Ensuring Quality and Fairness in the Asian EFL context: Challenges and Opportunities

Dr Jessica Wu
Language Training & Testing Center (LTTC)
Taipei, Taiwan

2014 ALTE
Introduction

- **Quality and Fairness** are the overriding concerns in all aspects of assessment.
  
  AERA/APA/NCME Standards (1999)
  ILTA Code of Ethics (2000)
  ALTE Principles of Good Practice (2001)

- No such professional standards have been specifically developed in Asia.

- However, Ensuring test quality and fairness is a highly regarded topic in Asia, too.
Without codes of ethics and good practice specifically developed in Asia, how do testing bodies in Asia address issues in relation to fairness and quality in their testing programs or services?

What are the issues and challenges? How are they overcome?

The Asian EFL domain including China, Japan, Korea, and Taiwan (Confucian-heritage cultures)
Locally-produced EFL tests

- Tailored to the specific educational systems and the changing contexts of test use
- High-stakes exams on a very large scale

- EIKEN (Japan) since 1963; 2.3 millions
- CET (China) since 1987; 18 millions
- GEPT (Taiwan) since 2000; 0.6 million
- NEAT (Korea) since 2012; 50,000
Commonalities among the tests

- Including positive impact on English learning and education as a stated objective – meet each own specific needs and take cultural factors into consideration
- Having reported success at introducing positive washback
- Increasing language assessment literacy
- Encouraging research in LTA
- Understanding learners’ strengths and weaknesses
Challenges in ensuring test quality

International standards and codes of practice are inappropriate or too difficult to be implemented due to large test-taking populations.

A. In the case of oral assessment ...

- too costly and impractical to use face-to-face interviews
- semi-direct tests or a two-stage design
- the validity of the speaking test format is questionable
- empirical investigations of issues concerning the improvement in an effort to strike a balance between controllability and spontaneity.
Challenges

B. Overusing the multiple-choice (MC) format to cope with the large scale of the tests and also a consequence of the “psychometric-structuralist” approach (Spolsky, 1995).

Producing “good” multiple-choice items ???

More empirical research is required to provide evidence in support of validity
Challenges

Numerous practical constraints, e.g., limited human resources – qualified professionals (item writers, markers, statisticians)

Inadequacies in quality control procedures:
  - Pretesting (size and representativeness), marking (double-marking of constructed-response items, the monitoring of the marking process), and test equating.

A survey of English language testing practice of six EFL examination boards in China (Jin & Fan, 2013).

Much variation in the testing practices was identified.

The EFL testing in China is in urgent need of professional standards.
Challenges

- The findings are generalizable to the other Asian contexts.
- In addition to the principles which are general and universal, principles that are context-specific should also be developed.
- The reflection of local features in developing professional standards for EFL testing in China (Fan, 2013)
Ensuring Quality

- The primary responsibility of the test developer is achieving an appropriate balance among the four examination qualities: validity, reliability, impact and practicality (Bachman & Palmer, 2010, p. 433).

- The relationship between test validity and test fairness: Fairness should be treated as an aspect of validity - ‘A test has to be fair to be valid.’ (Xi, 2010)

- How do the developers of the four Asian-produced tests shoulder their responsibilities?

- Experience with the GEPT in Taiwan
A socio-cognitive framework for test development and validation (Weir, 2005)
Measured Construct

Construct as residing in the interactions between an underlying cognitive ability (trait), a context (of use) in which the task is performed, and a process of scoring.

“With an increased public expectation of transparent and explicit test specification in the late 20th century, a broader conceptualization of construct validity (i.e. qualitative as well as quantitative) was seen as necessary. Therefore, test providers need to satisfy the expectations of stakeholders (learners, employers, receiving institutions, professional bodies) concerning the comparability of the constructs measured by each test version in terms of both cognitive and contextual validity, and scoring validity.”

(Weir, 2013: 3-4)
Case 1: Construct of the GEPT Speaking
A Multi-dimensional Approach

Establishing the Parallel Tasks (Weir & Wu, 2006; Language Testing, 23(2))

Code complexity (lexical and syntactical difficulty), cognitive complexity (content familiarity), and communicative demand (time pressure).

Sources of data: task scores, responses to post-task questionnaires, interlanguage measures in the areas of accuracy, fluency, complexity, and lexical density.

Quantitative approach: Conventional statistical procedures, such as Correlation, ANOVA, and factor analysis, MFRM

Qualitative approach: Checklists of Task Difficulty, Dale-Chall Readability Formula, Checklists of Language Functions
Case 2: Construct of the GEPT Advanced Reading and Writing

“It is thus a positive development that the reading comprehension tests of the GEPT Advanced Level value both parts of the reading construct in equal measure.” Weir (2013)

GEPT test format and structure

<table>
<thead>
<tr>
<th>Paper</th>
<th>Part</th>
<th>Task Types</th>
<th>No of items</th>
<th>Time (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>1</td>
<td>Careful reading</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Skimming &amp; scanning</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Writing</td>
<td>1</td>
<td>Summarizing main ideas from verbal input and expressing opinions (250 words)</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Summarizing main ideas from non-verbal input and providing solutions (250 words)</td>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>
Validity by design

Demonstrating test quality through a posteriori validation

Investigating the context and cognitive validity of GEPT Advanced Writing
(a joint project between LTTC and CRELLA; Chan, et al., 2013)
Context validity & Cognitive processing

Context validity for a writing task addresses the particular performance conditions, the setting under which the task is to be performed (e.g. purpose of the task, input to be processed, time available, length required, specified addressee, known marking criteria as well as the linguistic demands inherent in the successful performance of the task).

Cognitive processing in a writing test never occurs in a vacuum but is activated in response to the contextual parameters set out in the wording of the writing task.
Real life

Language tests should, as far as possible, place requirements on test-takers similar to those they will meet in the non-test ‘real-life’ situations.
Research Questions

What are the relationships between the contextual parameters set in the GEPT Advanced Writing and those set in the real-life academic writing tasks in the Business School in a UK university? (both expert judgment and automated textual analysis were employed to examine the correspondence between the overall setting and input text features of the GEPT task and those of the target academic writing tasks in real-life academic writing tasks)

What are the relationships between the cognitive processing activities elicited from the GEPT Advanced Writing and those elicited from the real-life academic writing tasks in a UK university? (through a cognitive process questionnaire)

A close similarity between the test and real-life conditions supports the context and cognitive validity of the writing test.

The results have important implications for university admissions officers and other stakeholders to consider the test a valid option when considering writing tests for academic purposes.
Case 3: Criterion-related validity

“Criterion-related validity is a form of external evidence, which is defined as ‘a predominantly quantitative and a posteriori concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance with established properties.’” (Weir, 2005:35)

“it is essential that any examination board follows clearly defined and public quality standards and aligns its tests to internationally-recognised frameworks, particularly the Common European Framework of Reference, which is now used worldwide to explain levels of achievement in language learning.” (Interview with Dr Michael Milanovic, The Way of Language, 2013)
CEFR in Taiwan

- The Ministry of Education has used it as a common yardstick to help interpret learners’ proficiency in English.

- All English language tests are required to align to the CEFR.

The CEFR can be used as an external criterion.

GEPT-CEFR linking studies
Mapping GEPT with CEFR as a validity criterion (Wu & Wu, 2010)

External evidence

*Aligning Tests with the CEFR: Reflections on Using the Council of Europe’s Draft Manual* (pp. 204-224), CUP
Further studies

Comparing different levels of GEPT Reading in terms of contextual parameters and cognitive processing skills by automated textual analysis (VocabProfile, Coh-metrix) and expert judgment
Contextual parameters

Test Construct

general purpose

rhetorical organisation

lexical frequency

response format

syntactic complexity

communicative topic

text length

time constraint

readability

text abstraction

cohesion
Cultural knowledge

Subject knowledge topic familiarity
Cognitive operations across GEPT levels

- Word recognition
- Lexical access
- Establishing propositional meaning at clause and sentence level
- Syntactic parsing
- Inferencing
- Mental model building
- Creating a text level representation
- Creating an intertextual representation

Levels:
- Elementary
- Intermediate
- High-Intermediate
- Advanced
International Recognition of the GEPT

- Started promoting recognition of the GEPT internationally in 2010
- More than 60 universities around the world accept GEPT scores when considering Taiwanese students’ applications for admission.
  - Test quality (Reliability & validity).
  - Mapping with an international framework (CEFR)
  - Cross-test comparability with international English tests (Score conversion)
  - Predictive power (Correlation between GEPT scores and real-life academic performance)
# GEPT Research Grants (since 2010)

| University of Bedfordshire UK (completed) | Examining the Criterion-Related Validity of the GEPT Advanced Reading and Writing Tests: Comparing GEPT with IELTS and Real-Life Academic Performance |
| University of Melbourne Australia | Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR) |
| California State University USA (completed) | An Investigation into the Comparability of the GEPT Advanced Level and TOEFL iBT |
| Lancaster University UK (Completed) | Linking the GEPT Listening Test to the Common European Framework of Reference |
| Hong Kong Polytechnic University (completed) | A Register Analysis of Advanced GEPT Examinees' Written Production |
| University of Bedfordshire UK | Examining the Cognitive Validity of GEPT High-Intermediate and Advanced Reading: an Eye Tracking and Stimulated Recall Study |
| University of Bristol UK (completed) | A Comparability Study on the Cognitive Processes of Taking GEPT (Advanced) and IELTS (Academic) Writing Tasks Using Graph Prompts |
Striving for Fairness

- Test developers’ responsibilities do not end with test development.

- Greater professional and social responsibilities due to the changing context of test use - in the broader context of ‘test use’ (Shohamy, 2000)

- Intended uses (improving English, promoting positive washback)
  - Lack of assessment literacy (decision makers, teachers, test-takers)
  - Competitive culture

Unintended uses (selection for admission & employment, residential permit)
Negative Consequences

- A decline in moral standards (cheating, fake score reports)
- Teaching to the test (narrowing teaching content to what is tested and replacing classroom teaching with test preparation)
- Learning to the test (focusing on what is tested and taking mock tests of poor quality)
- The higher the stakes of the test are, the greater tension exists.
Teaching and Learning to the Test
Ranking of TOEIC scores in Asia

ETS多益台灣區代表於今日公布「2012年全球多益測驗排行榜」，指出2012年台灣考生多益成績在全球排名第五。台語成績在亞洲排名第六，並以全亞洲第80名的成績表現

台灣考生多益成績在亞洲排名第六，並以全亞洲第80名的成績表現。
Exam classrooms to get air conditioning: MOE

The national exams for college and high school entry might just be a little less stressful this year, after the Ministry of Education announced that for the first time, air conditioning would be allowed in the exam rooms. This measure is intended to cool heated minds at a temperatures of over 30 degrees Celsius.

The temperature is uniformly not going to exceed 30 degrees for both the college and high school entry exams, the ministry added.

The topic of allowing air conditioned exam rooms has been hotly debated since April last year. As Taiwan's weather in July tends to reach temperatures over 30 degrees Celsius, students have long suffered from the stress of the heat and pervasive body odors in addition to an already nerve-wracking test experience.

Public polls have shown that over 85 percent of parents are in favor of air conditioners running during national exams. MOE Deputy Minister Lin Tsong-ming said parental concerns have been heard and a test of air-conditioned classrooms will be held the week prior to exams to ensure that all coolers run smoothly.

The College Entrance Examination Center (CEEC) reminded students to bring light sweaters if they feared 26 degrees Celsius might be too chilly. Students who believe they work better at natural temperatures do have an option. In their exam registration form, to forgo an air conditioned room, the CEEC added.
Joint Responsibilities in striving for fairness

“Tests are not neutral but rather embedded in political, social and educational, ideological and economic contexts.” (Shohamy, 2001)

Increasing language assessment literacy and educating stakeholders (decision makers, teachers, test-takers) is one approach to ameliorating these changing or unintended consequences.

● Joint responsibilities of the test users and test developers in striving for fairness – ALTE Principles of Good Practice
Turning Problems into Opportunities

- Increased professionalism and measurement expertise among test development teams
- Continued commitment to developing and administering assessments whose uses can be justified (i.e., the use of the test will provide beneficial consequences for stakeholders)
- Willingness to persevere despite limited resources and an often capricious educational policy context

-Comments given by Prof. Lyle Bachman in the symposium on English language tests developed in Asia in a symposium in 2013 LTRC.
“In comparing international tests with locally-developed ones, it would be wrong to assume that the former, even though developed by native speakers of English, are always superior…… Global, multi-national, generic language tests taken by people around the world are unlikely to be particularly sensitive … to the needs of people within a particular society. In contrast, domestic tests can more easily be tailored to the local educational system and the needs of learners within a country.” (Weir, 2013)
Joints Efforts of Testing Bodies in Asia

- The Academic Forum on English Language Testing in Asia (AFELTA) - the 17th year since its establishment

- Eight institutional members, including CET in China, EIKEN in Japan, GEPT in Taiwan, NEAT in Korea, and others (Hong Kong Exam and Evaluation Authority, Singapore Examinations and Assessment Board)

- LTTC-GEPT offered workshops for Vietnam National University (Hanoi)

- Joint investigation of the testing practice of major test developers in and the development of the professional standards which connect to the Asian context.

- Standards being developed in China and ALTE’s experiences
Closing remarks

- We have the same responsibility to achieve fairness and quality, though we may adopt different methods to achieve our common goals.

- There will be more discussion and collaboration not only among the testing bodies in Asia, but among the global community of language testing.

- Actively participate in the development and revision of international standards and contribute our local knowledge to the development of language testing at an international level.
Thank You