

二度空間視覺化測驗之試題 產生算則的驗證與修正*

劉子鍵**

林世華***

梁仁楷****

(收稿日期：1997年12月15日，接受登刊日期：1998年2月20日)

本研究以「二度空間視覺化能力」為特定的研究領域，根據林世華、劉子鍵（民86）所提出的認知測量整合模式，以及林世華、劉子鍵和梁仁楷（民86）的研究結果，進一步地探討下列問題：（一）林世華、劉子鍵和梁仁楷（民86）之研究所採用的受試皆為師大學生，使得作答反應的變異性不大，有可能因此影響估計模式的配適度。本研究進一步地加入國中二年級至高中二年級的學生作為研究樣本，使得樣本的涵蓋面擴及國中程度至大學程度，以蒐集更具代表性的反應組型，並藉此驗證模式的配適程度。（二）經由新樣本的加入，增加可能的反應組型，藉此來觀察 LLTM 估計結果中各成分之相對重要性的穩定程度。

研究結果有數項發現異於前次研究結果。其中，針對篩選後的38題試題就整體樣本（包含大學生樣本及高國中生樣本）的資料進行 Rasch 模式考驗，結果發現整體樣本並不符合 Rasch 模式，而是符合3參數模式。此項結果與先前僅以大學生為樣本的分析結果（林世華、劉子鍵和梁仁楷，民86）大有出入。進一步針對篩選後的38題試題就高國中生樣本進行研究，卻發現該樣本資料符合 Rasch 模式。此一結果顯示大學生與高國中生此二組樣本在組內有相當的一致性，但在組間卻有差異性。本研究進一步針對篩選後的9題，對兩組樣本進行 LLTM 分析，希望瞭解兩組在各個成分上的表現是否亦有所不同。結果發現兩組樣本在基本參數之加權值的排序上有明顯的差異。此一差異將造成兩組建構出不同意涵的試題產生算則。最後，本研究基於研究結果與發現進行討論與建議。

關鍵詞：認知整合測量模式、空間能力、線性洛基斯蒂克測驗模式

壹、研究背景

長久以來測驗專家與心理學家在各自的領域中鑽研，雖都能開創出一片天地，但也因此

* 本文乃根據「第三屆兩岸心理與教育測驗學術研討會」中所發表的論文加以修改而成。在該研討會中，承蒙王振世教授與丁振豐教授的指正，在此致謝。

** 國立華僑實驗高級中學

*** 台灣師範大學教育心理與輔導系

**** 中央大學資訊工程研究所博士班研究生

各自侷促於一隅。Cronbach (1957) 就曾指出心理學的領域中包含兩種不同的科學訓練—相關心理學 (correlational psychology) 與實驗心理學 (experimental psychology)。其中，在研究心理特質或能力時，相關心理學所採取的是心理計量研究取向：建構理論、提出假設、設計工具（例：問卷、量表或是測驗等）、收集資料、分析資料（例：因素分析）、驗證假設、支持或修改理論。在此，相關心理學所關心的是受試者在試題變項上的個別差異是否有系統的變化，是否能因此萃取出有意義的潛在因素 (latent factor)。至於情境變異 (situational variance)，相關心理學將其視為誤差項。此一取向的心理學家致力於發現個別差異的組型，雖然有所貢獻，但也產生下列缺點：1. 從以抽象的心理學理論構念為起點，到以抽象之個別差異的組型為終點（例：探索性因素分析中因素的命名），使所測量之心理特質或能力的意義曖昧不明。2. 以「測驗」(test) 來表徵構念，以組型來驗證構念（例：因素分析中因素負荷量的大小）；至於「試題」(item) 的效度則以測驗的總分（或分量表的總分）作為「內在效標」，以分析個別試題得分與總分的相關。此種做法忽略了各個試題的意義，也忽略了受試者在答題時的心理歷程。

另一方面，實驗心理學的研究取向則是：建構理論、提出假設、設計實驗情境、操弄變項（不同的實驗處理）、收集資料（例：反應的時間等）、分析資料（例：變異數分析）、驗證假設、支持或修改理論等。實驗心理學者信守「最大最小控制原則」，假設「總變異量」是由「實驗變異量」、「無關變異量」(extraneous variance) 及「誤差變異量」等三個部份所組成。在實驗的過程中實驗者希望不同實驗處理之間的變異最大，控制無關變異量，並使誤差變異量最小。個別差異在實驗心理學的取向中是屬於無關變項，應予以控制（例：利用隨機抽樣等方法），無法控制者則歸為抽樣誤差。此一取向的優點在於將心理特質及能力具體化，並使解題歷程透明化。然而，實驗心理學的結果對測驗卻少有直接的貢獻，其理由是：1. 心理測驗的主要目的之一是測量某一重要構念上的個別差異，而非試題間的差異。實驗心理學重視實驗變異量，而忽略了個別差異。2. 實驗情境過於人工化且複雜，使外在效度降低且無法大量實施。

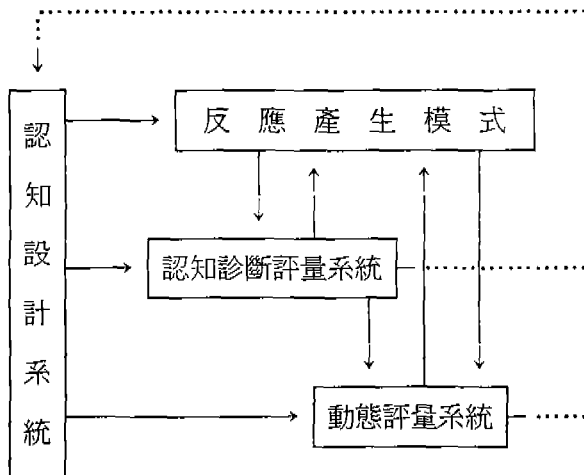
一般而言，測驗編製者所接受的是相關心理學的訓練，而非實驗心理學的訓練。因此，傳統測驗編製過程中建立建構效度的方法多為相關研究 (correlation studies)、內部一致性分析 (internal consistency)、因素分析 (factor analysis) 多特質多方法分析 (multitrait-multimethod analysis; MTMM) 等方法。

時至今日，實驗心理學和相關心理學雖然在本質上仍有所不同，但二者所使用的方法已經越來越相似，進而縮短這兩個領域間的鴻溝。尤其是認知心理學中的認知成分分析取向 (cognitive component analysis) 使實驗結果運用在測驗上的可能性增高 (Sternberg, 1991)。此一取向的做法是，1. 依據理論將受試者的訊息處理歷程分為數個成分，2. 依據不同成分的特性設計彼此相獨立的實驗，3. 將實驗過程中受試者的訊息處理的歷程化為數學模式，4. 考驗此模式並同時進行模式的參數估計，5. 研究這些成分在受試者間的相關以及與其他標準化心理計量測驗間的相關 (Sternberg, 1991)。因此可知數學模式在認知成分分析取向的研究中，是將成分與成分，以及成分與結果表現間的關係，根據有關的理論或實証研究，以數學方程式的形式加以模式化 (modeling)。而此數學模式除了具有考驗理論模式以及估計參數等功能外，也提供了認知成分分析取向與心理計量取向合作的契機。

近年來，有許多心理計量專家就掌握了此一契機，致力於心理計量與認知心理學的結合。其中，如 Fischer (1973) 所發展之線性洛基斯蒂克測驗模式 (linear logistic test model;

LLTM)、線性評定量表模式(linear rating scale model; LRSM) (Fischer, & Ponocny, 1995)、線性部分給分模式(linear partial credit model, LPCM) (Fischer, & Ponocny, 1995)以及 Embretson 所發展出的多成分潛在特質模式(multicomponent latent trait model; MLTM) (Whitely, 1980; Embretson, 1983, 1994)、一般化多成分潛在特質模式(general multicomponent latent trait model; GLTM) (Embretson, 1984; Embretson, Schneider, Roth, 1986)、測量改變的多向度潛在特質模式(multidimensional latent trait model for measuring change, MRMLC) (Embretson, 1992)、以及「MRMLC的拓展模式」(MRMLC+) (Embretson, 1995)等,皆在此一方向積極努力,並在模式的發展上不斷推陳出新。

基於此一趨勢的發展成熟,林世華、劉子鍵(民86)提出認知測量的整合模式,企圖結合認知設計系統(Cognitive Design System; CDS; Embretson, 1994)、反應產生模式(Response Generative Modeling; RGM; Bejar, 1993)、認知診斷評量(Cognitive Diagnostic Assessment; CDA; Nichols, 1994)、以及動態評量(Dynamic Assessment)等概念,並擬以電腦技術分別形成認知設計系統、反應產生系統、認知診斷評量系統以及動態評量系統等次系統,成為整合認知心理學、心理計量學及教學的理想模式。由於之前相關的研究已就此整合模式做了完整與詳細的說明(林世華、劉子鍵,民86;林世華、劉子鍵和梁仁楷,民86),因此以下不再贅述,僅以圖一來表示各個系統之間的關係。



圖一 整合認知心理學、心理計量學及教學的理想模式

必須說明的是認知設計系統是整合模式的核心,是以認知成分分析(Sternberg, 1984)為基礎,針對特定作業領域,依循 Embretson (1994) 認知設計系統的程序架構,依循著:確定測量的整體目標、確認試題的設計特徵、建構試題之解題歷程的認知模式、決定所欲操弄的試題內容特徵及其複雜度、產生設計規格相符的試題、將該認知模式轉換成心理計量模式、施測並評估試題的認知和心理特性、將試題的參數與能力參數標準化等邏輯程序,使試題具有以下特性: 1. 試題的意義與答題歷程的認知模式相連; 2. 可藉由內容特徵及其複雜度來明確界定各個試題的特性; 3. 可藉由試題的內容特徵及其複雜度來操弄試題的難易度。

另外,透過心理計量模式的估計,除可驗證模式的適合度,以瞭解該測驗的效度或模式

的合理性外，若驗證結果並未拒絕該模式，則可以進一步的以模式的估計結果來說明各試題的難度、各試題內容特徵的相對難度、個人的能力值以及三者之間的關係。依據上述估計結果所建立的「試題產生算則」(item-generation algorithm)可使所產生的試題在施測之前就具有已知的心理計量參數，此將有助於反應產生系統、認知診斷評量系統、以及動態評量系統等其他系統的運作。

由上述可知，認知模式與心理計量模式是否符合樣本資料，試題內容特徵的操弄是否正確等因素，皆維繫著認知評量整合模式的成敗與否。此皆有賴於依據施測之結果不斷地修正模式以及試題。本研究正是基於此一考量，繼前次研究僅以大學生為施測樣本之後，在原有樣本之中加入國高中生樣本，進一步分析模式的配適情形，以及解題歷程中各個成分之相對重要性的穩定程度。

貳、研究方法

以下分別就研究樣本、研究工具以及資料分析等三個部分來進行說明。

一、研究樣本

本研究除了包含林世華、劉子鍵與梁仁楷（民 86）的之研究中的樣本（研究樣本包括國立台灣師範大學教育心理與輔導學系、英文系、國文系等系的學生，共計 339 名）之外，並加入新的研究樣本，包括高中學生（國立華僑實驗高級中學一、二年級的聯考生）79 名，一般國中學生 75 名，合計研究樣本共有 493 名。

本研究中所進行分析的樣本涉及三種樣本組合，分別是整體樣本（包含大學生、高中及國中生）、大學生樣本，以及國高中生樣本。

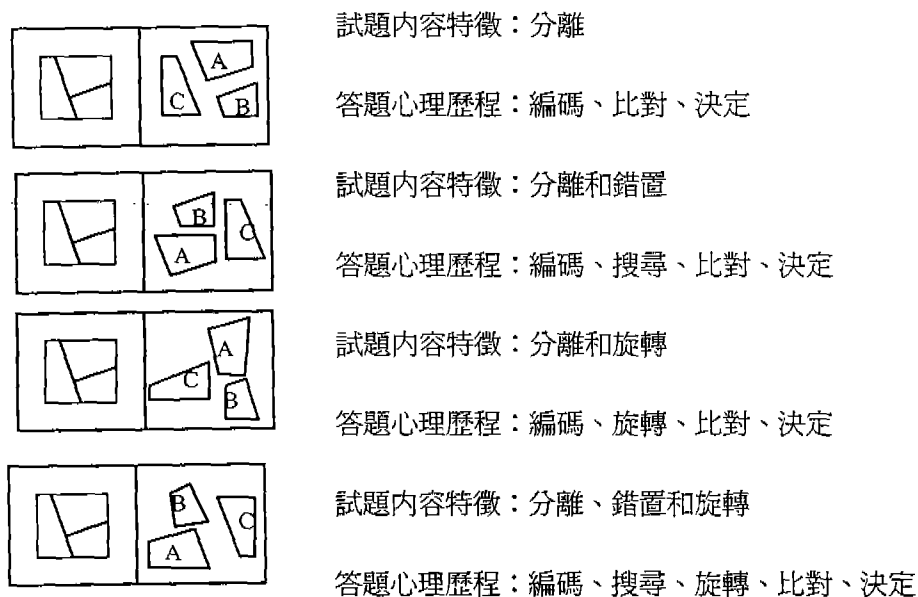
二、研究工具

本研究所使用的研究工具乃林世華、劉子鍵與梁仁楷（民 86）以 Pellegrino, Mumaw, & Shute (1985)發展的紙版測驗為藍本，利用試題產生輔助引擎（梁仁楷、劉子鍵，民 86）所編製的「二度空間視覺化能力測驗」。

Lohman (1979)在回顧一些嘗試整合空間能力測驗之因素分析研究後發現下列問題：(一)在不同的研究中，相同的測驗常採用不同的名稱；而不同內涵之測驗卻常採用相同的名稱。(二)測驗的形式或是答題之限制（例如：作答的時間）常常會影響因素結構的估計結果。(三)各研究所採用之因素萃取的方法以及因素轉軸的方法不盡相同，可能是導致研究結果不同的主要原因。據此，Lohman (1979)重新針對數個有關空間能力的研究再進行分析，試圖找出一組主要的空間因素。結果發現三個主要的因素，分別是空間定位 (spatial orientation)、空間關係 (spatial relation) 以及空間視覺化 (visualization) 等三個因素。一般而言，歸類於空間關係因素的測驗，作答比較容易，比較傾向於速度測驗；歸類於空間視覺化因素的測驗則較為複雜，傾向於難度測驗。而本研究所參考的測驗藍本，Pellegrino, Mumaw, & Shute (1985)所發展的紙版測驗，即歸類於空間視覺化向度。

本研究之「二度空間視覺化測驗」中所操弄的試題內容特徵包含：錯置、旋轉、是否錯

誤以及錯誤的類型（形狀錯誤或是鏡射）等。受試者在答題的過程中需經過編碼(encoding)、比對(comparison)、搜尋(search)、旋轉(rotation)以及決定(decision)等心理歷程。其中，錯置與旋轉此二項試題內容特徵所對應之受試答題時的心理歷程如圖二：



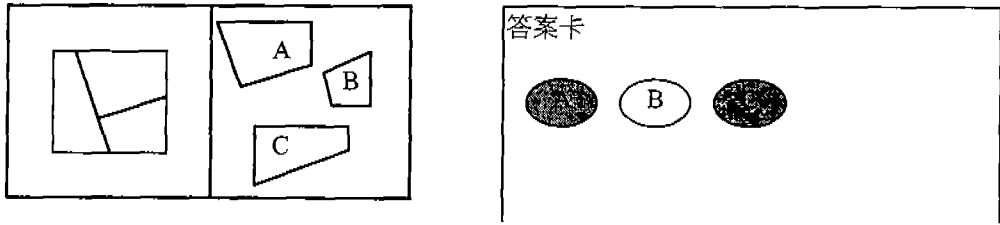
圖二 錯置與旋轉此二項試題內容特徵所對應之受試答題的心理歷程

另外，有無錯誤以及錯誤的類型（形狀錯誤或是鏡射）則是會影響受試在「比對」以及「決定」時的困難度。

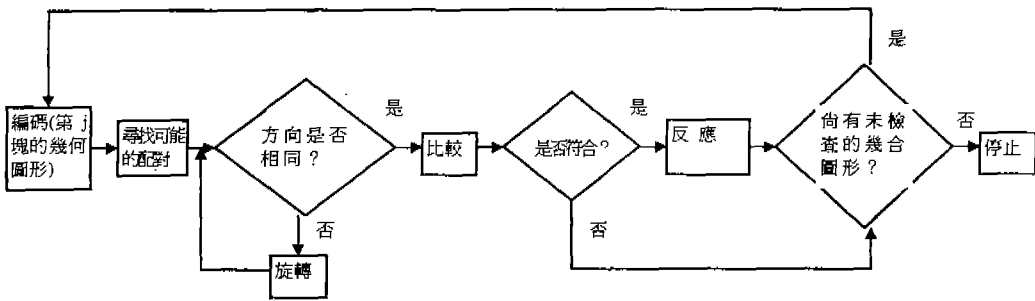
由於，Pellegrino 等人 (1985) 的相關研究中發現，當題目設計成右側的幾何圖形不能拼湊成左側的完整圖形時，作答者常常在發現該片不符的幾何圖形後採用「自我終結歷程的策略」(self-terminating processing strategy) (p.54)，亦即不再對剩餘的幾何圖形進行認知操弄，而導致所操弄的試題內容特徵無法有效地說明各個試題的難度值。

爲了避免上述現象的發生，本測驗對 Pellegrino 等人 (1985) 所設計的試題形式做了局部的修正。本測驗的典型試題如圖三。各試題的題幹（左側）皆是將一正方形切割成三個圖形，而作答者所要做的是：從右側的三個分開的圖形中找出與左側任何一個圖形相同者，並將該圖形之選項劃記於答案卡的相對位置上。當對同一試題的三個圖形的判斷皆正確時，該試題才算正確。

在圖三中，右側三個幾何圖形中的 A 與 C，與左側題幹中所分割的圖形相同，故正確答案是 AC。修正後之解題歷程的認知模式如圖四。



圖三 「二度空間視覺化測驗」中的典型試題



圖四 修正後之二度空間視覺化試題之解題歷程的認知模式

上述認知模式與 Pellegrino 等人所建構之認知模式最大的不同處在於強迫作答者對試題中的每片圖形（不論形狀一致與否）進行辨別。該機制促使答題者必須處理所有的幾何圖形，而不會中途終止。

本測驗將切割的片數固定為 3 片。並針對三項試題內容特徵採完全實驗設計，包括：錯置（有、無）、旋轉（旋轉角度固定為 90 度；旋轉的片數：0,1,2,3 片）及錯誤（錯誤的類型是形狀改變；錯誤的片數：0,1,2,3 片）等試題內容特徵，共計 32 題試題 (2 × 4 × 4)。另外，本測驗尚含 8 題錯誤類型為鏡射的試題，以供日後連結 (linking) 之用。因此，本測驗共包含 40 題試題，各試題的內容特徵如表一。

表一 各試題的內容特徵一覽表

	錯置的有無	旋轉的個數	錯誤的個數	錯誤的類型
1	0	0	0	0
2	0	0	1	0
3	0	0	2	0
4	0	0	3	0
5	0	1	0	0
6	0	1	1	0
7	0	1	2	0
8	0	1	3	0
9	0	2	0	0
10	0	2	1	0
11	0	2	2	0
12	0	2	3	0
13	0	3	0	0
14	0	3	1	0
15	0	3	2	0
16	0	3	3	0
17	1	0	0	0
18	1	0	1	0
19	1	0	2	0
20	1	0	3	0
21	1	1	0	0
22	1	1	1	0
23	1	1	2	0
24	1	1	3	0
25	1	2	0	0
26	1	2	1	0
27	1	2	2	0
28	1	2	3	0
29	1	3	0	0
30	1	3	1	0
31	1	3	2	0
32	1	3	3	0
33	0	0	1	1
34	0	0	3	1
35	0	2	1	1
36	0	2	3	1
37	1	0	1	1
38	1	0	3	1
39	1	2	1	1
40	1	2	3	1

註1：錯誤的類型該欄中，0代表錯誤的類型為圖形不符；1代表錯誤的類型為鏡射。

註2：本表僅供試題設計之用，實際施測時，試題的排序是採隨機編排。

三、資料分析

本研究之資料分別以 BILOG 3.08 與 LINLOG 程式處理。其中，BILOG 是用來驗證資料是否符合 Rasch 模式，並估計各試題的難度值。LINLOG 則是用以驗證 LLTM 的配適程度，以及估計各基本參數的相對加權值。其中，LLTM 中各題的難度值乃是以錯置的有無、旋轉的片數、錯誤的片數與錯誤的類型等內容特徵為基本參數的線性函數。其模式如式(1)。

$$P(X_{ij} = 1 | \theta_j, \eta_m, d) = \frac{\exp(\theta_j - (\sum_m c_{mi} \eta_m + d))}{1 + \exp(\theta_j - (\sum_m c_{mi} \eta_m + d))} \quad \dots(1)$$

$P(X_{ij} = 1 | \theta_j, \eta_m, d)$ 第 j 位受試在第 i 個試題上答對的條件機率

θ_j 第 j 位受試的能力參數

c_{mi} 第 i 題試題中，第 m 個內容特徵 ($m=1,2,3,4$; 分別代表錯置、旋轉、錯誤、與錯誤的類型) 的複雜度

η_m 是試題內容特徵 m 的加權值

d 常數

必須說明的是 LLTM 是 Rasch 模式的拓展模式，必須遵從 Rasch 的基本假設。因此，除了針對 9 題的分析之外（題目太少 BILOG 不提供配適指標），本研究在進行 LLTM 之前，皆會先確認該資料是否符合 Rasch 模式。

參、研究結果

首先，針對整體樣本的資料，以 LINLOG 進行 LLTM 的分析時，產生錯誤訊息。進一步分析資料後，發現第 33 題由於試題太簡單（試題的內容特徵為無錯置、無旋轉、無錯誤等），答對機率高達 99% 以上，因而 LINLOG 無法估計（LINLOG 採 CML 估計法）。同時，亦發現第 38 題因為試題特徵中的錯誤操弄得不明顯，導致許多作答者誤判，而使答對率偏低。因此決定將此二題刪除，而使本測驗的試題數目降為 38 題，此一現象與前次研究的結果相同。接著，應用 BILOG 針對篩選後的 38 題試題進行 Rasch 模式估計，估計結果顯示整體模式考驗達 .01 顯著水準 ($\chi^2 = 386.4$, $df = 237$, $p = .000$)，表示資料不符合單參數模式。隨後，進一步地以 2 參數模式及 3 參數模式進行考驗，結果發現 3 參數模式的整體模式考驗未達 .01 顯著水準 ($\chi^2 = 264.2$, $df = 236$, $p = .100$)（請參見表二）。此一結果與林世華、劉子鍵和梁仁楷（民 86）以大學學生為樣本的分析結果（應用 BILOG 針對同樣的 38 題試題進行 Rasch 模式估計，估計結果顯示整體模式考驗未達 .01 顯著水準； $\chi^2 = 257.4$, $df = 216$, $p = .0282$ ）並不一致。

為何大學生樣本的資料符合 Rasch 模式；而再加入高國中樣本之後，就不符合 Rasch 模式？針對該問題，進一步僅就高國中生樣本，應用 BILOG 就篩選後的 38 題試題進行 Rasch 模式估計，估計結果顯示整體模式考驗未達 .01 顯著水準 ($\chi^2 = 208.5$, $df = 164$, $p = .0107$)，表示高國中生樣本的資料符合 Rasch 模式。為便於分析比較，茲分別將以高國中生樣本資料進行 Rasch 模式估計的結果，和以大學學生樣本資料進行 Rasch 模式估計的結果，羅列於表三。

表二 整體樣本「二度空間視覺能力測驗」試題分析摘要表

	傳統分析		IRT 三參數模式				
	P	r_{bis}	a	b	c	χ^2	df
1	.751	.464	.660	-.918	.180	10.3	7.0
2	.491	.277	.427	.604	.148	6.8	8.0
3	.895	.754	.938	-1.795	.121	4.7	4.0
4	.594	.599	.868	-.255	.068	12.2	7.0
5	.728	.484	.651	-.865	.145	7.3	7.0
6	.874	.504	.579	-2.170	.126	7.2	6.0
7	.700	.504	.619	-.814	.105	5.5	7.0
8	.935	.636	.754	-2.556	.114	7.0	3.0
9	.509	.476	.632	.105	.067	10.9	8.0
10	.813	.540	.676	-1.376	.159	11.0	6.0
11	.890	.554	.666	-2.139	.126	5.8	5.0
12	.864	.535	.628	-1.980	.110	5.4	6.0
13	.897	.869	1.135	-1.697	.094	2.2	3.0
14	.757	.558	.640	-1.175	.086	7.0	7.0
15	.669	.622	.883	-.513	.101	3.9	7.0
16	.856	.566	.774	-1.570	.166	11.7	5.0
17	.706	.563	.726	-.743	.115	8.7	7.0
18	.746	.614	.809	-.930	.102	4.7	6.0
19	.710	.600	.783	-.752	.103	10.3	7.0
20	.767	.576	.747	-1.079	.110	5.2	6.0
21	.856	.606	.705	-1.792	.093	4.1	5.0
22	.566	.233	.313	.016	.138	6.7	9.0
23	.899	.662	.814	-1.945	.143	.9	5.0
24	.805	.479	.558	-1.615	.108	7.1	6.0
25	.824	.717	1.237	-.967	.219	6.0	5.0
26	.529	.385	.487	.090	.094	8.6	9.0
27	.834	.512	.603	-1.740	.117	2.4	7.0
28	.874	.640	.715	-1.939	.094	13.0	6.0
29	.850	.578	.697	-1.694	.127	4.7	6.0
30	.714	.595	.955	-.557	.177	6.1	6.0
31	.888	.584	.654	-2.160	.118	3.8	6.0
32	.795	.632	.863	-1.138	.121	4.1	6.0
34	.899	.692	.771	-2.087	.096	5.1	4.0
35	.588	.452	.733	-.041	.158	11.1	8.0
36	.771	.486	.538	-1.393	.105	8.3	7.0
37	.710	.579	.745	-.773	.104	8.9	7.0
39	.753	.635	.962	-.790	.154	5.6	6.0
40	.795	.547	.898	.656	1.525	.109	6.0

表三 高國中學生樣本與大學學生樣本「二度空間視覺能力測驗」試題分析摘要表

	高國中學生樣本					大學學生樣本				
	傳統分析		Rasch			傳統分析		Rasch		
	p	r _{bis}	b	χ^2	df	p	r _{bis}	b	χ^2	df
1	.797	.384	-1.524	6.1	5.0	.749	.541	-1.337	10.3	7.0
2	.378	.309	.545	4.9	6.0	.543	.212	-.197	21.5	8.0
3	.818	.702	-1.672	1.0	4.0	.929	.756	-3.077	5.2	2.0
4	.441	.655	.258	8.0	6.0	.664	.528	-.828	8.2	7.0
5	.706	.421	-.982	9.4	6.0	.743	.531	-1.299	4.7	6.0
6	.832	.452	-1.777	1.4	3.0	.894	.497	-2.569	1.9	4.0
7	.636	.481	-.628	5.7	5.0	.737	.457	-1.261	9.5	7.0
8	.923	.815	-2.706	1.4	1.0	.947	.492	-3.420	.3	4.0
9	.301	.527	.915	4.4	6.0	.599	.400	-.481	12.7	8.0
10	.685	.426	-.871	9.5	6.0	.864	.601	-2.245	13.0	5.0
11	.895	.792	-2.355	1.0	2.0	.903	.437	-2.681	6.2	5.0
12	.902	.862	-2.435	3.6	1.0	.855	.488	-2.158	7.4	6.0
13	.888	1.053	-2.279	13.9	1.0	.912	.772	-2.802	8.2	3.0
14	.783	.704	-1.432	3.0	5.0	.761	.509	-1.415	2.3	7.0
15	.545	.673	-.208	9.9	6.0	.735	.547	-1.243	7.8	7.0
16	.853	.453	-1.948	5.8	3.0	.864	.658	-2.245	16.1	4.0
17	.566	.634	-.303	5.6	6.0	.767	.479	-1.454	5.6	7.0
18	.573	.471	-.334	10.0	6.0	.829	.608	-1.921	3.2	5.0
19	.573	.559	-.334	6.3	5.0	.779	.529	-1.536	5.9	7.0
20	.748	.655	-1.217	1.5	5.0	.779	.578	-1.536	3.1	5.0
21	.825	.781	-1.724	8.4	4.0	.879	.481	-2.399	1.7	5.0
22	.483	.260	.073	9.6	6.0	.602	.177	-.496	21.3	8.0
23	.853	.727	-1.948	3.4	3.0	.917	.615	-2.888	1.8	4.0
24	.832	.649	-1.777	1.5	3.0	.808	.440	-1.754	5.2	5.0
25	.699	.563	-.945	12.4	5.0	.888	.742	-2.500	11.7	3.0
26	.371	.429	.577	8.6	6.0	.605	.288	-.511	10.2	8.0
27	.846	.606	-1.889	1.5	3.0	.835	.511	-1.971	1.9	6.0
28	.846	.742	-1.889	4.2	3.0	.897	.508	-2.606	3.4	5.0
29	.839	.596	-1.832	.4	3.0	.864	.566	-2.245	1.6	5.0
30	.510	.558	-.052	10.4	5.0	.805	.566	-1.731	5.0	5.0
31	.811	.598	-1.622	3.9	4.0	.923	.490	-2.979	.3	5.0
32	.706	.580	-.982	3.3	5.0	.847	.603	-2.075	7.8	5.0
34	.909	.983	-2.519	.1	.0	.903	.582	-2.681	3.3	5.0
35	.490	.463	.042	5.1	6.0	.631	.433	-.651	4.5	8.0
36	.699	.706	-.945	4.6	5.0	.805	.282	-1.730	9.5	7.0
37	.573	.476	-.334	7.8	6.0	.779	.577	-1.536	2.4	6.0
39	.622	.654	-.562	8.5	5.0	.817	.546	-1.824	4.2	6.0
40	.804	.684	-1.573	2.6	4.0	.802	.478	-1.708	8.4	6.0

再針對高國中生樣本資料就篩選後的 38 題試題進行 LLTM 的估計。結果顯示，以錯置（有、無）、旋轉（旋轉角度固定為 90 度；旋轉的片數：0,1,2,3 片）、錯誤（錯誤的類型是形狀改變；錯誤的片數：0,1,2,3 片）與錯誤的類型是否為鏡射等內容特徵為基本參數的線性函數，並不能產生與 Rasch 模式一致的難度估計值（ $\chi^2 = 454.97$, $df = 34$ ，達 .01 顯著水準）。而 LLTM 中各試題的難度估計值與 Rasch 模式中各試題的難度估計值的相關為 .73。針對此一結果，進一步繪製散佈圖加以分析，結果發現(1)當旋轉的片數為 2 至 3 片時，試題的難度並不一定會隨著旋轉片數的增加而提升、(2)當錯誤的片數為 2 至 3 片時，試題的難度並不一定會因為錯誤片數的增加而提升。此一結果與 Pellegrino (1985) 的研究發現近似，也與本文研究者前次的研究結果相吻合。據此，進一步地將內容特徵為錯置（有、無）、旋轉（旋轉角度固定為 90 度；旋轉的片數：0,1,片）、與錯誤的片數（錯誤類型包含形狀不一致與鏡射；錯誤的片數：0,1 片）等試題挑選出。其中，錯誤類型為形狀錯誤者共有 7 題（原本有 8 題，扣掉第 33 題），錯誤類型為鏡射者共有 2 題，故挑選後共有 9 題。直接就高國中生的樣本資料以 LLTM 進行估計。結果發現，以錯置（有、無）、旋轉（旋轉角度固定為 90 度；旋轉的片數：0,1,片）、錯誤的片數（錯誤的片數：0,1 片）與錯誤的類型是否為鏡射等內容特徵為基本參數的線性函數，可產生與 Rasch 模式一致的難度估計值（ $\chi^2 = 9.32$, $df = 5$ ，未達 .01 顯著水準），而 LLTM 中各試題的難度估計值與 Rasch 模式中各試題的難度估計值的相關為 .827。進一步地以巢狀模式 (nested model) 進行分析，結果發現錯置（有、無）的加權值為 0.37（ $\chi^2 = 24.86$, $df = 1$ ，達 .05 顯著水準），旋轉（旋轉角度固定為 90 度；旋轉的片數：0,1,片）的加權值為 0.44（ $\chi^2 = 29.27$, $df = 1$ ，達 .05 顯著水準），錯誤的片數（0,1 片）的加權值為 0.74（ $\chi^2 = 15.03$, $df = 1$ ，達 .01 顯著水準）與錯誤的類型是否為鏡射的加權值為 0.04（ $\chi^2 = .03$, $df = 1$ ，未達 .01 顯著水準），而截距項為 -.9037。

上述結果與前次研究以大學生為樣本分析的結果有顯著的不同。茲以表四與表五來對照說明。

表四 9題試題的LLTM之巢狀模式中各樣本在各模式的適合度及 χ^2 增加量

	基本參數（試題內容特徵）				模式適合度		增加量	
	錯置	旋轉	錯誤片數	錯誤類型	χ^2	df	χ^2	df
高 國 中 學 生	✓	✓	✓	✓	9.32	5		
		✓	✓	✓	13.71	6	4.39*	1
	✓		✓	✓	13.93	6	4.61*	1
	✓	✓		✓	21.76	6	12.44**	1
	✓	✓	✓		9.35	6	.03	1
大 學 學 生	✓	✓	✓	✓	7.13	5		
		✓	✓	✓	31.99	6	24.86**	1
	✓		✓	✓	36.40	6	29.27**	1
	✓	✓		✓	22.16	6	15.03**	1
	✓	✓	✓		22.90	6	15.77**	1

*代表達 .05 顯著水；**代表達 .01 顯著水準

由表四可發現大學生樣本的分析結果中，所有基本參數的加權值皆達 .01 顯著水準。而在高國中學生樣本的分析結果中錯誤類型此一基本參數的加權值未達 .05 的顯著水準，此代表就高國中生的樣本而言試題之錯誤的類型是否是鏡射並不會影響試題的難度值。

表五 二樣本在 LLTM 估計結果中基本參數之加權值的對照表 *

	基本參數的加權值			
	錯置	旋轉	錯誤片數	錯誤類型
高國中學生樣本	0.37 (3)	0.44 (2)	0.74 (1)	0.04 (4)
大學學生樣本	0.63 (3)	0.81 (1)	0.57 (4)	0.70 (2)

* () 中的數字為各加權值由大到小的排序，並未經過統計考驗

由表五可看出高國中生樣本與大學生樣本的分析結果中各基本參數之加權值的相對大小排序並不相同。綜合表四即表五的結果可知，對高國中學生樣本與大學學生樣本而言，各個對應於錯置、旋轉、錯誤片數與錯誤類型四個試題內容特徵的認知成分之相對重要性不盡相同。其中，又以二組樣本在錯誤類型上之估計結果的差異，最值得注意。

上述二組樣本在 LLTM 上分析結果的差異，延伸出另一值得思考的問題，即依二者分析結果所建構的試題產生算則將有所不同，並因此對於具有相同試題內容特徵的試題將預估出不同意涵的難度值（雖然在現有 LLTM 估計程序下，二組樣本之分析結果並未等化，因此二組樣本間的各參數估計值不能直接比較；然而，二組樣本所估計之基本參數的加權值大小排序不同，因此所建構出的試題產生算則的意涵也就不相同）。以高國中生樣本之 LLTM 分析結果建構的預測難度方程式，和以大學生樣本之 LLTM 分析結果所建構的試題產生算則分別如式(2)及式(3)。

$$b_i = 0.37C_{1i} + 0.44C_{2i} + 0.74C_{3i} - .9027 \quad \dots(2)$$

$$b_i = 0.63C_{1i} + 0.81C_{2i} + 0.57C_{3i} + 0.70C_{4i} - 1.24740 \quad \dots(3)$$

其中， b_i 代表第 i 個試題的難度值； C_1 代表錯置的有無； C_2 代表旋轉的片數（0,1 片）； C_3 代表錯誤的片數（0,1 片）； C_4 代表錯誤的類型（形狀不一致或鏡射）。基於式(2)及式(3)，便可在試題產生之初就藉由操弄試題的內容特徵來預估某試題的難度值。舉一試題說明，若某一試題中有錯置、有一片旋轉、有一片錯誤、而錯誤的類型是鏡射，因此該題的內容特徵矩陣是 [1,1,1,1]。將該試題的試題內容特徵矩陣帶入式(2)及式(3)，可分別得到一個難度預測值。其中，以國高中生樣本分析結果建構出的試題產生算則，其難度預測值是 0.6473；以大學生樣本分析結果建構出的試題產生算則，其難度預測值是 1.4626。

最後，將本研究的結果作歸納說明。本研究在原有的大學生的樣本中加入國中與中學生的樣本，以增加樣本的變異程度，藉此驗證模式的配適程度。其中，針對篩選後的 38 題試題就整體樣本資料進行 Rasch 模式考驗，結果發現整體樣本資料並不符合 Rasch 模式，而是符合 3 參數模式。此項結果與先前的研究結果大有出入。進一步針對篩選後的 38 題試題

就新加入之高國中生樣本進行研究，卻發現該樣本資料符合 Rasch 模式。顯示大學生與高國中生此二組樣本在組內有相當的一致性，但在組間卻有差異性。由於本研究所採用之測驗工具的編製是依據 Embretson (1994) 認知設計系統的程序架構所編製而成，試題中各個成分皆有其認知心理學上的意義，且經過系統的操弄。因此本研究進一步探討，希望瞭解兩組在各個成分上的表現是否亦有所不同。就篩選後的 9 題，對兩組樣本進行 LLTM 分析，結果發現兩組樣本在基本參數之加權值的相對排序上有明顯的差異。此一差異將造成兩組所建構的試題產生算則具有不同的意涵。

肆、結論與建議

本研究主要是依據林世華、劉子鍵與梁仁楷（民 86）的研究架構，在原有的大學生的樣本中加入國中與高中生的樣本，以增加樣本的變異程度，藉此來驗證 LLTM 的配適程度；並經由新樣本的加入，增加可能的反應組型，藉此來觀察 LLTM 估計結果中各成分之相對重要性的穩定程度。以下將對研究結果的利弊得失進行討論，並提出建議。

(一)新增加的樣本雖能增加樣本的變異程度，但無法有效提升 LLTM 的配適程度，甚至會影響 Rasch 模式的配適。進一步研究發現可能的原因是新加入的高國中生樣本的作答反應與大學生樣本的作答反應並不一致。此項結果延伸一個重要的議題，即模式分析結果的外在效度方面的問題。此項結果希冀間接建立一個放諸四海皆準的模式並不切合實際。認知評量整合模式的取向，是將測驗表現時潛藏的認知歷程和認知結構等訊息納入試題產生算則，利用此種方法讓所產生的試題具有已知的心理計量參數，而此一取向亦產生另一項有助於修正模式的副產品，就是在測驗實施的過程中便不斷地驗證有關反應歷程的知識。效度因此成為不斷進行的歷程，而非偶然的事件。然而，值得注意的是當樣本越來越大時，極有可能因此忽略了樣本中各個群體的特殊性。因此，即使以大範圍的樣本來建立模式，也要細心檢驗樣本中各群體的特殊性，甚至針對各個特定群體建構試題產生算則。以期在以簡馭繁和重視差異間求得平衡。

(二)當試題的內容特徵為錯置（有、無）、旋轉（旋轉角度固定為 90 度；旋轉的片數：0, 1, 片）、錯誤（錯誤的類型是形狀改變；錯誤的片數：0, 1 片）與錯誤的類型是否為鏡射時，對大學生與高國中生樣本進行 LLTM 分析，結果發現兩組樣本在基本參數之加權值的相對重要性上有所差異。其中，最值得注意的是在高國中生樣本中錯誤的類型（是形狀錯誤或是鏡射）並不會影響試題的難度；此一結果與大學生組的結果相左。由於錯誤的類型此一試題內容特徵與受試答題時的「比對」以及「決定」等心理歷程有關。因此，此項研究結果意味著：對高國中生之「比對」以及「決定」等心理歷程而言，試題之錯誤類型是形狀錯誤或是鏡射，所造成的困難度是相等的；但對於大學生的「比對」與「決定」等心理歷程而言，試題之錯誤類型為鏡射者遠比錯誤類型是形狀錯誤者來的困難。造成此一現象的原因是因為隨著年齡的增長，個體在「形狀錯誤」的比對與決定上，比在「鏡射」的比對與決定上發展得快；或是因為高能力者（此指學業成就，國立台灣師範大學學生是經過大學聯考篩選）在「形狀錯誤」的比對與決定上，比在「鏡射」的比對與決定上來得好；或是尚有其他原因。值得進一步研究。

(三)以往研究者在進行 IRT 分析發現資料與理論模式不符時，往往無法確定原因是出自於某些特定試題、或是特殊的反應組型、或是試題與反應組型的交互作用。之所以如此，是

因為傳統測驗的編製只著重於整體測驗的建構效度，而非重視個別試題的建構效度。因此在判斷試題是否合適時，試題本身並不能提供充分的資訊。且在判讀特殊反應組型時，也缺乏進一步的資訊來瞭解反應組型背後所隱藏的意義。本研究的過程中發現，當依據成分分析取向建構試題，在使用 IRT 進行分析時若遭遇資料與理論模式不符時，可依據當初試題特徵操弄的設計來逐一檢視各個試題。另外，利用 LLTM 進行後續的分析，也可以用來補充說明特定反應組型的可能成因。

(四)本研究分別對大學生與高國中生樣本進行 LLTM 分析，結果發現兩組學生在基本參數之加權值的相對重要性上有所差異。若只針對整體樣本進行分析，將忽略二組樣本的差異性。此一結果間接的指出 LLTM 的侷限性，即利用單一線性函數並無法有效說明樣本中特質互異之數個群體的實際狀況。事實上，此一侷限性在傳統迴歸分析中早已發現，階層線性模式 (hierarchical linear model; HLM) 就是在此需求下而發展 (劉子鍵、林原宏，民 86)。為提升 LLTM 的實用性，以及估計結果的正確性，如何依據 HLM 的概念拓展 LLTM 則是心理計量學者另一重要課題。

(五)雖然本研究修正 Pellegrino 等人所建構之認知模式，改變作答的形式，希望能促使答題者必須處理所有的幾何圖形，而不會中途終止。但研究發現，高國中生與大學生樣本的分析結果相似，即旋轉與錯誤對試題難度的影響並不會隨著片數的增加而直線上升。而此一結果與 Pellegrino 等人的研究結果類似。因此，有徹底檢討此一修正模式的必要。

(六)本研究所雖新增加高國中生樣本，但在取樣的過程中是採取立意取樣，因此在研究結果的推論上有其侷限性。後續研究有必要再修正認知模式、改良試題內容特徵的操弄 (如：第 38 題)、並就大學學生、高中學生、國中學生以及國小高年級學生作為施測的對象，以進一步驗證認知模式並瞭解各特定群體的表现。

參考資料

- 林世華、劉子鍵 (民 86)。整合認知心理學、心理計量學與教學的理想模式：結合認知設計系統、反應產生模式、認知診斷評量系統以及動態評量系統。**教育測驗新進發展趨勢學術研討會論文集 (pp.229-236)**。國立台南師範學院。
- 林世華、劉子鍵和梁仁楷 (民 86)。認知設計系統的建構與試題輔助產生引擎的運作—以二度空間視覺化測驗為例。**師大學報**。(印行中)。
- 梁仁楷、劉子鍵 (民 86)。試題產生輔助引擎 1.0 版。未出版。
- 劉子鍵、林原宏。(民 86)階層線性模式之理論與應用：以「影響自然科成績之因素的研究」為分析實例。**教育與心理研究**，20，1-22。
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen et al. (Eds.), *Test Theory for a New Generation of Test* (pp.323-358). Hillsdal, NJ: Lawrence Erlbaum Associates.
- Carroll, J. B., & Maxwell, S. (1979). Individual difference in ability. *Annual Review of Psychology*, 30, 603-640.
- Embretson, S. E. (1995). A measurement model for linking individual learning to process and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32(3), pp.277-294.

- Embretson, S. E. (1983). Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179-197.
- Embretson, S. E. (1984). A general latent trait model for response process. *Psychometrika*, 49, 175-186.
- Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp.195-218). New York: Academic Press.
- Embretson, S. E. (1992). Implication of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp.184-197). Washing, D.C.: APA
- Embretson, S. E. (1994). Applications of Cognitive Design Systems to test development. In C. R. Reynolds. (Eds), *Cognitive assessment: A multidisciplinary perspective* (pp.107-136). NY: Plnum Press.
- Embretson, S. E. (1995). A measurement model for linking individual learning to process and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32(3), pp.277-294.
- Embretson, S. E., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement*, 23(1), 13-32.
- Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175-193.
- Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Fischer, G. H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer & I. W. Molennar (Eds), *Rasch models foundations, recent developments, and applications* (pp.253-279). New York: Springer-Verlag.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Klvwer Nijhoff.
- Lohman, D. F. (1979). *Spatial ability: a review and re-analysis of the correlational literature*. Aptitude Research Project, Report gnostic assessments. *Review of Educational Research*, 64(4), 575-603.
- Pellegrino, J. W., Mumaw, R. J., & Shute, V. J. (1985). Analysis of spatial aptitude and expertise. In S. E. Embretson (Ed.). *Test design: Developments in psychology and psychometrics* (pp.45-76). New York: Academic Press.
- Sternberg, R. J. (1991). Cognitive theory and Psychometrica. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: theory and applications* (pp.367-394). Boston: Kluwer.
- Whitely, S. E. (1980). Modeling aptitude test validity from cognitive component. *Journal of Educational Psychology*, 72, 750-769.
- Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies:

A test theory approach. *Applied psychological Measurement*, 383-397.

Whitely, S. E., Schneider, L. M. & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning test. *Journal of Educational Measurement*, 23(1), 13-32.

Bulletin of Educational Psychology, 1998, 30(1), 177-193
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

The Item-Generation Algorithm of Two Dimension Spatial Visualization Ability Test: Testing and Modifying

Tzu-chien liu* Sieh-Hwa Lin** Steven Liang***

ABSTRACT

Based on the integrated cognitive assessment model proposed by Lin and Liu (1997a) and the research findings of Lin, Liu and Liang (1997b), this study aims to investigate the following questions: (1) whether adding subjects coming from lower age levels may increase the degree of model fit of LLTM, and (2) whether the response patterns, due to the new subjects, may enhance the stability of the weights of the LLTM components. Hence, unlike Lin, Liu and Liang (1997b) whose samples were college students only, this study includes additional samples from 8th grader to 11th grader students.

Compared with Lin, Liu and Liang (1997b), this study shows two significance difference. First, the data from the overall samples does not well fit the Rasch Model, in stead fits 3 pl model; the data from the high school subjects, however, fits the Rasch Model. The result suggests consistency within the same age groups but diversity between different age groups. Second, the LLTM analysis of the two age groups reveals different ranking of component weights. The difference results in constructing two difference item generation algorithms. Some discussions and suggestions are also given in the study.

Keywords: integrated cognitive assessment model, spatial ability, linear logistic test model

* National Overseas Chinese Experimental High School

** Department of Educational Psychology & Counseling, National Taiwan Normal University

*** Department of Computer Science and Information Engineering, National Central University