GEPT®
LTTC 全民英檢

Language unfolds worlds.
Testing sets standards.

LTTC-GEPT Research Reports RG-08

Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)

Ute Knoch
Kellie Frost

LTTC® 財團法人語言訓練測驗中心
THE LANGUAGE TRAINING & TESTING CENTER

Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)

LTTC-GEPT Research Reports
RG-08

Ute Knoch
Kellie Frost

This study was funded and supported by the Language Training & Testing Center (LTTC) under the LTTC-GEPT Research Grants Program 2014-2015

LTTC-GEPT Research Reports RG-08
Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)

Published by The Language Training and Testing Center
No.170, Sec. 2, Xinhai Rd., Daan Dist., Taipei, 10663 Taiwan (R.O.C)

First published July 2016

Foreword

We have great pleasure in publishing this report: *LTTC-GEPT Research Reports RG-08*. The study described in this report was funded by the 2014-2015 LTTC-GEPT Research Grants. Headed by Dr. Ute Knoch of the University of Melbourne, Australia, the study adopted an online asynchronous twin-panel approach and followed the suggested methods and procedures set out in the *CEFR Manual* to map the GEPT Writing Test suite onto the CEFR levels. The study not only provides empirical evidence of the relationship between the GEPT and the CEFR, but also offers useful recommendations for further improvement of the quality of the GEPT.

The GEPT, developed more than a decade ago by the LTTC to serve as a fair and reliable testing system for EFL learners, has gained wide recognition in Taiwan and abroad. It has generated positive washback effects on English education in Taiwan. As the GEPT has successfully reached out to the international academic community with remarkable success over the years, numerous studies and research projects on GEPT-related subjects have been conducted and published as technical monographs, conference papers, and refereed articles in books and journals. In view of the growing scholarly attention on the GEPT, and in order to assist external researchers to conduct quality research on topics related to the test, the LTTC has set up the LTTC-GEPT Research Grants Program, which offers funding to outstanding research projects.

The annual call for research proposals is publicized every October, attracting proposals from all over the world. A review board, which comprises scholars and experts in English language teaching and testing from Taiwan and abroad, evaluates the research proposals in terms of the following criteria:

- the relevance to identified areas of research
- the benefit of the research outcomes to the GEPT
- the theoretical framework, aims and objectives, and methodology of the proposed research
- the qualifications and experience of the research team
- the capability of the research outcomes to be presented at international conferences and published in journals
- the timeline and cost effectiveness of the proposed research

Complete and up-to-date information about the GEPT is available at https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm. Full research reports can be downloaded at https://www.lttc.ntu.edu.tw/lttc-gept-grants.htm.

We believe that with the further contributions from the external research community, the GEPT will continue to refine its quality and achieve wider recognition at home and overseas.

Hsien-hao Liao
Executive Director
LTTC

Author Biodata

Dr. Ute Knoch is the Director of the Language Testing Research Centre at the University of Melbourne. She has published in journals such as Language Testing, Language Assessment Quarterly, TESOL Quarterly, Applied Linguistics, Assessing Writing, Journal of Second Language Writing and Language for Specific Purposes. Her research interests are in the area of writing assessment, rating processes, assessing languages for academic and professional purposes, and placement testing. She is currently the Co-president of the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ). In 2014, Dr. Knoch was awarded the TOEFL Outstanding Young Scholar Award by the Educational Testing Service (Princeton, US), recognizing her contribution to language assessment.

Ms. Kellie Frost is a Research Fellow in the Language Testing Research Centre at the University of Melbourne. Her main interests are in language testing, particularly in the areas of speaking and listening assessment, and language assessment for professional and migration purposes. She is currently completing her PhD thesis, examining language test impact in the context of Australia's skilled migration policy.

摘要

◆ **研究團隊與研究目的**

本研究由澳洲墨爾本大學 Ute Knoch 博士主持，依照 *Relating Language Examinations to the CEFR: A Manual* (Council of Europe, 2009)建議的程序—包含 familiarization（熟悉 CEFR 分級）、specification（審試測驗品質與內容和 CEFR 級數的關聯）、standardization（標準設定，即判斷試題對應的 CEFR 級數），與 empirical validation（實證研究）等四階段—由台北與墨爾本兩地的兩個英語教學與評量專家小組（twin-panel），判斷全民英檢初級至高級寫作能力測驗對應 CEFR 的級數，研究結果為全民英檢寫作能力測驗提供更多的效度證據，且提供增進測驗品質的建議。

◆ **研究問題**

1. 參照全民英檢各級寫作能力測驗與 CEFR 級數。
2. 探討測驗內部人員與外部人員的觀點和不同標準設定方法所得到的結果。

◆ **研究方法摘要**

1. 測驗內容分析（specification）由墨爾本大學研究團隊與 LTTC 研究人員協力完成，過程主要分析全民英檢寫作能力測驗各級試題內容，並根據分析結果判定全民英檢各級所對應的 CEFR 級數。
2. 標準設定（standardization）由 15 位具英語教學與評量背景的老師與研究人員所組成的兩個專家小組分別在墨爾本與台北兩地進行。墨爾本組對於全民英檢較陌生，但對 CEFR 與其他國際英語測驗（例如劍橋英語能力測驗）都非常熟悉；台北組對於全民英檢較熟悉，且成員皆深入瞭解 CEFR 或曾參與類似的研究。標準設定的程序透過線上的方式進行，採用混合 Contrasting Group 與 Borderline Group 的方法，每位成員依試題內容與考生作答，根據 CEFR 寫作能力說明，判斷 CEFR 級數。

◆ **研究結果摘要**

1. 測驗內容分析與標準設定的結果皆顯示，全民英檢各級寫作能力測驗與所預期的 CEFR 級數相符，即初級、中級、中高級、高級分別對應 CEFR A2、B1、B2、與 C1 級。
2. 兩地專家小組與兩種方法的研究結果非常接近。
3. 根據專家小組判定的結果，研究團隊建議全民英檢寫作測驗各部份的通過標準微幅向下調整，以更符合所對應的 CEFR 標準。然而，測驗情境下的寫作須符合題目的要求（task requirements），但由於 CEFR 能力說明並沒有包含題目的要求，因此可能影響專家小組對於考生作答與 CEFR 能力說明的對應結果，且 CEFR 無對應句子組合、句子重組與翻譯相關的能力說明，本參照研究沒有包含初級、中級、與中高級寫作能力測驗的第一部份，本研究結果尚需進一步的研究佐證。

Abstract

The aim of this research project was to conduct an empirically-evidenced linking study to align the General English Proficiency Test (GEPT) writing tests to the Common European Framework of Reference (CEFR). Part one of the writing test at the Elementary, Intermediate, and High-Intermediate levels was excluded because the CEFR does not have scales relevant to the constructs. The study closely followed the suggested methods and procedures set out in the Manual on *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe, 2009) and therefore involved the following four stages: familiarization, specification, standardization and validation. The standard-setting study used a twin panel design, with a group of standard-setting judges in Australia and a group in Taiwan. Two standard setting methods were employed: the borderline method and the contrasting groups method. The standard setting process was conducted entirely online. Participants were provided online materials that they were able to access and interact with at times convenient to them.

The findings from the specification phase showed that, in terms of test content and task design, the different GEPT writing sub-tests aligned well with the intended CEFR levels. The results from the two standard-setting panels also supported alignment with the CEFR. Cut-scores resulting from the analysis of the panellists' judgments, however, indicate that the existing GEPT pass score for the sub-test may need to be set slightly lower in order to reflect the relevant CEFR level benchmarks. However, it should be noted that writing in the GEPT takes place under testing conditions in which 'task requirement' is a crucial criterion for evaluating candidates' performance, whereas 'task requirement' is not included in the CEFR can-do descriptors. This distinction may have influenced the resulting cut-scores. Results from the two panels and the two standard-setting methods were similar, adding validity to the findings.

Table of Contents

Table of Tables

1. Introduction

1.1 Overview

The aim of this research project was to conduct a study to link the GEPT writing sub-test to the Common European Framework of Reference (CEFR). The study was conducted according to the stages and methods set out in the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009). The following four stages were included: familiarisation, specification, standardisation and empirical validation. The study was designed to directly address the key research area of test validation by providing empirical evidence of a link between performance on the GEPT writing test and the CEFR.

1.2 The Common European Framework of Reference and language testing

The Common European Framework of Reference (CEFR) was developed by the Council of Europe in an attempt to provide common reference levels for teaching and learning for all languages in Europe. The CEFR divides learners into three broad divisions that can be further divided into six levels:

A Basic User
A1 Breakthrough
A2 Waystage

B Independent User
B1 Threshold
B2 Vantage

C Proficient User
C1 Effective Operational Proficiency
C2 Mastery

Language proficiency is described in a set of scales covering a range of skills, including reading, listening, writing and speaking in the publication *Common European Framework of Reference for Languages: Learning, teaching and assessment* (Council of Europe, 2001). Since its inception, the CEFR has become hugely influential in language assessment circles. In fact, the CEFR was designed with language testing in mind. The manual states that: 'one of the aims of the Framework is to help partners to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualification' (Council of Europe, 2001). The influence of the CEFR has been felt not only in Europe. The scales have become a set of standards adhered to across the world, and most major language testing agencies have already, are in the process of, or are feeling pressure to link their tests to the CEFR (see e.g. Milanovic & Weir, 2010). The GEPT in Taiwan has been cited as one such example. To help practitioners in the linking process, the Council of Europe piloted a set of procedures for linking tests to the CEFR in 2003, and a formal manual on the process was published in the *Manual on 'Relating language*

examinations to the Common European Framework of Reference for Languages: Learning, teaching, and assessment in 2009' (Council of Europe, 2009). While many institutions have conducted linking studies, most of these are unpublished or released as an internal report. One notable exception is the edited collection of linking studies 'Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual' (Martyniuk, 2010).

The Council of Europe's Manual proposes four distinct stages to be used in the linking process:

(1) Familiarisation: A panel of teachers take part in a number of activities designed to familiarise them with the CEFR and its associated scales and descriptors.
(2) Specification: Experts taking part in the linking process audit the coverage of the test to be linked to the CEFR and complete a number of forms to establish an overview of the test in relation to the CEFR.
(3) Standardisation: A standard setting meeting with a panel of trained judges is undertaken to establish the exact relationship between the CEFR and the test.
(4) Validation: Internal and external validation studies are undertaken to validate the claim of linkage to the CEFR.

1.3 The GEPT

The General English Proficiency Test (GEPT) is developed and administered by the Language Training & Testing Centre (LTTC), based in Taiwan. The general proficiency test is used for a range of purposes and by a range of English learners, including job applicants, employees and students. The test focuses on the skills of listening, reading, writing and speaking, and is accepted by a range of institutions both in Taiwan and further abroad. The test is offered at five levels (Elementary, Intermediate, High-Intermediate, Advanced and Superior). Writing sub-tests are part of all levels, but differ in format as outlined in more detail below in Table 1.

Table 1. GEPT writing tasks by test level

Level	**Task types**	**Number of items**	**Mins**
Superior	A 750-word essay based on a 10-15 min. Video/radio program and a 3000-word article	n/a	3 hours
Advanced	Summarizing main ideas from verbal input and expressing opinions	n/a	60
	Summarizing main ideas from non-verbal input and providing solutions	n/a	45
High-Intermediate	1 – Chinese-English Translation 2 – Guided Writing	2	50
Intermediate	1 – Chinese-English Translation 2 – Guided Writing	2	40
Elementary	1 – Sentence Writing 2 – Paragraph Writing	16	40

The specific level descriptions for writing at the five levels are described in Table 2, below:

Table 2. Skill area descriptions by GEPT writing test levels

Level	Skill area level description
Superior	• can express themselves with precision and clarity • can effectively carry out in-depth investigations into most subjects • can write with a logically-organized structure and demonstrate sophisticated rhetorical skills
Advanced	• can summarize articles on general and professional topics • can write well-organized and coherent essays, with appropriate lexical and grammatical usage • can express their opinions on a range of topics and discuss them in depth
High-Intermediate	• can write about topics related to daily life • can write about personal viewpoints on current events
Intermediate	• can use simple English to write feedback and comments • can write about their own experiences or about topics with which they are familiar
Elementary	• can write simple sentences and paragraphs

1.4 Research aims

The primary aim of the linking study was to establish if there was a correspondence between CEFR levels and the score levels of the GEPT writing test. The linking study focused on four levels of the GEPT writing suite: Elementary, Intermediate, High-Intermediate and Advanced. The Superior level was excluded from the current study because it is not administered on a regular basis. In order to link each GEPT test level to its corresponding level on the CEFR, the four stages set out in the Manual were followed: Familiarisation, Specification, Standardisation and Validation. A 'twin-panel' approach was adopted (see Brunfaut and Harding, 2014), to compare the judgments of those familiar with the GEPT (test 'insiders') with the views of those with little if any prior exposure to the GEPT (test 'outsiders') but with knowledge of the CEFR, and to provide a means of cross-validating panel decisions. The latter group consisted of seven judges, all based in Australia; the former group consisted of eight judges based in Taiwan and experienced in using the GEPT.

1.5 Research questions

Specifically, the linking study aimed to address the following two research questions:
1. How do the GEPT writing test levels and scores relate to the CEFR?
2. How did the judgments of test 'outsiders' compare to the judgments of test 'insiders'?

2. Overview of research design

The standard setting process was conducted entirely online over a three week period, with participants provided standard setting materials electronically, in the form of PDF documents, Word documents and PowerPoint presentations, which they were able to access and interact with at times convenient to them over the duration of the process. There was no face-to-face

interaction between individual panellists or between panellists and coordinators, and discussion took place asynchronously via a Google discussion forum and via group emails.

Conducting standard setting studies using online, internet-based tools is thought to provide a viable, cost-effective alternative to face-to-face standard setting meetings, which can be expensive to run and logistically difficult to arrange (Katz, Tannenbaum and Kannan, 2009; Katz & Tannenbaum, 2014). In the current study, the use of an internet-based method combined with flexibility in timing allowed greater participation and enabled input from two independent panels with different perspectives, which would otherwise have proven either overly costly, in terms of travel, or exceedingly difficult to schedule, due to diverse work-schedules and time differences between the two locations.

We acknowledge, though, that alongside these budgetary and logistic benefits, an online approach also gives rise to potential issues. As Katz, Tannenbaum and Kannan (2009) point out, "participants in a virtual team are more likely to become distracted, work on other parallel tasks, or drop out of the team altogether" (p. 20). Harvey (2000), in a comparison of onsite and online standard setting studies of the College Level Examination program in the United States, found that online panellists felt less involved and were less likely than onsite panellists to see themselves as active participants in discussions. On the positive side, however, Katz, Tannenbaum and Kannan (2009) also note that in being provided materials electronically to review in their own time, participants are able to benefit from repeated exposure to content.

As shown further below, the validity of the current approach in relation to these potential concerns was explored via questionnaire data collected from participants during and at the conclusion of the process.

3. Methods

3.1 Participants

As noted, a twin panel approach was adopted in the current study involving 15 participants in total across the two panels. Members of both panels were familiar with the CEFR. The panel based in Taipei, Taiwan was comprised of eight judges recruited by the LTTC. Two were male and six were female, and the age range was 31-50 years. All possessed either a Master or PhD level qualification, and were experienced English teachers who had worked extensively with the GEPT. Seven out of eight had previously participated in standard setting activities.

The second panel, based in Melbourne, Australia, was comprised of seven experienced English language teachers, three male and four female, recruited by the Language Testing Research Centre (LTRC) at the University of Melbourne. Three participants held Bachelor degrees with additional CELTA qualifications, and four held a Master level qualification. The age range for this group of panellists was between 31 and 60 years. The Australia-based group members were previously unfamiliar with the GEPT, as mentioned, but had extensive

experience with other internationally recognised English language tests, such as the Cambridge suite of exams and with the CEFR. Of these panellists, two had been involved in standard setting previously.

An overview of panel participants' English language teaching and testing experience is provided in Table 3, below.

Table 3. English language teaching and testing experience, by panel

	English teaching experience			English testing and assessment experience		
	Number of years		Type of experience	Number of years		Type of experience
	Range	Mean		Range	Mean	
Taiwan	3-24	10.4	Primary, secondary, higher education, adult education	3-23	9.2	test review, test design, item writing and test development, assessing, assessor training, test validation
Australia	2-20	15.1	Primary, secondary, higher education, pre-university transition, adult education	0-15	7.7	test design, item writing and test development, assessing, assessor training

3.2 Procedures

As already mentioned, in order to link each GEPT test level to its corresponding level on the CEFR, four interrelated stages were completed as part of the overall standard setting process, as set out in the Manual: Familiarisation, Specification, Standardisation and Validation. The project researchers conducted the specification stage in collaboration with staff from the LTTC, as detailed further below, and the familiarisation and standardisation stages involved the participation of the 15 panellists. Throughout the latter two stages, evidence was collected in support of the procedural and internal validity of the overall standard setting process (the validation stage).

Before detailing the procedures involved in each of the four stages of the standard setting process, as set out in the Manual, an overview is first provided below.

3.2.1 Overview of the standard setting process

As stated above, the standard setting process was conducted entirely online. Electronic versions of all materials were delivered to each panel member using secure Dropbox folders. A Google discussion forum was created to enable group discussion and mediation of judgments in the training sessions, to ensure that panellists were able to arrive at a shared understanding of how to apply CEFR descriptors to the GEPT writing tasks across each level of the test. In addition to the discussion forum, regular communication between session coordinators and panel participants was maintained via email. Broadly speaking, the standard setting process was divided into two parts and five individual sessions, as follows:

Part 1:

1. Preparatory session
2. Familiarisation session
3. Standardisation/ Benchmarking training session

Part 2:

4. Standard setting session (Round 1)
5. Standard setting session (Round 2 – final judgments)

Panellists were able to go through the materials and complete activities at times convenient to them within the days scheduled for opening and closing each part of the standard setting process. As indicated, sessions 1, 2 and 3 represented the first part of the process. These sessions were focused on providing panellists with preparation and training in using the CEFR and standard setting practices, and familiarity with the GEPT writing test. In sessions 4 and 5, the actual standard setting activities took place.

The five sessions were first conducted with the Melbourne group, and subsequently with the Taipei group. The Melbourne panel sessions ran for three weeks, from the 30th of November until the 21st of December, 2015. The Taipei group sessions opened on the 18th of January, 2016 and also ran for three weeks. Within the three weeks, one week was allocated for participants to complete sessions 1 to 3, one week to complete session 4 (round 1 of standard setting judgments) and four days to complete round 2 judgments. Of the remaining three days, two days were situated in-between sessions 3 and 4, to allow panellists time to comment and ask questions on the discussion forum before the actual standard setting sessions commenced. One day was situated between rounds 1 and 2 of the standard setting sessions, to allow time for the researchers to collate judgments and prepare feedback.

Each of the four stages involved in the standard setting process (familiarisation, specification, standardisation and validation) will now be described separately, below.

3.2.2 Familiarisation

The Familiarisation stage was undertaken at the beginning of both the specification and standardisation stages, by the project researchers in the case of the former stage, and by the 15 panellists who took part in the standard setting process for the latter. The Familiarisation stage was intended to ensure that the researchers and all participants in the standard setting process possessed an in-depth familiarity with the CEFR scales and descriptors, and to the extent possible, a shared understanding of how the CEFR level descriptors should be applied to GEPT writing scripts.

To this end, prior to completing the Specification stage, the project researchers re-familiarised themselves with the CEFR, in particular with the writing scales and the level descriptors. In addition, researchers reviewed exemplar writing task and response samples provided in the European Language Portfolio (ELP) by the Council of Europe (www.coe.int/en/web/portfolio),

illustrating the CEFR levels. The examples provided in the ELP were drawn from the Cambridge ESOL suite of examinations, and consisted of one task and response sample per CEFR level. Researchers also familiarised themselves with the GEPT by reviewing available information on the GEPT exam.

For the panellists, the Familiarisation stage was undertaken throughout the three sessions, outlined above: the preparatory session, familiarisation session, and benchmarking/ standardisation training session. Each session is outlined below:

1. *Preparatory session*

The preparatory session involved reading sections of the manual, including level descriptors, and accessing exemplar writing samples provided by the Council of Europe via the CEFTrain web-based training tool. Panellists were required to make practice judgments on the CEFR levels of each of the samples, and received immediate feedback on the consistency of their decisions with the consensus achieved by a group of CEFR experts on each sample, including explanations.

2. *Familiarisation session*

In the Familiarisation session, panellists were provided with two PowerPoint presentations, which they worked through individually upon completing the preparatory session materials. The first was an introductory presentation, aimed at familiarising or re-familiarising panellists with both the CEFR and the GEPT writing test. The second presentation involved a set of interactive familiarisation tasks, drawn from the range of recommended activities in the Manual. These activities included:

- Reconstructing or sorting CEFR scales, including the overall scale and the writing sub-scales (in particular Table C4 in the manual)
- Self-assessment of their own foreign or second language abilities

3. *Benchmarking/Standardisation training session*

To conclude the Familiarisation stage, a Benchmarking/Standardisation training session was conducted, by way of a third PowerPoint presentation in which illustrative samples across the relevant CEFR levels, provided by the Council of Europe, were shown to panellists with salient features highlighted. In addition, two sample GEPT scripts per test level (Elementary, Intermediate, High-Intermediate and Advanced) were provided to each of the panellists. They were asked to refer to the CEFR writing descriptors in order to individually assign a CEFR level to each sample. Panellists were then able to access the consensus judgments of three highly experienced CEFR users, including explanations on the sample scripts, and were invited to discuss areas of contention or confusion via the Google discussion forum. The forum was mediated by a session coordinator, with the aim of helping all participants arrive at a shared understanding of the writing qualities corresponding to each level of the CEFR.

3.2.3 Specification

The Specification stage was completed by the project researchers and colleagues in the LTRC at the University of Melbourne, in collaboration with LTTC staff. Project researchers and colleagues were all familiar with the CEFR and all possessed extensive experience in language test development, English language teaching, and standard setting.

As set out in the Manual, the Specification stage involved a comprehensive detailing of GEPT test materials and specifications, as well as an analysis of writing tasks at each of the test levels, in terms of content coverage, task types and assessment criteria. The specification forms provided in the Manual were used for this purpose. Three parallel writing tests for each of the four GEPT levels, Elementary, Intermediate, High-Intermediate and Advanced, provided by the LTTC, formed the basis of the analysis of test tasks. Completed specification forms are included in Appendix A of this report. LTTC staff assisted by completing sections of forms A1-8 and A14, where information was confidential or otherwise not available to the researchers.

The specification stage also included deriving an initial estimate of the alignment between the GEPT writing test levels and the CEFR levels, which would be scrutinised further throughout the standard setting process. The LTTC already provides details of estimated alignments between the GEPT and the CEFR levels on its website (www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/alignment.htm) for reading and listening, informed by existing linking studies (reading: Wu & Wu, 2010; Wu, 2014 and listening: Brunfaut & Harding, 2014).

3.2.4 Standardisation

Following the Familiarisation stage, which concluded with a benchmarking/standardisation training session, panellists were provided with an introductory PowerPoint presentation to begin the Standardisation stage. This presentation involved a general introduction, including explanations of the importance of standard setting, an overview of the reasons for mapping the GEPT against the CEFR, and an outline of the standard setting method and procedures. Standard setting materials for each level of the test were supplied as word documents, in addition to the PowerPoint presentation. The standard setting method and procedures and the standard setting materials are detailed separately, below.

1. *Standard setting method and procedures*

One of the most challenging matters confronting language test developers today is the setting of performance standards. Standard setting, or mapping test scores to descriptions of language skills expressed within a scale of levels or competencies, such as the CEFR, is a way of attributing meaning to test scores (Kane 2012) and of thereby translating test scores into descriptions of test takers abilities and/or achievements.

Extensive research has been carried out in the disciplines of educational measurement and psychometrics on standard setting (see Angoff, 1971; Cizek, 2001, 2012; Cizek & Bunch, 2007; Jaeger, 1982; Thurstone, 1927; Zieky, 2001), and a variety of methods exist to set cut scores for writing test performances (see Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007; Zieky et al., 2008; Cizek, 2001, 2012) and to situate the passing standards represented by cut scores within scales and level descriptors on an established framework, such as the CEFR.

In the current project, a combination of the 'Contrasting Group' and 'Borderline Group' methods was adopted, as described in the Manual (Council of Europe, 2009, p. 69). Generally, these methods are considered to be examinee-centred, as they rest on panellists being able to classify test takers who are known to them, based on their familiarity with these individuals' abilities in classroom contexts, for example. In this project, rather than basing judgments on knowledge of individuals, panellists classified test takers on the basis of their writing performances.

Examinee-centred methods, such as these, allow for a focus on the language produced by test takers as a basis means of judging their language ability level, and are thus well suited to standard setting on performance-based tests of writing and also speaking. Test-centred methods, by contrast, focus on evaluating the difficulty of test items in relation to the ability of test takers (Cizek & Bunch, 2007), and are not readily adaptable to performance tests where test takers are required to respond in writing or orally to a single prompt.

The Contrasting Groups method requires panellists to make judgments about the status of writing performances produced by examinees, as indicative of mastery or non-mastery at a particular target skill level. The Borderline Group method is very similar, except that the focus is on identifying performances that represent borderline cases, between mastery and non-mastery (Cizek & Bunch, 2007). In using the Borderline Group method in isolation, there is the risk of the number of scripts classified as borderline being small, which would result in large standard errors associated with mean test scores and by extension, tenuous cut scores. In combining the methods, this potential problem is minimised.

In applying a combined Contrasting Groups-Borderline Group approach, panellists are asked to classify writing scripts into three categories: mastery, borderline and non-mastery. Judging the mastery, non-mastery or borderline status of test takers on the basis of their language production is considered a familiar task for language educators, which, on the basis of their existing professional experience, can be expected to be undertaken with sufficient competence to ensure meaningful standard setting outcomes.

In the current study, each test taker was judged independently by 15 different panel members across the Australia-based and Taiwan-based panels as part of the standard setting procedures to link the GEPT writing test to the CEFR. Two rounds of judgments were conducted. In round one, panellists were informed of the CEFR level that each of the GEPT writing test levels was intended to capture. For each test level, they were then instructed to use the CEFR

written assessment grid and their knowledge of the CEFR level descriptors to judge each of the scripts provided to decide if:

- the script could be placed at the target CEFR level, for example, A2 in the case of the Elementary test (thereby indicating 'mastery' at that level); or
- the script was below the target CEFR level ('non-mastery'); or
- the script was borderline (in-between clearly at the target CEFR level and clearly below the target CEFR level).

This process was repeated across 30 scripts for each of the four GEPT test levels, 120 scripts in total. All panellists were asked to complete their judgments independently.

At the end of the first round of judgments, the researchers reviewed decisions and provided feedback to panellists concerning how their judgments compared with the judgments of the rest of the group across the four GEPT test levels.

In round two, panellists were asked to review scripts where their judgments deviated from the majority decision and to provide final judgments.

2. Standard setting materials

The LTTC provided researchers in the LTRC with two parallel test forms, writing scripts, rating scales and test scores for each script across each of the four GEPT writing test levels. The performances consisted of responses to both tasks for the Elementary and Advanced test levels (see Table 1, above). In the Intermediate and High-Intermediate tests, as set out in Table 1, above, test takers are required to complete a Chinese-English translation and a guided writing task. For the purpose of this linking study, only the guided writing task and associated test performances were included in the standard setting materials.

Thirty writing performances for each test level were selected by the researchers, 120 in total, for use in the Standardisation stage. Performances were selected in order to achieve an even distribution across the two parallel test forms and a spread of test scores across each level of the test. No test score information was provided to the 15 panellists, and selected scripts were not ordered according to score.

As mentioned, the 15 panellists were each provided with a standard setting pack via their individual online folders, accessible through DropBox. The packs were identical for all participants across the two panels, and contained sub-folders for each of the four test levels: Elementary, Intermediate, High-Intermediate and Advanced. Each of the four test level sub-folders contained two parallel GEPT test forms; 30 writing scripts, the CEFR written assessment grid, and a judgment form.

3.2.5 Validation

Two types of validity evidence were collected as part of this linking project: Procedural and Internal Validity. An investigation of external validity, which would involve comparing performances of test takers on the GEPT with performances on another writing proficiency test which has already been linked to the CEFR, was beyond the scope of the current project.

1. *Procedural validity*

Procedural validity was promoted in the current project by applying clear, systematic and rigorous criteria to the selection of panellists to ensure that all possessed extensive English language teaching experience, experience in using rating scale criteria to assess writing performances, and familiarity with the CEFR. In addition, the standard setting procedures and processes were fully specified and all materials were thoroughly reviewed prior to commencement.

In addition, the following procedural validity evidence was collected via questionnaires (see Appendix B) throughout the familiarisation and standardisation stages of the project:

- Evidence that the purpose of the standard setting process was clear, and that the instructions and procedures for each session were clear, understood and followed by panellists was collected via questionnaire.
- Evidence of the judges' familiarity with the CEFR after the Familiarisation stage was collected via questionnaire; and
- Evidence of the judges' confidence in the benchmarking process was collected via questionnaire.

2. *Internal validity*

Internal validity evidence was collected in the Standardisation stage of the project. This evidence included:

- the calculation of estimates of inter-judge consensus, to establish how well judges agreed with each other;
- cross-panel comparisons

3.3 Data analysis

Judgments by panellists from each of the two panels (Australia-based and Taiwan-based) across the 30 scripts and overall GEPT band scores for each script, provided by the LTTC, were entered into separate excel spreadsheets. As mentioned above, the GEPT rating scale for the Elementary, Intermediate and High-Intermediate tests consists of 6 bands (0-5), and 5 bands for the Advanced test (1-5). The GEPT band score range of the 30 selected scripts at each test level is as follows:

- Elementary: 2 - 5
- Intermediate: 2 - 5
- High-Intermediate: 2 - 5
- Advanced: 1 - 4

Mean GEPT scores and standard deviations were calculated for each of the three decision points (at target CEFR level, borderline, below target CEFR level) across each of the four test levels (Elementary, Intermediate, High-Intermediate, Advanced) for each panel group.

Judgments and GEPT scores for all panel members across both panel groups were then combined for each of the four test levels, and Rasch analyses were conducted using FACETS to check for misfitting scripts. In calculating potential cut scores misfitting scripts were removed.

Intraclass correlations and percentage agreement with the mode (Harsch & Martin, 2012) were calculated as measures of inter-judge agreement for each panel group, and for the combined panel group (all 15 participants). The mode is the most common rating of a performance, and the percentage of judges that agree with the mode is calculated for each script. The final measure is the average of the percentages across all of the scripts. The percentage agreement with the mode was calculated for each of the four GEPT test levels.

Two potential cut scores were derived from the combined panel group as follows:

- The mean GEPT test scores and standard deviations of scripts classified as borderline were calculated for each of the four GEPT test levels (according to the Borderline method).
- The average of the GEPT test score means of scripts classified as at the target CEFR level and below the target CEFR level were calculated (according to the Contrasting groups method).

4. Findings

The results are presented in four sections. Firstly, the specification results are presented which compare the CEFR descriptions with a content analysis of the test materials and arrive at an initial estimate of the CEFR level. We then present the judgments of the combined panel and outline how the cut scores would change if the judgments of all judges were adopted. We then show the results for our two sub-panels separately, first those of the Melbourne-based judges and then those of the Taiwan-based judges. The final section of the results relates to our findings relating to procedural and internal validity.

4.1 Specification results

On the basis of the comparison between the CEFR writing scale descriptors and the content analysis undertaken here within the specification phase, initial estimated alignments between GEPT writing test levels and CEFR levels were established. At the Specification stage, these

initial estimates were consistent with estimations for reading and listening, as provided by the LTTC. The estimated alignment between the GEPT writing test levels and the CEFR is shown in Table 4, below.

Table 4. Estimated alignment between GEPT writing test levels and CEFR levels

GEPT writing test level	CEFR level
Elementary	A2
Intermediate	B1
High-Intermediate	B2
Advanced	C1

Table 5 below shows the existing GEPT 'pass' scores on the GEPT rating scale for each of the four test levels. Linking the GEPT to the CEFR means that test takers who achieve the GEPT pass score will be deemed to possess a level of writing ability that corresponds to the target CEFR level to which each test is aligned.

Table 5. Existing GEPT cut scores by test level

GEPT writing test level	CEFR level	Existing GEPT pass score*
Elementary	A2	4
Intermediate	B1	4
High-Intermediate	B2	4
Advanced	C1	3

*GEPT rating scale consists of 6 band scores (0 to 5) for the Elementary, Intermediate and High-Intermediate test levels, and 5 band scores (1 to 5) for the Advanced test level.

As set out above in the research questions for the current project, the suitability of the current GEPT pass scores as indicators of particular CEFR level achievement will be scrutinised via the standard setting procedures set out below.

4.2 Findings from combined panel

As outlined in the methodology above, judges were asked to review writing samples from each of the test levels and judge whether the writing sample is at the relevant CEFR level for the test (e.g. A2 for the Elementary test, B1 for the Intermediate test etc), whether it is below that level or whether it is in between these two decision points (i.e. borderline). We then mapped the original GEPT scores against each of these decision levels and worked out the means and SD for each group of scripts. These mean GEPT test scores and standard deviations are presented for each GEPT test level in Table 6 below.

Table 6. Combined group mean GEPT scores across decision points

Elementary	At A2	Borderline	Below A2
Mean	4.22	3.11	2.17
SD	.72	.79	.38
Intermediate	At B1	Borderline	Below B1
Mean	4.13	3.06	2.38
SD	.79	1.02	.59
High-Intermediate	At B2	Borderline	Below B2
Mean	4.34	3.33	2.5
SD	.70	.62	.62
Advanced	At C1	Borderline	Below C1
Mean	3.00	2.51	1.83
SD	.82	.82	.71

Following expectations, the combined panel judgments resulted in mean scores that were lowest for scripts judged to be below the target CEFR level and highest for scripts judged to be at the target CEFR level, with the average GEPT score of the borderline scripts falling somewhere in-between. In addition, for all tests except for the Advanced level, there was a band score difference of one or more between the 'At CEFR level' category and the borderline category. It is clear from this table that the combined panel generally agreed with the broad alignment of the writing sub-tests to the CEFR, as it has also been found for other sub-tests of the GEPT.

Following from Table 6 above, Table 7 below outlines potential GEPT cut scores to CEFR levels if either the borderline method or the contrasting groups method is used. For ease of comparison, we have mapped the existing GEPT pass score against these two possible cut score methods.

Table 7. Comparison of existing GEPT pass score and potential cut scores by method

GEPT writing test level	CEFR level	Existing GEPT pass score	Borderline method score	Contrasting Groups method score
Elementary	A2	4	3.1	3.2
Intermediate	B1	4	3.1	3.3
High-Intermediate	B2	4	3.3	3.4
Advanced	C1	3	2.5	2.4

The table shows that the potential cut scores using these two methods of calculation are slightly different, with the contrasting groups method resulting in slightly higher cut scores for all tests apart from the Advanced level. We will discuss these two possible options in our discussion and recommendation section below.

4.3 Comparison of the two standard setting panels: Melbourne and Taiwan

In this section, we present the judgments of the standard setting panellists separately for the two sub-panels, one in Melbourne and one in Taiwan.

Mean GEPT test scores and standard deviations at each of the three decision points for the Melbourne panel are presented for each GEPT level in Table 8 below.

Table 8. GEPT mean scores across decision points – Melbourne panel

Elementary	At A2	Borderline	Below A2
Mean	4.11	3.00	2.11
SD	0.80	0.79	0.32
Intermediate	At B1	Borderline	Below B1
Mean	4.03	2.86	2.33
SD	0.85	0.99	0.48
High-Intermediate	At B2	Borderline	Below B2
Mean	4.33	3.26	2.28
SD	0.65	0.62	0.45
Advanced	At C1	Borderline	Below C1
Mean	3.04	2.44	1.71
SD	0.81	0.75	0.71

Across all four levels of the GEPT writing suite, mean scores were lower for scripts judged to be below the target CEFR level than those judged to be borderline, and the mean score for borderline scripts was lower than the mean score of scripts judged to be at the target CEFR level. Furthermore, at least one score band distinguished scripts deemed to be at the target CEFR level from those judged borderline or below for the Elementary, Intermediate and High-Intermediate tests, and while the score difference between the 'at level' and borderline categories was less than one band for the advanced test, there was a marked difference between the 'at level' and 'below level' categories. Taken together, the Melbourne panel judgments suggest that the four GEPT writing tests are effectively targeting the relevant CEFR levels, thereby supporting existing broad alignments with the CEFR.

The same judgments from the Taiwan-based panel can be found in Table 9 below.

Table 9. GEPT mean scores across decision points – Taiwan panel

Elementary	At A2	Borderline	Below A2
Mean	4.28	3.25	2.27
SD	0.66	0.75	0.45
Intermediate	At B1	Borderline	Below B1
Mean	4.17	3.29	2.41
SD	0.74	0.99	0.64
High-Intermediate	At B2	Borderline	Below B2
Mean	4.34	3.44	2.63
SD	0.74	0.62	0.67
Advanced	At C1	Borderline	Below C1
Mean	2.96	2.60	1.91
SD	0.83	0.89	0.70

Similar to the Melbourne panel judgments, mean scores across all four levels of the GEPT were lowest for scripts judged to be below the target CEFR level and highest for scripts judged to be at the target CEFR level. The mean score differences between adjacent categories were lower for the Taipei panel than for the Melbourne panel, however, with differences of less than one whole band score across all test levels apart from elementary. Still, there was a marked difference between mean scores for the 'at CEFR level' category and mean scores for the 'below CEFR level' category across the four GEPT tests, again suggesting support for the existing alignment with the CEFR.

Table 10 below sets out a comparison of the potential cut scores between the Melbourne and the Taiwan-based panels for both cut score calculation methods. Cut scores derived from the Taipei panel judgments were higher than those derived from the Melbourne panel judgments regardless of calculation method, except in the advanced test where the Contrasting Group method yielded the same score in both panels.

Table 10. Comparison of Melbourne and Taipei panel cut scores

Writing test level	BL cut score (Melbourne)	BL cut score (Taipei)	CG cut score (Melbourne)	CG cut score (Taipei)
Elementary	3.0	3.3	3.1	3.3
Intermediate	2.9	3.3	3.2	3.3
High-Intermediate	3.3	3.4	3.3	3.5
Advanced	2.4	2.6	2.4	2.4

BL=Borderline method, CG=Contrasting Groups method

Regardless of the calculation method or the judging panel, the cut scores found in this study are slightly lower than those currently used operationally by the GEPT and we will take this up in our discussion and recommendation section below.

4.4 Procedural validity

As mentioned above, questionnaire data were collected to evaluate the extent to which judges in both panels understood the standard setting process, felt familiar with the CEFR descriptors and felt able to follow the judgment procedures in the standard setting sessions. Three questionnaires were administered; one at the end of the familiarisation session, one at the end of the benchmarking/standardisation training session, and one at the end of the standard setting process.

The findings relating to the preparatory and familiarisation sessions can be found in Table 11 below. The findings show that all panel members felt well prepared and familiarised with the CEFR and the GEPT following the familiarisation sessions. Two judges, however, felt that they did not have sufficient time to complete the required sessions.

Table 11. Feedback questionnaire – Preparatory and Familiarisation sessions

Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
I have a good overview of the CEFR and the GEPT.	7	8	-	-	-
The sessions were clearly explained and I understood what I was being asked to do.	13	2	-	-	-
I was able to follow instructions and complete the activities.	12	3	-	-	-
The information and activities were useful.	12	3	-	-	-
The relevant CEFR levels and descriptors are clear to me.	12	3	-	-	-
I found the PowerPoint slides and supplementary files easy to work with.	5	10	-	-	-
I was able to complete the sessions within the suggested time required.	6	6	1	2	-

Table 12 below presents the findings from the questionnaire administered following the benchmarking session. Almost all judges found the benchmarking training useful and informative although not all found the discussion forum helpful. There was again one judge who felt that there was not sufficient time allocated to this activity.

Table 12. Feedback questionnaire – Benchmarking training session

Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
The session was informative and clearly presented.	11	3	1	-	-
The illustrative examples helped me understand the CEFR levels.	7	8	-	-	-
The CEFR descriptors were relevant to the GEPT writing samples.	7	7	1	-	-
I was able to follow instructions and complete the practice judgment form as required.	13	2	-	-	-
The discussion forum was helpful.*	1	4	7	-	-
I feel familiar with the CEFR descriptors.	8	7	-	-	-
I feel familiar with the GEPT writing tasks.	8	7	-	-	-
I was able to complete the session within the suggested time required.	8	5	1	1	-

*3 participants did not respond to this question

Table 13 below presents the questionnaire administered following the standard setting sessions. Although some participants selected neutral to some of the questions, the majority of

the panel felt that the standard setting process was clear and that they felt confident about their judgments. There was again one participant who felt that not enough time was allocated to the activity.

Table 13. Feedback questionnaire – Standard setting sessions

Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
The PowerPoint presentation slides provided me with a clear understanding of the purpose of the standard setting process.	10	5	-	-	-
The process instructions were clear and I understood what I was being asked to do.	13	1	1	-	-
Reviewing the writing tasks helped me understand the assessment.	11	4	-	-	-
The relevant CEFR levels and descriptors were clear to me.	6	8	1	-	-
The CEFR descriptors were relevant to the GEPT writing scripts.	4	8	3	-	-
I was able to follow instructions and complete the rating sheets as required.	10	5	-	-	-
Information showing how my judgments differed from the judgments of other participants was helpful.**	3	4	3	-	-
I am confident about the defensibility and appropriateness of the decisions I made about the CEFR levels of the writing samples.	3	9	3	-	-
I was able to complete the session within the suggested time required.	8	5	1	1	-

**As noted above, panellists were asked to review judgments where they deviated significantly from the rest of the group. Five participants were not asked to review any judgments and so responded with N/A to this question.

Qualitative judgments presented below show why three of the panellists were unsure about their judgments in relation to the CEFR levels. These comments all relate to the broad nature of the CEFR descriptors rather than to the procedures adopted for this study:

> *"The task was a useful exercise. I personally find the CEFR guidelines quite vague and as a result I can't help but feel my decisions were made by comparing students' answers against each other rather than against the CEFR." (Melbourne panellist)*

> *"For me, it is difficult to make entirely objective judgments because some of the CEFR descriptors are quite broad and general." (Taipei panellist)*

> *"I found it more difficult to use CEFR descriptors to judge high-level GEPT tests, especially the Advanced level. Perhaps it is because the CEFR descriptors tend to be broad." (Taipei panellist)*

4.5 Internal validity

While procedural validity is important to show that the participants understood and followed all standard setting procedures, it is equally important to show that the judges agreed with each other when grouping the writing samples. As mentioned in the methodology section, we calculated two methods of inter-judge reliability. The first method, the intraclass correlation is correlation-based and shows whether the judges were ranking the writing samples in a similar manner. The second, the percentage agreement with the mode, shows absolute agreement in judgments. Given that the panellists were asked to make relatively fine distinctions (in particular by introducing the borderline category into the judgment process), we expect the percentage agreement with the mode to be significantly lower than the intraclass correlation. Finally, we also conducted a Rasch analysis to investigate rater behaviour.

As shown in Table 14 below, the agreement level was higher in the Taipei panel for the Elementary test, but higher in the Melbourne panel for all other GEPT levels. Regardless of the panel membership, the results show that the judges were highly reliable in making their standard setting judgments, providing further validity evidence for the findings. The combined inter-judge statistics can be found in Table 15 below.

Table 14. Comparison of Melbourne and Taipei panel inter-judge agreement

Test level	Cronbach's Alpha		Agreement with mode (%)	
	Melbourne	Taipei	Melbourne	Taipei
Elementary	0.96	0.98	81.43	83.3
Intermediate	0.93	0.95	76.19	75.0
High-Intermediate	0.96	0.94	81.43	71.0
Advanced	0.93	0.91	73.33	69.6

Table 15. Combined panel group inter-judge agreement by GEPT level

Writing test level	Cronbach's Alpha	Agreement with mode (%)
Elementary	0.98	79.1
Intermediate	0.97	70.4
High-Intermediate	0.97	73.6
Advanced	0.96	67.6

5. Summary and recommendations

The aim of this research project was to link the GEPT writing sub-test to the Common European Framework of Reference (CEFR). The study was conducted according to the stages and methods set out in the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009). The Specification stage was conducted both by the researchers in Melbourne and by the GEPT staff in Taipei. The standard setting study used a twin panel design, with a group of standard setting judges in Australia and a group in Taiwan. The panels were trained online and submitted their judgments electronically.

The findings from the specification phase showed that based on a content review of the test materials, the different GEPT writing sub-tests are aligned with the CEFR according to expectations and in line with the findings of the linking studies of other GEPT sub-tests. The results from the standard setting panels also showed that the test is well-aligned with the CEFR, however, the cut scores resulting from the analysis of the panellists' judgments show that the cut scores may need to be set slightly lower, possibly nearly one score point lower (as indicated in Table 16 below).

Table 16. Combined panel group inter-judge agreement by GEPT level

GEPT writing test level	CEFR level	Existing GEPT pass score	Borderline method score	Contrasting Groups method score
Elementary	A2	4	3.1	3.2
Intermediate	B1	4	3.1	3.3
High-Intermediate	B2	4	3.3	3.4
Advanced	C1	3	2.5	2.4

One possible explanation for the latter finding is that the GEPT rating scale includes an additional criterion of 'task fulfilment', which does not align with any of the criteria on the CEFR Written Assessment Grid. As a consequence, it is possible that some of the writing samples included in the standard setting process were judged by panellists to be at a particular CEFR level in terms of language quality, but that these samples had GEPT scores below the existing pass mark due to failure to meet the task requirements or to adequately address the topic. Moreover, sentence completion, sentence combination, and unscrambling in the Elementary level GEPT writing test and Chinese-English translation in the Intermediate and High-Intermediate levels were excluded from the study because the CEFR does not have scales relevant to the constructs. These tasks, however, contributed to GEPT score associated with writing samples in the standard setting process. These factors represent potential limitations in the alignment process.

In terms of recommendations, how such cut scores would be operationalised depends on the scoring system used by the GEPT. If it is possible to use cut scores with decimal points (because the GEPT uses Rasch analysis and the fair averages could be used), then we recommend that the cut scores are set at the levels determined by the standard setting panels. However, if this is not possible due to operational reasons, we recommend that the GEPT consider lowering the cut scores by half or a full score point to reflect the alignment to the CEFR recommended by the standard setting panel. This may have a number of other operational or policy-related consequences which the GEPT team may have to consider, but these are beyond the scope of this study.

The study made use of two standard setting methods, the borderline method and the contrasting group method. The results of these two methods differed slightly. Across most sub-tests the cut scores for the contrasting group method were slightly lower, although this

was not the case for the Advanced sub-test. If the LTTC decides to change the current cut scores based on the results of this study, a decision needs to be made on which of these cut scores to adopt. One possible method to choose between these two methods would be to estimate how many of the writing samples used as part of this standard setting workshop would be placed differently according to these two methods. This would provide some estimate of the differences in impact of the two possible cut scores.

Comparing the cut scores of the two panels (Melbourne and Taipei), the results show that the judgments of the two groups were fairly similar. The Taipei panel displayed slightly lower inter-judge reliability. This may be explained by the fact that the judges in Taipei participated less in the discussion forum during the familiarisation and benchmarking phases. This fact may have resulted in any differences in opinion not being resolved sufficiently before the standard setting. However, the differences in reliability were not large and the overall combined reliability of the two panels more than acceptable and within expectations for such a study.

References

Angoff, W. H. (1971). *Educational measurement*. R. L. Thorndike (Ed.). Washington, DC: American Council on Education.

Brunfaut, T. & Harding, L. (2014). Linking the GEPT listening test to the Common European Framework of References. Taipei: The Language Training and Testing Center.

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

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. *Setting performance standards: Concepts, methods, and perspectives*, 3-17. New Jersey: Lawrence Erlbaum Associates.

Cizek, G. J. (2012). *Setting performance standards: foundations, methods and innovations, 2nd edition*. New York: Routledge.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A manual.* Retrieved 1 November 2011, from http://www.coe.int/t/DG4/Portfolio/documents/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf

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

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, 17(2), 228-250.

Harvey, A. L. (2000). Comparing onsite and online standard setting methods for multiple levels of standards. New Orleans: Paper presented at the annual meeting of the National Council of Measurement in Education.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis*, 461-475.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17.

Katz, I. R. & Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology*, 15(1), 1-17.

Katz, I. R., Tannenbaum, R. J., & Kannan, P. (2009). Virtual standard setting. *CLEAR Exam Review*, 20(2), 19-27.

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

Milanovic, M. & Weir, C. J. (2010). Series editors' note. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. viii-xx). Cambridge, England: Cambridge University Press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273.

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.

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.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. *Setting performance standards: Concepts, methods, and perspectives*, 19-51.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cut scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Appendices

Appendix A - Specification forms

Specification Forms – A1-A8
Test: General English Proficiency Test (GEPT) – writing

Form A1: General Examination Description

GENERAL EXAMINATION DESCRIPTION	
1. General Information	
Name of examination	The General English Proficiency Test (GEPT) – writing section Levels: Elementary / Intermediate / High-Intermediate / Advanced
Language tested	English
Examining institution	The Language Training & Testing Center (LTTC)
Versions analysed (June 2015)	Elementary (2010, 2012, 2014), Intermediate (2010, 2012, 2014), High-Intermediate (2010, 2011, 2014), Advanced (2008, 2011, 2014)
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 6 million (as at July 2014) 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	Advanced	**45 min**
	☐ **5** Spoken production	_____	_____
	☑ **6** Written production	Elementary	**40 min**
		Intermediate	**24 min (approx.)**
		High-Intermediate	**30 min (approx.)**
			_____
	☑ **7** Integrated skills	Advanced	**60 min**
	☐ **8** Spoken mediation of text	_____	_____
	☐ **9** Written mediation of text	_____	_____
	☐ **10** Language usage	_____	_____
	☐ **11** Other: (specify): _____	_____	_____

6. What is the weighting of the different subtests in the global result?	Elementary (EW): Part one (50%) Part two (50%) Intermediate (IW): Part one (not included in this benchmarking study) (40%) Part two (60%) High-Intermediate (HW): Part one (not included in this benchmarking study) (40%) Part two (60%) Advanced (AW): Part one (50%) Part two (50%)
7. Describe briefly the structure of each subtest	Elementary (EW): 2 parts, 16 items 1. Sentence writing 2. Paragraph writing (picture description – 50 words, approx.) Intermediate (IW): 2 parts, 2 items 1. Chinese-English translation (not included in this benchmarking study) 2. Guided writing: essay on familiar topic or personal experience (120 words, approx.) High-Intermediate (HW): 2 parts, 2 items 1. Chinese-English translation (not included in this benchmarking study) 2. Guided writing: essay on topic related to daily life/current events (150-180 words) Advanced (AW): 2 parts, 2 items 1. Essay based on information provided in two 400 word (approx.) written texts (at least 250 words) 2. Letter to the Opinion Section of a newspaper, based on non-textual information (at least 250 words)

8. What type(s) of responses are required?		Subtests used in (Write numbers above)
	☐ Multiple-choice	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ True / False	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Matching	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☑ Sentence writing	EW1 ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Sentence completion	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Gapped text / cloze, selected response	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Open gapped text / cloze	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Short answer to open question(s)	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☑ Extended answer (text / monologue)	EW2, IW2, HW2, AW1, AW2 ☐ ☐ ☐ ☐
	☐ Interaction with examiner	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Interaction with peers	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
	☐ Other	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
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 are 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: Elementary	**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: 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

Marking: Advanced	**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: Elementary	**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 Elementary Level Writing Tests is guided by National Curriculum Objectives of Junior High Schools in Taiwan. During the development stage, the research committee reached a consensus on the descriptions of the minimum acceptable level of writing proficiency for local junior high school graduates; hence, pilot-version of six-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 writing tests were administered to the sample candidates; a representative sample of the target population was selected from junior high school students; LTTC GEPT Intermediate Level no-pass candidates; and the general public. The writing performances were collected, and benchmark performances for each band score were selected based on the expert judgment and 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.

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 six-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 judgment 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 six-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.

Grading: Advanced	**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 Advanced 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 six-band rating scales (Band 0 to 5) for writing proficiency were developed, and the pass mark was set to be Band 3. In the piloting stage, the pilot-version tests were administered to the sample candidates; a representative sample of the target population was selected from English majors; candidates who took and passed LTTC GEPT High-Intermediate Level tests; and the native speakers. 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 the rating scale for use in the training of new raters 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 seven days 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 in 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: Advanced: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/Advanced.htm 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 Elementary: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/elementary.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		
IW: A1	**IW: B1**	**AW: C1**
☐	☐	☐
EW: 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 and Speaking: in progress

Appendix B – Questionnaires

Feedback questionnaire – Preparatory and Familiarisation sessions

	Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1	I have a good overview of the CEFR and the GEPT.					
2	The sessions were clearly explained and I understood what I was being asked to do.					
3	I was able to follow instructions and complete the activities.					
4	The information and activities were useful.					
5	The relevant CEFR levels and descriptors are clear to me.					
6	I found the PowerPoint slides and supplementary files easy to work with.					
7	I was able to complete the sessions within the suggested time required.					

Overall, I found the sessions…

Any other comments:

Feedback questionnaire – Benchmarking training session

	Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1	The session was informative and clearly presented.					
2	The illustrative examples helped me understand the CEFR levels.					
3	The CEFR descriptors were relevant to the GEPT writing samples.					
4	I was able to follow instructions and complete the practice judgment form as required.					
5	The discussion forum was helpful.					
6	I feel familiar with the CEFR descriptors.					
7	I feel familiar with the GEPT writing tasks.					
8	I was able to complete the session within the suggested time required.					

Overall, I found the session…

Any other comments:

Feedback questionnaire – Writing standard setting (Stage two)

	Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	
1	The PowerPoint presentation slides provided me with a clear understanding of the purpose of the standard setting process.						
2	The process instructions were clear and I understood what I was being asked to do.						
3	Reviewing the writing tasks helped me understand the assessment.						
4	The relevant CEFR levels and descriptors were clear to me.						
5	The CEFR descriptors were relevant to the GEPT writing scripts.						
6	I was able to follow instructions and complete the rating sheets as required.						
7	Information showing how my judgments differed from the judgments of other participants was helpful.						**N/A**
8	I am confident about the defensibility and appropriateness of the decisions I made about the CEFR levels of the writing samples.						
9	I was able to complete the session within the suggested time required.						

Overall, I found the stage two sessions…

Any other comments:

The Language Training and Testing Center (LTTC)
No.170, Sec.2, Xinhai Rd., Daan Dist.,
Taipei City, 10663 Taiwan(R.O.C.)
Tel: +886-2-2377-8071
Email: geptgrants@lttc.ntu.edu.tw
Website: www.lttc.ntu.edu.tw