

近代測驗設計

****鄭春蕓 *楊文金**

****國立嘉義師範學院數理教育學系**

***國立臺灣師範大學科學教育研究所**

壹、前言

認知心理學和教育及心理測驗，二十世紀後，各自發展，成為不同的領域。而近年來，教育測驗已有顯著的進步，經由試題反應模型，使測驗更具客觀性，對診斷與改進教學，更為有用。同時，認知心理學的認知訊息處理（Cognitive information Processing, CIP），可以分析測驗試題，瞭解答題時的心理過程以及所需的內容知識。

由於認知心理學和心理計量學的蓬勃發展，彼此互相衝擊，而使認知心理學對教育測驗的影響，表現更重要的角色，兩者密切聯合，成為近代測驗設計。對於認知的工作，包括學校裏的學習、教育測驗及心理測驗等，都能更瞭解其中的心理過程以及隱含於成就測驗背後之知識內容結構。

傳統的教育與心理測驗，大多注意受試者分數的變異量，而試題的變異量則受忽略。但近代的測驗設計，同時分析試題難度來源與受試者的變異量。所以本文介紹的近代測驗設計，不再視試題為一基本單位，而是可以再分解為更基本的認知工作。依次介紹分析試題為不同層面的層面理論、分析試題為基本認知成分的認知成分分析，以及計量學模型。

貳、層面理論

層面理論是 Guttman 根據接近律（principle of contiguity），假設“在概念定義上比較接近的項目，在實際經驗上也會比較接近。”層面理論基本上是一種後設分析（meta analysis），其研究方法需要根據先前的研究或先驅調查等，才能對研究主題之主要層面有確定的概念化，再做層面設計的分析。以下分別討論層面的定義及組成、層面理論的研究方法、層面化測驗與試題的難度。

一、層面的定義及組成

層面的定義根據 Shye (1978 ; from Brown, 1985) 的數學觀點，一個層面就是一個集合，實際上的意義就是對隱含於一組觀察的背後，作一個概念的歸類，將之標記為一個層面。層面研究要對層面進行一系列更清晰和精確的界定，就是系統化的界定層面的概念向度及組成元素。例如 Payne 等人 (1976 ; from Brown, 1985) 定義工作滿意度和組織氣氛的三個層面 A, B, C 及每一層面內的兩個元素：

層面 A 使用的分析單位

a₁ 個別的

a₂ 社會的集合體

層面 B 包含的分析成分

b₁ 工作

b₂ 組織

層面 C 概念的本質

c₁ 情意的

c₂ 敘述性的

上述例子，說明層面是由研究者提出，層面的組成元素是描述層面內的不同類別，關於層面與元素的界定，Runkel 和 McGrath (1972 ; from Brown, 1985) 提出下列幾點條件：

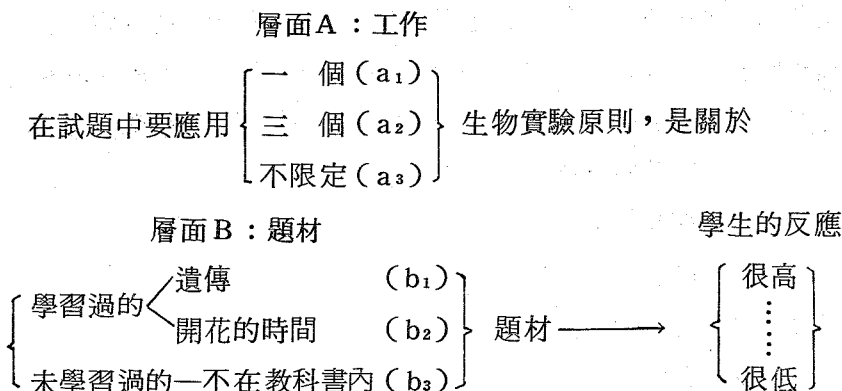
1. 文獻中的有關項目被歸類到這個領域的所有層面。
2. 每個層面包含盡可能有的元素。
3. 元素之間互相排斥。
4. 必須界定層面之間的邏輯關係。
5. 層面的集合要包含這個領域所關心的所有層面。

二、層面理論的研究方法

層面理論的研究方法分為兩個部分，一為層面設計，一為層面分析，分別說明如下：

1. 層面設計：一個層面就是實驗設計的一個變項，其中的元素就是這個變項的水準，例如算術測驗可設定第一個層面是加減乘除等四個元素（四則）的運算，就是四個水準。而第二個層面可設定為一位數、二位數、三位數等三個元素，也就是三個水準。有了層面和元素之後，再根據層面和元素來寫組合句。組合句是用來系統化地編寫試題，因為組合句有共適的聯合部分，相互間只有一個層面的元素不

同，所以能夠產生多樣而平行的測驗形式，例如 Blum (1989) 對 Bloom 的認知領域目標的“應用”作層面分析，分為兩個層面，工作和題材。至於學生的可能反應（即成就測驗的等級），則用箭頭分隔，以生物的一個例子，圖示如下：



在這例子中，可以產生的組合句題型 (structuples) 共有 9 個 ($a_{1-3} \times b_{1-3} = 9$)，即 $a_1b_1, a_1b_2, a_1b_3, a_2b_1, a_2b_2, a_2b_3, a_3b_1, a_3b_2, a_3b_3$ 等各種題型，而 Blum 寫了 24 道題，其中題型為 $a_1b_1, a_1b_2, a_2b_1, a_2b_2, a_3b_1, a_3b_2$ 各兩題，而 a_1b_3, a_2b_3, a_3b_3 則各有 4 題，使學習過的題材和未學習過的題材，各有 12 道試題，以比較不同題材的空間圖形之關係。

- 層面分析：層面之間的關係，一般是用 Guttman 發展的 SSA 程式 (Smallest Space Analysis) 來分析，對於實驗數據可以得到完整的空間圖形，圖 1 為 Blum 所得的空間圖形，每一個變數 (如上述例子的 a_1b_1, a_1b_2 等)，用多向度空間 (multidimensional space) 的一個點來表示，在空間越靠近的兩點，所代表的變數間之相關增加，因此可以作直接的解釋，並有利於整體的觀察。同時多向度空間圖形，被轉換為數個二度及三度空間的圖形，以觀察將某一層面固定不變而得到的分層結構。例如 Seddon (1978) 對 Bloom 的認知分類，將試題分為知識、理解、應用、分析、

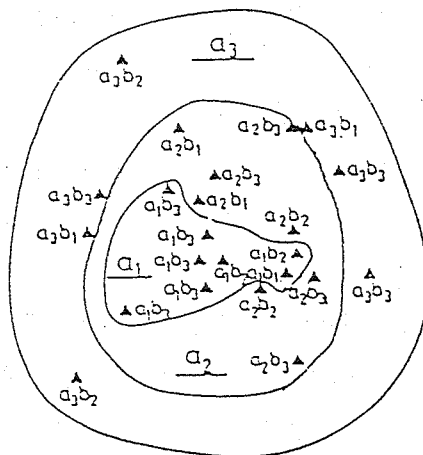


圖 1 由 SSA 繪出的空間圖形，曲線分隔試題工作的不同元素。(a_1, a_2, a_3) (From : Blum, 1989)

綜合、評鑑等六個階層，就是六個變項，得到圖形分爲下列三種：

(1) 線構 (simplex)：

一組變數的領域內容相同，只有複雜度不同，而得到直線序列的圖形。

如圖 2 的 1, 2, 3, 4, 5, 6 序列，代表認知的六個階層：知識、理解、應用、分析、綜合、評鑑。

(2) 圓構 (circumplex)：

一組變數的複雜度相同，但內容不同，例如生物和化學都有“理解”

的試題，理想的圖形，會落在圖 2 的第 2 圈之同心圓。

(3) 徑構 (radex)：

一組變數的內容和複雜度都不相同，其理想圖形爲圖 2 分隔的 A, B, C, D 區域，代表不同的領域，二變數的點不在同心圓的範圍裡。

三、層面化測驗與試題的難度

層面理論認爲構念可以被描述爲它的成分，前面提到的層面設計，就是由有限和有序的成分，來組織和定義這個構念的領域。例如 Blum 將“應用”這個構念，分爲工作和題材兩個成分，在工作層面，其成分難度隨需要利用的生物實驗原則之數目增加而增在，在題材方面，未學習的題材之成分難度高於學習過的題材。

Van de Vijver (1988)，根據層面理論設計試題，試題的難度可以用隱含於試題背後的層面加以描述，例如字母測驗的三個層面是母音字母的個數、字母群的產生法則，以及在第幾個字母開始變化。再根據 LLTM (Linear Logistic Trait Model，於第肆部分介紹) 估計每一層面各元素之難度。Van de Vijver 將受試者也當作一個層面，研究跨國文化的差異性，這個層面包括荷蘭人、土耳其人、尚比亞人，以比較文化的差異性。

對於 Gavurin (1967；from Snow & Peterson, 1985) 的拼字測驗，是對同一個人的層面測驗，對於空間能力高的人，心理操作比動手操作優越，但空間能力低的人，心理操作比動手操作差。由這個層面的設計，可以帶出成分過程的效應。此外，測驗的格式、測驗的情境，都可以設計爲層面化測驗的不同層面，提供個別差異與試題難

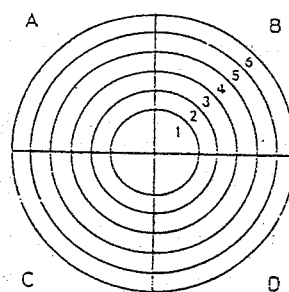


圖 2 線構、圓構和徑構的理想圖形。

線構 (simplex)：1, 2, 3, 4, 5, 6 的直線序列

圓構 (circumplex)：在同心圓的點。

徑構 (radex)：在 A, B, C, D 不同區域，且不是同心圓的點。

(From Seddon, 1978)

度的來源 (Schmitt-Crocker, 1981; from Snow & Peterson, 1985)。

參、認知成分分析

心理計量學研究的個別差異和認知訊息處理可以互補應用，由計量學發現受試者在測驗上的個別差異，再用認知訊息處理作為分析測驗試題的技術。認知成分分析的目的，在於確認隱含於一系列相關的訊息處理背後之心智操作的成分，並發現這些成分的關係與組織。對計量學而言，成分分析使推論的特質精緻化，可視為一個構念效化的詳細法則。而對認知訊息處理而言，認知成分分析可視為一組步驟，用來確定認知的基本成分及發展其組織的方法。

成分分析的基本單位就是成分，是一個基本的訊息處理，是對物體或信號的內在表徵之操作。這個認知成分可能是將感覺輸入轉化為概念的表徵，或將一個概念表徵轉化到另一個概念表徵，或是將概念表徵轉化為輸出的動作，所以認知成分分析與訊息處理分析的關係密切，其基本單位是過程，而不是靜態的表徵個別差異來源之構念。以下介紹其理論和模型，以及內含和外含的分析。

一、認知成分的理論和模型

認知成分的理論包括兩部分：(1)確認完成認知工作所需的成分；(2)界定這些認知成分的組合規則。

理論的不同，可能由於包含的成分不同，例如理論一的成分有 a, b, c 三種，理論二的成分只有 a, b 兩種，理論三的成分則多達 a, b, c, d 四種。

而兩種理論的成分雖然相同，也可能因為成分的組合規則不同，就成為不同的理論。例如理論一是加法的組合規則：反應時間 $(RT) = a + b + c$ ；理論二只是部分的加法組合規則： $RT = a + \text{Min}(b, c)$ ；理論三則不是加法的組合規則： $RT = \text{Max}(a, b, c)$ 。

一個認知成分的理論，並未考慮訊息處理的所有行為，需要由模型來進一步界定：(1)成分執行的次序，以及(2)成分執行的模式。例如一個理論有 a, b, c 三個成分時，其執行次序有 $3!$ 或 6 種之多。至於模式是關於作答的方式：可能是每一個答案都看過或找到答案就自動停止；序列進行或平行進行的；以及整體處理或個別處理方式之描述。

認知成分分析對理論和模型以及模型的模式之觀點列於表一（見下頁）。

二、內含和外含的分析 (intensive and extensive analysis)

認知成分分析最先獲得的是代表顯著特質 (manifest trait) 的“組合作或測

表一 認知成分分析的理論、模型、模式之觀點

理 論	確認認知成分	界定成分的組合規則
模 型	規定成分的執行次序	規定成分的執行模式
模 式	關於執行的描述	規定執行的處理方式

驗分數”，顯著特質是可觀察的行為，但認知成分分析所關心的是隱含於行為背後之“潛在特質”，潛在特質就是以成分分數為代表。在顯著與潛在兩個水準之間，有一個中間的水準，是以區段分數為代表。而建立區段分數和成分分數，是屬於內含的工作分析。集合這些內含分析，成為外含工作分析的基礎。內含工作分析是要盡可能地瞭解完成一項工作或測驗背後之認知成分，而外含分析的目的是要統整一系列相關的內含工作之發現。因此成分分析包括兩個部分，內含和外含的分析。

內含分析包括內部和外部的效化，內部效化是要將工作或測驗分解為隱含於背後的成分；外部效化則是要瞭解界定的認知成分，如何對外在工作或測驗產生個別差異。

在分解工作或測驗時，第一步先建立區段分數，就是將工作分解為一系列的次工作，每一個次工作比完整的工作，需要較少的訊息處理。通常的測驗，分為兩部分，第一部分是前示（precueing），第二部分為解題（solution），前示分數是對第一部分的反應；解題分數是對第二部分的反應，這兩種分數，都稱為區段分數，因為它們代表處理整個測驗工作的某一個特定區段。

區段分數的內部效化，是加成性的模型，假設區段分數是反映較少的訊息處理，而被包含於反映較多訊息處理的分數裡。這個假設是用線構模型（simplex model）來測試，所以線構模型提供區段分數的內部效化。至於區段分數的外部效化，是區段分數和參照能力測驗（reference ability test）的分數之相關。假如已建立良好的測驗或其他成就的效標與區段分數測量相同的能力，例如區段分數為類比推論時，則可選擇歸納推論作為參照的能力測驗。且要測量輻湊效度（convergent validity）和區別效度（discrimination validity），前者是與有關的測量有高的相關；後者是與無關的測量低相關。

至於成分分數是估計個別受試者的參數，與測驗分數、區段分數有下列三點不同：(1)測驗分數是測量公開的或顯著的特質，區段分數可視為等同於測驗分數，而成分分數則是測量潛在的特質。(2)測驗分數包含不同基本的訊息處理之混合，區段分數可以是不同基本的或不等數目的基本之訊息處理，但成分分數則只測量一個基本的訊息處理之時

間或難度。(3)測驗分數不需依賴一套測量行為的理論，區段分數則受加成性假設的限制，而成分分數是根據心理學的訊息處理理論，成分分數將隨理論或模型的改變而變。

至於成分分數的內部效化就是成分模型，以適合度考驗來評估模型，例如解釋變異量的百分比，而以最合適的模型用來做為計算成分分數的基礎。而成分分數的外部效化與區段分數相同，利用參照能力測驗分數，計算其相關，同時注意輻湊和區別效度。

對於成分分析，參照能力可看做成分的組織，因為每一個能力，不限一個測驗，參照的能力測驗就可以做為外部效化，將這些能力分解為隱含於背後的成分，並用成分分數之個別差異，來說明受試者在參照能力之個別差異。也就是受試者之間的變異量是用成分的組合來說明。

外含工作分析是由一系列相關的內含工作分析所組成，其主要目的是經由一系列的輻湊操作，以界定其內含分析的認知成分是心理學的實體 (psychological reality)，並將一個工作或測驗通則化到其他相關的工作或測驗，例如關於類比推論的外含分析之策略，將下列五種工作或測驗的變數，選擇作系統性變化，包括：(1)因變數：反應時間或正確百分率；(2)反應的格式：是非或選擇題；(3)內容的型態：語文或非語文；(4)難度：很容易、中等難度、很困難；(5)屬性的定義：定義明確 (well defined)、部分定義明確 (somewhat defined)、定義不明確 (poorly defined)。由上述這幾種系統的變化，例如將因變數由反應時間改變為正確百分率，這種輻湊的操作，就是外含工作的分析，將一個工作或測驗應用到其他工作或測驗，作通則化的驗證，而使內含分析的認知成分被接受為心理學的實體 (Sternberg, 1977)。

肆、計量學模型

在前一節認知成分分析所討論的，一道試題或工作，不只需要一種能力，而是需要一組的能力，所以可以將試題或工作分解，有兩種數學模式的分解方法，第一種是複雜度因素 (complexity factor)，第二種是次工作反應 (subtask response)，說明如下：

複雜度因素是將試題難度分解為幾個複雜度因素，例如圖3的幾何類比圖形，有下列兩個重要的訊息處理因素：

- (a) 編碼的複雜度 (encoding complexity)，是根據A的元素之數目，在這個例子，A有三角形和直線兩個元素。
- (b) 轉換的複雜度 (transformational complexity)，根據由A轉換到B所需

要轉換的數目，在這個例子，由A到B需要三個轉換：①外面元素的形狀改變；
②裡面元素數目的增加；③裡面元素90°的旋轉。

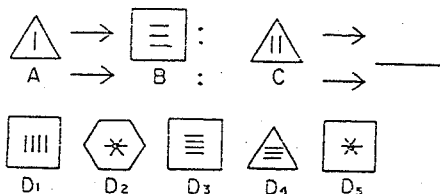


圖3 幾何圖形的類比測驗

(Whitely & Schneider, 1981, from Embretson, 1983)

每一題試題根據這兩個複雜度因素，計算其難度，其數學模式為：

$$P_i = \eta_1 q_{1i} + \eta_2 q_{2i} + a \dots\dots\dots(1)$$

P_i ：第 i 題的難度

q_{mi} ：第 m 個複雜度因素對試題難度的加權 (weight)

a ：常數

次工作反應是將試題或工作分解為一系列的次工作，例如表二的試題分為兩個次工作：規則建構與反應評估。

表二 語文推論的次工作

試題：
貓：老虎 :: 狗：_____
(a)獅子 (b)狼 (c)吠聲 (d)小狗 (e)馬
規則建構：
貓：老虎 :: 狗：_____
規則 _____ ? (規則：大的或野的犬科)
反應評估：
貓：老虎 :: 狗：_____
(a)獅子 (b)狼 (c)吠聲 (d)小狗 (e)馬
(From : Embretson, 1983)

兩個次工作成分的數學模式，預測答對全題的或然率為：

$$P_{iT} = P_{i1} P_{i2} \dots\dots\dots(2)$$

P_{iT} ：答對第 i 題的或然率

P_{i1} ：答對“規則建構”的或然率

P_{i2} ：答對“反應評估”的或然率。

現在心理計量學的方法，參照上述數學的模式，認為試題反應是來自於比較基本的理論構念，並視構念為“潛在特質”(latent trait)，是試題之間的一般因素，由主要向度的反應一致性來確定。上述數學模式只討論試題的參數，而心理計量學的模式包含試題和受試者兩個參數。以潛在特質預測受試者，正確解試題 i 的機率，在拉响模型(Rasch model)中，通過的機率決定於試題難度 b_i 和受試者能力 θ_j (代表受試者在隱含於測驗成就背後之潛在特質的位置)：

$$P(X_{ij} = 1) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} \dots\dots\dots (3)$$

方程式(3)有兩點重要的特點：①通過的機率 P 與參數的關係，包含在指數中，不是線性的。一個試題的通過機率對受試者能力之特性曲線，是類似 S 形的肩形(S-shaped ogive)。潛在特質的優點，在於由任何一組校正的試題，可能得到受試者可比較的能力估計。

拉响模型的兩個強假設為(1)局部獨立性(local independent)，即對於某一試題的反應，是獨立於其他試題。(2)單維測驗(unidimensionality)，只測量單一的潛在特質。不適合上述假設的試題，就要刪除，以增加測驗的效度。因為剩下的試題，對於“潛在特質”將會提供更可靠的訊息。

由拉响模型發展為多成分的潛在特質模型(multicomponent latent trait modeling)，由試題或工作分解方法去確認理論的參數，在歐洲和美國分別發展兩個家族的模型，一為線性羅吉特質模型(Linear Logistic Trait Model, LLTM)，在歐洲很有影響，另一為成分潛在特質模型(Component Latent Trait Model, CLTM)，為美國 Embretson 提出。本文介紹線性羅吉特質模型(LLTM)：

LLTM 是用複雜度因素的方法，由代表試題結構訊息的試題分數來確認試題的理論變數。LLTM 也包含受試者的參數 θ_j ，可以用來測量個別差異。但必需注意的是，LLTM 雖然具有多成分的處理變數，但它是單維的模式，只包含單維的能力參數。

LLTM 需要收集下列的數據：(a)對整個試題的反應；(b)對每一試題的一些變數之分數，這些變數代表理論的複雜度因素。例如圖 3. Mulholland et. al. (1980; from Embretson, 1983) 對每一試題的兩個變數計分，這兩個變數代表訊息處理的結構，即 A 的元素數目，和由 A 到 B 的轉換數目。用數學式子來表達，如式(4)的線性模式，試題難度是用複雜度因素 m 加以類化：

$$b_i = \sum_m \eta_m q_{im} + a \dots\dots\dots (4)$$

b_i ：第 i 題試題難度

q_{im} : 第 i 題操作複雜度因素 m 的數目

η_m : 處理操作複雜度因素 m 的難度

a : 正規化常數 (normalization constant)

所以 LLTM 的試題難度是訊息結構分數 (information structure scores), q_{im} 的線性函數。

考慮個別差異, 合併式(3)拉响模型和式(4), 成為完整的 LLTM 模型, 如式(5)。

$$P(X_{ij} = 1) = \frac{e^{(\theta_j - (\sum_m \eta_m q_{im} + a))}}{1 + e^{(\theta_j - (\sum_m \eta_m q_{im} + a))}} \dots\dots\dots (5)$$

式(5)中的參數, 可以用條件最大概度 (conditional maximum likelihood, CML) 加以計算。估計的電腦程式和估計的演算, 參見楊文金的論文 (民 80)。

Whitely & Schneider (1981; from Embretson, 1983) 應用 LLTM 研究幾何圖形類比的訊息處理結構, 提出三個模型, 用 χ^2 來比較模型的適合度考驗 (The goodness of fit of model)。根據其中一個, 他們得到的訊息處理結構的模型為式(6):

$$b_i = .01 q_{i1} - .27 q_{i2} + .96 q_{i3} - .73 \dots\dots\dots (6)$$

q_{i1} : 元素的數目

q_{i2} : 變形轉換 (distortion transformation) 的數目

q_{i3} : 位移轉換 (displacement transformation) 的數目。

在圖 3 中, 有兩個元素, 和三個轉換, 其中兩個轉換是變形的 (數目和形狀), 一個轉換是位移的 (90° 旋轉), 以下為成分分解的結果:

$$b_i = (.01) \times 2 + (-.27) \times 2 + (.96) \times 1 = .44$$

因為 LLTM 為潛在特質模型之一, 如同其他心理計量學的模式, 試題難度 b_i 可由任選的試題子集合來得到。而拉响模型, 亦可視為 LLTM 的特例, 就是每一試題只有一個參數時, LLTM 模型就是拉响模型。拉响模型可以說是 LLTM 的一個特例。

伍、結 論

近代測驗設計, 本文所介紹的層面理論、認知成分分析以及計量學模型, 都兼顧受試者和試題的變異量, 當做構念效度的來源, 而傳統測驗只有考慮受試者的變異量。

層面理論的測驗設計, 將試題內容、認知階層、試題格式、施測情況、受試者族群等都以層面加以界定, 找出各層面的元素, 預測試題的難度來源, 再以層面分析的空間

圖形判斷層面間各元素的關係，或以其他統計方法來驗證層面設計的結果，例如 Seddon (1978) 用層面設計考驗 Bloom 的認知階層之分類效度；以及 van de Vijver (1988) 用 LLTM 驗證試題難度可以用隱含於試題背後之層面加以描述。

成分分析是合併個別差異和訊息處理等兩種心理學的研究方法，在理論上，可以綜合二者的優點。成分分析在於確認心智操作的基本成分，這些成分可以分別用不同的限制條件加以組合。如果是個別差異的限制，則以受試者的差異為主，是外部效化的問題。如果是訊息處理的限制，則以試題之間的差異為主，是內部效化的問題。因此成分分析的內含工作分析包括內部和外部的效化，合併提供測驗的構念效化，並由外含性的工作分析，使認知成分成為心理學的實體。

而計量學的模型，將試題分解，建立可預測試題難度的成分模型。所以在技術方面需要先作試題或工作分解，再以 LLTM 考驗模型的預測能力，例如楊 (民80)，由閱讀理解探討形成假說測驗試題結構，不必使用效標效度，就能保證測驗的構念效度。

因此上述三種測驗設計，可以相互支援應用，對於題庫的建立，測驗的發展與測驗的構念效化，將更有助益。

參考文獻

1. 楊文金 (民80)，形成假說技能試題之結構分析研究。國立臺灣師範大學科學教育研究所博士論文。
2. Blum, A. (1989). Facet theory used to establish taxonomic subclasses for "application" with an example from science teaching. Science Education, 73 (1): 1-12.
3. Brown, J. (1985). An introduction to the uses of facet theory. In D. Canter (Ed.) Facet Theory--Approaches to Social Research, Springer Series in Social Psychology. NY: Springer-Verlag.
4. Embretson, S. E. (1983). Content validity: Construct representation versus nomothetic span. Psychological Bulletin, 93 (1): 179-197.
5. Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. Review of Educational Research, 48 (2): 303-323.

(下接第 59 頁)