

**APTIS–GEPT TEST COMPARISON STUDY:
LOOKING AT TWO TESTS FROM MULTI-PERSPECTIVES
USING THE SOCIO-COGNITIVE MODEL**

VS/2016/002

**Rachel Yi-fen Wu
Hsin-yi Yeh
Jamie Dunlea
Richard Spiby**

ABSTRACT

This present study investigates the comparability of Aptis, an English language assessment system developed by the British Council, with the GEPT (General English Proficiency Test), one of the most widely used English language tests in Taiwan.

To gather cross-test comparability evidence, both quantitative analysis of test-takers' scores on Aptis and the GEPT, and qualitative analysis of the textual features of the two tests were performed through CTT analyses, automated textual analyses, and expert judgment. Data were collected from 144 GEPT test-takers who had taken the Elementary, Intermediate, High-intermediate, or Advanced level of the test (equivalent to CEFR A2, B1, B2, and C1 respectively) within a six-month period before the administration of the Aptis test.

The results showed relatively high correlations between Aptis and GEPT scores across all subtests:

- the listening tests at .788 ($p < .01$)
- the reading tests at .733 ($p < .01$)
- the speaking tests at .842 ($p < .01$)
- the writing tests at .753 ($p < .01$).

A Principal Components Analysis also supported the conclusion that the different components of the two tests measure a common construct of language proficiency. Another area of comparison drew on the strong body of research relating each test to the CEFR. This would indicate that test-takers tended to achieve higher CEFR levels on the Aptis reading, listening, and writing tests than on the GEPT, and they tended to achieve lower CEFR levels on the Aptis speaking test than on the GEPT.

In terms of textual features, the two tests were distinct from each other in a number of ways, but were also consistent in general trends, such as a tendency for input texts to increase in the level of abstractness as the CEFR level increased. Two questionnaires were administered immediately after participants had taken the Aptis test. Findings from the two questionnaires administered immediately after participants had taken the Aptis test showed that most participants had positive attitudes toward the test. However, for the listening input, test-takers stated a preference for accents more familiar to the Taiwanese context.

In addition, most respondents recommended that more information be provided on the score report in order to enable them to understand their scores in comparison to other test-takers.

Authors

Rachel Yi-fen Wu

Rachel Yi-fen Wu has been closely involved in the research and development, test production, and validation of the General English Proficiency Test (GEPT). Her research interests include reading assessment, language test development and validation, and methodological approaches to linking examinations to the CEFR.

Hsin-yi Yeh

Hsin-yi Yeh graduated from the University of Newcastle upon Tyne with a MA in Linguistics and a MA in Chinese-English Translating and Interpreting. She has been involved in the research and test production of the General English Proficiency Test and other LTTC examinations.

Jamie Dunlea

Jamie Dunlea is a Senior Researcher for the Language Assessment Research Group at the British Council, based in London. He has worked on a range of language test development and validation projects with the British Council, as well as collaborating on projects with researchers, organisations, and ministries internationally. Jamie joined the British Council in 2013, and was previously Chief Researcher at the Eiken Foundation of Japan, a not-for-profit organisation which develops and administers EFL examinations in Japan.

Richard Spiby

Richard Spiby has been the Test Development Researcher for the receptive skills (reading and listening, together with grammar and vocabulary) since June 2016. His main responsibilities involve analysing operational data and revising and developing the receptive skills components of the Aptis test. Richard has previously worked in the UK and Turkey, mainly in the university sector, in test production, management and research. His particular areas of interest are cognitive processing in language assessment, strategy use in reading and listening and methods of testing vocabulary.

Acknowledgements

The authors wish to express their sincere appreciation to Dr Jessica Wu for her supervision. They also wish to thank Cecilia Liao, Matt Li and Judy Lo for their assistance with data analysis.

CONTENTS

1. BACKGROUND	6
1.1 Aptis	6
1.2 The GEPT	7
2. LITERATURE REVIEW	7
3. RESEARCH CONTEXT	9
4. METHODOLOGY	10
4.1 Participants	10
4.2 Instruments	11
4.3 Test administration	11
4.4 Data analysis	12
5. RESULTS	13
5.1 To what extent do Aptis scores correlate with GEPT scores?	13
5.1.1 Descriptive statistics	13
5.1.2 Correlation	14
5.1.3 Analysis of variance	14
5.1.4 Classification consistency	15
5.2 To what extent do the constructs of Aptis match those of the GEPT?	17
5.2.1 Factor analyses	17
5.2.2 Content analyses	18
5.3 What are test-taker perceptions of Aptis and the GEPT?	31
6. DISCUSSION AND CONCLUSIONS	34
7. LIMITATIONS AND RECOMMENDATIONS	36
REFERENCES	37
Appendix 1: Overview of Aptis and GEPT test components	39
Appendix 2: Test-taker Questionnaire A: Core+Reading+Listening	42
Appendix 3: Questionnaire B: Writing+Speaking+APTIS and GEPT	45
Appendix 4: Test-taker questionnaire: score report	48

LIST OF TABLES

Table 1: Composition of the sample	10
Table 2: Analysis methods utilised	12
Table 3: Descriptive statistics for Aptis and GEPT scores	13
Table 4: Correlations between the GEPT subtests and Aptis subtests	14
Table 5: Analysis of variance for Aptis and GEPT scores	15
Table 6: Agreement between GEPT and Aptis CEFR levels	16
Table 7: One-component PCA solution with Varimax rotation	18
Table 8: Criterial features addressed in the content analysis	19
Table 9: Number of tasks in the listening test at each level	20
Table 10: Results of the comparison between the GEPT and Aptis listening texts based on automated textual analyses	20
Table 11: Number of tasks in the reading test at each level	22
Table 12: Results of the comparison between the GEPT and Aptis reading texts based on automated textual analysis	22
Table 13: Number of tasks in the speaking test at each level	26
Table 14: Language functions of speaking tasks	28
Table 15: Number of tasks in the writing test at each level	29

LIST OF FIGURES

Figure 1: Scree plot of the relationship between the Eigenvalues and the number of components	17
Figure 2: Distribution of listening text domains	20
Figure 3: Degree of content knowledge specificity of listening texts	21
Figure 4: Degree of cultural specificity of listening texts	21
Figure 5: The nature of information in listening texts	22
Figure 6: Distribution of reading text domains	23
Figure 7: Distribution of reading text discourse modes	23
Figure 8: Degree of content knowledge specificity of reading texts	24
Figure 9: Degree of cultural specificity of reading texts	24
Figure 10: The nature of information in reading texts	25
Figure 11: Scope of text processing needed to respond to items	25
Figure 12: Distribution of speaking task domains	26
Figure 13: Degree of content knowledge specificity of speaking tasks	27
Figure 14: Degree of cultural specificity of speaking tasks	27
Figure 15: The nature of information in speaking tasks	27
Figure 16: Distribution of writing task domains	29
Figure 17: Degree of content knowledge specificity of writing tasks	30
Figure 18: Degree of cultural specificity of writing tasks	30
Figure 19: The nature of information in writing tasks	30
Figure 20: Participants' perception of which test is a better measure of their English ability	32
Figure 21: Participants' preference for tests that target the same or different levels of ability	32
Figure 22: Participants' preference for one-stage or two-stage tests	32
Figure 23: Participants' preference for an alternative mode of delivery of the GEPT	33
Figure 24: Participants' preference for an alternative mode of delivery of Aptis	33

1. BACKGROUND

As part of an effort to promote the Aptis test system to institutions in Taiwan, this study was supported by the British Council and the Language Training and Testing Center (LTTTC). The purpose of the study is two-fold. Firstly, to obtain evidence to support the use of Aptis test scores within Taiwan. This study, in collaboration with LTTTC and through comparison with the GEPT as an accepted local benchmark of EFL proficiency, thus provided the chance to not only obtain evidence to justify the uses of Aptis locally, but also to obtain feedback and insight into how the test might be further improved for the local context. Secondly, and more specifically, the study aimed to investigate the feasibility and usefulness of using Aptis to predict test-takers' performance on the General English Proficiency Test (GEPT). The GEPT is widely accepted for admission, placement and promotion purposes in Taiwan, and is a high-stakes, certificated proficiency exam, while Aptis is designed to maximise flexibility and efficiency in varied international contexts. Therefore, the two testing systems could potentially fulfil complementary roles for various assessment programs and goals. This study, thus, provided the chance to investigate the potential for such synergy between the two systems. To provide background for this study, this report will begin by introducing Aptis and the GEPT, and the institutions that have developed them: the British Council and the LTTTC.

1.1 Aptis

The Aptis test system has been developed by the British Council, the UK's international organisation for cultural relations and educational opportunities. However, it does not consist of a single test. It is rather an approach incorporating test design, development and delivery aspects within an integrated system to provide flexible English language assessment options for test users. This means that tests are developed within the Aptis system for various uses by different test users, but always according to the same theoretical principles of language test validation and quality assurance.

The main, or standard, variant within the Aptis test system is Aptis General, which is a test of general English language proficiency designed for test-takers who are 16 years old or more. The test is provided directly to organisations for internal use and is administered at times and locations decided by the test user. The test is not a certificated test and individuals do not apply to take a test directly. As no specific cultural background is specified in the test design, test content is developed to be appropriate for learners in a variety of contexts.

The test is computer-based (non-adaptive) and can measure all four skills, in addition to grammatical and vocabulary knowledge. Grammar and Vocabulary is offered as a core component in combination with other skills as required by the test user. Feedback includes scores reported on a 0–50 scale for each component, and proficiency level skill profiles for each of the four skill components on the six-level Common European Framework of Reference (CEFR) proficiency framework.

A key aspect of Aptis is that while it has been developed and validated for use in a number of English as a Second/Foreign Language (ESL/EFL) contexts, and is widely used internationally, the Aptis test development approach has from the outset emphasised the importance of understanding and, where necessary, adapting to the local context. An overview of the approach employed in the development and validation of the Aptis test system is given in O'Sullivan (2015).

1.2 The GEPT

The General English Proficiency Test (GEPT) has been developed and is administered by the Language Training and Testing Center (LTTC), a non-profit educational foundation registered with Taiwan's Ministry of Education. The stated mission of the LTTC is to meet the needs of Taiwan's social and economic development through research, development and administration in language training and testing. Accordingly, the LTTC is engaged in a variety of activities related to language learning, teaching and testing: publishing textbooks and test preparation materials; carrying out research; sponsoring and hosting workshops, seminars, and international conferences; and cooperating with local and international scholars and institutions.

The GEPT is a five-level criterion-referenced test used to identify whether test-takers have attained a specific level in English. The GEPT promotes a balanced English learning process, covering the four language skills of listening, reading, writing and speaking with the goal of improving the general English proficiency level of Taiwanese learners. This corresponds to Taiwan's English education framework, meeting the specific needs of English learners in Taiwan for self-assessment, and provides institutions or schools with a reference for evaluating the English proficiency levels of job applicants, employees, or students.

The first four levels (Elementary, Intermediate, High-Intermediate and Advanced) have been aligned to the CEFR (Common European Framework of Reference for Languages) levels A2, B1, B2 and C1 respectively (Green, Inoue & Nakatsuhara, forthcoming; Harding & Brunfaut, 2014; Knoch, forthcoming; Wu & Wu, 2010; Wu, 2014). At these four levels, the GEPT tests are administered in two modes. Test-takers can choose either to take all four components of the tests (listening, reading, writing and speaking) within a day, or to take the four components in two stages, listening and reading as the first stage, and writing and speaking as the second stage; in this case, they need to pass the first stage before moving onto the second. Most test-takers choose the latter.¹ A Superior Level integrated test is also available and administered upon request.

GEPT scores are recognised by more than 300 junior high schools and high schools, as well as hundreds of universities, private enterprises and government agencies in Taiwan. In recent years, it has also been adopted by an increasing number of universities worldwide, in Hong Kong, Japan, France, Germany, the UK and the U.S, as a means of measuring the English language ability of Taiwanese applicants for further study overseas. Since its first administration in 2000, the GEPT has tested more than 6.4 million EFL learners at over 100 different venues in Taiwan (www.gept.org.tw).

2. LITERATURE REVIEW

This literature review is intended to give a brief overview of the comparative test analysis and criterion-related validation studies most relevant for the research methodology and research questions in this study. In particular, attention will be paid to a number of studies comparing the GEPT to other English proficiency tests, since these studies represent those most directly related to the purpose of the present study. We will also touch on important work which has been done on deriving task specification frameworks, particularly in relation to the CEFR, as these frameworks provide an important source of overlap between the two tests. Finally, we will look briefly at the Aptis test development approach to identify those areas which have impacted directly on the methodology selected, particularly in relation to comparison of test constructs.

¹ At the time of the study, at these four levels, only the two-stage option was available. Test-takers needed to pass the first stage (listening and reading components) before moving onto the second (writing and speaking), and they needed to pass both stages to receive a certificate of achievement.

A seminal project in terms of developing a research framework for the qualitative and quantitative comparison of language tests is the Cambridge–TOEFL Comparability Study (Bachman, Davidson, Ryan & Choi, 1995). This study investigated the comparability of the FCE administered by Cambridge English Language Assessment and the paper-based version of the Test of English as a Foreign Language (TOEFL) administered by the Educational Testing Service (ETS). This research involved a qualitative content analysis of the test tasks and a quantitative analysis of test performance. The qualitative content analysis was conducted by expert judges using the Communicative Language Ability instrument, and the quantitative statistical analysis was conducted by analysing test-takers' performances. The results of the study suggested that the two tests generally measured similar language abilities.

Three studies on the comparability of the GEPT and other English language tests used quantitative research methods. Chin and Wu (2001) compared examinee performance on the GEPT and EIKEN, Japan's most widely used English-language testing program. Analysis of the test-takers' performance revealed that the first stages of the two tests were quite similar in terms of overall difficulty level. In addition, despite the apparent dissimilarities in the test formats of the EIKEN Grade 2 and GEPT Intermediate speaking tests, test-taker scores on the two tests were significantly correlated.

In another study, the LTTC (2003) conducted a concurrent validity study between the High-intermediate Level GEPT and two other major English tests: CBT TOEFL and the national test of English for university students in China (CET-6 or Band 6 of the College English Test). By comparing participants' scores on these tests, the researchers reached two conclusions. First, all subtests of the High-Intermediate GEPT correlated significantly with both CBT TOEFL and CET-6. Second, from a comparison of the relative difficulty of the High-Intermediate Level GEPT, CBT TOEFL, and CET-6, the researchers suggested that the listening test of CBT TOEFL was more difficult than that of both High-Intermediate GEPT and CET-6, while the High-Intermediate GEPT and CET-6 were comparable in terms of difficulty. As for the reading test, CBT TOEFL was considered to be the easiest, and CET-6 the hardest.

Weir et al. (2013) investigated the criterion-related validity of the reading and writing components of the Advanced level GEPT in terms of two types of evidence – cross-test comparability and predictive power. Cross-test comparability was investigated by analysing the relationships between the Advanced Level GEPT reading and writing scores and IELTS bands. The correlations obtained were found to be moderate to high. However, the findings suggested that it was harder for participants to pass the Advanced GEPT than to score 6.5 (indicative of CEFR C1 level) on IELTS. In terms of predictive validity, the Advanced level GEPT reading and writing scores were investigated with relation to test-takers' real-life academic performance on different writing tasks in their degree coursework and examinations. The analysis showed that the GEPT reading and writing scores correlated with the participants' real-life academic performances at .529 ($p < .01$), indicating that GEPT scores reflected real-life performance to a reasonable degree.

In contrast to the three studies described above, Wu (2014) adopted both qualitative and quantitative procedures, including automatic textual analysis tools, expert judgment, and an *a posteriori* empirical exploration of test performance, to compare the GEPT reading tests at CEFR B1 and B2 levels with the reading tests of the Cambridge main suite exams at B1 and B2 levels. The results indicated that the Intermediate GEPT and Preliminary English Test (PET), both of which target B1 level, were in general comparable, while the High-Intermediate GEPT and First Certificate in English (FCE), which target B2 level, exhibited greater differences, not only in terms of test results, but also in contextual features and cognitive processing operations. Test-takers also scored significantly higher on the High-Intermediate GEPT than on the FCE test. In this respect, both expert judgment and test-takers' self-reports suggested that the FCE was cognitively more challenging than the High-Intermediate GEPT. Wu's (2014) approach, outlined here, of comparing test results and the textual features of the tests was also adopted to investigate the relationships between Aptis and four levels of the GEPT in the present study.

Test construct and score performance were investigated by Brown et al. (2014), who used a range of quantitative analysis techniques to compare a set of EFL proficiency tests designed for a particular local context—the EIKEN tests in Japan—with a test of English for Academic Purposes (EAP) designed for use in diverse international contexts—the TOEFL iBT. The focus of their study was primarily to investigate techniques which would allow for the creation of a common score scale linking performance across the different EIKEN test levels for comparison with the single score scale of the TOEFL iBT. They used a Rasch model for equating and linking. However, they also employed a number of quantitative techniques relevant to this study, including correlation analyses and Principal Components Analysis to investigate the similarity of the underlying constructs across the different components of the two testing systems.

A number of studies in recent years have built on qualitative content analysis frameworks to facilitate the description and comparison of test tasks, particularly in relation to the CEFR. Alderson et al. (2006) proposed a series of content analysis grids for tests of reading and listening which have been adapted and expanded in a number of test evaluation and validation projects. These include Khalifa and Weir (2009) in relation to reading tasks and Geranpayeh and Taylor (2013) in relation to listening. The analysis frameworks proposed by Alderson et al. have been included in the Manual for Relating Examinations to the CEFR (Council of Europe, 2009), and thus facilitate aspects of content comparison in relation to the CEFR, and provide an important link to and source of comparison with the qualitative frameworks developed by Wu (2014).

The Aptis test system, in comparison to the GEPT, is still relatively recent, and the validation research agenda has focused on gathering evidence to justify the uses and interpretations of the test itself from contextual, cognitive and scoring validity perspectives, rather than on concurrent validity-type test comparison projects. For the purposes of this study, however, it is worth noting that an explicit model of test development, the socio-cognitive framework for language test development and validation (Weir, 2005; O'Sullivan & Weir, 2011), has driven task design and specification for Aptis, as described in O'Sullivan (2015). The analysis grids developed by Alderson et al. (2006) and the applications of the socio-cognitive model to task analysis in relation to the CEFR have all provided a rich source of task specification characteristics on which Aptis has been able to draw. Dunlea (2014) explained how these aspects have been applied to task specification in relation to the Aptis reading tests and described the overlap with many of the text analysis and qualitative analysis characteristics employed by Khalifa and Weir (2009), Geranpayeh and Taylor (2013) and Wu (2014).

3. RESEARCH CONTEXT

Currently in Taiwan, there is a growing need for individuals to demonstrate their level of English language proficiency for educational and occupational purposes. This is especially true for students at higher-education institutions. As part of an effort to improve the English proficiency of college graduates, the Ministry of Education (MoE) of Taiwan is encouraging universities and colleges to establish an exit requirement for which students must achieve a pass in a test of English in order to graduate. Moreover, the results of these tests have a further impact in that they are used to make judgments about the institutions themselves as well as the test-takers. Accordingly, the quality of a college or university is to be evaluated, taking into account the number of students who achieve a passing level in a test. The MoE specifies a score equivalent to at least CEFR B1 for university graduates, while students at technological and vocational colleges must demonstrate a minimum proficiency in English of CEFR A2 level (MoE, 2004). Following the MoE's announcement of this policy, several higher-education institutions in Taiwan have expressed the need for test instruments which will enable them to assign incoming first-year students to appropriate English classes. In this way, they can more effectively provide training for students to achieve the required English proficiency level. As a flexible and convenient means of placing students at an appropriate level with reference to the CEFR, the Aptis system is being considered for this purpose.

In order to enable institutions in Taiwan to make informed decisions about Aptis scores as a measure of L2 speakers' English proficiency for placement purposes, it is crucial that the criterion-related validity of Aptis be determined. For comparison, the GEPT was selected as the external criterion measure in this study. This is because it is widely used in Taiwan, and because it is among the few examinations supported by research in its claims of alignment with specific levels of the CEFR. It was envisaged that a systematic comparison of Aptis and the GEPT would support the use of Aptis for placement purposes for universities and colleges, as well as for a broader range of users in both the public and private sectors in Taiwan.

The research questions were as follows:

- RQ1. To what extent do Aptis scores correlate with GEPT scores?
- RQ2. To what extent do the constructs of Aptis match those of the GEPT?
- RQ3. What are test-taker perceptions of Aptis and the GEPT?

4. METHODOLOGY

The present study was designed to investigate the comparability of Aptis and the GEPT through a mixed-methods approach. To answer RQ1, test performances on all subtests of Aptis and the GEPT were compared through statistical analyses. For RQ2, the constructs of both tests were analysed using factor analysis and content analysis. RQ3, which concerned test-taker views on Aptis and the GEPT, was answered by examining the results of two post-test questionnaires.

4.1 Participants

A stratified sample of 2,086 test-takers who had taken the GEPT tests at different levels from the Elementary (CEFR A2) to the Advanced (CEFR C1) within a six-month period was invited to participate in the study. A total of 144 individuals volunteered to take part, comprising 31 at Elementary level (A2), 39 at Intermediate level (B1), 40 at High-Intermediate level (B2), and 34 at Advanced level (C1). Of the 144 participants, 33% had passed both the GEPT first stage (listening and reading) and second stage (writing and speaking) tests, 43% had passed the first stage but failed the second stage tests, and 24% had failed the first stage tests (see Table 1).

Table 1: Composition of the sample

GEPT level	No. of participants	Passed 1 st & 2 nd stages	Passed 1 st stage only	Failed 1 st stage
Elementary	31	13	12	6
Intermediate	39	15	16	8
High-Intermediate	40	14	17	9
Advanced	34	5	17	12
Total	144	47 (33%)	62 (43%)	35 (24%)

Analysis of the participant demographics revealed that 58% were male and 42% were female. The average age was 22, with the oldest being 53 years old and the youngest 13. More than half of the participants (52%) were between 15 and 19 years of age. Slightly less than half (48%) were senior high school students, 38% were college students or above, and 13% were junior high school students.

4.2 Instruments

A set of computer-based Aptis live test papers, one set of GEPT past test papers, from Elementary to Advanced level, and the specifications for each exam were used to investigate the relationship between the two exams in terms of the test-takers' performance and test constructs from CEFR A1 to C1 levels. The test format and structures of the two tests studied are described in Appendix 1.

In addition to the above-mentioned instruments, two post-test surveys were administered to the test-takers in Chinese. The first survey was designed to elicit participants' responses in four different sections and was conducted immediately after the test. Part A was related to computer use and preference for computer versus pen-and-paper tests; Part B to perceptions of the clarity, relevance and appropriacy of the Aptis grammar and vocabulary, reading and listening components; Part C to the same perceptions of the writing and speaking components; and Part D was related to overall preference for either Aptis or the GEPT. The second survey concerned participant attitudes towards the Aptis score report in terms of its clarity and towards their Aptis scores in terms of how representative they are of test-takers' actual ability. This survey was administered after the participants had received their score reports. All participants completed the first questionnaire, while 77 responded to the second survey.

4.3 Test administration

A computer-based Aptis General test, including listening, reading, speaking, writing, and grammar and vocabulary components, was administered to 144 participants on 18 May 2014, under operational testing conditions. Due to space limitations in the computer labs, the administration was separated into two sessions with approximately equal numbers of participants. In each session, the same Aptis papers were used.

The grammar and vocabulary, reading, and listening subtests were administered first. Then candidates took a 15-minute break after completing the attitudinal questionnaire on these subtests. The speaking and writing tests were administered in the second half of each session. Candidates then completed the attitudinal questionnaire on the writing and speaking tasks after those two tests.

The listening, reading, and grammar and vocabulary test responses were machine-scored while the speaking and writing test responses were single-rated by trained Aptis raters².

² All raters are required to pass accreditation following training. In addition, quality assurance is maintained through the use of Control Items (CIs). These items have been previously marked by a pool of senior raters. CIs are seeded into live performances by the computer system. When raters do not mark CIs within a predetermined tolerance range, they are automatically suspended from rating and their performance reviewed by the Examiner Network Manager. See O'Sullivan and Dunlea (2015) for more information on the quality control systems for ensuring scoring validity of the speaking and writing components in Aptis.

4.4 Data analysis

The study used both quantitative analysis and qualitative procedures to investigate the research questions. The analyses performed are listed in Table 2.

Table 2: Analysis methods utilised

Research questions	Focus	Data	Methods
RQ1: To what extent do Aptis scores correlate with GEPT scores?	Correlation between Aptis and GEPT scores	<ol style="list-style-type: none"> 1. GEPT and Aptis total test scores 2. GEPT and Aptis subtest scores 	<ol style="list-style-type: none"> 1. Descriptive statistics summary 2. Correlation analyses 3. Comparison of means by ANOVA 4. Classification consistency analyses
RQ2: To what extent do the constructs of Aptis match those of the GEPT?	Constructs of Aptis and the GEPT	<ol style="list-style-type: none"> 1. GEPT and Aptis scores 2. GEPT and Aptis specifications and test papers 	<ol style="list-style-type: none"> 1. Factor analyses 2. Content analyses
RQ3: What are test-taker perceptions of Aptis and the GEPT?	Test-taker perceptions of Aptis and the GEPT	Post-test questionnaire responses	Descriptive statistics summary

The quantitative analysis was performed using SPSS 20.0 to compare the test results on each Aptis component with candidates' GEPT test scores at the same CEFR level, and using ConQuest 2.0, which was employed to vertically link the GEPT Elementary, Intermediate, High-Intermediate and Advanced level listening and reading tests onto a common score scale.

The questionnaire responses were likewise analysed with SPSS 20.0. The qualitative analysis involved a comparison of the textual features of each test, employing WordSmith 5.0 (Scott, 2009), VocabProfile Version 4 (Cobb, 2002), and expert judgment.

In this study, the test specifications and one set of Aptis and GEPT test papers used in the study were analysed.

5. RESULTS

The following section will describe the results of the study, addressing each of the research questions in turn.

5.1 To what extent do Aptis scores correlate with GEPT scores?

5.1.1 Descriptive statistics

Descriptive statistics are reported in Table 3. To compare Aptis scores with GEPT scores from four different levels, the GEPT listening and reading scores were calibrated to a common vertical scale using the Item Response Theory (IRT) parameter estimates, and the raw scores of each participant were converted to logit scores. The GEPT speaking and writing scores were converted to levels on the CEFR scale. In this study, numeric scores of 1 to 5 were assigned to CEFR levels A1 to C1, respectively. Thus, test-takers who passed the Intermediate level of the GEPT speaking or writing test were categorised as B1 and were assigned a score of 3. Those who failed the level were categorised as one level below, A2, and were assigned a score of 2. The Aptis test results on each component were reported on the numerical scale of 0 to 50, which is used for reporting results on each of the Aptis components. Skewness and kurtosis for all the subtests shown in Table 3 were within ± 2 , indicating satisfactory univariate normal distributions.

The GEPT score data indicate that candidates generally scored higher on the listening test than on the reading test, and they scored higher on speaking than on writing. However, the results of the Aptis test indicate that the test-takers in this study received slightly higher scores on Aptis reading (40.21) than listening (37.69), and they received higher scores on Aptis writing (40.55) than speaking (31.93). It is also worth noting that, among all the Aptis subtests, the speaking component resulted in the lowest scores, and the writing component resulted in the highest scores.

Table 3: Descriptive statistics for Aptis and GEPT scores

	GEPT L	GEPT R	GEPT S	GEPT W	Aptis L	Aptis R	Aptis S	Aptis W	Aptis G&V
N	144	144	109	109	144	144	144	144	144
Mean	1.43	0.95	3.36	2.99	37.69	40.21	31.93	40.55	36.19
Median	1.36	0.97	3.00	3.00	40.00	44.00	31.50	42.00	37.00
Mode	2.26	0.13	3.00	2.00	44.00	50.00	30.00	42.00	35.00
S.D.	1.19	1.19	1.20	0.98	9.20	9.92	8.53	6.59	8.00
Skewness	0.03	0.20	-0.27	0.08	-0.86	-1.11	-0.14	-0.58	-0.72
Kurtosis	-0.21	-0.39	-0.82	-0.81	-0.21	0.48	-0.48	-0.23	0.08
Minimum	-1.347	-1.689	1	1	14	12	10	19	12
Maximum	4.129	4.081	5	5	50	50	48	50	50

5.1.2 Correlation

Table 4 shows the correlation coefficients for all possible combinations of the subtests of the GEPT and Aptis. The Aptis subtests correlated significantly with all the GEPT subtests at moderate to high levels. These results provide evidence to support concurrent criterion-related validity. It is interesting to note that Aptis speaking correlated highly with all four components of the GEPT; the correlations were even higher than those between Aptis and GEPT listening (.788 vs .802) and reading (.733 vs .806), and close to the correlation between Aptis and GEPT writing (.753 vs .737).

Table 4: Correlations between the GEPT subtests and Aptis subtests

	GEPT L	GEPT R	GEPT S	GEPT W	Aptis L	Aptis R	Aptis S	Aptis W	Aptis G&V
Aptis L	.788**	.715**	.735**	.589**	1				
Aptis R	.689**	.733**	.679**	.636**	.732**	1			
Aptis S	.802**	.806**	.842**	.737**	.693**	.684**	1		
Aptis W	.761**	.771**	.786**	.753**	.703**	.696**	.768**	1	
Aptis G&V	.775**	.835**	.811**	.805**	.670**	.777**	.790**	.766**	1

5.1.3 Analysis of variance

One-way ANOVA tests were conducted on four test components to find out if participants at one CEFR level, as determined by their Aptis test results, performed significantly differently from those at another level on the GEPT. However, no analysis was performed on the Aptis grammar and vocabulary scores because the GEPT does not have a comparable test component. The CEFR levels categorised by Aptis were treated as the independent variable, and the GEPT scores were treated as the dependent variable.

The results (Table 5) showed that there was a significant difference in GEPT mean scores across different CEFR level groups ($p < .001$), with the GEPT mean scores increasing as the CEFR level increased. To illustrate, 18 participants were awarded B1 on Aptis listening. Their average GEPT logit score was 0.07, which was higher than that of the A2 group, but lower than those of the B2 and C groups. However, post hoc multiple comparisons using the Scheffe test indicated that not all CEFR groups differed significantly from other groups in their GEPT mean scores. The Aptis listening, reading and writing tests could distinguish between C, B2, and B1 groups successfully, but the GEPT mean scores of the A1, A2, and B1 groups were not significantly different from one another. The speaking test could only distinguish among B2, B1, and A2.

These findings provided partial evidence in support of the proposition that Aptis can differentiate between learners of different levels of English proficiency.

Table 5: Analysis of variance for Aptis and GEPT scores

Aptis Subtests	Aptis		GEPT				Post hoc
	CEFR	N	Mean	SD	F	p	
Listening	A2	8	-0.47	0.51	52.62	<0.001	C>B2>B1=A2
	B1	18	0.07	0.74			
	B2	29	0.86	0.74			
	C1	89	2.06	0.88			
Reading	A1	4	-0.84	0.60	52.15	<0.001	C>B2>B1=A2=A1
	A2	12	-0.55	0.48			
	B1	23	-0.11	0.57			
	B2	33	0.62	0.74			
	C1	72	1.80	0.86			
Speaking	A1	4	1.25	0.50	49.13	<0.001	C=B2>B1>A2=A1
	A2	20	2.00	0.80			
	B1	49	3.27	0.73			
	B2	31	4.39	0.72			
	C1	5	5.00	0.00			
Writing	B1	42	2.14	0.61	74.51	<0.001	C>B2>B1
	B2	28	3.11	0.74			
	C1	38	3.87	0.58			

5.1.4 Classification consistency

Cross-tabulations were performed to investigate the extent to which the CEFR levels reported by the GEPT matched those reported by Aptis. The exact agreement between the two tests ranged from .17 to .54, and the agreement within adjacent levels ranged from .80 to .96, as shown in Table 6.

In comparison with the GEPT, test-takers achieved higher CEFR levels in Aptis listening, reading, and writing, but lower CEFR levels in Aptis speaking. This might in part be due to test-takers' progress during the six-month period, between their taking the GEPT and the Aptis test.

Given the fact that Aptis was administered by computer, whereas the GEPT is a paper-based test, the mode of administration may be another factor contributing to such discrepancies. However, further investigation would be needed to verify this.

Table 6: Agreement between GEPT and Aptis CEFR levels

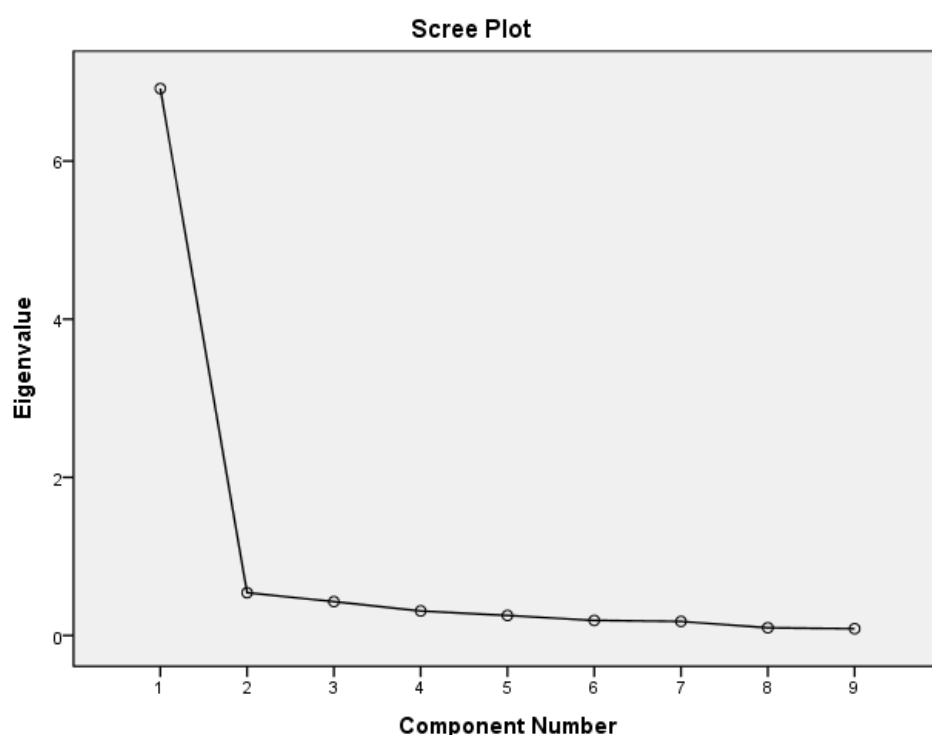
Test	GEPT CEFR	Aptis CEFR					Exact agreement rate across levels	Agreement rate within adjacent levels
		A1	A2	B1	B2	C		
Listening	A1	0	2	4	1	0	0.23	0.81
	A2	0	6	9	4	1		
	B1	0	0	5	19	18		
	B2	0	0	0	4	52		
	C1	0	0	0	1	18		
Reading	A1	1	4	2	2	0	0.23	0.80
	A2	3	8	17	10	0		
	B1	0	0	3	12	15		
	B2	0	0	1	9	45		
	C1	0	0	0	0	12		
Speaking	A1	3	5	0	0	0	0.54	0.96
	A2	1	11	6	1	0		
	B1	0	3	26	1	0		
	B2	0	1	15	14	0		
	C1	0	0	2	15	5		
Writing	A1	0	0	4	0	0	0.17	0.83
	A2	0	1	29	5	0		
	B1	0	0	8	16	9		
	B2	0	0	1	6	25		
	C1	0	0	0	1	4		

5.2 To what extent do the constructs of Aptis match those of the GEPT?

5.2.1 Factor analyses

To investigate the construct validity of Aptis, an exploratory factor analysis was performed to explore the variability among all the GEPT and Aptis subtests. Figure 1 suggests that a one-component solution would be appropriate.

Figure 1: Scree plot of the relationship between the Eigenvalues and the number of components



The one-component Principal Component Analysis (PCA) solution with Varimax rotation is shown in Table 7. These nine subtests accounted for 77.71% of variance, and all of the subtests exhibited factor loadings between .795 and .947 on one component. No loading was below 0.4, indicating that the GEPT and Aptis measure the same trait.

Table 7: One-component PCA solution with Varimax rotation

Subtests	Component 1
GEPT Listening	0.918
GEPT Reading	0.939
GEPT Speaking	0.947
GEPT Writing	0.883
Aptis Listening	0.804
Aptis Reading	0.795
Aptis Speaking	0.887
Aptis Writing	0.864
Aptis G&V	0.884

5.2.2 Content analyses

5.2.2.1 Content analysis proforma

A content analysis proforma was used to facilitate analyses of the textual features that affect the comprehensibility and difficulty of test tasks. The proforma was developed initially by the Aptis team, using the task specification grids used for item development and evaluation for operational Aptis tests, drawing on CEFR grids proposed by Alderson et al. (2006) and applications of the socio-cognitive model to test evaluation. This provided a great deal of overlap with the analysis frameworks developed and applied by Wu (2014). Thus, it was considered that the existing proforma would provide a principled basis for capturing criterial features of both tests relevant to the research questions, and ensure that both research teams would already have a number of precedents and a broad consensus on which to base their evaluation characteristics of their respective tests.

Two researchers each from the Aptis and GEPT teams were involved in evaluating their own tests using the proforma. An iterative, consensus-based approach was taken to refining the categories to be used and to interpreting those characteristics in practice. A face-to-face meeting was held in September 2014 between researchers from both teams to resolve differences in interpretation, discuss how the task characteristics to be evaluated through expert judgment were applied in practice for Aptis, and to select the most useful categories for the study. The GEPT team then reapplied the revised proforma to the GEPT tests, and the Aptis team to the Aptis test papers. The Aptis test content analysis was reviewed by the GEPT team to identify areas in which interpretations seemed to differ. Any such differences were resolved by the two teams through discussion and agreement via email.

It should be noted that, as the use of Aptis in the context of Taiwan was being examined, local interpretation, for example, in relation to cultural specificity, was prioritised. Of the original 42 features in the analysis, only 11 were used (Table 8) in this report since some features, such as “the number of speakers in the listening test”, and some others such as “the skills focused on in a test” were not amenable to a systematic comparison between the two test batteries.

The proforma was completed using both automated textual analysis tools and expert judgment. The automated tools, including WordSmith 5.0 (Scott, 2009) and VocabProfile Version 4 (Cobb, 2002), were used to analyse text length, sentence length and readability (Items 1, 2 and 4). The means of the three selected indices were reported and compared for the GEPT and Aptis test papers under review. The recording speed of listening tasks was based on the specifications of the tests.

Textual characteristics that were not measurable by the automated tools were analysed through expert judgment. The responses to the proforma were weighted on the basis of the tasks' contribution to the total score of the test in question. For the classification categories (Items 5, 6 and 11), percentages for each option were determined. For the rating categories (Items 7, 8 and 9), means of the four and five-point Likert scales were calculated while for Item 10, the raw results were reported.

Table 8: Criterial features addressed in the content analysis

	Listening	Reading	Speaking	Writing
1. Text length	V	V		
2. Average sentence length		V		
3. Speed (wps)	V			
4. Readability (FK Grade Level)	V	V		
5. Domain	V	V	V	V
6. Discourse mode		V		
7. Content knowledge	V	V	V	V
8. Cultural specificity	V	V	V	V
9. Nature of information	V	V	V	V
10. Language functions			V	
11. Scope of text content needed to process		V		

5.2.2.2 Content analysis results

The analysis was limited to four components: listening, reading, speaking, and writing. The grammar and vocabulary test of Aptis was not included since the GEPT does not have a comparable test component. The GEPT test items covered the A2, B1, B2 and C1 levels, and those of Aptis covered the A1, A2, B1 and B2 levels. Tasks with non-verbal input (i.e. Picture Description) or single-sentence input (i.e. Answering Questions and Sentence Completion) in the GEPT listening and reading tests were excluded from the examination of some criterial features, such as text length, average sentence length, readability, discourse mode, and the scope of text content that test-takers needed to process. The proforma included a category for CEFR level for each task.

As already noted, while GEPT tests are level-specific, with test content targeted at a specific proficiency band, Aptis is designed to cover a range of proficiency levels from CEFR A1 through to B2. The Aptis test specifications include detailed information for item writers on the specific criterial features and cognitive processing features for items at each CEFR level (see O'Sullivan and Dunlea, 2015, for detailed test specifications). In addition, during test construction, the number of tasks at each CEFR level is specified. For this review, however, the judges allocated each task to a CEFR level based on their own interpretation of the features in the task, rather than simply relying on the intended target level according to test construction documentation. This exercise was seen by the Aptis team as an opportunity to investigate the CEFR level allocation through expert judgment and the agreement with the original specification.

Listening

The GEPT listening tests included 19 to 38 tasks at each proficiency level while most Aptis listening tasks targeted A2 and B1 levels (Table 9). The number of tasks allocated to the B2 level in the Aptis listening test from a global expert judgment perspective was somewhat lower than the actual task and test specification, which generally calls for between 5–7 tasks at the B2 level (see O'Sullivan and Dunlea, 2015, for details).

Table 9: Number of tasks in the listening test at each level

	A1	A2	B1	B2	C1
GEPT	--	28	38	36	19
Aptis	5	7	12	1	--

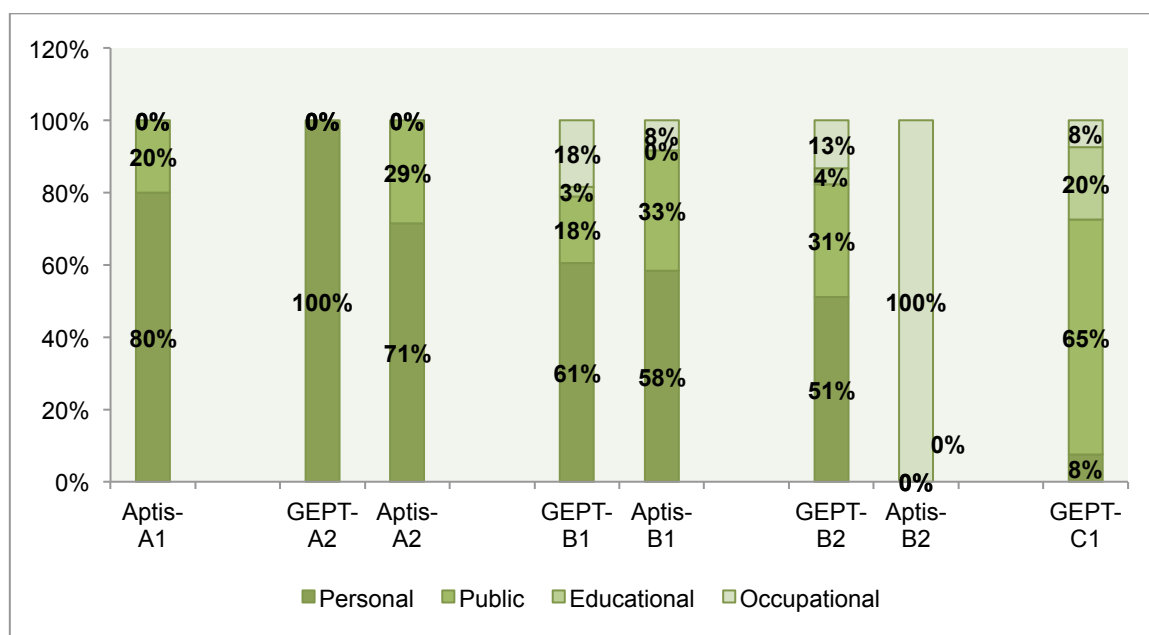
In general, the GEPT and Aptis used longer and more difficult texts at higher recording speeds as test level increased, with the exception of Aptis texts at B1 level. At A2 and B1 levels, the Aptis texts were much longer, their readability was lower, and the recording speed was faster than for the GEPT texts (see Table 10).

Table 10: Results of the comparison between the GEPT and Aptis listening texts based on automated textual analyses

	A1	A2		B1		B2		C1
	Aptis	GEPT	Aptis	GEPT	Aptis	GEPT	Aptis	GEPT
Text length	76.60	37.40	80.57	62.00	103.67	79.00	116.00	169.47
Flesch-Kincaid Grade Level	2.56	3.40	6.73	4.50	5.59	9.00	8.00	11.40
Speed (wps)	2.2–2.6	2.3	2.2–2.6	2.3	2.4–2.8	2.7–3.2	3.0–3.5	3.2

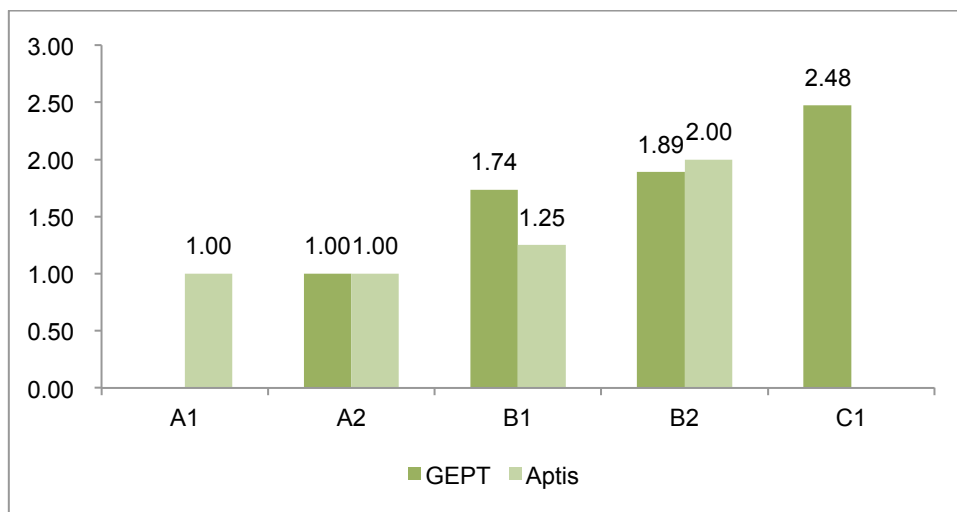
Based on the judges' responses to the content analysis proforma, the majority of the texts at A1 or A2 levels of both GEPT and Aptis were in the personal domain (Figure 2). But as the test level increased, the proportion of such texts decreased. For example, only 8% of the texts in GEPT-C1 were judged to be in the personal domain, while 65% of the texts were determined as being in the public domain. Overall, the GEPT encompassed texts in the personal, public, educational, and occupational domains but Aptis only included personal, public, and occupational texts.

Figure 2: Distribution of listening text domains



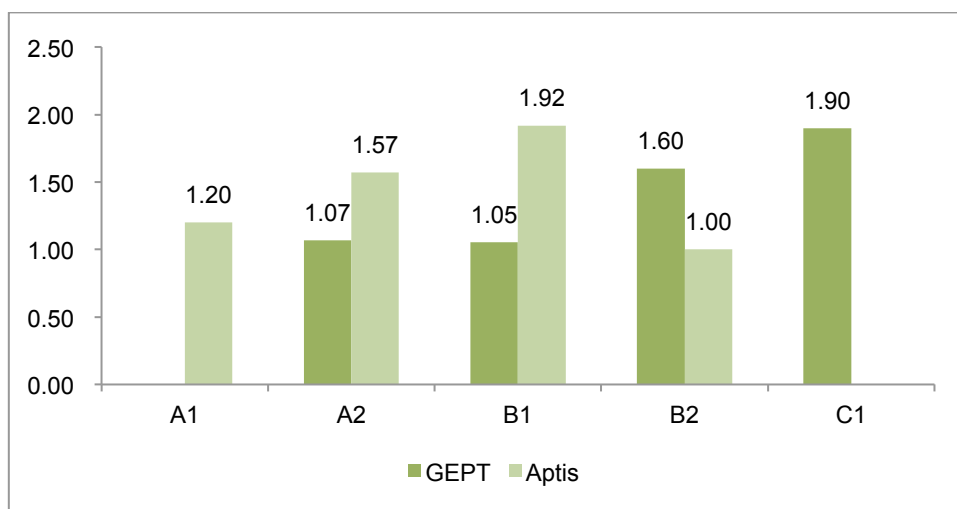
In terms of content knowledge specificity, cultural specificity and text abstractness (the extent to which information in the text refers to unobservable phenomena), responses from the expert judgment fell towards the lower end of the Likert scale. Generally speaking, as test level increased, both GEPT and Aptis texts became both more specific in content and more abstract (see Figure 3). However, compared with the GEPT at the same level, Aptis listening texts were more culturally specific at A2 and B1 levels and less specific at B2 level (Figure 4). Aptis texts were also considered to be more abstract than the GEPT texts at B1 and B2 levels (Figure 5).

Figure 3: Degree of content knowledge specificity of listening texts



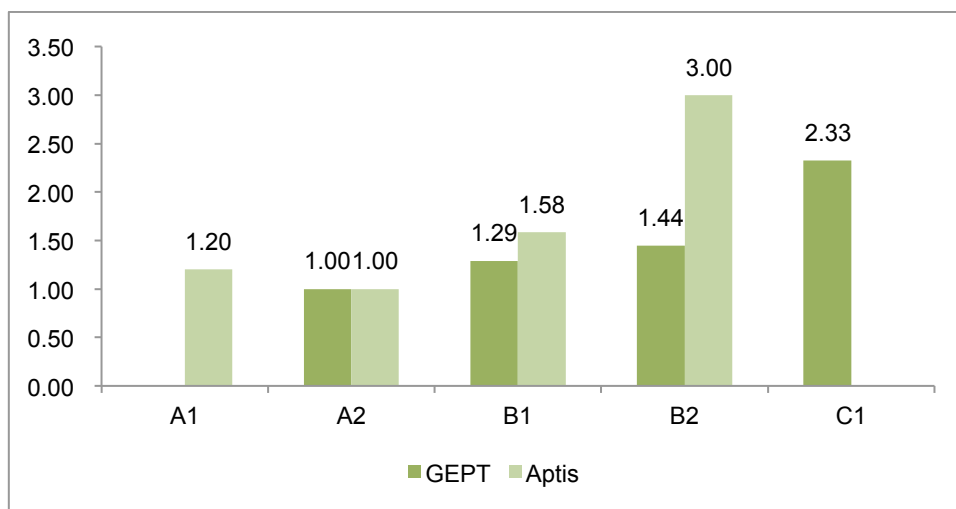
Key: 1 = general; 5 = specific

Figure 4: Degree of cultural specificity of listening texts



Key: 1 = general; 5 = specific

Figure 5: The nature of information in listening texts



Key: 1 = concrete; 4 = abstract

Reading

There were 7 to 22 tasks in the GEPT reading tests at each proficiency level, while each Aptis reading test included one task per level (Table 11).

Table 11: Number of tasks in the reading test at each level

	A1	A2	B1	B2	C1
GEPT	---	21	22	17	7
Aptis	1	1	1	1	---

The Flesch-Kincaid readability indices were similar for Aptis and GEPT reading texts at the same CEFR level, although the Aptis texts appeared to be more difficult than the GEPT texts at the B2 level (Table 12). One possible reason for this was that the Aptis B2 text was much longer than the GEPT texts. With the exception of the B1 text of Aptis, the average text length and sentence length in GEPT and Aptis passages rose as the CEFR level increased. Sentences used in Aptis were slightly longer than those in the GEPT at A2 and B2 levels.

Table 12: Results of the comparison between the GEPT and Aptis reading texts based on automated textual analysis

	A1	A2		B1		B2		C1
	Aptis	GEPT	Aptis	GEPT	Aptis	GEPT	Aptis	GEPT
Text length	99.00	116.33	97.00	142.00	136.00	217.57	697.00	765.14
Sentence length	8.00	10.95	14.00	13.79	13.60	16.28	19.20	17.94
Flesch-Kincaid Grade Level	3.10	5.80	5.50	9.30	9.10	10.90	13.10	12.00

Based on expert judgment, most GEPT reading texts belonged to the public domain, while Aptis included one personal text each at the A1 and A2 levels, one educational text at B1 level, and one public text at B2 level (Figure 6). The discourse modes of most GEPT texts were narrative or expository, whereas for Aptis they were descriptive and narrative (Figure 7). Neither the GEPT nor Aptis contained any occupational texts.

The expert judges considered the GEPT and Aptis reading texts similar to each other in content knowledge specificity, cultural specificity and text abstractness (Figures 8, 9 and 10, respectively). In general, the higher the test level, the more specific and abstract the texts were.

Figure 6: Distribution of reading text domains

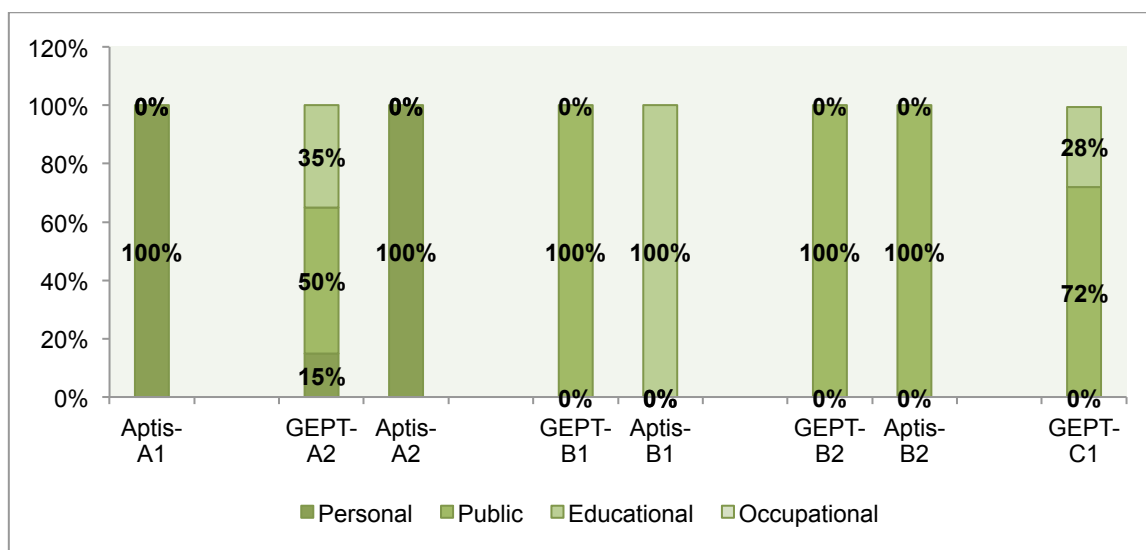


Figure 7: Distribution of reading text discourse modes

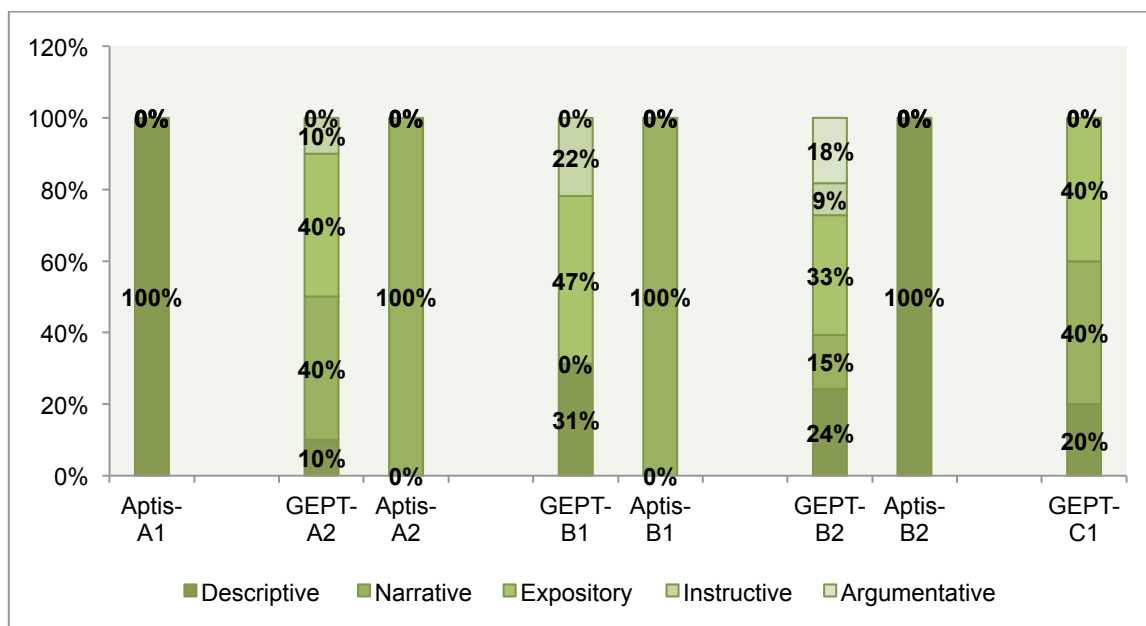
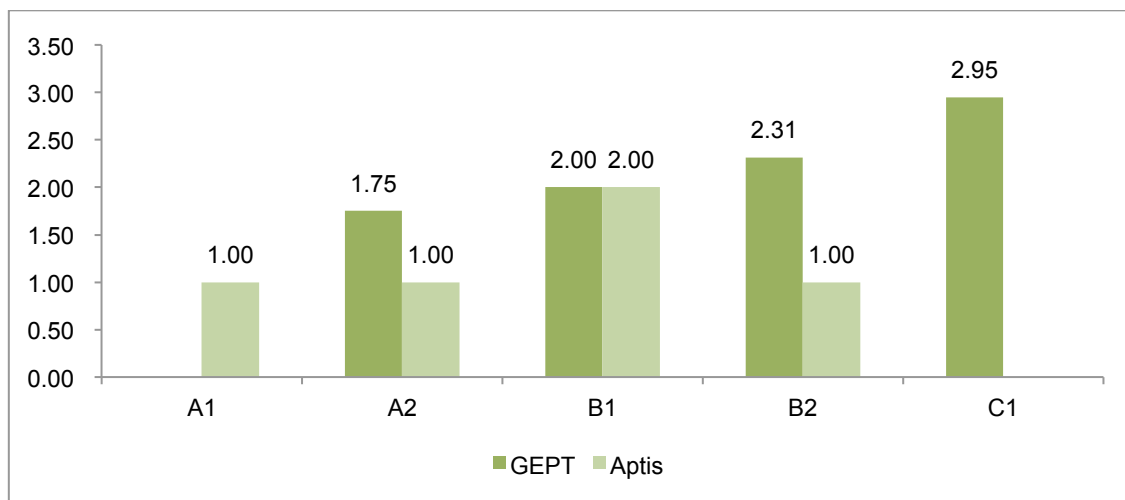
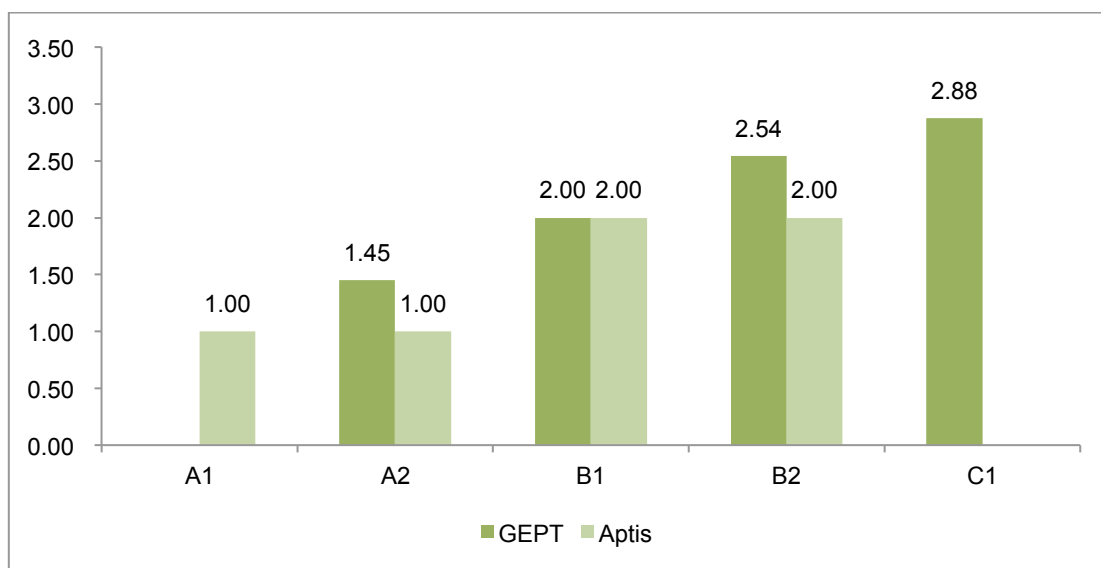


Figure 8: Degree of content knowledge specificity of reading texts



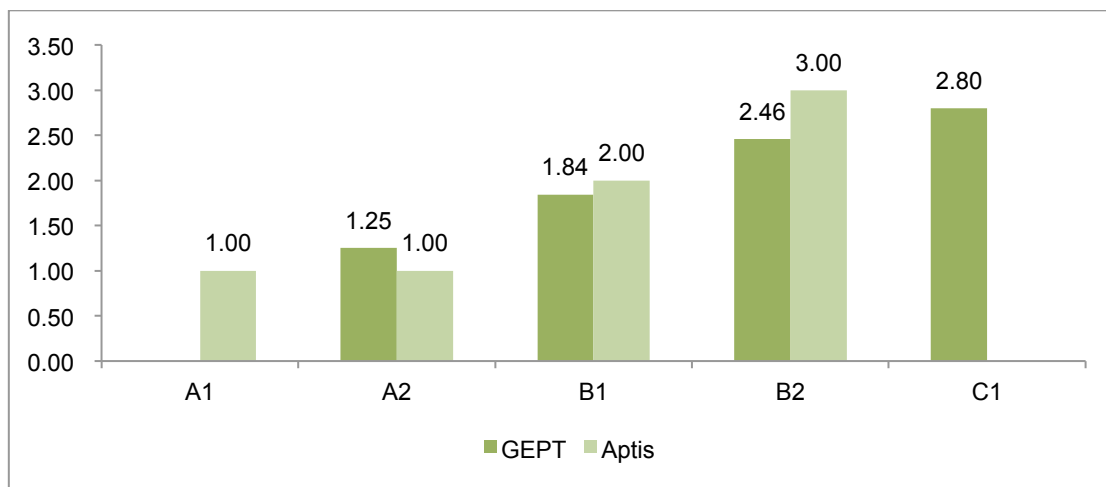
Key: 1 = general; 5 = specific

Figure 9: Degree of cultural specificity of reading texts



Key: 1 = general; 5 = specific

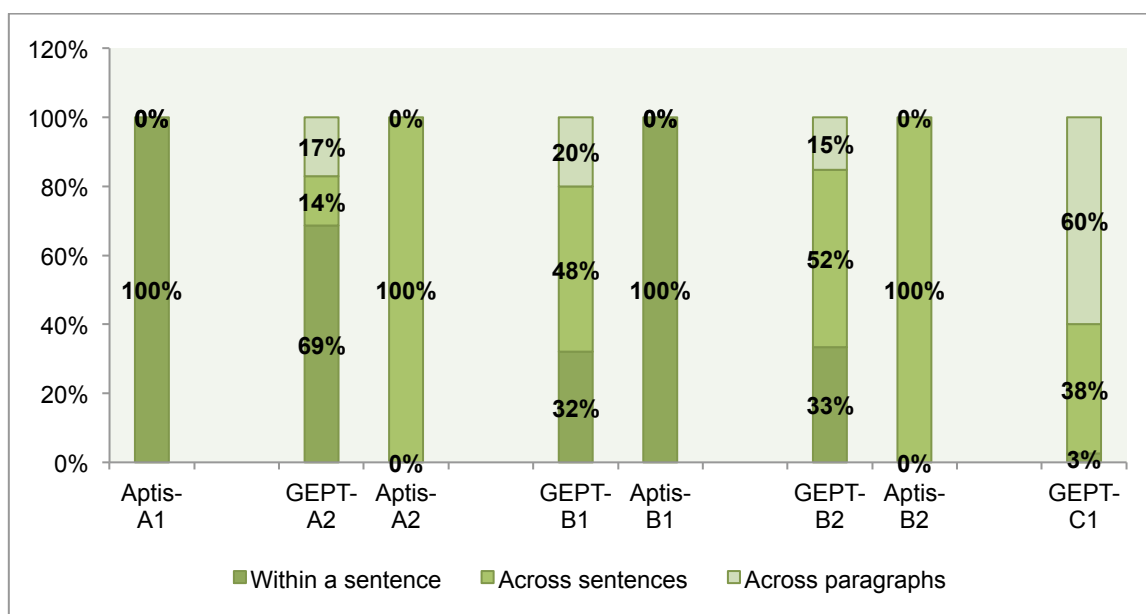
Figure 10: The nature of information in reading texts



Key: 1 = concrete; 4 = abstract

In terms of item dimension, the type of reading required to complete each task, the GEPT at A2 level contained more questions that required local comprehension, understanding content within a sentence; but at C1 level, only 3% required merely local comprehension and 60% involved comprehension across a number of paragraphs. In the Aptis reading test, all items for the B2 task required reading across sentences, but according to expert judgment, did not require reading across paragraphs. The A1 reading required only within-sentence comprehension to correctly answer each item, which is consistent with the task specification design. As with the results of the listening test content review, the expert analysis of the cognitive demands of the Aptis reading items differs somewhat from the task specifications used by item writers. The A2 and B1 items are designed to require across-sentence comprehension, but according to the judges, the items for B1 in this test form only required within-sentence comprehension to answer items correctly.

Figure 11: Scope of text processing needed to respond to items



Speaking

The GEPT speaking tests had 9 to 13 tasks at each proficiency level, while the Aptis speaking test contained two tasks at B1 level and one task each at A2 and B2 levels (Table 13).

Table 13: Number of tasks in the speaking test at each level

	A1	A2	B1	B2	C1
GEPT	--	13	12	10	9
Aptis	--	1	2	1	--

When the task domains were considered, the expert judges found that the GEPT contained a greater variety of speaking tasks as the test level increased. In contrast, the Aptis tasks were either personal or public in domain (Figure 12). On the other hand, the content knowledge specificity and cultural specificity of the two speaking tests fell towards the lower end of the five-point Likert scale at all levels. The nature of the information in both the GEPT and Aptis speaking tasks became more abstract as the CEFR level increased.

Figure 12: Distribution of speaking task domains

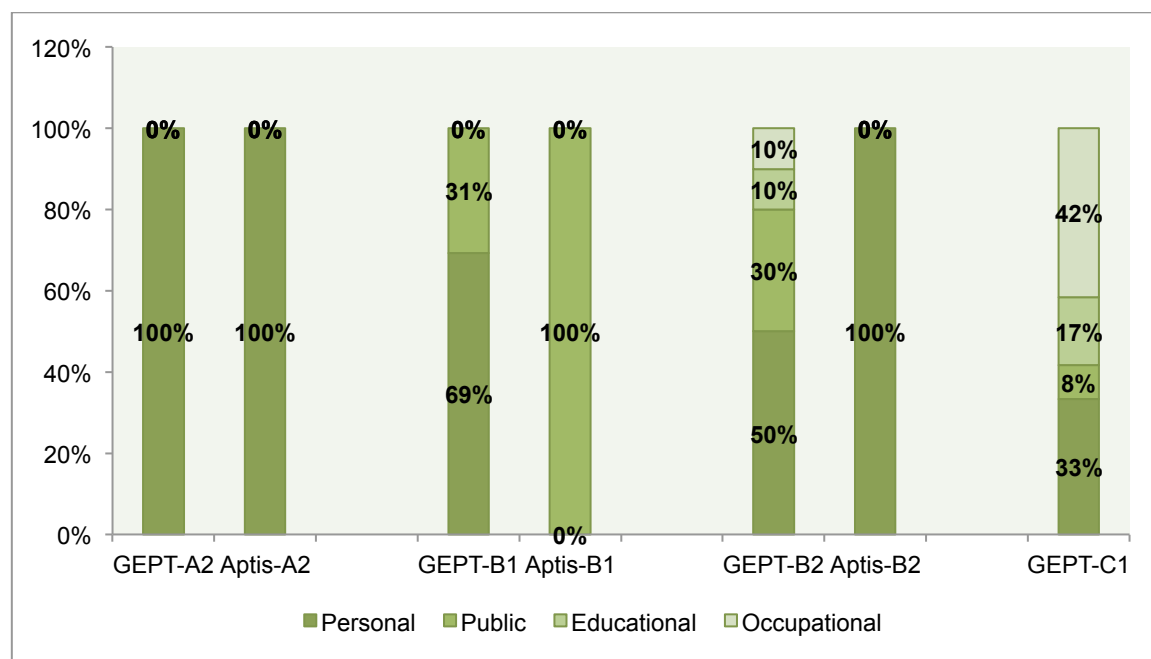
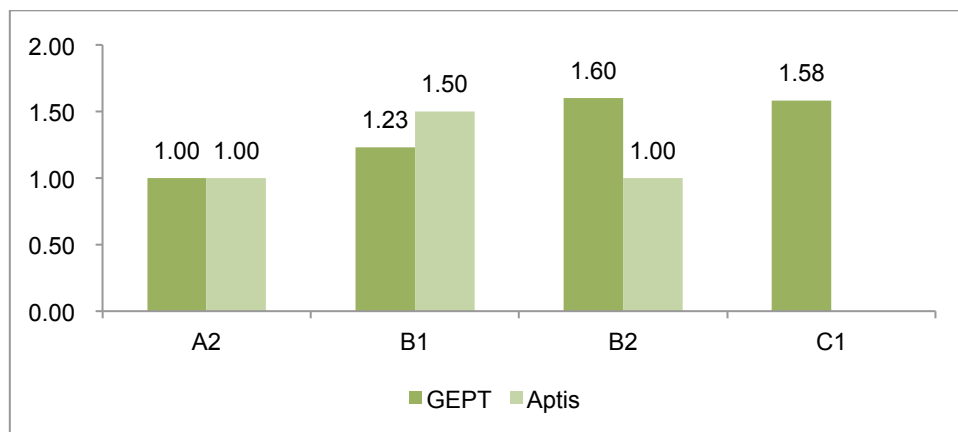
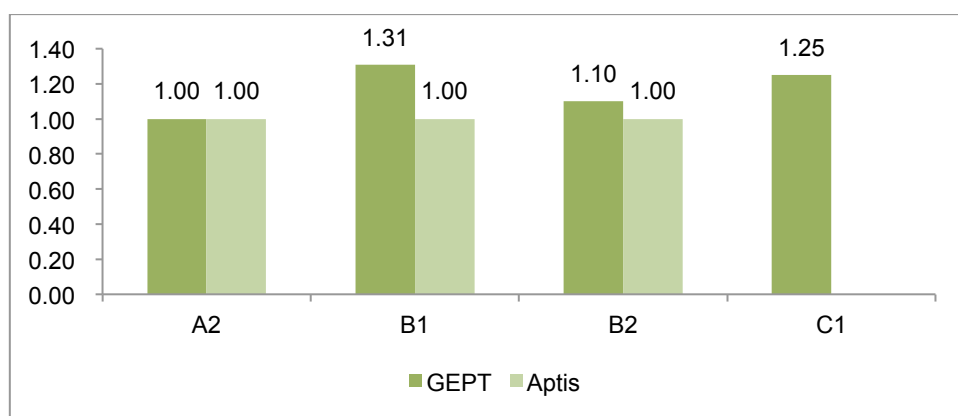


Figure 13: Degree of content knowledge specificity of speaking tasks



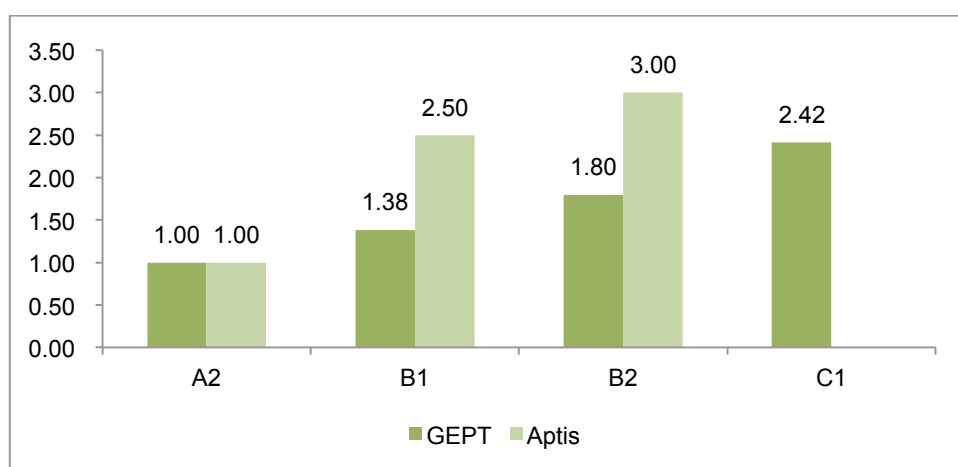
Key: 1 = general; 5 = specific

Figure 14: Degree of cultural specificity of speaking tasks



Key: 1 = general; 5 = specific

Figure 15: The nature of information in speaking tasks



Key: 1 = concrete; 4 = abstract

Although the GEPT and Aptis included different numbers of speaking tasks, it was still considered useful to investigate what language functions were elicited and assessed in each test. Subsequently, it was found that in addition to covering more language functions at each level, the GEPT assessed three macro types of language functions: informational, interactional, and interaction-managing, while Aptis focused on the informational functions only (Table 14).

Table 14: Language functions of speaking tasks

	A2		B1		B2		C1
	GEPT	Aptis	GEPT	Aptis	GEPT	Aptis	GEPT
Informational Functions							
1 Providing personal information	V	V	V		V		V
2 Explaining opinions/preferences	V		V	V	V	V	V
3 Elaborating	V		V	V		V	V
4 Justifying opinions			V		V		
5 Comparing				V		V	V
6 Speculating	V		V		V		V
7 Staging							
8 Describing		V	V	V	V	V	
9 Summarising							V
10 Suggesting							
11 Expressing preferences	V		V	V	V	V	V
Interactional Functions							
13 Agreeing							V
14 Disagreeing							V
15 Modifying/ commenting							
16 Asking for opinions							V
17 Persuading					V		V
18 Asking for information	V		V		V		V
19 Conversational repair							V
20 Negotiation of meaning							V
Managing Interaction							
21 Initiating							V
22 Changing topics							V
23 Reciprocating							V
24 Deciding							V
Number of functions	6	2	8	5	8	5	18

Writing

In the GEPT, there were two writing tasks each from B1 to C1 levels, while there were 16 tasks at A2 level. In contrast, the Aptis writing test contained one task per level (Table 15).

Table 15: Number of tasks in the writing test at each level

	A1	A2	B1	B2	C1
GEPT	---	16	2	2	2
Aptis	1	1	1	1	---

The content of the writing tasks was determined similarly to speaking. The GEPT writing tasks covered the personal, public and educational domains, while Aptis contained tasks in only the personal and educational domains (see Figure 16). Neither test paper under review contained tasks in the occupational domain. With respect to the specificity and abstractness of tasks, the GEPT was found to be more specific in content and cultural focus, as well as more abstract than Aptis at most levels, except for text abstractness at A2 level (Figures 17, 18 and 19). The GEPT writing tasks became more specific and abstract with increasing test level. In contrast, Aptis writing tasks did not differ in content knowledge specificity or cultural specificity at different levels.

Figure 16: Distribution of writing task domains

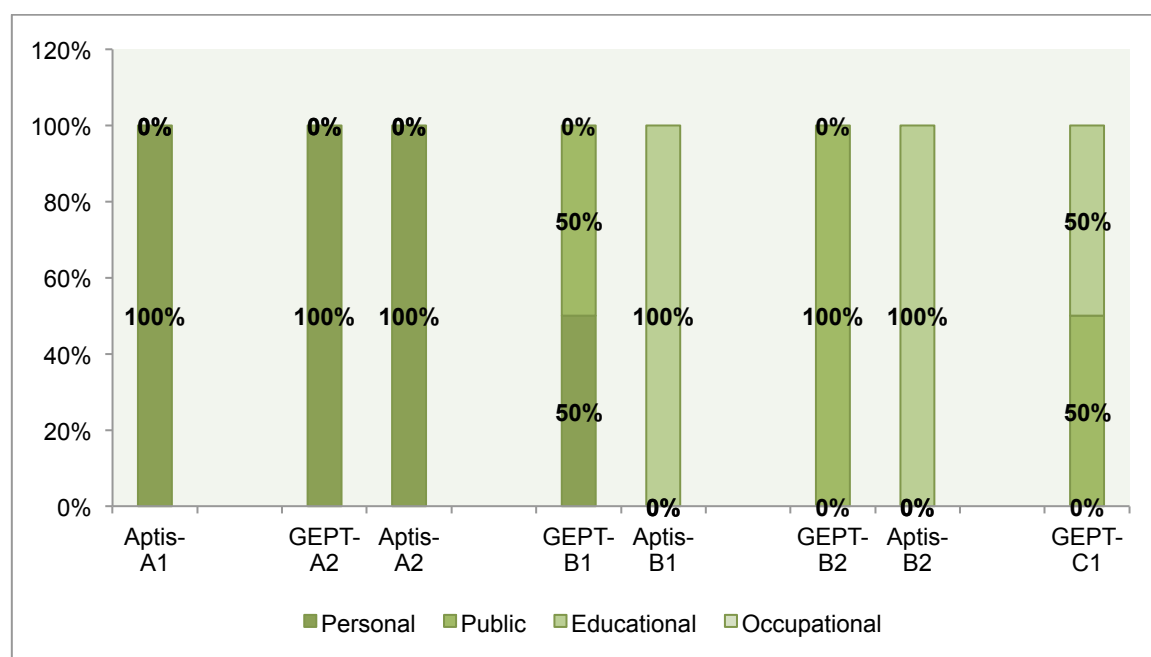
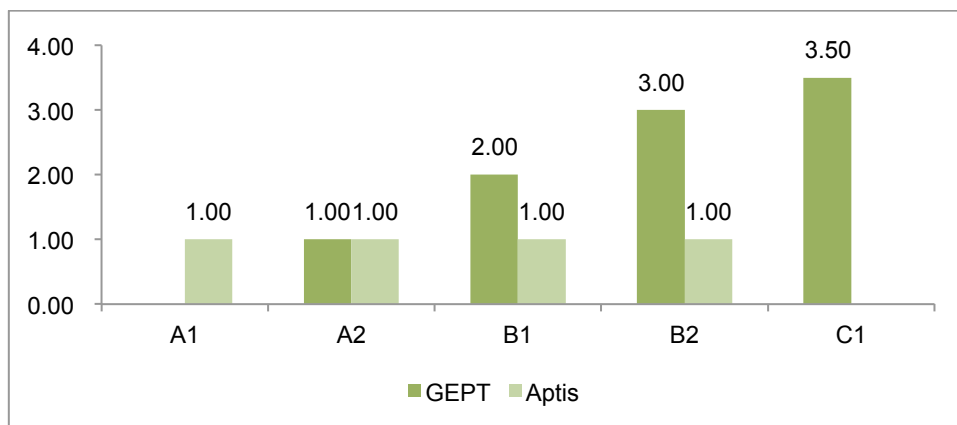
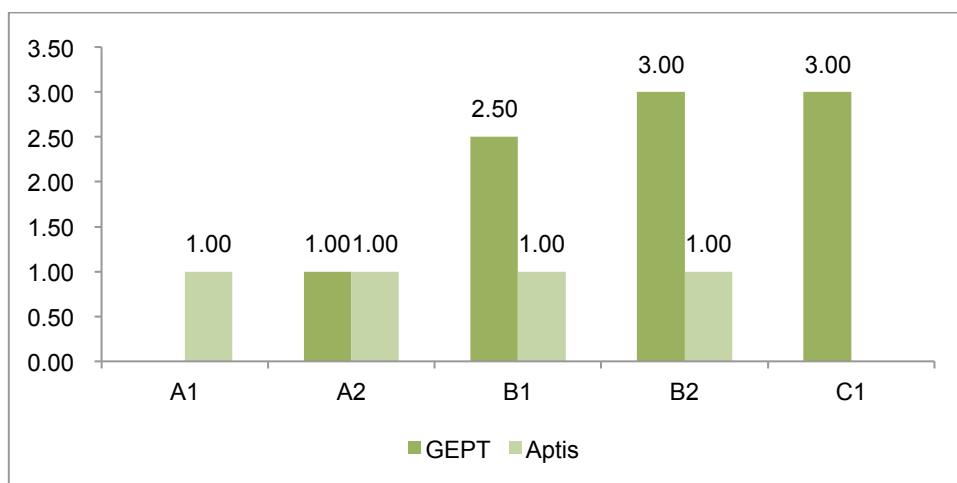


Figure 17: Degree of content knowledge specificity of writing tasks



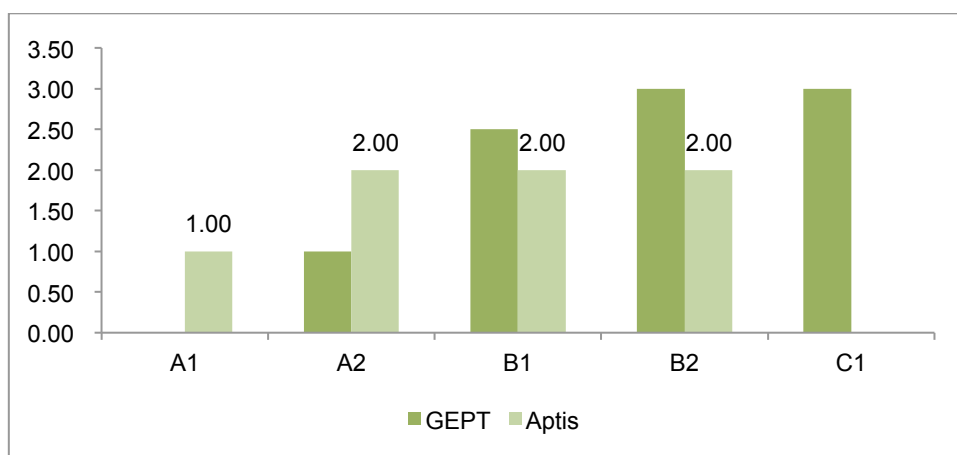
Key: 1 = general; 5 = specific

Figure 18: Degree of cultural specificity of writing tasks



Key: 1 = general; 5 = specific

Figure 19: The nature of information in writing tasks



Key: 1 = concrete; 4 = abstract

5.3 What are test-taker perceptions of Aptis and the GEPT?

This section summarises the results of the post-test surveys. The full data, as percentage response to each survey question, can be found in Appendices 2, 3 and 4.

Computer literacy

Responding to the section on computer use, 90% of the participants reported that they used computers frequently, but only 64% stated that they typed “often” in English. When participants were asked about their preferences for taking tests on computer rather than pen and paper, the number who preferred a computer test was 77% for grammar and vocabulary, 87% for listening, and 67% for writing, while a smaller percentage (53%) stated a preference for taking the reading test on computer. However, equal numbers of respondents expressed a preference for face-to-face (50%) and machine-recorded speaking tests (50%).

Feedback on Aptis

When respondents were asked their opinions about Aptis after taking the test, over 80% reported that the results of the grammar and vocabulary, reading, listening, and writing tests could reflect their English ability. For speaking, this figure was lower, with 74% stating that the speaking test could reflect their ability, possibly because of greater perceived difficulty of the speaking test as suggested by the lower scores for this component. Respondents generally felt Aptis to be relevant to them in content, with approximately 85% considering the topics of the tests and the vocabulary and sentence structure in the tests to be commonly used in daily life or the workplace. An even greater number of participants responded positively to the test instructions and the computer interface: approximately 90% found the test instructions clear and the interface user-friendly.

As for the number of items and the allotted response time, approximately 88% responded that these were appropriate. In addition, test-takers were asked about elements of the audio recordings. The speech rate of the listening input was deemed to be appropriate by 88% of test-takers. However, one difference worth noting is that the accents in the input for the speaking test were considered to be clear and easy to follow by 93% of respondents, whereas only 65% considered this to be true for the listening input. Finally, listening to the test items twice enhanced test performance according to nearly all (96%) test-takers.

Preference for Aptis or the GEPT

Test-takers were asked to state their preferences for different aspects of the test in comparison with each other. A total of 36% of test-takers considered the GEPT a better measure of their English ability, which was more than for Aptis, and those who regarded both to be equally good measures of ability (Figure 20). More test-takers also preferred the GEPT in that all items target the same level of ability, unlike Aptis, which contains tasks spanning a number of levels, while approximately one-third showed no preference (Figure 21). However, the majority of test-takers preferred Aptis in terms of being able to take all components in one sitting (Figure 22).

Figure 20: Participants' perception of which test is a better measure of their English ability

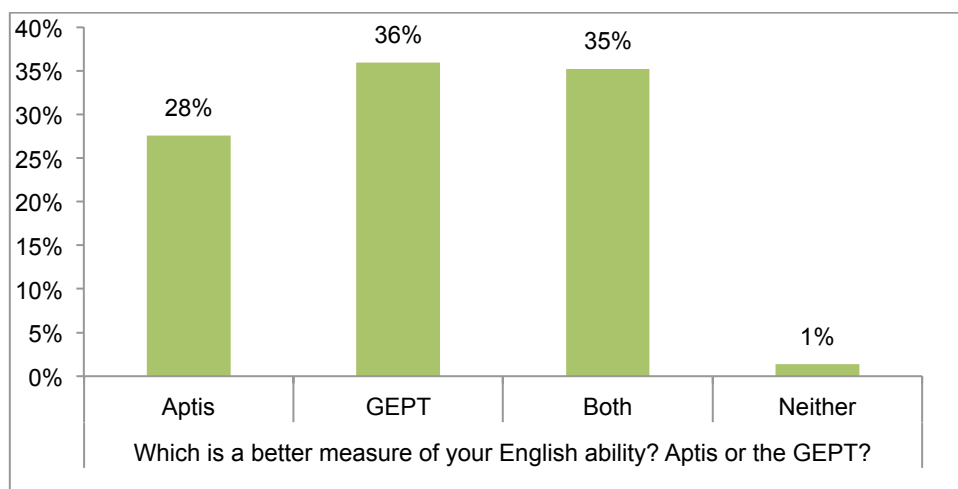


Figure 21: Participants' preference for tests that target the same or different levels of ability

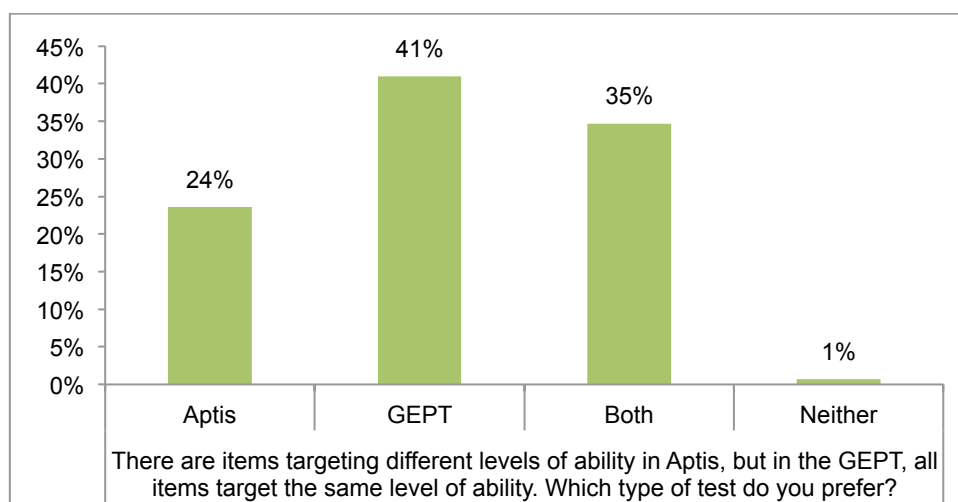
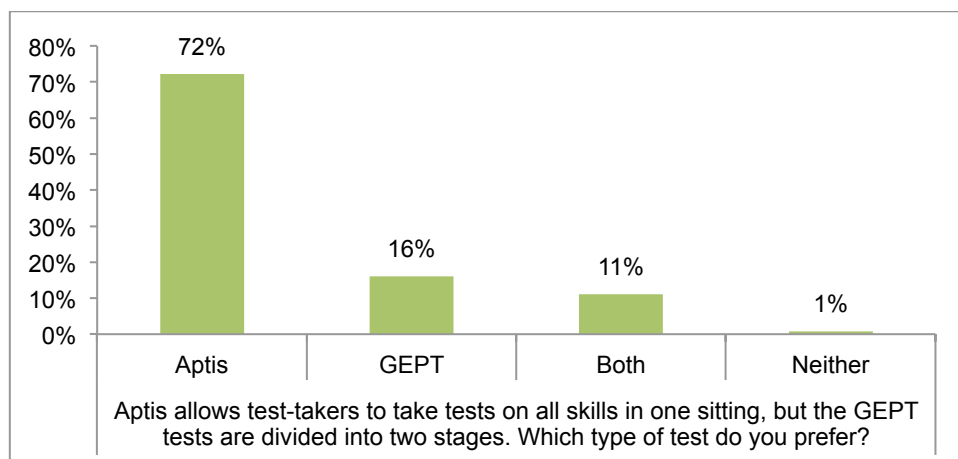
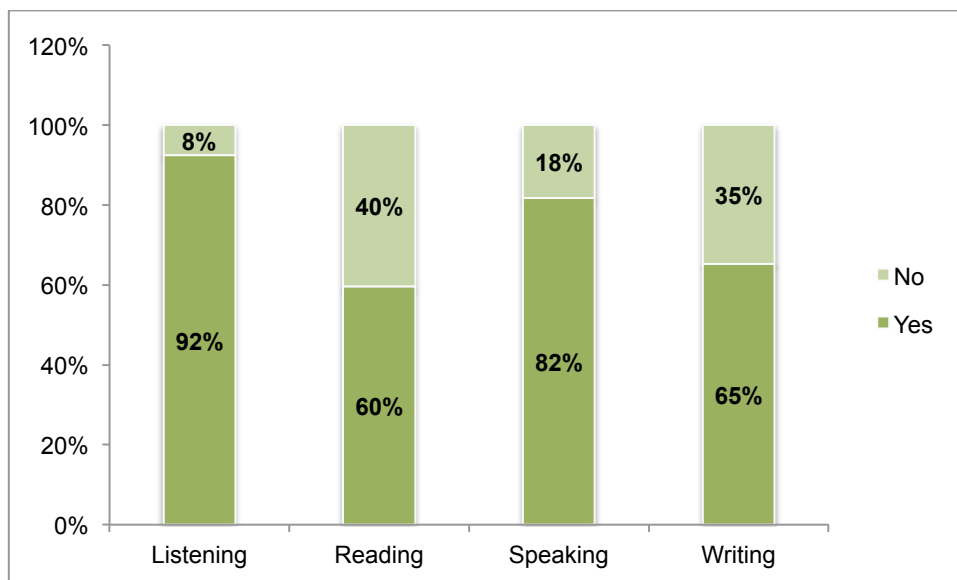


Figure 22: Participants' preference for one-stage or two-stage tests



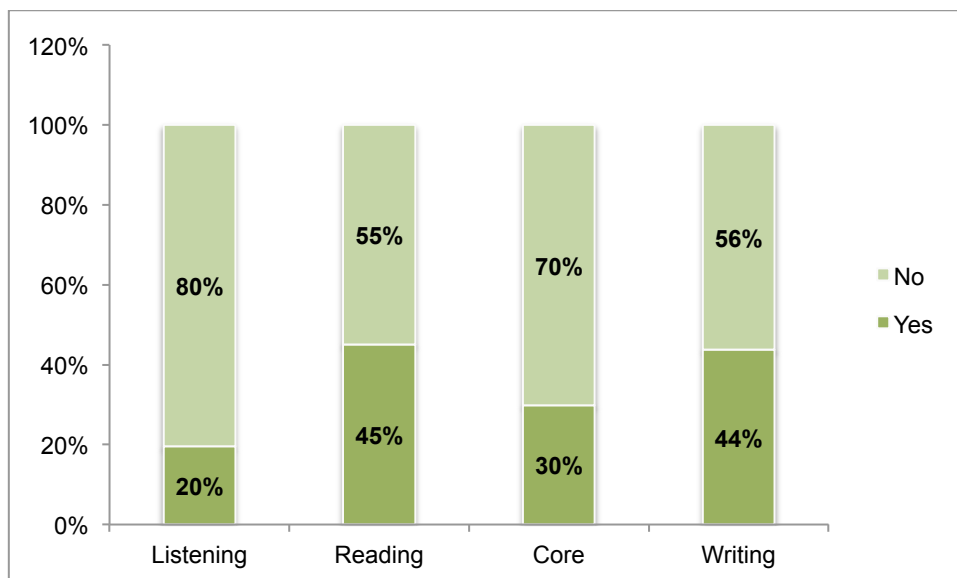
When asked about their preferred mode of delivery for taking the GEPT and Aptis tests, the majority of respondents reported that they would like to take each subtest on a computer. Candidate responses on each subtest are summarised in Figures 23 and 24.

Figure 23: Participants' preference for an alternative mode of delivery of the GEPT



If you could, would you prefer to take the GEPT on computer?

Figure 24: Participants' preference for an alternative mode of delivery of Aptis



If you could, would you prefer to take the paper-based Aptis?

Aptis scores and the score report

After receiving their score reports, test-takers were asked to what extent each component accurately reflected their ability. For all skills except speaking, the majority felt that their Aptis score provided a true estimate of their ability. The highest figure was for the reading test (71%), followed by the listening test (68%), the grammar and vocabulary test (65%), the writing test (61%), and then the speaking test (49%). This may be related to the fact that test-takers received lower scores on the speaking test than on the other components.

The next section of the questionnaire elicited responses concerning the content and design of the report. Roughly 90% of respondents considered the information in the score report, including the Scale Score, the CEFR Skill Descriptor, and the CEFR Skill Profile, to be clear and helpful.

Of those who considered the design of the score report to be unclear, some commented that there were too few scale levels (A1, A2, B1, B2 and C1) to properly differentiate the proficiency levels of test-takers, while others expressed a wish for the CEFR Skill Descriptor to be provided in Chinese. Similarly, most participants (64%) wanted the information on the score report to be printed in both English and Chinese.

In the final part of the questionnaire, respondents suggested what additional information they would like to be provided on the score report. Their responses can be summarised as follows: grade percentile rankings (86%); the score range for each CEFR band (74%); their Chinese name (42%); and their personal photograph (23%).

6. DISCUSSION AND CONCLUSIONS

The Aptis tests were compared with the GEPT tests to assess whether the two tests at the same CEFR level were comparable in terms of test-takers' performance and test content. In this study, the GEPT listening, reading, speaking and writing papers from A2 to C1 levels were selected as the external criterion because they are among the few exams that have made research-based claims about their relationship to the levels of the CEFR, and because GEPT scores are widely accepted for various purposes in Taiwan.

The comparison of participants' test performance on Aptis and the GEPT showed that mean GEPT scores increased as the CEFR level, as determined by Aptis, also increased, and this trend was consistent across all components. The differences were statistically significant between test-takers at B1 and B2 levels and C1 levels for listening, reading, and writing, and between B2 and B1 for speaking. However, the differences between means were not statistically significant between A2 and B1 for reading and listening, nor were they significant between A1 and A2 for speaking and reading. Given the relatively small number of participants in the study, and consequently even smaller numbers at each CEFR level, the lack of statistical significance is perhaps not surprising. Nonetheless, the findings are important considering the potential for using Aptis to identify readiness for taking the B1 and B2 levels of the GEPT tests, and this is considered further in Section 7 below. Participants' scores on the two exams were highly correlated at over 0.7. In addition, score agreement rate within adjacent levels was above 0.8 for all test components.

The result of the principal factor analysis suggested that the four GEPT test components and five Aptis test components measured the same trait, which could be called the overall English proficiency. However, it was easier for participants to achieve scores of higher CEFR levels on Aptis listening, reading and writing than on the GEPT, while participants tended to achieve scores of lower CEFR levels on Aptis speaking than on the GEPT speaking.

As regards the test content, Aptis and the GEPT were vastly different in terms of the number of items that each test component contained. This is not surprising, given the different test design and differences in uses and interpretations. The GEPT is a certificated examination with a strong criterion-referenced focus, whereas Aptis incorporates elements of both norm-referenced and criterion-referenced test construction, and it places a priority on flexibility and efficiency for institutional test users.

Analysis of the textual features also revealed that the two tests were distinct from each other in a number of respects. For example, Aptis listening texts were more difficult to comprehend than the GEPT at the same CEFR level. At the same time, it needs to be remembered that the comparison of features of tasks at each CEFR level employed the global, expert-judgment allocation of Aptis items to CEFR levels using the proforma. As already noted above in 5.2.2.2, the number of items in the listening test written to B2 specifications and pre-tested to confirm empirical difficulty is greater, and in fact, for the test form used in this analysis, six items were written to B2 specifications. All Aptis items are pre-tested and equated to a common Rasch-based scale for each skill, and only items complying with the intended hierarchy of difficulty are used for the construction of live test forms. Furthermore, live test forms are constructed so that they have the same mean Rasch item difficulty. The discrepancy between the empirical difficulty and level allocation through expert judgment seen in this analysis is something commonly noted in the literature on standard setting (e.g. Kaftandjieva, 2010), and confirms the concerns noted by Alderson et al. (2006) on the difficulty of achieving consensus on CEFR level allocation through expert judgment.

The results are, nonetheless, an important cross validation of the intentions of the original item specifications, and provide useful feedback for the Aptis research team to investigate what aspects, particularly the cognitive tasks demands, influenced the judges' decisions.

In comparing the demands of both tests, the Aptis and GEPT reading texts were similar in terms of readability, but the Aptis text at B2 level was more difficult than the GEPT, probably because the Aptis B2 text was much longer than the GEPT texts and Aptis used longer sentences than the GEPT. As for the speaking test, the topics used in Aptis were more abstract than those of the GEPT at B1 and B2 levels. In addition, the GEPT speaking test included a wider variety of task domains and was designed to elicit more language functions than Aptis. The GEPT writing tasks were considered cognitively more challenging than Aptis since the GEPT was more specific in content and cultural focus, as well as more abstract than Aptis at most levels.

7. LIMITATIONS AND RECOMMENDATIONS

The generalisability of the results obtained in this study is limited by its restricted scope. Although the study demonstrated that the CEFR score levels determined by Aptis and GEPT were comparable to some extent, in order to support the use of Aptis for placement purposes for universities and colleges in Taiwan, it would be useful to find local benchmarks for scoring decisions, rather than relying solely on CEFR levels. For that reason, further analyses on score data are recommended (e.g. regression analysis) to fine tune locally appropriate benchmarks on the Aptis score scale, which would be useful for predicting performance on the GEPT tests.

Since only one paper from each test was examined in the content analysis, the test items studied may not be wholly representative of those in operational tests. This has limited the generalisability of the results since textual differences between the two exams could be, to some extent, due to chance.

Furthermore, Aptis and the GEPT are very different in terms of the number of items that each test component contains. For example, in the test papers employed in this study, there were 67 reading tasks in the GEPT from the combined Elementary to Advanced levels; but there were only four tasks in Aptis. Future comparability studies between Aptis and the GEPT should include more than one Aptis test form in order to ensure that a more balanced selection of tasks is analysed.

Based on the questionnaire responses, the content of Aptis was considered to be generally appropriate for the local context. However, in the surveys, respondents expressed a preference for accents more familiar to test-takers in Taiwan, i.e. American English versions, for the listening input.

The score report was considered to be clear and helpful by most respondents, but they did recommend some amendments to the report, for example, putting the information on the score report in both English and Chinese, and providing the grade percentile rankings and information on the score range for each CEFR level, which would facilitate test-taker understanding of their scores compared with other test-takers. The above suggestions on the content and score report are based on the comments given by the participants in this study. It is recommended that additional trial tests be carried out to triangulate these survey results and determine to what extent these views are widely shared.

REFERENCES

- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct Project, *Language Assessment Quarterly*, 3(1), 3–30.
- Bachman, L. F., Davidson, F., Ryan, K. and Choi, I-C (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge TOEFL Comparability Study*, *Studies in Language Testing: Vol. 1*. Cambridge: Cambridge University Press.
- Brown, J. D., Davis, J. McE., Takahashi, C. & Nakamura, K. (2012). *Upper-level Eiken examinations: linking, validating, and predicting TOEFL iBT scores at advanced proficiency Eiken levels*. Society for Testing English Proficiency, Tokyo, Japan.
- Chin, J. & Wu, J. (2001). STEP and GEPT: A concurrent study of Taiwanese EFL learners' performance on two tests. *Proceedings of the Fourth International Conference on English Language Testing in Asia*, 22–44.
- Cobb, T. *Web Vocabprofile* [accessed 10 December 2014 from <http://www.lex tutor.ca/vp/>], an adaptation of Heatley, Nation & Coxhead's (2002) *Range*.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment: Manual*. Strasbourg: Council of Europe, Language Policy Division.
- Davis, A., Brown, A. Elder, C., Hill, K., Lumley, T. & McNamara, R. (1999). *Dictionary of Language Testing, Studies in Language Testing: Vol. 7*. Cambridge: Cambridge University Press.
- Dunlea, J. (2014). *Investigating the relationship between empirical task difficulty, textual features and CEFR levels*. Paper presented at the 11th Annual EALTA Conference, Warwick, United Kingdom.
- Geranpayeh, A. & Taylor, L. (Eds.) (2013). *Examining Listening: Research and practice in assessing second language listening*. *Studies in Language Testing* 35. Cambridge: Cambridge University Press.
- Green, A., Inoue, C. & Nakatsuhara, F. (forthcoming). *GEPT Speaking – CEFR Benchmarking, RG-09*. Taipei: LTTC.
- Harding, L. & Brunfaut, T. (2014). Linking the GEPT Listening Test to the Common European Framework of Reference, *LTTC-GEPT Research Reports*, RG-05. Taipei: LTTC. Retrievable from [<https://www.ltcc.ntu.edu.tw/ltcc-gept-grants/RReport/RG05.pdf>]
- Kaftandjieva, F. (2010). *Methods for Setting Cut-off Scores in Criterion-Referenced Achievement Tests: A Comparative Analysis of Six Recent Methods with an Application to Tests of Reading*. Arnhem: CITO and EALTA.
- Kane, M. (1998). Choosing between examinee-centred and test centred standard-setting methods. *Educational Assessment*, 5(3), 129–145.
- Khalifa, H. & Weir, C. J. (2009). *Examining Reading: Research and practice in assessing second language reading. Studies in Language Testing*, 29. Cambridge: Cambridge University Press.
- Knoch, U. (forthcoming). *Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)*, RG-08. Taipei: LTTC.
- Heatley, A., Nation, I. S. P. & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. Available at [<http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>]

Language Training and Testing Center. (2003). *Concurrent validity studies of the GEPT Intermediate level, GEPT High-Intermediate level, CBT TOEFL, CET-6, and the English test of the R.O.C. College Entrance Examination*. Taipei: Author.

Ministry of Education, Republic of China (Taiwan). (2004). 教育部未來四年施政主軸行動方案表[MOE Action Plan for Policy Initiatives for the Next Four Years]. Retrieved from <http://www.edu.tw/userfiles/url/20120921102842/a931022.doc/>

O'Sullivan, B. (2015). *Aptis Test Development Approach*. Aptis Technical Report TR/2015/001. London: British Council.

O'Sullivan, B. & Dunlea, J. (2015). *Aptis General technical manual version 1.0*. Aptis Technical Report TR/2015/005. London: British Council.

O'Sullivan, B. & Weir, C. J. (2011). Language Testing and Validation. In Barry O'Sullivan (Ed.) *Language Testing: Theory & Practice*, (pp.13–32). Oxford: Palgrave.

Scott, M. (2009). *Wordsmith Tools 5.0*. Oxford: Oxford University Press.

Taylor, L. (2004). Issues of test comparability. *Research Notes*, 15, 2–5.

Weir, C. J. (2005). *Language Testing and Validation: an evidenced-based approach*. Oxford: Palgrave.

Weir, C., 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 Reports RG-01*. Taipei: Language Training and Testing Center.

Wu, J. R. W. & Wu, R. Y. F. (2010). *Relating the GEPT Reading Comprehension Tests to the CEFR*, *Studies in Language Testing*: Vol. 33, 204–224.

Wu, R. Y. F. (2014). *Validating Second Language Reading Examinations: Establishing the Validity of the GEPT Through Alignment with the Common European Framework of Reference*, *Studies in Language Testing*: Vol. 41. Cambridge: Cambridge University Press.

Appendix 1:

Overview of Aptis and GEPT test components

Aptis test

Component	Part	Task type	No. of items/ Total		Time (mins)
Grammar & Vocabulary	1	Grammar	25	50	25
	2	Vocabulary	25		
Listening		Lexical recognition	5	25	50
		Identifying specific factual information	13		
		Meaning representation /inference	7		
Reading	1	Sentence comprehension (careful local reading)	6	27	30
	2	Inter-sentential text cohesion (careful global reading)	7		
	3	Short text comprehension (careful global reading)	7		
	4	Long text comprehension (expeditious global reading)	7		
Speaking	1	Personal information	3	10	12 (approx.)
	2	Picture description, expressing opinions, providing reasons	3		
	3	Describing, comparing and contrasting,	3		
	4	Integrating ideas on an abstract topic into a long turn	1		
Writing	1	Word-level form filling	1	7	50
	2	Short text writing	1		
	3	Short text writing. Responding to questions in a social media setting	3		
	4	Writing formal and informal email responses to an input text	2		

For detailed test and task specifications of the Aptis test, see O'Sullivan & Dunlea (2015).

GEPT Elementary, Intermediate, High-Intermediate and Advanced listening tests

GEPT level (Listening)	Part	Task type	No. of items/ Total		Time (mins)
Elementary	1	Picture description	5	30	20 (approx.)
	2	Answering questions	10		
	3	Conversations	10		
	4	Short talks	5		
Intermediate	1	Picture description	15	45	30 (approx.)
	2	Answering questions	15		
	3	Conversations	15		
High-Intermediate	1	Answering questions	15	45	35 (approx.)
	2	Conversations	15		
	3	Short talks	15		
Advanced	1	Short conversations and talks	15	40	45 (approx.)
	2	Long conversations	12		
	3	Long talks	13		

GEPT Elementary, Intermediate, High-Intermediate and Advanced reading tests

GEPT level (Reading)	Part	Task type	No. of items/ Total		Time (mins)
Elementary	1	Sentence completion	15	35	35
	2	Cloze	10		
	3	Reading comprehension	10		
Intermediate	1	Sentence completion	15	40	45
	2	Cloze	10		
	3	Reading comprehension	15		
High-Intermediate	1	Sentence completion	10	45	50
	2	Cloze	15		
	3	Reading Comprehension	20		
Advanced	1	Careful reading	20	40	50
	2	Skimming and scanning	20		20

GEPT Elementary, Intermediate, High-Intermediate and Advanced speaking tests

GEPT level (Speaking)	Part	Task type	No. of items/ Total		Time (mins)
Elementary	1	Repeating	5	18	10 (approx.)
	2	Reading aloud	6		
	3	Answering questions	7		
Intermediate	1	Reading aloud	2	13	15 (approx.)
	2	Answering questions	10		
	3	Picture description	1		
High-Intermediate	1	Answering questions	8	10	20 (approx.)
	2	Picture description	1		
	3	Discussion	1		
Advanced	1	Warm-up interview	4	9	25 (approx.)
	2	Discussion	3		
	3	Presentation	2		

GEPT Elementary, Intermediate, High-Intermediate and Advanced writing tests

GEPT level (Writing)	Part	Task type	No. of items/ Total		Time (mins)
Elementary	1	Sentence writing	15	16	40
	2	Paragraph writing	1		
Intermediate	1	Chinese–English translation	1	2	40
	2	Guided writing	1		
High-Intermediate	1	Chinese–English translation	1	2	50
	2	Guided writing	1		
Advanced	1	Summarising main ideas from verbal input and expressing opinions	1	2	60
	2	Summarising main ideas from non-verbal input and providing solutions	1		45

Appendix 2:

Test-taker Questionnaire A: Core+Reading+Listening

感謝您參加今天的 Aptis 測驗。請花 3-5 分鐘的時間完成本問卷。本問卷旨在了解您對 Aptis 測驗的意見，每題請塗黑作答。(Thank you for taking today's Aptis test. Please take 3 to 5 minutes to complete this questionnaire. The aim of this questionnaire is to collect test-takers' opinions about the Aptis test. Please read the statements and questions, and mark your answers.)

壹、電腦操作熟悉程度(A. Computer Familiarity)

	是 Yes		否 No	
1. 除了今天的 Aptis 測驗以外，我曾參加過其他電腦化英語測驗。I have taken other computer-based English tests before taking today's Aptis test.	41%		59%	
2. 我在今日考試前已在家上網試作練習題。I did the practice test online before taking today's Aptis test.	62%		38%	
	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree
3. 我經常使用電腦。I often use computers.	61%	29%	10%	0%
4. 我經常在電腦上閱讀文章。I often read on computer.	38%	39%	19%	4%
5. 我經常在電腦上打英文。I often type in English.	30%	34%	28%	8%
6. 與紙筆測驗相比，我更喜歡在電腦上考文法與字彙測驗。I prefer computer-based grammar and vocabulary tests to paper-and-pencil-based grammar and vocabulary tests.	28%	49%	16%	7%
7. 與紙筆測驗相比，我更喜歡在電腦上考閱讀測驗。I prefer computer-based reading tests to paper-and-pencil-based reading tests.	20%	34%	32%	15%
8. 與紙筆測驗相比，我更喜歡在電腦上考聽力測驗。I prefer computer-based listening tests to paper-and-pencil-based listening tests.	41%	46%	9%	4%

貳、Aptis 文法與字彙測驗、閱讀測驗、聽力測驗(B. Aptis Grammar and Vocabulary Test, Reading Test, and Listening Test)

	文法與字彙測驗(Grammar and Vocabulary)				閱讀測驗(Reading)				聽力測驗(Listening)			
	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree
1. 整體而言，今天的測驗結果可以反映出我的英語能力。Generally speaking, the results of today's test are able to reflect my English ability.	15%	68%	15%	1%	21%	63%	15%	1%	24%	60%	15%	1%
2. 測驗的內容或題材與我的日常生活或工作相關。The topics of the test are related to my life and/or working experiences.	17%	64%	17%	2%	12%	63%	24%	2%	24%	62%	12%	2%
3. 我認為試題的作答說明很清楚。I think the test instructions are clear.	54%	42%	3%	0%	51%	39%	10%	0%	56%	41%	3%	0%
4. 我認為測驗系統的介面友善，容易操作。I think the Aptis test system provides a user-friendly interface.	58%	38%	3%	1%	55%	39%	5%	1%	53%	41%	6%	0%
5. 我認為測驗的題數適中。I think the number of items of the Aptis test is appropriate.	31%	63%	6%	1%	28%	60%	10%	1%	26%	59%	13%	2%
6. 我認為測驗的作答時間適中。I think the time allotted for the Aptis test is appropriate.	27%	59%	13%	1%	28%	56%	13%	4%	32%	56%	11%	1%
7. 整體而言，我認為測驗所使用的句型、字彙及用語是我在日常生活中或工作場合常見的。Generally speaking, the sentence structures, vocabulary and phrases in the Aptis test are commonly used in daily life/workplace.	28%	60%	11%	0%	24%	61%	15%	1%	32%	62%	6%	0%

貳、Aptis 文法與字彙測驗、閱讀測驗、聽力測驗(B. Aptis Grammar and Vocabulary Test, Reading Test, and Listening Test)

	文法與字彙測驗(Grammar and Vocabulary)				閱讀測驗(Reading)				聽力測驗(Listening)			
	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree
8. 整體而言，我認為 聽力 試題的說話速度適中。 Generally speaking, the speech rate of the listening input is appropriate.									34%	54%	11%	1%
9. 整體而言，我認為 聽力 試題內容的口音清楚易懂。 Generally speaking, the accents of the listening input are clear and easy to understand.									18%	47%	28%	7%
10. 我認為 聽力 測驗時，每題聽兩遍能幫助我在聽力測驗的表現。I think listening to the test items twice enhances my performance.									51%	45%	4%	0%

☺問卷結束，感謝您的協助

This is the end of the questionnaire. Thank you.

Appendix 3:

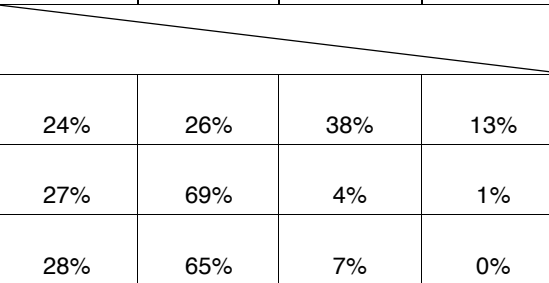
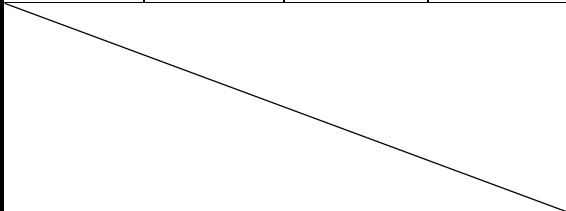
Questionnaire B: Writing+Speaking+APTIS and GEPT

感謝您參加今天的 Aptis 測驗。請花 3-5 分鐘的時間完成本問卷。本問卷旨在了解您對 Aptis 測驗的意見，每題請塗黑作答。Thank you for taking today's Aptis test. Please take 3 to 5 minutes to complete this questionnaire. The aim of this questionnaire is to collect test-takers' opinions about the Aptis test. Please read the statements and questions, and mark/write your answers.

參、Aptis 寫作與口說測驗(C. Aptis Writing Test and Speaking Test)

	寫作測驗 (Writing)				口說測驗 (Speaking)			
	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree
1. 整體而言，今天的測驗結果可以反映出我的英語能力。Generally speaking, the results of today's test are able to reflect my English ability.	19%	68%	12%	1%	15%	59%	26%	1%
2. 測驗的內容或題材與我的日常生活或工作相關。The topics of the test are related to my life and/or working experiences.	16%	69%	15%	0%	11%	69%	20%	1%
3. 我認為試題的作答說明很清楚。I think the test instructions are clear.	34%	57%	6%	3%	34%	59%	6%	1%
4. 我認為測驗系統的介面友善，容易操作。I think the Aptis test system provides a user-friendly interface.	41%	46%	10%	3%	48%	49%	3%	0%
5. 我認為測驗的題數適中。I think the number of items of the Aptis test is appropriate.	30%	64%	6%	0%	29%	61%	9%	1%
6. 我認為測驗的作答時間適中。I think the time allotted for the Aptis test is appropriate.	27%	62%	10%	1%	25%	60%	14%	1%

參、Aptis 寫作與口說測驗(C. Aptis Writing Test and Speaking Test)

	寫作測驗 (Writing)				口說測驗 (Speaking)			
	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree
7. 整體而言，我認為測驗所使用的句型、字彙及用語是我在日常生活中或工作場合常見的。Generally speaking, the sentence structures, vocabulary and phrases in the Aptis test are commonly used in daily life/workplace.	23%	71%	6%	0%	23%	65%	12%	0%
8. 與紙筆寫作測驗相比，我較喜歡在電腦上考寫作測驗。 I prefer computer-based writing tests to paper-and-pencil-based writing tests.	33%	34%	24%	9%				
9. 與錄音考試相比，我較喜歡用面試的方式考口說。 I prefer face-to-face speaking tests to machine-recorded speaking tests.					24%	26%	38%	13%
10. 整體而言，我認為口說測驗的說話速度適中。Generally speaking, the speech rate of the recording of the Aptis Speaking Test is appropriate.					27%	69%	4%	1%
11. 整體而言，我認為口說測驗的口音清楚易懂。Generally speaking, the accents of the input in the Aptis Speaking Test are clear and easy to follow.					28%	65%	7%	0%

肆、Aptis 與全民英檢(GEPT) (D. Aptis and the GEPT)

	Aptis		全民英檢 (GEPT)		兩者相同 Both		兩者皆否 Neither	
1. Aptis 和全民英檢，何者較能測量您的英文能力？ Which is a better measure of your English ability? Aptis or the GEPT?	28%		36%		35%		1%	
2. Aptis 各項測驗不分級，包含了各種難度的題目(例：初至高級的題目)；全民英檢是分級測驗，每級測驗的題目難易度相近(例：中級測驗內題目皆為中級)。您較喜歡哪一種？ There are items targeting different levels of ability in Aptis, but in the GEPT, all items target the same level of ability. Which type of test do you prefer?	24%		41%		35%		1%	
請說明原因。Please explain.								
	Aptis		全民英檢 (GEPT)		兩者相同 Both		兩者皆否 Neither	
3. Aptis 讓考生一次考完聽、說、讀、寫，及文法與字彙測驗；全民英檢分初試與複試，通過聽力閱讀測驗再考口說與寫作測驗。您較喜歡哪一種？ Aptis allows test-takers to take tests on all skills in one sitting, but the GEPT tests are divided into two stages. Which type of test do you prefer?	72%		16%		11%		1%	
請說明原因。Please explain.								
	Aptis		全民英檢 (GEPT)		兩者相同 Both		兩者皆否 Neither	
4. Aptis 聽力測驗每題可聽二次；全民英檢每題聽一次。您較喜歡哪一種？ Aptis allows test-takers to listen to each item twice, but the GEPT allows test-takers to listen to each item once. Which do you prefer?	82%		10%		8%		0%	
請說明原因。Please explain.								
	聽力測驗 Listening		閱讀測驗 Reading		口說測驗 Speaking		寫作測驗 Writing	
	Yes	No	Yes	No	Yes	No	Yes	No
5. 若可，您是否想在電腦上考全民英檢(GEPT)? 請分測驗回答。If you could, would you prefer to take the GEPT on computer?	92%	8%	60%	40%	82%	18%	65%	35%
	文法與字彙測驗 Core		閱讀測驗 Reading		聽力測驗 Listening		寫作測驗 Writing	
	Yes	No	Yes	No	Yes	No	Yes	No
6. 若可，您是否想選擇考紙筆的 Aptis? 請分測驗回答。If you could, would you prefer to take the paper-and-pencil-based Aptis?	30%	70%	45%	55%	20%	80%	44%	56%

☺問卷結束，感謝您的協助☺

This is the end of the questionnaire. Thank you.

Appendix 4: Test-taker questionnaire: score report

感謝您參加 5 月 18 日「台灣學習者國際英語能力調查研究」之 Aptis 電腦化英語測驗。本問卷旨在了解您對該測驗成績及成績單的意見，每題請點選最適合者回答。(Thank you for taking Aptis test on May 18. The aim of this questionnaire is to collect test-takers' opinions about their test scores and the Aptis score report. Please read the statements and questions, and mark your answers.)

	非常同意 strongly agree	同意 agree	不同意 disagree	非常不同意 strongly disagree
1. 我認為各測驗成績(Scale Score)呈現方式清楚明白。 I think the scale score for each test component is presented clearly.	42.9%	54.5%	2.6%	0.0%
2. 我認為 CEFR 級數的說明(CEFR Skill Descriptors)清楚明白。I think the CEFR Skill Descriptors are clear.	40.3%	55.8%	3.9%	0.0%
3. 我認為 CEFR 級數(CEFR Skill Profile)有助於我了解自己的英語能力。I think the CEFR Skill Profile helps me understand my English ability.	32.5%	61.0%	6.5%	0.0%
4. 我認為成績單整體設計佳、各項資訊呈現清楚。 The score report is well-designed and all information is presented clearly.	26.0%	71.4%	2.6%	0.0%

	比真實程度差 underestimates my true ability	比真實程度佳 overestimates my true ability	二者相近 estimates my true ability well	不清楚 not clear
5. 我認為聽力(Listening)測驗成績與我真實的聽力程度相較 In my opinion, my Aptis Listening Test _____	14.3%	16.9%	67.5%	1.3%
6. 我認為閱讀(Reading)測驗成績與我真實的閱讀程度相較 In my opinion, my Aptis Reading Test _____	11.7%	14.3%	71.4%	2.6%
7. 我認為口說(Speaking)測驗成績與我真實的口說程度相較 In my opinion, my Aptis Speaking Test _____	33.8%	11.7%	49.4%	5.2%
8. 我認為寫作(Writing)測驗成績與我真實的寫作程度相較 In my opinion, my Aptis Writing Test _____	18.2%	16.9%	61.0%	3.9%
9. 我認為文法與詞彙(Grammar & Vocab)測驗成績與我真實的文法與詞彙程度相較 In my opinion, my Aptis Grammar & Vocab Test _____	18.2%	11.7%	64.9%	5.2%

表單的底部

	英文 English	中文 Chinese	中英文並列 both English and Chinese	沒有意見 no comment
10. 我希望成績單的內容以何種語言呈現? I would like the information in the score report to be presented in _____	18.2%	3.9%	63.6%	14.3%

11. 除現有資訊外，我希望成績單還能提供 哪些資訊？（可複選） In addition to the existing information, I would like the score report to provide _____ (multiple choice)	41.6% 中文姓名 Chinese name	15.6% 性別 gender	18.2% 出生年月日 birth date
	23.4% 個人照片 personal photo	22.1% 母語 native language	19.5% 國籍 nationality
	9.1% 身分證/護照號碼 passport number	74.0% CEFR 各級數的成績範圍 score range for each CEFR level	85.7% 全體考生百分等級(PR 值) score percentile rankings
	0.0% 其他 other		

British Council Assessment Research Group

The Assessment Research Group was formed in 2013 to support the British Council's work in assessment and testing across the world. The team is responsible for ensuring that all new assessment products and new uses of existing products are supported by the most up-to-date research. They also continuously evaluate the quality of British Council assessment products.

APTIS-GEPT TEST COMPARISON STUDY: LOOKING AT TWO TESTS FROM MULTI-PERSPECTIVES USING THE SOCIO COGNITIVE MODEL

VS/2016/002

Wu, Yeh, Dunlea and Spiby

BRITISH COUNCIL VALIDATION SERIES

Published by British Council
10 Spring Gardens
London SW1A 2BN

© **British Council 2016**

The British Council is the
United Kingdom's international
organisation for cultural relations
and educational opportunities.

www.britishcouncil.org/aptis/research