25th International Conference of English Teaching and Learning
2008 International Conference on English Instruction and Assessment

Assessing English Proficiency at Advanced Level: Testers' Feedback to Teaching

Jessica R. W. Wu & Anita C. W. Lin
The Language Training and Testing Center
jw@lttc.ntu.edu.tw

Abstract

To achieve beneficial washback effects of a public language test in the EFL classroom, communication between the testers and those involved in teaching and learning is desirable. This paper reports the results of the 2007 GEPT Advanced Listening and Reading Tests, based on which the learners' strengths and weaknesses are discussed in relation to the test construct.

The GEPT Advanced Test is part of the five-level proficiency framework particularly developed for Taiwan's EFL learners. Those who master this level can communicate effectively and can handle academic or professional requirements. It is expected that university graduates who have majored in English can demonstrate this level of proficiency. However, only an average of 20% of test-takers have been able to pass the test annually, which is below the benchmark (30%) set by the GEPT research committee. Apparently, there is a gap between the test designers' expectation and the reality for the advanced learners. Therefore, the important questions are why such a difference exists and how it may be minimized or even resolved in the future.

This paper, by providing the assessment information, intends to render assistance in addressing these questions. Ultimately, it is hoped that a conscious feedback loop between the teaching and testing of English can be established.

Key words: GEPT, washback effect, listening test, reading test, factor analysis

Introduction

Testing and assessment are genuinely effective only if they inform pedagogy in order to improve it. Genesee and Upshur (1996) state that the most effective system will be one where assessment provides a feedback loop in which:

> (A)ssessment activities are motivated and shaped by instructional purposes, plans, and practices in the classroom, and the decisions that arise from the results of these activities, in turn, lead to reshaping of these instructional purposes, plans, and practices. (257)

Stobart (2003: 140) also notes that there are complex relationships between testing, teaching, and learning as testing is never a neutral process and always creates consequences. Within language testing and assessment, the effects of language tests on language teaching and learning, i.e. inside the classroom, have been referred to as *washback*. In Cheng's (2007)

recent review of the empirical research studies in washback conducted in the past 15 years, she recognizes the remarkable contribution made by Alderson and Wall (1993) and Wall and Alderson (1993) in developing the constructs of washback studies for the field of language testing and in exploring potentially positive and negative washback effects. Furthermore, they expanded their concern to test validity by questioning whether washback could be a property of test validity as suggested by Messick (1996). Therefore, by looking at the washback effect of a test, not only can we measure how the test has affected learners and teachers, but we can also evaluate the test validity.

In Taiwan, with increasing awareness of the importance of washback effect in recent years, the influences of the General English Proficiency Test (GEPT), an English language testing system specifically designed for Taiwan's EFL learners, on teaching and learning have been widely discussed. A number of studies investigating the GEPT washback have been conducted (e.g. Wu & Chin, 2006; Wu, 2007); however, they mostly deal with the lower levels of the GEPT. There are only very few studies focusing on the higher levels (High-Intermediate and Advanced) of the GEPT (e.g. Wu, 2002; Wu, 2003), which may be due to the fact that they have many fewer test-takers.

To aid in achieving a more comprehensive understanding of the GEPT's washback effect and validity, this study looks into the Advanced level, which is the highest level of the GEPT available in the market, though it has the smallest number of test-takers among the other GEPT levels. Although a complete account of the GEPT Advanced Test should include the speaking and writing tests, in order to offer readers a thorough investigation into the washback effects, this paper will focus only on the listening and reading tests. This study serves as a starting point from which to search for answers to the question: *what is the washback of the GEPT-Advanced?* Therefore, the present study primarily focuses on the aspects of test-takers' profile and performance in the GEPT-Advanced Listening and Reading. Based on the performance data obtained from the operational test in 2007, the advanced learners' strengths and weaknesses are discussed in relation to the test construct. Such concrete assessment information might be able to help establish a conscious feedback loop between the teaching and testing of English in Taiwan.

GEPT-Advanced Test

Test-takers

To promote "lifelong learning," the GEPT Advanced Test made its debut in 2002 after one and a half years of research and several pilot tests of different scales (Wu et al., 2001; LTTC, 2002). Designed to cater to learners of high-level language abilities, the test specifically targets those who use English for academic and professional purposes that require effective flexibility and spontaneity in communication.

Throughout the last six years, data on the test-takers of the GEPT Advanced Test have

been collected to provide insights into their characteristics. What follows are the main traits that sketch out the profile of these learners:

Age. The average age is 27.17, and learners between 20 and 30 make up 65% of the total number of test-takers.

Gender. About 66% of test-takers are female.

Employment. More than 80% of test-takers are students. However, there has been a significant increase in the number of non-student test-takers since 2004.

Education. Of the 35% of test-takers who are at or above the college level, 8% are English/foreign languages majors.

Reasons for taking the test. Nearly 70% of test-takers took the test for "self-assessment," and another 26% of test-takers for academic needs.

Level Criteria & Test Content

Test-takers who are awarded the certificate of the GEPT Advanced Test take all four sub-tests (Table 1). They have to score at least 80 out of 120 in both listening and reading in the first stage in order to proceed to the second stage, and the pass mark is set at band 3 for both writing and speaking.

Table 1 Test Formats and Structure

Stage	Component	Part	Item Type	Time (mins.)
First	Listening	1	Short Conversations & Talks	45 (approx.)
		2	Long Conversations	
		3	Long Talks	
	Reading	1	Careful Reading	50
		2	Skimming & Scanning	20
Second	Writing	1	Summarizing main ideas from verbal input and expressing opinions	60
		2	Summarizing main ideas from non-verbal input and providing solutions	45
	Speaking	1	Warm-up Interview	25 (approx.)
		2	Information Exchange	
		3	Presentation	

In general, test-takers who pass this level have English language abilities which enable them to communicate fluently, with only occasional errors related to language accuracy and appropriacy, and to handle academic or professional requirements and situations. Their English ability is roughly equivalent to that of a graduate of a local university who majored in English, or to that of someone who has received a degree from a university or graduate school in an English-speaking country. The table below gives further descriptions of the target situations that those who pass the GEPT Advanced Test are able to handle.

Table 2 Skill-Area Descriptions for the GEPT Advanced Level

Listening	Can understand conversations on all sorts of topics as well as debates, lectures, news reports, and TV/radio programs. At work, when attending meetings or negotiations, he/she can understand reports and discussions.
Reading	Can understand all sorts of written English from a wide variety of sources, including magazine and newspaper articles, literature, professional periodicals, and academic publications.
Writing	Can use English appropriately in writing several text types—such as reports, essays, news items or summaries of general/professional topics—and to be able to translate news articles or excerpts from books on general topics. He/she can express opinions on different topics and discuss them in depth.
Speaking	Can participate in discussions on, and fluently express his/her opinions about all sorts of issues. He/she can give reports or express his/her opinions in general meetings or professional seminars.

The multifarious contexts outlined above indicate that test-takers at this level should understand a wide range of language functions and to extract complete meanings from different types of extended discourse. They should also be able to draw upon a large array of strategies when they encounter different communicative themes and tasks despite possible complexity or unfamiliarity of topics.

Overview of the GEPT-Advanced Listening and Reading Tests

The GEPT Advanced Test has the distinctive feature of utilizing longer texts (Tables 3~6) to confirm whether test-takers can achieve an understanding of a text as a whole as they gradually integrate scattered ideas and sort out these ideas' relations with one another, as well as their particular functions in relation to the entire text (Alderson, 1996; Nuttall, 1996). Therefore, the most significant characteristic of the GEPT Advanced Test can be said to be the process of establishing a macrostructure of texts/discourses. In the case of listening and reading, it is worth mentioning that they are the first of the GEPT tests to apply constructed response method[1] with an aim to diversify response formats. In a test that underscores academic skills, a constructed response method such as short answer questions helps measure sophisticated language abilities more effectively (Khalifa & Weir, 2007).

Listening Test

Construct. The listening test comprises three parts, Short Conversations & Talks, Long Conversations, and Long Talks. Its items are designed to test how well test-takers can identify the purpose and main ideas of a discourse, and if they can understand details, recognize important contextual features (e.g., settings, relationships of speakers), determine the attitude of speakers, draw conclusions, and make inferences.[2] Specifically in the second and third parts, longer utterances are crucial to engaging test-takers' "discourse knowledge," such as their understanding of the cohesion among different paragraphs, when they process connected discourse (Buck, 2001).

Task types. The listening test contains 40 items in total (Table 3). The first part has 15 items, and the second and third have 12 and 13 respectively. The test is about 45 minutes in length. In Part I, test-takers hear short conversations and talks. To start the test with a task that is familiar to learners, multiple-choice questions are the response format employed in this part. Though multiple-choice questions are not entirely absent in Parts II and III, the design of Long Conversations and Long Talks is meant to approach real-world applications as closely as possible by introducing constructed response items, including short answer questions and notes-completion. The adoption of such items provides an integrated approach to assessing listening ability, for test-takers need to produce their own answers, whereas choices are listed in multiple-choice questions and no justification of the test-takers' answers are required.[3]

Table 3 Overall Design of the GEPT-Advanced Listening Test

<table>
<tr><th>Part</th><th>Passage Type</th><th>Recording Information</th><th>Response method</th><th>Number of Items</th></tr>
<tr><td>I</td><td>15 Short Conversations & Talks
(40~70 words/each)</td><td rowspan="3">➢ Played only once
➢ Speech rate: 190 wpm</td><td>4-option multiple-choice questions (MCQs)</td><td>15</td></tr>
<tr><td>II</td><td>2 Long Conversations
(500~600 words/each)</td><td rowspan="2">Notes-completion & Short-answer questions (SAQs)</td><td>12 (5~7 items for each passage)</td></tr>
<tr><td>III</td><td>2 Long Talks
(500~600 words/each)</td><td>13 (6~7 items for each passage)</td></tr>
</table>

Aside from short dialogues, Part I of the listening test also features various genres including personal narratives, commercials, lectures, and review extracts (Table 4). Part II covers an assortment of discourses such as discussions of social issues, extracts from TV interviews, radio shows, etc. Similarly, Part III bases its contexts on authentic TV programs and news reports, and draws from lectures in professional or academic fields.

The GEPT Research Report on the Advanced Level (LTTC, 2002) investigates its topics and discourse types and indicates that not only does the test incorporate descriptive, narrative, and instructive texts, it also relies heavily on expository and argumentative texts[4] to evaluate how well test-takers can prioritize information and determine the relationship among individual components of a connected discourse.

Table 4 Text/Task Design of the GEPT-Advanced Listening Test

Part	Passage Type	Discourse type & Examples	Test focus
I	Short Conversations & Talks	**Descriptive, narrative, expository:** Daily conversations, work-related discussions and transactions, short narrations, descriptions, reviews, commercial messages, etc.	To assess learners' ability to comprehend an extended discourse: 1. understand the gist, main ideas, and the framework 2. understand contextual features 3. understand important facts and details 4. make inferences based on available information, such as speakers' tone/attitude or implied meanings
II	Long Conversations	**Descriptive, narrative, expository, argumentative:** Interviews, long discussions, radio broadcasts, etc.	
III	Long talks	**Expository, argumentative:** Speeches, presentations, news reports, radio broadcasts, etc.	

Reading Test

Construct. Constructed to elicit a thorough reading of presented materials, Part I, Careful Reading, is devised to evaluate learners' ability to understand not only local or explicit information, such as important details, but also rhetorical devices and organizational functions, such as relationships between ideas or contrastive viewpoints in an article. The inclusion of more complex constructs of reading ability can be said to account for the longer texts in the GEPT-Advanced Reading compared with reading tests at lower levels.

Given that expeditious reading is a much-overlooked dimension in conventional reading tests, Part II of the GEPT-Advanced Reading requires test-takers to adjust their reading speed and employ different strategies according to disparate reading purposes.[5] The inclusion of Expeditious Reading as a separate task in the test acknowledges that the ability to quickly grasp the gist of a paragraph (Skimming) and locate specific information in a text (Scanning)[6] are just as essential to everyday reading as Careful Reading, in which the tasks intend for readers to arrive at in-depth comprehension.

Task types. Each part of the reading test has 20 questions (Table 5). Test-takers have 50 minutes to complete the first part of the test, and 20 minutes for the second part. The first three texts in Part I (Careful Reading) are followed mostly by short answer questions, and occasionally, multiple-choice ones. However, a summary paragraph containing six gaps follows the last text of the Careful Reading. Test-takers have to fill those gaps in the summary with either a word or a phrase. In Part I of the test, test-takers are expected to seek in-depth comprehension of texts through slow, linear word-by-word reading whether the items are testing ideas at a local or at a global level (LTTC, 2002: LR5).

Part II (Expeditious Reading) includes two kinds of reading activities, namely, Skimming and Scanning. Having the same number of items as Part I but a much tighter time constraint, Expeditious Reading is constructed to elicit reading selectively (Urquhart & Weir,

1998). To complete the Skimming task, test-takers choose from a list of headings those that are appropriate for each of the paragraphs in the two passages (each passage has six missing headings). Reading selectively allows test-takers to omit insignificant details or skip unnecessary information, focus on obtaining an overall impression of a paragraph, and determine which heading best expresses the main idea of the paragraph.

Test-takers are also expected to engage in reading quickly and selectively for the Scanning task. They are first instructed to read questions before proceeding to reading three thematically related texts. Presumably, having a particular goal in mind, test-takers then approach the three texts by looking only for specific information that they assume pertinent to answering the questions correctly. Unlike most test items in Part I, all of the items in Part II are of selected response formats.

Expository and argumentative texts play an important role in the GEPT-Advanced Reading (Table 6) in that these texts render more room for sophisticated structures, such as multiple points of view, comparison/contrast, and problem/solution etc., which are commonly seen in academic texts. Nevertheless, Expeditious Reading is predominantly composed of descriptions and narratives. This is especially true for Scanning, since test-takers are invested in searching for specific words or phrases and no text-level understanding is required (Khalifa & Weir, 2007). While the test focus of the GEPT-Advanced Reading is similar to that of other GEPT lower-level reading tests, short answer questions require test-takers to transform their understanding of written input into words of their own, and hence call for reading and interpreting a text in a more comprehensive manner.

Table 5 Overall Design of the GEPT-Advanced Reading Test

Part		No. of Texts & Length	Response Method	No. of Items	Time (min.)	Required Reading Speed
I	Careful Reading	4 long texts (600~900 words/each)	➢ 4-option MCQs ➢ SAQs ➢ Summary completion	20 (4~6 items /each text)	50	80~100 wpm
II	Skimming	2 long texts (600~800 words/each)	Heading matching	12 (6 items /each text)	20	150~200 wpm
	Scanning	3 thematically related texts (250~300 words/each)	3-way multiple matching	8		

Table 6 Text/Task Design of the GEPT-Advanced Reading Test

Part		Text Type	Discourse Type	Test Focus
I	Careful Reading	Newspapers, magazines & journals (articles/reports/ columns), book/film reviews	➢ Descriptive ➢ Narrative ➢ Expository ➢ Argumentative	Processing a text thoroughly to comprehend main ideas, supporting details and implied meanings
II	Skimming	Newspapers, magazines & journals (articles/reports/ historical accounts)	➢ Descriptive ➢ Narrative ➢ Expository	Processing a text quickly and selectively to get the gist of the text
	Scanning	Informational sources (guides/tourist pamphlets)	➢ Descriptive ➢ Narrative	Processing a text quickly and selectively to locate specific information

Test Results

Descriptive Statistics

A total of 698 test-takers sat for the 2007 GEPT-Advanced Listening and Reading Tests. A total of 106 of the test-takers (15.2%), whose scores for both listening and reading tests were 80 or above, were determined to have passed Stage I of the tests. The two test components correlate reasonably at a coefficient alpha of 0.75. The overall performance of the test-takers is summarized below in Table 7. From the higher mean p value and mean score in test-takers' performance on the listening test, it can be noted that test-takers performed better on the listening test than on the reading test. The distribution curves in Figure 1 clearly depict the difference between test-takers' performance on the two test components.

Table 7 Descriptive Statistics

	Listening Test	Reading Test
Number of Items	40	40
Mean P (Facility)	0.61	0.41
Mean Score	70.03/120	58.13/120
SD	19.4	20.81
Maximum	116	120
Minimum	15	11
Alpha	0.85	0.77

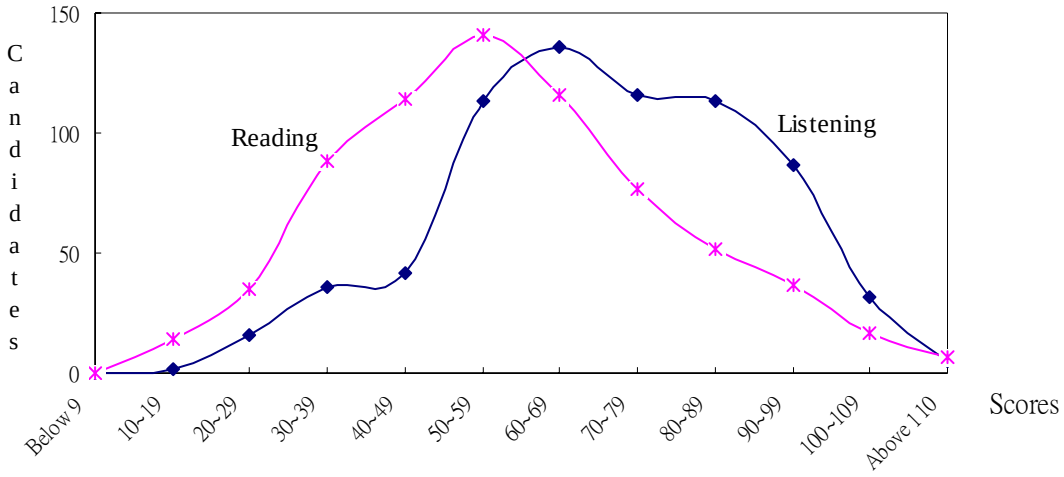

Figure 1

To have a deeper look at the test-takers' performance, the test results for the listening and the reading tests were examined separately at both subtest level and task level. In addition, the test results were compared between two different response types: the selected response items (SR) versus the constructed response items (CR).

Listening Test

Statistical information about test-takers' performance in the listening test is provided in Table 8. First, at the subtest level, test-takers performed with the highest facility (0.63) in Short Conversations and Long Talks, which was followed by Long Conversations with a facility of 0.57. Second, at the task level, the best performance, which bears the highest facility index, was found in Long Talks – Radio (0.74), followed by Short Conversations (0.63), Long Conversations – Discussion (0.59), Long Conversations – Interview (0.55), and Long Talks – Lecture (0.52).

Table 8 Listening Performance

Subtest/Task	Topic/Discourse Type	Items	Facility (p)	SD
Part I: Short Conversations	-	1-15	0.63	0.20
Part II: Long Conversations	-	16-27	0.57	0.20
Discussion	Services Science/Expository	16-22	0.59	0.21
Interview	Architecture Critic/ Descriptive & Expository	23-27	0.55	0.25
Part III: Long Talks	-	28-40	0.63	0.16
Radio	Whale Watching/Narrative	28-32	0.74	0.19
Lecture	Eighteenth-century Paris/ Expository	33-40	0.52	0.20

The result of the comparison of test-takers' performance in the listening test between the SR and the CR items is shown in Table 9. With the higher facility index of the SR items, it is evident that the SR items clearly outperformed the CR items.

Table 9 Listening Performance by Response Type

	No of Items	Facility	SD
Listening	40	0.61	0.16
SR Items	23	0.67	0.20
CR Items	17	0.55	0.16

Reading Test

Statistical information about test-takers' performance in the reading test is provided in Table 10. First, at the subtest level, test-takers' performance (from the strongest to the weakest) was found in the following order: Expeditious Reading – Scanning (0.52), Expeditious Reading – Skimming (0.44), Careful Reading – Articles I-III (0.44), and Careful Reading – Summary Cloze (0.24). Second, at the task level, ranked by facility index (from the highest to the lowest), the easiest task for the test-takers is Skimming – Heading Matching A (0.59), which is then followed by Careful Reading – Article II (0.58), Scanning (0.52), Careful Reading – Article I (0.44), Careful Reading – Article III (0.31), Skimming – Heading Matching B (0.30), and Careful Reading – Summary Cloze (0.24).

Table 10 Reading Performance

Part	**Subtest**/Task	Topic/Discourse Type	Items	Facility (p)	SD
Part I: Careful Reading (Items: 1-20)	**Article I-III**	-	1-14	0.44	0.18
	Article I	Media: Reality TV/ Argumentative	1-4	0.44	0.25
	Article II	Environment/Expository	5-9	0.58	0.26
	Article III	Museology: Islamic art/ Argumentative	10-14	0.31	0.22
	Summary Cloze	Technology: Video Game Therapy/ Expository	15-20	0.24	0.24
Part II: Expeditious Reading (Items: 21-40)	**Skimming**	-	21-32	0.44	0.20
	Heading Matching A	Natural Sciences: Landslides/ Expository	21-26	0.59	0.25
	Heading Matching B	History: Bounty Mutineers/ Narrative	27-32	0.30	0.24
	Scanning	Education/ Descriptive	33-40	0.52	0.24

The result of the comparison of test-takers' performance in the reading test between the SR and the CR items is shown in Table 11. Similar to what was found in the listening test, test-takers performed better on SR items than on CR items.

Table 11 Reading Performance by Response Type

	No of Items	Facility	SD
Reading	40	0.41	0.14
SR Items	22	0.49	0.16
CR Items	18	0.35	0.16

Factor Analysis

Statistical procedures such as factor analysis allow for testing whether or not different constructs are addressed by the subtests as the GEPT-Advanced Listening and Reading were designed to do. Kinnear and Gray (1995:78) describe factor analysis as: "… a set of methods designed to identify the latent psychological independent factors thought to underlie the correlations among a set of variables." Therefore, if the subtests function in a similar manner, it is expected that they will load on the same factor and should be considered to share the same trait. If, on the other hand, the subtests function statistically in different manners and load on different factors, it is evident that there are multi-traits in the test.

In the present study, factor analysis was performed separately with all the listening test items and all the reading items. In the case of the listening test, a two-factor solution was chosen for the Varimax Rotation procedure. Factor 1 and Factor 2 accounted for a total of 19.43% of the total variance, with Factor 1 (15.52%) and Factor 2 (3.91%). As shown in Figure 2, all the subtests in the listening test except Long Talks – Radio (Items 28-32) loaded on both factors, which means that most of the items in the listening test function similarly. Therefore, the results do not show a clear multi-trait construct in the listening test, which may be due to the fact that all the intended traits (e.g. global understanding, making inferences) are included evenly in all of the subtests. However, it is noted that Long Talks – Radio is the only subtest that loaded on one single factor, i.e., Factor 2. The difference may imply that Long Talks-Radio is distinct from that of the remaining subtests.

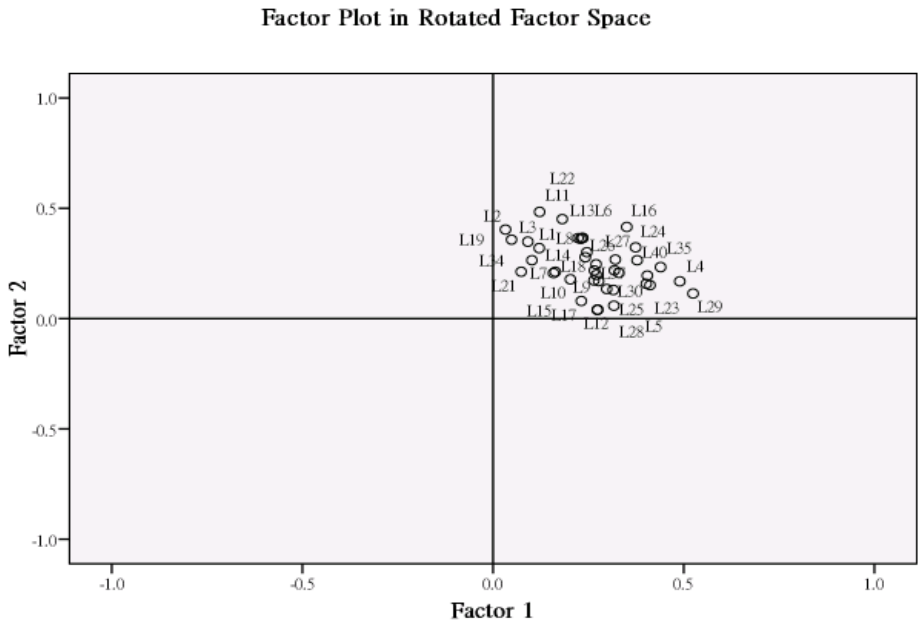

Figure 2 Varimax Rotation – Listening

In the case of the reading test, a four-factor solution was chosen for the Varimax Rotation procedure. The four factors accounted for a total of 16.22% of the total variance, with Factor 1 (10.80%), Factor 2 (5.02%), Factor 3 (4.4%), and Factor 4 (3.90%), respectively. As clearly shown in the statistical output in the appendix, each of the subtests in the reading test functioned statistically in a different manner. To be more specific, the items for Careful Reading – Articles I-III (1-14) mostly clustered on Factor 1; Skimming (21-32) on Factor 2; Careful Reading – Summary Cloze (15-20) on Factor 3; Scanning (33-40) on Factor 4. This phenomenon evidently supports the original intent of the GEPT-Advanced Reading Test to include multi-traits and to measure a different construct in each subtest.

Discussion

Beneficial washback effects of a public language test in the EFL classroom rely on communication between the testers and the professionals involved in teaching and learning. This paper hopes to provide teaching professionals with information about Taiwan's advanced EFL learners' strengths and weaknesses by examining the results of the 2007 GEPT-Advanced Listening and Reading in relation to the test construct. There are several directions for discussion:

Difficulty Level & Learners' Performance

Learners' performance in a language test is a function of their language ability and the difficulty level of the test. Overall, the results of the analyses show that the test-takers'

performance in the 2007 GEPT-Advanced Listening and Reading was affected by a number of variables which may influence difficulty as was assumed in the original design of the GEPT-Advanced Listening and Reading. These variables include response type, text length, text genre, task type, time constraints, topic familiarity, etc., which are all commonly identified in the testing literature (e.g. Buck, 2001).

CR Items vs. SR Items

The phenomenon that Taiwanese test-takers performed better in the SR items of the GEPT-Advanced reported earlier (Wu, 2002) was found again in this study. The gap between test-takers' performance in the SR items and the CR items supports the need to include items of both response types in the design of the GEPT Advanced Level Test. In addition, in viewing the poorer performance in the CR items of the advanced learners, it is not difficult to imagine that a similar gap between learners' performance in the SR items and CR items at the lower levels would also exist if CR items were employed. For the sake of practical concerns, particularly in a large-scale test, SR items are often chosen to be used as the only response type. Historically, SR items have been predominately used in Taiwan's large-scale testing, for example, the high school and the college entrance examinations, which has led English teaching and learning to incorporate intensive practice and drilling on SR items and consequently hampered the validity of those tests. Therefore, the drawbacks of SR items must be recognized and teachers can try to familiarize learners with a variety of response types in both English learning and testing tasks.

The Example of Long Talks-Radio

The results of the factor analysis show that unlike the other tasks, Long Talks – Radio uniquely fell on Factor 2 by itself. Having looked at this in combination with the test-takers' performance in different task types, interestingly, we found that the task of Long Talks – Radio was not equally difficult as its counter task (Long Talks – Lecture) and both tasks did not fall on the same factor. Thus, it is speculated that the difference may be associated with the genre of the listening input, in this case, radio talk versus lecture. To be more specific, the task in Long Talks – Radio was about a radio journalist introducing a unique holiday tour, which was written in a narrative style. On the other hand, the other task in Long talks concerned a brief lecture on Paris in the 18th century, which used an expository text. Although further studies are required to confirm this speculation, it is a good idea to expose learners to a variety of text genres to improve their listening ability.

Careful Reading & Expeditious Reading

Similarly, the results of the factor analysis show that there are four different constructs (Careful Reading – Articles I-III, Careful Reading – Summary Cloze, Skimming, and Scanning) in the reading test as it was originally designed. Moreover, test-takers' performance varied from subtest to subtest, for example, test-takers performed best in Scanning, followed by Careful Reading – Articles I-III, the most conventional task type; on the other hand, test-takers were much weaker in the items requiring expeditious reading skills (i.e., Skimming: Items 27-32) and those requiring integration of reading comprehension and information restructuring (i.e. Careful Reading – Summary Cloze: Items 15-20). Therefore, learners should be made aware of different reading skills and strategies (careful reading versus expeditious reading) and how to apply these appropriately depending on the reading purpose.

Weak Performance in Summary Cloze

Learners' low scores for the GEPT-Advanced Listening and Reading may be attributed to the fact that their English proficiency was below the benchmark set in the GEPT Advanced Level Test. Having said that, however, learners' weak performance in the test may have been caused by construct underrepresentation and construct irrelevant variance (Messick, 1996), or it could be a function of these two possibilities. Despite the fact that the construct of the GEPT-Advanced Listening and Reading was intended to be representative of learners' ability of listening and reading at the advanced level and was arrived at after rigorous pretesting and research during the developmental stage, the GEPT developer is endeavoring to investigate whether test-takers' low test scores on the GEPT-Advanced Listening and Reading resulted from factors that are irrelevant to the construct the test is designed to measure (Cheng, 2007). To be more specific, it was found that the test-takers performed the weakest in Summary Cloze. Research (e.g. test-takers' verbal report during or after the task performance) should be conducted to discover what the task actually measures and to see if Summary Cloze should be replaced with a different task type that represents the intended construct and at the same time elicits test-takers' best performance.

Conclusion

The GEPT statistics show that only 20% of the test-takers who have taken the advanced level can pass the test. Apparently, there is a wide gap between the standard set in the GEPT and the actual English ability of advanced learners in Taiwan. This poses two questions to both teaching and testing: (1) Why does such a mismatch exist? (2) How may the mismatch be minimized or even resolved in the future? It is believed that when both teaching and testing endeavor to answer these questions, a conscious loop between the teaching and testing of English will be successfully established. It is hoped that the present study has demonstrated one way to help bridge the gap between teaching and testing.

References

Alderson, J.C. (1996). The Testing of Reading. In C. Nuttall, *Teaching reading skills in a foreign language* (pp.221-229). (Second ed.). Oxford: Heinemann English Language Teaching.

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment: reading and listening: final report of the Dutch Construct Project*. Available on request from the Project Coordinator, J. Charles Alderson, c.alderson@lancaster.ac.uk.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics,* 14, 115-129.

Buck, G. (2001). *Assessing Listening.* Cambridge: Cambridge University Press.

Cheng, L. (2007). What does washback look like? Paper presented at the sixteenth International Symposium on English Teaching, Taipei.

Genesee, F. & G. Upshur. (1996). *Classroom-based Evaluation in Second Language Education.* Cambridge: Cambridge University Press.

Khalifa H & Weir, C. J. (forthcoming) *Examining reading: Reserach and practice in assessing second language reading.* Cambridge: Cambridge University Press.

Kinnear, P.R.S., & Gray, C.D. (1995). SPSS for Windows Made Simple. Norwood, NJ: Ablex Publisher.

LTTC (2002). *Pretest Reports on the General English Proficiency Test (GEPT) Advanced Level.*

Messick, S. (1996). Validity and washback in languagae testing. *Language Testing*, *13*(3), 241-256.

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. London: Heinemann.

Pugh, A.K. (1978). *Silent Reading.* London: Heinemann Educational.

Research & Development Program Office. (2002). *GEPT Research Report: Advanced Level Test.* Taipei: The Language Training and Testing Center.

Stobart, G. (2003). The Impact of Assessment: Intended and unintended consequences. *Assessment in Education*, 16(2), 139-140.

Urquhart, A.H. & Weir, C. J. (1998). *Reading in a second language: process, product and practice*. Harlow: Longman.

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing, 10,* 41-69.

Weir, C. J. (2005). *Language Test Validation: an evidence-based approach*. Oxford: Palgrave.

Weir, C. J. (1993). *Understanding and developing language tests.* New York: Prentice Hall.

Weir, C. J., & Wu, J. (2006). Establishing test form and individual task comparability: a case study of semi-direct speaking test, *Language Testing*. 23 (2): 167-197.

Weir, C. J., Yang, H. & Jin, Y. (2000). An empirical investigation of the componentiality of L2 reading in English for academic purposes. *Studies in language testing*, 12. Cambridge: Cambridge University Press.

Wu, J. (2002). *Assessing English Proficiency at Advanced Level: The Case of the GEPT*, The International Conference on 'Language Testing and Language Teaching,' Shanghai.

Wu, J., Wu, R., Wei, E., & Kuo, K. (2001). *A Progress Report on GEPT Advanced Level Development*, The Tenth International Symposium on English Teaching, Taipei.

Wu, R. (2003). *Assessing Advanced-Level English Writing: A Report on the Case of the GEPT*, Proceedings of the Sixth International Conference on English Language Testing in Asia, Seoul.

Wu, R., & Chin, J. (2006). *An Impact Study of the Intermediate Level GEPT*. Proceedings of the Ninth International Conference on English Testing in Asia, Taipei.

Footnotes

[1] After each test, sample answer sheets are reviewed by both the internal research staff and external committee to examine possible variations in the correct answers and further to decide on the acceptable range for full credit and partial credit.

[2] The test focus is based on the framework of listening comprehension proposed by Weir (1993), which includes the following:

Direct meaning comprehension:

--Listening for gist, main ideas or important information

--Listening for specifics, involving recall of important details

--Determining speaker's attitude/intentions toward listener/topic where obvious from the text

Inferred meaning comprehension:

--Making inferences and deductions

--Relating utterances to the social and situational context in which they are made

--Recognizing the communicative function of utterances

--Deducing meaning of unfamiliar lexical items from context

Contributory meaning comprehension (microlinguistic):

--Understanding phonological features (stress, intonation, etc.), grammatical notions (such as comparison, cause, result, degree, purpose, etc.), discourse markers, syntactic structure of the sentence and clause, grammatical cohesion (particularly reference), lexis and lexical cohesion, etc.

[3] Weir (2005) has cautioned test developers against the possible interference of short answer questions with the measurement of the intended construct, since writing ability would seem to be among the skills required to complete an SAQ task. This concern can nevertheless be eased by increasing precision of the wording when formulating questions. Training of the raters and post-test moderating sessions to standardize judgments can also help minimize the effect of this drawback.

[4] The discourse types are based on the classification offered by Alderson et al (2004:46).

[5] Such design finds its basis in the model of reading comprehension raised by Urquhart & Weir in 1998 and the empirical study of the development of reading tests of Weir et al. in 2000.

6 It is important, however, to point out here that both the task of Skimming and Scanning can involve another type of expeditious reading, namely, search reading. In the case of Skimming, it is very likely that test-takers simply try to locate information on "predetermined topics" regardless of the macrostructure of the entire text (Urquhart & Weir, 1998: 103). As far as Scanning is concerned, test-takers might go beyond lexical matching and search for "various words in a similar semantic field to the topic," since the extraction of information could require more close attention than the usual scanning (Pugh, 1978: 53).

Appendix

Rotated Component Matrix – Reading

	Factor1	Factor 2	Factor 3	Factor 4
R1	0.27			
R2			0.25	
R3	0.31			
R4	0.30			
R5	0.58			
R6	0.38			
R7	0.32		0.22	
R8	0.34			
R9	0.58			
R10	0.59			
R11	0.39		0.23	0.27
R12	0.38			0.21
R13	0.41		0.39	
R14	0.36			
R15	0.28		0.36	
R16			0.43	
R17			0.72	
R18			0.57	
R19			0.56	
R20			0.66	
R21		0.29		
R22		0.39		
R23		0.44		
R24		0.30		
R25		0.51		
R26		0.57		
R27		0.53		
R28		0.42		
R29		0.41	0.21	
R30		0.47		
R31		0.29		
R32		0.54		
R33				0.45
R34				0.36
R35				0.23
R36				0.54
R37				0.60
R38	0.23			0.46
R39				0.46
R40				0.51