華語文能力測驗電腦與紙筆測驗試題難度比較研究

藍珮君　　林玲英

martinalan@sc-top.org.tw　　hopelin.top@gmail.com

國家華語測驗推動工作委員會

摘要

本研究目的為探討在紙筆測驗與電腦化測驗兩種不同的施測介面下，考生於華語文能力測驗(TOCFL)基礎、進階、高階以及流利級四個等級的測驗成績以及試題難度參數的可比性(comparability)。

研究設計採用共同組法(single group design)以及對抗平衡法(counterbalanced design)，收集考生在紙筆測驗與電腦化測驗的作答表現，進行兩種施測介面的測驗總分、聽力與閱讀成績 t 考驗、試題難度參數差異比較，以及試題難度參數積差相關分析。

研究結果發現：1. 四個測驗等級中，僅基礎和進階級的聽力理解測驗得分 t 考驗達到顯著差異，但效果量相當小，顯示測驗介面對於考生成績的影響很小；2. 試題難度方面，基礎、進階、高階以及流利級測驗難度參數差異較大的試題佔總題數比例均不到一成；3. 四個測驗等級共八個分測驗的紙筆測驗和電腦化測驗試題難度參數間有高度正相關存在。整體來說，華語文能力測驗(TOCFL)電腦與紙筆介面之間具有可比性，考生的成績表現幾乎沒有不同，也僅有少數試題在不同介面的難度參數有所變動。最後，研究者針對難度參數變化較大的試題，諮詢並整理專家意見，提出可能的解釋與建議，供未來命題做為參考。

關鍵字：紙筆測驗、電腦化測驗、試題難度參數

前言

21 世紀的今日，電腦和網際網路逐漸成為生活中不可或缺的一部份。國際電信聯盟(International Telecommunications Union, ITU)估計，2010 年已開發國家有 71.0%的家庭擁有電腦；開發中國家有 22.5%的家庭有電腦，而已開發和開發中國家分別有 65.6%和 15.8%的家庭擁有網路。根據 comScore 在 2008 年 12 月針對 15 歲以上網路族(在家中或公司上網者)所做的調查，目前全球網路人口已超過十億大關(引自行政院研究發展考核委員會，2010)。顯現隨著科技的進步，電腦以及網際網路的使用已日漸普及。

在教學或教育的影響上，學校與教學單位開始採用電腦或多媒體作為教學的輔助工具；在評量方面，過去常用的紙筆測驗不再是唯一考量，特別是大型考試。由於電腦化測驗(computer-based tests，CBT)的許多優點，如：無須大量印刷題本及運送、可即時計分、施測程序更為標準化，已有許多大型語言測驗發展並實施電腦化測驗或電腦適性化測驗(computerized adaptive tests，CAT)，如：TOEFL-iBT、GRE、BULATS 等。

由國家華語測驗推動工作委員會(簡稱華測會)研發之華語文能力測驗(Test of Chinese as a Foreign Language，簡稱 TOCFL，前稱為 TOP)，自 2003 年 12 月在台灣第一次舉行正式考試以來，至今在台灣以及海外各地已累積二萬多名考生。在維持並提升試題品質的同時，華測會也不斷思考如何提供給考生更為便利的服務和更佳的施測品質。從 2009 年起，委託國立台灣師範大學心理與教育測驗研究發展中心，開發華語文能力測驗的電腦化測驗系統，經過多次的測試及預試，已於 2011 年 5 月於台灣舉行電腦化測驗第一次正式考試。

雖然台灣地區目前預試和正式考試均採用電腦化測驗，但海外地區受限於施測方式為區域網路，非透過網際網路傳輸，短期內仍需採用紙筆測驗進行施測。且未來有些地區可能因場地與電腦設備問題，僅能實施紙筆測驗。在同時提供紙筆和電腦兩種施測形式的狀況下，必須確認電腦測驗與紙筆測驗試題難度參數是否相當，考生的測驗表現不會受到測驗介面的影響，以確保考生權益及測驗的公平性，故進行本研究。

本研究欲探討以下問題：

1. 基礎、進階、高階以及流利級測驗，考生在紙筆與電腦介面的測驗成績是否有顯著差異？
2. 基礎、進階、高階以及流利級測驗，紙筆與電腦介面的試題難度參數是否有差異？
3. 如果有介面效應(mode effects)存在，可能是哪些因素造成的？

文獻探討

隨著電腦和資訊科技的進步，電腦測驗的許多好處使得測驗發展者或測驗機構紛紛設計與開發電腦測驗提供給考生使用。舉例來說，只要設定完成，電腦測驗在施測上比紙筆測驗更加容易，還能提供即時計分的功能(Bugbee & Bernt, 1990；Inouye & Bunderson, 1986；引自 Bodmann & Robinson, 2004)；電腦測驗施測情境標準化，試題呈現的順序容易操控(Inouye & Bunderson, 1986；引自 Bodmann & Robinson, 2004)。此外，Noyes 和 Garland(2008)也整理出電腦測驗的一些優點：電腦測驗可以使用彩色圖片、影片等多媒體素材多元呈現試題內容，測驗介面比傳統紙筆測驗來的豐富許多；電腦測驗可以測量作答反應時間等知覺表現；電腦測驗可以透過網際網路，讓不同地點的考生參加測驗，場地較不受侷限，讓測驗的使用群體更廣泛。

不過即便如此，電腦測驗仍有一些缺點或限制有待克服與處理。像是電腦測驗要使用硬體和軟體，一旦考試進行中電腦突然當機時，可能因為重新開機或更換電腦而浪費時間；試題較多的測驗，考生長時間觀看電腦螢幕的情形下，在電腦上作答會比在紙本上更容易感到疲倦；另外是試題呈現的問題，由於電腦螢幕大小的限制，電腦測驗很難達到與紙筆測驗完全相同的試題呈現方式。而在測驗操作上，一般來說，使用紙筆測驗的考生較為容易瀏覽試題和往前或往後翻頁(Noyes & Garland, 2008)。

也因為電腦測驗和紙筆測驗試題呈現和作答方式較為不同，若測驗單位採用紙筆測驗和電腦測驗並行的機制，學者 Parshall、Spray、Kalohn 和 Davey(2002)提出警告，基於紙筆測驗得到的試題校正參數(item calibration)可能不能代表同樣的試題在電腦施測介面上的表現。Parshall 等人同時建議要進行相關研究釐清使用不同介面對於考生成績和試題難度有無造成差異，以確保測驗分數的可比性(comparability)。

美國心理學會(APA)於 1986 年公布的「電腦化測驗方針和解釋(Guidelines for computer-based tests and interpretations)」，也提到解釋來自傳統測驗(conventional tests)電腦化版本的分數時，在使用從傳統測驗取得的常模或是通過門檻前，應該要先建立並提供電腦版本分數等值(equivalence of scores)的文件或資料[1](引自 Mead & Drasgow, 1993)。

過去二、三十年來，已有不少研究針對電腦測驗與紙筆測驗的可比性，或是測驗形式、介面對考生的影響進行探討。但至今研究結果仍相當分歧，有的研究發現紙筆測驗考生表現較佳，如美國大學認證測驗(CLEP)的數學科和英文科，紙筆測驗得分優於電腦(Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein ,1991；引自 Clariana & Wallace, 2002)；大學生課堂上微生物學測驗的表現，紙筆測驗平均得分高於電腦(Lee & Weerakoon, 2001)；Al-Amri(2008)也發現考生在 TOEFL 閱

[1] 原文爲 When interpreting scores from the computerized versions of conventional tests, the equivalence of scores form computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests.

讀測驗的紙筆測驗成績顯著高於電腦測驗。有些研究結果則指出電腦測驗對考生比較有利，Parshall 和 Kromrey(1993)對 GRE 做的研究即發現，考生在電腦測驗語文、計量和分析寫作三向度的表現都比紙筆測驗佳；大學生計算機概論課程的測驗成績，電腦介面表現優於紙筆介面(Clariana & Wallace, 2002)。

但也有其他研究的結果顯示，考生的電腦和紙筆測驗成績沒有差異。Mead 和 Drasgow(1993)對認知能力測驗電腦與紙筆介面的可比性研究進行後設分析，最後發現施測介面對於能力測驗的影響很少，幾乎可以說，紙筆測驗和電腦化測驗測量到的是相同的構念，特別是在測驗架構良好，作答時間充裕，且包含不同難度題目的前提下。Bodmann 與 Robinson(2004)對大學生進行研究，結果指出學生在課堂紙筆測驗和電腦測驗的平均得分沒有顯著差異。Kim 和 Huynh(2008)對大型全州英語測驗(large-scale statewide end-of-course English examination)的電腦和紙筆成績進行試題和測驗層次的分析，整體來說，從紙筆和電腦介面得到的分數可以相互比較，不過在內容領域層次，閱讀理解測驗可能較容易受到施測介面的影響。驗證性因素分析的結果也顯示施測介面並未改變測驗的構念。

另外還有一些研究本身的結果就不一致，Choi、Kim 與 Boo(2003)研究首爾國立大學英語能力測驗(TEPS)紙筆和電腦版本之間的可比性。採用語料庫語言學技術(corpus linguistic techniques)、相關分析、變異數分析以及驗證性因素分析(CFA)等統計方法進行測驗內容分析和建構效度的檢驗。儘管 Choi 等人最後宣稱研究結果支持 TEPS 分測驗(聽力理解、文法、詞彙和閱讀理解)紙筆和電腦版本之間的可比性，但變異數分析結果顯示聽力理解、詞彙和閱讀理解的成績在測驗形式(administration mode)有主要效果，作者解釋閱讀和聽力的主要效果可以接受，因為紙筆和電腦測驗在這兩個分測驗的圖片訊息呈現方式不太一樣，惟詞彙分測驗的呈現方式是相同的。

Pommerich(2007)對 3000 多名 11 和 12 年級學生評估相同內容的測驗在紙筆和電腦介面得到的試題難度參數差異，測驗內容以段落為主，領域為閱讀和科學推理(science reasoning)，電腦介面再分為換頁(paging)和捲軸(scrolling)兩種操作方式。t 考驗結果顯示，紙筆施測介面的閱讀平均成績雖高於兩種電腦施測介面，但未達到顯著差異水準；科學推理測驗方面，電腦施測介面的考生平均得分高於紙筆施測介面，但只有換頁組達到顯著水準。至於相關分析結果，兩種電腦施測介面均與紙筆測驗的試題難度參數達到 0.9 以上的高度正相關。

Kingston(2009)對 1997-2007 年之間美國 1 到 12 年級學生共 81 篇數學、閱讀、英語語言藝術(English Language Arts)、科學和社會研究(Social Studies)學科的電腦和紙筆可比性研究進行後設分析，瞭解年級或科目是否對可比性有影響。結果顯示年級沒有影響，但科目有影響，電腦考試的學生在英語語言藝術和社會研究兩科稍微有利；而紙筆測驗在數學學科稍微有利，不過效果量(effect size)相當小，均小於 0.20。

由於目前的研究發現仍莫衷一是，吾人無法輕易下結論電腦與紙筆介面的施測形式對於考生成績究竟有無介面效應。是故，在已預期短期內華語文能力測驗

(TOCFL)的電腦測驗與紙筆測驗將同時並行的情況下，華測會研發人員實有必要針對兩種測驗介面的考生作答表現進行可比性研究。此外，Pommerich(2004)提到評估介面效應時，不能單從測驗總分層次去檢視可比性，也要從試題層次進行檢視。因為個別的試題可能有很強(strong)的介面效應，但在整體分數層次可能會相互抵銷。因此在本研究中，除了測驗平均數的比較外，也針對個別試題難度進行探討，確認試題難度是否會因為介面的不同而有明顯的變化，以及考生的成績是否會受到影響。

至於測驗介面效應的影響因素，許多研究均提到閱讀長篇文本是造成電腦和紙筆測驗成績有差異的原因之一。Mourant、Lakshmanan 和 Chantadisai(1981)的研究發現，學生在電腦上閱讀文本會比在紙本上閱讀相同文本來得疲勞；Haas 和 Hayes(1986)亦指出當與試題有關的文本篇幅超過一頁時，考生的電腦測驗成績會比紙筆測驗低，顯然是由於在螢幕上閱讀較長文本有困難(均引自 Clariana & Wallace, 2002)。其他學者的研究也發現，當試題所有訊息無法完整在螢幕呈現，考生需要使用滑鼠滾輪瀏覽試題全貌時，這些測驗通常有顯著的介面效應(Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Bergstrom, 1992；引自 Pommerich, 2004)。Peak(2005)則宣稱已有充分的證據可以顯示，電腦施測的方式不會顯著影響學生表現，除非測驗內容包含長篇的閱讀段落(引自 Pommerich, 2007)。

對此，Pommerich(2004)提出可能的解釋為，紙筆介面的考生也許比電腦介面考生容易產生位置記憶(positional memory)，能記得段落中某個訊息的位置，因為紙本頁面中段落均在固定的位置。使用捲軸的電腦考生反應有時候在段落中找尋訊息會遇到困難。

除了閱讀長篇文本外，電腦操作介面的彈性和複雜度也是學者認為造成介面效應的可能原因。Spray、Ackerman、Reckase 和 Carlson(1989)提出假設，軟體的彈性(flexibility of the software)是造成介面效應的主因。測驗若設計能讓考生回頭看之前的作答反應，也能夠修改，則跨介面的成績就可以互相比較(引自 Mead & Drasgow, 1993)。Mason、Patry 和 Bernstein(2001)也發現有研究證據支持測驗介面效應是由於彈性的差異(引自 Bodmann & Robinson, 2004)。其他研究結果同樣指出，電腦測驗的施測介面越複雜，介面效應可能越大；所有試題訊息都能完整呈現在電腦螢幕上的測驗通常有小或不顯著的介面效應(Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Hetter, Segall, & Bloxom, 1997; Bergstrom, 1992; Spray, Ackerman, Reckase, &Carlson, 1989；引自 Pommerich, 2004)。

最後，研究者參考 Mead 與 Drasgow(1993)提出可比性研究內部效度的威脅：(1)不同施測介面採取不同的施測流程；(2)實施的版本沒有使用對抗平衡法；(3)考生群組沒有隨機分派；(4)考生參加不同版本測驗的動機不同。在研究設計上採取隨機分派、對抗平衡等設計，以避免上述問題。

研究方法

一、研究對象與研究設計

本研究共分為二階段進行，第一階段樣本為參加 2010 年 9 月預試之母語非華語考生，但由於基礎級和流利級報考考生人數過少，故第一階段先針對進階級與高階級測驗結果進行分析，並於 2011 年 3 月預試進行第二階段研究，對基礎級與流利級測驗再次收集樣本。

本研究設計採用共同組法(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；另一組為早上進行紙筆測驗，下午進行電腦測驗。由於紙筆與電腦測驗內容完全相同，為避免考生刻意記憶試題，事先未告知考生。

最後完成紙筆與電腦測驗的基礎級、進階級、高階級與流利級測驗有效考生人數分別為 292 人、187 人、143 人以及 275 人。

二、研究材料

華語文能力測驗(TOCFL)是一專為母語非華語之語言學習者設計的華語測驗，針對「聽力」與「閱讀」兩種能力進行評量。2011 年 5 月在台灣將正式推出「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。除基礎級外，其它三級測驗變更部分題型，題數則從原 120 題減為 100 題，測驗時間為 110 分鐘。基礎級測驗時間為 80 分鐘，測驗題目共 80 題。四個等級的題型分為聽力及閱讀兩大部份。測驗題目皆為單選題，每題一分；答錯不倒扣。應試者可依自己的學習背景或語言能力選擇合適的等級應考。 新版題型與題數對照見表 1。

表 1　新版華語文能力測驗測驗等級、題型與題數分佈表

測驗等級	聽力理解	閱讀理解
基礎	看圖回答 10 題	單句理解 10 題
	問答理解 10 題	看圖釋義 10 題
	對話理解 10 題	選詞填空 10 題
	完成對話 10 題	完成段落 10 題
進階	單輪對話 20 題	選詞填空 20 題
	雙輪對話 15 題	材料形式 15 題
	段落 15 題	短文 15 題

表 1(續)

測驗等級	聽力理解	閱讀理解
高階	短對話 20 題 長對話 15 題 段落 15 題	選詞填空 15 題 材料形式 10 題 短文 25 題
流利	短對話 10 題 長對話 20 題 段落 20 題	選詞填空 15 題 短文 35 題

因目前華語文能力測驗(TOCFL)有紙筆測驗和電腦測驗兩種介面，以下分別對兩者進行介紹：

(一)紙筆測驗

紙筆測驗每頁均為 A4 大小，字體大小為標楷體 14 號字，30 個字一行。因華語文能力測驗共分為四個等級，加上題型設計的不同，故每頁可呈現的題數也有所差異。基礎測驗聽力理解測驗部分，看圖回答、問答理解與對話理解三大題一頁為 5 題；完成對話題型一頁有 10 題。閱讀理解測驗部分，單句理解和選詞填空一頁約有 5 題；看圖釋義一頁約 2-5 題；完成段落一頁為 10 題。

進階、高階和流利級測驗的聽力理解測驗方面，單輪對話、雙輪對話和短對話一頁約 5-6 題；長對話和段落聽力題型一頁約 3-4 題。閱讀理解測驗方面，選詞填空一頁有 5 題；材料題一頁約有 1-2 題；短文閱讀題型一頁約 2-4 題。短文的閱讀篇章，皆排版在同一頁內讓考生閱讀，沒有跨頁的情形。

(二)電腦測驗

電腦測驗的螢幕規格為 19 吋，解析度設定為 1024×768，若寬螢幕則為 1280×800，字形與紙筆測驗相同，為標楷體，字體則考量長時間觀看螢幕較容易疲勞，因此設定較紙筆測驗略大，為 16 號字。華語文能力測驗(TOCFL)的電腦測驗設計之初，為了盡可能讓電腦考試和紙筆測驗試題呈現和作答方式盡量一致，段落編排方面，為一行 29 個字，一次呈現一道試題。在試題內容的配置及排版上，除了閱讀測驗「選詞填空」的選項排列方式受限於螢幕大小，從紙本的「由上至下」排列改為「由左至右」外，其餘題型均與紙筆測驗相仿。

而為避免考生因不熟悉操作介面而影響其作答表現，有許多設計與安排。像是在正式進入考試前有考試系統操作影片、聽力以及閱讀測驗說明影片，向考生說明如何操作電腦，如：滑鼠的使用、如何拉動捲軸、圖表放大鏡功能等等；每一大題之前也有大題說明影片，說明每一個題型的作答方式。在電腦操作方面，考生登入帳號密碼後，全程皆使用滑鼠；在聽力理解測驗可以自行調整耳機音量大小聆聽試題內容。在閱讀理解測驗，考生可以選擇先跳過某些試題，之後再回頭作答或檢查，也可以修改答案。

三、研究程序

收集考生於電腦及紙筆測驗之作答反應後，以 IRT 軟體 Winsteps，採用同時估計法，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，聽力與閱讀測驗採分開估計。估計完難度參數後，再進而比較同一道試題在電腦與紙筆介面的難度差異，由於試題標準誤平均值約為 0.25，故以標準誤的二倍 0.5 作為判斷標準。若難度相差超過 0.5 logit (logit 為 IRT 難度參數單位)，則視為差異較大之試題，需進一步探討可能原因。

為了解相同的試題為何在不同的施測介面下，試題難度參數會產生差異，研究者舉行諮詢會議，邀請華語文教學及測驗領域相關之專家學者與會共同討論可能的影響因素。研究者邀請 16 位專家進行書面與會議二階段的諮詢工作，部分專家學者礙於時間因素無法參與，最後共計有 10 位專家參與，有 1 名華語文教學領域教授及 1 名資深華語教師提供書面建議；其餘 8 名專家全程參與二階段諮詢工作，其中華語文教學領域專家共 5 位，測驗計量專家有 2 位，資深華語教師 1 位。

研究結果與討論

一、完成作答的比例

考生分別在紙筆與電腦測驗作答完成的比例如表 2 所示，在此作答完成的標準是指完成最後一題作答，若考生於最後一題沒有作答，則視為未完成測驗。考生在高階級和流利級紙筆測驗的完成比例略高於電腦測驗；基礎級和進階級測驗則是電腦測驗的作答完成比例略高，但兩者的差異並不大，差異最大為高階級測驗，也僅有 2.8%。

研究者推測基礎級和進階級測驗電腦介面完成作答比例較高的原因，可能由於電腦測驗考生作答只需直接以滑鼠點選選項前的圓圈即可完成作答，紙筆測驗考生則要使用 2B 鉛筆在答案卡上將選項方格塗黑、塗滿，所花費的時間較長。在高階級和流利級測驗，此一作答優勢可能受到閱讀測驗閱讀段落篇幅較長的影響而削弱，在紙本上直接閱讀長篇文本對於考生來說較為容易，在電腦螢幕上閱讀長篇文本需要以滑鼠拉動捲軸進行瀏覽。Muter 和 Maurutto(1991)、Muter(1996)的研究也指出在電腦上的閱讀速度會比在印刷紙本上來得慢，不論是一般閱讀或略讀(skimming)(引自 Pommerich, 2004)。這可能導致考生在作答時需要耗費較多時間。

然而整體看來，四等測驗考生在不同施測介面測驗完成的比率是差不多的。

表 2　基礎、進階、高階與流利級測驗紙筆與電腦介面考生完成作答的比例

測驗等級	人數	施測介面	測驗完成比率
基礎級	292	紙筆	95.5%
		電腦	96.2%
進階級	187	紙筆	94.7%
		電腦	97.3%
高階級	143	紙筆	95.8%
		電腦	93.0%
流利級	275	紙筆	96.0%
		電腦	93.8%

二、測驗總分比較

考生在華語文能力測驗(TOCFL)四個等級的紙筆與電腦測驗平均得分如表 3 所示。基礎級測驗，考生在紙筆測驗的測驗總分和閱讀理解測驗平均得分略高於電腦測驗，聽力理解測驗則是電腦測驗表現較佳。進階級測驗和流利級測驗，考生在電腦測驗的測驗總分和聽力理解測驗平均得分略高於紙筆測驗，閱讀理解測驗則是紙筆測驗的表現略佳。至於高階級測驗，測驗總分和聽力及閱讀理解測驗平均得分均為電腦測驗的平均分數略高。

以成對樣本 t 考驗進行平均數差異分析，結果顯示，僅基礎級與進階級測驗聽力理解測驗的平均得分達到顯著差異水準，兩者均顯示考生在電腦測驗的表現優於紙筆測驗。進一步檢視此差異的效果量(effect size)，採用 Cohen(1988)提出的 d 係數，基礎級測驗和進階級測驗的 d 係數分別為 0.065 和 0.149。根據 Cohen 的標準，若其值小於 0.2 表示實際顯著性為低，介於 0.2 至 0.5 表示實際顯著性為低至中等，而 0.5 至 0.8 表示實際顯著性為中至高等，高於 0.8 表示具有相當大的實際顯著性(引自 Kingston, 2009)。本研究得到的 d 值均小於 0.2，表示雖然考生在基礎級測驗和進階級測驗聽力理解測驗的紙筆和電腦測驗成績的平均得分有差異，但效果量均相當小。

此一結果顯示，除了基礎級和進階級測驗聽力成績外，考生在其他測驗等級或測驗項目的表現均不因施測介面的不同而有所差別。平均成績達到顯著的基礎級和進階級聽力測驗，效果量也相當小，表示測驗介面對於成績的影響並不大。

表 3　基礎、進階、高階與流利級測驗紙筆與電腦介面考生測驗成績 t 考驗

測驗等級	施測介面	平均數	標準差	*t* 值
基礎級 (N=292)	電腦總分	64.37	13.302	-0.124$^{n.s.}$
	紙筆總分	64.40	12.750	
	電腦聽力	34.31	5.377	2.335*
	紙筆聽力	33.96	5.423	
	電腦閱讀	30.06	8.906	-1.704$^{n.s.}$
	紙筆閱讀	30.45	8.389	
進階級 (N=187)	電腦總分	76.29	14.262	1.541$^{n.s.}$
	紙筆總分	75.29	14.120	
	電腦聽力	37.91	7.611	2.679**
	紙筆聽力	36.71	8.436	
	電腦閱讀	38.38	7.834	-0.600$^{n.s.}$
	紙筆閱讀	38.58	7.397	
高階級 (N=143)	電腦總分	74.27	12.846	0.543$^{n.s.}$
	紙筆總分	73.88	12.721	
	電腦聽力	39.35	6.398	0.839$^{n.s.}$
	紙筆聽力	38.99	6.242	
	電腦閱讀	34.92	8.245	0.084$^{n.s.}$
	紙筆閱讀	34.89	8.513	
流利級 (N=275)	電腦總分	69.27	16.631	0.666$^{n.s.}$
	紙筆總分	68.87	17.680	
	電腦聽力	36.60	8.661	1.424$^{n.s.}$
	紙筆聽力	36.07	9.288	
	電腦閱讀	32.67	9.442	-0.382$^{n.s.}$
	紙筆閱讀	32.80	9.828	

$^{n.s.}$ $p > 0.05$, * $p < 0.05$, ** $p < 0.01$

三、試題難度參數差異分析

表 4 為華語文能力測驗(TOCFL)基礎、進階、高階以及流利級測驗，聽力與閱讀理解測驗不同施測介面(紙筆、電腦)試題難度參數的皮爾森積差相關分析以及差異比較結果。相關分析結果顯示，四個測驗等級共八個分測驗的紙筆測驗和電腦測驗試題難度參數之間均有高度正相關存在。從圖 1 至圖 4 也可看出，相同試題在不同施測介面估計出的難度參數相當接近，落點均在對角線附近。

試題難度參數差異比較的結果，相差大於 0.5 logit 以上的試題，在基礎級測驗共有 7 題，均為聽力理解測驗；進階級測驗有 16 題，其中 11 題為聽力，5 題為閱讀理解測驗；高階級測驗共有 12 題，聽力和閱讀理解測驗各有 8 題和 4 題；流利級測驗則有 3 題，均為聽力理解測驗。

諮詢會議中，心理計量專家提到部分難度參數差異超過 0.5 logit 之試題，通過率相差不到 5%，可能因試題較容易之故，估計得到的難度參數差異較大，此類試題可以忽略。以此一原則再篩選後，基礎級和流利級測驗難度差異較大試題僅有 2 題，均為聽力題；進階級測驗分別有 2 題聽力題和 4 題閱讀題難度差異較大；高階級測驗難度差異較大的試題共 5 題，2 題為聽力，3 題為閱讀題。

整體來說，試題難度受到介面影響的比例，均不到一成，也就是有 90%以上的試題，不因施測介面為紙筆或電腦而在難度上產生很大的變化。

表 4　基礎、進階、高階與流利級測驗紙筆與電腦介面試題難度相關及差異比較

測驗等級	測驗內容	題數	相關值	難度差異較大題數	扣除通過率相差小於 5%試題後	電腦有利	紙筆有利
基礎級	聽力	40	0.98**	7 (17.5%)	2 (5.0%)	2	0
	閱讀	40	0.98**	0 (0.0%)	0 (0.0%)	0	0
進階級	聽力	50	0.95**	11 (22.0%)	2 (4.0%)	2	0
	閱讀	50	0.95**	5 (10.0%)	4 (8.0%)	3	1
高階級	聽力	50	0.96**	8 (16.0%)	2 (4.0%)	2	0
	閱讀	50	0.97**	4 (8.0%)	3 (6.0%)	1	2
流利級	聽力	50	0.96**	3 (6.0%)	2 (4.0%)	2	0
	閱讀	50	0.99**	0 (0.0%)	0 (0.0%)	0	0

** $p < 0.01$

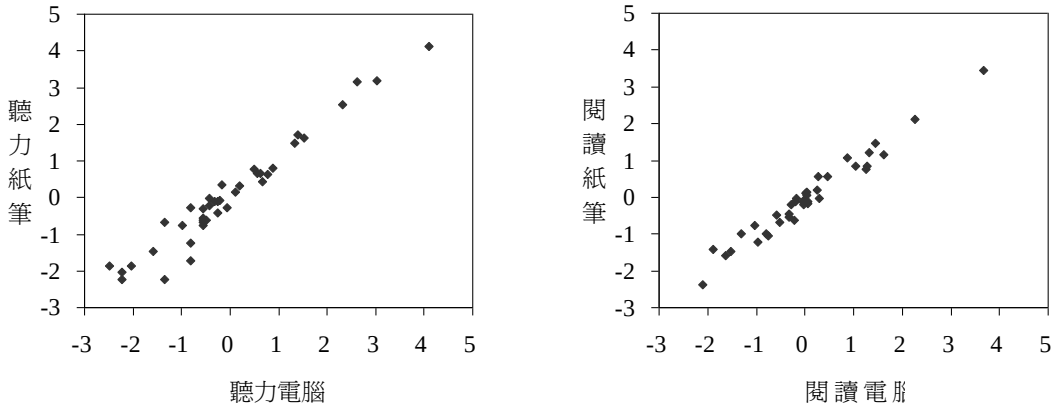

圖 1　基礎級測驗聽力與閱讀理解測驗試題難度參數相關圖

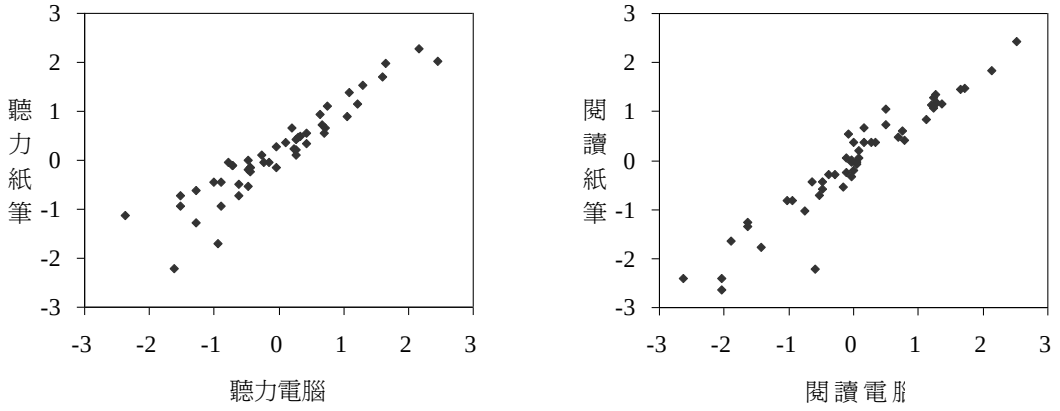

圖 2　進階級測驗聽力與閱讀理解測驗試題難度參數相關圖

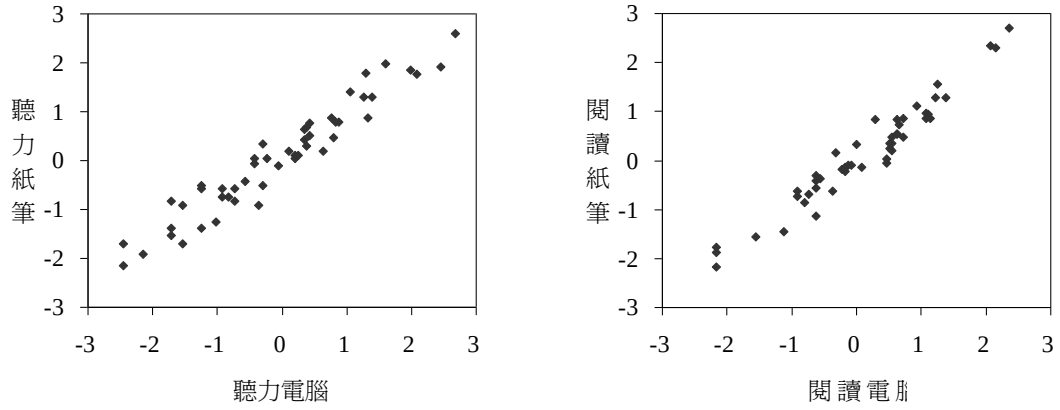

圖 3　高階級測驗聽力與閱讀理解測驗試題難度參數相關圖

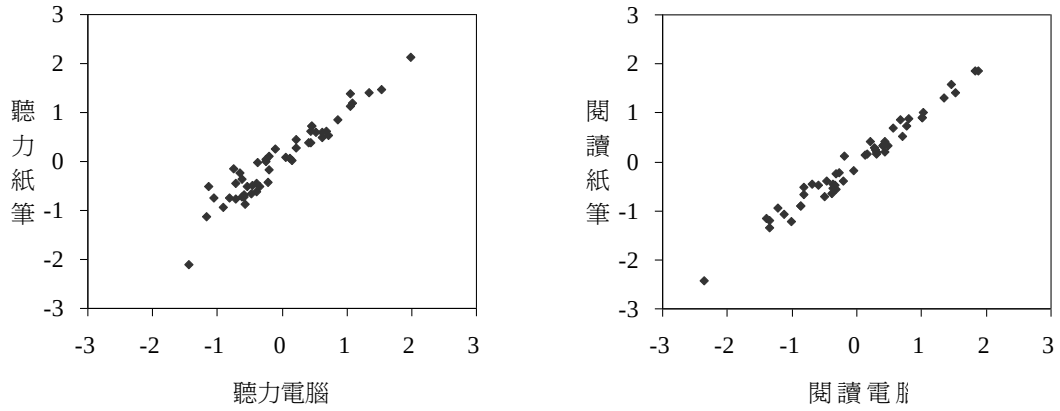

圖 4　流利級測驗聽力與閱讀理解測驗試題難度參數相關圖

研究者認為，測驗平均數和試題難度差異受到介面影響程度很小的原因，應與電腦測驗系統有很大的關係。前面已提及在設計之初，為避免考生需要重新適

應電腦測驗系統，因此規劃在電腦螢幕上的試題呈現盡可能維持與紙筆測驗一致；作答方式上，也增加許多彈性，如閱讀測驗就如同在紙本作答一樣，考生可以選擇跳過某些試題，之後再回頭作答，不確定答案的試題，也可以在系統上自行註記。分析的結果也符合 Spray 等人(1989)所假設的，軟體的彈性是造成介面效應的主因(引自 Mead & Drasgow, 1993)，由於華語文能力測驗電腦介面在操作上具備相當大的彈性，因此考生表現幾乎不因介面不同而受到影響。

四、諮詢會議結果

經過專家的書面以及會議諮詢討論後，研究者整理造成少部分試題難度有介面效應的可能原因包括以下三點：

(一) 聽力設備

紙筆測驗施測方式多半使用教室內公播系統公開播放聲音檔，考生可能由於施測時座位安排的不同，對於某些試題內容有聽不清楚的可能性。而電腦測驗全面使用個人耳機，且考生可以自行操作每一題聲音檔的大小，勢必比紙筆測驗考生聽得更清楚。可能由於聽力設備上的差異，在某些聽力試題上，電腦測驗的答對率高於紙筆測驗。

(二) 圖片較大被切割，段落篇章需拉動捲軸才能閱讀

針對紙筆測驗通過率較高，較電腦測驗容易的閱讀理解測驗試題，專家提出電腦螢幕能呈現的圖片大小有限，在部分試題考生需要點選「放大鏡」功能，才能完整瀏覽圖片進行作答；而紙筆測驗則是可以直接觀看到圖片全貌，因此影響試題的難度。

長篇的閱讀篇章，也因為篇幅較長，考生需要使用滑鼠拉動捲軸以閱讀整篇文章，可能因此影響作答，使得電腦測驗介面的試題比紙筆測驗來得困難。

此一看法與 Bridgeman 等人(2003)、Choi 和 Tinkler(2002)、Bergstrom(1992)的研究一致，試題訊息無法完整在螢幕呈現，考生需要操作滑鼠捲軸以瀏覽試題全貌時，通常會有顯著的介面效應(引自 Pommerich, 2004)。

(三) 注意段落切割處是否恰巧為題目重要線索

進階級閱讀理解測驗有一題短文理解，因篇幅較長被切割，恰巧螢幕所顯示的最後一句為第一個子題答題的關鍵，可能提供了考生重要線索，電腦測驗介面考生無需從第一段開始閱讀，即可快速找到標準答案，因而造成電腦測驗比閱讀測驗簡單。

除了上述三項可能因素外，尚有四題閱讀理解測驗試題，經由專家討論過後，仍無法提出可能的原因，有三題為在電腦測驗介面較為有利，一題為在紙筆測驗較為有利。

結論與建議

在測驗總分層次，本研究以 t 考驗分析基礎、進階、高階以及流利級測驗，考生在紙筆和電腦介面的測驗總分、聽力與閱讀理解測驗得分差異，結果僅基礎和進階級的聽力理解測驗得分有顯著差異，考生於其他測驗等級或測驗項目的表現均不因施測介面不同而有所差別。即使是平均成績差異達到顯著的基礎級和進階級聽力測驗，效果量也相當小，顯示測驗介面對於考生成績的影響很小。

在試題難度層次，基礎級與流利級測驗難度參數差異較大的試題僅有 2 題，均為聽力題；進階級測驗分別有 2 題聽力題和 4 題閱讀題難度參數差異較大；高階級測驗難度參數差異較大的試題共 5 題，其中 2 題為聽力，3 題為閱讀題。整體來說，試題難度受到介面影響的比例，均不到一成。相關分析結果亦顯示，四個測驗等級共八個分測驗的紙筆測驗和電腦測驗試題難度參數間有高度正相關存在，相關係數均在 0.95 以上。

從測驗總分和試題難度兩個層次的分析結果，可以看出華語文能力測驗(TOCFL)電腦與紙筆介面之間具有可比性，考生在兩種測驗介面的成績表現幾乎沒有不同，也僅有少數試題在不同介面的難度參數有所變動。

而諮詢會議的結果，專家討論出造成試題難度介面效應的可能原因為：聽力設備、圖片較大與閱讀篇章較長，以及段落切割點，並建議電腦介面的邊框(邊界)可以再調整，增加試題訊息可呈現的範圍，以減少考生使用滑鼠拖曳或放大鏡功能的必要性。未來本會研發人員也會依據諮詢結果，進一步修正命題方向與電腦施測介面，以及提升紙筆測驗的錄音檔案播放設備，確保參與測驗考生的成績不受介面影響，維持本測驗之公平性。

參考文獻

一、中文文獻

99 年個人家戶數位落差調查報告(民 99 年 11 月)。台北市：行政院研究發展考核委員會。民 100 年 9 月 22 日，取自：http://www.rdec.gov.tw/public/Attachment/0121411543371.pdf

二、英文文獻

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.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational computing research, 31*(1), 51-60.

Choi, I. C., 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.

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.

International Telecommunications Union. (2010). *The World in 2010: ICT Facts and Figures*. Available form International Telecommunications Union Web site, http:// www.itu.int/ITU-D/ict/

Kim, D. H., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*(4), 554-570.

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.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352-1375.

Parshall, C. G., & Kromrey, J. D. (1993). Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED363272). Retrieved from Education Resources Information Center [ERIC] Web site: http://www.eric.ed.gov

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2*(6). Available from http://www.jtla.org

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *The Journal of Technology, Learning, and Assessment, 5*(7). Available from http://www.jtla.org