國立臺灣師範大學教育心理與輔導學系
教育心理學報，民 95，38 卷，2 期，195 — 211 頁

能力估計方法對多向度電腦化適性測驗測量精準度的影響

陳 柏 熹
國立台灣師範大學
教育心理與輔導學系

本研究旨在分析不同能力估計方法對多向度電腦化適性測驗（multidimensional computerized adaptive testing, MCAT）測量精準度的影響。研究分爲兩階段：第一階段先找出在 MCAT 中貝氏期望後驗法（expected a posteriori, EAP）的最佳節點數（quadrature point）；第二階段是比較最大概似法（maximum likelihood, ML）、期望後驗法（EAP）與最大後驗法（maximum a posteriori, MAP）在不同向度（二向度與四向度）及不同相關性（低相關與高相關）的情況下，進行不同題數（20 題、40 題、60 題、80 題）MCAT 時的能力估計信度、偏誤（bias）以及均方根誤（root mean square of error, RMSE）。階段一的結果顯示，隨著 EAP 節點數的增加（從 5~30 點）與能力向度的增加，其選題所需的時間會明顯地增加。在考量到選題時間又不致影響到測量精準度的情況下，在 MCAT 中將 EAP 的節點數訂爲 10 是理想的選擇。階段二的結果顯示，MAP 法與 EAP 法比 ML 法的能力估計信度高，均方根誤較低。在平均偏誤方面此三種方法則無明顯差異，不過 MAP 法會有明顯的迴歸性偏誤。這些現象在能力間相關較高、能力向度數量較多以及題數較少時會更明顯。整體而言，三種方法各有其優缺點，其中 MAP 法的迴歸性偏誤、EAP 法的選題時間以及 ML 法的信度與測量誤差是未來進行 MCAT 時需要改善的問題。

關鍵字：多向度電腦化適性測驗、最大概似法、最大後驗法、期望後驗法

電腦化適性測驗（computerized adaptive testing, CAT）主要是利用電腦的快速運算能力，根據受試者的答題反應立刻估計出其能力，並且挑選出適合該受試者能力的下一道試題讓受試者作答。由於受試者所接受到的試題都很接近其能力水準，因此只要用較少的題數就可以達到與傳統測驗相同的測量精確度（Sand, Water, & McBride, 1997; Wainer et al., 1990; Weiss, 1985）。而 CAT 理論基礎主要是源自於試題反應理論（item response theory, IRT）。基於 IRT 的單向度（unidimensionality）假定，以及受試者的能力估計的不變性（invariance），使接受不同試題的受試者能力可以被放在同一個尺度上互相比較（Hambleton & Swaminathan, 1985; Wainer et al., 1990）。然而，單向度假定也限制了的 CAT 的應用，使 CAT 大都侷限在單向度能力的測量上。對於人格量表、多元性向測驗、綜合能力測驗…等多向度測驗，以及一些含有多向度題的測驗而言，目前尚無法用 CAT 來進行。

爲了突破單向度 IRT 的限制，學者們紛紛提出多向度試題反應理論（multidimensional item

response theory, MIRT），並將之應用到多向度電腦化適性測驗（multidimensional computerized adaptive testing, MCAT）的程序中（陳柏熹、王文中，民 89a，民 89b；Hattie, 1981; Luecht, 1996; Mckinley & Reckase, 1983; Reckase & Mckinley, 1991; Segall, 1996; Sympson, 1978）。由於 MCAT 同時包含了多向度分析與適性程序的優勢，因此其測量精準度又比單向度電腦化適性測驗（UCAT）更高。然而，測量精準度與能力估計方法有密切的關係：根據過去單向度電腦化適性測驗的研究顯示（洪碧霞、吳鐵雄、黃千綺、江秋坪、許宏彬，民 81; Bock & Mislevy, 1982; Weiss & Mcbride, 1984），最大概似估計法（maximum likelihood, ML）的均方根誤（root mean square of error, RMSE）較高，但是比較沒有偏誤（bias）；而貝氏最大後驗法（Bayesian maximum a posteriori, MAP）與貝氏期望後驗法（Bayesian expected a posteriori, EAP）的均方根誤較小，但會有迴歸性的偏誤。而這三種能力估方法在 MCAT 中測量精準度的差異有多大，目前尚不清楚。本研究的目的之一就是想比較這三種方法在 MCAT 中的測量精準度，以作爲後續發展 MCAT 時的參考；又由於在 EAP 方法中所選取的節點數量與 EAP 的效果有關，因此本計畫也將探討不同節點數量的 EAP 對 MCAT 測量精準度的影響。

一、單向度電腦化適性測驗的能力估計

CAT 主要的理論基礎是 IRT。在 IRT 中，每個考生在每個試題上的答對機率主要是受到考生能力與試題參數所影響。其基本假設有兩項，第一是單向度，也就是所有題目都是測量同一向度；第二是局部獨立性（local independency），意指對同能力水準者而言，答對某一題的機率與答對其他題目的機率是無關的。藉由 IRT 的模式與局部獨立性的假設，可以計算出受試者在整份測驗上的反應概似函數（likelihood function），並藉此估計出每個受試者在接受不同測驗試題後的能力。只要這些測驗試題都測量相同的能力並符合 IRT 的模式與基本假設，則接受不同試題的受試者其能力估計值就可以互相比較。

CAT 就是使用反應概似函數來估計受試者的能力。估計方法是先由 IRT 的反應模式算出 k 個人在 n 個試題上的概似函數，再找出此函數的最佳解。尋找最佳解的常見方法有三種：第一種是找出能使概似函數最大化的能力值，稱爲最大概似估計法（ML）。爲了加速找到能使概似函數最大化的能力值，通常以牛頓 - 約佛森（Newton-Raphson）法來進行疊代。

第二種是以受試者的事前能力分布 $f(\theta)$ 作爲加權值，形成事後機率密度函數，並找出能使此事後機率密度函數 $f(\theta \mid \mathbf{U})$ 最大化的能力值，稱爲貝氏最大後驗法（MAP）。事後機率密度函數計算方式如公式（1）所示：

$$f(\theta \mid \mathbf{U}) = \frac{L(\mathbf{U} \mid \theta) f(\theta)}{f(\mathbf{U})} \quad , \tag{1}$$

其中，$f(\theta)$ 爲受試者的事前能力分布，$L(\mathbf{U} \mid \theta)$ 是能力值爲 θ 者的反應概似函數，$f(\mathbf{U})$ 是受試者的邊際機率，是由 $L(\mathbf{U} \mid \theta) f(\theta)$ 從 $-\infty$ 到 ∞ 積分所得。爲了加速找到能使事後機率密度函數最大化的能力值，通常也是以牛頓 - 約佛森（Newton-Raphson）法來進行疊代。

第三種與第二種方法類似，只是所尋找的能力值是事後機率密度函數的期望值，稱爲期望後驗法（EAP）。如公式（2）所示：

$$\theta_{EAP} = \sum_{q=1}^{k_q} \theta_q f(\theta_q \mid \mathbf{U}) = \sum_{q=1}^{k_q} \theta_q \frac{L(\mathbf{U} \mid \theta_q) f(\theta_q)}{\sum_{q=1}^{k_q} [L(\mathbf{U} \mid \theta_q) f(\theta_q)]} \quad , \tag{2}$$

其中 $f(\theta)$ 、 $L(\mathbf{U} \mid \theta)$ 的說明如公式（1）所示。而 q 是計算能力的期望值時所切割成的節點，k_q 爲最高節點數。k_q 點愈大，EAP 的計算結果就愈精確。

在各種能力估計方法的比較上（洪碧霞等人，民 81; Bock & Mislevy, 1982; Weiss & Mcbride,

1984），MAP 法與 EAP 法的均方根誤較小，但是會有迴歸性的偏誤，估計時需要用到受試群體能力的先驗分布資訊（平均數與變異數），對各種答題反應的受試者皆可進行估計。兩種方法的估計效果差不多，但 EAP 法的迴歸性偏誤較小。ML 比較沒有迴歸性偏誤，但均方根誤較大，且受試者的答題反應中必須有答對也有答錯的反應才能進行估計，全部答對或全部答錯者無法進行。

二、多向度試題反應理論

由於目前大部分的 CAT 都是建立在 IRT 單向度假定的基礎上，亦即測驗中所有試題都是在測量同一種特質，可稱爲單向度電腦化適性測驗。然而，實際生活情境中有許些問題並非靠單一能力或潛在特質就能解決的，測驗的作答也是如此（Kelderman, 1996）。當受試者答對試題的機會受到不只一種能力所影響時，已經違反了單向度 IRT 的理論假設，因此不應該進行單向度 CAT。Ackerman（1991）的研究顯示，當試題測量不只一種能力時，如果以單向度 IRT 來進行參數估計會使鑑別度較大的能力向度被擴大、鑑別度較小的向度被縮小或忽略掉，產生偏差的試題參數估計值，而且所估計出來的能力其意義已經模糊了，不適合放在同一個向度上互相比較。

再從測量精確度的角度來看，單向度 IRT 無法藉著各向度能力的相關性來提升對各向度能力估計的精確性，因此每個向度都需要很多題，才能到達某個信度水準。這也是爲什麼目前大部分的人格量表，興趣量表或性向測驗的題數都這麼多。多向度 IRT 在估計能力時，會將向度間的相關性納入估計程序中，提升了各向度能力估計的精確性，因此每個向度只要少數幾題就能使各向度具有高信度了。Wang, Chen 與 Cheng（2004）的研究顯示，當向度之間爲高相關時，多向度 IRT 分析可以大幅提高各向度的信度，由原本的 0.6（單向度 IRT 分析）提昇至 0.8。

爲了使 CAT 的測量信度能夠更加提升，並突破試題只能測量一種能力的限制，近幾年來 MCAT 的概念漸漸被提出來（陳柏熹、王文中，民 89a，民 89b；Li & Schafer, 2005; Luecht, 1996; Segall, 1996; Wang & Chen, 2004）。MCAT 的發展需要仰賴多向度試題反應理論（MIRT），以及能力估計、訊息量、選題等相關算則的改進。

近代學者們提出來的多向度 IRT 模式大多是單向度 IRT 模式的衍生模式。例如：Mckinley 與 Reckase（1983）所發展的多向度二參數模式（multidimensional two parameters model）是二參數 IRT 的衍生模式（簡稱爲 M2PL），如公式（3）所示：

$$P_i(x_{i_j}=1\,|\,\mathbf{a}_i,b_i,\boldsymbol{\theta}_j)=\frac{1}{1+\exp[-(\mathbf{a}_i'\boldsymbol{\theta}_j-b_i)]}\ , \qquad (3)$$

其中 x_{ij} 爲受試者反應型態，答對該題時記錄爲 1，答錯時記錄爲 0。$\mathbf{a}_i$ 爲試題鑑別度向量，b_i 爲試題難度，$\boldsymbol{\theta}_j$ 爲能力向量。此模式是將原本的受試者能力值 $\boldsymbol{\theta}_j$ 與試題鑑別度 $\mathbf{a}_i$ 擴展爲向量 $\boldsymbol{\theta}_j$、$\mathbf{a}_i$，如此就能將多向度的能力同時包含在模式中，也就是答對試題的機率會受到多種能力所影響。當能力和鑑別力都只有一個向度時，就變成純量而非向量，M2PL 就簡化爲單向度的二參數模式（Birnbaum, 1968）。

Hattie（1981）的模式與 Sympson（1978）的模式相當類似，都是將 Birnbaum（1968）單向度三參數模式中的能力參數與鑑別度參數改成向量的型式所產生的（簡稱爲 M3PL）。其反應模式如公式（4）所示：

$$P_i(x_{ij}=1\,|\,\mathbf{a}_i,b_i,c_i,\boldsymbol{\theta}_j)\equiv c_i+\frac{1-c_i}{1+\exp[-\mathbf{a}_i'(\boldsymbol{\theta}_j-b_i\mathbf{1})]}\ , \qquad (4)$$

其中 c_i 爲試題的猜對率，$\mathbf{a}_i$ 爲試題鑑別度向量，$\mathbf{1}$ 是爲了使試題的難度 d_i 成爲向量，這樣才能與能力向量相減。$\boldsymbol{\theta}_j$，$\mathbf{a}_i$ 的意義與上述 Reckase 與 Mckinley（1991）的模式相同，這兩種模式的概念相當接近。Segall（1996）的 MCAT 程序就是用此模式發展出來的。此外，Ackerman（1994）也曾依上述兩

種模式提出二向度試題訊息量圖示法。

Adams、Wilson 與 Wang（1997）等人所提出來的多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM）爲 Rasch 模式的衍生模式。其反應模式如公式（5）所示：

$$f(X_{ik}=1;\boldsymbol{\xi}\,|\,\boldsymbol{\theta})=\frac{\exp(\mathbf{b}_{ik}'\boldsymbol{\theta}+\mathbf{a}_{ik}'\boldsymbol{\xi})}{\sum_{u=1}^{Ki}\exp(\mathbf{b}_{iu}'\boldsymbol{\theta}+\mathbf{a}_{iu}'\boldsymbol{\xi})}\quad , \tag{5}$$

其中 $f(X_{ik}=1;\boldsymbol{\xi}\,|\,\boldsymbol{\theta})$ 表示能力向量爲 $\boldsymbol{\theta}$ 的受試者在第 i 題上答出第 k 類別反應的機率；ξ 爲試題參數向量；X_{ik} 爲受試者反應型態，K_i 爲第 i 試題的計分類別數。其中 $\mathbf{b}_{ik}$ 爲第 i 題在第 k 個反應類別上的計分向量；$\boldsymbol{\theta}$ 爲受試者能力向量；$\mathbf{a}_{ik}$ 爲第 i 題中第 k 個反應類別的設計向量。舉例來說，若一份測驗中測量到了 D 種能力，分別爲 θ_1, θ_2, θ_3, …, θ_D 等；而受試者在第 i 個試題回答出第 k 個反應類別時，在這 D 個向度能力上所得到的分數爲 $\mathbf{b}_{ik}=(b_{ik1}, b_{ik2},\ldots b_{ikD})'$，此計分向量應該根據當初設計試題時的建構來決定的，也就是試題設計的理論建構與計分方式必須一致。在 ξ 方面，若測驗試題共有 3 題，第 1 題的計分爲 0~2 的部份給分，其餘兩題爲 0,1 的二元計分，則第一題需要估計 2 個參數（難度）、其餘兩題各估計 1 個參數，共估計了 ξ_{11}, ξ_{12}, ξ_2, ξ_3 等 4 個參數，即 $\xi=(\xi_{11}, \xi_{12}, \xi_2, \xi_3)'$。$\mathbf{a}_{ik}$ 稱爲設計向量（design vector），是估計每個試題參數時所使用的係數，也就是描述了第 i 題的第 k 類別反應是否要用來估計某個參數 ξ_{ik}，這可以根據研究者的目的自行設計，詳見 Wu、Adams 與 Wilson（1998）。

陳柏熹（民 90）曾對這三類 MIRT 模式進行比較。簡單來說，Reckase 與 Mckinley（1991）的優點是引入了多向度鑑別度以及其方位角的概念，這些資訊可以幫助理解多向度試題的訊息量。然而，該模式只適用在二元計分以及各向度能力互相獨立的情境中。而且其能力估計與訊息量計算方式也沒有考慮到能力間的相關性。Ackerman（1994）修改了該模式，發展出可以容許各向度能力有相關性的模式，以及納入此相關性的能力估計與訊息量校正公式。但是當能力向度超過兩個時，Ackerman 所發展出來的訊息量校正公式變得相當複雜。Hattie（1981）以及 Sympson（1978）的三參數模式雖然已經由 Segall（1996）發展成 MCAT 的程序，但也只能用在二元計分的模式中。上述兩種模式在進行試題參數估計時還會遇到無法界定（unidentifiable）的問題。因此在 Segall（1996）的研究中只能用單向度 IRT 估計出來的參數來進行 MCAT。

MRCMLM 功能較多且包容性較廣，舉凡最原始的 Rasch 模式（Rasch, 1960; Wright & Stone, 1979）、Fischer（1973）的對數潛在特質模式（logistic latent trait model; LLTM）、Andrich（1978）的評定量尺模式（rating scale model）、Master（1982）的部份給分模式（partial credit model）…等，都是它的特例。就計分方式來說，MRCMLM 可以適用在二元計分、多元計分、評定量尺等計分方式上；就能力向度來說，可以用來估計單向度與多向度能力，而且各向度間可以允許有相關性。該模式目前已經發展出參數估計軟體 ConQuest（Wu et al., 1998）。因此，本研究以該模式作爲理論基礎來發展 MCAT 的能力估計與選題算則。

三、多向度電腦化適性測驗的能力估計

根據多向度試題反應模式可以發展出多向度的電腦化適性測驗。根據 MRCMLM，受試者在 n 個試題上的概似函數如公式（6）所示：

$$L(x_{v_1}, x_{v_2},\ldots\,|\,\boldsymbol{\theta})=\prod_{i\in V}P_i(\boldsymbol{\theta})^{u_i}Q_i(\boldsymbol{\theta})^{1-u_i}\quad , \tag{6}$$

其中 V 爲被選到的適性試題，$P_i(\boldsymbol{\theta})$ 表示能力爲 $\boldsymbol{\theta}$ 的受試者在第 i 題答出某個反應的機率，如同公式（5）中的 $f(X_{ik}=1;\boldsymbol{\xi}\,|\,\boldsymbol{\theta})$；而 $Q_i(\boldsymbol{\theta})=1-P_i(\boldsymbol{\theta})$。各種能力估計法的推導過程如下：

（一）最大概似估計

最大概似估計就是找出能使公式（6）最大化的能力向量。爲了加速找到此能力向量，可以先對此概似函數取自然對數，再以牛頓—約佛森程序來進行疊代。其作法是先求出概似函數的一階微分向量（Wang, 1994）：

$$\frac{\partial}{\partial\theta}\ln L(\theta\,|\,\mathbf{u})=\begin{bmatrix}\dfrac{\partial}{\partial\theta_1}\ln L(\theta\,|\,\mathbf{u})\\ \dfrac{\partial}{\partial\theta_2}\ln L(\theta\,|\,\mathbf{u})\\ \vdots\\ \dfrac{\partial}{\partial\theta_k}\ln L(\theta\,|\,\mathbf{u})\end{bmatrix}$$

其元素爲

$$\frac{\partial \ln f(\boldsymbol{\theta}\,|\,\mathbf{u})}{\partial\theta_k}=\frac{\partial}{\partial\theta_k}\ln L(\mathbf{u}\,|\,\boldsymbol{\theta})=\sum_{i\in V}[b_{ik}-E_i(\boldsymbol{\theta})]\ ,\qquad(7)$$

其中 $E_i(\theta)=\sum_{k=1}^{K}b_{ik}f_{ik}(\theta)$，而 $f_{ik}(\boldsymbol{\theta})$如公式（5）所示。V 爲選到的題目。再求出二階微分矩陣：

$$\mathbf{J}(\theta)=\begin{bmatrix}\partial^2\ln L(\theta\,|\,\mathbf{u})/\partial\theta_1^{\,2} & \partial^2\ln L(\theta\,|\,\mathbf{u})/\partial\theta_1\partial\theta_2 & \ldots & \partial^2\ln L(\theta\,|\,\mathbf{u})/\partial\theta_1\partial\theta_p\\ & \partial^2\ln L(\theta\,|\,\mathbf{u})/\partial\theta_2\partial\theta_2 & \ldots & \partial^2\ln L(\theta\,|\,\mathbf{u})/\partial\theta_2\partial\theta_p\\ & & & \vdots\\ & & & \partial^2\ln L(\theta\,|\,\mathbf{u})/\partial\theta_p\partial\theta_p\end{bmatrix}$$

其而二階微分矩陣中第（k, l）元素的公式爲：

$$\frac{\partial^2 \ln f(\boldsymbol{\theta}\,|\,\mathbf{u})}{\partial\theta_k\partial\theta_l}=-\sum_{i\in V}[\sum_{k=0}^{K}b_{ik}b_{ik}{}'f_{ik}(\boldsymbol{\theta})-E_i(\boldsymbol{\theta})E_i(\boldsymbol{\theta})']\ ,\qquad(8)$$

接著再以下列算式進行疊代：

$$\boldsymbol{\theta}^{(j)}=\boldsymbol{\theta}^{(j-1)}+\boldsymbol{\delta}^{(j)}\ ,\qquad(9)$$

其中，$\boldsymbol{\theta}^{(j)}$ 爲受試者在在第 j 次疊代的能力估計值，而

$$\boldsymbol{\delta}^{(j)}=[\frac{\partial^2 \ln f(\boldsymbol{\theta}\,|\,\mathbf{u})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}]^{-1}\times\frac{\partial \ln f(\boldsymbol{\theta}\,|\,\mathbf{u})}{\partial\boldsymbol{\theta}}\ ,\qquad(10)$$

估計 $\boldsymbol{\theta}^{(j)}$ 時必須反覆將新的能力值代入公式（9）計算，直到 $\boldsymbol{\delta}^{(j)}$ 收斂到某個值爲止，如此便得到暫時的能力估計值 $\boldsymbol{\theta}^{(j)}$，再依此能力值選出下一題。

（二）貝氏期望後驗估計

貝氏期望後驗法是以事後機率密度函數的期望值作爲能力向量的最佳估計值。多向度能力的事後機率密度函數 $f(\theta\,|\,\mathbf{U})$ 爲：

$$f(\boldsymbol{\theta}\,|\,\mathbf{U})=\frac{L(\mathbf{U}\,|\,\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{U})}\ ,\qquad(11)$$

其中，$f(\theta)$ 爲受試者的事前能力分布，$L(\mathbf{U}\,|\,\theta)$ 是能力值向量爲 $\boldsymbol{\theta}$ 者的反應概似函數，$f(\mathbf{U})$

是受試者的邊際機率，是由 $L(\mathbf{U}|\theta)f(\theta)$ 在各能力向度從 $-\infty$ 到 ∞ 積分所得。而其期望值 $\boldsymbol{\theta}_{EAP}$ 爲：

$$\boldsymbol{\theta}_{EAP}=\sum_{q=1}^{k_q}\boldsymbol{\theta}_q f(\boldsymbol{\theta}_q|\mathbf{U})=\sum_{q=1}^{k_q}\boldsymbol{\theta}_q\frac{L(\mathbf{U}|\boldsymbol{\theta}_q)f(\boldsymbol{\theta}_q)}{\sum_{q=1}^{k_q}[L(\mathbf{U}|\boldsymbol{\theta}_q)f(\boldsymbol{\theta}_q)]}\ , \tag{12}$$

其中 q 爲各向度的節點數量。在進行多向度估計時，q 的數量會隨著能力向度呈指數遞增。例如：當我們設定每個能力向度分成 n 個節點來求期望值時，則五向度能力的節點總數爲 n^5。向度數量愈多，則能力估計時所需要的時間就愈久；若降低每個向度的節點數，又會因爲能力取樣數量太少使能力估計的精確度變差。因此實際應用在多向度能力的估計時，較常以貝氏最大後驗法來進行。

（三）貝氏最大後驗估計

貝氏最大後驗法的事後機率密度函數與貝氏期望後驗法相同。爲了加速找到事後機率密度函數的最大值，該方法比照 ML 法依牛頓 - 約佛森程序來進行。首先將 1n $f(\theta|\mathbf{U})$ 分別對 k 個能力向度進行偏微分。MRCMLM 事後機率度函數的一階偏微分向量中的元素爲（陳柏熹，民 90；Wang, 1994）：

$$\frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{u})}{\partial\theta_k}=\frac{\partial}{\partial\theta_k}\ln L(\mathbf{u}|\boldsymbol{\theta})-\frac{1}{2}\frac{\partial}{\partial\theta_k}[(\boldsymbol{\theta}-\boldsymbol{\mu})'\Phi^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})]\ ,$$

$$\sum_{i\in V}[b_{ik}-E_i(\boldsymbol{\theta})]-[\frac{\partial}{\partial\theta_k}(\boldsymbol{\theta}-\boldsymbol{\mu})']\Phi^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\quad , \tag{13}$$

其中 $E_i(\theta)=\sum_{k=1}^{K}b_{ik}f_{ik}(\theta)$，而 $f_{ik}(\boldsymbol{\theta})$ 如公式（5）所示。V 爲選到的題目，$\boldsymbol{\mu}$ 爲 $\boldsymbol{\theta}$ 的平均數向量，$\boldsymbol{\Phi}$ 爲 $\boldsymbol{\theta}$ 的共變數矩陣。接著再算出二階偏微分矩陣，其二階微分矩陣中第（k, l）元素的公式爲：

$$\frac{\partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})}{\partial\theta_k\partial\theta_l}=-\sum_{i\in V}[\sum_{k=0}^{K}b_{ik}b_{ik}'f_{ik}(\boldsymbol{\theta})-E_i(\boldsymbol{\theta})E_i(\boldsymbol{\theta})']-\phi^{kl}\quad , \tag{14}$$

其他程序則比照最大概似估計法來進行。

在 MCAT 的相關研究方面，Segall（1996）曾以 M3PL 模式發展出 MCAT 的程序，並以電腦化軍旅性向測驗（Computerized Adaptive Testing version of the Armed Served Vocational Aptitude Battery; CAT-ASVAB）的九個分量表進行 MCAT 以及 UCAT 的模擬研究，結果發現，用 MAP 法進行能力估計時，MCAT 的能力估計的精準度比 UCAT 好，以要達到相同的能力精準度來看，MCAT 可以比 UCAT 節省 1/3 的題數。Luecht（1996）將 Segall 的模式應用在醫學證照考試的題庫中，用最大概率法進行 MCAT 與 UCAT 的模擬研究，並且在選題時加上向度題數比例的條件限制。他發現從每個向度的分數與信度來看，MCAT 所得到的各向度信度都比 UCAT 高，而且這種測量信度的優勢在題數較少或信度較差的向度中更加明顯。Li 與 Schafer（2005）的研究則指出，使用多向度 MCAT 的能力估計精準度比 UCAT 好，尤其是對那些高能力與低能力者來說效果更好，而且還可以減少題庫中未被使用的試題比率。陳柏熹與王文中（民 89a，民 89b）以模擬資料來進行題間多向度 MCAT（BMCAT）與題內多向度 MCAT（WMCAT），結果發現當向度數量愈多或各向度之間的相關愈高時，MCAT 的信度比 UCAT 高出愈多，其中 WMCAT 又比 BMCAT 的信度更高。

Segall（1996）與 Luecht（1996）的研究都是以單向度試題參數與模擬資料來進行 MCAT 研究，Segall 將之稱爲多向性 UCAT（multi-unidimensional CAT）。但這在邏輯上是自相矛盾的。既然是使用 MCAT，就應該以多向度試題參數來進行，如此試題參數的取得與 MCAT 才是在相同的理論基礎上進行的，其結果才可靠，這也是爲何本研究採用 MRCMLM 模式的原因，因爲該模式已經發展出可以直接估計多向度試題參數的軟體 ConQuest。

在 MCAT 的能力估計方法的比較方面，Tseng（2001）曾使用 M3PL 模式比較 ML 法、加權最大概似法（WLE）、EAP 法與 MAP 法在三個向度的 MCAT 中的能力估計精準度，結果發現後面三種方法與 ML 法的信度與測量偏誤都差不多，這與過去單向度 CAT 所得的結果不太相同（洪碧霞等人，民 81; Bock & Mislevy, 1982; Weiss & Mcbride, 1984）。筆者深究其原因主要是使用了 M3PL 模式（如公式（4）所示）。由於在該模式中，各向度能力之間必須限制爲獨立，因此能力的相關係數矩陣無法提升測量精準度。本研究將改用 MRCMLM 來進行。當使用 MRCMLM 時，不論是貝氏 EAP 或 MAP 法都能利用能力事前分布的共變數矩陣來提升對各向度能力的估計，而 ML 法只有多向度適性選題的優勢，能力的事前分布的資訊並沒有被使用到其能力估計程序中，因此 EAP 或 MAP 法的能力估計精準度應該會比 ML 法高。此外，由於 EAP 的能力估計精準度會受到計算時分割的能力節點數量所影響，而能力節點數量與能力向度的數量太多又會造成能力估計所需的時間大量增加，因此，在使用 EAP 進行能力估計時，需先分析不同節點數量與不同向度數量對能力估計精準度及估計時間的影響，在找出較適當的 EAP 節點數量後，再進行後續研究了解這三種不同能力估計方法對 MCAT 能力估計精準度的影響。

方 法 與 結 果

研究一、EAP 節點數量對 MCAT 能力估計精準度與估計時間的影響

（一）研究設計

本研究是探討在 EAP 估計方法中，能力節點數量、向度數量與向度間相關程度對 MCAT 能力估計精準度與選題時間的影響。以找出在進行 MCAT 時最適當的 EAP 節點數。本研究的設計如下：

1. 自變項

a. 節點數量：將各向度節點數量定爲 5、10、20 與 30 點，共四種。

b. 能力向度數量：有兩種，分別爲二向度與四向度。

c. 向度間的相關程度：有兩種，分別爲低相關（< 0.4）與高相關（> 0.7）。

2. 依變項

本研究的依變項有兩項，分別爲各向度能力估計的平均信度與平均每進行一題 MCAT 的時間。其中能力估計的信度爲能力估計值與眞實能力值的相關係數平方（Segall, 1996）。

（二）研究程序

1. 產生模擬資料

先根據研究變項中的向度數量與向度間相關程度，產生四組相關係數矩陣。再以這四組相關係數矩陣來產生受試者的能力值。四組相關係數矩陣的數值如表一所示。產生方式是以 Fortran 中 CHFAC 與 RNMVN 等函數配合上述的相關係數矩陣來隨機產生 1000 筆多變項常態分布（平均數爲 0、標準差爲 1）的能力值。另外再針對每個向度隨機產生 100 題難度爲均等分布（UD（-3.0~3.0））的模擬題目。接著以 MRCMLM 來產生受試者的反應資料，也就是將受試者的能力與題目難度代入 MRCMLM 中算出每個受試者在每個題目上的答對機率，再將此答對機率與從 0.0~1.0 的均等分布中產生出來的隨機值相比較，如果隨機值大於答對機率，則判斷受試者答錯該題，否則就判斷受試者答對該題。

2. 估計多向度試題參數

本研究使用 ConQuest 軟體（Wu et al., 1998）對上述資料進行題間多向度（每個題目只測量到所屬的單一向度）試題參數估計。估計方法是分別對四組資料（兩種向度數量配合兩種相關程度）進行 MIRT 估計。MIRT 估計的優點是可以將各向度的相關也納入估計過程中，提高參數估計的精確性。

表一　產生模擬資料的四組相關係數矩陣

二向度，低相關				二向度，高相關			
1.00				1.00			
0.28	1.00			0.87	1.00		
四向度，低相關				四向度，高相關			
1.00				1.00			
0.29	1.00			0.89	1.00		
0.35	0.28	1.00		0.85	0.88	1.00	
0.22	0.27	0.31	1.00	0.92	0.87	0.91	1.00

3. MCAT 執行程序

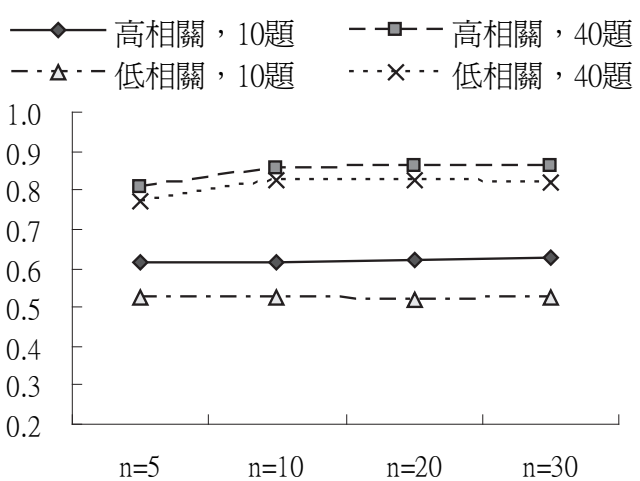
將上述估計出來的試題參數當作四組資料題庫中的試題參數，配合程序 1 所產生出來的受試者模擬反應，使用 EAP 能力估計方法來進行 MCAT，進行時分別將各向度節點數量定爲 5、10、20 與 30 點，每組資料皆進行四種不同節點數量的 MCAT。每種程序分別進行總題數爲 10 題（低題數）與 40 題（高題數）的 MCAT。

（三）資料分析

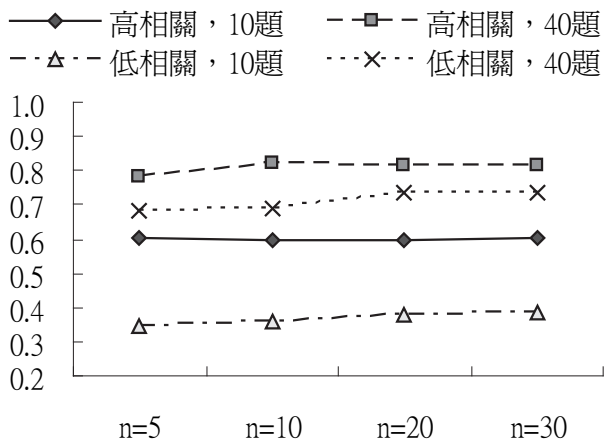
分別針對兩個依變項，描繪以各種節點數量、向度數量與向度間相關程度爲自變項的折線圖，以了解這些因素對 MCAT 測量精準度的影響，以及隨著 MCAT 測驗題數的要求不同時，兩項依變項指標的變化情形。

（四）研究結果

圖一與圖二是分別在能力爲二向度或四向度的情況下，以不同節點數的 EAP 進行 MCAT 的能力估計信度。從圖一可以看出，不同節點數對 EAP 的能力估計信度並沒有明顯的影響，只有當 MCAT 的題數較多（40 題）時，且每個向度的節點數量是 5 點時，以 EAP 法進行 MCAT 的信度會比較低。當 MCAT 題數較少時，或是節點的數量在 10 點以上時，以 EAP 進行 MCAT 所得到的信度都差不多。圖二也呈現出類似的情形，但是當各向度間爲低相關且節點數是 10 點或 5 點時，所得的信度比節點數是 20 點或 30 點時略低一些。而影響 MCAT 信度的最主要因素是題數與向度間的相關高低，這與過去 MCAT 的相關研究果一致（陳柏熹、王文中，民 93）。



圖一　以不同節點數之 EAP 進行二向度 MCAT 的能力估計信度



圖二　以不同節點數之 EAP 進行四向度 MCAT 的能力估計信度

表二是在能力爲二向度與四向度的情況下，以各節點數進行 EAP 時，平均每執行 1 題 MCAT 所需的能力估計加上選題的時間，筆者所使用的電腦配備是 Pentium 2.8 GMHz 的 CPU 以及 256M 的 RAM。從表中可以看出在二向度時，所有節點數量其平均每執行 1 題 MCAT 所需的時間均少於 1 秒；但是在四向度且節點數爲 20 點時，平均每執行 1 題 MCAT 就需要 3 秒；而節點數爲 30 點時，平均每執行 1 題 MCAT 就需要 15 秒，這是相當久的。從表中亦可明顯看出當能力向度數量與 EAP 節點數的增加，平均每執行 1 題 MCAT 所需的時間會大幅提高。因此，在同時考量信度與時間的情況下，筆者建議若要以 EAP 法執行四向度以內的 MCAT 時，其各向度的節點數應該至少要在 10 點以上，但是最好不要超過 20 點，否則隨著能力向度的增加，選題與能力估計時間會呈現指數倍增加。

表二　在不同節點數時，以 EAP 法平均每執行 1 題 MCAT 所需的時間

	二向度	四向度
節點 =5	0.001 秒	0.007 秒
節點 =10	0.005 秒	0.200 秒
節點 =20	0.015 秒	3.000 秒
節點 =30	0.033 秒	15.000 秒

研究二、不同能力估計方法對 MCAT 測量精準度的影響

（一）研究設計

研究二是探討最大概似估計（ML）、貝氏最大後驗估計（MAP）與貝氏期望後驗估計（EAP）三種能力估計方法對 MCAT 能力估計精準度的影響。研究設計如下：

1. 自變項

a. 能力估計方法：有三種，分別爲 ML 法、EAP 法與 MAP 法。

b. 能力向度數量：有兩種，分別爲二向度與四向度。

c. 向度間的相關程度：有兩種，分別爲低相關（< 0.4）與高相關（> 0.7）。

2. 依變項

本研究的依變項有三項，分別爲各向度能力估計的平均信度、平均偏誤（bias）與均方根誤（root mean square of error, RMSE）。其中能力估計的信度爲能力估計值與眞實能力值的相關係數平方（Segall, 1996）。其他項指標的算法分述如下：

$$bias(\hat{\theta}_k) = \frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_0)}{n}$$

$$RMSE(\hat{\theta}_k) = \sqrt{\frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_0)^2}{n}}$$

其中 θ_0 爲受試者眞實能力值，而$\hat{\theta}_k$是由 MCAT 估計出來的受試者能力值，n 爲資料筆數，本研究中爲 1000 筆資料。

（二）研究程序

1. 產生模擬資料

模擬產生方式與研究一相同。

2. 估計多向度試題參數

試題參數估計方法與研究一相同

3. MCAT 執行程序

將上述估計出來的試題參數當作四組資料題庫中的試題參數，配合程序 1 所產生出來的受試者模擬反應，使用 ML、EAP、MAP 三種不同的能力估計方法來進行 MCAT。其中 EAP 的節點數量爲 10 點，此乃依研究一的結果，以不影響 EAP 估計精準度且每進行一題 MCAT 的時間不超過 1 秒的原則來決定。每種程序皆進行 20 題、40 題、60 題與 80 題的 MCAT。

（三）資料分析

分別針對三種能力估計法、四種不同的 MCAT 總題數、兩種不同的向度數以及兩種不同的能力相關係數，計算三個依變項在各向度的平均值，以了解這些因素對 MCAT 測量精準度的影響，以及隨著 MCAT 總題數增加時的變化情形。

（四）研究二結果

1. 二向度

表三是在二向度的情況下，以 ML、MAP 與 EAP 法進行不同題數 MCAT 的各向度能力平均信度、平均偏誤、均方根誤。在信度方面，當能力間爲低相關時，這三種方法所得到的能力估計信度很接近，只有在總題數 20 題時（約每向度 10 題），MAP 法與 EAP 法的信度略高於 ML 法；但當總題數超過 40 題時，這三種方法的信度幾乎相同。但是當能力間爲高相關時，MAP 與 EAP 法的信度明顯高於 ML 法，這個現象在題數較少時更明顯。而 MAP 與 EAP 法的信度只有在總題數 20 題時有些微差異，在總題數 40 題以上時這兩種方法的信度幾乎相同。

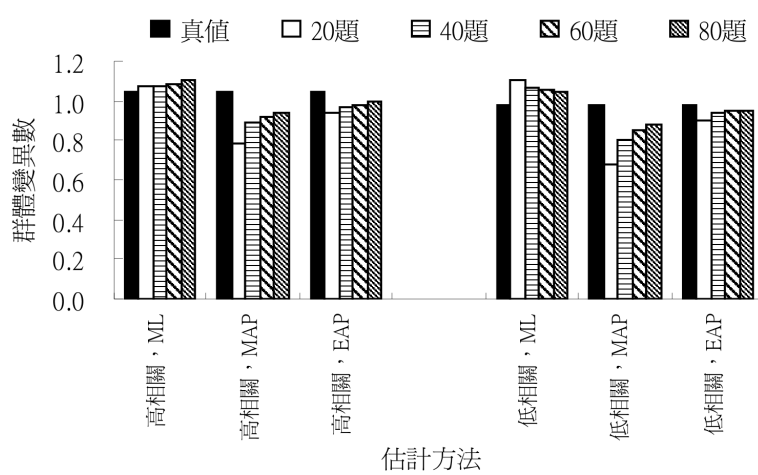
在平均偏誤方面，從表三中可以看出，這三種方法的能力估計偏誤都在 ± 0.03 以內。顯示出這三種能力估計方法的平均偏誤都很低。由於平均偏誤不易呈現出迴歸性偏誤的情形（正向與負向偏誤在相加時會互相抵消），因此本研究進一步分析以這三種估計法進行二向度 MCAT 時，受試群體之能力估計值變異數（見圖三）。如果在 MCAT 能力估計過程中產生迴歸性偏誤（高估低能力者的能力且低估高力者的能力），會使受試群體能力估計值的變異數變小。圖三的結果顯示，ML 法估計出來的受試群體變異數最接近原始眞值，而 MAP 與 EAP 法都會使受試群體能力估計值的變異數降低。這種

迴歸性偏誤在 MCAT 總題數較少時更加明顯；隨著總題數增加，MAP 與 EAP 的迴歸性偏誤逐漸減少。而 MAP 法的迴歸性偏誤又比 EAP 法嚴重。

在均方根誤（RMSE）方面，表三的結果顯示：ML 法的 RMSE 比 MAP 法及 EAP 法大，此差異在 MCAT 總題數較低時更爲明顯。而 MAP 法與 EAP 法的 RMSE 很接近，當題數在 40 題以上時，兩種方法的 RMSE 幾乎相同。在選題時間方面，二向度的 MAP 與 ML 法平均每執行 1 題 MCAT 所需的時間皆小於 0.001 秒，而 EAP 法（節點數 =10）約爲 0.005 秒。

表三　以三種估計法進行二向度 MCAT 之平均信度、平均偏誤與均方根誤 (RMSE)

MCAT 總題數	高相關，ML 法	高相關，MAP 法	高相關，EAP 法	低相關，ML 法	低相關，MAP 法	低相關，EAP 法
平均信度						
20 題	0.66	0.79	0.76	0.66	0.68	0.69
40 題	0.82	0.87	0.86	0.82	0.83	0.83
60 題	0.88	0.90	0.90	0.87	0.88	0.88
80 題	0.90	0.92	0.92	0.90	0.90	0.90
平均偏誤 (bias)						
20 題	0.00	0.00	0.00	0.00	0.03	-0.01
40 題	0.00	-0.01	-0.01	0.01	0.00	0.00
60 題	-0.01	-0.01	-0.01	0.00	0.01	0.01
80 題	-0.01	-0.01	-0.01	0.00	0.00	0.00
均方根誤 (RMSE)						
20 題	0.64	0.47	0.51	0.82	0.56	0.57
40 題	0.45	0.37	0.39	0.68	0.41	0.42
60 題	0.37	0.32	0.33	0.62	0.35	0.35
80 題	0.33	0.29	0.30	0.58	0.31	0.31



圖三　以三種估計法進行二向度 MCAT 時，受試群體之能力估計值變異數

2. 四向度

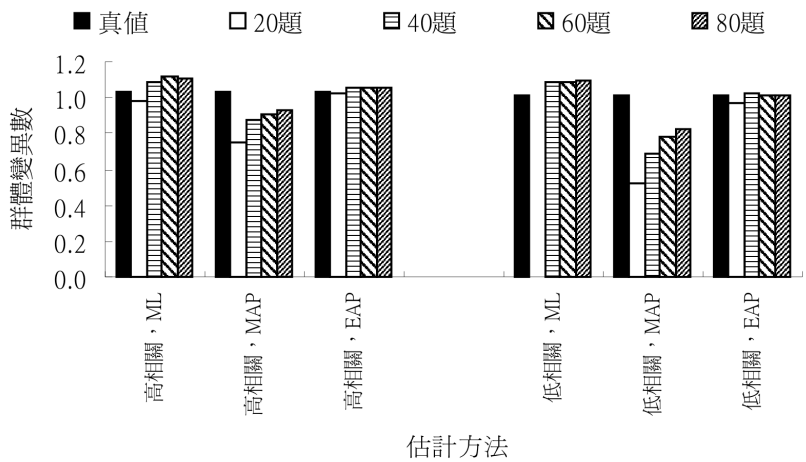
表四是在四向度的情況下，以 ML、MAP 與 EAP 法進行不同題數 MCAT 的各向度能力平均信度、平均偏誤、均方根誤。在信度方面，當能力間爲低相關時，這三種方法所得到的能力估計信度很接近；其中當總題數 20 題時，使用 ML 會因受試者在某些向度上全部答對或達錯而無法估計出其能力，因此在表中無法計算出來。但是當能力間爲高相關時，MAP 法與 EAP 法的信度明顯高於 ML 法，這種優勢在題數較少時更加明顯。而 MAP 法的信度略高於 EAP 法一些。

在平均偏誤方面，從表四中可以看出，除了當總題數爲 20 題時以 ML 法進行能力估計其平均偏誤會稍微大一點外，在其餘各種研究情境中，這三種方法的能力估計偏誤都在 ± 0.03 以內。顯示出這三種能力估計方法的平均偏誤都很低。由於平均偏誤不易呈現出迴歸性偏誤的情形（正向與負向偏誤在相加時會互相抵消），因此本研究進一步分析以這三種估計法進行四向度 MCAT 時，受試群體之能力估計值變異數（見圖四）。如果在 MCAT 能力估計過程中產生迴歸性偏誤（高估低能力者的能力且低估高力者的能力），會使受試群體能力估計值的變異數變小。圖四的結果顯示，ML 法與 EAP 法估計出來的受試群體變異數較接近原始眞值，而 MAP 法會使受試群體能力估計值的變異數變小。這種迴歸性偏誤在 MCAT 總題數較少時更加明顯；隨著總題數增加，MAP 的迴歸性偏誤逐漸減少。而在四向度 MCAT 中，MAP 法的迴歸性偏誤比在二向度 MCAT 中還要嚴重。

在均方根誤（RMSE）方面，表四的結果顯示：ML 法的 RMSE 比 MAP 法及 EAP 法大，此差異在 MCAT 總題數較低時更爲明顯。而 MAP 法的 RMSE 略低於 EAP 法一些。在選題時間方面，二向度的 MAP 與 ML 法平均每執行 1 題 MCAT 所需的時間皆小於 0.001 秒，而 EAP 法（節點數 =10）約爲 0.20 秒。

表四　以三種估計法進行四向度 MCAT 之平均信度、平均偏誤與均方根誤 (RMSE)

MCAT 總題數	高相關，ML 法	高相關，MAP 法	高相關，EAP 法	低相關，ML 法	低相關，MAP 法	低相關，EAP 法
平均信度						
20 題	0.43	0.75	0.73	--	0.53	0.53
40 題	0.66	0.84	0.82	0.68	0.70	0.69
60 題	0.76	0.88	0.86	0.77	0.78	0.78
80 題	0.82	0.90	0.89	0.82	0.83	0.82
平均偏誤 (bias)						
20 題	0.08	-0.00	-0.03	--	-0.00	-0.00
40 題	0.03	-0.00	-0.01	0.03	0.00	-0.00
60 題	0.02	-0.01	-0.01	0.02	-0.01	-0.01
80 題	0.01	-0.01	0.00	0.01	-0.01	-0.01
均方根誤 (RMSE)						
20 題	0.84	0.51	0.55	--	0.69	0.73
40 題	0.64	0.40	0.44	0.61	0.55	0.58
60 題	0.53	0.36	0.38	0.51	0.47	0.49
80 題	0.45	0.33	0.35	0.45	0.42	0.43



圖四　以三種估計法進行四向度 MCAT 時，受試群體之能力估計值變異數

討　　論

最大概似法、貝氏期望後驗法與貝氏最大後驗法是電腦化適性測驗中最常用的三種能力估計方法，它們的效果在單向度電腦化適性測驗中被研究的相當多（洪碧霞等人，民 81; Bock & Mislevy, 1982; Weiss & Mcbride, 1984），結果大都顯示出：MAP 與 EAP 法的能力估計信度較高且誤差較小，但是會有迴歸性的偏誤；而 ML 比較沒有迴歸性偏誤，但均方根誤較大，且全部答對或全部答錯者無法進行（在表四中，以 ML 法對四向度低相關能力進行 20 題的 MCAT 就因此而無法估計）。不過這三種能力估計方法在多向度電腦化適性測驗中的效果至今尚無定論。

本研究將三種能力估計方法在 MCAT 中的效果做有系統的比較。研究結果顯示，MAP 法與 EAP 法的能力估計信度比 ML 高，而均方根誤（RMSE）也比 ML 法低，而且隨著能力間的相關性愈高、能力向度數量愈多，或是當 MCAT 的總題數較少時，MAP 與 EAP 相對於 ML 的測量優勢更明顯。造成這樣的結果主要是因爲 MAP 與 EAP 法使用了各向度能力的先驗分佈（prior distribution），其共變數矩陣爲這兩種貝氏法提供了部份訊息量，幫助進行能力估計。當能力間相關愈高或是能力向度數量愈多時，此共變數矩陣所能提供的訊息愈高，因此這兩種能力估計法也比 ML 法的信度高出較多。不過其先決條件是先驗分布的共變數矩陣是正確的，如果受試者明顯不適用此共變數矩陣時，則這兩種貝氏估計法反而會讓能力均方根誤變大。

在 MCAT 總題數較少時，由於對受試者的能力估計均方根誤很大，因此大多仰賴來自於共變數矩陣的訊息，而 ML 法並沒有這方面的訊息，所以 MAP 法與 EAP 法在測驗總題數較低時其信度會比 ML 法高出許多；不過這樣的測量優勢在測驗題數較多時會漸漸消失，因爲此時受試者的作答反應愈來愈多，作答反應所提供的訊息漸漸增加，愈來愈不需要仰賴受試群體的先驗分佈。而且 MAP 法與 EAP 法的信度也已經相當高了，產生些微的天花板效應，因此 MAP 與 EAP 法的信度只略高於 ML 的信度一點點。

雖然 MAP 與 EAP 法可以提高信度，但是卻產生了能力估計的迴歸性偏誤。三種方法在平均偏誤方面並沒有明顯的差異；但是 MAP 法卻呈現出嚴重的迴歸性偏誤，而且在測驗題數較低時，這種迴

歸性偏誤更加明顯（見圖三與圖四）。EAP 法的迴歸性偏誤不像 MAP 法這般嚴重，可能是因爲本研究中所使用的 EAP 節點數僅爲每向度 10 點，也是因爲如此，在本研究中 EAP 的信度略低於 MAP。

在部份 IRT 軟體（例如 Bilog）中建議 EAP 的節點數最好爲 30 點，這對單向度 IRT 及 CAT 或許是可行的，但是對 MIRT 及 MCAT 是不可行的。因爲當向度數量增加到四向度時，如果節點數爲 30 點，平均每進行一題 MCAT 就需要花費 15 秒以上的時間來進行能力估計，更別說是向度數量超過四個向度的 MIRT 或 MCAT 了。因此，如果要以 EAP 法進行 MCAT，需要發展一些改良的作法來降低能力估計時間，並且又要保有其測量優勢。

在研究效度方面，本研究主要是以模擬資料的方式來進行。在本模擬資料研究中，題庫參數是完美的均勻分布，受試者的能力是多變量常態分佈且符合先驗分布，而其答題反應是依 MIRT 模式模擬產生的。雖然以模擬資料的方式進行研究可以對研究變項做較佳的控制，但在實際測驗情況中，題庫參數不一定是完美的均勻分布，受試者的能力也不見得符合先驗分布，答題反應也不見得符合 MIRT 模式。當題目參數不是均等分布時，或受試者能力不符合先驗分布時，三種方法的 MCAT 的能力估計結果是否還會如本研究所述，目前尚無所知。例如：過去有研究顯示在進行貝氏估計法時，先驗分布的正確性對能力估計結果的影響不大，尤其在題數愈來愈多時（洪碧霞等人，民 81；Chang & Twu, 2001; Robert, Donoghue, & Laughlin, 1999; Wainer & Thissen, 1987）。但這方面研究大多是以單向度 IRT 或 CAT 爲主。因此，未來還需要更近一步的研究或使用實際的 MCAT 資料來驗證。

整體來說，這三種方法在 MCAT 中各有其優缺點，雖然從整體信度與測量誤差來看，MAP 法是比較好的，但其迴歸性偏誤也是最嚴重的。EAP 的估計精準度與 MAP 差不多，但當向度數量較高時，選題與能力估計所需的時間太長。ML 則是 MCAT 中測量精準度較差的方法。因此，未來在 MCAT 中，研究者建議當能力向度數較少時（例如：少於四個向度）可以使用 EAP，以避免迴歸性偏誤的問題；但是當能力向度達到四個或四個以上時，最好使用 MAP 來進行，不過需注意其迴歸性偏誤的問題，最好能發展一些調整的作法來改良它，以降低其迴歸性偏誤。另一種折衷的作法是，先以 MAP 來進行選題與能力估計，最後的階段再用 EAP 來重新進行能力估計。這樣的好處是在選題過程中不會太慢而影響測驗的進行，而在最後又能以 EAP 來減少能力估計的迴歸性偏誤。

參 考 文 獻

洪碧霞、吳鐵雄、黃千綺、江秋坪、許宏彬（民 81）：能力估計法、題庫特質及終止標準對 CAT 考生能力估計影響之研究。測驗年刊，39 輯，249-267 頁。

陳柏熹、王文中（民 89a）：測驗組之題間多向度電腦化適性測驗。中華心理學會主辦「中華心理學會第三十九屆年會」宣讀之論文（台北）。

陳柏熹、王文中（民 89b）：題間與題內多向度電腦化適性測驗。中國測驗學會主辦「教育與測驗學術研討年會」宣讀之論文（台北）。

陳柏熹（民 90）：題數限制與曝光率控制對多向度電腦化適性測驗之測量精確性與試題曝光率的影響。國立中正大學心理學研究所博士論文。

陳柏熹、王文中（民 93）：曝光率控制對多向度電腦化適性測驗能力估計信度之影響：以 2001 年國中基本學力測驗資料爲例。教育與心理研究，27 卷，1 期，181-213 頁。

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 13,* 113-127.

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18,* 257-275.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Chang, S. H., & Twu, B. Y. (2001). Effects of changes in the examinees' ability distribution on the exposure control methods in CAT. *Psychological Testing, 48,* 167-189.

Fischer, G. H. (1973). The Linear logistic model as an instrument to educational research. *Acta Psychologica, 37,* 359-374.

Hambleton, R. J., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Klvwer Nijhoff.

Hattie, J. (1981). *Decision criteria for determining unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: The University of New England, Center for Behavioral Studies.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement, 20,* 155-168.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29,* 3-25.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20,* 389-404.

Master, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mckinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation, 15,* 389-390.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D., & Mckinley, R. L. (1991). The discrimination power of items that measure more than one ability. *Applied Psychological Measurement, 15,* 361-373.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1999). *Estimating parameters in the generalized graded unfolding model: Sensitivity to the prior distribution assumption and the number of quadrature points used.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, PQ, Canada.

Sand W. A., Water, B. K., & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: from inquiry to operation.* Washington, DC: American Psychological Association.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-345.

Sympson, J. B. (1978). A model for testing with the multidimensional items. In D. J. Weiss (Ed.). *Item response theory and computerized adaptive testing conference proceedings.* MN: University of Minnesota press.

Tseng, F. L. (2001). *Multidimensional adaptive testing using the weighted likelihood estimation: A comparison of estimation methods.* Unpublished doctoral dissertation, University of Pittsburgh, PA.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., et al. (Eds.) (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates publish.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339-368.

Wang, W. C. (1994). *Implementation and application of the multidimensional random coefficients multinomial logit model.* Unpublished doctoral dissertation, University of California, Berkeley, CA.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9,* 116-136.

Wang, W. C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized testing. *Applied Psychological Measurement, 28,* 295-316.

Weiss, D. J., & Mcbride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8*(3), 273-285.

Weiss, D. J. (Ed.) (1985). *Item response theory and computerized adaptive testing conference proceedings.* MN: University of Minnesota press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *Acer ConQuest.* Melbourne, Victoria, Australia: Australian Council for Educational Research press.

收 稿 日 期：2006 年 03 月 08 日
一稿修訂日期：2006 年 08 月 02 日
二稿修訂日期：2006 年 10 月 02 日
接受刊登日期：2006 年 10 月 03 日

Bulletin of Educational Psychology, 2006, 38 (2), 195-211
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

The Influences of the Ability Estimation Methods on the Measurement Accuracy in Multidimensional Computerized Adaptive Testing

Po-Hsi Chen

Department of Educational Psychology
and Counseling
National Taiwan Normal University

The goal of the research was to investigate the influences of ability estimation methods on multidimensional computerized adaptive testing. In stage 1, different quadrature points of the Baysian expected a posteriori (EAP) estimation were manipulated in order to find out the appropriate quadrature point of EAP in multidimensional computerized adaptive testing (MCAT). In stage 2, the maximum likelihood (ML) estimation, the Bayesian maximum a posteriori (MAP) estimation, and the EAP estimation methods were used in two kinds of ability dimensions (two and four dimensions) and two kinds of correlations between dimensions (high correlations and low correlations). The target item numbers of MCAT were 20, 40, 60, and 80. The dependent variables were the average reliability, bias, and the root mean square of error (RMSE) in all ability dimensions. Results in stage 1 indicated that the higher the quadrature point and the ability dimensions, the much higher the estimation time of MCAT. Ten points was appropriate in less than 4 dimensions of MCAT when the estimation time and the reliability of ability estimation were taken into consideration. Results of stage 2 indicated that MAP and EAP methods resulted in higher reliability and lower RMSE than ML method, especially in the conditions of high correlation between abilities, more ability dimensions, and fewer MCAT items. There were advantages and disadvantages in the three estimation methods. The regression bias of MAP, the estimation times of EAP, and the reliability and RMSE of ML were the problems that should be resolved when executing MCAT.

KEY WORDS: Bayesian expected a posteriori, Bayesian maximum a posteriori, maximum likelihood, multidimensional computerized adaptive testing