Powerful Tests, Powerful Test Designers?

— New Challenges Facing the College English Test
Testing in China

► Imperial Examination system

► Testing (考: Kao) in present-day China

- 中考 (Zhōng Kāo): high school entrance
- 高考 (Gāo Kāo): college/university entrance
- 考研 (Kǎo Yán): graduate program entrance
- 考博 (Kǎo Bó): doctoral program entrance

... …

- TEM, CET, PETS ...
- TOEFL, IELTS, TOEIC, BEC, FCE, BULLETS ...
The power of tests

Tests have a very powerful role in modern societies.

Shohamy 2001: 113

Their results have detrimental effects on individuals as they can create

- winners and losers,
- successes and failures,
- rejections and acceptances.

Gāo Kǎo: college/university entrance
The power of tests

- **levers**
  of instructional reform
  (Chapman and Snyder 2000)

- **lever**
  for change
  (Pearson 1988: 101)

- **catalyst**
  for curricular innovation

- **strategy**
  to promote curricular innovation

- **mechanism**
  to drive instruction

Anders 2004: 39-42

- **ferocious master**
  of the educational process
  (Madaus 1988: 84-85)

- **disempowers**
  the great majority of teachers
  (Shepard 1991: 27)

- **unethical and undemocratic way**
  of making policy
  (Shohamy 2000: 11)
### Large-scale: What does it mean in China?

<table>
<thead>
<tr>
<th>Test</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gao Kao</td>
<td>8.67 million</td>
<td>9.50 million</td>
</tr>
<tr>
<td>Kao Yan</td>
<td>1.17 million</td>
<td>1.27 million</td>
</tr>
<tr>
<td>CET</td>
<td>10.5 million</td>
<td>12 million</td>
</tr>
<tr>
<td>PETS</td>
<td>800,000</td>
<td>900,000</td>
</tr>
<tr>
<td>TOEFL</td>
<td>120,000</td>
<td>21960</td>
</tr>
<tr>
<td>IELTS</td>
<td>100,000</td>
<td>130,000</td>
</tr>
</tbody>
</table>
CET-4: 1987-2006

CET-6: 1989-2006
CET: past, present and future

Part 1  How powerful is the CET?
  ▪ A brief history of the CET

Part 2  What has been done to tackle the power?
  ▪ Recent reform of the CET

Part 3  What can be done to further improve the washback of the CET?
  ▪ The way forward for the CET
Questions to be focused on

How powerful can a test be?

Do test designers have any power over the use of tests?

What are the responsibilities of test designers?
1. CET: A brief history

Test purposes

… tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or of society at large.

Bachman (1990: 279)
The intended purposes of the CET

A pressing need for university graduates with a higher level of English language proficiency

College English Teaching Syllabus

1985/1986: First national syllabuses

1999: Revised version

2004: Curriculum requirements
College English Teaching Syllabus

- **1985/1986:**
  Ability to read in English; Use English as a medium to access information related to their major areas of study

- **1999:**
  Reading + listening, speaking and writing abilities

- **2004:**
  Further develop listening and speaking abilities; Exchange information effectively through both spoken and written channels
The intended purposes of the CET

- Measure the **English language proficiency** of college and university students
- Promote the teaching and learning of English as a foreign language in colleges and universities in China (the implementation of the NCETS)
Beneficial effects

- **CE teaching**
  - more attention
  - more resources
  - more effective teaching:
    - feedback
    - information for improving teaching

- **CE learning**
  - more attention
  - more resources
  - more effective learning:
    - feedback
    - information for improving learning

- **Society**
  - objective evaluation of a student’s overall English proficiency
  - information for making educational policies
The National CET Committee

- An academic organization

  - 12 language teaching and testing experts

- National CET Committee (1994-now)
  - 3rd NCETC: 25 professors from 23 universities
  - leading professors at the forefront of applied linguistics research in China
Operational structure of the CET (1987-2005)

- Higher Education Department
  Ministry of Education

- Higher Education Department
  Provincial/Municipal Education Commission

- National CET Committee

- CET Administration Office
  - Test design
  - Test administration
  - Quality control
  - Score report

- Test Centre 1
- Test Centre 2
- Test Centre 3

- CET Item Writers
- CET Marking Centres
- Colleges and Universities

Jin 2007 in Cheng et al, forthcoming
Operational structure of the CET (since 2006)

- Higher Education Department
  - Ministry of Education

- National Education Examinations Authority

- National CET Committee

- Provincial/Municipal Education Examinations Authority

- CET Administration Office
  - Test design
  - Quality control
  - Score report

- Item writers
- 11 marking centres (2000-3000 markers)
- More than 1000 colleges/universities

Jin 2007 in Cheng et al, forthcoming
Efforts to maintain the professional standards of the CET

► Test validity: CET validation study
  - Componential view of the language construct
  - Content representativeness and coverage
  - Concurrent validity, predictive validity, face validity…

► Test reliability: A strict system of quality control
  - intra-marker consistency
  - inter-marker consistency
  - inter-centre consistency

Yang and Weir 1998
<table>
<thead>
<tr>
<th>p value</th>
<th>&lt; 0.1</th>
<th>~ 0.2</th>
<th>~ 0.3</th>
<th>~ 0.4</th>
<th>~ 0.5</th>
<th>~ 0.6</th>
<th>~ 0.7</th>
<th>~ 0.8</th>
<th>~ 0.9</th>
<th>~ 1.0</th>
<th>Mean= 0.653</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of items</td>
<td>1</td>
<td>10</td>
<td>19</td>
<td>25</td>
<td>25</td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rbis value</th>
<th>&lt; 0.1</th>
<th>~ 0.2</th>
<th>~ 0.3</th>
<th>~ 0.4</th>
<th>~ 0.5</th>
<th>~ 0.6</th>
<th>~ 0.7</th>
<th>~ 0.8</th>
<th>~ 0.9</th>
<th>~ 1.0</th>
<th>Mean= 0.430</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of items</td>
<td>3</td>
<td>10</td>
<td>17</td>
<td>31</td>
<td>25</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Technical data (CET-4 Prototype)

<table>
<thead>
<tr>
<th></th>
<th>Listening</th>
<th>Reading</th>
<th>Voc &amp; Str</th>
<th>Cloze</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>0.563</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voc &amp; Str</td>
<td>0.539</td>
<td>0.615</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloze</td>
<td>0.467</td>
<td>0.531</td>
<td>0.626</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>0.387</td>
<td>0.359</td>
<td>0.470</td>
<td>0.404</td>
<td>1.000</td>
</tr>
<tr>
<td>Total</td>
<td>0.792</td>
<td>0.892</td>
<td>0.802</td>
<td>0.708</td>
<td>0.581</td>
</tr>
</tbody>
</table>

Marking approach/scheme
Range-finders and sample scripts
Training of directors

Test Centre 1
3 marking centres

Directors
Training of supervisors
Spot-checking
Supervisors
Training of markers
Spot-checking
Markers

Test Centre 2
4 marking centres

Directors
Training of supervisors
Spot-checking
Supervisors
Training of markers
Spot-checking
Markers

Test Centre 3
3 marking centres

Directors
Training of supervisors
Spot-checking
Supervisors
Training of markers
Spot-checking
Markers

Cheng et al, forthcoming
Jin 2007

National CET Committee
System of CET essay marking
(since 2006)

Jin 2007 in Cheng et al, forthcoming
Remarks by the Vice Minister of Education

… The fact that such a large-scale test has been developing steadily in the past 17 years was in itself solid evidence to show that the CET has met the social needs, won social recognition, produced beneficial effects on society and contributed significantly to the continued improvement of the quality of College English teaching in China.

Wu Qidi
2nd News Conference, 2005, Ministry of Education
The National CET Committee

► Exchanges with UCLES, ETS, HKEAA, CEEC, LTTC, STEP, KELTA, etc.

► AFELTA: Academic Forum on English Language Testing in Asia
  ▪ Initiator/organizer and host of the 1\textsuperscript{st} and 7\textsuperscript{th} AFELTA

► International conferences
  ▪ Symposium at 2000 LTRC, Vancouver
  ▪ 1\textsuperscript{st} International Conference on English Language Testing in China, 2002, Shanghai
  ▪ “Big Tests” Symposium, 2005 AILA, Wisconsin
2. Reform of the CET: Challenges facing the NCETC

Uses of the CET for purposes not intended

- Requirement for college/university graduation
- Requirement for Bachelor’s degree
- Prerequisite to getting a job
- Prerequisite to getting residential certification
The power of high-stakes tests

Results of the test “are seen—rightly or wrongly—by students, teachers, administrators, parents, or the general public, as being used to make important decisions that immediately and directly affect them”

Madaus, 1988: 87
Negative effects

CE teaching

X teaching to the test
- use coaching materials
- narrow the teaching content
- stop teaching after students pass the test

CE learning

X learning for the test
- use coaching materials
- narrow the learning content
- aim at a narrow pass and stop learning after a narrow pass

Society

X inappropriate use of CET certificate
- e.g. when recruiting new employees for jobs that do not require English

Negative effects
What’s the point of a powerful language test ... if the test cannot have a good and happy effect on millions of EFL learners in China?
Reform of the CET: What has been done

Changes over the past twenty years

- 1990: separate paper for WRITING
- 1996: new item forms (CD, SAQ, translation, etc)
- 1997: Average Graded Score
- 1998: a minimum score requirement for CET-4 writing
- 1999: CET-SET
- ...
## Recent reform of the CET: Changes in the score reporting system

<table>
<thead>
<tr>
<th>Starting from</th>
<th>Mean</th>
<th>Sd</th>
<th>Score report</th>
<th>Score interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987/1989</td>
<td>72</td>
<td>12</td>
<td>Pass</td>
<td>Percentile position in the norm group (students in the six top universities in China)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Distinction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>June 2005</td>
<td>500</td>
<td>70</td>
<td>No pass or fail; Reporting total and profile scores</td>
<td>In the process of setting up a new norm</td>
</tr>
<tr>
<td>December 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

J in 2007 in Cheng et al, forthcoming
Changes in the score reporting system

► New score reporting system: Why?
  - Encourage more appropriate use of the CET test results
  - Reduce social pressure on the CET

► New score reporting system: How?
  - No pass or fail
  - Issue a score report instead of a certificate
  - Report both the total and profile scores
Recent reform of the CET:
Changes in the test content and format

► More emphasis on listening ability
► More attention to contextualized language use instead of context free knowledge of language
► A larger proportion of constructed response items
## CET test content

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td>20%</td>
<td></td>
<td>20%</td>
<td></td>
<td>35% Short conversations</td>
</tr>
<tr>
<td></td>
<td>• Short conversations</td>
<td>• Listening passages</td>
<td>• Short conversations</td>
<td>• Listening passages or Passage dictation</td>
<td>• Short conversations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35% Long conversations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35% Listening passages</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35% Passage dictation</td>
</tr>
<tr>
<td><strong>Reading</strong></td>
<td>40%</td>
<td></td>
<td>40%</td>
<td></td>
<td>35% 25% Careful reading</td>
</tr>
<tr>
<td></td>
<td>• Careful reading</td>
<td></td>
<td>• Careful reading</td>
<td></td>
<td>• 10% Skimming and scanning</td>
</tr>
<tr>
<td><strong>Vocabulary &amp; structure</strong></td>
<td>15%</td>
<td>• 7.5% Vocabulary</td>
<td>15%</td>
<td>• CET-4: 10% Voc. and 5% Structure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• 7.5% Structure</td>
<td>• CET-6: 15% Voc.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Writing</strong></td>
<td>15%</td>
<td>120-140 words in 30 minutes</td>
<td>15%</td>
<td>120-140 words in 30 minutes</td>
<td>20% 130-150 words in 30 minutes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Translation (Chinese to English)</td>
</tr>
<tr>
<td><strong>Integrated tasks</strong></td>
<td>10%</td>
<td>• Cloze or Error identification and correction</td>
<td>10%</td>
<td>• Cloze or Error identification and correction</td>
<td>10% • Cloze or Error identification and correction</td>
</tr>
<tr>
<td><strong>Speaking (optional)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Monolog</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Dialogue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Group discussion</td>
</tr>
</tbody>
</table>

In China, CET-4 is equivalent to the Japanese Eiken 3, while CET-6 reflects the English Fourth Year Proficiency Test. Jin 2007 in Cheng et al, forthcoming
## CET test format

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td>• MCQ</td>
<td>• MCQ</td>
<td>• MCQ</td>
<td>• MCQ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Compound Dictation</td>
<td>• Compound Dictation</td>
<td></td>
</tr>
<tr>
<td><strong>Reading</strong></td>
<td>• MCQ</td>
<td>• MCQ</td>
<td>• MCQ</td>
<td>• MCQ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• SAQ</td>
<td>• SAQ</td>
<td>• SAQ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Translation (English to Chinese)</td>
<td>• Translation (English to Chinese)</td>
<td></td>
</tr>
<tr>
<td><strong>Vocabulary &amp; structure</strong></td>
<td>• MCQ</td>
<td>• MCQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Writing</strong></td>
<td>• Essay writing</td>
<td>• Essay writing</td>
<td>• Essay writing</td>
<td>• Essay writing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Translation (Chinese to English)</td>
</tr>
<tr>
<td><strong>Integrated tasks</strong></td>
<td>• Cloze</td>
<td>• Cloze</td>
<td>• Cloze</td>
<td>• Cloze</td>
</tr>
<tr>
<td></td>
<td>• Error identification and correction</td>
<td>• Error identification and correction</td>
<td>• Error identification and correction</td>
<td></td>
</tr>
<tr>
<td><strong>Speaking (optional)</strong></td>
<td></td>
<td></td>
<td></td>
<td>• Q&amp;A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Individual presentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Group discussion</td>
</tr>
</tbody>
</table>

*Jin 2007 in Cheng et al, forthcoming*
3. The way forward

From academic endeavour to social ethics: the NCETC’s new mission statement

... it has already been established that simply changing test contents and methods will NOT necessarily bring about direct and desirable changes in education as intended through a testing change.

Alderson 2004
CET: The way forward

► Intrinsic power of a test
► Extrinsic power of a test

Li 2000

Strengthening the intrinsic power of the test

Taking and sharing the responsibility for the consequences of the test
Strengthening the intrinsic power

Why?
► Eliminate potential sources of negative washback
  ▪ construct under-representation
  ▪ construct-irrelevant variance
  (Messick 1996:252)

How?
► Maintain the highest possible quality of the test
► Develop a common level framework for Chinese EFL learners
► Explore the possibility of a criterion-referenced CET
► Develop Computer- or Internet-based CET
► …
Dilemmas facing test designers

... negative effects can result from a properly developed, correctly targeted test that is being implemented in line with all available knowledge about best practice ...

Hamp-Lyons 1997:297

... It should not be assumed that a ‘good’ test will automatically produce good effects in the classroom, or that a ‘bad’ test will necessarily produce negative ones.

Dianne Wall 2000:505-506
the more likely the test is to be misused

the higher the stakes

the more power the test has

the higher its reputation

the better the quality of a test
There are limits to what a test developer can achieve ...

J C Alderson in Cheng et al. 2004: foreword
Is the test developer responsible for uses and misuses of tests?

What is the role of the tester once s/he notices misuses?

Or, perhaps it is not the responsibility of the tester at all to worry about the test takers after a test had been handed to the users.

E Shohamy 2001: 115
What is washback?

1993 J.C. Alderson and Dianne Wall

**Does washback exist?**

2004 J.C. Alderson in Cheng et al (Foreword)

**What does washback look like?**

**What brings washback about?**

**Why does washback exist?**
What is washback?

Messick 1989, 1996 → Consequential validity

Bachman and Palmer 1996
test usefulness framework → Impact

McNamara 2000 → Impact and washback

c McNunnan 2004

t est fairness framework → Social consequences
There was a growing awareness of the importance of factors other than test design in determining whether tests would have the impact that was desired.

Dianne Wall 2000: 499
Tests, therefore, are not neutral but rather embedded in political, social, educational, ideological and economic contexts.

...the powerful uses of tests in society as tools for mostly political, social, educational and ideological agendas

E Shohamy 2001: 113-114
Taking and sharing the responsibility

1. Investigate the social consequences
   - Identify the **CET stake-holders/factors involved**
   - Investigate **the washback on teaching/learning** and the impact on society (In which direction? In what way?)

2. Make other users aware
   - Disseminate relevant information
   - Introduce better practices (e.g. performance assessment)

3. Share responsibilities
   - Engage all the stake-holders in the mission: Every stake-holder has a role to play!
Code of Ethics for ILTA

1. Respect for test takers
2. Confidence
3. Research
4. No professional misuse
5. Continue to develop professionally
6. Uphold profession
7. Improve professional quality
8. Obligations to own society
9. Consequential effects

“Language testers shall regularly consider the potential effects, both short and long term, on all stakeholders of their projects, reserving the right to withhold their professional services on the grounds of conscience.”
The future of the CET

... lies in its role as an impetus to English language teaching and learning at the tertiary level in China. The job of the CET Committee is not simply of an academic nature ...

The widening of the lens through which we look at tests brings with it great ethical or moral problems about the responsibility of the language tester, and of language testing as a profession.

Hamp-Lyons 2000: 587
Thank you for your attention!
yjinyin@sjtu.edu.cn
Levels of examinations in the Imperial system
Penalties for cheating

代笔者称“枪手”，考试时替人代笔称“枪替”，图为枪手被查出后
于贡院大门前示众同罪。
CET-4 and CET-6

Semester 1, Semester 2, Semester 3, Semester 4, Semester 5, Semester 6

College English Course:
- Band 1
- Band 2
- Band 3
- Band 4
- Band 5
- Band 6

College English Test:
- CET-4
- CET-6
CET stakeholders: Who are they?

- College English teachers
- University students

At the micro levels:
- teaching, learning,
- curriculum and material,
- test preparation and
- attitude, etc.

- Parents
- The general public
- Employers
- School administrators
- Educational policy makers

At the macro levels:
- educational context,
- social/political/economic
- factors
Questions to be addressed

EFL teaching

- In what way and to what extent is college English teaching affected by the reform of the CET?

EFL learning

- In what way and to what extent is college English learning affected by the reform of the CET?

Educational policies

- In what way and to what extent are educational policies affected by the reform of the CET?

Other users

- In what way and to what extent are other users of the test affected by the reform of the CET?