JLTA 2015 Keynote Speech

A socio-cognitive approach to assessing speaking and writing: the GEPT experience

Jessica R. W. WU
The Language Training and Testing Center, Taiwan

A socio-cognitive approach to assessing speaking and writing: the GEPT experience

Jessica R W Wu
The Language Training and Testing Center, Taiwan
jw@lttc.ntu.edu.tw

Abstract

The General English Proficiency Test (GEPT) is a 5-level, criterion-referenced EFL testing system implemented in Taiwan to assess the general English proficiency of EFL learners at all levels. The GEPT was designed as a skills-based test battery assessing both receptive (listening and reading) and productive (speaking and writing) skills. Since its first administration in 2000, the GEPT has been taken by more than 6.5 million learners, and has become the largest-scale standardized English test in Taiwan. In the wake of the introduction of productive skills to the university entrance examination system in Japan, this talk aims to share the GEPT experience. Several key issues about speaking and writing tests will be presented in relation to the socio-cognitive framework for validation (Weir, 2005). A number of examples about GEPT validation will be illustrated to demonstrate that both *a priori* and *a posteriori* validity evidence are required to establish test quality. The paper also emphasizes the importance of facilitating communication between test developers and stakeholders when introducing new assessment.

1. Introduction

In the wake of the introduction of productive skills to the university entrance examination system in Japan, this paper aims to share Taiwan's GEPT experience with assessing speaking and writing in relation to the socio-cognitive framework for validation (Weir, 2005). This paper consists of three parts. First, it gives an introduction to the context in Taiwan in terms of its university entrance system and the use of GEPT scores for university admission. Second, a number of GEPT Speaking and Writing validation cases are illustrated to demonstrate that both *a priori* and *a posteriori* validity evidence are required to establish test quality. Third, it suggests what test developers can do to enhance positive washback and impact.

2. Taiwanese Context

Before 2001, we had the Joint College Entrance Examination (JCEE), which was administered only once each year. Based on years' of research on educational reform and testing development, the JCEE was replaced by two separate exams: the General

Scholastic Ability Test (GSAT) and the Advanced Subjects Tests (AST). All senior high school graduates are now required to take the GSAT during the winter break. The AST, however, is optional. It is administered in early July.

Only reading and writing skills were initially assessed in both the GSAT and AST; listening wasn't even included until 2014. So far, it has not been possible to include speaking mainly due to the large test population that is expected. Currently, there are about 120,000 candidates for the GSAT and 70,000 for the AST annually. A high school graduate can be assigned to a university based on his/her AST scores. Alternatively, students can simply submit their GSAT scores to apply to a university. In most cases, students are required to provide supporting documents for application, of which a proof of English ability is required by 80% of the universities in Taiwan. As a result, the GEPT Intermediate and High-Intermediate certificates, which are equivalent to CEFR B1 and B2, respectively, are the most convincing evidence of students' English ability as they assess both receptive and productive skills, thus meeting universities' needs to screen applicants.

3. The GEPT

The GEPT is a five-level criterion-referenced EFL testing system, which targets English learners in Taiwan at all levels, from junior high school upwards. The development of the GEPT was started as an in-house project of the Language Training and Testing Center (LTTC). Later, it was partially funded by Taiwan's Ministry of Education with the aims of promoting life-long learning and introducing positive washback effect on the learning and teaching of English. Since its launch in 2000, the GEPT has been administered independently by the LTTC. Currently, the GEPT is the largest standardized English language test in Taiwan, which is taken by approximately 500,000 test takers at over 100 test sites around the country each year. As a result, GEPT scores are also recognized as proof of English ability by government offices, schools, and employers.

The test content of the GEPT is not only linked to the local English curriculum, but also takes account of local cultural and social references. We hope that by providing a testing context that is suitable for local learners linguistically, visually, and cognitively, we can increase learners' motivation to learn and help them demonstrate their best performance in the test. The levels of the GEPT, which are also linked with the CEFR (Council of Europe, 2001) empirically, are roughly equivalent to CEFR A2-C1 (e.g. Wu & Wu, 2010; Wu, 2011). More details about the GEPT and associated research are available at www.gept.org.tw.

Next is a quick overview of GEPT speaking and writing specifications. For speaking, a semi-direct method is adopted at the first three levels. Yet, at the higher levels, a direct method is employed, with an interlocutor who interacts with 2 to 3 test-takers. For writing, both Chinese-to-English translation and essay writing are assessed at the lower levels. However, at the higher levels, read-to-write skills are assessed. Candidates are required to perform two tasks: summarizing the main points from two texts before writing an argumentative essay and interpreting two charts before writing a persuasive letter.

4. The socio-cognitive framework for validation

It is test developers' responsibility to examine validity continuously. According to the APA Standard (1999), validity is an argument based on evidence to support score interpretation and use. As we know, the major sources of validity evidence include test content, response processes, internal structure, relations to other variables, and testing consequences. Yet, how to collect validity evidence from various sources in a more comprehensive way or how to carry out validation more effectively are also important issues. As the test developer of the GEPT, the LTTC decided to adopt the socio-cognitive framework for validation (Weir, 2005) on which to lay out our research agenda. One of the merits of the framework is that it covers both the development stage *(a priori)* and the appraisal stage *(a posteriori)*. The former focuses on the work before the test is actually operationalized. At the stage of developing test specifications, we need to address different facets of validity, including construct, consequential, and criterion-related. *A priori* validity evidence can be collected during the development stage. Nevertheless, in the real world the test may not be realized as designed. As a result, we must seek *a posteriori* cumulative evidence to further establish validity after the test is administered.

The notion of construct validity (Weir, 2013) is well suited to the assessment of productive skills. In Weir's view, construct resides in the interactions between cognitive ability, the context of use in which the task is performed, and scoring, which often involves human raters. He further reminds test developers of the importance of satisfying the expectations of stakeholders with regard to the comparability of the constructs measured by each test version in terms of both cognitive and contextual validity, and scoring validity. In line with this notion, due attention to the underlying properties of speaking should be paid when we develop a speaking test. For example, under context validity, we need to specify task requirements in terms of linguistic input and output. Also, due to the nature of oral communication, we need to specify interlocutor-related factors such as the number of interlocutors, the relations between interlocutors and test-takers, speech rate, etc. For writing assessment, on the other hand, we need to consider the writer-reader relationship. Since both speaking and writing tests elicit constructed responses, issues such as rating criteria, rating procedures, raters, and rater standardization are all closely associated with test validity. .

In the next section, I will highlight a number of GEPT studies to illustrate how to examine the validity of speaking and writing tests.

Case One – The GEPT speaking construct at the Intermediate level

This is a multi-dimensional approach to investigating the construct of the GEPT speaking test at the intermediate level. A paper based on the study was published in *Language Testing* (Weir &Wu, 2006).

Due to its large test population, the GEPT employs the semi-direct method in its speaking component at the first three levels. Each test administration consists of multiple test sessions, each requiring a different test paper. Therefore, it is essential to establish parallel-form reliability and to demonstrate that the tests are comparable.

To establish parallel-form reliability in the speaking test, we compared the construct of three different test papers in terms of code complexity (lexical and syntactical difficulty), cognitive complexity (content familiarity), and communicative demand (time pressure). Data from different sources, including task scores and interlanguage measures in the areas of accuracy, fluency, complexity, and lexical density, were analyzed. By means of both qualitative (expert judgments of task difficulty and language functions) and quantitative analyses (correlation, ANOVA, factor analysis, Multi-Faceted Rasch Measurement), the results support the claim that the test papers can be considered parallel.

Case Two – The GEPT Writing tests at the Advanced level

The concept of 'validity by design' is evident in the GEPT. As the following table shows, the GEPT Advanced Writing Test requires candidates to summarize main ideas from both verbal and non-verbal inputs and express opinions. It provides a simulation of writing tasks for academic purposes, i.e., reading-to-write and writing for a specified purpose to a specific audience. Sample test tasks can be found at www.lttc.ntu.edu.tw/GEPT1/Advanced/writing/writing.htm.

Table 1 GEPT-Advanced Writing Test format and structure

Part	Task types	Time (mins)
1	Summarizing main ideas from verbal input and expressing opinions (250 words)	60
2	Summarizing main ideas from non-verbal input and providing solutions (250 words)	45

However, these descriptions can only be considered *a priori* evidence. We also need to demonstrate the test quality through *a posteriori* validation. Chan, Wu and Weir (2014), in a collaborative project between the LTTC and Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, investigated the context and cognitive validity of GEPT Advanced Writing Task 1. In a writing test context, context validity addresses the particular performance conditions under which the task is to be performed (e.g. purpose of the task, input to be processed, time available, length required, marking criteria as well as the linguistic demands inherent in the successful performance of the task, etc.), and cognitive validity is what test takers will activate cognitively in response to the contextual parameters set out in the performance conditions. However, a more pertinent question for us to ask is whether or not all the requirements we place on test takers when they perform in the writing test are similar to those they will meet in non-test 'real-life' situations. In other words, we need to investigate the similarity between test takers' test performance and their actual performance in real-life tasks in terms of context and cognitive validity. Therefore, Chan et al. (2014) addressed two research questions:

1. What are the relationships between the contextual parameters set in the GEPT Advanced Writing Test and those set in the real-life academic writing tasks in a business school at a UK university?
2. What are the relationships between the cognitive processing activities elicited from the GEPT Advanced Writing Test and those elicited from real-life academic writing tasks in a UK university?

Both expert judgment and automated textual analysis such as VocabProfile (Cobb 2010) and Coh-Metrix (Graesser, McNamara, Louwerse & Cai 2004) were employed to examine the degree of correspondence between the overall task setting and input text features of the GEPT task and those of the target academic writing tasks in real-life university business courses in the UK. As for cognitive validity, this study examined the cognitive processes elicited by the GEPT task in comparison to the real-life academic writing tasks through a cognitive process questionnaire. The demonstration of a close similarity between the test and real life conditions in the findings supports the context and cognitive validity of GEPT Advanced Writing Task 1, an integrated reading-into-writing task. In addition, the results have important implications for university admissions officers and other GEPT score users when considering whether the test is a valid option for assessing English writing for academic purposes.

Case Three – Enhancing scoring validity

Another dimension of construct validity is scoring validity, which can be established and enhanced through rigorous quality control procedures, from recruitment and training of raters to monitoring and evaluation. The rating processes should be examined in terms of speed, quality, and both inter- and intra-rater reliability. During each GEPT scoring session, raters' rating speeds are monitored, which helps us to identify the raters who may have worked too fast or too slow. Thus, we can provide assistance to the raters in a more effective manner. In addition, scoring reliability is also measured, and a monitor report is produced on the scores awarded by each rater, including mean score, standard deviation, and score distribution across score bands. Such monitor reports, generated two to three times daily during the rating session, help us to identify raters who may have problems with scoring, allowing necessary action to be taken to resolve the problems in due course.

Most GEPT raters are school teachers who are non-native speakers of English. They tend to focus more on linguistic accuracy, so they tend to be stricter about lexical and grammatical errors and more lenient when test-takers use more complex vocabulary and structures. Also, raters tend to judge the quality of the writing or speaking performance by comparing across performances or using their own evaluation criteria without applying the GEPT rating scales. These are behaviors that we've observed among GEPT raters in the past years. You may, however, find very different behaviors among the raters in Japan.

Case Four – How raters interact with rating scales

We understand that productive skill performance is always rated against a set of evaluation criteria. However, the application of the criteria is ultimately dependent upon how raters interpret them. The next example to be shared is an investigation of the rating processes using an analytic rating scale developed for the GEPT Advanced Writing Test (Wu & Ma, 2013). In this study, we aimed to answer the following questions:

- What do raters attend to when rating an essay?
- How frequently do raters refer to scoring criteria when making scoring decisions?
- Do raters base their decisions on features which do not directly reflect the scoring criteria?

The raters were asked to provide a verbal report of their rating processes, which were transcribed and then analyzed with reference to the evaluation criteria specified in the rating scale (coherence & organization; grammatical use; content relevancy, LU for lexical use). The results suggest that although the raters scored reliably and the degree of inter-rater reliability is acceptable, they may award similar scores for different reasons, suggesting that raters have different interpretations of the rating scale. Based on the findings of this study, we decided to improve the clarity of the wording of the scale and rater training materials in order to enhance scoring validity.

5. Enhancing positive washback

It cannot be denied that test performance data can inform teaching and learning and help achieve positive washback. For example, writing is the most difficult among the four skills for Taiwanese learners to acquire. To describe the features of GEPT performance and to better understand Taiwanese learners' writing difficulties, we've constructed a learners' corpus, which consists of 2 million words, based on GEPT writing performance. A corpus website provides public access to this data, allowing teachers and researchers to conduct keyword and collocation searches.

Also, an online writing learning and assessment system, named Dr. Writing, has been constructed. This system is not intended to aid test-preparation; instead, it is designed to facilitate learning. Although learners practice writing by completing a GEPT writing task, and their performance is evaluated according to the GEPT criteria, they receive individualized feedback in addition to a score. We hope this resource can be utilized to support formative assessment in classrooms.

Through the use of the Dr. Writing system, we are able to identify the strengths and weaknesses of learners' writing across levels. The data suggest that six major types of mistakes make up of over 90% of the errors. Also, most of the errors are caused by the differences between Chinese, learners' first language, and English. Based on the findings, a self-assessment checklist was later developed to help learners examine their own performance against a list of the features of an effective expository essay. This example demonstrates how test performance data can be transformed to provide pedagogical resources for EFL writing classes.

6. Conclusion

As Japan is launching the new productive skill assessment in the university entrance examination, I'd like to conclude the paper by highlighting the importance of facilitating communication with stakeholders.

First, it is essential to promote language assessment literacy and communicate with stakeholders. However, this should be done in a less technical manner. As assessment professionals, we tend to use very technical language when communicating with stakeholders. In reality, even school teachers have difficulty understanding the technical aspects of language assessment, not to mention lay persons.

Second, test developers should increase the transparency of their work. To achieve this, validation reports should be published periodically. Validation studies