

國立臺灣師範大學教育心理學系
教育心理學報, 民 76, 20 期, 131—182 頁

潛在特質理論與其應用於 適性測驗之評估研究

林 世 華

本研究主要的目的在於探討：(一)潛在特質理論與傳統測驗理論二者之間的關係。(二)潛在特質理論應用於適性測驗時，對於心理測量所產生的助益。

研究者使用電腦模擬的方式，模擬受試者對項目的反應，並就適性測驗 θ 已知與 θ 未知的兩種情況模擬反應的過程，從事系列性的追蹤觀察。

結果發現：

一、潛在特質理論的 a 參數、 b 參數及 θ 參數分別與傳統測驗理論的 r_{bis} 、 P 及個人得分等統計數之間均有很高的相關。

二、 θ 已知時的適性測驗路徑，根據 b 參數進行羣聚分析結果可以得到三種類型。這三種類型的路徑，也就是高、中、低三種不同能力水準受試者的路徑。這些路徑主要是依照能力與項目難度相匹配的原則所形成。就測量的功能而言，使用題庫中不到50%的題目所測量的結果與使用所有題目所測量的結果是一樣的。而且隨著題數逐步增加，測量結果也逐漸趨於穩定。

三、對 θ 未知的適性測驗路徑進行羣聚分析的結果顯示：根據 a 參數可以得到兩類，根據 b 參數可以得到三類，根據 θ 參數則可以得到四類。每一種類型的路徑都代表某個能力水準範圍內受試者的路徑。路徑的形成主要還是依照能力與項目難度相匹配的原則。惟，在測量的功能上則呈不穩定的現象；測量的結果並不會因題數的增加而趨於穩定。

根據本研究的結果可以得到本研究的結論是：

一、潛在特質理論與傳統測驗理論二者之基本觀念上是相通一致的。

二、將潛在特質理論應用於適性測驗是可行的，而且可以改善測量工作。但是因為目前使用的方法產生的偏差仍舊相當大，所以在技術上仍有待進一步研究突破。

本文最後研究者根據結論提出，估計 θ 有關技術的改善建議，以利適性測驗的實際應用。

對潛在特質理論的出現，測驗學界是以「革命」性的字眼來加以描述 (Marco, 1977; Warm 1978)。顯然的，這反映出傳統測驗理論遭到困難，無法再滿足測量上的需要。林一真 (民71) 指出傳統測驗理論的兩大困難：第一是在估計項目難度時受到樣本程度干擾。第二是在估計個人能力時受到項目難度影響。這兩個困難在測量工作上會形成下列兩個現象：

(一) 一個測驗工具往往是針對某一特定的受試組羣而編製。如為國中生編製的智力測驗，在高中、在小學都無法直接使用。當然也可以編製一個貫穿小學、國中、高中的智力測驗，但是這個測驗恐怕會長得不切實際。

(二) 兩位受試者的測驗結果要比較，必須做同一個測驗。例如「小學生與大學生的智力誰高？」這個問題就很難直接去回答，因為小學生與大學生，很難有一個二者均適用的測驗。

學測驗的人很快就會發現，上述兩種現象有例外，那就是個別智力測驗——比西量表。但嚴格的

說，比西量表的編製，並不是傳統測驗理論的概念，反而是潛在特質理論的概念。因為不同的受試者，有可能作不同的題目，但結果是可比較的。

測量學者負有改善測量工作的責任。這改善包括使測量更簡便、測量更正確。潛在特質理論的興起，已為此種改善測量工作奠定深厚基礎。適性測驗就是潛在特質理論在實際應用上改善測量工作的具體方式。

上述傳統測驗理論的困難，導致現行測驗出現瓶頸的現象，那就是測驗種類繁複，非但測量未改善，而且使用上有更繁瑣之處。

就是因為潛在特質理論有希望突破傳統測驗理論的瓶頸，驅使研究者進行本研究。

基於上述動機，研究者期望本研究的完成能達成下列四項目的：

第一、了解潛在特質理論的內涵。

第二、比較傳統測驗理論與潛在特質理論之間的關係。

第三、探討適性測驗在潛在特質理論下如何運作。

第四、探討適性測驗在潛在特質理論下如何改善測量工作。

壹、文獻探討

一、潛在特質理論：

(一) 潛在特質理論與項目反應理論

潛在特質理論 (Latent trait theory, 簡稱 LTT) 與另一個名詞，項目反應理論 (Item response theory, 簡稱 IRT)，在心理測量領域中有交互混用的現象，隨着學者的習慣不同，而有所不同，如 Urry (1977), Weiss (1983) 較常用潛在特質理論，而 Lord (1980a), Hambleton & Cook (1977)，等則較常用項目反應理論，然而基本上，這兩個詞所指的東西，並沒有什麼差別。Weiss (1983) 曾指出這兩個名詞的關係，他認為將 LTT 應用於能力測驗或是成就測驗上，便是IRT。IRT 強調的是測驗項目 (test item) 與受試者反應 (response of examinees)。而 Hambleton (1985) 則認為 LTT 與 IRT 並無不同，差別在於 LTT 容易與因素分析 (Factor analysis)，多向度量尺分析 (Multidimensional scaling) 及潛在結構分析 (Latent structure analysis) 等研究潛在架構的方法產生混淆，因為畢竟 LTT 是在研究測驗項目與個人反應之間的關係，所以他較傾向使用 IRT，然而這些說法並未引起心理測量學者們的爭執。事實上 Weiss (1983) 自己也承認混用 LTT 及 IRT 這兩名詞。因此本研究所指 LTT 與 IRT 並無不同。那麼不管使用 LTT 或使用 IRT，這個測驗理論最常使用於能力測驗，尤其是選擇式 (multiple-choice) 能力測驗 (Warm, 1978)，乃是個不爭的事實，因此在本研究中所指之 LTT，亦僅針對 LTT 在能力測驗上應用或 IRT 在能力測驗上之應用而言。或許也正因為 LTT 大部份的研究均用於能力測驗上，所以也就沒有過分去辨別 LTT 及 IRT 的必要。

(二) 潛在特質理論與潛在特質模式

潛在特質理論包含着一組數學模式 (a family of mathematical models) (Hambleton & Cook, 1977; Weiss, 1983)。Lord (1980a) 亦指出 IRT 的觀點是從數學陳述出發來說明個人反應與項目特性的關係，並進而推知個人潛在特質與個人反應的關係，它們之間存在的一種數學函數關係。由於數學函數關係的簡明及其映射關係的模式化，學者們亦常用潛在特質模式 (Latent trait model 簡稱 LTM) 替代 LTT，實際上 LTT 及 LTM 所指亦是完全相同的東西。

(三) 什麼是潛在特質理論

所謂潛在特質理論是用來解釋測驗項目，受試者反應及個人特質三者之間相互關係的一種理論架構 (Hambleton, 1985)。整個理論中包含了一組完整的命題，這些命題是用來說明個人對測驗項目所做的反應，在心理測量上的意義 (Hulin, Drasgow & Parsons, 1983)。以下分別就 LTT

組成之諸元素，逐項加以闡述：

1. 測驗項目：在 LTT 中，測驗項目是一種刺激變項；研究者將這些變項呈現給受試者或刺激受試者，以獲得可被觀察及可被紀錄的反應活動 (Weiss, 1983)。

從實驗心理學操弄刺激變項的觀點來看，測驗學者是根據什麼測驗項目特性 (item characteristics) 來操弄測驗項目呢？也就是說一個測驗設計者是會根據什麼測驗項目特性或怎樣的測驗項目特性來安排測驗呢？在 LTT 中描述測驗特性，觀念上仍就是以項目難度 (item difficulty)、鑑別度 (discrimination) 等指數來描述一個項目的特性。在 LTT 的觀念中，每一個測驗項目均有其「不變的」指數來加以界定。所謂「不變」(invariant) 指的是這些用來描述測驗項目的指標，並不因為樣本不同，而有所改變。因此，在 LTT 中會將這些指數稱為項目參數 (item parameter) (Lord, 1980a) 所採的是一種絕對性的觀點。這顯然與過去傳統測驗理論的相對性觀點不相同。

以下研究者就 LTT 中使用的項目參數，逐一加以探討。在討論項目參數時，我們必須先瞭解 LTT 中另外兩個重要的名詞，那就是項目反應函數 (item response function, 簡稱 IRF) 和項目特性曲線 (item characteristic curve, 簡稱 ICC)，此外，還須瞭解兩個在 LLT 中與個人反應及個人特質之間關係的概念，其中一個是個人對測驗項目，反應答對的機率，通常是以 P 來表示；另一則是個人能力在一個連續線上所在的位置，通常是以 θ 來表示。這裏，先探討 IRT 及 ICC 二個概念。IRF 指的是從個人能力映射到個人反應答對機率的一種數學函數關係 (Lord, 1980a)。這個函數關係的定義域是個人能力 θ ，理論上可定義在 $+\infty$ 到 $-\infty$ 。而其對應域是個人反應答對的機率 P，理論上對應範圍乃介於 0 與 1。函數關係則是以 $P(\theta)$ 來表示，整個 IRF 如圖 1 所示。

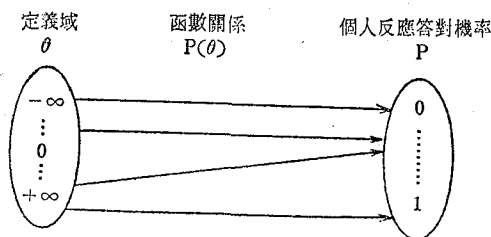


圖 1 項目反應函數圖

由圖 1 可以看出：一個 θ 透過 $P(\theta)$ 恰可以對映出一個 P，不同的 θ ，可能對映出相同的 P；但是相同的 θ 不可能對映不同的 P。IRF 函數 $P(\theta) = P(u = 1)$ 。u = 1 表示通過或答對，u = 0 表示答錯或未答 (Lord, 1980a)。IRF 則是被以下所要討論的項目參數所定義，至於在 LTT 項目參數有那些，其又如何定義 IRF 以及定義出來之函數關係的種類，是 LTT 的核心所在，都是以下陸續要討論的重點。至於 ICC 是將 IRF 的函數關係表達於正交座標平面圖上，以橫軸 (X 軸) 表示個人能力 θ ，以縱軸 (Y 軸) 表示個人反應答對機率 P 時， θ 與 P 所形成的數對在座標平面上對應點聯結而成的曲線 (林一真，民 71；Hambleton, 1985；Hambleton & Cook, 1977)。當 θ 越大時 P 值也越大，呈現一種單調性遞升 (monotone increasing) (Lord, 1980a)，簡單的說 ICC 便是 IRF 之函數圖。各種不同的項目特性曲線可以反映出不同模式之 LTT。

$$P(\theta_i) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (\text{公式 1})$$

公式 1 的函數 (以下簡稱函數 1) 便是 LTT 中一個 IRF 的例子，其中 θ 指的便是個人能力在

一連續線上的位置。而 a, b, c 所指的就是項目參數，也就是用以描述測驗項目的指數。圖 2 之實線部分曲線，便是 a, b, c 決定後所形成之 ICC (Lord, 1980a)。以下研究者將藉用函數 1 的 IRF 及圖 2 的 ICC，說明測驗項目及用以描述測驗項目之項目參數。

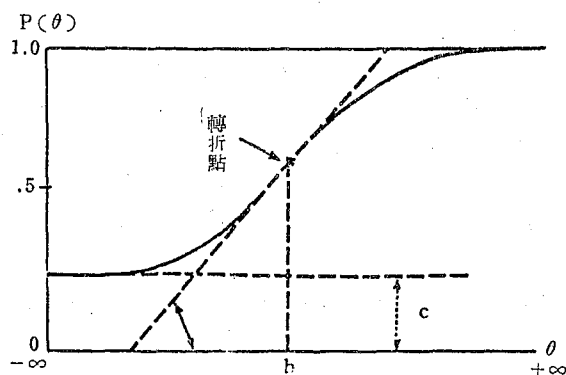


圖 2 項目特性曲線

(1) 參數 b ：在 LTT，函數 1 之中的 b 值是難度的指數，每一個測驗項目均有其 b 值 (Warm, 1978)，它決定了 ICC 的位置， b 值越大，表示測驗項目越難。從圖 2 可以看出實線部分之 IC-C 有一個轉折點，在此一轉折點之前 ICC 下凹，之後 ICC 上凸，那麼此一轉折點的橫軸坐標值，是一個 θ 值便是 b 值 (Lord, 1980a)。在 ICC 上，通過此一轉折點的切線斜率也是最大 (Hambleton & Cook, 1977)。另外當測驗項目不可能被猜對時，則在這一轉折點的縱軸坐標值，是一個 P 值，將會是 0.5 (Hambleton, 1985; Lord, 1980a)，若測驗項目有可能被猜中的話，則此一轉折點的縱軸坐標值，將會超過 0.5。在 LTT 中， b 值一般均會落在 -2 及 $+2$ 之間 (林一眞，民 71； Hambleton, 1985)， b 值越接近 -2 ，則該測驗項目越簡易，ICC 會偏左些，個人反

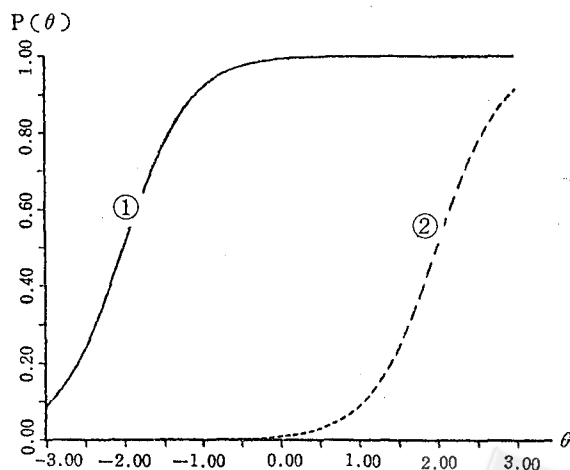


圖 3 不同 b 值不項目特性曲線

(① $b = -2.0, a = 1.39, c = 0$)
(② $b = +2.0, a = 1.39, c = 0$)

應答對的機率 P 值會大些，如圖 3，實線部分，第一測驗項目的 b 值便是 -2 。若 b 值越趨近 $+2$ ，則該測驗項目便較難些，ICC 也會偏右些，如圖 3，虛線部分，第二測驗項目的 b 值便是 $+2$ ，個人反應答對的機率也會小一點。

(2) 參數 a ：在 IRF 函數 1 中的 a 值，是項目鑑別度的指數。在圖 2 ICC 上轉折點的切線斜率與 a 值是成正比例的 (Hambleton, 1985; Lord, 1980a; Warm, 1978)。實際上轉折點上切線斜率值等於 $.425a(1-c)$ ，此地的 c 以後再加說明。當個人能力 θ 值逐次改變時，個人反應答對機率 P 值也隨之在改變；當 θ 改變，而 P 值改變的程度，便可以從 a 值反映出來 (Lord, 1980a)。從理論上看 a 值可以界定在 $-\infty$ 到 $+\infty$ 之間，但通常我們對負的 a 值不感興趣。另外實際上也不易求得一個大於 2 的 a 值，因此通常使用的 a 值多半會介於 0 與 2 之間 (Hambleton & Cook, 1977)。通常 a 值接近於 2 時，則表示該測驗項目鑑別性能佳，ICC 呈現險昇陡坡，如圖 4 第二測驗項目（虛線部分）的 a 值便是接近 $+2$ 。 a 值若接近 0，小於 0.8，則這個測驗項目就不可說是好的項目了 (Warm, 1978)，那它的 ICC 也會呈現緩升坡，如圖 4 第一測驗項目（實線部份）的 a 值就是接近 0 的。

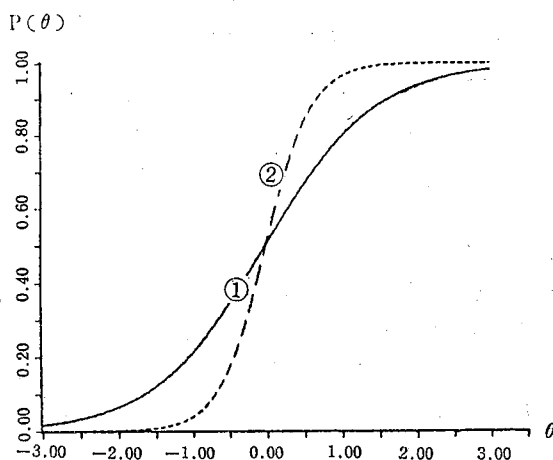
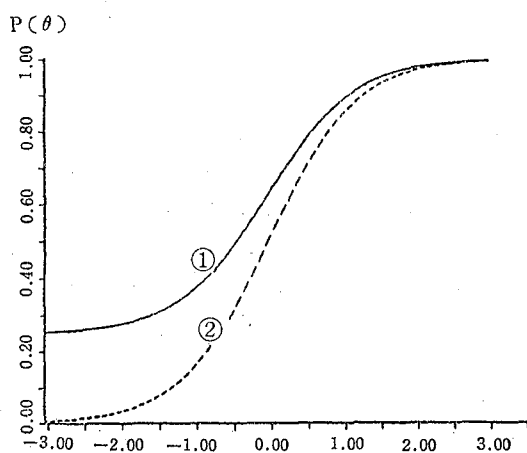


圖 4 不同 a 值之項目特性曲線

(① $b=0.0, a=0.79, c=0$;
② $b=0.0, a=1.90, c=0$)

(3) c 參數：在 IRF 函數 1 中的 c 值，指的是當個人能力 θ 值等於 $-\infty$ ，即是一種完全缺乏某種能力時，他仍能答對測驗項目的機率 (Lord, 1980a)，個人也許根據部分知識而猜測 (Lord & Novick, 1968)，也可能命題本身提供了某些暗示的線索，或是由於配置選項不當提示了個人而答對測驗項目，因而增加了個人反應答對機率 P 值的提高 (Hambleton & Cook, 1977)，因此 c 參數有時亦被稱為猜測參數 (guessing parameter)，但 c 值的存在及升高，並不全是個人對測驗項目的隨機猜測所導致，尚有其它上述的可能性，因此 c 參數有時也被稱為假性機會分數 (pseudo-chance score level) (Lord, 1980a)。從圖 2 的 ICC 來看， c 值所指的是 ICC 在往低走向時，逐漸靠近 $-\infty$ 的 P 值。大部份的 c 值是界於 0 與 0.4 之間，當測驗項目 c 值大於或等於 0.3 時，此測驗項目，不是一個好的項目， c 值越大表示測驗項目愈差，如圖 5，實線部份第一個測驗項目的 c 值便是接近於 0.3。 c 值越小越好，測驗項目的性能亦越佳。 c 值為 0 是一個最理想的狀況。如圖 2—5 虛線部份第二測驗項目的 c 值便是為 0 的。

圖 5 不同 c 值之項目特性曲線

- (① $b=0.0, a=0.99, c=0.25$;
 ② $b=0.0, a=0.99, c=0.0$)

(4) δ 參數：也有學者稱 r 參數 (Hambleton, 1985)，即使是個人能力 θ 值趨近於 $+\infty$ ，也不一定說個人反應答對機率 P 值必然是 1，只能說是接近 1。因為個人畢竟時有筆誤、粗心的機會，或甚至未解或誤解測驗項目的意義 (Barton & Lord, 1981; Hambleton, 1985)。當測驗項目的 δ 值越大，表示該測驗項目有問題的可能性也越大。理想狀況是 δ 值趨近於 0。圖 2 實線部份的 ICC，其 δ 值便是趨向於 0。倘若此 ICC 高走向時，當 θ 接近 $+\infty$ 時的 P 值未接近 1，則它與 1 之間的距離便是 δ 值，簡言之， $\delta = 1 - P(\theta = +\infty)$ 。

以上分別就幾個項目參數 b 、 a 、 c 、 δ 加以探討，這些項目參數正是 LTT 中用以描述測驗項目的指數。由各種項目參數不同的值做出不同的組合，便可形成各種不同測驗項目的 IRF。若再將 IRF 形成正交座標圖，便可得到各式各樣 ICC。應用 LTT 的測驗學者，便是操弄這些項目參數來從事測驗研究或是測驗編製工作。

2. 受試者反應：在 LTT 中，個人對測驗項目的反應通常以 u 來做記錄，如在第 i 個項目的反應，則以 u_i 記錄。當答對測驗項目，即個人反應正確，則 $u_i=1$ ，若個人反應不正確時則 $u_i=0$ ；當個人對 k 個測驗項目做反應，則可記錄為向量 $(u_1, u_2, \dots, u_i, \dots, u_k)$ ，此地 u_i 不是 1 便是 0。此一由 1 或 0 組成的向量亦可稱為項目反應組型 (pattern of item response) (Lord & Novick, 1968)。在 LTT 中，受試者反應的兩種機率，一是個人反應答對測驗項目機率，即前述的 $P(\theta)$ ；表示當個人特質 θ 值已知，答對某測驗項目的機率，亦可寫成 $P(u_i=1|\theta)$ 條件機率的形式。這是針對某一特定測驗項目而言。另一個是個人反應通過某些測驗項目的機率，表示當個人能力 θ 值已知時，答對某些測驗項目的機率，亦即 θ 值已知時，個人反應形成某種項目反應組型的機率，可以寫成 $P(u_1, u_2, \dots, u_k|\theta)$ (Hambleton, 1985)。從理論上看，不管 θ 值為何，只要 θ 已知，對於 k 個測驗項目而言，個人反應會有 2^k 個不同形式的項目反應組型。任何形式都有機會得到，只是得到的機率不同而已。例如，當個人能力 θ 值接近 $+\infty$ ，那麼他的項目反應組型的向量有可能是單元向量，也有可能是零向量，亦即可能答對全部測驗項目，也可能全部答錯。只不過這時候我們會說項目反應組型得到單元向量的機率會比得到零向量要多一些。這是 LTT 中一個重要的觀念。若是我們將此觀念反過來看，那就顯得更接近事實，而且更為有用了。即已知的部份由原先的 θ ，變為項目反應組型。這是符合事實的，我們是很難先知道 θ 的，測量中我們會先得到項目反應組型的。一

個固定的項目反應組型，亦有可能被任何的 θ 所得到，只是不同的 θ ，得到的機率也不同。例如當一個人得到的項目反應組型為單元向量，亦即通過全部測驗項目，則我們可能比較願意相信他的 θ 是接近 $+\infty$ ，而不會去相信他的 θ 是接近 $-\infty$ 。

3. 個人特質：在 LTT 中，個人特質所扮演的是中介變項的角色，是一個心理建構 (psychological construct)，幾乎所有從事心理學研究工作者，所感興趣的便是這個變項，尤其是認知心理學家。對於人類行為的理解，在心理建構的基礎之上，才不致於破碎及無能。當然心理建構也有它的特點，如它是一種假設性的，它是無法直接觀察的，物理特性上它是不存在的，自然它也就可能有真有偽。也正因它可能是正確，也可能是錯誤，所以就更具可研究性。

LTT 中的個人特質的特性就如同上述，所以它更常被稱為潛在特質 (latent trait) 及潛在變項 (latent variable)，看不見摸不着的，到底存在與否還是個問題，但它經常被拿來對個人行為做心理學上的描述 (Anastasi, 1982; Hulin, Drasgow & Parsons, 1983; Lord, 1980a; Weiss, 1983)。然而心理測量學者，所關心的並不是個人特質，因為這東西他們畢竟無法去面對它，更遑言去度量它。他們所真正感興趣的是呈現測驗項目，觀察紀錄個人反應，進而推估個人特質。他們欲瞭解個人特質，却只是依附個人特質，大事研究如何去安排測驗項目及如何去從個人反應中推估個人特質 (Weiss, 1983)。因此在 LTT 中，個人特質也只以一個 θ 參數來加以描述。這不同於以 a 、 b 、 c 、 δ 參數來描述測驗項目。本研究所欲研究的個人特質是「能力」這一個心理建構，所以 θ 所指的通常是個人能力的指數。理論上 θ 參數可以是在 $\pm\infty$ 之間，但是沒有自然零點及單位，因此習慣上是標準化的，亦即以 0 為平均數，1 為標準差，以使 θ 易於理解，因此 θ 便會經常界於 $+3$ 與 -3 之間，但 θ 仍舊有可能超出 ± 3 之外 (Warm, 1978)。

4. LTT 的理論模式：前述的三項 LTT 組成的基本元素——測驗項目、受試者反應及個人特質，三者之間的關係如何串起來，亦即整個理論架構如何運作起來，關鍵就在所採用的理論模式。LTT 的理論模式串聯了測驗項目與受試者反應之間，再從受試者反應到個人特質之間 (Hulin, Drasgow & Parsons, 1983)。串聯這三者的，經常是一些數學函數，所以整個 LTT 的理論模

表 1 LTT 的理論模式

處理資料性質	理	論	模	式
二 分 類 資 料	潛	在	線	性 模 式 (Latent Linear)
	完	全	量	尺 模 式 (Perfect Scale)
	潛	在	距	離 模 式 (Latent Distance)
	常	態	肩	形 模 式 (One-, Two-, Three- Parameter Normal Ogive)
	對	數	模	式 (One-, Two-, Three- Parameter Logistic)
	四	參	數	對 數 模 式 (Four-Parameter Logistic)
多 分 類 資 料	名	義	反	應 模 式 (Nominal Response)
	等	級	反	應 模 式 (Graded Response)
	局	部	給	分 模 式 (Partial Credit Model)
連 續 性 資 料	連	續	反	應 模 式 (Continuous Response)

(摘自 Hambleton, 1985)

式大抵均是數學模式 (Hambleton, 1985)，這是 LTT 的核心所在 (Weiss, 1983)，是 LTT 研究工作者爭議最多，也是工作努力最多之所在。那麼整個理論架構複雜的起源也就是在這裏。

Hambleton (1985) 將 LTT 的理論模式根據處理資料的性質，將理論模式分為三大類，詳如表 1。由表 1 可以看出目前的 LTT 的理論模式，大抵均集中於二分類資料的理論模式。所謂二分類資料，指的就是 0 與 1 的資料，即受試者反應若正確，以 1 表示，否則以 0 表示。

研究者僅列舉常態肩形模式及對數模式：

(1) 一、二、三參數常態肩形模式：

Lord (1980a) 指出：假設有一潛在變項 Y_i' ，它的大小決定了個人對第 i 個測驗項目反應的正確與否。另有一常數 r_i ，它是指對第 i 個測驗項目的常數。當個人的潛在變項 Y_i' 大於常數 r_i ，則它對第 i 個測驗項目反應正確，記為 $u_i=1$ 。當 Y_i' 小於 r_i ，則表示反應錯誤，記為 $u_i=0$ 。在這裏， Y_i' 潛在變項的內涵相當複雜，因為它的大小直接決定了個人對第 i 個測驗項目的正確與錯誤。因此前述個人能力 θ 值與 Y_i' 會有相當密切的關係，因為決定個人反應正確與否，個人能力 θ 是相當重要的因素，但却不是唯一的決定因素。換句話說， Y_i' 的組成有一大部分是 θ ，另亦有一部份其

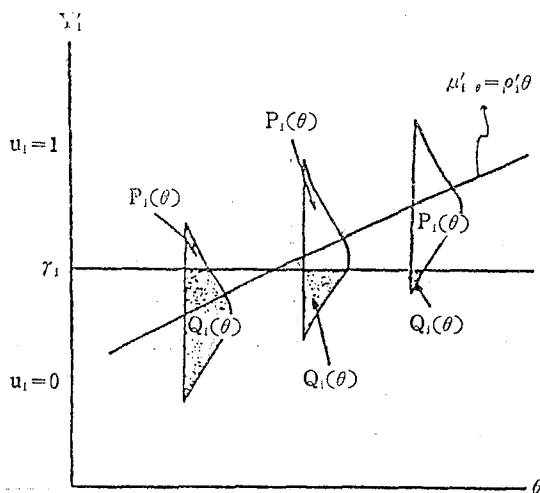


圖 6 Y_i' 在三個 θ 水準上之條件分配及 $\mu'_{i|_0}$ 之迴歸線， $Y_i'=r_i$ 之直線
(取自 Lord, 1980a)

它未必可預知的因素，沒有系統的在影響 Y_i' 。因此假定個人 θ 固定，則可能形成的 Y_i' ，也會有相當大的變異。基於上述，Lord 再進一步假定：①在 θ 的連續線上，每一個 θ 上，均會形成許許多多的 Y_i' ，而它們的平均數 $\mu'_{i|_0}$ ，也就是這 Y_i' 集中的位置；此 $\mu'_{i|_0}$ 與 θ 之間呈現線性關係。②在每一個 θ 上， Y_i' 的離散情形也是同質的，即在每個 θ 上 Y_i' 的條件變異量 $\sigma^2_{i|_0}$ 都是相同的，可用 $\sigma^2_{i \cdot 0}$ 表示。③在每一個 θ 上 Y_i 的條件分配均是常態分配。如圖 6 所示。從圖中我們可以看出個人能力 θ 下，其對第 i 個測驗項目之反應正確機率 $P(\theta) = P(Y_i' > r_i | \theta)$ ，它是等於標準化常態分配曲線下的一塊區域面積。現若將 r_i 標準化為 $(r_i - \mu'_{i|_0}) / \sigma_{i \cdot 0}$ ，用 $-L_i$ 表示，則 $-L_i$ 在 θ 固定狀態下，會成為標準化常態分配。為方便討論，將 θ 及 Y_i' 均標準化為平均數 0，標準差 1 的量尺，則根據 θ 預測 Y_i' 的迴歸方程式便是 $\mu'_{i|_0} = \rho'_{i \cdot 0} \cdot \theta$ ，此 $\rho'_{i \cdot 0}$ 指的便是 θ 與 Y_i' 的相關係數，此迴歸線的估計變異誤便是 $1 - \rho_i'^2$ ，也就是前述的 $\sigma^2_{i \cdot 0}$ 。現在再將 $-L_i = (r_i - \mu'_{i|_0}) / \sigma_{i \cdot 0}$ 改寫一下使成公式 2：

$$-L_i = \frac{\gamma_i - \rho'_i \cdot \theta}{\sqrt{1 - \rho_i'^2}} \quad (\text{公式 2})$$

$$a_i = \frac{\rho_i'}{\sqrt{1 - \rho_i'^2}} \quad (\text{公式 3})$$

$$b_i = \frac{\gamma_i}{\rho_i'} \quad (\text{公式 4})$$

其次令 a_i 、 b_i 分別如 (公式 3)、(公式 4) 則 $-L_i = a_i(b_i - \theta)$ ， $L_i = a_i(\theta - b_i)$ 。由圖 6 也可以看出每一個 θ 上 Y_i' 的標準化常態分配曲線上 Y_i 大於 γ_i 的那一區域面積便是 $P_i(\theta)$ ，表示個人

$$\begin{aligned} P_i(\theta) &= \int_{-L_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_{a_i(b_i - \theta)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \end{aligned} \quad (\text{公式 5})$$

$$\begin{aligned} &= \int_{-\infty}^{L_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \end{aligned} \quad (\text{公式 6})$$

對第 i 個測驗項目反應正確的機率。當 θ 上升時，這一塊面積也有增大的趨勢，亦即當 θ 上升時，其反應答對機率也必然逐漸上升，而這個反應答對機率 $P_i(\theta)$ ，便可以函數 5 及函數 6 的形式表示出來 (Lord, 1980a)。上述 $\mu'_i|_{\theta} = \rho_i' \theta$ 中的 ρ_i' 指的是 $\mu'_i|_{\theta}$ 與 θ 的相關係數，是迴歸線的斜率，它與測驗項目的鑑別度有密切的關係。另外 $Y_i = \gamma_i$ 此一平行於 θ 軸 (X 軸) 的直線，其 γ_i 所在的位置，與測驗項目的難度有直接的關係存在。

$$P_i(\theta) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (\text{公式 7})$$

上述推演結果形成的函數 5 或函數 6，指的是二參數常態肩形模式，因為從函數 3 及函數 4，得知其涉及的项目參數有 a 參數及 b 參數。 a 參數是測驗項目的鑑別度指數， b 參數是難度指數。現在令 $a_i = 1$ ，則形成函數 8，便是單一參數常態肩形模式。

$$P_i(\theta) = \int_{-\infty}^{\theta - b_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (\text{公式 8})$$

此一模式只是將二參數肩形模式中 a 參數模式均設定為 1。另外若在二參數模式中亦考慮 c 參數，即前述的猜測的指數，則可形成如函數 7 的三參數常態肩形模式。

LTT 的發展中，以常態肩形模式發跡較早，從 1943 年起至 1970 年有相當多的學者如 Lawley (1943; 1944), Tucker (1946), Lord (1952), Bock & Lieberman (1970) 及 Kolakowski & Bock (1970), (Hulin, Drasgow & Parsons, 1983)]，從事常態肩形模式之 LTT 的研究，過了七十年代似就乎少之又少了，原因可能是對數模式的提出。

(3) 一、二、三參數對數模式

Birnbaum (1968) 提出對數分配函數 (logistic distribution function)，如函數 9：

$$\begin{aligned} \psi(x) &= e^x / (1 + e^x) = 1 / (1 + e^{-x}) \\ &(-\infty < x < \infty) \end{aligned} \quad (\text{公式 9})$$

與累積常態分配函數 (cumulative normal distribution function) $\Phi(x)$ ，如函數 10：

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (-\infty < x < \infty) \quad (\text{公式 10})$$

二者的分配函數圖形差異極小。此地累積常態分配函數便是前述常態肩形模式所採用的數學函數。Birnbau 並指出 $\Psi(1.7x)$ 與 $\Phi(x)$ 之間的關係是：

$$|\Phi(x) - \Psi(1.7x)| < 0.01$$

表示二者之間差異極小。 $\Psi(1.7)$ 與 $\Phi(x)$ 之間差異關係在 Warm (1978)，及 Birnbau (1968) 均有詳細討論，此處只引用其結論： $\Phi(x)$ 與 $\Psi(1.7x)$ 的分配情形是極其相似的。然而在應用上對數分配函數要比累積常態分配函數簡便得多，因為累積常態分配函數涉及積分的問題，在數理處理上麻煩得多 (Hambleton, 1985; Hulin, Drasgow & Parsons, 1983; Warm, 1978)。

以下研究者將逐一將三種對數模式加以探討。

① 單一參數對數模式：

在西元 1966 年，丹麥數學家 Georg Rasch 獨立研究測驗理論，便已提出單一參數模式的理論。另外跟進研究的學者有 Anderson, Kearney & Everett (1968); Wright (1968, 1977); Wright & Panchapakesan (1969); Wright & Stone (1977) (Hambleton, 1985)。至今仍有相當多的學者擁護單一參數對數模式，尤其是美國芝加哥大學的 Benjamin D. Wright。畢竟它有它吸引人的長處，例如它涉及的參數較少，易於處理。再者，在估計參數時所遭遇的問題顯然少於其它的對數模式 (Hambleton & Cook, 1977; Hambleton, 1985)。

由於 Georg Rasch 的關係，單一參數對數模式也被稱為 Rasch 模式 (Rasch Model)。它的 IRF 如函數 11：

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (\text{公式 11})$$

其實若回到函數 1 之上，令 $c_i = 0$ ， $a_i = \bar{a}$ 常數，則函數 1 便可形成函數 11。顯然地它假定一個測驗當中所有測驗項目鑑別度都一樣。另則假定無猜測因素影響個人反應 (Hambleton, 1985)。換言之，影響個人反應答對機率的大小，除了項目參數以外，一切便由個人能力 θ 所決定。這樣的假定在實際狀況下似乎是相當困難，因為在一般狀況下的測驗很難符合所有測驗項目鑑別度都一致的基本假定。此外，個人反應答對機率完全由個人能力決定也是不容易達到，因為影響個人反應答對機率的，還有其它如猜測、動機等其它系統或非系統影響的因素 (Hulin, Drasgow & Parsons, 1983)。

② 二參數對數模式。

這是 Birnbau (1968) 所提出的，它的 IRF 如函數 12：

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (\text{公式 12})$$

它比單一參數對數模式多考慮一個項目參數 a ，即鑑別度的參數，在二參數對數模式裏不再是常數 \bar{a} ，而是每一測驗項目均可能不同的 a_i 。但若比起函數 1，它仍是令 $c_i = 0$ ，也是不考慮猜測的因素。因此二參數對數模式是較為適合於開放性作答的測驗，而較不適合於選擇式的測驗，因為只有開放性作答，猜測因素影響個人反應答對的機率，可以降到較接近 0，比較可能符合二參數對數模式 (Hulin, Drasgow & Parsons, 1983)。

③ 三參數對數模式：

1968 年 Birnbau 亦提出了三參數對數模式。與前述二參數對數模式之區別是，多使用了一

個項目參數 c 來描述測驗項目，其 ICC 如函數 1，圖 2 中實線部分曲線，便是一個標準的三參數對數模式的 ICC。當 θ 趨向於 $-\infty$ 時， $P(\theta)$ 到底是否等於 0，是三參數對數模式涉及的一項重要問題。在前述的單一參數對數模式及二參數對數模式中，實際上均是假定 c 參數為 0，而三參數對數模式中，所提出之 c 參數是可為 0，也可能是大於 0。 c 參數的介入使得 $P(\theta)$ 的全距由原先的 0 到 1，縮小為 c 到 1，對 ICC 會有壓扁的現象，如圖 5。 c 參數影響公式 4 中的 r_1 ，當 c 上升，其實 r_1 也要上升，若要保持 b_1 不變，則必須縮小 ρ_1' ，所以說當 c 參數上升時，ICC 的壓扁現象，也就是 ICC 上升的坡度減緩 (Warm, 1978)，實際上是減低了測驗項目的鑑別度。

在實際的測驗資料中，尤其是選擇式測驗項目，當一個能力極差的，即 θ 趨近 $-\infty$ 者，他對測驗項目反應答對率 $P(\theta)$ ，未必見得等於 0，換言之，以單一參數或二參數對數模式是無法解釋的。這樣的現象，在 LTT 未發展出來之前已受到學者的關心。最早的解釋是將其歸於隨機猜測，則認定 c 參數為 $1/m$ ， m 是提供的選項數。然而在實際的測驗狀況之下，即使是一個人對某一測驗項目完全不懂，他也不會隨機猜測 (Lord & Novick, 1968)。Lord (1974) 指出估計的 c 參數值通常小於 $1/m$ ，理由是命題者通常會配置一些誘惑性很高的錯誤選項，能力極差的個人，也比較容易選上這些錯誤答案。Warm (1978) 的研究指出，四個選項 A、B、C、D 的測驗，標準答案為 C 的測驗項目，估計出來的 c 參數會高一些。這現象說明了，命題者傾向以 C 答案做為正確答案，而個人反應亦有傾向選 C 的答案。Warm (1978) 的解釋認為命題者傾向於將正確答案隱藏在中間，當個人不知道正確答案時，其反應亦傾向於選擇隱藏在中間的選項做為正確答案。

c 參數的解釋，正如同上述的多樣性，顯示出在三參數對數模式中的 c 參數並不如 a 、 b 二參數那樣有系統地在變化。故而引發許許多多學者，如 Lord (1969, 1970, 1974, 1975, 1980)，Hambleton & Traub (1971)，Marco (1977) 的研究 (Hulin, Drasgow & Parsons, 1983)。甚至有的學者持反對意見。Wright (1977) 認為 c 參數介入對數模式，大大破壞了實際測量工作的邏輯，而且最主要的問題是 c 參數的介入並不能滿足實際測量工作上的需要；他引用 Lord 在 1968 年的研究指出 Lord 在三參數對數模式中參數的估計並不是聚斂得很理想，而且結果並不十分穩定。其實這現象至今仍存在，許多學者也正努力從事改善的研究工作 (Jones, 1982, 1983；Lord, 1981, 1982, 1984)。這只是一個技術上的問題，一時可能還無法完全滿足實際測量上的需要。但就整個三參數對數模式的概念上來看，它頗符合一般測驗的概念，尤其是針對選擇式的測驗 (Hambleton, 1985；Hulin, Drasgow & Parsons, 1983)。因此它仍被絕大多數學者所接納，例如 Warm (1978) 便指出三參數對數是符合事實的。LTT 的大師級學者如 ETS (Educational Testing Service) 的 Lord；USCSC (United States Civil Service Commission) 的 Urry 以及明尼蘇達大學的 Weiss，大抵均是從事三參數模式的研究。最近的實證性研究 (Jones, Wainer & Kaplan, 1984；Thissen & Wainer, 1985) 也以實際測驗結果符合三參數對數模式，而進一步支持三參數對數模式的可靠性。因此在本研究中 LTT 所採用的便是此三參數對數模式。

④四參數對數模式：四參數對數模式比上述三參數對數模式又增加一個項目參數 δ ，用以描述測驗項目。在單一參數及二參數對數模式中，當 θ 趨向於 $+\infty$ 時，其 $P(\theta)$ 都等於 1；當 θ 趨向於 $-\infty$ 時， $P(\theta)$ 等於 0。換言之，在單一及二參數對數模式中，個人對某測驗項目反對答對的機率，除了 a 、 b 參數之外，便完全由能力 θ 所決定。當個人能力極高時，即 θ 等於 $+\infty$ 時，則可保證答對該測驗項目，因為 $P(\theta)$ 等於 1。當個人能力極差時，即 θ 等於 $-\infty$ 時，則可確定無法通過測驗項目，因為 $P(\theta)$ 等於 0。前述三參數對數模式中，項目參數 c 的介入，便是討論當 θ 趨向 $-\infty$ 時， $P(\theta)$ 未必是 0 的問題。此地的四參數對數模式中， δ 參數所涉及的是 ICC 的另一端，即當個人能力 θ 趨向 $+\infty$ 時， $P(\theta)$ 也未必是 1 的問題 (Hambleton, 1985)。其 IRF 如函數 13：

$$P_i(\theta) = c_i + \frac{\delta_i - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (\text{公式 13})$$

Barton & Lord(1981) 的研究中使用了四參數對數模式，結果發現：多了 δ 參數對於個人能力 θ 的估計並無助益。因此， δ 參數始終沒有引起學者的興趣。

二、潛在特質理論與傳統測驗理論

乍看之下 LTT 的觀念與傳統測驗理論 (Classical Test Theory, 簡稱 CTT) 的觀念截然不同。然而 Weiss (1983) 認為 CTT 的觀念中早就隱含着 LTT 的觀念，甚至指出 CTT 就 LTT 的一個簡單模式。Hulin 等 (1983) 也認為此二者之間是局部重疊的。以下研究者擬簡要比較 CTT 及 LTT 二者，探討其異同：

(一) 能力參數 θ 與真正分數

CTT 的最重要目的就是要根據受試者反應組型計分所得的觀察分數 (observed score) 或稱實得分數 (obtained score)，來推估受試者的真正分數 (true score, 以 τ 符號)。正如同在 LTT 中，根據受試者反應組型估計個人能力 $\hat{\theta}$ ，以便推估受試者真正的 θ 所在。

在 CTT 中的 τ ，從理論上看，指的是對同一受試者實施同一個或複本測驗無限多次，得到無限多個觀察分數，這些觀察分數的期望值便是 τ 。因此若以 x 代表觀察分數，則 τ 與 θ 的關係便可如公式 14。在 LTT 中已知 $P(\theta)$ 會隨上升而遞升，亦可推知 $\sum_{i=1}^n P_i(\theta)$ 也是隨 θ

$$\begin{aligned} \tau &= E(x) = E\left(\sum_{i=1}^n u_i\right) = \sum_{i=1}^n E(u_i) \\ &= \sum_{i=1}^n [1 \times P(u_i = 1 | \theta) + 0 \times P(u_i = 0 | \theta)] \\ &= \sum_{i=1}^n P_i(\theta) \end{aligned} \quad (\text{公式 14})$$

上升而遞升的。所以說 τ 與 θ 的關係也是當 θ 上升 τ 亦會上升的 (Hulin, Drasgow & Parsons, 1983)。至此 τ 與 θ 的關係一目了然。難怪 Weiss (1983) 指出 CTT 的 τ 與 LTT 的 θ 是類似的，所不同的只是它們各自使用不同的量尺罷了。

其實 θ 與 τ 真正的差異是在功能上。若有兩個測量相同能力的測驗，而其中一個測驗項目較為艱難，另一個比較簡易，現一受試者同時接受此二測驗，理論上看，受試者能力在兩測驗上是一致的，但所得的二個 τ 却不一定會一樣的。很有可能在前一個測驗所得的 τ 會小些，因為測驗項目艱難，不易答對。這裏顯現一個 Lord (1980a) 所指出的 CTT 之缺點，那就是 CLL 中 τ 的量尺是被所選用的測驗項目所左右。這也使得 CTT 在實際應用上面臨一些難題，那就是測驗重覆的現象 (Warm, 1978)。往往基於標準化使用的理由，測量相同能力的測驗，經常會不只一個。因此從 CTT 看， τ 的使用是有很大的限制。

上述 τ 的限制，在 LTT 的 θ 參數，並不存在。因為 $\hat{\theta}$ 的估計不只和反應答對與否有關，更與項目參數有直接的關係。也就是，在 CTT 答對一題就是一題，不管是怎樣的一題；而在 LTT 中答對一題是一題，但尚得看看是怎樣的一題，是艱難的？抑或是簡易的？答對的意義不同。也正因為此一特性，使得適性測驗 (tailor testing) 更加有實際意義。因為適性測驗正是根據每個受試者的能力水準，選擇適當難度的測驗項目給予實施，因此每位受試者不一定接受相同的測驗項目，但在 LTT 下，他們的結果是可比較的 (Hulin, Drasgow & Parsons, 1983) 這在 CTT 是辦不到的。

(二) 項目參數與項目統計數

前面已提到，在 LTT 中是以項目參數 a 、 b 、 c 等來描述測驗項目。在 CTT 中則是以項目

統計數 (item statistics) 來描述測驗項目。Lord (1980a) 指出 CTT 是以反應正確受試者之百分比，做為難度指數 (以 P 表示)，是以項目分數與觀察分數的點二系列相關係數 (以 ρ'_{1x} 表示) 或二系列相關係數 (以 ρ_{1x} 表示) 做為鑑別度指數。根據前述二參數常態肩形模式中 θ 與 Y_1' 的關係，Lord (1980a) 指出 LTT 中的 a 、 b 項目參數與 CTT 中的 P 及 ρ_{1x} 的關係分別如公式 15 及公式 16：

$$a_1 = \frac{\rho_{1x}}{\sqrt{1 - \rho_{1x}^2}} \quad (\text{公式 15})$$

$$b_1 = \frac{\gamma_1}{\rho_{1x}} \quad (\text{公式 16})$$

公式 16 中的 γ_1 指的是圖 6 中的 γ_1 ；CTT 中的 P 便是 γ_1 以上的常態曲線內的面積，當 γ_1 上升時， P 便會減小。可見 LTT 中的 b 參數與 CTT 中的反應正確受試者百分比 P 是呈互為消長的關係。

Warm (1978) 指出 CTT 中所使用 P 與樣本能力有關係。當受試者能力偏高時， P 值會升高；反之則 P 會降低。除此之外，Lord (1980a) 指出 P 的矛盾現象；二個測驗項目 P 值的高低順序，會因取樣不同而改變順序，如圖 7 所示。就樣本 A (能力偏低者) 而言，第 1 題比第 2 題難；就樣本 B (能力偏高者) 第 2 題比第 1 題難。Warm (1978) 認為這現象不是取樣誤差的問題，而是 P 本身不是一個適當的難度指數。

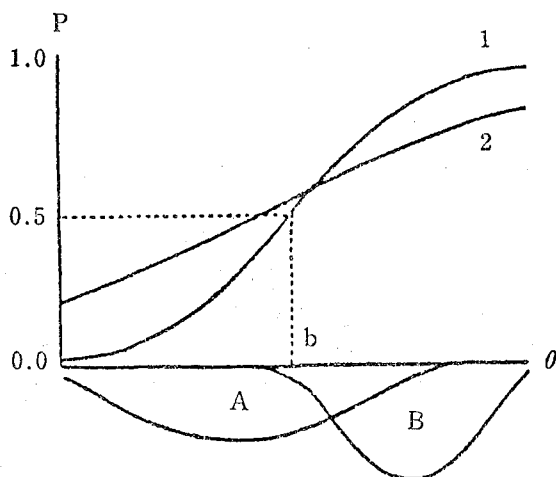


圖 7 A、B 兩個樣本與兩個項目的關係

Hambleton (1985) 亦指出 CTT 中所用的鑑別度指數 ρ_{1x} 與所取樣本能力分佈的情形有密切關係，當樣本能力分佈廣時， ρ_{1x} 亦較有升高的可能。

在 CTT 中所使用的 P 、 ρ'_{1x} 或 ρ_{1x} 均會受取樣所影響，也正因此，CTT 的難度及鑑別度指數，只堪稱項目統計數。

LTT 的二大目的，其一是估計個人能力參數 θ ，其二便是尋求不變的項目參數。所謂「不變」指的就是不因樣本改變而改變的意思。從 LTT 的理論上看，項目參數 b 、 a 、 c 是不會隨樣本而改變的。

(三) 測量的精確性

Hunlin (1983) 指出在 CTT 中，最主要的二個測量精確性指數，一個是測驗信度（以 ρ_{xx}' 表示），另一是測量變異誤（ σ_e^2 ）。從公式17測驗信度定義上看來， ρ_{xx}' 與 σ_e^2 根本就是同一回事。其中 σ_e^2 是指真正分數的變異數。值得注意的是，CTT 的 ρ_{xx} 與 σ_e^2 的變異均來自於全體樣本

$$\rho_{xx}' = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} \quad (\text{公式 17})$$

，而且也是將此指數反應回全體樣本。換句話說，對於全體樣本的測量，無論其能力水準如何， ρ_{xx}' 及 σ_e^2 都是同一個。CTT 無法說明測驗在某能力水準中的測量精確性如何，或甚至針對某一受試者而言，測量精確性如何。事實上，任何測量工具的精確性都會因所測量特質之程度而有所不同，正如同我們拿天平稱毛豬，天平的精確性會不理想。又，當以極艱難的測驗項目去測量一羣能力極差的受試者，結果沒有一個受試者答對，則此一項目精確性一定低。但我們用此同一測驗項目去測量能力偏高者，情況就會不同。顯然 CTT 所使用的精確性指數，是無法勝任上述的工作。

CTT 的 ρ_{xx}' 常是以重測信度、折半信度、庫李信度（Kuder-Richarson formula-20）來加以估計。顯然的，這些估計用的係數，也會隨著取樣不同而有不同。

LTT 中引用項目訊息函數（item information function, 簡稱 IIF）的觀念，來說明測量精確性的問題。根據 Warm (1978) 的說明，項目訊息與 ICC 每一個 θ 點上的斜率有密切的關係。從它的定義公式 18 看，項目訊息是 $P_1(\theta)$ 對 θ 第一階導數的平方 $P_1'(\theta)$ 除以答對反應機率 $P_1(\theta)$ 再除以反應答錯機率。 $Q_1(\theta) = 1 - P_1(\theta)$ 如公式 18

$$I(\theta, u_1) = P_1'(\theta) / P_1(\theta) Q_1(\theta) \quad (\text{公式 18})$$

這樣的一個精確性指數有它的特性如：

1. 項目訊息函數不像 CTT 的 ρ_{xx}' 或 σ_e^2 是整個測驗的指數，它沿襲 LTT 的觀念，是針對單一測驗項目而言。
2. 它不像 CTT 的 ρ_{xx}' 或 σ_e^2 是單一的指數，而是 θ 的函數。顯然每一個 θ 均會對應到它自己的 IIF。
3. ICC 上斜率最大的 θ 點，也最靠近 IIF 最大值的點。
4. LTT 模式中的 a 參數上升，則 IIF 也會上升，因為參數 a 與 ICC 的斜率有正相關。
5. LTT 模式中的 c 參數上升，則 IIF 會下降，因為參數 c 的上升，會降低 ICC 的斜率。
6. IIF 與 b 參數的關係比較微妙因 b 參數直接關係着 $P(\theta)$ ，而 $P(\theta)$ 在 IIF 的定義中扮演的是校正的角色。若 ICC 上二個不同的 θ 點斜率相同則 $P(\theta)$ 接近 0.5，它的 IIF 會大； $P(\theta)$ 接近 0 或 1，則 IIF 會小。

由於 IIF 具備上述的諸多特性，使得 LTT 在測驗的研究上及應用上產生了突破性的發展，如測驗編製選題、測驗對等（equating）、測驗偏差研究等，尤其是在適性測驗上的應用。因為 LTT 一反 CTT 測驗全體項目共赴使命的觀點，改採針對測驗中每個單一項目的觀點，處理測量問題的能力遠比 CTT 強得多。

（四）理論模式的基本假定

Hambleton (1985) 指出 CTT 的觀念直接了當，其基本假定較弱，且多數測驗資料均可符合其基本假定，所以較易於接受。而 LTT 則必須符合較強的基本假定，也正因此 LTT 會有較強的能力。換言之，促使 LTT 具有較強的功能，乃是源自其基本假定（Hambleton, 1985）。

三、適性測驗與潛在特質理論

（一）適性測驗的性質

Lord (1980b) 認為所謂適性測驗是指對受試者實施測驗的一種方式，它是針對一特定受試者，

根據其先前的反應來選取最適合此一特定受試者的測驗項目，做為下一題要實施用的項目；每實施一題便有可能對此一特定受試者評分；如此，一次選取一題，實施一題，評分一次，週而復始，直到預定題數達到，或預定的測量精確水準達到為止。Weiss (1983) 更具體指出：適性測驗是根據一套法則，在題庫 (item pool) 中選取項目難度與受試者能力相匹配的項目來實施。上述二者的觀點差異在於測驗項目選取的邏輯。Weiss 的觀點集中在項目難度與個人能力的相匹配，此一觀點為多數學者所採用 (Hambleton, 1985; Hulin, Drasgow & Parsons 1983; Urry, 1977)。而 Lord 的觀點則保留了較多 LTT 的觀點；以 Warm (1978) 的用語來說，測驗項目選取的邏輯是選取有助於對受試者能力估計的項目。

據 Lord (1980a) 指出適性測驗一詞乃西元1951年學者 William W. Turnbull 所提出使用。它與 adaptive testing, branch testing, individualized testing, programmed testing, sequential item testing, response contingent testing 等詞，所指的是相同的概念。其實適性測驗也不是什麼新觀念。根據 Weiss (1983) 所指，以前就有適性測驗的觀念了：

1. 西元1905年法國比奈 (Alfred Binet) 所發展的第一套智力測驗，便具備了適性測驗的特徵：(1)每位受試者是根據其年齡來決定開始實施的材料。(2)立即評分，並據以選取往下繼續實施的材料。(3)受試者不必做整個測驗即可結束。換言之，每個受試者可能都接受了不同測驗材料組合而成的測驗，而測驗結果是可比較的。

目前的適性測驗當然是複雜得多，但仍具備上述特性。

2. 適性測驗的概念在心理物理學上早就用在感覺閾限的測量。實驗心理學家測量感覺閾限所用的方法，如極限法 (The method of limit)、調整法 (The method of adjustment)，也就是使用適性測驗的不同概念而已。

所以適性測驗也只不過是舊瓶新裝的舊觀念，只是以新方法來處理而已。

(二) 適性測驗的基本立論：

1. 在 CTT 的標準化測驗中，為使測驗範圍擴大，即能力範圍擴大，必然的測驗項目難度也擴大，當然也加長了測驗長度。Kreitzberg & Jonse (1980) 引述 Weiss 的研究指出高能力的受試者對於 CTT 的測驗中簡單項目會厭煩，而影響測驗結果。低能力的受試者則會對 CTT 測驗中艱難的測驗項目感到挫折，產生焦慮。Weiss 研究亦指出低能力受試者對於 CTT 的測驗較常猜測，致使測驗結果較不正確。適性測驗基本上就是依據個人能力不同，而選擇難度相匹配的測驗項目來實施，因此一般說來高能力受試者做低難度項目，低能力受試者做高難度項目的機率均很小，因此便可杜絕上述厭煩及挫折的問題，進而改善測量的正確性。

2. 在 CTT 的測驗，經常是所謂單峯測驗 (peaked test) 指的是大部份測驗項目是屬於中難度，而偏高及偏低的極端難度只有少數，因此從 LTT 的觀點 CTT 的測驗是較適於測量中等能力受試者，而不適於測量高與低能力的受試者。適性測驗往往是在一個廣大的題庫中搜尋適當的題目，因此當測量極端能力的受試者，適性測驗亦可實施相當多極端難度的項目，也正因此 Lord (1968, 1980b) 纔說適性測驗對於極端能力受試者的測量優於 CTT 的測驗。Kreitzberg & Jones (1980) 引述 Lord (1970), McBride (1976) 及 Symposn (1970) 的研究指出：當受試者答對機率是在.50到.65時，則該項目的項目訊息最大，答對機率太高或太低的測驗項目，對測量均無太大助益。適性測驗則企圖增進測量正確性，降低測量標準誤。

(三) 適性測驗的發展

Weiss (1983) 指出1950年代紙筆式的適性測驗便已出現，但因其實施複雜，而終告放棄。直到1970年代左右，適性測驗再度受到注意。其原因有二：

1. 高速電腦的問世：Green (1970) 預測測驗擺脫不了電腦出現的影響。適性測驗有時也被稱為

電腦化測驗 (Lord, 1980a)。Weiss (1974) 更指出電腦化測驗比 CTT 的測驗更不受主試者影響，更符合標準化測驗的原則。而高速電腦問世對於適性測驗直接影響是電腦可以承擔適性測驗中複雜的實施程序。

2. LTT 的發展是促進適性測驗發展的主因：國內外測驗學者 (林一貫，民71；Anastasi, 1982; Urry, 1977; Weiss, 1983) 均指出 LTT 的出現，為適性測驗建立了良好的理論基礎。而且 LTT 的優點特性，也最容易表現在適性測驗的應用上。

Kreitzberg & Jones (1980) 指出 Angoff & Huddleston 於1958年便試圖應用 CTT 來發展適性測驗，但 CTT 應用於適性測驗存有三個重大困難：

1. 計分的問題：由於適性測驗的結果，不同的受試者接受不同安排的測驗項目，測驗內容可能不同，甚至測驗題數也不同。而 CTT 使用答對題數來計分易導致測驗結果解釋上及比較上的困難。

2. 項目參數的問題：在適性測驗選取適合受試者測驗項目的過程中，項目參數必須具備參數不變性，即項目參數不會隨着樣本改變而改變。在 CTT 中，Gulliksen (1950)指出項目參數是以羣體資料來加以界定，項目參數值的大小，會隨樣本而變動。

3. 研究比較的問題：CTT 中比較不同測驗實施方式，常用的是信度、效度之類的相關指數，由於這類指數並未具備不變性，故而並不適用於適性測驗研究使用。

上述三大困難阻礙了適性測驗在 CTT 中的發展。直到 LTT 的出現才解決了 CTT 中的三大困難。在 LTT 之下，不同的受試者所接受的測驗項目不同，而據以估計所得的能力參數是可比較的，而且理論上它是具備參數不變性。

LTT 的項目參數理論上也是具備參數不變性。LTT 並提供測驗訊息的觀念使適性測驗研究工作上，具體了許多。Weiss (1983) 更指出 LTT 提供了受試者反應與其能力參數 θ 之間的理論建構使得電腦模擬測驗實施及受試者反應的模擬研究變得可能。因此許多測驗的評估研究工作，便可以在 LTT 的模式下和模擬的情境下進行。這種測驗研究方式的特點是快速且省時經濟；研究結果常具備理論導向的作用。適性測驗研究便經常採取這種研究方式。研究者也是應用 LTT 的特性採用電腦模擬適性測驗進行本研究。

(四) 潛在特質理論應用在適性測驗

Hulin 等 (1983) 說明適性測驗的主要程序有三：第一、決定起始點：是指適性測驗如何開始測量的工作。第二、估計能力與項目選取：估計能力也就是前述的計分工作。第三、結束測量的標準。

上述三者是適性測驗的三大程序，其中又以第二項估計能力及項目選取為最重要，因為這兩項工作是 CTT 中最感困難的。而 LTT 的優點也正是表現在這兩項工作。大多數適性測驗研究主題也集中在此。下面就這三個程序再加以說明：

1. 決定起始點：Hulin 等 (1983) 指出適性測驗選擇第一題的方式有兩種，一是選取適中難度的項目，另一是以受試者有關資料如教育水準的高低，來選取較為符合某教育水準之適中難度的項目。簡單的說，前一種是不管是那一位受試者接受適性測驗，第一題都一樣。後者則可能是同一個年級的受試者，所做的第一題相同：Lord (1980a) 指出在1970年前後適性測驗研究，相當重視起始點的研究，因為當時的適性測驗項目選取策略主要是上下法 (up-and-down) 及羅一門二氏法 (Robbins-Monro)。這類的方法主要是依據項目難度來選取項目，受試者答對，則下一題會難一些；答錯，則下一題會簡單些。因此所選取的第一題，若項目難度與受試者能力相差太大時，則往往在開始階段要耗費較多測驗項目，才能大致估計出受試者能力。對於低能力的受試者，這開始階段可能會耗費更多的測驗項目，因為低能力受試者的反應，有相當的部份是由猜測因素所決定。所以那時的適性測驗研究會重視起始點。直到1977年 Lord 自己的一項電腦模擬適性測驗研究，在個人能力參數

已知的狀況下，安排第一題，使第一題的項目難度與個人能力之間的差距，在控制下進行研究，結果發現在受試者完成25題的情況下，不管第一題如何安排其測量精確的程度，大致是一樣的。Lord的結論是：起始點的安排是無關緊要的。研究者認為這個研究不夠詳盡，因為它是在25題的狀況下評估，並未對25題之前的結果做交待，也許不必25題，結果也會穩定下來。總之，對於25題之前未逐一探討，是浪費資料，也使問題的真相模糊不清。不過大致上，Lord的結論仍可看出，第一題的安排方法對測量精確度並沒有多大的影響。

2. 項目選取與估計能力：根據 Hulin 等 (1983) 指出，適性測驗的進行，其項目選取的方式有下列四種：

① 個人能力與項目難度的匹配：即根據前面的反應組型估計出一個暫時的能力參數 ($\hat{\theta}$)，然後再據此 $\hat{\theta}$ ，選取尚未用過的項目中其難度與 $\hat{\theta}$ 最接近者，再實施之。

② 考慮猜測因素、個人能力與項目難度匹配：當無猜測因素影響時，此一方式與第一種方式是相同的。而此方式是根據 $\hat{\theta}$ ，再考慮猜測的因素，然後選擇項目難度是大於或等於 $\hat{\theta}$ 的題目。

③ 最大項目訊息：是根據 $\hat{\theta}$ 算出尚未使用項目的訊息，以最大者為下一題實施之。

④ 貝氏估計的項目選取：根據貝氏估計 (Bayesian estimation) 估計出 $\hat{\theta}$ ，然後再計算尚未使用的項目的降不確定 (reduction in uncertainty) 指數。這個指數用以表示貝氏估計方法對於能力估計的貢獻程度，越大表示其對於能力估計越有助益。與項目訊息指數有類似的意義。而貝氏估計在適性測驗項目選取上，是選取降不確定指數最大者的項目，為下一題實施之。

上述的四種選取項目的方式，除了第一種方式可以不在 LTT 進行，其餘三種均是在 LTT 之下，才有進行的可能。

由於適性測驗下，每個受試者所做的測驗項目都不同，如何計分，才有比較上的意義，這也是適性測驗上一個特殊的問題。Weiss (1974), Hulin 等 (1983) 指出適性測驗對於個人能力估計方法，大致也有四種：

① 以所做最後一題的項目難度做為最後的 $\hat{\theta}$ 。

② 以做完最後一題，再選一題適合的項目，但不對受試者實施，而以該項目難度做為最後的 $\hat{\theta}$ 。

③ 使用最大可能性法估計 (maximum likelihood estimate, 簡稱 MLE) $\hat{\theta}$ 。

④ 以貝氏估計方法，估計 $\hat{\theta}$ 。

以上四種估計能力的方法的共同特性，是 $\hat{\theta}$ 均建立在相同的量尺上，即使不同的受試者做的是不同測驗項目，估計出來的 $\hat{\theta}$ ，仍是可比較的。其中第一、二種方法，可以不在 LTT 之下進行，而第三、四種方法，則完全依附在 LTT 之下。

3. 結束測量：適性測驗一步一步估計能力，選取下一題再做，週而復始，如何停止，也是適性測驗上一個特殊的問題。

適性測驗結束的方法，主要有三 (Hulin, Drasgow & Parsons, 1983; Warm, 1978):

① 當指定的題數達到時便可以停止。

② 當測量標準誤已低於預定標準時可以停止。

③ 當未使用項目中，無法再提供有意義的項目訊息時便可停止。

通常適性測驗有可能因為方法上的限制，而使適性測驗所必須完成一樣數量但內容不一樣的測驗項目，因此它們並未涉及結束測量的問題。這種狀況在此不予討論。研究者要討論的是測量結束不定的問題。假定有一個能力測驗題庫有 200 題，它的項目參數事先均已校準 (calibrated)，項目參數都是已知，則可按前述的方法進行適性測驗的程序。如果我們選用的是一個測量結束不定的方法，則適性測驗也可能是在做了第 200 個項目後才結束。若果真如此，則與在 CTT 裏要做完整個測驗的結果是一樣的，所不同的是項目安排的順序。CTT 的安排，難度通常是由簡入難，而適性測驗的安

排則是從適合的項目到不適合的項目。因此，研究者要提出第一個適性測驗研究的概念：適性測驗研究其實就是研究如何安排測驗項目的順序，使測量工作達到最佳的狀況。所謂最佳狀況指的是用的題數少，測量精確性高。

上述結束測量的方法，在實際情況下，有它的必要性。但從測驗研究的角度上看，這些結束測量的方法，基本上必須假定當結束測量條件符合時，則測量已經到達最佳狀況，也就是說如果有機會再往下做，測量精確性不會再降低，但可能還會上升。這樣的假定，在適性測驗實際狀況下運作，可能有相當的風險。理由是真實測驗情境下，受試者的能力並無法事先預知。換言之，在適性測驗研究中，冒然決定測量結束的條件，可能會使研究結果產生誤解。因此面對這樣的基本假定，適性測驗無論是研究或是實際運作，其結束測量的問題，就不得不格外的慎重小心，而且有待進一步研究澄清的必要。

研究者基於上面提出的適性測驗研究的概念與上述基本假定，擬以電腦模擬研究，以適性測驗項目選取的方法，安排受試者按順序逐一接受適合的項目，直到題庫中的所有項目均被用過，同時記錄受試者適性測驗全程的所有變化，藉此澄清適性測驗測量結束問題。

(四) 適性測驗的實際具體策略：

根據 Weiss (1974) 及 Hulin 等 (1983)，適性測驗策略大致可區分為二。其一是兩段式策略 (two-stage strategy)，另一是多段式策略 (multi-stage strategy)，多段式策略又可分為固定分支模式 (fixed branch model) 及可變分支模式 (variable branch model)。

1. 兩段式策略：指的是讓受試者先做一個前導測驗 (routing test) 然後再根據前導測驗立即計分的結果，從幾個測量測驗 (measurement test) 中挑選一個適合的給受試者做。每個受試者做的都是同一個前導測驗，但測量測驗就不一定是同一個。這種策略計分方法有二，第一是以受試者答對題目之難度平均來計分，另一則是以 LTT 最大可能性的方法，估計個人能力參數。兩段式策略是適性測驗中最為簡單的策略。

2. 固定分支模式多階段策略：指的是根據受試者前一個測驗項目的答對答錯，來決定下一題。答對則難度升高，答錯則難度降低。至於升高難度多少或降低難度多少，是一個重要課題。前述上下法便是這一類的策略，升高降低難度的水準都一樣，而且前後也一致。高低法 (H-L method) 是上升小降低大，但前後一致；這是考慮猜測因素的策略。羅一門二氏法則是前面項目難度上升下降較大，而後面項目則上升下降難度趨小。這種策略的計分方法除了上述二種之外，尚可以最後一題的難度指數來估計個人能力。大部份的適性測驗都是屬於這種策略。

3. 可變分支模式多階段策略：主要的有最大項目訊息策略與貝氏估計策略。前者是以 $\hat{\theta}$ 計算項目訊息，選擇最大者做為下一題。後者是以 $\hat{\theta}$ 計算降不確定指數，亦是選擇最大者。這種策略直接着眼於測量觀點。它的計分方法主要的有最大可能性法及貝氏估計法。

上述三種適性測驗策略，若以它們應用 LTT 的程度大小排列，由小而大應該是二階段策略、固定分支模式多階段策略、可變分支模式多階段策略。前二者在項目選取上主要是在個人能力與項目難度匹配上做考慮與第三種策略是以 LTT 的另兩個指數做考慮。

上述三種策略的另一種差異是：前二者的題庫先依難度給予結構化，所以某種測驗反應順序產生某特定適性測驗路徑 (path)。題庫的結構化，導致適性測驗路徑固定而有限。這種現象在第三種策略不會有的 (Hambleton, 1985; Weiss, 1974)。Weiss 的這個說法恐怕有點疑慮，因為即使第三種策略，題庫不須結構化，它的適性測驗路徑恐怕也是固定而有限的。只是種類多，變化多一點罷了，在本研究中也試着去探究這些路徑問題。

(五) 適性測驗評估研究的特徵

LTT 的發展為適性測驗建立良好的理論基礎。研究適性測驗的學者也紛紛建立模式，從事模擬

研究或實證研究。為學者所最熱衷的研究主題，應該就是依適性測驗的結果與傳統測驗 (Conventional test) 的結果相比較，試圖透過比較以顯示適性測驗的優越性能。最具代表性的一個研究是美國測驗服務社 F. M. Lord 所領導的一個自 1977 年至 1980 年的計畫 (Kreitzberg & Jones, 1980)。這個研究的適性測驗策略是最大訊息策略，以 25 題做為結束測量點。它的結果主要是與傳統測驗中的單峯測驗 25 題結果相比較，以測驗訊息指數為依變項。結果發現在各水準能力上，適性測驗所得訊息指數是傳統測驗訊息指數的 2 倍，表示適性測驗的測量標準誤，僅是傳統測驗的 $1/\sqrt{2}$ 而已；測量是精確了些。

事實上，Lord (1968) 早已在 LTT 剛出現時 (Lord & Novick, 1968) 便以 LTT 從事適性測驗研究。結果以相對效能 (relative efficiency) 為指標，指出適性測驗對於高能力及低能力的受試者測量得更好。Lord (1977) 說明了何以適性測驗會比較適用於極能力的受試者。理由是傳統測驗尤其是單峯測驗的設計，通常最適合作為測量中等能力的受試者，而適性測驗可以在各種能力水準都做有效的測量。換言之，適性測驗適合極端能力受試者的測量，是因為傳統測驗在極端能力受試者測量得較差的緣故。因此，Lord 原先樂觀的想法，並不是建立在適性測驗對於各個能力水準上的比較，而是與傳統測驗比較之相對性觀點。因此，適性測驗比較適用於極端能力受試者測量的想法，有待進一步研究。

適性測驗研究的另一特徵是對於結束測量點的取決不太重視。從早期二階段策略，及固定分支模式多階段策略 (Hulin, Drasgow & Parsons, 1983; Lord, 1980a; Weiss, 1974)，根本不必去討論結束測量的問題。因為一旦題庫結構固定，適性測驗的題數就固定了，沒有必要取決何時結束測量。影響所至，使可變分支樣式多階段策略，亦都固定題數，做為取決結束測量的依據。例如 Kreitzberg & Jones (1980) 是 25 題，Stocking (1984) 是 20 題，Hulin 等 (1983) 是 25 題，McBride & Martin (1983) 是 30 題，Reckase (1983) 的 20 題，Urry (1977) 的 30 題。事實上 Warm (1978) 指出適性測驗最大的好處就是能夠在不減低測量精確性的狀況下，以最少的題數來測量出一個人的能力。Warm (1978) 也指出適性測驗只須要傳統測驗題數的 10% 到 50% 便可以獲得與傳統測驗同樣精確的效果。然而這題數的決定到底是多少呢？一個預定的統一的標準，適當嗎？適性測驗對於各個能力水準受試者測量性能一致嗎？都有待研究進一步澄清。另一個問題是研究者在前面結束測量一節所提的：用固定題數來決定結束測量。這樣的說法，在測量的穩定性上，是無法保證的，因為也許再多做一題，精確性又要降低了。

歸結上述，適性測驗研究特徵所產生的困境，研究者以為主要是因為它的研究方法所導致。前述的研究多半採橫斷研究，亦即在統一的條件下比較適性測驗與傳統測驗之間的不同。而對於適性測驗本身做系列性 (sequential) 的研究則較少。Green (1970) 早已推介以系列方法研究適性測驗，以了解全貌。

貳、研究問題與假設

一、研究問題

基於上述研究者對 LTT 與適性測驗上的探討與認知，研究者擬透過本研究回答下列諸項有關 LTT 與它在適性測驗應用上的問題：

1. LTT 的項目參數 (a , b , c 參數) 及個人能力參數 θ 與 CTT 中項目統計數 P , r_{bi} 及個人得分之間是否一致？
2. θ 已知的適性測驗的測驗路徑結構為何？
3. θ 已知的適性測驗的測量精確穩定性如何？
4. θ 未知的適性測驗的測驗路徑結構為何？
5. θ 未知的適性測驗的測量精確穩定性如何？

二、研究假設

根據上述所列問題，研究者提出下列研究假設，以供考驗分析之用。

- 1—1 LTT 的 θ 參數與估計的 $\hat{\theta}$ 有正相關存在。
- 1—2 LTT 的 θ 參數與 CTT 的個人得分有正相關存在。
- 1—3 LTT 的 a 參數與 CTT 的 r_{bi} 值有正相關存在。
- 1—4 LTT 的 b 參數與 CTT 的 P 值有負相關存在。
- 1—5 LTT 項目中 a ， b 參數相同， c 參數不同的二組測驗項目，其 r_{bi} 值有差異。
- 2—1 θ 已知時，受試者可以根據適性測驗路徑的 a 參數加以分類。
- 2—2 θ 已知時，受試者可以根據適性測驗路徑的 b 參數加以分類。
- 2—3 θ 已知時，受試者可以根據適性測驗路徑的 c 參數加以分類。
- 3—1 θ 已知時，適性測驗裏做完題庫所估得的 $\hat{\theta}_F$ 與每一步驟所估得的 $\hat{\theta}_1$ 之間差的絕對值呈單調性遞降趨近於 0。
- 3—2 θ 已知時，適性測驗裏做完題庫計分所得的答對百分比 P_F 與每一步驟計分所得的答對百分比 P_1 之間差異的絕對值呈單調性遞降趨近 0。
- 3—3 θ 已知時，適性測驗每一步驟估計 $\hat{\theta}_1$ 時，疊代法聚斂的受試者數佔全體受試者之百分比呈單調性遞升趨近於 100。
- 3—4 θ 已知時，適性測驗每一步驟估計所得測驗訊息 $I(\hat{\theta}_1)$ ，佔做完題庫所估得測驗訊息 $I(\hat{\theta}_F)$ 之百分比，呈單調性遞升趨近於 100。
- 4—1 θ 未知時，受試者可以根據適性測驗路徑的 a 參數加以分類。
- 4—2 θ 未知時，受試者可以根據適性測驗路徑的 b 參數加以分類。
- 4—3 θ 未知時，受試者可以根據適性測驗路徑的 c 參數加以分類。
- 5—1 θ 未知時，適性測驗做完題庫所估得的 $\hat{\theta}_F$ 與每一步驟所估得的 $\hat{\theta}_1$ 之間差的絕對值，呈單調性遞降趨近於 0。
- 5—2 θ 未知時，適性測驗做完題庫計分所得的答對百分比 P_F 與每一步驟計分所得的答對百分比 P_1 之間差異的絕對值，呈單調性遞降趨近於 0。
- 5—3 θ 未知時，適性測驗每一步驟估計 $\hat{\theta}_1$ 時，疊代法聚斂的受試者，佔全體受試者百分比，呈單調性遞升趨近於 100。
- 5—4 θ 未知時，適性測驗每一步驟估計所得之測驗訊息 $I(\hat{\theta}_1)$ ，佔做完題庫所估得測驗訊息 $I(\hat{\theta}_F)$ 之百分比，呈單調性遞升趨近 100。

方 法

一、研究架構

為解答本研究所提的問題，並驗證本研究各項假設，研究者根據三參數對數模式，利用預先設定項目參數的 130 個測驗項目所組成的「題庫」及預先設定個人能力的 610 位「受試者」，以電腦模擬方式產生每一受試者對 130 個項目之反應組型。並藉此 610 位受試者對 130 個項目的反應結果，從事下列分析：

(一) 估計個人能力 $\hat{\theta}$ ，計算個人得分，並對 130 個項目進行項目分析，以求得 P ， r_{bi} 等項目統計數，並研究比較其與預先設定之各項參數的關係。

(二) 使每一受試者在 θ 已知的情況下，逐一接受適性測驗，亦即根據 θ 選取項目，安排適性測驗步驟，以分析 θ 已知的適性測驗路徑結構及其測量精確穩定性。

(三) 使每一受試者在 θ 未知的情況下，逐一接受適性測驗，並根據每一步驟所估計的 $\hat{\theta}$ ，選取項目

，安排適性測驗步驟，以分析 θ 未知的適性測驗路徑結構及其測量精確穩定性。

基本上，本研究是在三參數對數模式，假想各類參數均為已知的狀況下，模擬受試者對測驗的反應組型，然後再根據模擬的資料，反估計各參數的估計值，以研究已知的參數與估計的參數之間的關係，這是 LTT 模擬研究的一個共同特色。

二、研究變項之操作定義

茲將本研究中各變項之操作型定義分列如下。

(一) LTT：係指三參數對數模式。其 IRF 如函數 1。

(二) θ ：是 LTT 中個人能力的參數。 $\hat{\theta}$ 是個人能力參數 θ 的估計值。在本研究中， θ 的範圍是 -3 到 +3，間格是 0.1，因此，自 -3，-2.9，-2.8...至 2.8，2.9，3.0 有 61 種能力水準。 θ 越高代表受試者能力越高。

(三) b ：是 LTT 項目難度參數。在研究中， b 的範圍是 -3 到 +3，間格是 0.5。因此，自 -3，-2.5，-2...至 2，2.5，3 有 13 種項目難度水準。 b 越高，項目越難。

(四) a ：是 LTT 中項目鑑別度參數。在本研究中， a 的範圍是 0 到 2，間格是 0.5，因此有 5 種項目鑑別度水準，分別是 0、0.5、1.0、1.5、2.0。 a 越高，表示項目鑑別性能越好。

(五) c ：是 LTT 中項目猜測參數。在本研究中， c 只有 0 及 0.25 二個水準。 c 越高，表示項目越容易因猜測而答對。

(六) 個人得分：是 CTT 中個人能力的指數，以受試者答對測驗的題數表示。在本研究中，個人得分是由 0 到 130，個分越高表示個人能力越高。

(七) r_{bis} ：為 CTT 中項目鑑別度指數，是指受試者在測驗項目上得分與其個人得分之間的二系列相關係數。 r_{bis} 越高表示項目鑑別性能越好。

(八) P ：為 CTT 中項目難度指數，是指全體受試者答對測驗項目的百分比。 P 越高，表示項目越簡易。

(九) 適性測驗路徑：是指受試者在接受適性測驗每一步驟所接受之項目所形成的記錄。本研究的題庫有 130 題。若將 130 題全按適性測驗原則呈現給受試者，則可能的適性測驗路徑有 130！種。

(十) θ 已知的適性測驗：是指受試者 θ 參數已知，以最大訊息法選取項目，以最大可能性法估計 $\hat{\theta}$ 的適性測驗。只是在每一步驟估計項目訊息時，是以已知的 θ 來估計。因此，只要 θ 一樣的受試者，其適性測驗路徑也必定一樣。

(十一) θ 未知的適性測驗，是指受試者 θ 參數未知，也是以最大訊息法選取項目，以最大可能性法估計 $\hat{\theta}$ 的適性測驗。在每一步驟估計項目訊息時，是以 $\hat{\theta}$ 來加以估計，因此只要受試者反應組型一樣，其適性測驗路徑也必定一樣。

(十二) $\hat{\theta}_F$ 與 $\hat{\theta}_i$ ：這裏 $\hat{\theta}_F$ 是指以適性測驗安排到第 130 步驟時，即用完題庫的 130 題時所估計的 $\hat{\theta}$ 。又 $\hat{\theta}_i$ 是指安排到第 i 步驟時，所估計到的 $\hat{\theta}$ 。而 $|\hat{\theta}_F - \hat{\theta}_i|$ 是 $\hat{\theta}_F$ 與 $\hat{\theta}_i$ 差的絕對值，用以表示安排到第 i 個測驗項目時，其估計的 $\hat{\theta}$ 與最後用完題庫所估計的 $\hat{\theta}$ ，符合的程度， $|\hat{\theta}_F - \hat{\theta}_i|$ 越小表示符合程度越高。

(十三) P_F 與 P_i ： P_F 是指以適性測驗安排到第 130 步驟，即用完題庫的 130 題，所得答對題數佔 130 題的百分比。 P_i 是指安排到第 i 步驟時，答對題數佔 i 題的百分比。 $|P_F - P_i|$ 是 P_F 與 P_i 差的絕對值，用以表示安排到第 i 個測驗項目，其答對題百分比，與最後用完題庫所得答對題百分比相符合的程度 $|P_F - P_i|$ 越小表示符合程度越高。

(十四) 項目訊息： $I(\theta, u_i) = P_i'(\theta)/P_i(\theta)Q_i(\theta)$ ，如公式 18，是只對某一指定項目 i 而言；其中 $P_i(\theta)$ 是個人答對項目 i 的機率， $Q_i(\theta) = 1 - P_i(\theta)$ ； $P_i'(\theta)$ 是 $P_i(\theta)$ 對 θ 的第一階導數，是指 $P_i(\theta)$ 斜率函數。本研究採用三參數對數模式，故 $I(\theta, u_i)$ 的計算公式如公式 19，其中

D^2 是常數，約等於 2.89。

$P_i = P_i(\theta)$, $Q_i = Q_i(\theta)$ ，而 a_i 及 c_i 均是項目參數。

$$I(\theta, u_i) = D^2 a_i^2 Q_i (P_i - c_i)^2 / (1 - c_i)^2 P_i \quad (\text{公式 19})$$

(ㄅ) 測驗訊息： $I(\theta, \hat{\theta})$ 是項目訊息的累加和，如公式 20，經證明其計算公式如公式 21，其中 n 表示測驗由 n 個項目所組成。

$$I(\theta, \hat{\theta}) = \sum_{i=1}^n \frac{P_i'^2(\theta)}{P_i(\theta) Q_i(\theta)} \quad (\text{公式 20})$$

$$= \sum_{i=1}^n D^2 a_i^2 Q_i (P_i - c_i)^2 / (1 - c_i)^2 P_i \quad (\text{公式 21})$$

(ㄆ) $I(\hat{\theta}_F)$ 與 $I(\hat{\theta}_i)$ ： $I(\hat{\theta}_F)$ 是指以適性測驗安排到 130 步驟，即用完題庫的 130 題，估計的「測驗訊息」。 $I(\hat{\theta}_i)$ 是指安排到第 i 步驟，所估計的測驗訊息。 $I(\hat{\theta}_i)$ 對 $I(\hat{\theta}_F)$ 的百分比，用以表示安排到第 i 個測驗項目時，測量精確的程度，越接近 100%，表示測量越精確。

(ㄇ) 最大可能性法：用來估計 $\hat{\theta}$ 的方法，本研究係使用牛頓拉福森 (Newton Raphson) 數值分析的疊代法。主要有下列計算步驟：

1. 起始值： θ_{0a} 是指第 a 個受試者估計 $\hat{\theta}$ 時的起始值。其計算公式如公式 22，其中 n 是測驗項目

$$\theta_{0a} = \ln[r_a / (n - r_a)] \quad (\text{公式 22})$$

數， r_a 是第 a 個受試者答對題數。

2. 接近值： h_m 是指疊代法第 m 個步驟，所用的接近值，它的計算公式如公式 23：

$$h_m = \frac{D \sum_{i=1}^n a_i (u_{ia} - P_{ia}) (P_{ia} - c_i) / P_{ia} (1 - c_i)}{D^2 \sum_{i=1}^n a_i^2 (P_{ia} - c_i) (u_{ia} - P_{ia})^2 Q_{ia} / P_{ia} (1 - c_i)^2} \quad (\text{公式 23})$$

其中 n 是指測驗項目數， D 是常數 1.7， a_i 、 c_i 均是項目參數。 u_{ia} 是指第 a 個受試者答第 i 個項目的個分， P_{ia} 是指第 a 個受試者答對第 i 個項目正確反應的機率。

3. 估計值： $\hat{\theta}_{m+1}$ 是指疊代法第 $m+1$ 個步驟時，所估計的 $\hat{\theta}$ ，它的計算公式如公式 24，其中 $\hat{\theta}_m$ 及 h_m 分別是第 m 個步驟時，所估計的 $\hat{\theta}$ 及接近值。

$$\hat{\theta}_{m+1} = \hat{\theta}_m - h_m \quad (\text{公式 24})$$

4. 疊代法聚斂：在本研究中，當 h_m 的絕對值小於 .001 即表聚斂，否則要重覆上述 2、3 步驟，直到 h_m 的絕對值小於 .001 為止。但本研究中所用的方法，並不一定會聚斂，也就是 h_m 的絕對值並不一定會遞減，而且降到 0.001 以下。如果不聚斂，則 $\hat{\theta}$ 便無法估計。

(ㄏ) 最大訊息法：本研究選擇適性測驗的項目時所採用的方法。在 θ 已知的情況下，就以 θ 計算其它未使用項目的項目訊息，然後選取項目訊息最大者做為下一個題目。在 θ 未知的情況下，是以適性測驗每一步驟所估計的 $\hat{\theta}$ 來計算項目訊息，然後選取最大者，做為下一題。

三、研究程序

(一) 研究工具的選取

本研究中受試者反應資料的產生，及適性測驗的模擬均在電腦協助下進行。

1. 硬體選取：由於電腦模擬研究所耗 CPU 時間相當龐大，和計算機中心正常開機及關機時間不足且無法配合，乃選擇師大教育心理系的 IBM PC-XT 個人電腦，做為模擬研究使用的機器。

2. 軟體語言的選擇：無論是模擬資料或是適性測驗的模擬，均有龐大且複雜的計算工作。因此研

研究者選擇 Fortran 77 做為模擬程式撰寫的主要語言。以 MS-Fortran 3.20 做為編譯程式，並在個人電腦上裝置 Intel 公司的 8087 輔助處理器，以便聯結 8087. LIB 的程式庫，以增加計算的速度及準確度。

3. 亂數副程式的選擇：本研究受試者反應資料的產生部分，必須使用齊次分配亂數。研究者乃選用 LINENUM 及 STRNUM (取自 Fortran scientific subroutine library, John Wiley & Sons 公司, 1984) 兩個齊次亂數產生副程式。並令其各自產生 1000 個界於 1 到 10 的整數，再以卡方適合度考驗，考驗其是否符合齊次分配。同樣的工作，進行二十次，以觀察亂數產生的穩定性。結果如表 2：

表 2 齊次分配亂數產生副程式模擬資料之適合度考驗表

次 數	STRNUM	χ^2	LINENUM	次 數	STRNUM	χ^2	LINENUM
(1)	4.14		9.52	(11)	8.54		7.56
(2)	12.62		2.00	(12)	9.38		13.64
(3)	7.68		15.00	(13)	2.34		11.14
(4)	13.02		7.16	(14)	3.74		17.76*
(5)	6.50		8.00	(15)	13.16		12.12
(6)	11.58		17.56*	(16)	14.82		13.08
(7)	2.96		9.34	(17)	12.20		11.76
(8)	11.82		11.54	(18)	8.00		6.44
(9)	6.36		8.84	(19)	10.38		10.24
(10)	23.50*		6.04	(20)	7.08		11.20
				合 併	6.353		37.803*
df=9				* P < .05			

由表 2 的結果可知 STRNUM 副程式顯得較穩定，故本研究採用 STRNUM 做為齊次分配亂數產生副程式。

(二) LTT 理論模式的選取

研究者考慮配合實際上選擇式測驗的情境，因而選擇三參數對數模式做為本研究的理論模式。在估計個人能力 θ 時，是使用最大可能性法；在適應測驗的項目選取上，是使用最大訊息法，其計分方式是採最大可能性法估計 $\hat{\theta}$ 。

(三) 題庫的結構

本研究所使用的測驗題庫是一假想性的題庫，而且題數及項目參數均由研究者事先設定，不須再行校準。

根據 Warm (1978) 所說，a 參數通常界於 0 及 +2，b 參數通常界於 -3 及 +3，c 參數通常界於 0 及 +0.3。因此在本研究中，採用的測驗項目均是在上述範圍之內。正如操作定義中所提，本研究中 a 參數有 5 個水準，b 參數有 13 個水準，c 參數有 2 種。它們之間的組合會產生 $5 \times 13 \times 2 = 130$ 種，故本研究所用題庫，是 130 題，其詳細項目參數如表 3。

表 3 題庫項目參數表

題 號	a	b	c	題 號	a	b	c
1	2.00	3.00	0.00	66	2.00	3.00	0.25
2	2.00	2.50	0.00	67	2.00	2.50	0.25
3	2.00	2.00	0.00	68	2.00	2.00	0.25
4	2.00	1.50	0.00	69	2.00	1.50	0.25
5	2.00	1.00	0.00	70	2.00	1.00	0.25
6	2.00	0.50	0.00	71	2.00	0.50	0.25
7	2.00	0.00	0.00	72	2.00	0.00	0.25
8	2.00	-0.50	0.00	73	2.00	-0.50	0.25
9	2.00	-1.00	0.00	74	2.00	-1.00	0.25
10	2.00	-1.50	0.00	75	2.00	-1.50	0.25
11	2.00	-2.00	0.00	76	2.00	-2.00	0.25
12	2.00	-2.50	0.00	77	2.00	-2.50	0.25
13	2.00	-3.00	0.00	78	2.00	-3.00	0.25
14	1.50	3.00	0.00	79	1.50	3.00	0.25
15	1.50	2.50	0.00	80	1.50	2.50	0.25
16	1.50	2.00	0.00	81	1.50	2.00	0.25
17	1.50	1.50	0.00	82	1.50	1.50	0.25
18	1.50	1.00	0.00	83	1.50	1.00	0.25
19	1.50	0.50	0.00	84	1.50	0.50	0.25
20	1.50	0.00	0.00	85	1.50	0.00	0.25
21	1.50	-0.50	0.00	86	1.50	-0.50	0.25
22	1.50	-1.00	0.00	87	1.50	-1.00	0.25
23	1.50	-1.50	0.00	88	1.50	-1.50	0.25
24	1.50	-2.00	0.00	89	1.50	-2.00	0.25
25	1.50	-2.50	0.00	90	1.50	-2.50	0.25
26	1.50	-3.00	0.00	91	1.50	-3.00	0.25
27	1.00	3.00	0.00	92	1.00	3.00	0.25
28	1.00	2.50	0.00	93	1.00	2.50	0.25
29	1.00	2.00	0.00	94	1.00	2.00	0.25
30	1.00	1.50	0.00	95	1.00	1.50	0.25
31	1.00	1.00	0.00	96	1.00	1.00	0.25
32	1.00	0.50	0.00	97	1.00	0.50	0.25
33	1.00	0.00	0.00	98	1.00	0.00	0.25
34	1.00	-0.50	0.00	99	1.00	-0.50	0.25
35	1.00	-1.00	0.00	100	1.00	-1.00	0.25
36	1.00	-1.50	0.00	101	1.00	-1.50	0.25
37	1.00	-2.00	0.00	102	1.00	-2.00	0.25
38	1.00	-2.50	0.00	103	1.00	-2.50	0.25
39	1.00	-3.00	0.00	104	1.00	-3.00	0.25
40	0.50	3.00	0.00	105	0.50	3.00	0.25
41	0.50	2.50	0.00	106	0.50	2.50	0.25
42	0.50	2.00	0.00	107	0.50	2.00	0.25
43	0.50	1.50	0.00	108	0.50	1.50	0.25
44	0.50	1.00	0.00	109	0.50	1.00	0.25
45	0.50	0.50	0.00	110	0.50	0.50	0.25
46	0.50	0.00	0.00	111	0.50	0.00	0.25
47	0.50	-0.50	0.00	112	0.50	-0.50	0.25
48	0.50	-1.00	0.00	113	0.50	-1.00	0.25
49	0.50	-1.50	0.00	114	0.50	-1.50	0.25
50	0.50	-2.00	0.00	115	0.50	-2.00	0.25
51	0.50	-2.50	0.00	116	0.50	-2.50	0.25
52	0.50	-3.00	0.00	117	0.50	-3.00	0.25
53	0.00	3.00	0.00	118	0.00	3.00	0.25
54	0.00	2.50	0.00	119	0.00	2.50	0.25
55	0.00	2.00	0.00	120	0.00	2.00	0.25
56	0.00	1.50	0.00	121	0.00	1.50	0.25
57	0.00	1.00	0.00	122	0.00	1.00	0.25
58	0.00	0.50	0.00	123	0.00	0.50	0.25
59	0.00	0.00	0.00	124	0.00	0.00	0.25
60	0.00	-0.50	0.00	125	0.00	-0.50	0.25
61	0.00	-1.00	0.00	126	0.00	-1.00	0.25
62	0.00	-1.50	0.00	127	0.00	-1.50	0.25
63	0.00	-2.00	0.00	128	0.00	-2.00	0.25
64	0.00	-2.50	0.00	129	0.00	-2.50	0.25
65	0.00	-3.00	0.00	130	0.00	-3.00	0.25

(四) 受 試 者

本研究所指的受試者，亦是一假想性的受試者。他們的能力參數 θ 及人數，亦由研究者事先設定。本研究中 θ 的範圍，如操作型定義所指是界於 -3 及 +3，間格為 0.1，共計有 61 個能力水準。本研究並設定每一個能力水準均有 10 位受試者，故共計模擬有 610 位受試者。受試者的能力水準、組別

詳如表 4。

(戊) 受試者反應組型

本研究受試者反應組型，意指本研究中的 610 位受試者，對題庫中 130 個項目的作答結果的組型。此反應型是經由電腦模擬產生。其步驟如下：

1. 將受試者的 θ 及項目參數代入三參數對數模式的函數求得 $P(\theta)$ 。
2. 利用 STRNUM 齊次分配亂數產生副程式，隨機產生一個界於 0 及 1 之間的亂數，若此亂數小於或等於 $P(\theta)$ ，則便算此受試者答對。若此亂數大於 $P(\theta)$ 則算答錯。

舉例來說，本研究中的第一位受試者的 $\theta = 3$ 。他對題庫中的第一題（它的三個參數是 $a = 2.0$

表 4 受試者能力水準、組別及編號表

θ	能力水準	受試號碼	θ	能力水準	受試號碼
3	1	1 ~ 10	-0.1	32	311 ~ 320
2.9	2	11 ~ 20	-0.2	33	321 ~ 330
2.8	3	21 ~ 30	-0.3	34	331 ~ 340
2.7	4	31 ~ 40	-0.4	35	341 ~ 350
2.6	5	41 ~ 50	-0.5	36	351 ~ 360
2.5	6	51 ~ 60	-0.6	37	361 ~ 370
2.4	7	61 ~ 70	-0.7	38	371 ~ 380
2.3	8	71 ~ 80	-0.8	39	381 ~ 390
2.2	9	81 ~ 90	-0.9	40	391 ~ 400
2.1	10	91 ~ 100	-1.0	41	401 ~ 410
2.0	11	101 ~ 110	-1.1	42	411 ~ 420
1.9	12	111 ~ 120	-1.2	43	421 ~ 430
1.8	13	121 ~ 130	-1.3	44	431 ~ 440
1.7	14	131 ~ 140	-1.4	45	441 ~ 450
1.6	15	141 ~ 150	-1.5	46	451 ~ 460
1.5	16	151 ~ 160	-1.6	47	461 ~ 470
1.4	17	161 ~ 170	-1.7	48	471 ~ 480
1.3	18	171 ~ 180	-1.8	49	481 ~ 490
1.2	19	181 ~ 190	-1.9	50	491 ~ 500
1.1	20	191 ~ 200	-2.0	51	501 ~ 510
1.0	21	201 ~ 210	-2.1	52	511 ~ 520
0.9	22	211 ~ 220	-2.2	53	521 ~ 530
0.8	23	221 ~ 230	-2.3	54	531 ~ 540
0.7	24	231 ~ 240	-2.4	55	541 ~ 550
0.6	25	241 ~ 250	-2.5	56	551 ~ 560
0.5	26	251 ~ 260	-2.6	57	561 ~ 570
0.4	27	261 ~ 270	-2.7	58	571 ~ 580
0.3	28	271 ~ 280	-2.8	59	581 ~ 590
0.2	29	281 ~ 290	-2.9	60	591 ~ 600
0.1	30	291 ~ 300	-3.0	61	601 ~ 610
0.0	31	301 ~ 310			

， $b = 3.0$ ， $c = 0.0$ ），代入模式函數中，得 $P_1(3) = 0.50$ 。此時倘若 STRNUM 產生之亂數小於或等於 0.50，則第一位受試者答對第一題。同樣的方法，第一位受試者答第二題， $P_2(3) = 0.85$ 。因此，如果 STRNUM 產生的亂數小於或等於 0.85，便算答對第二題。如此反覆產生 $P_i(\theta)$ 及齊次分配亂數，便可決定本研究中 610 受試者在題庫中 130 個項目的反應組型，做為本研究的基本資料。

(己) 模擬研究資料蒐集

1. 研究假設一的資料蒐集

(1) 將上述受試者的反應組型，以最大可能性法估計 $\hat{\theta}$ ，得 610 個 $\hat{\theta}$ 的資料。

(2) 將反應組型資料以項目分析的方法求得 130 個項目的難度指數 (P) 及鑑別度指數 (r_{bis})。

2. 研究假設二、三的資料蒐集 (θ 已知的適性測驗)

(1) 將受試者已知的 θ 代入項目訊息函數，求取所有項目的項目訊息，選取最大者做為第一題。然後從受試者的反應組型中取出對應題號的反應結果，做為適性測驗的反應。記錄題號及計算答對題百分比 P_1 及第一步驟的測驗訊息 ($I(\hat{\theta})$)。

(2) 同樣的，以 θ 代入項目訊息函數，求取所有項目的項目訊息，選取未使用過的最大訊息者，做為第二題（或第 k 題， $k \geq 2$ ）。然後從受試者的反應組型中，取出對應題號的反應結果，

做為適性測驗的反應。記錄題號及計算答對百分比 P_k 及第 k 步驟的測驗訊息。在第二題以後，只要適性測驗結果出現有對有錯時，便可以最大可能性法估計 $\hat{\theta}_k$ ，若估計 $\hat{\theta}_k$ 時，疊代法不聚斂，則視同資料喪失。

(3) 重覆第(2)步驟，直到 $k=130$ 為止。記錄題號計算 P_F 、 $\hat{\theta}_F$ 及 $I(\hat{\theta}_F)$ 等資料，以備資料分析之用。

3. 研究假設三、四的資料蒐集 (θ 未知的適性測驗)

(1) 適性測驗的第一題使用題庫中的第7題，其 $a=2.0$ 、 $b=0.0$ 、 $c=0.0$ 。然後取出受試者反應組型中的第7題結果，做為適性測驗的反應，記錄題號並計算 P_1 。

(2) 適性測驗的第二題是根據第一題的結果來決定。第一題答對，則第二題使用題庫中的第6題，其 $a=2.0$ 、 $b=0.5$ 、 $c=0.0$ 。若第一題答錯，則使用題庫中的第8題，其 $a=2.0$ 、 $b=-0.5$ 、 $c=0.0$ 。然後從受試者反應組型中，取出對應題號的反應結果，做為適性測驗的反應，記錄題號，計算 P_2 。倘若前二題的結果是一對一錯，則可使用最大可能性法估計 $\hat{\theta}_2$ ，並計算測驗訊息 $I(\hat{\theta}_2)$ 。

(3) 若第一、二題都答對或都答錯，則第三題的選取，如同第(2)步驟。如果一直答對，則項目選取分別是題庫中的第5、4、3、2、1、14、27、40、53、66、79、92、105、118題。如果一直答錯，則項目選取分別是題庫中的第9、10、11、12、13、26、39、52、65、78、91、104、117、130。若受試者一直答對或一直答錯，而且超出上述的範圍則視同資料喪失，不再估計 $\hat{\theta}$ 。

(4) 若在第 k 題出現有對有錯的反應，則可使用最大可能性法估計 $\hat{\theta}_k$ ，並計算 $I(\hat{\theta}_k)$ 及 P_k ，和記錄題號，並以 $\hat{\theta}_k$ 代入項目訊息函數，求取所有項目的項目訊息；選取未使用過的最大訊息，做為下一題。

(5) 若在第 $k+1$ 題以後，估計 $\hat{\theta}_{k+1}$ 時不聚斂，則視同資料喪失。沒有 $\hat{\theta}_{k+1}$ ，便無法以最大訊息法，選取第 $k+2$ 題。在本研究中，處理方法是以公式 $\hat{\theta} \pm \frac{1}{2} |\hat{\theta} - b_{k+1}|$ 暫代 $\hat{\theta}_{k+1}$ ；若第 $k+1$ 題答對則公式中用+，答錯則用-，據此暫代 $\hat{\theta}_{k+1}$ 。然後，代入項目訊息函數，求得所有項目的項目訊息；選取未使用過的最大訊息者，做為下一題。

(6) 重覆第(4)(5)步驟，直到 $k=130$ 為止。記錄題號，並計算 P_F 、 $\hat{\theta}_F$ 及 $I(\hat{\theta}_F)$ 等資料，以備資料分析之用。

4. 電腦模擬程式

(1) 第一項資料蒐集，約耗電腦時間2小時。

(2) 第二項資料蒐集，約耗電腦時間120小時。

(3) 第三項資料蒐集，約耗電腦時間150小時。

四、資料處理

本研究的資料處理流程，大致可分為五個步驟：

(一) 資料蒐集及初步整理：均在師大教育心理系的 IBM 個人電腦進行，所使用的模擬程式及項目分析程式均由研究者自行撰寫及測試完成。

(二) 假設一的資料分析：研究者將個人電腦中所蒐集的資料上傳 (upload) 到師大電子計算中心的 PRIME-750，並使用 SPSSX 1.1 中的 PEARSON CORR 及 T-TEST 兩副程式進行假設一中的各項考驗。

(三) 資料轉換：研究者將適性測驗中所記錄的資料上傳到師大電子計算中心的 PRIME-750，並自行編寫程式從事下列的資料轉換工作：

1. 計算610位受試者在適性測驗每個步驟的 $|\hat{\theta}_F - \hat{\theta}_1|$ 、 $|P_F - P_1|$ 、 $I(\hat{\theta}_1) / I(\hat{\theta}_F) \times 100$ 。

2. 將610位受試者適性測驗路徑的題號轉換為項目參數，每位受試者便可得到三種參數的路徑，

分別是 a 的路徑 b 的路徑、及 c 的路徑。

3. 將 610 位受試者的 610×3 個路徑按能力水準相同者給予合併，得 61×3 個路徑。

4. 並將資料轉換的結果下傳 (download) 到個人電腦。

(四) 假設二、四的資料分析：

將上述的轉換資料再上傳到教育部電子計算機中心的 IBM 4341。使用 SPSSX 2.1 的 CLUSTER 副程式選擇完全連鎖法 (complete linkage method) 及歐氏距離 (Euclidean distance)，將受試者的適性測驗路徑，分別以路徑中的 a 參數、b 參數、c 參數進行路徑的羣聚分析，並繪出樹狀圖 (dendrogram)。

(五) 第二次資料轉換及假設三、五的資料分析。

將 $|\hat{\theta}_F - \hat{\theta}_I|$ ， $|P_F - P_I|$ ， $I(\hat{\theta}_I) / I(\hat{\theta}_F) \times 100$ 及 61×3 種適性測驗路徑，再依照羣聚分析的分類結果予以合併，簡化為少數幾類，以利假設三、五的分析比較。第二次資料轉換工作是在 PRIME-750 之下完成。第二次轉換的資料再上傳到工業技術學院電子計算機中心的 VAX-780 之下的 CALCOMP 965 繪圖機繪圖，完成分析。

結 果

一、LTT 各項參數與 CTT 各項統計數的關係

(一) 設定的個人能力參數 θ 與用最大可能性法估計的 $\hat{\theta}$ 積差相關係數是 .9936， $P < .01$ 。分佈圖如圖 8 所示。結果支持研究假設 1—1。

(二) 設定的個人能力參數 θ 與 CTT 的個人得分，積差相關係數是 .9841， $P < .01$ ，分佈圖如圖 9 所示。結果支持研究假設 1—2。

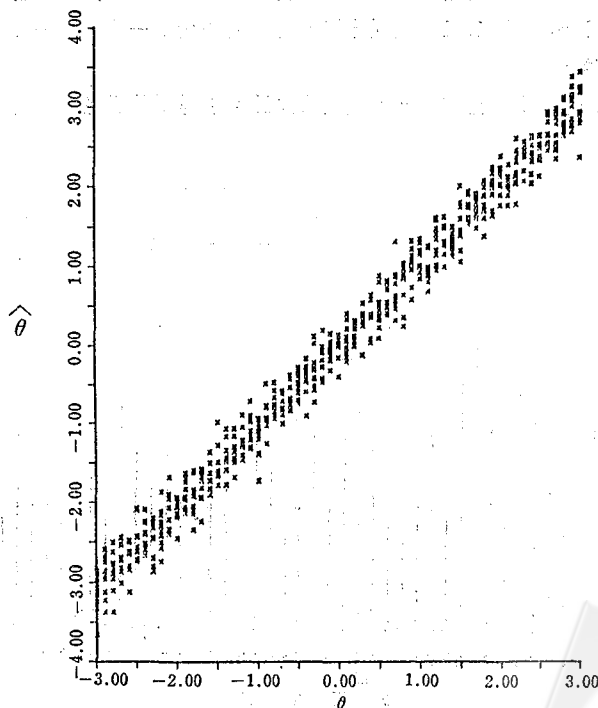


圖 8 $\hat{\theta}$ 與 θ 之分佈圖

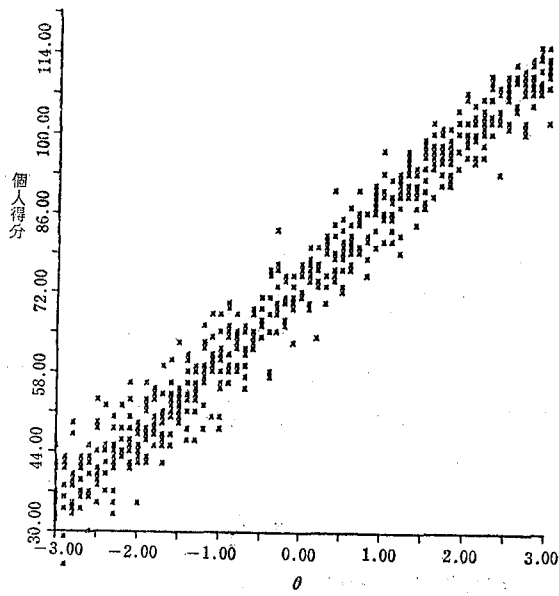


圖 9 θ 與個人得分之分佈圖

(三) 項目參 a 與 CTT 的 r_{bis} ，積差相關係數為 .7001， $P < .01$ 。參數 b 與 CTT 的 P ，積差相關係數為 $-.8356$ ， $P < .01$ 。參數 c 為 0.25 的項目，其 r_{bis} 的平均數是 0.6308，標準差是 0.360； c 為 0.00 的項目，其 r_{bis} 的平均數是 0.7466，標準差 0.303，二個 r_{bis} 之間差異 t 考驗，前者顯著地高於後者， $t(128) = 2.64$ ， $P < .01$ 。結果分別支持研究假設 1—3，1—4，1—5。

二、「 θ 已知」的適性測驗

所謂「 θ 已知」主要是指使用已知的 θ ，代入公式 19。求得項目訊息，以便選取項目進行適性測驗。

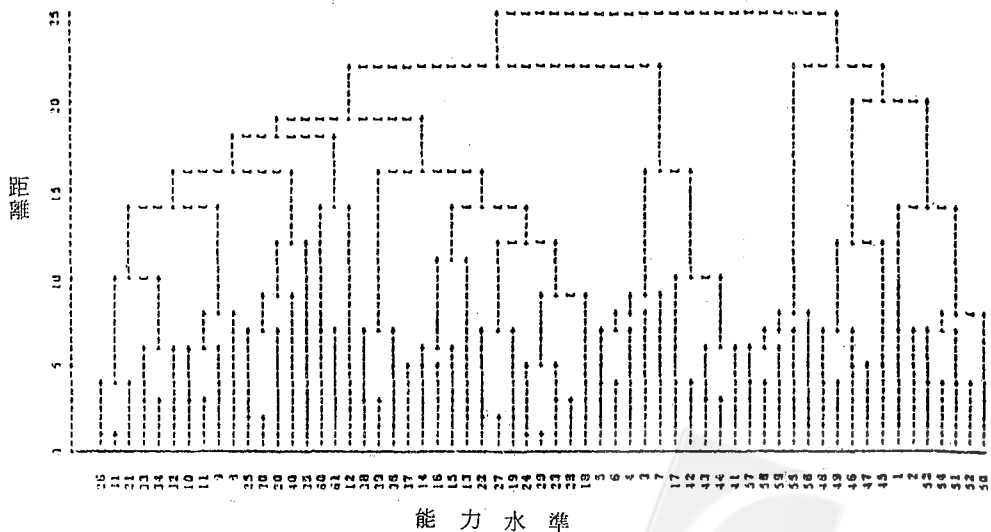


圖 10 θ 已知的適性測驗路徑分類樹狀圖 (以 a 參數分類)

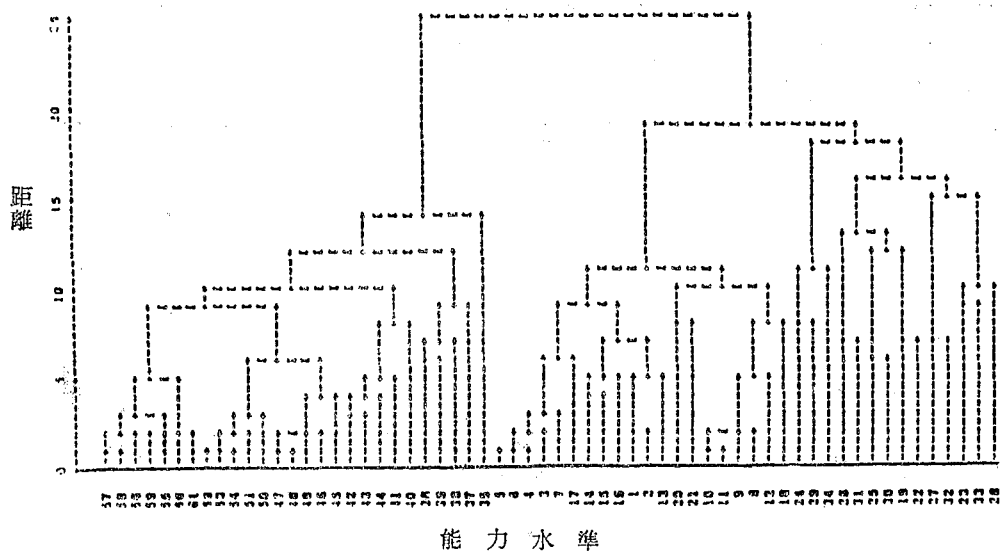


圖 11 θ 已知的適性測驗路徑分類樹狀圖 (以 b 參數分類)

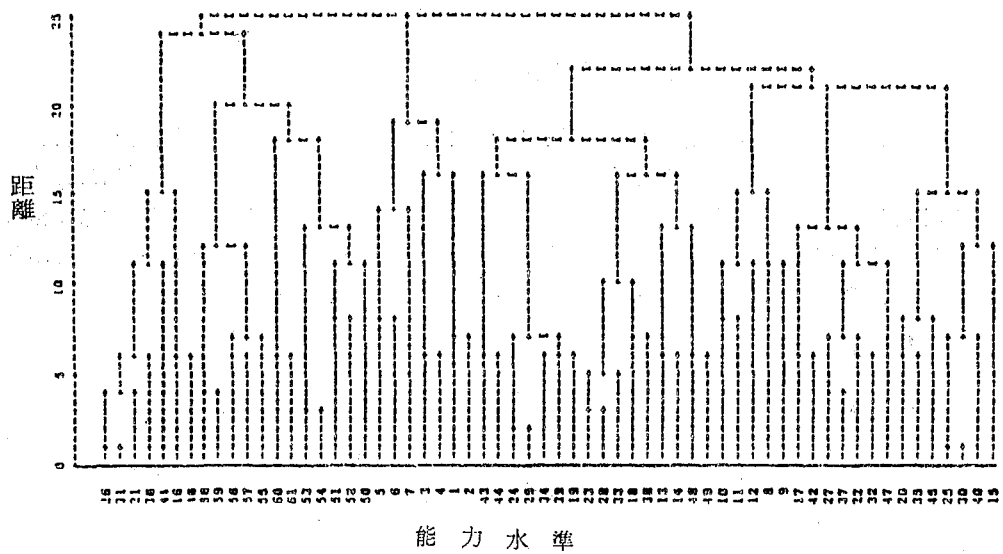


圖 12 θ 已知的適性測驗路徑分類樹狀圖 (以 c 參數分類)

(一) 適性測驗路徑的羣聚分析

將61個能力水準的受試者分別根據其路徑中的130項目參數進行羣聚分析，結果得樹狀圖如圖10、圖11、圖12所示。

由圖10和圖12羣聚分析結果所畫的樹狀圖看出： θ 已知的適性測驗路徑，以參數 a 和以參數 c 分類時的結果並不簡明。只有以參數 b 分類時 (圖11) 結果較為清楚，可以將適性測驗路徑分為三種類型。第一類是能力水準1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21 等20個能力水準所組成， θ 是界於1.0與3.0為高能力組。第二類是能力水準19,22,23,24,25,26,27,28,29,30,31,32，

33,34等14個能力水準所組成， θ 是界於-0.3與1.2為中能力組。第三類是能力水準35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61等27 個能力水準所組成， θ 界於-0.4與-3.0為低能力組。

至於這三種能力組受試者的適性測驗路徑究竟如何呢？進一步就 b 參數分析其變化如圖13所示。

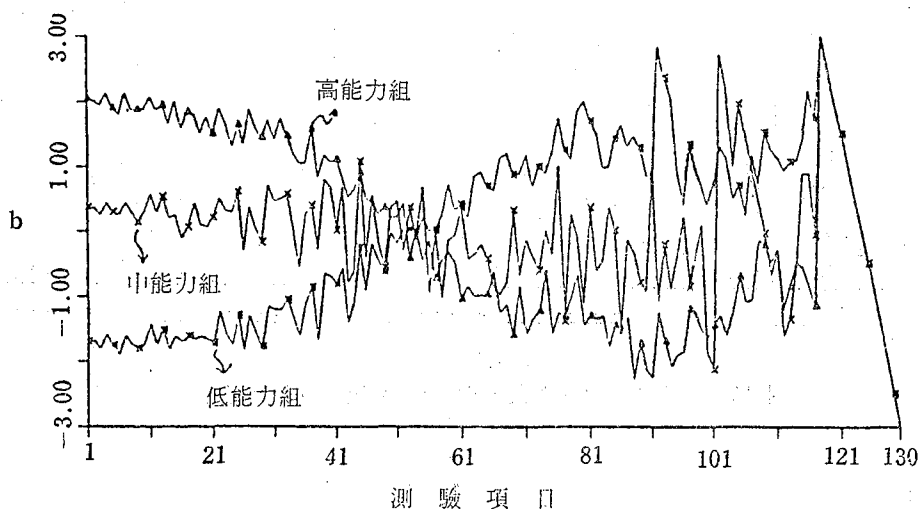


圖 13 三類適性測驗路徑在 b 參數的變化圖

由圖13可以看出，高能力組的適性測驗路徑在 b 參數上呈現由大而小的走向；低能力組則成由小而大的走向；中能力組的路徑則是在 0 上下跳動。另外這三類適性測驗在 a 參數及 c 參數的變化情形，如圖14、圖15所示。結果三種能力組的路徑在參數 a 及參數 c 上是分不清楚的。但是，當適性測驗在安排到118題以後，所選取的項目不論是那一類路徑，其 a 參數一定是 0，而 c 參數一定是 0.25。

上述結果只能部分支持研究假設 2—2，而未能支持假設 2—1 及 2—3。

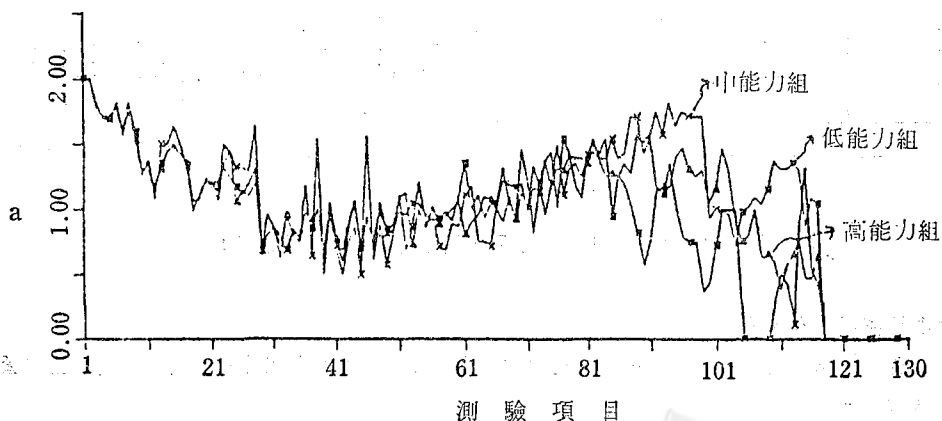


圖 14 三類適性路徑在 a 參數的變化圖

(二) 適性測驗精確穩定性分析

根據羣聚分析結果所得到的三類 θ 已知時的適性測驗路徑，進一步分析其測量精確穩定性如下：

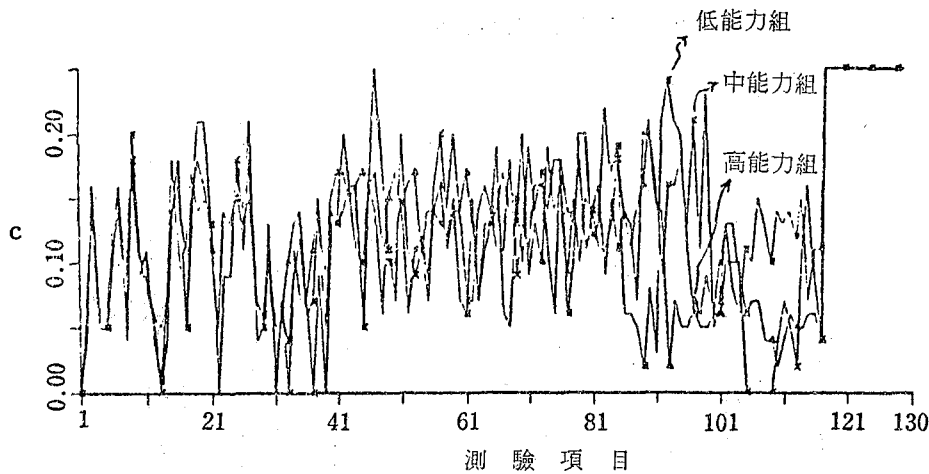


圖 15 三類適性路徑在 c 參數的變化圖

1. $|\hat{\theta}_F - \hat{\theta}_I|$ 的結果如表 5 及圖 16 所示。圖中顯示三條曲線相當接近，表示不論那一類適性測驗路徑，在每一步驟下所估計的 $\hat{\theta}_I$ 與 $\hat{\theta}_F$ 的差距，都是相近的。但是，在安排到 11 題以前，高能力組與低能力組的 $|\hat{\theta}_F - \hat{\theta}_I|$ 均呈現上下跳動的現象。過了 11 題，三組均呈現遞降趨近於 0。此

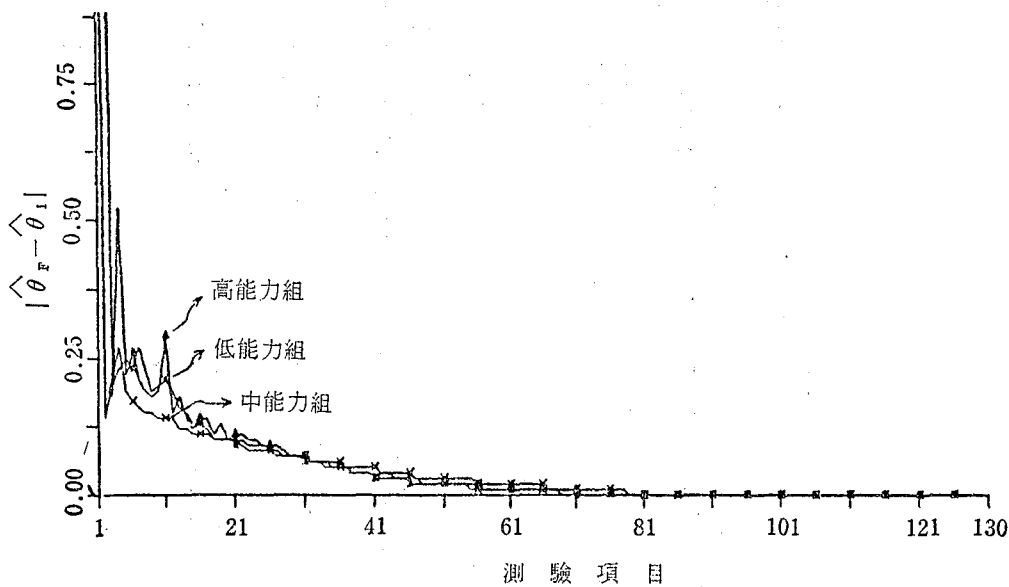


圖 16 三類適性測驗路徑在 $|\hat{\theta}_F - \hat{\theta}_I|$ 的變化圖

結果可支持研究假設 3-1。

2. $|P_F - P_I|$ 的結果如圖 17 所示。圖中顯示：高能力組及低能力組的路徑在每一步驟所估計的 P_I 與 P_F 的差距，都比中能力組的此項差距大。然而三條曲線，在安排第三題以後，均呈現遞降並趨向於 0 的現象。結果可以支持研究假設 3-2。



表 5 θ 已知的適性測驗 $|\hat{\theta}_F - \hat{\theta}_I|$ 值

項目	高能力組	中能力組	低能力組	項目	高能力組	中能力組	低能力組
1				66	0.01	0.02	0.01
2				67	0.01	0.01	0.01
3	0.18	0.21	0.20	68	0.01	0.01	0.01
4	0.52	0.27	0.23	69	0.01	0.01	0.00
5	0.22	0.19	0.25	70	0.01	0.01	0.00
6	0.26	0.17	0.23	71	0.01	0.01	0.00
7	0.21	0.16	0.27	72	0.00	0.01	0.00
8	0.19	0.16	0.23	73	0.00	0.01	0.00
9	0.18	0.15	0.19	74	0.00	0.01	0.00
10	0.19	0.14	0.20	75	0.00	0.01	0.00
11	0.29	0.14	0.21	76	0.00	0.01	0.00
12	0.15	0.14	0.19	77	0.00	0.01	0.00
13	0.18	0.12	0.16	78	0.00	0.01	0.00
14	0.14	0.12	0.16	79	0.00	0.00	0.00
15	0.12	0.11	0.13	80	0.00	0.00	0.00
16	0.14	0.11	0.13	81	0.00	0.00	0.00
17	0.14	0.11	0.12	82	0.00	0.00	0.00
18	0.11	0.10	0.10	83	0.00	0.00	0.00
19	0.13	0.10	0.10	84	0.00	0.00	0.00
20	0.10	0.10	0.10	85	0.00	0.00	0.00
21	0.11	0.10	0.09	86	0.00	0.00	0.00
22	0.11	0.09	0.10	87	0.00	0.00	0.00
23	0.10	0.08	0.09	88	0.00	0.00	0.00
24	0.10	0.08	0.09	89	0.00	0.00	0.00
25	0.09	0.08	0.09	90	0.00	0.00	0.00
26	0.09	0.08	0.08	91	0.00	0.00	0.00
27	0.09	0.07	0.08	92	0.00	0.00	0.00
28	0.08	0.07	0.07	93	0.00	0.00	0.00
29	0.07	0.07	0.07	94	0.00	0.00	0.00
30	0.07	0.07	0.07	95	0.00	0.00	0.00
31	0.07	0.07	0.06	96	0.00	0.00	0.00
32	0.06	0.06	0.06	97	0.00	0.00	0.00
33	0.06	0.06	0.06	98	0.00	0.00	0.00
34	0.06	0.06	0.05	99	0.00	0.00	0.00
35	0.05	0.06	0.05	100	0.00	0.00	0.00
36	0.05	0.06	0.05	101	0.00	0.00	0.00
37	0.04	0.05	0.05	102	0.00	0.00	0.00
38	0.04	0.05	0.04	103	0.00	0.00	0.00
39	0.04	0.05	0.04	104	0.00	0.00	0.00
40	0.04	0.05	0.04	105	0.00	0.00	0.00
41	0.03	0.05	0.03	106	0.00	0.00	0.00
42	0.04	0.04	0.03	107	0.00	0.00	0.00
43	0.04	0.04	0.03	108	0.00	0.00	0.00
44	0.03	0.04	0.03	109	0.00	0.00	0.00
45	0.03	0.04	0.03	110	0.00	0.00	0.00
46	0.03	0.04	0.02	111	0.00	0.00	0.00
47	0.03	0.03	0.02	112	0.00	0.00	0.00
48	0.02	0.03	0.02	113	0.00	0.00	0.00
49	0.02	0.03	0.02	114	0.00	0.00	0.00
50	0.02	0.03	0.02	115	0.00	0.00	0.00
51	0.02	0.03	0.02	116	0.00	0.00	0.00
52	0.02	0.03	0.02	117	0.00	0.00	0.00
53	0.02	0.03	0.02	118	0.00	0.00	0.00
54	0.02	0.03	0.02	119	0.00	0.00	0.00
55	0.02	0.03	0.01	120	0.00	0.00	0.00
56	0.02	0.02	0.01	121	0.00	0.00	0.00
57	0.01	0.02	0.01	122	0.00	0.00	0.00
58	0.01	0.02	0.01	123	0.00	0.00	0.00
59	0.01	0.02	0.01	124	0.00	0.00	0.00
60	0.01	0.02	0.01	125	0.00	0.00	0.00
61	0.01	0.02	0.01	126	0.00	0.00	0.00
62	0.01	0.02	0.01	127	0.00	0.00	0.00
63	0.01	0.02	0.01	128	0.00	0.00	0.00
64	0.01	0.02	0.01	129	0.00	0.00	0.00
65	0.01	0.02	0.01	130	0.00	0.00	0.00

3. 估計 $\hat{\theta}$ 的疊代法聚斂受試的百分比 (NP) 如圖18所示。圖中顯示，安排在55題以後，無論在那一種路徑，其估計 $\hat{\theta}$ 的疊代法均可完全聚斂，也就是說可以順利估計出 $\hat{\theta}$ 。然在55題以前，中能力組的路徑在估計 $\hat{\theta}$ 時會聚斂的受試者百分比，遠比高能力組及低能力組要高得多。此外，大致上三曲

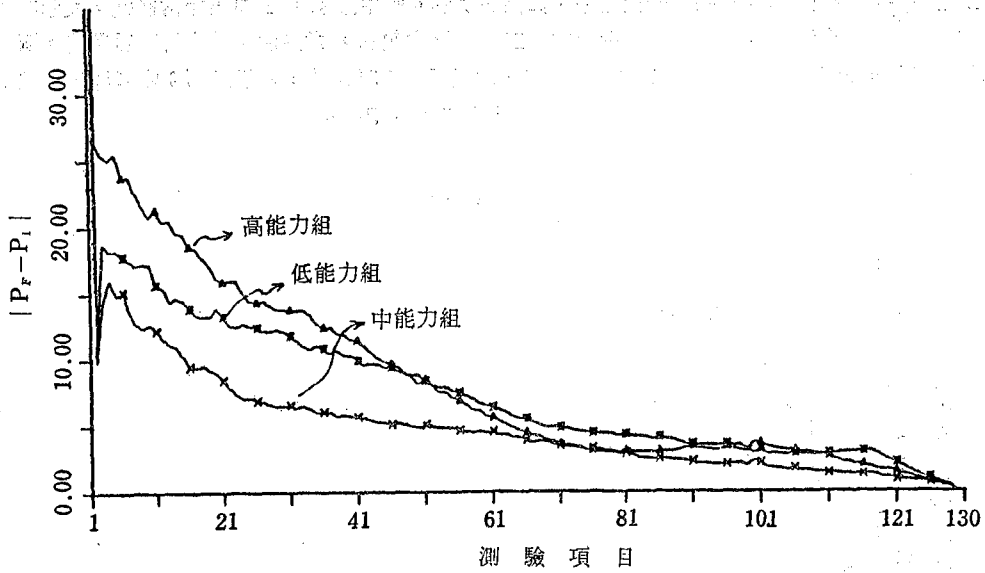


圖 17 三類適性測驗路徑 $|P_F - P_I|$ 的變化圖

線均呈遞升接近100。這些結果支持假設 3-3。

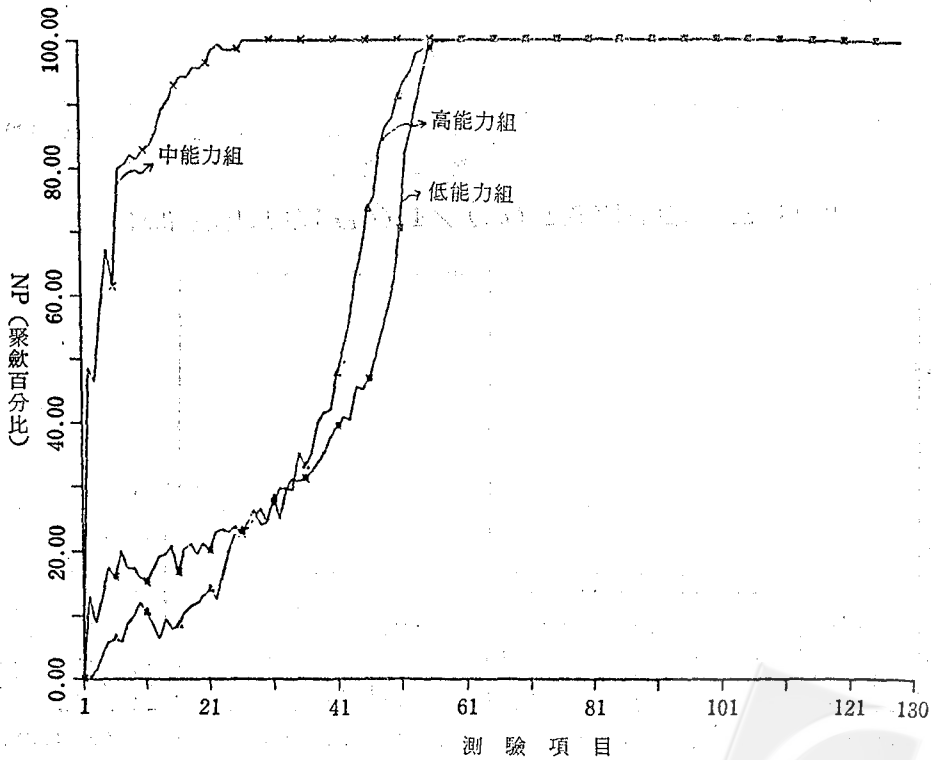


圖 18 三類適性測驗路徑在 NP (聚斂百分比) 上的變化圖

4. $I(\hat{\theta}_i) / I(\hat{\theta}_F)$ 的百分比結果，如圖19所示。圖中顯示，三曲線相當接近，表示在每一步驟的 $I(\hat{\theta}_i) / I(\hat{\theta}_F)$ 百分比是接近的。三條曲線均顯示，當適性測驗安排到11題時，測驗訊息已是佔可能最大測驗訊息的60%；安排到21題，約為80%；31題時，約為90%；當61題時，幾乎是100%。三曲線均呈遞升接近 100。這結果可支持本研究假設 3-4。

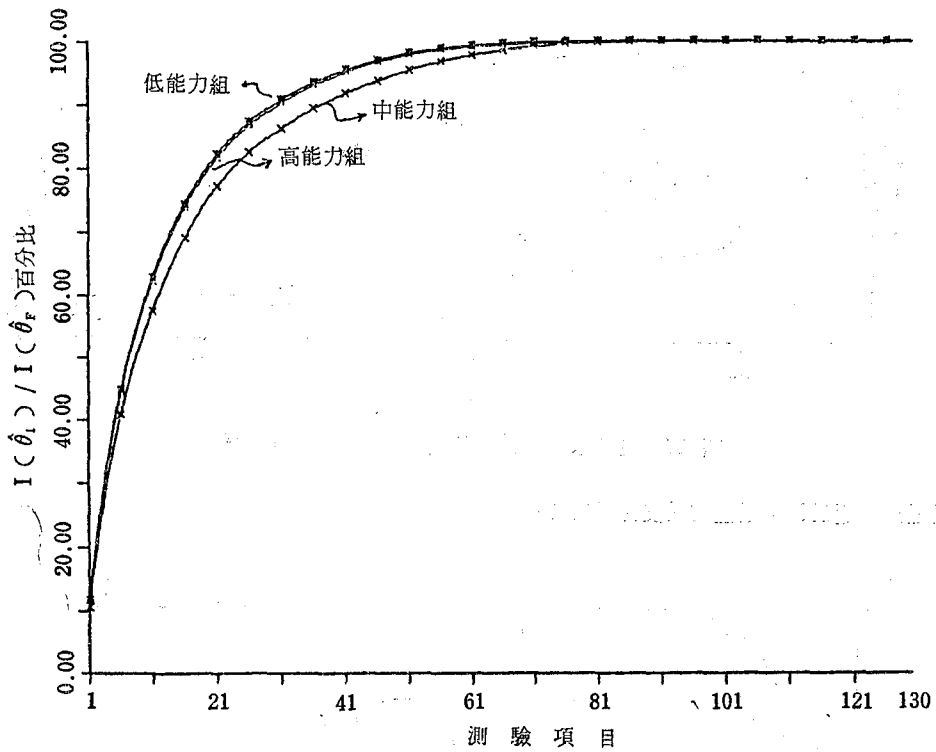


圖 19 三類適性測驗路徑 $I(\hat{\theta}_i) / I(\hat{\theta}_F)$ 百分比上的變化圖

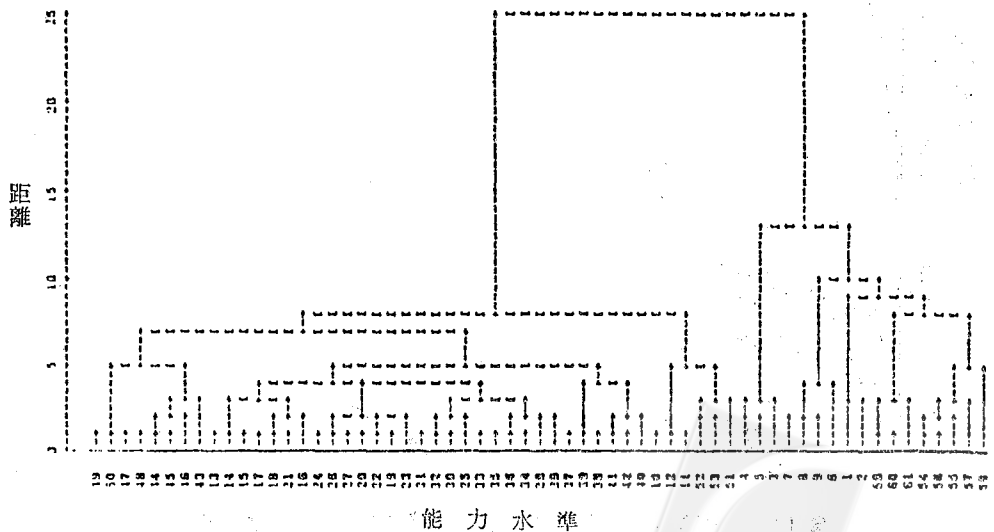


圖 20 θ 未知的適性測驗路徑分類樹狀圖 (以 a 參數分類)

三、「 θ 未知」的適性測驗

(一) 適性測驗路徑的羣聚分析

1. 以 a 參數分類

將61個能力水準的受試者，以其130個項目的 a 參數的路徑進行羣聚分析，結果得樹狀圖如圖20所示。圖中顯示 θ 未知的適性測驗路徑，可以用 a 參數清楚地分成兩類路徑。第一類是由1,2,3,4,5,6,7,8,9,54,55,56,57,58,59,60,61,等17個能力水準所組成， θ 分別是界於-2.3與-3.0, 2.2與3.0是為極端能力組。第二類是由能力水準10到53等44個能力水準所組成， θ 是界於-2.2與2.1是為中能力組。結果支持研究假設4-1。

至於上述究竟是那兩類路徑呢？圖21是 a 參數在二類路徑中的變化圖。結果顯示：兩類路徑在安排到105題以前 a 參數並沒明顯的差別。只是在105題以後，中能力組的適性測驗路徑， a 參數是接近0的；極端能力組的路徑，在105題以後， a 參數並未穩定地接近0，而且有大於1的現象。

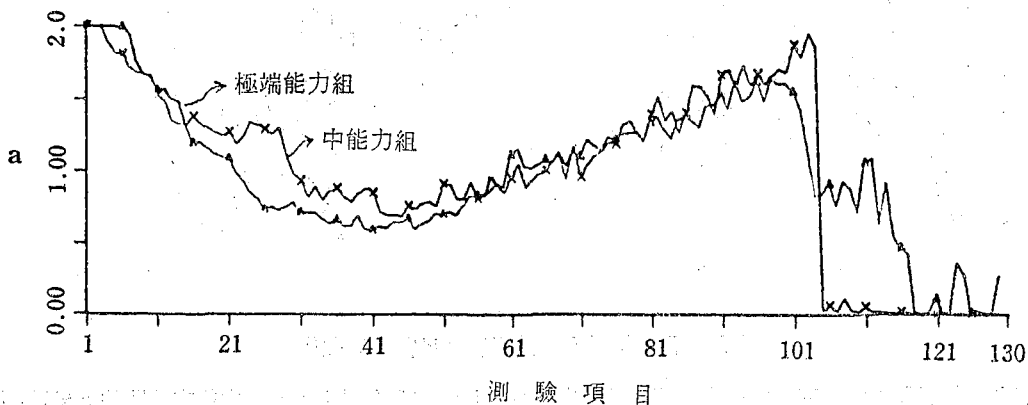


圖 21 兩類適性測驗路徑在 a 參數的變化圖

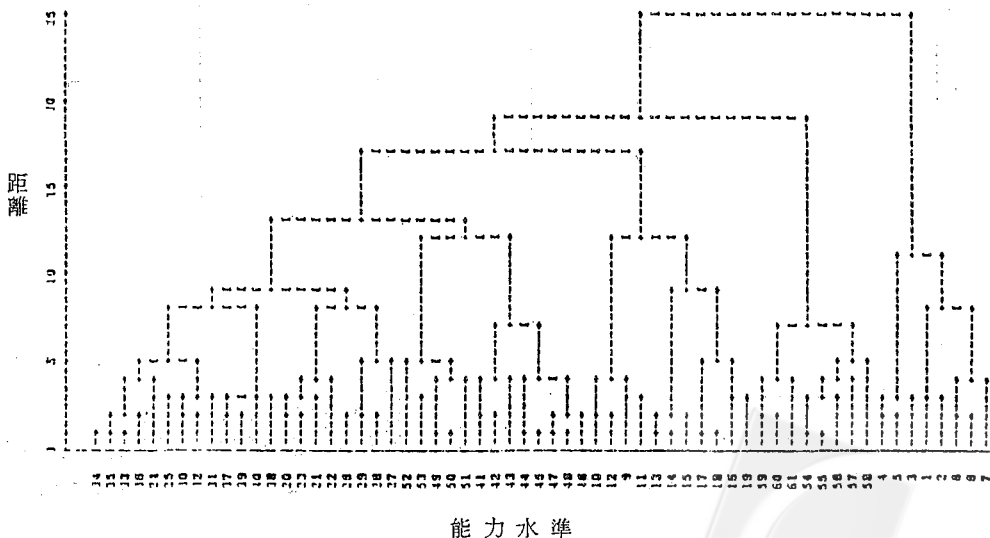


圖 22 θ 未知的適性測驗路徑分類樹狀圖 (以 c 參數分類)

2. 以 c 參數分類

將61個能力水準的受試者，根據路徑中的 c 參數羣聚分析，其樹狀圖如圖22所示。圖中顯示： θ 未知的適性測驗路徑，可以用 c 參數分成四種類型。第一類是由能力水準 1 到 8 等 8 個能力水準所組成， θ 界於 2.3 與 3.0 是為高能力組。第二類是由能力水準 9 到 19 等 11 個能力水準所組成， θ 界於 1.2 與 2.2，是為中上能力組。第三類是由能力水準 20 到 53 等 34 個能力水準所組成， θ 界於 -2.2 與 1.1，是為中能力組。第四類是由能力水準 54 到 61 等 8 個能力水準所組成， θ 界於 -2.3 與 -3.0，是為低能力組。結果支持研究假設 4-3。

上述四類路徑在 c 參數的變化如圖23所示。結果顯示：四類路徑在 c 參數變化的第一個差異是在

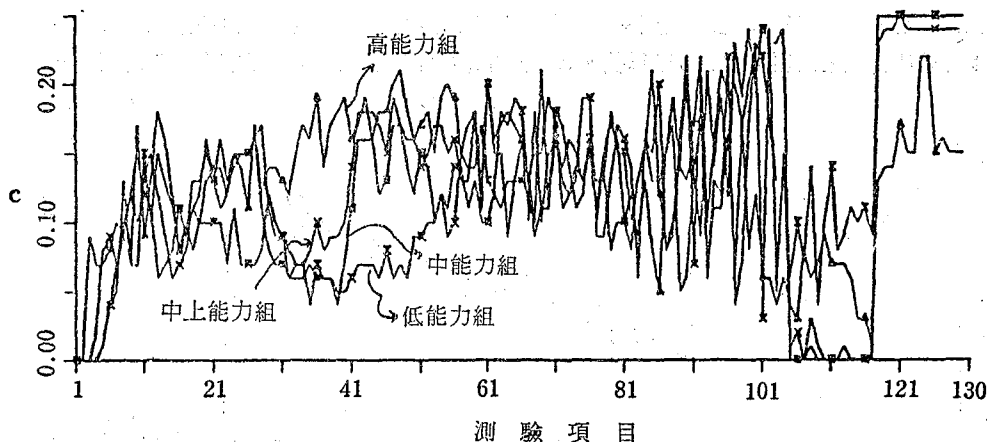


圖 23 四類適性測驗路徑在 c 參數的變化圖

第30題與第50題之間，高能力組在此間 c 參數在偏高，且都在 1.0 以上。低能力組在此間的 c 參數則偏低，且均在 1.0 以下。此四類路徑 c 參數的變化的第二個差異是在第115題以後，中能力組與低能力組在第115題以後， c 參數穩定地接近 0.25，而中上能力組與高能力組的 c 參數仍舊未穩定地接近 0.25

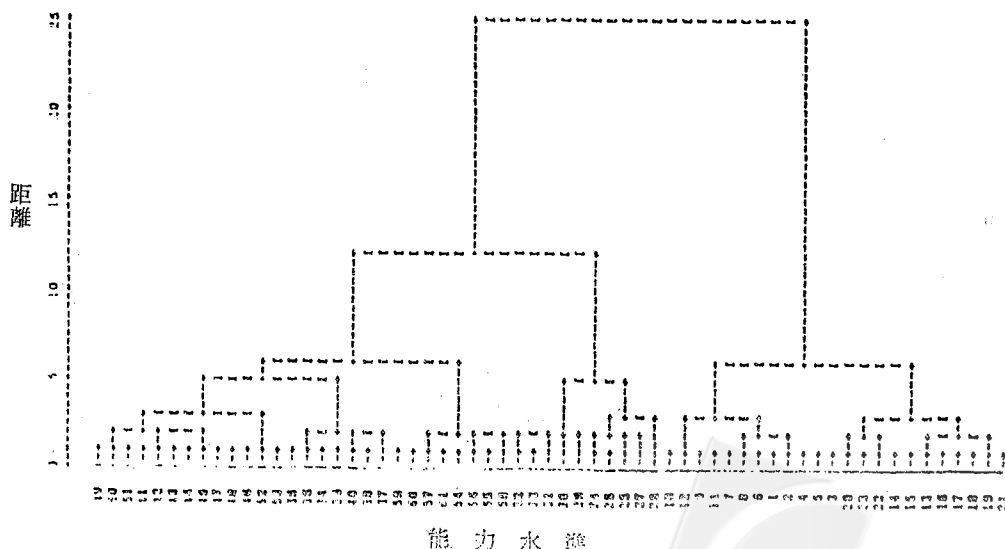


圖 24 θ 未知的適性測驗路徑分類樹狀圖 (以 b 參數分類)

3. 以 b 參數分類

將61個能力水準的受試者，以路徑中 130 個項目 b 參數進行羣聚分析，得到樹狀圖結果如圖24所示。圖中指出 θ 未知的適性測驗路徑，可以用 b 參數清楚地分成三路徑。第一類是由能力水準 1 到23等23個能力水準所組成， θ 界於0.8與3.0，是為高能力組。第二類是由能力水準24到33等10個能力水準所組成， θ 於-0.2與0.7，是為中能力組。第三類是由能力水準34到61等28個能力水準所組成， θ 界於-0.3與-3.0，是為低能力組。結果支持研究假設 4-2。

以 b 參數所分出之三類路徑，進一步分析其 b 參數在三類路徑中的變化，結果如圖25所示。由圖可以清楚看出：高能力組的路徑是 b 參數在前10題遞升，而10題之後則呈由大而小的走向。低能力組

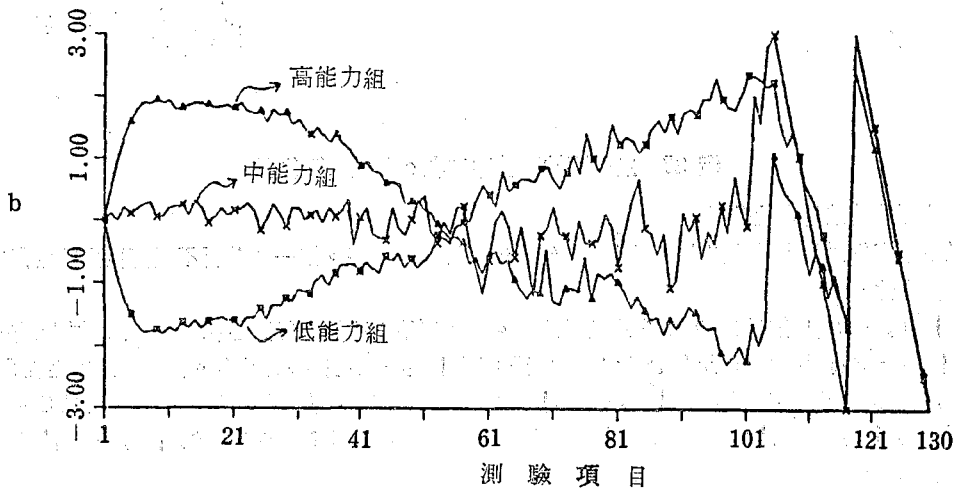


圖 25 三類適性測驗路徑在 b 參數的變化圖

在前10題是遞降，10題以後則呈由小而大的走向。中能力組的路徑，則是在 0 上下跳動。另外這三類路徑，在 a 參數上的變化，如圖26所示。三類路徑的 a 參數在105題之前差別並不清楚，而在105題以後，中能力組路徑的 a 參數，已穩定地接近 0。但在高能力組及低能力組路徑的 a 參數，並未穩定地接近 0。另外，此三類路徑，在 c 參數上的變化，如圖27所示。圖中顯示三類路徑的 c 參數，在105題以前差別也不清楚。而在106題與 118題之前，中能力組路徑穩定地接近 0，而在119題以後則呈穩定接近於0.25；這種 c 參數的現象，在高能力組與低能力組的路徑都沒有產生。三類路徑在 a 參數及

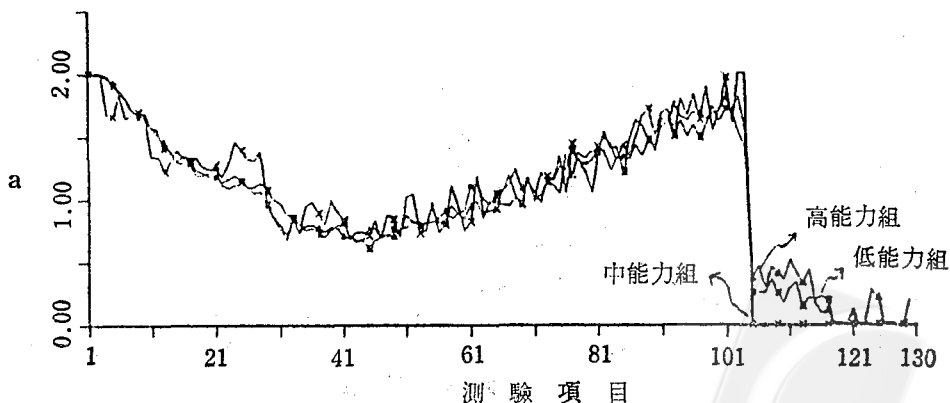


圖 26 三類適性測驗路徑在 a 參數的變化圖

c 參數的變化，與前述以 a 參數分成兩類路徑及以 c 參數分成四類路徑的結果是大同小異的。

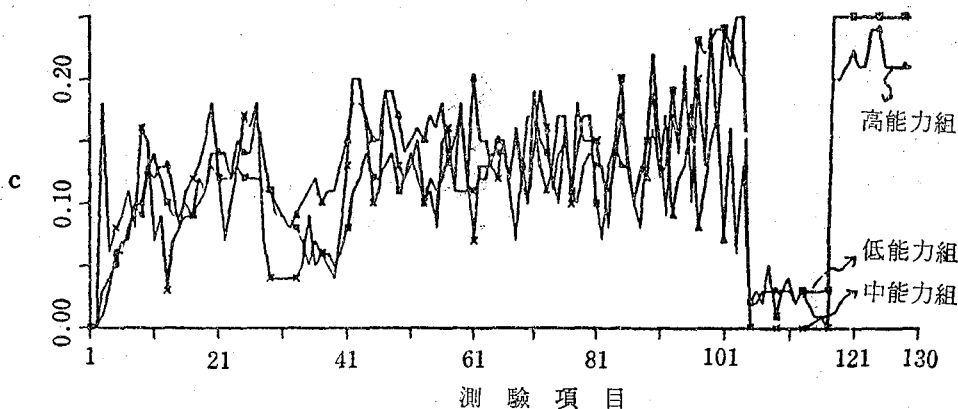


圖 27 三類適性測驗路徑在 c 參數的變化圖

(二) 適性測驗精確穩定性分析

以下筆者仍以 b 參數分出之三類「 θ 未知」的適性測驗路徑，進一步分析不同適性測驗路徑的測量精確穩定性。

1. $|\hat{\theta}_F - \hat{\theta}_I|$ 的結果如表 6 及圖 28 所示。結果顯示：高能力組路徑在 21 題到 72 題之間有一起伏， $|\hat{\theta}_F - \hat{\theta}_I|$ 最大差距是 39.85。72 題以後，則 $|\hat{\theta}_F - \hat{\theta}_I|$ 接近於 0。低能力組路徑在 21 題到 66 題之間亦有一稍小的起伏， $|\hat{\theta}_F - \hat{\theta}_I|$ 最大差距是 6.67，在 66 題以後則差距接近於 0。中能力組路徑的 $|\hat{\theta}_F - \hat{\theta}_I|$ ，過了第 2 題以後，即呈遞降而接近於 0；16 題以後的 $|\hat{\theta}_F - \hat{\theta}_I|$ 已小於 .10。結果僅部份支持研究假設 5—1。

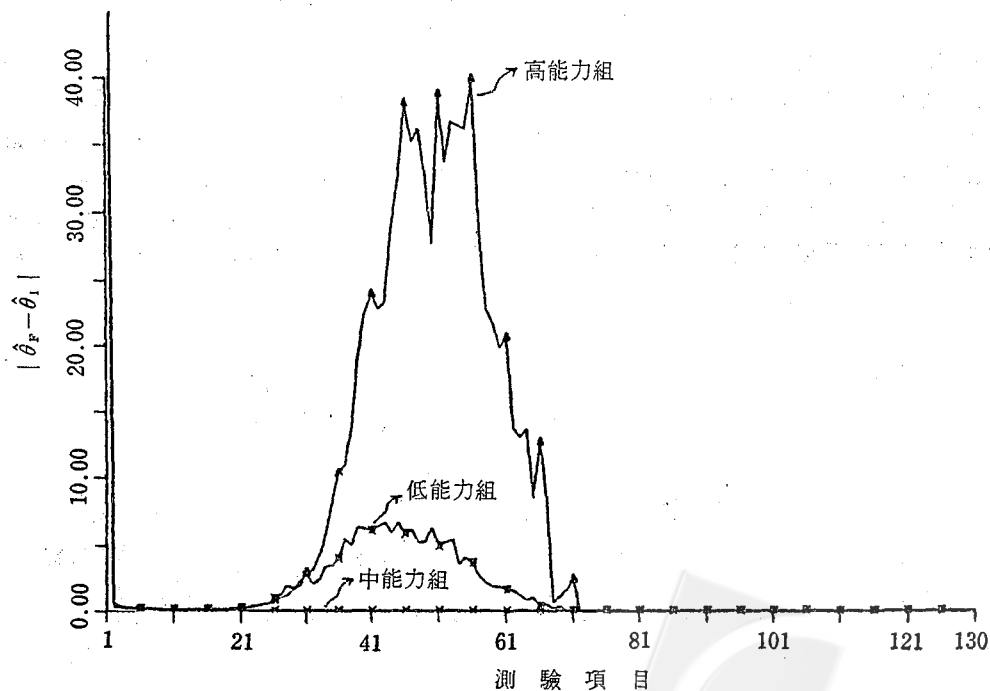


圖 28 三類適性測驗路徑在 $|\theta_F - \theta_I|$ 的變化圖

表 6 θ 未知的適性測驗 $|\hat{\theta}_F - \hat{\theta}_I|$ 值

項目	高能力組	中能力組	低能力組	項目	高能力組	中能力組	低能力組
1				66	12.62	0.01	0.36
2	0.67	0.27	0.52	67	8.24	0.01	0.55
3	0.41	0.26	0.35	68	0.72	0.01	0.27
4	0.32	0.25	0.30	69	1.09	0.01	0.35
5	0.31	0.24	0.28	70	1.61	0.01	0.01
6	0.26	0.22	0.24	71	2.42	0.01	0.01
7	0.24	0.20	0.22	72	0.01	0.01	0.00
8	0.21	0.17	0.20	73	0.01	0.01	0.01
9	0.19	0.15	0.19	74	0.01	0.01	0.00
10	0.17	0.14	0.18	75	0.01	0.01	0.00
11	0.16	0.14	0.17	76	0.02	0.01	0.00
12	0.15	0.13	0.19	77	0.04	0.01	0.00
13	0.15	0.12	0.18	78	0.00	0.01	0.00
14	0.16	0.12	0.17	79	0.00	0.01	0.00
15	0.17	0.11	0.19	80	0.00	0.00	0.00
16	0.19	0.10	0.19	81	0.00	0.00	0.00
17	0.18	0.09	0.23	82	0.00	0.00	0.00
18	0.22	0.09	0.23	83	0.00	0.00	0.00
19	0.22	0.09	0.24	84	0.00	0.00	0.00
20	0.29	0.09	0.25	85	0.00	0.00	0.00
21	0.31	0.09	0.29	86	0.00	0.00	0.00
22	0.34	0.08	0.36	87	0.00	0.00	0.00
23	0.47	0.08	0.46	88	0.00	0.00	0.00
24	0.53	0.08	0.67	89	0.00	0.00	0.00
25	0.63	0.08	0.67	90	0.00	0.00	0.00
26	0.80	0.07	0.96	91	0.00	0.00	0.00
27	1.04	0.07	1.34	92	0.00	0.00	0.00
28	1.23	0.07	1.97	93	0.00	0.00	0.00
29	1.63	0.07	1.81	94	0.00	0.00	0.00
30	2.27	0.07	1.94	95	0.00	0.00	0.00
31	3.03	0.06	2.74	96	0.00	0.00	0.00
32	3.10	0.06	2.10	97	0.00	0.00	0.00
33	4.26	0.06	2.42	98	0.00	0.00	0.00
34	5.75	0.06	3.39	99	0.00	0.00	0.00
35	8.06	0.05	3.60	100	0.00	0.00	0.00
36	10.30	0.05	4.00	101	0.00	0.00	0.00
37	10.93	0.05	5.46	102	0.00	0.00	0.00
38	13.52	0.05	5.04	103	0.00	0.00	0.00
39	18.95	0.04	6.34	104	0.00	0.00	0.00
40	22.47	0.04	6.25	105	0.00	0.00	0.00
41	23.95	0.04	6.09	106	0.00	0.00	0.00
42	22.85	0.04	6.44	107	0.00	0.00	0.00
43	23.29	0.04	6.67	108	0.00	0.00	0.00
44	28.74	0.04	5.99	109	0.00	0.00	0.00
45	32.42	0.03	6.63	110	0.00	0.00	0.00
46	38.04	0.03	5.82	111	0.00	0.00	0.00
47	35.21	0.03	6.13	112	0.00	0.00	0.00
48	36.12	0.03	5.19	113	0.00	0.00	0.00
49	32.66	0.03	5.30	114	0.00	0.00	0.00
50	27.61	0.03	6.26	115	0.00	0.00	0.00
51	38.68	0.03	4.91	116	0.00	0.00	0.00
52	33.63	0.03	5.25	117	0.00	0.00	0.00
53	36.64	0.03	5.41	118	0.00	0.00	0.00
54	36.40	0.03	3.63	119	0.00	0.00	0.00
55	36.12	0.02	4.07	120	0.00	0.00	0.00
56	39.85	0.02	3.65	121	0.00	0.00	0.00
57	28.77	0.02	2.71	122	0.00	0.00	0.00
58	22.80	0.02	2.13	123	0.00	0.00	0.00
59	21.74	0.02	1.88	124	0.00	0.00	0.00
60	19.79	0.02	1.86	125	0.00	0.00	0.00
61	20.58	0.02	1.61	126	0.00	0.00	0.00
62	13.67	0.01	1.58	127	0.00	0.00	0.00
63	13.07	0.01	1.21	128	0.00	0.00	0.00
64	13.60	0.01	0.86	129	0.00	0.00	0.00
65	8.42	0.01	1.01	130	0.00	0.00	0.00

2. $|P_F - P_I|$ 的結果如圖29所示。結果顯示：高能力組路徑在 9 到 65 題之間有一起伏，最大之 $|P_F - P_I|$ 為12.86；在87題到113題又有一小起伏， $|P_F - P_I|$ 值均小於 5。低能力組路徑，亦有類似的現象。在13題到 64 題之間有一起伏，最大之 $|P_F - P_I|$ 為 8.88；在91題到 110 題又有一小起伏，

$|P_F - P_I|$ 值均小於 5。中能力組路徑呈遞降趨近於 0。結果僅部份支持研究假設 5—2。

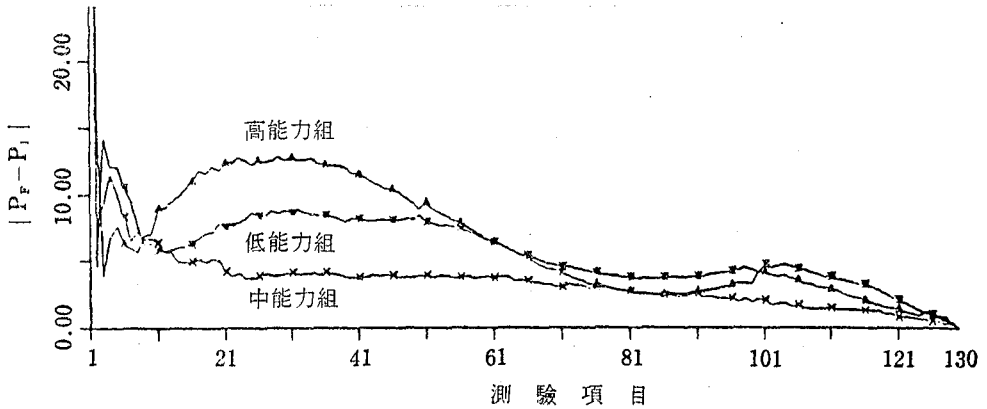


圖 29 三類適性測驗路徑在 $|P_F - P_I|$ 的變化圖

3. 估計 $\hat{\theta}$ 疊代法聚斂受試者的百分比 (NP) 之結果如圖 30 所示。圖中顯示：高能力組路徑在安排到 13 題時，NP 已是 100，只是在 49 題到 77 題有一小起伏，NP 最低降到 91.74。低能力組路徑在 14 題時，NP 已是 100。但是從 29 題到 71 題之間，NP 有一小起伏；最低降至 94.29。至於中能力組路徑，NP 是呈遞降趨近 100，而且是在 4 題以後 NP 就是 100。結果部份支持研究假設 5—3。

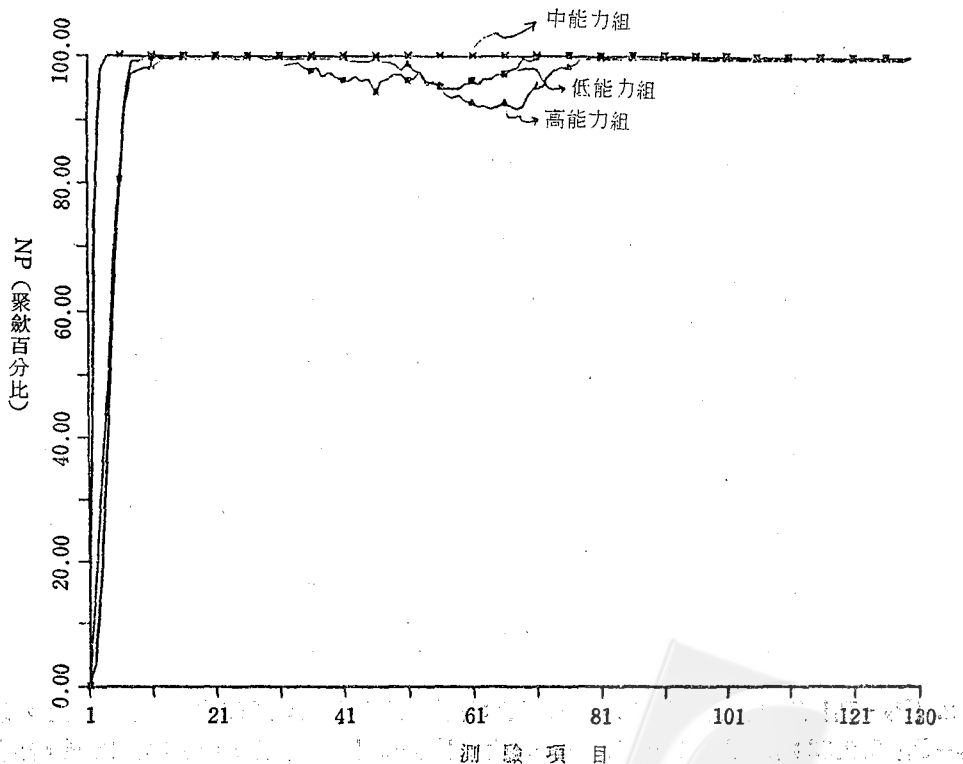


圖 30 三類適性測驗路徑在 NP (聚斂百分比) 上的變化圖

4. $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 的百分比結果，如圖31所示。圖中顯示：高能力組路徑在23題時 $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 的百分比已是70.8，只是在24題與49題之間有下降現象，最低降到60.5。50題以後又呈現遞升，直到87題時 $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 的百分比已是100。低能力組路徑，在22題時 $I(\hat{\theta}_i) / I(\hat{\theta}_r)$

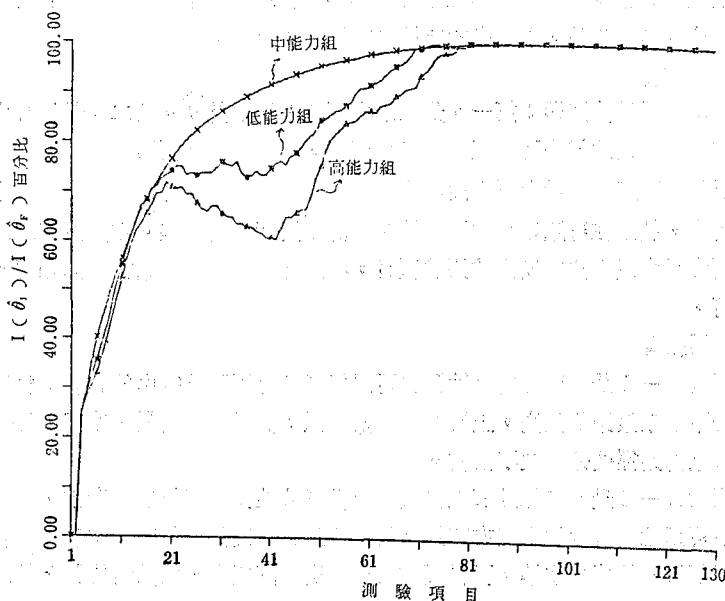


圖 31 三類適性測驗路徑 $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 百分比的變化圖

$I(\hat{\theta}_r)$ 的百分比是75.0，只是在23題到41題有下降的現象，最低降到72.7。42題以後又呈現遞升，直到82題時， $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 的百分比已是100。中能力組路徑， $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 的百分比呈遞升趨近100；在24題時， $I(\hat{\theta}_i) / I(\hat{\theta}_r)$ 的百分比是80.2，38題時是90.1，89題時是100。結果僅是部份支持研究假設5—4。

討 論

一、LTT各項參數與CTT的各項統計數的關係

本研究的假設1—1是LTT的 θ 參數與估計的 $\hat{\theta}$ 有正相關存在。結果發現二者之間積差相關係數高達.9936，與Ree (1981)的研究中所得的.927 (Traub & Lam, 1985) 結果極為一致。這樣的結果在本研究中可以同時說明兩件事：第一、本研究所模擬產生的610位假想受試者的反應組型是理想的，且符合LTT的。第二、估計 $\hat{\theta}$ 所採用的最大可能性法，在130個項目的條件下，估計結果是相當穩定靠可的，因為估計的 $\hat{\theta}$ 與真正的 θ 是相當接近的。

本研究的假設1—2是LTT的 θ 參數與CTT的個人得分有正相關存在。結果發現二者之間積差相關係數高達.9841，與Lord (1968) 用2862位受試者在SAT語文測驗上的結果，所求得個人得分與 $\hat{\theta}$ 之間的關係，相當類似。假設1—3是LTT的 a 參數與CTT的 r_{bi} 值有正相關存在。結果發現二者之間積差相關係數是.7001，這個關係可以與Lord (1980a) 在前述公式15的關係式相互印證。另外，假設1—4是LTT的 b 參數與CTT的 P 值有負相關存在，結果是二者

積差相關係數是 -0.8356 ，這結果可以與 Lord (1980a) 在公式16中的關係式相互印證。假設 1—5 是：LTT 項目中 a 、 b 參數相同、 c 參數不同的二組測驗項目，其 r_{b1a} 有差異，說明了猜測因素影響項目鑑別度。

上述結果說明了 LTT 與 CTT 在測驗上的基本觀念是相通的。LTT 所面對的測量問題，也是老問題。所討論的觀念，也並非完全是新觀念。因此 Hulin 等 (1983) 及 Weiss (1983) 才會認為二者是重疊或是包容的關係。本研究的結果也支持這種想法。

二、 θ 已知的適性測驗

適性測驗有兩個關鍵性問題：第一、是以最大可能性法估計 $\hat{\theta}$ 。第二、是以最大訊息法做為項目選擇的依據。這兩個問題的解決及處理均易產生偏差，而使得適性測驗結果也跟著產生的偏差。於是便無法看出真正原因到底是在最大可能性法估計 $\hat{\theta}$ ，或是在最大訊息法項目選擇。本文這一部分 θ 已知的適性測驗研究，就是在排除以最大可能性法估計 $\hat{\theta}$ 所產生的偏差。在適性測驗的過程中，直接代以 θ ，而純粹研究最大訊息法所安排的測驗項目，在適性測驗所形成路徑結構如何，及其在測量上的精確穩定性如何。

(一) 適性測驗路徑結構

本研究之假設 2—1 是 θ 已知的適性測驗可以根據適性測驗路徑中的 a 參數加以分類。結果發現在樹狀圖10中的下層羣聚距離偏高，上層羣聚距離偏低，表示羣聚內異質性高，且分類不清楚。意指以路徑中 a 參數來對路徑做分類並不恰當。

本研究之假設 2—2 是 θ 已知的適性測驗可以根據適性測驗路徑的 b 參數，加以分類。結果發現：在樹狀圖11中可以比較清楚地看出有三類路徑，而且三類路徑的組成受試者分別各是高能力、中能力、及低能力三個層次的能力水準。再從圖13可以看出，高能力組受試者的 b 參數路徑是一種由高而低的型態；低能力組是由低而高的型態；至於中能力組的 b 參數路徑，則一直是在中難度上下走動，只是 b 參數路徑的振幅，是由小而逐漸變大。然而，同樣的三組受試者的適性測驗路徑中的 a 參數（如圖14）及 c 參數（如圖15）並沒有如此清楚的差異。

本研究之假設 2—3 是 θ 已知的適性測驗可以根據適性測驗路徑的 c 參數加以分類。結果發現在樹狀圖12中的下層羣聚距離偏高，上層羣聚距離偏低，表示羣聚內異質性高，且分類不清楚。意指以路徑中 c 參數來對路徑做分類亦不恰當。

以上的結果可以說明本研究所採用的最大訊息法適性測驗，受試者的適性測驗路徑是可以分類的，至少可以根據難度來分成三類。而且所謂最大訊息法之項目選擇，也只是選擇難度與能力相匹配的項目而已，對 a 參數與 c 參數加以考慮似並沒太大的作用。所以，基本上最大訊息法與上下法，H-L 的項目選擇，在作法上可能也是大同小異，因此將題庫的項目，依難度予以結構化，將有助於適性測驗實施。這結果與 Hambleton (1985) 及 Weiss (1974) 的觀念稍有差異，因為兩位學者曾經指出，最大訊息法適性測驗的特色是題庫勿須結構化，而且路徑並不會太固定有限。但是本研究結果却指出了路徑不但固定有限，同時也指出難度與能力的匹配是最重要的項目選擇依據。因此題庫的結構化可便於實施。換言之，本研究的適性測驗與上下法或是 H-L 法的適性測驗，實際上是類同的。

(二) 適性測驗在測驗上的精確穩定性

適性測驗與傳統測驗的目的一樣，也是希望能有效地測量出受試者的能力。上述的三組受試者在接受適性測驗的過程中，測驗精確穩定性的變化情形可逐一討論如下：

1. 本研究之假設 3—1 是： θ 已知的適性測驗，其 $|\hat{\theta}_F - \hat{\theta}_I|$ 呈單調性遞降趨近於 0。圖16所得的結果，大致上是呈現單調性遞降，下降速度很快，而且三類路徑的狀況十分類似；到 15 題時， $|\hat{\theta}_F - \hat{\theta}_I|$ 已降到 0.1 以下。可見在不考慮最大可能性法估計 $\hat{\theta}$ 的偏差時，以 15 題來估計 $\hat{\theta}_{15}$ ，已經與 $\hat{\theta}_F$ 相當接近，而且隨著題數增加， $\hat{\theta}_I$ 會越靠近 $\hat{\theta}_F$ 。其實只要大約做 80 題左右所估計 $\hat{\theta}_{80}$ ，實際

上與 $\hat{\theta}_F$ 已經是完全相等了。81題以後根本可以不必做了。因此在不考慮最大可能性法估計 $\hat{\theta}$ 的偏差時，適性測驗可以同時兼顧到測量經濟性及其精確性。

另外，從其單調性降低的現象可以看出：此適性測驗所估計的 $\hat{\theta}_I$ 呈現相當穩定的現象。此一結果可以支持使用固定題數來結束測量的做法 (Kreitzberg & Jones, 1980 等)。

上面的結果可支持最大訊息法適性測驗的理論，也就是假定估計 $\hat{\theta}$ 的最大可能性法沒有問題，適性測驗是可行的。

另外一個意外發現是：15題以前的 $|\hat{\theta}_F - \hat{\theta}_I|$ 呈現偏高，及不穩定現象，原因正可能是最大可能性估計 $\hat{\theta}$ 的系統偏差所造成。從圖32的三類路徑在 $\hat{\theta}_F - \hat{\theta}_I$ 的變化可以看出：對高能力組而言， $\hat{\theta}$ 的估計有高於 θ_F 的情況，即有高估的現象；但對低能力組，則又呈現低估的現象，這是一個系統現象，值得進一步研究。

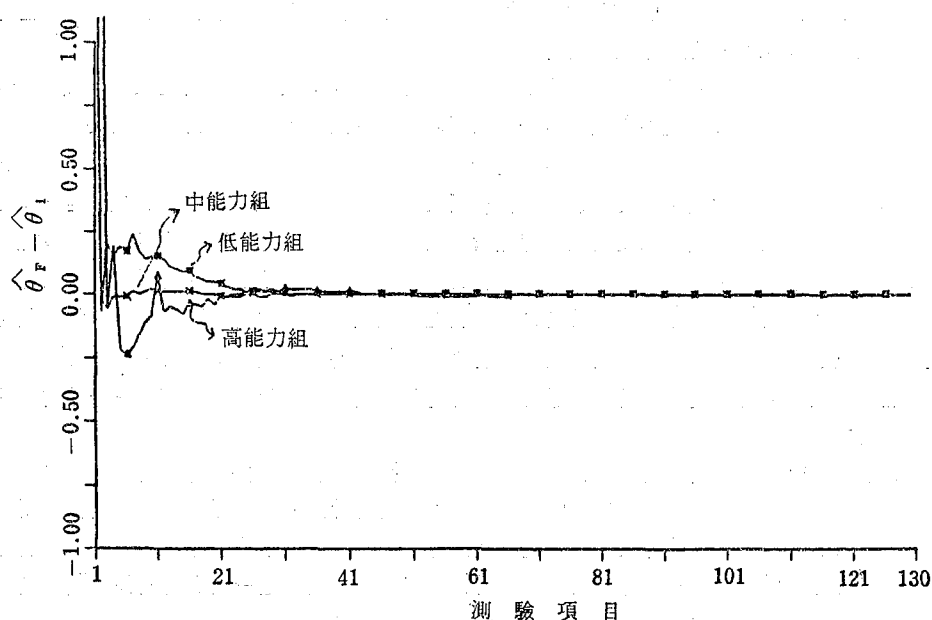


圖 32 三類適性測驗路徑在 $\hat{\theta}_F - \hat{\theta}_I$ 的變化圖

2. 本研究之假設 3—3 是 θ 已知的適性測驗，以疊代法（最大可能性法）估計 $\hat{\theta}$ 時，聚斂的受試者百分比呈單調性遞升趨近100。圖 18 的結果顯示：大致上就是呈單調性遞升。表示隨著適性測驗題數增加，最大可能性法使用起來也越加穩定，這個結果與 Hamdleton (1985) 所說，測驗項目越多時，使用最大可能性法估計 $\hat{\theta}$ ，會越容易聚斂，而求出 $\hat{\theta}$ 。然而，上升的速度三類路徑不完全一致。顯然，中能力組受試的 $\hat{\theta}$ 是比較容易估計出來的，只要做10題，就有.80以上的機率聚斂；而且聚斂百分比上升頗快。但在高能力組及低能力組受試的 $\hat{\theta}$ ，在題數不多時，估計常常不易聚斂；當做10題時，只有.1到.15的機率會聚斂。Lord (1980a) 指出題數在20以上，最大可能性法估計 $\hat{\theta}$ 便可相當穩定。在本研究中，只有中能力組的 $\hat{\theta}$ 估計，可能符合 Lord 的說法。此項結果不能有力支持學者們以固定題數做為結束測量的依據 [(Kreitzberg & Jones, 1980 等)]。因為受試者在完成一固定題數的項目，無法保證估計 $\hat{\theta}$ 時，一定聚斂。

3. 本研究之假設 3—2： θ 已知的適性測驗其 $|P_F - P_I|$ 呈現單調性遞降趨近於0。答對百分比

是 CTT 中常用的一個能力指數。它沒有聚斂不聚斂的問題，估計方法本身很穩定。因此值得探討其估計的 P_i ，精確穩定如何？

圖17的結果顯示： $|P_F - P_i|$ 大致上呈現單調性遞降，只是三組路徑下降的速度不完全一致。高能力組下降的最慢，低能力組次之，中能力組稍快些。雖然下降情形相當穩定，但在做完題庫之前，毫無接近 0 的情形，且距離 0 還有 2 到 10 的百分比。

雖然估計答對百分比方法很穩定，但 P_i 的精確性不理想。因此在適性測驗上，它是毫無意義的。

4. 本研究之假設 3—4： θ 已知的適性測驗， $I(\hat{\theta}_i) / I(\hat{\theta}_F)$ 百分比呈單調性遞升趨近於 100。圖19的結果完全呈單調性遞升。上升速度很快，而且三類路徑的狀況十分類似。因為 $I(\hat{\theta}_i)$ 的倒數就是測量變異誤，所以當 $I(\hat{\theta}_i) / I(\hat{\theta}_F)$ 的百分比上升，測量變異誤就下降。當做 21 題時，百分比已升到 80，測量變異誤已經很小。當進行到 80 題時，百分比已升到 100，表示與做完題庫所得的 $I(\hat{\theta}_F)$ 是一樣的，測量誤差也將降低到最小，因此在不考慮最大可能性法估計 $\hat{\theta}$ 的偏差時，適性測驗是可同時兼顧測量的經濟性及精確性。

另外，從其單調性升高可以看出，所估計的 $I(\hat{\theta}_i)$ 是相當穩定的。這樣的結果可以支持結束測量使用固定標準之測驗訊息或測量變異誤的作法 (Hulin, Drasgow & Parsons, 1983)

從以上的討論得知，在不考慮最大可能性法估計 $\hat{\theta}$ 時會產生偏差的條件下（即本研究所謂 θ 已知的狀況），本研究的結果大致上可以支持適性測驗使用 LTT 觀點以最大訊息法來選取項目的做法。無論是從使用題數經濟性，或是從測量精確穩定性看，最大訊息法所選取出來的項目。可以有題數少、誤差小、正確性高的效果。這正是適性測驗所期望的。然而，所謂 LTT 觀點的最大訊息實際上在選取項目時仍舊著重在受試者能力與項目難度之間的匹配，並沒有特別之處。另外還有一個特別值得注意的地方是這種適性測驗在每一步驟使用最大可能性法估計 $\hat{\theta}$ 時，對能力偏高及偏低的受試者而言，聚斂情形並不太理想。主要是要耗費較多題目，才可能所有受試都聚斂，而得到 $\hat{\theta}$ 一計分的結果。

三、 θ 未知的適性測驗

θ 未知的適性測驗，是接近真實測驗的情境，因為一個受試者在測驗之前，通常是無法預知其 θ ，而必須要透過測驗，然後再根據受試者的反應，估計 $\hat{\theta}$ ，以推估受試者的真正 θ 。本研究就是要以適性測驗的方式，來進行測驗。但是適性測驗是逐一選擇適合的項目，而選擇的標準與受試者的能力（ θ ）的高低有密切的關係。在前一部份適性測驗路徑的羣聚分析也說明這一明顯的關係。既然選擇項目與 θ 有關，而這一部份的研究又是 θ 未知，則整個適性測驗的首要工作，就是必須設法去估計 $\hat{\theta}$ ，以推估受試者真正的 θ ，以 $\hat{\theta}$ 做為計算項目訊息之用，和做為選擇項目之用。在這一部份的研究裏，所估計的 $\hat{\theta}$ 兼具兩個功能：一是用來做適性測驗的計分，另一是計分的結果（即估計的 $\hat{\theta}$ ）做為選擇下一題時用來計算項目訊息。本研究這一部分所使用估計 $\hat{\theta}$ 的方法，主要還是最大可能性法。因此這一部份 θ 未知的適性測驗研究是同時考慮最大可能性法估計 $\hat{\theta}$ 與最大訊息法選擇項目。換言之，將二者合併以後，使作用在適性測驗上，以探討其適性測驗路徑結構及其在測驗上的精確穩定性。

(一) 適性測驗路徑結構

本研究之假設 4—1 是： θ 未知的適性測驗可以根據適性測驗路徑中的 a 參數，加以分類。結果從樹狀圖 20 中可以清楚地看出，有兩種類型的路徑。其中一種是高能力或低能力受試者所接受的適性測驗路徑；另一種是中能力受試者所接受的適性測驗路徑。圖 21 中顯示：大致上 a 參數在極端能力組受試者的適性測驗路徑上影響作用比較小，因為在路徑末端 a 參數並沒有穩定地接近 0，而且還有大於 1 的現象。顯然的，在路徑的前端曾使用過低鑑別度的項目，原因可能是在項目選擇時，已優先匹配能力與難度的關係。

本研究之假設 4—3 是 θ 未知的適性測驗可以根據適性測驗路徑中的 c 參數，加以分類。結果從

樹狀圖22中清楚地看出可以分出四種類型的路徑。從圖23中也可以看出，在適性測驗路徑中 c 參數作用並不大，尤其是高能力組的受試者，因為在高能力組受試者的適性測驗路徑末端， c 參數並未穩定地接近0.25。顯然的，在路徑的前端曾使用過高猜測因素的项目，其原因也可能是項目選擇時，優先匹配能力與難度的關係。

本研究之假設 4—2 是： θ 未知的適性測驗可以根據適性測驗路徑中的 b 參數，加以分類。結果發現，從樹狀圖24中，可以清楚地分出三種類型的路徑。而且三種類型的路徑，分別是隸屬高能力、中能力、低能力三種能力水準的受試者，再從圖25中也可以看出三類路徑與 θ 已知的適性測驗的三類路徑，型態上十分類似；高能力組受試者適性測驗路徑中的 b 參數，是由高而低的走向；低能力組是由低而高的走向；至於中能力組，則一直是在中難度上下跳動，振幅也是由小變大。這更顯示出適性測驗的项目選擇，是根據能力與難度的匹配。

再看同樣的三組受試者其適性測驗中的 a 參數（如圖26）與 c 參數（如圖27），結果與 θ 未知時的以 a 參數及 c 參數的分類結果類同。

θ 未知的適性測驗路徑可以用路徑中 a 參數，清楚地分成兩類路徑，也可以用路徑中 c 參數，清楚地分成四類路徑。但是 θ 已知的適性測驗，却沒有這種現象。理由極可能是估計 $\hat{\theta}$ 所使用的方法，產生系統性偏差所導致，又間接再影響到適性測驗项目的選擇。

(二) 適性測驗在測量上的精確穩定性

適性測驗的目的是期望使用較少的项目，在不增加測量變異誤的條件下，正確地測量出受試者的能力。上述的三組受試者，在接受適性測驗後，其測量精確穩定性的變化情形可逐一討論如下：

1. 本研究之假設 5—1 是： θ 未知的適性測驗，其逐步的 $|\hat{\theta}_F - \hat{\theta}_I|$ 呈單調性遞降趨近於 0。研究結果（如圖28）發現：中能力組受試者下降速度快，且穩定地逼近 0。但是高能力組及低能力組的結果就不同：在路徑前端的20題，其結果相當穩定；但是隨後就產生了很不穩定的現象，尤其是高能力組的 $|\hat{\theta}_F - \hat{\theta}_I|$ 有高達40以上者。顯然這是一個大得離譜的偏差，在測驗上發生這種現象，簡直與瞎猜無異。所幸在70題左右又恢復穩定。這樣不穩定的結果已經十分明確地指出使用固定題數來終止測量的危險性。

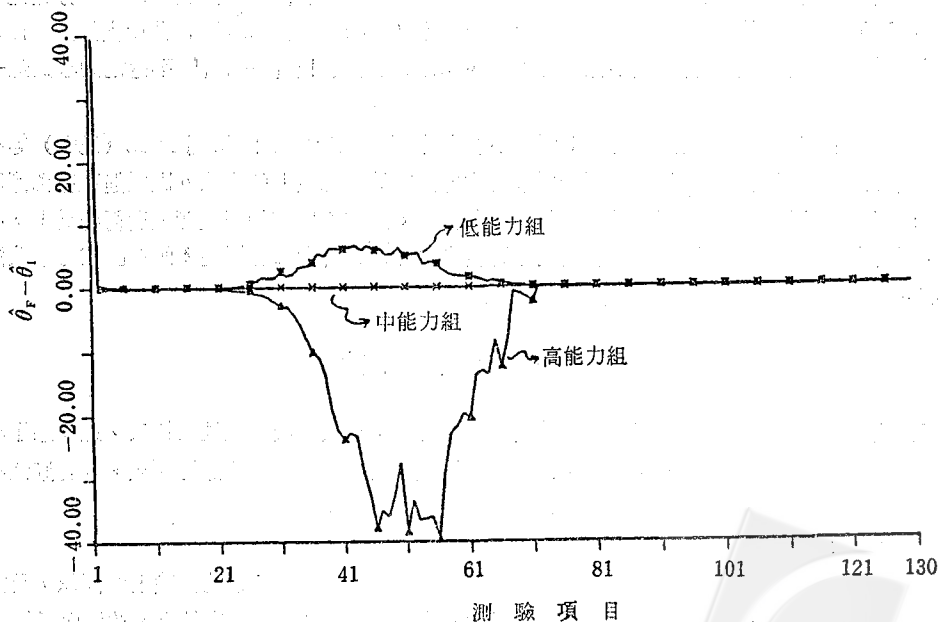


圖 33 三類適性測驗路徑 $\hat{\theta}_F - \hat{\theta}_I$ 的變化圖

爲何導致這種結果，其可能的原因有二：

第一是以最大可能性法估計 $\hat{\theta}$ 時，產生系統偏差。由圖33的三組受試者的路徑在 $\hat{\theta}_F - \hat{\theta}_I$ 的變化情形可以看出，最大可能性法估計 $\hat{\theta}$ ，對高能力的受試者有高估太多的趨勢，對於低能力的受試者則有低估的現象。這個現象顯示出：最大可能性法不一定是題數增加所估計的 $\hat{\theta}$ 就越近真正 θ （Hambleton, 1985）。

第二是題庫本身的限制。這可以從適性測驗的前20題穩定結果看出。因為題庫是由各種難度水準的項目所組成，能適合測驗某能力水準的項目通常也不會太多。因此在使用完適合的項目以後，所剩的就不是太適合的項目了。題庫的另一個限制是由於上述估計過高及計過低的系統偏差所致。因為當估計的 $\hat{\theta}_I$ 高估或低估太大時，那麼要根據 $\hat{\theta}_I$ 去計算項目訊息時，用來選取的適合項目在一般的題庫是不容易找到的。上述這種不穩定的現象在 θ 已知的適性測驗中，並沒有發現。主要的是因為它不是用 $\hat{\theta}_I$ 去計算項目訊息而是使用已知的 θ ，所以才不致於不穩定。雖然題庫的限制，可能導致適性測驗結果不穩定，但其真正基本的原因仍應是最大可能性法估計 $\hat{\theta}$ 的偏差。

2. 本研究之假設5—3是 θ 未知的適性測驗，以疊代法（最大可能性法）估計 $\hat{\theta}$ 時，聚斂受試者百分比呈單調性遞升趨近於100。結果（如圖30）除了中能力組聚斂百分比支持假設外，高能力組與低能力組聚斂百分比也有不穩定的現象。它的原因與 $|\hat{\theta}_F - \hat{\theta}_I|$ 不穩定的原因，可能是一樣。

值得一提的是：比較圖18與圖30可以看出，適性測驗路徑的前段。 θ 未知的適性測驗聚斂的反面比 θ 已知的適性測驗理想。原因可能是 Hambleton (1985) 指出的反應組型不正常，例如同一受試者答對困難的項目，却答錯簡單的項目。在 θ 已知的適性測驗裏，從開始就以難度匹配能力的項目來實施，因此有較多機會出現反應組型不正常；但是在 θ 未知的適性測驗裏，一開始是以 $b = 0$ 的項目實施，不一定是能力與難度的匹配，所以較不易產生不正常的反應組型，因而使用最大可能性法會聚斂得比較理想。

3. 本研究之假設5—2是： θ 未知的適性測驗，其逐步的 $|P_F - P_I|$ 呈單調性遞降趨近於0。結果與 θ 已知的適性測驗的結果類似，所以在適性測驗中使用答對百分比來計分，並不理想。

4. 本研究之假設5—4是： θ 未知的適性測驗，其逐步的 $I(\hat{\theta}_I) / I(\hat{\theta}_F)$ 的百分比是呈單調性趨近於100。結果（如圖31）除了中能組受試者的 $I(\hat{\theta}_I) / I(\hat{\theta}_F)$ 百分比支持假設以外，高能力組與低能力組的百分比也呈現不穩定的現象。其原因也與 $|\hat{\theta}_F - \hat{\theta}_I|$ 不穩定的原因是一樣的。

從以上對 θ 未知適性測驗的討論可知：研究不能支持 Kreitzberg & Jones (1980) 等所提「適性測驗可以使用固定題數或是固定測量標準誤的方法來結束測量」的說法。因為追蹤整個適性測驗路徑，所得測驗量結果並不會隨題數的增加而趨向穩定。這與 θ 已知的適性測驗所得的結果，不完全一樣。因為在 θ 未知的適性測驗中，有最大可能性法估計 $\hat{\theta}$ 時所形成的偏差；而 θ 已知的適性測驗中則沒有。

結 論 與 建 議

本研究爲試探性研究，目的在了解潛在特質理論的內涵，並與傳統測驗理論比較，進而想了解潛在特質理論在實際應用層面的意義。爲此，乃選擇最能表現潛在特質理論的適性測驗，探討適性測驗如何在潛在特質理論下運作以及適性測驗如何改善測量上的問題。

一、研究結論：

爲了達成上述本研究的目的，回答本研究的各項有關潛在特質理論及適性測驗上的問題，研究者採用模擬研究與系列研究的觀點，進行研究資料蒐集、分析，以獲具體之研究結果，並進而討論分析各項結果。現僅將結論列述如下：

(一) 從潛在特質理論中使用的參數與傳統測驗理論使用的項目統計數，比較分析結果，可知二者基本觀念是相通的。潛在特質理論的基本觀念是可以被接受的，正如同 Hulin等 (1983) 及 Welss (1983) 所說那樣，二者有重疊與包容的關係。

(二) 從適性測驗路徑的羣聚分析結果可知：適性測驗的實施其實就是以個人能力匹配項目難度的測驗實施方式。

(三) 從 θ 已知的適性測驗測量的結果得知，從理論上看適性測驗確實可以改善下列測量的問題：

1. 大約只須完成題庫的不到50%的項目，便可得到與做完題庫完全一致的結果。因此，測驗長度可以縮短。

2. 在測驗長度縮短後，測量標準誤並不會增加。

3. 測量工作及測驗方式是穩定可靠的。

(四) 由 θ 未知的適性測驗測量的結果得知：從實際進行適性測驗結果並不樂觀，因為整個測量工作及測驗方式並未穩定。但如果能實施50%以上的項目，結果仍會慢慢穩定的。

就適性測驗的理論架構及模擬結果，研究者可以肯定的說：應用潛在特質理論於適性測驗，是絕對有助於測量的。但是，在實際應用上，仍未達完全成熟的階段，主要困難是技術層面的問題，像最大可能性法在估計 $\hat{\theta}$ 的偏差問題便是。

根據以上的結論及研究者對於整個研究的認知，提出下列建議：

二、理論探討方面的建議：

(一) 潛在特質理論十分風行，但正如 Anastasi (1982) 所說潛在特質理論至今仍處形成階段，存有許多爭議，多數研究均是示範性質的研究。但是，不可否認的，它正在快速地影響整個心理測驗界。這方面的研究值得留意與投資。

(二) 模擬研究的弊端是易流於紙上談兵，建議從事真實測驗資料的研究。

(三) 要從事潛在特質理論方面的研究，首先必須充實研究設備及工具，尤其電腦的硬體及軟體。師大教育心理系甫自美國測驗服務社引進 LOGIST (Wingersky, Barton & Lord 1982)，正是為這一類研究工作鋪路。

三、在適性測驗方面的建議：

測量工作者均職志於改善測量工作，適性測驗便是重要的一環。以下研究者試著列出一套適性測驗可行的步驟，以供參考：

(一) 準備階段

1. 選擇並預先評估理論模式：根據測量的目的及資料特性選取理論模式，例如本研究所使用的三參數對數模式是適用於選擇式的測驗。

2. 建立題庫：根據本研究結果知道題庫的限制是適性測驗不穩定的原因之一。因此一個廣大的題庫是適性測驗所必須具備的。所以，應及早依照測量的目的，開始命題工作。命題本身是一件專業性的工作，宜有專人從事專門命題。尤其是極端難度的項目，命題更是困難，這也是研究者從事模擬研究的原因之一。命題也須事先考慮將來適性測驗進行的方式，是團體方式進行，是個別測驗方式，或是以電腦輔助進行。若是採用電腦輔助，則測驗項目可以使用遊戲或動畫式的呈現 (Kreitzberg & Jones, 1980)

3. 校準測驗項目參數：以題庫的項目從事預試及項目分析之類的工作。在潛在特質理論中使用的方法是這樣的：第一步，將題庫拆開成幾個部份，但每部份之間均有幾個重疊的題目。第二步是將各個部份分別對不同的受試者進行預試。第三步是將預試的資料使用項目參數估計的方法，分別估計出各部分的項目參數。第四步是利用重疊項目，使用測驗對等的方法將整個題庫的項目參數建立在相同的量尺之上。如此，校準項目參數工作即告完成。

4. 實際評估理論模式：分析預試資料，評估所得結果與理論模式之間符合的程度。若符合則往下繼續進行適性測驗階段的步驟；若不符合則再重複 1, 2, 3 的步驟，調整其中部份，直到第 4 步驟顯現實際資料與理論模式相符合為止。

(二) 適性測驗階段

1. 決定起始點：在本研究中是使用中難度項目，來做為第 1 題。結果發現很快地便可各自到達各能力水準的難度項目。Kreitzberg & Jones (1980) 是使用受試者年級水準，來決定第 1 題的。

2. 項目選擇策略：第 1 題做完，接著往下的第 2, 3, 4 …… 題，到底要使用那一種選擇的方法呢？在本研究中是使用最大訊息法，在 θ 已知的適性測驗中發現，它只是難度與能力匹配選擇而已。其它還有上下法、H—L 法、羅一門二氏法、貝氏估計法等。

3. 估計能力：這是測量工作的真正目的所在。要將受試者對項目的反應組型資料，用能力估計的方法推估受試者的能力。在本研究是使用最大可能性法。結果發現這個方法在適性測驗中會產生系統的偏差和方法上聚斂與否的問題。因此必須設法改善這方法。Jones (1982) 則使用強韌的方法 (robust method) 來改善不正常反應組型的估計。本研究的發現亦可提供改善最大可能性法之用。但也可以使用其它方法估計。

4. 結束測量：結束測量工作是適性測驗所必須的工作。目前常使用的是固定題數及固定測驗訊息的方法。但是本研究發現使用這種方法，不恰當，所以有待進一步研究。

5. 評估適性測驗：選擇效標評估適性測驗效度，再反覆 1, 2, 3, 4 步驟直到效度達到理想為止。

以上是研究者就所知，列出適性測驗的可行步驟。在每一此驟中，均存有許許多多待解答的問題，是一塊值得開採的領域。

參考文獻

- 林一真 (民71)：潛在特質理論簡介——測驗編製的發展新趨勢。載於路君約等編：我國測驗的發展。臺北市，中國行為科學社，61~70頁。
- 林邦傑 (民70)：集羣分析及應用。國立政治大學教育與心理研究，4 期，31~57頁。
- 林清山 (民74)：羣聚分析的理論和統計方法以及應用羣聚分析的實徵性研究。中國測驗學會測驗年刊，32輯，155~180頁。
- Ambrose, M. L. (1983). *Application of item response theory to the detection of item bias between males and females in a test of mathematical aptitude*. Unpublished master's thesis, University of Illinois, Urbana, IL.
- Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp.508-600). Washington DC: American Council on Education.
- Baker, F. B. (1971). Automatioin of test scoring, reporting, and analysis. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 202-234). Washington DC: American Council in Education.
- Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Report 81-20). Princeton, NJ: Educational Testing Service.

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 625-670). Washington DC: American Council on Education.
- Green, B. F., Jr. (1970). Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance* (pp. 184-197). New York: Harper & Row.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.
- Hambleton, R. K. (1985). *Item response theory: Principles and application*. Boston: Kluwer-Nijhoff.
- Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.
- Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 130-159). Washington DC: American Council on Education.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones Irwin.
- Jones, D. H. (1982). *Tools of robustness for item response theory* (Research Report 82-41). Princeton, NJ: Educational Testing Service.
- Jones, D. H. (1984). *Bayesian estimators robust estimator: A comparison and some asymptotic result*. (Research Report 84-42). Princeton, NJ: Educational Testing Service.
- Jones, D. H., Wainer, H., & Kaplan, B. (1984). *Estimating ability with three item response models when the model are wrong and their parameter are inaccurate* (Research Report 84-26). Princeton, NJ: Educational Testing Service.
- Kreitzberg, C. B., & Jones D. H. (1980). *An empirical study of the broad range tailored test of verbal ability* (Research Report 80-5). Princeton, NJ: Educational Testing Service.
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance* (pp. 139-183). New York: Harper & Row.
- Lord, F. M. (1974). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 2. Measurement, Psychophysics, and neural information processing* (pp. 104-126). San Francisco:

- W. H. Freeman.
- Lord, F. M. (1980a). *Applications of item response theory to practical testing problems*. Hillsdal, NJ: Lawrence Erlbaum Associates.
- Lord, F. M. (1980b). Some how and which for practical tailoring testing. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 189-205). Chichester: John Wiley & Sons.
- Lord, F. M. (1981). *Unbiased estimators of ability parameters of their variance and of their parallel-forms reliability* (Research Report 81-50). Princeton, NJ: Educational Testing Service.
- Lord, F. M. (1984). *Maximum likelihood and Bayesian parameter estimation in item response theory* (Research Report 84-30). Princeton, NJ: Educational Testing Service.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lord, F. M., & Wingersky, M. S. (1982). *Sampling variances and covariances of parameter estimates in item response theory* (Research Report 82-33). Princeton, NJ: Educational Testing Service.
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
- McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.
- Norusis, M. J. (1985). *SPSS-X advanced statistics guide*. New York: McGraw-Hill.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-257). New York: Academic Press.
- SPSS INC. (1983). *SPSS-X user's guide*. New York: McGraw-Hill.
- Stocking, M. L. (1984). *Two simulated feasibility studies in computerized adaptive testing* (Research Report 84-15). Princeton, NJ: Educational Testing Service.
- Thissen, D., & Wainer, H. (1985). *Some supporting evidence for Lord's guideline or estimating 'c'* (Research Report 85-15). Princeton, NJ: Educational Testing Service.
- Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.
- Traub, R. E., & Lam, R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology*, 36, 19-48.
- Urry, V. W. (1977). Tailoring testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 14, 181-196.

- Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 63-81). New York: Academic Press.
- Warm, T. A. (1978). *A primer of item response theory* (Technical Report CG-941278). Oklahoma City, OK: U. S. Coast Guard Institute.
- Weiss, D. J. (1974). *Strategies of adaptive measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Psychometric Method Program, Department of Psychology.
- Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *Logist user's guide*. Princeton, NJ: Educational Testing Service.
- Wood, R. (1974). Response-contingent testing. *Review of Educational Research*, 43, 529-544.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.



Bulletin of Educational Psychology, 1987, 20, 131—182
Taiwan Normal University, Taipei, Taiwan, China.

STUDIES ON LATENT TRAIT THEORY AND ITS APPLICATION IN ADAPTIVE TESTING

SIEH-HWA LIN

ABSTRACT

The purposes of this study were:

- (1) to investigate the relationship between latent trait theory (LTT) and classical test theory (CTT), and
- (2) to evaluate the efficiency of latent trait theory applied to the adaptive testing.

Computer simulation was used in this study. The response patterns of the "examinee" were first generated, and the procedures of adaptive testing were simulated, supposing θ known and θ unknown. The testing paths of each examinee were traced, recorded, and analyzed.

The results were summarized as follows:

- (1) The a , b , θ parameters in LTT were highly correlated with r_{b1s} , P , and obtained score in CTT respectively.
- (2) when θ was supposed to be known, the adaptive testing paths could be classified into three patterns according to b parameter in testing path. These three paths patterns belonged to examinees of high, middle and low ability level respectively. Apparently, the paths were formulated by matching examinee's ability and item difficulty. when less than 50% of items in the pool were taken, the efficiency of measurement was the same as when all items in the pool were taken.
- (3) when θ was supposed not to be known, the adaptive testing paths could be classified into two patterns according to a parameter, into three patterns according to b parameter, and into four patterns according to c parameter in path. The Paths were also formulated by matching examinee's ability and item difficulty. However, the efficiency of measurement was not stable.

Based on the results of this study, the conclusions were drawn:

- (1) The basic concepts of LTT and CTT were similar and consistent.
- (2) It was practicable to apply LTT to adaptive testing, however the techniques of estimating θ should be improved further.