International Journal of TESOL Studies (2022)
Vol. 4 (1) 169-186 https://doi.org/10.46451/ijts.2022.01.11

Article

Using Self-Assessments to Investigate Comparability of the CEFR and CSE: An Exploratory Study Using the LanguageCert Test of English

Wen Zhao
Jinan University, China

David Coniam*
LanguageCert, UK

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

Abstract

This paper reports on an exploratory comparability study between the Common European Framework of Reference for Languages (CEFR) and the China Standards of English (CSE). Established equivalences are exhibited via the LanguageCert Test of English of reading and language use for the CEFR and a comparable test of reading and language use produced by a top-tier China university. In the study, a large sample of test takers took part, first sitting the two comparable tests of reading and language use, and subsequently completing a number of self-assessment Can-Do statements related to the CEFR and the CSE.

Validity of the dataset was established by linking both tests and sets of self-assessments to a single frame of reference using a third test whose robustness and values had been previously established. While there were some divergences between how the two frameworks aligned – more notably towards the lower ends of the scales – correspondences which emerged between the CEFR and CSE frameworks were broadly in accordance with those reported in other studies referenced in the current paper. The current study therefore sets the groundwork for determining the correspondence between LanguageCert Tests, aligned to the CEFR, and the CSE.

Keywords

Self-assessment, CEFR, CSE, Rasch, comparability

1 Introduction

The current study is the first step in aligning LanguageCert's different tests – which are currently aligned to the CEFR – to other key frameworks or assessments, in this case the CSE. To frame the study, the following section presents detail on methods of establishing comparability between such assessment

*Corresponding author. Email: david.coniam@PeopleCert.org

instruments as Can-Do self-assessments. Background to the CSE and CEFR is then presented, along with a description of studies which have investigated the correspondence between the frameworks.

2 Self-assessment of Language Abilities

Over the past two decades, self-assessment has been shown to be of value in assisting learners to evaluate their language ability (Bailey, 1998). The benefits of self-assessment (SA) have been explored in a number of studies and shown to make worthwhile contributions in both learning and assessment. In the context of learning, for example, Butler (2018) illustrated the value of SA in the self-regulated learning process, Babaii et al., (2016) showed how SA aided self-awareness in learning, Dann (2002) showed its value in promoting learner autonomy, and De Saint-Leger (2009) demonstrated how SA was associated with learner confidence and hence performance.

In the area of language assessment, SA has been shown to offer a range of potential benefits. Bachman & Palmer (1996) demonstrated how SA permitted learners to self-assess themselves in an interactive, yet low-anxiety, manner. Oscarson (1989) showed how SA could help expand the range of assessment, emphasising the fact that assessment should be the responsibility of both learners and teachers. Of relevance to the current study, Liu & Brantmeier (2019) reported a study of young learners in China who were able to quite accurately self-assess their abilities in reading and writing. As outlined below, Peng et al. (2021) explored the alignment of the CSE and the CEFR frameworks, in large part through the use of self-assessment descriptors.

Jones (2014) presents a description and analysis of the large-scale use of 'Can-Do' self-assessment descriptors [Note 1] established in the 1990s to provide common levels of proficiency across European languages via the ALTE (Association of Language Testers in Europe) Framework. Jones concludes that, despite there being some variation across different educational systems in Europe, students of different languages were, on the whole, reasonably accurate in estimating their relative ability. The use of instruments such as Can-do statements in self-assessment has been validated in a number of other studies (see e.g., Brown et al., 2014; Summers et al., 2019).

3 The CSE and the CEFR

For the past two decades, the CEFR has been accepted as illustrating standards of language ability by many stakeholders: policy makers, publishers, 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.

Figure 1
CEFR and CSE Levels

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

Jin et al. (2017) describe the development of the "Common Chinese Framework of Reference for English (CCFR-E): Teaching, Learning, Assessment" which began in 2014. The CCFR-E was finalised in 2018, being released as the "China Standards of English" (CSE). The CSE has three major level stages, each subdivided into three sublevels. Figure 1 illustrates.

4 Previous CSE / CEFR Equivalence Studies

Alderson (2017) discusses a range of studies exploring the CSE and its correspondence to the CEFR. This is supported by the discussion by Jin et al. (2017) and by research by Zhao et al. (2017), investigating the linking of College English vocabulary levels with the CEFR. Figure 2 presents a summary of the results of the different studies.

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 Figure 2 below illustrates, CSE L2 appeared to correspond to CEFR A1, CSE L3 to A2, CSE L4 / L5 to CEFR B1, CSE L6 / L7 to CEFR B2, CSE L8 to CEFR C1 and CSE L9 to CEFR C2.

Peng and associates have undertaken a number of studies investigating correspondences between CEFR and CSE levels. Level A0, it should be noted, denotes a level below CEFR 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. Peng et al. (2021) report CSE L1 as corresponding to CEFR A0, CSE L2 to CEFR A1, CSE L2 / L3 to CEFR A2, CSE L4 / L5 to CEFR B1, CSE L6 / L7 to CEFR B2, CSE L7 / L8 to CEFR C1, and CSE L9 to CEFR C2.

Figure 2

CFR/CSE Comparative Mappings from Previous Studies

Dunlea et al. (2019) All skills

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

Peng et al. (2021) All skills

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

Peng (2021) Writing

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

Peng & Liu (2021) Listening

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

In another study, Peng (2021) investigated level alignments between the CSE and CEFR writing descriptors. Results indicated a general correspondence between CSE and CEFR levels. While there was some overlap, CSE L1 / L2 corresponded to CEFR A1, CSE L3 to CEFR A2, CSE L4 / L5 to CEFR B1,

CSE L6 to CEFR B2, CSE L7 to CEFR C1, CSE L8 to CEFR C1 / C2, and CSE L9 to CEFR C2. In a further study, Peng & Liu (2021) attempted to align CSE listening skill levels with those of the CEFR. Results indicated that CSE listening descriptors tended to spread across several adjacent CEFR levels. CSE L1 corresponded to CEFR A1, CSE L2 to CEFR A2, CSE L3 to CEFR A2 / B1, CSE L4 to CEFR B1, CSE L5 to CEFR B1 / B2, CSE L6 to CEFR B2 / C1, CSE L7 / L8 to CEFR C1, and CSE L9 to CEFR C2.

The different studies outlined in Figure 2 contribute to the level alignment between the CSE and the CEFR. As may be seen, while there is a degree of agreement in the correspondence between the two studies, there are also divergences which may result from a number of factors: the samples; the tests; the judges used in the ratings.

5 Current Study

This section briefly outlines the background and make-up of the tests and the self-assessment ratings which test takers completed. The methodology employed in the current study differs from that used in the Dunlea et al. (2019) and Peng et al. (2021) studies. The principal methodology in the latter two involved the use of expert ratings. In the current study, a large sample of test takers took a live LanguageCert test, which was then calibrated in a single frame of references with the self-assessment ratings.

5.1 Test material

In late 2020, approximately 2,500 Year 1 non-English major college students took a 65-item multiple-choice reading and language use test prepared by experts from the university involved in the current study. Three months later, this same set of students took a 53-item multiple-choice reading and language use test adapted from existing and 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 representing the spectrum of difficulty across the six CEFR levels.

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

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 has been 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).

Appendix 1 provides a comparative analysis of the make-up of the two reading and usage tests. As may be seen, the CET test is slightly longer than the LTE test; also, all CET items are 4-option multiple-

choice whereas the LTE items are 3-option multiple-choice. Despite these differences, the content of the two tests, and even the order in which the different sections of the test appeared to test takers, exhibit a great deal of similarity.

5.2 Can-do self-assessment descriptors

Both the CEFR and the CSE contain large arrays, for all skill areas, of *Can-Do* descriptors (see e.g., https://www.cultofpedagogy.com/can-do-ell/ for examples of how such descriptors help classroom teachers understand what learners at different levels of proficiency should be able to do).

To reflect the focus of the current study, two sets of Can-Do self-assessment descriptors were assembled for reading and language use for each framework. A set of 22 Can-Do statements related to the CSE was compiled by the China university staff who designed the CET test used in the current study. Another set of 16 Can-Do statements related to the CEFR was compiled by members of the LanguageCert research and assessment team. All Can-Do statements were framed as Yes/No questions so that test takers rated themselves dichotomously (i.e., as **can** / **cannot**) on each statement. The relevant Can-do statements may be found in Appendices 2 and 3.

The composite set of 38 items were then intermingled. This was intended to forestall respondents trying to guess where their own estimated ability level might terminate.

5.3 Test and self-assessment profile administration

The first test (the CET) was administered in late 2020. In early 2021, the second test (the LTE) was administered. Immediately after the second administration, test takers completed both sets of Can-Do self-assessments. These were all presented bilingually in both English and Chinese.

5.4 Self-assessment can-do statements and research questions

Against the backdrop outlined above, the current study pursued two main Research Questions.

RQ1: To what extent can self-assessment Can-Do statements 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?

6 Statistical Analysis: 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, a test taker's score is not derived solely from the raw score. Rather, the test taker's theoretical probability of success in answering items is gauged, with the resulting probabilistic score emerging from the calculations. 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 considered, 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 (e.g., Coniam et al., 2021), the established values can be used to equate different test forms.

In interpreting Rasch, the key statistic involves the 'fit' of the data in terms of how well obtained values match expected values (Bond et al., 2020). A perfect fit of 1.0 indicates that obtained mean square values match expected values one hundred percent. Acceptable ranges of tolerance for fit range from 0.5 to 1.5 (Lunz & Stahl, 1990).

6.1 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), defined by Humphry (2006) as "compris[ing] a class of persons responding to a class of items in a well-defined assessment context."

In order to calibrate the four datasets in the current study onto the LanguageCert Item Difficulty (LID) scale (see Table 1), a previously calibrated test (henceforth referred to as "Test 3") from the Coniam et al. (2021) study was incorporated into the data. As a subset of Test 3, 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 Test 3. Following this, the other datasets in the study – the CET test and the two sets of self-assessments – could then be calibrated against Test 3 onto the LID scale. This resulted in all five assessment datasets being included into one single FOR.

6.2 Analysing within a single frame of reference

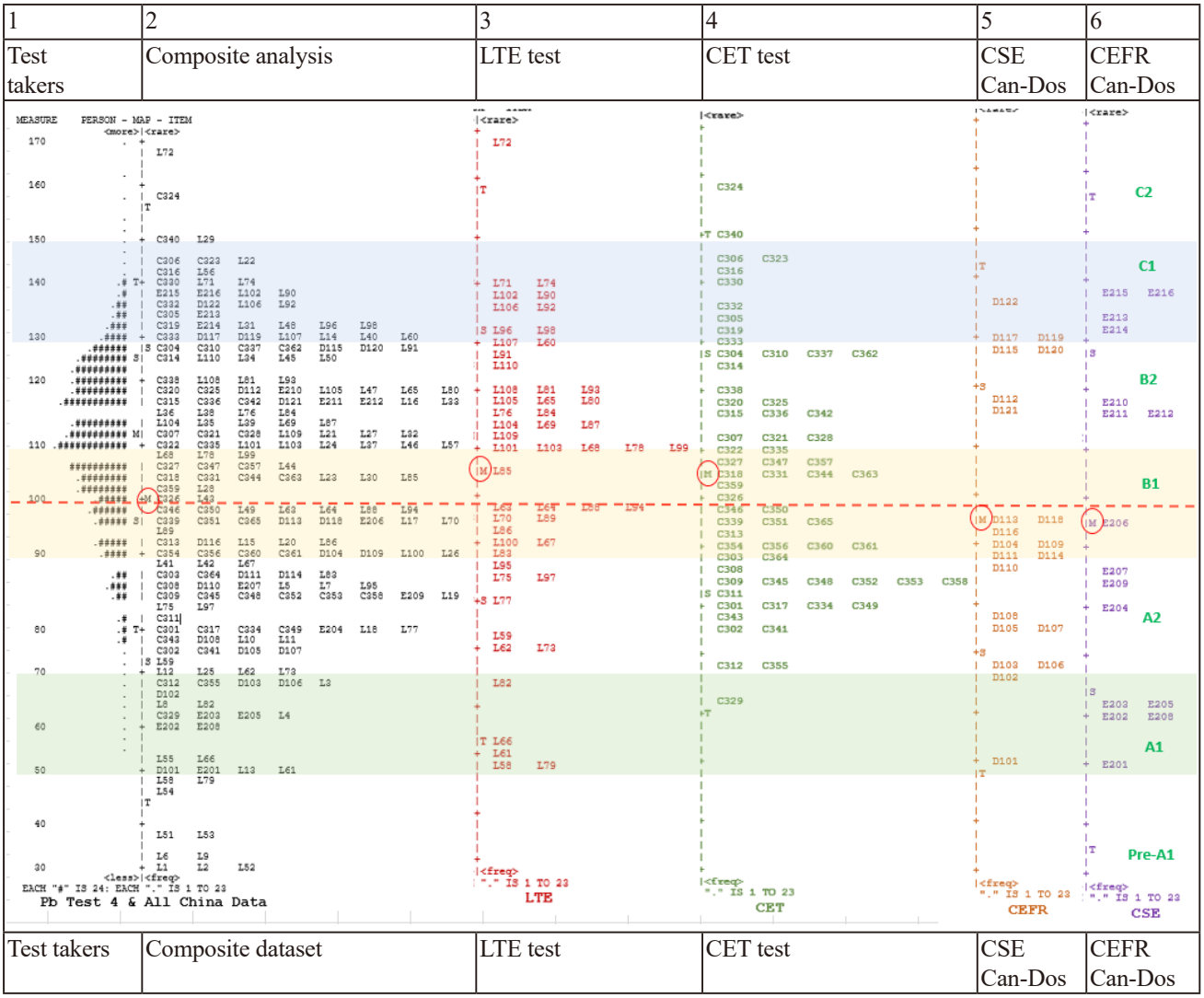
As mentioned, Test 3 was the anchoring frame, having been previously anchored to the LID scale. Against this backdrop, the composite analysis is presented in Figure 3 below.

In Figure 3, Column 2 contains the analysis of the amalgamated five datasets of 158+PB4 items. Column 3 contains the 53-item LTE test, Column 4 the 65-item CET test, Column 5 the 22 CSE-referenced Can-Do ratings, and Column 6 the 16 CEFR-referenced Can-Do ratings.

To recap, item links in the overall dataset are established between the 53 items in the LTE test and Test 3. Person links are established via the two tests and the two sets of self-assessments. All five datasets may therefore be seen to be within an overall FOR – the composite analysis to the far left of the person-item map in Figure 3. Against the overall picture of calibration, which is centred at 100, the mid-point of B1, it may be seen that the means for the two tests are slightly higher than the overall mean. Tables 2 and 3 present fit and reliability details on the two tests.

Figure 3

Composite Analysis of Three Tests and Two sets of Self-Assessments



Tables 2 and 3 show that the two tests fit the Rasch model well, with mean infit and outfit figures well within the 0.5 to 1.5 range, and high reliability figures. The means of both tests are very comparable, a quarter of a logit above the overall mean of 100. The LTE test mean was 105.33, and the CET test 104.28.

Table 2

Summary Analysis: 53-Item LTE Test

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2281.1	4218.7	105.33	.78	1.06	3.77	1.11	4.23
MAX.	4034.0	4265.0	168.63	1.38	1.23	9.90	1.51	9.90
MIN.	272.0	4137.0	48.18	.65	.93	-4.28	.78	-5.42
MODEL RMSE .80 TRUE SD 26.19 SEPARATION 32.77 ITEM RELIABILITY 1.00								
S.E. OF ITEM MEAN = 3.63								

Table 3

Summary analysis: 65-item CET test

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	1411.6	2328.2	104.28	1.01	.99	.20	1.00	.46
MAX.	2150.0	2335.0	158.32	1.58	1.18	9.90	1.35	9.90
MIN.	271.0	2293.0	62.50	.86	.87	-9.90	.68	-9.90
MODEL RMSE	1.02	TRUE SD	22.23	SEPARATION	21.82	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN = 2.78								

Tables 4 and 5 now present fit and reliability details for the two sets of self-assessments.

Table 4

Summary Analysis: 22 CSE Can-Do Statements

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2484.2	3888.5	95.29	.82	.89	-6.04	.83	-6.74
MAX.	3739.0	3980.0	136.74	1.36	.95	-1.35	.92	-2.55
MIN.	908.0	3823.0	50.61	.69	.82	-9.90	.72	-9.90
MODEL RMSE	.83	TRUE SD	23.90	SEPARATION	28.67	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN = 5.22								

Table 5

Summary Analysis: 16 CEFR Can-Do Statements

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2389.4	3877.2	95.66	.89	.90	-4.90	.80	-6.26
MAX.	3724.0	3954.0	138.02	1.39	.96	-1.43	.93	-2.21
MIN.	874.0	3836.0	49.67	.69	.84	-9.90	.61	-9.90
MODEL RMSE	.91	TRUE SD	30.67	SEPARATION	33.59	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN = 7.92								

From Tables 4 and 5, it can be also be seen that the two sets of self-assessments fit the Rasch model; mean infit and outfit figures are within the 0.5 to 1.5 range, and reliability figures are again high. The means of both two sets of self-assessments are again comparable, although this time a quarter of a logit below the overall mean of 100 – both being around 95. This slightly lower score is indicative that, on the self-assessments, test takers have tended to slightly over-rate themselves – a not uncommon phenomenon (Kruger & Dunning, 1999; Dunning et al., 2004).

The difference between the item means of the Can-Do ratings, and the LTE and CET assessment results are within half a logit (10 LID scale points): a difference which is generally accepted within Rasch measurement as being non-significant (Zwick et al., 1999). The conclusion that may be drawn is that test takers can be considered sufficiently objective in their self-assessments to permit tentative correspondences to be drawn between CSE and CEFR levels. The next section explores the correspondences.

6.3 Establishing correspondences between CSE and CEFR levels

Given that the two sets of self-assessments have been established as valid and broadly comparable, the current section presents sets of tables – one at each CEFR level – which incorporate Can-Do statements within corresponding CEFR and CSE levels. Tables are presented one at a time for each CEFR level, in

line with LID score ranges for the corresponding CEFR level. The tables are laid out such that the left-hand half of the table includes the detail for the CEFR level: the relevant Can-do statement, the LID value assigned in the current single FOR calibration, and the CEFR level for the Can-do, as laid down in formal documentation. The right-hand half of the table then includes corresponding detail for the CSE level: Can-do statements and their CSE level which fall into the LID value range for the CEFR level.

Table 6 presents the joint analysis for CEFR level C1, the LID range for which is 131-150 scale points.

Table 6

CEFR and CSE Can-Do Statement Level Comparisons: C1 (131-150)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.	138.02	C1			
I can understand specialised articles and longer technical instructions, even when they do not relate to my field.	137.77	C1			
			L7	136.74	I can comprehend academic papers or scientific and technical literature in relevant fields of study and evaluate the research methods.
I can understand long and complex factual and literary texts, appreciating distinctions of style.	133.82	C1			
I can extract necessary information and the points of the argument from articles and reference materials in my specialised field without consulting a dictionary.	130.74	C1			

Within the C1 CEFR LID range of 131-150, four CEFR C1 self-assessment were found, along with one CSE Level 7 self-assessment. The fit would appear to be CEFR C1 → CSE L7.

Table 7 presents the joint analysis for CEFR level B2, the LID range for which is 111-130 scale points.

Within the B2 CEFR LID range of 111-130, three CEFR C1 self-assessment 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 would appear to be broader, i.e., CEFR B2 → CSE L5-L7.

Table 8 presents the joint analysis for CEFR level B1, the LID range for which is 91-110.

Within the B1 CEFR LID range, one CEFR B1 self-assessment was found, along with four CSE self-assessments, of which one was at L4, one at L5, and two at L6. The B1 CEFR / CSE fit would therefore also appear to be quite broad, i.e., CEFR B1 → CSE L4-L6.

Table 7

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.
			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			

Table 8

CEFR and CSE Can-Do Statement Level Comparison Chart: B1 (91-110)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L4	95.4	I can extract the key information in practical forms of writing (e.g. memos or notes).
I can understand the plot of longer narratives written in plain English.	95.15	B1			
			L6	94.63	I can infer the author's attitudes with the help of diction or rhetorical devices.
			L6	93.86	I can understand and summarise the main features of the objects in expository writing.
			L5	90.93	I can extract detailed information (e.g. characters, scenic spots) from prose essays.

Table 9 presents the joint analysis for CEFR level A2, the LID range for which is 71-90.

Within the A2 CEFR LID range, three CEFR A2 self-assessment were found, along with seven CSE self-assessments, of which one was at L3, four at L4, and two at L5. The A2 CEFR / CSE fit would therefore appear to be mainly CEFR A2 → CSE L4-L5.

Table 9

CEFR and CSE Can-Do Statement Level Comparison Chart: A2 (71-90)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L4	89.84	I can analyse the authors' viewpoints on familiar social phenomena in short, simple pieces of argumentative writing.
			L5	89.07	I can read arguments on common topics and commentary on familiar topics.
			L5	88.32	I can generalise duly from what has been read while reading.
I can search the internet or reference books, and obtain school- or work-related information, with the help of a dictionary.	87.43	A2			
			L4	86.37	I can discover the key information or details by skimming, scanning, and/ or browsing.
I can understand clearly written instructions (e.g. for playing games, for filling in a form, for assembling things).	83.72	A2			
I can understand the main points of English newspaper and magazine articles adapted for educational purposes.	79.90	A2			
			L4	77.65	I can understand details (e.g. time, character, place) in travel notes.
			L4	75.88	I can read short, simple stories, prose essays, and expository writing.
			L3	75.61	I can understand the authors' viewpoints in short, simple letters.

Table 10 presents the joint analysis for CEFR level A1, the LID range for which is 51-70.

Within the A1 CEFR LID range, four CEFR A1 self-assessments were found, along with three CSE self-assessments, of which one was at L2, and two at L3. The broad A1 CEFR / CSE fit would appear to be CEFR A1 → CSE L3.

Finally, below CEFR A1, there was one fit between the CEFR and CSE. Table 11 presents.

In this mapping, low A1 ("A0") fitted with CSE L1.

From the above set of tables with the comparative fit of the CEFR and CSE levels, it is now possible to produce an overall tentative mapping of how the CEFR scale, as represented by the LTE, may be mapped against the CSE. Table 12 presents the match. It should be noted that there was insufficient data to calibrate CEFR level C2.

Table 10

CEFR and CSE Can-Do Statement Level Comparison Chart: A1 (51-70)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L3	68.48	I can improve my understanding with reference to key words or topic sentences.
			L3	68.23	I can understand linguistically simple stories.
			L2	67.23	I can pick out the key information in notes or notices.
I can understand the main points of texts dealing with everyday topics (e.g. life, hobbies, sports) and obtain the information I need.	62.61	A1			
I can understand short narratives and biographies written in simple words.	60.80	A1			
I can understand texts of personal interest (e.g. articles about sports, music, travel, etc.) written with simple words.	60.28	A1			
I can understand very short reports of recent events such as text messages from friends' or relatives', describing travel memories, etc.	59.61	A1			

Table 11

CEFR and CSE Can-Do Statement Level Comparison Chart: A0 (below 51)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L1	50.61	I can understand short, linguistically simple articles on daily life.
I can understand very short, simple, everyday texts, such as simple posters and invitation cards.	49.67	A0			

Table 12

CEFR / CSE Fit in LTE Study

CEFR	China 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 12, as might perhaps be expected, 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.

Figure 4 below, presents a reworking of Figure 2, which included the alignments proposed in the Dunlea et al. (2019) [henceforth the 'Donlea' study] and Peng and associates' (2021) studies [henceforth the 'Peng' studies], together with the alignments as they have emerged empirically in the current study.

Figure 4

Formal CEFR / CSE Mapping

LC Mapping		Dunlea et al. (2019)		Peng et al. (2021)		Peng (2021)		Peng & Liu (2021)	
Reading & Language Use		All skills		All skills		Writing		Listening	
CSE	CEFR	CSE	CEFR	CSE	CEFR	CSE	CEFR	CSE	CEFR
	C2	L9	C2	L9	C2	L9	C2	L9	C2
L7	C1	L8	C1	L7-L8	C1	L8	C1-C2	L7-L8	C1
L5-L7	B2	L6-L7	B2	L6-L7	B2	L7	C1	L6	B2-C1
L4-L6	B1	L4-L5	B1	L4-L5	B1	L6	B2	L5	B1-B2
L3-L5	A2	L3	A2	L2-L3	A2	L4-L5	B1	L4	B1
L2-L3	A1	L2	A1	L2	A1	L3	A2	L3	A2-B1
L1	A0	L1		L1	A0	L1-L2	A1	L2	A2
								L1	A1
									A0

The results of the current study can be seen to echo the mappings of the previous studies, although the mappings which have emerged suggest a slightly more lenient fit than that reported in other studies (see below) – as for example with CEFR C1 being located against CSE L7 in the current study as against CSE L7 / L8 in the Peng studies and CSE L8 by Dunlea. This is mirrored at the lower end of the scale, where the current study does not suggest direct one-to-one matches. There are a number of possible reasons for these divergences. A key difference is that the current study empirically matched *levels* against *performance*, as opposed to an expert-rater-focused methodology. Another reason may be attributable to the fact that only one skill – essentially reading – has been explored in the current study, whereas the other studies examined all four skills. A third is that the sample was limited at the top end of the ability spectrum to C1-level test takers.

7 Conclusion

The current study was pursuing two Research Questions. The first research question was that self-assessment Can-Do statements may be validly used to establish correspondences between the CEFR and CSE frameworks. As was illustrated, from a comprehensive analysis of both test and Can-Do self-assessment responses, respondents tended to slightly over-estimate their abilities on both the CEFR and the CSE. These over-estimations were minimal, however, in that mean values were only a quarter of a logit higher than might have been expected. Secondly, the over-estimations were consistent with the scales for both frameworks.

The second research question was that correspondences between the CEFR and CSE frameworks would be broadly in accordance with those proposed by previous studies. While there have been some divergences, more notably towards the lower end of the scales, the correspondences proposed in the current study broadly echo those reported in previous studies.

A range of correspondences may well be expected from different studies, exploring different assessment instruments. Difficulties in accurate alignment have been commented on by other researchers (Papageorgiou et al., 2015; North & Piccardo, 2018). Peng (2021) insightfully comments that "the CSE is a local standard with granular levels reflecting Chinese learners' requirements and progress [….] while the CEFR is a framework for reference with broad bands of proficiency and is intended to be adapted or further developed for specific contexts and uses". In the current study, the assessment context has focused on reading and language use, whereas the Dunlea et al. (2019) and the Peng et al. (2021) studies examined all four language skills, as well as writing and listening, which Peng (2021) and Peng & Liu (2021) respectively explored.

From a wider, and methodological, perspective, the use in the current study of a single frame of reference to calibrate self-assessment ratings directly against performance adds to the armoury of tools available to assessment professionals in linking exercises such as those between two different tests, or by providing a larger perspective between two different assessment frameworks.

The approach adopted in the current study may be useful for other assessment situations, where Can-Do ratings may be incorporated at the end of an assessment session. This may even be done in a user-friendly manner where individual candidates rate subsets of Can-Do ratings, which are then linked via common items to cover a range of Can-Do aspects.

A limitation of the current study was that the investigation of test types was limited to reading and language use. Future studies will broaden this by extending the investigations conducted in the current study to other language skills.

Notes

1. The CEFR framework comprises descriptors laying out what a student can do as a particular skill when they have completed a given level. A descriptor for Reading at A2, for example, is: "I can understand short narratives and biographies written in simple words."

Appendix 1
Component Analysis of CET and LTE Tests

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

Appendix 2
CSE Can-Do Statements used in the Study

CSE No.	No.		Level
CSE0201	1	I can understand short, linguistically simple articles on daily life.	CSE2
CSE0202	2	I can pick out the key information in notes or notices.	CSE2
CSE0301	4	I can understand linguistically simple stories.	CSE3
CSE0302	5	I can extract detailed information (e.g. characters, scenic spots) from prose essays.	CSE3
CSE0303	7	I can understand the authors' viewpoints in short, simple letters.	CSE3
CSE0304	8	I can improve my understanding with reference to key words or topic sentences.	CSE3
CSE0401	10	I can read short, simple stories, prose essays, and expository writing.	CSE4
CSE0402	11	I can understand details (e.g. time, character, place) in travel notes.	CSE4
CSE0403	13	I can analyse the authors' viewpoints on familiar social phenomena in short, simple pieces of argumentative writing.	CSE4
CSE0404	15	I can discover the key information or details by skimming, scanning, and/or browsing.	CSE4
CSE0501	17	I can read arguments on common topics and commentary on familiar topics.	CSE5
CSE0502	19	I can understand the common figures of speech in stories pertaining to social life written in relatively complex language.	CSE5
CSE0503	21	I can extract the key information in practical forms of writing (e.g. memos or notes).	CSE5
CSE0504	23	I can generalise duly from what has been read while reading	CSE5
CSE0601	25	I can understand novels and argumentative texts comprised of relatively complex language.	CSE6
CSE0602	27	I can understand and summarise the main features of the objects in expository writing.	CSE6
CSE0603	29	I can understand the terminology of operational texts in related professional areas.	CSE6
CSE0604	31	I can infer the author's attitudes with the help of diction or rhetorical devices.	CSE6
CSE0701	33	I can understand linguistically complex novels and materials related to culture and appraise their linguistic features.	CSE7
CSE0702	35	I can understand book reviews in relevant fields of inquiry.	CSE7
CSE0703	36	I can infer the content of an entire book or text by scanning the table of contents.	CSE7
CSE0802	37	I can comprehend academic papers or scientific and technical literature in relevant fields of study and evaluate the research methods	CSE8

Appendix 3
CEFR Can-Do Statements used in the Study

CEFR No.	No.		Level
CEFR01	3	I can understand very short, simple, everyday texts, such as simple posters and invitation cards.	A1
CEFR02	6	I can understand very short reports of recent events such as text messages from friends' or relatives', describing travel memories, etc.	A1

CEFR03	9	I can understand texts of personal interest (e.g. articles about sports, music, travel, etc.) written with simple words.	A1
CEFR04	12	I can understand short narratives and biographies written in simple words.	A2
CEFR05	14	I can understand the main points of texts dealing with everyday topics (e.g. life, hobbies, sports) and obtain the information I need.	A2
CEFR06	16	I can understand the main points of English newspaper and magazine articles adapted for educational purposes.	B1
CEFR07	18	I can understand clearly written instructions (e.g. for playing games, for filling in a form, for assembling things).	B1
CEFR08	20	I can search the internet or reference books, and obtain school- or work-related information, with the help of a dictionary.	B1
CEFR09	22	I can understand the plot of longer narratives written in plain English.	B1
CEFR10	24	I can read texts dealing with topics of general interest, such as current affairs, without a dictionary, and can understand multiple points of view.	B2
CEFR11	26	I can understand in detail specifications, instruction manuals, or reports written for my own field of work	B2
CEFR12	28	I can scan through rather complex texts, e.g. articles and reports, and can identify key passages.	B2
CEFR13	30	I can extract necessary information and the points of the argument from articles and reference materials in my specialised field without consulting a dictionary.	B2
CEFR14	32	I can understand long and complex factual and literary texts, appreciating distinctions of style.	C1
CEFR15	34	I can understand specialised articles and longer technical instructions, even when they do not relate to my field.	C1
CEFR16	38	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.	C2

References

Alderson, J. C. (2017). Foreword to the special issue "The Common European Framework of Reference for languages (CEFR) for English language assessment in China" *Language Testing in Asia, 7*, 20.

Babaii, E., Taghaddomi, S., Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing, 33*(3), 411–437.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions and directions*. Heinle & Heinle.

Bond, T., Yan, Z., & Heeney, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences.* Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how?. *CBE—Life Sciences Education, 15*(4): rm4.

Brown, G. T., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. *Assessment in Education: Principles, Policy & Practice, 22(*4), 444-457.

Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals, 47*(2), 261-285.

Butler, Y. G. (2018). The role of context in young learners' processes for responding to self-assessment

items. *The Modern Language Journal, 102*(1), 242–261.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Brigham Young University Testing Services and the Department of Instructional Science. https://testing.byu.edu/handbooks/betteritems.pdf

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

Council of Europe. (2009). Relating language examinations to the Common European Framework of References for languages: Learning teaching, assessment. Council of Europe

Educational Testing Service. (2012). *The official guide to the TOEFL test (4th Ed.).* McGraw-Hill.

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. Routledge.

De Saint-Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals, 42*(1), 158-178.

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.

Dunlea, J., Spiby, R., Wu, S., Zhang, J., & Cheng, M. (2019). China's standards of English language ability: Linking UK exams to the CSE. https://www.britishcouncil.org/sites/default/files/linking_cse_to_uk_exams_5_0.pdf.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69–106.

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

Educational Testing Service. (2015). Test and score data summary for TOEFL iBT Tests: January 2014 – December 2014 test data. https://www.ets.org/s/toefl_itp/pdf/toefl-itp-test-score-data-2014.pdf.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66(*4), 477–485.

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

Humphrey, 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.

Jones, N. (2014). *Multilingual frameworks (Vol. 40).* Cambridge University Press.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology, 77*(6), 1121-1134.

Lee, T., Milanovic, M., Coniam, D., & Pike, N. (2021). *Externally-referenced anchoring: Equating expert judgement and Rasch measurement values in LanguageCert IESOL English language tests.* LanguageCert.

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.

Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System, 80*, 60-72.

Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession, 13,* 425-444.

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

Ability. Ministry of Education.

North, B., & Piccardo, E. (2018). Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of References (CERF). Centre for Canadian Language Benchmarks.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. Language Testing, 6(1), 1-13.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels. Research Memorandum No. RM-15-06).*: Educational Testing Service.

Peng, C. (2021). Aligning the CSE with the CEFR: Level alignment in writing ability. *Foreign Language World, 5,* 84-93.

Peng, C., & Liu, J. (2021). The listening skill level alignment of the CSE with the CEFR. *Foreign Language Educator, 5,* 43-50.

Peng, C., Liu, J., & Cai, H. (2021). Aligning China's standards of English language ability with the Common European Framework of Reference for Languages. *The Asia-Pacific Education Researcher,* . https://doi.org/10.1007/s40299-021-00617-2.

Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (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.

Wright, B. D. (1997). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.

Zhao, W., Wang, B., Coniam, D., & Xie, B. (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.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28.

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. ORCID: 0000-0003-4965-0146

David Coniam 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. ORCID: 0000-0003-4480-1742