

驗證性因素分析與概化部份計分模式 之模擬比較

李昭璫

國立政治大學教育研究所研究生

摘要

本研究主要在探討因素分析模式和概化部份計分分析模式對潛在特質估計的影響，並將因素分數、能力值視為潛在特質，進行比較。此外，亦並探討不同的鑑別度、資料合併需求與模式的交互影響，研究者透過比較各模式相關、RMSD估計的精確，及變異數分析，得到如下的結論。1、在嚴格控制均勻的鑑別度、資料合併需求下，若只求相關值，概化部份計分並無較因素分析估計精準。2、透過變異數分析，不論模式、長度、鑑別度、資料合併需求皆達顯著。3、由因素分析所得的資料隨著因素負荷量、鑑別度、長度之上升與資料合併需求之下降，相關值規律上升，RMSD 規律下降。4、本研究仍未能證明因素分析優於概化部份計分模式，因為在一份真實的測驗，鑑別度、資料合併需求並不會完全均等。

關鍵詞：因素分數、因素分析、概化部份計分模式

The Comparison of Simulation between Confirmatory Factor Analysis and Generalization Partial Credit Model

Chao-Yun Lee
Graduate Student, Department of Education
National Chengchi University

Abstract

The main objective in the study is to analyze whether the use of different analytic model have an impact on latent trait estimation for confirmatory factor analysis and generalization partial credit model (GPCM). In the study, ability and factor score are regarded as latent trait. In addition, the study also examines whether test length discrimination and the need for collapsing interact with the models. For this reason, ANOVA is used. The cutoff point of the difference of RMSD and correlation between true latent trait and estimation which is obtained from stimulation is also used. By these ways, some results are inferred. First, under the same discrimination and need for collapsing, GPCM never surpasses factor analysis. Second, by ANOVA, it is significant regardless of any discrimination need for collapsing length and model. Third, the correlation grows and RMSD decline with the increase of discrimination, length and factor loading and decrease of need for collapsing by the data which is analyzed by factor analysis but not by GPCM. Lastly, in this research it is not proven that factor analysis excels GPCM, because in the true test, the discrimination and need for collapsing is not the same with each other.

Key words: factor score, factor analysis, generalization partial credit model

壹、前言

因素分析是測驗編製的基礎，也是結構方程式中測量模式的基礎（余民寧，2006），然而，近來隨著現代測驗理論，即 IRT 的發展，其是否能測得潛在特質，備受質疑。許多學者紛紛探究 IRT 理論是否能應用於教育心理界最常用的 Likert 量表上，因此從而發明了單向度的多元計分 IRT，及各種多向度的 MIRT（余民寧，2009）。然而，並非每一種現代測驗理論都有辦法立刻付諸實現。多元計分的 IRT 中，有些受限於軟體尚未普及穩定，如連續比率模式，在 BILOGMG、Parscale、Conquest 皆無法估計；有些則是理論尚未成熟，如多向度多元計分（余民寧，2009），亦有理論模式在實驗室的架構下早已被證明不佳的如等級反應模式（Park, 2000），因此。本文所探討的單向度多元計分 IRT，其焦點會集中於已有普遍穩定軟體及理論成熟的概化部份計分模式（generalized partial credit model, GPCM），比較其與傳統因素分析之異同。

另外，比較因素分析與概化部份計分法，不只是因為概化部份計分的能力值可以視為因素分數，更重要的是，如果證明概化部份計分可在估計潛在特質上較因素分析模式優越，則可使用「混合模型」，結合概化部份計分與迴歸模型，建立一套新的「測量-迴歸模型」。因此，比較這兩種模型深具意義。

本研究之研究目的，主要在由概化部份計分模式產生的資料，經由因素分析與概化部份計分分析所得之潛在特質與真實潛在特質之相關與平均平方差之根（root mean square difference, RMSD），並進行比較。因此，本研究目的如下：

- 1、以概化部份計分模式模擬產生之五點式量表得分，以概化部分計分模式分析及因素分析估計所得之潛在特質，與真實之潛在特質之平均平方差之根（RMSD）及相關。
- 2、以因素分析模擬產生之五點量表得分，以概化部分計分模式及因素分析估計所得之潛在特質，與真實之潛在特質之平均平方差之根（RMSD）及相關。
- 3、比較兩個模式分析出來之 RMSD 與相關之差異。
- 4、探討模式、長度、鑑別度、資料合併需求（need for collapsing），對分析結果的影響及其交互作用。

貳、文獻探討

文獻探討共分五個部份，分別為驗證性因素分析、因素分數的計算、多元計分的 IRT 模式與模式的選擇、影響部份計分模式的因素、跨模式的比較。茲敘述如後，以作為研究之基石。

一、驗證性因素分析

因素分析分成兩種，一種為探索性因素分析，一種為驗證性因素分析；兩者之差異在於一個有理論基礎，一個沒有理論基礎，因此驗證性分析可以事先設定潛在特質個數，而探索性因素分析無法事先設定潛在特質個數。驗證性因素分析公式如下（陳順宇，2007）：

$$\text{Var}(x) = \Lambda_x \Phi \Lambda_x' + \epsilon_\delta \quad \dots\dots\dots (\text{公式一})$$

公式一中 $\text{Var}(x)$ 為測量模式， Φ 特徵值矩陣（彼此可能為斜交或直交）， Λ_x 為

特徵向量矩， ε_{δ} 為殘差矩陣， x 為觀察變項。

二、因素分數的計算

因素分數，簡單而言是透過因素負荷矩陣及迴歸的原理，產生各因素的分數。其公式如下（蔡坤宏，1996）：

$$B = R^{-1}A \dots\dots\dots (公式二)$$

$$F = ZB \dots\dots\dots (公式三)$$

在上述的公式中， B 代表的是相關矩陣之反矩陣及因素負荷矩陣相乘的結果，即因素分數係數， Z 則代表各受試者在該題標準化的分數， F 為因素分數因此，因素分數所代表的意義，即是該題標準化分數與因素分數係數的乘積，易言之，即可代表該受試者在該因素之標準化得分，其平均數為0，標準差為1。

本文擬用因素分數當作該受試者在該因素之潛在特質，並與現代測驗理論的潛在特質進行比較。

三、多元計分的IRT模式與模式的選擇

IRT模式眾多，從二元計分到多元計分，從單向度到多向度皆有，然而本研究主要探討的是情意測驗的IRT模式，而在情意測驗中，單向度多元計分的IRT模式仍是主流，因此亦僅將焦點集中於此。而對單向度多元計分的IRT模式的分類，有兩種分類方法，一種方法乃Park（2000）之分類法，共分兩類。另一種則為Falenchuk（2006）的分類法，共分三類，本研究以Falenchuk（2006）的分法為準，輔以Park（2000）對單向度多元計分IRT模式之分類，茲分敘如下：

（一）累積模式（cumulative model）

累積模式在Park（2000）理論中又稱為差異模式（difference model），其意義乃是將兩個相鄰的累積試題反應曲線相減，而得類別反應曲線（item response category characteristic curve），而在此模式中，最有名的是由Samejima（1969）所發展的等級反應模式（graded response model）。其模式公式如下：

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{ix+1}^*(\theta) \dots\dots\dots (公式四)$$

其中 $P_{ix}^*(\theta)$ 與二參數的對數型模式類似。易言之，此理論即是將每一個選項當作一個試題，而每一個試題都有其鑑別度、難度，而從每一個選項跨至另一個選項，即是跨越一個難度值，而一位受試者會選擇該選項，即是其在該選項有較高的機率，因此，整個試題所構成的曲線，便是類別反應曲線。

（二）相鄰模式（adjacent model）

本模式在Park（2000）理論中又稱全體分割模式（divided-by-total model），易言之，即是一種條件機率模式，此模式中最著名的即為部份計分模式和概化部份計分模式。

部份計分模式由Master（1982）提出，之後Muraki（1993）則據以進一步認為，同一試題其鑑別度應相同，遂提出概化部份計分模式（GPCM），其公式如下：

$$P_{ix}(\theta) = \frac{\exp [\sum_{j=0}^x a_i(\theta - b_{ij})]}{\sum_{k=0}^m [\exp \sum_{j=0}^x a_i(\theta - b_{ij})]} \dots\dots\dots (公式五)$$

而在相鄰模式中，除了部份計分及概化部份計分模式外，還有 Yu (1991) 所提出的二參數部份計分模式及 Andrich (1978) 所提出的評定量表模式。

(三) 連續比率模式 (continuation ratio model)

連續比率模式事實上是將每個類別視為獨立事件，因此跨過每一個類別即相乘一次，其最有名的模式為 Tutz (1990) 所提出的序列模式 (sequential model)，其公式如下：

$$\prod_{k=1}^c \frac{1}{1 + e^{a_i(\theta - b_{ik})}} \dots\dots\dots (公式六)$$

從這個公式可以得知，其將五個選項視為獨立事件而言，然而除非是成就測驗，否則很難認定五個選項具有如此關係，而事實上 Falenchuk (2006) 對多元計分 IRT 的分類亦是以成就測驗的觀點，而非以情意測驗的觀點分類。

綜上所述，此三種模式中，以累積模式發展最早，相鄰模式次之，連續比率模式又次之，然而，在實際應用上，仍以累積模式與相鄰模式較為成熟，尤其是部分計分模式應用更廣，不論是 PARSCALE 或 CONQUEST 皆有部份計分模式列為其中。至於連續比率模式，由於時間較晚，且其理論將每個類別視為獨立事件，顯與情意量表之狀況有所出入，而較接近於成就測驗，因此，在態度測驗上是不大適用的。

此外，根據 Park (2000) 所進行的模擬實驗，研究結果顯示，在嚴格控制同一份測驗的鑑別度、資料合併需求下，概化部份計分模式優於等級反應模式，然而，依照 Falenchuk (2006) 的研究則指出，概化部分模式不見得優於等級反應模式，兩篇結論之所以不一致，研究者認為主要在於參數的設定上，Park (2000) 的參數設定乃是依理論設定，但是 Falenchuk (2006) 的研究則是依據實際的參數設定，而本研究，由於採取 Park (2000) 的方法，亦是依理論設定參數，不去再去測試實際性的參數 (已由 Falenchuk 研究過，但是實際參數不可窮盡，研究者本處不擬在研究)；因此，本研究乃將研究焦點集中於鄰近模式的概化部分計分模式，而不討論 Park (2000)、Falenchuk (2006) 已研究過的等級反應模式與概化部份計分模式的理論與實際模型的比較。

四、影響部份計分模式的因素

本研究乃在探討因素分析對由概化部份計分產生之資料強韌性之檢定，而為了達成此研究目的，必須瞭解概化部份計分之潛在特質估計受何種因素影響，以便操弄此因素，方能對其強韌性進行檢定。因此，本研究先探討影響部份計分模式能力估計精準度的幾個因素，茲整理如下：

(一) 測驗長度

測驗長度對 IRT 模式的估計精準性之重要性不言可喻，不論是多向度試題反應 (謝佳穎，2009) 或者是等級反應模式 (陳雁芳，2006) 其測驗長度，其估計精準度愈高。另外，依古典測驗理論，只要題目同質性高，測驗長度增加，信度亦會增加 (余民寧，2002)。事實上，根據 Macken-Ruiz (2008)、Falenchuk (2006)、Dickenson (2005)、Park (2000) 對部份計分的研究，當測驗長度增加時，IRT 模式的估計精準性亦增加，因此測驗長度應列入強韌性檢定之考量。

(二) 樣本大小

雖然在 Falenchuk (2006) 的研究中分別以 500、1000、10000 的模擬樣本進行研究,但是其研究結果顯示,當樣本達到一定規模時(事實上,此皆是大樣本),樣本大小對能力值的估計影響不大。因此,本研究在以理論參數值進行模擬研究時,亦以大樣本進行模擬,但不對樣本數進行探討。

(三) 鑑別度

在 Park (2000) 分別將鑑別度設為 0.6、0.9、1.4,三種高、中、低的鑑別度,並進行模擬研究,研究結果顯示鑑別度提高將使能力值估計更加準確,不論是在何種長度,及多少類別下。因此,鑑別度的大小亦需納入考量。

(四) 難度值之分布

Park (2000) 對難度值的設計,他稱之為資料合併需求 (need for collapsing),易言之,即是難度值愈低,其負值愈大,需要合併類別的需求就愈大。根據此概念,將資料合併需求分為高、中、低三種資料合併需求,高資料合併需求的試題難度參數介於-1.25~-0.75,中資料合併需求的試題難度參數介於-0.75~-0.25,低資料合併需求的試題難度參數則介於 0.25~-0.25,每個類別的閾值寬度設為 1,而其閾值分別為難度值加上-1.5、-0.5、0.5、1.5。如此的資料合併需求設計,主要是因為資料合併需求具有左右對稱的性質,因此當考慮一端,即不須要再考慮另一端,只要反之類推即可。而其研究的結果顯示,在五點量表的情況下,當高資料合併需求時,其 RMSD 最大,當在低資料合併需求時,其殘差最小。但是在其他的量表並不全然如此,例如在三點量表時,RMSD 則隨著資料合併需求的升高而升高,因此必須視情況而定。

(五) 類別多寡

類別的多寡對潛在特質估計的精確性亦有影響,不過其影響往往與資料合併需求呈交互作用,易言之,並非類別少其 RMSD 就高,必須看在什麼資料合併需求下。如同上述,五點量表和三點量表的狀況並不一樣,在此不再贅敘。另外,由於一般而言,態度量表多半是五點量表,本研究亦以五點量表為研究對象,因此類別的多寡並未被視為考量的對象。

除了上述五類外, Park (2000) 的探測性研究則指出,閾值 (threshold) 的寬度亦會影響估計的情況,不過,閾值的寬度到底需要多寬,仍無理論支持,欲圖探討這類問題,多半以實徵資料所得的參數進行探討。因此綜合上述,本研究必須考慮的因素包含測驗長度、鑑別度、資料合併需求等三種不同之影響。

五、跨模式的比較

進行單向度多元計元 IRT 跨模式的比較,這種研究在國內並不風行,不論碩博士論文或者期刊都無相關文獻,但在國外已有三篇主要的博士論文。茲將與本研究有關整理如下。

(一) 三種現代測驗與因素分析比較

Falenchuk (2006) 以實際分析所得之參數,運用概化部份計分、等級反應、連續比率模式及因素分析產生四種模擬資料,形成二個三點量表及一個四點量表(分別合併最高或最低類別),並以現代測驗模式中之概化部份計分、等級反應、連續比例模式進行分析,以比較其強韌性,研究結果得到如下的結論:

- 1、當測驗長度增至五十題,不論是用何種模式產生的資料,及用何種分析,其平均平方差的根 (RMSD)、偏誤 (Bias)、相關均非常相近,但是當在 10 題時,其結果則視各種情況而不同。
- 2、由因素分析所產生的資料,由概化部份計分模式分析,比由概化部份計分產生,再由概化部份計分分析其與真正能力值之相關性更高,RMSD、Bias 更小。

事實上，由因素分析所產生的資料，再由概化部份計分模式、等級反應模式、連續比率模式分析，皆優於其各自模式產生資料再由本身模式分析。顯然的，由因素分析產生的資料具有強韌性。

- 3、各模式除了因素分析產生的資料外，其餘的，由該模式產生的資料，再由該模式進行分析，皆得最佳效果。事實上，Falenchuk (2006) 認為這三種模式互為獨立不相統屬，並無所謂的通式可言，此點與 Master (1982) 認知不合，反而更近於 Park (2000)，研究者亦以為然。
- 4、研究的結論指出無法證實哪一個理論是最有效的，或優於其他的模式。

(二) GPCM 與 GRM 的比較

Park (2000) 將單向度多元計分的現代測驗分為全體分割模式 (divided-by-total model) 和差異模式 (difference model)，以過去文獻探討的模擬方法，分別以概化部份計分、等級反應模式產生模擬資料，得到如下的結論：

- 1、概化部份計分模式能力值之估計較等級反應精確，模式較佳。
- 2、題目愈長，估計愈佳。
- 3、類別的多寡影響概化部份計分較等級反應為多。
- 4、資料合併需求的高低對概化部份計分影響更大。
- 5、題目愈少，兩個模式差異愈大。

(三) GGUM、GPCM、PCM、factor analysis

Sgammato (2009) 分別以概化部份計分模式 (GPCM)、部份計分模式 (PCM)、因素分析模式 (factor analysis)、概化等級展開模式 (GGUM) 等四種模式分析增權賦能 (empowerment) 量表和領導 (leadership) 量表，並就能力值和參數估計進行評判。研究結果顯示：

- 1、Sgammato (2009) 將因素分數 (factor score) 視為潛在特質，即現代測驗中的能力值，並將因素分數與增權賦能量表上與概化部份計分模式、概化等級展開模式的能力值進行相關分析，發現因素分數與概化部份計分模式與概化等級展開模式之能力值有高相關，皆為 0.956，但是與部份計分的能力值有較低的相關，只有 0.873，事實上，部份計分模式之能力值與概化等級展開模式、概化部份計分模式之能力值的相關亦皆只有 0.888，因此可得知，因素分析與概化部份計分、概化等級展開模式在此量表上確有高度相關。
- 2、在領導量表上因素分數與概化部份計分、概化等級展開模式、部份計分模式的能力值皆有高度相關，分別為 0.958、0.958、0.941，但是，部份計分模式與其他模式能力值之相關仍較低。
- 3、增權賦能量表不能滿足單一向度，其相關亦降低。
- 4、展開模式 (即概化等級展開模式) 與累積模式 (即部份計分模式、概化部份計分模式相似) 有高度的相似性。但在某些題目上，展開模式較不適用。

(四) 綜合討論

綜合上述三篇的研究，研究者歸納出下列幾點對本研究的啟示：

- 1、在因素分析中，因素分數事實上即是潛在特質，亦即現代測驗中的能力質，易言之，因素分數與能力值是可進行比較並求相關。更進一步言，因素分數與現代測驗的能力值都是平均為 0，標準差為 1 的常態分配。
- 2、在 Falenchuk (2006) 的研究中已指出由因素分析所製造四點、三點量表，經由各模式分析 (其中並不包含使用因素分析進行分析)，皆具有較佳效果，顯然的，因素分析所產生的模擬資料具強韌性，那麼，從這一個角度出發，因素分析亦有可能對各模式所產生的資料具有強韌性。

而回顧現代測驗之所以有辦法取代古典測驗成為主流，原因是古典測驗以總分當作能力值，然而卻不考慮各試題的難易度、鑑別度；然而因素分數與總分的概念不同，它是對每題進行加權的結果，這樣的觀念頗似現代測驗中承認各個題目有不同的價值。

3、不論是 Park (2000)、Falenchuk (2006) 的研究可知，當長度增長，幾乎所有的模式所估出的能力值都相似。

參、研究方法

在研究方法上共分成三個部份，第一個部份先討論資料產生方式，第二部份討論資料分析，最後再介紹研究工具，茲分述如下。

一、資料產生方式

本研究資料產生方式共分兩種，一種為概化部份計分模式，一種為因素分析模式，茲將這兩種資料產生方式說明如下。

(一) 概化部份計分模式之資料產生

概化部份計分模式之參數設定方式及潛在特質，採用 Park (2000) 模擬概化部份計分五點量表方式，產生數值，並將每一種類型，重複模擬五十次，其模擬方式分敘如下：

1、潛在特質的參數

依 Sgammato (2009) 的理論，將因素分數與能力值，皆視為潛在特質，而依 Park (2000) 的理論，則認為能力值服從常態分布，為平均數為 0，標準差為 1 之常態分配，因此本研究乃將潛在特質，包含因素分數及能力值，皆視為平均數 0，標準差為 1 之常態分配，而模擬所使用之潛在特質亦依如上方式產生。

2、試題參數

試題參數的設定決定模擬方式，本研究一方面根據 Park (2000) 的方式產生試題參數，一方面則根據研究目的設定參數，茲列敘如下：

- (1) 鑑別度：依 Park (2000) 的方式設定鑑別度，並探討鑑別度對分析果影響，鑑別度為均勻分布，分別設定為 0.6、0.9、1.4。
- (2) 資料合併需求：Park (2000) 認為資料合併需求是指說，當一個難度過度偏向極端，勢必造成選填量表答案的分布不均。例如說，當一個五點量表，其四個難度值分別為 -4、-3、-2、-1，如此潛在特質高於 -1 者，包含潛在特質為 0~3 者皆會選填 5，而只有潛在特質小於 -3 者，選填 1 或 2，因此，1 和 2 便有和併的須需求，亦即 -4 這個難度值應被裁併，而將五點量表變成四點量表。另外，值得注意的是如果選填量表的受試過於極端，例如有很多受試者都選同一個數字（例如都選 1），則會使得概化部份計分不能估計，這是必須注意的。而依 Park (2000) 方式，設定資料合併需求，高資料合併需求，難度介於 -1.25~-0.75；中資料合併需求，難度介於 -0.25~-0.75，低資料合併需求，難度 0.25~-0.25。
- (3) 試題題數：依 Park (2000) 的方式設定，分別為五題和十五題。
- (4) 閾值寬度：依 Park (2000) 的方式設定，閾值的寬度固定為 (-1.5, -0.5, 0.5, 1.5)，亦即將難度加上閾值的寬度，即為閾值，例如難度值為 1，其四個閾值為 (-0.5, 0.5, 1.5, 2.5)。
- (5) 類別個數：本研究主要著重於五點量表，因此，類別個數暫不予考慮。

(三) 隨機參數

隨機參數設為均勻隨機分布，其值介於 0~1 之間，並與累積機率 (cumulative probability) 比較，以產生答案。舉例而言，當一個試題的鑑別度為 0.9，難度為 0.1932，閾值寬為 (-1.5, -0.5, 0.5, 1.5)，當一個能力值為 -0.07，在第一個選項的累積機率為 1，第二個選項的累積機率為 0.961646，第三個選項的累積機率為 0.77403，第四個選項的累積機率為 0.400898，第五個選項的機率為 0.099187，如果跑出的隨機參數大於 0.961646，那麼答案就是 1；介於 0.77403~0.961646，則答案是 2，介於 0.400898~0.77403，答案是 3，介於 0.099187~0.400898，則答案是 4，小於 0.099187，則答案是 5。依此類推。

(二) 因素分析模式產生資料方式

因素分析模式之參數設定方式及潛在特質，基本上有很多種(蔡坤宏, 1996; Falenchuk, 2006)，本研究為了與概化部份計分模式比較，採用 Falenchuk (2006) 模擬因素分析的方式，產生數值，並將每一種類型，重複模擬五十次，其模擬方式分敘如下：

1、模擬原理：

試題參數的設定決定模擬方式，本研究一方面根據 Falenchuk (2006)，但 Falenchuk (2006) 乃根據實際參數得來，而為了擁有比較有基準，亦以 Park (2000) 的方式建立實驗之標準模擬條件，茲將其原理闡述如下。

在因素分析並無機率這個概念，只有殘差的概念，因此，因素負荷量便決定殘差，殘差又決定是否能掉入理論值之答案；其公式如下：

$$y_{ij} = f_{i1}\theta_{1j} + \varepsilon_{ij} \dots\dots\dots (\text{公式七})$$

$$\begin{aligned} w_{ij} &= k, \text{ 假使 } r_{i,k-1} \leq y_{ij} < r_{i,k} \\ \varepsilon_{ij} &\sim N(0, \sigma^2) \dots\dots\dots (\text{公式八}) \\ \varepsilon_{ij} &= z_{ije} \sqrt{1 - f_{i1}^2} \end{aligned}$$

上述的公式 θ_{1j} 是受試者之原始潛在特質， y_{ij} 代表的是受試者表現在試題的潛在特值，兩者之所以有區別乃在於有殘差值 ε_{ij} ， ε_{ij} 之產生乃由一個隨機標準常態分布乘以 $\sqrt{1 - f_{i1}^2}$ 所得；而當 $f_{i1}\theta_{1j} + \varepsilon_{ij}$ 跨過閾值（此即現代測驗的閾值） $r_{i,k-1}$ 而未跨過 $r_{i,k}$ ，則答案落入該選項。

2、潛在特質的參數

潛在特質包含因素分數及能力值，其模擬如同上述，皆視為平均數 0，標準差為 1 之常態分配。

3、試題參數

- (1) 因素負荷量：一般而言，一般所謂的可採用之試題其因素負荷量應在 .7 以上，因此，本研究乃將因素負荷量以 .1 為間隔（蔡坤宏, 1996），分成兩個區段，第一個區段為 .7~.8；第二個區段為 .8~.9。
- (2) 資料合併需求及閾值寬度：為了與概化部份計分產生的資料能有比較之基準，將閾值寬度與資料合併需求之設法皆與前述（由概化部份計分產生之資料）相同，則可產生相同之閾值。
- (3) 試題題數：如同前述，分為五題和十五題。
- (4) 類別個數：如同前述為五點量表。

綜合上述，資料產生類型共有如下：

1、概化部份計分模式產生之資料：

三種鑑別度×三種資料合併需求×二種題數=18 種類型

2、因素分析模式產生之資料：

二種因素負荷量×三種資料合併需求×二種題數=12 種類型

易言之，本研究共可得三十種類型，研究者根據這三十種類型，每類型模擬五十次，模擬完後再用兩種模式進行分析。

二、資料分析

(一) 相關及其差異之檢定

判斷分析出來的潛在特質相關的方法有兩種，Park (2000) 採用簡單相關進行分析，Sgammato (2009) 則採用 tau-b 進行分析，原則上，tau-b 求的是其排序的一致性，但是簡單相關求的是標準化分數一致性，由於模擬的潛在特質、因素分數，與概化部份計分模式分析出來的潛在特質皆為平均數 0，標準差 1，因此更本研究採用簡單相關分析。

此外，為了瞭解簡單相關的差異是否達到顯著，本研究乃建立相關轉換為 Fish z 之檢定表，如表一。

表一 樣本數 1000 之 Fish z 之差之重要臨界值

Fish z 之差之臨界值	機率值
0.087786	p<.05 (1.96 個標準差)
0.073453	P<.1 (1.64 個標準差)

(二) RMSD 及其差異檢定

本研究亦欲分析由五點量表分析出來的潛在特質與原始模擬值的差距，而分析的方法有三種，一種為 Bias，第二種為平均平方差，第三種為 RMSD (平均平方差之根)，而事實上，求平均平方差與求平均平方差之根的意義是一樣的，只是一個有開根號一個沒開根號，因此三種分析方式實可歸類為 Bias 和 RMSD。茲將此二種之公式列之如下：

$$\text{Bias} = \frac{\sum_{j=1}^J \bar{\theta}_j - \theta_j}{J} \dots\dots\dots (\text{公式九})$$

$$\text{RMSD} = \sqrt{\frac{\sum_{j=1}^J (\bar{\theta}_j - \theta_j)^2}{J}} \dots\dots\dots (\text{公式十})$$

從以上的公式可得知，此公式的意義是將估算出的原始值減去原始的模擬值，再加以平方，而 Bias 和 RMSD 的最重大差異在於 Bias 的偏差基本上是可以互相抵消的，但是 RMSD 的偏差則不可以互相抵消，當選擇使用 Bias，其意義較為不明，因此，本研究選用 RMSD。

此外，由於 RMSD 非標準常態分配，因此乃將 RMSD 轉換成 Log (RMSD)，使之成為常態分配 (Harwell et al., 1996)。

另外，為了瞭解 RMSD 的差異是否有達到顯著水準，研究者模擬兩個隨機

之 Log (RMSD) 之差異，得到如下的檢定值：

表二 樣本數 1000 之 Log (RMSD) 之差之重要臨界值

Log (RMSD) 之差之臨界值	機率值
0.025893	p<.05 (1.96 個標準差)
0.02166	P<.1 (1.64 個標準差)

(三) 變異數分析

在 Park (2000) 的研究中，運用變異數分析的方法瞭解何項因子的影響。不過，由於他的研究涉及類別個數的改變，並以原來的參數進行資料合併，因此，他將類別個數與鑑別度、資料合併需求交互作用設為因子內變異，而本研究並沒有探討類別個數的改變，亦無進行資料合併，因此，本研究將所有的鑑別度、資料合併需求、模式、長度皆設為組間變異，並以之進行變異數分析。

三、研究工具

為了產生模擬資料及分析資料，本研究共使用 R 軟體、EXCEL、SPSS18、PARSCALE，其運用方式分述如下：

(一) R 軟體

- 1、運用 R 軟體模擬 RMSD 之差之重要臨界值。
- 2、運用 R 軟體模擬潛在特質之常態分布。

(二) EXCEL

- 1、運用 EXCEL 模擬概化部份計分、因素分析模式在五點量表之得分。
- 2、計算潛在特質之 RMSD、簡單相關及其轉換為常態分配之值。
- 3、求出各 RMSD、簡單相關之平均數。

(三) SPSS18

- 1、運用 SPSS18 求出模擬產生的五點量表之因素分數。
- 2、對轉換成常態分配的相關、RMSD，進行變異數分析，以瞭解各因子之影響力。

(四) PARSCALE

- 1、以概化部份計分模式分析模擬所得之五點量表之潛在特質。

肆、結果與討論

本研究主要在探討由概化部份計分和因素分析所產出的資料，分別由概化部份計分，因素分析進行分析比較，結果茲分述如下。

一、概化部份計分產生資料之分析

為了對產生的資料有全面的分析，研究者乃先就整體概況進行分析，其次再進行模式比較，最後再採用變異數分析。

(一) 整體概況分析

研究者以原始模擬值與分析所得潛在特質求其 RMSD 和相關，分別討如下：

1、RMSD

從表三、表四可以發現，當長度、鑑別度增加，RMSD 皆為下降，但是資料合併需求對 RMSD 的影響則視情況而定，在中等鑑別度時，RMSD 隨著資料合併需求降低而降低，且兩模式趨勢一致。

另外，從表中亦可發現，因素分析充滿規律，只要試題長度增加、鑑別度提高、資料合併需求降低，其 RMSD 皆一致的降低，但概化部份計分則無此規律。

但是資料合併需求，不論試題長度為何，不同的模式有不同的狀況，茲分述如下：

- (1) 因素分析模式，其 RMSD 隨著資料合併需求降低而減少，其規律不因鑑別度、長度有所不同。
- (2) 概化部份計分模式
 - (a) 在高鑑別度下，不論試題長度為五題或十五題，中資料合併需求 RMSD 最大，高資料合併需求次之，低資料合併需求 RMSD 最小。
 - (b) 低鑑別度下，試題長度五題，RMSD 隨著資料合併需求降低而降低，但是在試題長度十五題情況下，高資料合併需求的 RMSD 最大，低資料合併需求次之，中資料合併需求 RMSD 最小。

將本研究與 Park (2000) 對概化部份計分模擬所做的研究進行比較，其整體的趨向與本研究相近，鑑別度、長度愈高，其 RMSD 減少，但是在資料合併需求上，其研究指出，資料合併需求愈高，不論鑑別度為何，其 RMSD 皆一致增加，此點與本研究不同。而之所以不同，研究者認為可能是實驗誤差或軟體不同所致。

表三 試題長度五題之 RMSD 之平均

	低鑑別度 0.6		中鑑別度 0.9		高鑑別度 1.4	
	FA	GPCM	FA	GPCM	FA	GPCM
高資料合併需求	0.640059	0.912703	0.536274	0.649113	0.44198	0.423523
中資料合併需求	0.614976	0.806989	0.499931	0.598451	0.40779	0.443974
低資料合併需求	0.607308	0.791691	0.492044	0.577492	0.397471	0.420807

表四 試題長度十五題之 RMSD 之平均

	低鑑別度 0.6		中鑑別度 0.9		高鑑別度 1.4	
	FA	GPCM	FA	GPCM	FA	GPCM
高資料合併需求	0.425534	0.513815	0.359442	0.486932	0.31386	0.308647
中資料合併需求	0.399204	0.388352	0.325473	0.339304	0.268435	0.318818
低資料合併需求	0.287379	0.422084	0.31258	0.310277	0.251666	0.240821

2、相關

相關與 RMSD 的關係互為表裡，通常相關高，RMSD 則低。本研究亦顯示如此的狀況。與 RMSD 剛好相反，當長度、鑑別度增加時，相關即增加，此外，當資料合併需求降低時，相關亦呈現規律的增加。唯一的例外是在長度為十五題低鑑別度時，概化部份計分模式在中資料合併需求相關最高，低資料合併需求次之，高資料合併需求最低。

以 Park (2000) 對概化部份計分模擬所做的研究進行比較，發現本研究的概化部份計分模式之整體趨勢與 Park (2000) 相似，亦即當長度、鑑別度增加時，相關即增加。

表五 試題長度五題相關之平均

	低鑑別度 0.6		中鑑別度 0.9		高鑑別度 1.4	
	FA	GPCM	FA	GPCM	FA	GPCM
高資料合併需求	.794885	.762481	.854793	.833772	.902234	.897573
中資料合併需求	.810654	.793491	.8749	.860926	.916789	.906069
低資料合併需求	.815344	.798088	.878815	.869416	.920952	.915022

表六 試題長度十五題之相關之平均

	低鑑別度 0.6		中鑑別度 0.9		高鑑別度 1.4	
	FA	GPCM	FA	GPCM	FA	GPCM
高資料合併需求	.909382	.900027	.935369	.906628	.95074	.957333
中資料合併需求	.920261	.92144	.947021	.947688	.963986	.966633
低資料合併需求	.965387	.920231	.951139	.952334	.968355	.97024

(二) 模式比較

過去研究對模式比較有二種主要方法，第一種是藉由圖表分析其趨勢之差異 (Falenchuk, 2006)，第二種是運用模擬的方法，判定其差異是否達顯著 (Park, 2000)。本研究乃採用此後一種方法進行比較，並以 $p=.05$ 的水準，以表一、表二的關鍵值，判定顯著與否。

1、因素分析模式與概化部份計分模式在 RMSD 之差異

為了瞭解因素分析模式與部份計分模式的優劣，研究者以模擬所得的表三、表四、表五、表六，比較在每一個情況下的差異，歸納如下：

(1) 因素分析模式顯著優於概化部份計分模式之情境：

- a、試題長度五題的情況，所有的低鑑別度、中鑑別度，因素分析模式皆顯著優於概化部份計分模式。
- b、試題長度十五題的情況，則僅有高鑑別度中資料合併需求；中鑑別度高資料合併需求；低鑑別度的高資料合併需求、低資料合併需求，因素分析模式優於部份計分模式。

(2) 兩模式無顯著差異之情境：

- a、試題長度五題的情況下，高鑑別度上，因素分析與概化部份計分模式無顯著差異。
- b、試題長度十五題的情況下，高鑑別度的高資料合併需求、低資料合併需求；中鑑別度中資料合併需求、低資料合併需求；低鑑別度中資料合併需求，兩模式皆無顯著差異。

2、因素分析模式與概化部份計分模式在相關之差異

(1) 因素分析顯著優於概化部份計分之情境：

- a、試題長度十五題下中鑑別度中資料合併需求；低鑑別度低資料合併需求。

(2) 兩模式無顯著差異之情境：

- a、試題長度五題的情況下，兩模式無顯著差異。
- b、試題長度十五題的情況下，低鑑別度高資料合併需求、中資料合併需求；中鑑別度。中資料合併需求、低資料合併需求；及所有高鑑別度皆無顯著差異。

(三) 變異數分析

依循 Park (2000) 的變異數分析方法，先將相關與 RMSD 轉換成常態再進行分析，結果如下。

以 RMSD 進行四因子變異數分析，結果顯示模式、長度、鑑別度、資料合

併需求，皆達顯著，所有交互作用亦皆達顯著。不過除了考慮是否達顯著外，亦須考慮其效果值，結果顯示，長度的效果值為.911，鑑別度的效果值為.850，資料合併需求的效果值為.453，模式的效果值為.428，其大小分別為長度>鑑別度>資料合併需求>模式，這無疑說明了長度最能影響 RMSD，亦即估計的精準性。

再進行相關的分析，結果亦如同 RMSD 的分析，所有交互作用均達顯著，模式、長度、鑑別度、資料合併需求，亦達顯著，其效果量之大小分別為長度.974、鑑別度.943、資料合併需求.709、模式.186。不過值得注意的是，以相關作為依變項，模式的效果量遠不如以 RMSD 作為依變項，這說明，當如果僅只要求相關，而不是要求出正確的潛在特質，選擇何種模式顯然並不是那麼重要，此外，資料合併需求的相關之效果值相較於以 RMSD 進行的變異數分析的效果值大很多，這顯示當求相關時，資料合併需求有極大的影響。

二、因素分析產生資料之分析

本部份分析方法比照前述的方式，先進行整體之分析、再進行模式比較及變異數分析。

（一）整體概況分析

研究者分別就 RMSD 與相關所得之結果分析如下。

1、RMSD

從表七、表八可以發現在控制其他條件下，只要當因素負荷量提高，其 RMSD 必然降低。然而其他條件則沒有顯現如此強烈的必然性。不過如同以概化部份計分產生的資料，在以因素分析方法產生的資料中仍然可以發現，因素分析一如其原理-線性組合，其結果充滿著規律，當資料合併需求減低、試題長增加，因素分析之 RMSD 皆下降，但是概化部份計分則無此規律。

在概化部份計分方面，其規律性即不如因素分析。研究者發現隨著試題增加、資料合併需求減少、概化部份計分估計之 RMSD 並沒有一致的規律；只有因素負荷量的增加，必然有助於 RMSD 的減少。

表七 試題長度五題之 RMSD 之平均

	因素負荷量 0.7~0.8		因素負荷量 0.8~0.9	
	FA	GPCM	FA	GPCM
高資料合併需求	0.462964	0.459371	0.345361	0.269565
中資料合併需求	0.442456	0.540003	0.299386	0.3039
低資料合併需求	0.433485	0.465995	0.284674	0.275246

表八 試題長度十五題之 RMSD 之平均

	因素負荷量 0.7~0.8		因素負荷量 0.8~0.9	
	FA	GPCM	FA	GPCM
高資料合併需求	0.285182	0.529182	0.261713	0.260633
中資料合併需求	0.248699	0.382826	0.201542	0.268604
低資料合併需求	0.235809	0.235274	0.187154	0.186162

2、相關

從表九、表十可以發現在其他條件控制下，資料合併需求減少，因素負荷量增加，其相關皆會增加；但是長度增加，其規律性則依模式之不同而不同，在因

素分析模式下，隨著長度增加，相關亦增加，但是概化部份計分模式，只有在高因素負荷量下，其相關才會隨著長度增加而增加；從這裡又再度發現，因素分析充滿著規律。

表九 試題長度五題之相關之平均

	因素負荷量 0.7~0.8		因素負荷量 0.8~0.9	
	FA	GPCM	FA	GPCM
高資料合併需求	0.892718	0.883427	0.940337	0.945504
中資料合併需求	0.902003	0.888958	0.955186	0.950803
低資料合併需求	0.905948	0.90231	0.959489	0.956403

表十 試題長度十五題之相關之平均

	因素負荷量 0.7~0.8		因素負荷量 0.8~0.9	
	FA	GPCM	FA	GPCM
高資料合併需求	0.959344	0.742347	0.965772	0.96067
中資料合併需求	0.969099	0.967639	0.979732	0.982307
低資料合併需求	0.972227	0.97203	0.982533	0.982777

(二) 模式比較

在模式比較上，研究者仍承襲前述的方法，並以表一、表二的關鍵值作為比較的基準，茲分述如下。

1、因素分析模式與概化部份計分模式在 RMSD 之差異

以表一、表二的關鍵值進行判準，共可得三種結果。

(1) 因素分析模式顯著優於概化部份計分模式之情境

a、試題長度五題的情況：「因素負荷量 0.7~0.8 下，中、低資料合併需求」，因素分析模式皆優於概化部份計分模式。

b、試題長度十五題的情況：在「因素負荷量 0.7~0.8 下，高、中資料合併需求」，及「因素負荷量 0.8~0.9 下，中資料合併需求」因素分析模優於概化部份計分模式。

(2) 概化計分模式顯著優於概化部份計分模式之情境

a、試題長度五題的情況：在「因素負荷量 0.8~0.9 下，高資料合併需求」概化部份計分模式優於因素分析模式。

(3) 兩模式無顯著差異之情境：

a、試題長度五題的情況：「因素負荷量 0.7~0.8、高資料合併需求」及「因素負荷量 0.8~0.9，中、低資料合併需求」，因素分析模式與概化部份計分模式無顯著差異。

b、試題長度十五題的情況：「因素負荷量 0.8~0.9、高資料合併需求」及所有低資料合併需求，因素分析模式與概化部份計分模式皆無顯著差異。

2、因素分析模式與概化部份計分模式在相關之差異

(1) 因素分析顯著優於概化部份計分模式之情境：

在「十五題、因素負荷量 0.7~0.8、高資料合併需求」下，因素分析模式顯著優於概化部份計分模式。

(2) 兩模式無顯著差異之情境

除了上述的情況，其餘兩模式皆無顯著差異。

(三) 變異數分析

變異數分析亦可分成兩個部分，一個部份以 RMSD 進行變異數分析，另一個部份則以相關進行變異數分析。

以 RMSD 進行四因子變異數分析，結果顯示模式、長度、因素負荷量、資料合併需求，皆達顯著，而交互作用「模式、因素負荷量」、「模式、資料合併需求」、「模式、長度」、「因素負荷量、長度」、「因素負荷量、資料合併需求」、「長度、資料合併需求」、「模式、因素負荷量、資料合併需求」、「模式、因素負荷量、長度」、「模式、資料合併需求、長度」皆達顯著。再考慮效果值的大小。結果顯示，因素負荷量的效果值為.661，模式的效果值為.023，長度的效果值為.752，資料合併需求的效果值為.257，其大小分別為長度>因素負荷量>資料合併需求>模式，這說明了長度最能影響 RMSD，即估計的精準性。

再進行相關的分析，結果顯示交互作用、模式、長度、鑑別度、資料合併需求，皆達顯著，其效果量之大小分別為「長度 (.978)」、「因素負荷量 (.941)」、「資料合併需求 (.747)」、「模式 (.019)」。

從上述的數字再度得知長度是影響相關估計最重要之因子；再者，模式在相關的影響力相較於 RMSD，明顯下降很多，顯示若只求相關，模式重要性不高。最後，資料合併需求的相關之效果值相較於以 RMSD 進行的變異數分析的效果值大很多，這表示當求相關時，資料合併需求有極大的影響。

伍、結論與建議

根據上述模擬研究，研究者提出以下的結論與建議，茲分述如下。

一、結論

本研究分析之資料主要來自兩種模式所產生，一種為概化部份計分，另一種為因素分析；以是之故，在作結論，勢必先對這兩種資料分別下結論，最後再就兩種資料之共通處及資料產生之模式進行反思，茲分成三個部份敘述如下。

(一) 由概化部份計分產生之資料

1、在控制相同鑑別度、資料合併需求、閾值寬度，題數十五題以下之測驗，概化部份計分模式無法優於因素分析模式。

從上述的研究可以發現，現代測驗中的概化部份計分模式並沒有測量不變性的特質，亦未全然優於傳統的因素分析。

回顧能力測驗，現代測驗之所以能取代古典測驗，乃在於古典測驗以總分當作能力值，而現代測驗則是將常態分配進行黎曼積分得到試題特徵曲線函數，進而由多個試題特徵曲線函數獲得能力參數 (Masters, 1982)。如此，自然超越了以總分當能力值的古典測驗。然而這種優勢在五點量表的因素分析中不復存在，因素分析所得的潛在特質其概念雖然類似能力值，但是確非總分的概念，而是有加權的概念。事實上，因素分數係數即是承認題題不等值，這與現代測驗中當鑑別度不同，即題題不等值概念是類似的，此與將總分當能力值是不一樣的。

再者，在能力測驗中要估計的試題的參數，有鑑別度、猜測度、難度值，但是在概化部份計分中，要估計試題的參數則有鑑別度、四個難度值，其估計數足足多了二個，然而估計參數在電腦的運算中並不是以解微分方程的方式解出，而是先假設能力值，讓試題參數收斂至某個值而解出 (林玠毅, 2008)，因此，並非試題參數愈多，估計就愈準確，相反的，試題參數愈多，反而可能使試題參數估計不準確，而導致能力值估計偏差。

第三，本研究亦發現在題數較少（5 題），且中低鑑別度的情況下，透過概化部份計分模式所產生的模擬資料答案，經由因素分析，亦產生較少的 RMSD 及高相關，由此可知，因素分析在中低鑑別度，且題數較少的情況中較概化部份計分為優。

然而，以上的研究仍無法證明當資料是概化部份計分產生下，因素分析優於概化部份計分，原因是本研究是設定在控制相同的鑑別度、資料合併需求下，然而，就一般的測驗而言，其鑑別度、資料合併需求不可能如此的均勻不變，一定是混合式的，包含不同類型鑑別度、資料合併需求，這和嚴格控制這些要素的模擬測驗，其結果不見得相同。這也是為何 Falenchuk（2006）與 Park（2000）對等級反應和概化部份計分模式比較的研究有所不同。事實上，因素分析中，當大多數題目相關高，那麼因素分數係數必然給那些佔測驗大多數題目較高的加權（例如：當 1~12 題彼此高相關，13~15 彼此為高相關，因素分數係數會給 1~12 題較高的加權），當這種加權是合理的，那麼其估計自然準確。在本研究中，由於控制了鑑別度、資料合併需求，大部份的因素分數係數的值都相當接近，這種加權方式自然合理，但是當題目是混合不同的鑑別度、資料合併需求時，若對少數高鑑別度卻與其他題目有較低的相關，而給予較低的加權，如此的加權即為不合理，則其估計精準度是否能夠勝於概化部份計分則是有疑義的。

綜上所述，因素分析並非傳統的總分概念，概化部份計分亦未能如現代測驗理論在能力測驗中發揮所長，在嚴格控制鑑別度、資料合併需求，十五題下以之測驗，概化部份計分確實不會優於因素分析，甚至在五題且中低鑑別度的短測驗中，因素分析顯著優於概化部份計分，然而，在實際混合各類型鑑別度、資料合併需求、甚至不同寬度的閾值下，其估計精準度，因素分析是否仍優於概化部份計分，則充滿疑義，有待後來的學者繼續研究與闡釋。

2、概化部份計分模式在更長的測驗且更高鑑別度下有勝過因素分析的趨勢

雖然本研究顯示概化部份計分模式在任何嚴格控制鑑別度、資料合併需求下，其估計精準度皆無法顯著優於因素分析。然而，從整體的趨勢仍然可以發現，概化部份計分模式在高鑑別度且較長的測驗長度下，有逐漸勝過因素分析的趨勢。因此，研究者推測當鑑別度更高且測驗長度更長的情況下，概化部份計分模式其對潛在特質的估計精準度有機會勝過因素分析。而這樣的推測，大致上可以從資料的產生方式看出端倪。

以例子來說，當其四個閾值分別為 -1.5、-0.5、0.5、1.5，受試者潛在特質為 0.2 時，理論上，受試者應該選 3，但是當鑑別度有所不同時，該生選 3 的機率有所不同。以鑑別度 0.6 為例，該生選答 1~5 的累積機率分別為（1、0.902337、0.662126、0.337874、0.0977），易言之，受試者選 1~5 的機率分別為（0.097663、0.240211、0.324252、0.2402、0.0977）；而當鑑別度提高至 1.4 時，其累積機率分別為（1、0.971245、0.73643、0.26357、0.0288），其選 1~5 的機率分別為（0.028755、0.234815、0.47286、0.2348、0.0288）。由此可知，當鑑別度提高時，其掉入理論上答案的機率亦增加，這自然使得估計準確度提昇。因此，當鑑別度高時，自然大大有助於概化部份計分的估計。而這種情況，在因素分析較為不明顯，而概化部份計分卻大幅提高（這個問題將在下面說明）。

另外，再就長度來看，誠如 Falenchuk（2006）的研究指出，當測驗長度變長時，不論是概化部份計分模式、等級反應模式、概化等級展開模式其結果差異均不大。這樣的結論自然不能推理到因素分析與概化部份計分模式的比較，然而，從本研究的模擬言，當長度增加時，任何模式的估計精準度皆會增加，其彼此的

差異自然會縮小（因為估計精準度有其極限），因此在測驗長度增加下，任何模式差異逐漸縮小，而倘若再搭配高鑑別度，則概化部份計分模式，確有勝過因素分析之可能性。

綜上所述，研究者推測縱然題目均嚴格控制均勻鑑別度、資料合併需求，只要鑑別度夠大（可能需要達到2）、長度夠長（超過二十題），概化部份計分模式亦將勝過因素分析。

（二）由因素分析產生之資料

1、就相關言，當控制因素負荷量在.7 以上，穩定的資料合併需求、閾值寬度，題數十五題以下之測驗，概化部份計分模式無法優於因素分析模式。

承如前述，概化部份計分雖然承襲自現代測驗理論，但現代測驗理論的優勢並沒有如期的在五點量表發揮；因素分析亦優於古典測驗「總分」的概念；而透過本次的模擬，研究者發現在十五題以下，只要嚴格控制其因素負荷量、資料合併需求、閾值寬度，就相關而言，因素分析估計模式不劣於概化部份計分。

然而亦如前述的問題，這些題目都是實驗室產生的題目，在混合各種因素負荷量、資料合併需求是否能得到一致的答案，仍有待商榷，需往後的研究者多方面蒐集各種實際數據，進行模擬分析，才能確定模式優劣。

2、同時提高長度與因素負荷量，概化部份計分有勝過因素分析的趨勢

本研究雖然指出在相關上，概化部份計分無法優於因素分析，但是從整體趨勢來看，只要同時提高到足夠的長度、因素負荷量，概化部份計分仍可能會勝過因素分析。

這個結論亦可從資料產生分析得到結論。在資料產生中，因素負荷量是決定殘差之重要因子，因素負荷量愈高，殘差便愈少，其落在應該落的位置之機率便愈高；事實上，因素負荷量其功用可等同於現代測驗的鑑別度（這在下一個部份將詳細說明），在現代測驗理論中，當鑑別度愈大，便會促成掉入理論答案機率大增，而因素負荷量則在古典測驗扮演同樣的角色；因此，當因素負荷量提高，殘差便減少，答案便落在理論的位置，部份計分法便能更精準的估計。

當然，同樣的，當因素負荷量升高，因素分析亦較為準確，然而從模擬的結果來看，準確率之提高，不若概化部份計分模式（這在下一個部份將詳細說明）。

另外，就長度言，由於 Falenchuk（2006）的研究可得知，只要長度變長，任何模式的差異將變小，本研究亦指出類似的趨勢（除了長度十五題、低資料合併需求、因素負荷量.7 其相關較長度五題低資料合併需求、因素負荷量.7 為低）。因此，只要提高長度及因素負荷量，則概化部份計分模式將可能勝過因素分析模式。

（三）不同資料產生方式之反思及總結

1、因素負荷量是古典測驗中的鑑別度；殘差即非理論答案的期望值

回顧整個現代理論，現代測驗可以說是依照所謂的強勢假設而來的，即是認為落入正確答案的機率是常態分布的累積機率，而概化部份計分，基本上亦運用此原理，只不過每一個選項皆有其類別反應曲線，但其曲線來自「常態分布機率」這個事實自始自終皆存在（余民寧，2009）；但是因素分析的源頭-迴歸的概念則是假設資料呈現常態分配（黃俊英，2007），且估計標準誤亦呈現等分散性（林清山，1997），因此在因素分析中強調是估計標準誤的等分散性，易言之，其殘差必呈現等分散性。這兩者之常態分配的觀點是不一樣的；在概化部份計分中，所謂的「殘差」即是落入其他選項的機率（在現代測驗中原則上並不使用迴歸之殘差的概念，所以研究者乃加上括弧），因此當因素負荷量愈高，則相對的，所

有受試者落入理論答案機率便提高，落入其他其選項的機率便愈低，其「殘差」便愈小，此外，概化部份計分亦認為每一個人的「殘差」不是相等的期望值，但是相對的，迴歸估計的意義則是認為每個受試者擁有相同的估計標準誤，其殘差之大小受估計標準誤分配之影響。綜合上述，殘差可視為現代測驗中非理論答案的期望值，但是其背後並無「估計標準誤等分散性」在操縱，而是受各題項機率之操縱。

鑑別度在現代測驗中具有加權的效果，這一點亦如同因素負荷量在產生因素分數產生上具有加權的效果（蔡坤宏，1996）；另外，本模擬研究亦指出提高鑑別度亦可提高選擇該理值之機率及因素負荷量，同樣的提高因素負荷量亦可提高鑑別度及選擇該理論值之機率；因此，因素負荷量可以視為古典測驗中的鑑別度。

舉一個例子來說（以概化部份計分產生資料），假設一個受試者之潛在特質為 0.565028，門檻值分別為（-1.59171、-0.59171、0.408286、1.408286），在鑑別度 0.6 下其選擇 1~5 的機率分別為（0.040311、0.147035、0.294334、0.323358、0.194962），其選擇非理想答案的期望值為 0.904299；而在鑑別度 1.4 下其選擇 1~5 的機率分別為（0.00341、0.069832、0.352669、0.439206、0.134883），其選擇非理想答之期望值為 0.637445（可視為殘差）。而當受試者潛在特質為 -1.70784，在鑑別度為 0.6，其機率值為（0.389875、0.363636、0.186137、0.05229），其選擇非理想答案的期望值為 0.925027；而在鑑別度 1.4 下，其選擇 1~5 的機率分別為（0.00341、0.069832、0.352669、0.439206、0.134883），其選擇非理想答之期望值為 0.605834。但是在因素分析下，在鑑別度 0.6 下，因素負荷量為 0.687，各題目間的相關介於 0.26~0.365 間，不管潛在特質是 0.565028 或者是 -1.70784，其二個標準誤之殘差期望值皆為 0.501323；在鑑別度 1.4 下，題目間的相關介於 0.524~0.589 間，因素負荷量為 0.809，其二個標準誤之殘差期望值皆為 0.40553；從上述可發現三件事，第一、鑑別度的提升對概化部份計分估計準確度大於因素分析模式；此外，亦可發現對於較極端的能力值，鑑別度的提升對於概化部份之估計更是重要，更能有效的估出正確值，第二，因素負荷量與鑑別度是連動的，當鑑別度提昇，題目之間的相關自然提升，而因素負荷量亦隨之提升，惟受因素負荷量影響之估計標準誤則是等分散性的。第三，同理，當因素負荷量提升時，鑑別度自然提升，概化部分計分模式的精準性自然提升，且提昇之效果較因素分析佳。

2、提升鑑別度（或因素負荷量）與長度有助於概化部份計勝過因素分析模式

如同前述，鑑別度或因素負荷量提升對概化部份計分模式的提升大於因素分析。而一般而言，三參數之現代測驗必須在二十題以上才會穩定（余民寧，2009），而邏輯上，概化部份計分亦須二十題以上才會穩定，因此當長度夠長（超過二十題），鑑別度提高（可能要非常接近 2），從模擬的結果及理論推測，概化部分計分可能會勝因素分析模式。

3、因素分析分析各種模式產生之資料之結果相當規律

從模擬的結果來看，不論是何種方法產生之資料，由因素分析所得之結果皆相當規律。以概化部份計分產生的資料，只要長度增加，資料合併需求降低，鑑別度提升，因素分析所得的結果其相關皆提升，RMSD 皆下降；同樣的，以因素分析產生的資料，只要長度增加，資料合併需求降低，因素負荷量增加，因素分析所得的結果其相關皆提升，RMSD 皆下降，由此可知，因素分析所得之結果具有規律性，這種規律性是概化部份計分所沒有的，研究者推測此可能與因素分析

是線性模式有關。

4、當只求相關時，嚴格控制的實驗情境，概化部份計分法不會優於因素分析法

綜合兩種資料產生模式，研究者發現在十五題以下的實驗情境中，若只求與原始能力值之相關，概化部份不會優於因素分析模式，且如結果與討論中所陳述，因素分析模式在許多情境下顯著優於概化部份計分模式；更進一步言，概化部份計分只有在在一項 RMSD 的比較上勝過因素分析模式（因素分析模式產生之資料，五題、高資料合併需求、高因素負荷量），而若結合 Park（2000）所得之結論，則可以獲知，在十五題以下之嚴格實驗情境下，若只求相關，等級反應法不會優於概化部份計分法，概化部份計分法不會優於因素分析。

5、資料可能是透過概化部份計分模式產生

從模擬實驗來看，資料是由何種模式產生的，研究者尚不敢斷言。然而從模擬的資料來看，概化部份計分只有在「由因素分析產生之資料，其試題長度五題，因素負荷量 0.8~0.9 下」之 RMSD，才會勝過因素分析，依此邏輯，因素分析模式確實比較可能是資料產生之方式，然而，當鑑別度、長度、因素負荷量增加、資料合併需求降低，易言之，當整份試題愈優質、詳細（長度夠長），概化部份計分卻逐漸顯出其優點；因此，研究者推測當潛在特質其反應在試題作答上，很可能是如現代測驗所描述的狀況；但是當試題不佳又不夠詳細時，這種強勢假設便無展現，因此只有在優良且詳細的試題下，概化部份計分模式才能顯現優勢。

6、長度是影響估計最重要的因子，其次是鑑別度（或因素負荷量）

就模擬的資料而言，不論是由因素分析或者是概化部份計分產生之資料，要求得精確之相關或較小之 RMSD；透過變異數分析，皆會發現，效果值是長度>鑑別度（或因素負荷量）>資料合併需求>模式；這如同前述說明，長度和鑑別度是決定一個試題優量且詳細的重要關鍵，因此，從「務實」的角度來看，要獲得精確的估計，提升長度即可獲得甚佳之效果，也是最簡捷易行的。

二、建議

1、以實證資料之參數進行模擬與分析，瞭解不同情況下，不同模式的精準性

誠如上述，在嚴格控制鑑別度、資料合併需求下，十五題以下的測驗，若只求與原始模擬值的相關，概化部分計分模式確實不會優於因素分析，而大部份的測驗，尤其是結構方程式，測量一個潛在特質的題目多半不會超過十五題。就這一個觀點來看，將因素分析用於獲得潛在特質似乎有助於估計的精準。然而，大部份的測驗，其鑑別度、資料合併需求、閾值寬度，都不是嚴格控制的，都是題題不一樣，不均勻的，因此，往後的研究者可針對實證所得的參數繼續研究，測試是否因素分析優於概化部份計分模式。事實上，研究者認為，欲圖精準的測量潛在特質，或許不是找尋何種模式俱有普遍的精準性，而是去瞭解每一個模式，其侷限性與優越性。如同本研究的模擬所指出，除非是份優良且詳細的試題，具高鑑別度與一定的長度，在探測相關上，傳統因素分析是優於現代測驗中的概化部份計分模式。此外亦如 Falenchuk（2006）給予的研究啟示，他採用實際估算的參數，再運用這些參數進行模擬，產生不同的資料，之後用各種不同的模式進行分析，以驗證其強韌性，進行判準。易言之，當要追尋一個測驗何者是最精準的估量時，或許問題並不是哪一個模式是通式，而是當其用該模式估算出來的參數，再以此參數代入各種模式進行模擬時，是否仍具有估計的精準，這才是重點。舉例來說，當一個測驗，運用概化部份計分模式估計出鑑別度、四個難度值，那麼為了瞭解本模式是否精準估計，可再以這些參數，產生模擬資料，並運用不同的模式，如等級反應模式、因素分析、概化等級展開等估算其潛在特質，如果以

概化部份計分模式所估計出的潛在特質之相關、RMSD 的精確，不若以其他模式獲得之估計值精準，那麼便有合理的理由懷疑，概化部份計分模式不是該測驗測量潛在特質的最佳模式，其他的模式可能更適合，於是可以用同樣的方法對其他的模式進行測試，如此，或許可以找到最佳的測量方法。當然，這種做法相當不經濟的，但研究者認為這可能是求得最佳測量的可行之道。

2、進行多向度的現代測驗理論與因素分析的比較研究

本研究的重心仍在於單向度，然而因素分析的主要功用，並非求出單向度的因素分數，而是能依據文獻理論，求出多向度的潛在特質。因此，將因素分析與概化部份計分模式進行比較，仍有所不足。研究者認為後繼的學者，應致力發展多向度多元計分的理論與應用程式，使之能與因素分析作更深入的比較。

參考文獻

中文部份

- 余民寧 (2006)。潛變項模式。台北。高等教育。
- 余民寧 (2009)。試題反應理論 IRT 及其應用。台北：心理。
- 余民寧 (2002)。教育測驗與評量：成就測驗與教學評量。台北：心理。
- 林玠毅 (2008)。BILOG-MG 與 NOHARM 在二向度測驗試題參數估計之比較。國立彰化師範大學，彰化縣。
- 林清山 (1997)。心理與教育統計學。台北：東華。
- 陳順宇 (2007)。結構方程模式。台北：心理。
- 陳雁芳 (2006)。等級反應模式下等化效果之探究 (未出版之碩士論文)。國立台中教育大學，台中市。
- 蔡坤宏 (1996)。因素分析產生方式對線性迴歸模式係數估計影響之比較。測驗年刊，43 輯，373-384。
- 謝佳穎 (2009)。多向度試題反應理論用於次級量尺分數估計之模擬研究 (未出版之碩士文)。國立台中教育大學，台中市。

外文部份

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Dickenson, T. S. (2005). *Comparison of various ability estimates to the composite ability best measured by the total test score* (Unpublished doctoral dissertation). University of South Carolina, Columbia, Carolina
- Falenchuk, O. (2006). *A study of unidimensional IRT models for item scored in multiple ordered response categories* (doctoral dissertation, Department of Curriculum, Teaching and Learning Ontario Institution for Studies in education of the University of Toronto, Toronto, Canada). Retrieved from <http://pqdd.sinica.edu.tw/>
- Harwell, M., Stone, C., Hsu, T. & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101-125.
- Park, H. (2000). *Comparison of irt models for ordered polytomous response data*. (Unpublished doctoral dissertation). University of Minnesota, Twin Cities, Minnesota.
- Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas, Austin, Texas.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174
- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17, 351-363
- Rasch, G. (1980). *Probability models for some intelligence and attainment tests*.

Chicago: The University of Chicago Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychological Monograph*, No17.

Sgammato, A. N. (2009). *An application of unfolding and cumulative response theory models for noncognitive scaling: Examining the assumptions and applicability of the generalized graded unfolding model* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill, North Carolina.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39-55.

Yu, M. N. (1991). *A two-parameter partial credit model* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign, Illinois.