

## 華語文聽讀能力測驗改版歷程

林玲英 藍珮君

### 一、改版緣由

華語文能力測驗（前稱為 Test Of Proficiency-Huayu，現英文名稱更名為 Test of Chinese as a Foreign Language, 簡稱 TOCFL）自 2001 年開始研發，2003 年正式施測以來，考生人數逐年增加；2006 年正式推上國際，海外考點從 5 個國家，擴展到目前的 21 國 39 個城市。

為使本測驗能夠與國際語言標準接軌，提供考生能夠衡量其語言能力的國際評量工具，華測會於 2008 年積極著手研擬「新版華語文能力測驗」，並於 2011 年 5 月正式推出。

### 二、改版方向

改版範圍僅限於原初等、中等和高等三等測驗，主要的改版方向是參考歐洲理事會所研發的「歐洲共同語文參考架構」(The Common European Framework of Reference for Languages，以下簡稱 CEFR)，微調測驗內容以符合新版測驗的能力指標。

測驗改版要點如下：

- 一、新版測驗將刪除部份舊題型，包括單句聽力及單句閱讀。
- 二、在聽力部分，增加了雙輪對話題及多輪對話題，部份段落聽力試題之內容長度增加至 550 字左右，並將題組的子題數由 2 題增加到 4 題不等。
- 三、將原本的詞彙語法題，改成克漏字形式，以測試考生於篇章之中的詞彙語法運用能力。
- 四、初等測驗增加了短文閱讀題型，字數大約介於 100 至 200 字之間。中等及高等測驗之短文閱讀題型則增加了部份篇幅較長的文章，介於 300 至 500 字之間。
- 五、基礎測驗維持 80 題，初中高三等測驗題數皆為 100 題。

### 三、新舊版本比較

基礎測驗將維持原有題型與題數，而初等、中等和高等測驗之詳細改版情形，請見表一至表三。

表一 初等（進階級）測驗改版方式

	聽力理解測驗				閱讀理解測驗					合計 題數
	單句	單輪對話	雙輪對話	段落	詞彙	語法	單句	材料	短文	
舊版	20	20	—	10	20	20	10	20	—	120
新版	—	20	15	15	20		—	15	15	100

表二 中等（高階級）測驗改版方式

	聽力理解測驗					閱讀理解測驗					合計 題數
	單句	單輪對話	雙輪對話	多輪對話	段落	詞彙	語法	單句	材料	短文	
舊版	15	20	—	—	15	10	20	10	10	20	120
新版	—	10	10	15	15	15		—	10	25	100

表三 高等（流利級）測驗改版方式

	聽力理解測驗					閱讀理解測驗				合計 題數
	單句	單輪對話	雙輪對話	多輪對話	段落	詞彙	語法	單句	短文	
舊版	15	20	—	—	15	20	10	10	30	120
新版	—	—	10	20	20	15		—	35	100

### 四、新版等級名稱

為配合測驗改版，以及寫作測驗與口語測驗的推出，華語文能力測驗各等名稱也隨之改變。初等測驗更名為「進階級」，相當於 CEFR 的 B1 程度；中等更名為「高階級」，相當於 B2 程度；高等更名為「流利級」，相當於 C1 程度。詳如下表：

表四 華語文能力測驗等級名稱

華 語 文 能 力 測 驗	舊版	新版	對應 CEFR
	基礎	基礎級	A2
	初等	進階級	B1
	中等	高階級	B2
	高等	流利級	C1

## 五、新版等級能力描述

表五 華語文能力測驗能力描述

進階級 (B1)：著重在日常生活的一般簡易溝通能力	
聽力	閱讀
當談話內容為與工作、學習、娛樂相關的熟悉話題，且講話人口齒清晰、語音標準時，能了解內容大意和重要細節。	能讀懂個人感興趣的主題或與專攻領域相關的文章；前提是文章以淺白、平鋪直敘的方式寫作而成。
高階級 (B2)：著重在語言段落的理解分析能力	
聽力	閱讀
對於具有一定篇幅且以標準語表達的話語，包括專攻領域的技術性討論，不論內容抽象與否，都能聽懂要點大意。	閱讀具有相當大的自主性，懂得為了不同目的，採用不同方法和速度閱讀不同的文章，並能選擇適合使用的參考書。具備廣泛且可隨時提取的閱讀詞彙，但對於不常見的慣用語，可能有理解上的困難。
流利級 (C1)：著重在語言使用的廣度與精熟度	
聽力	閱讀
能聽懂各種抽象或複雜主題的話語內容，即使話語結構或關聯性可能不夠清楚、明確；但在不熟悉說話人口音的情況下，可能需要特別確認部分細節。	在有機會重新閱讀困難部分的情況下，不論主題是否與個人專攻領域相關，都能讀懂長篇複雜文本的各項細節。

## 六、紙筆與電腦難度比較研究

因應未來華語文能力測驗實施電腦測驗時，部分區域(如：海外地區)受限於場地與設備，僅能採用紙筆測驗進行施測，為確保考生權益及公平性，需確認電腦測驗與紙筆測驗試題難度參數可共用，故進行本研究。

本研究採用共同組法(single group design)以及對抗平衡法(counterbalanced design)，於考生報名預試時，限制每位考生只能報名一個等級，並告知為進行研究，必須參加上午、下午兩場考試(一次為電腦測驗、一次為紙筆測驗)，兩場測驗皆完成者，可免費參加一次正式考試。考生報名結束後，隨機分成兩組，一組

早上進行電腦測驗，下午進行紙筆測驗；另一組則相反。考生事先並不知道兩次考試的題目完全相同，以防止考生刻意記下考題內容，影響實驗結果。

收集考生電腦及紙筆測驗作答反應後，以 IRT 軟體 Winsteps，採同時估計法，先將電腦與紙筆測驗試題視為不同題目，估計試題難度參數，聽力與閱讀測驗分開估計。估計完難度參數後，比較同一道試題在電腦與紙筆測驗的難度差異，若難度相差超過 0.5 logit (logit 為難度參數單位)，則視為差異較大之試題。為了解同樣的試題為何在不同的施測介面下，試題難度參數會有所差異，將進一步邀請專家學者共同討論可能的原因。

基礎級測驗 40 道聽力題中，有 7 道難度參數相差大於 0.5 logit，但其中有 5 道試題通過率大於 0.9，經專家學者建議，通過率較極端之試題其難度參數估計誤差較大，因此可予以忽略，不需討論。依此原則，基礎級測驗只有 2 道聽力題難度參數差異較大，閱讀題的差異則皆在容許範圍內，因此在兩種考試介面下，基礎級測驗全測驗有 97.5% 的試題難度相差不大。

在進階級測驗方面，50 題聽力題中，有 11 題難度參數相差大於 0.5 logit，但有 9 題通過率偏高；50 題閱讀題中，有 5 題難度相差大於 0.5 logit，其中有 2 道題目通過率偏高。因此，進階級測驗全測驗有 95% 的試題難度相差不大。

高階級測驗 50 題聽力題中，有 8 題難度參數相差大於 0.5 logit，其中有 6 題通過率偏高；50 題閱讀題中，有 4 題難度相差大於 0.5 logit，其中 1 題通過率偏高。因此，高階級測驗全測驗有 95% 的試題難度相差不大。

最後，流利級測驗 50 題聽力題中，有 3 題難度參數相差大於 0.5 logit，其中有 1 題通過率偏高；50 題閱讀題的難度差異則皆在容許範圍內。因此，流利級測驗全測驗有 98% 的試題難度相差不大。

經向專家學者諮詢後，造成聽力試題難度差異較大的原因應為聽力設備的不同。電腦測驗的考生皆須戴上耳機應試，而紙筆測驗試場則以公開廣播的形式播放聽力考題，因此造成相同的試題在電腦測驗時顯得較為容易。

而造成閱讀測驗難度差異大的原因則為兩種測驗形式的外觀不同。紙筆測驗考生可在題本上看到完整的試題，而電腦測驗則受限於螢幕畫面的大小，無法完整呈現題幹，考生須以拖曳的方式閱讀隱於下方的文本。因此造成相同的試題在紙筆測驗時顯得較為容易。

## 七、新版測驗門檻訂定流程

由於改版之後試題數量由 120 題減為 100 題，過去的通過門檻勢必無法適用於新版測驗。為設定新版測驗之通過門檻，本會自 2010 年 3 月起進行為期二個

月的 Angoff 法標準設定，由於歐洲理事會語言政策部門於 2003 年發表之「測驗與 CEFR 連結手冊」中建議接收能力(receptive skill)的判斷先由閱讀能力開始，故本研究先進行閱讀測驗，再進行聽力測驗。研究程序分為以下三步驟，步驟一為熟悉 CEFR，步驟二為測驗介紹與訂定最低能力者能力描述，步驟三為 Angoff 標準設定。

研究者每次給予小組成員 50 道試題，於小組成員實際作答後，提供標準答案，小組成員再獨自預估每一題最低能力者答對的可能性，完成後繳交給研究者。研究者整理所有與會人員每一題的評估結果後，召開會議，提供小組成員每一題答對可能性的評估資料，以及每一題的試題參數。小組成員針對第一輪答對可能性評估結果歧異較大，或其他成員認為需要討論的試題進行討論，然後再進行第二輪的評估。研究者整理小組成員第二輪的評估結果，求出每一位成員判斷此 50 題答對率的平均，待聽力理解分測驗及閱讀理解分測驗皆完成此步驟，再與小組成員進行微調，即可得出該等級測驗的通過門檻。

每一個等級都要分聽力及閱讀測驗兩部分進行同樣的步驟：評估-討論-評估-結果統整，大約歷時 2 個月的時間，最後訂定新版測驗通過門檻，如表六。其中聽力理解及閱讀理解分測驗門檻之訂定依據，是將歷次預試之考生依成績表現高低依序排列，觀察表現較佳之前 80% 考生之聽力及閱讀成績，依此訂定。

表六 華語文能力測驗新舊版成績對照表

測驗類別	等級分數範圍		
	聽力理解	閱讀理解	總分
進階級	25-50	25-50	66-100
高階級	25-50	24-50	64-100
流利級	25-50	23-50	62-100

## 八、未來發展方向

上述三項因應改版的而進行研究雖有初步成果，但仍須進一步驗證其可靠性，因此未來將透過施測收集更多資料，以進行後續研究。以新舊版分數對照為例，目前的研究成果來自於新版預試及舊版正式考試所提供的資料，由於預試試卷並非按照一固定的難度比例組成，試卷難度較不易掌握。由於新版正式考試已開始實施，未來將比對新版正式考試與舊版正式考試的考生表現，再微調新舊版分數對照表。

此外，新版華語文能力測驗未來將朝向電腦適性化測驗 (Computerized

adaptive testing, CAT)發展, CAT 乃根據試題反應理論(item response theory, IRT)而來, 其特色為試題的呈現順序與應試考生其自身的能力水準息息相關, 而非一成不變, 若該考生答對目前試題則呈現難度較難的下一題, 反之, 則呈現較簡單試題, 如此便能在較短時間內, 以較少之試題量測得考生的能力。為了發展適性化測驗, 除了必須大幅提高題庫中的試題量, 以及足夠的預試考生量以估計試題難度外, 尚須擬定合理之選題規則, 以確保能以最有效率的方式估計出考生能力, 並盡量降低估計誤差。

目前華語文能力測驗已對應至 CEFR 之 A2、B1、B2、C1 四個等級, 但海外施測的結果顯示, 很大一部份之海外外籍人士華語能力之程度, 大致落在入門(A1)等級或不到 A1 水準, 因此, 本會將著手規劃 CEFR-A1 級別測驗的研發工作, 以切合海外考生之需求。

## 參考文獻

### 中文部分

- 吳宜芳、鄒慧英、林娟如(2010)。標準設定效度驗證之探究—以大型數學學習成就評量為例。《測驗學刊》, 第 57 輯, 第 1 期, 1-27。
- 林宜臻(2010)。TASA2009 國小四、六年級數學領域學習成就標準設定。2010 NAER「永續教育發展-創新與實踐」國際學術研討會, 台北。
- 謝進昌(2006)。精熟標準設定方法的歷史演進與詮釋的新概念。《嘉義大學國民教育研究學報》, 16, 157-193。

### 英文部分

- Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: A comprehensive approach to examining the comparability of testing modes. *Essex Graduate Student Papers in Language & Linguistics*, 10, 22-44.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59-88.
- Choi, I., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320.
- Cizek, G. J. (2006). Standard setting. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp.225-258). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage

Publications.

- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593-602.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment* (chap.1 & chap.4). Retrieved January 17, 2007, from [http://www.coe.int/t/dg4/linguistic/Source/Framework\\_EN.pdf](http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf)
- Council of Europe. (2003, Sep). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF).Manual: A preliminary pilot version*. Retrieved March 1, 2007, from <http://www.coe.int/T/DG4/Portfolio/documents/Manual%20for%20relating%20Language%20Examinations%20to%20the%20CEF.pdf>
- Kim, D., & Huynh, H. (2007). Comparability of computer and paper- and- pencil versions of algebra and biology assessments. *Journal of Technology, Learning and Assessment*, 6(4), 4-29.
- Kingston, N. M. (2009). Comparability of computer-and paper-administered multiple-choice tests for k-12 populations: A synthesis. *Applied Measurement in Education*, 22, 22-37.
- Lee, G., & Weerakoon, P. (2001). The role of computer-aided assessment in health professional education: A comparison of student performance in computer-based and paper-and-pen multiple-choice tests. *Medical Teacher*, 23(2), 152-157.
- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Rep. No. RR-08-34). Princeton, NJ: Educational Testing Service.
- Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standard of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.