國立臺南大學
「教育研究學報」第 47 卷第 1 期（民國 102 年）：113~134
DOI：10.3966/199044282013044701006

題組效果對參數估計精確度之影響

賴佩伶

高雄市美濃區吉洋國小 教師

涂柏原[1]

國立臺南大學教育學系 教授

摘要

本研究旨在探討題組效果對參數回復性（item parameter recovery）的影響，以模擬資料探討傳統 IRT 模式、TRT 計分模式及 bi-factor 模式等三種模式在不同題組（testlet）條件下，對受試者能力以及試題參數估計回復性的比較，同時也探討 Yen（1984）的Q_3統計數對於試題局部依賴性(local item dependence)的偵測。每一個測驗皆包含 6 個題組，每個題組皆為 5 題，而三個題組效果（0、0.6、1.2）以及兩種樣本人數（500 人、1000 人）被操弄來產生模擬的試題反應向量，每種條件組合重複 100 次，所產生的試題反應資料分別以上述的三種校準估計的模式進行分析，並計算Q_3值。本研究主要發現如下：（1）在各種模擬條件下，以題組模式(testlet-based model)的試題參數回復性表現良好；以 TRT 模式的估計最佳，其次為 bi-factor 模式估計，再者為傳統 IRT 模式。（2）人數越大，參數估計的回復性也較好。隨機題組斜率影響越大，估計造成的 RMSE 也越大。（3）假定有相依關係的題組內試題配對的Q_3值明顯大於假定是獨立的題組間的試題配對所得到的Q_3值，顯示對於偵測存於試題之間的相依情形，Q_3的表現不錯。

關鍵字：題組、局部依賴性、Q_3、題組反應模式、雙因素模式

[1] 通訊作者 e-mail：bortwu@gmail.com

壹、緒論

廣泛應用在教育與心理測驗上的試題反應理論（item response theory[IRT]），可根據受試者的潛在變項與試題參數（如難度、鑑別度）之關係的函數，來解釋受試者在某一試題作答反應的表現（Wang, Cheng, & Wilson, 2005）。利用 IRT 模式估計試題參數時，不受考生能力的影響，適合用來發展題庫；考生的能力估計也不受試題特性影響，適合用來進行能力分數的等化，而訊息量（information）概念更可以反映出測驗對不同能力者的不同測量精準度，IRT 儼然已成爲當代測驗發展時所主要依賴的方法。

受試者在進行測驗時，接受題目刺激需要相當多的心理程序和時間，一組使用相同刺激的試題，可減少收集訊息時的時間，因此題組式的試題早在數十年前就常被用在教育測驗。在 IRT 的範疇中，題組這個名詞的概念是 Wainer 和 Kiely（1987）所提出來的，從此在心理計量的領域中受到許多的關注與討論。題組可視爲是一個小型測驗，小至一個題組僅包含一個試題，最大可以一個測驗中僅有一個題組（Wainer & Kiely, 1987; Wainer & Lewis, 1990; Lee, Dunbar, & Frisbie, 2001）。一般認爲題組式的測驗能評量受試者高層次思考及解決問題的能力，並提升測驗的建構效度（Allen & Sudweeks, 2001; Zeniskey, Hambleton, & Sireci, 2002）。

近年來，題組式的測驗被廣泛應用在大型測驗上，如國內的國中基測和大學指考、著名的托福、PISA 和 NAEP 等。以單一向度 IRT 的架構來分析資料，需符合單向度（unidimensionality）和局部獨立（local independence）的假設，但許多研究（Chen & Thissen, 1997; Lee, 2000; Wainer, Bradlow, & Du, 2000; Wainer & Kiely, 1987; Wainer, & Wang, 2000; Wainer & Wilson, 2005; Yen, 1993）皆指出，題組式的測驗違反了 IRT 的基本假設。題組試題除依賴一個共同刺激外，還會受到其他因素影響（如專門知識的主題、錯誤圖解的刺激、疲勞等等），因此試題間並不是局部獨立的（Yen, 1993）。爲了解決題組式試題測驗的計分問題，Wainer、Bradlow 和 Wang（2007）等人延伸傳統試題反應理論，加上隨機題組效果 $\gamma_{id(j)}$，提出題組試題反應理論（testlet response theory[TRT]）。DeMars（2006）指出 TRT 理論基本上是 bi-factor 模式（Gibbons & Hedeker, 1992）的一個特例，因此亦有學者利用 bi-factor 模式來分析題組的效果。

過去許多研究者發現，當試題不符局部獨立時卻使用獨立試題的模式，所估計的試題參數和能力值估計都可能不是正確的（Yen, 1993; Bradlow, Wainer & Wang, 1999; Wainer, & Wang, 2000）；隨著測驗欲測量更多、更複雜的行爲反應，要精準的估計受試者的表現，選用適當地作答反應模式的重要性亦不容小覷。忽略局部獨立性會導致高估信度，且低估能力估計的標準誤（Wainer, 1995; Wainer & Wang, 2000; Yen, 1993），以致於錯估試題參數，影響受試者能力估計的精確度。當一組試題存有局部依賴性時，難度估計依舊良好，但試題鑑別度會時而高估或低估（Ackerman, 1987; Bradlow, Wainer, & Wang, 1999; Wainer & Wang, 2000）。Wainer 等人也表示，當模式中忽略題組依賴性時，能力和難度的回復性（recovery）會比鑑別度和猜測度（低漸近線）來的好；而題組模式試題鑑別度的回復性又比傳統的三參數試題反應模式來好（Glas, Wainer, & Bradlow, 2000）。Dresher（2004）指出，在多點計分題組或題組效果模式的比較中，若忽略題組試題依賴性時，估計所得的能力均方根誤差值（root mean square error[RMSE]）會較高；但在題組效果模式下考慮題組試題依賴性時，試題難度和鑑別度的 RMSE 值會較小。

綜上所述，題組試題中局部依賴性對參數估計的影響頗大，適當地使用正確的估計模式是重要的，當前研究多著重在探討獨立試題模式和多點計分題組試題的比較（Lee, Brennan, & Frisbie, 2001; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Yen, 1993），或是獨立試題模式和題組效果模式的比較（Bradlow et al., 1999; Glas et al., 2000; Wainer et al.,

2000; Wainer & Wang, 2000）。Li、Bolt 和 Fu（2006）與 DeMars（2006）學者等人應用 bi-factor 模式於題組試題模式的相關比較，但尚未針對試題依賴性、總人數等題組相關特性做探討，故本研究欲分兩大部分進行探討。首先，以模擬資料操控題組相關特性，探討在傳統 IRT、題組試題反應理論（TRT）以及 bi-factor 模式估計下，能力參數與試題參數的回復性；其次，探討 Q_3 統計數（Yen, 1984）是否能檢測出存在於題組內試題之間的相依情形。

貳、文獻探討

一、題組相關概念

試題是構成測驗的核心部份，其性能的優劣會影響整個測驗的品質，測驗試題的類型繁多，如選擇題、是非題、簡答題、填充題和配合題等，要使測驗有效的評量到欲測量的特質，命題者在命題時需要了解各種題型的優缺點。爲了達到測量高層次的認知能力，並有效評量學生在概念理解的完整性，近年來有越來越多測驗傾向使用「題組型」試題（郭生玉，1998）。

過去學者對「題組型」試題賦予多種不同名稱，如 Cureton（1965）的超級試題（superitems; 引自 Haladyna, 1992）、Wilson & Adams（1995）的題集（item bundle）、Ferrara, Huynh & Baghi（1997）的題束（item clusters）、Yen（1993）的段落（passages），以及 Wainer 等學者的題組（testlet）（Wainer & Kiely, 1987; Wainer & Lewis, 1990）等，其中題組似乎是目前最廣被接受和使用的。題組的定義也隨不同學者觀點而有所差異，Wainer 和 Kiely（1987）應用在電腦適性測驗上，認爲若測驗中的試題是透過共同刺激材料（stimulus material）的題幹（stem）、試題架構（item structure）或試題內容（item content）加以連結，成爲一群相關的試題，這樣的試題群稱之爲「題組」。Wainer 與 Lewis（1990）進一步將題組定義爲小測驗（small tests），小至讓研究者可以操弄，又大至可涵蓋題組本身的內容。Lee、Brennan 和 Frisbie（2000）將題組定義爲某一測驗中試題的子集合（subset of items），這集合在測驗結構、實施和計分時被視爲測量的單位。無論題組的定義爲何，就測驗建構觀點而言，題組必須包含一段落、圖表、或其他刺激材料，並在刺激後跟隨一群試題，受試者必須依賴此相同刺激作答相關試題（Lee, 2000）。

實際上，題組中的試題可以是二元計分和多元計分試題（像是數學或科學測驗，在題組形式下可能包含建構反應試題，需多點計分）。而以刺激爲本位的題組，試題必須依賴在同一刺激材料，在現代理論觀點下，有可能導致試題間依賴性的問題。但由於題組形式可以測量高層次思考，並應用於多種題型上，因此，目前許多大型的標準化成就測驗或國家證照考試，皆採用此種測驗類型測驗來評量學生的能力。例如：美國國家教育進展評量(National Assessment of Educational Progress[NAEP])、國際閱讀素養進展研究(Progress in International Reading Literacy Study[PIRLS])、國際學生評量計畫(Programme for International Student Assessment[PISA])等大型評量；托福（Test of English as a Foreign Language[TOEFL]）或英語檢定測驗。此外，我國的國中生基本學力測驗和大學學測和指考等，也都使用了題組的測驗形式。這類測驗成績多爲學生申請入學的必備條件之一，故如何對題組式測驗進行正確且適當的分析是相當重要的議題，因此瞭解題組試題

相關特性所造成的影響，如何正確分析題組式試題在理論與實務上，均具有其研究的價值。

當測驗中的試題根據一個共同文章或共同刺激因素時，往往存在有條件依賴性，因而違反局部獨立性的假設（Lee & Frisbie,1999; Yen,1993）。若測驗中存有違反局部獨立性的試題，利用傳統的 IRT 模式，可能會高估受試者的能力，得到有偏差的試題參數（Chen & Thissen,1997; Yen,1993）。測驗若是違反局部獨立性，將對試題的信效度、試題訊息量（item information）、參數估計等產生影響（Chen & Thissen, 1997; Yen, 1984, 1993）。在應用傳統的 IRT 模式進行試題分析時，判斷試題間是否違反局部獨立性是重要的。

Yen（1981）提出Q_1統計量，爲一傳統卡方考驗比較適配度所用的統計數指標；van den Wollenberg（1982）提出一種 Rasch 模式的適配度測量，稱爲Q_2統計量，適用於單參數模式；稍後 Yen（1984）又提出了Q_3統計量，以表示同一能力水準兩兩試題分數間的關係；Yen 並以模擬和實徵資料比較Q_1、Q_2、Q_3三種統計量的表現，她發現Q_1無法有效判斷參數模式的適配度，而Q_2易受到多向度的影響，只有Q_3是較客觀的評估統計法，不受樣本大小波動的影響（Yen, 1984）。Chen 和 Thissen（1997）指出Q_3在潛在局部相依（underlying local dependence）和表面試題相依（surface local dependence）情況下的表現都較其他指標良好。

常見的偵測指標還包括 Pearson χ^2 和概似比（likelihood ratio）G^2 等。Chen 和 Thissen（1997）以χ^2、G^2、Q_3和ϕ_{diff}等統計量偵測試題局部獨立性和多向度，發現χ^2和G^2的值越大，表示違反局部獨立性愈高，其中ϕ_{diff}可指出關聯的趨勢，正值表示有較大依賴性。林欣怡（2007）探討 Kim、Cohen 與 Lin（2005）所提出的多元計分反應模式下的四種偵測指標χ^2、G^2、Q_3和Z_d，在多元計分試題的資料下和二元計分試題的資料中之表現情形，結果發現這四個指標在多元計分的情形下，皆能偵測出違反局部獨立性之試題配對，而在二元計分速度測驗資料的情境中亦同，其中G^2和χ^2的表現近似，Q_3與Z_d則較不受題數影響。

Q_3統計數主要是從試題分數中移除受試者能力θ的影響，計算試題殘差分數間的相關（Yen, 1984）。詳細地說，若令$d_{ik} = u_{ik} - \hat{P}_i(\hat{\theta}_k)$，其中$u_{ik}$是受試者$i$在試題$k$的分數，$\hat{P}_i(\hat{\theta}_k)$是受試者$i$在試題$k$的期望分數，$d_{ik}$即爲受試者$i$在試題$k$中觀察分數與期望分數間的差異（即爲殘差），則$Q_3$的公式如下：

$$Q_{3jk} = r_{d_j d_k} \tag{1}$$

其中，$r_{d_j d_k}$表試題j和試題k殘差分數的相關。若符合單向度假定，Q_3值幾乎等於零；若局部獨立但不符合單向度假設，Q_3呈現負值表示測量到不同潛在構念，也就是測驗資料是多向度的，Q_3若呈現正值表示試題間測量的構念是相同的，但是局部獨立這個假定爲獲得滿足，表示有局部相依的現象存在；若其絕對值越大，表示違反「單一向度且局部獨立性」這個假定的可能性越高（Habing, Finch, & Roberts, 2005）。當局部獨立爲真

時，對Q_3值進行 Fisher 的z轉換後，Q_3期望值會接近$-1/(K-1)$，其中K爲試題長度（Yen, 1993）。因此本研究將應用Q_3統計量檢驗資料間局部獨立性的關係，以探討產生模擬的資料時，研究者所操弄的題組效果是否可以由Q_3統計量加以檢測出來。

二、題組反應理論

試題反應理論以數學模式來描述試題參數和受試者的能力與其答題反應之間的關係。由於 IRT 的能力估計不受試題參數影響，試題參數的估計也不受受試者能力所影響，因此可以使接受不同測驗的受試者能力可以放在相同的量尺上進行比較（陳柏熹，2005）。過去二十年來，IRT 被應用在許多方面，包括測驗(量表) 編製、分數等化、題庫建置、電腦化適性測驗等。依試題計分方式，IRT 的模式可分爲二元計分（dichotomous）模式與多點計分（polytomous）模式。依估計的試題參數的個數又可分爲單參數模式、二參數模式、三參數模式（Birnbaum, 1968; Lord, 1980）。以三參數模式（three-parameter model[3PL]）爲例，其公式如下（Birnbaum, 1968）：

$$P(y_{ij}=1)=c_j+(1-c_j)\frac{1}{1+e^{-1.7a_j(\theta_i-b_j)}} \tag{2}$$

其中，y_{ij}爲受試者i在試題j的分數，$P(y_{ij}=1)$是受試者i回答試題j正確的機率，θ_i爲受試者i的能力，a_j、b_j、c_j分別爲試題j的鑑別度參數、難度參數、及猜測參數。

基於IRT的單向度（unidimensionlity）與局部獨立性（local independency）假設，只要試題符合IRT模式，接受不同難度試題的受試者之能力是可以互相比較的（余民寧, 1992; 陳柏熹, 2005; Hambleton & Swaminathan, 1985）。但根據相同刺激材料的題組試題，包含了題組對受試者的影響力，使受試者答題反應違反IRT局部獨立的假設（Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, & Kiely, 1987;Wang & Wainer, 2005; Yen,1993），傳統的IRT理論似乎無法有效地處理這些新型的試題，因此能夠處理題組資料的IRT模式遂漸漸被提出（Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2004）。

Wainer、Bradlow 和 Du（2000）將傳統試題反應理論加以延伸，加入一隨機效果γ用以表示題組試題間的相關程度，主要目的爲解決違反局部獨立性的情形，把題組試題間局部依賴的程度加以模式化，因此他們在公式（2）中增加了一個描述受試者和題組間互相影響的隨機題組效果，成爲新的題組模式：

$$P(y_{ij}=1)=c_j+(1-c_j)\frac{1}{1+e^{-1.7a_j(\theta_i-b_j-\gamma_{id(j)})}} \tag{3}$$

其中$\gamma_{id(j)}$表示受試者i和題組$d(j)$間的交互作用，其餘參數的意義如前所述。

爲使參數估計容易，Wainer、Bradlow 和 Wang（2000）等學者把三參數題組模式嵌入較大的階層貝氏架構（hierarchical Bayesian framework）中，主要的參數被假定具有下列的先驗分配（prior distribution）：

$$\theta_i \sim N(0,1)$$

$$\gamma_{id(j)} \sim N(0, \sigma^2_{\gamma_{d(j)}})$$

$$\log(a_j) \sim N(\mu_a, \sigma^2_a)$$

$$b_j \sim N(\mu_b, \sigma^2_b)$$

$$\log\left[c_j / (1-c_j)\right] \sim N(\mu_c, \sigma^2_c)$$

其中$\sigma^2_{rd(j)}$指的是題組$d(j)$的題組效果總和。當$\sigma^2_{rd(j)}$越大，表示題組造成的分數變異程度較大。這架構可運用在所有受試者下試題與題組的訊息，將數值的不確定性加以模式化，之後再以馬可夫鏈蒙地卡羅法（Markov chain Monte Carlo methods[MCMC]）運算方式迭代取得各參數的估計值。

受試者根據同樣的材料作答多個試題，各試題間除了共同受到文章材料的影響，還共同受到潛在特質的影響，因此題組試題反應也可視爲多向度模式的一種，主要的特質是試題所要測量的考生能力，而次要特質是受到共同文章材料特定內容影響的部分。Bi-factor 模式最適合用於這類的內容（DeMars, 2006; Li, Bolt & Fu, 2006），在 bi-factor 模式，每一試題反應是一主要特質和其中一種次要特質的函數；各個試題的次要特質彼此間，或者和主要特質之間，是互相獨立的（Gibbons & Hedeker, 1992）。題組內各個試題間的相依情形（或題組特徵），基本上可以用次要特質來加以模式化，因此 bi-factor 模式可用下面的多向度 3PL 模式（M3PL）的公式來加以表示（DeMars, 2006）：

$$P = c_j + (1-c_j)\frac{1}{1+e^{-1.7\mathbf{a}_j(\boldsymbol{\theta}_i - \mathbf{b}_j)}} \tag{4}$$

其中，$\mathbf{a}_j$、$\boldsymbol{\theta}_i$和$\mathbf{b}_j$都是向量。若將傳統的 3PL 模式加入一隨機題組效果，則 TRT 模式即可寫成：

$$P = c_j + (1-c_j)\frac{1}{1+e^{-1.7a_j(\theta - b_j - \gamma_{g(j)})}} \tag{5}$$

Li、Bolt 和 Fu（2006）認爲公式（5）中題組特徵$\gamma_{g(j)}$和主要能力θ是互相獨立的，使用相同的的鑑別度參數可能會有問題，因此建議把主要能力和題組特徵的鑑別參數區分開來。

重新整理公式（5），把隨機效果$\gamma_{g(i)}$量尺化爲平均數 0，標準差 1，且以$\theta_{g(i)+1}$來表示，其中$\alpha_{g(j)}$爲一題組常數（testlet constant），如果$\alpha_{g(j)}$等於$\gamma_{g(j)}$的標準差，則a_j和$\alpha_{g(j)}$的相乘積是題組斜率（testlet slope）。公式（5）可以改寫成底下的型式（DeMars, 2006）：

$$P = c_j + (1-c_j)\frac{1}{1+e^{-1.7a_j(\theta_1-\alpha_{g(j)}\theta_{g(j)+1}-b_j)}} \quad (6)$$

這是 TRT 的另一種型式。基本上，公式（6）與公式（4）近似，可以說公式（6）是公式（4）的一個特殊型式。因此 Li 等人（2006）認爲 Wainer、Bradlow 和 Du（2000）的 TRT 模式可以用 bi-factor model 來加以表示，且 Li 等學者（2006）認爲，在 bi-factor 模式下，題組斜率和主要斜率（primary slope）是互相獨立的，因此，題組模式可說是 bi-factor 模式的一個特例（DeMars, 2006）。

Wainer 和 Wang（2000）以題組模式來分析 86 個 TOEFL 題組（50 題閱讀理解測驗、36 題聽力理解測驗），每一題組得到一個變異數 γ。經由與傳統模式所得的標準試題參數來進行比較，可發現難度 b 在兩種不同模式估計（BILOG 與 MCMC with Gibbs Sampler）中皆有良好的估計。但鑑別度 a 和猜測度 c 的估計在條件獨立是錯誤的假定時有偏差，最重要的是，在條件獨立是錯誤的假定時，測驗訊息量被大大的超估。

Lee (2000) 對 1995 年四、七年級愛荷華基本技能測驗（Iowa Test of Basic Skills[ITBS]）之受試者資料進行分析，來說明 item-based 方法和 testlet-based 方法間的差異，Lee 以字彙測驗和模擬的資料爲基準，與題組組成的測驗做比較（閱讀理解、地圖、圖表測驗），用以比較五種模式估計測量標準誤的方法，對題組組成的測驗採用題組定義來估計，調查其適合度和應用性。結果發現 item-based 方法比 testlet-based 方法提供較低的 SEM 估計，item-based 方法會低估測量標準誤（standard error of measurement[SEM]），測驗模式中違反假設的程度會影響 SEM 的估計，而 item-based 估計方法的誤差隨條件依賴增加而越大。

Li、Bolt 和 Fu（2006）利用模擬的和閱讀測驗實徵資料來比較以貝氏方法估算傳統二參數模式、bi-factor 模式和 Rasch 模式的參數時之適配情形，他們發現 bi-factor 模式有較佳的適配度，分析的結果並可提供研究者檢測哪些試題或類型對哪些受試者存有缺失；因此 Li、Bolt 和 Fu 建議以 bi-factor 的概念分別處理整體能力和題組因素的鑑別度參數，是較爲適切的。DeMars（2006）操弄 2 種不同的試題長度（25 及 50 題）、5 種題組斜率等因素，並分別以不同模式產生模擬的資料，進行 bi-factor 模式、題組效果模式、多點計分模式以及獨立試題模式等四種模式能力估計精確度的比較。結果發現題組效果模式嵌在 bi-factor 模式內，獨立試題模式嵌在題組效果模式內。對極端程度的受試者而言，估計模式的不同會造成差異，又以多元計分的差異最大，這和過去研究指出多元計分題組會造成某些訊息遺失的結果相似（Wainer & Wang, 2000; Zenisky, Hambleton, & Sireci, 2002）；另外，試題長度越長，估計得到的能力值差異越小。DeMars 並以 PISA 2000 年公開的實徵資料進行分析，發現不同模式產生的估計能力值關聯性極高（至少達 0.99），假如題組效應小，使用不同的模式所得的信度將無差異。

目前關於題組效果的研究大多著重在題組效果被忽略或題組效果的大小對 IRT 能力估計的影響，因此本研究同時探討題組效果對於試題與能力參數估計回復性的影響。Glas、Wainer 和 Bradlow（2000）應用 5000 筆反應資料比較 MCMC 和 MML 的估計結果，發現 MML 和 MCMC 的試題參數估計在題組模式下呈現高相關，而在 3PL 模式和 3PL 題組模式下的 MML 試題參數估計相關則較低（請參見表 1）。

表 1　試題參數間的相關

參數	MCMC	Testlet MML	BILOG
a			
MCMC	1.00	.95	.86
Testlet MML	.95	1.00	.82
BILOG	.86	.82	1.00
b			
MCMC	1.00	.95	.98
Testlet MML	.95	1.00	.93
BILOG	.98	.93	1.00
c			
MCMC	1.00	.87	.75
Testlet MML	.87	1.00	.85
BILOG	.75	.85	1.00

註：取自” MML and EAP estimation in testlet-based adaptive testing.”, by A. W. Glas, H. Wainer & E. T. Bradlow, 2000, *Computerized adaptive testing: Theory and practice*, p. 276..

在這個研究中雖然使用的軟體可能運用不同的估計方法，但是由這個研究的結果來看，在題組模式下，MML 和 MCMC 所估計得到的試題參數之間相關頗高，不同估計方法對題組試題估計結果的影響應該不大。

參、研究方法

本研究主要目的在探討當試題中含有題組效果而違反 IRT 局部獨立性的假定時，對於試題參數以及個人能力參數估計回復性的影響，以及由 Yen（1984）所提出來的 Q_3 統計數是否能夠檢測出研究者所操弄的題組效果，整個研究的架構大致如圖 1 所示：

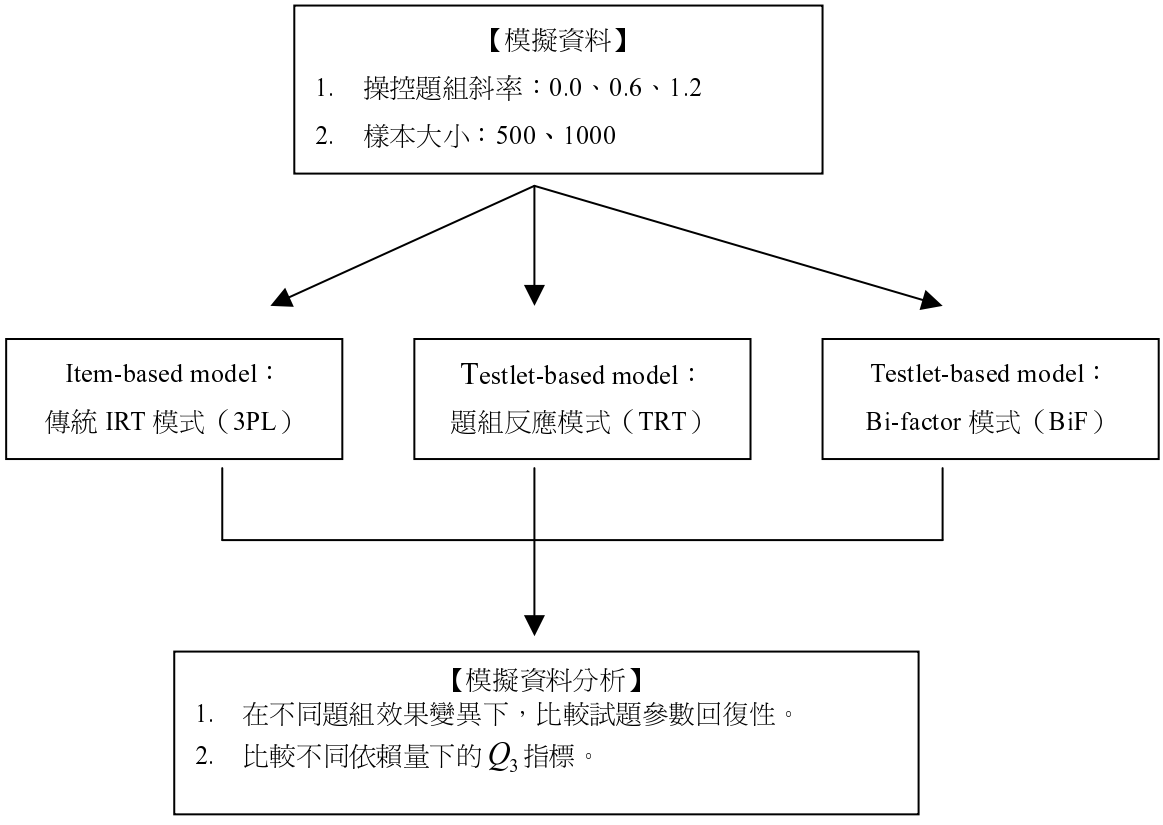

圖 1　研究架構圖

一、模擬資料設計

本研究所有的資料都是利用模擬的方式產生的，因此參考 Li、Bolt 與 Fu（2006）和 Wainer、Bradlow 與 Du（2000）之作法，將各種參數的真值所具有的分配設定如下：$\theta \sim N(0,1)$，$a \sim N(0.8,0.2^2)$，$b \sim N(0,1)$，$c \sim N(0.2,0.03^2)$，以及 3 種題組斜率（0、0.6 和 1.2），以產生具有題組效果的資料。根據研究設計，爲方便起見，每一個資料集都具有 6 個題組，每個題組有 5 個題目，因此一共有 30 個試題；另外操弄 2 種受試者人數（500 和 1000 人），所有組合條件各模擬 100 次，以探討不同題組相關特性在三種模式下的反應，以及對試題參數估計的精確度情形。整個模擬的情形，如表 2 所示。

表 2　**模擬資料分配表**

題組數量	每題組試題數	人數	設定題組斜率
6	5	500 人 1000 人	0、0.6、1.2

DeMars（2006）指出，獨立試題模式基本上是受限或巢套於（nested in）題組模式內，題組模式又受限於 bi-factor 模式內，複雜度高的模式，其適配度優於約束模式（reduced model），當數據取自約束模式時，雙方差異性並不顯著，故本研究使用 bi-factor 模式的參數設定方式產生模擬資料，並利用 SPSS 的語法來完成。

在單一試題計分的 IRT 模式下，使用二元計分 3PL 模式校準估計，採 MML 程序，以 BILOG-MG 軟體進行資料分析；在題組效果模式下，TRT 計分模式使用三參數題組估計，參數估計使用貝氏結構下 MCMC 估計，以 SCORIGHT 3.0 軟體進行估計分析，迭代次數設定爲 4000 次，burn-in 3000 次迭代後收斂；而 bi-factor 模式，採 MML 估計程序，使用 TESTFACT 軟體估計。模式設計整理如表 3：

表 3　**模擬模式設計整理**

估計模式	傳統 3PL model	題組反應模式 TRT	Bi-factor model
簡稱	3PL	TRT	BiF
使用軟體	BILOG-MG	SCORIGHT	TESTFACT
參數估計	MML	MCMC	MML

二、分析方法

爲了解試題參數和能力參數的回復性，每一情境下的參數估計值與真值間的均方根誤差（root mean square error[RMSE]）被計算，其公式如下：

$$RMSE(\hat{\varsigma}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\varsigma}_r - \varsigma)^2} \tag{7}$$

其中，R 表示重複次數，ς、$\hat{\varsigma}$ 分別代表試題參數或是能力參數的真值與估計值。RMSE 值和估計精確度成反比例，可視其爲估計精確度的指標（Wang & Wilson, 2005）。

此外，本研究以 Q_3 值來檢視試題間局部依賴性的表現，Q_3 值爲觀察值與估計值兩

兩試題間殘差的相關，計算方法如前述的公式（1）。根據由各種情境資料所得到之能力參數與試題參數估計來求得期望反應，並與觀察值比較以計算殘差；當各題的殘差之間的相關係數絕對值較大時，相關係數的樣本分配會呈現明顯的偏態，不適合採用傳統的 t 檢定的方法來進行假設檢定。因此，習慣上會將 Q_3 值加以 Fisher 的 z 轉換，然後利用公式（8）來進行轉換過的 Q_3 值之假設檢定：

$$z = \frac{z_r}{\frac{1}{\sqrt{N-3}}} \tag{8}$$

其中 z_r 爲轉換後的 Q_3 值，N 爲試題配對數，而 z 近似標準常態分配，因此當顯著水準訂爲.05 時，±1.96 就是判斷虛無假設是否被保留或是被拒絕的臨界值，當虛無假設被拒絕時，表示 Q_3 的值顯著地不等於 0，也就是該對中的試題是相依的。

肆、研究結果與討論

一、不同估計模式下試題參數回復的情形

因爲 TESTFACT 軟體不估計試題猜測度（c）參數，如欲將 c 值放入模式中，必須由研究者自行提供，故本研究不探討猜測度的回復性，僅針對鑑別度（a）和難度（b）以及能力值（θ）的回復性進行探討。在各種條件組合下，重複 100 次模擬所估計得到的能力參數以及試題參數之平均數之描述統計結果，呈現在表 4 之中。由表 4 平均數那一欄的數據來看，題組模式（TRT、BiF）估計下得到的估計平均值都較爲相似，差異不大，但在忽略題組影響的獨立試題模式（3PL）估計下，試題鑑別度和難度的平均數都較題組模式估計來得大。

表 4　各個校準方法估計得到之參數描述統計表

估計模式	人數	參數	最大值	最小值	平均數	標準差
3PL	500	a	2.420	0.331	1.076	0.344
		b	2.994	-2.167	0.644	0.964
		θ	2.328	-2.472	0.000	0.874
	1000	a	2.348	0.046	1.024	0.328
		b	2.782	-1.851	0.572	0.957
		θ	2.930	-1.823	0.009	0.817
TRT	500	a	3.274	0.197	0.999	0.355
		b	2.511	-3.017	0.121	1.014
		θ	2.3826	-2.4867	0.001	0.869
	1000	a	2.734	0.174	0.993	0.351
		b	2.230	-2.526	0.143	1.008
		θ	2.377	-1.469	0.000	0.866
BiF	500	a	3.975	-0.093	0.919	0.544
		b	1.589	-2.285	0.079	0.905
		θ	2.336	-1.650	0.000	0.874
	1000	a	2.464	-0.013	0.968	0.390
		b	1.394	-2.104	0.074	0.896
		θ	2.336	-1.650	-0.007	0.861

註：a 表示鑑別度，b 表示難度，θ 表示能力值，平均數、標準差爲模擬 100 次後的平均值。

以 1000 人爲例，比較三種模式下試題參數估計值與真值的表現，可看到在圖 2 和圖 3 之中，題組效果模式（TRT、BiF）下的鑑別度估計值較相近，當忽略題組效果而改以傳統單一試題估計（3PL）時，對鑑別度有高估的情形，尤其當 a 的值小於 1 時。比較難度值在三種模式下的關係，由圖中可發現，難度在不同模式估計下，估計結果較爲相似，此研究結果與 Wainer 和 Wang（2000）的實徵研究相吻合。

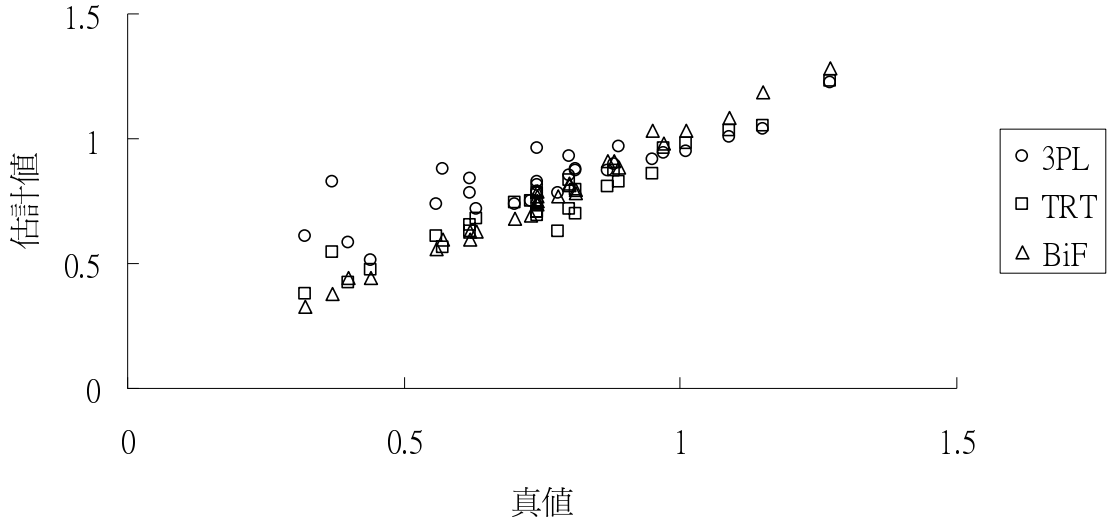

圖 2　1000 人下三種模式估計的鑑別度 a 值關係比較圖

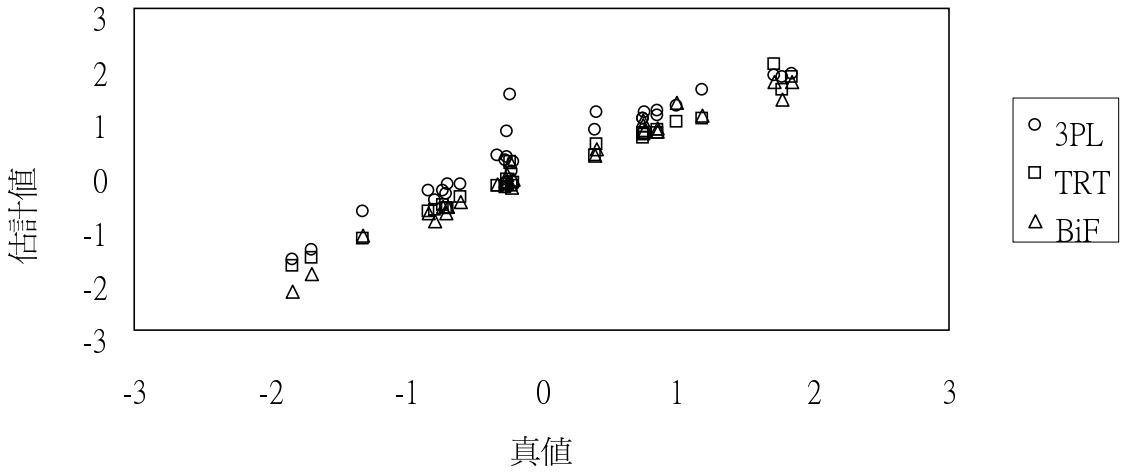

圖 3　1000 人下三種模式估計的 b 值關係比較圖

要瞭解參數估計的回復性，最容易的作法是計算參數真值與估計值之間的相關，此項相關的結果呈現在表 5 之中。表 5 顯示當人數由 500 人提升爲 1000 人時，估計值與真值之間的相關也稍微提高，但差異不大，a 值相關達 0.8 以上，b 值相關也達 0.9 以上，顯示參數估計的情形良好。從表 6 中 500 人時的鑑別度回復性來看，TRT 模式的 a 值估

計之 RMSE 值最小，是表現最好的；而傳統單一試題估計模式（3PL）下 RMSE 的執筆 TRT 和 BiF 的稍微高一些，其值爲 0.26，可知忽略題組效果估計對試題鑑別度會有影響。當樣本人數提高爲 1000 人時，參數估計的 RMSE 明顯降低，舉例來說，RMSE 從 0.225~0.269 降低爲 0.174~0.183，此結果可與 Wang 與 Wilson（2005）以 Rasch model 估計的研究中提到，人數增加會使 RMSE 下降的結果相呼應。

表 5　不同估計模式下，試題參數真值與估計值相關比較

模式	500 人		1000 人	
	r_a	r_b	r_a	r_b
3PL	.80	.91	.82	.91
TRT	.88	.97	.90	.98
BiF	.88	.96	.91	.97

註：r_a 代表 a 值間的相關，r_b 代表 b 值間的相關；表中數字是各種組合重複 100 次之　後的平均值。

表 6　不同人數在不同模式估計下的參數回復性（**RMSE 值**）

模式	a 值		b 值	
	N =500	N =1000	N =500	N =1000
3PL	0.269	0.183	0.343	0.321
TRT	0.225	0.174	0.262	0.197
BiF	0.226	0.181	0.250	0.225

註：表中數字是各種組合重複 100 次之後的 RMSE 平均值。

由表 6 中的 b 參數回復情形可發現，題組模式對難度估計的誤差較小，RMSE 值皆於 0.197~0.262，三種估計模式中以 3PL 模式的誤差最大，RMSE 值達 0.343。從前述表 5 中的估計相關來看，b 參數的估計值和真值的相關頗高，故分別以三種模式下的 b 參數的 RMSE 值做比較，由圖 4 和圖 5 可發現在不同人數下，3PL 的估計 b 值誤差較分散，TRT 模式和 BiF 模式對難度估計的表現較爲雷同，相當接近 0，但當難度值較大時，估計的誤差也較大；故可推論，在有題組效果下，難度適中的試題，以題組模式估計的表現最爲理想。

三種校準估計方法對於能力估計的回復性結果如圖 6 所示，在估計的過程中，皆設定用常態尺度（normal metric），相當於平均數爲 0、標準差爲 1 的常態分配；以 500 人的情境爲例，在不同估計模式下，RMSE 值分部幾乎相同，尤其是中間能力值區域，並未因不同估計模式而多大變化。

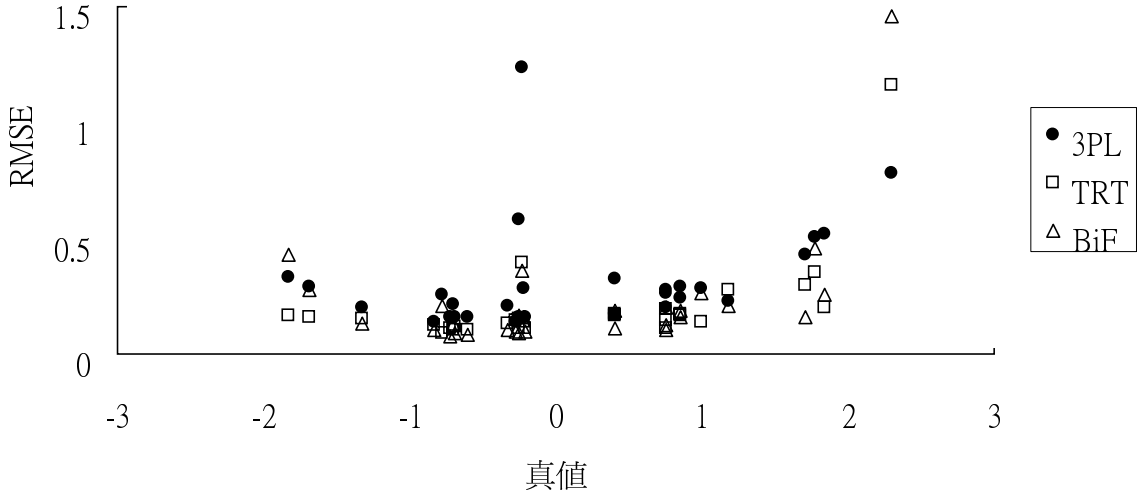

圖 4　500 人下 b 值的估計 RMSE 與真值間的關係圖

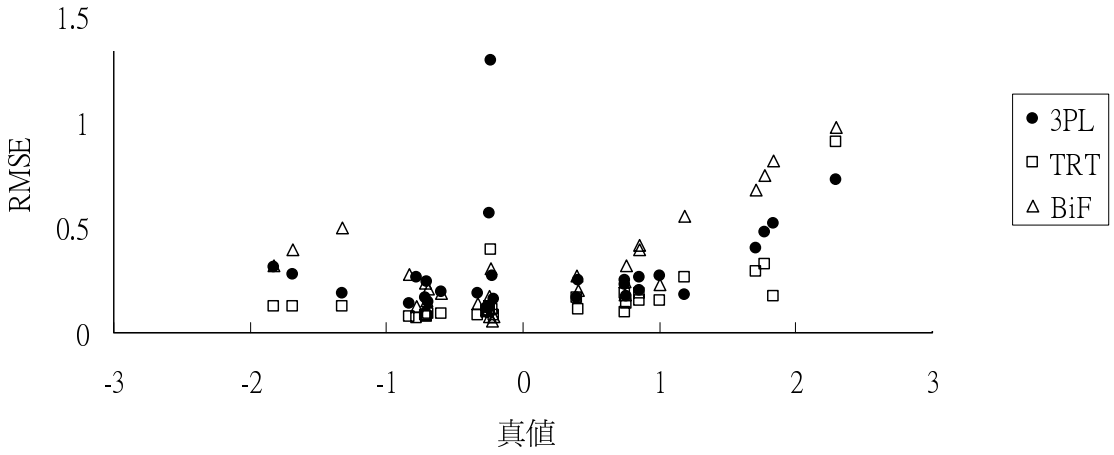

圖 5　1000 人下 b 值的估計 RMSE 與真值間的關係圖

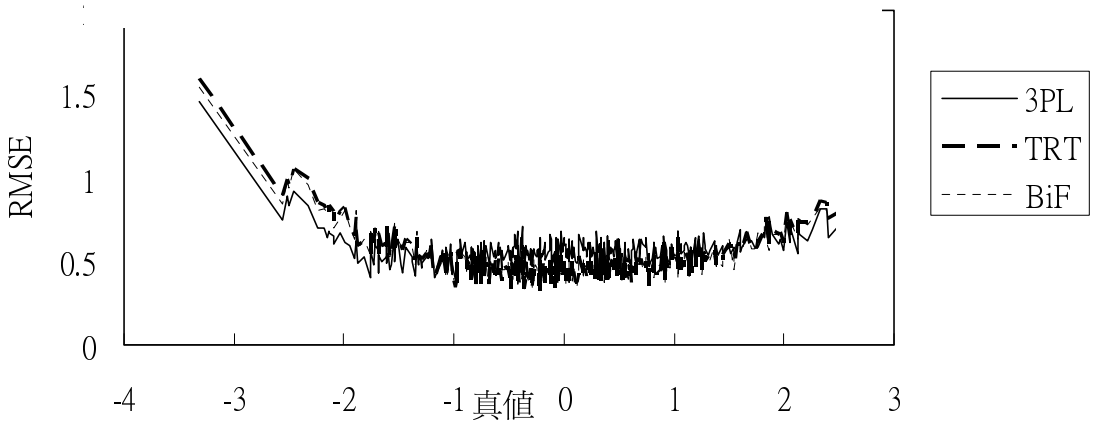

圖 6　500 人情境下，能力值估計 RMSE 的分佈

無論使用哪一個校準估計模式，題組中非獨立試題都會降低能力估計的信度，因爲題組參數會使隨機誤差增加。古典測驗理論中的信度定義之一，是估計值與實際值之間相關的平方，因爲資料由模擬得到，真實能力是已知的，可以計算真實的能力值與估計得到的能力值之間的相關，這個相關的平方另一方面也是一個決定係數，因此在本研究中，同時以決定係數和信度係數估計做爲估計精確度的判斷依據。決定係數具有「降低誤差比例（proportional reduction in error）」的意涵，可由決定係數來看進行預測時，減少預測誤差發生的機率有多大。舉例來說：如果 r^2 爲 0.81 時，表示可減少 81% 的預測誤差。決定係數越高的話表示效標變項中預測變項解釋成分越多。

在表 7 中所呈現的數據包括真實能力值與估計值之間的相關的平方（亦即表 7 中的決定係數），以及利用 Cronbach 的 alpha 所計算得到的信度係數。由表 7 中的數據可以看到，題組的資料以 3PL 模式來校準估計時，所得到的決定係數的值低於由 TRT 和 BiF 所得到的結果，而 Cronbach 的 alpha 係數似乎比 TRT 和 BiF 的高。顯示當資料爲題組所組成時，以 3PL 模式進行參數估計，對於真實能力值變異量的解釋量比較低一點點；信度係數估計卻高出由題組模式所得到的，表示忽略題組效果影響會導致能力信度的高估，此結果與 DeMars（2006）的研究結果相呼應。

表 7　能力真值與估計值間決定係數與相關係數表

估計模式	3PL		TRT		BiF	
人數	決定係數	信度估計	決定係數	信度估計	決定係數	信度估計
500	0.738	0.921	0.766	0.918	0.761	0.915
1000	0.762	0.928	0.773	0.921	0.772	0.919

註：此資料由題組模式所產生，細格內決定係數爲估計值與真值間 100 次校正後的決定係數平均值，信度估計爲該模式針對能力進行 100 次 Cronbach's alpha 係數的平均值。

二、不同題組效果對試題參數的影響

在前一小節的討論中，基本上是探討不同估計方法之間的差異，並未關心到題組效果的影響，因此在這一小節中，將題組效果一併納入考慮。在不同的題組效果（或題組斜率）下，利用 3PL 模式、TRT 和 BiF 所估計得到的 a 值之 RMSE 呈現於表 8，而估計值與真值之間的相關則呈現在表 9。由表 8 的數據來看，人數爲 500 人時，利用傳統的 3PL 模式進行估計，當題組斜率由 0 變成 0.6 時，RMSE 基本上可以說是沒有改變，而題組斜率變成 1.2 時，RMSE 由.250 變成.295，稍微提高了一點；在 TRT 或 BiF 的估計方法中，當人數爲 500 人時，三種方法的 RMSE 結果接近，TRT 和 BiF 的稍微小了一點，但當人數變成 1000 人時，三種方法的 RMSE 值都下降，題組效果愈大時，RMSE 值也傾向於比較大，而 3PL 模式的 RMSE 似乎比較明顯地大於 TRT 和 BiF 的值。。

以鑑別度估計值與真值的相關來看，表 9 的數據說明了當題組效果爲 0 時，無論人數是 500 人或是 1000 人，估計的 a 值與真實的 a 值之間的相關，由三種方法所得到的結果都類似，皆在.90~.94 之間。當題組斜率變大時，可看得到 3PL 的結果明顯比 TRT 和 BiF 的差；而人數愈多時，參數估計的回復性似乎也比較好。

表 8　試題鑑別度參數估計的 **RMSE**

題組數	人數	斜率	3PL	TRT	BiF
6	500	0.0	0.260	0.230	0.207
		0.6	0.250	0.219	0.229
		1.2	0.295	0.225	0.241
	1000	0.0	0.159	0.168	0.160
		0.6	0.171	0.173	0.179
		1.2	0.221	0.180	0.186

表 9　不同題組斜率下鑑別度的相關

斜率		0.0			0.6			1.2		
相關	人數	3PL	TRT	BiF	3PL	TRT	BiF	3PL	TRT	BiF
r_a	500	0.90	0.93	0.93	0.82	0.88	0.88	0.71	0.84	0.85
	1000	0.94	0.93	0.92	0.83	0.89	0.93	0.70	0.90	0.88

註：r_a 代表 a 參數的估計值與真值間的相關。

在難度參數的部分，當人數爲 500 人時，由表 10 中可以看到隨著題組斜率愈大，RMSE 有隨著變大的趨勢，當題組斜率由 0 變成 0.6 時，RMSE 的改變不明顯，但是當題組效果變成 1.2 時，除了由 3PL 估計方法的所得到的 RMSE 以外，RMSE 都明顯地增大，原因爲何，值得進一步探究。當人數增加爲 1000 人時，RMSE 變化的組型與 500 人相同，基本上三種估計模式隨題組斜率越高，所得到的 RMSE 的值皆比 500 人時相對應的值來得低。由表 11 的結果亦可看到，雖然隨著題組效果的增大，估計的難度值與真值之間的相關有下降的趨勢，但是 TRT 與 BiF 的結果都比 3PL 來得好。

表 10　試題難度參數估計值與真值的 **RMSE**

題組數	人數	斜率	3PL	TRT	BiF
6	500	0.0	0.329	0.237	0.203
		0.6	0.330	0.231	0.191
		1.2	0.371	0.318	0.357
	1000	0.0	0.303	0.155	0.173
		0.6	0.327	0.177	0.162
		1.2	0.333	0.259	0.271

表 11　不同斜率難度估計相關比較

斜率		0.0			0.6			1.2		
相關	人數	3PL	TRT	BiF	3PL	TRT	BiF	3PL	TRT	BiF
r_b	500	0.92	0.99	0.98	0.90	0.97	0.97	0.90	0.92	0.93
	1000	0.93	0.99	0.98	0.91	0.97	0.98	0.90	0.94	0.94

註：r_b 代表 b 值間的相關。

二、Q_3 統計數的偵測結果

Yen（1984）的 Q_3 值在本研究中被用來檢測所模擬的資料是否如研究者所希望的那樣，亦即，題組內各個試題間存在有條件相依的情形。本研究所模擬的資料無論人數多

少或題組斜率爲何，每一筆資料皆有 30 個試題，因此一共有 435 個試題配對，每一筆資料皆有 435 個 Q_3 值，所以每一種條件組合有 435 個 Q_3 平均值。如在研究方法那一小節最後所提到的，$z = \pm 1.96$ 是判斷試題之間是獨立的這個假設是否被拒絕的臨界值，根據公式（8）我們可以求得此臨界值相對應的 z_r 值爲 ±0.0943 ，因此當這 435 個 Q_3 的值經過 Fisher 的 z 轉換之後的值如果大於 0.0943 或是小於-0.0943 的話，則試題之間是獨立的這個虛無假設將被拒絕。由表 12 的結果可以看到，當題組效果爲 0 時，利用 TRT 和 BiF 來估計校準資料的結果顯示仍有部分試題配對之 Q_3 值的檢定結果是虛無假設被拒絕；當題組效果爲 0.6 時，根據 TRT 和 BiF 所得到的結果，Q_3 值拒絕獨立性假設的比率約在 90%，但是由 3PL 模式所得到的虛無假設被拒絕的比率似乎比較偏低；當題組效果爲 1.2 時，則可發現三種校準估計方法獨立性被拒絕的比率類似。因爲當初 Yen 提出 Q_3 這個統計數時是以 3PL 模式爲基礎的，因此當題組效果（斜率）爲 0.6 時，以 3PL 所得到的結果來看，值得進一步探討 Q_3 用來代表是否有局部相依存在時之限制，也就是當題組效果比較低時，Q_3 是否有足夠的敏感度？而換另一種角度來看，或許是當題組效果不大時，3PL 模式仍是足夠強韌的（robust），而 Q_3 也能反應出這一點；對於此一現象，值得進一步探討其原因爲何。

表 12 Q_3 值拒絕虛無假設的比率

模式	人數	斜率= 0	斜率= 0.6	斜率=1.2
3PL	500	0.00	0.25	0.90
	1000	0.00	0.00	0.90
TRT	500	0.30	0.85	0.80
	1000	0.15	0.95	0.90
BiF	500	0.00	0.90	0.90
	1000	0.25	0.90	0.95

註：虛無假設爲試題獨立，拒絕虛無假設即表示有依賴性存在。

根據 Yen（1993）的推導，當局部獨立成立時，Q_3 的期望值爲 $-1/(K-1)$，K 爲題數，在本研究中 $K = 30$，因此 Q_3 的期望值等於 –0.034，這個值相當於 0。本研究中由各個資料所計算得到的 Q_3 統計數之描述統計結果分別呈現於表 13 和表 14 之中。表 13 所呈現的是 500 人的結果，由表 13 中的數據可以清楚看到無論是利用 3PL、或 TRT、或 BiF 哪一種校準估計的方法，也無論題組效果是 0.6 或 1.2，題組內各個試題殘差之間的相關（也就是 Q_3 值）皆明顯地比題組間試題的配對所得到的高，這種現象是我們所樂於見到的，意謂著對於研究者所設定具有條件相依的試題之配對 Q_3 統計數可以反映出其相依性；而題組間試題的配對，本來就未設定有相關存在，Q_3 的值也反應出它們之間的獨立性。表 14 呈現的是由 1000 人所得到的結果，由表中的數據可以看到，表 14 中的結果組型與表 13 類似。另外，由表 13 和表 14 中也可以看到，當題組斜率變大時，Q_3 的值也隨之變大。

表 13 **500 人下各模式間 Q_3 描述統計**

模式	斜率		平均數	標準差	最小值	最大值	正值比率	負值比率
3PL	0	題組間	0.003	0.043	-0.100	0.110	48.80	40.80
		題組內	0.011	0.042	-0.070	0.090	65.00	35.00
	0.6	題組間	0.008	0.045	-0.100	0.140	52.40	38.00
		題組內	0.123	0.053	0.050	0.240	100.00	0.00
	1.2	題組間	0.004	0.049	-0.100	0.140	49.20	44.40
		題組內	0.248	0.049	0.150	0.360	100.00	0.00
TRT	0	題組間	0.002	0.045	-0.110	0.120	65.20	30.40
		題組內	0.062	0.048	-0.020	0.130	85.00	10.00
	0.6	題組間	0.002	0.049	-0.110	0.130	47.20	45.20
		題組內	0.291	0.106	0.110	0.520	100.00	0.00
	1.2	題組間	-0.002	0.044	-0.110	0.110	41.20	49.60
		題組內	0.472	0.174	0.110	0.680	100.00	0.00
BiF	0	題組間	0.007	0.043	-0.140	0.110	53.20	39.60
		題組內	0.046	0.051	-0.060	0.120	80.00	20.00
	0.6	題組間	0.006	0.044	-0.110	0.120	52.00	38.00
		題組內	0.292	0.092	0.140	0.470	100.00	0.00
	1.2	題組間	0.005	0.046	-0.140	0.120	49.20	45.60
		題組內	0.460	0.155	0.140	0.630	100.00	0.00

註：正、負值比率爲 Q_3 值在該情境下大於或小於零的比率。

表 14 **1000 人下各模式間 Q_3 描述統計**

模式	斜率		平均數	標準差	最小值	最大值	正值比率	負值比率
3PL	0	題組間	0.002	0.032	-0.090	0.070	52.00	41.20
		題組內	0.028	0.039	-0.030	0.120	70.00	25.00
	0.6	題組間	-0.002	0.031	-0.080	0.070	40.80	46.80
		題組內	0.074	0.029	0.020	0.130	100.00	0.00
	1.2	題組間	0.001	0.031	-0.090	0.070	45.60	46.40
		題組內	0.222	0.039	0.150	0.310	100.00	0.00
TRT	0	題組間	0.006	0.034	-0.100	0.080	54.40	35.60
		題組內	0.099	0.055	0.010	0.200	100.00	0.00
	0.6	題組間	-0.002	0.036	-0.110	0.150	40.40	48.40
		題組內	0.247	0.059	0.150	0.340	100.00	0.00
	1.2	題組間	-0.003	0.039	-0.110	0.150	40.40	52.80
		題組內	0.445	0.155	0.120	0.610	100.00	0.00
BiF	0	題組間	0.012	0.032	-0.080	0.090	60.40	32.40
		題組內	0.123	0.050	0.020	0.190	100.00	0.00
	0.6	題組間	0.003	0.040	-0.110	0.150	48.40	43.20
		題組內	0.284	0.089	0.140	0.400	100.00	0.00
	1.2	題組間	-0.005	0.039	-0.110	0.150	38.40	53.20
		題組內	0.454	0.150	0.140	0.610	100.00	0.00

註：正、負值比率爲 Q_3 值在該情境下大於或小於零的比率。

當試題爲局部獨立時，Q_3 期望值爲一相當接近 0 的負值，根據殘差相關的概念，若呈現負值越大，表示測量到不同潛在構念，若呈現正值則表示試題間測量到的構念是相同的，也就是有相依性存在，故可由 Q_3 的正、負值，做爲判斷試題間是否符合單向度獨立性假設的參考。由表 13 和表 14 的結果來看，題組間的負值比率，都有 30%以上，遠

遠高於題組內的負值比率（題組內負值比率幾乎等於 0），而題組內正值的比率除了 3PL 題組效果爲 0 時爲 70%以外，其餘的皆爲 100%；這項結果支持了研究者在模擬資料時原先所設定的題組效果組型。附帶說明的是在表 13 和 14 之中，Q_3 等於零的未被列入計算，所以部分組合之正負值比率之和不等於 100，例如表 14 中第一橫列正值比率與負值比率分別爲 52%和 41.20%，其和爲 93.20%。

伍、結論與建議

題組在許多測驗情境中都經常被使用，無論是傳統紙筆測驗或電腦適性測驗，這種題組型試題能夠測量到較高層次的分析、比較、綜合能力，是未來測驗的發展趨勢，因此本研究以題組作爲研究主題，希望藉由模擬研究，瞭解題組效果對於參數估計回復性的影響。

依據前面的研究結果實驗發現，本研究結論如下：

一、TRT 和 BiF 校準估計模式對於 b 參數估計的回復性明顯地比 3PL 來得好，當人數增加時，雖然由 3PL 模式所得到的 RMSE 也降低，但是 TRT 和 BiF 降低的幅度更大。相對於此一組型，在 a 參數方面較不明顯，雖然 TRT 和 BiF 的 RMSE 也傾向於比 3PL 來得低，人數增加時 RMSE 也降低，但是方法間之差異沒有 b 參數那麼明顯。在能力參數方面，三種校準估計方法之間的差異不大。

二、使用 Q_3 來偵測試題之間是否存有相依的現象，在本研究中的結果顯試題內各試題配對所得到的 Q_3 值顯著高於題組間試題配對所得到的結果，顯示 Q_3 對於題組效果（也就是局部相依）的檢測有一定的能力。

根據研究結果，對未來發展的建議如下：（1）在本研究中所使用的資料是二元計分試題的作答反應結果，目前許多測驗皆爲二元計分試題與多元計分試題所混合組成的，未來研究應將多元計分試題包含在內。（2）本研究僅討論六個題組下的表現，往後研究可朝不同題組數、不同題組長度、更多題組效應、不同題組數佔總試題比率等方面進行探討，比較結果是否會有所不同。（3）因爲 SCORIGHT 執行需要花費時間頗長，在本研究中，MCMC 程序的 burn in period 設爲 3000 次，然後利用後續的 1000 次迭代進行抽樣，迭代次數可能稍嫌不足，可以考慮較多次的迭代迴圈。（4）本研究僅進行模擬研究，比較題組斜率設定下的影響，缺乏實徵資料加以輔助證明，未來可進行實徵資料的分析，以對本研所探討的議題有更清楚的瞭解。

參考文獻

林欣怡（2007）。**GRM 模式之試題局部獨立性偵測指標模擬研究**（未出版之碩士論文）。國立臺中教育大學，台中市。

郭生玉（1998）。**心理與教育測驗**。臺北：精華書局。

陳柏熹（2005）。電腦化適性測驗的理論與應用。**國家精英季刊**，**1**(1)，157-174。

陳柏熹（2006）。**IRT 在量表（測驗）編製上的應用**。測驗專業工作坊講義，2010 年 12 月 31 日，取自：http://www.rcpet.ntnu.edu.tw/download.htm.

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence* (ACT Research Report 87-14). Iowa City, IA: ACT, Inc.

Allen, S., & Sudweeks, R. R. (2001, April). *Identifying and managing local item dependence in context-dependent item sets*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

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

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets *Psychometrika, 64*, 153-168.

Chen, W. H. & Thissen, D.(1997). Local dependence indexes for item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement*, *25*(2), 327-346.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*, 145-168. doi: 10.1111/j.1745-3984.2006.00010.x

Dresher, A. R. (2004, April). *An empirical investigation of LID using the testlet model: A further look.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego , CA .

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education*, *10*(2), 123-144.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436

Glas A. W., Wainer H., & Bradlow E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271-287). Dordrecht, Netherlands: Kluwer.

Habing, B., Finch, H., & Roberts, J. S. (2005). A Q_3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement, 28*(6), 457-471. doi: 10.1177/0146621604279550

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice*, *11*, 21-25.

Kim, S. H., Cohen, A. S., & Lin, Y. H. (2005). *LDIP:A Computer Program for Local Dependence Indices for Polytomous Items* [Software and Manual]. Athens, GA: University of Georgia.

Lee, G. (2000). Estimating conditional standard errors of measurement for tests composed of testlets. *Applied Measurement in Education, 13*, 161-180. doi: 10.1207/S15324818AME1302_3

Lee, G. (2000). A comparison of methods of estimating conditional standard errors of

measurement for testlet-based test scores using simulation techniques. *Journal of Educational Measurement, 36*, 91-112. DOI: 10.1111/j.1745-3984.2000.tb01078.x

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001).The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement, 61*, 958-975. doi:10.1177/00131640121971590

Lee, G. (2002). The Influence of Several Factors on Reliability for Complex Reading Comprehension Tests. *Journal of Educational Measurement*, *39*(2), 149-164.

Lee, G., Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, *12*, 237-255.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and practice, 19*, 9-15.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21. doi: 10.1177/0146621605275414

Lord, F. M. (1980). *Applications of items response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Thissen, D., Steinberg, L., & Mooney J. A. (1989). Trace Lines for testlets: a use of multiple-categorical response models. *Journal of Educational Measurement*, *26*, 247-60.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123-140.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157-187.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wainer H., & Wang X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203-220. doi: 10.1111/j.1745-3984.2000.tb01083.x

Wainer, H., Bradlow E. T., & Du Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Dordrecht, Netherlands: Kluwer.

Wainer, H., Bradlow E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wang, X., Bradlow, E. T., Wainer H. (2002). A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement, 26*, 109-128. doi:

10.1177/0146621602026001007

Wang, X., Bradlow, E. T., & Wainer, H. (2004). A user's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis (ETS Technical Report RR-04-49). Princeton, NJ: Educational Testing Service.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149. doi:10.1177/0146621604271053

Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*, 5-27. doi: 10.1177/0013164404268676

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181-198.

Wollack, J. A., Suh, Y., & Bolt, D. M. (2007, April). *Using the testlet model to mitigate test speedeness effects*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Yen, W. M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Zeniskey, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*(4), 291-309. doi: 10.1111/j.1745-3984.2002.tb01144.x

投稿日期：2011 年 01 月 20 日
修正日期：2012 年 08 月 03 日
接受日期：2012 年 09 月 24 日

The Influence of Testlet Effect on the Accuracy of Parameter Estimation

Pei-Ling Lai

Teacher, Jiyang Elementary School, Meinung District, Kaoshiung City

Bor-Yaun Twu[2]

Associate Professor, Department of Education, National University of Tainan

Abstract

This study investigated the influence of testlet effect on the recovery of item and person parameters. Simulated data was used to compare the parameter recovery given by the traditional item response theory (IRT) model, the testlet response theory (TRT) model and the bi-factor model. The Q_3 index was computed for all the data sets to serve as a tool for detecting the local item dependency. There are six testlets in each test with 5 items for each testlet. Three testlet slope (0.0, 0.6, & 1.2), and two sample sizes (500 & 1000 examinees) were manipulated for simulating the item response vectors. For each combination of the conditions, item response data was simulated 100 times. The simulated item response data was calibrated by using the three models described above separately, and Q_3 was calculated for each data set. The main findings are as the following: (1) The accuracy of item parameter recovery was higher for the testlet-based models than that of traditional item-based model (i.e., 3PL model). Among the three models used, the TRT model performs best, followed by the bi-factor model and then the IRT model; (2) The greater sample size is, the better accuracy of item parameter recovery is gained. Similarly, the greater testlet slope is, the greater estimation error is found; (3) The Q_3 indexes calculated for the paired items from the same testlet were found larger than those for the item pairs from different testlets. This indicates that the Q_3 index performs well in terms of detecting the local dependence between items.

Key words：testlet, local item dependence, Q_3, TRT, bi-factor model

[2] Corresponding Author's e-mail: bortwu@gmail.com