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