行政院國家科學委員會專題研究計畫　成果報告

電腦適性測驗試題參數即時(線上)估計之相關研究

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-018-003-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立彰化師範大學數學系暨研究所

計畫主持人：李信宏

計畫參與人員：黃珮漩

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 10 月 31 日

中文摘要

在電腦適性測驗（Computerized Adaptive Testing, CAT）整個施測過程中，如果某些試題被過度選用，就有可能造成試題外流，所以通常必須藉由試題曝光控制方法來刪除使用頻率過高的題目，以確保一個適用的題庫。不過隨著測驗的進行，題庫中可用的試題數會愈來愈少，因此必須即時加入新的試題，同時也要即時估計新試題的試題參數，讓新題目的參數和原題庫的參數維持在同一量尺上。在電腦適性測驗中，即時使用受試者的作答結果來更新原題庫的試題參數，或者作為新增加試題的參數估計之用，這個過程稱為即時（線上）估計。即時估計的最大優點乃是藉由新作答資料持續不斷的加入，可以增進試題參數的估計準確性，意即降低估計誤差。當測驗的答題結果是完整而且固定時，可以利用最大概似估計法等來估計試題參數。目前常用的一些軟體，例如：BILOG 和 LOGISTS 等即是如此。當考慮進行即時（線上）估計時，因為受制於有限的作答資料，上述方法可能導致參數估計值有偏差或者有較大的估計誤差，甚至無法獲得穩定收斂的估計值。本研究計畫提出以貝氏估計為核心的 Markov Chain Monte Carlo（MCMC）方法運用到試題參數即時估計。MCMC 是一種結合了馬可夫鏈和 Monte Carlo 積分兩個步驟的估計方法，首先應用 Metropolis-Hastings algorithm 生成一個平穩分配為所欲估計參數之後驗分配的馬可夫鏈，再依據 Monte Carlo 積分，利用樣本平均數去逼近母體平均數以得到最後的估計值。本計畫將透過大量模擬研究的施行，分別生成原始題庫以及新題目題庫提供施測，先利用原始題庫中的題目估計受試者能力，再利用得到的能力估計值與受試者對新題目作答情形，即時估計新題目的試題參數，其中關於能力之設定又分為三種情況來討論 MCMC 的表現。研究中是以模擬和估計的試題特徵曲線間之差異來決定 MCMC 估計的精確度，並討論在哪些試題參數條件下，MCMC 會有較佳的表現。本研究希望在預試題庫較小以及受試人數較少的情況之下，實施線上即時估計仍然能獲得穩定有效的參數估計值。

關鍵詞：電腦適性測驗、試題曝光控制、即時參數估計、Markov chain Monte Carlo（MCMC）、Metropolis-Hastings algorithm、試題特徵曲線。

Abstract

In computerized adaptive testing, updating item parameters using adaptive testing data is generally called on-line calibration. Currently, a number of packages have been developed for calibrating item parameters when the response data are fixed. For example, BILOG (Mislevy & Bock, 1982) and LOGISTS (Wingersky, Barton, & Lord, 1982). These procedures are based upon the maximum likelihood estimation. However, in the early stage of on-line calibration, there is too little information in the testing data to estimate all parameters. The application of those packages tends to produce large bias and standard error in item parameter estimates. This research project utilizes Bayesian procedure - Markov chain Monte Carlo（MCMC）as new methods for on-line item calibration. A full-scale simulation study is designed to investigate the performance of MCMC, especially for the cases with small size of item bank and small number of examinees. The standard errors of estimates as well as the distance between estimated ICCs and true ICCs are the major criteria used to evaluate the procedure. It hopes that new approaches will perform well in on-line item calibration for computerized adaptive testing.

Keywords: computerized adaptive testing, on-line calibration, Bayesian estimation, MCMC

一、前言

電腦適性測驗（computerized adaptive testing, CAT）的施測過程，簡略地說，首先是從題庫中挑選一道難易適中的題目給考生，再依考生答對該題與否的結果估計其能力，接著再根據某種選題方式，由電腦選擇下一道最“適當”的試題以供施測，再次依照作答情形來估計考生能力，然後依選題、施測、估計等方式重復進行測驗，直到某種終止測驗的準則達到為止。其中試題的來源—題庫（item bank）之建立，乃是先經過預試（pre-test）階段，收集部分考生的作答資料，然後估計所有試題的相關參數，如：鑑別度參數、難度參數和猜測參數等，再依測驗目標、試題內容、有無 DIF 等其他指標篩檢後而形成。所以，在實施 CAT 之前，題庫已建立完備隨時可用，而其中試題的參數都是已知且固定不變的。

由上述可知，電腦適性測驗的實施，需要不斷選擇難易度大致符合受試者能力的試題來施測，因此在整個測驗過程中，有些題目就有可能被多次重複地挑選出來，以供不同的受試者作答。而這些試題如果被過度選用，就可能造成試題外流，進而影響到測驗的安全性與公平性。所以藉由試題曝光控制方法（item exposure control）來刪除使用頻率過高的題目，乃是維護題庫相當重要的一環。不過從另一方面而言，在控制試題曝光率之後，隨著測驗的進行，題庫中可用的試題數目會愈來愈少，因此必須即時補充加入新的試題，同時即時估計（on-line calibration）新試題的試題參數，才能確保測驗順利進行。另外，由於新題目和題庫中原有題目施測所得到的能力估計值必須維持在同一量尺上，所以新題目的參數估計值也必須和原題庫試題的參數值保持在同一量尺上，因此即時參數估計可說是一個重要的課題。

所謂即時（線上）估計（on-line calibration）是指在CAT施測過程當中，將受試者的每一題作答結果，即時地應用參數估計方法，重新更新每一道試題的所有參數。On-line calibration的優點乃是藉由持續不斷加入的新作答資料，可以增進試題參數估計的準確性，獲得較穩定的參數估計值，意即降低估計誤差。此外，從CAT選題的角度而言，以最大訊息量（maximum information）為例，試題訊息（item information）是由能力和試題參數所共同決定的函數值。在one-, and two-parameter logistic models中，試題訊息的最大值發生於當能力估計值等於該試題的難度參數時。因此經過on-line calibration之後，CAT選題必會受到影響。同理，目前常用的試題曝光控制方法（item exposure control）也常使用到試題訊息及參數來建立試題曝光參數，因此，改變試題參數值也會同時改變試題曝光率的計算。歸納而言，即時（線上）估計不但得以更新較合適的試題參數，對於電腦適性測驗的其他施測環節也會有所影響。

二、研究目的

現階段已有一些關於試題參數即時估計的方法，例如：Multiple-EM Cycle Method（Ban et al., 2000）、Stocking’s Method A 和 Stocking‘s Method B（Stocking, 1988）等。這些方法中，有些估計誤差大，有些則需要較長的測驗長度或者較大的受試者樣本。另外，由文獻中可知，目前也有不少研究將 MCMC 應用於傳統紙筆測驗來估計試題參數，結果發現 MCMC 在試題作答理論上的應用有優於其他估計方法之處（Patz & Junker, 1999），尤其是當模式趨於複雜或者作答資料不完全時時，使用 MCMC 在計算上會較容易。基於以上論點，因此考慮應用 MCMC 方法到即時參數估計上，以解決不易計算的問題。本研究主要目的是針對兩種不同生成方式的題庫，各搭配兩種不同的建議分配，再以三種不同能力估計方式，討論 MCMC 用在即時參數估計上的收斂情形以及參數估計的準確性，研究中以 weighted

integrated squared error (WISE)作為判斷估計方法優劣的準則。

三、研究方法

本計畫將進行模擬研究，藉由各種不同條件來製造電腦適性測驗的情境，以獲得受試者的作答資料並進行試題參數之估計。模擬研究所得參數估計值將和實際的參數值比較，以了解 MCMC 在 on-line calibration 方面的誤差和其它性質。以下僅簡略說明模擬 CAT 設計時所需要考慮的因素。

1. 試題作答模式：使用 three-parameter logistic model 模擬作答結果。
2. 題庫：規模大小分為兩種，首先是當做電腦適性測驗之用的原始題庫（operational items），此題庫共有 360 題，其中的試題參數值皆為已知；另有一個非電腦適性測驗題庫，而是試題參數待估計題目的題庫（pretest items），其試題題數設計為 20 題和 60 題兩種. 為了模擬真實測驗情形, 兩種題庫的試題參數以 ACT-Math（Chang, 1998）的試題參數估計值模擬生成, 其生成方式也分成兩類, 第一種題庫的鑑別度參數 a 以較常見的常態分配 $N(1.018, 0.108)$ 生成，難度參數 b 以 $N(0.193, 1.073)$ 生成，猜測參數 c 則以 $Beta(4,18)$ 生成，第二種題庫鑑別度參數 a 以 $lognormal(\log(1.018), 0.108)$ 生成，難度參數 b 與猜測參數 c 的生成方法則與第一種方式相同。研究中也特別注意其鑑別度和難度參數的分佈情形。
3. 受試者樣本：假設考生的能力服從標準常態分配，其樣本大小為 3000。
4. 能力估計法：每次作答後使用 Expected a posterior estimation (EAP, Bock & Aitken, 1981)估計受試者能力，測驗終止後再以 MLE 估計。
5. 選題：根據 maximum Fisher information 選取下一道被施測的題目。
6. 測驗長度：原始題庫選 30 題作答，試題參數待估計題目的題庫則選擇 10 題作答。
7. 參數比較準則：計算 estimated ICC 和模擬 true ICC 之間的差距, 以 WISE 表示如下：

$$\text{WISE}=\int [P(\theta \mid a,b,c) - P(\theta \mid \hat{a},\hat{b},\hat{c})]^2 \, w(\theta) d\theta$$

，其中 $w(\theta)$ 是標準常態分配 $N(0,1)$ 的權重。實際計算時則將能力值限制在 $(-4, 4)$ 之間，以數值方法求取積分。

至於在 MCMC 的設定部份，本研究中將討論三種不同的能力值設定方法，其目的想瞭解三種不一樣的能力假設方式，對整個即時參數估計結果的影響。第一種是參考 Stocking's method A 的想法，直接將電腦適性測驗中得到的受試者能力估計值視作真實能力值，代入 MCMC 過程，也就是說只使用 MCMC 估計試題參數。第二種方法是在得到能力估計值後，以此能力估計值做為 MCMC 中受試者能力的先驗分配之期望值, 利用受試者所有的作答結果（包括原始題庫和新題庫裡的題目），同時估計受試者能力（重新再估計）和新題庫的試題參數。第三種方法源自於 BILOG/Prior 與第二種方法雷同，但是將受試者能力的先驗分配之變異數設的很小（strong prior）。因此能力估計值會侷限在先前電腦適性測驗得到的能力估計值附近。此法充分利用電腦適性測驗裡得到的能力估計值，同樣以受試者所有的作答結果（包括原始題庫和新題庫裡的題目），同時估計受試者能力和新題庫的試題參數。

此外，由 MCMC 的理論可知建議分配對整個 MCMC 的估計結果以及收斂情形可能有顯著的影響, 因此本研究中針對產生方式相異的兩種題庫, 各選擇不一樣的建議分配搭配，其目的乃是想瞭解建議分配的影響程度。 最後, 則是決定 MCMC 抽取樣本的樣本數大小，經過許多次的測試，大約 3000 次可以達到收斂的效果，所以把運算次數設為 3000 次，而取後 1000 次的平均值做為估計之用。

四、結果與討論

研究結果發現三種不同能力的設定方式對估計的精確度並沒有太大的影響，基本上以方法一的誤差 WISE 大了一點，但因為研究中 WISE 乃是乘上 10000 倍後的值，所以真正的差異並不算顯著。另外，不論是採用哪一種方法，小題庫情形下的估計結果都比大題庫的估計所得要精準，這應該是因為在小題庫時，每一題的作答人數較多，所得答題資訊較完整，因此估計結果也比較精準。

在參數估計值的收斂方面，研究中也發現建議分配對收斂有些許影響，因此不同的題庫要搭配不同的建議分配，才能有較好的收斂結果，若鑑別度參數是以常態分配生成，則參數建議分配全部使用常態分配，會比猜測參數的建議分配改為均勻分配的情況更適當；如果鑑別度參數是以 lognormal 分配產生的，則參數建議分配分別採用 lognormal、常態和均勻分配比全部使用常態分配來得好，但是後者有個其他情形都沒有的優點：在鑑別度參數估計讓沒有低估的問題。另外，建議分配中的變異數的確會影響 MH 法中的機率值α，基本上變異數的設定要讓 MCMC 抽取的樣本有適當的移動空間才恰當。

接下來討論 MCMC 的部分。MCMC 中設定的起始值雖會對剛開始抽取的樣本有較顯著的影響，但是只要建議分配的變異數不要設定的過小，讓連續抽取的兩個樣本有足夠的移動空間，那麼 MCMC 過程中所抽取的樣本很快的就會脫離起始值的影響，進而變成趨向從平穩分配中所抽取的樣本。不過從另一方面而言，適當的起始值設定還是有助於加速收斂速度的。另外在 Patz 和 Junker 的研究中，所抽取的樣本數都在 7000 筆以上，本研究中經過多次的實驗結果，發現在此模擬研究條件下取 3000 個樣本數，就已有不錯的收斂結果。其實只要當 MCMC 進入收斂狀態，一個很長的馬可夫鏈對估計結果並沒有太大的影響。研究中儘管有些少數題目在 3000 次時還沒進入收斂狀態，其抽取的樣本數值有越來越大或越來越小或者來回振動不停的情況，不過如果在建議分配抽取候選樣本時，限定其最大值與最小值的範圍，可能會對收斂情況有所改善。另外，研究中抽取樣本的方式為全部選取，但是 MCMC 方法也可以間隔選取樣本，例如每 5 個樣本選取一個樣本作為估計之用，將來可以試試這樣的方式是否能增加估計精確度。

在本研究中，每種模擬研究的情況，都僅做了一次 MCMC 估計，可能仍有不足的部分，如果重複 MCMC 多次，以數次結果之平均值作為估計值，那麼估計誤差可能會更小。另外，在三種能力方法的比較，只用了圖形與 WISE 數值大小作為依據，可再進一步的以統計方法檢定三種方法是否有所差異。最後，在研究中方法二與方法三的 strong 與平坦的先驗分配區別並不大，可能是造成三種能力方法估計的結果差異不明顯的原因，或許可以給定一個更加平坦的先驗分配，再做模擬研究，觀察是否有顯著的差別。

最後，在研究中發現大部分難度偏低的題目，其估計誤差都較小，是否以 MCMC 估計，難度低的題目其估計精確度都較高仍待進一步的證實。而鑑別度參數除了第二類題庫搭配建議分配全為常態分配的情形外都有被低估的現象，但此情形卻出現一些估計誤差較大的試題。而鑑別度低估的原因，目前尚未有合理的推論，也需再進一步探討。

五、參考文獻

Ban, J. -C., Hanson, B. H., Wang, T., Yi, Q., & Harris, D. J. (2000). *A comparative study of online pretest item calibration methods in computerized adaptive testing*, (ACT Research Report 00-11). Iowa City, LA: ACT, Inc.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Chang, S. W. (1998). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. Doctoral dissertation, University of Iowa, Iowa City.

Flaugher, R. (1990). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 41-63). Hillsdale, NJ: Lawrence Erlbaum Associates.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1998). *Markov chain Monte Carlo in practice.* London: Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika, 57,* 97-109.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Chicago: Scientific Software, Inc.

Patz, R .J. & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Roberts, G. O. (1995). Markov chain concepts related to sampling algorithms. In *Markov chain Monte Carlo in practice.* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 45-57. London: Chapman & Hall.

Stocking, M. L. (1988). Scale drift in on-line calibration (Research Report 88-28). Princeton, NJ: ETS.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). A*nnals of Statistics, 22,* 1701-1762.

Tierney, L. (1995). Introduction to general state-space Markov chain theory. . In *Markov chain Monte Carlo in practice.* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 59-74. London: Chapman & Hall.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Services.