

# 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)

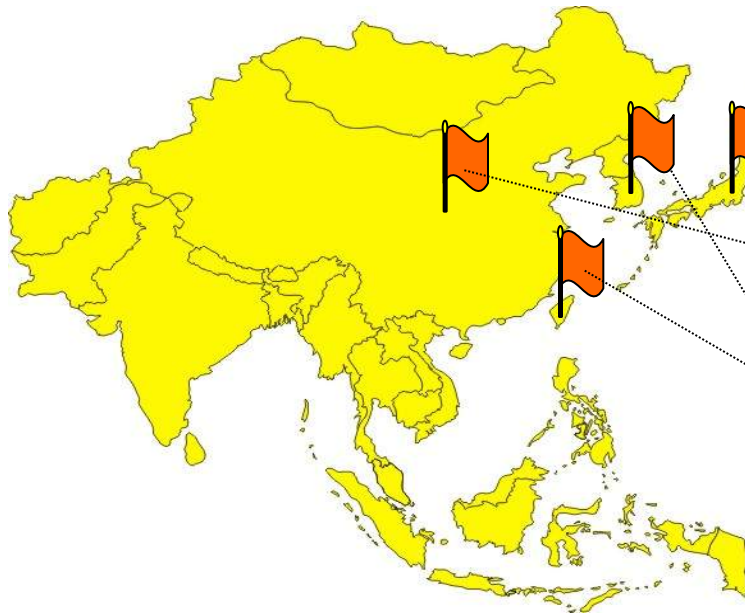
- No such professional standards have been specifically developed in Asia.
- However, Ensuring test quality and fairness is a highly regarded topic in Asia, too.

- 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

English as a Foreign Language

# 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)

Locally  
appropriate

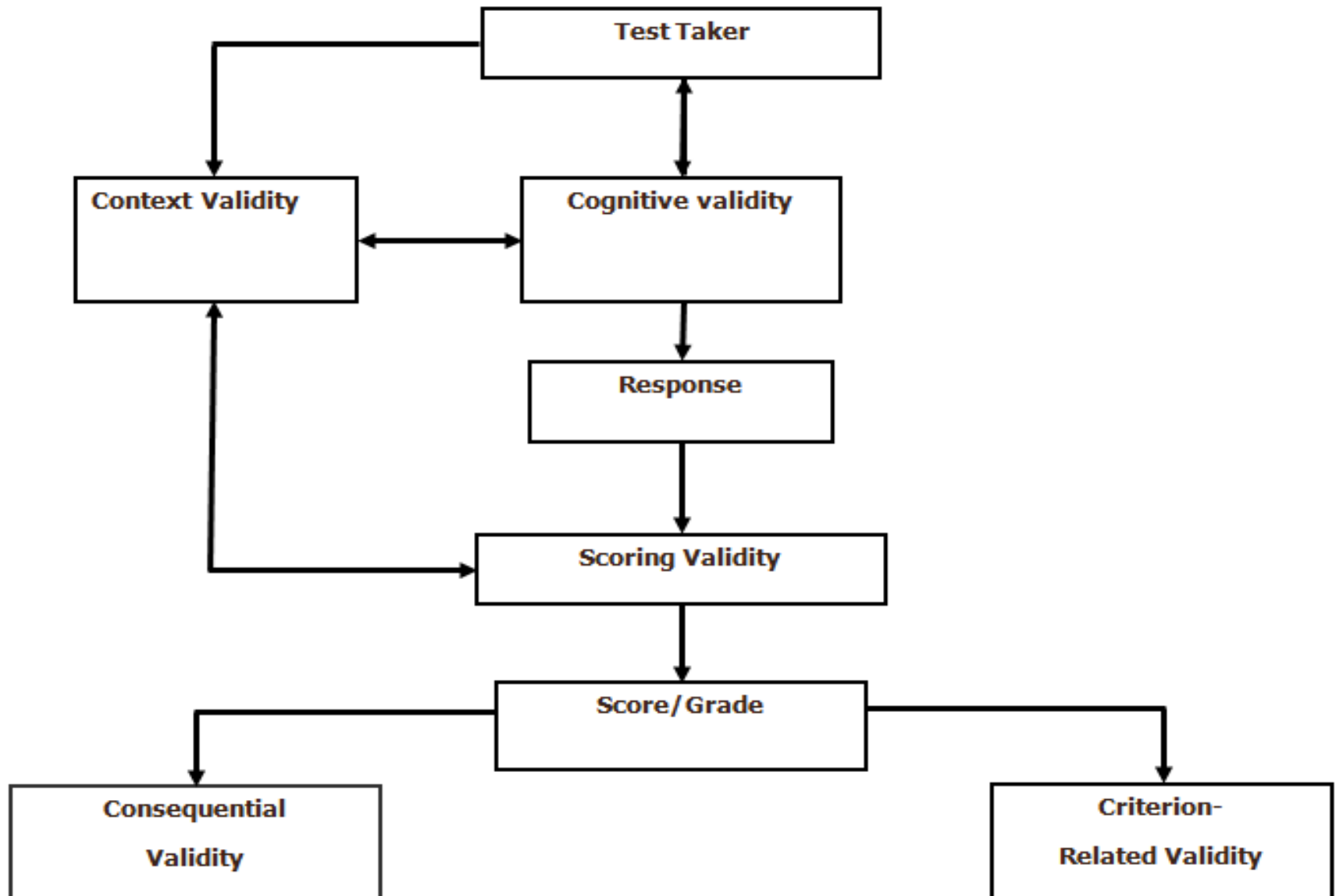


Universal

# 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

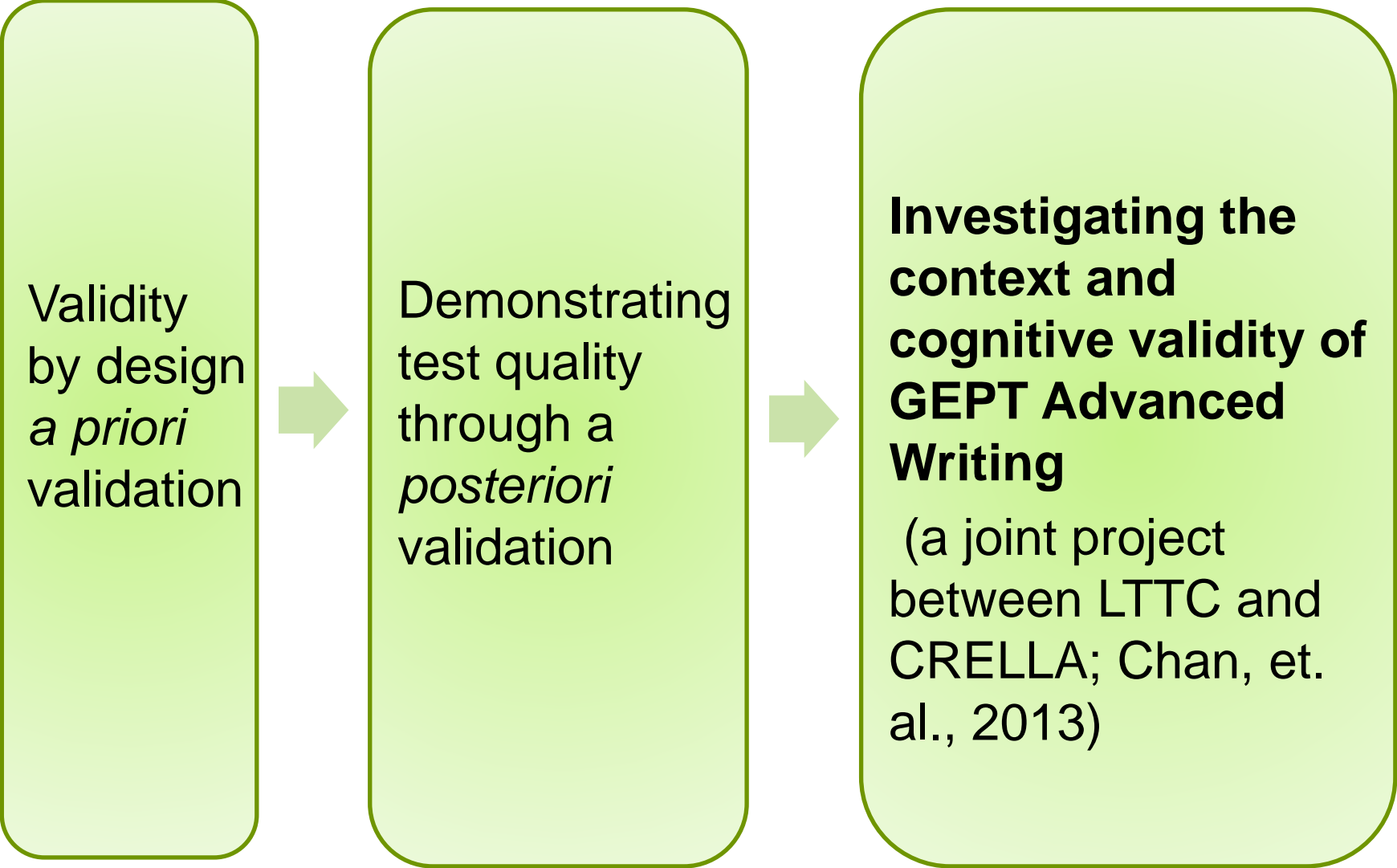
# Case 2: Construct of the GEPT Advanced Reading and Writing

“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

Paper	Part	Task Types	No of items	Time (mins)
Reading	1	Careful reading	40	50
	2	Skimming & scanning		20
Writing	1	Summarizing main ideas from verbal input and expressing opinions (250 words)	60	60
	2	Summarizing main ideas from non-verbal input and providing solutions (250 words)		45

Validity  
by design  
*a priori*  
validation



```
graph LR; A[Validity by design  
a priori  
validation] --> B[Demonstrating  
test quality  
through a  
posteriori  
validation]; B --> C[Investigating the  
context and  
cognitive validity of  
GEPT Advanced  
Writing  
(a joint project  
between LTTC and  
CRELLA; Chan, et.  
al., 2013)];
```

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.

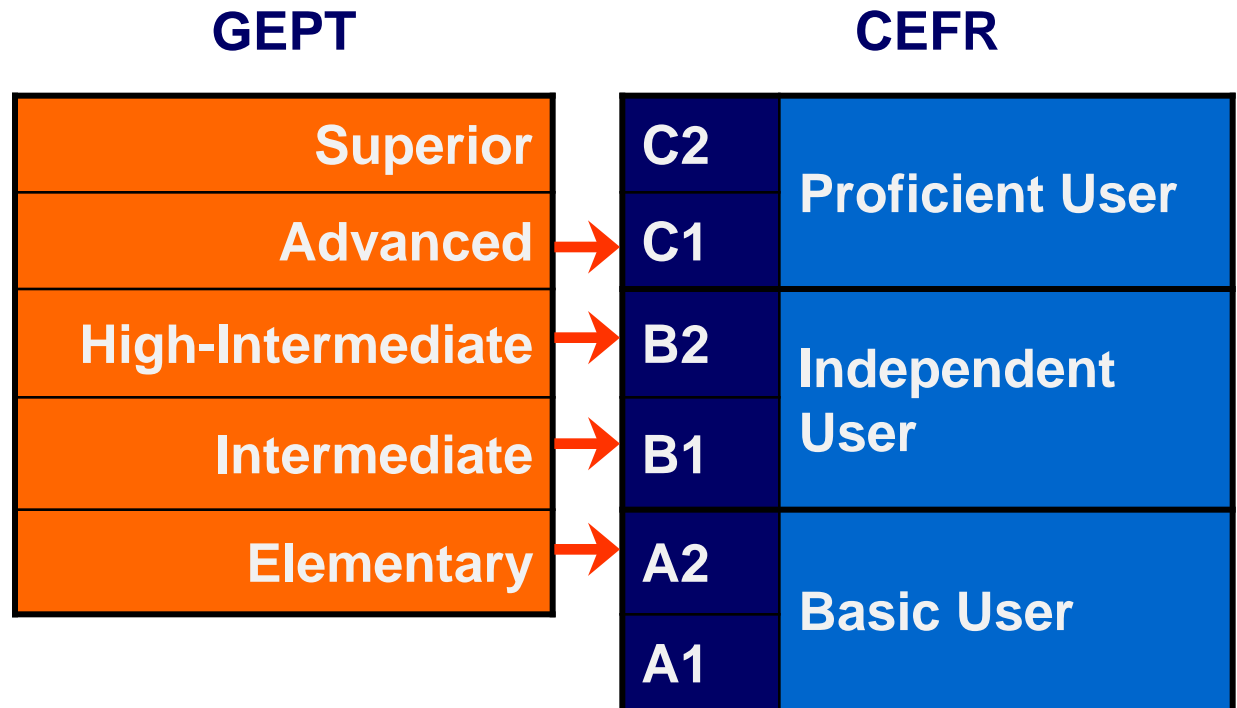
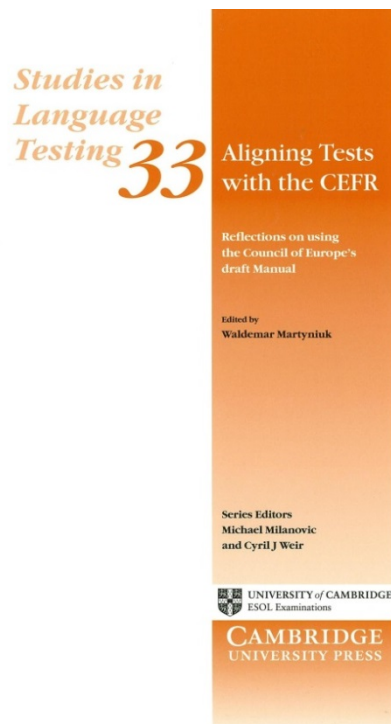


GEPT-CEFR linking studies

# Mapping GEPT with CEFR as a validity criterion (Wu & Wu, 2010)

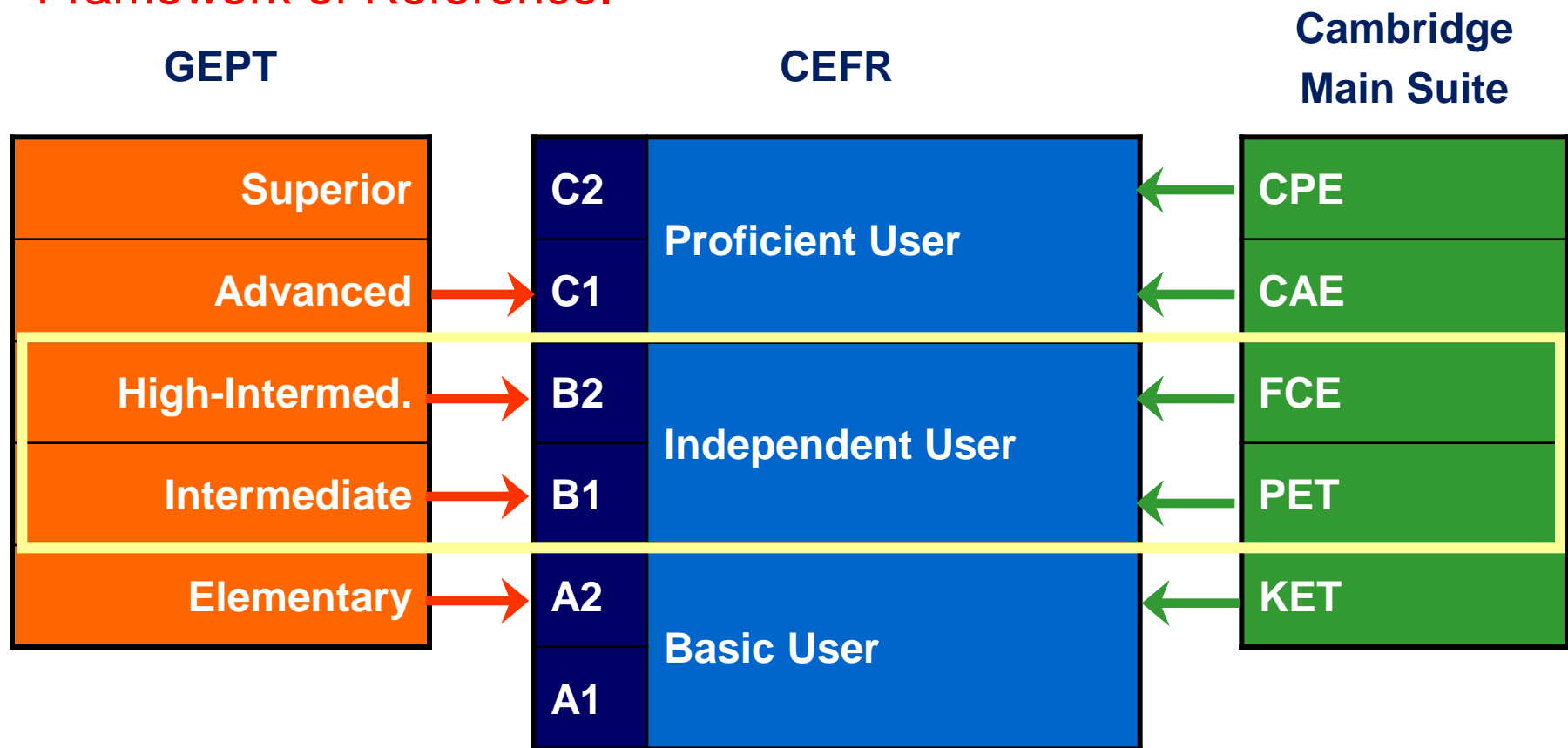
## External evidence

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



# Further studies

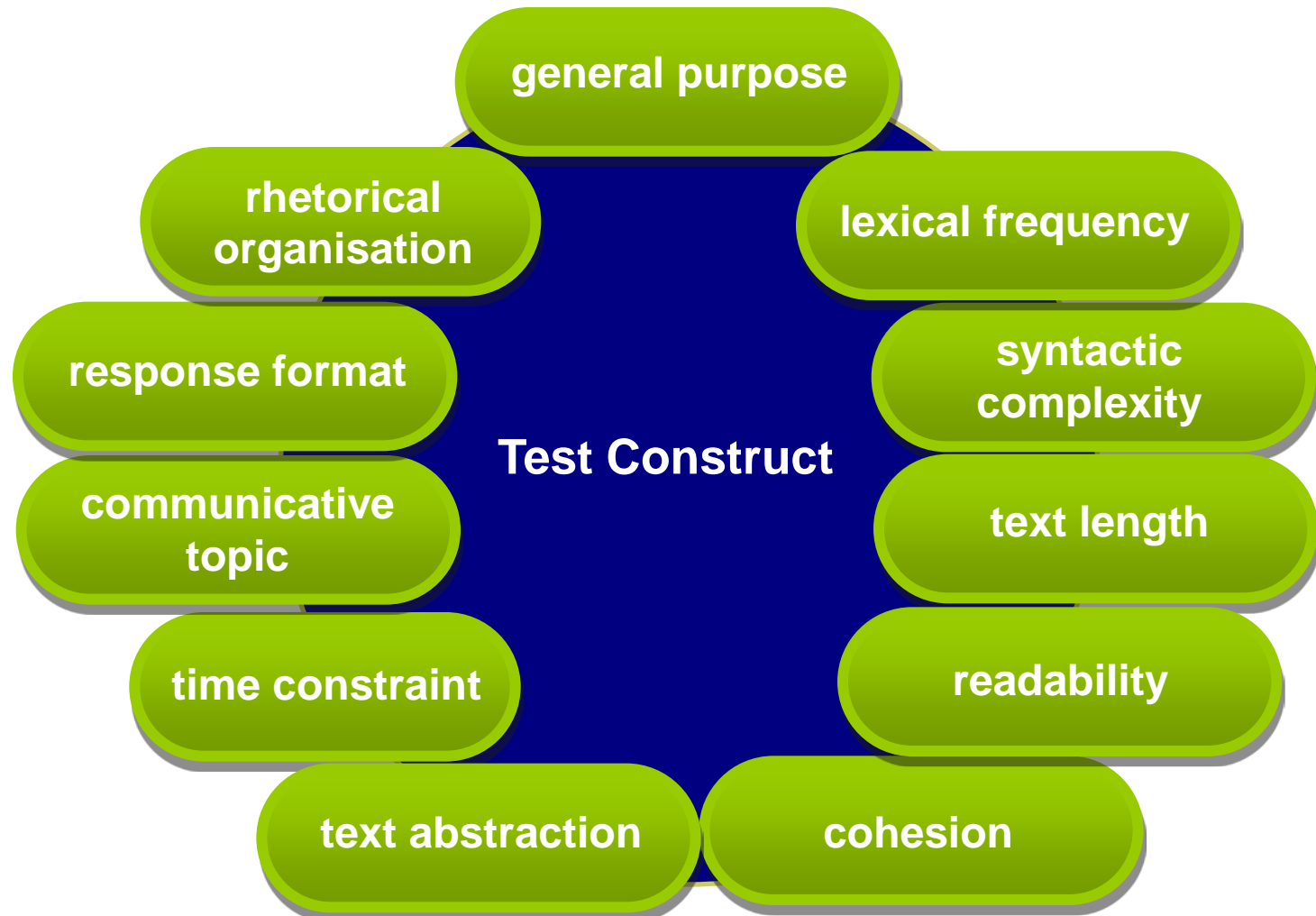
Wu, R. Y. F. (2011). Establishing the Validity of the General English Proficiency Test Reading Component through a Critical Evaluation on Alignment with the Common European Framework of Reference.



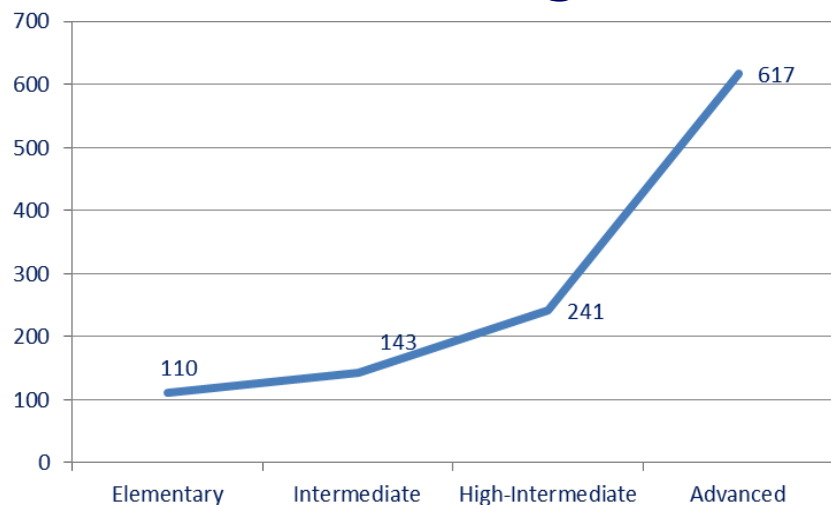


**Comparing different levels of GEPT  
Reading in terms of contextual  
parameters and cognitive processing  
skills by automated textual analysis  
(VocabProfile, Coh-metrix) and expert  
judgment**

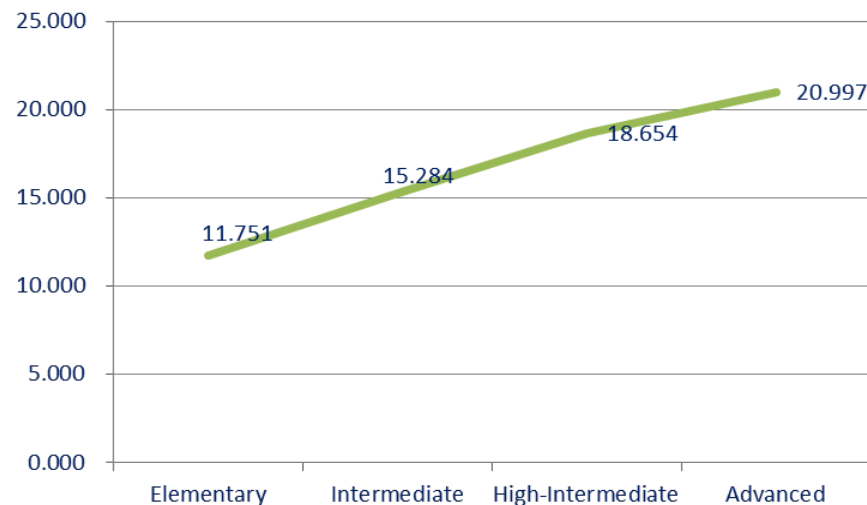
# Contextual parameters



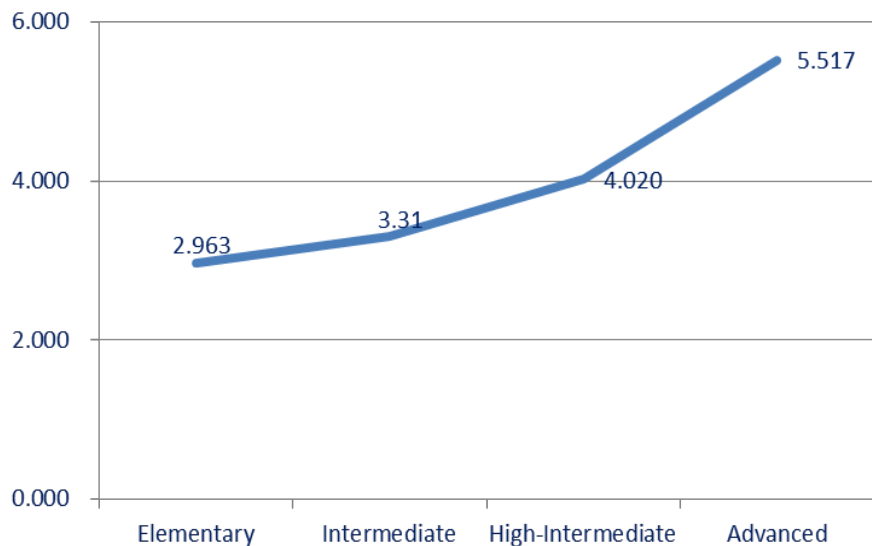
## Text length



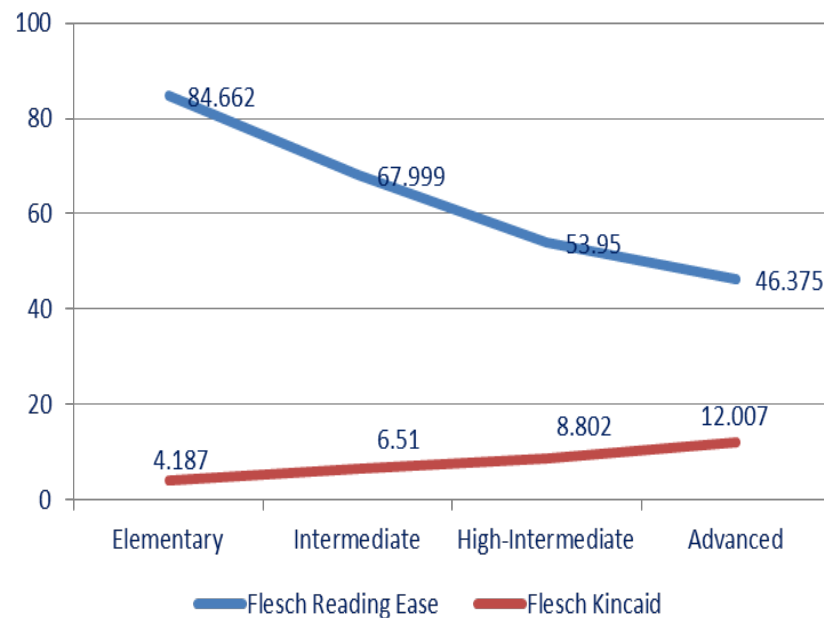
## Sentence length



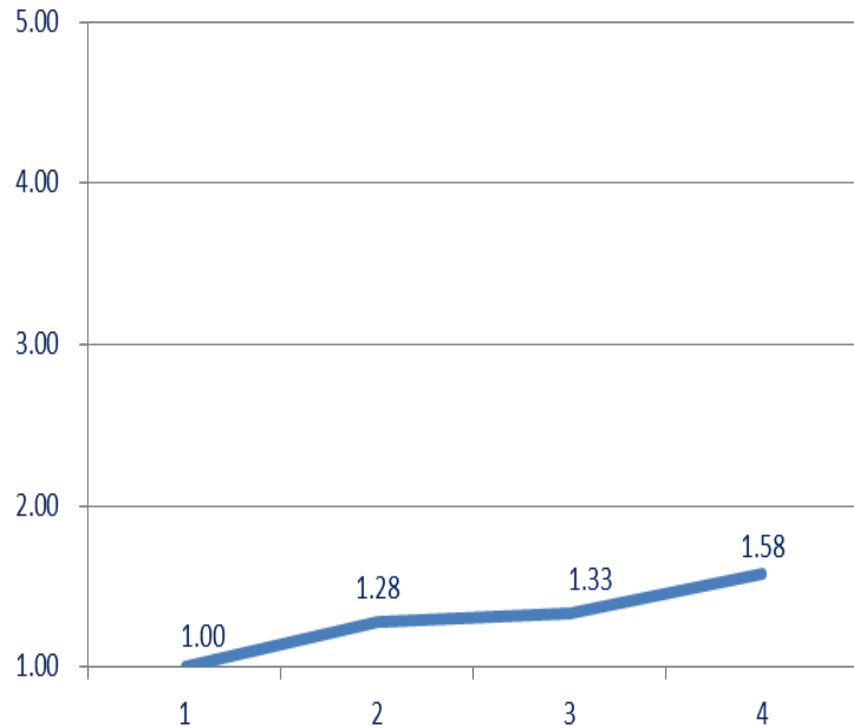
## Words before main verbs



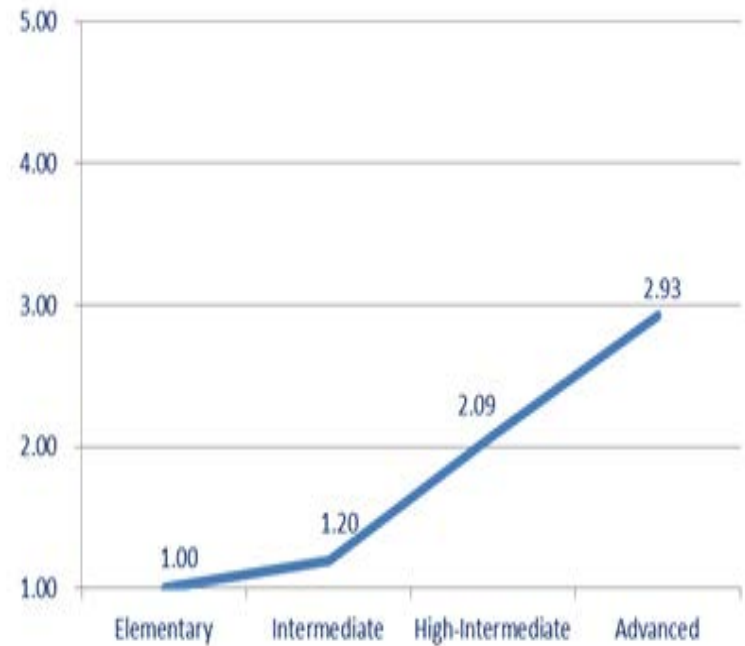
## Readability



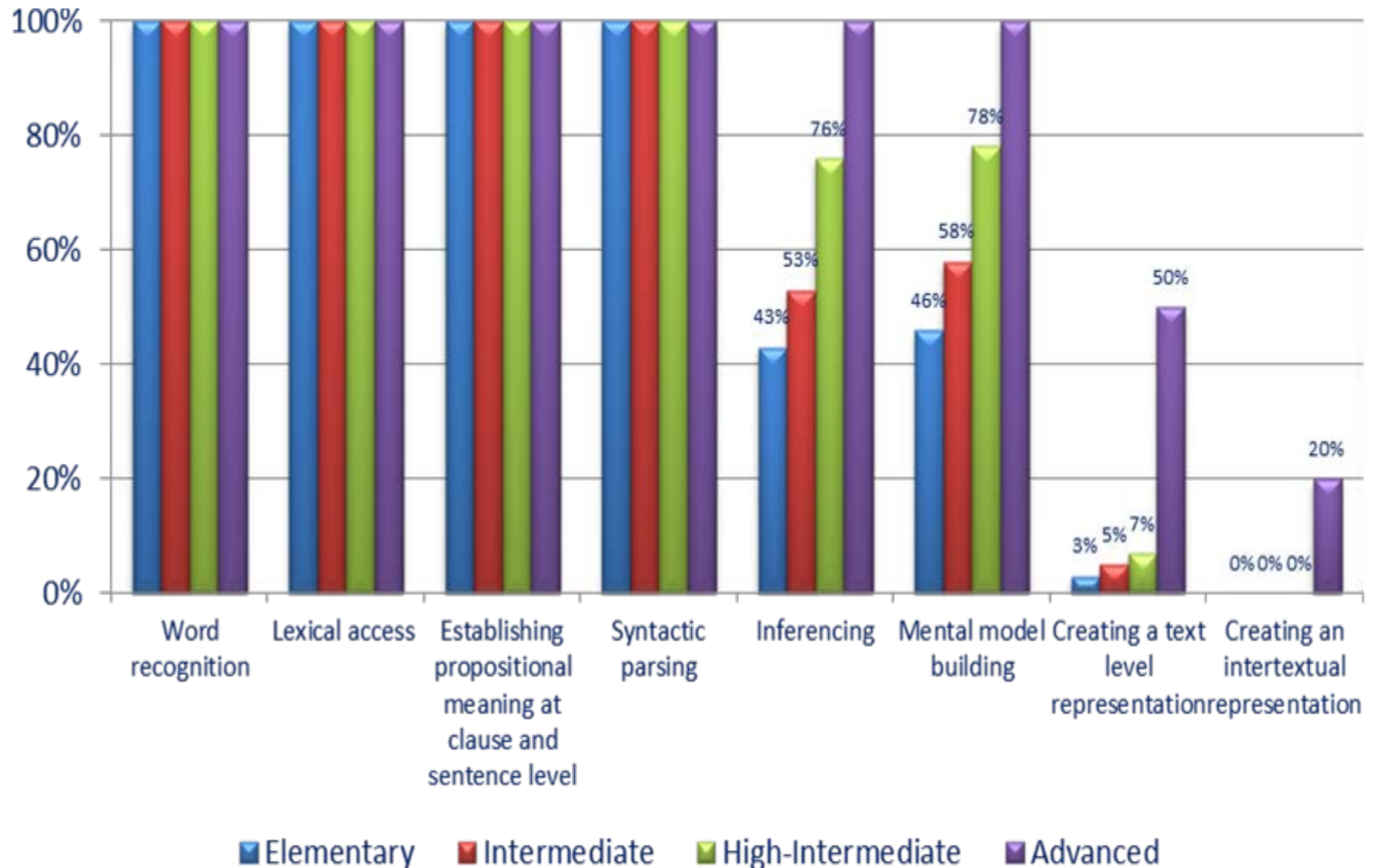
# Cultural knowledge



# Subject knowledge topic familiarity



# Cognitive operations across GEPT levels



# 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	Examining the Cognitive Validity of GEPT High-Intermediate and Advanced Reading: an Eye Tracking and Stimulated Recall Study
University of Melbourne Australia	Linking the GEPT Writing Sub-test to the Common European Framework of Reference (CEFR)
California State University USA	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 (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 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

聯合新聞網 意見評論

2013/5/7 星期二

熱門關鍵字: udn on LINE | 珍愛媽咪 | 折扣季 | 大贏盤

udn / 意見評論 / 民意論壇

最新 | 發燒 | 哇新聞

請選擇... <民意論壇> 相關新聞

## 補出的成績 / 多益排名 意義何在

【聯合報／吳若愚／語言訓練測驗中心研發長（台北市）】 2010.11.11 02:05 am

ETS多益台灣區代表十日發布「2009年台灣與國際產學英語能力差距報告」指出，2009年台灣考生多益平均較前一年只進步6分；台灣多益成績在亞洲排名第6，並以兩岸三地的多益成績（台灣、香港129分，但大幅落後中國大陸171分）作為英語力的比較。本人認為，只憑藉英語測驗成績來探討不同地區人民的英語能力或競爭力，似過於簡化。

日前在一場討論不同國家托福成績比較意義的論壇上，許多語言測驗學者認為，類似的分數比較忽略各國學習者、英語教育制度、社會環境等因素，且用一標準化測驗的成績來評比不同國家英語力、國際競爭力過於武斷。

可見在了解台灣民眾的英語能力時，僅靠英語測驗的成績是不足的。以多益成績為例，需要先瞭解並比較不同地區的英語學習者參加多益測驗的目的、態度。

在這次多益台灣區代表發布的報告中，台灣也明顯落後南韓73分，但韓人自己卻質疑多益高分的真正意義。

韓國學者指出韓人每年有超過二百萬人次報考多益，日、韓兩國多益考生佔全球多益總數量的九成。多益自1979年引進南韓，原為評量職場人士英語溝通能力，但近年來因韓國人偏好國際測驗而擴大其原測驗目的，例如學校入學與畢業門檻。因此，除職場人士外，學生族群也報考此測驗。

這位學者也指出，學生為拿高分，紛紛參加補習學校的考前準備班，且重考率頗高。另一位學者更指出南韓多益考生中，近四成密集重考達四次以上。考前惡補的確使得南韓學生多益表現不差，但南韓學者及媒體紛紛質疑多益高分者是否真的具有英語能力。為降低對國際測驗的過度依賴與減緩測驗的不良影響，南韓政府已自行研發英語能力測驗，預計2012年正式施測。

政府多年來積極設法提升台灣民眾的英語能力，當又聽到台灣的英語測驗表現落後他國的訊息，無疑令人感到挫敗。但當我們理性冷靜地思考不同地區多益使用情形、考生人數、考生群落、學習背景、考試動機、考前準備等相異之處，應該可以用較平常心來看待。

美國國家評量理事會暨教育協會呼籲：「要能辨識不當的評量方法與評量資料的誤用」。長久以來，台灣社會存在「考試領導教學」的觀念，近年在政府提升國人英語能力的政策下，英語標準化測驗更被視為推動的工具，測驗成績即被用作招募、甄選、評鑑之條件。英語測驗成績意義已有被擴大解釋，甚至有可能被誤用的情形，值得關注。

※延伸閱讀>

• 進步6分！台灣多益成績 進步緩慢

瀏覽人數1760 | 字級: A A A A

多益 大陸第10我37

【聯合報／記者沈育如／台北報導】

2013/10/20

主辦多益考試的ETS，昨天公布去年全球考生成績，台灣在全世界四十六個國家中，排名第卅七名，略優於日本的第卅九名，但遠落後中國的第十名及南韓的第十五名。

二〇一二年亞洲地區多益的閱讀平均分數為兩百五十五，聽力為三〇九分，總分是五百六十四；台灣的閱讀平均分數為兩百四十四，聽力為兩百九十五，總分是五百卅九，都在平均分數以下。

台灣近三年的多益平均成績年年下降，去年廿六萬四千多人次報考，學生占六成五最多，高中生的分數比大學生還好。

ETS台灣區代表忠欣公司總經理王星威分析，很多知名大學會要求申請入學的高中生提供多益等英檢成績，這些高中生，原本就屬英文程度較好的那一群，分數自然拉高。

TOEIC 2012年 全球多益 測驗排行榜		
國家	總分	排名
孟加拉	899	1
斯里蘭卡	893	2
尼泊爾	879	3
中國	747	10
南韓	628	25
台灣	539	37
香港	513	38
日本	512	39

註／多益總分990  
資料來源／ETS

製表／沈育如

聯合報

圖／聯合報提供

f 分享

超低價!! 我的媽媽咪啊  
母親節特惠檔期

訂閱電子報

請輸入您的e-mail

☒ 聯合電子報

馬不  
上凡  
觀結  
局完  
整由  
版你

做出  
你的  
不凡

# Test Fairness (News Report)

*The China Post*

[www.ChinaPost.com.tw](http://www.ChinaPost.com.tw)

## Exam classrooms to get air conditioning: MOE

Wednesday, February 16, 2011  
The China Post news staff

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 will be provided in exam venues, cooling heated minds at a time when temperatures are rising.

The temperature is uniformly rising across the country, covering the budget for both air conditioning and heating at NT\$30 million. Exam registration ministry added.

**Should air-conditioning be provided in a testing venue?**

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 humidity 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.

