行政院國家科學委員會專題研究計畫成果報告

電腦適性測驗試題曝光率和能力估計之相關研究
Item Exposure Control and Ability Estimation in Computerized Adaptive Testing

計畫編號：NSC 90-2118-M-018-002
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：李信宏　彰化師範大學數學系
E'mail: li@math.ncue.edu.tw
計畫參與人員：陳建榮　彰化師範大學數學系

一、中文摘要

試題曝光率和測驗重複率的控制是電腦適性測驗不可缺少的環節。所謂試題曝光率是指試題被重複使用的頻率，而測驗重複率是指兩個或兩個以上的試題同時出現的次數。有效地控制試題被重複選取的機會，可以避免試題因為過度“曝光”而影響到測驗的保密性及安全性，進而達到確保測驗公平的要求。目前已知有數種方法證明能夠將試題曝光率控制在預設值之下，並且同時產生最低的測驗重覆率，例如：the Sympson and Hetter procedure，the Davey and Parshall method，以及 the Stocking and Lewis conditional multinomial procedure等。然而在另一方面，應用試題曝光率之時也會增加考生能力估計值的誤差，這是因為選題受到曝光率限制，無法選到“最佳”試題所必須付出的代價。本研究主要是計畫在實施試題曝光率控制的前提下，配合Fisher information的選題方式，討論WLE和OWEN + MLE兩種方法在電腦適性測驗（CAT）估計能力的表現。研究中考慮三個不同試題特性的題庫，分別觀察二者估計受試者能力的情形，並且以偏誤(bias)的平均值和RMSE來評估二者的表現。根據模擬研究的結果顯示：在三種形式的題庫中，OWEN + MLE產生的偏誤之平均值幾乎都比WLE所產生的更接近0；在RMSE方面，OWEN + MLE也比WLE稍微小，但是二者差距是否達到統計顯著尚待進一步檢驗。另一方面，對於每位受試者固定施測30題的過程中，OWEN + MLE 總是比較快速地得到穩定的能力估計值，可見OWEN+MLE在能力估計方面確實比WLE更有效率。綜合這些模擬研究的結果，就試題曝光控制下的CAT而言，使用OWEN + MLE進行能力估計比使用WLE更有效率，準確度也高，所以較符合電腦適性測驗所欲達成的目標，也就是在兼顧測驗安全及公平的原則之下，也能夠提高估計考生能力的精確度。

關鍵詞：試題曝光率 測驗重複率 電腦適性測驗 偏誤 WLE OWEN + MLE RMSE

Abstract

The topic of controlling item exposure rate and test overlap rate has attracted many researchers in recent years. To enhance test security in computerized adaptive testing (CAT), the goal is to have as little overlap as possible between sets of items that are administrated at several different test forms. The general approach to reducing the usage of some frequently appearing items is called “exposure control”. Currently, a number of procedures have been designed for controlling the item exposure rate to a desired value that is specified in advanced of testing. For example, the Sympson and Hetter procedure (Sympson ＆Hetter, 1985), the Davey and

Parshall method (Davey & Parshall,1995), and the Stocking and Lewis conditional multinomial procedure (Stocking & Lewis, 1995). However, when algorithms are utilized to control item exposure, the measurement precision of examinees' ability level is reduced because the most informative items are always not selected. That is, there exits a trade-off problem between trait estimation and item exposure in CAT. For this purpose, estimation techniques based on weighted MLE (WLE) and OWEN + MLE are proposed and carefully studied with the usage of item exposure control in adaptive testing. A simulation study is carried on to investigate the performance of proposed methods. The bias and RMSE (root mean square error) are major criteria used to evaluate the procedures.

Keywords: **Item Exposure Rate, Test Overlap Rate, Computerized Adaptive Testing (CAT), bias, WLE, OWEN + MLE, RMSE**

二、計畫緣由與目的

控制試題曝光率是保護測驗安全的必要措施，但是同時也要付出能力估計值精確度降低的代價。不論電腦適性測驗的選題方式為何，其目的乃是希望能選擇最有效地、精確地估計考生能力的試題，提供施測；然而加入了曝光率的限制以後，選題的結果就很有可能不會選到“最佳”的試題，因而造成能力估計的偏誤。Chang（1998）以五種試題曝光率的控制方法，配合 maximum information procedure 選題，以及其他因素（如：題庫大小、測驗長度和考生人數等）進行大規模的模擬研究；而在能力估計方面，則是先應用OWEN於最初的一些試題，最後在使用MLE。其研究結果顯示五種試題曝光率的控制方法都會影響能力估計的精確度（與沒有實施任何曝光率控制相比），而其中以 the Stocking and Lewis conditional multinomial procedure (S & L-C)所呈現的偏誤與均方差(mean square error)較小。另一方面，Warm (1989)在其所提出WLE（weighted likelihood estimate）估計方法中指出，使用WLE能夠比使用常態先驗分配的貝氏估計法或最大概似估計法，產生比較小的偏誤、標準差以及均方差；而且，使用WLE的電腦適性測驗只需要比較少的測驗試題，就能達到測驗終止的標準。此外，Cheng and Liou（2000）則是在電腦適性測驗的相關研究中，以MLE 和 WLE 兩種方法配合 Fisher information、Kullback-Leibler information 和 optimal item difficulty 等選題方法進行模擬研究。結果顯示以WLE 和 Fisher information的組合所得θ之估計值有較小的誤差，不過此研究並未包含試題曝光率的控制在內。基於上述Chang, Warm, Cheng & Liou 對WLE的相關研究討論，本研究的主要目的在於：藉由實施S & L-C曝光控制的電腦適性測驗，比較二種能力估計方法OWEN + MLE和WLE的準確性。研究中採用偏誤及RMSE（root mean square error）作為判斷估計方法優劣之準則。

三、結果與討論

研究中考慮三種不同形式的題庫: real item pool、ideal-a item pool與ideal-b item pool，分別模擬作答結果以觀察能力估計方法OWEN+MLE和WLE的表現。題庫大小分為150題及300題兩種，而評比的標準則是以偏誤平均值、RMSE以及偏誤和RMSE的收斂情形四個角度，來比較二種能力估計方法的優劣。

首先針對real item pool所模擬的題庫研究，在大型測驗（300題）中CAT施測的結果，WLE和OWEN + MLE兩種能力估計方法分別在7個能力點上得到1000筆偏誤的平均值。除了θ=-1.0之外，其餘的能力點部分，OWEN + MLE產生的偏誤比WLE產生的偏誤更接近0，顯示OWEN + MLE的估計較不具偏誤。不過，兩者在低能力水準處具有高估（overestimate）的情形，至於低估（underestimate）的狀況則出現於高能力水準處。對WLE而言，此一現象也許可以歸因於能力估計值在能力水準二端

處校正效果較差所導致；此外，由於貝氏估計法具有向平均數迴歸的現象(Weiss & McBride 1984)，所以不難理解OWEN + MLE為什麼也會有發生高估及低估的情形。至於RMSE方面，在介於-1.5和1.5之間的能力點上，OWEN +MLE產生的RMSE，明顯地WLE產生的RMSE微小。另外研究結果也顯示OWEN+MLE產生的偏誤比較快速地得到穩定的能力估計值，意即OWEN + MLE只需要比較短的測驗長度就能獲得穩定的能力估計值進而縮短施測時間及節省測驗成本。同時，在整個電腦適性測驗的過程中（從第1題到第30題），在大部分的能力點上，OWEN + MLE產生的偏誤，都比WLE產生的偏誤更靠近0。然而，當θ=-1.5和-1.0時，OWEN + MLE得到最後能力估計值的偏誤，具有突然增大的現象，可能是因為加入試題曝光控制的條件之後，使得被施測的試題未必具有最大訊息量，最後導致校正MLE之效果不彰。至於以小型測驗去模擬CAT施測，其研究結果與前述是類似的。

歸納而言，比較OWEN + MLE和WLE在試題數不同的測驗中之表現。首先這兩種能力估計方法都具有高估低能力及低估高能力的現象，並不會因為改變測驗長度而有所不同。所以即使變動題庫大小，這兩種估計方法優劣表現也是一致的。其次以估計值的的收斂情形來看，不管是以偏誤或者RMSE為判斷指標，OWEN + MLE在大題庫及小題庫中幾乎都以很快地達到穩定的值，明顯地較WLE來得優勢。不過，OWEN + MLE的偏誤在估計的最後階段有突然增加的現象，這在小題庫中出現的次數較多，可見在施測CAT中使用較大型的題庫似乎可以提高能力估計的準確性。

本研究接著以ideal-a item pool與ideal -b item pool為模擬題庫之依據，結果加以分析之後，這三種形式的題庫之模擬研究都一再顯示了：OWEN + MLE的表現比WLE更為出色。比較特別的是，WLE和OWEN + MLE都會在能力水準兩端處出現高估及低估的現象，但是採用第三種形式的題庫（亦即ideal-b item pool）進行模擬CAT施測，能夠讓OWEN + MLE進行估計能力時有效地減少偏誤，進而緩和向平均數迴歸的態勢，尤其是使用規模越大的題庫效果愈明顯。至於可能是由於校正MLE效果失效而在測驗結束之後出現突然增大的現象，顯然在第二種形式的題庫（亦即ideal-a item pool）中比較嚴重，因為在第一種形式的題庫（亦即real item pool）中不存在RMSE忽然加大的情形，而在第三種形式的題庫（亦即ideal-b item pool）中則缺乏偏誤具有突然增加的狀況。根據以上模擬研究的分析與討論，顯示OWEN + MLE在電腦適性測驗估計能力方面，比WLE具有較好的效率與準確度。

四、計劃成果自評

在尚未證實OWEN + MLE產生的偏誤和RMSE均**顯著地**低於WLE之前，不宜斷定OWEN + MLE的能力估計準確性優於WLE，但是模擬研究結果發現OWEN + MLE的確比MLE較有效率。因此當測驗效率成為最主要關心的議題時，如果能夠克服界定受試者能力先驗分配的困難，則可以選擇OWEN + MLE做為估計受試者能力的方法。

五、參考文獻

[1] Chang, S. (1998). *A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

[2] Cheng, P.E. & Liou, M. (2000). Estimation of Trait Level in Computerized Adaptive Testing. *Applied Psychological Measurement*, 24 (3), 257-265.

[3] Davey, T., & Parshall, C. G. (1995). *New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

[4] Hetter, R. D., & Sympson, J. B. (1997). Item Exposure Control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, DC: American Psychological Association.

[5] Owen, R.J. (1975). A Bayesian Sequential

Procedure for Quantal Response in the Context of Adaptive Mental Testing, *Journal of the American Statistical Association*, 70(350), 351-356.

[6] Sympson, J. B., & Hetter, R. D. (1985). *Controlling Item Exposure Rate in Computerized Adaptive Testing*. Paper presented at the annual meeting of the Military Testing Association. San Diego, CA: Navy Personal Research and Development Center.

[7] Warm, T.A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, 54, 427-450.

[8] Weiss, D.J. & McBride, J.R. (1984). Bias and Information of Bayesian Adaptive Testing. *Applied Psychological Measurement*, 8(3), 273-285.