緒論

研究測驗理論的學者通常會把測驗分爲古典測驗理論（classical test theory）及現代測驗理論（modern test theory）兩大學派。古典測驗理論以真實分數模型（true score model）爲理論依據，測驗的結果容易受到不同試題的影響，相關的試題評估分析，亦常常因不同的受試者，而有不同的結果。現代測驗理論主要以試題反應理論爲其理論的架構，考生的能力估計不會受到不同試題的影響。相較於古典測驗理論，現代測驗理論具有嚴謹的學理根據，在近幾十年來受到更多的重視。

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）在最近幾年來已逐漸取代傳統的紙筆測驗，成爲現代測驗的新趨勢，並廣泛地應用在證照考試或是專業檢定上，如GRE、GMAT及TOEFL等。這些相關的測驗考生往往來自不同群體，因此在時間、地點及測驗成本的考量上，常常無法在同一時間施測，因此測驗單位就必須顧慮到題目的外洩，以及必須建立多個題本來防止測驗不公平的情況。測驗結束後，測驗分數的比較便爲一個重要的課題。在「試題反應理論」（item response theory, IRT）的假設下，不同的測驗結果，需要建立共同量尺，方能進行分數間的比較。而共同量尺的建立，則需借助測驗等化（test equating）的技巧方得以完成。所謂測驗等化係指利用統計方法，將兩份或是兩份以上試卷的測驗分數進行轉換，等化的目的是在校準測驗難度之差異，而非測驗內容之差異（Kolen & Brennan, 2004）。藉由等化處理技巧，不同測驗的結果可以轉換到同一量尺上，因此等化後的分數可以直接進行比較。在IRT的假設下，測驗結果爲考生能力估計值，在有限的測驗題數下所得之能力估計值，常常會因許多不同的因素對測驗結果造成影響（郭伯臣、王暄博，2008），藉由等化的

校準，可以調整所得估計值的差異。

不同測驗間的分數欲進行等化，需要在各測驗中包含一份共同試題，稱爲定錨試題（anchor items），以便作爲測驗間的連結之用。定錨試題的品質，對於後續的等化過程，扮演極爲重要的角色。因此，施測者在蒐集等化資料的同時，可以根據不同的研究需要，採用不同的測驗等化設計。進行測驗等化時通常假設定錨試題的參數爲已知，換言之，施測者通常會利用校準完成的試題。然而，在線上測驗中，題庫中的題目會存在過度曝光或是試題消耗量過大的問題。因此，我們必須不斷地補充新題目，並同時對新題目進行試題校準工作。傳統的作法，須經過預試階段，藉由考生的作答結果，估計所有新試題的參數。然而，這樣的作法會耗費許多時間及成本。因此，線上校準（online calibration）便成爲一種經濟且有效率的作法。所謂線上校準，係指讓考生進行線上測驗時，同時進行新試題的校準工作。CAT的目的爲估計考生的能力，而線上校準爲當考生進行考試時，將未完成試題校準的試題給予考生施測，其目的爲估計試題參數，其用意爲利用目前進行測驗的考生作爲預試人選，達到節省成本的目的。

在線上校準的問題中，必須事先估計考生的能力值，以便根據最佳設計的結果，選擇所需考生來進行試題校準。許多研究學者提出了不同的試題選取法則，如Chang與Ying（1999）。Chang（2004）並且證明出當測驗題數爲無窮多時，考生能力的估計值具有強收斂的性質。但是在實際的測驗環境中，考試的時間是有限的，我們只能用有限的考題去估計考生的能力，所以用來作爲試題校準的考生，其能力估計將會存在估計誤差。相關研究發現，當自變數存在測量誤差時，參數的估計會存在偏誤（Chang, 2011; Clark, 1982; Li & Hsiao, 2004; Michalik & Tripathi, 1980）。因此，利用線上校準所得到的試題參數估計亦必然存在偏誤。相關研究發現，若利用試題參數存在估計誤差的試題施測，將對測驗結果造成影響（Spray & Reckase, 1987）；在