The Challenge of Large-Scale English Language Testing in China

Yan Jin,
Shanghai Jiao Tong University
INTRODUCTION

English language testing in China
## English language testing in China

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Education level</th>
<th>Test</th>
<th>Proficiency level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission</td>
<td>High-school</td>
<td>Zhong Kao</td>
<td>Band 5</td>
</tr>
<tr>
<td></td>
<td>College or university</td>
<td>Gao Kao (NMET)</td>
<td>Band 8</td>
</tr>
<tr>
<td></td>
<td>Graduate school</td>
<td>GSEEE</td>
<td>/</td>
</tr>
<tr>
<td>Program exit</td>
<td>Vocational English</td>
<td>PRETCO-B</td>
<td>Level B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRETCO-A</td>
<td>Level A</td>
</tr>
<tr>
<td></td>
<td>College Non-English major</td>
<td>CET-4</td>
<td>Band 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CET-6</td>
<td>Band 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CET-SET</td>
<td>Grade A, B, C, D</td>
</tr>
<tr>
<td></td>
<td>College English major</td>
<td>TEM-4</td>
<td>Band 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEM-8</td>
<td>Band 8</td>
</tr>
<tr>
<td>Proficiency</td>
<td>Public English language test</td>
<td>PETS</td>
<td>Band 1, 2, 3, 4, 5</td>
</tr>
</tbody>
</table>
The scale of English language tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM</td>
<td>0.50m</td>
</tr>
<tr>
<td>GSEEEE</td>
<td>1.65m</td>
</tr>
<tr>
<td>PETS</td>
<td>2.04m</td>
</tr>
<tr>
<td>NMET</td>
<td>9.15m</td>
</tr>
<tr>
<td>CET</td>
<td>17.98m</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>......</td>
<td>0.50m</td>
</tr>
</tbody>
</table>
The high-stakes purposes of EFL tests

- College admission
- Graduate program admission
- College graduation
- Employment and career
- ......
Modern Language Testing at the Turn of the Century: Assuring What We Count Counts

“a strong program of test validation that includes considerations of ethical test use”

Bachman (2000, p. 1)
Test validation

“steers between the Scylla and Charybdis of what Messick called construct under-representation, on the one hand, and construct-irrelevant variance, on the other”

McNamara & Roever (2006, p. 18)
Ethical test use

Language testing needs to go beyond the cognitive and psychological models and explore a model that encompasses a clear social dimension

McNamara & Roever (2006)
Challenges

- Construct underrepresentation: native speaker norms
- Construct-irrelevant variance: Computer-based language testing
- Ethical language test use
CHALLENGE

Native-speaker norms?
The relevance of native speaker norms

As the use of English in international communication increasingly involves only nonnative speakers in many settings, the relevance to this communication of relatively distant native speaker norms is increasingly questioned. The topic raises complex sociolinguistic, policy, cultural, and political issues that are only beginning to be explored...

McNamara & Roever (2006: 252)
A paradigm shift for English language teaching within the Asian region that is responsive to

“the increasing use of English as a language of contact between non-native speakers across national boundaries”

Wang & Hill (2011, p. 206)
Three ways of using English as an international language

1. **World Englishes (WEs) approach**
   Using a variety of Englishes in the lectures and dialogues that form the stimulus material in a test—an approach currently being used in international tests.

2. **English as a lingua franca (ELF) approach**
   Elder and Davies (2006): testing it with language input that is standard English but designing test accommodations that modify the test delivery system in order to make it accessible and fair for ELF users without changing the construct; Or, testing ELF as a construct onto itself while favoring strategic competence.

3. **Locally defined English as an international language (EIL) approach**
   Decisions are based on carefully considered local needs for English including its international uses. Such local needs will typically be based on a thorough needs analysis of the EIL language and context involved in a particular local English learning situation.

J.D. Brown, 2012
## Accents in listening tests

<table>
<thead>
<tr>
<th>Test taker</th>
<th>Accent</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM</td>
<td>English major</td>
<td>British &amp; American, Australian, Australian, New Zealand, Canadian, Canadian</td>
</tr>
<tr>
<td>GSEEE</td>
<td>College graduate</td>
<td>British &amp; American</td>
</tr>
<tr>
<td>PETS</td>
<td>English learner</td>
<td>British &amp; American</td>
</tr>
<tr>
<td>NMET</td>
<td>High school leaver</td>
<td>British &amp; American</td>
</tr>
<tr>
<td>CET</td>
<td>Non-English major</td>
<td>British &amp; American</td>
</tr>
<tr>
<td>IB-CET</td>
<td>Non-English major</td>
<td>British &amp; American</td>
</tr>
<tr>
<td>TOEFL iBT</td>
<td>High school leaver</td>
<td>British &amp; American, Australian, New Zealand, Canadian</td>
</tr>
<tr>
<td>IELTS</td>
<td>High school leaver</td>
<td>British &amp; Australian, American, New Zealand</td>
</tr>
<tr>
<td>PTE-A</td>
<td>High school leaver</td>
<td>British &amp; American, non-native accents</td>
</tr>
</tbody>
</table>
## Example 1: CET Listening

### Paper-based
- Fully scripted
- Recorded in a studio
- Native speaker with a standard accent
- Preferably English radio/TV broadcaster or teacher/editor/writer
- Male + Female
- British + American

### Internet-based
- Audio clips (radio program)
- Video clips (TV, movie, online materials)
- Possible presence of NNS or NS with a local accent
- Questions on NS with a standard accent
Example 2: IELTS Listening

Listening
30 minutes

You will listen to four recorded texts, monologues and conversations by a range of native speakers and write your answers to a series of questions.

These include questions which test your ability to understand main ideas and detailed factual information, ability to understand the opinions and attitudes of speakers, ability to understand the purpose of an utterance and the ability to follow the development of ideas.

A variety of voices and native-speaker accents are used and each section is heard only once.
Example 3: TOEFL iBT Listening

New Native-speaker English Accents Added to the TOEFL iBT® Test

Beginning in March 2013, the Listening and Speaking sections of the TOEFL iBT® test will include other native-speaker English accents in addition to accents from North America. Test takers may hear accents from the United Kingdom, New Zealand or Australia.

Visit TOEFL iBT Test Content for more information and to hear examples.

Posted: March 1, 2013

http://www.ets.org/toefl/important_update/english_accents_added
Example 4: PTE-Academic Listening

Furthermore PTE Academic is the only academic English test to include non-native English accents. This reflects the diversity of English that students are likely to experience at any university where English is the medium of instruction, and where students will be taught by professors and teaching assistants who are not native speakers of English.

This use of international varieties of English and non-native accents demonstrates that PTE Academic is highly relevant to today’s modern international academic institutions.

More research of WEs and language testing

The attitude toward WEs in language testing will have a marked effect on the construct representation: test design, testing processes, rating criteria, rater training, score interpretation ...

JD Brown (2013):

7 recommendations on how language testers and the WEs community can start working together.

Recommendation No. 4:

Base tests on context, needs and decision purposes.
“… nothing important be left out of the assessment of the focal construct"
Computer-based language testing
Computer-based language tests


B. TOEFL iBT: since 2005 ([www.ets.org/toefl/ibt](http://www.ets.org/toefl/ibt))

C. IB-CET: since 2008 ([www.ccets.org](http://www.ccets.org))


E. CB-CETSET: since 2012 ([www.cet.edu.cn](http://www.cet.edu.cn))
Validity of CBLT/CALT

Growing concern about the potential threat to validity of computerized tests in the context of high-stakes testing:

Validation studies of large-scale computer-based language tests center on whether the use of technology will confound the measurement of test takers’ language proficiency with computer proficiency, resulting in test bias and unfairness.
Empirical studies

- Taylor et al. 1998: Computer familiarity and TOEFL CBT
- Sawaki 2001: PB and CB L2 reading assessment
- Choi et al. 2003: PB and CB TEPS in Seoul
- Brown 2003: Handwritten vs. word-processed IETLS essays
- Breland et al. 2004: Handwritten vs. word-processed TOEFL writing
- Wolfe & Manalo 2005: TOEFL composition medium & score quality
- Weir et al. 2007: CB IELTS vs. IELTS writing

Cross-modal validity aiming at establishing equivalence

- Mixed results: enhance / impede / have no effect on test performance

Assumption: Computer as a source of CIV
Cognitive processing (Weir 2005):

- (Jin, Wu and Yan, 2012): Cognitive processing of writing
  *Is computer literacy construct-relevant in a language test in the 21st century? (LTRC 2011)*

- (Jin & Zhang, L. 2012): Communicative strategies in speaking
  *The impact of test mode on the use of communication strategies in the paired discussion task (AFELTA 2012)*

- (Jin & Zhang, X. 2013): Cognitive processing of integrated and independent tasks
  *A comparative study of PTE-Academic and IB-CET (LTRC 2013)*
What is the role of computer literacy in the conceptualization of the construct of a CBLT?

(Is the ability to use the computer part of the construct to be measured?)
What do we want to measure in a CBLT?

**Computer operation in IB-CET**

| ✓ Click (MCQ) | ✓ Double click (de-select) |
| ✓ Drag and drop (match) | ✓ Type (SAQ, composition) |
| ✓ Cut and paste (composition) | ✓ Talk to the microphone |
| ✓ Highlight while reading | ✓ Scroll while reading & writing |
| ✓ Check and adjust equipment | ✓ Automatic spelling check |
| ✓ Check and adjust equipment |

| ? Automatic grammar check | ? Automatic capitalization |
| ? Online dictionary | ? Online resources |
| ? Cut & paste (summary writing) | ? Pauses of video clips (dictation) |
| ? Repeat (listening to repeat) | ? Previewing listening questions |
Social-cognitive construct representation

A global construct:
‘ability–in language user’ + ‘language user–in context’
Incorporates interaction from an individual-focused cognitive perspective (a psycholinguistic ability model)

A local, context-bound construct:
ability–in language user–in context
Adopts a social interactional perspective: individual ability and contextual facets interact in ways that change them both

(Chalhoub-Deville 2003: 369-383)
... future L2 construct exploration should address the ‘abilities – in language users – in contexts’, as well as the connections that enable language users to transfer relevant schemes to appropriately engage in a variety of communicative events.

(Chalhoub-Deville, 2003: 378)
A paradigm shift for the conceptualization of the construct of a CBLT/CALT:

- Specify computer-mediated TLU situations
- Design CBLT tasks and interfaces to elicit best performance
- Investigate
  - The interaction involved in language use among attributes of the test taker;
    between the test taker and task characteristics.
  - ......
We spend a lot of time worrying about the technical reliability and technical excellence of our instruments, and too little worrying about the validity for the purposes for which they are intended (or about the ethical justification for those uses).

(Spolsky, 1995: 9)
... once the assessment-based interpretations are actually used for the purpose for which they are intended, the assessment can develop a life of its own. It may be lured out of the well-described domain where the test developer intended it to reside, and other test users may co-opt it for uses beyond those for which it was developed

(Bachman & Palmer, 2010, p. 429)
College English Test (CET)

CET: A history of over 25 years
CET: the first decade

Psychometric-structuralist approach to language testing

- 100,000 in 1987 → 1 million in the mid-1990s

Ensuring the psychometric and technical qualities of this large-scale test: “A good test would automatically produce good effects in the classroom.”

(Wall, 2000, p. 505-506)
The social dimension of language testing (McNamara & Roever, 2006; Yang & Gui, 2007)

- **9.5 million** in 2005 → **18 million** in 2012

The large scale and the high stakes place greater professional and social responsibilities on the test developer. However, ...

“Is it possible for test developers to take the responsibility for the consequences of the test without being given sufficient power over the use of the test?”
CET: the intended purpose

Based on the requirements stipulated in the national College English teaching syllabus
(State Education Commission, 1985; 1986)

“check whether the English language proficiency of college students has met the requirements set in the national College English teaching syllabus”
(College English Test Working Group, 1987; 1988)

Promote the implementation of the teaching syllabus
CET: uses for other purposes

- Graduate program admission
- Employment opportunity
- Residential permit
- Diploma/degree
- .....
Consequences of overuse

- The De Facto curriculum
  - Teaching to the test
  - Learning to the test
- Cheating, fake certificates
  - Concern with test fairness
  - Decline in moral standards
- Corrupting the test
  - When a quantitative indicator is used for social decision-making, it distorts and corrupts the indicator itself and the social process it was intended to monitor
  (Campbell, 1975, in Madaus et al., 2009: 155)
Who should answer for the consequences?

**Critical language testing**
- A hidden political agenda? (Shohamy, 2001, p. 113-114)
- Gatekeeping function of language tests (Spolsky, 1997)

The CET is used mainly for initial screening purposes.
- A fiercely competitive culture in which opportunities are insufficient or unevenly distributed

**Ethical language testing**
- The test developer and the test user should shoulder the responsibilities of test development and use.
  (Bachman and Palmer, 2010; Davies, 1997, 2004)
STAGES

1 Initial Planning

2 Design

3 Operationalization

4 Trialing

5 Assessment Use

Test developer’s responsibility

Decision maker’s responsibility

Needs to understand

Primary responsibility

Bachman & Palmer, 2010, p. 432
A responsible test developer/decision maker...

A responsible test developer needs to convince the decision maker that
- the assessment records are consistent,
- the assessment-based interpretations are meaningful, impartial, generalizable, relevant, and sufficient ...

A responsible decision maker needs to convince the other stakeholders that
- the decisions are values sensitive and equitable,
- the consequences are beneficial ...

(Bachman & Palmer, 2010, p. 433)
Cooperation among stakeholders

Fruitful cooperation toward a common goal is possible only if all the stakeholders are sufficiently equal in power and ability (Mathew, 2004, p. 123-134)

Empower test developers:
- Adequate funding for quality test development
- Truth about the intended purpose of the test
- Express expert views on how tests impact stakeholders

Educate test users:
With a good knowledge of the test, test users will be in a better position to justify their decisions and the consequences of the decisions.
Developing an “ethical milieu”

- Contribute to the development of an “ethical milieu” for language testing. (Davies, 1997b, p. 336)
  - Improve stakeholders’ assessment literacy.
  - Teachers: make the best use of a standardized language test and avoid teaching to the test.
  - Students: understand what is being assessed and how to improve performance on the test.
  - Develop a code of practice for language assessment in China.
Limits of language testing and language tests

The test developer: reconcile users’ high expectations with the ethics of test use

The test user: understand the test-based interpretations and the intended use of the test
Language testing: A weak profession

“... unlike medicine and the law, there is no quasi-legal body controlling entry and authorizing the right to practice.”

(Boyd & Davies, 2002, p. 307)

There are limits to what language tests can tell us about test takers and what test developers can do in their professional role as language testers.

Collaborate to strengthen their roles as promoters of professionalism and ethics in language testing.
Local English language tests

Global English language tests
## Strengths and Weaknesses

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contextualization:</strong> e.g., needs, culture, policy</td>
<td><strong>Theoretical model</strong> underpinning the test design</td>
</tr>
<tr>
<td><strong>Alignment to curriculum</strong></td>
<td><strong>Evidence supporting validity arguments:</strong> e.g., cognitive processing,</td>
</tr>
<tr>
<td><strong>Resources:</strong> e.g., human, facility, finance</td>
<td>meaningfulness of scores...</td>
</tr>
<tr>
<td><strong>Accessibility</strong></td>
<td><strong>Code of practice:</strong> e.g., piloting, equating, accommodation, post-test</td>
</tr>
<tr>
<td><strong>Cost effectiveness</strong></td>
<td>analysis, transparency</td>
</tr>
</tbody>
</table>

... ...
# Linking to the CEF

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IELTS</strong></td>
<td>-</td>
<td>-</td>
<td>4-5</td>
<td>5.5-6.5</td>
<td>7-8</td>
<td>8.5-9</td>
</tr>
<tr>
<td><strong>BEC-Higher</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>≥C</td>
<td>-</td>
</tr>
<tr>
<td><strong>PTE-Academic</strong></td>
<td>-</td>
<td>-</td>
<td>43-58</td>
<td>59-75</td>
<td>76-84</td>
<td>-</td>
</tr>
<tr>
<td><strong>TOEFL iBT</strong></td>
<td>-</td>
<td>-</td>
<td>57-86</td>
<td>87-109</td>
<td>110-120</td>
<td>-</td>
</tr>
<tr>
<td><strong>TOEIC_R/L</strong></td>
<td>60-109</td>
<td>110-274</td>
<td>275-399</td>
<td>400-489</td>
<td>490-495</td>
<td>-</td>
</tr>
<tr>
<td><strong>TOEIC_Speaking</strong></td>
<td>50-89</td>
<td>90-119</td>
<td>120-159</td>
<td>160-199</td>
<td>200</td>
<td>-</td>
</tr>
<tr>
<td><strong>TOEIC_Writing</strong></td>
<td>30-69</td>
<td>70-119</td>
<td>120-149</td>
<td>150-199</td>
<td>200</td>
<td>-</td>
</tr>
</tbody>
</table>

www.britishcouncil.org.cn/newdirections
# English language testing in China

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Education level</th>
<th>Test</th>
<th>Proficiency level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission</td>
<td>High-school</td>
<td>Zhong Kao</td>
<td>Band 5</td>
</tr>
<tr>
<td></td>
<td>College or university</td>
<td>Gao Kao (NMET)</td>
<td>Band 8</td>
</tr>
<tr>
<td></td>
<td>Graduate school</td>
<td>GSEEE</td>
<td>/</td>
</tr>
<tr>
<td>Program exit</td>
<td>Vocational English</td>
<td>PRETCO-B</td>
<td>Level A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRETCO-A</td>
<td>Level B</td>
</tr>
<tr>
<td></td>
<td>College</td>
<td>CET-4</td>
<td>Band 4</td>
</tr>
<tr>
<td></td>
<td>Non-English major</td>
<td>CET-6</td>
<td>Band 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CET-SET</td>
<td>Grade A, B, C, D</td>
</tr>
<tr>
<td></td>
<td>College</td>
<td>TEM-4</td>
<td>Band 4</td>
</tr>
<tr>
<td></td>
<td>English major</td>
<td>TEM-8</td>
<td>Band 8</td>
</tr>
<tr>
<td>Proficiency</td>
<td>Public English language test</td>
<td>PETS</td>
<td>Band 1, 2, 3, 4, 5</td>
</tr>
</tbody>
</table>
A PROFOUND NEED FOR RESEARCH

Plan
Use
Design
Trial
Operate
The End
Thanks For Your Attention

yjin@sjtu.edu.cn