Linking the GEPT Writing Subtest (Part 1) to the Common European Framework of Reference (CEFR)

Jason Fan
Ute Knoch
Ivy Chen

Language Testing Research Centre, University of Melbourne

Final report - January 2021

Contents

Abstract

Previous research has aligned the GEPT reading, listening, speaking, and writing (Part 2) subtests to the Common European Framework of Reference (CEFR) (Brunfaut & Harding, 2014; Green & Inoue, 2017; Knoch & Frost, 2016; Wu & Wu, 2010). No attempt, however, has been made to link Part 1 of the GEPT writing subtest (Chinese-English translation) to the CEFR levels. In view of the recent publication of the illustrative descriptors for mediation in the companion volume to the CEFR (Council of Europe, 2018), this study aimed at linking Part 1 of the GEPT writing subtest at the intermediate and high-intermediate level to the CEFR, following the four stages of familiarisation, specification, standardisation, and validation set out in the CEFR manual (Council of Europe, 2009). Twelve panellists participated in this alignment study, with eight test 'insiders' based in Taipei and four test 'outsiders' based in Melbourne. Two examinee-centred standard-setting methods (i.e. the 'Contrasting Groups' and 'Borderline Group' methods) were used in combination in the alignment process. In addition to aligning the GEPT translation tasks to the CEFR, this study also explored, through a think-aloud study, the processes through which the panellists linked the GEPT translation scripts to the CEFR levels. In their study aiming to link Part 2 of the GEPT writing subtest (i.e. guided writing) to the CEFR, Knoch and Frost (2016) recommended that the pass scores of the intermediate and high intermediate level be lowered by one point based on their linking results. The findings of this project are generally consistent with Knoch and Frost (2016), suggesting that the cut scores for both levels be lowered from 4 to 3.5, if half point scores could be used in the GEPT score reports. Based on our experiences while undertaking this study, we also provide a few recommendations for future researchers linking translation tasks in language tests to the CEFR.

1. Introduction

This project aims to link Part 1 of the General English Proficiency Test (GEPT) writing subtest at intermediate and high-intermediate levels to the Common European Framework of Reference (CEFR). In this part of the GEPT writing test, test takers are asked to translate a short passage from Chinese into English. According to Wu (2012), these two levels attract a large proportion of the GEPT test takers. Previous linking research has aligned the GEPT to the CEFR, including reading (Wu & Wu, 2010), listening (Brunfaut & Harding, 2014), Part 2 of the GEPT writing subtest (i.e. guided writing, Knoch & Frost, 2016), and speaking (Green & Inoue, 2017). No research, however, has been conducted to link Part 1 of the GEPT writing subtest to the CEFR levels.

Since its inception, the CEFR has been highly influential, both in Europe and globally. Major international language testing agencies have either already aligned their tests to the CEFR levels (e.g., Fleckenstein, Keller, Krüger, Tannenbaum, & Köller, 2020; Kecker & Eckes, 2010; Lim, Geranpayeh, Khalifa, & Buckendahl, 2013) or are feeling the pressure to do so. This study is particularly meaningful against the backdrop of the recent publication of the CEFR companion volume where illustrative descriptors for mediation are provided to CEFR users (Council of Europe, 2018; see also North & Piccardo, 2016). It is yet unclear what specific difficulties might be encountered when linking translation tasks to the new mediation descriptors. As such, the objectives of this proposed study are two-fold: 1) linking Part 1 of the GEPT writing subtest to the CEFR levels; and 2) exploring the panellists' processes in linking translation tasks to the mediation descriptors in the CEFR.

1.1 Linking language tests to the CEFR

Developed by the Council of Europe, the CEFR represents one of the major initiatives by the Council of Europe to provide common reference levels for teaching and learning of all languages in Europe. Specifically, the CEFR aims to promote and facilitate cooperation among educational institutions in different countries, provide a basis for the mutual recognition of language qualifications, and assist language learners, teachers, course designers, and examination agencies to situate and coordinate their efforts (Council of Europe, 2001, p. 25). The CEFR consists of six reference levels across three bands:

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

The six common reference levels in the CEFR aim to provide a common metalanguage for the language education profession and to facilitate the mutual recognition of language qualifications, as indicated by courses taken or examinations passed. In the CEFR, language proficiency is described in a set of scales covering a range of skills including reading, listening, writing, and

speaking, as well as a range of communicative competences, with illustrative 'can-do' descriptors provided in *Common European Framework of Reference for Languages: Learning, teaching, and assessment* (Council of Europe, 2001).

Motivated by requests from CEFR users to continue to develop illustrative descriptors of second/foreign language proficiency, the Council of Europe recently published the companion volume (Council of Europe, 2018), which updated the scales and descriptors of language proficiency for online interaction, mediation, plurilingual and pluricultural competence, signing competence, and young learners. Particularly relevant to this project are the new scales and descriptors for mediation, a term which covers translation and interpretation. We will explain the notion of mediation in the CEFR in the next section.

The enormous impact of the CEFR has not only been felt in Europe, but indeed globally. In the field of language assessment, efforts have been made to align major international language tests to the CEFR levels (Milanovic & Weir, 2010), including the IELTS (Lim et al., 2013), TOEFL iBT (e.g., Fleckenstein et al., 2020; Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015), and Pearson Test of English Academic (De Jong & Zheng, 2016). To assist test providers in mapping their language tests to the CEFR levels, the Council of Europe piloted a set of recommended linking procedures and subsequently published a manual for relating language examinations to the CEFR (Council of Europe, 2009). According to the manual, four stages should be followed in linking a language test to the CEFR, namely a) familiarisation, b) specification, c) standardisation, and d) validation. These four stages are illustrated in Figure 1 below.

During the familiarisation stage, a range of activities can be designed and organised to help the panel of judges who will be involved in the linking process gain an in-depth understanding of aspects of the CEFR relevant to the linking purpose. The specification stage involves an analysis of the test content, tasks and assessment criteria in relation to the relevant categories of the CEFR. Next, in the standardisation stage, the panel of judges participate in a standard setting procedure to map test takers' performances on a language test to the CEFR levels. Finally, the validation stage aims to provide a range of evidence to support the linking claims.

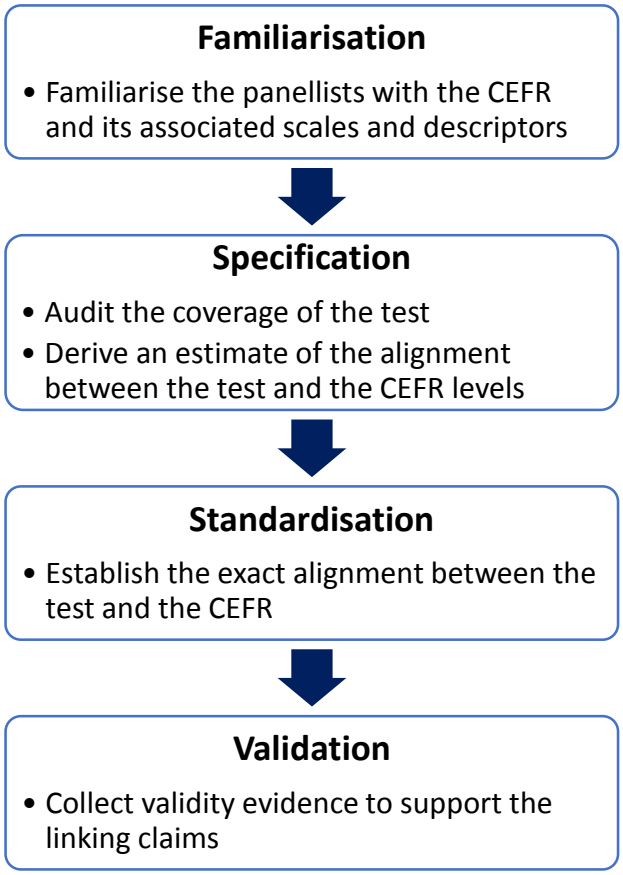

Figure 1. The four stages of aligning language examinations to the CEFR

1.2 Mediation in the CEFR companion volume

Mediation is not a novel concept in the CEFR. In the previous editions of the CEFR, it was considered to be one of the four major communication modes, together with reception, production, and interaction. According to the 2001 edition of the CEFR (Council of Europe, 2001), a language user mediates when he or she acts as an intermediary between interlocutors who are unable to understand each other directly, most likely because they speak different languages. Spoken interpretation and written translation are two typical mediation activities. Despite the significance of mediation in language use in a plurilingual and pluricultural society, the previous editions of the CEFR did not provide a detailed conceptual treatment of this notion; neither did they provide separate scales or illustrative descriptors to delineate mediation competence. The companion volume published in 2018 fills this gap by providing scales and descriptors of mediation competence.

It should be noted, however, that the companion volume adopts a broad conceptualisation of the notion of mediation (North & Piccardo, 2016). In the companion volume, mediation is conceived as a complex process where the learner acts as a social agent who creates bridges and helps to

construct or convey meaning, sometimes within the same language, sometimes from one language to another. The descriptive scheme and illustrative descriptors attest to this conceptualisation. In the descriptive scheme (see Figure 2), mediation falls into two broad categories: mediation activities and mediation strategies. The former encompasses mediating texts and concepts, whereas the latter includes strategies to explain a new concept and to simplify a text. The mediation activity the present study, that is, written translation, falls under the category of 'mediating a text', where the descriptors of 'translating a written text in writing' (see Appendix I) are most relevant to this project.

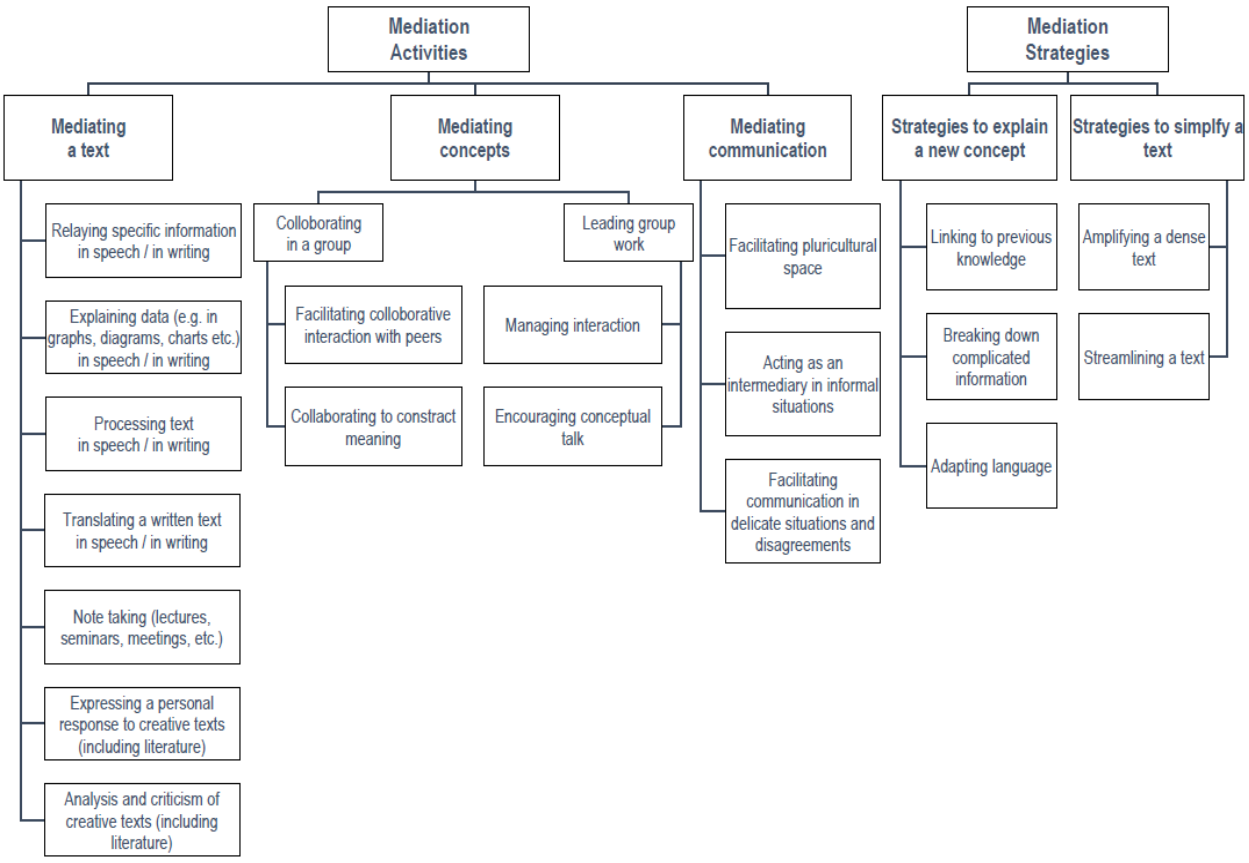

Figure 2. Descriptive scheme of mediation in the companion volume to the CEFR (from Council of Europe, 2018, p. 104)

While these descriptors were employed in this study, we also decided to include some writing descriptors in the linking process because the number of translation descriptors at each CEFR level was very limited. For example, there are only two descriptors at B1 and B2, the two target CEFR levels of this study in the translation scale, making it difficult for panellists to link the GEPT translation task to the CEFR levels. The inclusion of the writing descriptors in this study can also be justified by examining the scoring rubrics for the translation tasks at the two GEPT levels (see Appendix II) which are explained in more detail in the next section. The scoring rubrics include some aspects of performance which are typically used to assess a writing performance, such as

organisation, coherence, and grammatical and lexical accuracy. This also explains why the translation task is a part of the GEPT writing subtest. As such, both the translation and writing scales in the CEFR were employed in this linking study.

1.3 General English Proficiency Test

The GEPT is developed and administered by the Language Training and Testing Centre (LTTC), based in Taiwan. The test made its debut in 2000 designed to help individuals gauge their English proficiency and assist employers and educational institutions with selection and placement. Furthermore, it was implemented with the agenda of fostering lifelong education and enhancing people's English proficiency in Taiwan (Roever & Pan, 2008). The launch of the GEPT was a major event in English language education in Taiwan (Shih, 2010).

The GEPT consists of five levels: elementary, intermediate, high-intermediate, advanced and superior. Writing is assessed at all levels but in different formats. Table 1 outlines in more detail the writing subtest at the different test levels of the GEPT. The focus of this study is the Chinese-English (C-E) translation task in the GEPT writing subtest at the intermediate and high-intermediate levels. At both levels, test takers are required to translate a short Chinese paragraph into English. The paragraph at the intermediate level has approximately 90-100 Chinese characters; that at the high-intermediate level is slightly longer, with 120-130 Chinese characters. For the intermediate level, the topics in the paragraph are relevant to the life and cultural experiences of Taiwanese learners of English and the level of difficulty is appropriate for average senior high school graduates in Taiwan; for the high-intermediate level, topics are drawn from the life and cultural experiences of Taiwanese learners of English as well as contemporary issues in a Taiwanese context, and the level of difficulty is appropriate for average university non-English major graduates in Taiwan. At both levels, the translation task accounts for 40% of test takers' writing scores. For the translation task at both levels, a 6-point holistic rating scale (from 0-5) is adopted to rate test takers' performance (see Appendix II). The rating scale mainly focuses on a) the correspondence in meaning between the original text and the translation; b) organisation and coherence; and c) accuracy in terms of vocabulary, grammar, spelling, punctuation, etc.

Table 1. GEPT writing tasks at different test levels

Level	Task types	No of items	Duration
Superior	A 750-word essay based on a 10-15 min. video/radio program and a 3000-word article	1	3 hours
Advanced	Summarising main ideas from verbal input and expressing opinions	1	60 mins
	Summarising main ideas from non-verbal input and providing solutions	1	45 mins
High-intermediate	Chinese-English translation	1	20 mins*
	Guided writing	1	30 mins
Intermediate	Chinese-English translation	1	16 mins*
	Guided writing	1	24 mins
Elementary	Sentence writing Paragraph writing	16	40 mins

*The estimated duration based on the weighting of this task.

1.4 Research questions

This study aims to link Part 1 of the GEPT writing subtest to the CEFR levels. Following Brunfaut and Harding (2014), Knoch and Frost (2016), and Green & Inoue (2017), this study adopted a 'twin-panel' approach to compare the judgements of those familiar with the GEPT (the Taipei Group, or test 'insiders') with the judgements of those with little if any prior exposure to the GEPT (the Melbourne Group, or test 'outsiders'), thus providing a rigorous means of cross-validating the panellists' judgements. In addition, we were also interested in exploring the processes through which the panellists linked the GEPT translation tasks to the CEFR scales and descriptors.

Specifically, this study investigated the following three research questions:

(1) How do the score levels of the GEPT writing subtest (Part 1) relate to the CEFR levels?
(2) How do the judgements of test 'outsiders' compare to these of test 'insiders'?
(3) What are the processes through which the panellists linked the GEPT translation tasks to the CEFR scales and descriptors?

2. Methodology

To investigate the three research questions, both qualitative and quantitative data were collected in this study. With regard to the first research question, we followed the four stages set out in the

CEFR manual to align language examinations to the CEFR (Council of Europe, 2009): familiarisation, specification, standardisation, and validation (see Figure 1). The second research question was addressed through many-facets Rasch analysis of the judgements of the two groups of panellists as well as a comparison of their linking results at the group level. The last research question was explored through an introspective think-aloud study. In what follows, we provide details concerning the participants, procedures and materials, and data analysis in this study.

2.1 Participants

Twelve panellists participated in this study, including eight test 'insiders' based in Taipei and four test 'outsiders' based in Melbourne. All participating panellists were native speakers of Mandarin and were proficient in both Chinese and English. Table 2 below shows some of the background details of the 12 panellists.

Table 2. Participating panellists in this study

Panellist	Gender	Age	Group	Occupation	Familiarity with the CEFR
A	Male	31-40	Insider	LT R&D*	Familiar
B	Female	31-40	Insider	LT R&D	Familiar
C	Female	31-40	Insider	LT R&D	Familiar
D	Female	31-40	Insider	LT R&D	Familiar
E	Female	41-50	Insider	LT R&D	Familiar
F	Female	41-50	Insider	LT R&D	Familiar
G	Female	41-50	Insider	LT R&D	Familiar
H	Female	31-40	Insider	LT R&D	Familiar
I	Male	51-60	Outsider	Translator/Interpreter	Unfamiliar
J	Male	21-30	Outsider	Translator/Lecturer	Somewhat familiar
K	Male	51-60	Outsider	Translator/Interpreter	Unfamiliar
L	Female	31-40	Outsider	Translator/Interpreter	Unfamiliar

* Language testing research and development

As indicated in this table, there were eight females and four males; except for Panellist J, all other panellists were in the age groups of 31-40 (n = 6), 41-50 (n = 3), or 51-60 (n = 2). The eight test 'insiders' all worked in the area of language testing research and development; the four test 'outsiders' worked as translators or interpreters. One of them was also working as a university lecturer when the data was collected. The participants were asked to report their self-assessed levels of familiarity with the CEFR. Table 1 indicates that all test 'insiders' were familiar with the CEFR. For the four test 'outsiders', however, only one of them was somewhat familiar with the CEFR; the rest of the group reported that they were unfamiliar with the CEFR.

We also asked the participants about their prior experience in language teaching and testing. All participants reported that they had experience in teaching English as a Foreign Language (EFL), though they had worked at different educational levels, including secondary schools, universities,

and private tutoring schools. Also, the amount of teaching experience varied significantly, ranging from 1 to 20-plus years. All participants had experience in language testing and assessment. For the test 'insiders', most of them had experience working as item writers, test designers, markers or researchers on EFL tests; most test 'outsiders', on the other hand, had experience working as markers for translation tests involving Chinese and English.

2.2 Procedures and materials

As mentioned previously, we followed the four stages set out in the CEFR linking manual, each of which is detailed below.

2.2.1 Familiarisation

A familiarisation workshop was conducted to help the panellists gain in-depth knowledge and understanding of the CEFR scales and descriptors. Before the workshop, the panellists were instructed to go through some preparatory activities, including a) reading the section in the 2001 CEFR publication on the salient features of the six reference CEFR levels; and b) visiting the CEFR training website (www.helsinki.fi/project/ceftrain/index.php.66.html) where they were asked to view the writing samples at different CEFR levels.

The workshop was conducted on Zoom, an online conferencing platform. One of the researchers from the Language Testing Research Centre, University of Melbourne (LTRC) facilitated the workshop. The first part of the workshop was dedicated to an introduction to the CEFR, including its purposes, the common reference levels, and the illustrative descriptors. We also briefly introduced the companion volume published in 2018 and covered how mediation competence was conceptualised in the volume, as it was particularly relevant to this project. The next part of the workshop focused on the GEPT writing test, and in particular, the C-E translation task at the intermediate and high-intermediate levels. The LTTC, the developer of the GEPT, provided the sample C-E tasks which were shown to the panellists. We also briefly explained the rating scales that were used to score test takers' performance on the translation task at the two GEPT levels.

Following the introduction to the CEFR and the GEPT writing tasks, the panellists were divided into groups of three to four to work on a few of the familiarisation activities recommended by the CEFR linking manual. The first activity asked participants to assign descriptors which were in jumbled order to each of the common reference levels in the CEFR. Participants were also asked to highlight the key elements at each level in the descriptors. In the next activity, the participants were asked to self-assess their proficiency of English or any other foreign language that they had studied, using the self-assessment scale in the CEFR. Given that the focus of this study was C-E translation, the panellists were then instructed to reassemble individual descriptors in the CEFR translation and writing scales in the correct order. The workshop lasted for three and a half hours. After the workshop, we asked each panellist to fill out an online questionnaire, aiming to gauge the effectiveness of the familiarisation workshop.

2.2.2 Specification

The purpose of the specification stage is to analyse the content of the test in this study in order to profile it in relation to CEFR categories and levels. A range of specification forms in the CEFR linking manual were used to analyse content coverage, task type, and the assessment criteria of the C-E translation task in the GEPT writing subtest. The specification stage was completed by researchers at the LTRC in collaboration with a colleague based in the LTTC; all participants were familiar with the CEFR. In total, four parallel C-E translation tests, two each at intermediate and high-intermediate levels, were analysed. The completed forms (A1-A8) are attached as Appendix III of this report. The focus of each form is outlined below:

- A1: general description of the GEPT and the C-E translation task at the intermediate and high-intermediate levels
- A2: test development and item writing
- A3: marking test takers' performance
- A4: grading and establishing pass marks
- A5: reporting results to test takers
- A6: analysis of test data and test review procedures
- A7: rationales for decisions made about test takers and test revisions
- A8: initial estimation of overall test level

An initial estimate of the alignment between the GEPT and the CEFR levels was derived from the specification stage, which has been supported by previous linking studies on the GEPT reading subtest (Wu & Wu, 2010), listening subtest (Brunfaut & Harding, 2014), and the guided writing task in the writing subtest (Knoch & Frost, 2016). The linking results are also available on the GEPT website (https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/alignment.htm), indicating that the GEPT intermediate level corresponds to B1 and the high-intermediate level to B2 on the CEFR. These initial estimates were scrutinised through this linking study.

2.2.3 Standardisation

Standard setting is a variety of systematic processes which entail the assignment of interpretative meaning to performances on language tests (Kenyon & Romhild, 2014); it aims to 'establish one or more cut scores on tests' (Cizek & Bunch, 2007, p. 13), thus creating categories of performance and a classification of examinees. Standard setting plays a significant role in language assessment development and validation because it relates to how a test taker's performance is interpreted (Kane, 2013), which in turn affects the decisions that are made about test takers (Bachman & Palmer, 2010). Test users such as schools, universities, and businesses rely on the standard setting results to make important decisions such as selection, placement, and recruitment.

A range of standard setting methods are available (e.g., Angoff methods, bookmark methods, body of work methods), each with its own strengths and weaknesses (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kenyon & Romhild, 2014), as illustrated in the CEFR linking manual

(Council of Europe, 2009). Researchers should select and implement linking methods based on their situation and intended functions (Fleckenstein et al., 2020).

In this project, we adopted a combination of two examinee-centred standard setting methods, that is, the contrasting group method and the borderline group method. According to the CEFR linking manual (Council of Europe, 2009, p. 67), these two linking methods are well suited to the context where the students are known to the panellists. In this study, instead of relying on panellists' familiarity with students in the classroom context, they were asked to classify test takers based on their performance on the C-E translation task in the GEPT writing subtest. These two methods were successfully employed by Knoch and Frost (2016) in linking the guided writing task in the GEPT writing subtest to the CEFR.

In the contrasting group method, panellists classify students into two groups: 'masters', that is, students are clearly above a particular performance level and 'non-masters', that is, students are clearly below that level. As mentioned previously, instead of basing their judgements on their knowledge of the test takers, the panellists in this study examined their translation scripts to determine whether they should be classified into the groups of 'masters' or 'non-masters'. At both the intermediate and high-intermediate levels, the LTTC provided the translation scripts from two parallel test forms and at different score levels. Table 3 below provides the details of the scripts. Each translation script had a score which had been awarded by certified GEPT raters based on the scoring rubrics (see Appendix II) but the participating panellists were not made aware of these scores. After the panellists made their judgements, the test score distributions of the two groups (that is, 'masters' and 'non-masters') were analysed to determine a cut score (Eckes, 2012).

Table 3. The translation scripts used in this study

Intermediate			High-intermediate		
Score level	N		Score level	N	
	Form A	Form B		Form A	Form B
2	4	4	2	4	4
3	4	4	3	4	4
4	4	4	4	4	4
5	4	4	5	4	4
Total	16	16	Total	16	16

When using the contrasting group method, panellists are expected to make clear, unambiguous judgements about the member status of the students being evaluated; however, this is unlikely to be always possible because some students may not fall clearly into the two categories of 'masters' and 'non-masters'. To address this problem, another linking method called the borderline group method was developed. When applying this method, panellists are asked to identify students who should be classified into the borderline group. The test score distribution

of the borderline group is then analysed to yield the cut score (Eckes, 2012; Council of Europe, 2009). Despite the intuitive appeal of this linking method, a major limitation is that the number of students that are classified into the borderline group in many practical circumstances is very small, thus making it difficult to generate a stable estimate of the cut score. Furthermore, compared with placing students into the 'master' and 'non-master' groups, it seems much more difficult for panellists to reach agreement over classifying the students into the borderline group (Eckes, 2012). In an effort to mitigate the limitations of each method, we employed both in this project.

The standardisation stage consisted of three steps, as illustrated by Figure 3 below. First, a standardisation training workshop was conducted. Before the workshop, we prepared eight GEPT translation scripts at different score levels, four at each of the two levels in the study. We invited two CEFR experts who also had extensive experience in C-E translation to examine the scripts and determine whether a script at the intermediate level should be classified into 'below B1', 'borderline', or 'at B1'. They were also asked to provide justifications for their judgements. The same procedure was repeated for the four scripts at the high-intermediate level. During the standardisation workshop, the panellists worked in groups of three or four; they reviewed the same scripts, put them into the categories, and discussed their judgements. One of the researchers facilitated this workshop, which lasted about three hours. Afterwards, we asked each panellist to fill out an online questionnaire to gauge the effectiveness of the workshop.

During the second step (the first round of standard setting), the panellists were instructed to work independently after the workshop to classify 32 translation scripts at the intermediate level (see Table 2) into two categories: 'below B1'and 'at B1', and a further 32 scripts at the high-intermediate level (see Table 2) into another two categories: 'below B2' and 'at B2'. A judgement form was used to facilitate the classification. If the panellists felt that it was difficult to assign a script into the given binary categories and thus that it was a borderline case, they were instructed to add this script to a separate judgement form (see Appendix IV for a sample judgement form). The panellists used both the translation descriptors and the CEFR Written Assessment Grid (i.e. the writing scale used in this study, see Appendix I) in the judgement process. After the panellists finished the first round of standard setting, they sent the judgement forms via email to one of the researchers who examined their judgement results, identified those judgements that apparently deviated from the rest of the group, and provided feedback to the panellists. During the last step (the second round of standard setting), the panellists were instructed to review their judgements and revise them if they felt it appropriate.

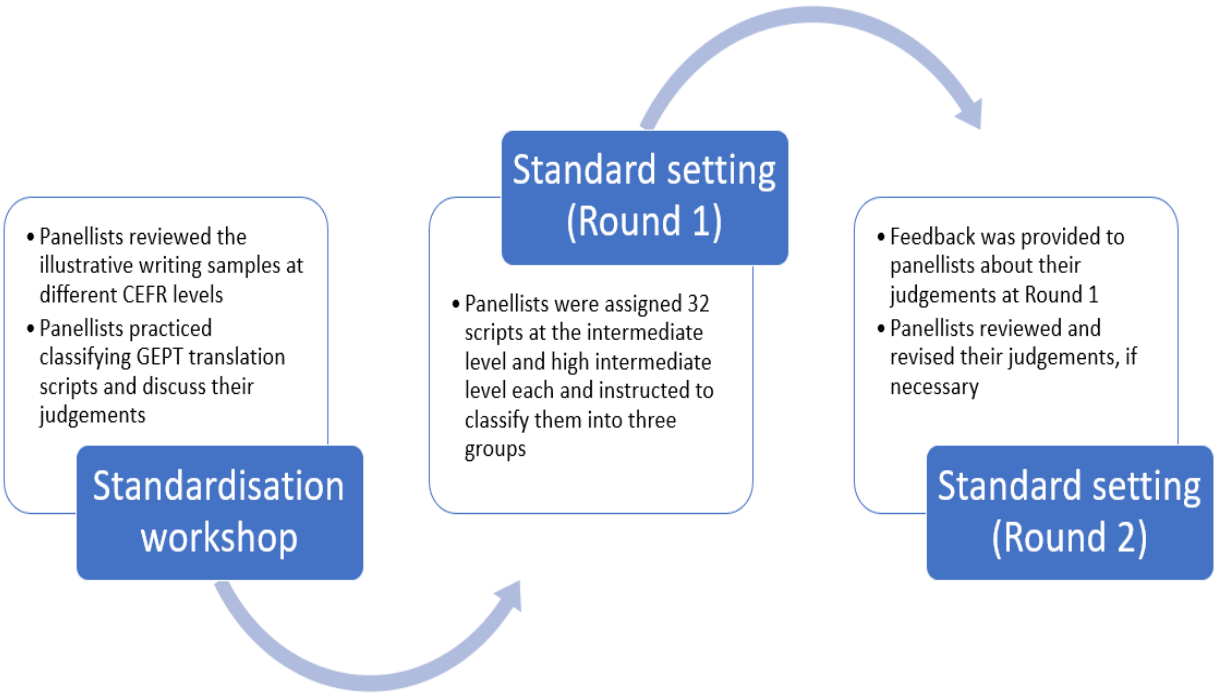

Figure 3. The standard setting procedures in this project

2.2.4 Validation

Standard setting is a complex, judgement-based process (Cizek & Bunch, 2007). To ensure the standard setting results are robust, reliable and useful, systematic measures need to be implemented and followed to support the linking claims. In this project, both procedural and internal validity evidence was collected. Each type of evidence is detailed below.

Procedural validity

Two measures were taken to ensure the procedural validity of this standard setting study. First, transparent and rigorous criteria were applied in selecting the panellists for this project. As explained previously, the two groups of participants (i.e. test 'insiders' and 'outsiders') had experience in English language teaching and assessment; several of them also had experience of participating in previous linking projects. In addition, all test 'insiders' were familiar with the GEPT and the CEFR. Though most test 'outsiders' had originally indicated a lack of familiarity with the CEFR, the familiarisation workshop activities helped them gain an in-depth understanding of the CEFR scales and descriptors. When organising group discussions in both the familiarisation and standardisation workshops, we made conscious efforts to mix test 'insiders' and 'outsiders' so that they had ample opportunities to exchange views on the CEFR as well as their linking processes.

Second, as indicated previously, all panellists were asked to fill in an online questionnaire after attending the familiarisation and standardisation workshops. Both questionnaires were designed to elicit the panellists' perceptions of the effectiveness of the workshops. For both questionnaires, the panellists were asked to indicate their level of agreement with statements about the workshops on a five-point Likert scale (5 - Strongly Agree, 4 - Agree, 3 - Neutral, 2 - Disagree, 1 - Strongly Disagree). Specifically, the questionnaires were intended to shed light on a) whether the purpose of each workshop and the instructions for each activity were clearly explained; b) the usefulness of the activities in each workshop; and c) the overall effectiveness of each workshop. The items in the questionnaires are presented in Tables 8 and 9.

Internal validity

To support the internal validity of this study, we first of all calculated the intraclass correlation coefficients of the combined panel as well as of the two groups of test 'insiders' and 'outsiders' to investigate the reliability of the panellists' judgements. Second, we performed many-facets Rasch analysis of the panellists' rating data to compare the average severity levels of the two groups of panellists; we also compared the linking results derived from the judgements of the two groups of panellists. Finally, a think-aloud verbal protocol study was conducted to explore the processes used by the two groups of panellists when linking the GEPT translation scripts to the CEFR scales and descriptors.

2.2.5 Verbal protocol analysis

Verbal protocol analysis (VPA) is a research methodology which has been utilised extensively in language research. In the field of language testing, this method is typically employed to investigate the cognitive processes that test takers engage with when responding to test tasks or items, thus generating important evidence about test validity (Green, 1998). Verbal reports may also provide insights into the reasoning processes that underlie the cognition, response, and decision making of learners and test takers (Cohen, 2000). Despite the multiple advantages of using VPA in language research, systematic procedures need to be developed to ensure that the data are reliable and valid (Douglas & Hegelheimer, 2007). Some guidelines have been provided for researchers intending to employ VPA in their research, such as providing concrete prompts that elicit detailed information, explaining the purposes of the introspective or retrospective accounts to the participants, and including sufficient context when writing up the study to help readers understand the conclusions that are drawn from the data (Greene & Higgins, 1994).

In this project, four panellists, comprising two test 'insiders' (Panellists A & F, see Table 2) and two test 'outsiders' (Panellists J & L, see Table 2), were invited to participate in the think-aloud verbal protocol study after they finished the standardisation workshop and were about to start the first round of standard setting. The two test 'insiders' had quite extensive experience in EFL teaching and testing and were familiar with both the GEPT and the CEFR; the two test 'outsiders', on the other hand, were experienced C-E translators with experience of working as markers for

C-E translation tests. Eight GEPT translation scripts (four each at both intermediate and high-intermediate levels) at different score levels from two test forms were used in the VPA study.

All VPA sessions were conducted via Zoom. At the beginning of each session, the researcher began with a short training session for the panellists, following the training procedure that we drafted (see Appendix V) by referring to the guidelines in conducting VPA research (e.g., Greene & Higgins, 1994) as well as previous studies using the VPA method (e.g., Douglas & Hegelheimer, 2007). This procedure had already been piloted on one participant and revised based on her feedback. Following the training, the participating panellists were given two GEPT scripts and had to report how they linked them to the CEFR levels using the written translation scale and the Written Assessment Criteria Grid. The researcher and the panellists then discussed any issues that emerged from this process. After that, the panellists repeated the same procedures and finished the remaining six scripts. Each VPA session lasted for about one and a half hours. The sessions were audio-recorded, and the panellists' reports were subsequently transcribed verbatim for analysis.

2.3 Data analysis

The panellists' ratings were analysed using many-facets Rasch analysis. Misfitting scripts were identified and removed from further analysis. The Rasch analysis was implemented using the FACETS 3.80.0 software (Linacre, 2017). To address the first research question, we derived cut scores from both the contrasting group method and the borderline group method. For the contrasting group method, we calculated the means and standard deviations (SDs) of the GEPT scores of the scripts that were classified into the groups of 'at' or 'below' the target CEFR levels. For the borderline group method, we calculated the means and SDs of the scripts that were placed into the borderline group. The data was analysed in SPSS (IBM, 2012). To answer the second research question, cut scores derived from the two methods were compared. In addition, we also included group membership as a dummy facet in the many-facets Rasch analysis to investigate whether the two groups of panellists differed significantly in their severity (Linacre, 2017).

To investigate the third research question, we followed guidelines for analysing VPA data (Kasper, 1998). We coded the verbal reports in NVivo 12 (QSR, 2012). Due to the exploratory nature of this study, we employed an open-coding method, which means the transcripts were reviewed line-by-line. After coding for larger themes, we performed detailed coding to identify sub-themes and relationships between codes. We will elaborate on the themes and subthemes that emerged from the data in the Findings section. One researcher coded all the data and developed the coding scheme; another one coded half the data using the same coding scheme. Inter-coder reliability was verified by means of kappa statistics (k = 0.85). Discrepancies in the coding process were resolved through discussion.

3. Findings

3.1 Many-facets Rasch analysis

In this Rasch analysis, we included four facets: script, test form, panellist, and group membership, among which test form and group membership were specified as dummy facets. In many-facets Rasch analysis, dummy facets are not used for measuring main effects, and all the elements of a dummy facet are anchored at 0 (Linacre, 2017). Test form was included as a dummy facet because we intended to ascertain whether the variable 'test form' impacted the judgement results; panellist group membership was also included as a dummy facet because we were interested in examining whether the two groups of panellists differed in their severity levels when classifying the scripts into the three categories (i.e. below or at a target CEFR level and borderline group). In what follows, we report the Rasch analysis results for the intermediate and high-intermediate levels respectively.

3.1.1 Intermediate level

Figure 4 presents the variable maps for the analysis results at the intermediate level. The variable map provides a wealth of useful information about this standard setting study. As shown in Figure 4, eight scripts are at the very top of the 'script' column, representing the most competent test takers in the sample. Indeed, the participating panellists reached the consensus that all these eight scripts should be classified into the group of 'at B1'; similarly, there are eight scripts at the bottom of the 'script' column, representing the least competent test takers in the sample. The participating panellists unanimously put these eight scripts into the group of 'Below B1'.

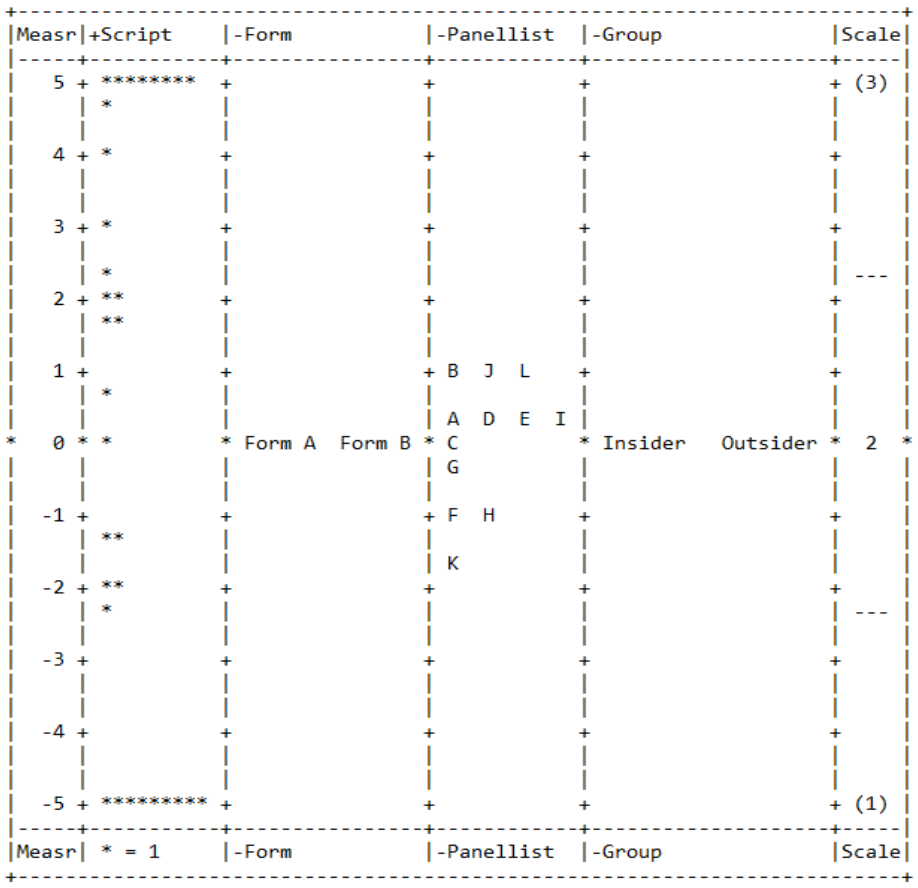

Figure 4. Wright map (intermediate level)

Next, we examined the fit statistics of each script, as indicated by the mean square values and their associated standardised Z statistics, with a view to identifying the misfitting scripts, which would be excluded from subsequent analysis. We adopted the range of infit and outfit mean square values from 0.6 to 1.4 as representing sound measurement qualities (e.g., Bond & Fox, 2015; McNamara, Knoch, & Fan, 2019). Having said that, we believe that overfit does not constitute a serious concern for this study as it might just represent high agreement of the panellists concerning the classification of a script. As a result, three scripts (15, 21, and 28) were identified as misfitting, as they all had mean square values beyond the recommended range. These three scripts were therefore removed from the subsequent analysis from which the cut scores were derived.

The measurement report on the test form (see Appendix VI) indicates that the panellists' judgements were similar across the two parallel test forms (χ^2 = 0.00, df = 1, p = 1.00). In other words, the use of scripts from two parallel test forms did not affect the classification results. The Chi-square test also indicates that there was no significant difference between the two groups of panellists (i.e., test 'insiders' and 'outsiders') in terms of their average severity when classifying the scripts into the three levels (χ^2 = 0.00, df = 1, p = 1.00). The measurement report on the panellists

(see Appendix VI) indicates that one panellist ('L') did not fit the Rasch model satisfactorily (Infit MnSq = 2.72; Outfit MnSq = 2.56). This was probably caused by some of this panellist's judgements differing quite significantly from those of the rest of the group. At Round 2 of Standard Setting (See Figure 3), we sent our feedback to this panellist. In our feedback, we highlighted the judgements on six scripts which clearly deviated from those of other panellists. After carefully deliberating each judgement that we highlighted, he indicated that he would leave these judgments intact because each of them was accurate from his perspective and represented his values. As such, it was decided not to remove this panellist's judgements from our subsequent data analysis. This panellist also participated in the think-aloud study. We will draw on the analysis of the verbal protocol data to explore his rating behaviors (see Section 3.3 on panellists' linking processes).

3.1.2 High-intermediate level

Figures 5 presents the variable maps for the analysis results at the high-intermediate level. As shown in this figure, four scripts are at the very top of the 'script' column, representing the most competent test takers in the sample; another four scripts are at the bottom of the 'script' column, representing the least competent test takers in the sample.

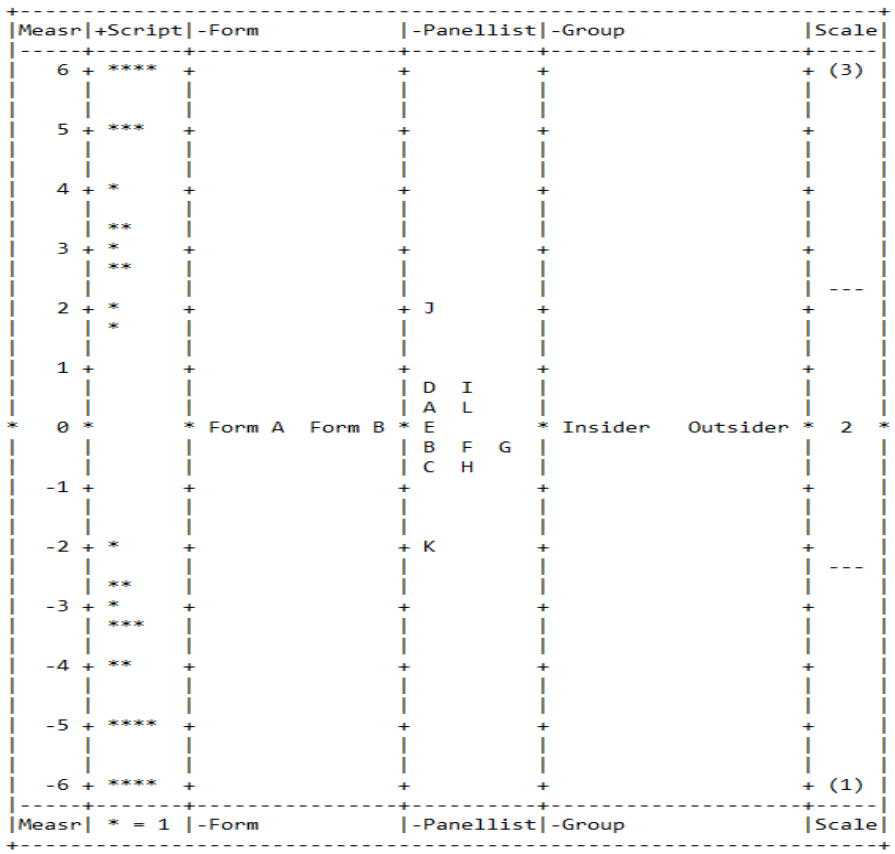

Figure 5. Wright map (high-intermediate level)

As with the results at the intermediate level, we examined the fit statistics of each script that was included in the analysis. Two scripts (20 and 23) were found to be misfitting and were therefore eliminated from subsequent analysis. We also examined the measurement reports for each facet in this analysis. Analysis results for the test form (see Appendix VI) indicate that the use of scripts from two parallel test forms did not affect the classification results ($\chi^2 = 0.00$, df = 1, $p = 1.00$), while these for group membership (see Appendix VI) indicate that the two groups of panellists on average applied similar level of severity when making their judgements ($\chi^2 = 0.00$, df = 1, $p = 1.00$). The measurement report on the panellists (See Appendix VI) indicates that three of them underfit the Rasch model (i.e. Panellists J, K, and L). At Round 2 of Standard Setting (see Figure 3), we sent them our feedback highlighting the judgements which clearly deviated from the rest of the group. They all carefully deliberated their judgements and made revisions where they felt appropriate. As such, their judgements were included in our subsequent data analysis. It is worth mentioning that all three underfitting panellists are test 'outsiders'. We will further explore the judgements of the two groups of participants, that is, test 'insiders' and 'outsiders' through the think-aloud study reported in Section 3.3.

3.2 Relating the GEPT writing subtest (Part 1) to the CEFR

In this section, we report the linking results at the two levels in the study. At each level, we report the results of the combined panel. In addition, we compare the linking results from the two groups of panellists, that is, test 'insiders' and 'outsiders'.

3.2.1 Intermediate level

To link the panellists' judgements with the script scores, we mapped the classification result of each script with its original score awarded by certified GEPT raters. Since we intended to compare the linking results between test 'insiders' and 'outsiders', we also included the group membership of the panellists as a variable when organising the data. Then we computed the means and SDs of the original script scores that were classified into each of the three groups: 'below B1', 'borderline', and 'at B1'. We also computed the results as a function of the panellists' group membership. Table 4 presents the results from both the combined panel and the two groups of panellists. A comparison of the linking results from test 'insiders' and 'outsiders' is also illustrated in Figure 6. As indicated in Table 4, there is a clear progression of test scores from 'below B1' to 'borderline' to 'at B1'. For all three groups, there is at least a difference of one score between adjacent levels (i.e., from 'below B1' to 'borderline' and from 'borderline' to 'at B1'). In addition, the results indicate that the differences in the judgements by the two groups of panellists are only minimal.

Table 5 presents the target CEFR level, the pass mark, and the GEPT scores based on the two standard setting methods, that is, the contrasting group method and the borderline group method. As indicted in Table 5, for the combined panel, the results for both methods are identical (both 3.4); for the two groups of panellists, there are only very slight differences (e.g., for test 'insiders',

the result based on the borderline group method is 3.5 as compared with 3.4 based on the contrasting group method). We will discuss these results further in the Summary and Recommendations section.

Table 4. Linking results (intermediate level)

Group	Below B1		Borderline		At B1	
	Mean	SD	Mean	SD	Mean	SD
Combined panel	2.27	0.51	3.40	0.54	4.59	0.63
Test insiders	2.26	0.47	3.45	0.54	4.58	0.63
Test outsiders	2.30	0.61	3.31	0.54	4.60	0.65

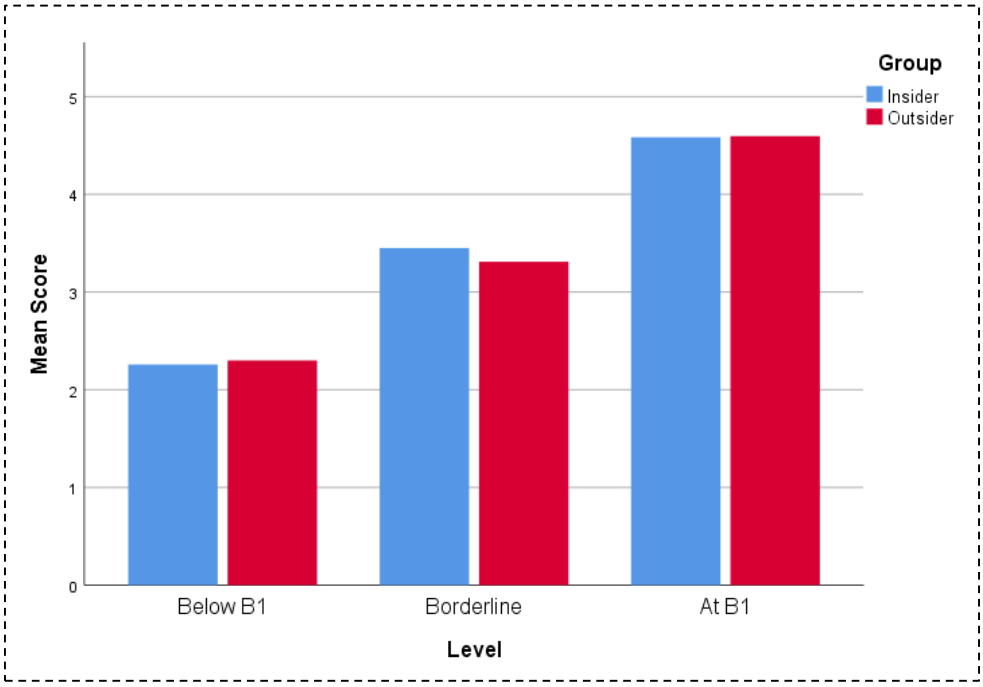

Figure 6. A comparison of the linking results between the two groups (intermediate level)

Table 5. Linking results based on the two standard setting methods (intermediate level)

Group	Target CEFR level	Pass mark	Contrasting group	Borderline Group
Combined	B1	4	3.4	3.4
Test insiders	B1	4	3.4	3.5
Test outsiders	B1	4	3.5	3.3

3.2.2 High-intermediate level

The same procedures were repeated for the data at the high-intermediate level. Table 6 below presents the results from both the combined panel and the two groups of panellists. As indicated in this table, there is a clear progression of test scores from 'below B2' to 'borderline' to 'at B2'. For all three groups, there is a difference of one or nearly one score between adjacent levels (i.e., from 'below B2' to 'borderline' and from 'borderline' to 'at B2'). The results also indicate that the score means of test 'outsiders' across the three levels are slightly higher than those of test 'insiders', suggesting that test 'outsiders' were marginally more severe in their judgements than their 'insider' counterparts, particularly when it came to the two levels of 'borderline' and 'at B2' (see also Figure 7).

Table 6. Linking results (high-intermediate level)

Group	Below B2		Borderline		At B2	
	Mean	SD	Mean	SD	Mean	SD
Combined panel	2.51	0.58	3.49	0.78	4.72	0.45
Test insiders	2.50	0.58	3.44	0.56	4.70	0.46
Test outsiders	2.53	0.58	3.53	0.96	4.77	0.43

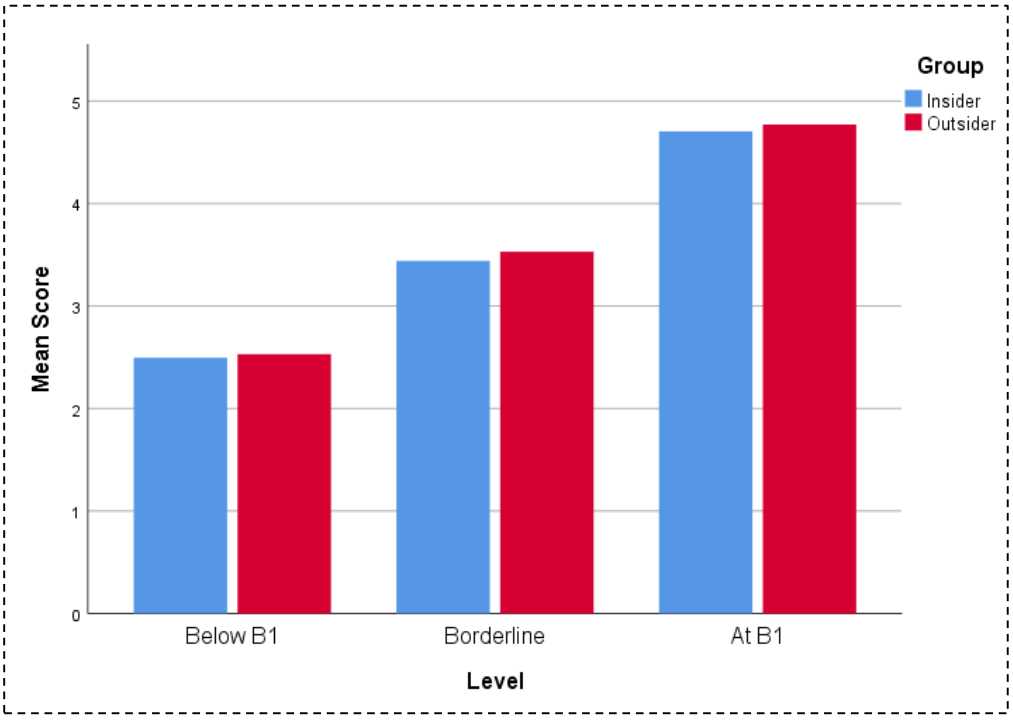

Figure 7. A comparison of the linking results between the two groups (high-intermediate level)

Table 7 presents the target CEFR level, the pass mark, and the GEPT scores based on the two standard setting methods. As indicated in this table, for the combined panel, there is a slight difference between the two methods (3.6 vs. 3.5); when looking at the two groups individually, slightly larger differences can be observed (e.g., for test 'insiders', the result based on the borderline group method is 3.4 as compared with 3.6 based on the contrasting group method). We will further discuss relevant findings in the Summary and Recommendations section.

Table 7. Linking results based on the two standard setting methods (high-intermediate level)

Group	Target CEFR level	Pass mark	Contrasting group	Borderline
Combined	B2	4	3.6	3.5
Test insiders	B2	4	3.6	3.4
Test outsiders	B2	4	3.7	3.5

3.3 Panellists' linking processes

The coding scheme in Table 8 shows the themes and subthemes that emerged from the verbal protocol data. The participating panellists primarily focused on four aspects when reporting their linking processes: the language and the translation quality of the translations scripts that were being scrutinised, the procedures that they followed in comparing the scripts to the CEFR scales and descriptors, and the challenges or difficulties that they experienced in the linking process. As indicated in Table 8, each main theme consists of several subthemes. For example, when the panellists commented on the language quality of a translation script, they mainly focused on the aspects of accuracy, coherence, range, and style; when examining the translation quality of a script, their comments mostly focused on content alignment, that is, whether the translation faithfully reflected the original Chinese text. They also mentioned the influence from Chinese, the test takers' native language, on translation quality.

As explained previously, two scales were adopted in this linking study: the CEFR translation scale and the Written Assessment Criteria Grid (see Appendix I). When examining the translation scripts against the CEFR levels, the panellists reported that they were comparing the scripts with the descriptors in both the translation scale and the writing scale; they also reported making some initial judgements and recalibrating their decisions as they took a closer look at each script. Finally, the panellists identified the challenges that they experienced in the linking process, including making comments on the translation scale in general and describing how their background might have affected their judgements. They also mentioned the difficulties of evaluating and classifying borderline scripts and the tensions that arose from using two scales which seemed to tap into two different constructs, that is, translation and writing ability. In what follows, we will briefly report the findings under each of the main themes.

Table 8. The coding scheme

Themes	Subthemes
1) Language quality	a) Accuracy b) Coherence c) Range d) Style
2) Translation quality	a) Content or meaning b) Wrong translation c) Influence from Chinese
3) Linking process	a) Comparing the script with the scale b) Making initial judgements
4) Challenges	a) The translation scale b) Influence of the panellist's background c) Difficulty with borderline cases d) Tensions between translation and writing

3.3.1 Language quality

Language quality was clearly the aspect on which the panellists commented most frequently. This is consistent with our expectations because language quality is the very focus of the CEFR writing scale; it also features quite prominently in the CEFR translation scale, especially at B1 level (see Appendix I). When commenting on language quality, the panellists mainly focused on the aspects of a) accuracy, b) coherence, c) range, and d) style. In the CEFR writing scale, accuracy, coherence, and range are all essential aspects of writing performance. As such, it is not surprising that the panellists focused on these aspects when assessing the language quality of the translation scripts.

The comments on accuracy, in most cases, surrounded the errors that test takers made in their translation, and whether these errors impeded readers' understanding of the scripts. The comments also touched upon the accuracy of grammar, vocabulary, and mechanics specifically. Excerpt 1 below illustrates the comment made by Panellist A, a test 'insider', on the errors in a script that he was examining:

Excerpt 1

> I think this test taker made so many mistakes that it was not possible for him or her to express the meanings in the original text clearly. According to the translation scale, 'although linguistic errors may occur, the translation remains comprehensible.' In my view, the mistakes that this person made have to some extent impeded my comprehension of the translation. (Panellist A, test 'insider')

In another excerpt below, Panellist L, a test 'outsider', commented on the inaccurate choice of words in the translation script.

Excerpt 2

> In the last sentence, the original text says that 'Emily treated him beef steaks.' However, this test taker had no idea about how to translate '*niupai*' (steak) in English; he simply used the word 'meat', which is not accurate; in addition, he also apparently didn't know how to translate '*qingke*' (treat) which was translated into 'pay the bill for the meat'. Maybe it is comprehensible, but it is not accurate. (Panellist L, test 'outsider')

In addition to accuracy, coherence was another aspect that featured prominently in panellists' comments, mainly in relation to the use of connecting words or devices.

3.3.2 Translation quality

Translation quality is the second main theme that emerged from the panellists' verbal reports. When evaluating translation quality, the panellists tended to focus on whether the meaning of the original Chinese text was faithfully translated into English, and whether the translation included all the details in the original text. Both observations are well aligned with the descriptors in the translation scale. For example, at the B1 level, one of the descriptors requires that a language user at this level 'can produce approximate translations from (Language A) into (Language B).' The following comment made by Panellist J serves to illustrate this point.

Excerpt 3

> In the original Chinese text, there are several important points. First, there was a traffic accident, which caused the delay of the train. Because of the delay, Emily waited for her cousin for half an hour. Since neither of them had dinner, she treated her cousin beef steaks. Can you see the development and logical progression of the story? I think this test taker's translation successfully depicted what had happened, and the meaning of the original text was well conveyed. (Panellist J, test 'outsider')

Conversely, in Excerpt 4 below, Panellist L made scathing comments on a script that he was examining, arguing that the translation almost completely failed to reflect the meaning of the original text.

Excerpt 4

> From the translation point of view, I would say that this script is a train wreck. Don't you think so? He almost failed completely to translate the original text into English. Even though I was trying to make wild guesses, I could hardly make sense of what he or she meant. For example, in the first sentence, the translation goes 'Taiwan had many different race'. This is a radical departure from the original text. Clearly, this is a very unsuccessful transfer of information! (Panellist L, test 'outsider')

Specifically, the panellists highlighted some wrong or inaccurate translations in the scripts. Interestingly, most of comments concerning incorrect translations were made by test 'outsiders', that is, those who had experience working as professional translators or interpreters. Also, as we will demonstrate shortly, the two groups of panellists exhibited quite different tolerance of errors. Another interesting observation is that the test 'outsiders' tended to consider the potential impact of a wrong or inaccurate translation on readers while this was not clear in the comments by test 'insiders' who were more experienced in language testing and teaching. The comment below by Panellist J illustrates this observation.

Excerpt 5

> When this test taker translated '*Taiwan de lupao shengxing*' into 'Taiwan had many different race' in the first sentence, the first thing that appeared in a reader's mind wouldn't be a sports event. Rather, you would be thinking about ethnicity, you know, groups of people… And for someone who has already got the wrong idea in the first sentence, you would be even thinking about the Black Lives Matter movement, because this test taker was talking about racial issues. (Panellist J, test 'outsider')

3.3.3 Linking processes

When explaining how they identified the CEFR levels for each script, it is not surprising that the panellists kept comparing the scripts against the two CEFR scales that were employed in this study, that is, the translation scale and the writing scale. In most cases, they used the translation scale before using the writing scale. During the mapping process, they usually read the script carefully to determine whether it was up to the quality as specified in the descriptors in the scale. The excerpt below exemplifies this process: Panellist F was examining a script at the high intermediate level against the descriptors in the translations scale.

Excerpt 6

> If you take a look at what is required in the B2 descriptors, one of them states that the translation needs to 'closely follow the sentence and paragraph structure of the original text'. This script fails to fulfil this requirement; in addition, it also fails the requirement of 'conveying the main points of the source'. Therefore, I don't think it is up to the B2 level. This is quite clear in this case. In terms of language quality, it doesn't reach that level, either. (Panellist F, test 'insider')

Another observation is that the panellists usually referred back to the script on which they had based their initial judgement about whether the current text was below or at the target CEFR level or a borderline case. In some cases, they might recalibrate their judgement after taking a closer look at the script. It is very likely that language quality played a significant role when the panellists approached a script and made the initial judgement. In the excerpt below, Panellist J remarked on the poor language quality of a script which enabled her to reach the judgement that it was clearly below the target CEFR level.

Excerpt 7

> This test taker has been making lots of errors, well, small errors that probably don't harm the meaning. But they come out very frequently and very routinely… Well there are problems in almost every sentence… So in that sense, I think she is kind of a borderline case because at B1, it says 'occasionally makes errors that readers usually can interpret correctly on the basis of the context'. With the context which I think is the key as well as the task that I have already seen, I can probably understand the translation. But there are too many errors. The errors are so frequent. Maybe the errors do not impede comprehension. Therefore, I think it's a borderline case. (Panellist J, test 'outsider')

In some cases, the panellists made their initial judgements based on the translation quality which, in their view, was clearly below or above the target CEFR level. In the excerpt below, Panellist F noticed that quite a lot of details in the original text were missing in the translation script. In consequence, he made the decision that it was clearly below B2 level.

Excerpt 8

> Panellist: This one is clearly below B2.
>
> Researcher: Why? Could you explain the reason?
>
> Panellist: Look, so many details are missing in the English translation. For example, he didn't translate '*jin nian lai*' (in recent years) in the first sentence; he translated '*tongguo meiti dafu baodao*' (through massive media coverage) simply into 'they use online'. So many details are missing! (Panellist F)

3.3.4 Challenges

Several challenges arising from the linking process were mentioned in the panellists' comments. First, the panellists commented negatively on the translation scale, holding the view that it was difficult for raters to use to evaluate a translation performance or in a linking study. The comments on the scale mainly touched on the following three aspects: a) the scale focused more on the complexity of the original text than the quality of translation; b) the construct of translation was not clear at all based on the descriptors, thus making it extremely difficult to use in practice; and c) the translation scale was very broad and lacked many important details. In the excerpt below, Panellist L was criticising the translation scale on the grounds that the construct of translation ability was far from clear.

Excerpt 9

> I think a serious problem with the translation scale is that the construct of translation is not clearly spelt out in the descriptors at all. According to the scale, good translation means the correspondence between the original text and the translation; however, the notion of correspondence seems very vague to me. The descriptor seems to suggest that the translation should not be overinfluenced by the original text, but that is it! (Panellist L, test 'outsider')

The second challenge that the panellists mentioned in their comments was that their professional training and background seemed to have some effect on their evaluation of a translation script. This was manifested in the observation that the panellists who had been trained in translation and worked as professional translators were much harsher when it came to some mistakes which they believed significantly affected the meaning and readers' understanding of the original text. When asked why this had happened, Panellist J explained:

Excerpt 10

> Translation is different from writing. We tend to believe that if a person can translate, that means he or she has reached a quite high level of proficiency in both languages. Therefore, I would assume that those from the translation background would be much harsher than those who are from EFL teaching background when it comes to the mistakes. This is how I understand this question. (Panellist J, test 'outsider')

Compared with test 'insiders' who did not have a background in translation, test 'outsiders' had far less tolerance of some mistakes that test takers made in their translation; they argued that these mistakes caused serious problems or even failures in conveying messages in the original text. For example, in one script, a test taker translated '*pengyou*' (friend) into 'boyfriend'. This was considered as a fatal mistake, as explained by Panellist L below:

Excerpt 11

> I know that it might just be a minor linguistic error; however, the consequence is that the communication broke down. Therefore, I don't think such mistakes can be tolerated when we evaluate translation. The translation is a failure though you may say that the mistake is not a serious one from the linguistic point of view. (Panellist J, test 'outsider')

The panellists also mentioned that whereas it was much easier to classify some scripts into the categories of below or at the target CEFR level, it was much more difficult to place scripts into the borderline category. Another challenge that the panellists mentioned repeatedly was the tension between the constructs reflected in the translation and writing scales. The panellists reported that it was difficult to make a judgement when a test taker was stronger in one construct but weaker in the other. Panellist F explained this conundrum below:

Excerpt 12

> If I look at the translation scale, I think this script is up to B1 because he has basically translated the original text into English. Having said that, he made quite a few mistakes. If you look at the writing scale, you would hesitate to classify it into the category of at B1; it is more like a borderline case. Should I move it down to the borderline category? (Panellist F, test 'insider')

3.4 Validity evidence

As mentioned previously, two types of validity evidence were collected to support the linking claims. In what follows, we present the evidence related to procedural and internal validity.

3.4.1 Procedural validity

Following each workshop, questionnaires were distributed to the participating panellists to solicit their perceptions of the effectiveness of the workshop. Tables 8 and 9 show the survey results from the familiarisation and standardisation workshops respectively. As indicated in Table 8, nine panellists participated in the familiarisation workshop survey. All participants expressed either strong agreement or agreement with the seven statements that were included in the questionnaire, suggesting the overall effectiveness of the workshop. For example, most participants strongly agreed that the activities in the workshop were useful and helped them understand the descriptors in the CEFR; most participants also strongly agreed that the workshop was well-conducted overall.

Table 9. Results of the questionnaire survey (familiarization workshop, n = 9)

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1) I have a clear understanding of the purpose of this workshop.	7	2	0	0	0
2) I have a clear understanding of the purpose of this study.	7	2	0	0	0
3) I have a good overview of the CEFR.	6	3	0	0	0
4) I have a good overview of the Chinese-English translation task in the GEPT writing test.	6	3	0	0	0
5) The activities help me understand the descriptors in the CEFR scales.	7	2	0	0	0
6) The activities in the workshop are useful.	6	3	0	0	0
7) Overall, I feel that workshop is well-conducted.	7	2	0	0	0

Similar findings can be derived from the survey results of the benchmarking workshop. As indicated in Table 9, most participants either strongly agreed or agreed with the nine statements in the questionnaire. There was one participant who was unsure about the group discussion activity in the workshop. The responses to the open-ended question at the end of the questionnaire revealed that one or two participants were not used to group discussions in the breakout rooms in Zoom. Except for one participant, the others all strongly agreed that the workshop was well-conducted. This finding was also corroborated by the participants' responses

to the open-ended question, supporting the overall satisfactory effectiveness of the benchmarking workshop. To sum up, the results from the two surveys lend support to the procedural validity of this linking study.

Table 10. Results of the questionnaire survey (benchmarking workshop, n = 9)

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1) I understand the purpose of this workshop.	6	3	0	0	0
2) I understand what I was asked to do for the activities in the workshop.	8	1	0	0	0
3) The illustrative examples help me understand the CEFR levels.	5	3	1	0	0
4) The activities in the workshop are useful.	7	1	1	0	0
5) The discussions are helpful.	4	4	1	0	0
6) I feel familiar with the GEPT writing tasks.	9	0	0	0	0
7) I feel familiar with the CEFR writing descriptors.	6	3	0	0	0
8) I feel familiar with the CEFR translation descriptors.	5	4	0	0	0
9) Overall, I feel that workshop is well-conducted.	8	0	1	0	0

3.4.2 Internal validity

To interrogate the internal validity of this linking study, three types of evidence were collected. First, intraclass correlations were computed to examine the reliability of the panellists' judgements. We computed the intraclass correlation coefficients of the judgement results of both the combined panel and the two groups of test 'insiders' and 'outsiders'. Second, the average severity measures of the two groups of panellists were compared. The group-level severity measures were generated by the many-facets Rasch analysis of the panellists' rating data, wherein their group membership was specified as a dummy facet. In addition, we compared the linking results of the two groups of panellists based on both the contrasting group method and the borderline group method. Finally, the linking processes of the two groups of panellists were examined to shed light on the internal validity of this linking study.

The intraclass correlations, together with their 95% confidence intervals (CI), are shown in Table 10. As indicated in this table, the panellists as a combined group exhibited very satisfactory levels of consistency in their judgements at both the intermediate (Cronbach's alpha = 0.989, CI = 0.983-

0.994) and high-intermediate level (Cronbach's alpha = 0.985, CI = 0.975-0.992). Each of the two groups of panellists also demonstrated very high levels of consistency in their judgements. The results indicate that the panellists were highly consistent when making classification decisions about the translations scripts.

Table 11. Intraclass correlations

	Combined panel (95% CI)	Test 'insiders' (95% CI)	Test 'outsiders' (95% CI)
Intermediate	0.989 (0.983-0.994)	0.987 (0.979-0.993)	0.959 (0.927-0.979)
High-intermediate	0.985 (0.975-0.992)	0.985 (0.976-0.992)	0.922 (0.857-0.960)

As demonstrated in Section 3.1, many-facets Rasch analysis results indicate that the two groups of panellists had very similar average severity measures (see Figures 4 and 5), an observation which was confirmed by the not significant Chi-square test. The linking results derived from the judgements of test 'insiders' and 'outsiders' are very similar (see Section 3.2). These findings suggest that the two groups of panellists as a whole applied similar severity levels when determining whether a script was below or at the target CEFR level or a borderline case, thus lending further support to the internal validity of this linking study.

Coding of the panellists' think-aloud verbal protocols revealed that they frequently compared the translation scripts against the descriptors in both the translation scale and the writing scale before making their judgements (see Section 3.3). In addition, both language and translation quality featured prominently in their comments on the translation scripts (see Table 8). Though the panellists' background seemed to come into play in their interpretations of the CEFR scales and descriptors and their evaluation of the translation scripts, they applied similar severity levels in making their judgements at the group level, as evidenced by, for example, the results generated by Rasch analysis. Overall, the analysis of their linking processes supports the internal validity of this linking study.

4. Summary and recommendations

This study aimed to link the C-E translation task, that is, Part 1 of the GEPT writing subtest at the intermediate and high-intermediate levels to the CEFR. Two groups of panellists were invited to participate in this study: test 'insiders' who were based in Taipei (n = 8) and test 'outsiders' who were based in Melbourne (n = 4). In addition to linking the GEPT translation tasks to the CEFR, we also explored the processes through which the panellists' linked the GEPT translation scripts to the CEFR levels. In what follows, we briefly summarise the major findings of this study and provide recommendations to the LTTC, the GEPT provider, based on these results. We also

provide some recommendations to those who are planning to link translation tasks in language tests to the CEFR.

In this linking study, we followed the four stages recommended in the CEFR linking manual: familiarisation, specification, standardisation, and validation. Though most of the panellists indicated that they were familiar with the CEFR, all of them participated in the familiarisation workshop, which helped them gain an in-depth understanding of the CEFR scales and descriptors. During the specification stage, we analysed the translation tasks in the study by filling out a number of forms as recommended in the CEFR linking manual. As a result, an initial estimate of the alignment between the GEPT and the CEFR levels was derived, that is, the intermediate GEPT level at B1 and the high-intermediate GEPT level at B2. These initial estimates were further scrutinised in this study.

Next, two examinee-centred standard setting methods were adopted to align the GEPT translation tasks to the CEFR levels: the contrasting group method and the borderline group method. We employed these two methods to cross-validate the linking results. Two rounds of standard setting were implemented. The panellists were instructed to classify translation scripts at the intermediate level into one of the three categories of 'below B1', 'borderline', and 'at B1', and those at the high-intermediate level into 'below B2', 'borderline', and 'at B2'. The linking results in Table 11 indicates that the translation tasks at the two levels in the study are generally well aligned with the target CEFR levels; however, it is recommended that for the intermediate level, the cut score be set slightly lower, possibly by one score point; for the high-intermediate level, the current cut score could be maintained. To facilitate the comparison of the linking results between Knoch and Frost (2016) focusing on Part 2 of the GEPT writing subtest (i.e. guided writing) and this project, we include the linking results from both studies in Table 12.

Table 12. Linking results

Level	CEFR level	Pass mark	Contrasting group method	Borderline group method
Intermediate	B1	4	3.4 (3.3)	3.4 (3.1)
High-intermediate	B2	4	3.6 (3.4)	3.5 (3.3)

Notes. The values in the brackets represent the linking results from Knoch and Frost (2016).

As indicated in Table 12, Knoch and Frost (2016) arrived at similar linking results concerning the guided writing task at the intermediate level. In this study, the cut scores resulting from the panellists' judgements are slightly higher in comparison. The cut score based on the borderline method is 3.4, as compared to 3.1 in Knoch and Frost (2016), and that based on the contrasting group method is 3.4, as compared to 3.3 in Knoch and Frost (2016). Based on the linking results, we suggest that the cut score could be set at 3, as opposed to the current cut score of 4 for the intermediate level. Similar to the intermediate level, the linking results at the high-intermediate level in this study are slightly higher than those in Knoch and Frost (2016) based on both linking

methods. For example, the cut score based the contrasting method in this study is 3.6 as compared with 3.3 in Knoch and Frost (2016). The linking results derived from this study at the high-intermediate level suggest that the current cut score of 4 is generally appropriate. Having said that, if half point scores could be used in score reports in the future, we would suggest 3.5 as the cut score for both levels.

Unlike previous linking studies such as Knoch and Frost (2016), this study also explored the panellists' linking processes using think-aloud data. The results indicate that the panellists frequently referred to the descriptors in the two scales that were employed in this study, that is, the written translation scale and the Written Assessment Criteria Grid. When approaching a translation script, they focused on both the language and translation quality. Furthermore, the findings shed light on the challenges that they experienced when linking the translation scripts to the CEFR levels. For example, they felt that the translation scale was too broad and lacking in important details, thus making it difficult for raters to use. Another important observation was that the construct of translation was not clearly spelt out in the descriptors in the translation scale. As such, the panellists raised doubts over the overall usefulness of the translation scale. Finally, though the two groups of panellists applied similar levels of severity when making their judgements, as demonstrated by quantitative analysis results, their background seemed to play a quite significant role in their evaluation process. For example, those with a background in translation were found to be harsher and less tolerant of some mistakes that test takers made in their translation as compared with those with a background in language test development and EFL teaching.

Given that the CEFR companion volume was published only recently (Council of Europe, 2018), we anticipate more research endeavours in the future to link translation tasks to the CEFR levels. In view of our experience with this linking study, we would like to provide a few recommendations to those intending to link translation tasks in language tests to the CEFR. First, mediation in the CEFR is a much broader concept than translation; in the CEFR descriptive scheme on mediation, written translation is one of the activities under the category of 'mediating a text' (see Figure 1). As such, it is not surprising that there are only a few descriptors in the translation scale. Linking researchers will find it difficult to solely rely on the translation scale in a linking study. Depending on the constructs that are assessed in the translation task, linking researchers are likely to use the translation scale in combination with other scales in the CEFR, such as the writing scales, as demonstrated in this study. However, using different scales leads to tensions between the two seemingly different constructs, as this study revealed, though admittedly, the two constructs share some common aspects. Linking researchers need to make informed assessment of the relative importance of each construct and this should be made transparent to all participating panellists.

Second, no translation examples are currently available at different CEFR levels, thus making it difficult to train panellists on the salient features of the translation performance at each CEFR

level. It is advisable that linking researchers work together with experts who are knowledgeable about both the CEFR and translation to select sample translation performances at the relevant CEFR levels and for different languages. These performances can then be used for training panellists participating in the linking study. Next, it is necessary to involve panellists with background in language teaching and testing and those who have experience in translation. As demonstrated in this study, panellists' backgrounds may affect their orientations and linking processes. In the training process, these two groups of panellists could be mixed in a way that facilitates their understanding of each other's orientations through focusing on their evaluation of a translation script. Finally, we believe that using the two linking methods in this study, that is, the contrasting group method and the borderline method, helps to enhance the validity and rigour of the linking results.

We acknowledge a few limitations with this study. First, the sample size is relatively small, particularly considering that the participating panellists were divided into two groups of test 'insiders' and 'outsiders'. Next, as revealed by the think-aloud part of this study, the translation scale in the CEFR companion volume is broad with only a small number of descriptors, making it difficult for panellists to use in a linking study. In addition, translation samples at different levels of the CEFR are unavailable, hence exacerbating the challenges of training panellists in the linking process. Finally, the use of two scales (i.e. the translation and writing scale), as suggested in this study, raises concerns over tensions of the constructs reflected in the two scales. Except for the sample size, the other limitations were caused by constraints beyond our control. Future linking researchers should anticipate these constraints and take actions accordingly when planning a study to link the translation task in a language test to the CEFR.

5. References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

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

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

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

Cohen, A. D. (2000). Exploring strategies in test taking: Fine-tuning verbal reports from respondents. *Learner-directed assessment in ESL, 127*, 150.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Strasbourg, France: Author.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, and assessment*. Strasbourg, France: Author.

Council of Europe. (2018). *Common European Framework of Reference for languages: Learning, teaching, assessment (Companion volume with new descriptors)*. Strasbourg, France: Author.

De Jong, J., & Zheng, Y. (2016) Linking to the CEFR: validation using a priori and a posteriori evidence. In Banerjee, J. & Tsagari, D. (eds.), *Contemporary second language assessment* (pp. 83-100). London: Bloomsbury Academic.

Douglas, D., & Hegelheimer, V. (2007). *Strategies and use of knowledge in performing new TOEFL listening tasks* (Unpublished research report). Iowa State University.

Eckes, T. (2012). Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling, 54*(3), 257.

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

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

Green, A., & Inoue, C. (2017). *GEPT speaking–CEFR benchmarking (LTTC–GEPT Research Report No. RG-09)*. Taipei: The Language Training and Testing Center.

Greene, S., & Higgins, L. (1994). Once upon a time: The use of retrospective accounts in building theory in composition. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 115-140). Thousand Oaks: Sage.

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

IBM. (2012). *IBM SPSS statistics version 21*. Boston, MA.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kasper, G. (1998). Analysing verbal protocols. *TESOL quarterly, 32*(2), 358-362.

Kecker, G., & Eckes, T. (2010). Putting the Manual to the test: The TestDaF–CEFR linking project. *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual*, 50-79. Cambridge: Cambridge University Press.

Kenyon, D., & Romhild, A. (2014). Standard setting in language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-14). Hoboken, NJ: John Wiley & Sons.

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

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

Linacre, J. M. (2017). *Facets computer program for many-facet Rasch measurement* (version 3.80.0). Beaverton, Oregon: Winsteps.com. Retrieved from www.winsteps.com

Martyniuk, W. (2010). Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual. Cambridge: Cambridge University Press.

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

Milanovic, M., & Weir, C. J. (2010). Series editors' note. *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*, viii-xx.

North, B., & Piccardo, E. (2016). Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR): A Council of Europe project. *Language Teaching, 49*(3), 455-459.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels.* New Jersey: Educational Testing Service.

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

Roever, C., & Pan, Y.-C. (2008). Test review: GEPT: General English proficiency test. *Language Testing, 25*(3), 403-408.

Shih, C.-M. (2010). The washback of the General English Proficiency Test on university policies: A Taiwan case study. *Language Assessment Quarterly, 7*(3), 234-254.

Wu, J. R. W. (2012). GEPT and English language teaching and testing in Taiwan. *Language Assessment Quarterly, 9*(1), 11-25.

Wu, J. R. W., & Wu, R. Y. F. (2010). Relating the GEPT reading comprehension tests to the CEFR. In W. Martyniuk (Ed.), *Aligning Tests with the CEFR* (Vol. 33, pp. 204-224). Cambridge: Cambridge University Press.

Appendix I: The translation scale and the Written Assessment Criteria Grid

TRANSLATING A WRITTEN TEXT IN WRITING	
C2	Can translate into (Language B) technical material outside his/her field of specialisation written in (Language A), provided subject matter accuracy is checked by a specialist in the field concerned.
C1	Can translate into (Language B) abstract texts on social, academic and professional subjects in his/her field written in (Language A), successfully conveying evaluative aspects and arguments, including many of the implications associated with them, though some expression may be over-influenced by the original.
B2	Can produce clearly organised translations from (Language A) into (Language B) that reflect normal language usage but may be over-influenced by the order, paragraphing, punctuation and particular formulations of the original.
	Can produce translations into (Language B, which closely follow the sentence and paragraph structure of the original text in (Language A), conveying the main points of the source text accurately, though the translation may read awkwardly.
B1	Can produce approximate translations from (Language A) into (Language B) of straightforward, factual texts that are written in uncomplicated, standard language, closely following the structure of the original; although linguistic errors may occur, the translation remains comprehensible.
	Can produce approximate translations from (Language A) into (Language B) of information contained in short, factual texts written in uncomplicated, standard language; despite errors, the translation remains comprehensible.
A2	Can use simple language to provide an approximate translation from (Language A) into (Language B) of very short texts on familiar and everyday themes that contain the highest frequency vocabulary; despite errors, the translation remains comprehensible.
A1	Can, with the help of a dictionary, translate simple words and phrases from (Language A) into (Language B), but may not always select the appropriate meaning.
Pre-A1	*No descriptors available*

	Overall	Range	Coherence	Accuracy
C2	Can write clear, *highly accurate and* smoothly flowing complex texts in an appropriate and effective *personal* style *conveying finer shades of meaning. Can use* a logical structure which helps the reader to find significant points.	Shows great flexibility in *formulating* ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Can create coherent and cohesive texts making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Maintains consistent *and highly accurate* grammatical control of *even the most complex language forms. Errors are rare and concern rarely used forms.*
C1	Can write clear, well-structured *and mostly accurate* texts of complex subjects. *Can underline* the relevant salient issues, *expand and support* points of view at some length with subsidiary points, reasons and relevant examples, and *round* off with an appropriate conclusion.	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. *The flexibility in style and tone is somewhat limited.*	Can produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices.	Consistently maintains a high degree of grammatical accuracy; *occasional errors in grammar, collocations and idioms.*
B2	Can write clear, detailed *official and semi-official* texts on a variety of subjects related to his field of interest, synthesising and evaluating information and arguments from a number of sources. *Can make a distinction between formal and informal language with occasional less appropriate expressions.*	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, using some complex sentence forms to do so. *Language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic.*	Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some "jumpiness" in a longer text.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.
B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. *The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading.*	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Can link a series of shorter discrete elements into a connected, linear text.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more *common* situations. *Occasionally makes errors that the reader usually can interpret correctly on the basis of the context.*
A2	Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". *Longer texts may contain expressions and show coherence problems which makes the text hard to understand.*	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information mainly in everyday situations.	Can link groups of words with simple connectors like "and", "but" and "because".	Uses simple structures correctly, but still systematically makes basic mistakes. *Errors may sometimes cause misunderstandings.*
A1	Can write simple isolated phrases and sentences. *Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand.*	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Can link words or groups of words with very basic linear connectors like "and" and "then".	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. *Errors may cause misunderstandings.*

Appendix II: Scoring rubrics for the translation task at the intermediate and high-intermediate level

GEPT Intermediate Level Writing Test

Rating Scale

Part I: Chinese-English Translation (40%)

Band Score	Description
5	The translation fully conveys the content of the original text and is well organized and coherent. Errors are rarely found in vocabulary, grammar, spelling, punctuation, or capitalization.
4 (pass)	The translation adequately conveys the content of the original text and is generally organized and coherent. Errors are sometimes found in vocabulary, grammar, spelling, punctuation, or capitalization, but these do not impede satisfactory completion of the task.
3	The translation does not adequately convey the content of the original text and lacks sufficient organization and coherence. Errors in vocabulary and grammar, spelling, punctuation, or capitalization impede satisfactory completion of the task.
2	The translation only partially conveys the content of the original text and displays little control of sentence structure. Most sentences are erroneous, incomprehensible, and incoherent. The text demonstrates a limited range of vocabulary. Errors are frequently found in grammar, spelling, punctuation, or capitalization.
1	The translation neither conveys the content of the original text nor shows understanding of sentence structure. The text is incomprehensible, lacks coherence, and demonstrates a very limited range of vocabulary. Serious errors are frequently found in grammar, spelling, punctuation, or capitalization.
0	No answer/ Non-ratable*

*Non-ratable: e.g. the translation is too short (< 25 words) to be marked.

GEPT High-Intermediate Level Writing Test

Rating Scale

Part I: Chinese-English Translation (40%)

Band Score	Description
5	**Demonstrates full competence in translation** The translation fully conveys the content of the original text, is well organized and coherent, and demonstrates effective control of sentence structure. Errors are rarely found in vocabulary, grammar, spelling, punctuation, or capitalization.
4 (pass)	**Demonstrates fair competence in translation** The translation adequately conveys the content of the original text, is generally organized and coherent, and demonstrates sufficient control of sentence structure. Errors are sometimes found in vocabulary, grammar, spelling, punctuation, or capitalization, but these do not impede satisfactory completion of the task.
3	**Demonstrates limited competence in translation** The translation does not adequately convey the content of the original text, lacks sufficient organization and coherence, and demonstrates limited control of sentence structure. Errors in vocabulary and grammar, spelling, punctuation, or capitalization impede satisfactory completion of the task.
2	**Demonstrates little competence in translation** The translation only partially conveys the content of the original text, lacks organization and coherence, and demonstrates little control of sentence structure. Most sentences are incomprehensible. Serious errors are found in grammar, spelling, punctuation, or capitalization.
1	**Lacks competence in translation** The translation does not convey the content of the original text, lacks control of sentence structure, and is incomprehensible. Serious errors are frequently found in grammar, spelling, punctuation, or capitalization.
0	**No answer/ Non-ratable**

Appendix III: Specification forms

Form A1: General Examination Description

GENERAL EXAMINATION DESCRIPTION	
1. General Information	
Name of examination	The General English Proficiency Test (GEPT) – writing section Levels: Intermediate/ High-Intermediate
Language tested	English
Examining institution	The Language Training & Testing Centre (LTTC)
Versions analysed ()	Intermediate (), High-Intermediate ()
Type of examination	☑ International ☑ National ☐ Regional ☐ Institutional
Purpose	Measuring general English writing proficiency level of Taiwanese learners (source: https://www.lttc.ntu.edu.tw/e_lttc/E_GEPT.htm)
Target population	☐ Lower Sec ☑ Upper Sec ☑ Uni/College Students ☑ Adult
No. of test takers per year	Over 7 million (as at October 2016) since its launch in 2000 (source: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/recognition.htm)
2. What is the overall aim?	
Testing general writing skills with the aim of promoting learning, improving the general writing proficiency of Taiwanese learners and providing institutions/schools with a reference for evaluating the English proficiency level of their job applicants, employees, or students (source: https://www.lttc.ntu.edu.tw/e_lttc/E_GEPT.htm)	

3. What are the more specific objectives? If available describe the needs of the intended users on which this examination is based.

- Evaluation of the general English writing proficiency of English learners in junior high schools, high schools, universities and private enterprises in Taiwan.
- Evaluation of the general English writing proficiency of high school applicants in Taiwan and for university applicants in universities in Taiwan as well as institutions around the world (including in Asia, Europe, and the USA), with the purpose of school and university entry, student placement and as a criterion for university graduation.
- Evaluation of the general English writing proficiency of job applicants and employees in the general and government employment sectors, and for career advancement.

4. What is/are principal domain(s)?	☑ Public ☑ Personal ☑ Occupational ☑ Educational		
5. Which communicat-ive activities are tested?		**Name of Subtest(s)**	**Duration**
	☐ **1** Listening comprehension	_____	_____
	☐ **2** Reading comprehension	_____	_____
	☐ **3** Spoken interaction		
	☐ **4** Written interaction		
	☐ **5** Spoken production	_____	_____
	☑ **6** Written production	Intermediate	**16 min (approx.)**
		High-Intermediate	**20 min (approx.)**
	☐ **7** Integrated skills	_____	_____
	☐ **8** Spoken mediation of text	_____	_____
	☑ **9** Written mediation of text	Intermediate	**16 min (approx.)**
	☐ **10** Language usage	High-Intermediate	**20 min (approx.)**
	☐ **11** Other: (specify): _____	_____	_____
		_____	_____

6. What is the weighting of the different subtests in the global result?	Intermediate (IW): Part one (40%) Part two (not included in this benchmarking study) (60%) High-Intermediate (HW): Part one (40%) Part two (not included in this benchmarking study) (60%)
7. Describe briefly the structure of each subtest	Intermediate (IW): 2 parts, 2 items 1. Chinese-English translation: translate a Chinese paragraph of approximately 90-100 characters into English; topics are relevant to the life and cultural experiences of Taiwanese learners of English; the level of difficulty appropriate for average senior high school graduates. 2. Guided writing: essay on familiar topic or personal experience (120 words, approx.) (not included in this benchmarking study) High-Intermediate (HW): 2 parts, 2 items 1. Chinese-English translation: translate a Chinese paragraph of approximately 120-130 characters into English; topics are relevant to the life and cultural experiences of Taiwanese learners of English as well as contemporary issues in Taiwanese contexts; the level of difficulty appropriate for average university non-English major graduates. 2. Guided writing: essay on topic related to daily life/current events (150-180 words) (not included in this benchmarking study)

8. What type(s) of responses are required?	☐ Multiple-choice ☐ True/False ☐ Matching ☐ Sentence writing ☐ Sentence completion ☐ Gapped text / cloze, selected response ☐ Open gapped text / cloze ☐ Short answer to open question(s) ☑ Extended answer (text / monologue) ☐ Interaction with examiner ☐ Interaction with peers ☐ Other	Subtests used in (Write numbers above) ☐ IW1, IW2, HW1, HW2 ☐
9. What information is published for candidates and teachers?	☑ Overall aim ☑ Principal domain(s) ☑ Test subtests ☑ Test tasks ☑ Sample test papers ☐ Video of format of oral	☑ Sample answer papers ☑ Marking schemes ☐ Grading schemes ☑ Standardised performance samples showing pass level ☑ Sample certificate
10. Where is this accessible?	☑ On the website ☑ From bookshops ☑ In test centres ☑ On request from the institution ☐ Other	_____
11. What is reported?	☑ Global grade ☐ Grade per subtest	☐ Global grade plus graphic profile ☐ Profile per subtest

Form A2: Test Development

Test development	**Short description and/or references**
1. What organisation decided that the examination was required?	☑ Own organisation/school ☐ A cultural institute ☑ Ministry of Education ☑ Ministry of Justice ☑ Other: specify: the Central Personnel Administration of the Executive Yuan acknowledges the GEPT as a criterion for the promotion of civil servants; a number of private enterprises and government agencies; many high schools and universities. For more information, go to https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/recognition.htm
2. If an external organisation is involved, what influence do they have on design and development?	☑ Determine the overall aims ☐ Determine level of language proficiency ☐ Determine examination domain or content ☐ Determine exam format and type of test tasks ☐ Other: specify:
3. If no external organisation was involved, what other factors determined design and development of examination?	☑ A needs analysis ☑ Internal description of examination aims ☑ Internal description of language level ☑ A syllabus or curriculum ☑ Profile of candidates
4. In producing test tasks are specific features of candidates taken into account?	☐ Linguistic background (L1) ☑ Language learning background ☑ Age ☑ Educational level

	☐ Socio-economic background ☑ Social-cultural factors ☐ Ethnic background ☑ Gender
5. Who writes the items or develops the test tasks?	Native and non-native item writers, specialized in English teaching and testing fields and familiar with local English learning environments
6. Have test writers guidance to ensure quality?	☑ Training ☑ Guidelines ☑ Checklists ☑ Examples of valid, reliable, appropriate tasks: ☐ Calibrated to CEFR level description ☐ Calibrated to other level description: _____
7. Is training for test writers provided?	☑ Yes ☐ No
8. Are test tasks discussed before use?	☑ Yes ☐ No
9. If yes, by whom?	☑ Individual colleagues ☑ Internal group discussion ☑ External examination committee ☐ Internal stakeholders ☐ External stakeholders
10. Are test tasks pretested?	☑ Yes

	☐ No
11. If yes, how?	Items are selected and compiled into pre-test papers which conform to the test specifications. Pilot papers are administered to a representative sample of target population.
12. If no, why not?	
13. Is the reliability of the test estimated?	☑ Yes ☐ No
14. If yes, how?	☑ Data collection and psychometric procedures ☐ Other: specify: _____
15. Are different aspects of validity estimated?	☑ Face validity ☑ Content validity ☐ Concurrent validity ☐ Predictive validity ☑ Construct validity
16. If yes, describe how.	Questionnaires are distributed to stakeholders to check if the tests meet the current standards of public expectations in regard to the format and content of the test. To ensure that the test content is a fair reflection of the construct, specifications of each task is used as the basis for selection of the elements to be included in the test form. Criterion-related validitiy (https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG01.pdf) and context and cognitive validity (https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG03.pdf) are also investigated.

Form A3: Marking

Marking: Intermediate	**Complete a copy of this form for each subtest.** **Short description and/or reference**
1. How are the test tasks marked?	For receptive test tasks: ☐ Optical mark reader ☐ Clerical marking For productive or integrated test tasks: ☑ Trained examiners ☐ Teachers
2. Where are the test tasks marked?	☑ Centrally ☐ Locally: ☐ By local teams ☐ By individual examiners
3. What criteria are used to select markers?	Raters have to be in-service English teachers.
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers/raters ☑ Moderating sessions to standardise judgments ☑ Using standardised examples of test tasks: ☐ Calibrated to CEFR ☑ Calibrated to another level description ☐ Not calibrated to CEFR or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	☑ One holistic score for each task ☐ Marks for different aspects for each task ☐ Rating scale for overall performance in test ☐ Rating Grid for aspects of test performance

	☑ Rating scale for each task ☐ Rating Grid for aspects of each task ☑ Rating scale bands are defined, but not to CEFR ☐ Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	☐ Single rater ☐ Two simultaneous raters ☑ Double marking of scripts / recordings ☐ Other: specify:_____
7. If double rated, what procedures are used when differences between raters occur?	☑ Use of third rater and that score holds– in the case that the discrepancy between the two marks is significant ☐ Use of third marker and two closest marks used ☑ Average of two marks ☐ Two markers discuss and reach agreement ☐ Other: specify:_____
8. Is inter-rater agreement calculated?	☑ Yes ☐ No
9. Is intra-rater agreement calculated?	☑ Yes ☐ No

Marking: High-Intermediate	**Complete a copy of this form for each subtest.** **Short description and/or reference**
1. How are the test tasks marked?	For receptive test tasks: ☐ Optical mark reader ☐ Clerical marking For productive or integrated test tasks:

	☑ Trained examiners ☐ Teachers
2. Where are the test tasks marked?	☑ Centrally ☐ Locally: ☐ By local teams ☐ By individual examiners
3. What criteria are used to select markers?	Raters have to be in-service English teachers.
4. How is accuracy of marking promoted?	☑ Regular checks by co-ordinator ☑ Training of markers/raters ☑ Moderating sessions to standardise judgments ☑ Using standardised examples of test tasks: ☐ Calibrated to CEFR ☑ Calibrated to another level description ☐ Not calibrated to CEFR or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	☑ One holistic score for each task ☐ Marks for different aspects for each task ☐ Rating scale for overall performance in test ☐ Rating Grid for aspects of test performance ☑ Rating scale for each task ☐ Rating Grid for aspects of each task ☑ Rating scale bands are defined, but not to CEFR ☐ Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	☐ Single rater ☐ Two simultaneous raters ☑ Double marking of scripts / recordings ☐ Other: specify:_____

7. If double rated, what procedures are used when differences between raters occur?	☑ Use of third rater and that score holds– in the case that the discrepancy between the two marks is significant ☐ Use of third marker and two closest marks used ☑ Average of two marks ☐ Two markers discuss and reach agreement ☐ Other: specify:_____
8. Is inter-rater agreement calculated?	☑ Yes ☐ No
9. Is intra-rater agreement calculated?	☑ Yes ☐ No

Form A4: Grading

Grading: Intermediate	**Complete a copy of this form for each Subtest.** **Short description and/or reference**
1. Are pass marks and/or grades given?	☑ Pass marks ☑ Grades
2. Describe the procedures used to establish pass marks and/or grades and cut scores	The content of LTTC GEPT Intermediate Level Writing Test is guided by National Curriculum Objectives of Senior High Schools in Taiwan. During the development stage of the test, the research committee reached a consensus on the descriptions of the minimum acceptable level of writing proficiency for local senior high school graduates; hence, pilot-version of five-band rating scales (Band 0 to 5) for writing proficiency were developed, and the pass mark was set at Band 4. In the piloting stage, the pilot-versions of writing tests were administered to a representative sample of the target population. The writing performances were collected, and benchmark performances for each band score were selected based on the expert judgement of the raters in conjunction with the descriptions in the rating scale.
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	Not applicable
4. If grades are given, how are the grade boundaries decided?	Not applicable

5. How is consistency in these standards maintained?	After each test administration, range-finding sessions are held to select benchmark performances for each band score from the responses of the candidates to the live test, based on both the rating scale and the benchmark samples of the previous test session, for use in the training of new raters training and in tune-up sessions. Before the marking sessions, all raters are requested to attend the tune-up and trial-marking sessions.

Grading: High-Intermediate	**Complete a copy of this form for each Subtest.** **Short description and/or reference**
1. Are pass marks and/or grades given?	☑ Pass marks ☑ Grades
2. Describe the procedures used to establish pass marks and/or grades and cut scores	The content of LTTC GEPT High-Intermediate Level Writing Test is based on results of textbook analyses, and surveys of stakeholders' needs, collected from college teachers, target candidates and target test users using questionnaires and interviews. During the development stage, the research committee reached a consensus on the descriptions of the minimum acceptable level of writing proficiency for local university graduates; hence, pilot-versions of five-band rating scales (Band 0 to 5) for writing proficiency were developed, and the pass mark was set to be Band 4. In the piloting stage, the pilot-version tests were administered to the sample candidates; a representative sample of the target population was selected from college students; and candidates

	who took and passed LTTC GEPT Intermediate Level operational tests; and the general public. The writing performances were collected, and benchmark performances for each band score were selected based on the descriptions of the rating scale for future use in training, tune-up and trial-marking sessions for raters.
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	Not applicable
4. If grades are given, how are the grade boundaries decided?	Not applicable
5. How is consistency in these standards maintained?	After each test administration, range-finding sessions are held to select benchmark performances for each band score from the responses of the candidates to the live test, based on both the rating scale and the benchmark samples of the previous test session, for use in the training of new raters training and in tune-up sessions. Before the marking sessions, all raters are requested to attend the tune-up and trial-marking sessions.

Form A5: Reporting Results

Results	**Short description and/or reference**
1. What results are reported to candidates?	☑ Global grade or pass/fail ☐ Grade or pass/fail per subtest ☐ Global grade plus profile across subtests ☐ Profile of aspects of performance per subtest
2. In what form are results reported?	☐ Raw scores ☐ Undefined grades (e.g. "C") ☑ Level on a defined scale ☐ Diagnostic profiles ☐ Scaled scores
3. On what document are results reported?	☐ Letter or email ☑ Report card ☑ Certificate / Diploma ☑ Online score report: It cannot be used as a substitute for the official score report. Individual candidates can check their own scores on the LTTC and GEPT websites during the period of ten days (or seven workdays) after the official score reports have been mailed.
4. Is information provided to help candidates to interpret results? Give details.	Level descriptors and the pass mark are provided to the general public. Institutions or organizations which register their students or employees as a group receive a score roster, a report with descriptive analyses, and grouped analyses based on information which the candidates provided on their backgrounds in the registration forms.
5. Do candidates have the right to see the corrected and scored examination papers?	No
6. Do candidates have the right to ask for remarking?	Yes

Form A6: Data Analysis

Data analysis	**Short description and/or reference**
1. Is feedback gathered on the examinations?	☑ Yes ☐ No
2. If yes, by whom?	☑ Internal experts (colleagues) ☑ External experts ☐ Local examination institutes ☑ Test administrators ☑ Teachers ☑ Candidates ☑ Parents
3. Is the feedback incorporated in revised versions of the examinations?	☑ Yes ☐ No
4. Is data collected to do analysis on the tests?	☑ On all tests ☐ On a sample of test takers: How large?: _____. How often?:_____ ☐ No
5. If yes, indicate how data are collected?	☑ During pretesting ☑ During live examinations ☑ After live examinations
6. For which features is analysis on the data gathered carried out?	☑ Difficulty ☑ Reliability ☑ Validity ☑ Descriptive analysis
7. State which analytic methods have been used (e.g. in terms of psychometric procedures).	The CTT (including descriptive and correlation) and IRT analysis.

8. Are performances of candidates from different groups analysed? If so, describe how.	Performances of candidates are grouped and analysed based on information that the candidates provided on their backgrounds in the registration forms.
9. Describe the procedures to protect the confidentiality of data.	All information collected is protected under Personal Information Protection Act. Also, a hierarchy of user levels regulates access to the computers designated for scoring.
10. Are relevant measurement concepts explained for test users? If so, describe how.	Yes. The relevant information, such as difference between norm-referenced and criterion-referenced testing and marking procedures, is published on the LTTC website and candidate handbooks.

Form A7: Rationale for Decisions

Rationale for decisions (and revisions)	**Short description and/or reference**
Give the rationale for the decisions that have been made in relation to the examination or the test tasks in question.	Candidates who pass the GEPT Writing Test are certified to have the abilities described in the GEPT level descriptors. GEPT level descriptors are available online: High-Intermediate: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/hi_intermediate.htm Intermediate: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/intermediate.htm
Is there a review cycle for the examination? (How often? Who by? Procedures for revising decisions)	Yes. The reviewing procedures are conducted from time to time to monitor reliability and validity so that adjustments to the tests can be made when necessary.

Form A8: Initial Estimation of Overall Examination Level

Initial Estimation of Overall CEFR Level		
☐ **A1**	**IW: B1**	☐ **C1**
☐	☐	☐
☐ **A2**	**HW: B2**	☐ **C2**
☐	☐	☐

Short rationale, reference to documentation

Information on the GEPT-CEFR alignment is provided by the LTTC on their website:

https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/alignment.htm

GEPT-CEFR alignment studies have been undertaken for reading and listening:

Reading:

Wu, J. R. W. & Wu, R. Y. F. (2010). Relating the GEPT reading comprehension tests to the CEFR. *Studies in Language Testing*, 33, 204-224.

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*. Cambridge, England: Cambridge University Press.

Listening:

Brunfaut, T. & Harding, L. (2014). Linking the GEPT listening test to the Common European Framework of Reference. LTTC-GEPT Research Report RG-05.

Writing (subtest 2):

Knoch, U. & Frost, K. (2014). Linking the GEPT writing sub-test to the Common European Framework of Reference (CEFR). LTTC-GEPT Research Report RG-08.

Speaking:

Green, A., Inoue, C., & Nakatsuhara, F. (2017). *GEPT speaking–CEFR benchmarking (LTTC–GEPT Research Report No. RG-09)*. Taipei: The Language Training and Testing Center.

Appendix IV: Sample judgement forms

Intermediate Test Judgments (Contrasting Group)

Intermediate test (targets CEFR level B1)

Decisions (place an 'X' in the appropriate box):

Script	Below B1	B1
IA1		
IA2		
IA3		
IA4		
IA5		
IA6		
IA7		
IA8		
IA9		
IA10		
IA11		
IA12		
IA13		
IA14		
IA15		
IA16		
IB1		
IB2		
IB3		
IB4		
IB5		
IB6		
IB7		
IB8		
IB9		
IB10		
IB11		
IB12		
IB13		
IB14		
IB15		
IB16		

Intermediate Test Judgments (Borderline)

Intermediate test (targets CEFR level B1)

Decisions (place an 'X' in the appropriate box):

Script	Borderline
IA1	
IA2	
IA3	
IA4	
IA5	
IA6	
IA7	
IA8	
IA9	
IA10	
IA11	
IA12	
IA13	
IA14	
IA15	
IA16	
IB1	
IB2	
IB3	
IB4	
IB5	
IB6	
IB7	
IB8	
IB9	
IB10	
IB11	
IB12	
IB13	
IB14	
IB15	
IB16	

Appendix V: Think aloud training procedures

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

- How you evaluate the translation sample in focus
- How you use the CEFR translation and writing scales
- How you relate the features of the translation sample to the CEFR scales.

In other words, we are interested in understanding your reasoning for deciding whether a GEPT translation sample is below B1, borderline, or B1 or above (for intermediate level), and below B2, borderline, or B2 or above (for high-intermediate level). This may seem strange at first. With a little practice, I am sure that you will feel more comfortable talking out loud about what you are thinking.

Before you start working on each translation sample, please say aloud its ID number so that we can associate what you report about the linking process to a particular sample afterwards. Take IA3. Please say aloud 'This sample is IA3' before you start reporting your linking process.

It is important to talk as much as possible. As mentioned previously, we are interested in understanding your reasoning when linking the GEPT translation samples to the CEFR levels. We can only know what you are thinking about if you talk out aloud as you work on a sample. If you are silent for some time, I might say "keep talking" in order to remind you to talk. We understand that your thoughts might be in both Chinese and English. In the think aloud session, please feel free to speak in either Chinese or English, as you see appropriate.

To reiterate, we want to know not only *what* you are doing, but *why* you are doing it. As you think out aloud, I may ask you to explain what you are thinking further if it is not clear from what you are reporting. For example, as you go through a translation sample you say out aloud, "This is below B1." I will ask you, "Why do you think so?" After a little practice, you should understand what we are asking you to do. I will model for you an example of someone thinking out aloud as they read the translation sample and try to link it to the CEFR.

Below is an example:

Participant: 这份 I3。我先读一下 script。我觉得这一份感觉是 B1 or above。我觉得他写得挺好的。我先看一下 translation scale，B1 这个级别具体的内容，比如说 can produce approximate translations from Language A into Language B of information contained in short, factual texts…我觉得这里是在描述翻译的任务，就是翻译的题目，所以我会继续往下看。Closely following the structure of the original…这里是跟他具体翻译的结果是相关的。我看了一下他英文的翻译和中文的内容是否一一对应。这篇文章的内容都是对应的，而且很完整。因此 B1 这一条是 okay 的。这里谈的是语言，Although linguistic errors may occur, the translation remains comprehensible…我觉得这条标准是比较宽松的。也就是说它允许有一些语言错误，但是翻译是能够被读懂的。我

觉得这篇 script 没有什么大的语言错误，而且完全能够看懂。这篇翻译还有一些英文的思维，例如使用了从句之类，还有语序的转化等。我觉得信息量最大的是 closely follow the structure of the original，即使是有语言错误，翻译也是能够被读懂的。这些标准我个人感觉还是比较 broad。

然后我看语言质量，我要看一下 CEFR writing scale。我看到 B1。有几个层面，例如 overall, range, coherence 等。我先看一下 B1 这个等级的描述语。我觉得 overall 的描述语比较抽象。我觉得这篇文章读起来很顺畅。下面提到 but occasional unclear expressions and/or inconsistencies may cause a break-up in reading…我没有觉得这篇文章有什么 inconsistency，也没有感觉到阅读的时候会有 break up，因此这一条我觉得是符合的。接下来我看了一下 range。他的语言还是挺好的。例如他没有不停地使用一些很简答的词汇；他的词汇 range 还是比较好的，例如 travel around, scenery, pick up 等，都还是比较地道的。然后下面一个是 coherence。我觉得他这篇翻译上下文的衔接是比较好的，用了一些衔接词，例如 but, because of, since 等，逻辑还是比较清楚的。而且这篇文章的时态应用得也挺好的。Accuracy 就是关于用词和时态等方面了。Uses reasonably accurately a repertoire of frequently used “routines” and patterns associated with more common situations。因此从这些方面来看，这篇翻译感觉是达到了这个要求。

Appendix VI: Measurement reports

Script measurement report (intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Script
30	12	2.50	2.51	2.26	.60	.76	-.8	.72	-.9	1.51	.58	.39	1	1
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	2	2
13	12	1.08	1.07	-4.87	1.07	.69	-.1	.29	-.4	1.23	.62	.26	3	3
19	12	1.58	1.60	-1.88	.59	1.33	1.1	1.35	1.0	.43	-.17	.40	4	4
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	5	5
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	6	6
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	7	7
24	12	2.00	2.00	-.01	.64	.13	-2.7	.11	-2.7	1.64	.00	.36	8	8
34	12	2.83	2.86	4.00	.79	.83	-.1	.57	-.4	1.22	.47	.26	9	9
35	12	2.92	2.93	4.83	1.05	.82	.0	.43	-.1	1.15	.43	.19	10	10
26	12	2.17	2.15	.79	.63	1.17	.4	1.21	.5	.87	.48	.39	11	11
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	12	12
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	13	13
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	14	14
29	12	2.42	2.41	1.90	.60	1.60	1.7	1.72	1.9	-.10	.27	.40	15	15
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	16	16
28	12	2.33	2.32	1.54	.60	.78	-.5	.74	-.6	1.31	.46	.40	17	17
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	18	18
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	19	19
18	12	1.50	1.50	-2.23	.60	.90	-.2	.90	-.2	1.20	.39	.40	20	20
28	12	2.33	2.32	1.54	.60	1.43	1.1	1.48	1.2	.41	.43	.40	21	21
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	22	22
32	12	2.67	2.69	3.02	.64	.66	-1.2	.58	-1.0	1.68	.69	.35	23	23
29	12	2.42	2.41	1.90	.60	.57	-1.5	.54	-1.5	1.72	.76	.40	24	24
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	25	25
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	26	26
21	12	1.75	1.77	-1.17	.60	.84	-.3	.82	-.3	1.21	.19	.37	27	27
19	12	1.58	1.60	-1.88	.59	1.87	2.5	1.85	2.2	-.65	.02	.40	28	28
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	29	29
12	12	1.00	1.02	(-6.19	1.86)	Minimum					.00	.00	30	30
21	12	1.75	1.77	-1.17	.60	.95	.0	.92	.0	1.09	.05	.37	31	31
36	12	3.00	2.98	(6.12	1.85)	Maximum					.00	.00	32	32
24.7	12.0	2.06	2.06	.25	1.26	.96	-.1	.89	-.1		.18		Mean (Count: 32)	
9.6	.0	.80	.79	4.70	.60	.42	1.3	.49	1.3		.26		S.D. (Population)	
9.7	.0	.81	.80	4.78	.61	.43	1.3	.51	1.3		.26		S.D. (Sample)	

With extremes, Model, Populn: RMSE 1.40 Adj (True) S.D. 4.49 Separation 3.21 Strata 4.62 Reliability .91
With extremes, Model, Sample: RMSE 1.40 Adj (True) S.D. 4.57 Separation 3.27 Strata 4.69 Reliability .91
Without extremes, Model, Populn: RMSE .69 Adj (True) S.D. 2.40 Separation 3.47 Strata 4.95 Reliability .92
Without extremes, Model, Sample: RMSE .69 Adj (True) S.D. 2.49 Separation 3.59 Strata 5.12 Reliability .93
With extremes, Model, Fixed (all same) chi-square: 349.8 d.f.: 31 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 32.1 d.f.: 30 significance (probability): .36

Test form measurement report (intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Form
402	192	2.09	2.10	A .00	.24	.97	-.1	.80	-.7	1.06	.93	.92	1	Form A
388	192	2.02	2.10	A .00	.21	1.01	.1	.98	-.1	1.00	.90	.90	2	Form B
395.0	192.0	2.06	2.10	.00	.23	.99	.0	.89	-.4		.92		Mean (Count: 2)	
7.0	.0	.04	.00	.00	.02	.02	.1	.09	.3		.01		S.D. (Population)	
9.9	.0	.05	.00	.00	.02	.02	.2	.13	.4		.02		S.D. (Sample)	

Model, Populn: RMSE .23 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .23 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 1 significance (probability): 1.00

Panellist group measurement report (intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Group
528	256	2.06	2.10	A .00	.20	.80	-1.8	.72	-1.8	1.28	.93	.91	1 Insider
262	128	2.05	2.10	A .00	.28	1.39	2.1	1.24	.8	.55	.90	.91	2 Outsider
395.0	192.0	2.05	2.10	.00	.24	1.09	.1	.98	-.5		.92		Mean (Count: 2)
133.0	64.0	.01	.00	.00	.04	.29	2.0	.26	1.4		.02		S.D. (Population)
188.1	90.5	.01	.00	.00	.06	.41	2.8	.37	1.9		.02		S.D. (Sample)

Model, Populn: RMSE .24 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .24 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 1 significance (probability): 1.00

Panellist measurement report (intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Panellist
65	32	2.03	2.05	.27	.54	1.31	1.0	1.20	.5	.66	.90	.92	1 A
62	32	1.94	1.88	1.16	.56	.81	-.4	.81	-.1	1.20	.93	.91	2 B
66	32	2.06	2.10	-.03	.55	.59	-1.5	.50	-1.3	1.58	.95	.92	3 C
65	32	2.03	2.05	.27	.54	.46	-2.2	.40	-1.7	1.74	.96	.92	4 D
65	32	2.03	2.05	.27	.54	.57	-1.6	.49	-1.3	1.61	.95	.92	5 E
69	32	2.16	2.31	-.96	.58	.61	-1.1	.51	-.9	1.45	.94	.92	6 F
67	32	2.09	2.16	-.33	.55	1.28	.9	1.15	.4	.73	.91	.92	7 G
69	32	2.16	2.31	-.96	.58	.78	-.5	.67	-.4	1.29	.93	.92	8 H
65	32	2.03	2.05	.27	.54	1.05	.2	.92	.0	1.03	.92	.92	9 I
63	32	1.97	1.94	.86	.55	.84	-.4	.76	-.3	1.20	.95	.91	10 J
71	32	2.22	2.49	-1.66	.62	.84	-.2	.72	-.1	1.15	.93	.91	11 K
63	32	1.97	1.94	.86	.55	2.72	3.7	2.56	2.4	-1.25	.82	.91	12 L
65.8	32.0	2.06	2.11	.00	.56	.99	-.2	.89	-.3		.92		Mean (Count: 12)
2.6	.0	.08	.17	.81	.02	.58	1.5	.56	1.1		.04		S.D. (Population)
2.7	.0	.09	.18	.84	.02	.61	1.6	.58	1.1		.04		S.D. (Sample)

Model, Populn: RMSE .56 Adj (True) S.D. .58 Separation 1.04 Strata 1.72 Reliability .52
Model, Sample: RMSE .56 Adj (True) S.D. .63 Separation 1.12 Strata 1.83 Reliability .56
Model, Fixed (all same) chi-square: 23.1 d.f.: 11 significance (probability): .02
Model, Random (normal) chi-square: 8.2 d.f.: 10 significance (probability): .61

Script measurement report (high-intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Script
28	12	2.33	2.32	1.50	.59	1.04	.2	1.09	.3	.94	.10	.42	1	1
15	12	1.25	1.22	-3.43	.70	1.15	.4	2.85	2.3	.46	-.12	.36	2	2
13	12	1.08	1.06	-4.85	1.07	.62	-.2	.25	-.4	1.26	.63	.25	3	3
36	12	3.00	2.98	(6.20	1.87)	Maximum					.00	.00	4	4
13	12	1.08	1.06	-4.85	1.07	.62	-.2	.25	-.4	1.26	.63	.25	5	5
35	12	2.92	2.94	4.87	1.08	.60	-.2	.24	-.4	1.27	.66	.27	6	6
32	12	2.67	2.69	2.97	.65	1.20	.6	1.10	.3	.73	.16	.40	7	7
19	12	1.58	1.59	-1.86	.59	.85	-.4	.81	-.5	1.28	.42	.42	8	8
13	12	1.08	1.06	-4.85	1.07	.62	-.2	.25	-.4	1.26	.63	.25	9	9
36	12	3.00	2.98	(6.20	1.87)	Maximum					.00	.00	10	10
15	12	1.25	1.22	-3.43	.70	1.15	.5	.95	.1	.88	.21	.36	11	11
33	12	2.75	2.78	3.43	.71	.67	-.8	.53	-.8	1.44	.68	.38	12	12
14	12	1.17	1.14	-3.99	.81	.78	-.2	.70	-.1	1.19	.48	.32	13	13
36	12	3.00	2.98	(6.20	1.87)	Maximum					.00	.00	14	14
17	12	1.42	1.40	-2.58	.61	1.06	.3	1.04	.2	.89	.25	.40	15	15
33	12	2.75	2.78	3.43	.71	.67	-.8	.53	-.8	1.44	.68	.38	16	16
17	12	1.42	1.40	-2.58	.61	.93	-.1	.86	-.2	1.18	.40	.40	17	17
12	12	1.00	1.02	(-6.17	1.86)	Minimum					.00	.00	18	18
34	12	2.83	2.86	4.00	.82	1.33	.7	3.88	2.3	.47	-.28	.35	19	19
31	12	2.58	2.60	2.57	.61	1.43	1.4	1.88	2.0	.01	-.22	.41	20	20
14	12	1.17	1.14	-3.99	.81	1.12	.3	.86	.0	.95	.23	.32	21	21
12	12	1.00	1.02	(-6.17	1.86)	Minimum					.00	.00	22	22
31	12	2.58	2.60	2.57	.61	1.42	1.4	1.24	.7	.15	.51	.41	23	23
35	12	2.92	2.94	4.87	1.08	.60	-.2	.24	-.4	1.27	.66	.27	24	24
13	12	1.08	1.06	-4.85	1.07	1.23	.5	2.05	1.0	.76	-.15	.25	25	25
35	12	2.92	2.94	4.87	1.08	.60	-.2	.24	-.4	1.27	.66	.27	26	26
16	12	1.33	1.31	-2.98	.64	1.00	.0	1.88	1.7	.75	.16	.38	27	27
12	12	1.00	1.02	(-6.17	1.86)	Minimum					.00	.00	28	28
29	12	2.42	2.41	1.85	.59	.63	-1.3	.59	-1.3	1.67	.65	.42	29	29
15	12	1.25	1.22	-3.43	.70	.78	-.4	.66	-.4	1.30	.54	.36	30	30
36	12	3.00	2.98	(6.20	1.87)	Maximum					.00	.00	31	31
12	12	1.00	1.02	(-6.17	1.86)	Minimum					.00	.00	32	32
23.2	12.0	1.93	1.93	-.33	1.06	.92	.0	1.04	.2		.27		Mean (Count: 32)	
9.8	.0	.82	.83	4.43	.49	.28	.6	.88	1.0		.31		S.D. (Population)	
10.0	.0	.83	.84	4.50	.50	.29	.7	.90	1.0		.31		S.D. (Sample)	

With extremes, Model, Populn: RMSE 1.17 Adj (True) S.D. 4.27 Separation 3.65 Strata 5.21 Reliability .93
With extremes, Model, Sample: RMSE 1.17 Adj (True) S.D. 4.35 Separation 3.72 Strata 5.29 Reliability .93
Without extremes, Model, Populn: RMSE .82 Adj (True) S.D. 3.56 Separation 4.37 Strata 6.16 Reliability .95
Without extremes, Model, Sample: RMSE .82 Adj (True) S.D. 3.64 Separation 4.47 Strata 6.29 Reliability .95
With extremes, Model, Fixed (all same) chi-square: 556.9 d.f.: 31 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 30.7 d.f.: 30 significance (probability): .43

Test form measurement report (high-intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Form
402	192	2.09	2.10	A .00	.24	.97	-.1	.80	-.7	1.06	.93	.92	1	Form A
388	192	2.02	2.10	A .00	.21	1.01	.1	.98	-.1	1.00	.90	.90	2	Form B
395.0	192.0	2.06	2.10	.00	.23	.99	.0	.89	-.4		.92		Mean (Count: 2)	
7.0	.0	.04	.00	.00	.02	.02	.1	.09	.3		.01		S.D. (Population)	
9.9	.0	.05	.00	.00	.02	.02	.2	.13	.4		.02		S.D. (Sample)	

Model, Populn: RMSE .23 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .23 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 1 significance (probability): 1.00

Panellist group measurement report (high-intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Group
499	256	1.95	1.91 A	.00	.18	.77	-2.5	.59	-2.7	1.32	.93	.91	1 Insider
243	128	1.90	1.91 A	.00	.27	1.43	2.3	1.94	2.0	.49	.88	.90	2 Outsider
371.0	192.0	1.92	1.91	.00	.22	1.10	-.1	1.27	-.4		.91		Mean (Count: 2)
128.0	64.0	.03	.00	.00	.04	.33	2.4	.68	2.4		.02		S.D. (Population)
181.0	90.5	.04	.00	.00	.06	.47	3.5	.96	3.4		.04		S.D. (Sample)

Model, Populn: RMSE .23 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .23 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 1 significance (probability): 1.00

Panellist measurement report (high-intermediate level)

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Panellist
61	32	1.91	1.88	.19	.51	.52	-1.9	.37	-1.6	1.59	.95	.91	1 A
63	32	1.97	1.98	-.33	.51	.72	-1.0	.55	-1.0	1.39	.93	.91	2 B
64	32	2.00	2.02	-.58	.50	.63	-1.5	.46	-1.2	1.50	.94	.91	3 C
59	32	1.84	1.76	.73	.53	.73	-.8	.50	-.8	1.32	.94	.92	4 D
62	32	1.94	1.93	-.07	.51	.95	.0	.82	-.2	1.07	.92	.91	5 E
63	32	1.97	1.98	-.33	.51	.97	.0	.76	-.4	1.12	.92	.91	6 F
63	32	1.97	1.98	-.33	.51	.86	-.4	.63	-.7	1.26	.93	.91	7 G
64	32	2.00	2.02	-.58	.50	.76	-.8	.60	-.8	1.35	.93	.91	8 H
59	32	1.84	1.76	.73	.53	.91	-.1	.63	-.5	1.18	.93	.92	9 I
55	32	1.72	1.45	1.95	.58	1.47	1.1	2.20	1.2	.47	.88	.90	10 J
69	32	2.16	2.30	-1.87	.52	1.42	1.2	2.78	1.7	.34	.86	.89	11 K
60	32	1.88	1.82	.46	.52	1.90	2.4	2.17	1.8	-.04	.85	.92	12 L
61.8	32.0	1.93	1.91	.00	.52	.99	-.2	1.04	-.2		.91		Mean (Count: 12)
3.3	.0	.10	.20	.90	.02	.39	1.2	.80	1.1		.03		S.D. (Population)
3.5	.0	.11	.21	.94	.02	.41	1.3	.83	1.2		.03		S.D. (Sample)

Model, Populn: RMSE .52 Adj (True) S.D. .73 Separation 1.40 Strata 2.20 Reliability .66
Model, Sample: RMSE .52 Adj (True) S.D. .78 Separation 1.50 Strata 2.33 Reliability .69
Model, Fixed (all same) chi-square: 32.8 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 8.8 d.f.: 10 significance (probability): .55