壹、緒論

近幾年，不同測驗分數連結（linking）大部分皆透過統計分布的方法進行比較，雖然有許多研究耗費相當多的時間與精力進行測驗分數資料的蒐集，以及藉由不同等化方法評估測驗等化（equating）之成效，卻沒有重視在測驗分數連結的過程中是否達到既定的目標。Lord（1980）指出若等化分數成效降低，則代表等化是不可能或不必要的事，因此，提供評估等化測驗分數敏感性（sensitiveness）的準則。然而，導致等化分數成效降低之結果，大都來自於練習的原因，因此，目前已廣泛應用多組測驗題本來獲取受試者資料，並透過等化程序使得不同測驗題本分數能進行比較（Angoff & Cowell, 1986; Harris & Kolen, 1986）。Dorans與Holland（2000）修正Lord提出的準則，以測驗等化是否有達到群體不變性（population invariance）的性質，做爲評估不同測驗分數連結成效之標準。Dorans與Holland（2000）指出群體不變性在評估等化成效時扮演一個重要角色，並認爲在不同次群體（subpopulation）中用來連結兩個測驗分數的等化函數若沒有符合不變性，則這兩個測驗分數應該不能被視爲相等（equatable）。群體不變性是指不同次群體的測驗分數，經過相同的等化程序後，其轉換後之量尺分數應該相同。也就是說，爲了要滿足測驗公平性的原則，等化在進行時必須符合群體不變性之需求（Kolen & Brennan, 2004）。

近年來，許多研究使用不同等化方法驗證大型測驗（large-scale assessments）是否符合群體不變性的性質（Dorans, Liu, & Hammond, 2008; Liu & Holland, 2008; von Davier & Wilson, 2008; Yang & Gao, 2008; Yi, Harris, & Gao, 2008）。例如：von Davier與Wilson（2008）使用跳級安置計畫（the Advanced Placement Program, AP Program）舉辦的微積分測驗，評估其測驗等化效果，並比較連結百分位數等化（chained equipercentile equating）（Braun & Holland, 1982）、Tucker（Gulliksen, 1950）、真實分數等化（true score equating）（Lord & Wingersky, 1984）等方法的等化成效。研究結果發現，以試題反應理論（item response theory, IRT）模式爲基礎的真實分數等化方法有較好的等化效果；Yang與Gao（2008）使用IRT真實分數等化方法（Lord & Wingersky, 1984），並以性別爲次群體變項，檢驗CLEP（College-Level Examination Program）題組測驗之量尺分數是否符合群體不變性。

隨著資訊科技快速進步、測驗形式的改變及需求量的快速增加，大型測

驗也發展更多元的測驗題型。除了傳統的選擇式反應試題（selected-response items）與建構式反應試題（constructed-response items）之外，國內外的教育心理測驗與大型標準化成就測驗（standardized achievement educational test）皆使用題組（testlet）試題來評量學生的學習成效，包括：托福測驗（The Test of English as a Foreign Language, TOFEL）、學術評量測驗（Scholastic Assessment Test, SAT）、美國國家教育進展評量（National Assessment of Educational Progress, NAEP）、國際閱讀發展研究（Progress of International Reading Literacy Study, PIRLS）、臺灣學生學習成就評量（Taiwan Assessment of Student Achievement, TASA）等。Yang與Gao（2008）研究使用Rasch測量模式進行題組測驗等化群體不變性之成效評估，而非使用題組反應理論（testlet response theory, TRT）模式。然而，以IRT模式分析測驗資料，必須符合單向度（unidimensionality）與局部獨立性（local independence）假設，但Rosenbaum（1988）指出題組測驗結構違反了局部獨立性的假設，因此，當題組測驗資料使用IRT模式分析時，忽略題組試題內的相關，將產生受試者能力參數高估與試題參數偏誤的情形（Wainer, 1995; Wainer & Lukhele, 1997; Wainer, Sireci, & Thissen, 1991; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993）。此外，使用TRT模式分析題組測驗資料，也可避免以多點計分模式來分析二元計分題組測驗資料所造成的缺失，並保留試題參數的概念，得到更精準的參數估計（Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005）。

在等化方法的選擇上，許多研究皆指出以IRT模式爲基礎的等化方法比以古典測驗理論（classical test theory, CTT）模式爲基礎的等化方法有較佳的等化效果（von Davier & Wilson, 2008）。Wang、Lu、Kuo與Cheng（2011）、Lord與Wingersky（1984）等人的研究則指出IRT真實分數與觀察分數等化（observed score equating）方法的估計結果差異不大。此外，Wang等人比較平均數與標準差法（mean/sigma）、平均數法（mean/mean）、特徵曲線法（characteristic curve method）等常見的量尺轉化方法（scale transformation methods），研究結果顯示特徵曲線法的估計結果較穩定，較不易受測驗分數分布的影響而產生不穩定的結果。有鑑於此，本研究將以IRT真實分數等化方法爲基礎，並採用特徵曲線的量尺轉換方法，檢驗TASA 2007英語文題組測驗資料之群體不變性，以評