Linking the BEST Test of English Proficiency (BESTEP) Writing Test to the Common European Framework of Reference (CEFR)

LTTC-BESTEP Research Report
BESTEP-01

Jason Fan, Ivy Chen, and Ute Knoch

Language Testing Research Centre, University of Melbourne, Australia

January 2024

Table of Contents

Abstract

This study linked the BEST Test of English Proficiency (BESTEP) writing test to the Common European Framework of Reference (CEFR), following the four stages recommended by the CEFR Linking Manual (Council of Europe, 2009): familiarisation, specification, standardisation, and validation. The BESTEP is an English proficiency test developed by the Language Training and Testing Centre (LTTC) in Taiwan; it assesses college students' readiness for academic English as the medium of instruction (EMI) programs in Taiwan's tertiary education. The BESTEP writing test consists of three tasks: Answering questions (Task 1); Expressing opinions (Task 2), and Writing an integrated essay (Task 3).

The study involved 15 panellists, including six test 'insiders' from Taipei and nine test 'outsiders' based in Australia. Compared with the test 'outsiders', the test 'insiders' were more familiar with the relevant English teaching and learning context, the test takers and the BESTEP writing test. The test 'outsiders', in contrast, had little background knowledge about the BESTEP, yet they possessed considerable knowledge of and experience in academic writing and the CEFR. Notably, this study not only linked the score levels of the BESTEP writing test to the CEFR levels, but also explored the panellists' cognitive processes through a think-aloud study. Three research questions were investigated in this study: RQ1. How do the score levels of the BESTEP writing test relate to the CEFR levels? RQ2. How do the judgements of test 'insiders' compare to those of test 'outsiders''? RQ3. What are the panellists' mental processes when linking the BESTEP writing scripts to the CEFR levels? Are there any differences between test 'insiders' and test 'outsiders' in their linking processes?

To address RQ1, we followed the four-stage linking process recommended by the CEFR Linking Manual. Specifically, we employed the Body of Work (BoW) method for standard setting, which is suitable for holistic judgements of performance on different task types. We also collected multiple types of evidence to support the validity of the linking results. To investigate RQ2, we applied the many-facets Rasch model (MFRM) to analyse the data from the panellists' judgements. The findings offered important insights into the panellists' severity levels in their evaluations as well as their fit to the Rasch model. RQ3 was explored through a think-aloud study, involving three test 'insiders' and three test 'outsiders'. The data were analysed thematically to identify the key themes in the participants' verbal protocols.

MFRM results revealed that as a group, test 'outsiders' tended to be more lenient in their judgements compared to test 'insiders', although all panellists fit the Rasch model. A few misfitting scripts were eliminated from the subsequent linking analysis. As part of the BoW method, a series of logistic regression analyses were implemented to determine the cut scores at different CEFR levels. The analysis of the think-aloud data identified three broad categories: linking process, linking strategies, and the challenges encountered. The findings suggest that both test 'insiders' and 'outsiders' engaged in a dynamic and iterative linking process, focusing on similar aspects of test takers' performance and adopting similar strategies. This linking study is significant in several aspects. The linking results provide the LTTC and the BESTEP users with credible evidence regarding the alignment of the BESTEP writing test to the CEFR levels, hence facilitating the interpretations of test takers' scores on the BESTEP writing test. In addition, this

study has implications for future linking research, including the engagement of panellists from different backgrounds and the use of the Rasch model as a quality control mechanism to enhance the validity of the linking results.

List of Tables

List of Figures

1. Introduction

This study aimed to link the BEST Test of English Proficiency (BESTEP) writing test to the Common European Framework of Reference (CEFR). Developed by the Language Training and Testing Centre (LTTC) in Taiwan, the BESTEP assesses college students' readiness for academic English in English as the medium of instruction (EMI) programs in Taiwan's tertiary education, as an important part of the Program on Bilingual Education for Students in College (a.k.a., the BEST Program) launched by Taiwan's Ministry of Education to cope with the challenges of globalisation. The BESTEP was also implemented with the express purpose of promoting the teaching and assessment of English skills in Taiwan's tertiary education. The test is designed to reflect the language skills and abilities required in Taiwan's EMI learning context, covering A2 to C1 on the CEFR.

The CEFR represents one of the major initiatives by the Council of Europe to provide common reference levels for teaching and learning of all languages in Europe (Council of Europe, 2001, 2018). The CEFR consists of six reference levels across three bands:

1) A - Basic user, including A1 (Breakthrough) and A2 (Waystage)
2) B - Independent user, including B1 (Threshold) and B2 (Vantage)
3) C - Proficient user, including C1 (Effective operational proficiency) and C2 (Mastery)

The six common reference levels in the CEFR aim to provide a common metalanguage for the language education profession and to facilitate the mutual recognition of language qualifications, as indicated by courses taken or examinations passed. In the CEFR, language proficiency is described in a set of scales covering a range of skills, including reading, listening, writing, and speaking, as well as a range of communicative competences, with illustrative 'can-do' descriptors provided in *Common European Framework of Reference for Languages: Learning, teaching, and assessment* (Council of Europe, 2001) and the recently published CEFR companion volume (Council of Europe, 2018). Since the publication of the CEFR, its enormous impact has not only been felt in Europe, but indeed globally.

This study has two objectives: (a) linking the BESTEP writing test to the CEFR levels; and (b) exploring the panellists' cognitive processes when linking the BESTEP writing samples to the CEFR levels, particularly in terms of the challenges related to specific writing tasks. The findings of this study are expected to provide robust evidence to the BESTEP stakeholders (e.g., the test provider, test takers, teachers, and policymakers) regarding the alignment of the BESTEP score levels with the CEFR levels, thus facilitating test score interpretation and use. This will, in turn, further promote the implementation of EMI programs in Taiwan. Additionally, by delving into the panellists' cognitive processes, the findings will offer useful insights regarding the use of the CEFR writing scales and descriptors for future alignment studies.

2. Literature Review

2.1 The four stages of linking language tests to the CEFR

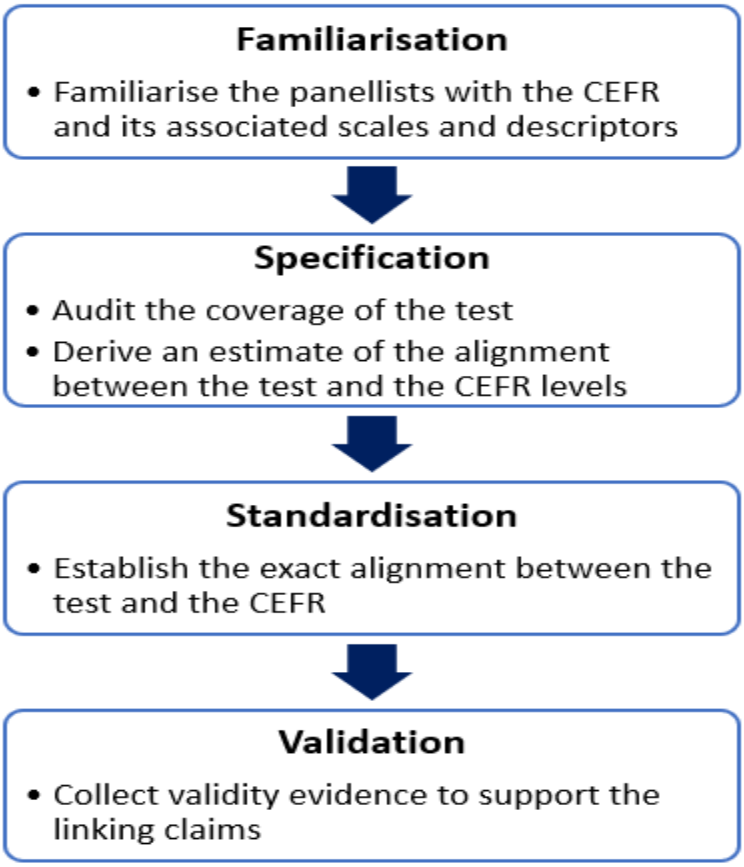
To assist test providers in mapping their language tests to the CEFR levels, the Council of Europe piloted a set of recommended linking procedures and subsequently published a manual for relating language examinations to the CEFR (Council of Europe, 2009). According to the Manual, four stages should be followed in linking a language test to the CEFR, namely (a) familiarisation, (b) specification, (c) standardisation, and (d) validation (see Figure 1), each of which is briefly explained below.

The first stage, familiarisation, aims to ensure that every participating panellist has a good understanding of the CEFR, particularly the CEFR scales and descriptors related to the target skill(s). This step is crucial for an alignment study because 'many of the language professionals in a linking project start with a considerably lower level of familiarity with the CEFR than they think they have' (Council of Europe, 2009, p. 17). Additionally, this stage can also assist panellists in acquainting themselves with the target test, including its tasks and assessment criteria. The Manual recommends a variety of activities to help panellists develop their knowledge of the CEFR, such as reading the CEFR descriptors to identify the salient features for each level, engaging in self-assessment using the CEFR scales and descriptors, and sorting individual descriptors from a CEFR scale. For researchers undertaking an alignment project, it is important to document the familiarisation activities and their outcomes as this is an integral part of supporting the validity of the alignment results (British Council et al., 2022).

The second stage, specification, involves a detailed description or delineation of the content, skills and abilities that the test aims to measure in relation to the CEFR categories and levels. According to the Manual, it is important for this process to be supported by evidence of the reliability and validity of the test as well as evidence demonstrating adequate quality control procedures in test development and administration. The Manual provides a range of specification forms and activities that assist linking researchers in analysing a language test, covering a range of areas such as a general description of the test, test development process, marking, grading, reporting the results, data analysis, and rationale for decisions. As highlighted by the Manual, the purpose of this stage is to profile the various aspects of a test in relation to the relevant CEFR scales and descriptors. In addition, linking researchers should also arrive at a rough estimate of the CEFR level for a test based on the analysis at this stage.

Figure 1

The four stages of aligning language examinations to the CEFR



The third stage is known as standardisation. This stage involves a group of panellists evaluating a language test and/or test-taker performances in relation to the CEFR scales and descriptors. The purpose is to build a consensus regarding what a test taker can do at a given CEFR level and whether this corresponds to the level of the test (British Council et al., 2022). The Manual recommends four steps in the process of standardisation: (a) carry out the CEFR familiarisation activities; (b) work with illustrative examples which have already been aligned to the CEFR to achieve an adequate understanding of the CEFR levels; (c) develop an ability to align the test tasks and performances to the CEFR levels; and (d) ensure all parties share the understanding (see also British Council et al., 2022, p. 40). Quite a few questions should be carefully considered at this stage, including the number of panellists, their background, skills, subject knowledge and expertise, as well as how they should be trained. This process is prerequisite to benchmarking or standard setting.

A range of standard setting methods are available (e.g., Angoff methods, bookmark methods, Body of Work methods), each with its own strengths and weaknesses (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kenyon & Römhild, 2013). The Manual recommends that

researchers select and implement the linking methods based on their situation and intended functions. In this study, the Body of Work method (BoW, Kingston et al., 2001) was employed, which is suitable for holistic judgements of performance on different task types. According to Cizek and Bunch (2007, p. 117), the BoW method is 'perhaps the most widely used of the holistic methods'. When applying the BoW method, panellists examine students' responses to different tasks in the test and match their response set to performance level categories (Kingston et al., 2001; Kingston & Tiemann, 2012). The BoW method encompasses several important steps in the standard setting process. Before the standard setting panel meeting, it is important to ensure that all panellists are familiar with the writing descriptors in the CEFR and the target test. In addition, three types of student work folders need to be created: (a) pinpointing folders, referring to an initial set of folders that includes samples of students' work at different score levels; (b) range-finding folders, where a sample of students' work is selected from the pinpointing folders (highest- and lowest-scoring performances); and (c) training folders, referring to a small set of response sets covering a range of test scores used for training the panellists in the standard setting study (Kingston & Tiemann, 2012). Several possible ways can be used to analyse the data collected in the BoW method, the most common of which is logistic regression, which models the relationship between a continuous variable (e.g., a test score) and the probability of being in a binary category (e.g., at CEFR level B1 or not at B1) (Kingston et al., 2001).

The last stage, validation, pertains to the building of a compelling argument from the evidence collected in the standard setting process to support the claims made of the standard setting results (British Council et al., 2022). This stage involves synthesising the evidence gathered across the different stages to back the interpretation and use of the standard setting results. Different types of evidence can be collected. For example, at the familiarisation stage, evidence may include documentation of the familiarisation activities and a survey of the panellists' views on the effectiveness of the familiarisation process. At the standard setting stage, evidence may include documentation of how the panellists are selected, the standard setting method that is used, as well as the rigour of the statistical analysis. In this study, we mainly focused on procedural and internal validity. Procedural validity was assessed by evaluating panellists' familiarity with the CEFR writing descriptors and the BESTEP writing test, and their ability to follow the judgement procedures. Internal validity was examined by gathering evidence demonstrating the agreement of panellists when grouping the writing samples.

2.2 Studies linking language tests to the CEFR

In the field of language assessment, a profusion of research has been conducted to link language tests to language proficiency frameworks, most notably the CEFR due to its global influence. Fleckenstein et al. (2020) reported on a study aiming to link TOEFL iBT writing rubrics to CEFR levels. The study was conducted in the context of two European countries: Germany and Switzerland. Specifically, the study served two purposes: (a) determining the cut scores for

writing profiles of students in the two countries in upper secondary education; and (b) analysing the congruity between the TOEFL iBT, the CEFR and the educational standards for upper-secondary education in the two countries. The standard setting method adopted by this study was a modified version of the 'examinee paper selection method' or 'performance profile method'.

Similarly, Papageorgiou et al. (2019) mapped the TOEFL iBT to China's Standards of English Language Ability (CSE), a locally developed language proficiency framework, following a series of steps, which included establishing recommended cut scores for test takers at each local proficiency level, to make test score interpretation meaningful in the Chinese context. Cut scores were set via three different methods: (a) standard setting with a panel of local experts (i.e., teachers from varying school levels), (b) congruence between test-takers' test scores and their teacher's evaluation of their CSE levels, and (c) cut-scores derived from a TOEFL iBT-IELTS score concordance study. The standard setting method employed for linking the constructed response test sections (i.e., writing and speaking) was a variation of the 'performance profile method'.

For PTE Academic, De Jong et al. (2014) linked the test to the CEFR. Unlike most linking studies where tests are linked to language proficiency frameworks, this study describes how the CEFR was also considered during the test development process. Item writers and item reviewers familiar with the CEFR were asked to indicate CEFR levels for each item. Statistical linking procedures were also followed afterwards; an examinee-centred approach was used for essay writing, where Rasch analysis of CEFR ratings were used to determine cut-scores. Lim et al. (2013) conducted a standard setting study linking IELTS Academic to the CEFR. A modified 'analytic judgement' method was used for constructed response tests to determine cut scores. Findings were compared with an external criterion validation study linking IELTS and CAE test scores, where CAE has known links to CEFR levels.

Knoch and Frost (2016) used a twin panel design to link the General English Proficiency Test (GEPT) writing subtest to the CEFR, with a group of panellists from Taiwan (i.e., those familiar with the GEPT) and a group from Australia. Two examinee-centred standard setting methods were employed: the 'borderline' method and the 'contrasting groups' method. The resulting cut scores were similar for both panels and across the two methods. Fan et al. (2021) linked Part 1 of the GEPT writing test (translation task) to the CEFR, as it was not able to be linked by Knoch and Frost (2016) due to relevant CEFR mediation scales not being released until 2018 in the CEFR companion volume (Europe, 2018). This project similarly followed a twin panel design with the same two standard setting methods. It also included a think-aloud study to explore the processes through which the two groups of panellists linked the GEPT translation scripts to the CEFR levels.

A review of recent studies on linking language tests to the CEFR indicates that most studies follow the four steps recommended by the linking manual (Council of Europe, 2009). Different standard setting methods have been adopted, depending on the target test to be linked to the CEFR and the research context. Some studies adopted more than one linking method to cross-validate the linking results (Knoch & Frost, 2016). Very few studies, however, have examined the

panellists' mental processes when linking the language test to the CEFR, except for Fan et al. (2021), who explored the panellists' mental processes through a think-aloud study. Their findings indicate that test 'insiders' (i.e., those who were familiar with the test and the CEFR but with limited or no experience in translation theory and practice) and test 'outsiders' (i.e., those who had a significant amount of experience in translation theory and practice but were less familiar with the test and the CEFR) were characterised by different orientations, with the former group focusing more on the accuracy of language use (e.g., grammar, vocabulary) and the latter group focusing more on the quality of translation (e.g., fidelity of the translation to the source text, completeness of the translation).

This study employed a similar think-aloud study to examine the panellist's cognitive processes when linking the BESTEP writing samples to the CEFR levels. This think-aloud study was motivated by several factors. First, this study used Body of Work (BoW) as the standard setting method, which involves a holistic evaluation of a test taker's performance on the three tasks in the BESTEP writing test. As such, it would be interesting to explore how each panellist approached a writing sample and which aspects of writing performance they prioritised when evaluating a test taker's performance across the three tasks. Second, the think-aloud study would also shed light on the challenges that the participating panellists encountered in the linking process. Finally, consistent with Fan et al. (2021), we were also interested in exploring the differences, if any, in the cognitive processes between the two groups of panellists.

2.3 The BESTEP writing test

As noted, the BESTEP is an academic English proficiency test developed by the Language Training and Testing Centre (LTTC) in Taiwan. It assesses college students' readiness in academic English as the medium of instruction (EMI) programs in Taiwan's tertiary education. This test was developed as a component of the Program on Bilingual Education for Students in College (a.k.a., the BEST Program) launched by Taiwan's Ministry of Education to cope with the challenges of globalisation. The test also aims to promote the teaching and learning of academic English proficiency in Taiwan's tertiary education. It is a multilevel English proficiency test, targeting those who are at A2 to C1 on the CEFR. This study focuses on the BESTEP writing test.

The BESTEP writing test consists of three tasks. In the first task (Answering questions), test takers are required to read a poster and compose short answers to three questions in about 5 minutes. Test takers' responses should be approximately 25 words in total. In the second task (Expressing opinions), test takers are required to write a short email of approximately 80 words based on a prompt to express their opinions on an issue or phenomenon. Test takers are expected to complete this task in 15 minutes. In the third task (Writing an integrated essay), test takers are required to write an essay of 120-150 words based on a prompt that consists of two graphs. Test takers are expected to complete this task in 30 minutes. More details about the writing test, including sample

tasks and responses are available on the BESTEP official website (https://bestep.tw/eng/Resource/page?id=c1db2b929ff6480fb7204b914dbb0b41).

Task-specific holistic rating criteria are used to evaluate test takers' performance on the BESTEP writing section. For Task 1 (Answering questions), a 6-point (from 0 to 5) holistic scale is used to evaluate students' performance, covering (a) content, focusing on the relevance and completeness of their responses; and (b) language quality, including grammar, vocabulary, and mechanics (i.e., spelling, punctuation). The 6 points aim to correspond to different levels in the CEFR (see below):

- 0: A1 or below
- 1-2: A1 (1 – A1.1; 2 – A1.2)
- 3-4: A2 (3 – A2.1; 4 – A2.2)
- 5: B1 or above

For Task 2 (Expressing opinions), a 7-point holistic scale (from 0 to 6) is used to evaluate students' performance, covering (a) content relevance and completeness, (b) organisation, (c) cohesion and coherence, (d) grammar and vocabulary, and (e) mechanics (i.e., spelling, punctuation). The 7 points aim to correspond to different levels in the CEFR (see below):

- 0: below A2
- 1-2: A2 (1 – A2.1; 2 – A2.2)
- 3-4: B1 (3 – B1.1; 4 – B1.2)
- 5: B2
- 6: C1 or above

For Task 3 (Writing an integrated essay), a 7-point holistic scale (from 0 to 6) is used to evaluate students' performance, covering (a) content relevance and completeness, (b) organisation, (c) cohesion and coherence, (d) grammar and vocabulary, and (e) mechanics (i.e., spelling, punctuation). The 7 points aim to correspond to different levels in the CEFR (see below):

- 0: below A2
- 1: A2
- 2-3: B1 (2 – B1.1; 3 – B1.2)
- 4-5: B2 (4 – B2.1; 5 – B2.2)
- 6: C1 or above

Raw scores of the three parts are weighted in commensurate with their relative difficulty level. The more difficult the task is, the more weight it is assigned. The scores on each part are adjusted according to its weighting and aggregated to yield a scale score ranging from 0 to 360. The score conversion table is presented in Table 1 below:

Table 1
Conversion from the raw scores on the BESTEP writing test to scale scores

Scale score	CEFR	Band score		
		Part 1 (22%)	Part 2 (36%)	Part 3 (42%)
150	Above C1			6
130	C1		6	
120	B2.2			5
110	B2.1		5	4
90	B1.2		4	3
80	B1.1	5	3	2
60	A2.2	4	2	1
50	A2.1	3	1	
40	A1.2	2	0	0
30	A1.1	1		
20	Below A1	0		
0	Responses are inadequate or completely off topic.			

2.4 Research questions

Following Brunfaut and Harding (2014), Knoch and Frost (2016), and Fan et al. (2021), this study adopts a 'twin-panel' approach to compare the judgements of those familiar with the target teaching and learning context, the test takers and the BESTEP (the Taipei Group, or test 'insiders') with those with little background knowledge about the BESTEP but with considerable knowledge of and experience in academic writing and the CEFR (the Melbourne Group, or test 'outsiders'), thus providing a rigorous means of cross-validating the panellists' judgements.

The following three research questions were investigated in this study:

RQ1. How do the score levels of the BESTEP writing test relate to the CEFR levels?

RQ2. How do the judgements of test 'insiders' compare to those of test 'outsiders'?

RQ3. What are the panellists' mental processes when linking the BESTEP writing scripts to the CEFR levels? Are there any differences between test 'insiders' and test 'outsiders' in their linking processes?

3. Methodology

In this study, we collected both quantitative and qualitative data to explore the three research questions. To investigate RQ1, we followed the four steps recommended by the CEFR linking manual (Council of Europe, 2009): familiarisation, specification, standardisation, and validation (also see Figure 1). The quantitative data collected comprised the panellists' ratings of the BESTEP writing samples during the standard setting process, along with the questionnaire data obtained after the familiarisation and standardisation sessions. The panellists' ratings also enabled us to investigate RQ2, which aimed to compare the judgements of 'test insiders' to 'test outsiders' to ascertain any significant differences in their scores. The findings of this research question also provide a means to cross-validate the linking results. Lastly, to address RQ3, qualitative data were collected through a think-aloud study aiming to delve into the cognitive processes of the panellists while they aligned the BESTEP writing samples to the CEFR levels.

3.1 Participants

Fifteen panellists participated in this study, with six test 'insiders' based in Taipei and nine test 'outsiders' based in Australia. Table 2 below presents some of the background details of the 15 panellists. As indicated in this table, the panellists included four males and 11 females. The ages of six panellists fell within the 31-40 range, three within the 41-50 range, five within the 51-60 range, and one was over 60. Regarding educational qualifications, ten panellists held master's degrees, four had doctoral degrees, and one had a Diploma in English Language Teaching to Adults (DELTA), a credential designed for advanced educators in English as a foreign language (EFL) or English as a second language (ESL), awarded by Cambridge English Language Assessment, affiliated with the University of Cambridge. All participants reported familiarity with the CEFR, except I-5 who indicated less familiarity with the framework. Six of the panellists (three from each group) also participated in a subsequent think-aloud study, aiming to explore their cognitive processes when linking the BESTEP writing samples to CEFR levels.

Table 2

Participating panellists in this study

Panelist	Group	Gender	Age	Highest degree	Familiarity with the CEFR	Think aloud study
I-1	Insider	Female	51-60	Master	Familiar	No
I-2	Insider	Female	31-40	Master	Familiar	No
I-3	Insider	Female	31-40	Master	Familiar	Yes
I-4	Insider	Male	31-40	Master	Familiar	Yes
I-5	Insider	Female	31-40	Doctorate	Less familiar	Yes
I-6	Insider	Male	31-40	Master	Familiar	No

O-1	Outsider	Female	61 or above	Doctorate	Familiar	No
O-2	Outsider	Female	51-60	DELTA	Familiar	No
O-3	Outsider	Female	51-60	Doctorate	Familiar	No
O-4	Outsider	Female	51-60	Master	Familiar	No
O-5	Outsider	Female	41-50	Doctorate	Familiar	No
O-6	Outsider	Female	41-50	Master	Familiar	Yes
O-7	Outsider	Male	41-50	Master	Familiar	Yes
O-8	Outsider	Male	31-40	Master	Familiar	Yes
O-9	Outsider	Female	51-60	Master	Familiar	No

When it comes to their current occupation, all test 'insiders' reported that they were involved in various aspects of language testing, including test development, item writing, and test validation research. On the other hand, test 'outsiders' reported occupations primarily within language education, holding positions such as lecturer, curriculum and assessment manager, director of English language institute, assessment specialist, among other roles. While test 'insiders' showed a range of teaching experience (from three to four years to over 30 years), test 'outsiders' uniformly reported extensive experience in English language teaching, ranging from 10 to 35 years (mean = 22.89, SD = 9.12). It should be noted that all test 'outsiders' had experience teaching academic English in higher education. In addition, most panellists reported experience in English language assessment, in areas of item writing, test development, scoring, classroom assessment, and test validation. Most participants also reported their knowledge of the CEFR, with several indicating the experience of working in teaching programs that were mapped to the CEFR, or using the materials or resources that were developed based on the CEFR. Some panellists also had experience with research linking language tests to the CEFR.

3.2 Procedures and materials

As indicated in Figure 1, this study consists of four stages: familiarisation, specification, standardisation, and validation. The participants for each stage, along with the data collection and analysis methods, are detailed below.

3.2.1 Familiarisation

The participants at this stage were 15 panellists, with six 'test insiders' based in Taipei and nine test 'outsiders' from various parts of Australia. Following the CEFR linking manual (Council of Europe, 2009), all panellists went through a familiarisation stage, including a self-paced preparatory session and an online familiarisation workshop on Zoom. The preparatory session aimed to help the panellists get familiar with the CEFR scales and descriptors, especially those related to writing. The panellists were required to work on their own to complete several activities, including a careful review of the writing examples that we provided at different CEFR levels. At the end of the preparatory session, they were asked to complete an online questionnaire about

the effectiveness of this session. Following the preparatory session, an online workshop was set up, during which the panellists engaged in a series of familiarisation activities. For example, they were asked to work in small groups to sort randomised CEFR descriptors and discuss their sorting results. When the panellists worked in groups, conscious efforts were made to mix 'test insiders' and 'test outsiders', following the recommendations from previous linking studies (Fan et al., 2021). At the end of the workshop, the panellists were asked to complete an online questionnaire to report their experience and perceived effectiveness of the workshop.

3.2.2 Specification

The purpose of the specification stage is to analyse the content of the test to be linked in order to profile it in relation to CEFR categories and levels. The research team worked collaboratively with the team at the Language Training and Testing Centre (LTTC), the developer and provider of the BESTEP writing test. The specifications forms in the CEFR linking manual (i.e., A1-A8) were used to analyse content coverage, task types and assessment criteria of the BESTEP writing section. The focus of each form is outlined below:

- A1: general description of the BESTEP writing section
- A2: test development and item writing
- A3: marking test-taker performances
- A4: grading and establishing pass marks
- A5: reporting results to test takers
- A6: analysis of test data and test review procedures
- A7: rationales for decisions made regarding test takers and test revisions
- A8: initial estimation of overall test level

The completed specification forms can be found in Appendix I below.

3.2.3 Standardisation

Prior to linking the BESTEP writing samples to the CEFR levels, a benchmarking workshop was conducted over Zoom with a view to helping the panellists get familiar with the judgement procedures and more importantly, reach an agreement on aligning the BESTEP writing samples to the CEFR levels. The workshop focused on the BESTEP writing test in terms of the task format, the writing constructs, and the rating criteria. Next, the panellists reviewed a few writing samples from a different writing test and assigned a CEFR level to each sample. They then discussed the rationales for their individual level assignment in small groups. Similar to the familiarisation workshop, we made conscious efforts to mix 'test insiders' with 'test outsiders' in these groups. This was followed by a review of a few BESTEP writing samples from different score levels that were already assigned CEFR ratings by the LTTC. Panellists assigned a CEFR level to each writing sample and once again discussed the rationales for their level assignment in small groups, before the LTTC-assigned CEFR ratings were disclosed. In the last part of the workshop, panellists were assigned eight BESTEP writing samples from the training folder. They first worked individually on assigning a CEFR level to each sample before working in small groups to discuss the levels

they had assigned and their rationale. After the benchmarking workshop, the panellists were asked to complete an online questionnaire to report on their experience and the perceived effectiveness of the workshop.

The LTTC provided 80 BESTEP writing samples at different score levels from a recent administration of the BESTEP writing test. When selecting these samples, the LTTC endeavoured to exclude those with a very uneven score distribution on the three tasks (e.g., a high score on Task 2 but a much lower score on Task 3). We set up 16 pinpointing folders, aiming to include five samples in each folder (see Table 3), though this was not possible at the highest and lowest test-score ranges, due to a lack of very low-scoring and high-scoring test takers. This was not an issue, as those extreme scores were expected to be outside the CEFR levels relevant to Taiwan's EMI learning context (A2 to C1).

Table 3
Writing samples included in the pinpointing folders

Folder	Score range	Number
1	350-360	3
2	335-345	7
3	320-330	5
4	305-315	5
5	290-300	5
6	275-285	5
7	260-270	5
8	245-255	5
9	230-240	5
10	215-225	5
11	200-210	5
12	185-195	5
13	170-180	5
14	155-165	5
15	140-150	7
16	135-135	3

Next, we selected a total of 48 samples from the pinpointing folders to set up the range-finding folder. The standard setting in this study followed two rounds. During the first round, the 15 panellists were required to evaluate the 48 samples and assign a CEFR level to each sample based on the descriptors in the CEFR writing assessment grid. After they finished the first round of evaluation, we reviewed all ratings carefully and identified a few outlier ratings, which were sent back to the panellists to review and revise if they deemed it necessary. All outliers were revised,

as panellists reported them being errors (mostly typos, with fewer errors in judgement); all revisions resulted in ratings in line with those of the other panellists.

Based on the evaluation results during the first round, we selected another 28 samples from the range-finding folder which were used for the second round of evaluation. As with the first round, panellists were required to review these samples and assign a CEFR level to each sample. A similar review and revision process was implemented. The data from both rounds were used in the logistic regression analysis to determine the cut scores.

3.2.4 Validation

As noted, we focused on procedural and internal validity in this study. With regard to procedural validity, we meticulously documented the activities in which the panellists engaged at each stage of this standard setting study. Additionally, we distributed three questionnaire surveys to gauge the panellists' perceptions of the effectiveness of the preparatory session, the familiarisation workshop, and the benchmarking workshop. Regarding internal validity, we used many-facets Rasch analysis (e.g., Eckes, 2015; McNamara et al., 2019), a powerful statistical analysis method, to assess the panellists' severity levels when assigning CEFR levels to the BESTEP writing samples. We also computed the intraclass correlation coefficient (ICC) for the combined panel as well as separately for the two groups of test 'insiders' and test 'outsiders' to investigate the reliability of the panellists' judgements.

3.2.5 Think-aloud study

A think-aloud study was conducted to explore the panellists' cognitive processes in mapping the BESTEP writing samples to CEFR levels. Six panellists, three test 'insiders' and three test 'outsiders', participated in this part of the study (see Table 2). Think-aloud, also known as verbal protocol analysis (VPA), is a research methodology that has been widely used in language assessment research to delve into test takers' cognitive processes when engaging with test tasks, thus providing insights into test validity (Green, 1998). Heeding the recommendations from previous research using the think-aloud method, systematic data collection procedures were developed in this study (see Appendix III) to enhance data reliability and validity (Douglas & Hegelheimer, 2007).

Prior to data collection, we trialled the think-aloud guidelines and procedures through a pilot study. Each think-aloud session began with a short training/introductory session to help participants get familiar with the think-aloud procedures. This was followed by a trial evaluation session, during which the participant assigned a CEFR level to a BESTEP writing sample while reporting their reasoning and linking processes. After this initial trial, the participant discussed any issues that emerged in a debrief session with the facilitator. Then, they proceeded to evaluate three samples while reporting their cognitive processes, followed by another debrief session where they shared their thoughts or reported any challenges that they had encountered in the linking process. This was repeated for another six samples across two sessions. All think-aloud sessions were conducted over Zoom and were recorded. The recordings were subsequently

transcribed for coding and analysis. The durations of the think-aloud sessions vary, ranging from 1.5 to 3 hours.

3.3 Data analysis

3.3.1 Many-facets Rasch analysis

We first conducted many-facets Rasch analysis (MFRA) of the panellists' ratings (i.e., the CEFR levels that the panellists assigned to the BESTEP writing scripts). The many-facets Rasch model (MFRM) is an extension of the basic Rasch model where more parameters are added to the analysis (known as 'facets') (Bond & Fox, 2015). This model has been widely used in the field of language assessment, providing a rigorous means for examining rater severity in performance-based language assessment (McNamara et al., 2019). The purpose of this analysis was two-fold: (a) to investigate the panellists' severity levels when assigning CEFR levels to the BESTEP writing scripts; and (b) to detect whether there were any scripts and/or raters that misfit the Rasch model. The data associated with the misfitting scripts and/or raters would be removed from the subsequent linking analysis. Additionally, the analysis results could also help us identify the patterns of severity levels associated with the panellists' group membership (i.e., test 'insiders' and test 'outsiders'). All Rasch analyses were implemented in the Rasch program Facets (Linacre, 2017).

3.3.2 Logistic regression analysis

Several possible ways have been proposed to analyse the Body of Work (BoW) data to determine the cut scores at different levels (Council of Europe, 2009). Among these, logistic regression has been recognised as an effective method for analysing the BoW data (Kingston & Tiemann, 2012). This regression model is used for situations where one or more independent variables are used to predict an outcome, such as 'yes' or 'no', 'success' or 'failure' (Field et al., 2012). Unlike linear regression where the dependent variable is continuous, logistic regression predicts the probability of a binary outcome. It is particularly useful in scenarios where the relationship between the independent variables and the dependent variable is not linear but follows a logistic curve. The model implements the logistic function, which can map any input value to a value between 0 and 1, representing the probability of the dependent variable.

In this study, we used students' scores on the BESTEP writing test as the predictor (or independent) variable to predict the probability of a script being rated 'at or above' or 'below' a specific CEFR level. Once the relationship between the predictor and outcome variable is determined through the model, we set the probability threshold at 0.5 (or 50%) for a script to be classified as 'at or above' or 'below' a certain CEFR level. This threshold was then used to calculate the script's score based on the regression model, which served as the cut score for that CEFR level. Table 4 below outlines the data used in the logistic regression analysis to determine the cut score at each CEFR level. As noted, the data associated with the scripts that misfit the Rasch model were eliminated from this regression analysis. As Table 4 shows, we used the rating data from

six pinpointing folders (i.e., Folders 1-6) for the C1 level analysis, with a total of 405 ratings of writing samples with test scores ranging from 275 to 360. Similarly, the data from another six pinpointing folders (i.e., Folders 5-10) were analysed for the B2 level, with 420 ratings and sample scores from 220 to 300. For the B1 level, the data from eight pinpointing folders (i.e., Folders 8-15) were analysed, with 510 ratings and scores from 140 to 255. Finally, the analysis for the A2 level was conducted using data from four pinpointing folders (i.e., Folders 13-16), encompassing 225 ratings with scores between 135 and 175. All regression analyses were performed in Minitab (Minitab, LLC., 2023).

Table 4

Rating data used for logistic regression analysis at each level

Level	Pinpointing folder	No of ratings	Score range
C1	1-6	405	275-360
B2	5-10	420	220-300
B1	8-15	510	140-255
A2	13-16	225	135-175

3.3.3 Think-aloud data analysis

The think-aloud data was coded inductively and iteratively, following several steps in qualitative data analysis (Miles et al., 2014; Richards, 2014), including 'three concurrent flows of activity: data reduction, data display, and conclusion drawing/verification' (Miles & Huberman, 1994, p. 10). First, we read and re-read the transcripts of the think-aloud data carefully to identify the themes or aspects that emerged from the panellists' verbal protocol reports, such as their linking processes and strategies as well as the challenges they had navigated. As our coding progressed, some themes were merged, and some were categorised under broader themes. This iterative process continued until data saturation was reached, resulting in the final coding scheme (see Section 4.3). Two researchers then applied this coding scheme to code the data collected from one test 'insider' and one test 'outsider', covering about 30% of the total data gathered. To ensure coding consistency, inter-coder reliability was assessed, revealing a high level of agreement (Cohen's kappa = 0.85). Any discrepancies in coding were resolved through discussion. Finally, one researcher coded the remaining data. All qualitative data was analysed in NVivo 14 (QSR, 2012).

3.3.4 Questionnaire data analysis

The purpose of the questionnaires was to investigate the panellists' perceptions of the effectiveness of the three sessions set up for this study: the preparatory activities, the familiarisation workshop, and the benchmarking workshop. The findings provide evidence regarding procedural validity of the linking results. Since the questionnaires were constructed using a five-point Likert scale (i.e., 1 = Strong Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 =

Strongly Agree), we calculated the frequency of the panellists selecting each of the five categories. Due to the small sample size (n = 15), we didn't perform any inferential statistical analyses.

4. Results

4.1 Many-facets Rasch analysis results

The panellists' ratings were analysed using the many-facets Rasch model (MFRM), implemented in Facets (Linacre, 2017). As noted in the methodology section, the ratings were conducted in two rounds (i.e., Rounds 1 and 2). Prior to conducting MFRM on the combined data from both rounds, we also applied MFRM separately to the ratings from each round. The purpose of this separate analysis was to detect any patterns in the severity levels that the two groups of 'insider' and 'outsider' panellists applied when assigning CEFR levels to the BESTEP writing scripts. The analysis would also help us identify the misfitting panellists and scripts, that is, the panellists and scripts whose data failed to fit the expectations of the Rasch model.

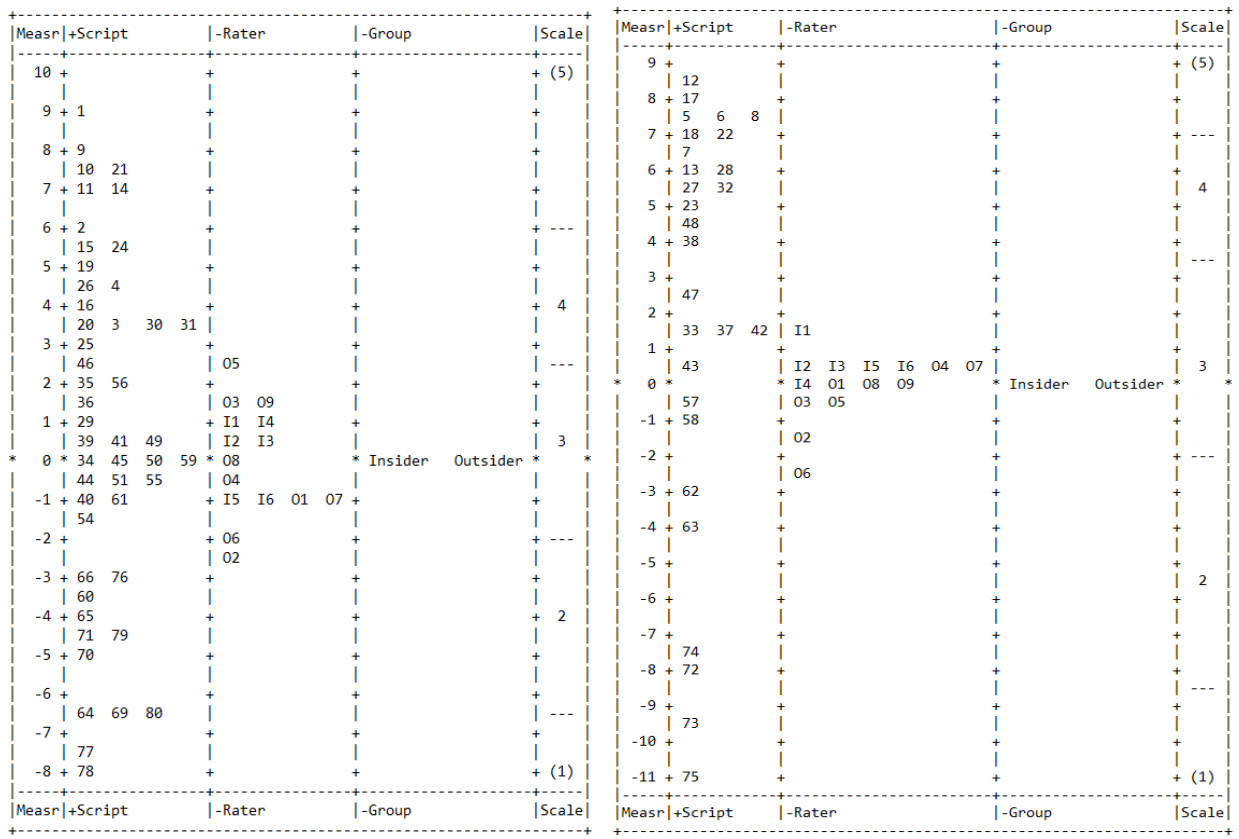
We specified three facets for all three analyses: (a) script (representing the student's writing ability on the BESTEP writing test), (b) rater (representing the panellist's severity when assigning CEFR levels to BESTEP writing scripts), and (c) group (i.e., 'test insiders' and 'test outsiders'). It should be noted that the third facet, 'group', was included in the analysis as a dummy facet, which was constrained to have the same average measure of 0 in Facets. As a dummy facet, it was not included in the main analysis, but was used only for interaction analysis (Linacre, 2017). In this section, we first briefly outline the findings based on the data from Rounds 1 and 2, respectively. We then report the analysis results of the combined data from both rounds in more detail because these data were used to set the cut scores in the subsequent linking analysis.

4.1.1 Analysis of Rounds 1 and 2

Figure 2 presents the Wright maps for the MFRA of the ratings in Rounds 1 and 2. Since the third facet, 'group', was set as a dummy facet, we didn't include it in the Wright maps. As noted, 48 BESTEP scripts were evaluated in Round 1 and 28 in Round 2. The left side of Figure 2 displays the Wright map for Round 1, with the right side depicting that for Round 2. A few observations emerge from these two Wright maps. First, the scripts (representing students' writing ability) cover a wide range of proficiency levels, as indicated by their distribution in the left column of both Wright maps. This is not surprising, given that scripts at different score levels were included in this study. Second, the participating panellists exhibited a range of severity levels when assigning CEFR levels to the BESTEP scripts, as indicated by the distribution of the raters in the right column of each map. Third, compared with 'test insiders', 'test outsiders' appear to exhibit more differences in their severity levels; this is consistent across both groups. In the Wright maps, 'test insiders' (denoted by 'I' in the map) tend to group closely, whereas 'test outsiders' (denoted by 'O' in the map) show more dispersion from each other.

Figure 2

The Wright maps (Rounds 1 and 2)



Notes. I = Insider; O = Outsider.

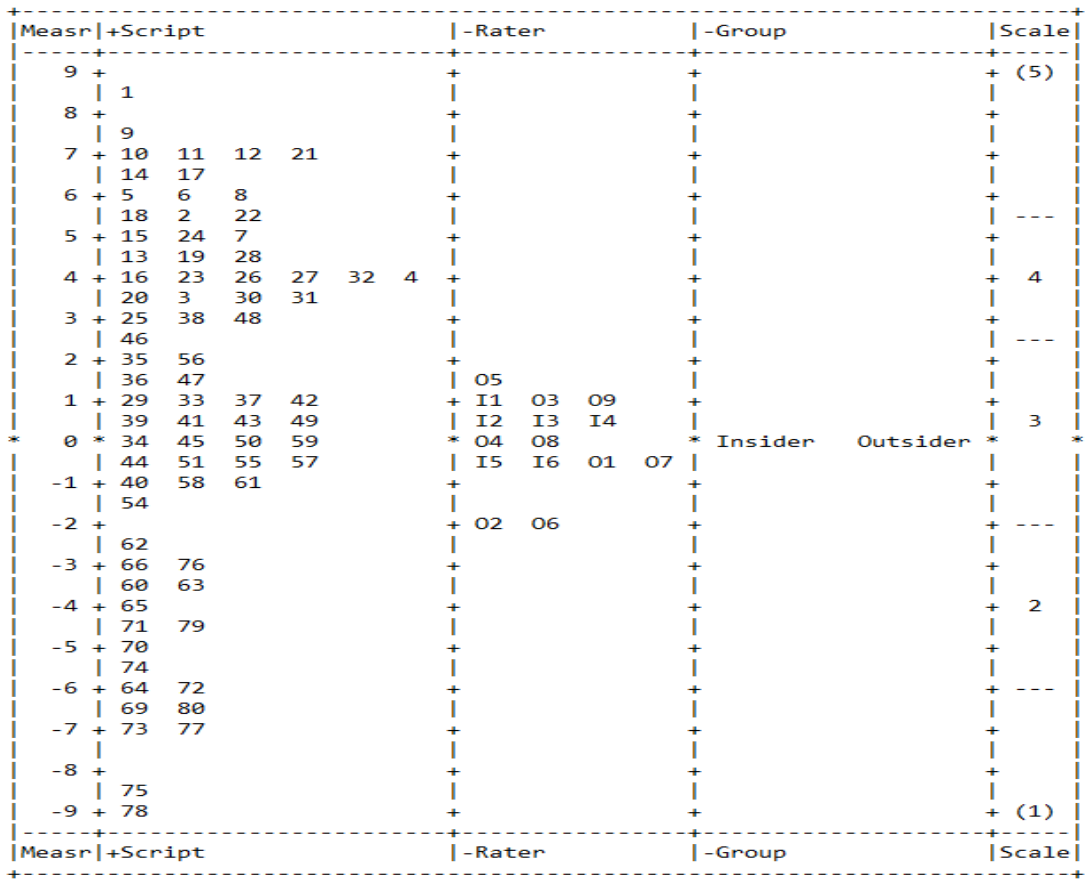
No panellists were found to misfit the model during Round 1, although one rater (O7: Infit MnSq = 0.47) overfit the model. Several scripts were identified as misfitting, with their Infit Mean Square (MnSq) statistics exceeding 1.5 (i.e., 70: MnSq = 1.90, 77: MnSq = 1.90, 29: Infit MnSq = 1.71, and 55: Infit MnSq = 1.56) (Bond & Fox, 2015; McNamara et al., 2019). We also performed interaction analysis between the panellists' group membership (i.e., 'test insider' and 'test outsider') and script; no significant interactions were identified between the two facets. Similar findings were identified for the ratings in Round 2. No panellist was found to misfit the model; however, the infit MnSq statistics of two scripts exceeded 1.5 (i.e., 73: MnSq = 1.59 and 75: MnSq = 1.52), suggesting that they were misfitting. No significant interactions were detected between the panellists' group membership and the script.

4.1.2 Analysis of combined data from both rounds

Figure 3 displays the Wright map from the analysis of the combined data from both rounds. Overall, the findings align well with those based on the data from each round of evaluation, as outlined in the previous section. As is shown in the Wright map, the panellists' severity levels differ significantly. In addition, the severity levels of 'test insiders' appear more homogeneous compared to those of 'test outsiders', who exhibit a wider range of severity levels. This is particularly evident in the case of two 'outsider' panellists (i.e., O2 and O6), who are located at the lower end of the group.

Figure 3

The Wright map (both rounds)



Notes. I = Insider; O = Outsider.

Table 5 below presents the panellists' severity measures, along with their infit and outfit mean square (MnSq) statistics. As indicated in this table, the panellists' severity levels range from -2.20 logits (most lenient) to 1.35 logits (most severe). The infit and outfit MnSq values for all panellists are within the acceptable range of 0.5-1.5, suggesting a satisfactory fit of all panellists to the Rasch model. Notably, one panellist (O7) who failed to fit the Rasch model in Round 1, shows a

reasonably good fit to the model when the data from both rounds are combined. We calculated the average severity measures of the two groups and found that the 'insider' group (0.35 logits) is more severe than the 'outsider' group (-0.23 logits). The primary reason for this disparity is attributed to two raters (O2 and O6, see Figure 3), who were the most lenient in the group based on their severity measures (-2.08 and -2.2 logits). However, since they both fit the Rasch model, we included their data in the subsequent linking analysis. A few scripts were identified as misfitting (29: MnSq = 2.02, 73: MnSq = 1.79, 70: MnSq = 1.72, 77: MnSq = 1.77, 27: MnSq = 1.73, 55: MnSq = 1.54, and 12: MnSq = 1.60). Consequently, these scripts were removed from the subsequent linking analysis.

Table 5

Panellists' measures and fit statistics

Panellist	Measure	Model S.E.	Infit		Outfit	
			MnSq	ZStd	MnSq	ZStd
I1	1.14	0.23	1.11	0.70	1.07	0.40
I2	0.55	0.23	0.68	-2.20	0.67	-2.00
I3	0.55	0.23	0.71	-1.90	0.72	-1.60
I4	0.61	0.23	0.99	0.00	1.01	0.10
I5	-0.50	0.24	0.77	-1.40	0.73	-1.40
I6	-0.27	0.24	1.15	0.80	1.41	1.90
O1	-0.61	0.24	0.86	-0.80	0.89	-0.40
O2	-2.08	0.25	0.93	-0.30	0.87	-0.30
O3	0.77	0.23	1.00	0.00	1.16	0.90
O4	-0.05	0.24	0.84	-0.90	0.81	-1.00
O5	1.35	0.23	1.42	2.40	1.40	2.00
O6	-2.20	0.25	1.06	0.40	1.22	0.70
O7	-0.27	0.24	0.76	-1.40	0.73	-1.40
O8	0.06	0.24	1.25	1.40	1.16	0.80
O9	0.93	0.23	0.95	-0.30	0.87	-0.60

4.2 Logistic regression analysis results

We report the logistic regression analysis results for each level, that is, C1, B2, B1, and A2. The cut score for each level was determined by using the regression model, with the probability fixed at 0.5. Table 6 provides a summary of the logistic regression analysis results across the four levels. Regarding the C1 level, as Table 6 indicates, the BESTEP score significantly predicted whether a BESTEP script should be classified as 'at or above C1' or 'below C1' ($p < 0.01$). The regression model for this analysis is as follows:

Table 6
Logistic regression results

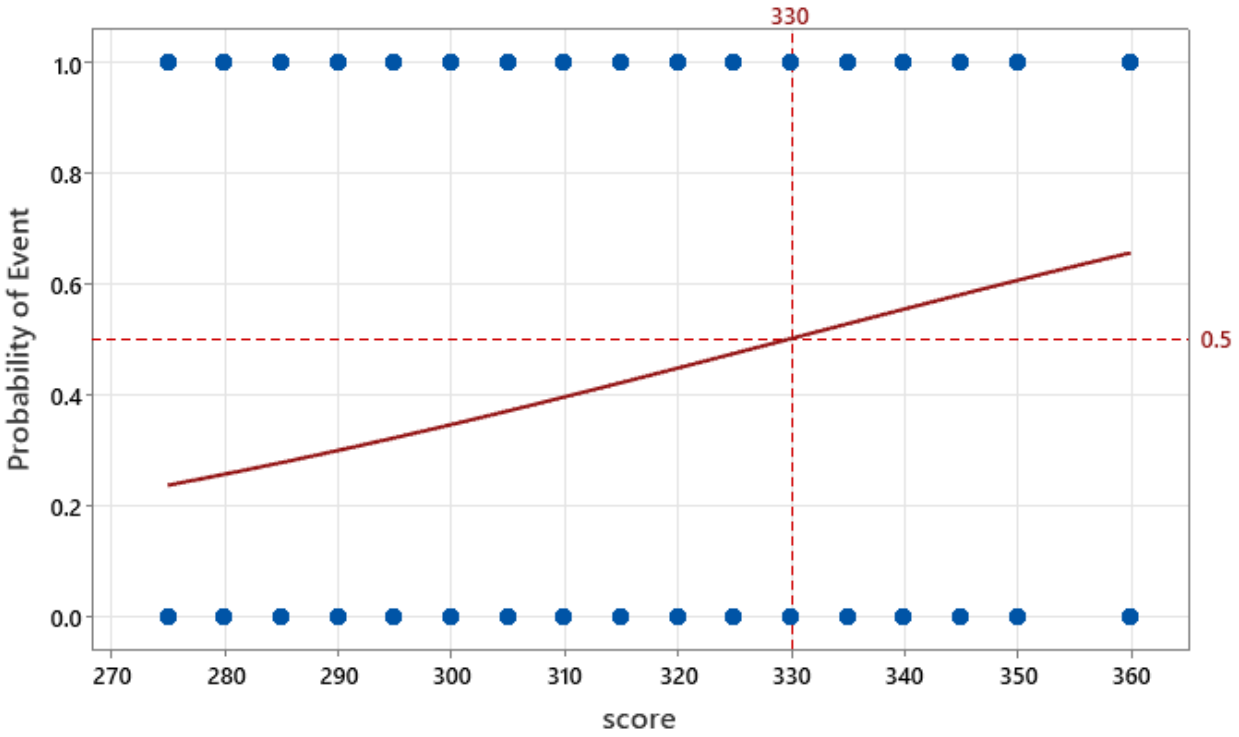
Variable	Coefficient (B)	Standard error	Odds ratio (OR)	95% confidence interval	*p*-value
C1 level					
Intercept	-7.07	1.47			
BESTEP score	0.020	0.004	1.02	[1.01, 1.03]	< 0.01
B2 level					
Intercept	-12.10	1.30			
BESTEP score	0.046	0.005	1.05	[1.04, 1.06]	< 0.01
B1 level					
Intercept	-10.48	0.94			
BESTEP score	0.055	0.005	1.06	[1.05, 1.07]	< 0.01
A2 level					
Intercept	-8.08	1.66			
BESTEP score	0.058	0.011	1.06	[1.04 1.08]	< 0.01

$$P(1) = \exp(-7.07 + 0.02144 * \text{BESTEP score}) / (1 + \exp(-7.07 + 0.02144 * \text{BESTEP score}))$$

As noted, the cut score was determined by setting the probability (P) at 0.5, representing the inflection point when a script is equally likely (i.e., 50% probability) to be classified as 'at or above C1' or 'below C1.' This enabled us to calculate the BESTEP score through the regression formula, which is the cut score for the C1 level (i.e., 7.07 / 0.02144 = 329.75 ≈330). Figure 4 The cut score at C1/B2 below illustrates this relationship, with the probability fixed at 0.5 and the corresponding score of 330 on the *x*-axis.

Figure 4

The cut score at C1/B2



Notes. The blue dots on the x-axis represent scores on the BESTEP writing section, while the probability on the y-axis indicates the likelihood of a script being classified at or above C1.

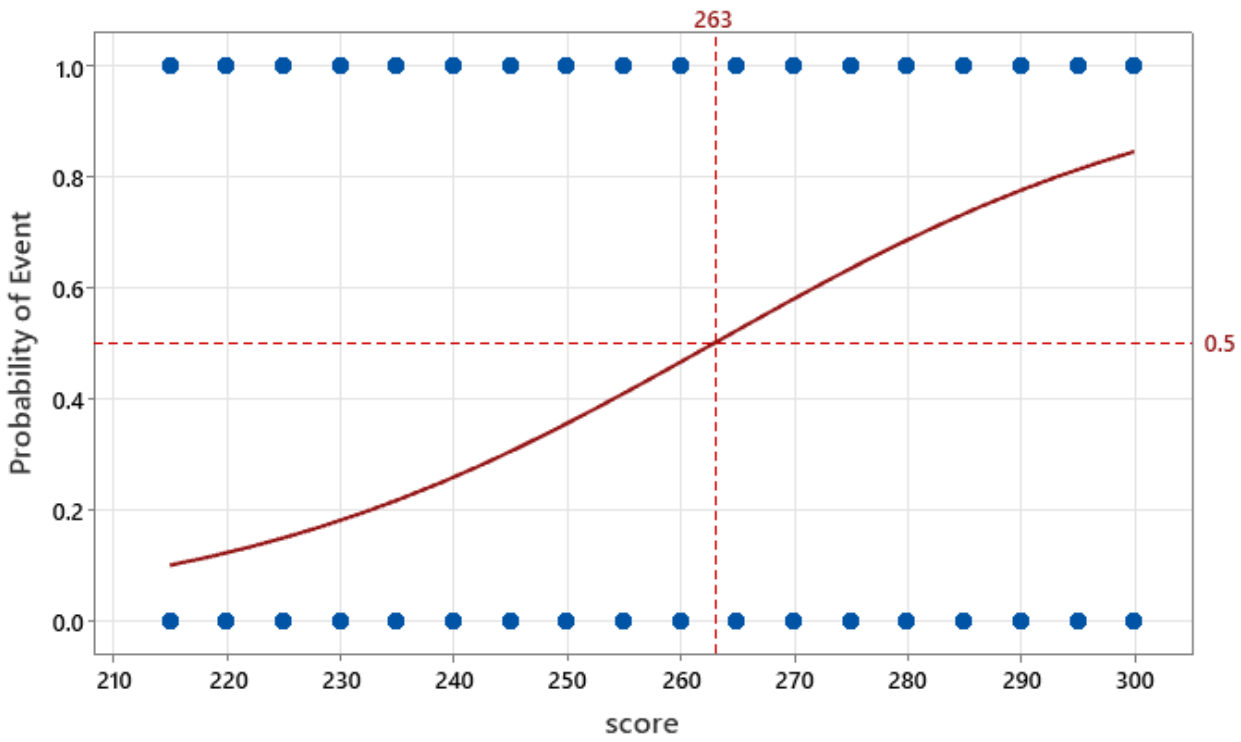
Table 6 indicates that the BESTEP score is a significant predictor for classifying a BESTEP script as 'at or above B2' or 'below B2' ($p < 0.01$). The regression model is presented as follows:

P (1) = exp (-12.10 + 0.04602*BESTEP score) / (1 + exp (-12.10 + 0.04602*BESTEP score)

Figure 5 below depicts this relationship and shows that when the probability is set at 0.5, the corresponding score on the x-axis is 263. Since the BESTEP score increments in intervals of five marks, the cut score for the B2 level is established at 265.

Figure 5

The cut score at B2/B1



Notes. The blue dots on the x-axis represent scores on the BESTEP writing section, while the probability on the y-axis indicates the likelihood of a script being classified at or above B2.

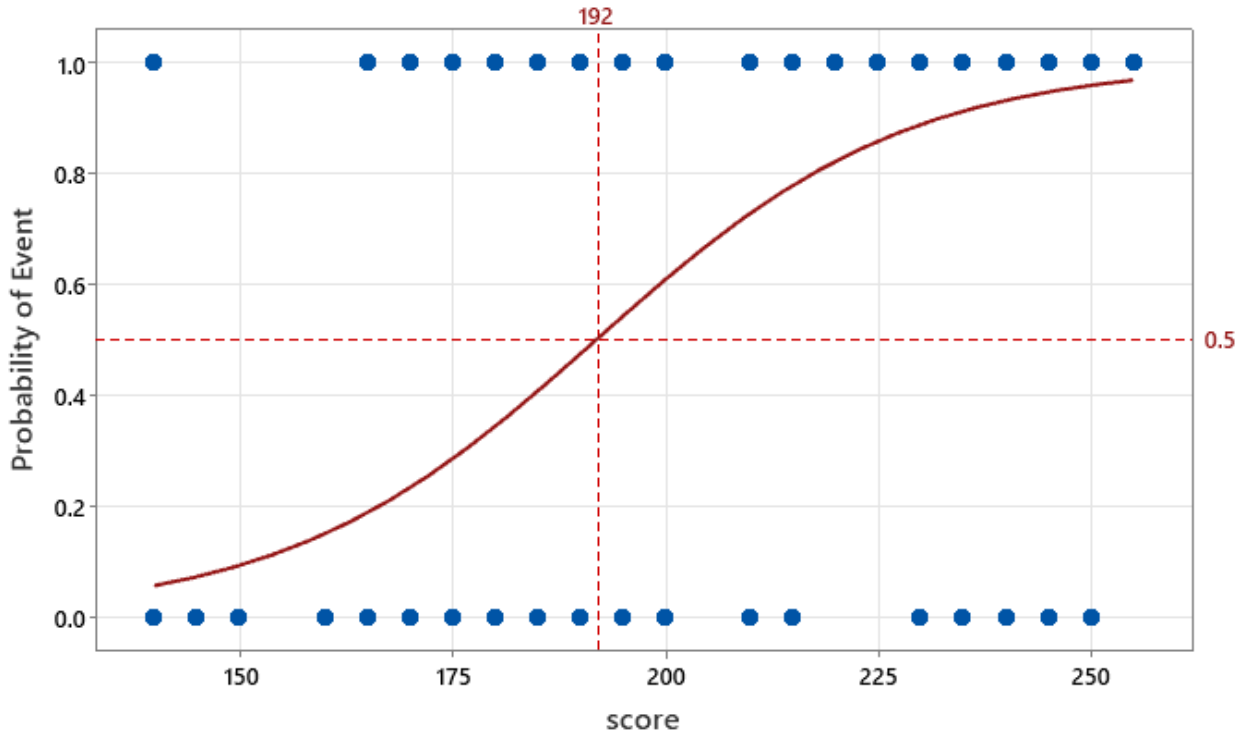
When it comes to the B1 level, Table 6 similarly shows that the BESTEP score is a significant predictor for determining whether a BESTEP script is classified as 'at or above B1' or 'below B1' ($p < 0.01$). The regression model is presented below:

$$P(1) = \exp(-10.48 + 0.05459*\text{BESTEP score}) / (1 + \exp(-0.48 + 0.05459*\text{BESTEP score})$$

Figure 6 below depicts the relationship between the BESTEP score, and the probability of a script being classified as 'at or above B1'. With the probability set at 0.5, the corresponding score on the *x*-axis is 192.

Figure 6

The cut score at B1/A2



Notes. The blue dots on the x-axis represent scores on the BESTEP writing section, while the probability on the y-axis indicates the likelihood of a script being classified at or above B1.

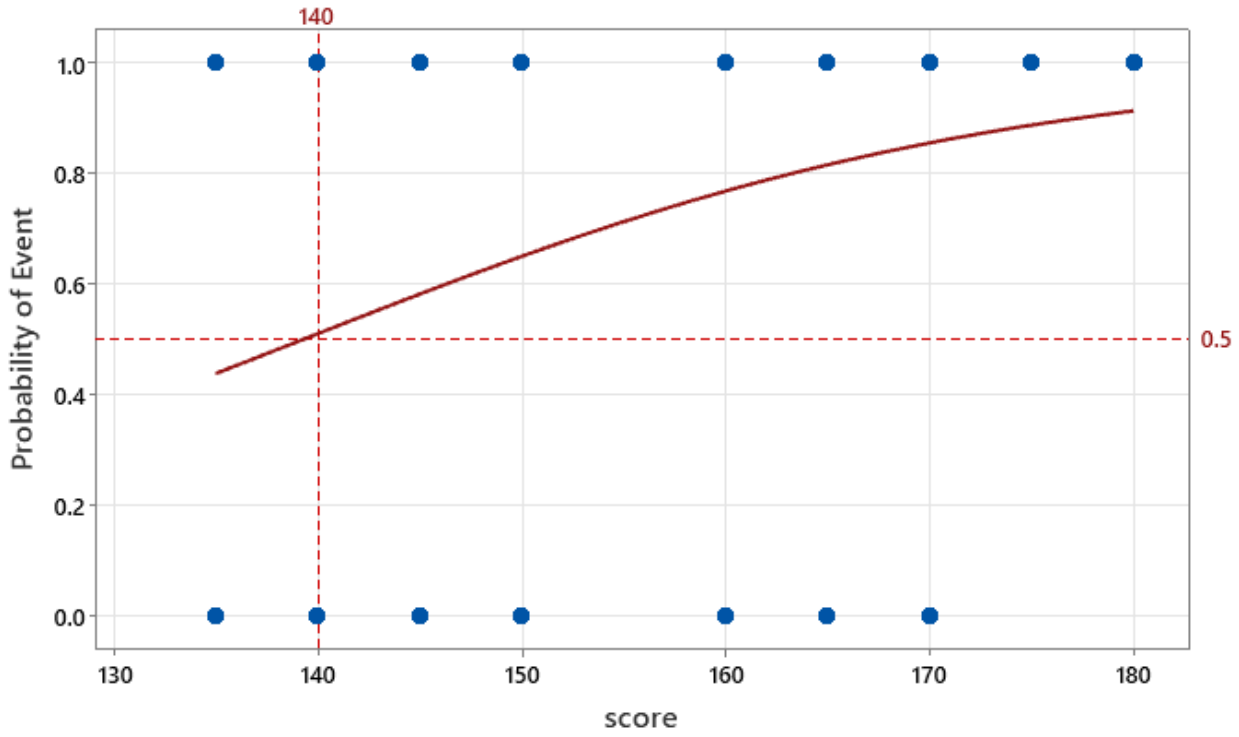
Finally, regarding the A2 level, Table 6 similarly indicates that the BESTEP score is a significant predictor for classifying a BESTEP script as 'at or above A2' or 'below A2' ($p < 0.01$). The regression model for this analysis is presented as follows:

$$P(1) = \exp(-8.08 + 0.0580*\text{BESTEP score}) / (1 + \exp(-8.08 + 0.0580*\text{BESTEP score}))$$

Figure 7 below depicts this relationship. As illustrated in this figure, when the probability is set at 0.5, the corresponding score on the *x*-axis is 140.

Figure 7

The cut score at A2/A1



Notes. The blue dots on the x-axis represent scores on the BESTEP writing section, while the probability on the y-axis indicates the likelihood of a script being classified at or above A2.

In addition to determining the cut score for each level based on the panellists' rating data, we also calculated the 95% confidence interval for the cut scores at each level, adhering to the steps recommended by Kingston et al. (2001). First, we used logistic regression analysis to determine the cut score for each level based on the two rounds of rating data from each of the 15 participating panellists. Next, we computed the mean and standard error of the cut scores for each individual panellist. By so doing, we were able to calculate the 95% confidence interval for the cut score at each level. Table 7 below presents the cut scores set at the four CEFR levels, along with 95% confidence intervals with scores rounded to the nearest 5 points to match BESTEP test score increments.

Table 7

Cut scores for the BESTEP writing test at different CEFR levels

CEFR level	Cut score	95% confidence interval
C1	330	325-350
B2	265	255-290
B1	190	180-220
A2	140	135-145

4.3 Results of the think-aloud study

Table 8 below outlines the coding scheme that we developed, including the themes and subthemes that we identified in the coding process. As indicated in this table, the coding scheme encompasses three broad themes: linking processes, linking strategies, and challenges, each of which is detailed below.

Table 8

The coding scheme of the think-aloud data

Themes	Subthemes	
Linking processes	1) Adjusting and finetuning CEFR levels	a) Make an initial level estimate based on Task 1
		b) Adjust the level based on Task 2
		c) Finalise the level based on Task 3
	2) Task 1 – Answering questions	a) Grammar
		b) Fulfil task requirements
		c) Vocabulary
		d) Mechanics
	3) Task 2 – Expressing opinions	a) Grammar
		b) Fulfil task requirements
		c) Vocabulary
		d) Coherence
		e) Mechanics
	4) Task 3 – Writing an integrated essay	a) Fulfil task requirements
		b) Grammar
		c) Vocabulary
		d) Coherence
		e) Use language from the prompts
		f) Mechanics
Linking strategies	1) Refer to CEFR writing scale descriptors	

	2) Refer to task prompts or requirements 3) Compare with previous samples 4) Demonstrate familiarity with test takers and language use contexts
Challenges	1) Evaluate fulfilment of task requirements 2) CEFR writing assessment grid 3) Task design

As far as linking processes are concerned, our data clearly indicate that the panellists engaged in a dynamic and iterative linking process, typically characterised by three steps of (a) making an initial estimate of the CEFR level based on the test taker's performance on Task 1; (b) adjusting the CEFR level based on their performance on Task 2; and (c) finalising the CEFR level after evaluating their performance on Task 3. While most panellists tended to follow this sequence from Task 1 to Task 3, our data suggest that the process in many cases is not strictly linear. Instead, panellists frequently referred back to the test taker's performance on earlier tasks when assigning a CEFR level. It is important to note that certain aspects of a test taker's performance, such as their use of grammar and vocabulary, were essential in prompting them to revise their initial CEFR level assignments. For example, in Excerpt 1 below, Panellist O-6 revised the CEFR level upward because she was impressed by the complex structures in the test taker's writing.

Excerpt 1 - Panellist O-6

> *Well, that's very complex grammar. That far exceeds what a B2 learner do. So, this person has exposure to C1 grammar and can apply it accurately. The ideas flow together very well. There is a good range of grammar, present perfect, present perfect continuous, and so forth. And the last sentence, 'Thank you for reading my concerns, and I hope you sincerely consider my proposals.' So somewhere at the moment, probably B2 or C1. I am moving up to C1 now, but the last one will probably decide [the CEFR level].*

This excerpt highlights that Panellist O-6 was impressed by the 'very complex grammar' exhibited by the test taker in the email writing task, leading to the conclusion that such performance 'far exceeds what a B2 level learner do [sic].' In addition to complexity, the panellist was also struck by the extensive range of grammatical usage, demonstrated through the use of various tenses and aspects. As a result, she adjusted her initial level estimate to C1, while

acknowledging that this test taker's performance on Task 3 would ultimately determine the final CEFR level.

In Excerpt 2 below, Panellist I-4 revised the CEFR level that he had assigned previously (B1) after evaluating the test taker's performance on the first two tasks. This decision was prompted by the grammatical inaccuracies and errors in vocabulary use that he observed in his evaluations of this test taker's performance on Task 3. Consequently, he downgraded the level from B1 to A2. Our data indicates that this re-evaluation and lowering of the CEFR level, triggered by the panellists' identification of language inaccuracies and errors in test takers' writing performance, is a frequent occurrence.

Excerpt 2 - Panellist I-4

> *I think… so my thinking on this is that perhaps it does satisfy the overall descriptors at B1, but I actually think some of the grammatical inaccuracy and errors in word choice, I think, are too much to justify the B1 level. Yeah. So, I think what I would do, therefore, is I would put this at A2.*

The other three subthemes under 'linking processes' are related to the aspects that the panellists focused on when approaching and evaluating test takers' performance on each task. The aspects in Table 8 are ordered based on their reference frequencies in the coding results. As indicated by Table 8, 'grammar' and 'fulfil task requirements' are the two aspects that the panellists most frequently focused on when evaluating the three tasks. When it comes to 'grammar', the panellists typically commented on the inaccuracies or errors in the test takers' grammatical use, although in some cases, they also commented positively if a test taker demonstrated a strong performance in grammar. The panellists also referred to the task requirements to examine the extent to which a test taker's performance fulfilled these requirements, and the evaluation, in turn, was used to guide their decision-making with regard to the CEFR level. The excerpt below demonstrates how Panellist O-8 evaluated a test taker's performance on Task 3, with the focus on 'fulfil task requirements.'

Excerpt 3 - Panellist O-8

> *So,* the first figure shows the most important ability that students think they have learned from college training. *Some sort of, awkward phrasing which impedes understanding a little bit... I'm not sure if they've summarized at least the 2nd figure well. And then finally, state your own opinion… So, I mean, I would be looking for some sort of, I guess, in terms of language, something that indicates an opinion, which I guess you could argue is evident…[continues reading the test taker's writing]* which I think is the reason that the huge difference pops up to initiatives… Because many students want to make sure they are competitive enough for the societies... *Okay. So, I'm not sure if they necessarily meet the task fulfillment for the 2nd part. Mm. They haven't given their opinion on the topic.*

In this excerpt, Panellist O-8 was scrutinising the test taker's performance, particularly assessing whether the requirement of 'state your own opinion' was fulfilled. Upon evaluation, the panellist observed that the test taker primarily described the discrepancies in the figure without explicitly stating an opinion on the topic. As noted by the panellist, this lack of adherence to the task requirements influenced his decision-making process regarding the assignment of the CEFR level to this performance. The panellists also evaluated aspects such as vocabulary, coherence, and mechanics. For Task 3, whether the test taker copied the language in the task prompts was also one of the considerations.

It is worth noting that Task 1 (Answering questions) is characterised by a rather constrained design and test-taker responses are typically short. As a result, some panellists commented that it was not particularly effective in differentiating test takers' writing ability. Although most panellists started their evaluation with this task, they seemed to give more emphasis to Tasks 2 and 3. Indeed, one panellist (Panellist I-5) chose to bypass Task 1 entirely when assessing a test taker's performance and when assigning CEFR levels.

The second major theme identified in the coding process is 'linking strategies' (see Table 8). Specifically, the panellists predominantly employed four strategies. Not surprisingly, they frequently referred to the CEFR writing assessment grid and descriptors during their decision-making process, comparing a test taker's performance with the descriptors at different proficiency levels. They also paid close attention to task prompts or requirement, assessing the extent to which a test taker had fulfilled these requirements, as an important part of their evaluations of the writing samples that were assigned to them. As noted previously, 'fulfil task requirements' was a key focus in panellists' evaluations throughout all three tasks. Additionally, they sometimes compared the current sample under review with those they had previously evaluated and assigned CEFR levels to.

Excerpt 4 below illustrates that in assessing a writing sample, Panellist I-4 referenced the descriptors at C1 and C2 in the CEFR writing assessment grid. His approach began with reviewing the descriptors at the C1 level, comparing the sample's performance against these

criteria. After concluding that the sample met the C1 criteria, he contemplated whether it could qualify for C2. To this end, he examined the C2 descriptors carefully. However, he ultimately determined that there was no sufficient evidence in the performance indicative of a C2 level.

Excerpt 4 - Panellist I-4

> *Let's see. For argument, at C1 level,* can write clear, well-structured expositions of complex subjects underlining irrelevant salient issues and expand the support …point of view with some subsidiary points, reasons, and examples. *Yeah. I think it's pretty clear that this is a C1 level. Yeah. It is. And, okay, I'll take a look at C2. It* can produce clear, smoothly flowing, complex reports, articles, and essays which present a case or give critical appreciation of... *I feel like this could really be C2. Well, C1, for sure, but really C2? Okay. I'm, I'm kind of in a debate with myself here. Could it be C2? Well, an issue for me is that I don't think I have seen a lot of examples to be confident enough to say that this person is at C2…*

The last major theme relates to the challenges that the panellists encountered during the linking process. Specifically, they noted the difficulty in evaluating 'task fulfilment', and its impact on their decision-making process. This difficulty is exemplified by the test taker's inconsistent performance across the three tasks [despite our efforts to select samples with relatively consistent performance on the three tasks based on task scores and their estimated correspondence to CEFR levels, by referencing Table 1], with notable discrepancies in Tasks 2 and 3. For instance, how should a test taker be holistically evaluated and assigned a CEFR level if they excel in Task 2 but perform poorly in Task 3? Additionally, some test takers displayed jagged profiles in their performance, such as excelling in areas like grammar and vocabulary, but falling short in others like coherence and argument. How to reconcile the disparities in performance across different aspects presented yet another thorny issue for the panellists in their decision-making process.

The panellists offered insights on the CEFR writing assessment grid, and the difficulties in assigning CEFR levels to writing samples that were associated with the grid. For example, one panellist highlighted the lack of descriptors regarding 'relevance of the response to task requirements'. She pointed out that 'a response could be very logical and appropriate but is completely irrelevant to a topic' (Panellist I-3). Essentially, she argued that the CEFR descriptors did not adequately or explicitly cover how well test takers' responses adhered to the given topic or task. This aspect, she noted, was often a critical element in the rating scale that the testing agency used to evaluate a test taker's performance. Another panellist commented on the lack of descriptors for 'pragmatics', thus making it difficult to evaluate the tone and voice of test takers' performance on Task 2 – Expressing opinions/Email writing.

Several panellists provided comments on the design of the BESTEP writing tasks, including their prompts. In the excerpt below, Panellist O-6 critiques the overall design of the BESTEP writing test. She considered Task 1 as unsuitable for an academic writing task, while Task 3 aligned more

closely with academic writing tasks. With regard to Task 2, which involves email writing, this panellist noted from her evaluation experience that while some writing samples exhibited characteristics of academic writing, others appeared more as informal emails and lacked academic features. She recommended that the prompts should more clearly specify whether the task is to write a formal or informal email, and that the rating criteria be adjusted to reflect this distinction.

Excerpt 5 - Panellist O-6

> *Also, you know, like, general academic context. I think there should be something in there about… Now we're hoping for an academic writing for the 3rd task, not for the first one. Maybe for the second one. We've seen some academic and some more actually informal emails. That's fine. But it doesn't say formal or informal. You know, they could actually be specific. Write a formal letter. Write an informal email. But if they do that, then they'll have to adjust the criteria and the genre in the instructions or something. So, the instructions could be clearer.*

We didn't observe notable differences between test 'insiders' and 'outsiders' in their reported linking processes, strategies, and challenges. The only distinction identified was that while test 'outsiders' focused on what test takers at particular CEFR levels were able to do more broadly, the panellists from Taipei (i.e., test 'insiders') tended to highlight specific characteristics of Taiwanese English learners as an L2, such as typical mistakes common to this group and demonstrate their in-depth understanding of the test takers as well as their language learning and use context.

In the excerpt below, Panellist I-5, a test 'insider,' suggested that her shared linguistic and cultural background with the test taker, who translated directly from their first language, Mandarin Chinese, in their writing, facilitated her understanding of the sentence. The shard background enabled her to recognise the nuances of 'direct translation'. She observed that if she 'read them in really an English way', the writing wouldn't make sense. This excerpt indicates that the linguistic and cultural backgrounds of test 'insiders', along with their teaching and assessment experience with Taiwanese learners, equip them to identify unique language usage patterns characteristic of this group of L2 learners.

Excerpt 6 - Panellist I-5

> *So, this is I think it's a little bit of the direct translation from Chinese. I can guess the meaning if I switch to the Chinese mode. But if I want to read them in really an English way, then it doesn't really make sense…*

4.4 Validity evidence

4.4.1 Procedural validity

As noted, we conducted an online survey following the preparatory activities, the familiarization workshop, and the benchmarking workshop. All panellists completed the questionnaire surveys following the first two events, while 12 of them responded to the one related to the benchmarking workshop. Table 9 indicates that all panellists either strongly agreed or agreed that the materials gave them a clear understanding of the objectives of the preparatory session. Most panellists either strongly agreed or agreed that the instructions for each activity were clear, and that the two activities in the preparatory session helped them get familiar with the CEFR scales and descriptors. Furthermore, there was a strong agreement among the panellists regarding the effectiveness of the preparatory session.

Table 9

Results of the questionnaire survey (preparatory activities, n = 15)

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1) The introduction to the preparatory session is clear.	8	7	0	0	0
2) I understand the purpose of this study.	9	6	0	0	0
3) I understand what I was asked to do for each activity.	11	3	1	0	0
4) Activity 1 (that is, Reading Section 3.6 in the CEFR) helps me understand the characteristics of the CEFR levels.	10	4	1	0	0
5) Activity 2 (that is, Reading the writing samples from the Cambridge English tests at different CEFR levels) helps me understand the salient characteristics of the CEFR writing scales.	10	4	1	0	0
6) Overall, the preparatory session is useful for me to understand the CEFR levels.	12	3	0	0	0

Table 10 below presents a summary of the survey findings regarding the familiarisation workshop. As indicated by the frequency statistics in this table, all panellists either strongly agreed or agreed that the workshop offered them a clear understanding of the purpose of this study and a good overview of the CEFR. Most panellists also agreed that the workshop provided

them with a good overview of the BESTEP writing test. Feedback on the activities during the workshop was unanimously positive, with all panellists acknowledging their value. Finally, all panellists strongly agreed that the familiarisation workshop was well conducted.

Table 10

Results of the questionnaire survey (familiarisation workshop, n = 15)

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1) I have a clear understanding of the purpose of this workshop.	10	5	0	0	0
2) I have a clear understanding of the purpose of this study.	8	7	0	0	0
3) I have a good overview of the CEFR.	10	5	0	0	0
4) I have a good overview of the BESTEP writing test.	6	7	1	1	0
5) I understand what I was asked to do for the activities during the workshop.	9	6	0	0	0
6) The activities help me understand the descriptors in the CEFR writing scales.	12	3	0	0	0
7) The activities in the workshop are useful.	9	6	0	0	0
8) Overall, I feel that workshop is well conducted.	10	5	0	0	0

Table 11 below provides a summary of the survey results regarding the benchmarking workshop. As indicated in this table, all panellists understood the purpose of the workshop, with the majority of them having a clear understanding of their expected contributions to each activity during the workshop. Most of the panellists recognised the usefulness and effectiveness of the activities in the workshop, particularly the group discussions. Most panellists reported that the workshop helped them get familiar with the CEFR writing descriptors and the BESTEP writing test. They also agreed on the overall effectiveness of the workshop. These findings were corroborated by the qualitative feedback left by the panellists. For example, one panellist noted that 'the group discussions helped a lot to clarify different concerns from different perspectives'. Another remarked that she really appreciated the facilitator joining the breakout rooms 'to have a chat in order to capture our impressions of the writing.'

Table 11

Results of the questionnaire survey (benchmarking workshop, n = 12)

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1) I understand the purpose of this workshop.	10	2	0	0	0
2) I understand what I was asked to do for the activities in the workshop.	7	4	1	0	0
3) The illustrative examples help me understand the CEFR levels.	6	6	1	0	0
4) The activities in the workshop are useful.	8	4	1	0	0
5) The discussions are helpful.	10	1	1	0	0
6) I feel familiar with the BESTEP writing tasks.	6	5	1	0	0
7) I feel familiar with the CEFR writing descriptors.	5	6	1	0	0
8) Overall, I feel that workshop is well-conducted.	8	3	1	0	0

4.4.2 Internal validity

We computed the intraclass correlation coefficients (ICC), along with their 95% confidence intervals (CI) to evaluate the consistency of the panellists' ratings. The results suggest that the panellists demonstrated a high level of consistency in their evaluations of the BESTEP writing scripts. As a combined group, the panellists achieved satisfactory consistency in their judgements, as indicated by the ICC for the whole group (Cronbach's alpha = 0.982, CI = 0.974-0.988). When examining the two groups separately, the ICC for test 'insiders' was 0.965 (CI = 0.950-0.977), and that for test 'outsiders' was 0.967 (CI = 0.949-0.979), both suggesting a high level of consistency.

5. Discussion and conclusions

This project aimed to link the BESTEP writing test to the Common European Framework of Reference (CEFR), following the four linking stages and good linking practices recommended by the CEFR linking manual (Europe, 2009). Despite the extensive body of research in the field of language assessment on linking language tests to the CEFR (Fleckenstein et al., 2020; Knoch & Frost, 2016; Papageorgiou et al., 2015), two innovative features have distinguished this linking study from previous ones. First, in this study, we utilized the many-facets Rasch model (MFRM) to interrogate the technical qualities of the panellists' evaluation results. The Rasch analysis served as an important quality control mechanism through which the misfitting writing samples were identified and eliminated from the subsequent linking analysis. Furthermore, the Rasch analysis results provided important evidence to support the validity (or lack thereof) of the linking results. Additionally, the interaction analysis in the MFRM also helped us detect any potential interactions between the panellists' group membership (i.e., test 'insiders' and 'outsiders') and the test takers' writing samples. The linking study underscores the MFRM as a rigorous and powerful means to investigate the panellists' rating behaviours and results in a linking study.

Second, in addition to linking the scores of the BESTEP writing test to the CEFR levels, this study also explored the panellists' cognitive processes as they assigned CEFR levels to the BESTEP writing samples through a think-aloud study, including their linking strategies and challenges they encountered. The findings offer valuable insights into the panellists' linking processes, providing useful validity evidence from a unique perspective. This is particularly true for this study in which two groups of panellists were involved, that is, test 'insiders' and 'outsiders', from different linguistic and professional backgrounds.

Rasch analysis results indicate that compared with test 'outsiders', test 'insiders' appeared to be more homogenous in their ratings. This finding may not be entirely surprising. Compared with test 'outsiders', the test 'insiders' were from a similar professional background. For example, they all worked for the LTTC, the provider of the BESTEP writing test. All of them were also involved in the BESTEP project, with comparable experience in language test development and validation. In contrast, test 'outsiders' came from quite diverse backgrounds, ranging from roles such as academic English program coordinators or academic language advisors at Australian universities to language assessment specialists. In addition, compared with test 'outsiders', test 'insiders' were much more familiar with test takers' writing performance on the BESTEP writing test, due to their direct involvement in the BESTEP project. It is important to highlight, however, that despite the broader variance in the evaluation results of test 'outsiders', none of the panellists misfit the Rasch model when we put the two rounds of rating data together. As such, the evaluation data from all panellists were included in the linking analysis, although a few misfitting writing scripts were excluded. Therefore, the MFRA served as an effective quality control mechanism, enhancing the validity of the linking results.

Another noteworthy finding is that compared with test 'insiders', test 'outsiders' were more lenient in their evaluations. This observation aligns with existing research on the rating behaviours of native versus non-native speaker raters in assessing language performance (e.g., Marefat & Heydari, 2016; McNamara, 1996; Zhang & Elder, 2011), suggesting that the two groups prioritised different aspects when applying the rating criteria. Non-native speaker raters, in contrast to their native speaker counterparts, often placed greater emphasis on linguistic accuracy, most notably grammar and vocabulary. Our findings based on the think-aloud study clearly indicate that 'grammar' and 'vocabulary' were the panellists' key focal points across the three writing tasks in the BESTEP writing test. Consequently, it appears that test 'insiders', who are non-native English speakers, focused more on the grammatical and lexical accuracy in test takers' performance, as compared with their native speaker counterparts. Our think-aloud data also revealed that inaccuracies in grammar and vocabulary were the key factors which prompted the panellists to lower their CEFR level. These reasons might explain the relatively higher severity levels of the test 'insiders' in their evaluations.

We conclude this report by summarising the final linking results, the validity evidence supporting these results, and offering recommendations for the LTTC, the provider of the BESTEP writing test, and potentially, other testing agencies planning to link their language tests to the CEFR. As noted, the Body of Work (BoW) method was adopted to link the BESTEP writing test to the CEFR. As part of this standard setting method, the logistic regression analysis was implemented to determine the cut scores at different CEFR levels: 140 for A2/A1, 190 for B1/A2, 265 for B2/B1, and 330 for C1/B2 (see Table 7). The validity of these results was supported by multiple sources of evidence, addressing both procedural and internal validity. Procedural validity was supported by careful documentation of the activities conducted at each stage of this linking study. Internal validity was supported by several types of evidence. For example, intraclass correlation coefficients (ICC) for the combined panel and for each of the two panellist groups demonstrated a high level of consistency of the panellists' ratings. Moreover, the insights from the think-aloud study offered valuable evidence of internal validity from a process-oriented perspective.

Based on our findings, we offer a few recommendations for the LTTC, the provider of the BESTEP writing test, to consider for the future development of this important academic writing test in Taiwan's higher education. Firstly, despite the various types of validity evidence supporting the linking results, we advise some caution in interpreting and using these cut scores. Despite the panellists' qualifications and their reported familiarity with the CEFR, this study involved only 15 panellists. Ideally, a linking study using the body of work (BoW) method should involve a larger sample of panellists. Furthermore, as highlighted by Panellist I-3, 'the CEFR descriptors did not adequately or explicitly cover how well test takers' responses adhered to the given topic or task.' This suggests that the CEFR descriptors overlook the significance of relevance, a crucial aspect in BESTEP scoring. As such, we advise that these linking results should not be taken as definitive figures. Instead, the LTTC should consider the 95% confidence intervals along with the

linking results (see Table 7) when setting the cut scores for each level. In light of our findings, the current cut scores on the BESTEP website for Levels A2, B2, and C1 appear appropriate, although we recommend the LTTC consider lowering the cut score for the B1 level.

Our next recommendation concerns the first task, that is, Answering questions in the BESTEP writing section. As noted by several panellists in this study, this task may not contribute significantly to the variance of test takers' scores on the BESTEP writing test. Indeed, one of the panellists bypassed it entirely during the linking process. In light of this finding, the LTTC may consider re-evaluating the role of this task in the overall BESTEP writing test. Of course, we make this recommendation only from the perspective of linking the BESTEP writing test to the CEFR. This task may be fully justifiable to ensure construct coverage or accommodate the diverse proficiency levels of the target test takers. Secondly, feedback from a panellist in the think-aloud study suggests that there could be improvements to the instructions or prompts for Task 2 (Expressing opinions/Email writing). It might be useful to specify the expected style of the email: Should it adhere to a relatively formal tone to align with the academic writing construct? Such clarification would not only reinforce the academic focus of the test but also provide test takers with a clearer understanding of the expected output for this task.

This study also has a few implications for future linking research in language assessment. Firstly, depending on the nature and purpose of the test, we believe it is advisable, where practical, to include panellists from different backgrounds (e.g., linguistic, cultural, and professional). An ideal panel would include panellists who are familiar with the CEFR and possess experience in language education and/or assessment. In the case of this study, the inclusion of panellists from both Taipei and Australia has strengthened the validity of the linking results. Nevertheless, the involvement of panellists from different backgrounds demands extensive and rigorous training to foster a deep understanding of the CEFR scales and descriptors, the target test, and a consensus on the standards to assign CEFR levels to language performance samples.

Secondly, we recommend the use of the Rasch model to calibrate rater severity levels, and as a quality control mechanism to ensure the statistical qualities of the evaluation results before proceeding to the analysis of the linking data. As demonstrated by this study, the MFRM results can identify misfitting panellists and writing scripts, which should be eliminated from the linking analysis, thereby bolstering the validity of the linking results. Finally, we recommend the investigation of the panellists' linking processes through an introspective and/or retrospective think-aloud study. Such investigations may provide the linking researchers with valuable insights and nuances regarding the linking results, which can be used to support the validity of the linking results.

6. Appendices

Appendix I. Specification Forms – A1-A8

Form A1: General Examination Description

GENERAL EXAMINATION DESCRIPTION	
1. General Information Name of examination	The BESTEP writing test Levels: CEFR A2 to B2
Language tested	English
Examining institution Versions analysed ()	The Language Training & Testing Centre (LTTC)
Type of examination	☐ International ☑ National ☐ Regional ☐ Institutional
Purpose	The BESTEP provides evidence for students' readiness to communicate at CEFR levels A2 to B2 in English in academic contexts, such as in English-mediated instructed (EMI) courses.
Target population No. of test takers per year	☐ Lower Sec ☐ Upper Sec ☑ Uni/College Students ☐ Adult

2. What is the overall aim?

- To cultivate good English communication skills among college and university students so that they can meet the demands of the workplace or international environments.
- To evaluate students' English learning outcomes during their school years to enable educational institutions to continuously improve their teaching quality and provide more effective language instruction.
- To assess students' English proficiency before they enter English-medium instruction (EMI) courses to ensure their smooth adaptation to a learning environment where English is the language of instruction.
- To serve as an important indicator of applicants' English language proficiency for corporations when recruiting and selecting employees.
- To predict candidates' English performance in a work environment to ensure their ability to adapt to the needs of international communication or cooperation.

<table>
<tr><td colspan="4">3. What are the more specific objectives? If available describe the needs of the intended users on which this examination is based.
• To understand the improvement in students' English proficiency after taking English courses at school.
• To assist students in understanding their English proficiency before taking EMI courses.
• To help schools understand the students' basic prerequisite skills for entering EMI courses.</td></tr>
<tr><td>4. What is/are principal domain(s)?</td><td colspan="3">☐ Public
☐ Personal
☐ Occupational
☑ Educational</td></tr>
<tr><td>5. Which communicat-ive activities are tested?</td><td>☐ 1 Listening comprehension
☐ 2 Reading comprehension
☐ 3 Spoken interaction
☐ 4 Written interaction
☐ 5 Spoken production
☑ 6 Written production
☐ 7 Integrated skills
☐ 8 Spoken mediation of text
☐ 9 Written mediation of text
☐ 10 Language usage
☐ 11 Other: (specify): _____</td><td>Name of Subtest(s)
_____
_____

_____
BESTEP writing test

_____
_____
_____
_____</td><td>Duration
_____
_____

_____
50 min

_____
_____
_____
_____</td></tr>
<tr><td>6. What is the weighting of the different subtests in the global result?</td><td colspan="3">Part one: Answering questions (22%)
Part two: Expressing opinions (36%)
Part three: Writing an integrated essay (42%)</td></tr>
<tr><td>7. Describe briefly the structure of each subtest</td><td colspan="3">Part 1. Answering questions
Briefly respond to input related to school life and learning. Test takers are required to answer 3 questions in a total of about 25 words. The suggested response time is 5 minutes.

Part 2. Expressing opinions
Write a short text of email to express an opinion or exchange ideas on learning-related topics. Test takers are required to write about 80 words in approximately 15 minutes.

Part 3. Writing an integrated essay
Summarise the main points from textual and visual inputs on academic topics and express personal opinions. Test takers are required to write about 120-150 words in approximately 30 minutes.</td></tr>
</table>

8. What type(s) of responses are required?	☐ Multiple-choice ☐ True/False ☐ Matching ☐ Sentence writing ☐ Sentence completion ☐ Gapped text / cloze, selected response ☐ Open gapped text / cloze ☐ Short answer to open question(s) ☑ Extended answer (text / monologue) ☐ Interaction with examiner ☐ Interaction with peers ☐ Other	Subtests used in (Write numbers above) ☐
9. What information is published for candidates and teachers?	☑ Overall aim ☑ Principal domain(s) ☑ Test subtests ☑ Test tasks ☑ Sample test papers ☑ Instructional video on test content	☑ Sample answer papers ☑ Marking schemes ☑ Grading schemes ☑ Standardised performance samples showing pass level ☑ Sample certificate
10. Where is this accessible?	☑ On the website ☐ From bookshops ☑ In test centres ☑ On request from the institution ☑ Other: on request from the students who have completed registration	_____
11. What is reported?	☑ Global grade ☑ Grade per subtest	☑ Global grade plus graphic profile ☑ Profile per subtest

Form A2: Test Development

Test development	**Short description and/or references**
1. What organisation decided that the examination was required?	☐ Own organisation/school ☐ A cultural institute ☑ Ministry of Education ☐ Ministry of Justice ☐ Other: specify:
2. If an external organisation is involved, what influence do they have on design and development?	☑ Determine the overall aims ☑ Determine level of language proficiency ☐ Determine examination domain or content ☐ Determine exam format and type of test tasks ☑ Other: specify: National Taiwan Normal University was involved in the early stage of test development, determining performance descriptors and test focuses.
3. If no external organisation was involved, what other factors determined design and development of examination?	☑ A needs analysis ☑ Internal description of examination aims ☑ Internal description of language level ☑ A syllabus or curriculum ☑ Profile of candidates
4. In producing test tasks are specific features of candidates taken into account?	☐ Linguistic background (L1) ☑ Language learning background ☑ Age ☑ Educational level ☐ Socio-economic background ☑ Social-cultural factors ☐ Ethnic background ☑ Gender
5. Who writes the items or develops the test tasks?	Native and non-native item writers, specialised in English teaching and testing fields and familiar with local English learning environments
6. Have test writers guidance to ensure quality?	☑ Training ☑ Guidelines ☑ Checklists ☑ Examples of valid, reliable, appropriate tasks: ☐ Calibrated to CEFR level description ☐ Calibrated to other level description: _____
7. Is training for test writers provided?	☑ Yes ☐ No
8. Are test tasks discussed before use?	☑ Yes ☐ No
9. If yes, by whom?	☑ Individual colleagues ☑ Internal group discussion ☑ External examination committee ☐ Internal stakeholders ☑ External stakeholders
10. Are test tasks pretested?	☑ Yes ☐ No

11. If yes, how?	A small scale pilot test (40 test takers) and nationwide pretests (1,007 test takers) were implemented in September 2022.
12. If no, why not?	
13. Is the reliability of the test estimated?	☑ Yes ☐ No
14. If yes, how?	☑ Data collection and psychometric procedures ☐ Other: specify: _____
15. Are different aspects of validity estimated?	☑ Face validity ☑ Content validity ☑ criterion-related validity ☐ Predictive validity ☑ Construct validity
16. If yes, describe how.	Face validity: The results of questionnaires showed that the majority of test takers found the test materials clear and easily understandable, including instructions and charts. About 80% of participants recognised the familiarity of test topics. Additionally, 90% believed the test effectively measured the English skills they were learning. Adequate writing time was noted by 80% of test takers. Overall, the feedback suggested positive impressions of test relevance and design. Content validity: A panel of educators and assessment specialists, most of whom were university professors in language related fields, formed the advisory committee and reviewed the test tasks of the BESTEP. They considered that the test adequately covers an adequate range of language functions and contextual features. Criterion-related validity: LTTC's research team conducted a study where they administered the BESTEP to 202 university students. They also gathered data on their performance in other language proficiency tests, including GEPT, TOEFL, and IELTS. The correlation coefficients ranged from .45 to .48 ($p < .05$). Construct validity: In 2022, a questionnaire survey and small-scale follow-up interviews were conducted with EAP/EMI teachers and students in higher education institutions to validate the construct validity of the BESTEP performance descriptors and test focuses. The results revealed that the assessed abilities in BESTEP corresponded to the skills required in various academic learning contexts within the country. Furthermore, the speaking and writing

	performance indicators and test focuses were effective in distinguishing between university students of different English proficiency levels.

Form A3: Marking

Marking:	**Complete a copy of this form for each subtest.** **Short description and/or reference**
1. How are the test tasks marked?	For receptive test tasks: ☐ Optical mark reader ☐ Clerical marking For productive or integrated test tasks: ☑ Trained examiners ☐ Teachers
2. Where are the test tasks marked?	☑ Centrally ☐ Locally: ☐ By local teams ☐ By individual examiners
3. What criteria are used to select markers?	Markers are selected and assigned when they meet the following criteria: (1) They are teachers with a background in teaching English for academic purposes (EAP) in universities in Taiwan. (2) They have attended and completed training sessions before the official marking session.
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers/raters ☑ Moderating sessions to standardise judgments ☑ Using standardised examples of test tasks: ☑ Calibrated to CEFR ☑ Calibrated to another level description ☐ Not calibrated to CEFR or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	☑ One holistic score for each task ☐ Marks for different aspects for each task ☐ Rating scale for overall performance in test ☐ Rating Grid for aspects of test performance ☑ Rating scale for each task ☐ Rating Grid for aspects of each task ☐ Rating scale bands are defined, but not to CEFR ☑ Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	☐ Single rater ☐ Two simultaneous raters ☑ Double marking of scripts ☐ Other: specify:_____
7. If double rated, what procedures are used when differences between raters occur?	☑ Use of third rater and that score holds– in the case that the discrepancy between the two marks is significant ☐ Use of third marker and two closest marks used ☑ Average of two marks ☐ Two markers discuss and reach agreement ☐ Other: specify:_____
8. Is inter-rater agreement calculated?	☑ Yes ☐ No
9. Is intra-rater agreement calculated?	☑ Yes

	☐ No

Form A4: Grading

Grading:	**Complete a copy of this form for each Subtest. Short description and/or reference**
1. Are pass marks and/or grades given?	☐ Pass marks ☑ Grades
2. Describe the procedures used to establish pass marks and/or grades and cut scores	Part scores are given by markers in BESTEP based on rating scales referencing descriptions of CEFR levels. Each part is assigned a specific scoring range (0 to 5 points for part one and 0 to 6 points for parts two and three).
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	
4. If grades are given, how are the grade boundaries decided?	Points obtained in all three parts of BESTEP are converted into scale scores and summed together. The resulting total corresponds to a particular CEFR level. This comprehensive approach ensures that the individual part scores are not just numerical values but are linked to the broader CEFR framework, offering a meaningful interpretation of a test taker' s overall language proficiency level based on their performance across the various test tasks.
5. How is consistency in these standards maintained?	During the official marking session, performances of markers are evaluated based on summary statistics, including the mean, standard deviation, distribution of band scores, and correlations. The inter-rater agreement and the intra-rater agreement are calculated to serve as additional statistics for evaluating markers' performances. If the difference between an individual marker's average ratings and the overall rating statistics exceeds an acceptable range (e.g., greater than 2 band scores on a scale of 0 to 5 or 6), the marker is flagged. Individual markers are also flagged if their inter-rater agreements are significantly lower than the average agreement of other markers. The flagged markers then receive a notice to confer with "scoring leaders", who are members of the LTTC R&D teams. In addition, when scores given by two independent markers on the same test taker's response differ by more than 2 band scores, their ratings are considered discrepant. Resolution of discrepancies by a senior marker is required before scores are reported. Scoring leaders monitor markers during the official marking process; if markers are consistently producing discrepancies in marking

	and their scores for certain test takers exceed the acceptable range, their marking scores for those affected test takers will be cancelled and test responses will be rescored by a third marker.

Form A5: Reporting Results

Results	**Short description and/or reference**
1. What results are reported to candidates?	☐ Global grade or pass/fail ☐ Grade or pass/fail per subtest ☑ Global grade plus profile across subtests ☐ Profile of aspects of performance per subtest
2. In what form are results reported?	☐ Raw scores ☐ Undefined grades (e.g. "C") ☐ Level on a defined scale ☐ Diagnostic profiles ☑ Scaled scores
3. On what document are results reported?	☑ Letter or email ☑ Report card ☑ Certificate / Diploma (when applied for) ☑ Online score report:
4. Is information provided to help candidates to interpret results? Give details.	Yes. Test takers will find on their score reports their total score of the test, scores of each part, how well they' ve done on each part in percentages (e.g. if one gets 90 out of 150 on part three, 60% for part three is shown on the report), and the CEFR level of their overall performance.
5. Do candidates have the right to see the corrected and scored examination papers?	No.
6. Do candidates have the right to ask for remarking?	Yes. An experienced marker will undertake the remarking when applied for.

Form A6: Data Analysis

Data analysis	**Short description and/or reference**
1. Is feedback gathered on the examinations?	☑ Yes ☐ No
2. If yes, by whom?	☑ Internal experts (colleagues) ☐ External experts ☐ Local examination institutes ☑ Test administrators ☐ Teachers ☑ Candidates ☐ Parents
3. Is the feedback incorporated in revised versions of the examinations?	☑ Yes ☐ No
4. Is data collected to do analysis on the tests?	☑ On all tests ☐ On a sample of test takers: How large?: _____. How often?:_____ ☐ No
5. If yes, indicate how data are collected?	☑ During pretesting ☑ During live examinations ☐ After live examinations
6. For which features is analysis on the data gathered carried out?	☑ Difficulty ☑ Reliability ☑ Validity ☑ Descriptive Analysis
7. State which analytic methods have been used (e.g. in terms of psychometric procedures).	Descriptive analysis is done to summarise the overall patterns of test takers' performances. Reliability of the assessment is estimated by calculating the Pearson product-moment correlations of the marking scores given by two independent markers of each part. Multifaceted Rasch analysis is also used to investigate candidate ability, marker severity, part task difficulty, the overall reliability, and whether different test forms are parallel.
8. Are performances of candidates from different groups analysed? If so, describe how.	Performances of candidates are analysed in terms of gender (male and female), educational background (graduate and undergraduate), academic year (freshman, sophomore, etc.), educational system (university and vocational college), and department.
9. Describe the procedures to protect the confidentiality of data.	Secure test distribution: the BESTEP ensures confidentiality by securely distributing test materials directly to authorised testing centres. This prevents unauthorised access to test content before the scheduled test date. Individualised test IDs: Each test taker is assigned a unique identification code instead of using personal information. This ensures that individual identities remain confidential during the scoring process.

	Strict proctoring guidelines: the BESTEP provides clear guidelines to test proctors to maintain a controlled testing environment. Proctors are instructed to prevent any communication among test takers during the exam, ensuring that answers and content remain confidential. Limited access to scoring data: After the tests are completed, access to scoring data is restricted to authorised personnel only. This prevents the dissemination of individual test scores without proper authorisation. Secure storage and disposal: Test materials, including answer sheets and test booklets, are securely stored and later disposed of using confidential shredding processes to prevent any leakage of sensitive information.
10. Are relevant measurement concepts explained for test users? If so, describe how.	Test takers are aware that their responses will be scored by two trained and qualified markers, and that the focus for the scoring includes topic relevance and language performance (vocabulary, grammar, fluency, organisation, coherence, etc.). An official guide including sample questions and sample answers of target levels is published along with a full set of mock test. Test takers can also find the electronic version of the guide as well as additional learning resources on the official website of BESTEP.

Form A7: Rationale for Decisions

Rationale for decisions (and revisions)	**Short description and/or reference**
Give the rationale for the decisions that have been made in relation to the examination or the test tasks in question.	The three parts of the writing test are designed to evaluate the test takers' English proficiency in academic contexts. The scoring weightages are determined based on the difficulty of each part. Part one (80 points out of a total of 360 points) assesses test takers' ability to answer simple questions; part two (130 points) focuses on expressing opinions with higher language complexity; part three (150 points) involves data synthesis and advanced writing skills. This weighting aligns with the progression of cognitive demands and the significance of skills in academic and professional contexts.
Is there a review cycle for the examination? (How often? Who by? Procedures for revising decisions)	Yes. The reviewing procedures are conducted from time to time to monitor reliability and validity so that adjustments to the tests can be made when necessary.

Form A8: Initial Estimation of Overall Examination Level

Initial Estimation of Overall CEFR Level		
☐ **A1** ☐	☑ **B1** ☐	☑ **C1** ☐
☑ **A2** ☐	☑ **B2** ☐	☐ **C2** ☐
Short rationale, reference to documentation		

Appendix II. BESTEP-CEFR Linking Study Panelist Background Questionnaire

Q1 Name (will be treated confidentially):

_____

Q2 Highest degree:

- ○ Doctorate
- ○ Master's Degree
- ○ Bachelor's Degree
- ○ Others (please specify) _____

Q3 Please briefly describe your current workplace(s) and role(s):

_____

Q4 Years of experience in teaching English:

_____

Q5 Please briefly describe your type of experience in teaching English (e.g., primary education, secondary education, higher education, adult education):

_____

Q6 Years of experience in English language testing/assessment:

_____

Q7 Please briefly describe your type of experience in English language testing/assessment (e.g., item writing, test design, test validation, marking)

_____

Q8 Are you familiar with the CEFR?

- ○ Yes
- ○ No

Q9 If 'Yes', please provide details:

_____

Q10 Are you familiar with the BEST Test of English Proficiency (BESTEP) ?

- ○ Yes
- ○ No

Q11 If 'Yes', please provide details:

_____

Q12 Do you have previous standard setting experience?

- ○ Yes
- ○ No

Q13 If 'Yes', please provide the details:

_____

Q14 Your age (please select):

- ○ 21-30
- ○ 31-40
- ○ 41-50
- ○ 51-60
- ○ 61 and over

This is the end of the questionnaire. Thank you!

Appendix III. Think-aloud procedures

In this part of the study, we would like to understand what you are thinking as you link the BESTEP writing samples to the CEFR levels. I am going to ask you to think-aloud and describe the mental processes that you engage in your judgement process, including:

- How you evaluate the writing sample in focus
- How you use the CEFR writing scale
- How you relate the features of the writing sample to the CEFR writing scale

In other words, we are interested in understanding your reasoning for deciding the CEFR level for a BESTEP writing sample. This may seem strange at first. With a little practice, I am sure that you will feel more comfortable talking out loud about what you are thinking.

Before you start working on each writing sample, please say aloud its ID number so that we can associate what you report about the linking process to a particular sample afterwards. Take BESTEP 3. Please say aloud 'This sample is BESTEP 3' before you start reporting your linking process.

It is important to talk as much as possible. As mentioned previously, we are interested in understanding your reasoning when linking the BESTEP writing samples to the CEFR levels. We can only know what you are thinking about if you talk out aloud as you work on a sample. If you are silent for some time, I might say "keep talking" in order to remind you to talk.

To reiterate, we want to know not only *what* you are doing, but *why* you are doing it. As you think out aloud, I may ask you to explain what you are thinking further if it is not clear from what you are reporting.

For example, as you go through a writing sample you say out aloud, "This is B2." I will ask you, "Why do you think so?" After a little practice, you should understand what we are asking you to do.

I will model for you an example of someone thinking out aloud as they read the writing sample and try to link it to the CEFR.

7. References

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

British Council, UKALA, EALTA, & ALTE. (2022). Aligning language education with the CEFR: A handbook. https://www.britishcouncil.org/exam/aptis/research/publications/cefr-handbook

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference*. https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG05.pdf

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment (Companion volume)*. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, and assessment (CEFR): A manual*. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Council of Europe. (2018). *Common European Framework of Reference for languages: Learning, teaching, assessment (Companion volume with new descriptors)*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

De Jong, J., Becker, K., Bolt, D, & Goodman, J. (2014). *Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages*. https://assets.ctfassets.net/yqwtwibiobs4/591XSxw7jHUYbtil6j1TfH/4c5485a779d07de94201bcb914017294/Aligning_PTE_Academic_Test_Scores_to_the_Common_European_Framework_of_Reference_for_Languages.pdf

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, *27*, 115-132.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. Peter Lang.

Fan, J., Knoch, U., & Chen, I. (2021). *Linking the GEPT Writing Subtest (Part 1) to the Common European Framework of Reference (CEFR).* Available at: https://www.lttc.ntu.edu.tw/lttc-grants/doc/108report/GEPT_linking_study_Final_report.pdf.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 100420.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge University Press.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. *Educational Measurement*, *4*(1), 433-470.

Kenyon, D. M., & Römhild, A. (2013). Standard setting in language testing. In A. J. Kunnan (Ed.), *The companion tolanguage assessment* (pp. 944–961). John Wiley & Sons, Inc.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 233-262). Erlbaum.

Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The Body of Work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 201-223). Routledge.

Knoch, U., & Frost, K. (2016). *Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)*. https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG08.pdf

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, *13*(1), 32-49.

Linacre, J. M. (2017). *Facets computer program for many-facet Rasch measurement, version 3.80.0*. In http://www.winsteps.com

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing*, *27*, 24-36.

McNamara, T. (1996). *Measuring second language proficiency*. Longman.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and langauge assessment*. Oxford University Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook. (3rd ed.)*. Sage.

Minitab, LLC. (2023). *Minitab [Computer software]*. https://www.minitab.com/

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. Educational Testing Service.

Papageorgiou, S., Wu, S., Hsieh, C. N., Tannenbaum, R. J., & Cheng, M. (2019). Mapping the TOEFL iBT® test scores to China's Standards of English language ability: Implications for score interpretation and use. *ETS Research Report Series*, *2019*(1), 1-49. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/ets2.12281?download=true

QSR. (2012). *NVivo qualitative data analysis software*. QSR International Pty Ltd.

Richards, L. (2014). *Handling qualitative data: A practical guide*. Sage.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31-50.