the first large scale comparability studies was the Cambridge-TOEFL comparability study conducted in the late 1980s (Bachman et al. 1995). This study, which investigated the comparability of the First Certificate of English (FCE) and the Test of English as a Foreign Language (TOEFL), is notable for two reasons. Firstly, it established a baseline for comparability studies and secondly, it initiated comparability studies for different high-stakes tests.

Bachman et al. (1995) set the standard in terms of test taker samples, test types and scoring procedures selected for analysis. Against these robust background measures, conclusions could be drawn about comparability between the two tests: "score comparisons across tests are justified and could be made in a meaningful way" (Bachman 1990: 48).

There have since been numerous studies investigating and establishing comparability between tests. Some of these have been robust studies; some less so, merely claiming comparability. Some of these studies are discussed below.

3.1 Evidence-based equivalence-establishing studies

Detailed recommendations regarding procedures and methods for comparing tests are outlined in the Council of Europe manual (2009) relating language examinations to the CEFR document. According to the manual, establishing large-scale test comparisons requires a considerable amount of data, resources, and analysis, as discussed in the studies below.

Apart from the Bachman et al. study (1995), there have been other large-scale studies that have investigated the comparability of two different tests. Such studies broadly follow procedures described in Bachman et al. (1995); i.e., conducting an analysis of both test content and test results.

Taiwan's General English Proficiency Test (GEPT) assesses learner proficiency across the range of Taiwan's English education framework. A number of robust studies have been conducted, comparing the GEPT with the CEFR, with these studies, for the most part, adhering to the guidelines in the Council of Europe manual (2009), and using expert judgement panels. A brief commentary is provided below.

Brunfaut & Harding (2014) conducted a study investigating the comparability of the GEPT and CEFR listening tests. They concluded that GEPT listening test levels 1-4 largely corresponded to CEFR levels A2 to C1. Green & Inoue (2017) investigated the comparability of GEPT and CEFR speaking test levels. Their analysis indicated that the GEPT speaking tests were generally aligned well with CEFR levels. Knoch & Frost (2016) explored the alignment of GEPT writing tests to the CEFR. While results indicated that the GEPT writing tests aligned with CEFR levels, a slight lowering of GEPT pass scores was recommended in order to better align with CEFR levels.

In a data-driven comparability study, Kunnan & Carr (2017) explored the comparability of GEPT and Internet-Based Test of English as a Foreign Language reading and writing tasks via tests administered to test takers in Taiwan and the USA. They concluded that the two tests were broadly comparable. While the two tests generally assessed the same reading constructs, the reading focus was slightly different in each test.

3.2 Evidence-based CSE/CEFR comparability studies

Alderson (2017) discussed a range of studies exploring the CSE and its correspondence to the CEFR. These have been augmented by the work of Jin et al. (2017) and Zhao et al. (2017), investigating the linking of College English vocabulary levels with the CEFR.

Further studies were conducted, showing comparability between the CSE and CEFR. Dunlea et al. (2019) describe a comprehensive study involving all four language skills that explored the relationship between the British Council's Aptis test and IELTS with the CSE. The methodology involved expert judgement of items against CSE and CEFR levels and the assignment of CSE descriptors against tasks. Following this, the proposed levels were field tested in an "external evaluation" exercise, where Chinese teachers rated their own students against the proposed matched levels, as illustrated in Table 2 below.

Table 2. ***Level match between the CSE and the CEFR (Dunlea et al. 2019)***

CSE	CEFR
L9	C2
L8	C1
L6/7	B2

L4/5	B1
L3	A2
L2	A1

Peng et al. (2021) report on a study attempting to establish level correspondences between CEFR and CSE levels using difficulty estimates of all published descriptors (467 for the CEFR and 1,051 for the CSE) of ratings by English language teachers and students. While there was close correspondence at the top and bottom ends of the scale, there was overlap in the middle levels. Table 3 elaborates.

Table 3. ***Level match between the CSE and the CEFR (Peng et al. 2021)***

CSE	CEFR
L9	C2
L7/L8	C1
L6/L7	B2
L4/L5	B1
L2/L3	A2
L2	A1
L1	A0

As may be seen, while there is a good degree of agreement in the correspondence between the two studies that tested all four skills, there are also divergences. These may be due to a number of factors: the samples, the tests, the judges used in the ratings. These factors will, over time, be investigated by studies that focus on one factor at a time so that, in spite of a good degree of agreement, the divergences may also be explored.

3.3 Comparability studies with little supporting evidence

The discussion above has reported on studies claiming evidence of comparability between tests. Comparability has been claimed for many tests, usually with the CEFR. Often, such claims have been made – and are still made – on the basis of little or no apparent evidence – see Table 4 below.

The (now inaccessible) AWEMAP project from the early 2000's laid out numerous tables indicating apparent equivalence[1]. As Green (2012) comments, however, on certain of AWEMAP's equivalences, "convincing evidence of the relationship" for certain scales was simply not available (2012: 87). Such 'mappings' listed by AWEMAP's included Ordinate's Phone Pass, mapped to the CEFR Scale by Ordinate; DynEd Dynamic Education's Placement Test, mapped to the ILR OPI Scale and TOEIC by DynEd. The latter, mapping "exam correlations", can be found at https://www.dyned.com/media-library/correlations-intl/.

Current claims about comparability, usually with the CEFR scales – for which no evidence is provided —may be seen in the claims of numerous education bodies. Table 4 presents a small sample of specious claims.

1. AWEMAP was the Worldwide English LET EFL ESL EALLEP ESOL Assessment Scales and Tests Mapping Project. It was last available at http://www.geocities.com/esolscale/index.html?200510.

Table 4. ***Apparent mappings to the CEFR***

Education Body	URL
Express Publishing's Vocational English Certificate (VEC)	http://ecahe.eu/w/index.php/English_language_test_equivalency_table
EF Education First's Standard English Test (EF SET)	https://www.ef.com/wwen/english-tests/test-comparison/score-converter/

4 Overarching purpose: Exploring comparability between the CEFR and the CSE

As stated in the abstract, the overarching purpose of the research conducted in the two studies reported here involved exploring how comparability in reading and language use tests (such as the LanguageCert Test of English) might be established between the CEFR and the CSE.

4.1 Data

A variety of test taker/self-assessment/test/expert judgement data was collected from a Chinese university from late 2020 to May/June 2021. The principal focus was Year 1 CET students. Given that the university admits a considerable number of overseas Chinese students, it could be taken that there would be a considerable student ability spread – ranging across the CEFR and CSE. Table 5 presents a picture of the data collected.

Table 5. ***Project data***

Sample	Instrument	Timeframe
2,498 Year 1 CET students	65-item in-house CET Reading and Language Use placement test	Oct 2020
4,128 Year 1 CET students	53-item LTE Reading and Language Use test	Mar 2021
4,128 Year 1 CET students	16 CEFR self-assessment Can Do statements	May 2021
4,128 Year 1 CET students	22 CSE self-assessment Can Do statements	May 2021
8 ESL university professors	Expert judgement of 23 discrete-point LTE items	May 2021
8 ESL university professors	Expert judgement of 30 discrete-point CET items	Jun 2021
2,311 Year 1 students	Official CET 4 test	Oct 2021

In late 2020, approximately 2,500 Year 1 CET students took a 65-item multiple-choice reading and language use test prepared by experts from the university. Approximately three months later, this same set of students took a 53-item multiple-choice LanguageCert reading and language use test constructed from the LTE item bank. The LTE items were selected on the basis of having been calibrated to represent the spectrum of difficulty across the six CEFR levels. The items had been adapted from the material validated in Coniam et al. (2021a). The composition of the tests is described in the following section.

4.2 Content analysis of the tests

This section presents a comparative analysis of the make-up of the two reading and language use tests. Table 6 elaborates.

Table 6. ***Component Analysis of CET and LTE Tests***

CET	LTE
Cloze: 15 items One 15-item cloze passage assessing grammar, syntax, discourse, vocabulary	Cloze: 30 items Three 5-item cloze passages assessing grammar, syntax, discourse, vocabulary
Discrete items: 30 items 30 items assessing grammar, syntax, vocabulary, language use	*Discrete items: 23* items 23 items assessing grammar, syntax, vocabulary, language use
Reading comprehension: 20 items Four 5-item reading comprehension passages assessing a range of reading comprehension skills	Reading comprehension: 15 items Three 5-item reading comprehension passages assessing a range of reading comprehension skills
65 items	53 items

As Table 6 illustrates, the CET test is slightly longer than the LTE test; also, all CET items were four-option multiple-choice whereas the LTE items were three-option multiple-choice. Despite these minor differences, the content of the two tests, and even the order in which the different sections of the test were presented to test takers, exhibit a broad amount of similarity.

4.3 Test administration and expert judge study

Test takers took the CET 65-item test in late 2020 as part of their university course whereas the official CET is usually taken at the end of the academic year.

In 2021, the same group of students took the LanguageCert 53-item test. Test takers subsequently completed two Can Do self-assessment profiles. One profile consisted of 16 Can Do statements drawn from the CEFR and the second 22 statements drawn from the CSE. The composite set of 38 items were all presented bilingually in both English and Chinese, with CEFR and CSE items and levels intermingled in an attempt to reduce the chances of respondents trying to guess where their own estimated ability level finished.

The focus for the expert judge study was the discrete items in the two tests. There were 30 such items in the CET test and 23 items in the LTE test, testing grammar, syntax, vocabulary and language use: the second component of the test presented in Table 6 above. There were eight expert judges, professors from the Foreign Studies Department, all of whom had been involved in setting CET items for their students at the university.

Before rating took place, training and standardisation sessions were conducted for the expert raters participating in the study. The purpose of these sessions was to increase rater reliability and familiarity with the less known CEFR framework. First, they rated sample CET items using the nine-level CSE. They then rated sample CEFR items using the six-level CEFR. Following this, the expert raters were given the 30 CET items to rate against the nine CSE levels and the 23 LTE items against the six standard CEFR Levels.

5 Statistical analysis

In the current study – to gauge test fitness for purpose, and to link two different tests to a common scale —both Classical Test Statistics (CTS) and Rasch measurement have been used and are briefly outlined below. CTS analysis reports test mean, standard deviation and test reliability. Rasch measurement facilitates the calibration of different facets within and between tests.

5.1 Classical Test Statistics (CTS)

The test mean for a proficiency test tends to be within a range of 60-70% (Heaton 1990)[2]. This will depend, however, on where the pass mark is set by the exam body concerned, and the purpose for which the test is intended. A test mean of around 60-70% suggests that the test is generally appropriate to the level of a 'typical' test taker (Burton et al. 1991). Such a mean in general indicates that most test takers managed to finish the test and that test takers may be assumed to have done their best.

In terms of test reliability – where levels of reliability are associated with test length (Ebel 1965) – a desirable level is generally taken as 0.7 for tests with 65 or more items.

5.2 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 often likened to 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 difficulties also estimated independently from the sample (Bond et al. 2020).

Since test taker measures and item difficulties are placed on an ordered continuum in Rasch, 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 taking different tests; and were used in the current study. Once a test, or scale, has been calibrated, the established values can be used to equate different test forms.

In the current study, the LTE test has been compiled, as mentioned, from robust test material validated in Coniam et al. (2021a).

5.3 CTS analyses

CTS item analyses are presented in Table 7. Results for the whole test are presented first for all test takers. Second, since expert judgement of difficulty was judged against the sets of discrete items from both tests, an analysis of the exact-same set of test takers is also presented for purposes of direct comparison.

A 'good' item is defined by Falvey et al. (1994) as one with a facility index of 30%-80%. A 'reliable' item is defined by Falvey et al. (1994) is one with a discrimination index greater than 0.3.

2. Heaton (1990) states, in the case of means, that with proficiency tests, test scores should ideally be spread out "over the whole range of the scale" (p. 171) (i.e., with a mid point of around 0.6), which aids in discriminating among test takers. The corollary is that proficiency tests will have a slightly lower mean than classroom tests. This is reflected in, for example, the mean performance score on key proficiency tests. To exemplify, for 2019, the overall mean on IELTS Academic was 6.08/9 and IELTS General Training 6.59/9. Cf., https://www.ielts.org/for-researchers/test-statistics/test-taker-performance.

Table 7. *Item analyses*

	CET Whole test	LTE Whole test
Test takers	2,498	4,128
Items	65	53
Mean	57.31 (60.3%)	28.31 (53.4%)
SD	12.1 (12.7%)	6.6 (12.5%)
Reliability KR20	0.83	0.72
Good items	39/53 (74%)	51/65 (78%)
	CET Discrete items	LTE Discrete items
Test takers	2,318	2,492
Items	30	23
Mean	15.71 (52.3%)	13.15 (57.2%)
SD	3.9 (13.1%)	3.38 (14.7%)
Reliability KR20	0.63	0.62
Good items	21/30 (70%)	20/23 (87%)

As Table 7 indicates, whole test analyses were broadly comparable. Standard deviations and reliability (as measured by the Kuder-Richardson KR20 statistic) were very close on both tests. Both test means were in the 'desirable' range – in the 50-60% range, suggesting that the tests broadly fit the target population, and that most test takers finished the test and had given it their best shot. Test reliability for both tests was close to or above 0.8, indicating that the tests may be assumed to have been well constructed.

The more focused picture with the discrete items showed an even closer match between the two tests, suggesting that comparisons may be seen as generally valid.

While CTS gives a baseline indication of comparability, if tests are to be linked so that they may be directly compared, Rasch measurement needs to be applied, because only then can analysis be carried out to determine comparability between two non-linked and separate tests.

5.4 Data and frame of reference

To recap, there are four sets of assessment data in the current study: the 65-item CET test, the 53-item LTE test, 22 CSE-referenced Can Do ratings and 16 CEFR-referenced Can Do ratings. 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) (Humphry 2006)[3].

In order to calibrate the four datasets in the current study onto the LanguageCert Item Difficulty (LID) scale (Table 1), a previously calibrated test (henceforth referred to as the "anchor test") from the Coniam et al. (2021a) study was incorporated into the data. As a subset of the anchor test, the LTE test in the current study provides a set of *item links* (Boone 2016). With sets of both person links and item links established, the LTE test could then be linked to the anchor test. Following this, the other datasets in the study – the CET test and the two sets of self-assessments – could then be calibrated against the anchor

3. 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." In a given frame of reference, a number of disparate datasets may then be calibrated together and aligned to each other.

test onto the LID scale. This resulted in all five assessment datasets being included into one single FOR.

For analysis and calibration purposes, 100 has been taken as the mid-point of the scale (see Table 1 above). To this end, Rasch logit values are rescaled to a mean of 100 and a standard deviation of 20.

5.5 Single frame of reference analysis

As mentioned, the *anchor test* had been previously anchored to the LID scale. Against this backdrop, the composite analysis is presented in Appendix 1.

To recap, item links in the overall dataset were established between the 53 items in the LTE test and the anchor test. Person links were established via the two tests and the two sets of self-assessments. All five datasets may therefore be seen to be within an overall frame of reference – the composite analysis to the far left of the person-item map in Appendix 1. In the analysis of the data, the two tests fit the Rasch model well, with mean square infit and outfit figures within the 0.5 to 1.5 range, and high reliability. The means for both tests were very comparable, both approximately a quarter of a logit above the overall mean of 100. The two tests emerged as being of comparable difficulty, if slightly more demanding than the mean calibration point. In contrast, respondents tended to slightly overestimate their abilities on both the CEFR and the CSE, with self-assessment mean values slightly higher than their actual results indicated.

5.6 External test reference point

Official CET-4 scores were obtained in November 2021 for 2,311 of the test takers. Table 8 presents the Pearson correlations between the official CET-4 test for reading, the LTE reading and language use (RLU) test and the China university CET reading and language use test used in the current study.

Table 8. ***Pearson correlations in CET tests***

Test	Correlation detail	Official CET-4 reading test	CET RLU test
LTE RLU test	r	0.71	0.73
	p level	< .001	< .001
CET RLU test	r	0.74	
	p level	< .001	

As can be seen, the tests inter-correlate significantly at the 0.7 level. The highest correlation, as might be expected, is between the official CET-4 reading test and the in-house CET RLU test at 0.74. Given that the accepted correlation for tests of 65 items is around 0.7 (Ebel 1965), the inter-test correlations in the current study are an indication that the tests are broadly assessing similar constructs.

In sum, it may therefore be seen that the two tests are similar in construction. Test means, standard deviations and reliability figures were broadly comparable at two levels: at the whole test level and on the reading and language use subtests. The two cloze subtests correlate at the 0.7 level, with the shared variance overlap indicating that the two cloze subtests are potentially measuring approximately 50% of the same construct.

6 The two studies

As mentioned, two studies were undertaken in addition to the large-scale testing of candidates on a CET-type and LTE-type test. These will now be further discussed as a basis for subsequently triangulating the results of those two formal tests. Without a process of triangulation, it is difficult to accurately determine comparisons between the two separate tests.

The first study involved a set of China English language professors in expert rating. The professors first rated the CET reading and language use items against the CSE scales; second, they rated the LTE reading and language use items against the CEFR levels.

The second study involved a series of self-assessment Can Do statements describing English language competences. Test takers who took the two tests later self-evaluated using the Can Do statements after finishing the two tests. One set of Can Do statements generated self-assessments based on the CSE levels, while the other set generated self-assessments based on the CEFR levels. The use of instruments such as Can Do statements in self-assessment has been validated in a number of studies (see e.g., Brown et al. 2014; Summers et al. 2019). These two datasets were analysed along with the two tests.

In order to assist the reader, key issues in and outcomes from the two studies are reproduced in the following section.

7 Study one: Expert judge ratings

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. The research question pursued in the study was:

(RQ1) To what extent are expert judges more accurate in their judgement of item difficulty when rating test items against a framework with which they are very familiar, as opposed to rating test items against a framework with which they are less familiar.

Against this backdrop, a high level of agreement between student test scores and expert-rated values was hypothesised for the CET items (i.e., a 'strong agreement' [0.8] in Kappa statistic terms (Landis and Koch 1977). Conversely, with the LTE items, only a moderate level of agreement ('substantial agreement' [0.6] in Kappa terms) was hypothesised between student test scores and expert-rated values.

The study involved eight expert judges, professors from the Foreign Studies Department, all of whom had set CET items for their students at the university. Given the relevance and status of the CSE in China, the eight expert judges had a clear picture of standards in the CSE, confirmed by senior staff at the university. Given the fact that they were all English language professionals, most had also knowledge of, albeit not in-depth familiarity with, the CEFR.

In standardisation sessions, the eight judges trial-rated sample CET items against the nine CSE levels, and sample CEFR items against the six CEFR levels. The judges then rated the 30 discrete CET items against the CSE and the 23 discrete LTE items against the CEFR.

Following the training and standardisation, test taker mean scores and expert judge mean ratings of the discrete items for each cloze subtest were equated by aligning the differences between the means and standard deviations of both sets of scores. Table 9 presents the findings which emerged following analysis of test takers' scores on the tests and from judges' ratings of item difficulty. Both tests, as mentioned, were anchored at 100 – the mid-point of the LanguageCert scale at which all LanguageCert tests are anchored (see Lee et al. 2022).

Table 9. ***Test taker mean scores and expert judge mean ratings of CET and LTE items***

Subtest type and mean	Items	Mean LID value	SD	Reliability
LTE test taker mean	23	102.96	32.10	0.97
LTE expert mean rating	23	104.89	32.77	1.00
CET test taker mean	30	104.28	22.43	1.00
CET expert mean rating	30	96.25	20.22	0.84

As a baseline, verification of reliability and Rasch fit statistics were first conducted. Reliabilities for all four elements of the dataset were high, above 0.8. Rasch infit and outfit figures were within acceptable levels (0.5-1.5).

With the LTE 23 discrete LTE items, the test taker mean score was 102.96; the expert judge mean rating was 104.89, a difference of 1.93.

With the 30 discrete CET items, the test taker mean score was 104.28; the expert judge mean rating was 96.25, a difference of 8.03. Differences between test taker mean scores and expert judge mean ratings were less than 10 points, the half-logit difference generally accepted as non-significant (Zwick et al. 1999).

Standard deviations (SD) were broadly comparable within each pair of tests. However, as was the case with the mean scores, the SDs differed between tests. To smooth out these differences, means and SDs needed to be aligned into a single frame of reference (see Linacre 2022). Following the 'smoothing-out', or alignment, directly comparable values were then able to be computed for each LTE and CET item by subtracting the expert judge rated item mean from the test taker item mean.

With all LTE and CET items in the same frame of reference, expert judge-rated and test taker mean item values could finally be mapped to CEFR levels. On both tests, for each item, the CEFR/CSE level match between the expert judge mapped level and the test taker score level was examined. A tally was made of whether the match was exact, or whether the difference was lenient or strict by one or more CEFR levels. The results are presented in Table 10.

Table 10. ***Fit of expert judge mapped levels to test taker scores levels***

	CET items	LTE items
Number of items	N=30	N=23
Over-rated by one level	0	0
Exact fit	27 (90.0%)	5 (21.7%)
Under-rated by one level	3 (10.0%)	18 (78.3%)
Kappa	0.92 (p<.001)	0.40 (p<.001)

With the CET items, 27/30 (90%) of the expert ratings matched the test taker mean score values; three items were under-rated by one level. With the LTE items, however, expert ratings matched test takers' 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. This indicates that the expert judges do not have as good an understanding of the CEFR as they do the CSE levels, hence the under-rating of many items against the CEFR levels. The implications that can be drawn from these findings are that, in future studies and practices of comparability, additional standardisation and training should be given to raters in the frameworks with which they are less familiar.

After recording expert judge-mapped levels against test taker mean score levels (as 1-6, where A1=1 and C2=6), Kappa was calculated, with the results presented in the final row of Table 10. With the CET items, a Kappa of 0.92 (p<.001) emerged – a 'strong' agreement between the two variables. With the LTE items, a Kappa of 0.40 (p<.001) emerged between the two variables – only a 'fair' agreement.

The conclusion drawn from the expert rating study was 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 results in the expert rating study might appear to be somewhat self-evident, the results lend support to the argument that expert judgement is a methodology that may be reliably utilised in test validation provided the raters are completely familiar

with and experts in that specific test. This lends credence to the fact that the assessments in the current study may be taken as reliable. One implication is that raters who are less familiar with one of the tests require further standardisation and training.

8 Study two: Test taker self-assessment study

Against the backdrop outlined above, the second published study, an essential component of the triangulation that eventually took place, pursued two questions.

(RQ1) To what extent can self-assessments be validly used to establish correspondences between the CEFR and CSE frameworks?

(RQ2) To what extent are correspondences between the CEFR and CSE frameworks in line with those reported in previous studies?

Subsequent to having taken both tests, test takers responded to 38 Yes/No-framed Can Do statements. 22 Can Do statements related to the CSE and 16 Can Do statements to the CEFR. The results were analysed in the same frame of reference as the two tests, where the anchor point was 100, the LID scale mean. Table 11 presents the results for the test means and the self-assessment means. Appendix 1 presents the big picture.

Table 11. ***Test taker test self-assessment mean scores***

Subtest type and mean	Items	Mean LID value
LTE test mean	53	105.33
CET test mean	65	104.10
CEFR Can Do assessment mean	16	95.29
CSE Can Do assessment mean	22	95.66

As can be seen, the means of both two sets of self-assessments are comparable. The fact that both sets of self-assessments are five LID scale points (a quarter of a logit) below the anchored mean of 100 is indicative of test takers tending to slightly over-rate themselves – a not uncommon phenomenon (Dunning 2006).

The fact that means for both tests and self-assessment ratings were acceptably within half a logit (Zwick et al. 1999), suggested that test takers could be considered sufficiently objective in their self-assessments to permit tentative correspondences to be drawn between CSE and CEFR levels. Correspondences were then drawn up between the two sets of self-assessments for each CEFR level. To exemplify, Table 12 below presents the CEFR and CSE Can Do statements for the LanguageCert B2 level (111-130 LID scale points).

Table 12. ***CEFR and CSE Can Do statement level comparison chart: B2 (111-130)***

CEFR			CSE		
CEFR Can Do Statements	*LID value*	*CEFR level*	*CSE level*	*LID value*	CSE Can Do Statements
			L7	129.72	I can understand linguistically complex novels and materials related to culture and appraise their linguistic features.

CEFR			CSE		
			L6	128.73	I can understand the terminology of operational texts in related professional areas.
			L7	127.85	I can understand book reviews in relevant fields of inquiry.
			L6	127.27	I can understand novels and argumentative texts comprised of relatively complex language.
I can scan through rather complex texts, e.g. articles and reports, and can identify key passages.	118.74	B2			
			L5	117.63	I can understand the common figures of speech in stories pertaining to social life written in relatively complex language.
I can understand in detail specifications, instruction manuals, or reports written for my own field of work	116.58	B2			
			L5	116.41	I can infer the content of an entire book or text by scanning the table of contents.
I can read texts dealing with topics of general interest, such as current affairs, without a dictionary, and can understand multiple points of view.	115.69	B2			

Within the B2 CEFR LID range of 111-130, three CEFR C1 self-assessments were found, along with six CSE self-assessments, of which two were at L5, two at L6 and two at L7.

The B2 CEFR / CSE fit was therefore interpreted as being CEFR level B2 fitting quite broadly against CSE levels L5-L7.

Conclusions regarding RQ1 were that, while respondents tended to slightly overestimate their abilities on both the CEFR and the CSE, such overestimations were minimal, in that mean values were only a quarter of a logit higher than might have been expected. Overestimations were also consistent with the scales for both frameworks. The premise that self-assessments could be used to establish correspondences between the CEFR and CSE frameworks was then accepted.

Regarding RQ2, correspondences which emerged between the CEFR and CSE frameworks were broadly in accordance with those proposed by previous studies. While there were some divergences,

more notably towards the lower end of the scales, the correspondences proposed broadly echo those reported in previous studies.

The results reveal that – with the exception of CEFR level C2, for which there was insufficient data to perform a calibration, it was possible to produce an overall tentative mapping of how the CEFR scale, as represented by the LTE, might be mapped against the CSE scale as represented by the CET-based assessment. Table 13 presents the fit.

Table 13. ***Current study CEFR / CSE fit***

CEFR	CSE
C2	N/A
C1	L7
B2	L5-L7
B1	L4-L6
A2	L3-L5
A1	L2-L3
A0	L1

As can be seen from Table 13, while there is not a one-to-one match between the levels in the two frameworks, as one moves up the scale, there is a graduated fit between the CEFR and the CSE.

Table 14 below extends the picture of alignments presented in Table 13, and includes the alignments proposed in Dunlea et al. (2019) [the 'Dunlea' study] and in Peng et al. (2021) [the 'Peng' study].

Table 14. ***Extended CEFR / CSE mapping***

Current study		*Dunlea study*		*Peng study*	
Reading & Language Use		All skills		All skills	
CSE	*CEFR*	CSE	CEFR	*CSE*	*CEFR*
	C2	L9	C2	L9	C2
L7	C1	L8	C1	L7-L8	C1
L5-L7	B2	L6-L7	B2	L6-L7	B2
L4-L6	B1	L4-L5	B1	L4-L5	B1
L3-L5	A2	L3	A2	L2-L3	A2
L2-L3	A1	L2	A1	L2	A1
L1	A0	L1		L1	A0

The different mappings revealed both similarities and differences. For readability sake, these are listed below.

- The current study mapped A0 onto L1, as did the Peng study.
- The current study mapped A1 against L2/L3. The Dunlea study mapped A1 to L2, and the Peng study mapped A1 to L2.
- The current study mapped A2 more broadly against L3-L5. The Dunlea study mapped A2 to L3 while the Peng study mapped A2 against L2/L3.
- The current study mapped B1 against L4/L6. The Dunlea and Peng studies mapped B1 against L4/L5.

- The current study mapped B2 against the bottom end of L5 to L7. The Dunlea study mapped B2 against L6/L7 and the Peng study mapped B2 against L6/L7.
- The current study mapped C1 at L7. The Dunlea study mapped C1 at L8 while the Peng study mapped C1 at L7/L8.
- There was no data for C2 in the current study.
- The results of the current study can therefore be seen to broadly reflect the mappings of the previous Dunlea et al. (2019) and Peng et al. (2021) studies. As mentioned above and shown in Table 14, both Dunlea et al (2019) and Peng et al 2021) recently published the results of comparative studies. These studies mapped the results of all four skills at appropriate levels within the CEFR and CSE frameworks. They are useful because they reveal both similarities and differences in the findings. The contribution of the current study is that it too has mapped results against the CSE and CEFR frameworks. These findings broadly reflect the mappings of both Dunlea et (2019) and Peng et al (2021) so we have a broader perspective of these comparison studies.

9 Conclusion

The paper has investigated comparability between the CEFR and the CSE. Comparisons have been explored from the perspective of tests of reading and language use, as produced by LanguageCert for the CEFR and a comparable test of reading and language use from a China CET (College English Test) produced by a university in China.

Datasets in the study comprised a large cohort (N=2,500) of Year 1 students at a prestigious university in China. These students took two tests of reading and language use: one, initiated, developed, trialled and standardised by ESL professors within the university and the other derived from the LanguageCert Test of English. Analyses of both whole tests and reading and language use sections were seen to be broadly comparable. Classical Test Statistics indicated that both tests were reliable, with comparable means and standard deviations. A Rasch analysis showed that the two tests fit the Rasch model well, with acceptable mean square infit and outfit figures, and high reliability figures. Inter-test correlations between the two tests and the official CET-4 test results emerged in the 0.7 range. Such a level of correlation is considered reasonable, indicating that it is possible to come to conclusions about the amount of comparability between the two different tests.

Against this backdrop of two comparable and reliable tests, two different studies were undertaken, to investigate CEFR/CSE comparability issues.

In the first study, expert judges rated CET items against the CSE, with which they were very familiar, and LTE items against the CEFR, with which they were not as familiar. A high level of agreement between test taker mean score values and expert-judge-rated values on the CET discrete items was obtained. This suggested that values obtained between the cross-calibration of the CET and LTE tests could be seen as robust, and could support the other analyses conducted.

In the second study, test takers completed sets of Can Do self-assessments related to CEFR and CSE scale levels. Test takers tended to slightly overestimate their abilities, and although there was not a one-to-one match between the levels in the two frameworks, correspondences between the CEFR and CSE frameworks were nonetheless broadly in accordance with those proposed by previous studies (Dunlea et al. 2019; Peng et al. 2021).

The implications for stakeholders (students, teachers, administrators and universities) are that broad comparisons may be seen between CSE-aligned tests and CEFR-aligned tests. The mapping for reading and language use between the CEFR and the CSE as it emerged in the current study was reported in Table 13 above.

As reported earlier, as one moves up the scale, there is a graduated fit between the CEFR and the CSE. While there are some divergences, more notably towards the lower end of the scale, the correspondences

broadly echo those reported in previous studies. This study therefore provides a useful first stage in the comparison between the LanguageCert Item Difficulty scale and, hence, LanguageCert tests, and the CET-type tests linked to the China Standards of English. It also contributes to the research literature on how comparability between separate non-linked tests can be investigated and established. In summary, both a theoretical and practical contribution have been made by this study: theoretically, the findings add to the knowledge provided by different comparison studies of these two important assessment frameworks and the exams that have been centred within them; practically, students in China who take both the CET and the LTE can use those results to inform and enlighten admission tutors in both the UK and the rest of Europe.

10 References

Alderson, J. Charles. 2017. Foreword to the special issue "the common European framework of reference for languages (CEFR) for English language assessment in China" of language testing in Asia. Language Testing in Asia, 7(1), 1-9.

Bachman, Lyle. F. 1990. A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL test batteries. *Issues in Applied Linguistics 1*(1). 30-54.

Bachman, Lyle F., Fred Davidson, Katherine Ryan & Inn-Chull Choi. 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, Trevor, Zi Yan & Moritz Heene. 2020. *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9780429030499.

Boone, William. J. 2016. Rasch analysis for instrument development: why, when, and how? *CBE-Life Sciences Education* 15(4), rm4. https://doi.org/10.1187/cbe.16-04-0148.

Brown, N. Anthony, Dan P. Dewey & Troy L. Cox. 2014. Assessing the validity of can□do statements in retrospective (then-now) self-assessment. *Foreign Language Annals 47*(2). 261-285. https://doi.org/10.1111/flan.12082.

Brunfaut, Tineke & Luke Harding. 2014. *Linking the GEPT listening test to the Common European Framework of Reference*. Taipei, Taiwan: Language Training and Testing Center.

Burton, Steven J., Richard R. Sudweeks, Paul F. Merrill & Bud Wood. 1991. *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services and The Department of Instructional Science: Utah. https://testing.byu.edu/handbooks/betteritems.pdf.

Choi, Inn-Chull & Youngsun Moon. 2018. A comparability study of two standardized English as a foreign language tests. (Language Research. https://doi.org/10.30961/lr.2018.54.2.277).

Choi, Inn-Chull. 1995. A comparability study on SNUCREPT and TOEIC. (Language Research).

Coniam, David, Tony Lee, Michael Milanovic & Nigel Pike. 2021a. Validating the LanguageCert Test of English scale: The paper-based tests. London, UK: LanguageCert.

Coniam, David, Tony Lee, Michael Milanovic & Nigel Pike. 2021b. Validating the LanguageCert Test of English scale: The adaptive test. London, UK: LanguageCert.

Coniam, David, Wen Zhao, Tony Lee, Michael Milanovic & Nigel Pike. 2022. The role of expert judgement in language test validation. *Language Education and Assessment, 5*(1), 18–33.

Council of Europe. 2009. *Relating language examinations to the Common European Framework of References for Languages: Learning teaching, assessment*. Strasbourg: Language Policy Division.

Deygers, Bart, Koen Van Gorp & Thomas Demeester. 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.

Dunlea, Jamie, Richard Spiby, Sha Wu, Jie Zhang & Mengmen Cheng. 2019. *China's standards of English language ability: Linking UK exams to the CSE*. British Council. https://www.britishcouncil.org/sites/default/files/linking_cse_to_uk_exams_5_0.pdf.

Dunning, David. (2006). Strangers to ourselves. The Psychologist, 19(10), 600-603.

Dyned. 2022. Exam correlations. https://www.dyned.com/media-library/correlations-intl/.

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

Falvey, Peter, Jack Holbrook & David Coniam. 1994. *Assessing students*. Hong Kong: Longman.

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

Green, Anthony & Chihiro Inoue. 2017. *Relating the GEPT Speaking Tests to the CEFR*. Taipei, Taiwan: Language Training and Testing Center.

Green, Anthony. 2012. *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range (Vol. 2).* Cambridge: Cambridge University Press.

Gu, Mini. 2018. An introduction to China's college English test (CET). *World Education News and Reviews*. https://wenr.wes.org/2018/08/an-introduction-to-chinas-college-english-test-cet.

Heaton, J. B. 1990. Writing English language tests. Harlow, UK: Longman.

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

Jin, Yan, Zunmin Wu, Charles Alderson & Weiwei Song. 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.

Knoch, Ute & Kellie Frost. 2016. *Linking the GEPT Writing Sub-test to the Common European Framework of Reference.* Taipei, Taiwan: Language Training and Testing Center.

Kunnan, Antony John & Nathan Carr. 2017. A comparability study between the General English Proficiency Test-Advanced and the Internet-Based Test of English as a Foreign Language. *Language Testing in Asia 7*(1). 1-16. https://doi.org/10.1186/s40468-017-0048-x.

Landis, J. Richard & Koch, Gary G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 159-174. https://doi.org/10.2307/2529310.

Lee, Tony, Michael Milanovic & Nigel Pike. 2022. Equating Rasch values and expert judgement through externally-referenced anchoring. *International Journal of TESOL Studies* 4(1). 96-112. https://doi.org/10.46451/ijts.2022.01.12.

Linacre, John M. 2012. *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Little, David. 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.

Peng, Chuan, Jianda Liu & Hongwen Cai. 2021. Aligning China's Standards of English language ability with the Common European Framework of Reference for Languages. Asia-Pacific Education Researcher. Springer. https://doi.org/10.1007/s40299-021-00617-2

Summers, Maria M, Troy L. Cox, Benjamin L. McMurry & Dan P. Dewey. 2019. Investigating the use of the ACTFL Can Do statements in a self-assessment for student placement in an Intensive English Program. *System 80.* 269-287. https://doi.org/10.1016/j.system.2018.12.012.

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

Zhao, Wen, Boran Wang, David Coniam, & BIngxue Xie. (2017). Calibrating the CEFR against the China Standards of English for College English vocabulary education in China. *Language Testing in Asia* 7(1), 1-18. https://doi.org/10.1186/s40468-017-0036-1.

Zhao, Wen & David Coniam. (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. doi.org/10.46451/ijts.2022.01.11.

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.

11 Biographies

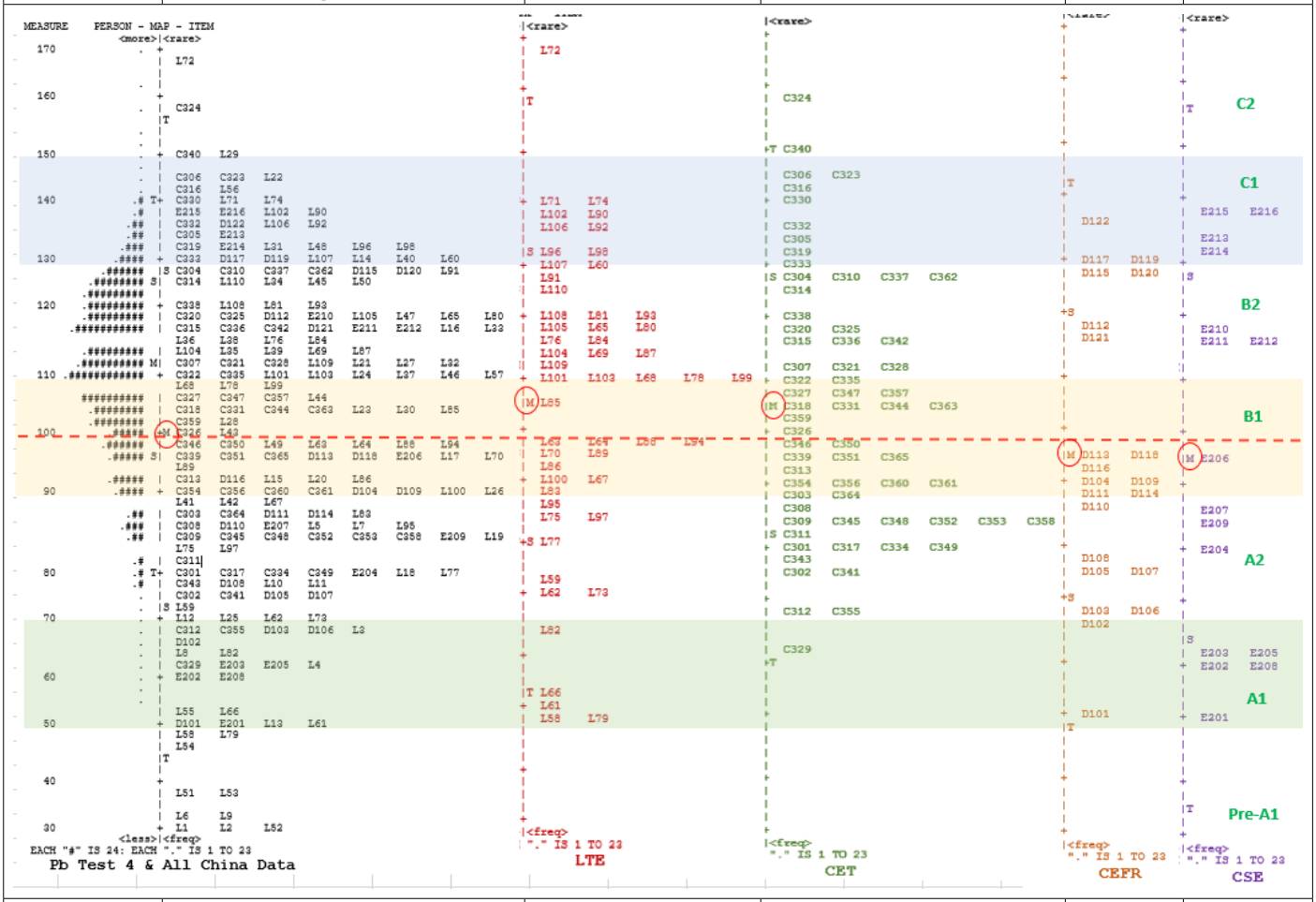
David Coniam [corresponding author] is Head of Research at LanguageCert. He has been working and researching in English language teaching, education and assessment for almost 50 years. His main publication and research interests are in language assessment, language teaching methodology and academic writing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

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.

Wen Zhao is Dean of the School of Foreign Studies at Jinan University, Guangzhou. Her main publications 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.

Appendix 1: Composite analysis of diverse assessment elements

The figure below should be read as follows: Column 2 contains the analysis of the amalgamated five datasets of 158 items plus the anchor items. Column 3 contains the 53-item LTE test, Column 4 the 65-item CET test, Column 5 the 22 CSE-referenced self-assessment ratings, and Column 6 the 16 CEFR-referenced self-assessment ratings.

1	2	3	4	5	6
Test takers	Composite analysis (including *anchor test*)	LTE test	CET test	CSE Can Dos	CEFR Can Dos
					
Test takers	Composite dataset	LTE test	CET test	CSE Can Dos	CEFR Can Dos