「教育研究學報」第 46 卷第 1 期（民國 101 年）：119~137

成就測驗組合分數議題探討[1]

涂柏原

國立臺南大學教育學系 副教授

盧思丞

國立臺南大學教育學系 博士生

摘要

本研究根據一個成就測驗題庫建置計畫的試題參數產生模擬資料來探討組合分數計算的相關議題，文獻中曾提到的組合分數的加權數計算方法多達十餘種，本研究主要探討信度加權法和 IRT 加權法；另外以某市一般智能資優學生甄選資料來探討多元迴歸和主成份分析法。結果發現在信度加權法的部分，用原始分數之間的相關矩陣或是用各科能力值之間的相關矩陣所計算得到的各科分數之加權數類似，且二者所得到的組合分數之信度係數雷同。利用四個科目的能力值所計算得到的主成份分數和組合分數與將四科資料視為一科時利用 3PL 模式所估計得到的能力值之間的相關係數，對五年級來說，分別為.969 和.982，而對六年級來說，則是.971 和.990。最後，利用團體智力測驗的三個分測驗的資料計算主成份分數，或利用多元迴歸分析預測個別智力測驗的分數，結果發現主成份分數與個別智力測驗分數之間的相關，和用多元迴歸所得到的預測分數與真正的個別智力測驗分數之間的相關，二者的值非常接近，都接近.95。

關鍵字：組合分數、信度加權、IRT 加權、主成份分析、多元迴歸

[1] 本文部分結果是取自 NSC 95-2413-H-024-009-的研究案，部分結果曾發表於 2007 年 11 月中國測驗學會年會。

壹、前言

實務上許多國內外的大型測驗計畫常會將考生在各個分測驗所得到的分數加以加權，以得到一個加權的分數來代表考生在該測驗上面的整體表現，並將該分數報導給考生知道，以作爲考生在該測驗的得分（Peterson, Kolen & Hoover, 1989）。比如說，Law School Admission Test（LSAT）的分數是閱讀理解（Reading Comprehension）、分析推理（Analytical Reasoning）和邏輯推理（Logical Reasoning）三個部份的總和，而 SAT II Writing Subject Test 的分數則是選擇題部份的分數與申論題得分之加權總和；ACT 的組合分數則是 English, Math, Science 和 Reading 等四個分測驗的量尺分數之平均（Kolen & Hanson, 1989, p. 39）。因此，這個加權總分的計算就成了測驗編製者在建立量尺分數時一個重要的工作，在測驗的文獻中，這個加權的總分一般稱爲 composite score，在本文中譯爲組合分數。

筆者參與一項大型成就測驗題庫建置的工作，該項成就測驗題庫的內容涵蓋二至七等六個年級，包含國語文、數學、自然、社會與英文等五個學科或領域的試題，其中英文只用在七年級。該項測驗題庫建置計畫也考慮將學生在各個科目上的表現，利用線性加權的方式來得到一個組合分數，因而組合分數計算的議題就吸引了筆者的注意。

一般而言，組合分數是將原始分數或量尺分數加以線性組合得到的，很少用非直線轉換的方式（Peterson, Kolen, & Hoover, 1989）。根據 Wainer 和 Thissen（2001）的用語，以線性的方式得到之組合分數可以定義如下：

$$z_c = \sum w_v z_v \qquad \text{(公式 1)}$$

其中 z_c 是組合分數，w_v 是成份 v 的加權係數，z_v 是成份 v 的標準化分數。以筆者所參與的題庫建置工作來說，成分 z_v 指的是國語文、數學、社會或是自然等學科領域之任何一個，w_v 是在這些學科領域上得分的加權係數。當得到考生的組合分數之後，組合分數的信度係數（reliability of composite score）也可以被計算出來。一般來說，組合分數的信度之定義與一般加總分數之信度的定義一樣：

$$\rho_c = 1 - \frac{\sigma_{e_z}^2}{\sigma_z^2} \qquad \text{(公式 2)}$$

其中 ρ_c 爲組合分數的信度係數，σ_z^2 爲組合分數的變異數，$\sigma_{e_z}^2$ 是組合分數的誤差變異數（Feldt & Brennan, 1989）。在大部分的實務應用中，$\sigma_{e_z}^2$ 和 σ_z^2 的值可以從僅施測一次的測驗樣本資料中計算得到；因此，我們可以在沒有施測平行測驗的情形之下估計 ρ_c（Wainer & Thissen, 2001）：

$$\rho_c = 1 - \frac{\sum w_v^2 \sigma_{e_v}^2}{\sum_v w_v^2 \sigma_{z_v}^2 + \sum_v \sum_{v'} w_v w_{v'} \sigma_{z_v z_{v'}}} \qquad \text{(公式 3)}$$

利用標準分數的便利性，$\sigma_{e_v}^2 = 1-\rho_v$，帶入成份 v 信度係數的樣本估計數 r_v，而 $r_{vv'}$ 是成份 v 和成份 v' 分數的相關，則組合分數信度之估計數變成

$$\rho_c = 1 - \frac{\sum w_v^2(1-r_v)}{\sum_v w_v^2 + \sum_v \sum_{v'} w_v w_{v'} r_{vv'}} \text{。} \qquad \text{(公式 4)}$$

Wainer 和 Thissen（1993）、Feldt 和 Brennan（1989）、以及 Lord 和 Novick（1968）都提供了組合分數詳細的介紹，Gulliksen（1950/1987）更是最早提供組合分數詳細處理的資料來源；McDonald（1968）針對加權的問題提出一個統整的處理程序，是個可以得到變項的線性組合之一般化程序，主成份分析、多元迴歸及最大信度法等皆爲該法程序之特例，Wang 和 Stanley（1970）也提供了計算組合分數多種方法詳細的回顧。Rudner（2001）在探討分數成份加權這個議題時，將計算組合分數的方法分成兩大類：隱含的（implicit）和外顯的（explicit），其中隱含的方法包含了原始分數加總（adding raw scores）和試題反應理論法（item response theory, IRT）兩種；而外顯的方法包含難度加權（weighting by difficulty）、信度加權（reliability weighting）和效度加權（validity weighting）。

在上述的文獻中，最常被提起的組合分數計算的方法包括信度加權法、效度加權法和迴歸分析法等，而近年來試題反應理論（item response theory, IRT）普遍被運用來作爲試題分析與計分的方法。考慮到筆者所參與的題庫建置計畫資料的特性，要利用其他的方法來計算組合分數可能比較不易，所以才擬以信度加權法、IRT 方法和迴歸分析等來進行此研究，另外，考慮到主成份是觀察變項的線性組合，於是在本研究中一併探討主成份分析法。因此底下簡要說明在本文中所用到的信度加權法、估計效度加權法、IRT 的方法、多元迴歸法及主成份分析等方法。

一、信度加權法

本文前面曾提到了組合分數的信度係數可以由公式（4）來估計得到，如果一個測驗是由兩個成份（component）所組成的，那麼公式（4）可以改寫成下式

$$r_c = 1 - \frac{w_1^2(1-r_1) + w_2^2(1-r_2)}{w_1^2 + w_2^2 + +2w_1 w_2 r_{12}} \qquad \text{(公式 5)}$$

Wainer 和 Thissen（2001）舉了一個例子來說明如何利用這個公式計算組合分數的信度係數，如果一個假想的寫作技能測驗是由一個測驗時間 40 分鐘而 α 係數爲.85 的單一選擇題和測驗時間 20 分鐘的申論題所組成；其中申論題只有一題，由兩個評分者加以評分，而申論題的複本信度係數爲.60，選擇題和申論題之間的相關大約爲.43。這個測驗的總分是由選擇題的標準分數與申論題的標準分數加權組合而成的，選擇題的加權係數爲 2，申論題的加權係數爲 1。之所以選用 2:1 的加權數是因爲這兩個部份測驗時間的比率爲 2:1。這種以作答時間的比率來作爲組合分數加權係數是目前許多測驗所採用的（Wainer & Thissen, 2001）。

$$r_c = 1 - \frac{2^2(1-0.85)+1^2(1-0.60)}{2^2+1^2+2\times 2\times 1\times 0.43} \approx 0.851$$

將這個例子中的數據帶入上面的公式，我們得到亦即根據 2:1 來進行加權時，這個例子中所提到的測驗組合分數之信度係數爲.851。

但是這種以作答時間之比率爲加權係數的方式，未必能使所得到的組合分數之信度係數最大化，因此如何使得組合分數的信度係數最大化，是許多學者曾討論的（Wainer & Thissen, 2001, 1993; Kane & Case, 2004）。Wainer 和 Thissen （1993）將以上的組合分數計算的方法稱爲信度加權（reliability weighting）。

在上面例子中所呈現的組合分數信度之計算公式假定加權係數 w 是已知的，假如不想事前預先決定一組加權係數，而是想從資料中找到最適合的加權係數的話，眾所皆知的作法是將組合分數與代表測驗目的某一個效標變項之間的關係加以最大化。可惜的是，可靠的效度效標實在是很少，在缺乏效標變項的情形下，一般人就轉向內部一致性信度係數求救，因此內部一致性係數提供了一個合理的加權係數選擇的基礎。

當要從資料中找出組合分數的加權係數時，爲了讓問題變得比較容易處理，一般的作法是將公式（5）加入 $w_1 + w_2 = 1$ 這個限制，因此 w_2 可以改寫成 $1 - w_1$ ，公式（5）就變成

$$r_c = 1 - \frac{w_1^2(1-r_1)+(1-w_1)^2(1-r_2)}{w_1^2+(1-w_1)^2+2w_1(1-w_1)r_{12}} \qquad \text{(公式 6)}$$

於是除了用統計的方法來找到最佳的加權係數外，畫出組合分數信度係數與 w_1 的函數圖形，也可以用來協助我們找出最佳的值。就前面 Wainer 和 Thissen 所提供的例子而言，當 $r_1 = .85$ ， $r_2 = .60$ ， $r_{12} = .43$ 時，選擇題的加權數 w_1 與組合分數的信度係數之間的函數圖形可得到如下：

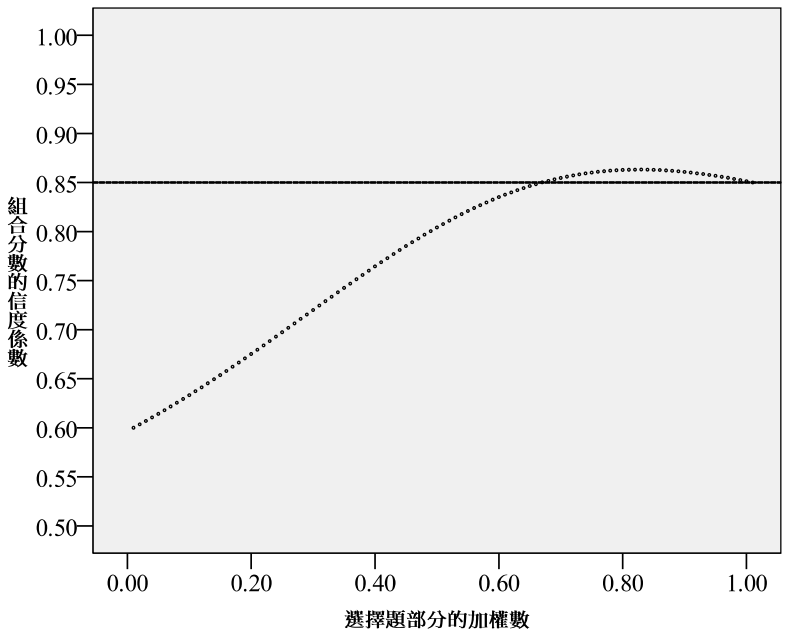

圖 1 組合分數信度與選擇題加權數之關係

從圖 1 之中我們可以藉著橫軸 w_1 值（選擇題部分的加權數）的變化，很容易地看出當 w_1 的值多大的時候，組合分數會有最大的信度係數。在此圖中，當 w_1 約等於 0.8 時，組合分數的信度係數最大。

當不止兩個成份時，用上面的圖要找出最佳的加權係數可能不是一件容易的事，必須藉助統計的方法。公式（6）可以改寫成矩陣型態以得到最佳的加權係數，詳細的做法請參見 Wainer 和 Thissen（2001）。當要利用信度加權的方法來計算該題庫建置計畫中所欲得到的組合分數時，從公式（4）中可以得知國語文、數學、社會和自然等各科測驗的信度係數以及各科之間的相關係數是必要的。

二、IRT 加權法

試題反應理論模式在估計參數時，基本上已經同時解決「加權」和「量尺化」的問題，所估計得到的考生精熟（proficiency）或能力（ability）參數 θ 已是個最佳解（Childs, Elgie, Gadalla, Traub, & Jaciw, 2004; Wainer & Thissen, 1993）。因此，如果用 IRT 的 2PL 或 3PL 模式來估計考生的能力參數，基本上就不必再考慮加權的問題，因爲 2PL 和 3PL 模式中的鑑別度參數 a 是二元計分試題用來計算加權組合分數時之最加權重（optimal weights; Lord, 1980）。Childs, Elgie, Gadalla, Traub 和 Jaciw（2004）在計算一個由選擇題、簡答題和問答題所組成的測驗之組合分數時，發現利用公式（1）將加權數設爲 2:2:6 時，原始分數之加權總和與從 IRT 模式所估計得到的 θ 值之間的相關係數高達.9 以上。Wainer 和 Thissen（1993）提到 IRT 裡面所隱含的加權係數基本上是信度係數的函數，因此 Childs 等人（2004）的發現就不令人驚訝。

Kolen, Wang, 和 Lee（in press）提到在兩種情形之下組合分數可能被建立或計算，以協助測驗分數的使用或解釋。第一種情形是不同的教育成就測驗的分數組成一個組合分數，以提供在二個或多個內容領域上一個單一的成就指標；另一種情形是一個測驗中包含數種題型，例如選擇題和建構反應試題等，也就是混合題型的測驗（mixed-format test），由不同題型所得到的分數被用來形成一個組合分數，作爲整個測驗的總分。Wainer 和 Thissen（2001）的例子是 Kolen, Wang 和 Lee 文章中所提到的第二類組合分數，而筆者所參與的題庫建置工作，則是屬於 Kolen 等人所說的第一種情形中的組合分數。像國中基測將各科的量尺分數加總得到的總分，實際上就是一個組合分數；而美國的 ACT Assessment 將 Math、English、Reading 和 Science Reading 等四個分測驗的量尺分數加以平均得到最後的測驗總分，也是屬於第一種情形的組合分數。在題庫建置的計畫中，二至六年級皆有四個學科，每一個科目的試題皆由選擇題所組成，因此要用 IRT 的方法來計算組合分數時，顯然與 Wainer 和 Thissen（1993）所提到的或是 Childs 等人（2004）所探討的問題不同。

與本文所探討的其他方法相較，IRT 的方法與其他的方法最大不同之處在於 IRT 的能力估計直式利用非線性的方法得到的，而其他的方法皆是利用線性的方法得到組合分數。因此如果所得到的結果之間有差異的話，必須將線性與非線性之間的不同加以考慮。

三、估計效度加權法

Wainer 和 Thissen（1993）提到的第三種方法爲估計效度加權法（predicted validity weighting）。如果能夠找到一個有效的效標變項，Wainer 和 Thissen（1993）認爲最合理的加權方式就是選擇能將估計效標效度最大化的那組加權係數，他們並將這個方法稱爲效標加權法（criterion weighting）。要用這種方法必須能夠找到一個好的效標，如果找不到合適的效標的話，將無法使用此方法。在題庫建置工作裡面，目前所用抽樣和收集資料的計畫當中，並未包含效標變項在內，因此若要用目前該案所蒐集到的資料來進行本研究，可能就無法使用這個方法。

四、多元迴歸法

如果效標變項 Y 可以取得，那麼形成組合分數的變項可以作爲預測變項，傳統的多元迴歸分析的公式能給予將組合分數與效標分數之間的相關加以最大化的加權數。一般而言，此迴歸公式可以寫成如下的形式：

$$\hat{z}_Y = \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_v z_v$$

其中 $\hat{z}_Y$ 是效標變項 Y 的標準化形式，$\beta_1, \cdots, \beta_v$ 是各個預測變項的加權數，而 $z_1, \cdots, z_v$ 如公式（1）所定義。

在解釋由多元迴歸法所得到的結果時必須小心謹慎，雖然預測變項的加權係數 $\beta_1, \cdots, \beta_v$ 最大化了樣本的複相關，然而會因爲抽樣誤差的緣故，造成複相關 R 的值虛假地高，因此由樣本所得到的複相關 R 不是母群的複相關之不偏估計數。

在本研究中，將以某是 2006 年一班智能資優學生甄選資料中的個別智力測驗分數作爲效標變項，以參加甄選學生在團體智力測驗三個分測驗的分數作爲預測變項，來探討此法的結果。

五、主成份分析法

主成份分析可以幫助找出一群變項中主要的決定成份爲何，也可以進一步將一群具有相同或不同單位的變數，透過賦於適當的權重，綜合成一決策性指標，以簡化決策（林師模、陳苑欽，2004）。主要目的是在找到適當的特徵向量，來將原來的觀察分數轉換成「成份分數」（component score），此種成分分數是各變項原始分數的線性組合，其變異數即爲各變項之相關矩陣的特徵值（eigenvalue），特徵值依大小順序排列，第一個特徵值最大，是所有變項的共同成份或因素，其餘各成份之重要性或所能解釋之變異量遞減，各成份之間彼此無關（吳裕益，2011）。將主成份分析法應用在組合分數的議題上，主要是以所計算得到的第一個主成份分數作爲組合分數。主成份分析的模式如下：

$$z_C = u_1 x_1 + u_2 x_2 + \cdots + u_p x_p$$

其中是所得到的第一個主成份分數，也是組合分數，$u_1, u_2, \cdots, u_p$ 是成份加權係數。在探討此法的結果時，同樣將以參加一般智能資優學生在團體智力測驗的三個分測驗上的得分作爲觀察變項，來計算組合分數以與學生在個別智力測驗上的得分進行比較。

在利用主成份分析法來計算組合分數時，涉及主成份分數的計算，在統計軟體 SPSS 之中，有多元迴歸法、Bartlett 法、和 Anderson-Rubin 法等三種主成份分數計算的方法可供選擇（Pett, Lackey, & Sullivan, 2003）。其中利用迴歸法所得到的因素分數之平均數爲 0，變異數等於被估計的因素分數與真正的因素值之複相關的平方，當因素彼此是直交時，因素分素可能仍會有相關存在；由 Bartlett 法所得到的因素分數之平均數等於 0，而各變項之特殊性因素離均差平方和被最小化，因素分數間仍然可能有相關存在；而 Anderson-Rubin 的方法是 Bartlett 法的修正，確保估計的因素之直交性，而計算得到的音死分數之平均數爲 0，標準差等於 1，且是無相關的。在本研究的情形中，因爲僅計算第一個主成份的分數，所以成份分數之間是否有相關的問題應當不是特別重要，再者，就此三種方法而言，以迴歸法的計算是最直接容易的，因此本研究以主成份分析法計算組合分數時，將以迴歸法爲之。

六、研究目的

本研究擬以筆者所參與的題庫建置案中的資料爲研究資料（包含模擬和實徵兩個資料集），探討這幾種組合分數計算的方法所得到的結果之異同。主要的研究目的如下：

（1）利用 Wainer 和 Thissen（2001）所介紹的信度加權法，來計算出各科的加權數以得到組合分數。

（2）利用 IRT 的方法中的 3PL 模式來估算考生在國文、數學、自然與社會等四科的能力值，然後利用主成份分析來得到這四個能力值的主成份分數，並將此分數與在（1）中所得到的組合分數進行比較。

（3）將考生的國文、數學、自然與社會四科之作答反應合併在一起，視爲一個學科，然後以 3PL 模式來估計考生的能力值，並將此能力值與由（1）所得到的組合分數和由（2）所得到的主成份分數等加以比較。

（4）使用 2006 年某市國小二年級升三年級的一般智能資優生甄選資料，以該項資料中團體智力測驗的三個分測驗的分數爲觀察變項，個別智力測驗的分數作爲效標變項，探討主成份法與迴歸分析法在組合分數計算上的應用，以及目前該題庫建置計畫所用的分數計算方法之可用性。

貳、研究方法

一、研究對象及資料

對該項題庫建置計劃而言，在進行常模建置時，考慮到要讓同一群考生在學期中用四節課的時間來同時作答國文、數學、社會與自然等四個科目，似乎會造成學生過多的負擔以及教師上課進度追趕的問題；因此當初在擬定抽樣計畫時，盡量讓考生僅作答一個科目的試題，只有部分考生回答兩個科目。即使如此，透過共同考生設計，勉強可以得到如表 1 和表 2 所呈現的各科分數的相關係數，雖然那些相關係數並不是由同一群人同時作答四個科目所得到的。爲得到由同一組人作答所有四個學科的資料，筆者擬以模擬的方式產生本研究第一個部分所需要之作答反應資料。另外 2006 年某市甄選小學二年級升三年級之一般智能資優學生甄選的資料，將被用來說明主成份法以及多元迴歸法

在組合分數計算上面的應用。選擇這個資料是因爲在該市一般智能資優學生甄選資料中，包含有參加甄試的學生在題庫建置計畫所編製用來作爲初選工具之團體智力測驗以及複選時所用的個別智力測驗所得到之分數。團體智力測驗包含語文推理、數量推理及圖形推理三個分測驗，適合用來計算代表智力總分之組合分數，而參加甄試的學生在複選的個別智力測驗分數可作爲效標，適合用來探討多元迴歸法以及主成份法之效果以及兩者之間的關連。

要模擬每個考生在國文、數學、社會和自然等四個科目上的答題反應，需要每個考生在各個科目上面的能力參數、以及各科的試題參數。本研究中，各科的試題參數取自該題庫建置案所提供的試題參數，爲了滿足各科能力值之間的相關與由觀察分數所得到的類似，亦即與表 1 或表 2 中的相關矩陣類似，能力參數的產生得利用 Morgan（1984, p. 86 & p. 88）所提到的方法。Morgan 的方法主要如下：假如相同考生在四個科目上得分的相關矩陣爲 $\underline{R}$（例如表 1 或表 2 中的相關矩陣），則對 $\underline{R}$ 進行 Cholesky 分解，可得到一個下三角形（或上三角形）的矩陣 $\underline{A}$。若隨機從 $N(0,1)$ 抽樣 1000 個能力值作爲考生的國文能力，得到一個 1000×1 的矩陣或向量 $\underline{\theta_1}$，依照此方式，我們可以得到數學能力 $\underline{\theta_2}$、社會能力 $\underline{\theta_3}$ 和自然能力 $\underline{\theta_4}$；其中 $\underline{\theta_1}$、$\underline{\theta_2}$、$\underline{\theta_3}$ 和 $\underline{\theta_4}$ 因是分別產生得到的，可以認定是相互獨立的。若 $\underline{\Theta}=[\underline{\theta_1},\underline{\theta_2},\underline{\theta_3},\underline{\theta_4}]$ 爲 1000 名考生在四個科目上面的能力值矩陣，$\underline{A}\underline{\Theta}'$ 的轉置矩陣就是用來產生 1000 名考生在四個科目的答題反應所需的能力參數，這些能力參數可以視爲模擬資料時考生之真實能力。有了用來產生答題反應資料的能力值之後，可利用各科的試題參數和這些能力值，以 3PL 模式來產生各科的答題反應，涂柏原（2008）亦是利用這個方法來產生模擬的資料。

表 1 五年級各科之間的相關矩陣

	國文	數學	社會	自然
國文	1			
數學	.660	1		
社會	.736	.576	1	
自然	.533	.634	.713	1

註：用來計算各科分數之相關的人數由 261 至 643 人不等。

表 2 六年級各科之間的相關矩陣

	國文	數學	社會	自然
國文	1			
數學	.643	1		
社會	.781	.741	1	
自然	.697	.720	.769	1

註：用來計算各科分數之相關的人數由 231 至 673 人不等。

在本研究中首先以表 1 所列的相關矩陣，依照前一段所描述的方法，產生 1000 名五年級的考生在國文、數學、社會和自然等四個學科測驗上的作答反應；在得到五年級的資料之後，另外再以表 2 的相關矩陣，依相同程序模擬 1000 名六年級考生的原始作答反應資料。這些模擬的資料主要用來探討以信度加權法和 IRT 方法所得到的組合分數之異同。

二、研究所要進行的分析

當模擬資料預備妥當之後，下列三個分析將被進行：（1）利用信度加權法估算考生的組合分數；（2）利用 IRT 的方法估算考生的組合分數；（3）利用主成份法估算考生的組合分數。在利用信度加權法來計算組合分數時，五、六年級考生各科之原始分數和能力值，將分別被用來計算組合分數。

要利用由模擬資料所得到的各科之 IRT 能力值來估算組合分數時，筆者先利用 3PL 模式來估計考生國文、數學、社會及自然等四科的能力值，然後利用信度加權法來計算組合分數；再者，將考生在四個學科上的答題反應，視爲在同一個試題很多的測驗上面的作答反應（也就是當作一個科目看待），然後利用 3PL 模式來估計考生的能力；在後面的這個作法中，每一個考生只有一個能力值被估計出來，這個能力值將會與前面所計算得到的組合分數一起比較。也就是如果直接將這四個學科視爲一個測驗組（test battery）的四個分測驗，對所有的資料一起估計的話，所得到的能力估計值（即組合分數）將會是如何？與前項用信度加權法所得到的組合分數之相關是否會如同 Childs 等人（2004）所觀察到的那麼高嗎？這是本研究將探討的。

該題庫建置計畫中除了二至七年級的國文、數學、社會及自然等四科（七年級另有英文）成就測驗之外，還有團體智力測驗一份兩式，包含語文推理、數量推理、圖形推理等三個分測驗。該團體測驗自從 2005 年之後，幾乎每一年都會被部分縣市用來作爲一般智能資優學生初試的工具；如前所示，2006 年某縣市的甄選資料將用來探討利用迴歸分析法以及主成份法計算組合分數的情形。在 2006 年的甄選中，該團體智力測驗被用來作爲初試的工具，而複試所用的是某商業化的個別智力測驗，因此這個部分的分析，也可作爲該題庫計畫所編製的團體智力測驗之效度證據。

在該團體測驗的編製過程中，語文推理、數量推理及圖形推理等三個分測驗的原始分數都先經常態化轉換成 T 分數，然後三個分測驗的 T 分數被加總在一起，得到一個 T 分數總分，這個總分再一次被常態化轉換成 T 分數，作爲團體智力測驗之智力分數。因爲有三個分測驗之分數及代表智力的總分，也有個別智力測驗的結果，因此筆者就嘗試利用多元迴歸和主成份分析，來探討組合分數的議題，檢驗由多元迴歸、主成份分析所得到的團體智力測驗之組合分數以及團體智力測驗所提供的智力總分（最後得到之 T 分數）之間的關係。其中，利用多元迴歸法時是以團體智力測驗的三個分測驗的 T 分數來估計個別智力測驗的分數，所得到的估計值作爲團體智力測驗的組合分數；而利用主成分分析法時，則是以團體智力測驗三個分測驗的 T 分數作爲觀察分數，來計算第一個主成分，以此作爲團體智力測驗的組合分數。

參、結果與討論

底下將分成四個小節來呈現本研究的結果，各科模擬資料之間的相關係數以及信度係數將先被呈現，然後利用信度加權法所得到的結果將在第二個小節中呈現，在第三個部分則是與 IRT 方法有關的結果，利用多元迴歸以及主成份分析法來計算組合分數所得

到的結果將最後呈現。

一、模擬資料的描述統計

根據本研究的設計所得到的各科模擬資料之間的相關以及信度係數，呈現於表 3 和表 4 之中。與呈現在表 1 和表 2 的數據比較，可發現所模擬的五年級各科資料間的相關係數比原先的相關係數稍低一些，其中的原因可能是數學、社會與自然三科模擬資料的信度稍微低了一些，尤其是數學。但是六年級的資料與原來的資料比較近似，整體而言，應當還不算太差，因此可以繼續進行原先所計畫的分析。

二、利用信度加權法估算考生的組合分數

表3 五年級各科原始分數間的相關、信度係數及題數（ $N = 1000$ ）

	國文	數學	社會	自然
國文	.914			
數學	.549**	.823		
社會	.629**	.438**	.871	
自然	.436**	.515**	.618**	.870
題數	75	42	69	59

註：主對角線上的信度係數。

** p<.01

表4 六年級各科原始分數間的相關、信度係數及題數（ $N = 1000$ ）

	國文	數學	社會	自然
國文	.922			
數學	.613**	.896		
社會	.744**	.664**	.905	
自然	.648**	.620**	.704**	.897
題數	80	46	76	63

註：主對角線上的信度係數。

** p<.01

表 5 信度加權法所得到各科之加權係數以及組合分數之信度係數

	五年級		六年級	
	RS 加權數	θ 加權數	RS 加權數	θ 加權數
國文	.38449	.38225	.30929	.30785
數學	.15165	.15386	.21234	.21306
社會	.24572	.24515	.25619	.25634
自然	.21814	.21873	.22215	.22276
組合分數信度係數	.95391	.95460	.96891	.96875

註：RS 加權數爲利用原始分數所得到的各科之加權係數，θ 加權數爲以各科 IRT 能力值所計算得到的加權係數。

根據 Wainer 和 Thissen（2001）所示範的方法，分別將表 3 和表 4 中的相關矩陣代入計算公式，即可以得到各科之加權數以及組合分數的信度係數；除此之外，筆者另外將各科利用 IRT 模式所估計得到的能力值之間的相關係數矩陣，也利用信度加權法，來計算各科的加權數以及組合分數之信度係數，所得到的結果整理呈現於表 5 之中。由表 5 中的數據來看，可以發現利用答對題數原始分數所得到的加權數和組合分數之信度係數與由 IRT 能力值所得到的結果類似，四個組合分數之信度係數在.95391~.96891 之間，算是非常的高。

三、利用 IRT 的方法估算考生的組合分數

在這裡所探討的與 Wainer 和 Thissen（2001）文中所提到的不同，在他們的研究報告中，所處理是同一個科目有不同題型的 mixed-format 測驗之組合分數問題。而在此處，筆者所探討的是如何將四個測驗的分數加總得到一個最佳的分數之問題；也就是 Kolen, Wang 和 Lee（in press）所提到的第一種情形的組合分數問題。

Wang(1985)利用單一向度的三參數羅吉斯模式(three-parameter logistic model, 3PL）模式來分析二個向度的資料，以檢視其估計誤差，發現在那種情況之下，3PL 模式所估計得到的θ值恰好反應了考生在那二個潛在變項的主成份（principle component）分數。她將此能力估計值稱爲 reference of composite。因此，如果我們將考生在四個學科上面的作答反應集合在一起，當作是單一科目的試題來看待，然後利用 3PL 模式來估計考生的能力值時，不知道所估計出來的θ值是否恰好是那四個科目能力值所計算得到的主成份分數？如果是的話，那麼將估出來的θ視爲考生的組合分數是可行的，必要的話，再將θ值加以直線轉換成比較容易溝通解釋的值，即可作爲測驗分數報告給考生。

表6 五年級各科能力值與主成份之間的相關（ $N=1000$ ）

	國文	數學	社會	自然	整體能力	組合(主成份)
國文	1					
數學	.568**	1				
社會	.637**	.461**	1			
自然	.449**	.531**	.631**	1		
整體能力	.850**	.722**	.821**	.753**	1	
組合(主成份)	.819**	.781**	.846**	.802**	.969**	1

註：整體能力表示將所有的試題視爲是同一個科目時所估計得到的能力值。

** $p<.01$

表7 六年級各科能力值與主成份之間的相關（ $N=1000$ ）

	國文	數學	社會	自然	整體能力	組合(主成份)
國文	1					
數學	.604**	1				
社會	.732**	.665**	1			
自然	.639**	.624**	.704**	1		
整體能力	.860**	.794**	.882**	.819**	1	
組合(主成份)	.862**	.834**	.901**	.858**	.971**	1

註：整體能力表示將所有的試題視爲是同一個科目時所估計得到的能力值。

** $p<.01$

表 8 五年級主成份分數、能力值與組合分數之相關（ $N=1000$ ）

	主成份1	主成份2	原始總分	整體能力	組合(信度)
主成份1	1				
主成份2	.990**	1			
原始總分	.987**	.996**	1		
整體能力	.969**	.962**	.968**	1	
組合(信度)	.982**	.991**	.999**	.969**	1

註：主成分1是利用各科能力值計算得到的主成份分數，主成份2是利用各科原始分數計算得到的主成份分數，原始總分是所有試題的答對題數原始分數，整體能力即將所有的題目視爲是同一個測驗的，所估計得到的能力值；組合(信度)是利用信度加權法所計算得到的組合分數。

** $p<.01$

表 9 六年級主成份分數、能力值與組合分數之相關（ $N=1000$ ）

	主成份1	主成份2	原始總分	整體能力	組合(信度)
主成份1	1				
主成份2	.992**	1			
原始總分	.990**	.999**	1		
整體能力	.971**	.968**	.969**	1	
組合(信度)	.990**	.999**	1.000**	.969**	1

註：主成分1是利用各科能力值計算得到的主成份分數，主成份2是利用各科原始分數計算得到的主成份分數，原始總分是所有試題的答對題數原始分數，整體能力即將所有的題目視爲是同一個測驗的，所估計得到的能力值；組合(信度)是利用信度加權法所計算得到的組合分數。

** $p<.01$

筆者首先以考生在四個科目上所得到各科的能力值，利用主成分析來計算各個考生之主成份分數；同時利用 3PL 模式來估算四科混合在一起之後的作答反應之能力值，以比較這二者之差異情形。以五年級考生在四個科目上所得到的能力值計算主成份分數時，國文、數學、社會和自然四科的加權係數分別爲.819、.781、.846 和.802；六年級的加權係數爲.862、.834、.901 和.858。五、六年級各科的能力值與利用各科能力值計算得到的主成份分數之間的相關分別呈現在表 6 和表 7 之中。可以看到的是，各科的能力值與主成份分數之相關大致都在.8 以上；將所有的試題皆視爲是屬於同一個科目時，兩個年級所估計得到的能力值與利用各科能力值所計算得到的主成份分數之間的相關都在.97 上下。此項結果比 Childs 人（2004）所得到的值.9 還要高，究其原因可能是 Childs 等人的研究中對具有最高信度的選擇題所給予的權重比簡答題還要低的緣故，若 Childs 等人給予選擇題較高的權重，或許他們所發現的相關值就會比.9 大。

在表 8 和表 9 中，也一併呈現利用信度加權法對原始分數所計算得到的組合分數（在表 8 和表 9 中稱爲「組合（信度）」）與其他分數的相關，根據 Wainer 和 Thissen 的作法，如本文公式（1）所示，各科的原始分數需要先加以標準化（即轉換爲 z 分數）之後，才與相對應的加權數相乘，加總後得到組合分數。由表 8 中可看到這個由各科原始分數所計算得到的組合分數與利用考生在各科的能力值計算得到的主成份分數之間的相關達到.982（ $p<.01$， $N=1000$ ），與將四個科目視爲一科所估算得到的能力值之間的相關

則爲.969（ $p<.01$, $N=1000$ ），由四個科目的能力值所計算得到的主成份分數（即組合分數）與所有的試題視爲一科來加以得到的能力值，應都可以視爲是 Wang（1985）的 reference of composite 之估計值，二者之間有高的相關，應當是合乎預期的。六年級的結果在表 9 之中，與表 8 的結果類似。

四、主成份與多元迴歸法

研究者利用某市 2006 年小學一般智能資優甄選的資料來進行主成份法與多元迴歸法的應用之分析，在這份資料中，有當年度參加甄選的學生之團體智力測驗和個別智力測驗的分數，而團體智力測驗爲語文推理、數量推理和圖形推理等三個分測驗所組成。在該資優生甄選的程序中，參加甄選的學生必須通過團體智力測驗初試的門檻，始得參加個別智力測驗的複試，因此參加初試的人數比複試的人數多。該項資料中各個（分）測驗之描述統計與相關等數據，分別呈現在表 10 與表 11 之中。

表10 團體智力測驗與個別智力測驗之描述統計資料

	人數	最小值	最大值	平均數	標準差
語文推理	1802	15	83	61.68	10.08
數量推理	1802	28	83	60.20	9.40
圖形推理	1802	23	82	61.30	9.74
團體智力測驗總分	1802	25	85	63.59	10.06
個別智力測驗	630	72.00	145.60	114.33	10.45

註：除個別智力測驗分數外，其餘皆爲T分數。

表11 團體智力測驗各個分測驗分數與個別智力測驗之相關

	語文推理	數量推理	圖形推理	團體智力測驗總分	個別智力測驗成績
語文推理	1				
數量推理	.566**	1			
圖形推理	.547**	.561**	1		
團體智力測驗總分	.843**	.835**	.834**	1	
個別智力測驗	.618**	.610**	.626**	.955**	1

註：除了那些與個別智力測驗有關的相關係數是用630人計算以外，其餘的相關係數皆利用1802人的資料計算的。

** $p<.01$

由表 11 中的數據可以發現團體智力測驗的總分與個別智力測驗分數之間的相關係數達.955（ $p<.01$, $N=630$ ），若以 Gulliksen（1950/1987, p. 137）書上第十一章的公式（18）進行校正，得到在 1802 人的情形下，相關係數可以達到.999。由這項數據可以看到參加甄選的學生在兩個測驗的得分之排序相當一致，因爲個別智力測驗是商業發行的標準化測驗，應是有充分的效度證據支持該測驗所測量的是智力，而考生在題庫建置研究計畫所編製的團體智力測驗之得分，應也可說是反應了考生的智力這個夠念，因此這項數據可作爲團體智力測驗一項構念效度中之效標關聯的效度證據。

利用這項資料，筆者以團體智力測驗的三個分測驗的分數來進行主成份分析，計算主成份分數以作爲組合分數，然後計算主成份分數與個別智力測驗和團體智力測驗總分之相關。另外以團體智力測驗的三個分測驗作爲預測變項來預測個別智力測驗的分數，以所得到的預測分數作爲組合分數，計算預測分數與個別智力測驗實得分數和團體智力測驗總分之間的相關。由這二種方法所得到的各個分測驗之加權係數呈現在表 12 之中，主成份分數與個別智力測驗分數和團體智力測驗總分之間的相關，以及多元迴歸預測的分數與實際個別智力測驗分數及團體智力測驗總分之間的相關，亦一併呈現在表 12 之中。此外，由主成份分析所計算得到的組合分數與由多元迴歸所得到的組合分數之間的相關達到.999（p<.01, N=630），相關這麼高的原因可能是因爲主成份分數是透過 SPSS 軟體利用迴歸法計算得到的。

表12 主成份分析的加權係數以及迴歸分析之係數（N=630）

	成份加權數	非標準化迴歸係數	標準化迴歸係數
語文推理	.839	.833	.512
數量推理	.845	.746	.422
圖形推理	.836	.799	.521
與團測分數之相關	.997**	.997**	.997**
與個測分數之相關	.946**	.948**	.948**

註：上半部是加權係數，而最後兩列是分別利用主成份分析和多元迴歸法所得到的組合分數與團體智力測驗和個別智力測驗分數之相關。

**p<.01

從表 12 的相關係數來看，由主成份法所得到的組合分數與個別智力測驗分數之相關係數爲.946（$p<.01$, $N=630$），而以多元迴歸所得到之預測分數與個別智力測驗分數之相關爲.948（$p<.01$, $N=630$），而試題庫建置計畫所用的方法所得到的團體智力測驗總分與個別智力測驗分數之間的相關爲.955（$p<.01$, $N=630$，呈現於表 11 之中），看起來這三種方式所得到的組合分數都是相似的。因爲語文推理、數量推理和圖形推理三個分測驗彼此之相關的值接很接近，因此各個分測驗主成份加權係數的值也接近，因而造成由主成份分析所得到的組合分數與團體智力測驗的總分有高的相關。而迴歸分析的部分，效標變項是學生在個別智力測驗上的得分，而如上面所描述的，學生在團體智力測驗所得到的分數與在個別智力測驗上面的分數之相關達到.95，校正後的相關幾乎爲 1，因此以個別智力測驗分數爲效標或以團體智力測驗分數爲效標所得到的結果基本上應當是近似的；利用最小平方法進行多元迴歸分析的特色是，預測的與實際的效標分數之間的相關會被最大化，因此以多元迴歸分析來計算組合分數時，所得到的組合分數與個別智力測驗分數之間的相關會被最大化，在此例中，相當於所得到的組合分數與觀察得的團智力測驗的分數之間的相關也被最大化。然而，團體智力測驗的分數本身就是由三個分測驗的分數加總在轉換爲 T 分數得到的，因此在此中，利用迴歸分析所得到的組合分數自然就與實際的團體智力測驗分數之間有極高的相關。

肆、結論與建議

本研究以模擬的資料來探討信度加權法和 IRT 法在計算組合分數上之表現，主要發現利用信度加權法來計算組合分數時，無論是用各個學科原始分數或是 IRT 能力值之間的相關矩陣，所得到的各個學科分數之加權係數皆十分接近，且所計算得到的組合分數與將四個學科的試題合起來視爲是同一個學科的試題來處理所估計得到的 IRT 能力值之間有高的相關存在，五年級和六年級的相關係數分別爲.969 和.971，顯示這二者所代表的構念相近似。在以某市一般智能資優生甄選的資料所進行的分析中，可以看見利用主成份分析法所得到團體智力測驗的組合分數與原先所用的測驗總分之間有接近完美的相關存在，顯示題庫建置計畫在編製團體智力測驗時所計算的測驗總分所代表的構念與由三個分測驗所計算得到的主成份之意義近似，此一結果支持了原來的團體智力測驗分數計算的方法，而題庫建置案所編製的團體智力測驗與商業發行的個別智力測驗分數之間的相關達.95，提供了團體智力測驗一個效度證據。

由本研究所得到的結果來看，若有一個良好的效標存在，迴歸分析法應不失爲一個計算組合分數的好方法，因爲可以很容易計算得到；就 IRT 的方法來說，若將所有學科的試題通通視爲是同一個科目的，而利用 3PL 模式來估算，所得到能力值與答對題數總分與由四個學科的能力值所得到之主成份亦有相當高的相關，顯示用這種方法來得到組合分數應該也是可行的。在本文中，是將所有的試題當做是同一個測驗的來看待，若是將所有的試題依舊視爲四個科目，且允許各科的能力值之間有相關存在，而要以 IRT 的方法來計算答對題數分數轉換成量尺分數的組合分數，則得用到 Kolen, Wang, 和 Lee（2012）所提議的方法。基本上，由各科的原始分數或是各科的能力值所計算得到的主成份值和將四科題目視爲一科所得到的 IRT 能力值，與利用最大信度法所得到的組合分數之間，彼此之間皆有幾近完美的相關存在。究其原因，如前面所提到的，Wainer 和 Thissen（1993）認爲 IRT 所隱含的加權係數基本上是信度係數的函數，而主成份分析中第一個主成份是解釋觀察變項之變異量之最大者，若將觀察變項的變異數和視爲觀察變項之變異數，而主成份所解釋的變異數視爲真分數變異數，那麼二者之比值等於古典測驗理論中信度的值；所以，主成份分數的計算也可以算是最大信度的取向。若由這樣的觀點來看表和表 9 中的結果，也許就可以理解爲何他們是如此的接近。

組合分數雖是個古老的議題，但是仍然有許多是值得探討的，除了本文所用到的三、四種方法，以及新興的 IRT 法外，Wang 和 Stanley（1970）文中所提到其他的方法，尚有一些方法是值得研究者去嘗試的。近年來有關組合分數研究的方向，主要有三：（1）計算組合分數的條件測量標準誤（conditional standard error of measurement, CSEM），例如，Kolen, Wang 和 Lee（in press）將利用 IRT 的方法來計算由答對題數原始分數所得到的組合分數之條件標準誤計算方法（Kolen, Zeng, & Hanson, 1996）延伸到多向度 IRT 的情境，Chang, Teng 和 Wu（2010）探討將 Kolen, Wang 和 Lee 等人的方法應用在國中基測（BCTEST）總分計算的可行性；（2）雖然分量表分數與組合分數乍看之下不同，但是由數個分測驗所組成的測驗之總分其實就是個組合分數，在許多情境中，分測驗分數的計算是需要的，因此近年來許多探討分量表分數（subscores）的文獻被提出來。例

如，Sinharay & Haberman（2008）將當時被用來計算分測驗分數的方法應用在數個測驗資料上，結果發現分測驗分數僅對少數的測驗能夠提供附加的價值；Haberman & Sinharay（2010a, 2010b）探討了如何利用 MIRT 來計算分測驗分數，Sinharay, Puhan & Haberman（2011）介紹了在文獻中被用來計算分測驗分數的五種方法，以及四種替代的方法，並提出一些建議工實務工作者參考；而 Sinharay（2010a, 2010b）評介了分測驗分數的計算對於測驗總分的報導增加了哪些價值，Sinharay & Haberman（2011）進一步探討分測驗分數的等化問題。（3）發展同時可以計算組合分數和分量表分數的 IRT 模式被發展出來，例如 higher order IRT（HO-IRT）模式（de la Torre & Song, 2009），以及 Sheng 和 Wikle（2008）的階層結構之多向度 IRT 模式或是 Brandt（2008）所提出來的次向度模式（Rasch model including subdimensions）模式等皆可同時計算全測驗與分測驗的分數。這些新近的發展方向，爲組合分數相關的研究注入了新的動力，也是研究者未來可以努力的方向。

參考文獻

吳裕益（2011）。**因素分析**。未出版。

林師模、陳苑欽（2004）。**多變量分析：管理上的應用**。台北市：雙葉書廊。

涂柏原（2008）。BCTEST 量尺分數轉換議題探討。**教育研究學報，42**(2)，67-82。

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In von Davier, M. & Hastedt, D. (Eds.) *Issues and methodologies in large-scale assessments*. IERI Monograph Series Vol. 1. Princeton, NJ: IEA-ETS Research Institute.

Chang, S.-W., Teng, S., & Wu, Y.-T. (2010, May). *Explorations of composite scores under the multivariate proficiency distribution using IRT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver.

Childs, R. A., Elgie, S., Gadalla, T., Traub, R., & Jaciw, A. P. (2004). IRT-linked standard errors of weighted composites. *Practical Assessment, Research & Evaluation*, *9*(13). Retrieved December 14, 2005 from http://PAREonline.net/getvn.asp?v=9&n=13.

de la Torre, J. & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620-639.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education and Macmillan.

Gulliksen, H. (1950/1987). *Theory of mental tests*. New York: John Wiley.

Haberman, S. J. & Sinharay, S. (2010a). *How can multivariate item response theory be used in reporting of subscores?* (ETS Research Report No. RR-10-09). Princeton, NJ: ETS.

Haberman, S. J. & Sinharay, S. (2010b). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209-227.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, *17*(3), 221-240.

Kolen, M. J., & Hanson, B. A. (1989). Scaling the ACT assessment. In R. L. Brennan (Ed.), *Methodology Used in Scaling the ACT Assessment and P-ACT+* (pp. 35-55). Iowa City, IA: ACT, Inc.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129-140.

Kolen, M. J., Wang, T., & Lee, W.-C. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing*, *12*(1), 1-20.

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

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

McDonald, R. P. (1968). A unified treatment of the weighting problem. *Psychometrika*, *33*(3), 351-381.

Morgan, B. J. T. (1984). *Elements of simulation*. New York: Chapman and Hall.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). New York, NY: American Council on Education/Macmillan Publishing.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications.

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice*, *20*(1), 16-19.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*(3), 413-430.

Sinharay, S. (2010a). *When can subscores be expected to have added value? Results from operatonal and simulated data*. (ETS Research Report No. RR-10-16). Princeton, NJ: ETS.

Sinharay, S. (2010b). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150-174.

Sinharay, S. & Haberman, S. (2008). *Reporting subscores: A survey*. (ETS Research Memorandum No. RM-08-18). Princeton, NJ: ETS.

Sinharay, S. & Haberman, S. (2011). *Equating of subscores and weighted averages under the NEAT design*. (ETS Research Report No. RR-11-01). Princeton, NJ: ETS.

Wainer, H., & Thissen, D. (1993).Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. Applied Measurement in Education, *6*(2), 103-118.

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Ed.), *Test Scoring* (pp. 23-72). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT*. Unpublished manuscript, University of Iowa.

Wang, M., & Stanley, J. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*, 663-705.

投稿日期：2011 年 04 月 17 日
修正日期：2011 年 12 月 25 日
接受日期：2011 年 12 月 31 日

Issues Related to the Composite Score of Achievement Tests

Bor-Yaun Twu

Associate Professor, Department of Education,
National University of Tainan

Szu-Cheng Lu

Doctoral Graduate Student, Department of Education,
National University of Tainan

Abstract

The purpose of this study is to survey four ways for computing the composite score of four tests from a test battery, which was simulated using the item parameters taken from a test item banking project. More than ten approaches of constructing composite scores can be found in the literature. Among them, reliability weighting, IRT weighting, principle component analysis and multiple regression were investigated in this study. With the data from Chinese, Math, Science, and Social Science achievement tests, the weights given by raw score and IRT trait score using reliability weighting method are very similar. Treating all items from four subject areas as items of a test and calibrating the trait level with 3PL model, the resulted trait scores has a Pearson correlation coefficient of .969 and .982, respectively, with the principal component score and composite score obtained from traits from that four achievement tests for the fifth grader, and .971 and .990, respectively, for the sixth graders.

Finally, both the principal component scores, obtaining from the three subscales from a group intelligence test, and predicted individual intelligence test score correlated highly with the observed score of the individual intelligence test, near .95 for both cases.

Key words: composite score, reliability weighting, IRT weighting, principle component analysis, multiple regression