

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

林世華 劉子鍵 梁仁楷

國立臺灣師範大學

國立中央大學

本研究以「二度空間視覺化能力」為特定研究領域，致力於落實林世華、劉子鍵（民86）所提出的認知測量整合模式，具體的目標包括：(一)參考 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，編製二度空間視覺化能力測驗。(二)建立「試題輔助產生引擎」達成下列階段性的任務。1. 擔任輔助命題工具：本研究需參照認知設計系統的程序架構，編製大量的試題，期望經由實證研究來建立「試題產生算則」。因此，「試題輔助產生引擎」所擔負的任務在輔助研究者有效地操弄試題內容特徵、有系統地設計出試題、簡易地修改與操弄試題，且方便列印成試卷。2. 當作「試題自動產生引擎」的測試版：藉由「試題輔助產生引擎」的應用結果，可清楚地瞭解本系統不成熟之處，可藉此修正改進。本研究最後將針對上述兩項目標的研究結果，評估認知測量整合模式的可行性，並提出該模式有待修正的地方。為顧及概念的完整性，全文將分成：認知測量的發展背景、認知測量的整合模式、研究方法與步驟、研究結果以及結論與建議等五個部分加以呈現。

認知測量的發展背景

Carroll 和 Maxwell (1979) 在 Annual Review of Psychology 中指出自從 25 年前首位學者就測驗之發展進行回顧以來，此一領域少有改變。然而，少有改變並不代表傳統測驗的發展已臻成熟。事實上，由於心理學理論和測驗分析技術的限制，使得早期的測驗編製者及使用者不得不去漠視或忽略一些早已存在的問題。例如：在心理學界尚以行為主義馬首是瞻的時候，實驗心理學家避而不談人類內心的世界，只在乎外在刺激與反應行為間的關係。因此，對於受試者答題時的心理內在歷程少有所知。另一方面，傳統心理計量取向，雖然重視抽象的心理特質（如：智力、動機等），且以相關、因素分析等統計技術來避開心理特質之量尺 (scale) 不明的窘境，但此種做法的效度問題早已受到廣泛的注意與批評。

依據傳統心理計量的做法，測驗的編製者利用所編製的測驗與其他效標測驗間的相關來支持該測驗的建構效度，此種做法並無法詳細說明受試者解題時所需的特定技能、知識與歷程。之所以如此，最主要的原因在於傳統心理計量取向強調以統計技術從測驗結果來推論受試者能力的因素組型，而未建立受試者實際答題歷程的理論基礎，因此無法建立有效的心理計量模式來說明測驗刺激之特徵與個人特質間的關係。

近年來，心理學理論的更迭和心理計量技術的革新對測驗發展的方向影響甚大。其中，認知心理學不再視人類的內心世界為黑箱子，也不再視作答的歷程為刺激反應的聯結。認知心理學中的認知成分分析 (cognitive component analysis) 以及 Vygot-



sky 的認知發展理論對測驗的發展具有關鍵性的影響。不同於以往心理學家以工作分析(task analysis)，依邏輯順序將某項作業分解成數個次作業，而後將次作業再細分的做法（例：Gagne 的學習階層(learning hierarchy)）；認知成分分析則改採分析答題歷程中所需的認知成分，以確實瞭解答題時所需之特定技能、知識與歷程，以及答題者所應用的策略。

另外，相對於傳統測驗強調學習者知道什麼(what)，但無法反應出學習者如何(how)知道；強調學習者「所能為者」的能力，而忽略了學習者「可能為者」的能力；只能描述學習者成功或失敗的情形，而無法針對其失敗的真正原因進行診斷。Vygotsky (1978) 認知發展理論中的近側發展區(zone of proximal development, ZPD) 概念對動態評量(dynamic assessment) 的影響甚大。Vygotsky 採「學習先於發展」的觀念，認為評量不僅應評估目前已發展的認知能力，也應評估正在發展的潛在認知能力。此外，Vygotsky 亦強調教學者與學習者間的互動(interaction)，認為評量應兼重學習結果與學習歷程，且學習歷程中的評量應提供有關教學處方的訊息。

最後，在心理計量分析技術上，亦因試題反應理論(item response theory) 的出現而有劃時代的改變。Warm (1978) 便指出試題反應理論的出現與發展對於心理計量學而言，好比是愛因斯坦相對論(Einsteinian theory of relativity) 在物理學上之重要性一樣。之所以如此，是因為試題反應理論提供了與古典測驗理論(classic test theory) 不同的假設與

觀點。其中，潛在特質與試題難度具有相同尺度(scale)、以試題(item) 為分析單位、以及在模式中允許將試題參數從能力參數中解離等特性，皆有別於古典測驗理論，也開拓了IRT 更廣大的應用空間。另外，近年來有一些心理計量的學者致力於將認知心理學與心理計量結合。諸如 Fischer (1973) 所發展之線性洛基斯蒂克測驗模式(linear logistic test model; LLTM) 和 Embretson (註一) 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，皆能有效地將當代心理學理論與心理計量模式結合。

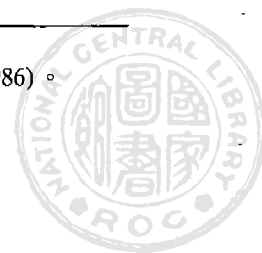
由於認知心理學理論與心理計量技術的發展，以及近年來重視測驗在教學診斷上的功能，Sternberg (1991) 指出未來測驗的發展方向應是結合認知心理學理論、心理計量學以及教學，使測驗的發展具有認知心理學的基礎，測驗的結果能提供有關訊息處理的診斷訊息，測驗的分數能反應出答題的歷程等。然而，截至目前為止，雖有專家學者提出動態評量、認知診斷評量等模式，但多為各自獨立，並無法有效整合認知心理學理論、心理計量學與教學。

認知測量的整合模式

林世華和劉子鍵(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

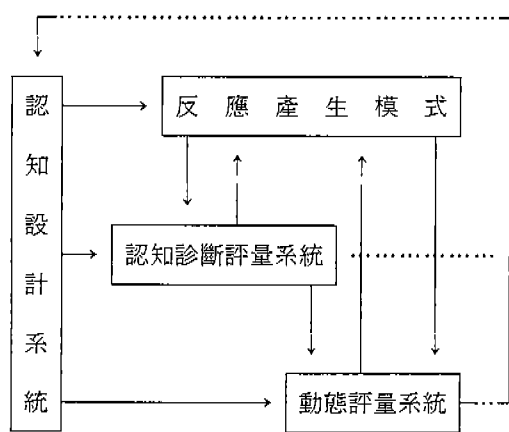
System; CDS; Embretson, 1994)、反應產生模式(Response Generative Modeling; RGM; Bejar,

註一：S.E. Embretson 曾以 S.E. Whitely 為名發表多篇關於 MLTM 的文章 (1980a, 1980b, 1980c, 1981a, 1981b, 1986)。



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。

以下將進一步介紹整合模式中的各個系統, 以及系統之間的關係 (如圖一)。



圖一 整合認知心理學、心理計量學及教學的理想模式

一、認知設計系統

整合模式之認知設計系統乃以認知成分分析 (Sternberg, 1984) 為基礎, 針對特定作業領域, 參考 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. 每個試題的特性能以內容特徵及其複雜度來明確界定; 3. 試

題的難度能藉由試題的內容特徵及其複雜度來操弄。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) 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。

二、反應產生系統

本系統主要是根據 Bejar (1993) 所提出的概念設計而成。可分為「試題產生算則」以及「試題自動產生引擎」二個部分。其中, 「試題產生算則」乃根據認知設計系統的估計結果 (包括試題的心理計量特徵 (難度)、試題的內容特徵 (可引發特定的認知成分)、個人特徵 (能力) 以及三者之間的關係) 等所建立的。

而「試題自動產生引擎」主要是利用先前建立的「試題產生算則」、加上根據認知診斷系統和動態評量系統的回饋所建立之作答者答題歷程資料庫、以及結合電腦的技術, 以達到下列目標: 1. 能自動產生具有不同表面結構, 但具有特定深層結構的試題。2. 能自動產生具有特定難度值的試題。3. 能根據認知診斷系統的回饋自動產生符合作答者解題歷程特徵的試題。4. 能配合認知診斷系統與動態評量系統的回饋, 針對某一特定認知成分組型自動產生一系列具有難度梯度及認知階層的試題。

三、認知診斷評量系統

本認知診斷評量系統是修正 Nichols (1994) 的架構而來。本認知適性診斷評量系統又可再細分成探測試題之命題機制、施測介面、偵查系統、評斷系統、回饋系統等等五個次系統。

其中, 探測試題之命題機制乃一組特定的命題



規則，能有系統地操弄試題的表面結構與深層結構。反應產生系統則依據該命題規則，有系統地產生探測試題，並藉由施測介面在電腦上呈現。

偵察系統則負責記錄作答者的反應組型，分析出作答者答題歷程的特徵，並與認知設計系統的資料庫相比對。若所獲得的訊息充份，則可推估出作答者的能力值，以及作答者在哪些試題內容特徵所組成的試題上的表現的較不精熟。若訊息不充份，則將結果回饋給探測試題之命題機制，縮小範圍，經由反應產生系統再產生另一組探測試題，透過施測介面進行施測。直至訊息充份為止。

當訊息充分時，評斷系統可根據偵查系統的偵察結果（所推估之作答者的能力值，以及作答者在哪些試題內容特徵所組成的試題上的表現的較不精熟）提出具體的建議及評語。

最後，回饋系統會累積每位作答者的反應資料，等到資料達到某一特定數量後，將該資料回饋至認知設計系統中，重新估計試題的難度、內容特徵的相對難度、作答者的能力值以及三者之間的關係等。並依照估計結果修正試題產生算則。預計，當樣本越來越多之後，各項估計值將更趨於穩定。

四、動態評量系統

此系統乃根據精熟學習 (mastery learning) 的精神以及蘇俄心理學家 Vygotsky (1978) 認知發展理論中近側發展區 (zone of proximal development, ZPD) 和鷹架 (scaffolding) 的概念加以發展而成。其主要分為鷹架系統及評量系統以及回饋系統等三個次系統。

其中，鷹架系統是參照認知診斷評量系統所提供的訊息（所推估之作答者的能力值，以及作答者

在哪些試題內容特徵所組成的試題上的表現的較不精熟）進行分析，將作答者不精熟的試題，分解成一系列具有難度梯度及認知階層的試題。並藉由反應產生系統產生難度由容易到困難、認知需求由簡單到複雜的試題組，如同鷹架般循序漸進地提供給作答者，使作答者在答題的歷程中學習到知識的深層結構。若作答者在某一難度或認知階層上的實作表現達到標準，則本系統會授命反應產生系統產生更高難度及認知階層的試題；若作答者在某一難度或認知階層上的實作表現未達到標準，則本系統會授命反應產生系統產生難度相同且認知階層相同但表層結構不同的試題。而評量系統可提供「起點試題」與「目標試題」間的難度差異與「起點試題」到「目標試題」之間所需提示試題的數量等訊息，藉由此二者的比值可估計出作答者學習的潛能。

此外，如同認知診斷評量系統中的回饋系統一般，動態評量系統中的回饋系統亦會累積每位作答者在動態評量系統中的反應資料，等到資料達到某一特定數量後，將該資料回饋至認知設計系統中，重新估計試題的難度、內容特徵的相對難度、作答者的能力以及三者之間的關係等。

上述整合模式的落實必須透過電腦技術的配合。以下利用圖二來說明整合模式的系統開發流程。

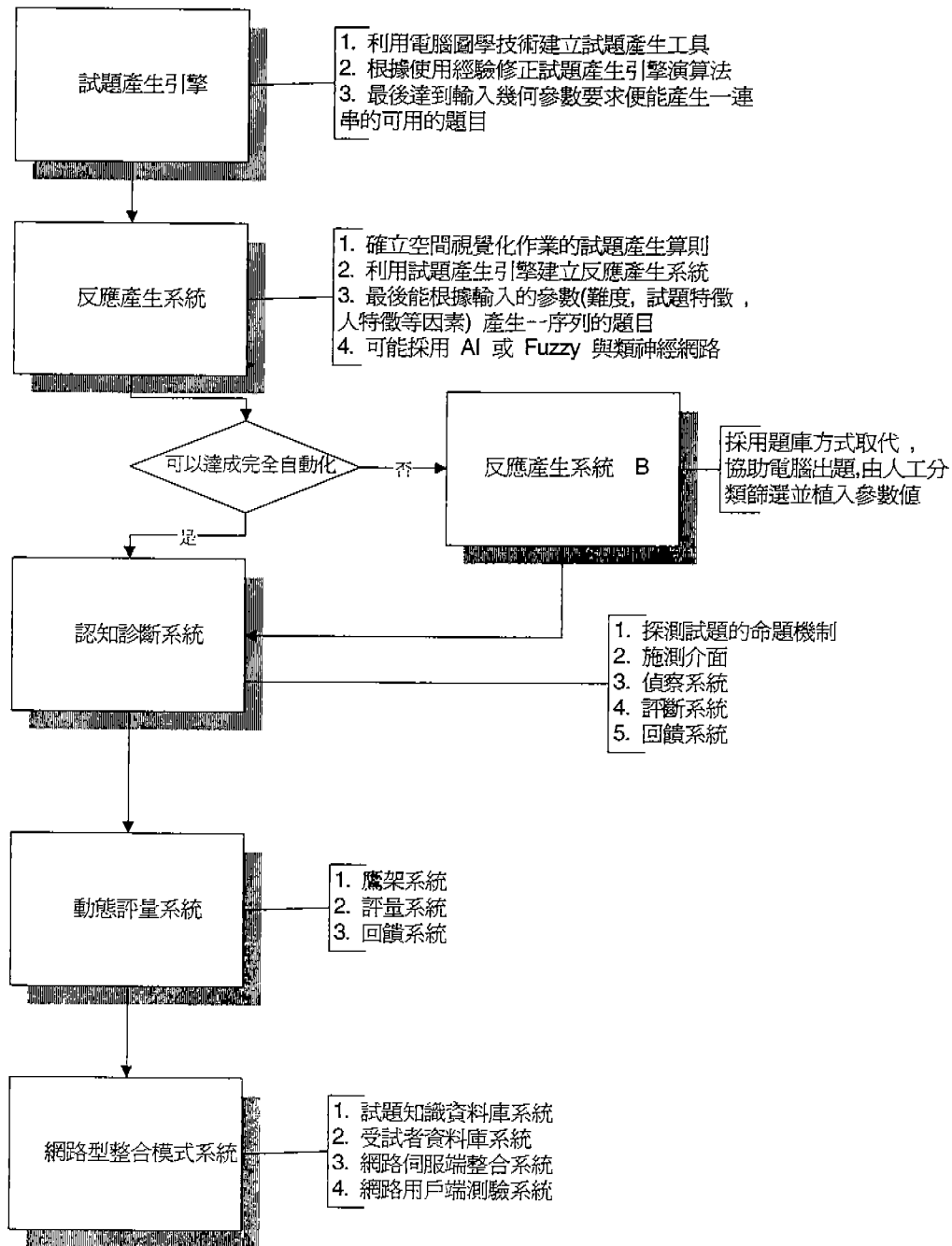
綜觀上述對整合模式的介紹可知，「認知設計系統」在整個模式中扮演著總攬全局的角色，其不僅關係著「試題」的效度，亦關係著「試題產生算則」的品質，而二者皆會嚴重影響其他系統的運作。此外，「試題產生系統」扮演著承先啓後的角色，尤其該系統中的「試題自動產生引擎」更是樞紐。藉由該機制根據其他系統傳來的訊息有效地產生特定的試題，本模式才能有效地整合。

研究方法與步驟

本研究乃針對「二度空間視覺化能力」，以認知設計系統的程序架構來設計試題，利用試題輔助

產生引擎來產生試題，藉以編製「二度空間視覺化能力測驗」，其目的是為整合模式進行奠基的工作。





圖二 系統開發流程



茲將本研究之研究方法與步驟分述如下：

一、「二度空間視覺化能力測驗」的編製

本研究主要是依據 Embretson (1994) 認知設計系統的程序架構，來編製「二度空間視覺化測驗」。

(一) 確立測量的整體目標

一般而言，空間能力測驗可分為空間關係 (spatial relation) 與空間視覺化 (spatial visualization) 兩類測驗，前者包含三度空間或二度空間的旋轉，後者則包含紙版測驗、平面圖測驗等 (Pellegrino, & Kail, 1982)。此外，空間關係測驗與空間視覺化測驗的特性亦不相同，前者因試題的內容特徵較為簡單，故較重視答題的速度；後者因試題的內容特徵較為複雜，故較強調解題的正確性。

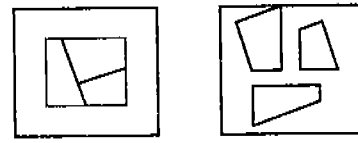
本研究所採用的主要是針對空間能力測驗中「二度空間視覺化作業」 (spatial visualization task) (Pellegrino, Mumaw, & Shute, 1985) 編製有關試題。所編製出的試題要能符合下列 5 項標準：(1) 個人在某一試題上的表現必須與該試題的內容特徵有關；(2) 試題的內容特徵具有理論假設；(3) 試題的內容特徵必須容易操弄；(4) 作答者在不同試題特徵上的表現不同；(5) 不同作答者在相同試題特徵上表現不同。

(二) 確認作業的設計特徵

本研主要以 Pellegrino, Mumaw 和 Shute (1985) 所發展的紙版測驗為藍本。該作業包含錯置、旋轉、錯誤等試題內容特徵。該測驗的典型試題如圖三。受試者在作業中需判斷右圖的幾何圖形能否拼湊成左圖的完整圖形。在答題的過程中答題者需經過編碼 (encoding)、比較 (comparison)、搜尋 (search)、旋轉 (rotation) 以及決定 (decision) 等心理歷程。

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 (右圖的三片幾何圖形中至少有一片與左圖不符)，作答者常常會在發現該片不符的幾何圖形

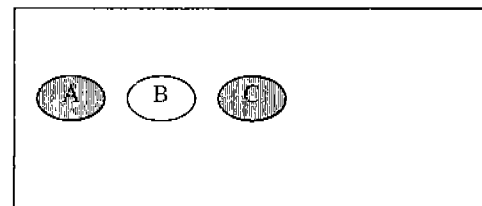
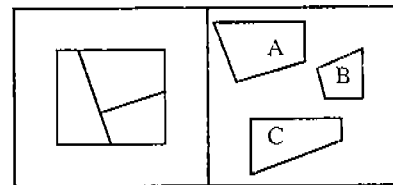
後採用「自我終結歷程的策略」 (self-terminating



圖三 Pellegrino, Mumaw 和 Shute (1985) 所發展的紙版測驗中的典型試題

processing strategy) (p.54)，亦即不再對剩餘的幾何圖形進行認知操弄。此一現象導致剩餘幾何圖形的試題內容特徵無法引出作答者相對應的答題歷程，使得作答者答題的心理歷程是隨意且無法控制的，進而使表徵認知成分的試題內容特徵無法有效說明各個試題的難度值，因而降低各個試題內容特徵的解釋力。

為避免此一現象的發生，本研究對上述試題形式加以修正。本研究的典型試題如圖四。「二度空間視覺化測驗」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。



圖四 本研究中的典型試題

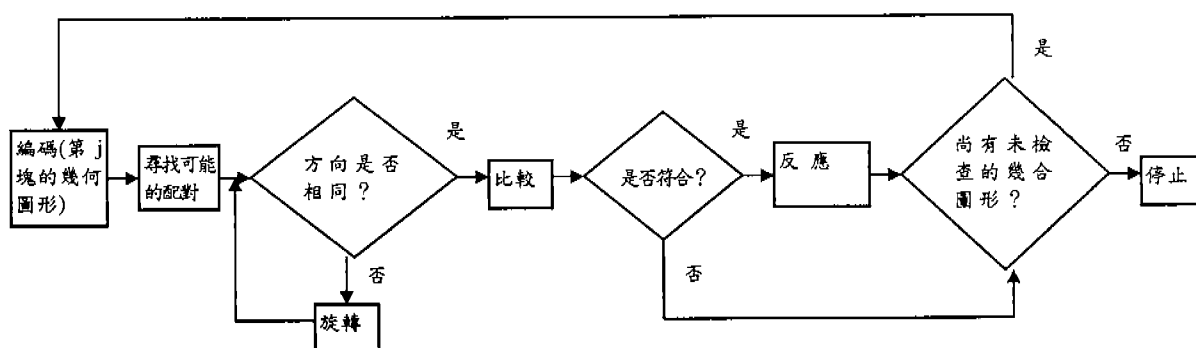


在圖四中，右側三個幾何圖形中的 A 與 C，與左側題幹中所分割的圖形相同，故正確答案是 A、C。

(三) 建構試題之答題歷程的認知模式

本研究主要參考 Pellegrino 和 Kail (1982)、Pellegrino, Mumaw 和 Shute (1985) 等人的理論及研究結果來建構本測驗中解題歷程的認知模式。上述文獻中值得注意的是 Pellegrino 等人曾針對二度

空間視覺化作業建構出答題歷程的認知模式 (Pellegrino, & Kail, 1982; Mumaw, Pellegrino, 1984; Pellegrino, Mumaw, & Shute, 1985)。但由於該模式所根據的二度空間視覺化作業因無法避免作答者使用「自我終結歷程策略」，而使所建構之認知模式無法解釋解題歷程。為避免此一現象發生，本研究除修正原試題之作答形式外，並據之修正解題歷程的認知模式。修正後之解題歷程的認知模式如圖五。



圖五 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。

(四) 決定所欲操弄的試題內容特徵及其複雜度

本研究所欲操弄的內容試題特徵可分為二類：

1. 與認知模式中的認知成分相對應的試題特徵。此意味著操弄試題特徵也就是在操弄作答者答題時的認知歷程。在本研究中此類的試題特徵包含：有無錯置、旋轉的片數、錯誤的片數與錯誤的類型等。

2. 與認知模式中的認知成分無直接關係的內容特徵。操弄此類試題特徵，雖然對答題時的認知歷

程並無直接的影響，但會改變試題的表面結構，而使試題更多樣化。在本研究中此類的內容特徵包括：幾何圖形的形狀等。

基本上，本測驗的編製只對上述第一種內容特徵進行結構性的操弄，而利用「試題輔助產生引擎」在產生試題時對第二種內容特徵進行隨機變化。另外，二度空間視覺化測驗的作答需要作答者完全集中注意力，若試題太過複雜，容易使作答者感覺不耐煩，因而導致情緒因素的介入。由於本測驗的整體目標是在測量空間能力，因此在試題的編製上並不考慮過份複雜的試題，以避免作答者的情緒干擾。故本研究將切割的片數設定為 3 片。

基於上述考量，本測驗的編製對第一種內容特



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，計畫產生 32 題試題（ $2 \times 4 \times 4$ ）。另外，本研究計畫產生 8 題錯誤類型為鏡射的試題，以供日後連結 (linking) 之用。因此，本測驗包含 40 題試題，各試題的內容特徵如表二。

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。

(五)產生設計規格相符的試題

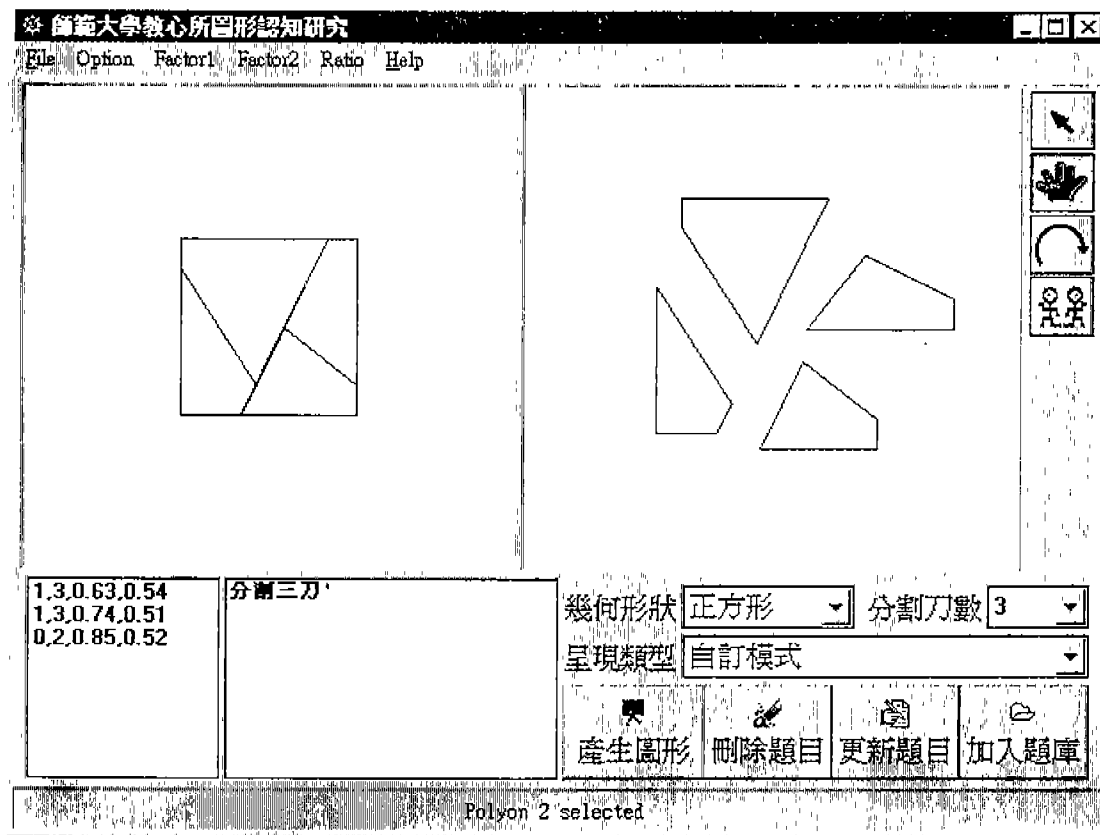
本步驟主要是發展「試題輔助產生引擎」來產生設計規格相符的試題，並進而形成測驗。之所以稱之為「試題輔助產生引擎」，是因為本階段「試題產生算則」尚待建立，「試題自動產生引擎」無法根據算則自動地產生合乎規格的試題。據此，「試題輔助產生引擎」在本研究中有下列階段性的任務。(1)擔任輔助命題工具：本研究在階段一需參照認知設計系統的程序架構，編製大量的試題，並且有效操弄所欲研究的試題內容特徵，期望經由實證研究來建立「試題產生算則」。因此，此一階段中「試題輔助產生引擎」所擔負的任務在輔助研究者有效地操弄試題內容特徵、有系統地設計出試題、簡易地修改與操弄試題，且方便列印成試卷。(2)當作「試題自動產生引擎」的測試版：藉由「試題輔助產生引擎」的應用結果，可清楚地瞭解本系統不成熟之處，可藉此修正改進。

本研究依照前述試題內容特徵與複雜度的設計上需要，設計試題輔助產生引擎（如圖六），藉以產生 40 題試題之「二度空間視覺能力測驗」。目前試題輔助產生引擎的運作，可依據指令自動將左側的完整圖形分成 $n(n=1,2,\dots,7)$ 片形狀大小不同的幾何圖形，並在右側形成相同的圖形。之後，試題編製者必須利用手動在電腦上作業，逐題依照實驗設計對右側的圖形進行操弄（錯置的有無、旋轉的片數、錯誤的片數以及錯誤的類型），逐題產生，並逐題儲存。待按照實驗設計產生了 40 題試題之後，試題輔助產生引擎可將所產生的試題以隨機方式排列並以題本的形式輸出。

(六)將認知模式轉換成心理計量模式

本研究將認知模式轉換成 LLTM。其中，LLTM 中的難度值乃以錯置的有無、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。其模式如方程式(1)。





圖六 試題輔助產生引擎的圖示

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

$P(X_{ij} = 1 | \theta_j, \eta_m, d)$ 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

(七) 考驗 LLTM 以及估計試題的認知特性

本研究所得之資料分別以 BILOG 3.08 與

LINLOG 程式處理之。實際的估計結果，將在研究結果該節中說明。

二、實際研究樣本

本研究以國立台灣師範大學一至四年級的學生為取樣對象。研究樣本包括心理與輔導學系、英文系、國文系等，共計 344 名。扣除無效樣本（答案卡填寫不清，無法以電腦閱卷者）以及全對者（因本資料採 LINLOG 估計，涉及 CML 估計法，要求作答者的反應組型不得有全對或全錯的現象）5 名，因此實際分析的樣本為 339 名。



研究結果

本部分將分成：(一)試題輔助產生引擎的運作；以及(二)LLTM的估計結果等兩個部分來分別說明。

一、試題輔助產生引擎的運作

本研究的試題編製者再熟悉試題輔助產生引擎的操作方式之後（熟悉的過程約花費15分鐘），依據設計中所欲操弄的試題特徵與複雜度，藉由試題輔助產生引擎總共產生53題試題。其中，有13題不合規格，其餘40題合乎規格的試題，所花費的時間約95分鐘。平均每產生一題合乎規格的題目需花費2.4分鐘。在上述13題不合規格的題目中，有4題是試題編製者操作的失誤，有9題是試題輔助產生引擎本身所產生的錯誤。在剔除測驗編製者的操弄錯誤之後，試題輔助產生引擎本身所造成的失誤率約0.18（49題中有9題不合格），亦即約每產生5題試題，會有1題不合格。

上述9題試題之所以不合格，是因為自動分割出來的3片幾何圖形彼此大小差異過大。在本研究中，希望控制幾何圖形大小差異的程度，以避免因差異的程度而影響作答者的判斷。至於其他所欲操弄的試題內容特徵（錯置、旋轉、鏡射等）皆能有效操弄。

至於將所設計出的題庫編印成試卷，本系統提供指定排序或隨機排序的方式。在本研究中是採隨機排序的方式，按照所需的規格，列印成題本。

綜合上述試題輔助產生引擎的實際運作結果，發現試題輔助產生引擎確實能有效地操弄試題內容特徵、有系統地設計出試題、簡易地修改與操弄試題，且方便列印成題本。

二、LLTM的估計結果

將上述40題合乎規格之試題的施測結果，以Linlog進行LLTM的估計時，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。應用Bilog程式進行Rasch模式估計，估計結果顯示整體模式考驗未達.01顯著水準（ $\chi^2=257.4$, $df=216$, $p=.0282$ ），表示資料符合單參數模式。另外，各試題估計的b參數平均數為-1.842，標準差為.801。詳如表二。

再利用Linlog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（ $\chi=785.97$, $df=34$ ，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.48。針對此一結果，進一步以SPSS進行繪圖分析，發現(1)當旋轉的片數超過2片時，試題的難度並不一定會隨著旋轉片數增加而提升、(2)當錯誤的片數超過2片以上時，試題的難度並不一定會因為錯誤塊數的增加而提升。此一結果與Pellegrino的研究發現近似。

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題一題），錯誤類型為鏡射者共有2題，故挑選後共有9題。直接利用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（ $\chi^2 = 7.13$, $df = 5$ ，未達 .01 顯著水準），且 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 .933。進一步地以巢狀模式 (nested model) 進行分析（如表三），發現上述基本參數的加權值皆達顯

著。其中，錯置（有、無）的加權值為 0.63（ $\chi^2 = 24.86$, $df = 1$ ，達 .01 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 0.81（ $\chi^2 = 29.27$, $df = 1$ ，達 .01 顯著水準），錯誤的片數（0,1片）的加權值為 0.57（ $\chi^2 = 15.03$, $df = 1$ ，達 .01 顯著水準）與錯誤的類型是否為鏡射的加權值為 0.70（ $\chi^2 = 15.57$, $df = 1$ ，達 .01 顯著水準），而截距項為 -1.24740。

表二 「二度空間視覺能力測驗」試題分析摘要表

	傳統分析		BILOG Rasch			Ling	
	P	r_{bis}	b	c^2	df	RASCH b	LLTM b
1	.75	.54	-1.337	10.30	7	.28	.49
2	.54	.21	-0.197	21.50	8	-.17	1.59
3	.93	.76	-3.077	5.20	2	.00	1.20
4	.66	.53	-0.828	8.20	7	.39	.98
5	.74	.53	-1.299	4.70	6	.47	.53
6	.89	.50	-2.569	1.90	4	-.41	-.71
7	.74	.46	-1.261	9.50	7	-.06	.57
8	.95	.49	-3.420	.30	4	-.64	1.53
9	.60	.40	-0.481	12.70	8	.22	1.32
10	.86	.60	-2.245	13.00	5	.05	-.39
11	.90	.44	-2.681	6.20	5	-.36	-.82
12	.86	.49	-2.158	7.40	6	-.02	-.31
13	.91	.77	-2.802	8.20	3	-.32	-.93
14	.76	.51	-1.415	2.30	7	-.30	.42
15	.74	.55	-1.243	7.80	7	.16	.58
16	.86	.66	-2.245	16.10	4	.11	-.39
17	.77	.48	-1.454	5.60	7	.41	.38
18	.83	.61	-1.921	3.20	5	.34	-.07
19	.78	.53	-1.536	5.90	7	.11	.30
20	.78	.58	-1.536	3.10	5	.00	.30
21	.88	.48	-2.399	1.70	5	-.47	-.54
22	.60	.18	-0.496	21.30	8	-.08	1.30
23	.92	.62	-2.888	1.80	4	.24	1.02
24	.81	.44	-1.754	5.20	5	-.35	.09
25	.89	.74	-2.500	11.70	3	.07	-.64
26	.61	.29	-0.511	10.20	8	.17	1.29
27	.84	.51	-1.971	1.90	6	-.15	-.12
28	.90	.51	-2.606	3.40	5	-.25	-.74
29	.86	.57	-2.245	1.60	5	-.13	-.39
30	.81	.57	-1.731	5.00	5	.58	.11
31	.92	.49	-2.979	.30	5	-.42	1.11
32	.85	.60	-2.075	7.80	5	.09	-.23
34	.90	.58	-2.681	3.30	5	-.49	-.82
35	.63	.43	-0.651	4.50	8	.45	1.15
36	.81	.28	-1.730	9.50	7	-.18	.11
37	.78	.58	-1.536	2.40	6	.33	.30
39	.82	.55	-1.824	4.20	6	.56	.02
40	.80	.48	-1.708	8.40	6	-.23	.13



表三 9題試題的 LLTM 之巢狀模式中各模式的適合度及增加量

基本參數 (試題內容特徵)			模式適合度		增加量	
錯置	旋轉	錯誤類型	χ^2	df	χ^2	df
✓	✓	✓	5.26	4		
	✓	✓	17.66**	5	12.40**	1
✓		✓	33.28**	5	28.02**	1
✓	✓		19.53**	5	14.27**	1

** .01代表達顯著水準

上述研究結果，可建構出下列難度的預測方程式

$$b_i = 0.63C_{1i} + 0.81C_{2i} + 0.57C_{3i} + 0.70C_{4i} - 1.24740 \quad (2)$$
 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)，便可預估某試題的難度值。以圖四中的典型試題為例，該試題中有錯置、有一片旋轉、有一片錯誤、錯誤的類型是鏡射。因此該題的內容特徵矩陣是〔1,1,1,0〕，帶入式(2)可得預測難度值0.0626。

結論與建議

本研究主要是以二度空間視覺化能力測驗為特定領域就認知測量整合模式中的認知設計系統與試題輔助產生引擎加以研究。以下將對研究結果的利弊得失進行討論，並提出建議。

(一)本研究以「試題輔助產生引擎」來擔任輔助命題工具，確實能有效地輔助測驗編製者編製測驗。其不但能縮減測驗編製者命題的時間，亦能根據所要操控的因子，有系統地產生試題，並且依指令形成測驗。上述優點在目前紙筆測驗的編製上對於結合認知心理學與心理計量，以及建立題庫等皆有幫助。

(二)本研究以「試題輔助產生引擎」當作日後自動化之試題自動產生引擎的測試版，發現「試題輔助產生引擎」在將左側正方形分成數片幾何圖形時，有時會產生各幾何圖形大小差距過大的現象。此一現象的產生有可能會影響試題的難易程度。雖然，此一現象出現的機率並不高（平均約每產生五題試題，就有一題試題如此），在本研究中以半自動方

式產生試題的狀態下，可輕易地以篩選的方式來解決。但在未來「試題自動產生引擎」以自動化的方式運作時，此一現象必須嚴密監控，不得發生。因此，「試題自動產生引擎」在未來發展尚需解決此一問題。

(三)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）與錯誤的類型是鏡射或是形狀不一致時，上述內容特徵是該試題難度的有效預測變項，有助於建立試題產生算則，以供反應產生系統針對特定作答者產生特定試題內容特徵與難度的試題。

(四)經由 Bilog 的估計結果發現，第 38 題的試題過於困難。經過重新檢討試題後發現這一題的試題內容特徵包含：無錯置、一片旋轉、兩片錯誤且錯誤類型為形狀不一致者。其中，由於在操弄錯誤時，形狀改變的程度並不大，導致作答者在答題時容易誤判。此一結果說明了：當所錯誤的類型為形狀不



一致時，形狀不一致的程度會影響試題的難度值。另外，尚有不易精準地操弄試題形狀的不一致、不易利用試題輔助產生引擎來控制試題形狀不一致的程度等客觀因素。因此，後續研究是否仍採用「形狀不一致」作為錯誤的類型，有必要進一步地探討。

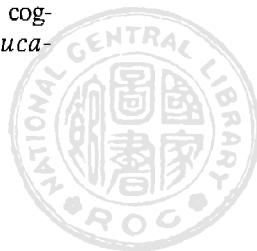
(五)雖然本研究修正 Pellegrino 等人所建構之認知模式，改變作答的形式，希望能促使答題者必須處理所有的幾何圖形，而不會中途終止。但研究結果發現，旋轉與錯誤對試題難度的影響並不會隨著

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。之所以如此，是因為作答型式的設計不佳或有其他原因（例如：試題的內容特徵間有交互作用），尚待進一步的分析。

(六)本研究所採用的受試皆為師大學生，使作答結果的變異性並不大。後續研究有必要以高中生、國中生作為施測的對象，以蒐集所有可能的反應組型。

參考文獻

- 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。
- 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). Hillsdale, 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). Washington, 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: Plenum 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.
- Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore, MD: University Park Press.
- Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff.
- Haywood, H. C., Brown, A. L. & Wingenfeld, S. (1990). Dynamic approaches to psychoeducational assessment. *School Psychology Review*, 19(4), 412-422.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12, 252-284.
- Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, 64(4), 575-603.



- Pellegrino, J. W., & Kail, R. V. (1982). Process analysis of spacial aptitude. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 1; pp.311-365). Hillsdale, NJ: Lawrence Erlbaum Associates.
- 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. (1984). *Advances in the psychology of human intelligence*. Hillsdale, NJ: Erlbaum.
- 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.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Warm, T. A. (1978). *A primer of item response theory*. Springfield, VA: National Technical Information Service.
- 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.

收稿日期：86年7月8日

修正日期：86年10月14日

接受日期：86年10月15日



The Operation Research For Cognitive Design System And Item-Generation Assistant Engine: An Implementation Case In A Two Dimension Spatial Visualization Ability Test

Sieh-Hwa Lin Tzu-chien Liu Steven Liang

National Taiwan Normal University

Center University

Abstract

To implement the integrated cognitive assessment model that Lin and Liu(1997a) proposed, the study focuses on using the basic ideas in cognitive design system and item-generation assistant engine in developing test of two dimension spatial visualization ability, to achieve the two goals: (1) applying the procedural framework of cognitive design system (Embretson, 1994) to develop two dimension spatial visualization ability test, (2) designing test-generation assistant engine to help researcher to effectively control item characteristics, to systematically write items, to conveniently modify items and print test form. Based on the results of the research, an improvement of the idea framework of integrated cognitive assessment model has been made.

Keywords: cognitive measurement, LLTM

