學生對電腦化測試的適應

梁成安

引言

隨著電腦科技的進步，教育上有很多的發展都繼而衍生出來。其中一項就是教育的測試及評估，傳統的教育測試及評估都是以紙筆測試爲主；相反，電腦化測試(Computerized Adaptive Test)無論在擬題、記錄及更新資料、即時匯報等各方面都能提供更多的彈性和創新。電腦化測試的測試時間通常比較短，不但可以即時紀錄資料，還可以即時匯報結果，這些都是明顯的好處。重要的是，電腦程式有互動和動態的功能，這些功能令電腦化測試能用較少的多項選擇題去測試學生，而得出差不多的結果。但另一方面，這些創新衍生另一些問題，就是測試的可靠性、內容、估計學生能力誤差等的問題(Drasgow and Olson-Buchanan 1999; Sands, Waters and McBride 1997; van der Linden and Glas 2000; Wainer 2000)。

在進行電腦化測試時，多項選擇題中每一題的難度都不同，而不同學生的能力亦有差異，題目反應理論（Item Response Theory）則能提供一個分析題目的基礎；與此同時，電腦能把每一個學生在電腦化測試中的每一項題目的反應時間都記錄下來，這比傳統的紙筆測試提供了更多的資料(Hambleton, Swaminathan & Rogers 1991)。

然而，在澳門進行電腦化測試的有關工作尚未完善，今次在澳門進行電腦化測試的研究則可爲未來吸取經驗。其中小五年級的語文科被選定作爲今次的研究項目，這樣便能避免因升讀中一時而要面對多一重壓力。

研究方法及步驟：

第一階段：建構題目

這次在澳門進行測試的科目是小學五年級的語文科，兩位研究助理在研究員的指導下設計題目，題目是從一些慣用的練習簿修改而來的，再諮詢現職小學老師，經過預試後再修改而成，具有一定內容效度，能測試學生中文語文能力。

整份卷共有 170 題多項選擇題，然後分成甲乙兩卷，每卷 85 題，兩卷的規格和難度都相同，即兩份卷爲「平行測卷」，題目設計由淺到深排列。

檢定測試在二〇〇一年三月於澳門的六間學校進行，甲乙兩卷於班房內以梅花間竹的進行分配，所以，最終甲乙兩卷都當能由兩批性質相同的學生所完成。

爲了連結甲乙兩卷的難易度，所以就設立丙卷，丙卷共有 18 題，其中 9 題選自甲卷，而另外 9 題選自乙卷，這個設計被稱爲「共同試題等化」（common item equating）。丙卷於同年六月十四日在一間小學裏進行。最後數據包括 170 條題目及 779 名學生，其中少於百分之一是遺漏數據（missing values）。

第二階段：題目的選擇及軟件開發

研究員運用 XCALIBRATE 軟件分析第一階段所收集得來的數據，統計的項目包括有困難度、區別度和題目總分相關系數。

經過題目分析後，其中有六條題目的題目總分相關系數爲負，於是它們在將來的研究中會被撇除，剩下來的 164 條題目會繼續以淺到深的排列，依據這個排列，會順序的每三題抽一題，共抽出 54 題，把這 54 條題目設定爲紙筆試卷(PPT)，並於下一個階段進行紙筆測試。

被選取的54條題目定爲紙筆試卷(PPT)，而剩下來的 110 條題目就用來進行電腦化測試(CAT)，這項研究的其中一個目的是比較 (PPT) 和 (CAT) 兩份平行卷。CAT的運作方式如下，在某一時段內被選取出來作測試的題目乃取決於之前所進行的測試結果，如果一名學生成功答對某一條問題，他（或她）會繼續有一條更難的問題，相反，如果一名學生答錯某一條問題，他（或她）會繼續有一條更容易的問題。但是，學生能否答對題目，除了本身的能力外，還有一些隨機的因素。因此，測試以五題一組來更替一題，以減低隨機性。由於在澳門進行一個完全的CAT亦有一定的技術困難，所以設計會有適當的調整。

整個CAT測試過程總共有四個步驟，每一步驟包含五條題目。第一步，從由淺到深排列 110 條題目，然後每 22 題就抽出一題，共抽出 5 題，這 5 題題目代表很難、很易及中等困難的題目。第一步完成後，每條題目答對得一分，得分分別爲 0 至 5 分，而剩下來的 105 題題目會作爲隨後之用。第二步：如第一步般從 105 條題目中選出 10 題（1 至 10），當一名學生在第一步得 0 分，第二步就測試第 1 至 5 條，若在第一步取 1 分，第二步就測試 2 至 6 題，如此類推，這個過程在步驟三和步驟四都重覆。

電腦化測試軟件在中文視窗 98 中開發，並沿用Authorware教育軟件爲基礎。畫面設計盡量爲學生舒服而設，在正式測試之前有三題熱身題，不計分而且容易，以便令學生容易投入；這些熱身題並不會影響隨後在正式電腦化測試的題目選擇。

第三階段：預試和主要測試

在正式測試之前，首先請幾位同學進行一個小規模的預試，當中學生產生的問題都要處理，例如：學生感到不適、疲倦、硬件不配合等。主要測試在兩間學校中進行，每個小五學生都要進行兩個測試：一份紙筆試卷(PPT)和一個電腦化測試(CAT)，兩個測試中都標明學生的姓名和學號，以便可以連結兩個測試的數據。研究助理計算學生做紙筆測試（PPT）的完成時間，大概爲 15 至 20 分鐘；

至於CAT，電腦會自動爲每一位學生記錄時間，而且大多數學生都能在 9 分鐘內完成。

爲有效推行電腦化測試，所有學生在學校內的電腦室同步地進行測試。軟件安裝在學校的內聯網，而且每位學生的位置都要清楚，以便連結紙筆測試的資料。

所有收集得來的數據會儲存在CD-ROM以策安全，並且作備份及日後再分析之用，今次的經驗和數據可作爲日後在澳門進行同類研究的參考。

結果：

每個進行PPT和CAT的學生都有原始分（答對題數）和IRT分（題目反應理論得分）。表一提供綜合的統計數字。在CAT中，每位學生都會測試首5題，所以匯報這5題的分數以作比較。

表一：PPT和CAT中原始得分與IRT得分的統計資料

	平均數	中位數	標準差	第一四份位數	第三四份位數
PPT 原始分數	38.76	40.00	7.54	34.00	45.00
PPT IRT 分數	-0.49	-0.55	1.02	-1.17	0.23
CAT 首 5 題原始分數	3.60	4.00	0.90	3.00	4.00
CAT 首 5 題 IRT 分數	0.49	-0.08	2.97	-1.52	0.55
CAT 20 題原始分數	11.07	11.00	3.11	9.00	12.00
CAT 20 題 IRT 分數	-1.70	-1.76	1.53	-2.84	-0.76

在表一中，54 題PPT答對題目的平均數是 38.76，正確率是 0.71(=38.76/54) ，即十條中大約有七條答對，換言之，這是一個比較易的測試；把首 5 題CAT作出相同的計算都得出大概相同的結果，正確率是 0.72(=3.6/5)；然而，把20題CAT作出相同的計算則得 0.55(=11.07/20)。這個結果是正常的，因爲CAT的題目相對地比起PPT的題目更接近學生的能力，如果學生們感到PPT容易，那麼他們會發覺CAT比較難。再這，0.55 的正確率顯示學生的答對題目的機會率是一半半，換句話說，這些題目比較接近他們的能力。

表二匯報PPT和CAT兩個得分的相關系數。就如預期所想的，無論在PPT或是CAT兩個測試中，原始得分和IRT得分兩者之間都有很高的相關系數，因爲原始得分和IRT得分只是在同一測試中評估學生能力的不同評分方式，即是說，在同一個測試中，原始分數和IRT分得出的結果相差不大。但是，無論是原始得分或是IRT得分，PPT和CAT之間對學生能力的評估的相關系數大約在 0.6 之間，數字明顯地比 0 大，但與 1 仍有一段距離，這表示PPT和CAT兩個的能力得分是正相關，在PPT得分越高，在CAT得分越高，但不能作預測之用。因此，在一個測試中取得高分的，在另一個測試中亦能取得高分；可是，我們不能以一個測

試的得分去預測另一個測試的得分。所以，PPT和CAT之間有一個基本的分別，在現時發展階段，CAT暫時未能代替PPT。

表二：匯報PPT和CAT的相關系數

	PPT 原始得分	PPT IRT 得分	CAT 20 題原始得分	CAT 20 題 IRT 得分
PPT 原始得分	1.000	-	-	-
PPT IRT 得分	0.975	1.000	-	-
CAT 20 題原始得分	0.593	0.611	1.000	-
CAT 20 題 IRT 得分	0.631	0.646	0.937	1.000

因爲在進行CAT時，不是每一位學生都做相同的題目，所以PPT中原始得分和IRT得分的相關系數都比CAT的高(0.975比0.937)，能力比較高的學生做較深的題目，能力較低的同學做較易的題目。因此在理論上，與IRT得分相比較之下，原始得分實際不能反映學生的能力；另外，因爲這個原因，CAT的IRT得分應該較接近真實分數，所以，它比CAT的原始分與PPT的原始分和IRT得分有更高的相關系數。

結論

在澳門教育界中，紙筆測試仍然是常用的測試方式，電腦化測試是一個用來測試學生能力的新方法。因此，這次經驗可作爲日後的參考，今次所得的數據可作爲將來研究和發展的資料。今次的結論得出PPT和CAT的相關系數非常顯著地大於零，表示在一個測試中有高成績，再另一個測試中也會有高的成績。但另一方面，相關系數雖然大於零，但和１亦有距離，表示在一個測試中所取的成績未能準確地預測在另一個測試的分數，這顯示到用電腦測試和用紙筆測試在某程度上有一點不同。另外，我們有記錄每位學生每題的反應時間，現時尚未用來進行分析，但可作日後研究之用。

註：本研究計劃獲澳門大學研究委員會資助，作者在此特意致謝。

參考資料

Drasgow, Fritz and Olson-Buchanan, Julie B. (1999). *Innovations in computerized assessment*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates, Publishers.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Sands, William A., Waters, Brian K., & McBride, James R. (1997). *Computerized*

adaptive testing: From inquiry to operation. Washington DC: American Psychological Association.

van der Linden, Wim J. & Glas, Cees A.W. (2000). *Computerized adaptive testing:Theory and practice*. Dordrecht, The Netherlands: Kluwer.

Wainer, Howard (2000). *Computerized adaptive testing: A primer* (Second Edition). Mahwah NJ: Erlbaum.