

華語文閱讀測驗信度效度分析與垂直等化研究*

藍珮君

陳柏熹

國家華語測驗推動工作委員會

國立臺灣師範大學
教育心理與輔導學系

摘要

本文旨在探討華語文閱讀測驗四個測驗等級：基礎級、進階級、高階級與流利級的信度與效度表現，並將四個等級試題難度連結至同一量尺上。樣本來自 2011 年 5 月與 11 月正式考試，及 2012 年預試之考生作答反應資料，以古典測驗理論與試題反應理論進行分析。研究結果顯示：1. 閱讀測驗信度良好，各等測驗 KR20 信度係數接近或達到 0.90 以上，IRT 估計標準誤換算後的信度數值皆達到 0.90 以上，且各測驗通過門檻的考生能力值亦有較高的測驗訊息量與較低的估計標準誤；2. 閱讀測驗具有建構效度，各等級因素分析結果抽出閱讀理解單一因素，解釋變異量在 66.91% 以上，且各等級試題與模式適配比例達 87.5% 以上；3. 四等測驗試題難度分佈良好；4. 進階與高階級測驗折半合併為一等測驗，通過門檻之測驗訊息量及估計標準誤，與原進階級測驗相當，略差於原高階級測驗，將此兩等級測驗合併為一等測驗在實務上應為可行，惟組卷時試題難度比例需再做調整。

關鍵詞：華語文能力測驗 信度 效度 試題反應理論 垂直等化

1. 緒論

由國家華語測驗推動工作委員會（簡稱華測會）所研發的華語文能力測驗（Test of Chinese as Foreign Language，簡稱 TOCFL）含括四種測驗類別：

* 本研究感謝教育部「國際與兩岸教育司」、「邁向頂尖大學計畫」與「國科會跨國頂尖研究中心計畫」（NSC 103-2911-I-003-301）支持。

聽力測驗、閱讀測驗、口語測驗以及寫作測驗。聽力與閱讀測驗於 2003 年 12 月在臺灣地區舉辦第一次正式考試，從 2006 年開始，華測會也陸續在海外各地舉辦聽力與閱讀測驗的正式考試和預試。2012 年已擴大至 26 國 40 個地區進行施測，截至 2013 年年底，臺灣與海外地區累積報考 TOCFL 聽力與閱讀測驗考試的考生人數已突破 15 萬人，詳見華測會官網（http://www.sc-top.org.tw/picture/tocfl_150K.png）。

華語文聽力與閱讀測驗開辦至今將近 10 年的時間，試題品質的控管與測驗信度、效度的結果一直為外界所關切，華測會為了讓華語教學及語言測驗等相關領域的專家學者了解華語文能力測驗研發現況，本研究將呈現華語文能力測驗信度與效度分析結果，盼能引發討論，藉以讓華語文能力測驗的發展更臻完善。惟華語文聽力與閱讀測驗雖同屬語言能力的接收能力，但本質上仍為不同的語言能力，測驗相關分析皆分別進行，聽力與閱讀測驗也各有各的試題難度量尺，故本研究先探討並報告閱讀測驗部分的信度與效度。

國內現行之華語文閱讀測驗為電腦化測驗（Computer based test，簡稱 CBT），未來目標為進一步發展至電腦適性化測驗（Computerized adaptive testing，簡稱 CAT）。相較於 CBT，CAT 可縮短測驗長度，有效節省考生測驗時間，且仍可提供穩定可靠的測驗分數。然而要發展 CAT，首要工作便是建立起橫跨四個等級的大題庫，目前進階、高階與流利測驗試題難度參數已連結完成，基礎級測驗參數則尚未併入此一難度量尺，故需進行垂直等化研究，將基礎級測驗的試題參數，串連至現有的進階、高階與流利級測驗難度參數量尺上，並檢視各等級測驗難度分佈是否恰當。

再者，目前華語文能力測驗分為基礎、進階、高階以及流利級四個等級，各等級測驗皆設置通過門檻。試務工作上，經常遇到考生反應不知該報考哪一測驗等級，有些考生一方面為了確保可獲得證書，另一方面又想探求自身能力上限，往往同時報考兩個等級，既花費金錢也耗費時間。華測會因而積極思考調整測驗實施方式的可行性，若合併皆為歐洲語言共同架構（The Common European Framework of Reference for Languages，簡稱 CEFR）獨立使用者（independent user）程度的進階與高階級兩等測驗為一等測驗，是否仍能維持原先測驗的信度水準，以達到精簡測驗等級，提升考試效率的目的。

綜上所述，本研究針對華語文閱讀測驗基礎級、進階級、高階級以及流利級四個測驗等級進行測驗信度、效度及垂直等化的相關研究，欲探討以下四項研究問題：

- (1) 華語文閱讀測驗基礎、進階、高階以及流利級測驗的信度表現。
- (2) 華語文閱讀測驗基礎、進階、高階以及流利級測驗的效度表現。
- (3) 華語文閱讀測驗各等級是否能合併成為大題庫，供未來跨等級比較，進一步發展為電腦適性化測驗。
- (4) 將原本華語文閱讀測驗之進階與高階級測驗分開施測，調整為合併兩等測驗為一等測驗的可行性。

2. 文獻探討

以下將先簡介華語文閱讀測驗架構、內容及各等級門檻分數，接著分別說明古典測驗理論(classical test theory)和試題反應理論(item response theory，簡稱IRT)測驗信度、效度的意涵，以及測驗等化的設計和參數估計方法。

2.1 華語文閱讀測驗

華語文閱讀測驗是專為母語非華語者所研發，為一套標準化的語言能力測驗。在臺灣地區採電腦化測驗進行施測，海外地區目前暫以紙筆測驗方式進行。2011年起華測會在臺灣正式推出「新版華語文能力測驗」，新版測驗有四個等級—基礎級、進階級（舊版初等）、高階級（舊版中等）、流利級（舊版高等），分別對應歐洲語言共同架構（CEFR）之A2、B1、B2及C1。

表1為華語文閱讀測驗各等級雙向細目表，說明各題型測得閱讀能力面向與題數分佈情形。除基礎級外，其它三級測驗變更部分題型，且測驗題數減少，由70題減為50題，測驗時間為60分鐘。基礎級測驗題目共40題，測驗時間亦為40分鐘。各級測驗題目皆為單選題，每題一分；答錯不倒扣。

表 1：華語文閱讀測驗各等級雙向細目表

等級	能力面向 題型	整體性的 閱讀理解	閱讀 書信	導向 閱讀	為資訊論證 而閱讀	語法 能力	合計
基礎級	單句理解	-	-	-	-	10	40
	看圖釋義	-	-	5	5	-	
	選詞填空	-	-	-	-	10	
	完成段落	10	-	-	-	-	
進階級	選詞填空	20	-	-	-	-	50
	材料閱讀	-	5	15	-	-	
	短文閱讀	-	-	-	15	-	
高階級	選詞填空	15	-	-	-	-	50
	材料閱讀	-	4	6	-	-	
	短文閱讀	-	-	-	25	-	
流利級	選詞填空	15	-	-	-	-	50
	短文閱讀	-	-	-	35	-	

四級測驗由於對應於 CEFR，各有其評量重點，基礎級側重在「與個人相關、主題具體下的簡易溝通能力」；進階級著重「在日常生活的一般簡易溝通能力」；高階級和流利級則分別為著重在「語言段落的理解分析能力」以及「語言使用的廣度與精熟度」上，表 2 與圖 1 至圖 4 為各級測驗通過者具備的閱讀理解能力描述以及模擬試題。

表 2：華語文閱讀測驗各等級能力描述

測驗等級	能力描述
基礎級	當文章簡短，且多為日常生活或工作的常用詞彙時，能理解內容。
進階級	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	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	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。



- (A) 李天明想學英文。
- (B) 李天明教別人英文。
- (C) 李天明沒有錢學中文。

圖 1 華語文閱讀測驗基礎級模擬試題

在我的國家，每年都有颱風，颱風有的大，有的小。如果來的是大颱風，對我們的生活就有嚴重的影響。今年八月有個大颱風，下大雨、吹大風，火車沒辦法開，飛機停飛，所有人不上班、不上課，大部分的人不敢出去，只能在家看颱風的新聞。好不容易等到颱風離開，很多屋子卻都進水了，許多樹、房子被吹倒了，連橋也被大水沖壞了。聽說這個週末又有一個颱風要來，真希望它是個小颱風。

22. 這篇文章主要在說什麼？

- (A) 颱風季節都是在八月
- (B) 大颱風對人們生活的影響
- (C) 如果颱風來，要準備的東西
- (D) 如果小颱風來，會發生的事情

圖 2 華語文閱讀測驗進階級模擬試題

有一首歌，歌名叫「煩」，已流行一段時間，其中有二句是「煩到無法呼吸，煩到沒有力氣」，話說回來，人生真的這麼煩嗎？基本上，每個人都有煩惱，可是為什麼我們的腦袋裡裝了那麼多煩惱？我們把事情看得太嚴重了，如果我們都能以平常心去看事情，去接受這個世界的好與壞，我們就會發現煩惱並沒有什麼了不起。

解決煩惱的第一步，就是消除一切消極的想法，把不安、抱怨、不樂觀、不感興趣、後悔、無信心、緊張、灰心等想法全部丟到腦外，不要去想它；第二是積極參加各式各樣外面的活動，例如登山、郊外旅遊、攝影、露營、看球賽、游泳、看電影等，多和別人聊天，向朋友訴說心事，聽聽別人意見；最後則是維持正常的生活時間表，讓生活規律，並且多多找時間運動，保持充分的體力與活力。

有一句話說「天塌下來，都有人撐著」，就是告訴我們不要自己尋找煩惱，不管發生什麼事，都一定有辦法度過，讓自己過得快樂一點兒比較重要。

50. 文章中「天塌下來，都有人撐著」是什麼意思？

- (A) 幫助別人等於幫助自己
- (B) 無法預先知道將發生什麼事
- (C) 隨時都要做好預防災害的準備
- (D) 無論發生任何事情都有方法解決

圖 3 華語文閱讀測驗高階級模擬試題

自兩億多年前的侏羅紀之初，體型巨大就是陸生動物的普遍特點。動物為什麼要長成巨無霸呢？有些理由是顯而易見的：身軀愈大愈不易被獵殺，而且還利於撲殺獵物。例如羚羊很容易淪為獅子、土狼和獵犬的獵物；但成象及犀牛則幾乎不受威脅，牠們的子女也因父母體型巨大而得到保護。對草食動物來說，長得巨大表示「高人一等」，可以吃到更高處的樹葉。長頸鹿及大象就有長到五公尺高的，而大象尚可利用龐大的身軀推倒高樹。

其他重要卻不明顯的理由有：運動消耗的能量會隨體型增大而減少，故一隻五公噸重的大象走一公里路所消耗的能量，比一群總重五公噸的瞪羚走同樣路程要少很多。此外，新陳代謝率也會隨體型增大而降低。這就是為什麼鮑鰻每天都得吞入超過其體重的食物，而大象只需吃進占體重 5% 的食物便足矣。不僅如此，龐大的身體也具有像隔熱板一樣的保溫作用，讓動物不受環境劇烈溫差的影響。

不過，長得巨大也是有害處的。因為大的動物吃得多，其總數自然有限。非洲的大象及犀牛數目，在人類獵捕之前也不過是以百萬計；相對的，齧齒類動物的數量則不知有幾十億。此外，小動物鑽地洞、上樹、飛天等動作，龐然大物也很難望其項背。

32. 最後一段從哪兩個角度來比較大型動物和小型動物？

- (A) 數量及靈活度
- (B) 出生率及死亡率
- (C) 飲食和居住習慣
- (D) 獵食及逃生技巧

圖 4 華語文閱讀測驗流利級模擬試題

2.2 測驗信度

信度是指測量結果的一致性程度。古典測驗理論中，信度基本上可以分為以下幾種不同的信度係數種類，包括：穩定係數、複本信度、內部一致性係數，以及評分者信度，各自代表不同的誤差來源。內部一致性係數只需一個題本的一次測量結果便能估計信度，所關心的是受試者在各評量項目上的表現一致的程度。因此，內部一致性係數的大小反映的是內容取樣（content sampling）的誤差，以及題目的同質性程度；前者指的是因為題目選擇的隨機因素所造成的分數變異，後者指的是試題是否測量相同的特質（張郁雯 2004），此法為一般能力測驗常使用的信度分析方法。內部一致性的分析方法中，較常見的有 KR20 公式（Kuder-Richardson 20）與 Cronbach's α 係數，其

中 KR20 公式是適用於二元計分的測驗，Cronbach's α 係數則適用於多元計分的測驗，由於本測驗皆為二元計分之選擇題，所以採用 KR20 公式。KR20 公式為 Kuder 和 Richardson (1937) 發展的信度分析方法，主要是依據受試者對整份測驗所有題目的反應，分析題目間的一致性，以確定測驗中的題目是否測量相同的特質（郭生玉 2000）。

在古典測驗理論中，一份測驗只有一個信度數值，因為假定接受同一份測驗的所有考生測量精準度都是相同的。然而在試題反應理論中，不論是單一題目或整份測驗，對不同能力的考生會提供不同的測量精準度；若題目難度越符合考生能力值時，就可以提供較高的測量精準度，若題目難度與考生能力值相差較大，此時的測量精準度就較差。試題反應理論以試題訊息量（item information）表示試題在不同能力點上的測量精準度，訊息量越高表示試題對該能力點的測量精準度越高。將一份測驗中各題的試題訊息量加總後即為測驗訊息量（test information），此概念與古典測驗理論的信度概念非常相似，惟測驗訊息量的高低會隨著考生能力值不同而改變，而古典測驗理論的信度並無此一特性（陳柏熹 2011）。由於本測驗以試題反應理論估計試題難度與組卷，藉由測驗訊息量可了解試題難度分佈是否與通過門檻的能力值相符，故亦以此結果作為評量測驗信度的指標。

2.3 測驗效度

效度是指測驗分數的正確性，也就是測驗能夠測量到所欲測量特質的程度。古典測驗理論中，效度通常可以分為以下三種不同的證據來源，包括：內容效度、效標關聯效度，以及建構效度。Anastasi (1982) 指出建構效度是一個範圍很廣的概念，涵蓋內容效度和效標關聯效度，是指測驗能夠測量到理論上的構念或特質的程度（引自郭生玉 2000）。用於獲得建構效度證據的方法很多，因素分析法是其中一種，分析試題在各因素上的因素負荷量是否符合理論架構的預期，確認收集到的資料與理論架構的符合程度。

大型語言測驗不乏探討測驗建構效度的研究，Sawaki, Stricker 與 Oranje (2009) 發現驗證性因素分析的高階因素模式（Higher-order Factor model）最能解釋 TOEFL iBT 的測驗結構，包含一整體因素（英語外語能力）與四個低階因素（聽、讀、說與寫）。柴省三 (2012) 以階層群集分析（Hierarchical Cluster Analysis，簡稱HCA）進行 HSK 初、中等測驗閱讀篇章的建構效度研究，指出考生對閱讀材料的理解程度主要反映的是其華語閱讀能力的高低，

HSK 閱讀理解測驗的分數解釋或使用具有較高的建構效度。符華均、張晉軍、李亞男、李佩澤與張鐵英（2013）對新版 HSK 五級進行因素分析，以最大概似法進行探索性因素分析的結果，從聽力、閱讀與書寫測驗八種題型中抽出一個因素（華語應用能力），解釋變異量為 64.12%；驗證性因素分析結果則指出，書寫一題型同時測得閱讀與書面表達能力，而其他題型測得之構念則與原先設計相符。由於本測驗根據閱讀理解能力測量面向的不同，區分為幾種題型，因此，本研究採用因素分析方法了解閱讀理解測驗的因素結構是否一致，各題型是否都反映出相同的潛在能力。

在試題反應理論中，試題與模式的適配程度（item fit）可以做為測驗的效度證據。試題反應理論基本假設之一為單向度，是指同一份測驗中的所有題目主要都是測量相同的某一項特質，或是受測者在測驗題目上的答題反應主要是受到單一特質所影響（陳柏熹 2011）。若使用單參數 Rasch 模式對測驗試題收集到的作答反應進行分析，大多數試題都與模式符合，沒有或不適配試題極少，就表示題目內容與單向度模式相符，此份測驗試題測量到相同潛在特質，具有建構效度。因此，本研究亦進行 Rasch 分析並以此結果做為試題反應理論的效度證據。

2.4 測驗等化

因華語文閱讀測驗未來要朝向 CAT 做規劃，並讓不同測驗等級考生成績可以互相比較，追蹤考生長期華語文學習成效，故要進行基礎級與進階、高階與流利級測驗試題難度的垂直等化，使閱讀測驗四個等級試題難度為一完整量尺。本節將介紹等化的種類及等化設計方法。

2.4.1 等化的定義與種類

等化（equating）是指使用統計方法，將一測驗的分數轉換至另一測驗分數量尺，以比較兩個測驗分數關係的過程，其目的是為了校準試題難度的差異。Hambleton 與 Swaminathan（1985）指出測驗等化可分為水平等化（horizontal equating）與垂直等化（vertical equating）兩種，介紹如後（引自張鈺卿 2007）：

（1）水平等化

水平等化是指對兩個以上測量相同特質、相同能力且難度相近的測驗，將其原始分數轉換至同一量尺的過程，常被應用在許多大型測驗，如：托福、

GRE，與基本學力測驗等考試，可在一年中實施多次複本測驗考試，藉由水平等化的過程，將不同複本測驗的成績轉換為同一量尺以進行比較。

（2）垂直等化

垂直等化是指對兩個以上測量相同特質、相同能力但難度不一的測驗，將其原始分數轉換至同一量尺的過程。垂直等化可以將測量特質或能力較為廣泛的測驗結果進行相互比較，當編製測量同一特質卻含有不同程度的測驗，且希望這不同水準的測驗能使用相同的計分量尺時，即適合使用此法。此類測驗的受試者的能力通常是屬於不同年齡或年級，如美國的加州成就測驗（California Achievement Tests, CAT）、愛奧華基本技能測驗（Iowa Test of Basic Skills）等，就是透過垂直等化的方式，將測驗與測驗之間的分數進行連結。

本研究測驗等化目的為將基礎級測驗試題難度連結至現有之進階、高階與流利級測驗難度量尺，乃是將測量相同特質但難度不同之測驗轉換至同一量尺，故屬於垂直等化。

2.4.2 等化設計

測驗等化設計是指收集等化資料的方法。一般常見的等化設計包括單組設計（single group design）、平衡對抗隨機組設計（counterbalanced equivalent groups design）、等群組設計（equivalent group design）、平衡不完全區塊設計（balanced incomplete block design，簡稱 BIB）、試題預先等化設計（item pre-equating design），以及定錨題不等組設計（non-equivalent groups with anchor test design，簡稱 NEAT）等（王寶墉 1995；余民寧 2009；Kolen & Brennan 1995）。

定錨題不等組設計為兩組不同考生作答兩份題本，題本之中放置相同試題，藉由相同試題將兩份題本的其他試題進行連結。另外有學者提出共同題等化設計（common item equating），並再分為複本測驗等化（alternate form equating）和跨樣本等化（across sample equating）。前者做法與定錨題不等組設計類似，同時分析共同題和新題（unique items）；後者則是同一份試卷中，包含難度參數已知的共同題與欲進行參數連結的新題，對一群考生施測後，固定共同題參數，估計新題的難度參數（Yu & Osborn Popp 2005）。

本研究採用的是共同題等化設計的跨樣本等化，此法的優點在於非同時估計共同題與新題難度參數，因此不需重新建立新的量尺，大型測驗經常以

此法建置與擴充測驗題庫。

3. 研究方法

3.1 信度與效度分析研究

3.1.1 分析資料來源

樣本取自 2011 年 5 月份與 11 月份臺灣地區正式考試，5 月份基礎級、進階級、高階級以及流利級測驗考生人數分別為 374 人、758 人、557 人，以及 388 人；11 月份基礎級、進階級與高階級測驗考生人數分別為 159 人、504 人以及 473 人¹。

3.1.2 測驗題本來源

自 2008 年起，進階、高階與流利級測驗試題難度已連結至同一量尺上，各等級測驗每份預試卷皆於各種題型放置 1/5 難度已知，且與單參數 Rasch 模式適配良好的定錨題。在完成預試樣本收集後，固定定錨題難度用以估計新題難度參數。基礎級測驗同樣在每份預試卷之中放置 1/5 定錨題作為試題難度連結之用，惟過去尚未與進階、高階與流利級試題進行參數等化，所使用的難度量尺與此三等級不同。於 2012 年進行垂直等化研究後，已將基礎級試題難度參數連結至進階、高階與流利級測驗量尺（請見 3.2 垂直等化研究一節），本研究使用的基礎級測驗參數為等化後的試題難度參數。本研究分析所使用之正式卷乃依照華測會建置之雙向細目表（如表 1 所示）進行組卷，試題來源為過去預試過，分析結果與單參數 Rasch 模式適配，適配標準為試題之訊息加權均方差（*infit MNSQ*）介於 0.7 至 1.3 之間，或訊息加權標準化殘差（*infit ZSTD*）介於 -3 至 3 之間，適配標準的訂定依據參見下一節資料分析。此外，試題鑑別度（*D* 值、點二系列相關）需達到 0.2 以上；因試題皆已預試過，故難度參數為已知。

3.1.3 資料分析

本研究在信度部分採用內部一致性 *KR20* 公式和 *IRT* 測驗訊息量二個面向進行分析。在內部一致性信度分析方面，使用統計分析軟體 *SPSS 21.0* 版，以 *KR20* 公式進行分析。*IRT* 測驗訊息量方面，以統計分析軟體 *SPSS 21.0* 版，固定正式卷已知試題參數，進行各能力點測驗訊息量及估計標準誤之計算並繪圖。

¹ 2011 年 11 月未舉辦流利級測驗正式考試。

效度部分則分為因素分析和 IRT 試題適配比例二個面向進行分析。在因素分析方面，使用統計分析軟體 SPSS 21.0 版進行探索性因素分析，以主成分分析（principle component factor analysis，簡稱 PFA）萃取特徵值大於 1 之因素；轉軸方式採用直交轉軸（Orthogonal rotation）的最大變異法（Varimax）。IRT 試題適配比例方面，以 IRT 分析軟體 Winsteps 3.68.2 版，採用單參數 Rasch 模式進行分析，檢驗試題與 Rasch 模式適配的數量與比例。

檢驗試題與模式適配與否主要參考二項指標：訊息加權均方差（infit MNSQ）與訊息加權標準化殘差（infit ZSTD），前者根據學者建議，將標準訂於 0.7 至 1.3 之間（Wright, Linacre, Gustafsson, & Martin-Loff, 1994，引自 Bond & Fox, 2007），大於 1.3 或小於 0.7 皆為不適配；後者則設定在 -3.0 至 3.0 之間。標準化殘差一般建議介於 -2.0 至 2.0 之間，但因為此一數值容易受到樣本數量影響，若樣本較大（如 Linacre 指出超過 300 人），即使觀察值和期望值差距很小，在統計上仍可能達到顯著（Lai, Chang, Bode, & Heinemann, 2003；Bond & Fox, 2007；Winsteps & Rasch measurement Software, 2013）。考量本測驗測驗預試樣本人數通常介於 200 至 500 人不等，故本會放寬標準化殘差之判斷標準。

3.2 垂直等化研究

3.2.1 分析資料來源

基礎級測驗試題連結至進階、高階與流利級測驗難度量尺之垂直等化研究的樣本來自 2012 年 7 月及 9 月份臺灣地區預試，7 月份考生人數為 76 人，9 月份考生人數為 157 人，合計有 233 名考生。由於正式考試為考生主動報名並繳交費用，較難要求考生參與相關研究，故以預試樣本進行垂直等化研究。且華語文閱讀測驗樣本為母語非華語之考生，來源取得不易，依據學者 Wright 與 Stone（1979）、陳柏熹（2011）建議用單參數 Rasch 模式估計試題參數時，至少需要 200 名樣本，能力值不會過度集中，就能估計出穩定的試題參數。因此華測會每年固定舉辦預試，每道試題均至少收集 200 名以上的樣本，預試分析的結果，若試題與單參數 Rasch 模式適配，且鑑別度達到 0.2 以上，即可輸入題庫做為未來正式考試組卷之用。

3.2.2 研究設計

等化設計上採用共同題等化設計的跨樣本等化，讓考生作答一份基礎級與進階級測驗混合題本。題本設計未挑選高階級和流利級試題的原因為，雖

然進階、高階與流利級測驗試題難度已連結至同一量尺，但因基礎級測驗對象為華語初學者，考量作答高階或流利級測驗試題，對於初學者過於困難，加上若作答完整的一份基礎與進階級題本，對於考生的體力與精神負荷過大，可能會因疲勞而影響試題難度等化效果，因此分別從基礎級與進階級測驗題庫挑選 20 題及 25 題閱讀試題，組成垂直等化研究用測驗卷，題數合計為 45 題。挑選原則為題庫中難易適中，與單參數 Rasch 模式適配良好，以及難度估計誤差較小之試題；此外，亦考量測驗內容的代表性，雖然題數減為原測驗 1/2，但各題型分佈比例維持不變。

3.2.3 資料分析

基礎級試題難度參數的估計上，分為二階段，使用 IRT 分析軟體 Winsteps 3.68.2 版。第一階段為收集預試考生於垂直等化研究用測驗卷的作答反應後，固定其中進階級試題難度參數，然後估計基礎級試題難度並檢核試題與 Rasch 模式適配情形。第二階段為完成垂直等化研究基礎級試題的難度參數估計後，進一步利用這些完成連結之試題，連結其餘過去已進行水平等化之基礎級閱讀試題難度參數。此步驟亦檢核試題與 Rasch 模式適配情形，對不適配試題進行討論。造成試題不適配原因可能為內容測量到其他能力，或是較為符合二參數或三參數模式，而本研究為單參數模式。對於不適配的試題，由統計分析人員彙整試題的訊息加權均方差、訊息加權標準化殘差數值、試題特徵曲線圖，及古典測驗理論數據，包含難易度、鑑別度與各選項選答人數等資料，提供審題人員參考，並針對內容進行審查。若試題難易度過於困難或過於容易，鑑別度低於 0.2 以下，則考慮刪除；或試題內容設計上有偏頗，如選項設計不佳，易使考生誤解，亦進行刪除。完成基礎級試題難度參數連結後，即建立起橫跨四個等級的大題庫，各等級試題可以互相比較，將以 2011 年 5 月正式考試試題為例，了解四等測驗難度分佈概況。

4. 研究結果

4.1 內部一致性信度

華語文閱讀測驗四個等級正式考試的 KR20 公式分析結果如表 3 所示，5 月正式考試的 KR20 信度係數介於 0.89 至 0.94 之間；11 月結果則介於 0.89 至 0.92 之間，僅基礎級 11 月正式考試有一題刪除後係數些微提高，但增加幅度相當小，僅 0.01，故不對此題內容作進一步的分析討論。

表 3：華語文閱讀測驗 KR20 內部一致性信度

測驗時間	測驗等級	題數	信度	刪除該題後信度係數提高 ²
2011 年 5 月	基礎級	40	0.89	無
	進階級	50	0.94	無
	高階級	50	0.90	無
	流利級	50	0.91	無
2011 年 11 月	基礎級	40	0.90	66 (0.91)
	進階級	50	0.89	無
	高階級	50	0.92	無

4.2 測驗訊息量

華語文閱讀測驗四個等級皆使用 IRT 難度參數進行正式卷之組卷工作，圖 5 至圖 8 為 5 月份與 11 月份正式卷各級測驗之測驗訊息量與估計標準誤（standard error of estimation）結果，並標示出通過門檻能力值相對應之測驗訊息量與估計標準誤，估計標準誤為測驗訊息量開根號之倒數。

參照下圖可知，閱讀測驗 5 月正式考試基礎級的測驗訊息量在能力值 -3.28 至 -3.21 間最大，數值為 7.91，估計標準誤為 0.356；進階級測驗訊息量在能力值為 -1.18 至 -1.13 之間最大，數值為 11.44，估計標準誤為 0.296；高階級測驗，能力值介於 0.17 至 0.24 之間的測驗訊息量最大，為 10.49，估計標準誤為 0.309；至於流利級測驗，則是能力值在 1.38 至 1.50 之間的測驗訊息量最大，數值為 11.04，估計標準誤為 0.301。

11 月正式考試結果，基礎級的測驗訊息量在能力值 -3.11 至 -2.98 間最大，數值為 8.73，估計標準誤為 0.338；進階級測驗訊息量在能力值為 -1.22 至 -1.13 之間最大，數值為 11.49，估計標準誤為 0.295；高階級測驗，能力值介於 0.07 至 0.18 之間的測驗訊息量最大，為 10.78，估計標準誤為 0.305。

此外，各級測驗訊息量之分佈型態符合單一切截分數之標準參照測驗要求，測驗訊息量均呈現單峰分佈。通過門檻的測驗訊息量與估計標準誤如圖 5 至圖 8 以及表 4 所示，基礎級、進階級、高階級以及流利級閱讀測驗通過門檻（基礎級以總分通過門檻 62 分除以 2 計算，閱讀門檻為 31 分）估計得到的考生能力值，均有較高的測驗訊息量與較低的估計標準誤，顯示試題難

² 因 2011 年聽力與閱讀測驗採合併施測，試題編號部分，聽力測驗在前，閱讀測驗在後。以基礎級測驗為例，1 至 40 題為聽力測驗，41 至 80 題為閱讀測驗。

度分佈與通過門檻能力值大致相符；惟與其他等級相較，基礎級測驗訊息量較低，估計標準誤亦較大。

表 4：華語文閱讀測驗各等級通過門檻測驗訊息量與估計標準誤

測驗時間	測驗等級	通過門檻	測驗訊息量	估計標準誤
2011 年 5 月	基礎級	31	5.63	0.421
	進階級	34	10.05	0.315
	高階級	32	9.76	0.320
	流利級	30	10.68	0.306
2011 年 11 月	基礎級	31	6.23	0.401
	進階級	34	10.06	0.315
	高階級	32	10.11	0.315

而由於各級測驗考生能力分佈接近標準常態分佈，以古典測驗理論測量標準誤和信度關係公式 ($SEM = SD\sqrt{1 - \rho_{xx}}$)，將各級測驗通過門檻考生能力值之估計標準誤及受測群能力值之標準差帶入，可得出相當於古典測驗理論的信度。換算後閱讀測驗四個等級 5 月正式考試信度係數由低至高依序為 0.91、0.95、0.93、0.93，11 月正式考試三個等級依序為 0.93、0.92 以及 0.95，顯示閱讀測驗皆具有良好的信度。

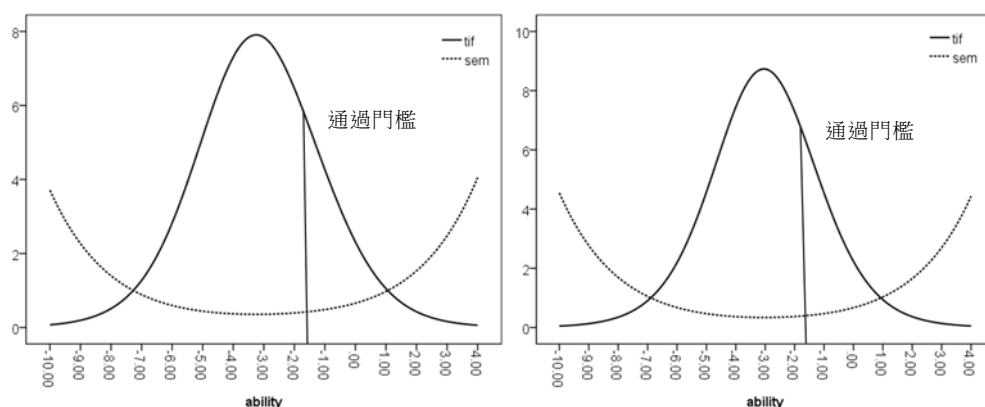


圖 5 基礎級測驗訊息量與估計標準誤（左：5 月；右：11 月）

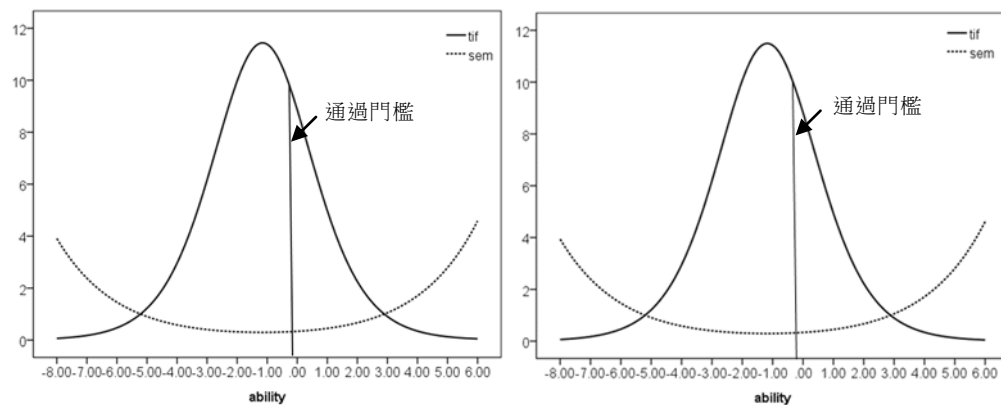


圖 6 進階級測驗訊息量與估計標準誤（左：5 月；右：11 月）

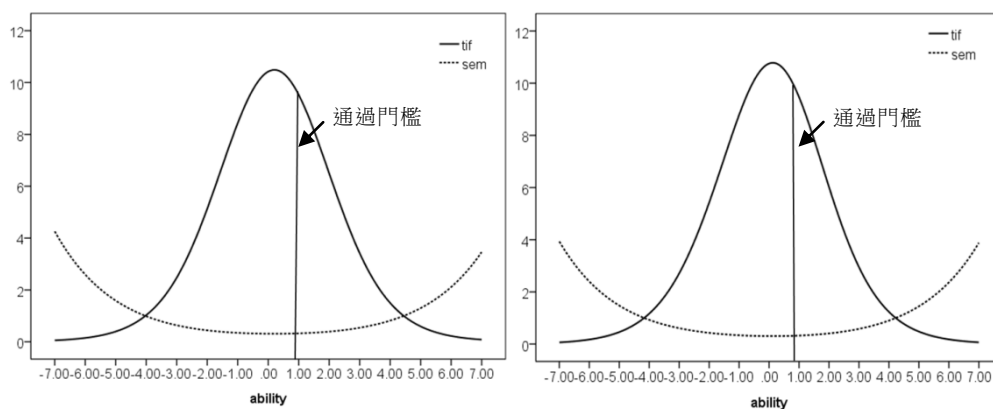


圖 7 高階級測驗訊息量與估計標準誤（左：5 月；右：11 月）

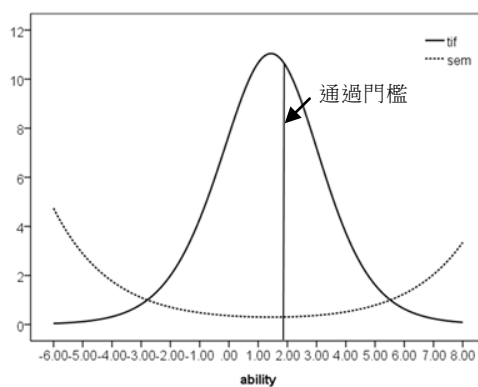


圖 8 流利級測驗訊息量與估計標準誤（5 月）

4.3 因素分析

華語文閱讀測驗答題反應為非對即錯的二分名義變項，不宜直接用於因素分析，因此研究者輸入資料時採用測驗各題型的答對題數，以符合因素分析的要求。在進行因素分析前，需要先做 KMO (Kaiser-Meyer-Olkin) 與 Bartlett's 球形檢定，以確認資料是否適合進行因素分析。2011 年 5 月與 11 月正式考試所有測驗等級的 KMO 數值都在 0.5 以上，根據 Kaiser (1974) 觀點，若 KMO 值小於 0.5，較不宜進行因素分析 (引自吳明隆，2000)；此外，Bartlett's 球形檢定均達顯著 ($p < 0.001$)，二項結果皆表示資料的相關矩陣有共同因素存在，適合進行因素分析，分析結果如圖 9 及表 5 所示。

從陡坡圖可判斷出因素數量，因篇幅有限，且 5 月與 11 月二次正式考試的分析結果相仿，故在此僅呈現 5 月份正式考試結果。從圖 9 可以看到，基礎、進階、高階以及流利級測驗各抽出一個因素後，曲線即下降至平坦處，且其他成分的特徵值皆小於 1。

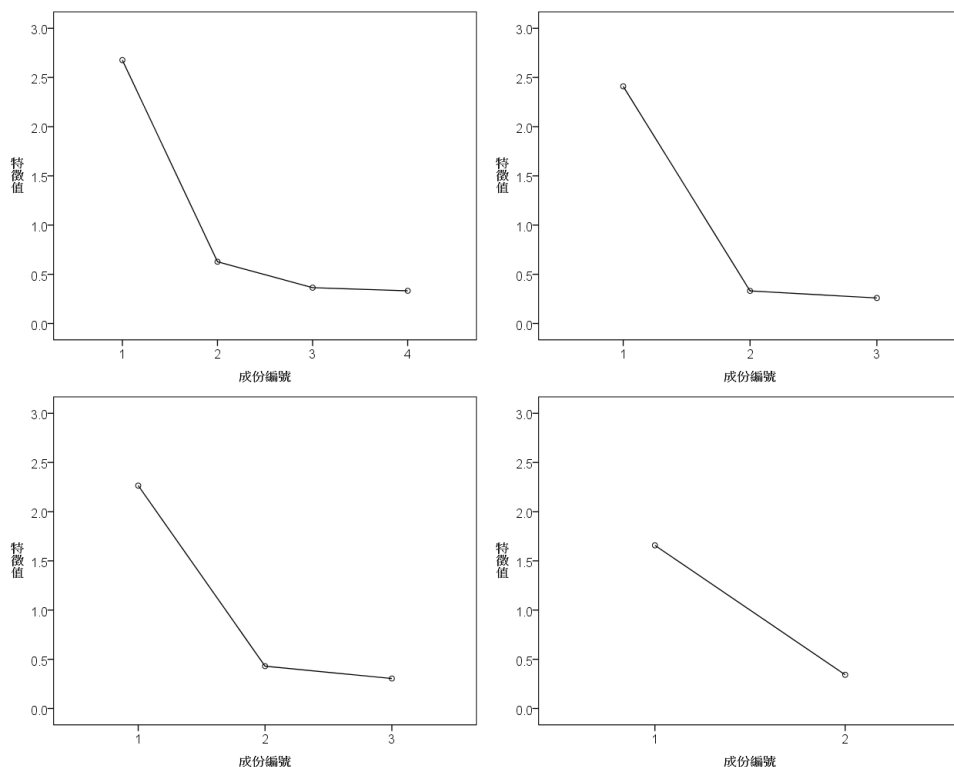


圖 9 華語文閱讀測驗陡坡圖 (由左至右，由上至下為基礎、進階、高階與流利級)

表 5 為因素分析摘要表，可以發現，華語文閱讀測驗各等級都只抽取出一個因素，也就是閱讀理解能力，解釋變異量皆達到 60% 以上，介於 66.91% 至 82.88% 之間，可見華語文閱讀測驗具有良好的建構效度。此外，各等級各題型的因素負荷量都高達 0.70 以上，也表示閱讀理解此一潛在因素對於各題型皆有良好的解釋力。

表 5：華語文閱讀測驗各等級因素分析摘要表

測驗時間	測驗等級	因素	特徵值	解釋變異量	題型	因素負荷量
2011 年 5 月	基礎級	閱讀理解	2.68	66.91%	單句理解	0.70
					看圖釋義	0.86
					選詞填空	0.85
					完成段落	0.85
	進階級	閱讀理解	2.41	80.33%	選詞填空	0.89
					材料閱讀	0.91
					短文閱讀	0.89
	高階級	閱讀理解	2.27	75.51%	選詞填空	0.85
					材料閱讀	0.86
					短文閱讀	0.90
	流利級	閱讀理解	1.66	82.88%	選詞填空	0.91
					短文閱讀	0.91
2011 年 11 月	基礎級	閱讀理解	2.91	72.81%	單句理解	0.81
					看圖釋義	0.84
					選詞填空	0.85
					完成段落	0.91
	進階級	閱讀理解	2.18	72.68%	選詞填空	0.84
					材料閱讀	0.86
					短文閱讀	0.86
	高階級	閱讀理解	2.31	76.99%	選詞填空	0.86
					材料閱讀	0.87
					短文閱讀	0.91

4.4 試題與模式適配度

華語文閱讀測驗基礎、進階、高階與流利級二次正式考試試題與單參數 Rasch 模式適配的比例如表 6 所示。試題與模式之適配標準為訊息加權均方差 (infit MNSQ) 介於 0.7 至 1.3 之間，訊息加權標準化殘差 (infit ZSTD) 介於 -3 至 3 之間，若試題二項指標皆不符合則判斷為與模式不適配 (unfit)。其中，訊息加權均方差大於 1.3，為低適配 (underfit)，小於 0.7 為過度適配 (overfit)，前者會損害估計的品質；後者可能造成膨脹的信度係數 (Bond & Fox, 2007)，相較之下低適配試題對於整個測量模式的影響較為嚴重。

5 月份各等測驗試題與模式適配比例介於 87.5% 至 100.0% 之間，以流利級測驗的適配比例最高，50 道試題皆與 Rasch 模式適配；11 月份則介於 88.0% 至 95.0% 之間，二次正式考試結果顯示各等級測驗試題與單參數 Rasch 模式適配的比例相當理想，絕大多數試題都測量到相同的潛在特質。不適配的 19 道試題中，有 14 題點二系列相關達到 0.20 以上，表示試題有良好的鑑別度，但因部分考生答題反應與模式預期不符或過於符合而造成不適配。組成正式考試的試題均已通過預試分析結果的篩選，但仍可能由於正式考試樣本與預試不同而造成估計結果的改變，本會研發人員將持續追蹤這些不適配的試題在其他正式考試卷的分析結果，若仍不適配，未來會評估是否修改後再重新預試，建立新的參數值。

表 6：華語文閱讀測驗各等級試題與單參數 Rasch 模式適配情形

測驗時間	測驗等級	不適配題數	低適配題數	過度適配題數	適配比例
2011 年 5 月	基礎級	5	2	3	87.5%
	進階級	1	1	0	98.0%
	高階級	1	0	1	98.0%
	流利級	0	0	0	100.0%
2011 年 11 月	基礎級	2	1	1	95.0%
	進階級	6	4	2	88.0%
	高階級	4	3	1	92.0%

4.5 垂直等化試題難度分佈

基礎級與進階、高階、流利級閱讀測驗完成垂直等化後，礙於篇幅僅呈現 2011 年 5 月正式卷之試題難度分佈，結果如表 7 所示。由於基礎級閱讀測

驗題數與其他三個等級不同，為便於比較，統一以百分比表示。由表 7 可知，基礎級難度介於-5.5 至-0.5 之間，有 52.5%比例分佈在-4.0 至-2.5 之間；進階級難度在-3.0 至 0.5 之間，有 76.0%的試題集中在-2.0 至-0.5 之間；高階級與流利級測驗試題難度分佈分別介於-2.0 至 2.0，以及-1.5 至 3.0 之間，前者有 44.0%的試題分佈在-0.5 至 0.5，後者有 70%的試題落在 0.5 至 2.0 之間。四個等級難度分佈雖有重疊，但主要約 50%的試題難度均適當區分開來，各自有較為集中的難度區間。由於本測驗四個等級測量不同閱讀理解程度的考生，各等級的難度分佈自然也需有差異才能區分出考生能力高低，故此一分佈符合預期。此外，本文依照 Angoff 標準設定方法（藍珮君、林玲英，2011）制定出各個等級通過門檻答對題數對應之能力值由低至高依序為-1.61、-0.34、0.89、1.87，皆落在該等級的試題難度分佈範圍內，亦顯示出試題難度分佈良好。除了流利級測驗外，其餘等級通過門檻能力值未落在試題分佈的主要集中範圍內，是因為通過門檻的答對題數約為總題數的六成（進階級為 34 題，高階級為 32 題；基礎級為 31 題，約為總題數八成），故通過門檻的能力值會略高於試題難度集中範圍。

表 7：華語文閱讀測驗各等級試題難度分佈情形

試題難度區間	基礎級	進階級	高階級	流利級
-6.0~-5.5	0.0%			
-5.5~-5.0	5.0%			
-5.0~-4.5	5.0%			
-4.5~-4.0	12.5%			
-4.0~-3.5	17.5%			
-3.5~-3.0	17.5%			
-3.0~-2.5	17.5%	2.0%		
-2.5~-2.0	10.0%	10.0%		
-2.0~-1.5	2.5%	16.0%	2.0%	
-1.5~-1.0	10.0%	28.0%	12.0%	2.0%
-1.0~-0.5	2.5%	32.0%	6.0%	0.0%
-0.5~0.0		8.0%	20.0%	6.0%
0.0~0.5		4.0%	24.0%	2.0%
0.5~1.0			14.0%	14.0%
1.0~1.5			14.0%	30.0%

試題難度區間	基礎級	進階級	高階級	流利級
1.5~2.0			8.0%	26.0%
2.0~2.5				10.0%
2.5~3.0				10.0%

4.6 兩等測驗合併為一等測驗之測驗訊息量與估計標準誤變化

基礎、進階、高階與流利級測驗分別對應至 CEFR 之 A2、B1、B2 及 C1 等級，若參照 CEFR 三等六級（A1、A2、B1、B2、C1、C2）完整架構將相鄰等級合併為一等測驗，則進階與高階級可合併為 B 等，故本研究針對 2011 年 5 月進階與高階級正式試卷題數折半再合併的測驗訊息量與估計標準誤分析。採用奇偶折半法進行試題折半處理，先取進階級和高階級閱讀測驗奇數題組成 50 題之試卷，再取進階級和高階級閱讀測驗偶數題組成 50 題試卷。此法為對原測驗所做的事後分析，因根據試題反應理論，只要已知試題難度參數和考生能力值，即可計算出測驗對不同能力考生可提供的測驗訊息量與估計標準誤。

進階與高階級測驗各自折半後組成 50 題試卷與原測驗之測驗訊息量、估計標準誤之比較如表 8 所示，在進階級方面，測驗訊息量和估計標準誤並無太大變化；高階級方面，測驗訊息量較原測驗略低一些，估計標準誤也略微提高一些。此一結果應與原測驗試題難度組卷原則為中等偏易有關，進階與高階級試題折半合併後，對於進階級通過門檻的考生來說，換了一些較難的試題，這些較難的試題仍能提供一些訊息量，所以影響不大；然而對於高階級通過門檻的考生來說，換了一些較容易的試題，所能提供的訊息量就降低了。

表 8：進階級與高階級原測驗與合併後之通過門檻測驗訊息量與估計標準誤比較

通過門檻	原測驗		奇數題		偶數題	
	測驗 訊息量	估計 標準誤	測驗 訊息量	估計 標準誤	測驗 訊息量	估計 標準誤
進階級	10.05	0.315	10.09	0.315	9.99	0.316
高階級	9.76	0.320	7.42	0.367	7.72	0.360

5. 討論與建議

統整基礎級、進階級、高階級以及流利級閱讀測驗的信度分析結果，內部一致性方法結果，四級測驗之 KR20 係數，在 5 月與 11 月正式考試，都接近或達到 0.90 以上；以 IRT 估計標準誤換算之相當於古典測驗理論信度之數值，在兩次正式考試，四級測驗皆達到 0.90 以上。此外，由圖 5 至圖 8 及表 4 可知，基礎級、進階級、高階級以及流利級閱讀測驗，在通過門檻分數與能力值均有較高的測驗訊息量和較低的估計標準誤，惟基礎測驗訊息量與估計標準誤數值比其他等級不理想，可能原因為題數較少，因測驗訊息量為試題訊息量的加總，基礎級閱讀測驗題數為 40 題，其他等級皆為 50 題。未來可以考慮增加基礎級題數，以提高測驗訊息量並降低估計標準誤。

根據學者 Gay (1992) 的觀點，任何測驗或量表的信度係數如果在 0.90 以上，表示測驗或量表的信度甚佳（引自吳明隆 2003）；陳柏熹（2011）則表示由於能力特質定義通常較具體清楚，且有明確答案或評分標準，因此能力測驗的信度最好能在 0.8 以上。而同樣測量華語文能力的新漢語水平考試（HSK），在 2011 年公布的分析結果顯示，一級聽力與閱讀測驗合計的信度係數介於 0.85-0.95 之間，三級聽讀合計的信度在 0.90-0.95 之間；五級聽讀合計的信度在 0.90-0.95 之間（張晉軍 2011）。其他大型語言測驗，如 TOEFL iBT 閱讀測驗近年公布的信度分別為 0.86 與 0.85（ETS, 2007; ETS, 2011）；TOEIC 於 2012 年的測驗介紹手冊宣稱聽讀測驗 KR20 信度係數接近或達到 0.90 以上（ETS, 2012）。依照上述學者提出之標準，華語文閱讀測驗四個等級的測驗信度皆為良好，信度係數都接近或達到 0.90 以上，此一信度分析結果也與相關語言測驗的數值相仿，甚至更佳。此外，本測驗各級測驗訊息量分佈型態亦符合單一切分數之標準參照測驗要求，基礎級、進階級、高階級以及流利級測驗通過門檻所對應的考生能力值皆有較高的測驗訊息量與較低的估計標準誤，由於測驗訊息量為試題訊息量的加總，測驗訊息量較高，表示一份測驗中有較多試題難度與通過門檻相符，提供較高的試題訊息量，對於通過門檻附近能力值的估計也會較為精確，得到較低的估計標準誤。整體而言，華語文閱讀測驗具有良好的測驗信度。

華語文閱讀測驗的效度證據方面，因素分析結果，各等級測驗均抽取出單一因素，即閱讀理解能力，可解釋變異量都達到 66.91% 以上，顯示具有良好之建構效度。試題與模式適配比例的分析結果，也指出各等級測驗適配的

比例相當理想，最低為 87.5%，最高甚至達到 100%，絕大多數試題都測量到相同的潛在特質，也就是閱讀理解能力，亦可做為閱讀測驗的效度證據。從古典測驗理論與試題反應理論二方面的分析結果都指出華語文閱讀測驗具有建構效度。

垂直等化研究結果，將基礎級測驗連結至既有之進階、高階與流利級測驗難度量尺，試題難度分佈良好，亦與單參數 Rasch 模式適配良好。以 2011 年 5 月正式卷來看，四個測驗等級，雖然在難度分佈上與相鄰等級有所重疊，但各自有不同的試題難度集中區間，主要 50% 的試題難度均適當區分開來，顯示四個等級確實測量到不同難度水準之華語文閱讀能力。完成基礎、進階、高階與流利級測驗試題難度連結後，未來可朝電腦適性化測驗做更進一步的發展。

最後，嘗試以奇偶折半法將進階級與高階級兩等測驗合併為一等測驗的分析結果，無論是奇數題或偶數題版本，通過門檻的測驗訊息量和估計標準誤，在進階級與原測驗的變化均不大，高階級測驗訊息量則略微降低。大體來看，若將進階級、高階級測驗合併為一個測驗，應為可行，但未來規劃合併等級後組卷的試題難度比例時，必須再作微幅調整，需特別注意高階級測驗通過門檻的測驗訊息量，可將試題難度集中在兩個等級通過門檻附近，更能確保通過門檻的測驗訊息量。合併測驗等級的方式，讓考生只要用和過去相同的時間和金錢參加考試，就能多了一次判斷華語能力水準的機會；對於華測會本身來說，精簡測驗等級，也可以減輕試務工作和研發試題的雙重壓力，可說是一舉數得。華測會將針對合併測驗等級相關細節進行研議，期待能提供考生品質與效率兼具的華語文能力測驗。

引用文獻

- Bond, Trevor G., and Christine M. Fox. 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Educational Testing Service. 2007. *TOEFL® iBT Score Reliability and Generalizability*. Retrieved Sep 26, 2013, from http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Reliability_Generalizability.pdf

- Educational Testing Service. 2011. Reliability and Comparability of TOEFL iBT® Scores(PDF). *TOEFL iBT Research Insight Series 1, Vol. 3*. Retrieved Sep 26, 2013, from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf
- Educational Testing Service. 2012. *TOEIC Examinee handbook listening & reading*. Retrieved Sep 26, 2013, from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf
- Kolen, Michael J., and Robert J. Brennan. 1995. *Test Equating: Methods and Practices*. New York: Springer-Verlag.
- Lai, J., D. Cella, , C. H. Chang, R. K. Bode, and A. W. Heinemann. 2003. Item banking to improve, shorten, and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue scale. *Quality of Life Research* 12: 485-501.
- Sawaki, Y., L. J. Stricker, and A. H. Oranje. 2009. Factor structure of the TOEFL Internet-based test. *Language Testing* 26.1:5-30.
- Winsteps and Rasch measurement Software. 2013. *Misfit diagnosis: Infit outfit mean-square standardized*. Retrieved from <http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm>.
- Wright, B. D., and M. H. Stone. 1979. *Best test design*. Chicago: MESA Press.
- Yu, Chong Ho., and Osborn Popp, Sharon E. 2005. Test Equating by Common Items and Common Subjects: Concepts and Applications. *Practical Assessment, Research & Evaluation* 10.4:1-19.
- 王寶壙。1995。《現代測驗理論》。臺北市：心理出版社。[Wang, Bao-Yong. 1995. *Modern Test Theory*. Taipei: Psychological Publishing Co., Ltd.]
- 余民寧。2009。《試題反應理論 IRT 及其應用》。臺北市：心理出版社。[Yu, Min-Ning. 2009. *Item Response Theory and Application*. Taipei: Psychological Publishing Co., Ltd.]
- 吳明隆。2003。《SPSS 統計應用實務》。臺北：松崗。[Wu, Ming-Lung. 2003. *SPSS Statistical Application and Practice*. Taipei: Unalis Corporation.]
- 柴省三。2012。〈關於 HSK 閱讀理解測驗構想效度的實徵研究〉，《世界漢語教學》，2012 年，第 2 期，243-253。北京市：北京語言大學。[Chai, Xing-San. 2012. An Empirical Research on the Construct Validity of the Reading Comprehension Test of HSK. *Chinese Teaching in the World* 2012.2:

- 243-253. Beijing: Beijing Language and Culture University.]
- 張郁雯。2004。〈信度〉，《教育測驗與評量：教室學習觀點（第二版）》，王文中、呂金燮、吳毓瑩、張郁雯及張淑慧合著，95-130。臺北：五南書局。
- [Chang, Yu-Wen. 2004. Reliability. *Educational Assessment: A Classroom Learning Perspective*(2nd). eds. by Wang, Wen-Chung, Lu, Chin-Hsieh, Wu, Yuh-Yin, Chang, Yu-Wen, and Chang, Shu-Hui, 95-130. Taipei: Wu-Nan Book Inc.]
- 張晉軍。2011。〈新漢語水準考試（HSK）品質報告〉。2013年9月26日，取自：http://blog.sina.com.cn/s/blog_53e7c11d0100v7lz.html [Zhang, Jin- Jun. 2011. The report of the new Hanyu Shuiping Kaoshi (HSK). Retrieved Sep 26, 2013, from http://blog.sina.com.cn/s/blog_53e7c11d0100v7lz.html]
- 張鈺卿。2007。《BIB 與 NEAT 設計在不同年度測驗連結效果之比較》。國立臺中教育大學教育測驗統計研究所碩士論文（未出版）。[Chang, Yu-Ching. 2007. The performances of BIB and NEAT designs for linking large-scale assessments in different years. Taichung: National Taichung University of Educations master thesis (Unpublished).]
- 符華均、張晉軍、李亞男、李佩澤、張鐵英。2013。〈新漢語水平考試 HSK(五級)效度研究〉，《考試研究》，2013 年，第 3 期，65-69。[Fu, Hua- Jun, Zhang, Jin- Jun, Li Ya- Nan, Li, Pei- Ze, and Zhang, Tie-Ying. 2013. The Validity Research of New HSK (Level 5). *Examinations Research*, 2013.3:65-69.]
- 郭生玉。2000。《心理與教育測驗（第 14 版）》。臺北：精華書局。[Guo, Sheng-Yu. 2000. *Psychological and Educational Testing* (14th). Taipei: Jing- Hua Book Company]
- 陳柏熹。2011。《心理與教育測驗-測驗編製理論與實務》。臺北：精策教育。[Chen, Po-Hsi. 2011. *Psychological and Educational Testing: Theoretical and Practical of Test Development*. Taipei: Planned Education Ltd.]
- 藍珮君、林玲英。2011。〈新版華語文能力測驗與 CEFR 之連結：標準設定方法的應用〉。論文發表於 ALTE 第四屆國際研討會，波蘭克拉科。2011 年 7 月 7-9 日。[Lan, Pei-Jiun, and Ling-Ying Lin. 2011. Cut-off Scores of the New Chinese Proficiency Test Based on the Angoff Standard Setting Method. Presented at the ALTE 4th International Conference, Krakow, Poland. 2011.7.7~9].

華語文教學研究

[審查：2013.8.9 修改：2014.2.19 接受：2014.2.21]

藍珮君

Pei-Jiun LAN

106 臺北市青田街 5 巷 6 號 6 樓

6F., No.6, Ln. 5, Qingtian St., Taipei City 106, Taiwan

martinalan@sc-top.org.tw

陳柏熹

Po-Hsin, CHEN

106 臺北市和平東路一段 162 號

No.162, Sec. 1, Heping E. Rd., Taipei City 106, Taiwan

chenph@ntnu.edu.tw

A reliability, validity and vertical equating study of the reading subtest of the Test of Chinese as a Foreign Language

Pei-Jiun LAN

**Steering Committee for the Test
of Proficiency-Huayu**

Po-Hsi CHEN

**Department of Educational
Psychology and Counseling,
National Taiwan Normal University**

Abstract

The purpose of this study is to investigate the reliability, validity and vertical equating of the Reading subtest of the Test of Chinese as a Foreign Language. Four levels are included in the reading section, they are Level 2, 3, 4, and 5, respectively. The analysis data was sampled from the formal version of the test administered in 2011 and pretest version in 2012. The results showed that, first, the coefficients of the Kuder-Richardson 20 were closed to or higher than .90. Moreover, large test information is provided to the value of cutoff which is determined an examinee is passed or failed. In other words, low standard error of estimation was obtained for the examinees. Second, the results of factor analysis showed that only one factor was extracted, which could account for above 66% of the variance. In addition, the results of Rasch analysis revealed that more than 87.5% of the items fit the model well. Third, there is a suitable range of difficulties for each level of test. Finally, standard error of estimation about the cutoff values were similar to Level 3 but lower than Level 4 when the items in Level 3 and 4 were split to assemble two tests (i.e., test information on the cutoff values for the even items included in Level 3 and 4, the odd items included in Level 3 and 4, and items in Level 3 and 4). That is these two adjacent levels can be combined to form a composite level of test in the future to reduce the burden for examinees and developers of the test. However, the item difficulty distribution of the composite test should be adjusted.

Keywords: mandarin test, reliability, validity, item response theory, vertical equating