國立臺灣師範大學教育心理與輔導學系
教育心理學報，2011，42 卷，4 期，613-630 頁

模式錯誤假設對電腦化測驗的影響*

盧宏益　　徐永豐　　薛國松
輔仁大學　　銘傳大學　　輔仁大學
統計資訊學系　　資訊管理學系　　應用統計研究所

試題反應理論被廣泛地使用在電腦化適性測驗上，其以機率的觀點，透過試題反應模式，解釋考生能力與試題間的關係。藉由所選擇的試題反應模式，施測者可以根據不同的測驗目的編製適合的測驗。然而在實際的測驗情境中，試題反應模式通常是未知且須事先認定的。本研究旨在探討試題反應模式錯誤假設對測驗結果造成之影響，研究結果顯示，在常模參照測驗中，試題反應模式錯誤假設對考生能力估計所產生的偏誤比在真實模式下來的大，並造成測驗成本的增加，尤以真實測驗模式爲 3PLM 時最爲嚴重。在效標參照測驗中，試題反應模式錯誤假設對分類結果影響不大，但會造成測驗題數的增加，浪費施測成本。

關鍵詞：試題反應理論、電腦化適性測驗、模式錯誤假設

一、研究動機與目的

教學與評量在教學歷程中扮演重要的角色，教師透過教學設計，對學生進行適當且有效的教導，達到知識傳遞的目的。而藉由測驗評量的施行，教師得以了解學生的學習狀況，進而調整教學策略，有效地達成教學目標。評量的結果，除了提供教師了解學生的學習成效外，更重要的是，由學生的錯誤發生中，診斷出學習困難的所在及觀念的迷失混淆，進一步進行補救教學。藉由評量---教學---再評量的循環歷程修正錯誤，增進學習效能。

測驗可依分數的解釋方式分爲常模參照測驗（norm-referenced test）與效標參照測驗（criterion-referenced test）。常模參照測驗的分數解釋爲學生在團體中的相對位置，測驗的結果爲考生能力的估計或判定；而效標參照測驗則是著重於學生個人的表現是否達到預設的通過標準或精熟程度，測驗的結果爲通過或不通過（精熟或非精熟），因此效標參照測驗亦稱精熟測驗（mastery test）。施測者可以依據不同的教學目標及測驗目的，選擇適當的測驗型式。

測驗理論是一種解釋測驗資料與受試者之間關係的理論，傳統的測驗理論以古典測驗理論（classical test theory）爲主，測驗結果容易受到測驗樣本的影響。現代測驗理論以試題反應理論

* 本篇論文通訊作者：盧宏益，通訊方式：069201@mail.fju.edu.tw。

（item response theory, IRT）爲主（Lord, 1952），與古典測驗理論不同之處，IRT 以機率的觀點來解釋受試者能力與試題間的關係，具有試題獨立及樣本獨立的優點。隨著電腦資訊及測驗理論的發展，電腦化適性測驗（computerized adaptive testing, CAT）在最近幾十年來逐漸受到重視，已經逐漸取代傳統的紙筆測驗，成爲現代測驗的新趨勢。相關的測驗如 GRE、GMAT 及 TOEFL 等。IRT 被廣泛地使用在電腦化適性測驗上，藉由電腦的輔助，系統可以依據不同的考生程度，選擇最適合的題目。透過適性測驗的完整記錄歷程，系統可以迅速的估計考生的能力；結合適性學習系統，電腦並可以迅速地分析學生問題的所在，增進補救教學的學習成效。許多研究發現 CAT 可以節省許多測驗時間及成本，並有效提升考生能力估計的準確度（Lord , 1971; Wainer, 2000）。

CAT 的施行前提，必須事先建立題庫（item bank），在題庫的建立過程，需要經過預試的程序，執行試題校準（item calibration）的工作，亦即估計所有試題參數，以利 CAT 的施行，施測者可以根據不同的需要編制符合各種目標的測驗。試題校準的過程，首要爲針對不同類型資料與研究問題的瞭解，挑選適當的試題反應模式（item response model）。余民寧（1992）指出使用試題反應理論時，須先檢定模式與資料間是否具有滿意的適合度（goodness-of-fit），以確定選用的模式能適用於分析的資料，若模式與資料間適合度很差的話，則所計算得到的試題參數估計值和試題訊息函數（item information function），將會產生誤導的作用。在實際的情境中，試題反應模式是未知的，決策者的主觀認定錯誤或是預試樣本的偏誤（如預試樣本人數過少或是樣本代表性不夠）等原因，將使得試題反應模式假設錯誤；亦或隨著課程綱要的修訂，或是題庫中試題隨著測驗次數增多而存在高度曝光的問題，新試題的補充與線上校準（online calibration），更容易造成試題反應模式的認定錯誤。當試題反應模式假設錯誤時，將造成試題參數的估計及推論發生錯誤，估計量的統計推論性質及效率性亦發生變化。利用錯誤試題反應模式所得到具有校準偏誤的試題進行測驗，測驗的公平性必將受到質疑。本研究擬探討試題反應模式錯誤假設對測驗結果造成之影響。在常模參照測驗中，應探討對考生能力估計造成之影響，其爲估計問題；而在精熟測驗中，所欲瞭解的是對判斷考生達到精熟與否所造成之影響，實爲分類問題。

文獻探討

一、試題反應理論

現代測驗理論以試題反應理論（item response theory, IRT）爲主（Lord），與古典測驗理論不同之處，IRT 以機率的觀點來解釋受試者能力與試題間的關係。其假設受試者在某一題目上的反應情形，可以藉由某些潛在變項來解釋，此關係透過圖形上的呈現即爲試題特徵曲線（item characteristic curve, ICC）。 IRT 具有試題獨立及樣本獨立的優點，換言之，相關的參數估計不會受到不同的測驗題目以及不同的受試群體，產生不同的結果，以統計的術語，即爲不變性（invariant）。試題反應模式依據記分方式的不同可以分爲多元計分（multicategory scoring）模式及二元計分（dichotomous）模式。多元計分模型如 Nominal response model（Bock, 1972），Graded response model（Samejima, 1969）及 Partial Credit model（Master, 1982）等；二元計分模型如 Latent linear model（Lazarsfeld & Henry, 1968），Normal ogive model（Lord, 1952）及 Logistic model（Birnbaum, 1968; Lord & Novick, 1968; Rasch, 1960）等。本研究探討二元反應模型，在眾多二元

反應模型中，三參數羅吉斯模型（three-parameter logistic model, 3PLM）是常被使用的模型（Chang, 2004; Hambleton & Swaminathan, 1985）。3PL 模型可表示如下：

$$p(\theta) = P(x=1|\theta,a,b,c) = c + (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}, \quad (1)$$

式中 a 爲試題鑑別度參數（discrimination parameter），b 爲試題難度參數（difficulty parameter），c 爲試題猜測度參數（pseudo-guessing parameter），其中 D 爲一常數，當 D = 1.7 時，$p(\theta)$ 近似於 Normal ogive 模型下所算得之機率（Haley, 1952）。x 爲二元反應變數，答對爲 1，答錯爲 0。3PL 模型描述能力值爲 θ 的考生答對參數爲 $\beta = (a,b,c)$ 的試題之機率。當 c=0 時，即試題沒有猜對機率時，3-PL 模型簡化爲 2-PL 模型；如果我們進一步假設鑑別度參數 a 爲一固定常數，意即所有題目皆有相等的鑑別度，則此模型稱爲 Rasch model（Rasch, 1960）。本研究旨在羅吉斯迴歸模型的假設下，探討模式錯誤假設對測驗結果造成之影響。

二、SPRT

逐次分析（sequential analysis）之概念開始於二次世界大戰期間，當時的研究目的在於針對新武器的研發與改良，如何利用有限的試射次數評估新武器的有效性。有別於固定樣本數的抽樣設計，逐次分析將樣本數視爲隨機變數，利用現有樣本的特性決定後續的抽樣設計，並進行相關的統計分析。Wald（1947）提出逐次機率比檢定（Sequential Probability Ratio Test, SPRT），使得逐次分析的概念及理論研究逐漸受到重視。相關理論逐漸發展，並應用至許多不同的研究領域，如生物統計、臨床試驗及品質管制等。SPRT 爲一統計決策理論，最原始的概念爲處理簡單虛無假設及簡單對立假設的檢定問題。SPRT 的統計假設如下：

$$H_0 : \theta = \theta' - \delta_0 = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta' + \delta_1 = \theta_1, \quad (2)$$

其中 θ 爲未知參數，θ' 代表切點，θ_0 及 θ_1 分別代表虛無假設及對立假設的參數值，介於 θ_0 及 θ_1 之間的區域（$\theta' - \delta_0, \theta' + \delta_1$）稱爲灰色地帶（indifference region）。灰色地帶的設計爲避免當真實參數與切點非常接近時，樣本數會趨近於無窮大（Siegmund, 1985）。型一錯誤及型二錯誤的決策錯誤發生機率分別設定爲：$\alpha = p$（接受 $H_1|H_0$爲真）及 $\beta = p$（接受 $H_1|H_0$爲真）。令 $X_k = (x_1, x_2, ..., x_k)$ 爲累積至第 k 個樣本的所有觀測值，其中 x_i 代表第 i 個樣本觀測值。假設 x_i 二元反應變數，則 X_k 的概似函數可以表示爲

$$L_k(\theta, X_k) = \prod_{i=1}^{k} L(\theta; x_i)$$

$$= \prod_{i=1}^{k} [f(x_i;\theta)]^{x_i} [1 - f(x_i;\theta)]^{1-x_i} \quad (3)$$

式中f（ $x_i;\theta$ ）代表x_i的機率分配函數。則在H_1爲真時的概似函數值相對於在H_0爲真時的概似函數值所得之概似比（likelihood ratio）爲

$$LR_k(X_k)=\frac{\prod_{i=1}^{k}L(\theta_1;x)}{\prod_{i=1}^{k}L(\theta_0;x_i)} \qquad (4)$$

SPRT 在累積到第 k 個樣本時的決策法則如下：

（1）當$LR_k(X_k)\geq(1-\beta)/\alpha$ 時，則停止抽樣，統計決策爲接受H_1。

（2）當$LR_k(X_k)\leq\beta/(1-\alpha)$ 時，則停止抽樣，統計決策爲接受H_0。

（3）當$\beta/(1-\alpha)<LR_k(X_k)<(1-\beta)/\alpha$時，則不做決策，繼續抽取下一個樣本。

從以上的決策法則可以得知，SPRT 的決策邊界函數爲兩條平行線，下界函數爲$\beta/(1-\alpha)$，上界函數爲$(1-\beta)/\alpha$。當概似比$LR_k(X_k)$大於等於上界函數或小於等於下界函數時，即停止抽樣，並作出統計決策，否則繼續進行抽樣。假設試題的選取方式是隨機的，則我們可以分別證明出在H_0及H_1爲真時，所需的期望樣本數爲

$$E_0(n)\cong\frac{1}{\mu_0}\{\alpha\log(\frac{1-\beta}{\alpha})+(1-\alpha)\log(\frac{\beta}{1-\alpha})\} \qquad (5)$$

及

$$E_1(n)\cong\frac{1}{\mu_1}\{(1-\beta)\log(\frac{1-\beta}{\alpha})+\beta\log(\frac{\beta}{1-\alpha})\}, \qquad (6)$$

式中 $\mu_i=E_i[\log(f_1/f_0)]$，i = 0 或 1，f_0及f_1分別代表在H_0及H_1爲真時的機率分配函數（Siegmund, 1985）。

在 IRT 的假設下，本研究假設$f(x_i;\theta)$爲一 3PL 模型。欲檢定學生學習成效是否已達精熟水準，θ代表考生能力值，θ' 代表通過標準的分界點，θ_0 及 θ_1 分別代表非精熟者及精熟者能力的上界及下界。當H_0成立時，代表考生未達精熟水準；當H_1成立時，則認定考生已達精熟水準。兩種決策錯誤機率：型一錯誤發生機率$\alpha=P$（接受$H_1|$ H_0爲真）表示將未達精熟程度的考生誤判爲精熟者；型二錯誤發生機率$\beta=$（接受$H_1|$ H_0爲真）則表示將已達精熟程度的考生誤判爲非精熟者。令$X_k=(x_1,x_2,...,x_k)$爲累積至第 k 題的所有作答反應結果，其中x_i代表第 i 題的作答結果，答對爲 1，答錯爲 0。

SPRT 在第 k 題的決策法則如下：

（1）當$LR_k(X_k)\geq(1-\beta)/\alpha$時，則測驗終止，統計決策爲接受$H_1$，判定考生爲精熟者。

（2）當$LR_k(X_k)\leq\beta/(1-\alpha)$時，則測驗終止，統計決策爲接受$H_0$，判定考生爲非精熟者。

（3）當 $\beta/(1-\alpha) < LR_k(X_k) < (1-\beta)/\alpha$ 時，則不做決策，繼續進行下一題測驗。

當受試者能力很高或很低時，施測者可以很容易地判斷出受試者爲精熟者或是非精熟者，換言之，所需的樣本數會較少；反之，若受試者能力很接近通過標準的分界點時，則施測者需要更多的樣本數才足以判斷受試者是否爲精熟者。

Reckase（1983）、Kingsbury 與 Weiss（1983）與 Spray 與 Reckase（1996）將 SPRT 應用於精熟測驗上，利用灰色地帶的概念，採用假設檢定的方法處理精熟測驗中的分類問題。研究發現，在固定的 α 及 β 下，當灰色地帶愈小，所需的樣本數愈大。當類別數多於兩類時，Spray（1993） 將多類別的檢定問題轉換爲兩兩成對的檢定問題，以 SPRT 處理成對檢定，並進一步將其應用於多類別的精熟測驗判斷。相關的研究發現，相較於傳統的紙筆測驗而言，應用 SPRT 於精熟測驗上，可以節省更多的樣本數，並得到更高的正確分類率。

三、考生能力估計方法

對於考生能力的估計方法，常採用的方法有最大概似估計法 （maximum likelihood estimate）及貝氏估計法 （Bayesian estimate）。

（一） 最大概似估計法

令 $X_k = (x_1, x_2, ..., x_k)$ 爲考生累積至第 k 題的所有作答反應結果，其中 x_i 代表第 i 題的作答結果，其爲二元反應變數，答對爲 1，答錯爲 0。則 X_k 的概似函數 $L_k(\theta, X_k)$ 可以表示爲式（3）。最大概似估計法意指尋找參數 θ 的估計量 $\hat{\theta}$，使得當 $\theta = \hat{\theta}$ 時，可以讓概似函數 $L_k(\theta, X_k)$ 達到極大值。換言之，考生能力的最大概似估計量可以表示爲

$$\hat{\theta}_{mle} = \arg\max_{\theta} L_k(\theta, X_k). \qquad (7)$$

（二） 貝氏估計法

貝氏估計法可分爲貝氏最大後驗法（maximum a posteriori, MAP）及貝氏期望後驗法（expected a posteriori, EAP）。貝氏估計法的基本概念，係利用考生能力值的事前（prior）分佈用以修正概似函數。MAP 以考生能力的事前分布 $f(\theta)$ 作爲加權值，得到事後（posterior）機率分配爲

$$f(\theta \mid X_k) = \frac{L_k(\theta, X_k) \cdot f(\theta)}{g(X_k)} \qquad (8)$$

式中 $g(X_k)$ 爲考生的邊際機率密度函數。MAP 意指尋找參數 θ 的估計量 $\hat{\theta}$，使得當 $\theta = \hat{\theta}$ 時，可以讓概似函數 $f(\theta \mid X_k)$ 達到極大值，意即

$$\hat{\theta}_{MAP} = \arg\max_{\theta} L_k(\theta \mid X_k) \qquad (9)$$

而 EAP 與 MAP 類似，但所尋找的能力值是事後機率密度函數的期望值 （相當於平均數），而不是最大值 （相當於眾數）。EAP 估計量定義爲

$$\hat{\theta}_{EAP}=\sum_{i=1}^{r}\theta_i\frac{L_k(\theta_i,X_k)\cdot f(\theta_i)}{\sum_{i=1}^{r}L_k(\theta_i,X_k)\cdot f(\theta_i)} \quad (10)$$

式中 r 爲能力區間分割的段數，而 θ_i 則爲切割而得的切點。

在實際測驗情境中，考生能力分布是未知的，因此本研究採用最大概似估計法估計考生能力。然而最大概似估計法在面臨考生作答結果爲全對或是全錯的情況下，無法進行估計，因此本研究在發生上述情形時，採用 Chang 與 Ying（2004）的方法，當作答結果全數答對時，選取較已做過題目難度更高的題目施測；反之全數答錯時則選取更簡單的題目施測，結果呈現將快速的解決無法估計的問題。

四、模式錯誤假設相關文獻

在線性模型中，模式設定誤差主要分成兩種情形，分別爲遺漏重要的解釋變數（omitted variable）與納入不當的解釋變數（irrelevant variable）。Gujarati（1992）探討使用最小平方法（ordinary least squares, OLS）來估計模型的參數，結果發現若遺漏某些解釋變數會使得估計量產生偏誤，導致估計量爲不一致估計量（inconsistent estimator）。反之，若納入不相關的解釋變數且同時使用最小平方法來估計模型參數，所求出的估計量仍爲不偏（unbiased）且一致（consistent）估計量，但會造成有效性的損失，亦即會得到不偏但較無效率的估計式。

另一種模式錯誤假設情形爲模型本身函數形式（incorrect functional form）的錯誤設定。當真正模式爲非線性（nonlinear），但估計模式卻爲線性（linear），會導致估計量產生偏誤且爲不一致估計量。當解釋變數及反應變數間關係設定錯誤時，其係數估計式爲有偏估計式（Gujarati, 1992；Pindyck & Rubinfeld, 1998）。

Begg 與 Lagakos（1990）探討在羅吉斯迴歸下模式錯誤設定對計分檢定（score test）有效性（efficiency）的影響。結果顯示，不管模式是否有無共變數，暴露變數的誤設皆不會對計分檢定的效度（validity）造成影響。而當模式有共變數時，暴露變數與共變數的誤設將會導致計分檢定不再是有效的。此外，若模式中有重要的共變數被忽略會導致處理效應（treatment effect）產生偏誤估計，亦會減低反應變數與暴露變數關聯性檢定的有效性。Begg 與 Lagakos（1992）亦指出在羅吉斯迴歸下，重要的暴露變數與共變數的錯誤設定，會對反應變數與暴露變數的關聯性檢定產生不好的影響。

Attfield（1983）的研究亦對模型設定誤差做探討，發現當工具變數（instrumental variable）的選擇出現錯誤時，其所估計之變數的係數依然符合一致性，但會對檢定力和變數的效率性造成影響。謝曜安（1993）探討多放及少放工具變數對模型所產生之影響，結果顯示，即使選取了錯誤的工具變數組合，對於模型的係數估計並不會造成影響，其係數估計式仍然爲不偏估計式。

在測驗的應用上，Kalohn 與 Spray（1999）探討模式錯誤假設（model misspecification）對 SPRT 決策正確性的影響。研究探討假設真正試題反應模式爲 3PLM 時，卻使用 1PLM 來估計試題的參數所造成之影響。結果顯示，相較於正確模式，1PLM 有較大的決策錯誤；當測驗題數在不受限制下，使用錯誤的模式可能會導致型二錯誤的增加；當測驗題數在受限制下（100~200 題），使用錯誤的模式可能會導致型一錯誤的增加；使用不正確的模式會降低正確分類的比率以及與正確模式分類一致性的比率。Jiao 與 Lau（2003）探討模型誤失（model misfit）對電腦化分類測驗的

影響。結果顯示，當真正模式爲 1PLM 或 2PLM 時，錯誤使用其他模式進行 SPRT 檢定，對於分類決策沒有很大的影響。但當真正模式爲 3PLM 時，使用錯誤模式 1PLM 進行 SPRT 會產生較大的型一錯誤，而使用錯誤模式 2PLM 進行 SPRT 會產生較大的型二錯誤。

有關測量誤差的相關文獻探討，Gleser（1981）探討當自變數有測量誤差時，利用所得到的應變數及自變數資料配適線性迴歸模型時，所得的最小平方估計量並非爲不偏估計量；Stefanski 與 Carroll（1985）的研究中發現，利用存在測量誤差的自變數，在羅吉斯迴歸模型的假設下，所得到的估計量將不具有一致性（consistent）的性質。

方　法

本研究在羅吉斯模型的假設下，探討模式錯誤假設對測驗結果造成之影響，羅吉斯模型分爲單參數（1PLM）、雙參數（2PLM）及三參數（3PLM）三種。在每種模式下，分別比較模型正確假設與錯誤假設（分爲真實模型參數個數多於假設模型參數個數之 underspecification 及真實模型參數個數少於假設模型參數個數之 overspecification）對測驗所造成之影響。以 3PLM 爲例，本研究分別在模式正確假設（3PLM）及模式錯誤假設（1PLM 及 2PLM）時，對題庫中的試題進行參數估計，意即進行試題校準（item calibration）工作，進而將產生三個試題校準完成的題庫。爾後進行 CAT 時，將針對不同測驗情境，同時利用三個試題校準完成的題庫進行測驗，探討對測驗結果造成之影響。

一、研究工具

本研究所使用的軟體爲 BILOG-MG 與 MATLAB 7.0。本研究分爲兩階段，第一階段爲建立題庫及預試階段，模擬三種常用測驗模型的題庫，分別爲 1PLM、2PLM 與 3PLM，題庫的試題總數皆爲 1000 題。進一步進行預試階段，本研究模擬預試考生分佈爲較符合實際情況的常態分佈 $N(0,1)$，預試考生人數爲 8000 人（本研究同時模擬不同預試考生人數，結果發現當預試人數超過 5000 人時，所得估計結果非常接近）。之後利用 BILOG-MG 分別在正確模式及錯誤模式的假設下對題庫的題目進行參數估計（換言之，共計有九種組合情形，三種正確模式假設組合及六種錯誤模式假設組合）。第二階段爲實際測驗階段，使用在第一階段所得的試題參數資料，模擬各種測驗情境，利用 MATLAB 7.0 軟體撰寫程式進行測驗結果分析。

二、參數設定

試題參數設定方面，本研究參考相關文獻資料（Drasgow, 1989; Mislevy & Stocking, 1989; Skaggs & Stevenson, 1989; Baker, 1990; Stone, 1992）的參數範圍，設定鑑別度參數 a 服從均勻分佈 $U(0.5,2.5)$，難度參數 b 服從均勻分佈 $U(-3.3,3.3)$，猜測度參數 c 服從均勻分佈 $U(0,0.2)$。而考生的能力範圍及分佈，本研究分別假設考生能力服從 $N(0,1)$ 與 $U(-3,3)$ 兩種情形。

在電腦化測驗中，本研究同時模擬隨機選題及適性選題兩種選題策略，其中適性選題策略中採用最大 Fisher information 法作爲選題法則。意即利用現階段之能力估計值$\hat{\theta}$，進一步在題庫裡，由該考生尚未做過的考題中，挑選具有最大訊息量的試題加以施測。爲避免因爲測驗題數過少所造成的偏誤，本研究設定出初始測驗題數$n_0 = 5$。由於測驗初期對考生能力毫無所知，因此由題庫中均勻挑選出n_0題不同難度的測驗題數施測，爾後即進行適性選題及停止條件的判定（如前所述，當採用最大概似估計法時，在面臨考生作答結果爲全對或是全錯的情況下，採用 Chang 與 Ying（2004）的方法解決估計問題）。在考生的能力估計過程中，採用累積試題訊息量作爲測驗終止條件，本研究採用 Chang（2001）的停止準則如下：

$$T_d = \inf\{n \geq 1 : I(\hat{\theta}_n) \geq (\frac{z_{\alpha/2}}{d^*})^2\} \qquad (11)$$

其中$I(\hat{\theta}_n)$爲考生估計能力值$\hat{\theta}_n$累積的試題訊息量，若測驗題數很大時，對於考生真實能力θ之信賴區間$\hat{\theta}_n \pm \frac{z_{\alpha/2}}{\sqrt{I(\hat{\theta}_n)}}$可包含$1-\alpha$覆蓋機率。本研究設定$d^*$爲 0.5，型一錯誤及型二錯誤發生機率設定爲$\alpha = \beta = 0.05$。

結　果

本研究分別比較模型錯誤假設對常模參照測驗 （估計） 及效標參照測驗 （分類） 所造成之影響。

一、對估計的影響

本節內容爲探討試題反應模式錯誤假設下，對受試者的能力估計與測驗長度影響之分析比較，本研究設定受試者人數爲 1000 人，模擬次數爲 100 次。本研究同時模擬二群不同能力分佈的考生，分別爲$N(0,1)$與$U(-3,3)$。

表 1 至表 3 分別爲當真實模式爲 1PLM，2PLM 及 3PLM 時，利用不同估計模式所得的試題參數，對考生能力進行估計所得的能力值均方根誤差與平均測驗長度之比較表。表 1 爲真實模式 3PLM 下，不同估計模式之結果比較表，粗體代表利用正確模式估計所得之結果。由表 1 得知，當考生能力爲標準常態分佈$N(0,1)$時，且選題策略爲隨機選題下，利用正確估計模式 3PLM，能力值均方根誤差 RMSE 爲 0.2513，較錯誤估計模式 1PLM、2PLM 小；平均測驗長度約爲 54 題左右，

亦較錯誤估計模式 1PLM、2PLM 小。由此可知，當真實模式爲 3PLM 時，在建立題庫時使用正確模式進行試題參數估計，可以使得考生能力的估計偏誤最小，測驗成本也最少。而使用錯誤模式（underspecification），將使得 RMSE 明顯增大，平均測驗長度也明顯的增加，尤以假設模式爲 2PLM 更爲嚴重。而使用適性選題策略所得的平均測驗長度明顯較低，但 RMSE 略爲增加，原因爲試題參數本身的估計偏誤原就會使得考生能力估計偏誤增大 （即使使用正確模式亦然），而適性選題策略會讓測驗題數明顯下降，因此估計偏誤更容易顯現。當考生能力爲均勻分佈 $U(-3,3)$ 時，RMSE 皆高於當考生能力爲標準常態分佈 $N(0,1)$ 的情況，原因爲能力介於兩端的考生（低能力與高能力）的比例較高，因爲題庫的參數範圍較小，適合兩端考生的題目也相對較少，因此 RMSE 較高。例如，能力爲 $N(0,1)$ 且隨機選題下，使用正確估計模式 3PLM，能力介於兩端的考生 （能力介於-3~-2 及 2~3 的考生）的 RMSE 分別爲 0.2914 與 0.2749，而整體受試者的 RMSE 爲 0.2513；而能力爲 $N(0,1)$ 且適性選題下，錯誤估計模式 2PLM，能力介於兩端的考生（能力介於-3~-2 及 2~3 的考生） 的 RMSE 分別爲 0.9491 與 0.7641，高於整體考生的 RMSE 爲 0.4504。

表 1　測驗模式 3PLM 下，不同估計模式之結果比較表

估計模式	真實模式 3PLM							
	受試者能力 $\theta \sim N(0,1)$				受試者能力 $\theta \sim U(-3,3)$			
	隨機選題		適性選題		隨機選題		適性選題	
	測驗題數	均方根誤差	測驗題數	均方根誤差	測驗題數	均方根誤差	測驗題數	均方根誤差
3PLM	**54.29 (19.14)**	**0.2513 (0.0070)**	**11.59 (2.23)**	**0.2973 (0.0120)**	**69.66 (53.90)**	**0.2496 (0.0072)**	**11.98 (2.51)**	**0.2917 (0.0080)**
1PLM	60.81 (11.32)	0.2957 (0.0050)	26.95 (2.31)	0.3608 (0.0096)	70.76 (20.84)	0.3258 (0.0065)	27.65 (3.15)	0.5346 (0.0072)
2PLM	92.19 (110.58)	0.4244 (0.0243)	20.98 (24.61)	0.4504 (0.0227)	180.61 (223.76)	0.7630 (0.0253)	41.66 (51.77)	0.7637 (0.0234)

註：表格內數字爲重覆 100 次實驗所得結果的平均值，括號內爲標準誤。

表 2 爲真實模式 2PLM 下，不同估計模式之結果比較表，粗體代表利用正確模式估計所得之結果。由表 2 中可以發現，當真實模式爲 2PLM 時，使用正確模式及錯誤模式所得考生能力估計 RMSE 相距不大，主要反應在平均測驗題數。當使用錯誤模式 3PLM 時（overspecification），平均測驗題數略高於正確模式，但差異不大，原因是 2PLM 爲 3PLM 的特例，因此不會有太大的差異；當使用錯誤模式 1PLM 時（underspecification），平均測驗題數則明顯的增加。

表 3 爲真實模式 1PLM 下，不同估計模式之結果比較表，**粗體**代表利用正確模式估計所得之結果。由表 3 中可以發現，當真實模式爲 1PLM 時，使用正確模式及錯誤模式所得考生能力估計

RMSE 及平均測驗題數相距不大。原因同上，係因 1PLM 爲 2PLM 及 3PLM 的特例。當使用錯誤模式爲 2PLM 時（overspecification），相較於使用正確模式所得之結果，RMSE 及平均測驗題數結果差異皆不大；當使用錯誤模式爲 1PLM 時（overspecification），平均測驗題數則略爲增加。由上述結果可知，當真實模式爲 1PLM 時，真實模式與錯誤模式的參數個數相差愈多時，測驗結果差異則愈明顯。

表 2　測驗模式 2PLM 下，不同估計模式之結果比較表

估計模式	真實模式 2PLM							
	受試者能力 $\theta \sim N(0,1)$				受試者能力 $\theta \sim U(-3,3)$			
	隨機選題		適性選題		隨機選題		適性選題	
	測驗題數	均方根誤差	測驗題數	均方根誤差	測驗題數	均方根誤差	測驗題數	均方根誤差
2PLM	**40.97（13.39）**	**0.2463（0.0082）**	**10.52（1.49）**	**0.2487（0.0058）**	**47.40（17.08）**	**0.2519（0.0072）**	**10.93（1.60）**	**0.2592（0.0058）**
1PLM	60.72（10.76）	0.2288（0.0039）	26.39（1.54）	0.2277（0.0063）	70.77（21.05）	0.2546（0.0047）	26.79（1.77）	0.3012（0.0099）
3PLM	43.57（26.51）	0.2464（0.0055）	10.48（1.85）	0.2572（0.0042）	61.23（46.35）	0.2645（0.0090）	11.75（2.85）	0.2933（0.0082）

註：表格內數字爲重覆 100 次實驗所得結果的平均值，括號內爲標準誤。

綜合來說，模式錯誤假設會導致考生能力估計偏誤及測驗成本的增加。當錯誤模型爲 underspecification 時，RMSE 及平均測驗題數皆明顯增加，尤以真實模式爲 3PLM 最爲嚴重；當錯誤模型爲 overspecification 時，估計結果影響較小，主要影響爲平均測驗題數的增加，而當真實模式與錯誤模式的參數個數相差愈多時，影響程度愈明顯。

二、對分類的影響

本節內容探討試題反應模式錯誤假設下，對效標參照測驗所造成之影響。本研究以 SPRT 作爲判定考生是否達到精熟標準的決策準則，分別針對正確模式與錯誤模式假設的情況，進行分類結果與測驗長度之分析比較。本研究設定考生人數爲 1000 人，模擬次數爲 100 次。本研究同時模擬

表 3　測驗模式 1PLM 下，不同估計模式之結果比較表

估計模式	真實模式 1PLM							
	受試者能力 $\theta \sim N(0,1)$				受試者能力 $\theta \sim U(-3,3)$			
	隨機選題		適性選題		隨機選題		適性選題	
	測驗題數	均方根誤差	測驗題數	均方根誤差	測驗題數	均方根誤差	測驗題數	均方根誤差
1PLM	**59.23 (9.68)**	**0.2521 (0.0052)**	**26.49 (1.68)**	**0.2542 (0.0067)**	**66.43 (16.38)**	**0.2506 (0.0065)**	**26.62 (1.72)**	**0.2575 (0.0040)**
2PLM	60.87 (9.77)	0.2488 (0.0055)	25.88 (1.60)	0.2549 (0.0077)	69.65 (18.28)	0.2493 (0.0039)	25.94 (1.63)	0.2624 (0.0042)
3PLM	66.43 (41.06)	0.2598 (0.0065)	25.74 (6.17)	0.2704 (0.0064)	80.58 (46.46)	0.2697 (0.0100)	30.52 (14.67)	0.3143 (0.0069)

註：表格內數字爲重覆 100 次實驗所得結果的平均值，括號內爲標準誤。

二群不同能力分佈的考生，分別爲 $N(0,1)$ 與 $U(-3,3)$。而 SPRT 的相關參數設定中，型一錯誤及型二錯誤的發生機率分別設定爲 $\alpha = \beta = 0.05$，精熟標準的門檻值（threshold） 爲 $\theta = 0$，亦即檢定 $H_0 : \theta \le 0$ vs. $H_1 : \theta > 0$，又根據式（2），則可將統計假設化簡爲簡單假設檢定，並設定 $\delta_0 = \delta_1 = 0.1, 0.3, 0.5$。

表 4 爲真實模式 3PLM 下，不同的灰色地帶設定，SPRT 分類比較表，粗體代表利用正確模式估計所得之結果。由表中可以發現 δ 的大小與平均測驗成反比，換句話說，欲達到愈精確的分類結果，則需要愈多的試題數方得以完成 （Wald, 1947）。當 δ 小時，代表精熟與非精熟的標準愈接近，因此需要更多的試題方得以正確地鑑別出受試者精熟與否 （從分類的觀點，當兩類別愈接近時，愈難區分出所屬的類別）。以 $\delta = 0.1$ 爲例，當真實模式爲 3PLM 時，在正確模式及錯誤模式的假設下，正確分類比率皆在九成以上，並無太大的差異。主要差異反應在平均測驗題數上，使用錯誤模式 （underspecification） 會使得平均測驗題數增加，造成測驗成本的提高，尤以 1PLM 更爲明顯，意即影響程度隨真實模式與錯誤模式的參數個數差增加而增加。

由表 4 中可以看出，當 $\theta \sim N(0,1)$ 時，所需的試題數較 $\theta \sim U(-3,3)$ 來的大，原因爲相較於 $\theta \sim U(-3,3)$ 而言，當考生能力分佈爲 $\theta \sim N(0,1)$ 時，能力值介於灰色地帶的考生比例高出許多，因此需要更多的試題方得以正確地鑑別考生達到精熟與否。實際的測驗情境中，測驗長度不可能太長，由表中可以看出不同大小的灰色地帶，其正確分類比率差異不大，主要反應在所需的測驗題數，因此建議施測者可以根據式（5）及式（6）考量實際測驗時間及成本，決定適合的參數。

表 4　測驗模式 3PLM 下，SPRT 精熟判定比較表

灰色地帶 δ	估計模式	真實模式 3PLM							
		$\theta \sim N(0,1)$				$\theta \sim U(-3,3)$			
		隨機選題		適性選題		隨機選題		適性選題	
		測驗題數	正確分類比率	測驗題數	正確分類比率	測驗題數	正確分類比率	測驗題數	正確分類比率
0.1	**3PLM**	**255.02 (261.26)**	**0.9776**	**61.64 (57.39)**	**0.9953**	**149.52 (195.09)**	**0.9905**	**43.63 (30.85)**	**0.9995**
	2PLM	274.89 (276.84)	0.9283	60.80 (79.75)	0.9607	162.40 (210.11)	0.969	41.20 (40.50)	0.9837
	1PLM	284.06 (275.60)	0.9593	107.92 (116.12)	0.9881	153.66 (212.03)	0.9821	60.10 (51.55)	0.9994
0.3	**3PLM**	**88.06 (119.94)**	**0.9401**	**26.36 (31.37)**	**0.9421**	**51.21 (83.22)**	**0.9752**	**20.18 (15.60)**	**0.9778**
	2PLM	98.50 (137.88)	0.9639	25.19 (41.20)	0.9586	57.71 (98.64)	0.984	18.04 (29.68)	0.9808
	1PLM	95.87 (123.20)	0.9142	42.43 (57.94)	0.9578	54.05 (88.54)	0.9655	27.18 (40.16)	0.9817
0.5	**3PLM**	**46.81 (57.14)**	**0.9019**	**15.91 (12.21)**	**0.9030**	**28.36 (40.92)**	**0.9579**	**13.91 (9.41)**	**0.9557**
	2PLM	52.09 (66.68)	0.9604	14.50 (17.75)	0.9674	31.54 (48.57)	0.9823	11.36 (10.98)	0.9848
	1PLM	49.28 (56.94)	0.8747	23.79 (23.48)	0.9209	29.22 (41.83)	0.948	16.09 (15.72)	0.9677

註：表格內數字爲重覆 100 次實驗所得結果的平均值，括號內爲標準誤。

表 5 爲真實模式 2PLM 下，SPRT 分類比較表。由表 5 可以看出當真實模式爲 2PLM 時，使用正確模式及錯誤模式時，正確分類比率差異亦不大。當使用錯誤模式爲 3PLM 時（overspecification），平均測驗題數與使用正確模式所需測驗題數差異不大；而當使用錯誤模式爲 1PLM 時（underspecification），所需測驗題數則明顯增加。而表 6 爲真實模式 1PLM 下，SPRT 分類比較表。由表 6 可以看出，當真實模式爲 1PLM 時，使用正確模式及錯誤模式時，正確分類比率差異亦不大。而當使用錯誤模式爲 2PLM 及 3PLM 時（overspecification），平均測驗題數與使用正確模式所需測驗題數差異不大。

表 5　測驗模式 2PLM 下，SPRT 精熟判定比較表

灰色地帶	估計模式	真實模式 2PLM							
		$\theta \sim N(0,1)$				$\theta \sim U(-3,3)$			
		隨機選題		適性選題		隨機選題		適性選題	
		測驗題數	正確分類比率	測驗題數	正確分類比率	測驗題數	正確分類比率	測驗題數	正確分類比率
0.1	**2PLM**	**201.42 (239.29)**	**0.9792**	**40.93 (55.88)**	**0.9934**	**110.27 (174.27)**	**0.9913**	**28.60 (35.14)**	**0.9976**
	3PLM	200.91 (239.82)	0.9735	40.80 (53.32)	0.9905	109.63 (171.50)	0.9889	34.02 (36.64)	0.9959
	1PLM	277.54 (277.60)	0.9795	99.86 (94.65)	0.9991	154.05 (209.01)	0.9938	65.62 (71.09)	0.9997
0.3	**2PLM**	**65.97 (93.93)**	**0.942**	**17.16 (24.46)**	**0.9393**	**36.48 (66.12)**	**0.9756**	**12.56 (18.03)**	**0.9747**
	3PLM	64.71 (91.55)	0.9356	16.78 (18.83)	0.9352	36.09 (65.17)	0.9702	13.97 (17.22)	0.9721
	1PLM	100.87 (142.06)	0.9457	42.71 (55.58)	0.9507	55.04 (100.95)	0.9776	27.57 (40.01)	0.9777
0.5	**2PLM**	**34.96 (43.88)**	**0.9037**	**11.10 (6.98)**	**0.9027**	**20.29 (32.16)**	**0.9578**	**9.05 (5.13)**	**0.9547**
	3PLM	34.23 (42.71)	0.8956	10.93 (6.97)	0.8996	20.07 (31.15)	0.956	9.35 (5.44)	0.9538
	1PLM	53.97 (71.62)	0.9076	25.82 (28.65)	0.9017	30.23 (50.36)	0.9611	17.16 (20.97)	0.9573

註：表格內數字爲重覆 100 次實驗所得結果的平均值，括號內爲標準誤。

表 6　測驗模式 1PLM 下，SPRT 精熟判定比較表

灰色地帶	估計模式	真實模式 1PLM							
		$\theta \sim N(0,1)$				$\theta \sim U(-3,3)$			
		隨機選題		適性選題		隨機選題		適性選題	
		測驗題數	正確分類比率	測驗題數	正確分類比率	測驗題數	正確分類比率	測驗題數	正確分類比率
0.1	**1PLM**	**274.61（275.57）**	**0.9747**	**102.65（98.40）**	**0.9965**	**153.66（208.31）**	**0.9902**	**66.91（68.95）**	**0.9983**
	3PLM	270.40（239.82）	0.9646	101.67（97.60）	0.9901	152.87（206.25）	0.9845	68.51（70.21）	0.9965
	2PLM	279.71（277.14）	0.9753	105.47（105.88）	0.9969	156.14（209.63）	0.99	68.36（76.63）	0.9992
0.3	**1PLM**	**96.11（131.12）**	**0.9403**	**43.47（54.45）**	**0.9436**	**52.79（92.13）**	**0.9751**	**27.85（39.76）**	**0.9749**
	3PLM	93.35（126.23）	0.9257	41.58（49.81）	0.9295	52.49（90.15）	0.968	27.66（34.88）	0.9681
	2PLM	98.09（134.51）	0.9407	44.34（58.15）	0.9448	54.66（95.54）	0.9751	28.02（42.14）	0.9759
0.5	**1PLM**	**50.34（62.20）**	**0.9006**	**25.19（27.05）**	**0.8995**	**28.70（44.06）**	**0.9576**	**16.86（18.37）**	**0.9560**
	3PLM	49.33（60.46）	0.8846	23.579（24.94）	0.8815	28.49（43.39）	0.9517	16.40（16.59）	0.9497
	2PLM	51.68（64.34）	0.9014	24.85（26.78）	0.9009	29.77（46.36）	0.9577	16.98（18.65）	0.9559

註：表格內數字爲重覆 100 次實驗所得結果的平均值，括號內爲標準誤。

綜合上述，可以發現無論題庫的真實模式爲何（3PLM、2PLM 或 1PLM），使用正確與錯誤的試題反應模式，對於 SPRT 分類結果並無顯著的影響。當使用錯誤模式爲 underspecification 時，會造成所需測驗題數明顯增加，影響程度隨真實模式與錯誤模式的參數個數差增加而增加。當使用錯誤模式爲 overspecification 時，所需的測驗題數則差異不大。

討　論

試題反應理論被廣泛地使用在電腦化適性測驗上，藉由所選擇的試題反應模式，施測者可以根據不同的測驗目的編制適合的測驗。然而在實際的情境中，試題反應模式是未知的，試題反應

模式的錯誤假設將使得試題參數的估計及推論發生錯誤，進一步影響測驗結果。本研究旨在探討試題反應模式錯誤假設對測驗結果造成之影響。

在常模參照測驗中，探討試題反應模式錯誤假設對考生能力估計造成之影響。綜合來說，模式錯誤假設會導致考生能力估計偏誤及測驗成本的增加。當錯誤模式爲 underspecification 時，RMSE 及平均測驗題數皆明顯增加；當錯誤模式爲 overspecification 時，估計結果影響較小，主要影響爲平均測驗題數的增加，且當真實模式與錯誤模式的參數個數相差愈多時，影響情形愈明顯。另外，從統計模型的角度思考，1PLM 及 2PLM 爲標準型羅吉斯模型，而 3PLM 則不是。當真實模式爲 3PLM 時，錯誤模型的假設爲模式型式的錯誤，因此造成的影響較真實模式爲 1PLM 及 2PLM 時爲大。

在效標參照測驗中，探討試題反應模式錯誤假設對判斷考生達到精熟與否所造成之影響。本研究使用 SPRT 作爲決策法則，SPRT 以概似函數衡量未知參數在各類別發生的可能性，利用概似比概念比較相對發生的可能性進行檢定，不涉及參數的估計。結果顯示，當模型錯誤假設時，對於 SPRT 分類結果並無顯著的影響。主要反應在測驗題數的增加。當使用錯誤模式爲 underspecification 時，會造成所需測驗題數明顯增加，影響程度隨真實模式與錯誤模式的參數個數差增加而增加。當使用錯誤模式爲 overspecification 時，所需的測驗題數則差異不大。

本研究在羅吉斯模型的假設下，探討試題反應模式錯誤假設對測驗結果造成之影響，建議後續研究者可以探討不同類型模型（如 Normal Ogive 等）的錯誤假設，亦或不同形式模式的誤設對測驗結果造成之影響。而關於考生的能力估計，本研究採用最大概似估計法，在考生能力分布的假設下，可採用貝氏估計法進行估計，可避免最大概似估計法所面臨的估計問題。再者，本研究採用最大 Fisher information 法作爲適性選題策略，亦可嘗試不同的選題策略，如 Fisher Interval Information, Fisher Information with a Posterior Distribution, Kullback-Leibler Information 及 Kullback-Leibler Information with a Posterior Distribution 等。而根據不同測驗目的，題庫參數的設計，以及選題策略中 item exposure control 和 content balance 的考量，亦爲本研究主題的後續研究重點。

參考文獻

余民寧 （1992）：試題反應理論的介紹（五）－模式與資料間的適合度。**研習資訊，9**（4），6-10。

謝曜安（1993）：**資金成本之模型誤設－台灣實證研究**。輔仁大學經濟學研究所碩士論文。

Attfield, C. L. F. (1983). Consistent estimation of certain parameters in the unobservable variable model when there is specification error. *Review of Economics and Statistics, 65*, 164-167.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement, 14*, 139-150.

Begg, M. D., & Lagakos, S. W. (1990). On the consequences of model misspecification in logistic regression. *Environmental Health Perspectives, 87*, 69-75.

Begg, M. D., & Lagakos, S. W. (1992). Effects of mismodeling on tests of association based on logistic regression models. *Annals of Statistics, 20*, 1929-1952.

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 tests scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Chang, Y.-C. I. (2001). Sequential confidence regions of generalized linear models with adaptive designs. *Journal of Statistical Planning and Inference*, *93*, 277-293.

Chang, Y.-C. I. (2004). Application of sequential probability ratio test to computerized criterion-referenced testing. *Sequential Analysis, 23*(1), 45-61.

Chang, Y.-C. I., & Ying, Z. (2004). Sequential estimation in variable length computerized adaptive testing, *Journal of Statistical Planning and Inference,121*(2), 249-264.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Gleser, L. J. (1981). Estimation in a multivariate "errors in variables" regression model: Large sample results. *The Annals of Statistics, 9*, 24-44.

Gujarati, D. N. (1992). *Essentials of econometrics* (2nd ed.). New York, NY: McGraw-Hill.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report No.15). Palo Alto, CA: Applied Mathematics and Statistics Laboratory, Stanford University.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Jiao, H., & Lau, A. C. (2003). *The effects of model misfit in computerized classification test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Kalohn, J. C., & Spray, J. A. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement, 36*, 47-59.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-Based adaptive mastery and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing* (pp. 258-288). New York, NY: Academic Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227-242.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (4th ed.). Boston, MA: Irwin, McGraw-Hill.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: The University of Chicago Press.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in Testing : Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

Siegmund, D. (1985). *Sequential analysis : Tests and confidence intervals.* New York, NY: Springer-Verlag.

Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. *Applied Psychological Measurement, 13*(4), 391-402.

Spray, J. (1993). *Multiple-category classification using a sequential probability ratio test* (Research Report 93-7). Iowa, IA: ACT.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405-414.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, *13*(4), 1335-1351.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wald, A. (1947). *Sequential analysis*. New York, NY: Dover.

收 稿 日 期：2009 年 09 月 21 日
一稿修訂日期：2009 年 11 月 20 日
二稿修訂日期：2010 年 03 月 02 日
接受刊登日期：2010 年 03 月 02 日

Bulletin of Educational Psychology, 2011, 42(4), 613-630
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

The Effect of Model Misspecification on Computerized Testing

Hung-Yi Lu
Department of Statistics and Information Science
Fu Jen Catholic University

Yung-Feng Hsu
Department of Information Management
Ming-Chuan University

Kuo-Sung Hsueh
Graduate Institute of Applied Statistics
Fu Jen Catholic University

Item response theory (IRT) has been widely applied in computerized adaptive testing (CAT) with the logistic type models most often used. IRT prescribes an item characteristic curve that provides the probability of an examinee correctly answering an item with a parameter of a given ability level. Examiners can develop various tests for different purposes based on a chosen item response model. However, in actual testing practice, the priori item response model is often unknown. The purpose of this study is to examine the effect of model misspecification on computerized testing. Using norm-referenced testing, results indicated that model misspecification has an effect on the estimate of examinees' abilities. Both the RMSE and test length significantly increased when the wrong item response models were used, especially when item bank belongs to three-parameter logistic model. In criterion-referenced testing, model misspecification has no effect on the accuracy of classification. However, it will increase test length and cost of testing.

KEY WORDS: computerized adaptive testing, item response theory, model misspecification