Linking the GEPT Listening Test to the Common European Framework of Reference

LTTC-GEPT Research Reports
RG-05

Tineke Brunfaut
Luke Harding

This study was funded and supported by the Language Training & Testing Center (LTTC) under the LTTC-GEPT Research Grants Program 2012-2013

LTTC-GEPT Research Reports RG-05
Linking the GEPT Listening Test to the Common European Framework of Reference

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 2014

Foreword

We have great pleasure in publishing this report: *LTTC-GEPT Research Reports RG-05*. The study described in this report was funded by the 2012-2013 LTTC-GEPT Research Grants. Headed by Dr. Tineke Brunfaut and Dr. Luke Harding of Lancaster University, UK, the study adopted an innovative asynchronous twin-panel approach and followed the procedures set out in the *CEFR Manual* to map the GEPT Listening 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 Tineke Brunfaut is a Lecturer in language testing at Lancaster University, UK. Her main research interests are language testing, and reading and listening in a second or foreign language. She has conducted research on factors that affect second language listening task difficulty, published in *Studies in Second Language Acquisition* (Révész and Brunfaut, 2013) and *TESOL Quarterly* (Brunfaut and Révész, forthcoming). Recent research on diagnosis in second and foreign language assessment is reported in the journal *Applied Linguistics* (Alderson, Brunfaut and Harding, 2014). Her recent work furthermore includes publications in *Language Assessment Quarterly* (Brunfaut, 2014; Brunfaut, forthcoming). She is the Director of Lancaster University's online *MA in Language Testing* and teaches post-graduate and short courses on language testing and applied linguistics topics, at Lancaster and abroad. She carries out consultancies in language test development and evaluation, working with ministries, universities, and examination boards in a variety of countries.

Dr Luke Harding is a Lecturer in the Department of Linguistics and English Language at Lancaster University. His research interests are in language testing, particularly in the areas of listening assessment, pronunciation and intelligibility, assessor decision-making, assessment literacy and the challenges of World Englishes and English as a Lingua Franca for language testing and teaching. His research has appeared in international journals such as *Language Testing*, *Language Assessment Quarterly* and *Applied Linguistics*, and he has published a book - *Accent and listening assessment* - through Peter Lang (2011). Luke has a background in test development and English language teaching, and carries out consultancies for educational organisations and private language testing companies.

摘要

◆ **研究團隊與研究目的**

本研究由英國蘭卡斯特大學（Lancaster University）Dr. Tineke Brunfaut 與 Dr. Luke Harding 主持，依循 *Relating Language Examinations to the CEFR: A Manual*（Council of Europe, 2009）的步驟—包含 familiarization（熟悉 CEFR 分級）、specification（審視測驗品質與內容和 CEFR 級數的關聯）、standard setting（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—由異地、不同背景的兩個專家小組（twin-panel），判斷全民英檢初級至高級聽力測驗對應 CEFR 的級數，同時也比較參照研究常用研究方法的優劣。

◆ **研究問題**

1. GEPT 聽力測驗與 CEFR 的關聯性為何？
2. 何者是參與本研究標準設定（standard setting）的測驗與教學專家所認為最適合用來進行本聽力測驗與 CEFR 參照研究的方法，"basket method"或"modified Angoff method"？
3. 雙專家小組法（twin-panel approach）的優點與缺點為何？

◆ **研究方法摘要**

1. 測驗內容分析階段（specification）由兩位研究者進行，分析 GEPT 各級聽力測驗內容，並據此判斷每題對應的 CEFR 級別，即 A1、A2、B1、B2、C1、C2。
2. 標準設定階段（standard setting）由不同背景的 12 位教學和測驗專家所組成的兩個專家小組分別在英國和台北進行。台北組較熟悉全民英檢，英國組較熟悉 CEFR，但對全民英檢的了解較少。每位專家依試題內容與 CEFR 聽力能力說明判斷 GEPT 聽力測驗每題的 CEFR 級別，並進一步區分試題的難度在該級中屬於難(H)、中等(M)、或易(L)。英國組於 102 年 7 月進行；台北組則於同年 10 月進行。

◆ **研究結果摘要**

1. 測驗內容分析(specification)與標準設定(standard setting)的結果都顯示，GEPT 初、中、中高級聽力測驗難度分別相當於 CEFR A2、B1、B2 級，但 GEPT 高級聽力測驗的難度則相當 B2+到 C1。
2. 研究者建議中高級和高級可加入更多真實聽力材料的元素(如多樣化的口音、背景聲音)、以及更接近真實的說話方式（如說話中重複字詞或自我修正），以提高聽力測驗真實度(authenticity)。
3. 專家認為 basket method 與 modified Angoff method 各有優劣，而多數專家表示集合兩者的優點而成的 modified basket method(即本研究所使用的方法)—除判定 CEFR 級別外，另需將試題難度細分為難(H)、中(M)、易(L)三種等級—對難度判定的過程更有幫助，是很實用的方法。
4. 雖然雙專家小組在判定試題難度的過程稍有差異，但所提供的判定結果仍相近。此外，此設計能讓參照研究能同時考量不同背景專家的意見與看法，且交叉驗證兩個專家小組的判定結果。

Abstract

This document reports on a linking study designed to provide empirical evidence on which to base claims about the connection between the listening test suite of the General English Proficiency Test (GEPT) and the Common European Framework of Reference (CEFR). The investigation was guided by the recommended methods and procedures set out in the manual *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe, 2009), and entailed the stages of familiarisation, specification, standardisation and empirical validation. In addition, the project involved an innovative research design marked by a) a pilot stage to evaluate the suitability of two different standard setting methods for a linking study with listening test materials (the basket method and the modified Angoff method), and b) an asynchronous "twin-panel" approach. This methodology allowed for the inclusion and comparison of both insider and outsider perspectives through a panel of six language testing/teaching/standard setting experts without prior knowledge of the GEPT suite and a separate panel of six teachers/researchers/developers with an intimate knowledge of the GEPT suite. Preliminary alignments and cut scores were established by the researchers at the specification stage.

Informed by the pilot study, a two-step adapted version of the basket method was developed which requires the standard setter to assign each item to a particular CEFR "basket" and, in addition, to decide whether a just-qualified, a mid, or a high test-taker at the particular CEFR level would already be able to answer the item correctly. After a familiarisation process with the CEFR, the GEPT, and the standard setting procedures, this modified basket method was used by the panels to judge the CEFR level of GEPT listening items. Validation checks were realised through judgement reliability analyses, judgement comparisons across panels, and qualitative analysis of recorded panel discussions.

Based on the full linking study–including individual and combined panel data analyses–recommendations are made for the alignment of the GEPT listening suite with the Common European Framework of Reference. In addition, a set of suggestions is formulated in order for the GEPT listening tests to be more easily articulated with the CEFR descriptors, and in so doing to increase the validity of alignment to the framework.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1. Overview

The aim of this research project was to conduct a linking study, relating the listening tests of the General English Proficiency Test (GEPT) to the Common European Framework of Reference (CEFR). This linking study was guided by the recommended methods and procedures set out in the manual *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe, 2009). Following the manual, the study involved four stages: familiarisation, specification, standardisation and empirical validation. The project was designed to provide empirical evidence on which to base claims about the connection between performance on the GEPT listening test suite and the CEFR.

1.2. Linking language tests to the CEFR

The Common European Framework of Reference (CEFR) is a common framework of reference for language teaching and learning. The CEFR consists of six reference levels across three bands: A1-A2 (basic user), B1-B2 (independent user) and C1-C2 (proficient user). Language proficiency is described at these levels across a range of skills, including reading, writing, listening and speaking, with full sets of scales provided in the seminal publication *Common European Framework of Reference for Languages: Learning, teaching and assessment* (Council of Europe, 2001).

Since its inception the CEFR has become enormously influential in language testing practice and research, not only in the European context, but also on a global scale. The scales have come to function as a set of de facto external language standards in many contexts, and have facilitated benchmarking and evaluation across a range of testing contexts (sometimes unsuitably, see Milanovic & Weir, 2010). Following their introduction there has been growing interest among test providers in "linking" or "aligning" their exams to the CEFR, and as a result of this the Council of Europe piloted a set of recommended procedures in 2003 for mapping language tests to the Framework (the results of many pilot studies have been published in the collection *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* edited by Martyniuk, 2010). These procedures were formalised in the 2009 document released by the Council of Europe *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching and assessment. A Manual* (henceforth referred to as the *Manual*).

The *Manual* recommends that a linking study should be conducted in four stages: (1) a "Familiarisation" stage, (2) a "Specification" stage, (3) a "Standardisation" stage, and (4) a "Validation" stage. Details on each stage are provided below:

Familiarisation: A panel of judges participate in a range of activities designed to make sure that they have a sufficiently deep knowledge of the CEFR scales and descriptors relevant to the linking purpose.

Specification: Key participants audit the coverage of the exam under investigation, and create a profile of how the content of texts and tasks relate to CEFR.

Standardisation: A panel of judges take part in standard setting during which the claims for linking to the CEFR are substantiated.

Validation: A series of internal and external "checks" on the linking procedure are undertaken to strengthen the linking claim.

These four stages of the *Manual* functioned as an overarching framework for the design of the linking study (see Research Design, below).

1.3. The GEPT

The General English Proficiency Test (GEPT) is developed and administered by the Taiwan-based Language Training and Testing Centre (LTTC), and its scores are recognised by a large number of institutions (in Taiwan, but increasingly also abroad). It is a suite of English language proficiency exams which consists of five levels, each assessing all four skills (reading, writing, listening and speaking). At one of these five levels, the superior level, listening is assessed in an integrated manner, in combination with writing and speaking. The present study, however, focuses on the remaining four levels – Advanced, High Intermediate, Intermediate, and Elementary – at which listening is assessed in an isolated manner, in accordance with the level descriptions as presented in Table 1 (*all information retrieved from http://www.lttc.ntu.edu.tw).

Table 1: GEPT level descriptions for listening

GEPT level	**Skill-area level descriptions for listening***
Advanced	An examinee who passes this level can understand conversations on all sorts of topics as well as debates, lectures, news reports, and TV/radio programs. At work, when attending meetings or negotiations, he/she can understand reports and discussions.
High Intermediate	An examinee who passes this level can understand conversations in social settings and grasp the general meaning of lectures, news reports, and TV/radio programs. At work, he/she can understand brief reports, discussions, product introductions, and operating instructions.
Intermediate	An examinee who passes this level can understand general conversation in daily life situations and grasp the general meaning of public announcements, weather forecasts, and advertisements. At work, he/she can understand simple product introductions and operating instructions. He/she can catch the general meaning of native English speakers' conversations and inquiries.
Elementary	An examinee who passes this level can understand simple conversation related to daily life on such topics as prices, time, and places.

As shown in Table 2, the listening tests include a range of tasks and several items at the different levels (*all information retrieved from http://www.lttc.ntu.edu.tw).

Table 2: GEPT listening tasks per level

GEPT level	Listening parts & task types	Number of items	Time (mins.)*
Advanced level	1 – Short conversations & talks 2 – Long conversations 3 – Long talks	40	45
High Intermediate	1 – Answering questions 2 – Conversations 3 – Short talks	45	35
Intermediate	1 – Picture description 2 – Answering questions 3 – Conversations	45	30
Elementary	1 – Picture description 2 – Answering questions 3 – Conversations 4 – Short talks	30	20

1.4. Research questions

The overarching research question which guided the linking study was:

RQ1. How do the GEPT listening test levels relate to the CEFR?

In breaking down this broad question, specific aims guided each stage of the linking study, and these will be detailed in the sections below. However, the project also involved an innovative research design marked by a pilot stage to evaluate the suitability of different standard setting methods for this context, and an asynchronous "twin-panel" approach at the Standardisation stage to provide a wider range of perspectives on the relationship between listening test materials and CEFR levels. As a result, two further research questions – of interest to the broader field – were also addressed in the study:

RQ2. Comparing the "basket method" and the "modified Angoff method", which is perceived by standard setters to be the most suitable for standard setting a suite of listening tests?

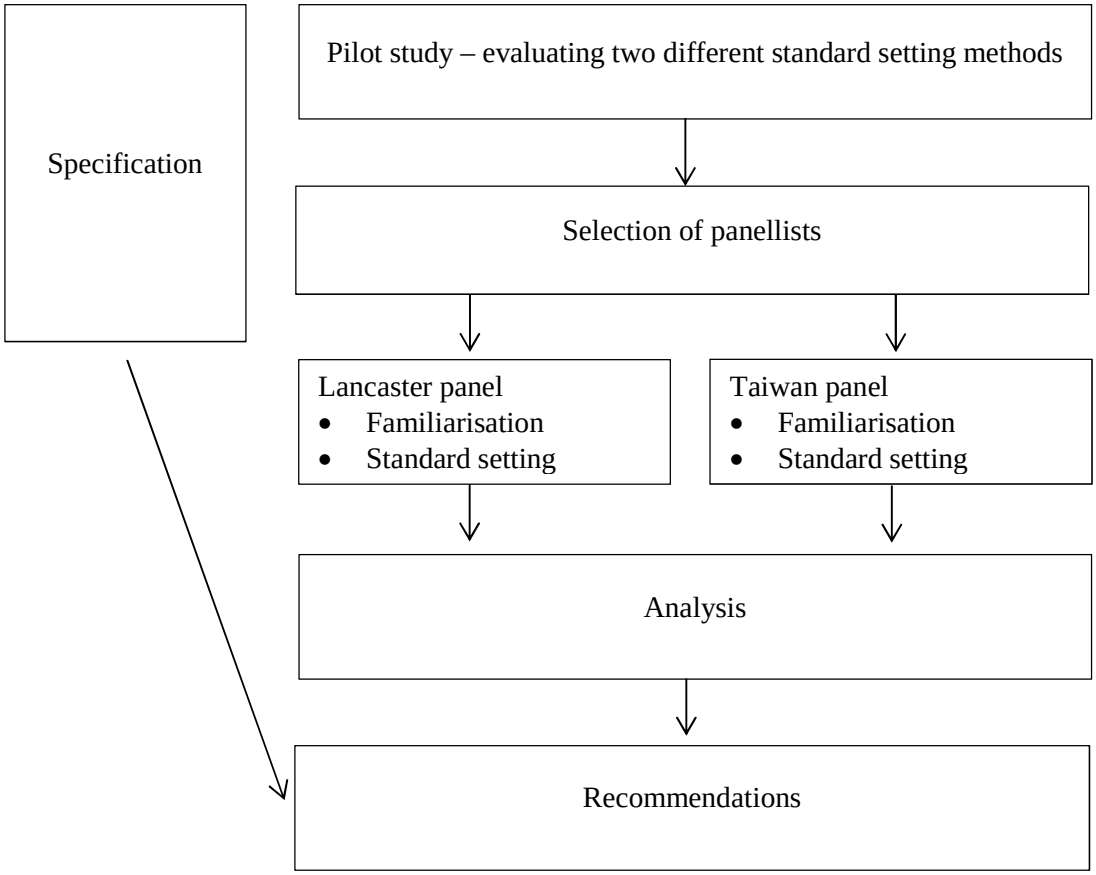
RQ3. What are the strengths and weaknesses of a twin-panel linking approach?

2. Research Design

The research design followed the four-stage design which has been recommended in the (2009) *Manual*: Familiarisation, Specification, Standardisation and Validation. Within these stages, the linking study followed procedures which have been recommended in the *Manual*; however, it also drew upon innovations in standard setting research which have been proposed in recent literature (e.g., Tannenbaum, 2011). A key feature of the study was that it involved "twin" standard setting panels conducted in Lancaster and in Taipei. The advantages of the twin panel approach were estimated to be that (a) panels would include both language testing/standard setting experts as well as teachers/researchers/developers who have an intimate knowledge of the GEPT suite (providing both outsider and insider perspectives), and (b) the twin panels would provide a means of cross-validating the decisions of each panel (facilitating an in-built means of performing an external check at the Validation stage) (see Hambleton, 2001).

Figure 1 provides an overview of the various stages of the present study. This diagram shows that the specification stage – a form-filling exercise which relied upon expert judgement – was conducted separately from those aspects of the study which involved participant judges: the pilot study, and the twin panels which undertook familiarisation activities and then performed standard setting. At the analysis stage, internal and external validity checks were also performed. Finally, the outcomes of the linking study will be reported, and a number of recommendations will be made at the end of this document.

Figure 1: Overview of the research design stages



3. Specification

The specification stage was conducted independently by the researchers, using the forms made available in the *Manual*. The completed forms are included in Appendix A of this report. A considerable proportion of the information in Forms A1-A8 was completed by staff from the Language Training and Testing Centre. More detailed information was not available to the researchers because, presumably, this information is confidential and primarily shared within the test development team. However, sufficient information was available to be able to make estimates about representative tests from each of the four levels of the GEPT suite under focus.

Prior to undertaking the specification stage analyses, the researchers (who have 6-10 years of experience of working with the CEFR and 2-3 years in standard setting) re-familiarised themselves with the CEFR and the listening descriptors in particular, and also with sample listening items made available by the Council of Europe (2005) and illustrating the common reference levels. In addition, the researchers acquainted themselves with all information available on the GEPT exam.

Of primary importance to this project, during the specification stage an initial estimate of the CEFR levels was made, and this was then reviewed based on more specific information provided about the GEPT listening tests, and then a detailed analysis of test materials. The *initial* estimate of the CEFR level of each of the four tests is shown in Table 3.

Table 3: Initial estimate of alignment between GEPT levels and CEFR levels

GEPT level	**CEFR level**
Elementary	A2
Intermediate	B1
High intermediate	B2
Advanced	C1

This initial estimate was based on information that was available in the public domain concerning the GEPT, which had been derived from studies of the reading section of the exam (Wu & Wu, 2010; Wu, 2011) and from a GEPT/IELTS comparability study (Weir, Chan & Nakatsuhara, 2013).

A second step in the specification stage was to match information gleaned from GEPT listening test specifications (in the public domain) with descriptors from the CEFR Overall Listening Comprehension descriptors. As a result of this analysis, the hypothesised CEFR levels were revised. Table 4 shows the revised estimates (detailed justifications are provided in Appendix A, Form A9).

Table 4: Revised estimate of alignment between GEPT levels and CEFR levels

GEPT level	**CEFR level**
Elementary	A2
Intermediate	A2/B1
High intermediate	B1/B2
Advanced	B2/C1

The final step in the specification stage concerned a detailed content analysis of a test version of each of the four GEPT levels under focus (see below for more details on the test versions). This process involved entering information about each task, text and item into a grid by specifying a range of options derived from the CEFR (see Appendix A, section D). During this process, a number of observations about the nature of the items and the ease with which these could be related to CEFR descriptors were also noted (see Appendix A, section C, notes per test level).

As a result of this analysis, a preliminary set of cut scores was established, as specified in Table 5.

Table 5: Preliminary cut scores

GEPT level	**Preliminary cut scores**			
Elementary	<A1: 0	A1: 1-18	A2: 19-30	
Intermediate	<A2: 0	A2: 1–24	B1: 25-45	
High intermediate	<B1: 0	B1: 1-18	B2: 19–42	>B2: 43-45
Advanced	<B2: 0	B2: 1–15	C1: 16-39	>C1: 40

The specification stage analyses suggested that the Elementary GEPT listening level can be aligned with CEFR A2 level, and that the Intermediate, High Intermediate and Advanced levels can be situated at the borderline of the adjacent CEFR levels A2/B1, B1/B2, and B2/C1 respectively. The next step in the linking process concerned the verification of these preliminary alignments through a larger-scale standard setting activity.

4. Standard Setting

4.1. Pilot study

4.1.1. Overview

Given the large range of standard setting procedures available (see e.g. Cizek & Bunch, 2007 for an overview), prior to the standardisation phase a pilot study was undertaken to select the method for standard setting of the GEPT listening tests. In particular, the suitability and efficiency of the "basket method" and the "modified Angoff method" to standard setting of listening tests was evaluated through a comparative study. Both methods have been perceived as useful for judging listening items (see e.g. Kecker & Eckes, 2010 for the "basket method" and Tannenbaum & Wylie, 2008 for the "modified Angoff"). However, as described by the *Manual* (Council of Europe, 2009), each standard setting method has its advantages and disadvantages in terms of accuracy and ease of use.

The "basket method" to standard setting requires judges to estimate the difficulty of test items according to CEFR levels (thus placing individual items into a "basket"). By averaging judges' decisions across each test, a set of standards can be derived. The "modified Angoff", on the other hand, requires judges to estimate the probability that a "minimally competent" test-taker would answer an item correctly. The sum of each judge's estimations across test items is then taken to represent a cut score, and these individual cut scores are then averaged to find the panel's overall cut score. Usually judgements are made in several rounds, with discussion encouraged in order to reach consensus as a group.

4.1.2. Participants

To evaluate the usefulness of these two methods for linking the GEPT listening tests to the CEFR, a pilot standard setting activity was run. The participants were five members of the Language Testing Research Group at Lancaster University, all of whom were language assessment specialists familiar with the CEFR. Three had live-test standard setting experience; the two other participants had standard setting experience within a research context. Two participants were English native speakers, three were expert users of English as an additional language. All pilot study participants were based in Lancaster (UK), but they had European, Middle Eastern, Asian and Australian backgrounds.

4.1.3. Procedures

The comparative standard setting pilot study was conducted in three stages. The first stage involved a familiarisation activity similar to the one described in detail below for the main study. In short, participants were familiarised with the research project, the exam suite under investigation, and the CEFR. Most time was dedicated to familiarisation activities on the listening scales of the CEFR. In addition, participants' written consent was sought at this stage.

The second and third stage focussed on one standard setting method each; first the basket method was trialled on a sample of GEPT listening items, than the modified Angoff method. At the start of each session, the participants were given information on the specific standard setting method used during the session and were familiarised with the relevant standard setting activities and instruments. Each session was led by one of the two main investigators. The items were sampled from the different sections of a version of the High Intermediate GEPT listening test, which represents the mid-range proficiency level of the full GEPT exam suite. The High Intermediate test has been suggested to be situated at CEFR B2 level (http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/hi_intermediate.htm).

For the basket method, the participants were asked to consider the question 'At which CEFR level can a test-taker already answer the item correctly?' and to provide their CEFR-level judgement of each item on a judgement sheet designed for this purpose. The modified Angoff method required the participants to consider a different set of questions and to make three types of judgements. First, they were asked to decide whether a just-qualified B2 learner can answer the item correctly. To answer this question, the participants had to circle 'Yes' or 'No' on a judgement sheet designed for this purpose. Secondly, they had to judge the probability that a just-qualified B2 learner will answer the item correctly. This judgement had to be indicated on a 0-100 probability scale with 10% increments (except for one 5% increment at the lowest end of the scale). Thirdly, the standard setters were asked to report their level of confidence about their first judgement ('Can a just-qualified B2 learner answer this item correctly?') on a 0-100 scale.

4.1.4. Findings

The discussions held during both pilot standard setting activities were recorded, and the pilot study was rounded-up with a general discussion on the user-friendliness of each method from the standard setters' perspective. Although overall the participants found both methods workable, they had an outspoken preference for the basket method. This approach was considered much more straightforward and thought to put less cognitive strain on the standard setter. Since the main study was scheduled to involve several days of standard setting of four different levels of the GEPT exam suite, the participants expressed concerns over the more demanding nature of the modified Angoff method. They believed this method may be associated with higher fatigue risks, negatively affecting judgement consistency. At the same time, the researchers observed the time-consuming nature of the modified Angoff method as compared to the basket method, i.e. 35% more time was needed to judge the same number of items using the modified Angoff method. On the other hand, the judges appreciated the more detailed probability-correct judgement of the modified Angoff method. With the basket method, they felt they missed being able to further qualify their judgement, particularly when items were considered to be borderline cases of adjacent levels.

4.1.5. Modified basket method

Taking into consideration the pilot study participants' views, the practicalities of the main study, and their own expert evaluations of the overall usefulness of both methods, the investigators decided to opt for the basket method for the main study. However, to accommodate for some of the comments made by the participants, a modified basket method which allows for a second, refining judgement was developed by the researchers. The adapted version does not only require the standard setter to assign each item to a particular CEFR 'basket', but also to decide whether a just-qualified, a mid, or a high test-taker at the particular CEFR level would already be able to answer the item correctly. This modified basket method was later presented to the pilot study participants for feedback, and was informally tried out by the researchers on a few GEPT items. The method was positively evaluated by all parties, and thus it was decided to adopt it for the main study.

4.2. Main study panellists

Having completed the pilot study and selected a suitable method, panellists were recruited for the main standard setting study. The panel consisted of 12 people in total. As explained earlier, a twin panel approach was followed, whereby two groups of participants conducted the standard setting activity independently of one another. One group of six did the standard

setting at Lancaster University in Lancaster (UK). The other group of six carried out the activity at the Language Training and Testing Centre in Taipei (Taiwan).

The Lancaster participants were all based in European countries and travelled to Lancaster from within the UK (two participants) or abroad (four participants) for the purpose of the standard setting activity. Two of the participants were native speakers of English. The first language backgrounds of the other four participants were German, Greek, Sinhala, and Slovene. Participants on the Taipei panel were all based in Taiwan. One participant was an English native speaker. The other five participants had a Chinese first language background.

Detailed information on the panellists' language teaching, language testing, and standard setting experience is presented in Table 6– for the two individual panels as well as the panel as a whole. Data is provided on the panels' length and type of experience, as well as the languages in which the participants have teaching, testing or standard setting experience.

Table 6: Standard setting participants' personal background experience

	Lancaster panel	**Taipei panel**	**Full panel**
Language teaching experience			
Length (years)	Range = 2 - 37 M= 19.3	Range = 3 – 20 M = 12.5	Range = 2 - 37 M = 15.9
Languages (number of participants)	English (6) German (1)	English (6) Mandarin (1)	English (12) German (1) Mandarin (1)
Type (number of participants)	Primary education (4) Secondary education (5) Higher education (6) Adult education (4)	Primary education (1) Secondary education (2) Higher education (4) Adult education (2)	Primary education (5) Secondary education (7) Higher education (10) Adult education (6)
Language testing experience			
Length (years)	Range = 2 - 18 M = 11	Range = 2 – 20 M = 8.1	Range = 2 - 20 M = 9.55
Languages (number of participants)	English (6) French (1) Greek (1) Italian (1) Spanish (1)	English (6)	English (12) French (1) Greek (1) Italian (1) Spanish (1)
Type (number of participants)	examiner/rater (6) item writer (6) researcher (1) test developer (6) trainer (2)	examiner/rater (6) item writer (5) test developer (5)	examiner/rater (12) item writer (11) researcher (1) test developer (11) trainer (2)
Standard setting experience			
Length (years)	Range = 1 - 6 M = 3.4	Range = 0 – 10 M = 2.8	Range = 0 - 10 M = 3.1
Languages (number of participants)	English (6) French (1) Greek (1) Italian (1) Spanish (1)	English (2)	English (8) French (1) Greek (1) Italian (1) Spanish (1)
Type (number of participants)	participant (5) organizer (2)	participant (1) organizer (1)	participant (6) organizer (3)

The standard setting sessions were led by three language testing experts from Lancaster University and assisted by one English language teacher. The Lancaster panel was led by the two project researchers. One researcher was in control of the procedural aspects of the panel and primarily managed the panel discussions; the other researcher was mainly responsible for data entry and data reporting during the standard setting activity. Both convenors have extensive experience of working with the CEFR (10 and 6 years, respectively) and have participated in and led standard setting panels (3 and 2 years, respectively).

The Taipei standard setting sessions were run by the third language testing expert, who overlooked, led and managed the entire process. The expert was assisted by a language teacher for data entry. The Taipei convenor has worked with the CEFR since the framework's conception, and has participated in and led standard setting panels for 15 years.

4.3. Schedule of panels

4.3.1. Schedule

Standard setting was conducted over a period of four days, Monday to Thursday. The dates were determined by the availability of the different panellists and the convenors. The Lancaster-based sessions were held from 15th until 18th July 2013. The Taipei-based sessions took place during the period 21st - 24th October 2013. These were full-day sessions, from 9am to 5pm, but with regular breaks to avoid fatigue effects. The first half day was dedicated to introductions and familiarisation activities (see below for details). After this, the judgment process was started and continued until the end of the third day (see the section on Judging Procedures for further details). Judgements were made on four listening tests, covering the different GEPT levels under focus in the present study. To ensure high concentration levels, a small set of items was taken at a time (5 to 15 items, organized per task type/test section). On the fourth day, the panellists were presented with a mixed set of items sampled from test versions of all four GEPT levels. This was undertaken as part of the standard setting validation checks (see the Analysis section below for further details). The closing activity of the standard setting week consisted of a final, plenary discussion in which the panellists were given the opportunity to share their overall thoughts on the entire standard setting experience and process (see the Qualitative Findings section for further details). The full timetable of the standard setting sessions is provided in Appendix B.

4.3.2. Test materials

For each of the four test levels under focus, one complete listening test was selected from a set of three test versions (per test level) that had been made available to the researchers by the Language Training and Testing Centre. These four tests were chosen more or less randomly from the pool of test papers; all three test versions of each test level (and in fact all twelve papers in the pool) showed a similar spread of facility values, almost the exact same mean test difficulties and standard deviations, and contained the same item types.[1] Judgements were made on all items of each of the four tests (i.e., ranging between 30-45 items, depending on the test level). The reliability coefficients of the listening test versions used in the study ranged from .81 to .87.[2]

[1] Note that for reasons of confidentiality the actual facility range, mean and SDs cannot be reproduced in this report.

[2] The reliability indices provided to the researchers by the LTTC are based on performance data from live administrations.

The mixed set of items judged on the last day was sampled from one of the two remaining test versions per level (7 Elementary, 10 Intermediate, 9 High Intermediate, and 8 Advanced level items were extracted), representing the various test levels and task types. These items were randomly ordered, and level-identifying information was removed.

4.4. Familiarisation stage

4.4.1. Overview

The first stage of the linking process (as described in the *Manual*) consists of training activities to familiarise participants with different aspects of the linking activity. Therefore, a range of familiarisation activities were undertaken during which participants were familiarised with the research project, the exam suite under investigation, the CEFR, and the standard setting activities and instruments. The first half day of the standard setting activity was dedicated to this stage.

Given the differences in expertise between the participants at the two locations (see participant descriptions), different elements of the familiarisation activities were covered more or less extensively depending on the panel. For example, although CEFR familiarisation activities were conducted with both panels, more extensive CEFR familiarisation was conducted with the Taipei panel. The Lancaster panellists all have extensive experience in working with the CEFR in their day-to-day language teaching, testing and standard setting contexts, whereas the Taipei participants had more limited or no experience of working with the CEFR. The Taipei panellists, however, were familiar with the GEPT exams, the target population, and the Taiwanese language learning and testing context. The Lancaster panel had no prior knowledge of the exam suite or the main test context, and therefore more time was taken to introduce the GEPT exams to this group.

The following familiarisation activities were undertaken:

4.4.2. Introduction to the standard setting project and the exam suite

By means of a PowerPoint presentation, the participants were explained the general aim of the research project, i.e. to perform a linking study to relate the GEPT listening exams to the CEFR. A description was given of the different procedures typically followed during standard setting.

In addition to this, the participants were provided with a general description of the GEPT exam suite. Information was given on the overall purpose, population, and use of the GEPT, and on the test development institution. The content of the exams and the different exam levels were also described.

4.4.3. Familiarisation with the CEFR

Theoretical as well as practical CEFR familiarisation activities were conducted. First, the convenors provided the panellists with general information on the CEFR. They explained the nature of the framework, how it is used, and what its aims are. They also described the CEFR's six different proficiency levels. Next, all participants were given a CEFR familiarisation activity, focusing on listening. A set of randomly ordered, edited CEFR descriptors of salient characteristics of the CEFR illustrative listening scales was given to the panellists. They were asked to decide which CEFR level each descriptor represented. The solutions were then shared with the panellists and deviating descriptor allocations were

discussed. The panellists were furthermore asked to carefully review and consider the features of those descriptors which they had wrongly allocated to a particular CEFR level.

Following this activity, the panellists were provided with the CEFR illustrative scales for listening. More specifically, they were asked to carefully read through and reflect on the following scales:

- The Common Reference Level global scale (Council of Europe, 2001, p.24)
- The illustrative scales for aural receptions: Overall Listening Comprehension (Council of Europe, 2001, p.66), Understanding Conversation Between Native Speakers (Council of Europe, 2001, p.66), Listening As A Member of a Live Audience (Council of Europe, 2001, p.67), Listening to Announcements And Instructions (Council of Europe, 2001, p.67), Listening to Audio Media And Recordings (Council of Europe, 2001, p.68),
- Illustrative scale for audio-visual reception: Watching TV And Film (Council of Europe, 2001, p.71)

Given their more limited experience of the CEFR and linking exams to the CEFR, the Taipei-based panellists conducted an additional familiarisation activity. They were given nine listening tasks that have been standard set and have been made available by the Council of Europe specifically for familiarisation purposes (Council of Europe, 2005). The panellists were asked to judge the CEFR level of each listening task and to justify their judgements with reference to the CEFR listening descriptors. A group discussion was held, during which the standard set level was revealed and the rationale of the original standard setters was shared and discussed with the panellists.

4.4.4. Familiarisation with the standard setting process

The final steps in the familiarisation phase involved a description of the standard setting procedures and instruments. The convenors described the typical four-stage process of linking exams to the CEFR, and the activities that would be undertaken during this specific standard setting. A timetable for the standard setting sessions was shared with the participants, and they were talked through the different activities planned for the week and the general procedures that would be followed. The panellists were also provided with a copy of the judgement sheet they would be using to note down their CEFR judgements, and the convenors explained how the sheet had to be completed.

4.4.5. Ethical procedures and consent

In addition to having been orally explained the standard setting aims and procedures, the participants were provided with a written information sheet detailing the nature of the project, their involvement, and contact details of the researchers and their Head of Department. Furthermore, a set of rules concerning confidentiality issues was agreed upon with the participants. Written consent to participate was obtained from all panellists.

4.5. Judging procedures

At each research site – Taipei and Lancaster – judgements were provided according to the following identical procedures:

1. Judges were provided with a copy of the test (with level-identifying information removed), and a judgement sheet for that particular set of materials (see Appendix C

for a sample judgement sheet). Judges retained the CEFR scales they had been given during the familiarisation stage. Each judge was also assigned a unique judge code (a single number); the connection between each judge and their code was known only to the judge.

2. The panel convenor played the audio recording for a specific test section, and judges were instructed to answer the items themselves (to experience the test from the perspective of a candidate).
3. The panel convenor then provided the correct answers to each item so that judges could check their answers.
4. The panel convenor then played the audio for a second time, and judges were asked to provide their initial judgements for that section of the test. Judges were encouraged to consult the CEFR descriptors as much as possible at this stage, to ensure judgements were directly linked (and could be justified) with specific statements from the CEFR. Judgements were provided in numeric format so that:

   A1 = 1
   A2 = 2
   B1 = 3
   B2 = 4
   C1 = 5
   C2 = 6

   Judges were also asked to circle L (low), M (mid) or H (high) in a box to the right of each item to indicate whether a "just-qualified", "mid" or "high" level test-taker within that level would be able to answer the item correctly (see Appendix C).
5. Once panel members had provided their judgements, papers were collected (this often coincided with a short break for panellists). Data were entered into an Excel spreadsheet which showed items in rows and judges' codes in columns. At this stage, only CEFR level judgements were entered (i.e., the "low", "middle" and "high" ratings were not shown owing to the difficulty involved in entering that amount of data in the available time). Aberrant ratings – those judgements which differed from the majority – were highlighted to facilitate discussion.
6. Judges were then shown their first-round ratings, which were projected onto a screen. The panel convenor talked through each item, asking – where ratings were split – for a justification of each level according to the CEFR descriptors. Facility values, made available by the LTTC, were also revealed to judges to provide additional, empirical input to consider. The panel convenor was careful not to influence judgements by voicing his/her own views about particular items. This discussion, always conducted in plenary, would ordinarily take 30 minutes, and was recorded by a digital audio device.
7. Once discussion had finished, rating sheets were re-distributed and judges were asked to revise their ratings – if they desired – by crossing out any judgements they wished to change and marking their new judgements clearly. Judges were also able to revise low, middle and high judgements at this stage.
8. Second-round judgements were collected and entered into the Excel spreadsheet.

This procedure was followed for each test section on each of the four level tests used in the standard setting study.

This procedure was also adhered to when judging a mixed-order set of items at the end of the standard setting week.

5. Analysis

5.1. Quantitative findings

Judgements were analysed for the Lancaster panel alone, for the Taipei panel alone, and for the combined panels. In analysing the judgements, the low, middle and high (henceforth LMH) judgements – indicating "compartments" of a basket – were used to create a continuous scale which provided a more nuanced view of each panellist's judgements, such that:

A1 low = .67
A1 mid = 1
A1 high = 1.33
A2 low = 1.67
A2 mid = 2
A2 high = 2.33
B1 low = 2.67
B1 mid = 3
B1 high = 3.33
B2 low = 3.67
B2 mid = 4
B2 high = 4.33
C1 low = 4.67
C1 mid = 5
C1 high = 5.33
C2 low = 5.67
C2 mid = 6
C2 high = 6.33

The rationale for this scale was that, if each CEFR level could be conceptualised as containing low, mid and high compartments, then the best way of classifying judgements at each compartment would be to divide the level into three equal parts and select the mid-point of each of these parts as the scale point. This is illustrated in Table 7 below for two levels only: A1 and A2.

Table 7: Scaling of low, mid and high judgements

.50 - .84	.85 – 1.17	1.18 – 1.50	1.51 – 1.84	1.85 – 2.17	2.18 – 2.50
Low A1 range Scale score = .67	Mid A1 range Scale score = 1.00	High A1 range Scale score = 1.33	Low A2 range Scale score = 1.67	Mid A2 range Scale score = 2.00	High A2 range Scale score = 2.33
A1			A2		

This scale was used in calculating the following overall analysis of judgements across the two panels, as will be explained below.

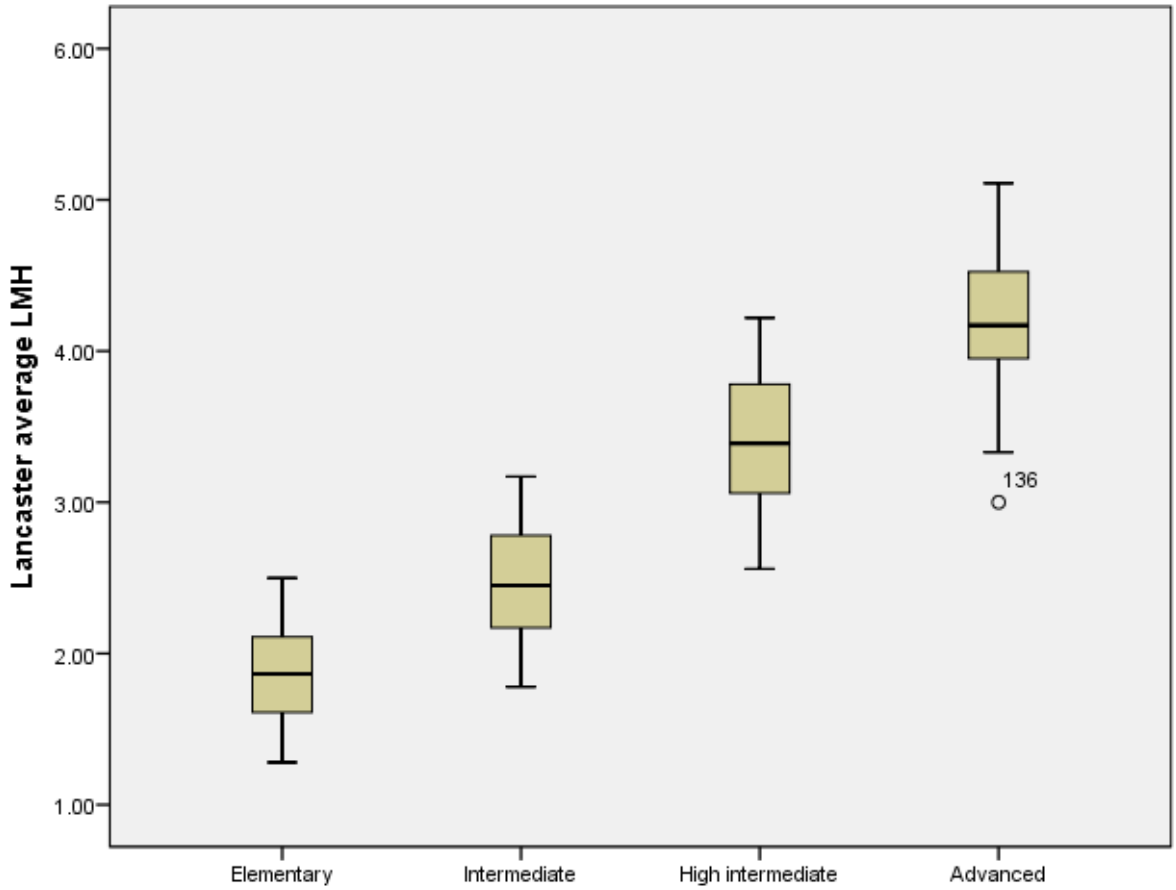
5.1.1. Lancaster panel judgements

Table 8 provides the descriptive statistics of (LMH) judgements given by the Lancaster panel at each test level; the means are represented visually in Figure 2.

Table 8: Lancaster panel – Average judgements by test level

GEPT listening test level	Mean	SD	Range
Elementary	1.86	0.36	1.28 – 2.50
Intermediate	2.45	0.36	1.78 – 3.17
High Intermediate	3.41	0.43	2.56 – 4.22
Advanced	4.22	0.43	3.00 – 5.11

*Figure 2: Lancaster panel – Average judgements by test level**



*Note: Here and throughout all similar charts: 1=A1, 2=A2, 3=B1, 4=B2, 5=C1, 6=C2.

This broad view was revealing, because it showed that judgements were somewhat lower than anticipated for the Intermediate, High Intermediate and Advanced level listening tests. As noted above, other GEPT research has shown a fairly straightforward mapping of Elementary to A2, Intermediate to B1, High Intermediate to B2 and Advanced to C1. However these mean figures suggest that while the mean judgement of Elementary test items was clearly in A2 range (M=1.86), the mean judgement for the Intermediate test was at the A2/B1 borderline (M=2.45), the mean judgement for the High Intermediate test was at the B1/B2 borderline (M=3.41), and for the Advanced level test, the mean judgement was approaching the high B2 level (M=4.22).

Two different validation checks were used to check for judge consistency. First, inter-rater reliability analyses were conducted. Cronbach's alpha and Intra-class correlation coefficients (ICCs) (average measures) were calculated using the LMH data for individual test levels and across all tests in the suite. Judgements were shown to be acceptably consistent, with levels above .7. The results are shown in Table 9.

Table 9: Lancaster panel – Judgement reliability

GEPT listening test level	α	ICC
Elementary	.784	.761
Intermediate	.737	.720
High Intermediate	.892	.880
Advanced	.879	.871
Full suite (all levels)	.969	.968

The reliability of judgements was also checked by considering the relationship between judgements at each test level and judgements made of items on a "mixed-order" test administered at the end of the standard setting. Judgements on the mixed-test showed that judgements were similar when using a different set of items sampled from all GEPT levels (in randomised order). Table 10 shows that means for the mixed-test were, on the whole, slightly higher than the means yielded by judgements on the full test. These mixed-test means, however, were clearly within the full-test range.

Table 10: Lancaster panel – Average judgements mixed-test

GEPT listening test level	Mixed-test mean	Full-test mean	Full-test range
Elementary	2.08 (N = 10)	1.86	1.28 – 2.50
Intermediate	2.59 (N = 7)	2.45	1.78 – 3.17
High Intermediate	3.61 (N = 9)	3.41	2.56 – 4.22
Advanced	4.41 (N = 8)	4.22	3.00 – 5.11

Cut scores for the Lancaster panel were calculated in two different ways. In the first method, mean (LMH) judgements for each item were re-categorised as a single CEFR level, so that <1.5 = A1; 1.51-2.5 = A2; 2.51-3.5 = B1; 3.51-4.5 = B2; 4.51-5.5 = C1; > 5.51 = C2 This led to the distribution of frequencies as shown in Table 11.

Table 11: Lancaster panel – Frequencies of CEFR level judgements based on (LMH) means

	Lancaster CEFR level					Total
	A1	A2	B1	B2	C1	
Elementary	**5**	**25**	0	0	0	30
Intermediate	0	**26**	**19**	0	0	45
High intermediate	0	0	**27**	**18**	0	45
Advanced	0	0	2	**28**	**10**	40
Total	5	51	48	46	10	160

The frequency table (Table 11) shows a noticeable split at the Intermediate, High Intermediate and Advanced level across two levels.

These frequencies also provide a means of setting cut scores, as the minimum number of items deemed answerable by a just-qualified (low) test-taker at a given level can be ascertained from a count of the number of items below that level + 1. Thus, for example, on the Elementary test the cut score for the A1/A2 boundary would be 5 + 1 = 6. In other words, a test-taker would have to be minimally proficient at the A2 level in order to answer 6 out of 30 items on this test. This, however, would represent a "just-qualified" candidate, and would be a somewhat conservative estimate as this assumes that a minimally proficient test-taker would answer all items correctly at the level below.

Another approach would be to consider the score which would indicate a test-taker at a "comfortable" level; that is, a candidate who is at the mid-point of a level. Following de Jong (2009), a more "comfortable" mid-level candidate might be expected to answer 50% of the items judged at that level correctly, and 80% of items at the level below (based on typical IRT probabilities, which was the method by which CEFR descriptors were scaled in development). So, for example, according the Lancaster panel judgements a comfortable A2-level test-taker would be expected to answer 5 x .8 + 25 x .5 items correctly, which would yield a cut score of 17 (rounded-up).

Based on these two methods, Table 12 shows recommended cut scores for each of the four level tests both for minimally proficient and comfortable CEFR levels.

Table 12: Lancaster panel – Cut scores

GEPT listening test level	**Minimally proficient**	**Comfortable**	**Total test score**
Elementary	A2 = 6 (20%)	A2 = 17 (57%)	30
Intermediate	B1 = 27 (60%)	B1 = 30 (67%)	45
High Intermediate	B2 = 28 (62%)	B2 = 31 (69%)	45
Advanced	B2 = 3 (8%) C1 = 31 (78%)	B2 = 16 (40%) C1 = 29* (73%)	40

* Note: This counter-intuitive result, where the comfortable cut score is lower than the minimally proficient cut score, is a consequence of (a) the conservative approach in setting the minimally proficient threshold, and (b) the high number of items judged B2 relative to those judged C1.

It is clear that, based on the Lancaster panel judgements, mid-range scores on the Elementary test (i.e. around the 50% mark) could be interpreted as being at a clear A2 level. Mid-range scores on the Intermediate test, however, might fall somewhere on either side of the A2+/B1- boundary, with candidates required to answer at least 67% of items correctly in order to be in the comfortable B1 range. A similar pattern is observed for the High Intermediate test, with candidates who achieve a 50% score still within the B1+ range, rather than the hypothesised B2 range. At the advanced level, candidates answering 50% of items correctly would be in the B2/B2+ range, and would need to answer over 70% of items correctly to be judged minimally proficient at the C1 level. This set of cut scores, therefore, reinforces the interpretation of mean judgements at each test level as signifying that: Elementary = A2 level; Intermediate = A2+/B1- level; High Intermediate = B1+/B2- level; Advanced = B2+ level.

5.1.2. Taipei panel judgements

The Taipei panel judgements were analysed, separately, in an identical way to the Lancaster judgements above. In the first instance, an overall view of the data is shown in Table 13 and Figure 3.

Table 13: Taipei panel - Average judgements by test level

GEPT listening test level	Mean	SD	Range
Elementary	1.91	0.39	1.06 – 2.50
Intermediate	2.68	0.42	1.95 – 3.55
High Intermediate	3.75	0.43	3.00 – 4.67
Advanced	4.57	0.45	3.39 – 5.56

At face-value, the mean judgements at each test level followed a similar pattern to the Lancaster data, though were somewhat higher, with the Elementary test firmly in the mid-A2 range (M=1.91), the Intermediate test judged on average at a low-B1 (M=2.68), the High Intermediate test at a low-B2 (M=3.75), and the Advanced test on the borderline of B2 and C1 (M=4.57).

As with the Lancaster panel, the validity of judgements was checked in two ways: by examining inter-rater reliability, and by considering the relationship between judgements at each test level and judgements made of items on a “mixed-order” test administered at the end of the standard setting.

Inter-rater reliability was at similar levels to that observed among the Lancaster judges, with alpha and ICC coefficients ranging from .688 to .859 for ratings on levels, coefficients of over .96 for the entire suite (see Table 14).

Figure 3: Taipei panel – Average judgements by test level

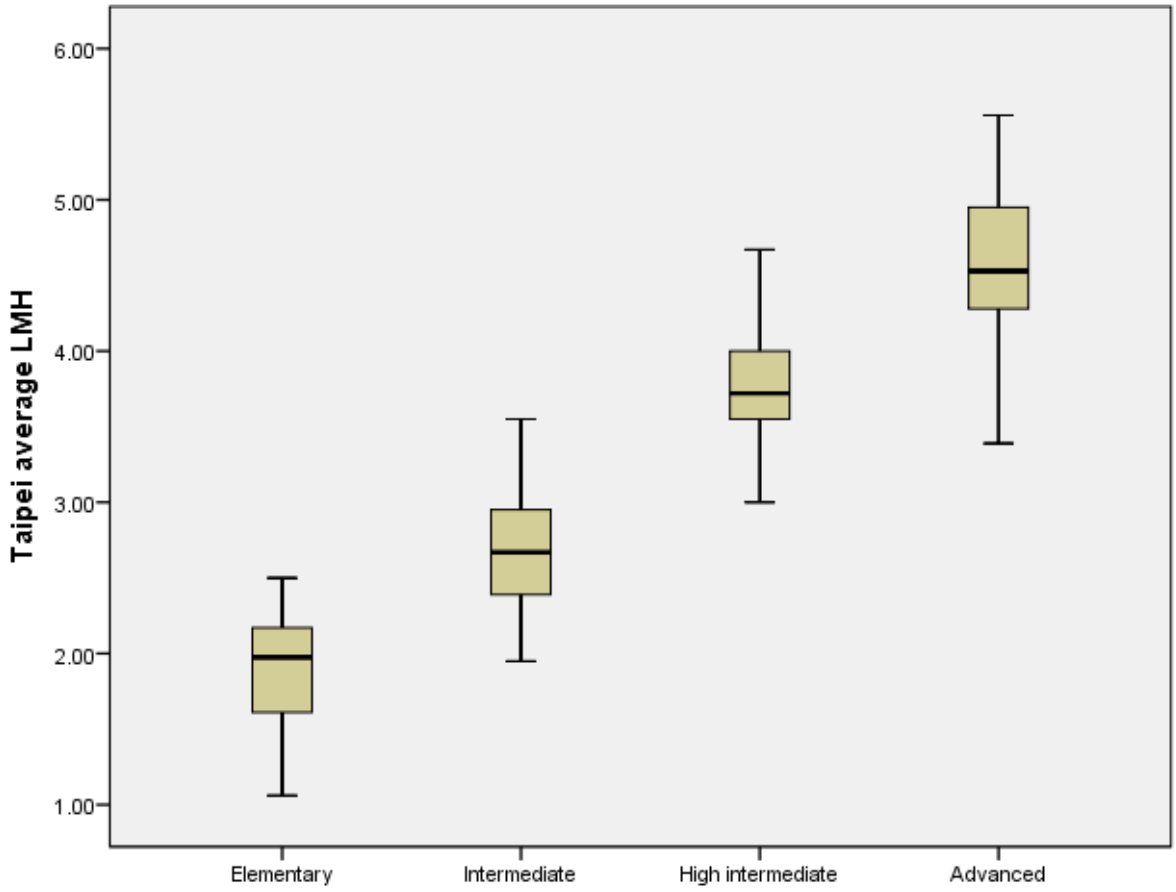


Table 14: Taipei panel – Judgement reliability

GEPT listening test level	α	ICC
Elementary	.816	.813
Intermediate	.715	.688
High Intermediate	.821	.809
Advanced	.859	.856
Full suite	.967	.965

These figures suggest a moderate- to high-level of agreement among judges, and show that the average measures were robust indicators of the judgements as a whole.

Judgements on the mixed-test also showed that judgements were broadly similar when using a different set of items sampled from all GEPT levels (in randomised order). Table 15 shows that means for the mixed-test were, on the whole, slightly lower than the means yielded by judgements on the full test. These mixed-test means, however, were clearly within the full-test range.

Table 15: Taipei panel – Average judgements mixed-test

GEPT listening test level	Mixed-test mean	Full-test mean	Full-test range
Elementary	1.85 (N = 10)	1.91	1.06 – 2.50
Intermediate	2.53 (N = 7)	2.68	1.95 – 3.55
High Intermediate	3.45 (N = 9)	3.75	3.00 – 4.67
Advanced	4.39 (N = 8)	4.57	3.39 – 5.56

Similar to the procedures followed for the Lancaster panel, cut scores were calculated in two different manners. Firstly, mean (LMH) judgements for each item were re-categorised as a single CEFR level, so that <1.5 = A1; 1.51-2.5 = A2; 2.51-3.5 = B1; 3.51-4.5 = B2; 4.51-5.5 = C1; > 5.51 = C2. The resulting distributions of judgements across CEFR levels on each test by the Taipei panel are presented in Table 16.

Table 16: Taipei panel – Frequencies of CEFR level judgements based on (LMH) means

	Taiwan CEFR level						Total
	A1	A2	B1	B2	C1	C2	
Elementary	**6**	**24**	0	0	0	0	30
Intermediate	0	**16**	**28**	1	0	0	45
High intermediate	0	0	**11**	**33**	1	0	45
Advanced	0	0	1	**19**	**19**	1	40
Total	6	40	40	53	20	1	160

A manifest split across two levels can be observed at the Advanced level (see Table 16), and a considerable split is also visible at the Intermediate level.

Secondly, the minimum number of items thought to be answerable by a just-qualified and a more comfortable test-taker at a given level were established following the same procedures as described for the Lancaster panel (see above). Table 17 shows the cut scores derived in this manner from the Taipei panel judgements.

Table 17: Taipei panel – Cut scores

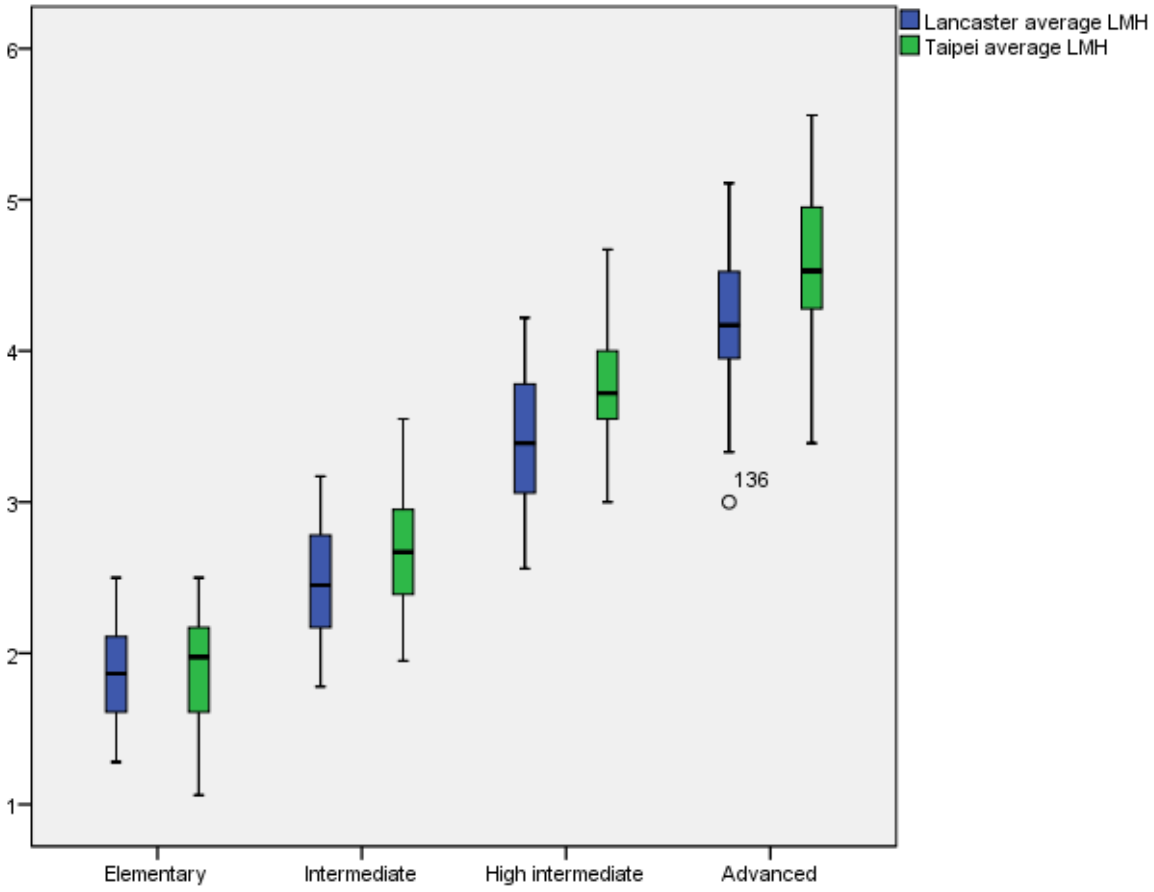
GEPT listening test level	Minimally proficient	Comfortable	Total test score
Elementary	A2 = 7 (23%)	A2 = 17 (57%)	30
Intermediate	B1 = 17 (38%)	B1 = 27 (60%)	45
High Intermediate	B2 = 12 (27%)	B2 = 25 (56%)	45
Advanced	C1 = 21 (53%)	C1 = 26 (65%)	40

The Taipei panel cut scores suggest that the Elementary test is comfortably at A2 level, the Intermediate test is pitched at a low- to mid-B1 level, and the High Intermediate test is at mid-B2 level. At the Advanced level, candidates need to answer more than half of the items correctly to be judged minimally proficient at the C1 level. The Advanced test is thus at the borderline of B2+/C1-.

5.1.3. Comparison of Lancaster and Taipei panels

Having analysed each panel's judgements separately, a decision needed to be made of whether or not it was feasible and valid to combine the judgements of the two sets of judgements. As noted above, the Taipei panel's judgements appeared to be slightly higher, on average, than the judgements of the Lancaster panel. Figure 4 illustrates this more clearly, suggesting also that the panels increasingly diverged in their mean judgements as test levels became more difficult. These apparent differences between panels were partly confirmed in a series of t-tests, which showed significant differences between the mean judgements of the two panels at Intermediate [$t(44) = -4.808$, $p < .001$, Cohen's $d = -0.57$], High Intermediate [$t(44) = -9.441$, $p < .001$, Cohen's $d = -0.81$], and Advanced [$t(39) = -5.401$, $p < .001$, Cohen's d = -0.81] levels. There was no significant difference in mean judgements on the Elementary test at $p < .05$. The increasing magnitude of the difference was also confirmed in the rising effect size from the Intermediate level to the High Intermediate and Advanced level. At these higher levels, there were differences between the mean judgements of each group of .346 (High Intermediate) and .353 (Advanced). This, effectively, meant that at these levels, the Lancaster panel were judging, on average, one LMH "step" below the Taipei panel (such that an item judged low-B2 by the Taipei panel would on average be judged high-B1 by the Lancaster panel). This is reflected in the differences in cut scores between the two panels, particularly at the higher levels.

Figure 4: Lancaster and Taipei panels – Average judgements by test level



Although the panels differed in the extent to which they agreed on a particular level, this does not mean that there was not a clear relationship between the two sets of ratings. Table 18 shows that averaged judgements of individual items were correlated at all levels, though with a somewhat weaker relationship on the Elementary tests (oddly, the only test where there was no significant difference between mean ratings).

Table 18: Correlations between Lancaster and Taipei panel judgements

GEPT listening test level	***r***	**Sig.**
Elementary	.420	.021
Intermediate	.689	<.001
High Intermediate	.836	<.001
Advanced	.556	<.001
Full suite	.941	<.001

These apparent differences between the behaviour of the two panels might be attributable to various factors. For one, it is likely that members of the Taipei panel were more familiar with the test, and with the typical test-takers, than the members of the Lancaster panel. Interestingly, when judging the mixed-test results and thus being less clearly cued on the GEPT level due to the randomized item order, the Taipei panel's average CEFR-judgements at a test level were slightly lower than when they had judged the items per test booklet (but

these differences were not vast and the judgements fell within the same range for the individual test level and mixed-test level items).

Also, several judges in Lancaster had many years' experience in using the CEFR, and may have interpreted and understood the descriptors (as they related to the panel materials) in different ways based on their experience of using and applying the framework in a European context. It was certainly the case that the Lancaster panel expressed difficulty in applying the CEFR descriptors to the GEPT listening tests (for reasons which will be explained below), and it is likely that this difficulty found expression in slightly lower judgements. Nevertheless, it was decided that the relationship between judgements was sufficient to combine judgements of across the twin panels to arrive at a final set of recommendations in order to reflect the mixture of expertise and experiences with the CEFR and with the GEPT main target population.

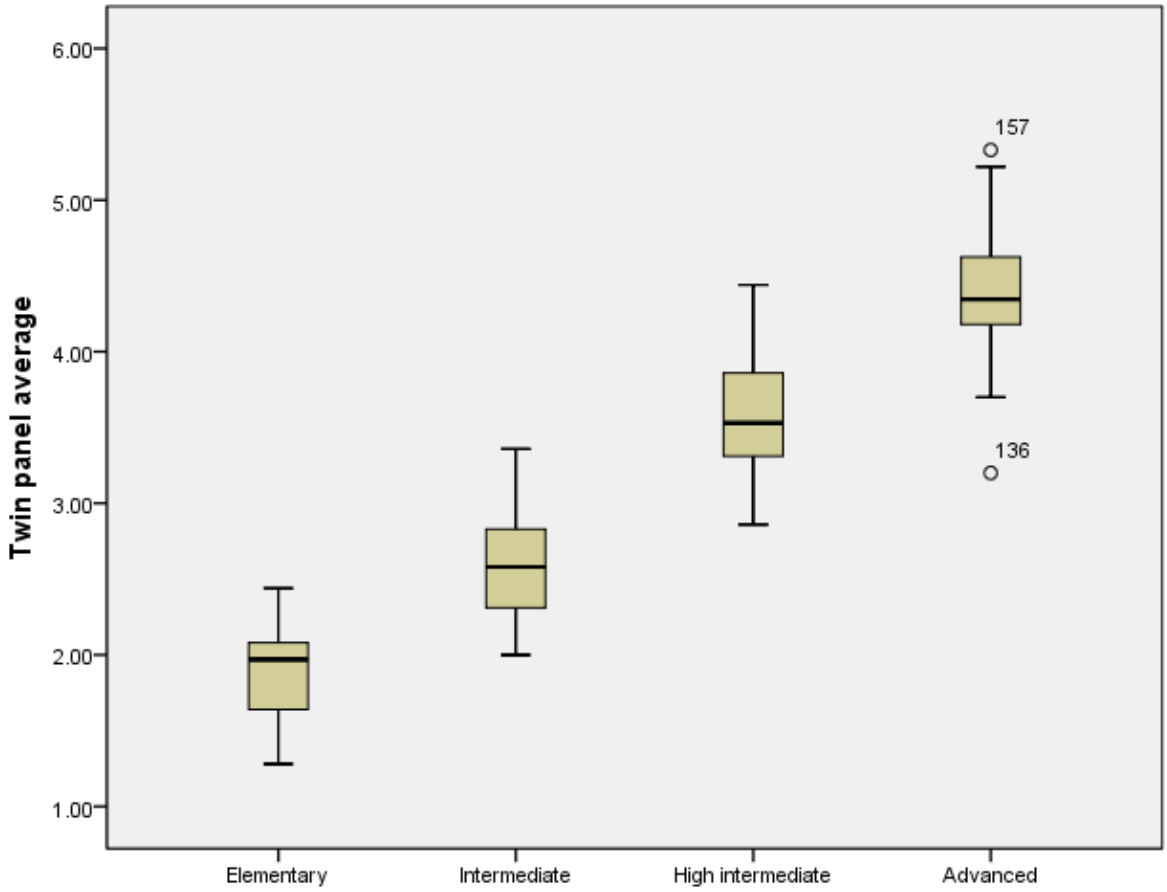
5.1.4. Combined Lancaster-Taipei results

The overall results of the combined Lancaster-Taipei panel are shown in Table 19 and Figure 5. These results were calculated by taking mean judgements of all twelve judges together. As such, they show what is effectively a mid-point between the judgements of the individual panels.

Table 19: Combined Lancaster-Taipei panel results

GEPT listening test level	Mean	SD	Range
Elementary	1.89	0.32	1.28 – 2.44
Intermediate	2.57	0.36	2.00 – 3.36
High Intermediate	3.58	0.41	2.86 – 4.44
Advanced	4.39	0.39	3.20 – 5.33

Figure 5: Combined panels – Average judgements by test level



These findings show a familiar separation across each of the four levels, and suggest that the Elementary test is firmly at the A2 level, the Intermediate test straddles the A2/B1 border, the High Intermediate test spans the B1/B2 level, and the Advanced test is placed more clearly at the high B2 level.

Inter-rater reliability was calculated for all judges combined, and the alpha and ICC levels were predictably higher than for each individual test by virtue of their being twice as many raters (see Table 20). These values, however, suggest that the mean judgements for the whole panel were robust indicators of the full set of judgements.

Table 20: Combined panels – Judgement reliability

GEPT listening test level	α	ICC
Elementary	.835	.828
Intermediate	.833	.812
High Intermediate	.920	.900
Advanced	.897	.878
Full suite	.982	.980

There were exceptionally close judgements between the combined means for different levels of the mixed-test and the four full tests (see Table 21). This suggests that the combined panel judgements were highly dependable.

Table 21: Combined panels – Average judgements mixed-test

GEPT listening test level	Mixed-test mean	Full-test mean	Full-test range
Elementary	1.97 (N = 10)	1.89	1.28 – 2.44
Intermediate	2.56 (N = 7)	2.57	2.00 – 3.36
High Intermediate	3.53 (N = 9)	3.58	2.86 – 4.44
Advanced	4.40 (N = 8)	4.39	3.20 – 5.33

In order to establish cut scores based on the combined panel data, items were categorised according to panel-mean CEFR judgements. Table 22 presents the resulting distribution of frequencies, which shows a noticeable split across two CEFR levels at the Intermediate and at the High Intermediate test level. In addition, the majority of Advanced level items have been judged to be B2 level by the overall panel.

Table 22: Combined panels – Frequencies of CEFR level judgements based on (LMH) means

	Taiwan CEFR level						Total
	A1	A2	B1	B2	C1	C2	
Elementary	**6**	**24**	0	0	0	0	30
Intermediate	0	**20**	**25**	0	0	0	45
High intermediate	0	0	**20**	**25**	0	0	45
Advanced	0	0	1	**25**	**14**	0	40
Total	6	44	46	50	14	0	160

Having followed the procedures outlined for the individual panels' cut-score analyses (see above), the findings on the minimum number of items deemed to be answerable by a just-qualified and a more comfortable test-taker at a given level (based on the combined panel judgements), are provided in Table 23.

Table 23: Combined panels – Cut scores

GEPT listening test level	Minimally proficient	Comfortable	Total test score
Elementary	A2 = 7 (23%)	A2 = 17 (57%)	30
Intermediate	B1 = 21 (47%)	B1 = 29 (64%)	45
High Intermediate	B2 = 21 (47%)	B2 = 29 (64%)	45
Advanced	B2 = 2 (5%) C1 = 27 (68%)	B2 = 13 (33%) C1 = 28 (70%)	40

As was observed in the separate panel analyses, the combined panel judgements indicate that the Elementary test is comfortably at A2 level. Scores around the 50% mark at this test level could be interpreted as being at a clear A2 level. The Intermediate and the High Intermediate tests are pitched at a low-B1 and a low-B2 level respectively, with candidates who obtain mid-range marks showing to be just-qualified at these levels. Similar to the findings based on the individual panels, the combined panel judgements suggest that the Advanced level test is pitched at the B2+ level. Candidates would need to answer approximately 70% of items correctly to be judged minimally competent at the C1 level.

However, the existing passing scores for the GEPT levels are comparatively high, and when mapped on to these passing scores made available by the LTTC (personal communication, but see also https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/PassingStandard.htm) it becomes clear that a test-taker who has passed the Intermediate and High-Intermediate levels would, in fact, have answered a sufficient number of items correctly to be judged B1 and B2 respectively according to the cut scores set in this study. Similarly, a test-taker who passed the elementary level test would be comfortably at an A2 level. However, even taking into account the pass-score for the Advanced level test, the level is still clearly at the B2 level (we would suggest B2+ based on the mapping of the passing score to the cut scores shown in Table 23). The relationship between pass scores and cut scores for each level of the suite is shown in Table 24:

Table 24: Cut scores mapped-on to existing pass scores

Level	Existing pass score	Combined panel cut score	Recommended CEFR alignment
Elementary	67%	57% (A2)	A2
Intermediate	67%	64% (B1)	B1
High Intermediate	67%	64% (B2)	B2
Advanced	63%	33% (B2) 70% (C1)	B2(+)

5.2. Qualitative findings

The group discussions which were held at the end of the panel sessions were useful for two reasons. Firstly, they provided a means of checking perceptions of the validity of procedures directly with participants. Secondly, they provided a means of collecting data on judges' perceptions of any difficulties they had in applying the CEFR, making their judgements, dealing with (for some) unfamiliar test materials, and general comments on the experience of standard setting.

5.2.1. Perceptions of validity of procedures

In the final recorded discussion, panel members in Lancaster commented extensively on the validity and ease of the standard setting procedures. For example, there was consensus that the judges felt comfortable with the tasks they had to perform:

> *I think that the whole process that you two were running was really smooth and personally saying I had enough time (yeah) so to look at the descriptors look at the task itself and it was, really I appreciate it that we had, that we could listen to every*

item twice, so that was good, so listening, just so doing it as the test taker first and then looking at the descriptors

Judges particularly felt that they had been given enough time to work with the descriptors:

… some of the booklets that we used Tuesday afternoon there were large sections to do at one time where one question ran over three items in one booklet it was great that you stopped to allow us time to apply before we moved on to the next one

The implementation of the low, mid and high level judgements was also commented on by one judge (and resonated by others), who found that this particular innovation made the judging "more useful":

I found for instance that judgment two [L-M-H] I mean I would be in touch about how you implement that in the analysis and I found that was a very neat approach to complement your ... second judgment that you have the level plus (yeah) I found that actually quite useful.

There was also a sense among judges on the Lancaster panel that they were confident with their own judgements. This was attributed to the amount of discussion which was allowed between first and second judgements:

I think what I felt very happy about with talks again with all discussions we we had … and I felt that the discussions were certainly solidly linked to the CEFR descriptors we may have disagreements in which table we were using and what descriptor we refer but generally between one band and another which is perfectly within the limits.

It is worth pointing out, though, that the panel members tended to treat the facility values (which were shown during the discussion between first and second round judgements) with a degree of scepticism:

the facility values were there but we really didn't look at them as much as we do in our tests

This is most likely because, early in the judging process, the Lancaster panel noticed that their judgements did not always reflect rises and falls in the facility values provided. It should be pointed out that this was one judge's opinion, and may not have reflected the way in which the facility values were utilised by other Lancaster panel judges in making their second round judgements.

The Taipei panel provided fewer comments on the validity of procedures, but the elements of the discussion seemed to reflect the Lancaster panel's views that the procedures were well-planned and easy to apply. One judge stated that the panel was an "enjoyable process". Another commented on the adequate amount of time that was provided for making judgements:

quite relaxed actually (ok) yeah I mean I mean the pace we wouldn't feel like you know we are pressured to do something this is like you know based on our own experience and we make judgements and then we still have some time to think and reflect

5.2.2. Challenges in making judgements

Despite a consensus that the standard setting procedures were sound, judges on both panels expressed a number of problems in relating this particular suite of listening tests to the Common European Framework of Reference. Two main issues arose across final panel discussions: (1) the deficiencies in the CEFR descriptors as they apply to listening tests, and (2) the fact that the test was not designed with CEFR principles in mind. As will be shown below, there was an imbalance in the degree to which these issues were discussed in each context.

5.2.2.1. CEFR-related issues

The Taipei panel members made a number of critiques of the CEFR descriptors which they had been using to set standards. Problems were identified with the coverage of descriptors with respect to particular aspects of listening:

> *… the descriptors haven't provided enough information on the pace [of the speaker]*.

However most comments concerned the general structure or quality of the scales:

> *… less information on the bottom and top levels; missing descriptors: especially for A1*

> *… but the descriptors are not consistent*

At other times, the judges problematized the application of the CEFR because of the vagueness of descriptors:

> *[There was an] unclear match between CEFR tables and listening texts: it's hard for us put it into any category for example does it belong to like TV or is it a conversation or like recorded everything is recorded of course but how do you define that*

One judge also expressed that the information contained in scales was not enough to make a clear judgement for borderline cases:

> *adjacent level distinctions can be difficult; specially when they are in the you know distinction between B1 from B2 … because it seems that there is not a clear cut just like a zone and sometimes you have to decide now which should you go for based on my experience or on my observation you know or my like probably imagination about my you know participants*

The Lancaster panel members did not make explicit references to deficiencies in the final discussion, although these had been raised throughout discussions during the entire standard setting process. Particularly, the Lancaster judges also pointed to issues with the vagueness of some descriptors (and particularly the lack of descriptors relating to cognitive processes at different levels of listening ability), and inconsistencies in features across levels.

5.2.2.2. Test-related issues

Panellists in both contexts pointed out that the GEPT suite had not been designed according to CEFR principles, and it was therefore not always easy to make judgements according to the descriptors provided. However this perceived problem was a much more prominent feature of

discussion among the Lancaster panel judges. For example, a member of the Lancaster panel made the following comment:

> *… I think all of us clearly expressed that certain points that we couldn't actually map a particular task to the CEFR because we felt that was not extended speech there was not enough conversation, there were too many other factors involved and we felt that they introduced lots of construct-irrelevant variance … we were not sure it was listening um I could have felt more comfortable for example if I could apply reading CEFR descriptors and listening CEFR descriptors but I had to apply only the listening ones.*

This comment points to a general concern that the Lancaster panel had in dealing with a test which was unlike many they had worked with previously. Some of the features of the GEPT approach to listening assessment which came under criticism during this discussion were:

- The high reading load in processing multiple-choice questions
- The mixing of item types in some tasks (e.g., MCQ and short-answer questions within the same listening passage)
- The memory load, particularly items where MCQ stems and options are delivered orally
- The level of language used in rubrics and options delivered orally (which was sometimes perceived to be above the level of the passages themselves)
- The use of the same two voices across numerous items in the same test (and also using those same voices to deliver instructions)
- The lack of variation in accent type and speech rate
- The lack of background noise in stimuli
- The lack of characteristics of natural speech (e.g., pauses, stress, intonation and articulation patterns) in the listening input

The Lancaster panel felt that these factors might have affected their decisions; for example, judges reported a "ceiling effect" owing to a lack of natural speech features in the input which prevented them from judging items at the C1 (or C2) level. This may be because speech does not often move beyond clear articulation of standard North American English (see Specification analyses in Appendix A, Section C), which is a feature of B1 level texts (see e.g., the illustrative scale "Understanding Conversation Between Native Speakers" (Council of Europe, 2001, p.66)).

The Taipei panel noted this general issue as well, but with much less elaboration. One judge, for example, noted that there were elements of the CEFR descriptors which did not occur in the GEPT suite:

> *if the listenings are to be based on the descriptors or not if they work then for example the table that understanding conversations between the speakers you see one mentioned group discussions and debates we didn't see in terms of that also listening to audio video and recordings C1 says identifying personal attitudes relationships between speakers the question that was really asking the test takers to do that B2 level viewpoints and attitudes may be occasionally or identifying speakers' mood, tone and etc.*

The critique of the test itself, however, was a much more prominent theme in the Lancaster panel discussion, leading to the bold statement from one judge during the discussion, voicing

concerns about what this judge perceived as a serious mismatch between the CEFR and the GEPT:

> *I'm also particularly concerned about test use and misuse and if our part in this is leading towards the misuse of the CEFR …*

These comments could be taken as indicating a disjunction between two "testing cultures". On the one hand, the CEFR-informed European approach to test development which is primarily oriented towards communicative language testing principles, and on the other hand, the more discrete-point, reliability-focused testing methods employed by the GEPT listening suite. However, this is a real concern as the GEPT is seeking to link to the CEFR framework, yet the approach to listening assessment taken across the suite does not match up easily with the CEFR descriptors. We have therefore included, in our recommendations below, some options that might be considered in any future revisions to the GEPT.

6. Summary and Recommendations

6.1. Summary

This study was designed with the aim of linking the GEPT listening test suite (Elementary, Intermediate, High Intermediate, and Advanced) with the Common European Framework of Reference. The study involved several stages, which were informed by procedures recommended in the document *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching and assessment. A Manual* (Council of Europe, 2009). Of particular importance, a standard setting study was conducted with an innovative twin-panel design, featuring judges in Lancaster and in Taipei. This process yielded data which allowed for a broad alignment of the GEPT listening test levels to the CEFR, with a key finding that existing, hypothesised levels might be slightly over-estimated. Recommendations were also made for cut scores, and these are reiterated in the section below. The study also addressed two sub-research questions designed to explore the usefulness of some of the innovative aspects of this study, including the choice of a method by which to standard set a suite of tests, and the validity of the twin-panel approach. With respect to the first question, judges' perceptions of the strengths and limitations of the basket method versus the modified Angoff method fed into the design of a "modified-basket" procedure which sought a more nuanced judgement of level by dividing CEFR baskets into "compartments" of low-, mid- and high-ability learners. Secondly, the twin-panels method yielded a set of comparable judgements which effectively cross-validated the standards set by each panel. While judgements were, quantitatively, comparable in nature, and panels were internally consistent to a similar degree, qualitative data showed that there were differences in the experiences of the panels in applying descriptors during the standard setting, with the Taipei panel tending to problematize CEFR descriptors and the Lancaster panel outlining in detail aspects of the GEPT listening tests which made it difficult to relate to the can-do descriptors.

6.2. Recommendations

Based primarily on the results of the standard setting panels, and reinforced by the specification stage carried out by expert judges, the following suggestions are made to the Language Training and Testing Centre.

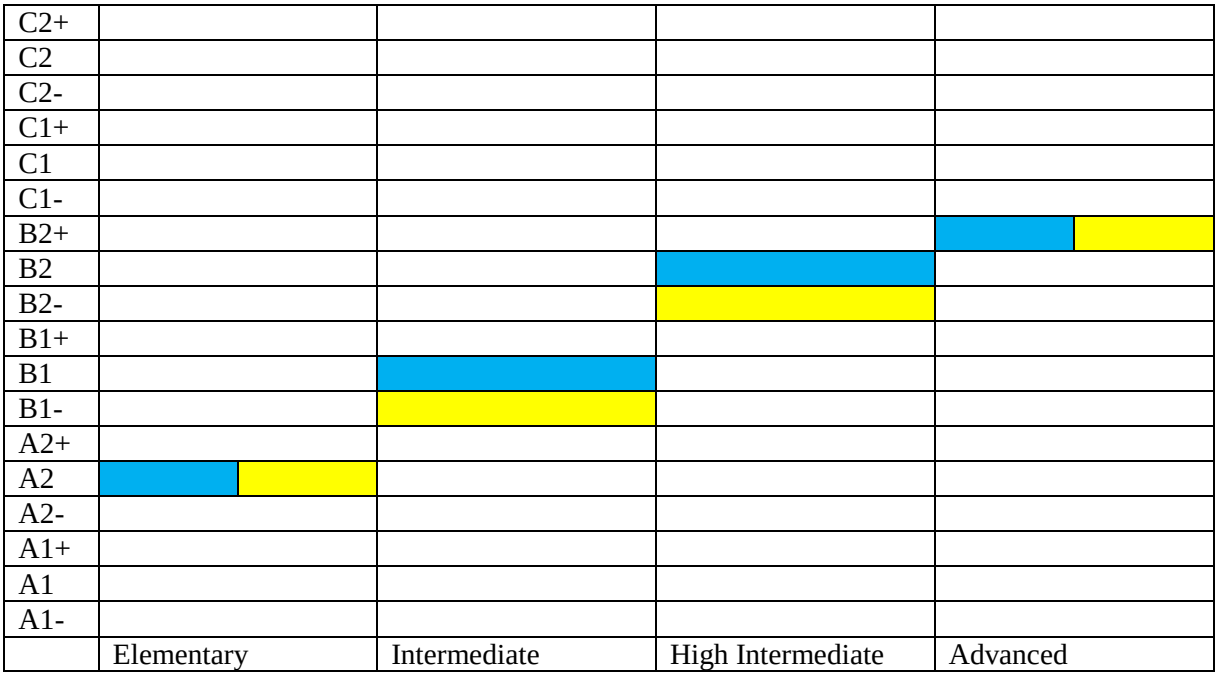
First, it is recommended that the combined Lancaster-Taipei cut scores be adopted, and from these that the "comfortable" level cut scores be selected, as they would theoretically be more stable. It is important to note that these cut scores will need to be interpreted with respect to the Standard Error of Measurement for each test, and so in fact might represent a range of two or three score points in practice, depending on the reliability of particular test versions.

Second, it is recommended that the GEPT listening test be reported as aligned to the Common European Framework according to the recommendations made in Table 24, which is reproduced below:

Level	Existing pass score	Combined panel cut score	Recommended CEFR alignment
Elementary	67%	57% (A2)	A2
Intermediate	67%	64% (B1)	B1
High Intermediate	67%	64% (B2)	B2
Advanced	63%	33% (B2) 70% (C1)	B2(+)

This recommended alignment is illustrated, for ease of understanding, in Figure 6. This chart has been developed to indicate (in blue shading) the comfortable CEFR that would be achieved by a passing test-taker, by mapping on the recommended cut-scores to the existing passing standards for the GEPT suite (according to Table 24). However, it is noteworthy that at the middle two levels the items were judged, on average, to be at a somewhat lower level (shaded in yellow – this information relates to the mean judgements shown in Table 19). This suggests that although test-takers can be considered to have achieved the CEFR level indicated by the blue shading, the test items on the whole are pitched at the level indicated by the yellow shading. The mismatch at the middle two levels suggests that in developing and revising the listening suite, item writers may aim to include more items at the desired level in order to ensure that multiple observations are made of the candidate's performance at the targeted level.

Figure 6: GEPT listening suite CEFR alignment



Third, there is scope for modifications to test specifications in order for the GEPT listening tests to be more easily articulated with the CEFR descriptors, and in so doing to increase the validity of alignment of the test levels to the framework. This is not to say that the GEPT listening tests are not of sufficient quality or usefulness as they stand; certainly, they fulfil an important role in the many contexts in which GEPT test scores are used. However, in order, particularly, to be more easily mapped at higher CEFR levels, and specifically to achieve the target of a C1 level Advanced level test, the following recommendations might be considered:

1. To include in the oral stimuli – particularly at the High Intermediate and Advanced levels – more naturalistic features of speech such as false starts, hesitations, repair, a variety of accents, a more extemporaneous style.
2. To include longer texts, particularly in the genre of conversation, and on more authentic topics and in more naturalistic contexts (e.g., background noise might be included behind listening texts at higher levels). In particular, the first sets of listening texts at the Elementary, Intermediate and High Intermediate levels might be broadened from orally presented multiple-choice stems and/or options to longer, authentic listening texts.
3. To include more variation in speakers, which is likely to be associated with more diversity in accents, pitch, speech rate, and other features of spoken language. In particular, a distinct pool of speakers for the task instructions versus the listening text input is suggested.
4. In summary, particularly at the higher test levels, to include wider variation in the various characteristics of the listening input, in order to reduce the emphasis on lexical complexity features for distinguishing between levels of difficulty.

Also, as a general recommendation:

5. The reading load of the exam could be reduced, and where instructions are given orally, these might also be written to reduce the memory load.

References

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

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

Council of Europe (2005). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR).Reading and listening items and tasks: Pilot samples illustrating the common reference levels in English, French, German, Italian and Spanish.* CD-Rom.

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

De Jong, J. H. A. L. (2009, June). *Unwarranted claims about CEF alignment of some international English language tests*. Paper presented at EALTA, Turku, Finland. Retrieved from http://www.ealta.eu.org/conference/2009/docs/friday/John_deJong.pdf

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology. *TOEFL iBT Research Report*, RR-08-34. Princeton, NJ: ETS. Retrieved 15 December 2011, from http://www.ets.org/Media/Research/pdf/RR-08-34.pdf

Weir, C. J., Chan, S. H. C., & Nakatsuhara, F. (2013). Examining the criterion related validity of the GEPT advanced reading and writing tests: Comparing GEPT with IELTS and real life academic performance. *LTTC-GEPT Research Report,* RG-01. Retrieved from https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG01.pdf

Wu, R. Y. F. (2011). *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference* (Unpublished doctoral dissertation). University of Bedfordshire, UK.

Wu, J. R. W., & Wu, R. Y. F. (2010). Relating the GEPT reading comprehension tests to the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 204-222). Cambridge: Cambridge University Press.

Appendix A – Specification Documentation

Please note that part of the information in Forms A1-A19 was completed by the Language Training and Testing Centre.

Form A1: General Examination Description

GENERAL EXAMINATION DESCRIPTION	
1. General Information	
Name of examination	General English Proficiency Test (GEPT) suite - listening section Levels: Elementary/Intermediate/High-intermediate/Advanced
Language tested	English;
Examining institution	Language Training and Testing Centre (LTTC)
Versions analysed (date)	Listening booklets: Elementary 1161, Intermediate 1062, High-Intermediate 1161, Advanced 1001
Type of examination	✓ International ✓ National ☐ Regional ☐ Institutional
Purpose	Measuring general English listening proficiency level of Taiwanese learners (source: http://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	5.4 million since its launch in 2000 (source: http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/recognition.htm)
2. What is the overall aim? Testing general listening skills with the aim of improving the general English listening proficiency of Taiwanese learners and providing institutions/schools with a reference for evaluating the English proficiency levels of their job applicants, employees, or students (source: http://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 listening proficiency of English foreign language learners in junior high schools, high schools, universities, and private enterprises in Taiwan – Evaluation of the general English listening proficiency of university applicants in Taiwan and in institutions around the world (including in Asia, Europe, and the USA) for university entry at under- and postgraduate level, for student placement, and as a criterion for university graduation. – Evaluation of the general English listening 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

<table>
<tr><td>5. Which communicat-ive activities are tested?</td><td>✓1 Listening comprehension

☐ 2 Reading comprehension
☐ 3 Spoken interaction
☐ 4 Written interaction
☐ 5 Spoken production
☐ 6 Written production
☐ 7 Integrated skills
☐ 8 Spoken mediation of text
☐ 9 Written mediation of text
☐ 10 Language usage
☐ 11 Other: (specify):</td><td>Name of Subtest(s)
Elementary
Intermediate
High-Intermediate
Advanced
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____</td><td>Duration
20 mins
30 mins
35 mins
45 mins
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____</td></tr>
<tr><td>6. What is the weighting of the different subtests in the global result?</td><td colspan="3">Equal weighting</td></tr>
<tr><td>7. Describe briefly the structure of each subtest</td><td colspan="3">Elementary (EL): 4 sections, 30 items
1. Picture description
2. Answering questions
3. Conversations
4. Short talks
Intermediate (IL): 3 sections, 45 items
1. Picture description
2. Answering questions
3. Conversations
High-intermediate (HL): 3 sections, 45 items
1. Answering questions
2. Conversations
3. Short talks
Advanced (AL): 3 sections, 40 items
1. Short conversations & talks
2. Long conversations
3. Long talks</td></tr>
</table>

8. What type(s) of responses are required?	✓ Multiple-choice ☐ True/False ☐ Matching ☐ Ordering ✓ Gap fill sentence ☐ Sentence completion ☐ Gapped text / cloze, selected response ☐ Open gapped text / cloze ✓ Short answer to open question(s) ☐ Extended answer (text / monologue) ☐ Interaction with examiner ☐ Interaction with peers ☐ Other	Subtests used in (Write numbers above) **EL 1, EL2, EL3, EL4, IL1, IL2, IL3, HL1, HL2, HL3, AL1, AL2, AL3** ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ **AL2, AL3** ☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ **AL2, AL3** ☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐
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 (writing test) ✓ 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 (part)

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: National Police Administration, high schools, colleges & universities, private

	sectors
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: External organizations are not involved in design or development of the test. They select a suitable level of the test based on their needs.
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 environment.
6. Have test writers guidance to ensure quality?	✓ Training ✓ Guidelines/Wordlists ✓ Checklists ✓ Examples of 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, together with tested anchor items, into pre-test papers which

	conform to the test specifications. Pre-test papers are administered to a representative sample of the target population. Through anchor items, new items can be linked to a common scale of difficulty. Pre-tested items which show sound item statistics go into the item bank for later use.
12. If no, why not?	n/a
13. Is the reliability of the test estimated?	✓ Yes (the reliability of the listening forms used in the study ranges from .81 to .87.) ☐ 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 ✓ Washback/Consequential 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 skill is used as the basis for selection of the elements to be included in the test form. • Score comparisons have been made of the GEPT Intermediate Level and High-Intermediate Level and of the CBT TOEFL, developed by ETS in the U.S. Moderate to strong correlations were demonstrated in the studies (MoE sponsored project, 2003). • A multitrait-multimethod (MTMM) matrix of correlations is calculated after every operational test to check the convergent validity and the discriminant validity. Theory-based validity such as metacognitive processing, and consequential validity are also investigated.

Form A3: Marking

Marking: Subtest	**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?	High school and university English teachers with extensive teaching experience, or English majors whose English proficiency levels are at or above C1, and who are familiar with the local English learning environment. In addition, all markers are trained in advance of marking.
4. How is accuracy of marking promoted?	✓ Regular checks by co-ordinator ✓ Training of markers/raters (AL) ✓ Moderating sessions to standardise judgments (AL) ✓ Using standardised examples of test tasks(AL) ☐ 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 ☐ 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: Subtest _____	**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	At each GEPT level, the listening component is combined with the reading component as one test, and the cut score is set for the combined test rather than for individual components. The content of the listening and reading tests 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 of the tests, the LTTC Research Committee reached a consensus on the level of proficiency in the test deemed to be satisfactory as equivalent to a pass in the test, i.e., that candidates had to answer at least two thirds of the test items correctly in order to pass the level. Hence the cut score for the combined listening and reading test at each GEPT level is set to be 160 out of a total of 240. Please note that the cut score of the Advanced Level test was revised downwards to 150 out of a total of 240 in 2013 based on the results of internal research and the study conducted by Weir, Chan, and Nakatsuhara (2013).
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	See Q2 above.
4. If grades are given, how are the grade boundaries decided?	See Q2 above.
5. How is consistency in these standards maintained?	The content of all tests is based on the specifications of each level test. In order to conform to the specifications, throughout the test production process, stringent guidelines are followed. Test items are pre-tested, and those with sound item statistics are compiled into operational tests to ensure that the difficulty of each test form remains stable.

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 ✓ On-line (Please note that the online score report 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 posted.)
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 personal background information which the candidates provided on 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

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 ✓ Discrimination ✓ Reliability ✓ Validity
7. State which analytic methods have been used (e.g. in terms of psychometric procedures).	Classical item analysis, IRT analysis, MTMM, ANOVA, and DIF are performed.
8. Are performances of candidates from different groups analysed? If so, describe how.	Performances of candidates are grouped and analysed based on personal background information that the candidates provided on the registration form.
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 designed for scoring.
10. Are relevant measurement concepts explained for test users? If so, describe how.	The relevant information, such as the difference between norm-referenced and criterion-referenced testing, and marking procedures, is published on the LTTC website and candidate handbooks.

Form A7: Rationale for Decisions

Rationale for decisions (and revisions)	**Short description and/or reference**
Give the rationale for the decisions that have been made in relation to the examination or the test tasks in question. Is there a review cycle for the examination? (How often? Who by? Procedures for revising decisions)	Decisions in relation to the development and revision of the GEPT and its test tasks are based on the following criteria: - **document analysis**: the analysis of curriculum guidelines, course books, and other learning materials - **needs analysis**: data collected from test takers, test administrators, teachers, test users, and other stakeholders through questionnaires after operational tests or teacher forums - **test data analysis**: after every operational test, test takers' performances are analysed and the test forms are reviewed

	There is a review cycle for the test based on the on-going analysis of the above documents and data. A few examples of test revisions are: the Elementary Level Listening Test in 2010; the High-Intermediate Level Reading Test in 2010; in addition, the Elementary Level Writing Test is now under review.

Form A8: Initial Estimation of Overall Examination Level

Initial Estimation of Overall CEFR Level		
☐ **A1**	**IL: B1**	**AL: C1**
☐	☐	☐
EL: A2	**HL: B2**	☐ **C2**
☐	☐	☐
Short rationale, reference to documentation Information on the GEPT-CEFR alignment is given on the following LTTC webpage: http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/alignment.htm A GEPT-CEFR alignment project was undertaken for the reading section of the exam, and has been published in Wu & Wu (2010) and Wu (2011).		

A. Specification: Communicative Language Activities

B.1. Reception

Form A9: Listening Comprehension

	Short description and/or reference
1 In what contexts (domains, situations, …) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	**EL/IL/HL/AL:** - Personal - Public - Occupational - Educational
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	**EL**: food and drink, places, health and body care, free time and entertainment, weather, relations with other people, work, shopping, travel, education **IL**: house and home, food and drink, places, health and body care, free time and entertainment, weather, relations with other people, work, shopping, travel, education **HL**: food and drink, places, health and body care, free time and entertainment, weather, relations with other people, work, shopping, travel, education, services, science, business **AL**: health and body care, free time and entertainment, work, travel, education, science, history, society, business (source: topic areas indicated in the answer keys)
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	**EL/IL/HL/AL:** Listening to public announcements Listening to radio, TV, and recordings Listening as a member of a live audience of meetings, lectures, entertainments Listening to conversations
4 What text-types and what length of text are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	**EL:** informal conversations **IL:** informal conversations; public announcements; weather forecasts; commercial passages; instructions; telephone messages **HL:** conversations in various social settings (e.g., chats, discussions, transactions) ;public announcements; weather forecasts; commercial passages; public service messages; instructions narratives; descriptions; news reports; extracts from lectures / presentation **AL**: conversations in various social settings (e.g., chats, discussions, transactions, interviews, debates); commercial passages; public service

	messages; narratives; descriptions; news reports / news features; extracts from TV/Radio programs (e.g., documentaries, commentaries); lectures / presentation
5 After reading the scale for Overall Listening Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for listening comprehension in CEFR 4.4.2.1 listed after the scale might be of help as a reference.	**Level & Justification (incl. reference to documentation)** **EL: A2** "Can understand simple English sentences, short conversations, and stories" (source: http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/elementary.htm) ➔ Cf. CEFR A2 descriptor Overall Listening Comprehension **IL: A2/B1** "Can understand general English conversations in daily life situations" ➔ Cf. CEFR B1 descriptor Overall Listening Comprehension "Can grasp the general meaning of announcements, advertisements, and broadcasts" (source: http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/intermediate.htm) ➔ Cf. CEFR A2 descriptor Listening to Announcements and Instructions **HL: B1/B2** "Can understand English conversations in social settings and workplaces" ➔ Cf. CEFR B1 & B2 descriptors Overall Listening Comprehension "can grasp the general meaning of lectures, news reports, and TV/radio programs" (source: http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/hi_intermediate.htm) ➔ Cf. CEFR B2 Listening as a Member of a Life Audience; CEFR B1 & B2 descriptors Listening to Audio and Media Recordings; CEFR B1 & B2 descriptors Watching TV and Film **AL: B2/C1** "Can understand conversations on all sorts of topics" ➔ Cf. CEFR B2 Overall listening comprehension "Can understand professional lectures, speeches, and news reports" (source: http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/Advanced.htm) ➔ Cf. CEFR C1 Listening as a member of a live audience; CEFR B2 Watching TV and Film

B. Specification: Communicative Language Competence

C.1 Reception

Form A19: Aspects of Language Competence in Reception

Linguistic Competence	Short description and/or reference
1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	**EL**: *Lexical:* GEPT Elementary Level Word List, which contains about 2000 words drawn from the following sources: --Collins Cobuild Bands 4 & 5 --MOE curriculum for junior high schools *Grammatical:* Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words. Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information. Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people, what they do, places, possessions etc. Has a limited repertoire of short memorised phrases covering predictable survival situations; frequent breakdowns and misunderstandings occur in non-routine situations. **IL**: *Lexical:* GEPT Intermediate Level Word List, which contains about 5000 words drawn from the following sources: --Collins Cobuild Bands 3~5 --MOE curriculum for senior high schools *Grammatical:* Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films. Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times.

	HL: *Lexical:* GEPT High-Intermediate Level Word List, which contains about 8000 words drawn from the following sources: --Collins Cobuild Bands 2~5 --CET (College English Test) Word List Levels 1~6 *Grammatical:* Can express him/herself clearly and without much sign of having to restrict what he/she wants to say. Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so. **AL**: *Lexical:* Collins Cobuild Bands 1~5 (about 14,000 words) *Grammatical:* Can select an appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say.

2 After reading the scale for Linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level & Justification (incl. reference to documentation)** **EL: A2** "Test-takers who pass this level have basic ability in English and can understand and use rudimentary language needed in daily life" (Source: Elementary Level general level descriptor, http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/elementary.htm) ➔ Cf. CEFR A2 descriptor General Linguistic Range **IL: B1** "Test-takers who pass this level can use basic English to communicate about topics in daily life" (Source: Intermediate Level general level descriptor, http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/elementary.htm) ➔ Cf. CEFR B1 descriptor General Linguistic Range **HL: B2** "Test takers who pass this level have a generally effective command of English" (Source: High-Intermediate general level descriptor, http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/elementary.htm) ➔ Cf. CEFR **B**2 descriptor General Linguistic Range **AL: B2/C1** "Test-takers who pass this level have English ability enable them to communicate fluently with only occasional errors related to language accuracy and appropriateness" (Source: Advanced Level general level descriptor, http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/elementary.htm) ➔ Cf. CEFR B2 & C1descriptor General Linguistic Range and Grammatical Accuracy
Socio-linguistic Competence	**Short description and/or reference**
3 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	✓ Unknown to the researchers
4 After reading the scale for Socio-linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level** ✓ Unknown to the researchers **Justification (incl. reference to documentation)** ✓ Unknown to the researchers
Pragmatic Competence	**Short description and/or reference**
5 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.	✓ Unknown to the researchers
6 After reading the scale for Pragmatic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level** ✓ Unknown to the researchers **Justification (incl. reference to documentation)** ✓ Unknown to the researchers

Strategic Competence	**Short description and/or reference**
7 What are the strategic competences that the test takers are expected to be able to handle? The discussion in CEFR 4.4.2.4. might be of help as a reference.	**At each level, test takers will need to draw on the following strategic competences to varying degrees:** Planning: setting up expectations Execution: identifying cues and inferring from them Evaluation: Hypothesis testing, matching cues to expectations Repair: Revising hypothesis
8 After reading the scale for Strategic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	**Level** ✓ Unknown to the researchers
	Justification (incl. reference to documentation) ✓ Unknown to the researchers

C. Content Analysis Grids

CEFR Content Analysis Grid for Listening

The *CEFR Content Analysis Grid for Listening & Reading*[3] allows test developers to analyse tests of reading and listening in order to relate them to the CEFR. Information about each task, text and item in the test is entered into the Grid by specifying their characteristics (e.g. text source, discourse type, estimated difficulty level, etc.) from a range of options derived from the CEFR.

[3] The Grid was produced by a working group consisting of J. Charles Alderson (Project Coordinator) Neus Figueras, Henk Kuijpers, Günther Nold, Sauli Takala and Claire Tardieu. With further funding from the Dutch Ministry of Education the group developed a computerised version which is available at www.lancs.ac.uk/fss/projects/grid A report on the project is available on request from the Project Coordinator at c.alderson@lancaster.ac.uk

Listening Comprehension in English – GEPT Elementary Level			
Test section number: 1		**Total test length:** 20 mins.	
Item types	MC		
Items	i1+i4	i2+i3	i5
Source	Unknown to the analyst		
Authenticity	scripted	scripted	scripted
Discourse type	Descriptive MC options	Descriptive MC options	Descriptive MC options
Domain	Public	Personal	Public
Topic	Food	Directions/Location	Sports/Visual literacy
Number of speakers	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete
Grammar	Simple	Simple	Simple
Vocabulary	Only frequent	Mostly frequent	Mostly frequent
Nr of listening	1	1	1
Input text compre-hensible at level	A2	A2	A2
Items comprehensible at level (item codes)			
A1			
A1/A2			
A2	i1, i4	i2, i3	i5
A2/B1			
B1			
B1/B2			
B2			
B2/C1			
C1			
C1/C2			
C2			

Listening Comprehension in English – GEPT Elementary Level										
Test section number: 2					**Total test length:** 20 mins.					
Item types	MC									
Item numbers	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	A question	A question	A descriptive statement	A question	A descriptive statement	A question	A question	A descriptive statement	A question	A question
Domain	Personal	Personal	Personal	Personal	Personal	Personal	Personal	Personal	Personal	Personal
Topic	Travel	Clothing	Entertainment	Clothing	Health	Interaction	Entertainment	Clothing	Work	Health
Number of speakers	1	1	1	1	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Simple	Simple	Simple	Simple	Simple	Simple	Simple	Simple	Simple	Simple
Vocabulary	Only frequent	Mostly frequent	Mostly frequent	Only frequent	Mostly frequent	Only frequent	Mostly frequent	Mostly frequent	Only frequent	Mostly frequent
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text comprehensible at level	A1	A1	A1	A1	A1/2	A1	A1	A1	A1	A2
Items comprehensible at level (item codes)										
A1	i6	i7				i11			i14	
A1/A2			i8	i9	i10		i12	i13		i15
A2										
A2/B1										
B1										
B1/B2										
B2										
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT Elementary Level										
Test section number: 3						**Total test length:** 20 mins.				
Item types	MC									
Item numbers	i16	i17	i18	i19	i20	i21	i22	i23	i24	i25
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation	Q&A conversation
Domain	Occupational	Personal	Personal	Personal	Personal	Personal	Personal	Personal	Educational	Personal
Topic	Work	Shopping	Entertainment	Family	Travel	Animal	Food & shopping	Clothing	School	School
Number of speakers	2	2	2	2	2	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Mostly simple	Simple	Mostly simple	Simple	Mostly simple	Mostly simple	Mostly simple	Simple	Mostly simple	Mostly simple
Vocabulary	Only frequent	Only frequent	Mostly frequent	Only frequent	Mostly frequent	Mostly frequent	Mostly frequent	Mostly frequent	Mostly frequent	Mostly frequent
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	A2	A2	A2	A1	A1/2	A2	A2	A2	A2	A2
Items comprehensible at level (item codes)										
A1		i17	i18		i20					
A1/A2	i16					i21	i22	i23		
A2				i19						i25
A2/B1									i24	
B1										
B1/B2										
B2										
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT Elementary Level					
Test section number: 4			**Total test length:** 20 mins.		
Item types	MC				
Items	i26	i27	i28	i29	i30
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	Descriptive	Descriptive	Descriptive	Descriptive	Descriptive
Domain	Personal	Public	Educational	Public	Public
Topic	Travel	Things	School	Entertainment	Entertainment
Number of speakers	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Simple	Mostly simple	Mostly simple	Mostly simple	Mostly simple
Vocabulary	Mostly frequent	Mostly frequent	Somewhat extensive	Somewhat extensive	Somewhat extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	A2	A2	A2/B1	A2	A2
Items comprehensible at level (item codes)					
A1					
A1/A2	i26				
A2		i27	i28	i29	i30
A2/B1					
B1					
B1/B2					
B2					
B2/C1					
C1					
C1/C2					
C2					

Elementary level preliminary cut-offs: <A1: 0; A1: 1-18; A2: 19-30

Notes on the Elementary Level items:

- Items 1- 5: Listening is limited to listening to the MC options.
- Items 6-15: Listening is limited to listening to the MC stem.
- Items 1-15: It is difficult to allocate topics & domains due to the very limited amount of listening input.
- Items 26-30: The picture description is spoken by one speaker, but the question and topic are introduced by another speaker. It is unclear whether this should be described as 1 or 2 speakers.

Listening Comprehension in English – GEPT Intermediate Level										
Test section number: 1						**Total test length:** 30 mins.				
Item types	MC									
Item numbers	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options
Domain	Public	Public	Personal	Public	Public	Public	Public	Public	Public	Public
Topic	Food	Food	Clothing	Shopping	Shopping	Entertainment	Entertainment	Entertainment	Entertainment	Directions
Number of speakers	1	1	1	1	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Simple	Simple	Simple	Simple	Mostly simple	Simple	Mostly simple	Simple	Simple	Simple
Vocabulary	Only frequent	Only frequent	Mostly frequent	Mostly frequent	Mostly frequent	Somewhat extensive	Mostly frequent	Somewhat extensive	Mostly frequent	Frequent
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	A2	A1/2	A2	A2/B1	A2/B1	B1	A2	A2/B1	A2/B1	A2
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2	i1	i2	i3					i8	i9	i10
A2/B1				i4	i5	i6	i7			
B1										
B1/B2										
B2										
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT Intermediate Level					
Test section number: 1			**Total test length:** 30 mins.		
Item types	MC				
Items	i11	i12	i13	i14	i15
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options	Descriptive MC options
Domain	Public	Public	Public	Professional	Occupational
Topic	Entertainment	Numbers	Comparatives	Numbers	Numbers
Number of speakers	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Simple	Mostly simple	Mostly simple	Mostly simple	Mostly simple
Vocabulary	Somewhat extensive	Mostly frequent	Mostly frequent	Extensive	Extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	A2/B1	A2/B1	A2/B1	B1/B2	B1/B2
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1	i11	i12	i13		
B1				i14	i15
B1/B2					
B2					
B2/C1					
C1					
C1/C2					
C2					

Listening Comprehension in English – GEPT Intermediate Level										
Test section number: 2							**Total test length:** 30 mins.			
Item types	MC									
Item numbers	i16	i17	i18	i19	i20	i21	i22	i23	i24	i25
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	A question	A question	A question	A question	A question	A question	A question + a descriptive statement	Exclamatory statement	A question	A descriptive statement
Domain	Personal	Personal	Personal	Public	Public	Public	Personal	Personal	Personal	Personal
Topic	Social life	Shopping	Entertainment	Sports	Social life	Transportation	Clothing	Clothing	Social life	School
Number of speakers	1	1	1	1	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Mostly simple	Mostly simple	Mostly simple	Simple	Somewhat complex	Mostly simple	Mostly simple	Simple	Mostly simple	Simple
Vocabulary	Only frequent	Mostly frequent	Mostly frequent	Mostly frequent	Somewhat extensive	Somewhat extensive	Mostly frequent	Only frequent	Only frequent	Mostly frequent
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	A1/A2	A2	A2	A2	B1	B1	A2/B1	A2	A2	A2
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2								i23		
A2/B1	i16	i17							i24	
B1			i18	i19	i20	i21	i22			i25
B1/B2										
B2										
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT Intermediate Level					
Test section number: 2			**Total test length:** 30 mins.		
Item types	MC				
Items	i26	i27	i28	i29	i30
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	A question	A descriptive statement	A descriptive statement	A question	A question
Domain	Personal	Personal	Educational	Public	Personal
Topic	Travel	Entertainment	School	Entertainment	Family
Number of speakers	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Simple	Mostly simple	Mostly simple	Somewhat complex	Simple
Vocabulary	Only frequent	Mostly frequent	Extensive	Somewhat extensive	Somewhat extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	A2	A2/B1	B1	B1	B1
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2	i26				
A2/B1					
B1		i27	i28	i29	i30
B1/B2					
B2					
B2/C1					
C1					
C1/C2					
C2					

Listening Comprehension in English – GEPT Intermediate Level										
Test section number: 3						**Total test length:** 30 mins.				
Item types	MC									
Item numbers	i31	i32	i33	i34	i35	i36	i37	i38	i39	i40
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Conversation	Conversation	Telephone conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation
Domain	Public	Personal	Public	Educational	Personal	Public	Public	Public	Personal	Personal
Topic	Clothing	Social life	Shopping	Entertainment	Sports	Entertainment	Transportation	Directions	Health	Social life
Number of speakers	2	2	2	2	2	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Mostly simple	Mostly simple	Mostly simple	Mostly simple	Mostly simple	Somewhat complex	Somewhat complex	Mostly simple	Mostly simple	Mostly simple
Vocabulary	Somewhat extensive	Mostly frequent	Mostly frequent	Mostly frequent	Mostly frequent	Somewhat extensive	Mostly frequent	Somewhat extensive	Somewhat extensive	Somewhat extensive
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	B1	A2/B1	A2/B1	A2/B1	A2/B1	B1	B1	A2/B1	B1	B1
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2										
A2/B1	i31	i32	i33		i35		i37		i39	
B1				i34		i36		i38		i40
B1/B2										
B2										
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT Intermediate Level					
Test section number: 3			**Total test length:** 30 mins.		
Item types	MC				
Items	i41	i42	i43	i44	i45
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	Conversation	Conversation	Conversation	Conversation	Conversation
Domain	Personal	Public	Occupational	Public	Public
Topic	Weather	Entertainment	Work	Social life	Food
Number of speakers	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Somewhat complex	Somewhat complex	Somewhat complex	Somewhat complex	Mostly simple
Vocabulary	Somewhat extensive	Mostly frequent	Extensive	Somewhat extensive	Somewhat extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	B1	A2/B1	B1	B1	B1
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1					
B1	i41	i42	i43		
B1/B2				i44	i45
B2					
B2/C1					
C1					
C1/C2					
C2					

Intermediate level preliminary cut-offs: < A2: 0; A2: 1–24; B1: 25-45

Notes on the Intermediate Level items:

- Items 1-15: Listening is limited to listening to the MC stem and MC options.
- Item 2: This item requires mathematical skills.
- Items 8-15: These items require reading skills.
- Items 16-30: The listening input is very limited – a question or a one-sentence statement. The listening is (relatively) easy, but the difficulty mostly lies in the reading of the MC options (or sometimes the lexis used in the oral question/statement). This made it very difficult to link the items to the CEFR.

Listening Comprehension in English – GEPT High-Intermediate Level										
Test section number: 1					**Total test length:** 35 mins.					
Item types	MC									
Item numbers	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	A question	A question	A descriptive statement	A question	A descriptive statement + a command	A descriptive statement + a question	A question	A question	A question	A question
Domain	Personal	Public	Personal	Occupational	Occupational	Public	Occupational	Personal	Public	Public
Topic	Entertainment	Shopping	Health	Work	Shopping	Food	Work	Social life	Entertainment	Weather
Number of speakers	1	1	1	1	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Somewhat complex	Somewhat complex	Mostly simple	Simple	Mostly simple	Simple	Somewhat complex	Mostly simple	Mostly simple	Somewhat complex
Vocabulary	Mostly frequent	Mostly frequent	Somewhat extensive	Mostly frequent	Somewhat extensive	Somewhat extensive	Extensive	Somewhat extensive	Somewhat extensive	Somewhat extensive
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	B1	B1	B1/B2	B1	B1/B2	B1/B2	B2	B1/B2	B1/B2	B2
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2										
A2/B1										
B1										
B1/B2	i1	i2	i3	i4	i5	i6	i7	i8		i10
B2									i9	
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT High-Intermediate Level					
Test section number: 1			**Total test length:** 35 mins.		
Item types	MC				
Items	i11	i12	i13	i14	i15
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	A question	A descriptive statement	A question	A question	A statement
Domain	Public	Occupational	Public	Public	Personal
Topic	Shopping	Work	Comparatives	Entertainment	Health
Number of speakers	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Somewhat complex	Somewhat complex	Somewhat complex
Vocabulary	Extensive	Extensive	Somewhat extensive	Somewhat extensive	Extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	B2	B2	B1/B2	B2	B2
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2			i13		
B2	i11	i12		i14	i15
B2/C1					
C1					
C1/C2					
C2					

Listening Comprehension in English – GEPT High-Intermediate Level										
Test section number: 2						**Total test length:** 35 mins.				
Item types	MC									
Item numbers	i16	i17	i18	i19	i20	i21	i22	i23	i24	i25
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation
Domain	Public	Public	Personal	Personal	Personal	Personal	Occupational	Occupational	Personal	Personal
Topic	Food	Science	Social life	Health	Shopping	Clothing	Work	Work	Weather	Health
Number of speakers	2	2	2	2	2	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Somewhat complex	Somewhat complex	Somewhat complex	Complex	Somewhat complex	Somewhat complex	Complex	Complex	Complex	Somewhat complex
Vocabulary	Extensive	Extensive	Somewhat extensive	Extensive	Extensive	Extensive	Extensive	Somewhat extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	B1/B2	B1/B2	B1/B2	B1/B2	B1/B2	B1/B2	B2	B1/B2	B2	B2
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2										
A2/B1										
B1										
B1/B2	i16	i17	i18		i20	i21		i23		
B2				i19			i22		i24	i25
B2/C1										
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT High-Intermediate Level					
Test section number: 2			**Total test length:** 35 mins.		
Item types	MC				
Items	i26	i27	i28	i29	i30
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	Telephone conversation	Conversation	Conversation	Conversation	Conversation
Domain	Occupational	Public	Personal	Educational	Personal
Topic	Work	Travel	Social life	Science	Weather
Number of speakers	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Somewhat complex	Somewhat complex	Complex	Complex
Vocabulary	Somewhat extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	B2	B2	B2	B2/C1	C1
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2	i26		i28		
B2		i27			
B2/C1				i29	
C1					i30
C1/C2					
C2					

Listening Comprehension in English – GEPT High-Intermediate Level										
Test section number: 3						**Total test length:** 35 mins.				
Item types	MC									
Item numbers	i31	i32	i33	i34	i35	i36	i37	i38	i39	i40
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Announcement	Announcement	Lecture	Lecture	Lecture	Telephone message	Telephone message	Telephone message	Commercial	Conversation
Domain	Public	Public	Educational	Educational	Educational	Public	Public	Public	Personal	Personal
Topic	Entertainment	Entertainment	People	People	People	Shopping	Shopping	Shopping	Health	Health
Number of speakers	2	2	2	2	2	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Somewhat complex	Somewhat complex	Complex	Complex	Complex	Complex	Complex	Complex	Complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	B2	B2	C1	C1	C1	B2	B2	B2	C1	C1
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2										
A2/B1										
B1										
B1/B2										
B2	i31	i32					i37	i38		
B2/C1			i33	i34	i35	i36			i39	i40
C1										
C1/C2										
C2										

Listening Comprehension in English – GEPT High-Intermediate Level					
Test section number: 3			**Total test length:** 35 mins.		
Item types	MC				
Items	i41	i42	i43	i44	i45
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	News report	News report	News report	Presentation	Presentation
Domain	Educational	Educational	Educational	Occupational	Occupational
Topic	Health	Health	Health	Work	Work
Number of speakers	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Complex	Complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	B2/C1	B2/C1	B2/C1	B2/C1	B2/C1
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2					
B2					
B2/C1	i41	i42		i44	
C1			i43		i45
C1/C2					
C2					

High Intermediate level preliminary cut-offs: <B1: 0; B1: 1-18; B2: 19–42; >B2: 43-45

Notes on the High Intermediate level items:

- Items 1-15: Listening is limited to listening to the MC stem. The listening input is very limited – a question. The listening is (relatively) easy, but the difficulty mostly lies in the reading of the MC options (or sometimes the lexis used in the oral question/statement). This made it very difficult to link the items to the CEFR.
- Items 16-30: The difficulty often resides largely in the lexis in the listening input. The difficulty also sometimes lies in the reading (lexis) of the MC options. At the same time, the input is spoken in a very clearly articulated, acted manner. This sometimes makes it difficult to link the items to the CEFR higher level descriptors.

Listening Comprehension in English – GEPT Advanced Level										
Test section number: 1						**Total test length:** 45 mins.				
Item types	MC									
Item numbers	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
Source	Unknown to the analyst									
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Conversation	Lecture	Advertisement	TV program
Domain	Public	Educational	Personal	Personal	Occupational	Public	Public	Educational	Public	Educational
Topic	Health	Work	Environment	Money	Work	Entertainment	Sports	Health	Money	Science
Number of speakers	1	1	1	1	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Somewhat complex	Complex	Complex	Complex	Complex	Complex	Complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1	1	1	1	1	1
Input text compre-hensible at level	B2	B2	C1	B2/C1	B2/C1	B2/C1	B2/C1	B2/C1	B2/C1	C1
Items comprehensible at level (item codes)										
A1										
A1/A2										
A2										
A2/B1										
B1										
B1/B2										
B2							i7			
B2/C1		i2		i4				i8		i10
C1	i1		i3		i5	i6			i9	
C1/C2										
C2										

Listening Comprehension in English – GEPT Advanced Level					
Test section number: 1			**Total test length:** 45 mins.		
Item types	MC				
Items	i11	i12	i13	i14	i15
Source	Unknown to the analyst				
Authenticity	Scripted	scripted	scripted	scripted	scripted
Discourse type	News report	Narrative recording	News report	Announcement	Lecture
Domain	Public	Personal	Public	Public	Educational
Topic	Travel	Work	Shopping	Entertainment	History
Number of speakers	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Complex	Complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	B2/C1	B2/C1	C1	C1	B2/C1
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2					
B2					
B2/C1	i11				
C1		i12	i13	i14	i15
C1/C2					
C2					

Listening Comprehension in English – GEPT Advanced Level					
Test section number: 2			**Total test length:** 45 mins.		
Item types	Short answer	Short answer	MC	Short answer	Short answer
Items	i16	i17	i18	i19	i20
Source	Unknown to the analyst				
Authenticity	scripted	scripted	scripted	scripted	scripted
Discourse type	Conversation	Conversation	Conversation	Conversation	Conversation
Domain	Public	Public	Public	Public	Public
Topic	Work	Work	Work	Work	Work
Number of speakers	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Somewhat complex	Complex	Complex	Complex	Complex
Vocabulary	Somewhat extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1
Input text comprehensible at level	B2	B2/C1	B2/C1	C1	C1
Items comprehensible at level (item codes)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2					
B2	i16				
B2/C1		i17	i18		
C1				i19	i20
C1/C2					
C2					

Listening Comprehension in English – GEPT Advanced Level							
Test section number: 2				**Total test length:** 45 mins.			
Item types	Short answer	Short answer	Short answer	MC	Short answer	Short answer	Short answer
Item numbers	i21	i22	i23	i24	i25	i26	i27
Source	Unknown to the analyst						
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Radio interview	Radio interview	Radio interview	Radio interview	Radio interview	Radio interview	Radio interview
Domain	Educational	Educational	Educational	Educational	Educational	Educational	Educational
Topic	Archeology	Archeology	Archeology	Archeology	Archeology	Archeology	Archeology
Number of speakers	2	2	2	2	2	2	2
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Complex	Complex	Somewhat complex	Somewhat complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1	1	1
Input text compre-hensible at level	C1	C1	C1	C1	C1	C1	C1
Items comprehensible at level (item codes)							
A1							
A1/A2							
A2							
A2/B1							
B1							
B1/B2							
B2							
B2/C1			i23				i27
C1				i24	i25	i26	
C1/C2		i22					
C2	i21						

Listening Comprehension in English – GEPT Advanced Level						
Test section number: 3			**Total test length:** 45 mins.			
Item types	Short answer	Short answer	MC	Short answer	Short answer	Short answer
Item numbers	i28	i29	i30	i31	i32	i33
Source	Unknown to the analyst					
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	News report	News report	News report	News report	News report	News report
Domain	Educational	Educational	Educational	Educational	Educational	Educational
Topic	Artefacts	Artefacts	Artefacts	Artefacts	Artefacts	Artefacts
Number of speakers	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Complex	Complex	Complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1	1
Input text compre-hensible at level	C1	B2/C1	C1	C1	C1	C1
Items comprehensible at level (item codes)						
A1						
A1/A2						
A2						
A2/B1						
B1						
B1/B2						
B2						
B2/C1		i29				
C1	i28		i30	i31	i32	
C1/C2						i33
C2						

Listening Comprehension in English – GEPT Advanced Level							
Test section number: 3				**Total test length:** 45 mins.			
Item types	Short answer	Short answer	Short answer	Short answer	Short answer	Short answer	MC
Item numbers	i34	i35	i36	i37	i38	i39	i40
Source	Unknown to the analyst						
Authenticity	scripted	scripted	scripted	scripted	scripted	scripted	scripted
Discourse type	Lecture	Lecture	Lecture	Lecture	Lecture	Lecture	Lecture
Domain	Educational	Educational	Educational	Educational	Educational	Educational	Educational
Topic	Smuggling	Smuggling	Smuggling	Smuggling	Smuggling	Smuggling	Smuggling
Number of speakers	1	1	1	1	1	1	1
Pronunciation	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE	Standard AmE
Content	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete	Concrete
Grammar	Complex	Complex	Complex	Complex	Complex	Complex	Complex
Vocabulary	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive
Nr of listening	1	1	1	1	1	1	1
Input text compre-hensible at level	C1	C1	C1	C1	C1	C1	C1
Items comprehensible at level (item codes)							
A1							
A1/A2							
A2							
A2/B1							
B1							
B1/B2							
B2							
B2/C1	i34				i38	i39	
C1			i36	i37			i40
C1/C2		i35					
C2							

Advanced level preliminary cut-offs: < B2: 0; B2: 1–15; C1: 16-39; >C1: 40

Notes on the Advanced level items:

- All items: The difficulty seems to reside largely in the length of the audio and the lexis and phraseology in the listening input. At the same time, the input is spoken in a very clearly articulated, acted manner. This makes it difficult to link the items to the CEFR higher level descriptors.

Appendix B – Standard Setting Timetable

Monday	
09:00 – 10:30	Introduction and familiarisation (1)
10:30 – 11:00	Tea/coffee break
11:00 – 12:00	Familiarisation (2)
12:00 – 13:30	Lunch
13:30 – 15:00	Standard setting listening test 1
15:00 – 15:30	Tea/coffee break
15:30 – 17:00	Standard setting listening test 1

Tuesday	
09:00 – 11:00	Standard setting listening test 2
11:00 – 11:30	Tea/coffee break
11:30 – 12:30	Standard setting listening test 2
12:30 – 14:00	Lunch
14:00 – 16:00	Standard setting listening test 3
16:00 – 16:30	Tea/coffee break
16:30 – 17:00	Standard setting listening test 3

Wednesday	
09:00 – 11:00	Standard setting listening test 3
11:00 – 11:30	Tea/coffee break
11:30 – 12:30	Standard setting listening test 4
12:30 – 14:00	Lunch
14:00 – 16:00	Standard setting listening test 4
16:00 – 16:30	Tea/coffee break
16:30 – 17:00	Standard setting listening test 4

Thursday	
09:00 – 11:00	Standard setting – mixed level item set
11:00 – 11:30	Tea/coffee break
11:30 – 12:00	Standard setting – mixed level item set
12:00 – 13:30	Lunch
13:30 – 15:30	Standard setting – mixed level item set
15:00 – 15:30	Tea/coffee break
15:30 – 17:00	Final plenary discussion

Appendix C – Sample Judgement Sheet

Write your ID number here:

Judgement sheet

Please provide your judgements on the CEFR level of each of the items in the table below. You will need to consider the questions:

1. **At which CEFR level can a test-taker already answer the item correctly?**
2. **Would a just-qualified (L), mid (M), or high (H) test-taker at that level already be able to answer the item correctly?**

Judgement 1: Please write down the CEFR level.
Remember to use the following codes:

A1 = 1
A2 = 2
B1 = 3
B2 = 4
C1 = 5
C2 = 6

Judgement 2: Please encircle.

Item	Judgement 1	Judgement 2		
1		**L**	**M**	**H**
2		**L**	**M**	**H**
3		**L**	**M**	**H**
4		**L**	**M**	**H**
5		**L**	**M**	**H**