Relating the BESTEP Speaking Test to the Common European Framework of Reference for Languages

LTTC-BESTEP Research Report
BESTEP-02

Anthony Green and Chihiro Inoue
Centre for Research in English Language Learning and Assessment
University of Bedfordshire, UK

Relating the BESTEP Speaking Test to the Common European Framework of Reference for Languages

Anthony Green and Chihiro Inoue, University of Bedfordshire

1. Introduction

This study was undertaken to corroborate the intended relationship between the Speaking component of a recently launched test of academic language use (BESTEP) and the Common European Framework of Reference (CEFR; Council of Europe 2001). The study employed the staged approach to linking recommended by the Council of Europe (2009a) which includes Familiarization, Specification, Standardization, Benchmarking/ Standard Setting and Validation.

1.1. The BEST program and BESTEP tests

The Program on Bilingual Education for Students in College (BEST, best.twaea.org.tw) was initiated by the Taiwanese Ministry of Education in 2021. The program promotes English as a medium of instruction (EMI) at universities and colleges, contributing to the government's ambitious long-term policy objective of building a bilingual environment in Taiwan by 2030. According to the Ministry, students with a level of English matching the B2 level of the Common European Framework of Reference for languages (CEFR) (Council of Europe 2001) will "possess essential abilities for taking EMI courses [i.e., courses taught through the medium of English], and their professional learning will not be affected when encountering teaching in unfamiliar languages" (Ministry of Education, Republic of China 2021, p.5). Although a B2 level in relevant aspects of language may be an essential baseline for students pursuing university or college-level study, difficulties attributable to language are likely to persist (Deygers et al., 2018). It should be noted that a B2 level provides no guarantee that students will be free from serious language-related challenges that may limit their chances of realizing their potential and achieving academic success (Harsch et al. 2017, Hamnes Carlsen 2018, Green 2018, Aizawa et al. 2023).

To support the BEST project, the Ministry funded the Language Training & Testing Center (LTTC) to develop and administer the BEST Test of English Proficiency (BESTEP). The BESTEP, launched in autumn 2022, is designed to assess English for general academic purposes and English communication skills in international academic and workplace situations. The test consists of four subtests, each assessing one of the traditional four skills of speaking, writing, listening, and reading.

1.2. The BESTEP Speaking test

This report focusses on the BESTEP Speaking test, a semi-direct speaking test that takes approximately fifteen minutes to administer with test takers allowed up to five and a half minutes in total to record their responses. Test takers are seated at workstations and record their responses to audio-visual prompts presented via headphones and on paper. There are three distinct Parts on the test, each intended to be more challenging than the last. In the first Part, test takers answer a series of six questions related to school life and learning. They have 15 seconds to record each response. In Part 2, they are asked to view a chart on a learning-related topic and, guided by three questions, have 90 seconds to prepare and then 90 seconds to respond. They are expected to explain key points from the chart and express their opinions. In Part 3 they have 150 seconds to prepare and then 150 seconds to record a short presentation. They are expected to integrate learning-related information from textual and visual inputs presented on screen and on paper, summarize key points, and compare key information. All responses are recorded by the software and stored for scoring. Responses to each Part are scored independently by trained examiners using a set of rating scales. There is a separate rating scale for each test Part. For Part 1 the rating scale has six levels or *Bands* (0 to 5). Parts 2 and 3 each have rating scales with seven Bands (0 to 6). Taking account of the increasingly demanding nature of the tasks, descriptors appearing at higher score levels in earlier Parts of the test recur at lower levels in later Parts. For example, *Speech is generally fluent* describes a Band 3 performance on Part 2, but a Band 2 performance on Part 3. For reporting purposes, the Bands are converted to a scale with a maximum of 360 points: 80 points is available for Part 1, 130 for Part 2 and 150 for Part 3. Reflecting the BESTEP emphasis on productive language use, the maximum score for Writing is also 360 points with the Reading and Listening test each contributing a maximum of 140 points. The rating scales and samples of other test material can be accessed at bestep.tw.

1.3. Score interpretations

BESTEP scores are intended to be interpreted in relation to the levels of the CEFR and LTTC has published a table displaying initial estimates of the correspondences between BESTEP scores and CEFR levels with Can Do descriptors explaining in general terms how test takers at each level might be expected to use English in academic contexts (Table 1).

Table 1 The interpretation of BESTEP Speaking test scores in relation to CEFR levels (bestep.tw)

CEFR		Score	Description
C1 (and above)		330 ~ 360	See descriptors for B2.
B2	B2+	310 ~ 325	Can give a coherent and well-structured presentation on academic topics in fluent, appropriate English. Can communicate effectively in discussing academic topics in fluent, appropriate English. Can present key points about integrated information and express personal opinions in fluent, appropriate English.
	B2	280 ~ 305	
B1	B1+	260 ~ 275	Can describe or talk about information or experiences related to academic learning in clear, intelligible

	B1	230 ~ 255	English. Can exchange information and ideas about familiar topics related to academic learning in clear, intelligible English.
A2	A2+	180 ~ 225	Can describe or talk about information or experiences related to academic learning in simple English. Can share/exchange information and opinions about familiar academic topics in simple English.
	A2	150 ~ 175	
A1	A1+	130 ~ 145	Does not reach the standard for A2.
	A1	120 ~ 125	
Below A1		0 ~ 115	

1.4. The Common European Framework of Reference for Languages

The aim of the Common European Framework of Reference for Languages or CEFR (Council of Europe 2001, 2020) is to encourage and facilitate reflection, communication and networking in language education. It is intended to "make it easier for practitioners to tell each other and their clientele what they wish to help learners to achieve, and how they attempt to do so" (Council of Europe 2001, p.4) and to "make their... objectives and methods clear and explicit for the benefit of those who use the products of their work" (p.5). There is no requirement or expectation that local curricula or assessments should be "aligned" to the framework in the same way that tests are often aligned to curriculum standards. Instead, reflection questions are posed throughout the framework document, prefaced by "Users of the Framework may wish to consider and where appropriate state…" (Council of Europe 2001, passim.). These are focused on how well the instrument in question meets the specific needs of the language learners making up the assessment population. The critical objective is to "facilitate transparent, coherent alignment" not to the framework, but "between the overall curriculum aims, the detailed objectives teachers use to implement the curriculum, and the assessment of achievement in relation to them" (North and Piccardo 2019, p.156). In the case of language examinations, the CEFR is intended to assist users of the framework to articulate their standards regarding content (the nature of the skills being tested) and performance (levels of proficiency). As explained in the CEFR Companion Volume (Council of Europe 2020), "The reason the CEFR includes so many descriptor scales is to encourage users to develop differentiated profiles. Descriptor scales can be used firstly to identify which language activities are relevant for a particular group of learners and, secondly, to establish which level those learners need to achieve in those activities in order to accomplish their goals." (p.38). For BESTEP, the CEFR thus provides a means of communicating the standards on which the test is based using a shared international terminology that facilitates comparisons with the standards represented by other tests.

Content coverage is addressed by the framework's Descriptive Scheme, explained in the CEFR as its "horizontal dimension" (p.16), a set of parameters for conveying standards relating to language use. The Descriptive Scheme embraces a range of potential content, conceptualizing language use from the perspectives of *Communicative Activities*, *Communication Strategies* and *Communicative Language Competences*. It also provides terms for describing contexts for language use: "constellation[s] of events and situational factors (physical and others), both internal and external to a person, in which acts of communication are embedded." (Council of Europe 2001, p.9). These are elaborated in the framework in Chapter 4, pp.44-56. The complementary "vertical dimension" (p.16) is represented by a set of Common Reference Levels that divide objectives for organized language learning into flexible levels of proficiency that can be further subdivided according to purpose to express different degrees of granularity (Council of Europe 2020, p.26). Bringing together aspects of the horizontal and vertical dimensions of the framework, by 2020 1,832 descriptors had been arranged into over 90 partly overlapping scales that reflect categories of activity, strategy and competence. The 2020 Companion Volume made it very clear that these descriptors are "not in themselves offered as standards" but that they serve as "one source for the development of standards appropriate to the context concerned" (Council of Europe 2020, p.41). It is acknowledged in the CEFR that although the framework is intended to be comprehensive, it is not exhaustive and additional categories, expanding or elaborating on the published framework, may be needed to supplement both the Descriptive Scheme and the Common Reference Levels to better capture learner needs and abilities in specific language teaching or assessment contexts.

1.5. Linking language examinations to the CEFR

In pursuit of its aim to "enhance the transparency… and… facilitate the mutual recognition of qualifications gained in different learning contexts" (Council of Europe 2001, p.1), the Council of Europe recommends that responsible testing agencies should follow a process they refer to as "linking" (Council of Europe, 2009a, p.1) to relate tests to the framework. To support this linking process, in 2009 they released the *Manual for relating language examinations to the CEFR* (Council of Europe, 2009a). In this report we refer to that document as the Manual. In common with many similar projects, the BESTEP Speaking linking study reported here was guided by the Manual.

Linking is explained as a means of expressing local policy intentions through content standards – what learners are expected to know – and performance standards – how much learners are expected to know. To achieve these ends, two stages set out in the Manual are of central importance. *Specification* involves setting the social context for the assessment, describing its content in relation to the CEFR's Descriptive Scheme and relating it to widely accepted quality standards for assessments. This also supports the transparency of comparisons between tests when, for example, judgements are made about their suitability for specific populations of test takers. The Manual realizes Specification through a self-audit by assessment developers of the coverage and quality of their assessments. The process is supported by sets of forms provided in an appendix to the Manual. Complementing Specification, *Standard Setting* involves establishing cut scores that divide performances into two or more CEFR levels, thus mapping local performance standards to the vertical dimension of the framework: the Common Reference Levels. Both stages require a preliminary *Familiarization* with the CEFR and are supported by *Validation* through the collection of evidence that supports the procedures followed and conclusions drawn.

To promote understanding of test results, it is at least as important to use the CEFR to "increase the transparency for teachers, testers,

examination users and test takers about the content and quality of the examination or test" (Council of Europe 2009a, p.27) as it is to label results with a CEFR level or range of levels. Both Specification and Standard Setting are of fundamental importance to the interpretation of results. However, the primary focus in the literature has tended to be on the vertical dimension of levels and Standard Setting. Often, Specification is used merely to support preliminary claims about performance standards rather than to communicate content standards. Among linking studies that have addressed tests of English in academic contexts or other tests used in Taiwan, some have made no mention of the forms and have simply pointed to commonalities between test rating scales and CEFR descriptors (Lim et al. 2013, Fleckenstein et al. 2020). Others have used the Specification forms provided in the Manual but treat these primarily as a basis for preliminary judgements of the proficiency level or levels as reflected in the test material (Brunfaut and Harding 2014, Knoch and Frost 2016, Fan, Knoch and Chen 2021). Little scrutiny has been given to how well the Manual forms fulfil their declared reporting function, helping to convey the principles informing test design to score users and other stakeholders. It has been suggested that the forms themselves are of limited value for this purpose as they convey little practical information about how listed aspects of the Descriptive Scheme are implemented in an assessment. Green, Inoue and Nakatsuhara (2017) proposed an alternative approach, reconfiguring the checklists from the Manual into textual templates, prompting all providers to use the common terms of the Descriptive Scheme to describe their assessments, but in a relatively accessible narrative format. These templates were employed in this study in preference to the original forms provided in the Manual.

2. Methods

2.1. Specification

The Specification procedures outlined in the Manual are intended to "to define and describe clearly the test that is going to be linked to the CEFR" (Council of Europe 2009a, p.27). Members of the BESTEP development team from LTTC were guided in building a description of the test using the Specification templates adapted by the University of Bedfordshire researchers (Green, Inoue and Nakatsuhara, 2017, Appendix A) from the forms provided in the Manual (Council of Europe 2009a, pp.122-180). Familiarisation with the CEFR is built into this process with notes provided in the templates referring users to relevant sections of the CEFR and Companion Volume. The LTTC team was invited to consult the following Council of Europe resources:

- Noijons, Bérešová, Breton, and Szabó (2011) Chapters 5, 6, and 7: a non-technical guide to linking produced by the European Centre for Modern Languages aimed at aimed at "policy makers, assessment experts at examination centres, curriculum developers, teacher trainers and other educationalists" (p.7).

- A checklist of questions from ALTE (2002), namely:

  Users of the Guide who are involved in drawing up specifications may like to consider and where appropriate state:

  - what type and level of language performance needs to be assessed
  - what type of test tasks are necessary to achieve this
  - what practical resources are available, e.g. premises, personnel, etc.
  - what political, social and/or economic issues are likely to influence test development
  - who should be involved in drafting test specifications and developing sample test materials, e.g. in terms of expertise, influence, authority, etc.
  - how the content, technical and procedural details of the test will be described in the specifications
  - what sort of information about the test needs to be given to users, and how, e.g. a publicly available version of the test specifications (p.12).

- ALTE/ Council of Europe (2011): a complement to the Manual that supports assessment development and implementation.

The test development team submitted the completed templates to the researchers who provided feedback on the extent to which they made consistent use of CEFR terms and provided comprehensive information concerning content and quality. After two rounds of feedback, the forms were finalized and circulated to the fourteen judges attending the Standard Setting workshop to inform them about the BESTEP Speaking test (Appendix A).

2.2. Standard Setting

To link performance standards to the CEFR, the Manual, drawing extensively on Cizek and Bunch (2007), suggests a variety of approaches and devotes Section B of its Reference Supplement to the choices available (Kaftandjieva in Council of Europe 2009b). The Manual authors suggest that what they refer to as the *Benchmarking* phase of the linking process (scoring local performance samples against scales derived from the CEFR) is a "special type of standard setting" (p.36) appropriate for use with direct assessments of productive skills, although alternative standard setting approaches are also outlined. Among these is the Body of Work (Kingston, Kahl, Sweeney, and Bay 2001; Kingston and Tiemann 2012) selected as the basis for the procedures followed in this study. The Body of Work is similar to the Benchmarking procedures described in the Manual in that it involves training a group of judges to assign CEFR levels to a selection of local performance samples, but, similar to the Bookmark standard setting method (Lewis, Mitzel, Mercado, and Schulz 2012); the performances are ordered in advance from the lowest to the highest scoring and the task for the judges is to identify the point on the assessment scale at which the quality of student or test taker performance passes from one CEFR level to the next higher level. Like many other standard setting procedures, it also involves a series of discussion rounds in which judges are encouraged to seek consensus on their cut-score recommendations. The Body of Work typically takes place in a series of three stages: the training round, the range-finding round, and the pinpointing round. The training round familiarizes the judges with the Body of Work procedure; the range-finding round identifies areas of the test scale where cut points will be placed; the pinpointing round finalizes or confirms the cut-score recommendations. Relative to the alternatives, the Body of Work is straightforward as it resembles familiar rating practices and avoids probabilistic judgements (Cizek and Bunch 2007).

Discussion and consensus building between the judges is central to the process and the Council of Europe Manual recommends convening a meeting for this purpose. In common with other linking studies on LTTC examinations (Brunfaut and Harding 2014, Knoch and Frost 2016, Wu and Wu 2010, Wu 2014, Green, Inoue and Nakatsuhara, 2017), this project involved convening a diverse panel of experts to act as judges of how the content and levels of the examination related to the CEFR. In this case, the panel included a total of 14 participants made up of six members of the LTTC team (four female and two male) who had worked on the test development. Following Brunfaut and Harding 2014, this group is referred to as 'insiders'. Two of these judges had previously participated in CEFR linking studies. The remaining eight judges, referred to as 'outsiders', were experts in Applied Linguistics or Language Assessment with no institutional affiliation to LTTC. Three (two female, one male) were from universities in Taiwan with experience of the local target context and five (four female and one male) were from the UK university sector with experience of teaching English for academic purposes and previous experience of CEFR linking studies. With the exception of one insider, all judges reported at least five years' experience working in English language education. In common with other recent studies such as Knoch and Frost (2016), because of the logistical challenges involved in setting up a meeting of researchers based in the United Kingdom with test developers and others working in Taiwan, a decision was taken that the panel of expert judges should meet virtually using videoconferencing software (Microsoft Teams).

In applying Body of Work procedures, the researchers faced the challenge that because this was a new test that was not yet in operation, there were comparatively few recordings available. Although it is preferable to include performances representing every possible score outcome, especially around the emergent cut-score recommendations, the limited number of performances available meant that there were areas of the score range that were not well represented. LTTC was able to provide 20 recordings for the researchers and of these, none had been awarded a score lower than 150 points, the initial suggestion for the borderline between A1 and A2. Only two had scores over 330, the initial suggestion for the borderline between B2 and C1. This necessarily restricted the scope for responding to the cut scores emerging from the early rounds by progressively narrowing the range of scores exemplified in later rounds and introducing additional samples with scores close to those cut points. As the interpretation of scores suggested by LTTC (Table 1) is based on performance on the Speaking test as a whole, in accordance with the Body of Work methodology, we asked the judges to assign test takers to a CEFR level based on a judgement of their performance across all three test Parts.

To confirm that the judges remained consistent with a broader international interpretation of the CEFR levels, we included scored recordings from other international tests of English for academic purposes that had been linked to the CEFR through documented procedures consistent with the Council of Europe Manual (Council of Europe 2009a). These included tests from Cambridge English (cambridgeenglish.org/exams-and-tests), IELTS (ielts.org) and LanguageCert (languagecert.org). Additionally, LTTC was able to provide scores on other tests that had been linked to the CEFR for each BESTEP test taker, thus providing an independent estimate of their CEFR levels, although based on different performance samples.

Before their first online meeting, the panelists completed a selection of Familiarization activities recommended for this purpose in the Manual (Council of Europe 2009a). As an initial phase of Standardisation Training, in the *CEFR Training Session* they assign CEFR levels to eight test takers on four video recordings of paired learners taking Cambridge English Speaking tests. These were taken from a DVD published by the Council of Europe with samples of spoken proficiency illustrating the A2 to C1 Common Reference Levels taken from Cambridge English speaking tests (www.coe.int/en/web/portfolio/materials-illustrating-the-cefr-levels). A *CEFR Questionnaire* designed by the researchers invited them to identify the level of 22 descriptors taken from the CEFR scales for *General linguistic range*, *Vocabulary range*, *Grammatical accuracy*, *Vocabulary control*, *Overall phonological control*, and *Sound articulation*. Additionally, 14 descriptors were included from the BESTEP rating scales, allowing direct comparisons between the CEFR and levels or *Bands* of the BESTEP Speaking rating scales. We used Google Forms to host this and all other questionnaires included in this study. Finally, in a modification of the Item-Descriptor standard setting method (Ferrara, Perie and Johnson 2002), judges identified which of the CEFR scales for *Oral Production* and *Oral Interaction* were most salient to each test Part and which levels of those scales best represented the tasks required of test takers.

Two weeks before the first panel meeting, all judges were sent a package of materials for Familiarisation and Standardisation training accompanied by a set of instructions that invited them to complete the following activities:

1. Watch a 15-minute video describing salient features showing samples of spoken English at each CEFR level.
2. Review the description of the salient levels from the Companion Volume (Council of Europe 2020, pp.173-5).
3. Download and review the CEFR and Companion Volume.
4. Assess their own ability to speak a language they had experience of learning using the self-assessment grid for oral interaction (Council of Europe 2020, p.179) and profile their abilities using the linguistic competence scales displayed on pages 129 to 136 of the Companion Volume.
5. *CEFR Training Session*: Via an online form, assign CEFR levels to a series of video clips of learners taking Cambridge English Speaking tests from the *DVD with samples of spoken proficiency illustrating the Common Reference Levels* available via www.coe.int/en/web/common-european-framework-reference-languages/spoken-interaction-and-production. The clips illustrated levels A2 to C1, but do not distinguish between criterion and plus levels. After making their judgements, the judges were shown the relevant explanation for the official ratings provided by the Council of Europe.
6. *CEFR Questionnaire*: Via an online form, assign CEFR levels to a set of 36 descriptors: 22 from the CEFR, 14 from BESTEP.
7. Review the *BESTEP-CEFR Test Description for Specification* prepared by the LTTC development team (Appendix A) and identify which CEFR scales and which level seemed to best represent the communicative language activities required by each Part of the test.

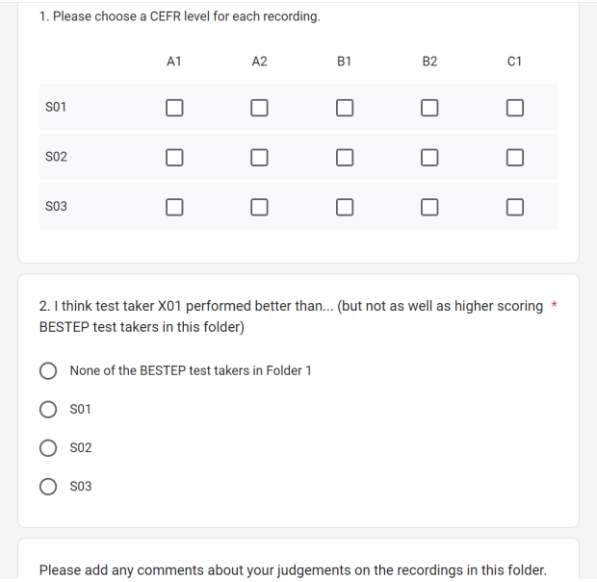
Following completion of the activities, the panel met on three occasions for 90 minutes each time with two or three days between meetings. The meetings were devoted to training activities and discussion while the judgement rounds were completed independently by the judges between meetings. In the meetings, judges were divided into 3 groups (of 4 or 5 members), each of which was chaired by one of the outsiders. This chair was responsible for moderating and recording the discussions, as well as reporting back to the whole group at the end of each meeting. The video recordings and transcripts were analysed for insights into why judges agreed or disagreed in assigning CEFR levels to

performances.

Results of the Familiarization activities were reviewed at the first panel meeting with a focus on the assignment of descriptors to levels as a basis for defining minimally A2, B1, B2 and C1 level performance. This was followed by Standardization Training, corresponding to the training round in the Body of Work methodology. This began with an opportunity during a plenary session to use the scale for qualitative features of spoken language (expanded with phonology) from Appendix 3 of the Companion Volume (Council of Europe 2020, pp.183-5) and the BESTEP rating scales to score two sample performances from the Council of Europe benchmarking seminar organized at the Centre International d'Etudes Pédagogiques (CIEP) in Sèvres, France (www.coe.int/en/web/common-european-framework-reference-languages/spoken-interaction-and-production). This was followed by a brief discussion of the scores awarded. The judges were then divided into three groups of four to five members to score and discuss a further set of four sample performances, again using both the scale for qualitative features of spoken language and the BESTEP rating scales. Although the Manual recommends the use of local samples for Standardisation training, as there was a limited number of BESTEP sample recordings available and to help maintain the connection to an international perspective with previously assigned CEFR levels, this part of the meeting involved the use of performances from other international tests of English that employed test tasks like those encountered on BESTEP.

Between meetings, judges were sent electronic folders containing sets of recordings to rate. The performances (complete, previously scored BESTEP speaking performances) were ordered by overall BESTEP score and these scores were shared with the judges. They were instructed to rate one recording at a time, and to provide a holistic CEFR level for each participant based on their performance across the three BESTEP tasks. As the scales provided in the Companion Volume (Council of Europe 2020) do not differentiate between the "criterion" levels (e.g., B1) and "plus" levels (e.g., B1+) and as our experience with Familiarization suggested that it would be challenging for them, the judges were not asked to make this distinction. Following the first meeting, the judges worked at home to assign CEFR levels to a set of 20 recordings (fifteen from BESTEP, five from other international test providers that were accompanied by CEFR ratings from those test providers. In this initial range-finding judgement round, the judges were asked to use the CEFR scale for qualitative features of spoken language (Council of Europe 2020, pp.183-5) to rate each of the 15 BESTEP performances and to compare these with the recordings from other tests: locating each recording between the two BESTEP performances closest to it in level. Figure 1 shows a screenshot of the online form used to collect responses to the judgement rounds between the online meetings.

Figure 1 Screenshot of the online form (Round 1 judgements)



Based on the results from the range-finding judgements, initial cut-score ranges were identified. In a second online meeting, the judges were shown summaries of the scores they had awarded to each performance. They then, as in the first meeting, divided into small groups of four or five and discussed any differences between them in their scoring decisions. They tried to reach consensus by justifying their scores with evidence from the performances and from the CEFR scales. In the final session of the meeting, they reported back on these discussions to the full panel.

Following the second meeting, the judges again worked individually, moving on to the next, pinpointing round of judgement. They were again presented with fifteen recordings from BESTEP, but three of the original (range-finding) set that fell outside the cut score ranges were replaced with new recordings drawn from the selection supplied by LTTC that had scores falling within the cut-score ranges identified in round one. The five recordings from international test providers were replaced by three different recordings, also drawn from international tests that had been linked to the CEFR.

In the final meeting the judges were shown a summary of the results from the pinpointing round before dividing into groups for discussion. As in the previous rounds, the groups reported back to the full panel. Following this meeting, the judges worked individually to confirm their final judgements. They were given a set of all the BESTEP recordings they had previously heard and were asked to place cut points between test takers for A2/B1, B1/B2, and B2/C1. As no test performance had been rated lower than A2 by more than two of the 14 judges, no attempt was made to identify a cut score to distinguish A1 from A2. Lewis et al. (2012), suggest that cut score recommendations may be calculated using

either logistic regression based on the results of the pinpointing round or by taking the midpoints between the highest scored test taker placed at one CEFR level and the lowest scored test taker at the next higher level. Taking into account the limited number of performances and the risk that outliers might influence the results of regression analysis, we chose to take the midpoints as the panel recommendations for cut scores.

Finally, after submitting their final recommendations, the judges were asked to complete a feedback questionnaire (see Table 8) based on the Sample Evaluation Form for Standard-Setting Participants in the Manual (p.62). Among other questions, this asked whether the training provided was helpful for the judges to understand how to perform their task; whether they had a clear understanding of the purpose of the standard setting meeting; and how confident they felt in the standard setting process and in the resulting cut scores.

3. Results

3.1. Specification

The completed *BESTEP-CEFR Test Description for Specification* templates are provided in Appendix A. In the context of this study, these were used to familiarize judges with the content and quality of BESTEP. Among the activities judges undertook before the first meeting, they used information from the templates to judge which CEFR scales and which levels seemed to best represent the communicative language activities required by each Part of the test. The results are summarised in Table 2.

Table 2 CEFR scales and levels identified by judges with each BESTEP Speaking test Part

CEFR scales	No. of judges			No. of judges				
Oral production	Part 1	Part 2	Part 3	A1	A2	B1	B2	C1
Sustained monologue: Describing experience	8	8	3	2	7	10	2	2
Sustained monologue: Giving information	6	9	6		5	9	4	1
Sustained monologue: Putting a case (e.g. in a debate)		9	13			6	11	3
Addressing audiences	1	1	11		1	4	10	2
Oral interaction	Part 1	Part 2	Part 3	A1	A2	B1	B2	C1
Understanding an interlocutor	8	1	1		7	4		
Information exchange	6	2	1		4	6	1	
Conversation	5	1			5	3		
Informal discussion (with friends)	5	2	1		2	7	3	
Formal discussion (meetings)		2	5			1	5	2

The columns in the section on the right of Table 2 show the number of judges identifying each CEFR scale with each of the three BESTEP test Parts. The columns in the section on the left show the number of judges who considered that each level of the relevant CEFR scale reflected the tasks included in the BESTEP Speaking test. They were not asked to pick out which of the descriptors presented for each level that they found most pertinent, but selection of a scale by five or more judges is taken as evidence for its salience. Details of the descriptors identified as relevant are listed in Appendix B.

The results of this exercise suggested a clear progression in the BESTEP speaking test from tasks that involve describing experience and giving information in Part 1 to putting a case and addressing audiences in Part 3 with Part 2 bridging the two. Part 1 was seen by judges more clearly to engage elements of oral interaction as described in the CEFR than Parts 2 and 3. The tasks were generally viewed as reflective of the A2 to B2 levels with movement from A2/B1 in Part 1 towards B2 in Part 3. This is consistent with the test developers' intention that the tasks should become more demanding as the test progresses. No CEFR scale other than those listed in Table 2 for *Oral Production* or *Oral Interaction* was suggested by more than three judges for any test Part. In other words, the *Oral Production* scale for *Public Announcements* and the *Oral Interaction* scales for *Goal-oriented co-operation, Obtaining goods and services, Interviewing and being interviewed*, and *Using telecommunications* were not generally considered relevant to the BESTEP Speaking tasks.

3.2. Familiarisation

One of the expert judges was unable to complete the online questionnaires, leaving 13 complete sets of responses. The results for the initial online *CEFR Training Session* undertaken before the first meeting are displayed in Table 3.

Table 3 Results of the online CEFR Training session

Panellist scores		Sample 4:1	Sample 4:2	Sample 1:1	Sample 1:2	Sample 3:1	Sample 3:2	Sample 2:1	Sample 2:2
		Score: A2	A2	B1	B1	B2	B2	C1	C1
	A1	1	1						
	A2	**10**	**10**	5	1				
	B1	2	2	**7**	**12**		1		
	B2			1		**13**	**10**	7	7
	C1						2	**6**	**6**

The eight test takers on the four video recordings sampled from the Council of Europe/ Cambridge DVD were rated by 13 judges giving a total of 104 ratings. Of these, CEFR levels were correctly assigned on 74 occasions (71.2% of the total), matching the Council of Europe ratings. 23 (22.1%) individual judgements were at the CEFR criterion level below the Council of Europe rating, seven (6.7%) at the level above. All incorrect CEFR level assignments were to a criterion level adjacent to the Council of Europe rating. Two judges (both outsiders: external experts rather than internal LTTC staff) assigned all eight performances to the correct level (correct assignments are displayed in bold type in Table 3). Two judges, also both outsiders, only assigned three performances to the correct level. Both of these judges seemed to display a degree of central tendency, avoiding assigning high or low CEFR ratings by assigning both C1 level performances to B2 and both A2 performances to B1. Overall, the activity suggested that the judges had developed a reasonable understanding of the Common Reference Levels by this stage with a majority assigning the correct CEFR level to each performance with the exception of the two C1 performances, which were both judged to be B2 by seven of the 13. The C1 level was therefore identified as an important focus for training in the initial meeting.

During the initial meeting, the judges viewed and discussed the four sample recordings of paired speaking tests from the Council of Europe/ Cambridge DVD. Judges were generally in agreement as to which levels the learners were at in both the CEFR and on the BESTEP Part 3 scales, clarifying why the two C1 level test takers were judged to be at that level. The outcomes of these discussions are summarized below.

Sample 1: Deciding prices publicity brochure – B1

- These candidates were both fluent and spontaneous (matching descriptors at B2), but the male candidate's phonological control was lower than B2 because he was not intelligible throughout. The female candidate's pausing and repair towards the end were evident (B1).
- Their use of more difficult words such as 'publicity' and 'reputation' made some judges feel their vocabulary range might be at B2. However, these were given in the task sheet, so were not necessarily being used spontaneously.
- On the BESTEP scale, they were Level 4. The descriptors for Levels 4 and 5 are almost identical (except for relevance to the topic), and judges felt that the candidates' responses were not good enough to reach the 'adequate' stipulated at Level 5.

Sample 2: Places to eat – A2

- These candidates can answer questions and respond to simple statements (A2), using simple sentences, short phrases or simple utterances to get by (A2)
- BESTEP Level 1 because they show some control of basic syntactic structures and a limited control of vocabulary, although they did not make many errors (as in the descriptors). They couldn't elaborate and their responses were only partially relevant.

Sample 3: Long distance travel and internet banking – C1

- These candidates produced clear, smooth-flowing, well-structured speech and they expressed themselves fluently and spontaneously, almost effortlessly (all C1 descriptors)
- BESTEP Level 6, matching all the descriptors.

Sample 4: Delegating work and attracting staff – B2

- These candidates were fluent but there were more pauses and self-repairs. They were hesitant at times as they searched for patterns and expressions (B2). They had a 'sufficient' (B2) rather than a 'broad' (C1) range of vocabulary.
- BESTEP Level 5, not as strong as Level 6, but both addressed the task (adequate and relevant).

The results of the online familiarisation activities that were given prior to the panel meetings also demonstrated a good degree of understanding of the CEFR levels and descriptors from the experts. A vast majority of the experts were able to identify a descriptor at the correct CEFR level or an adjacent level (for details, see Appendix C).

3.2.1. CEFR Descriptors

Overall, 60.5% of the 286 CEFR descriptor assignments were correct and 82.2% were within one level. 14 of the 22 CEFR descriptors were correctly assigned by the majority of judges. Of the remaining eight descriptors, more judges assigned four of these to the correct level than to any other level and one (*Prosodic features, B2: Employs prosodic features – e.g. stress, intonation, rhythm – to support the message they intend to convey, though with some influence from the other languages they speak*) was evenly divided between the correct level (B2) and the level below (B1+) with four judges choosing each. There were three descriptors that were assigned to an incorrect CEFR level by the highest proportion of judges. All three represented plus levels:

- General linguistic range, B1+ Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and film (assigned to B2 by five judges).
- Grammatical accuracy, B1+ Communicates with reasonable accuracy in familiar contexts; generally good control, though with noticeable mother-tongue influence. Errors occur, but it is clear what they are trying to express (assigned to B1 by seven judges).
- Prosodic features, A2+ Uses the prosodic features of everyday words and phrases intelligibly, in spite of a strong influence on stress, intonation and/or rhythm from the other language(s) they speak (assigned to A2 and B1 by five judges each).

3.2.2. BESTEP Descriptors

For BESTEP, the assignment of descriptors was broadly in line with the intentions of the test developers with CEFR level assignments increasing in line with the BESTEP Bands. Band 0 on Parts 2 and 3 was unambiguously associated with CEFR A1 and Band 6 on Parts 2 and 3 clearly aligned with CEFR C1. Bands 1 and 2 on Part 2 and Band 1 on Part 3 seemed to represent elements of both A1 and A2 suggesting that there might be scope to better differentiate these Bands in the BESTEP scales. Descriptors appearing at Bands 3 and 4 on Part 2 and at 2 and 3 on Part 3 were generally identified with B1. The exception was the descriptor Speech is generally fluent which most judges (seven of the 13) identified with B2, although one, an LTTC insider, placed it at A2. All descriptors appearing at Band 5 on the scale for Part 2 and at Bands 4 and 5 on Part 3 were assigned to B2 by the highest proportion of judges. On this evidence:

- On Part 1 of BESTEP Speaking, Band 5 reflects B1 on the CEFR.
- On Part 2 of BESTEP Speaking, Bands 1 and 2 reflects aspects of A1 and A2; Bands 3 and 4 mainly reflect B1, but with elements of B2, Band 5 reflects B2 and Band 6, C1.
- On Part 3 of BESTEP Speaking, Band 1 reflects A1 and A2, Bands 2 and 3 mainly reflect B1, but with elements of B2, Bands 4 and 5 reflect B2 and Band 6 reflects C1.

The judges were generally consistent in their judgements with at least six of the 13 (8.4 on average) agreeing on the assignment of each descriptor with ten or more (average 11.2) within one level above or below the majority choice.

Consistent with the evidence from the *CEFR Training Session*, the results of the *CEFR Questionnaire* suggested that judges had a good understanding of the CEFR Common Reference Levels by the time of the first panel meeting. One notable concern was that the judges were not confident in assigning descriptors to plus levels, doing so less than half as often on average (18 times each) as to criterion levels (46.4 times each). Of these, only 32.7% were correct. The same reluctance was even more marked for BESTEP descriptors which were assigned to plus levels just 9 times on average compared to 29.2 times for criterion levels. No BESTEP descriptor was identified by a majority of judges with a CEFR plus level. This suggested that it might be better to avoid the use of CEFR plus levels in judging test taker performance in this project.

3.3. Standard Setting Round 1 judgements

Table 4 BESTEP test takers' scores with judges' Round 1 assignments to CEFR levels

BESTEP Test Taker: Score	***Part 1***	***Part 2***	***Part 3***	***Alternative test***	***Score***	***CEFR level****	***A1***	***A2***	***B1***	***B2***	***C1***	***Panel mode***
S01: 150	2	1	1	IELTS	6.0	B1	2	9	3			A2
S02: 185	3.5	1	2	TOEFL iBT	17	B1		9	4	1		A2
S03: 200	3.5	2	2.5	GEPT Int	60	A2	1	12	1			A2
S04: 200	3	2.5	2	GEPT Int	70	A2+		7	7			A2/B1
S05: 215	3.5	3	2	GEPT Int	60	A2		5	9			B1
S06: 220	4	3	2	GEPT HI	70	B1+			11	3		B1
S07: 240	4	3.5	3.5	TOEIC SW	130	B1			5	9		B2
S08: 260	4	4	4	GEPT Int	80	B1+			6	8		B2
S09: 260	4	4	4	TOEFL iBT	16	B1		1	6	7		B2
S10: 270	5	4	3.5	TOEIC SW	150	B1			2	12		B2
S11: 300	5	5	4	TOEFL iBT	22	B2			2	10	2	B2
S12: 310	5	5	5	TOEFL iBT	24	B2			1	10	3	B2
S13: 310	5	5	5	TOEFL iBT	22	B2			1	7	6	B2
S14: 360	5	6	6	TOEFL iBT	28	C1					14	C1
S15: 360	5	6	6	TOEFL iBT	25	C1			1	6	7	C1

* Note that this is the CEFR level derived from performance on the alternative test, not the level assigned by the judges on the panel. As these are based on different test tasks and were obtained at different points in time, they might not be expected to align closely with the panelists' judgements.

Following the first online meeting, judges worked individually to make their first range finding round of judgements. These are displayed in Table 5. The judges as a group agreed with the ranking of performances suggested by the BESTEP scores, although there were some minor discrepancies. For example, test taker S07 with a BESTEP score of 240 was assigned to B2 by nine judges, but test taker S09 with the higher BESTEP score of 260 was only assigned to B2 by seven (with six choosing B1 and one A2). Where two test takers were awarded the same BESTEP score, their CEFR ratings might be relatively distinct. The panel generally agreed about S08 and S09 who scored 260 on BESTEP. These two test takers both divided the judges with most placing them at B2, but just one or two fewer placing them at B1. In contrast, S14 and S15 both scored 360, but where S14 was unanimously placed at C1, judges were equally divided over whether S15 satisfied minimum C1 requirements: one placing this test taker at B1 and six at B2.

With respect to the samples from other international tests, the A1 sample was judged by eight of the 14 judges (four insiders and four outsiders) to be weaker than all the BESTEP samples. Three located this between S01 (BESTEP 150) and S02 (BESTEP 185) while another three placed it between S03/ S04 (both had BESTEP scores of 200) and S05 (BESTEP 215). This result is consistent with the LTTC preliminary identification of the A1/A2 cut point with a BESTEP score of 150. The B1 sample was located below S04 (BESTEP 200) by one judge (an outsider), between S04 (200) and S05 (215) by three judges, but between S05 and S06 (220) by five and between S06 and S07 (240) by the remaining five judges. Again, this pattern of results is broadly consistent with the A2/B1 cut point of BESTEP 225 suggested by LTTC, although this recording would have been placed just below the B1 threshold by a majority of the judges. The first and weaker of two B2 level sample recordings was placed between S06 and S07 (BESTEP 220 and 240) by two judges (one insider and one outsider) between S07 and S08 (260) by two (both outsiders), between S08 and S09 (also BESTEP 260) by six, and between S09 and S10 (270) by the remaining four. A second B2 level recording from a different international test was placed by two judges (one insider and one outsider) between S09 and S10, between S10 and S11 (BESTEP 300) by three judges, between S11 and S12 (310) by five and between S12/ S13 (both BESTEP 310) and S014 (BESTEP 360) by the remaining four (two outsiders and two insiders). Again, these results show the judges substantially in agreement with other international linking studies but placing the B1/B2 cut point a little higher for these sample recordings than other international test providers: the weaker B2 level recording would fail to meet the LTTC suggested BESTEP score of 280. Four judges (two insiders and two outsiders) judged the C1 sample to be between S11 and S12/13 (BESTEP 310), six placed it between S13 and S14 (BESTEP 360), another one put it at the same level as S14/S15 (BESTEP 360) and the remaining three (all outsiders) placed it higher than all the BESTEP samples. This was consistent with the LTTC suggestion of 330 as the cut point for C1. The inclusion of this small selection of samples from other international tests of English suggested that the panel's interpretation of the CEFR Common Reference Levels was consistent with that of other international test providers but that the LTTC cut points may have been a little higher when assigning performance to the B1 and B2 levels.

The judges discussed the three cut points in three groups. reaching the following conclusions:

(1) **The cut point between A2 and B1 fell between test takers S03 and S04.** All the groups agreed on this cut point, as illustrated by this remark by P04 (in Group 1):

- [A]ctually the difference [between S03 and S04] is not very obvious, but there is a difference. S03 is A2, because especially the pause, false starts, and the reformulation—the frequency is pretty high and also there are more basic mistakes. As for S04, I think her performance was getting better toward Part 2 and Part 3…especially [in] Part 3, the task actually is a bit more difficult.

P06 in another group (Group 2) also added that they "agree with S04 as B1, as her Part 1 response was quite hesitant, but she picked up in Parts 2 and 3. She was able to use some complex sentences". Overall, S04 'can keep going comprehensively, even though pauses and reformulations are very evident' (B1).

(2) For the cut point between B1 and B2, the judges were in disagreement—whether it should be between test takers S06 and S07 or between S07 and S08. S07's hesitations in the latter part of her performance divided opinions as to whether she was a strong B1 or a very weak B2. For example, her hesitation may be attributed to the need for lexical and grammatical planning (as described in the Fluency category of 'qualitative features of spoken language' scale at B1), as described by P13:

- I think the hesitation…at the beginning, there's not a lot of hesitation. But as she [S07] goes on and she's attempting to formulate language which perhaps she hasn't been able to rehearse, then the longer hesitations start to occur. For instance, she's talking about the search system…of the library…and there's probably some inherent difficulty in in expressing. (P13)

Alternatively, S07 could have paused for "formulating ideas" (P11) and not for searching for language, and other aspects of her performance may be at B2 level:

- [compared to S06 (who was judged as B1)] Her language is much more natural and she rarely makes a mistake… her mistakes are not obvious, and then I think that her performance declines because she picked a difficult topic, not because then she doesn't have sufficient language. (P12)

(3) **The cut point between B2 and C1 falls between test takers S13 and S14.** The judges agreed on this cut point, as S14 produced 'effortless' response while S13 didn't:

- [S13] is able to produce a coherent passage, and it's generally quite fluent, but no, she uses expressions such as 'studying place'. She's not able to express herself with the kind of lexical range [i.e. a 'broad' range of vocabulary that is expected at C1] and clarity (P12)
- I've thought…it [S14's performance] was quite strong and it felt like he can express himself fluently and spontaneously, almost effortlessly. Phonological control wise, I mean the accent is retained, but it doesn't really affect intelligibility. (P14)

3.4. Standard Setting Round 2 judgements

Table 5 BESTEP test takers' scores with judges' Round 2 assignments to CEFR levels

BESTEP Test Taker: Score	***Part 1***	***Part 2***	***Part 3***	***Alternative test***	***Score***	***CEFR level***	***A2***	***B1***	***B2***	***C1***	***Mode***
S02: 185	3.5	1	2	TOEFL iBT	17	B1	14				A2
S03: 200	3.5	2	2.5	GEPT Int	60	A2	13	1			A2
S04: 200	3	2.5	2	GEPT Int	70	A2+	6	8			B1
S04x: 210	4	3	1.5	GEPT Int	70	A2+	3	11			B1
S05: 215	3.5	3	2	GEPT Int	60	A2		14			B1

BESTEP Test Taker: Score	*Part 1*	*Part 2*	*Part 3*	*Alternative test*	*Score*	*CEFR level*	*A2*	*B1*	*B2*	*C1*	*Mode*
S06: 220	4	3	2	GEPT HI	70	B1+		13	1		B1
S07: 240	4	3.5	3.5	TOEIC SW	130	B1		6	8		B2
S07x: 245	4	3.5	3.5	TOEIC SW	150	B1			14		B2
S08: 260	4	4	4	GEPT Int	80	B1+			14		B2
S09: 260	4	4	4	TOEFL iBT	16	B1			14		B2
S10: 270	5	4	3.5	TOEIC SW	150	B1			14		B2
S11: 300	5	5	4	TOEFL iBT	22	B2			14		B2
S12: 310	5	5	5	TOEFL iBT	24	B2			14		B2
S13: 310	5	5	5	TOEFL iBT	22	B2			14		B2
S13x: 310	5	5	5	GEPT HI	80	B2+			13	1	B2
S14: 360	5	6	6	S14: 360	TOEFL iBT	28			1	13	C1
S15: 360	5	6	6	S15: 360	TOEFL iBT	25				14	C1

For the second, pinpointing round of judgements, three sample recordings were added to the set. These were within the range of possible cut scores indicated in the first, range-finding round. In Table 6, these are labelled S04x, So7x and S13x. Again, recordings from other tests with scores linked to the CEFR by other international testing organisations were included. This time, there were three new recordings, one each at the A2, B1 and C1 CEFR levels.

It was apparent that the discussions following the first round of judgement had brought the judges closer together. In this round there was complete consensus on 10 of 18 test takers. However, recordings S04 and S07 continued to divide the panel with six placing S04 at A2, eight at B1; six placed S07 at B1, eight at B2. In next round of discussion, the judges showed higher degrees of agreement than in Round 1. The lowest scoring test takers at B1, B2, and C1 were generally identified as follows:

- **S04: Lowest scoring B1** because she was able to keep going comprehensively, even though she hesitated from time to time. She also demonstrated that she had enough language to get by, placing her at B1. Nevertheless, some judges felt that S04x, the test-taker placed after S04, did not consistently demonstrate features of B1.
- **S07: Lowest scoring B2** because the reason the candidate shows a sufficient range of language and in general, even tempo and high degree of grammatical control.
- **S14: Lowest scoring C1** because his performance was effortless, and while his accent was noticeable it did not impede understanding. The test-taker immediately below S14 (S13X) did not make the cut for C1 because the range of grammar and vocabulary seemed narrower and she was not always able to be spontaneous in her responses.

3.5. Standard Setting Round 3 judgements

Table 6 BESTEP test takers' scores with judges' Round 3 assignments to CEFR levels

BESTEP Test Taker: Score	*Part 1*	*Part 2*	*Part 3*	*Alternative test*	*Score*	*CEFR level*	*A2*	*B1*	*B2*	*C1*	*Mode*
S01: 150	2	1	1	IELTS	6	B1	14				A2
S02: 185	3.5	1	2	TOEFL iBT	17	B1	14				A2
S03: 200	3.5	2	2.5	GEPT Int	60	A2	14				A2
S04: 200	3	2.5	2	GEPT Int	70	A2+	7	7			A2/B1
S04x: 210	4	3	1.5	GEPT Int	70	A2+	7	7			A2/B1
S05: 215	3.5	3	2	GEPT Int	60	A2		14			B1
S06: 220	4	3	2	GEPT HI	70	B1+		14			B1
S07: 240	4	3.5	3.5	TOEIC SW	130	B1		9	5		B1
S07x: 245	4	3.5	3.5	TOEIC SW	150	B1			14		B2
S08: 260	4	4	4	GEPT Int	80	B1+			14		B2
S09: 260	4	4	4	TOEFL iBT	16	B1			14		B2
S10: 270	5	4	3.5	TOEIC SW	150	B1			14		B2
S11: 300	5	5	4	TOEFL iBT	22	B2			14		B2
S12: 310	5	5	5	TOEFL iBT	24	B2			14		B2
S13: 310	5	5	5	TOEFL iBT	22	B2			14		B2
S13x: 310	5	5	5	GEPT HI	80	B2+			14		B2
S14: 360	5	6	6	TOEFL iBT	28	C1				14	C1
S15: 360	5	6	6	TOEFL iBT	25	C1				14	C1

A third round of judgements (Table 7) was used to confirm the cut points arrived at in the pinpointing round. In this final round, judges were simply given the list of 18 BESTEP samples and asked to suggest which represented the lowest scoring test taker who satisfied the minimum requirements for each CEFR level. At this point, the panel was equally divided over whether the lowest scoring B1 test taker was S04 (BESTEP

200) or S04x (BESTEP 210). This largely confirmed the result from round 2 where S04 was assigned to A2 by 6 and B1 by eight judges. All assigned S05 (BESTEP 215) to B1. S07 was the only test taker to be moved from one level to another in round 3, now being placed at the B1 level by nine judges although a majority (eight) had placed them at B2 in round 2. There was unanimity that S14 and S15 were at C1 but that no test taker with a lower score should be assigned to that level.

As these findings suggested some discrepancies between the CEFR panelists' judgements and the original scores, LTTC carried out an investigation into possible causes. It was concluded that the differences in the results could be attributable to differences between the BESTEP scoring system and the Body of Work method used for this study. While BESTEP scores are calculated by combining the scores awarded separately to each test Part, the Body of Work method, like the scoring approaches used by many other international tests, involves panelists arriving at an overall judgement of a test taker's level that balances all the available evidence. Where candidates scored poorly on Parts 1 and 2 by giving incomplete or irrelevant responses, they were awarded low scores using the BESTEP approach, whereas the CEFR panelists appear to have given more credit to their relatively successful performance on Part 3. Candidate S05, for example, was awarded a score of 215 (A2+) on BESTEP made up of the following scores: Part 1, 3.5 (A2+), Part 2, 3 (A2+) and Part 3, 2 (B1). The panelists judged that this candidate met the requirements for B1. The differences may also reflect the use of an explicit task completion criterion in the BESTEP rating scales which involves points deductions when candidates miss two or more questions in Part 1. There are no such penalties in the CEFR scales.

Table 7 Average test taker scores on each BESTEP Speaking test Part by CEFR level

CEFR Level	***N***	***Part 1***	***Part 2***	***Part 3***
A2	4	3.00	1.63	1.88
B1	4	3.88	3.13	2.25
B2	8	4.63	4.44	4.25
C1	2	5.00	6.00	6.00

Table 8 shows the average score on each BESTEP Speaking test Part for test takers at each CEFR level. These are consistent with an increase in task difficulty between Part 1 and Part 2 with test takers generally obtaining higher ratings on the first Part of the test. Both test takers judged to be at C1 by a majority of panellists achieved the maximum of 5.0 points on Part 1. Test takers were also consistently awarded higher ratings on average on Part 2 than on Part 3. Although only limited data is available at this point, the results would be consistent with an A2 performance earning ratings of 2.5 or 3.0 on Part 1 and 1.5 on Parts 2 and 3; a B1 performance earning 4.0 on Part 1, 3.5 on Part 2 and 2.5 on Part 3; a B2 performance earning 5.0 on Part 1 and 4.5 on Parts 2 and 3; and a C1 performance earning maximum points across the three scales.

3.6. Validation of the linking process

Table 8 Responses to the BESTEP CEFR event feedback questionnaire

Item	Strongly agree	Agree	Disagree	Strongly disagree	Average
I understood the purpose of the meetings.	13				4.00
I understood how to carry out the CEFR familiarisation activities.	11	2			3.85
The resources provided in preparation were helpful.	12	1			3.92
The training provided helped me to understand the judgement process.	12	1			3.92
I understood the instructions for the discussion activities in the meetings.	9	4			3.69
I understood how to complete the Judgement Rounds using the online forms.	10	3			3.77
I felt I had sufficient understanding of the CEFR.	7	6			3.54
There was adequate time for reflection and discussion before making the judgements.	11	1	1		3.77
I was able to make my viewpoints known when we were in breakout groups.	13				4.00
Use of the Teams videoconferencing facility was effective.	9	4			3.69
I am confident in the decisions I have made.	6	7			3.46

The results of the online BESTEP CEFR event feedback questionnaire are shown in Table 9. Of the thirteen judges responding, all agreed with all but one of the statements: one judge responded that they did not feel there had been enough time for discussion. All judges strongly agreed that they understood the purpose of the meetings and that they were able to make their viewpoints clear during the discussions. Although all agreed that they were confident in their decisions, this was the only item for which a majority selected "agree" rather than "strongly agree".

Panellists were also given the opportunity to respond to three open-ended questions: What did you like or dislike, or find easy or challenging about this method of setting standards? Do you have any other comments about how we could improve our online standard setting seminars in the future? and Any additional comments? Although selecting "Agree" to the question timing, one further judge commented "I would have preferred a bit more time for discussions". Two judges mentioned a difficulty with the online form for the CEFR Questionnaire: "In the initial questionnaire, we had to scroll up and down to locate the corresponding level (ranging from A1 to C1; nine levels) for a long list of questions". Two others commented on challenges in using the video-conferencing platform for discussions, for example: "I found the [video-conferencing] platform and not seeing participants a bit challenging (e.g. in terms of turntaking)".

These results are encouraging and suggest that this exercise provides a firm basis for BESTEP-CEFR links. The Council of Europe (2009a) recommends that links should be reviewed on a regular basis and the limited amount of data available at this early stage in the life cycle of BESTEP suggests the need for further validation work as more performance samples become available. Post hoc validation will also be needed

to confirm the suitability of the B2 cut point represented by BESTEP scores as an entry criterion for EMI courses in Taiwan.

4. Conclusions and recommendations

Table 9 Summary of panel recommendations for cut scores

CEFR	BESTEP Speaking test
B2/C1:	335
B1/B2:	245
A2/B1:	205
A1/A2:	Insufficient evidence
Below A1/A1:	Insufficient evidence

Table 10 summarizes the panel's recommendations and these are discussed below.

A1 (BESTEP 120). No BESTEP performance samples were available with scores that would place them at the A1 level according to the initial cut point of 120 indicated by LTTC. The BESTEP descriptors that the panel assigned to A1 tended to be negatively worded: "too many errors", "unintelligibility in individual sounds" implying more a failure to reach A2 than positive evidence for an A1 level of performance. Our suggestion would be to reframe these scores as "Below A2" or "A1 or below".

A2 (BESTEP 150). A BESTEP performance sample with a score of 150 was rated as A2 by nine of the 14 judges and as B1 by a further three in the first, range finding round. The BESTEP sample with a score of 185 was judged to be A2 by all 14 judges but they were divided over the two samples with BESTEP scores of 200: a majority placed one at A2 and the other at B1. This supports the initial suggestion that a BESTEP score of 150 or higher is consistent with A2. It may be that the most appropriate A1/A2 cut point is lower than BESTEP 150, but as there were no samples scored below BESTEP 150, there could be no basis for further investigation. With regard to the external A2 sample (a sample A2 level performance taken from another international test linked to the CEFR), eight of the 14 panelists located this between S01 (BESTEP 150) and S02 (BESTEP 185). Another five placed it at BESTEP 200 (between S03 and S04). One put it between S04 and S04x (both BESTEP 210). These results were also consistent with retaining BESTEP 150 as the cut point between A1 and A2. Overall, we found no evidence to support any change to the initial suggestion of BESTEP 150 as the A1/A2 cut point, although this necessarily tentative conclusion should be reviewed in the future as more data becomes available.

B1 (BESTEP 230). In Round 3, all performance samples with scores of BESTEP 230 or higher were unanimously assigned by the panelists to B1 or above. The panelists were also unanimous in assigning all performance samples with BESTEP scores of 215 or higher to at least the B1 level. A majority also assigned to B1 a sample scored BESTEP 210 and one of the two with BESTEP scores of 200. The panel's recommendation based on these results would therefore be to place the A2/B1 BESTEP cut score at 205: the midpoint between sample S04 and S04x.

B2 (BESTEP 280). In Round 3, all performance samples with scores of 245 or higher were unanimously assigned to B2 or above. One sample with a score of BESTEP 240 was also placed at B2 by a majority of panelists in Round 2 but downgraded to B1 by nine of the 14 in Round 3. Externally sourced samples placed at B2 by other providers did not all reach the LTTC B1/B2 threshold used for BESTEP. The recommendation based on these results would therefore be to place the B1/B2 BESTEP cut score at 245. Again, any decision on whether to take a relatively cautious approach should take account of the needs of score users.

C1 (BESTEP 330). In Round 3, two BESTEP sample performances with scores of 360 were assigned to C1 by all 14 judges and the two samples with BESTEP scores of 310 were both unanimously assigned to B2. Consistent with locating the B2/C1 cut point between 310 and 360, the sample C1 performance from another international test was rated as higher than both of the BESTEP samples with scores of 310 by 10 of the 14, another three judging it to be higher than either of the BESTEP samples with scores of 360. Based on this evidence, the midpoint between 310 and 360 (335) is taken as the panel recommendation for the B2/C1 cut point. However, the initial suggestion of 330 is also consistent with our findings: there is insufficient evidence here to suggest that the C1 cut score should be changed.

Now that the test is in operational use, it will be possible to collect more extensive evidence to confirm our conclusions. As it was not possible in this study to determine whether the suggested 120 points on the BESTEP Speaking test is a suitable minimum for A1 and whether the cut point between A1 and A2 should be 150 points, although a score of 150 was agreed by all panellists to satisfy the requirements for A2. These questions should be a focus for further study. Similarly, there was relatively little evidence for the precise location of the B2/C1 cut point. If this is to become important for high stakes decision making, it too will require fuller investigation.

The possible impact of different scoring approaches on the estimation of CEFR levels raised by this study has not attracted much attention in the context of CEFR linking research. This is an area that also requires further investigation. Differences in the scoring approaches adopted by different test providers may add to the degree of uncertainty that attaches to the comparability of linking exercises.

More positively, our findings based on both quantitative and qualitative data are supportive of the current CEFR cut score recommendations for the BESTEP Speaking test with both the A2/B1 and B1/B2 cut scores on the test being well above the panel's recommended cut points and the results for B2/C1 also being consistent with current practice. In other words, if the cut scores put forward by the LTTC are retained, there is limited risk of false positive results at B1 and B2, although there is a greater risk that some test takers who do satisfy the criteria for B1 or B2 might be placed lower by the test.

5. References

Aizawa, I., Rose, H., Thompson, G., & Curle, S. (2023). Beyond the threshold: Exploring English language proficiency, linguistic challenges, and academic language skills of Japanese students in an English medium instruction programme. *Language Teaching Research, 27*(4), 837-861. DOI: 10.1177/1362168820965510

ALTE (2002), *Language Examining and Test Development*. Strasbourg: Council of Europe Language Policy Division.

ALTE (Association of Language Testers in Europe) (2011). Manual for language test development and examining. For use with the CEFR. Produced by ALTE on behalf of the Language Policy Division, Council of Europe.

ALTE/ Council of Europe (2011) *Manual for Language Test Development and Examining for use with the CEFR* Produced by ALTE on behalf of the Language Policy Division, Council of Europe. Council of Europe.

Brunfaut, T. and Harding, L. (2014). *Linking the GEPT Listening Test to the Common European Framework of Reference*, LTTC-GEPT Research Report No. RG-05. LTTC.

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

Council of Europe (2009a). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A manual*. Council of Europe.

Council of Europe (2009b). *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe.

Council of Europe (2011) *Manual for Language Test Development and Examining for use with the CEFR* Produced by ALTE on behalf of the Language Policy Division, Council of Europe. Council of Europe.

Council of Europe (2020), *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, available at www.coe.int/lang-cefr.

Deygers, B., Zeidler, B., Vilcu, D. & Hamnes Carlsen, C. (2018) One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies, *Language Assessment Quarterly, 15*(1), 3-15, DOI: 10.1080/15434303.2016.1261350

Fan, J., Knoch, U., & Chen, I. (2021). Linking the GEPT Writing Subtest (Part 1) to the Common European Framework of Reference (CEFR). LTTC.

Ferrara, S., Perie, M., & Johnson, E. (2002). *Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching procedure*. Washington, DC: American Institutes for Research.

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*.

Green, A.B. (2018). Linking Tests of English for Academic Purposes to the CEFR: The score user's perspective, *Language Assessment Quarterly, 15*(1), 59-74, DOI:10.1080/15434303.2017.1350685

Green, A., Inoue, C., & Nakatsuhara, F. (2017). *Relating GEPT speaking tests to the CEFR* (LTTC–GEPT Research Report No. RG-09). LTTC.

Hamnes Carlsen, C.H. (2018). The Adequacy of the B2 Level as University Entrance Requirement, *Language Assessment Quarterly, 15*(1), 75-89, DOI 10.1080/15434303.2017.1405962

Harsch, C., Ushioda, E., & Ladroue, C. (2017). Investigating the predictive validity of TOEFL iBT scores and their use in informing policy in a United Kingdom university setting. *ETS Research Report Series*, *2017*, 1–80.

Kingston, N., & Tiemann, G. (2012). Setting performance standards on complex assessments: The Body of Work Method. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.) (pp. 201–223). New York, NY: Routledge.

Kingston, N., Kahl, S., Sweeney, K., Bay, L. (2001). Setting performance standards using the Body of Work Method. In G. Cizek (Ed.), *Setting performance standards: Concepts. methods. and perspectives* (pp. 219–248). Mahweh, NJ: Lawrence Erlbaum Associates.

Knoch, U., and K. Frost. (2016). Linking the GEPT writing sub-test to the Common European Framework of Reference (CEFR). *LTTC-GEPT Research Reports RG-08. Taipei: Language Training and Testing Centre*.

Lewis, D., Mitzel, H., Mercado, R., & Schulz, M. (2012). The Bookmark standard setting procedure. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.) (pp. 225–253). New York, NY: Routledge.

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.

Ministry of Education, Republic of China (Taiwan) (2021) *BEST: The Program on Bilingual Education for Students in College.* Ministry of Education.

Noijons, J., Beresova, J., Breton, G., & Szabo, G. (2011). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Highlights from the Manual*, Council of Europe Language Policy Division.

North, B (2010). Linking Certification to the CEFR: Do we need standard-setting? In Mader, J. and Urkun, Z. (Eds.), *Putting the CEFR to good use: Selected articles by the presenters of the IATEFL Testing, Evaluation and Assessment Special Interest Group (TEA SIG) and EALTA Conference, Barcelona, Spain* 29-30 October 2010. IATEFL.

North, B (2014). Putting the Common European Framework of Reference to good use. *Language Teaching, 47*, 228-249 DOI:10.1017/S0261444811000206

Piccardo, E., & North, B. (2019). *The action-oriented approach: A dynamic vision of language education*. Multilingual Matters.

Wu, J. R. W. & Wu, R. Y. F. (2010). Relating the GEPT Reading Comprehension Test to the CEFR. In Martyniuk, W. (ed.) *Aligning Tests with the CEFR: Case studies and reflections on the use of the Council of Europe's Draft Manual* (pp. 204-224). Cambridge University Press.

Wu, R. Y. F. (2014). *Validating second language reading examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*. Studies in Language Testing 41. Cambridge University Press.

Appendix A: Specification template

Relating language assessments to the Common European Framework of Reference for Languages:
Describing your assessment (Specification)

The following forms support users with **Specification**: the first stage of linking an assessment to the CEFR.

Before starting to complete the forms, please read Noijons, J., Bérešová, J., Breton, G. & Szabó, G. (2011), *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Highlights from the Manual.* Graz: ECML, Chapters 5, 6, and 7.

This book is available as a free download here[1].

Specifications are documents that are used to guide the production of assessment material. Defensible linking of any assessment to the CEFR must be based on a clear specification of the purpose and content of that assessment (Council of Europe 2009a).

Assessment specifications contain:

- a description of the purpose and context of use of the assessment and the population of learners for which the assessment is appropriate.
- a description of the theoretical construct (the knowledge, skills and abilities being assessed) with details of how these are operationalized in, for example, domains, tasks, rating criteria, pass/fail boundaries), the weight given to each part of the assessment in calculating overall scores and the ways in which results are calculated and reported.

Specifications are an important tool to:

- ensure the quality of the test and to help to demonstrate that recommended interpretations and uses of assessment results are justified.
- ensure that different forms of the same assessment share the same basis in a defined content domain (such as a teaching syllabus or context for language use).

If your assessment does not already have specifications, these forms will help you to produce a set of specifications that meets international standards.

If your assessment does already have specifications, these forms will help you to present your assessment using the **common terms** of the CEFR and in a way that will support comparisons with other assessments that have been linked to the CEFR.

Before completing the forms, you can use the checklist of questions below (from ALTE, 2002. *Language Examining and Test Development*. Strasbourg: Council of Europe Language Policy Division, p.12). These questions are intended to help you to collect useful information and begin the process of writing or transposing the specifications for your assessment.

[1] www.ecml.at/Resources/ECMLPublications/tabid/277/PublicationID/67/language/enGB/Default.aspx

Appendix A: Specification template

CEFR checklist for drawing up test specifications

Users who are involved in drawing up test specifications may like to consider and where appropriate state:

- what type and level of language performance needs to be assessed?
- what type of test tasks are necessary to achieve this?
  what practical resources are available? e.g., premises,
  personnel, etc.
- what political, social and/or economic issues are likely to influence test development?
  who should be involved in drafting test specifications and developing sample test materials?
  e. g., in terms of expertise, influence, authority, etc.
- how will the content, technical and procedural details of the test be described in the specifications?
- what sort of information about the test needs to be given to users, and how?
  e. g., a publicly available version of the test specifications.

One resource that will be particularly useful (and that is referred to throughout the forms) is the Council of Europe (2011) *Manual for Language Test Development and Examining for use with the CEFR Produced by ALTE on behalf of the Language Policy Division, Council of Europe.* Strasbourg: Council of Europe (referred to here as the Manual for Language Test Development). This book is available as a free download here[1].

Notes for guidance on completing the forms

In the forms, you will find instructions and boxes to be completed with information about the assessment.

Text appears in four colours:

Blue text appears in the instructions. It indicates references to the CEFR and allied resources to help you complete the forms. **Do not change this text.**

Red text appears in the titles, instructions and content boxes. It represents common core elements that should appear in the instructions and specification for all assessments.

You should not generally change this text, except to add names or select options[2].

For example, if the assessment was called "The Socrates Test of Greek", you would change

"The purpose and use of [the assessment]" to "The purpose and use of The Socrates Test of Greek"

Black text appears inside the content boxes. This gives instructions, sets out key questions to be answered, or provides guidance on the information that should be included in the forms. Sometimes it provides options you can choose from to help you complete the form.

Black text can be adapted or replaced by the responses as needed, but please respond to all the questions or prompts.

Grey text offers examples of possible responses relating to specific assessments or sets of alternative options. **Grey text can be adapted or replaced as needed**, but please keep to the terms that are used in the CEFR.

Please **note** that the words used on this form may not be same as the ones you usually use in your assessment programmes. For example, you may usually use the word 'test' where the CEFR uses 'assessment', or you might use 'workplace contexts' where the CEFR uses 'Occupational domain'.

In such cases, it is better to use the CEFR words to fill in this form. This will make it easier to connect your work to that of other assessment providers.

[1] rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-thece/1680667a2b

[2] Items in [square brackets] may be replaced with names, numbers or terms as relevant. Where items are separated by a slash(/), this indicates a choice between options.

Appendix A: Specification template

Describing your assessment (Specification)

6. Preliminaries

1. The organisation

Which organisation has provided the information included in the following forms?

Name: The Language Training & Testing Center (LTTC)
Address: No. 170, Section 2, Xinhai Road, Da'an District, Taipei City, Taiwan (R.O.C.) 106 Website: www.lttc.ntu.edu.tw

2. The person responsible for the information in the forms

Who is the person at the organisation responsible for entering and updating the information?

Miranda Min-Hsin Chen
Researcher, Testing Editorial Department miranda@lttc.ntu.edu.tw

Hogan Ping-Hsuan Wang
Researcher, Testing Editorial Department hogan@lttc.ntu.edu.tw

7. A. The purpose and use of the BEST Test of English Proficiency (BESTEP)

For help in defining language abilities for assessment in relation to the CEFR, see the Manual for Language Test Development, Section 1.1, pp.10-13.

1. Overview of the BESTEP

The BESTEP is a national assessment taken by…

Students in higher education in Taiwan, including those enrolled in undergraduate or graduate programs of Taiwanese universities or colleges, mostly, though not exclusively, expected to speak Mandarin Chinese as their first language and coming from Taiwan.

The BESTEP provides evidence of…

Students' readiness to communicate at CEFR levels A2 to B2 in English in academic contexts, such as in English-mediated instructed (EMI) courses. The EMI courses herein refer broadly to Englishtaught modules within the local higher education system, which both graduate and undergraduate students in Taiwan can enroll in. Universities or other academic institutions in Taiwan may choose to use the BESTEP results in ways they prefer (e.g., diagnostic or placement for individual courses, modules, or degree programs).
The BESTEP is primarily concerned with the educational domain, but includes some tasks that involve the personal and public domains that are relevant to students in higher education. In addition, it involves linguistic competences, including vocabulary range and control, grammatical accuracy, phonological control (speaking) and orthographic control (writing), as well as pragmatic competences, including the ability to manage discourse in terms of thematic organization, coherence and cohesion, style and register, and rhetorical effectiveness.
It relates to English for academic purposes (EAP) contexts for language use.

The BESTEP aims to encourage…

A virtuous cycle between EAP learning, teaching, and assessment in Taiwan's higher education sector in which the test results enable 1) teachers to evaluate students' speaking and writing abilities and thereby adjust their pedagogy accordingly for enhanced teaching effectiveness; 2) students to track their own learning progress, set goals for future stages, and foster the ability for autonomous learning; 3) policy-making institutions to plan for supporting measures to enhance English proficiency, integrate resources, diffuse benefits, and systematically drive the development of English proficiency among college and university students.

The BESTEP is owned and administered by…

Owned by the Ministry of Education (MOE) of Taiwan, and developed and administered by the LTTC under the auspices of the Ministry of Education

Appendix A: Specification template

The BESTEP results should be used by… to…

By language learners and teachers to set individual language learning goals; by universities and colleges in Taiwan to evaluate students' readiness for EMI courses; by the MOE to distribute and allocate educational resources and support.

The BESTEP results are recognised by… for the purpose of…

The BESTEP is recognized by the MOE of Taiwan for higher education; by colleges and universities for evaluating students' readiness for EMI courses.

To find out more about the aims of the BESTEP and the LTTC, and about organisations that recognize the results…

Visit the official website at www.bestep.tw/eng

8. B. Producing the BESTEP

1. Development of the BESTEP

For help in reporting the development of an assessment, see the Manual for Language Test Development, Chapter 2, pp.20-25.

The BESTEP was developed by…

By the LTTC under the auspices of the Ministry of Education

The content of the BESTEP is based on…

Literature reviews of existing English tests and ability indicators, as well as research on EAP and EMI, and EAP and EMI courses in Taiwan were conducted in the early stage of test development to identify key aspects of can-dos, including topics and language use contexts, language functions, and language performance for speaking and writing tests (see Initiative for the National Speaking and Writing Assessment for College Students: Phase 1 Progress Report, 28 January 2022)[3]. A needs analysis was also conducted to understand 1) the texts and figures commonly seen in university or college lectures or other learning contexts in higher education; 2) event information university or college students are likely to encounter in their campus life (for the results of the needs analysis, see Initiative for the National Speaking and Writing Assessment for College Students: Phase 1 Progress Report, 28 January 2022)[4].
Based on the literature review and results of the needs analysis, the test tasks were selected and task-specific rating scales for speaking and writing were developed. After pilot testing (N=40) and pretesting (N=1002), the test results and questionnaire analysis indicated that the content relevance and difficulty of the test tasks were appropriate, and the scoring rubric was fine-tuned on the basis of feedback from teachers. The development process was assisted by a panel of
expert members and involved five meetings (see Initiative for the National Speaking and Writing Assessment for College Students, Phase 2 Progress Report (題庫與施測認證系統建置計畫第二次成果報告)[5].

To find out more about how the BESTEP was developed…

Consult the progress reports, phases 1 and 2 (unpublished; written in Mandarin Chinese with abstract in English). At the time of writing, all reports have been submitted to the MOE and are under review. The reports will be made publicly available on the official website www.bestep.tw/eng (under Research) at the time the MOE deems fit.

[3] Tung, S., Wu, J. R. W., & Wu, R. Y. F (Eds.), (28 January 2022) Initiative for the National Speaking and Writing Assessment for College Students, Phases 1 Progress Report (全國大專校院英語說寫評量檢測計畫：題庫與施測認證系統建置計畫第一次成果報告). Under review

[4] Tung, S., Wu, J. R. W., & Wu, R. Y. F (Eds.), (28 January 2023). Initiative for the National Speaking and Writing Assessment for College Students, Phase 2 Progress Report (題庫與施測認證系統建置計畫第二次成果報告). Under review.

Appendix A: Specification template

2. Writing the BESTEP

For help in reporting how assessments are written and assembled, see the Manual for Language Test Development, Chapter 3, pp.26-33.

The BESTEP writers are…

Professionals, both native and non-native speakers of English, with a graduate degree in language testing and a background in teaching English as a foreign language (EFL) or English as a second language (ESL) in Taiwan, who respond to the LTTC's needs for test items on a contractual basis. Item writers are first screened for their occupation, with those teaching EFL in cram schools or other after-school institutions rejected, in order to avoid any conflict of interests that may otherwise jeopardize the objectivity or confidentiality of the test items.

The BESTEP writers are given training…

To be considered for the role, candidates first submit sample items for review by LTTC editors. This determines whether or not they are commissioned. To be employed as writers, they need to demonstrate to the editors that they can write material that is judged appropriate in difficulty and suitable for the local learning context for use on the BESTEP. The writers are then given a handbook that details the test specifications for the BESTEP and guidelines to follow as to how to submit the finished assessment material via a secure portal. The review process entails, submission of a first draft; revision according to feedback from the LTTC editors; and resubmission of subsequent drafts until the items are deemed suitable for test purposes.

To ensure that the content follows the developers' intentions and to ensure that it is parallel across different versions of the BESTEP, when preparing material, the writers are required to…

Adhere to writer guidelines, use examples of good tasks, consult wordlists and specified language functions. Writers also maintain close contact with the LTTC editors to discuss the content and formats. The LTTC editors also feedback the writers using the post-test results (e.g., the task written by them was (unexpectedly) more difficult than assumed) and further training/advice for the future is given.

Assessment material is reviewed by…

In-house editors, both native and non-native speakers of English, who meet the minimum requirement of holding a graduate degree in a language-related field. The editors are familiar with the local EFL environment and have gone through rigorous training that involves familiarization with test development.
They review test questions and test forms in terms of language functions, topic and the expected difficulty specified in the item assignment.

In designing and developing tasks for use in each assessment, account is taken of… by…

The LTTC collects test takers' information, such as age, gender, type of school, major and discipline, etc. to discern any potential item biases that favor test takers of a particular demographic profile. In the EAP context, specifically, the editors consider whether the content favors test takers of a certain discipline by conducting interviews with test takers. Prospective test takers with physical, visual, hearing, speech, balance or multiple impairments may request special arrangements at registration, which include test papers with enlarged fonts, extended preparation and response time, accessible test rooms, additional proctors for assistance, etc. In 2022, these requests accounted for 0.17% - 0.41% of test takers per test administration. Further information is available on the BESTEP website at www.bestep.tw/eng (under Registration).

Appendix A: Specification template

Before being used on the BESTEP, materials are subject to the following procedures:

- Item review: After the items have been developed, they are reviewed by editors and reviewers to ensure that they are accurate, relevant, and appropriate for the intended purposes.
- Pilot testing: The test is administered to a small group of participants in order to identify any potential issues with the test administration, timing, scoring, and item performance.
- Item analysis: The data collected from the pilot test are analyzed in terms of difficulty and discrimination to assess the quality of the test items.
- Test revision: The test items are revised on the basis of the results of item analysis, which assesses task difficulty of individual tasks, as well as rater reliability (including percentages of rater agreement, inter-rater reliability, and intra-rater reliability) and rating scale analysis, which examines the distribution of scores across the rating scale categories and correlations among individual tasks. The analyzed test items are hen subjected to another round of pilot testing.
- Item banking: After the items have been reviewed and analyzed, they are assigned metadata, such as the topic area, level of difficulty, and type of question and stored in an item bank.
- Test assembly: The items are selected and assembled into a final test form according to the test blueprint. This involves arranging the items in a specified order, and ensuring that the test includes an appropriate mix of topics and difficulty levels.

To find out more about how [the assessment] is prepared…

Visit the official website at www.bestep.tw/eng. As of this writing, all reports have been submitted to the MOE and are under review. The reports will be made publicly available on the official website www.bestep.tw/eng (under Research) at the time the MOE deems fit.

3. Summary of the content of the BESTEP

For help in providing an overview of assessment formats, see the Manual for Language Test Development, Section 2.4, pp.21-23 and Appendix III The BESTEP includes…

Test Overview

Paper/Timing	Format		No. of Qs	Test focus	Band score	Scale score
Speaking Approx. 1 hour Computer-based (in a lab)	Part 1 Answering Questions	Each test taker hears a question and responds immediately for 15 seconds.	6	Deliver learning-related information and describe personal experiences	0-5	20-80
	Part 2 Expressing Opinions	Each test taker has 1½ minutes to prepare and 1½ minutes to discuss their experiences and opinions related to a chart.	3	Express opinions and exchange ideas on familiar, learning-related topics	0-6	40-130
	Part 3 Giving a Short Presentation	Each test taker has 2½ minutes to prepare and 2½ minutes to give a 'short presentation' based on a short passage and a chart.	2	Integrate information from different sources, summarize the main points clearly and accurately, and support arguments with details and evidence	0-6	40-150
Writing Approx. 75 minutes Paper-based	Part 1 Answering Questions	Test takers are required to answer, in short sentences, three questions based on one visual prompt.	3	Inquire about and respond to input related to school life and learning	0-5	20-80
	Part 2 Expressing Opinions	Test takers are required to write a message, letter, or email; 80-100 words.	1	Express opinions on learning-related topics	0-6	40-130
	Part 3 Writing an Integrated Essay	Test takers are required to write an essay based on two figures; 100-120 words.	1	Compose essays on academic topics, stating a position, presenting an argument, and supporting it with appropriate evidence	0-6	40-150

See Wu, Wu, & Lin. (2023)[6]

[6] Wu, J. R. W., Wu, R. Y. F., & Lin, A. C. W. (2023, June 5-9), *Co-constructing with stakeholders the performance descriptors for an English productive skills test* [Conference presentation]. 44th Language Testing Research Colloquium, New York City, NY, United States. https://ltrc2023.weebly.com/uploads/1/4/3/6/143613600/ltrc_2023_schedule_updated_6.2.23.pdf

Information on Assessment Tasks 1. The BESTEP Speaking test

Part 1

i. Rationale

This part of the speaking test involves responding to six prompts in the form of questions in English. The prompts are pre-recorded, and each is played twice. After hearing the prompts for the second time, the test taker is given 15 seconds in which to supply a response.

This task aims to elicit a response that is similar to classroom interactions between students and professors or students and classmates as well as classroom activities, such as: giving a class presentation; describing study and learning experiences; offering advice on test preparation to a fellow student; requesting information on course selection from the office.

The spoken instructions are given in English, while written instructions are given on the test paper in both Chinese and English. The spoken and written English instructions are the same.

In this part of the BESTEP speaking test, the test taker has 15 seconds to respond to each prompt.

ii. What the person being assessed reads/ listens to/ sees (the input)

The input includes spoken prompts and instructions, both spoken and written, designed especially for the test.

The input mainly relates to the educational domain.

Communication themes may include: - Education - Relations with other people - Services - Places

The input is prepared especially for the BESTEP.

Spoken input is recorded by professional voice actors in a private studio.

To make material suitable for this part of BESTEP, of which the difficulty level is set between CEFR A2 to B1, vocabulary and grammar appropriate to the A2 level are used in the questions.

The input is familiar and mostly concrete to university students: e.g. personal information, learning experiences, events regularly encountered in the educational domain.

The input is 50-70 words in total.
The vocabulary of the input includes a basic range of words and phrases.
The grammar of the input includes simple grammatical structures and sentence patterns.
The delivery is slow and clearly articulated in a standard accent.
There is minimal background noise or distortion in the recordings.
The recording involves only one speaker.
Test takers being assessed hear the recording twice.

The input is likely to be comprehensible to a language learner at CEFR level A2.

iii. What the person being assessed needs to do (the expected response).

The response involves answering spoken questions.

Responses are expected to be in the form of individual utterances and take a maximum of 15 seconds.

The main rhetorical functions expected include giving personal information, describing experiences, requesting information, giving advice, convincing and persuading, expressing opinions, and discussing hypothetical scenarios.

The response is expected to be in the form of short utterances.

The response is open-ended.

The main purpose of the response is informational or directive.

The vocabulary of the response is expected to involve a basic range of words and phrases.

The range of grammar in the response is expected to involve simple grammatical structures and sentence patterns.

The level of coherence and cohesion in the response is expected to involve linking groups of words with simple connectors like "and," "but", and "because".

In responding to this part of the speaking test, people are expected to draw on their daily life experiences and fulfill basic communication needs in the educational domain.

Example of Part 1:

Test takers see:

第一部分：回答問題

Part I: Answering Questions

共 6 題，問題經由耳機播放 2 次，不印在試卷上。題目播出 2 次後，請立即回答。每題回答時間 15 秒。

You will hear six questions. Each question is played twice. Please answer each question immediately after you hear it for the second time. You will have 15 seconds to answer each question.

Test takers hear (next page):

Part I: Answering Questions

You will hear six questions. Each question is played twice. Please answer each question immediately after you hear it for the second time. You will have 15 seconds to answer each question.

Question No. 1.	Please tell me about your major.
Question No. 2.	Why did you choose your major? Please explain.
Question No. 3	Your younger cousin Rita is considering majoring in the same field as you do. What advice will you give her?
Question No. 4.	Have you given a presentation before? What was the presentation about?
Question No. 5.	Your teacher asks you to work with a classmate for your next presentation. Invite your classmate Ray to join you.
Question No. 6.	Ray is afraid of speaking in front of the whole class. Discuss with him how to divide work on your presentation so that you can both do what you're best at.

Part 2

i. Rationale

This part of the speaking test involves expressing opinions in English. The test taker is given 1.5 minutes in which to supply a response to the prompt and questions given on the test paper.

This task aims to elicit a response that is similar to having a general discussion (including expressing opinions and exchanging information and ideas with classmates), such as: - offering advice on test preparation; - requesting information on course selection; - expressing opinions about joining a student club.

The spoken instructions are given in English, while written instructions are given in both Chinese and English on the test paper. The prompt, consisting of a chart and three related questions, is written in English on the test paper.

In this part of the BESTEP speaking test, the test taker has 1.5 minutes to read the prompt and another 1.5 minutes to respond.

ii. What the person being assessed reads/ listens to/ sees (the input)

The input includes a graph and three related questions designed especially for the test.

The input mainly relates to the educational domain.

Communication themes may include: - Education - Relations with other people - Services - Places

The input is prepared especially for the BESTEP.

To make material suitable for this part of BESTEP, of which the difficulty level is set between CEFR A2-B2, vocabulary and grammar appropriate to the A2 level and below are used in the prompt and questions.

The input is familiar and mostly concrete for university students, including 1) campus activities, such as looking for part-time jobs, 2) learning experiences, such as planning one's weekly class schedule, and 3) matters regularly encountered in the educational domain, such as reading for pleasure.

The written input is 50-70 words in total.
The input involves a sufficient range of vocabulary to express most topics pertinent to student life such as learning experiences, extracurricular activities, part-time jobs, and social life.
The grammar of the input includes simple grammatical structures and sentence patterns.

The input is likely to be comprehensible to a language learner at CEFR levels A2 and above.

iii. What the person being assessed needs to do (the expected response).

The response involves answering questions.

Responses are expected to be in the form of short monologues and take a maximum of 1.5 minutes in total.

The main rhetorical functions expected are describing data, giving information, and giving opinions.

The responses are expected to be in the form of a series of contributions to a discussion.

The responses are partially controlled by the prompt or input and partially open-ended.

The main purpose of the responses is informational.

The vocabulary of the responses is expected to include a sufficient range of words and phrases for routine, everyday transactions involving familiar situations and topics.

The range of grammar in the responses is expected to include simple grammatical structures and sentence patterns.

The level of coherence and cohesion in the responses is expected to involve linking a series of shorter, discrete simple elements into a connected, linear sequence of points.

In responding to this part of the speaking test, test takers are expected to draw on their daily life and common, general, non-specialized knowledge in the educational domain.

Example of Part 2:

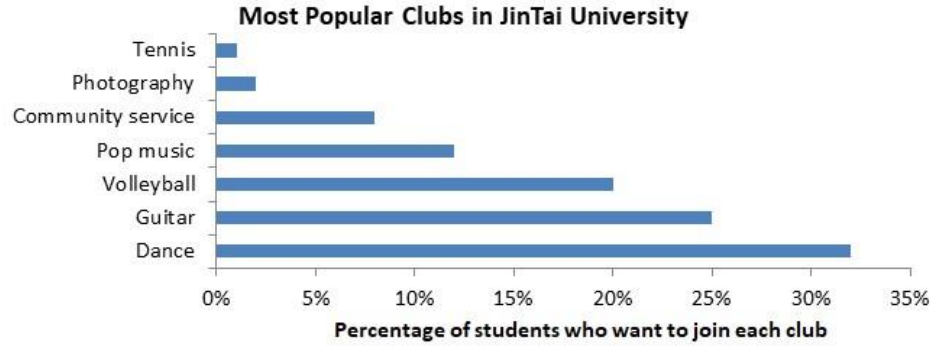
第二部分：表達意見

Part II: Expressing Opinions

下面有一張圖表及三個相關的問題，聽到提示開始作答的鈴響後，請在 1.5 分鐘內完成作答。
現在請先利用 1.5 分鐘的時間看圖表及問題，準備時請勿發出聲音。

On your test paper, you will see a chart and three related questions. First, you will have 1½ minutes to prepare your answers based on these materials. After that, you will hear a tone. You will then have 1½ minutes to record your answers. Please begin your preparation now.

The chart below shows the results of a survey on the most popular clubs at Jintai University.



You have 1½ minutes to answer the following three questions.

1. Are the most popular clubs in Jintai University also popular in your university? Please explain.
2. Have you joined a club like any of the clubs in the chart? Why or why not?
3. What is your opinion of joining a club in university?

Part 3

i. Rationale

This part of the speaking test involves giving a short presentation in English. The test taker is asked to respond within 2.5 minutes.

This task aims to elicit a response that is similar to giving a presentation in class that involves:

- presenting key points about an editorial or a school letter;
- describing a chart or a graph that illustrates an educational phenomenon;
- comparing information from multiple sources;
- expressing personal opinions on topics in the educational domain.

The spoken instructions are given in English, while written instructions are given in both Chinese and English on the test paper. The prompt, consisting of a graph, a short passage, and two related questions, is given in English on the test paper.

In this part of the BESTEP speaking test, the test taker has 2.5 minutes to prepare for the presentation and another 2.5 minutes to respond.

ii. What the person being assessed reads/ listens to/ sees (the input)

The input includes a graph, a short passage, and two related questions designed especially for the test.

The input mainly relates to the educational domain.

Communication themes may include:

- Education
- Relations with other people
- Services
- Places

The input is prepared especially for the BESTEP.

To make material suitable for this part of BESTEP, of which the difficulty level is set between CEFR B1 to B2, vocabulary and grammar appropriate to the B1 level and below are used in the questions.

The input is less familiar and somewhat abstract for university students, including academic or social issues regularly encountered in the educational domain, such as:

- the gender gap between different disciplines;
- the benefits of student volunteer work; - the trends of textbook formats

over time; - the needs of studying overseas.

The written input is 65-85 words in total.

The input involves sufficient range of vocabulary for most academic and social issues with varied formulation.

The grammar of the input may include some complex sentence forms.

The input is likely to be comprehensible to a language learner at CEFR level B1 or above.

iii. What the person being assessed needs to do (the expected response).

The response involves giving a presentation.

Responses are expected to be in the form of an oral presentation in the educational domain and take a maximum of 2.5 minutes.

The main rhetorical functions expected are summarizing, comparing, evaluating, and giving opinions.

The response is expected to be in the form of a presentation.

The response is partially restricted by the instructions, prompt, or input and is partially openended.

The main purpose of the response is informational.

The vocabulary of the response is expected to include a sufficient range of vocabulary for most topics in the educational domain with varied formulation.

The range of grammar in the response is expected to include some complex sentence forms.

The level of coherence and cohesion in the response is expected to involve a limited number of cohesive devices to link utterances into clear, coherent discourse (but with some "jumpiness" in a long contribution).

In responding to this part of the speaking test, test takers are expected to draw on their academic life and common, general, non-specialized knowledge in the educational domain.

Example of Part 3 (next page):

第三部分：摘要報告

Part III: Giving a Short Presentation

下面有一篇文章、一張圖表以及兩個相關的指示，聽到提示開始作答的鈴響後，請在 2.5 分鐘內完成作答。現在請先利用 2.5 分鐘的時間看文章及圖表，準備時請勿發出聲音。

On your test paper, you will see a passage, a chart, and two instructions. First, you will have 2½ minutes to prepare your presentation based on these materials. After that, you will hear a tone. You will then have 2½ minutes to record your presentation. Please begin your preparation now.

The passage and the chart below are about learning styles. The information contained in each does not entirely support the other.

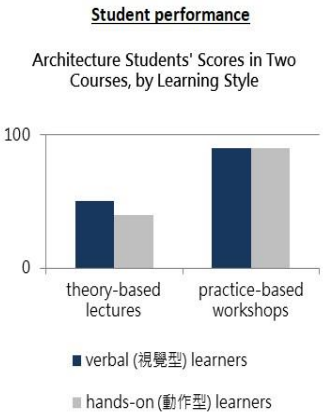
You have 2½ minutes to give your presentation. You should:

1. discuss the key difference between the passage and the chart;
2. explain whether you agree or disagree with the passage. You may draw examples from your own experience.

Editorial

Study in Your Own Way for Best Results

Each of us has a unique learning style. We learn best when we study according to that style. For example, in a course that focuses on theory, students who prefer reading input have a great advantage. However, when it comes to workshops, hands-on learners do much better than verbal learners.



4. Scoring Assessment Tasks

For each part of the BESTEP Speaking test, the response is marked by two independent trained raters in a central location. Raters are teachers with a background in teaching English for academic purposes (EAP) in universities in Taiwan.

The two scores given by independent raters are averaged to produce a mean score for each part. Researchers from the LTTC are responsible for managing the marking process.

To ensure the accuracy of marking, each official marking session begins with raters attending training sessions and completing a set of calibration tests to demonstrate that they understand and apply the scoring criteria accurately and consistently. During the official marking session, the BESTEP on-screen rating system randomly pairs two raters. Performances of raters 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 raters' performances. If the difference between an individual rater's average ratings and the overall rating statistics exceeds an acceptable range (e.g., greater than 2 band score on a scale of 0 to 5 or 6), the rater is flagged. Individual raters are also flagged if their inter-rater agreements are significantly lower than the average agreement of other raters. The flagged raters 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 raters on the same test taker's response differ by more than two band score, their ratings are considered discrepant. Resolution of discrepancies by a senior rater is required before scores are reported. Scoring leaders monitor raters during the official marking process; if raters are consistently producing discrepancies in marking and their scores for certain test takers exceed the acceptable range, their responses will be rescored by a third rater, and scores given by the third rater stands.

Rating scale:

Task-specific holistic rating criteria are used to evaluate test takers' performance on the BESTEP speaking section. A 6-point (from 0 to 5) holistic scale is used in part 1, and a 7-point (from 0 to 6) holistic scale is used in parts 2 and 3.

Part 1. Answering Questions

Band	Scale score	Description
5	80	• Most responses are relevant to the topic. • Most simple grammatical structures and vocabulary are used appropriately. Errors may occur but do not impede communication. • Speech is generally fluent. • Utterances are comprehensible, with easily recognizable sounds and intonation.
4	60	• Most responses are relevant to the topic. • Some simple grammatical structures are used appropriately. Errors do occur, but communication is not impeded. • Vocabulary is limited. • Speech is slow with occasional inappropriate pausing. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation; communication is impeded occasionally.
3	50	• Many responses are relevant to the topic. • Some simple grammatical structures are used appropriately. Errors do occur, but communication is not impeded. • Vocabulary is limited. • Speech is slow with occasional inappropriate pausing. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation; communication is impeded occasionally.
2	40	• Some responses are relevant to the topic. • Grammatical errors occur frequently, and utterances are mostly fragmentary. • Vocabulary is limited, and communication is impeded. • Frequent and inappropriate pausing interferes with communication and puts stress on the listener. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation, making the listener stressful.
1	30	• Few responses are relevant to the topic. • Grammatical errors occur frequently, and utterances are mostly fragmentary. • Vocabulary is limited, and communication is impeded. • Frequent and inappropriate pausing interferes with communication and puts stress on the listener. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation, making the listener stressful.
0	20 and below	• Responses are inadequate or completely off-topic. • Errors are frequent, making the responses difficult to understand.

Part 2. Expressing Opinions

Band	Scale score	Description
6	130	• Responses are relevant to the topic and more than adequate. • Responses show appropriate use of a wide range of vocabulary and syntactic structures. Errors are rare. • Communication flows smoothly, with logical organization and clear expression. • Utterances are fluent, with clear pronunciation and natural intonation.
5	110	• Responses are adequate and relevant to the topic. • Responses show a range of syntactic structures. Errors may occur, but they do not impede understanding. • Vocabulary range and accuracy are sufficient to complete the task. • Speech is fluent. Utterances are generally coherent and clear in meaning. • Utterances are fluent, with appropriate pronunciation and intonation.
4	90	• Responses are adequate and relevant to the topic. • Responses show a good control of basic syntactic structures and vocabulary. Errors do not impede understanding. Complex ideas are not explained succinctly. They are expressed by using simple language or paraphrasing. • Speech is generally fluent. • Utterances are comprehensible, with easily recognizable sounds and intonation.
3	80	• Responses are largely relevant to the topic. • Responses show a good control of basic syntactic structures and vocabulary. Errors do not impede understanding. Complex ideas are not explained succinctly. They are expressed by using simple language or paraphrasing. • Speech is generally fluent. • Utterances are comprehensible, with easily recognizable sounds and intonation.
2	60	• Responses are largely relevant to the topic. • Responses show some control of basic syntactic structures and a limited control of vocabulary. Errors occur frequently and impede understanding. • Speech is slow, accompanied by frequent inappropriate pauses. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation; communication is sometimes impeded.
1	50	• Responses are partially relevant to the topic. • Responses show some control of basic syntactic structures and a limited control of vocabulary. Errors occur frequently and impede understanding. • Speech is slow, accompanied by frequent inappropriate pauses. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation; communication is sometimes impeded.
0	40 and below	• Responses are either too short or completely irrelevant to the topic. • Responses contain too many errors and show a very limited control of vocabulary. Utterances are mostly unintelligible and incomprehensible. This results in poor expression of ideas.

Part 3. Giving a Short Presentation

Band	Scale score	Description
6	150	• Responses are relevant to the topic and more than adequate. • Responses show appropriate use of a wide range of syntactic structures and vocabulary. Errors are rare. • Communication flows smoothly, with logical organization and clear expression. • Utterances are fluent, with clear pronunciation and natural intonation.
5	120	• Responses are adequate and relevant to the topic. • Responses show a range of syntactic structures. Errors may occur, but they do not impede understanding. • Vocabulary range and accuracy are sufficient to complete the task. • Speech is fluent. Utterances are generally coherent and clear in meaning. • Utterances are fluent, with appropriate pronunciation and intonation.
4	110	• Responses are largely relevant to the topic. • Responses show a range of syntactic structures. Errors may occur, but they do not impede understanding. • Vocabulary range and accuracy are sufficient to complete the task. • Speech is fluent. Utterances are generally coherent and clear in meaning. • Utterances are fluent, with appropriate pronunciation and intonation.
3	90	• Responses are largely relevant to the topic. • Responses show a good control of basic syntactic structures and vocabulary. Errors do not impede understanding. Complex ideas are not explained succinctly. They are expressed by using simple language or paraphrasing. • Speech is generally fluent. • Utterances are comprehensible, with easily recognizable sounds and intonation.
2	80	• Responses are partially relevant to the topic. • Responses show a good control of basic syntactic structures and vocabulary. Errors do not impede understanding. Complex ideas are not explained succinctly. They are expressed by using simple language or paraphrasing. • Speech is generally fluent. • Utterances are comprehensible, with easily recognizable sounds and intonation.
1	60	• Responses are partially relevant to the topic. • Responses show some control of basic syntactic structures and a limited control of vocabulary. Errors occur frequently and impede understanding. • Speech is slow, accompanied by frequent inappropriate pauses. • Utterances show unintelligibility in individual sounds and inappropriate use of intonation; communication is sometimes impeded.
0	40 and below	• Responses are either too short or completely irrelevant to the topic. • Responses contain too many errors and show a very limited control of vocabulary. Utterances are mostly unintelligible and incomprehensible. This results in poor expression of ideas.

5. Reporting scores on the BESTEP as a whole

Scores on the BESTEP Speaking test are reported to test takers in the form of a letter or email. Test takers will receive score reports including part scores and the overall score. In addition, schools registering for a session will also receive a report on overall scores of candidates in that session.

The part scores are weighted on the basis of task difficulty for each part. They are then combined to produce an overall scaled score out of 360. Weights for Part 1, Part 2, and Part 3 are 22% (section score 80), 36% (section score 130), and 42% (section score 150), respectively.

A score equivalent to CEFR B2 level (280 and above) is recommended by the Ministry of Education (MOE) of Taiwan to be the threshold for taking EMI courses in universities. Students who reach a score equivalent to CEFR B1 level (230-275) are recommended to take EAP courses, while students who reach a score equivalent to CEFR A2 level are recommended to take EGP courses in universities. The ultimate goal is to prepare all students for EMI courses. For more information, please visit the official website for The Program on Bilingual Education for Students in College, BEST (https://best.twaea.org.tw) and BEST Test of English Proficiency, BESTEP (https://bestep.tw).

6. Assessment results and analysis

Test response data will be routinely collected and analyzed for the BESTEP.

Statisticians at the LTTC are responsible for collecting and analyzing test response data from the assessment.

Test takers' background information, including gender, age, and majors in universities or colleges, are also collected for further analysis.

The BESTEP Speaking test will be launched in September 2023, and subsequent results will be reported. For now, we report test results on the pretest version. The BESTEP Speaking pretest was administrated in September, 2022. Around 800 students from 15 universities participated in the pretest, including undergraduate and graduate students with a variety of majors. Forty-five percent of the participants are male and 55% are female. Most of these schools have sent their students to take other LTTC tests before. Based on their previous performances, participants' English proficiency was expected to be within CEFR A2 to B2 level, and therefore the pretest sample was considered as being representative.

The mean score and standard deviation of the BESTEP Speaking pretest are 240 (out of 360) and 58, respectively. The Speaking pretest score ranges from 35 to 360. Around 29% of test takers are considered to be at the CEFR B2 (or above) level; 37% of test takers are at the CEFR B1 level; and, 28% of test takers are at the CEFR A2 level.

Further analyses on the background information of test takers and their BESTEP Speaking pretest scores indicate that scores are not significantly different due to test takers' gender or majors.

Gender	Male	Female
Average Speaking pretest score (out of 360)	238	233

College	Science / Engineering	Liberal Arts	Management	Biology/ Agriculture / Medicine	Electrical Engineering / Computer Science
Average Speaking pretest score (out of 360)	249	261	262	251	260

A reliability coefficient is an index that describes the consistency of test scores across contexts such as different times, items, or raters. The higher a reliability coefficient, the more trustworthy a test score is. Types of reliability coefficients used in the BESTEP Speaking test are parallel-form reliability and inter-rater reliability. The parallel-form reliability of the BESTEP speaking test was obtained using a Multifaceted Rasch analysis on scores from two alternate forms. The overall reliability was .93, which indicated that the two test forms used in the pretest are closely comparable forms in terms of difficulty. The inter-rater reliability was estimated by correlating the marking scores given by two independent raters of each part; the resulting reliability estimates of Part 1, Part 2, and Part 3 were .93, .97, and .92, respectively. This suggested that raters in general produced very similar ratings in judging the speaking abilities of the same test-takers.

To find out more about how scores on [the assessment] are calculated and reported…

For more information, please see:

1. Tung, S., Wu, J. R. W., & Wu, R. Y. F (Eds.), (28 January 2022) Initiative for the National Speaking and Writing Assessment for College Students, Phase 1 Progress Report (全國大專校院英語說寫評量檢測計畫：題庫與施測認證系統建置計畫第一次成果報告)
2. Tung, S., Wu, J. R. W., & Wu, R. Y. F (Eds.), (28 January 2023). Initiative for the National Speaking and Writing Assessment for College Students, Phase 2 Progress Report (題庫與施測認證系統建置計畫第二次成果報告).
   Research reports have been written in Chinese (with an abstract in both Chinese and English) and so far for internal use only. They may be published at a later date with the permission of the Ministry of Education.
3. Wu, J. R. W., Wu, R. Y. F., & Lin, A. C. W. (2023, June 5-9), *Co-constructing with stakeholders the performance descriptors for an English productive skills test* [Conference presentation]. 44th Language Testing Research Colloquium, New York City, NY, United States.
   https://ltrc2023.weebly.com/uploads/1/4/3/6/143613600/ltrc_2023_schedule_updated_6.2. 23.pdf

Appendix B. CEFR levels and descriptors identified as relevant to the BESTEP Parts

Section 3.1

Under the heading of *Oral production*, eight of the 14 judges identified the CEFR scale for *Sustained monologue: Describing experience* as relevant to Parts 1 and 2 of BESTEP Speaking and three considered it relevant to Part 3. There are many descriptors on this scale, and this may partly explain why it proved such a popular choice for judges. 10 thought that the B1 level of the *Describing experience* scale reflected what the test takers were required to do. The B1 level includes the following descriptors: *Can clearly express feelings about something experienced and give reasons to explain those feelings. Can give straightforward descriptions on a variety of familiar subjects within their field of interest. Can reasonably fluently relate a straightforward narrative or description as a sequence of points. Can give detailed accounts of experiences, describing feelings and reactions. Can relate details of unpredictable occurrences, e.g. an accident. Can relate the plot of a book or film and describe their reactions. Can describe dreams, hopes and ambitions. Can describe events, real or imagined. Can narrate a story*). Seven judges found the A2 level of the scale relevant (*Can tell a story or describe something in a simple list of points. Can describe everyday aspects of their environment, e.g. people, places, a job or study experience. Can give short, basic descriptions of events and activities. Can describe plans and arrangements, habits and routines, past activities and personal experiences. Can use simple descriptive language to make brief statements about and compare objects and possessions. Can explain what they like or dislike about something. Can describe their family, living conditions, educational background, present or most recent job. Can describe people, places and possessions in simple terms. Can express what they are good at and not so good at (e.g. sports, games, skills, subjects). Can briefly describe what they plan to do at the weekend or during the holidays*). Only two judges suggested that the A1, B2 or C1 levels on this scale also closely matched what test takers were asked to do on the test.

Sustained monologue: Giving information was among judges' most popular choice of scale for all three test Parts. It was identified by at least six judges as relevant to each of the three Parts of the test with the largest number, nine, identifying it as relevant to Part 2. The B1 descriptors (*Can explain the main points in an idea or problem with reasonable precision. Can describe how to do something, giving detailed instructions. Can report straightforward factual information on a familiar topic, for example to indicate the nature of a problem or to give detailed directions, provided they can prepare beforehand*) were considered relevant by nine judges with five identifying the A2 descriptor (*Can give simple directions on how to get from X to Y, using basic expressions such as "turn right" and "go straight", along with sequential connectors such as "first", "then" and "next"*) and four the B2 level (*Can communicate complex information and advice on the full range of matters related to their occupational role. Can communicate detailed information reliably. Can give a clear, detailed description of how to carry out a procedure*).

Sustained monologue: Putting a case (e.g. in a debate) was considered relevant to Parts 2 and 3 of the BESTEP speaking test by nine and 13 judges respectively, although none identified it with Part 1. Panellists were most likely to select the B1 level (six judges: *Can develop an argument well enough to be followed without difficulty most of the time. Can give simple reasons to justify a viewpoint on a familiar topic.Can express opinions on subjects relating to everyday life, using simple expressions. Can briefly give reasons and explanations for opinions, plans and actions. Can explain whether or not they approve of what someone has done and give reasons to justify this opinion*) and B2 level (11 judges: *Can develop an argument systematically with appropriate highlighting of significant points, and relevant supporting detail. Can develop a clear argument, expanding and supporting their points of view at some length with subsidiary points and relevant examples. Can construct a chain of reasoned argument. Can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options*). Addressing audiences was particularly identified with Part 3 of the test: eleven judges listing it for this Part of the test compared with just one for each of Parts 1 and 2. The B2 level was chosen as relevant by 10 judges (*Can give a clear, systematically developed presentation, with highlighting of significant points, and relevant supporting detail. Can depart spontaneously from a prepared text and follow up interesting points raised by members of the audience, often showing remarkable fluency and ease of expression. Can give a clear, prepared presentation, giving reasons in support of or against a particular point of view and giving the advantages and disadvantages of various options. Can take a series of follow-up questions with a degree of fluency and spontaneity which poses no strain for either themselves or the audience*). Four judges picked out the B1 level (*Can give a prepared presentation on a familiar topic within their field, outlining similarities and differences (e.g. between products, countries/regions, plans). Can give a prepared straightforward presentation on a familiar topic within their field which is clear enough to be followed without difficulty most of the time, and in which the main points are explained with reasonable precision. Can take follow-up questions, but may have to ask for repetition if the delivery is rapid*).

Under the heading of *Oral interaction,* judges generally limited their choices of scale to Part 1 of the test, which requires test takers to respond to a series of questions, rather than Parts 2 and 3, both of which require a lengthier response to a single prompt. The exception was the scale for *Formal discussion (meetings),* perhaps because this includes descriptors relating to the expression of ideas or points of view. This scale was selected as relevant to Part 3 by five judges (two selected it for Part 2 and none for Part 1. Here it was the B2 level that was seen to be most salient (five judges: *Can keep up with an animated discussion, identifying accurately arguments supporting and opposing points of view. Can use appropriate technical terminology when discussing their area of specialisation with other specialists. Can express their ideas and opinions with precision, and present and respond to complex lines of argument convincingly. Can participate actively in routine and non-routine formal discussion. Can follow the discussion on matters related to their field, understand in detail the points given prominence. Can contribute, account for and sustain their opinion, evaluate alternative proposals and make and respond to hypotheses*). For Part 1, *Understanding an interlocutor* was selected by eight judges, *Information exchange* by six and *Conversation* and *Informal discussion (with friends)* were each selected by five. On the scales for *Understanding an interlocutor* and *Conversation*, the A2 level was most often chosen as relevant. For *Understanding an interlocutor,* seven judges chose A2 (*Can understand enough to manage simple, routine exchanges without undue effort. Can generally understand clear, standard speech/sign on familiar matters directed at them, provided they can ask for repetition or reformulation from time to time. Can understand what is said clearly, slowly and directly to them in simple everyday conversation; can be made to understand, if the interlocutor can take the trouble*). For *Conversation,* five judges chose A2 (*Can establish social contact (e.g. greetings and farewells, introductions, giving thanks). Can generally understand clear, standard language on familiar matters directed at them, provided they can ask for repetition or reformulation from time to time. Can participate in short conversations in routine contexts on topics of interest. Can express how they feel in simple terms, and express thanks. Can ask for a favour (e.g. to borrow something), can offer a favour, and can respond if*

someone asks them to do a favour for them. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of their own accord, though they can be made to understand if the interlocutor will take the trouble. Can use simple, everyday, polite forms of greeting and address. Can converse in simple language with peers, colleagues or members of a host family, asking questions and understanding answers relating to most routine matters. Can make and respond to invitations, suggestions and apologies. Can express how they are feeling, using very basic stock expressions. Can state what they like and dislike). Six judges selected the B1 level of the scale for *Information exchange* (*Can exchange, check and confirm accumulated factual information on familiar routine and non-routine matters within their field with some confidence. Can summarise and give their opinion about a short story, article, talk, discussion, interview or documentary and answer further questions of detail. Can find out and pass on straightforward factual information. Can ask for and follow detailed directions. Can obtain more detailed information. Can offer advice on simple matters within their field of experience*.) and seven chose B1 for *Informal discussion (with friends)* (*Can follow much of what is said around them on general topics, provided interlocutors avoid very idiomatic usage and articulate clearly. Can express their thoughts about abstract or cultural topics such as music or films. Can explain why something is a problem. Can give brief comments on the views of others. Can compare and contrast alternatives, discussing what to do, where to go, who or which to choose, etc. Can generally follow the main points in an informal discussion with friends provided they articulate clearly in standard language or a familiar variety. Can give or seek personal views and opinions in discussing topics of interest. Can make their opinions and reactions understood as regards solutions to problems or practical questions of where to go, what to do, or how to organise an event (e.g. an outing). Can express beliefs, opinions and agreement and disagreement politely*).

Appendix C. Results of online Familiarisation: CEFR Questionnaire (based on responses by 13 experts)

General linguistic range

Descriptor	CEFR	A1	A2	A2+	B1	B1+	B2	B2+	C1	% =	% +/-1
13. Produces brief, everyday expressions in order to satisfy simple needs of a concrete type (e.g. personal details, daily routines, wants and needs, requests for information).	A2	4	9							69.2%	86.7%
3. Has a repertoire of basic language which enables them to deal with everyday situations with predictable content, though they will generally have to compromise the message and search for words.	A2+		4	6	3					46.2%	86.7%
16. Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and film.	B1+				3	2	5	1	2	15.4%	66.7%
10. Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so.	B2					1	7	4	1	53.8%	80.0%

Vocabulary range

Descriptor	CEFR	A1	A2	A2+	B1	B1+	B2	B2+	C1	% =	% +/-1
7. Has a basic vocabulary repertoire of words and phrases related to particular concrete situations.	A1	6	3	2	2					46.2%	60.0%
5. Occasional minor slips, but no significant vocabulary errors.	C1				1	1	2	1	8	61.5%	60.0%
18. Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.	B1		2	2	8	1				61.5%	73.3%
15. Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.	B2			1	1	3	8			61.5%	73.3%

Grammatical accuracy

Descriptor	CEFR	A1	A2	A2+	B1	B1+	B2	B2+	C1	% =	% +/-1
1. Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	B1		1		12					92.3%	92.3%
9. Communicates with reasonable accuracy in familiar contexts; generally good control, though with noticeable mother-tongue influence. Errors occur, but it is clear what they are trying to express.	B1+			1	7	4	1			30.8%	92.3%
19. Good grammatical control; occasional ,'slips', or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	B2+				2		4	5	2	38.5%	84.6%
17. Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.	C1							2	11	84.6%	100.0%

Prosodic features

Descriptor	CEFR	A1	A2	A2+	B1	B1+	B2	B2+	C1	% =	% +/-1
28. Uses the prosodic features of a limited repertoire of simple words and phrases intelligibly, in spite of a very strong influence on stress, rhythm and/or intonation from the other language(s) they speak; their interlocutor needs to be collaborative.	A1	8	3	1	1					61.5%	84.6%
23. Uses the prosodic features of everyday words and phrases intelligibly, in spite of a strong influence on stress, intonation and/or rhythm from the other language(s) they speak.	A2+		5	2	5		1			38.5%	92.3%
26. Conveys their message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from the other language(s) they speak.	B1		2	1	9		1			69.2%	92.3%
22. Employs prosodic features (e.g. stress, intonation, rhythm) to support the message they intend to convey, though with some influence from the other languages they speak.	B2				1	4	4	2	2	30.8%	76.9%
27. Varies intonation and place stress correctly in order to express precisely what they mean to say.	C1						3		10	76.9%	76.9%

Sound articulation

Descriptor	CEFR	A1	A2	A2+	B1	B1+	B2	B2+	C1	% =	% +/-1
35. Articulates a limited number of sounds, so that speech is only intelligible if the interlocutor provides support (e.g. by repeating correctly and by eliciting repetition of new sounds).	A1	12	1							92.3%	100.0%

30. Pronunciation is generally intelligible when communicating in simple everyday situations, provided the interlocutor makes an effort to understand specific sounds.	A2		11	2						84.6%	100.0%
31. Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words they are less familiar with.	B1	1	5	1	6					46.2%	53.9%
33. Articulates a high proportion of the sounds in English clearly in extended stretches of production; is intelligible throughout, despite a few systematic mispronunciations.	B2				3	1	9			69.2%	76.9%
36. Articulates virtually all the sounds of English with a high degree of control. They can usually self-correct if they noticeably mispronounce a sound.	C1					1	2	2	8	61.5%	76.9%

BESTEP

Descriptor BESTEP:	Pt1	Pt2	Pt3	A1	A2	A2+	B1	B1+	B2	B2+	C1	Mode	Median
6. Makes too many errors and shows a very limited control of vocabulary.		0	0	13								A1	A1
11. Has a limited control of vocabulary. Errors occur frequently and impede understanding.		1/2	1	10	2		1					A1	A1
34. Utterances show unintelligibility in individual sounds and inappropriate use of intonation; communication is sometimes impeded.		1/2	1	8	4	1						A1	A1
24. Speech is slow, accompanied by frequent inappropriate pauses.		1/2	1	6	6	1						A1/A2	A2
4. Has some control of basic syntactic structures. Errors occur frequently and impede understanding.		1/2	1	3	10							A2	A2
32. Utterances are comprehensible, with easily recognizable sounds and intonation.		3/4	2/3				7	3	2		1	B1	B1
2. Good control of basic syntactic structures and vocabulary. Errors do not impede understanding.		3/4	2/3			1	7	4	1			B1	B1
8. Has a good control of basic syntactic structures and vocabulary. Errors do not impede understanding.		3/4	2/3				7	3	3			B1	B1
25. Speech is generally fluent.	5	3/4	2/3		1		2	3	7			B2	B2
14. Demonstrates a range of syntactic structures. Errors may occur, but they do not impede understanding.		5	4/5				2	3	8			B2	B2
20. Speech is fluent. Utterances are generally coherent and clear in meaning.		5	4/5				1		6	4	2	B2	B2
12. Makes appropriate use of a wide range of vocabulary and syntactic structures. Errors are rare.		6							3	1	9	C1	C1
29. Utterances are fluent, with clear pronunciation and natural intonation.		6	6					1	2	2	8	C1	C1
21. Communication flows smoothly, with logical organization and clear expression.		6	6						2		11	C1	C1

Table 4 shows the results for the Familiarisation exercise based on the online *CEFR Questionnaire.* This involved assigning descriptors from the CEFR and the BESTEP rating scales to a CEFR level. Table 4 is divided by scale with the BESTEP descriptors appearing in the final section. Each descriptor is numbered, reflecting the order in which they appeared in the online *CEFR Questionnaire*. The first two columns display the descriptors and their CEFR level. The columns to the right show the number of judges assigning each descriptor to each CEFR level and the final two columns show the percentage of correct descriptor assignments (%=) within each row and the percentage within one CEFR level (% +/-1). Here criterion and plus levels are treated as separate so that a B1+ descriptor assigned to B1 is treated as being one level below its correct location. Because each Part of the BESTEP Speaking test has its own rating scale, in the final section, which displays the BESTEP descriptors, the three columns to the right of the list of descriptors show the Part of the BESTEP test the descriptor relates to and the level on the scale at which it appears. For example, descriptor 11: *Has a limited control of vocabulary. Errors occur frequently and impede understanding* does not appear on the scale for BESTEP Part 1, is included at Bands 1 and 2 on the scale for Part 2, but only appears at Band 1 for Part 3. The final two columns show which CEFR level each descriptor was assigned to by the most judges (the mode) and by the seventh ranked judge (the midpoint, or median) when ordering judges from the one assigning the highest to the one assigning the lowest CEFR level to a descriptor.