成數個階段來進行，第一個階段是先發展一份具有全台灣區常模的測驗，以後再以較小樣本，採共同試題設計的方法，逐漸將新試題加入原有題庫。本研究即爲整個學術性向測驗電腦化題庫建立的第一階段工作。

貳、學術性向測驗編製的理論基礎

一、學術性向測驗之意義及與成就測驗之異同

成就測驗與性向測驗有某些相似的地方，但也有一些相異之處。一般認爲二者主要不同在於「成就測驗是測量學生已學得的知能，而性向測驗是測量學生學習新知能的能力」。此種觀點似乎能明晰區分二者之不同，但是卻過分簡化這個問題，而且未能指出二者某些重要的相似點和差異處。事實上，這兩類測驗均在測量學生已學得的知能，而且二者對能否成功學習新工作的預測也都很有用。成就與性向測驗二者主要的不同是在二者所測量的學習類型及最有用的預測類型稍有不同（Gronlund，1985）。

㈠所測量的學習類型：能力譜（the ability spectrum）

Cronbach（1984）曾以圖 2.1 的能力譜來說明成就和性向測驗所測量的各類學習結果之同異。在這個能力測驗譜中，成就測驗含括層次A和B，而性向測驗含括層次C，D，和E。此譜是依照測驗內容依賴特定學習經驗的程度來對各種測驗加以分類。最上端（層次A）是測量特定課程內容知識的「內容導向成就測驗」。最下端（層次E）是測量不易受直接訓練影響的那類學習的「文化導向非語文性向測驗」。當我們從層次A逐漸移至層次E，測驗內容就越來越不受任何特定學習經驗的影響。能力測驗譜上越是相近的兩類測驗，其所測量的學習類型也就越相近。例如，設計來測量一般教育發展的成就測驗（層次B）和以學校所學能力爲依據的學術性向測驗（層次C），二者所測量的能力很類似，因此，二者可預期有相當高的相關。同樣地，在能力測驗譜上距離較遠的兩個測驗，由於共同的部分較少，因此，其相關也就較低。此種資訊有助於性向測驗的選擇和使用。例如，我們可以預期層次C和D的性向測驗，對於學校成就的預測要優於層次E的性向測驗。另一方面，如果我們主要的興趣是在鑑定學生未開發的學習潛能（potential），那層次E的測驗就比較適合。圖 2.1

的能力測驗譜只是爲了分類上的方便而設計，並非每個能力測驗均可明確歸入某一層次，有某些測驗包括兩個以上的層次。 例如， Otis-Lennon School Ability Test ，雖然使用單一測驗分數，但其測驗內容包括了層次C，D，和E。

圖2.1 能力測驗譜各層次測驗試題內容之性質

層次	測驗型態	測驗內容之性質
A	內容導向成就測驗 (如，Tests of Achievement and Proficiency)	課程（如社會科、英語、數學、自然科學）教材知識。
B	一般教育發展測驗 (如，Iowa Tests of Educational Development，STEP)	很多課程共同的基本技能和複雜學習結果，如，原則應用和資料解釋。
C	學校導向性向測驗 (如，Cooperative School and College Ability Tests，SCAT)	語文、數字和一般問題解決能力，大多與「學校」所學的相似，如，字彙、閱讀、和算術推理。
D	文化導向語文性向測驗 (如，Cognitive Abilities Test，Verbal Battery and Quantitative Battery)	與層次C一樣是語文、數字和一般問題解決能力，但取自一般文化的多於取自一般學校經驗者。
E	文化導向非語文性向測驗 (如，Cognitive Abilites Test，Nonverbal Battery)	以圖形類推、圖形系列、圖形分類和其他非語文試題來測量的抽象推理能力。

*改自 L.J. Cronbach，Essential of Psychological Testing，4th ed. (NY：Harper & Row，1984)

㈡ 成就和性向測驗所做的預測類型

成就與性向測驗也可以根據其最有用的預測類型來區分。由於一個人過去的成就，常是預測未來成就的最佳預測變項，因此，層次A及B兩種成就測驗均可有效預測未來的學習。

一般而言，內容導向成就測驗（層次A）可預測學生在相同內容領域新知識的學習有多好，但是對於未來其他領域學習成就之預測則沒有很大的價值。例如，國中一年級上學期的英語成績是預測下學期的英語成績之有效變項，但卻無法有效預測歷地、數學、理化、國文或其他學科的成就。換句話說，能否做爲預測未來學習成就的良好預測變項，要視預測變項所測量內容與未來學習情境的內容之關聯強度而定。測量一般教育發展的測驗（層次B）要比內容導向測驗（層次A）更能有效預測未來的成就，這是因爲層次B的測驗是在測量各內容領域共同所需的智能（intellectual skills and abilities）。

雖然成就測驗已能有效預測未來的學習成就，但是性向測驗（層次C至E）仍有下列幾個成就測驗無法完全取代的重要特點，因此，學校仍有使用性向測驗之必要。

(1)性向測驗施測所需時間通常較短（某些測驗甚至可在20分鐘之內測畢），而綜合性的成就測驗組（comprehensive battery of achierement test）施測所需時間則長得多（幾個小時）。

(2)性向測驗較適合測量教育背景差異較大的學生。這是因爲性向測驗所測量的學習類型是大多數學生共同的學習經驗，比較不會有某些學生由於在過去的教育訓練中有某些特殊弱點而受到傷害。

(3)學校可以在學生未接受某一特定領域的任何訓練之前就使用性向測驗。例如，學生沒有學英語之前，我們無法用英語成就測驗來預測他未來英語學習成就，一定要學生學過某些英語之後，才有可能用目前英語成就來預測未來英語成就。但是性向測驗所測量的較一般性，對各學科均有某些程度的預測力。

(4)性向測驗（特別是文化導向的性向測驗，即層次D及E）所測量的能力，較不受學校學得能力之影響，因此，可用來區分已經充分發揮其能力的低成就者和未充分發揮其能力者之不同。如果使用較偏重學校學得能力的性向測驗（層次C）來鑑定上述低成就者，雖然還是有可能，但是卻比較沒有效率，這是因爲層次C的測驗所需的成就技能，很可能是低成就者的弱點。

綜括以上之分析和討論，可將成就與性向測驗之異同歸納如下：成就

測驗與性向測驗均在測量學得的能力，但成就測驗所測量的能力與特定的學校經驗較有直接關聯，而性向測驗所測量的較爲廣泛，同時包括校內與校外經驗。不過，這也只是程度上的差異，我們可以將兩類測驗同時安置在同一能力測驗譜上，從測量特定課程內容的測驗（層次A），到文化導向的學得能力測驗（層次E）。在能力測驗譜較中間部分（層次B和C），成就測驗與性向測驗變得非常相似。成就和性向測驗另一相似點是二者均可有效預測未來的成就。一般而言，性向測驗使用較方便，而且所能預測的未來經驗也較廣。和前述所測量學習結果的類型一樣，當所預測的經驗接近能力譜中間時，兩類測驗預測功能的差異就較不明顯了。

二、智力、學術性向、和學習潛力

傳統上將用來測量學習能力的測驗稱爲智力測驗（intelligence tests）。目前雖仍有很多個別測驗和某些團體測驗使用這個名詞，但已有漸少之趨勢。主要原因是：(1)很多人會將智力這個概念和遺傳能力相連結；(2)對於智力概念的意義以及智力所應包括的因素之看法越來越分歧；(3)此領域的測驗已逐漸用來預測而非用來描述。例如，如果我們主要的興趣是在預測成就，那學校學得能力的混合測驗要比一個更純粹的智力測驗更有用。目前已逐漸用心理能力測驗（mental ability tests）、學校能力測驗（school ability tests）、認知能力測驗（cognitive ability tests）、及學術性向測驗（scholastic aptitude tests）等名稱來取代「智力測驗」。在各級學校所使用的這類測驗，較多採用「學術性向測驗」這個名稱，因此，本研究所編製的測驗也稱之爲「學術性向測驗」。

學術性向測驗並非直接測量先天能力或學習潛能。性向測驗與學校所使用的其他測驗一樣，都是在測量已學得能力。就這個觀點而言，性向測驗可視爲一類特殊的成就測驗。任何根據學術性向測驗結果所作的有關學習潛力（capacity or potential for learning）的推論，只有在符合下列條件時才有效（gronlund , 1985）

1.所有學生均有同等的機會學習性向測驗所呈現的作業。

2.所有學生均有同樣強烈的動機回答測驗所呈現的問題。

3.所有學生均具備要在測驗上有最大表現所需的行爲能力（enabling

behaviors，如閱讀能力）。

4.沒有任何學生因爲受到考試焦慮、情緒問題、或其他不利行爲（disabling behaviors）等足以防礙其最佳表現的因素之影響，而降低其得分。

當然，上述那些條件幾乎不可能充分滿足。如果我們根據學術性向測驗分數來估計學生的學習潛能，則上述那些條件不符合的程度，決定了我們所犯錯誤的大小。有很多錯誤的解釋和誤用學術性向測驗，是因爲不了解那些條件對測驗結果以及根據測驗結果所作的推論之影響。

學術性向測驗分數較適切的解釋是將其視爲目前的學習能力量數。此分數必然反應某些程度的遺傳特徵，這幾乎是沒有人會反對的，但也必然反映個人的經驗背景、動機、考試技巧（test-taking skills）、持久性、自信、及情緒調適等人格特質。這些因素同時影響學術性向測驗分數和學業成就。但是這些因素中有很多是可以透過訓練而改善的，只要經過適切的訓練，學習能力和學業成就均可提高。如果我們將學術性向分數視爲是學習潛能的直接測量，或是視爲是無法改善的，那就很容易誤用測驗結果。

三、與學術性向測驗有關之智力理論

在人類諸多特質中，智力是心理學上研究最多的主題之一。自從Binet 和Simon 首創智力量表以來，目前已問世的智力測驗已不下數百種，且仍有不繼增加之趨勢。有關智力的各種研究論著尤難以數計，即以美國「心理學摘要」（Psychological Abstract）上歷年來所收錄的這方面文獻來說，估計已在一萬種以上，足見有關智力的研究在心理學領域中所佔的地位。而其研究方法上，亦從最初的以心理計量法（psychological approach）爲主要研究方法的局面，漸而兼採用發展理論（developmental theory）和信息處理學說（Information processing）等研究法。以下將對智力這一個構念的性質和有關問題作簡要的討論。

㈠ 智力的定義

關於智力的定義，研究者有許多不同的看法。如 Terman（1921）認爲智力乃是抽象思考的能力；Thorndike（1921）主張智力是對抽象事物作良好反應的能力；Colvin（1921）將智力界定爲一個人的學習的能力；

Pintner（1921）則認爲智力代表一個人對新環境適應的能力。 智力測驗的先驅者 Binet 籠統的將智力解釋爲多種能力（如、理解、判斷、注意、想像等）之總合以及個體適應環境之能力（Binter & Simon, 1916）。後來的學者企圖從不同的角度來定義智力，如 Stoddard（1943）相信智力是一個人解決當前問題，並爲未來問題預作準備的能力； Wechsler（1958）綜合各家的觀點將智力界定爲「有目的的行動、合理的思考、和有效地應付環境之總能力」；發展心理學者 Piaget 認爲智力是一個人思考和推理的能力（Elkind, 1969）。晚近的研究者對智力的理解則有偏重認知過程（cognitive processes）和問題解決（problem-solving）能力的傾向， Hunt(1976) 等人便主張應從認知過程的觀點去了解智力的性質；Simon (1976) 並提議在分析智力的因素時可以用電腦模擬的方法爲之；另外，Charl- esworth(1976) 認爲智力包括兩種特性：一是認知過程，另一爲適應環境的能力。

雖然智力的定義各家有異，但如果深入檢視，便會發現它們之間實有許多相同之處，其差異只在於每一定義所涵蓋的範圍和强調的重點有有不同。概括來說，歷年來有關智力的定義大致可分爲以下四類：一、智力是抽象思考和推理的能力；二、智力是學習的能力；三、智力是環境適應的能力；四、智力是問題解決的能力（葛樹人，心理測驗，民77）。

㈡智力的理論

傳統上有關智力理論的研究，主要以因素分析法爲基礎。早期的測驗學者大多認爲智力是許多心理官能（mental faculties）相結合的總能力（The gfactor）。後來Spearman發明了因素分析法，並利用此法來研究智力的組成和性質質等問題，才在智力理論的探索上開拓了一個新的視野。多年來鎮隨著因素分析法的演進，以此一方法爲基礎的智力理論亦益趨發達。底下就對這方面的幾個主要理論加以討論，至於其它次要的智力理論，如學習理論、發展理論、認知理論等，則不予以介紹。

因素分析的智力理論可概分爲兩大學派，一派强調智力的整體性，另一派則認爲智力可進一步分成不同的因素。前者包括Gatlon，Binet，Spearman等人；後者則以Thurstone，Cattel，Guilford等爲代表。以下就扼要討論幾個重要的理論，以做爲瞭解智力測驗的編製依據：

1.二因論

英國心理學家 Spearman 首先依據測驗分數的統計，提出二因論（two- factor theory）。他認爲智力是由一個普通因素（general factor，簡稱 g因素）和許多特殊因素（specific factor，簡稱 s 因素）所構的。普通因素是所有心智活動所共同必需的，心智活動愈複雜時其所需的普通因素也愈多，如學校各科目的學習都需要此種能力。Spearman 相信每個人皆有普通智力因素，唯其多寡的不同造成個體之差異現象，故 g 因素是決定一個人在智力測驗上表現的主要因素；特殊因素則是某種特殊的心智活動所必需的，如空間關係，運動協調等。這些因素並不爲所有的智力活動所共有，而僅存在於個別的智力測驗中，由特殊性的智力測驗加以測量。每個人在智力測驗上的表現乃是其普通因素和特殊因素作用的結果，但普通因素在智力測驗上佔有遠比特殊因素重要的地位。

2.群因論

群因論（group-factor theory）是由 Thurstone（1938）所提出的。他將當時常用的五十幾種智力測驗加以重複因素分析，結果得出了六種主要智力因素，並將它們命名爲主要能力（primary abilities），包括語文理解（verbal- comprehens-ion），數字運算（number），空間關係（space），語文流暢（word fluency），聯想記憶（associative memory），歸納推理能力(inductive reasoning) 等。這六種主要能力在智力的組成和測量上具有相當的重要性。根據此研究結果，Thurstone 發展出第一個多元性向測驗組（multiaptitude test batteries），稱爲基本心理能力測驗（Primary mental abilities Test），用來測驗這六種能力， 並利用測驗上的剖面圖（profile）來分析一個人的智力運作情形。基本心理能力測驗可測量上述六種基本心理能力，這些能力不像普通因素（g 因素）那樣普通，也不像特殊因素（s 因素）那樣獨特，它是屬於一群心智活動所共有的。 Thurston 在這方面的研究對後來的團體性向測驗產生了相當的影響（郭生玉，民 74）。

3.多因論

Thondike（1921）主張智力是由許多不同的能力所構的。他認爲智力可分爲三大類， 即社會的（social）、具體的（concrete）和抽象的（

abstract）。Thondike 的理論不是根據因素分析的結果， 他可能是提出多因論（Multiple- factor theory）的第一位心理學家。 他曾編製了一系列命名爲 CAVD 的測驗，用來測試語句完成、算術推理、字彙和遵令行事等四種學習上的主要能力。

後來，Guilford（1956，1967）應用因素分析法研究二十多年，提出一個智力的結構模式（the structure-of-intellect），說明各種不同的智力成分。在這個模式中，他依三個向度（three dimension）將人類的心理能加以分類：

(1)思考的內容（content）：個人思考運考的材料或資料，即測驗內容的種類。包括圖形的、符號的、語意的與行爲的等四種。

(2)思考的運作（operation）：處理上述各種資料所使用的思考方式，包括認知、記憶、擴散思考(divergent thinking)、聚斂思考（convergent thinking）和評價等五種。

(3)思考的結果（product）：對某種測驗內容運用不同的思考方式所獲得的結果，包括單位、類別、關係、系統、轉換與應用等六種（郭生玉，民74）。

依據四種不同的材料，採用五種不同的思考方式，得到六種不同的思考結果，這些分類的交互作用共產生 120 種獨特的智力因素（4*5*6）。如個人對圖形的類推測驗（內容），運用認知的方式思考（運作），由其關係行爲(結果)表現出來，這是屬於「認知-圖形-關係」(cognition-figural-relation）的能力。同樣的，對語文類推測驗，個人需認識其語句意義的關係，故屬於「認知-語意-關係」的能力。

Guilford 的智力結構理論將傳統智力的概念擴大了，在智力的了解和測量上有其相當的貢獻，但多數的心理學家認爲其模式的理論價值遠大於實用價值。Guilford（1977）將思考內容的「圖形」又細分成視覺與聽覺兩類，總共可得到150種（即5×5×6）不同的能力。

4.階層理論

階層理論（hierarchical theory）是由英國心理學家 Vernon（1960）所提出。最高層次是普通因素（g），在此因素之下有兩個主要群因素：一是語文與教育的能力（V：ed），另一是實用或機械的能力（K：m）。主要群因素之下又可分爲許多較小的群因素，例如語文與教育的能

力包括機械理解、動作能力與空間關係等。在較小群因素之下是屬於最低層次的特殊因素。此理論統合了上述Spearman 的 g 因素，Thurston 的群因素和Guilford 的多因論，因此最受美國心理學家的喜愛。

5.其他的理論

Cattell（1963，1971）的因素分析理論主張智力可分爲流體智力（fluid intelligence）和結晶智力（crystalized intelligence）兩種。流體智力主要來自遺傳，是各種不同領域所共有的。具有較多流體智力的人能夠做好許多不同的工作。這種智力與一個人的學習和速度有高度的正相關，但它不受環境經驗或文化背景的影響。結晶智力則相反，它是一種經學習和經驗而逐漸累積得來的智力，可被看成是一個人智識和專門性技能之總體。結晶智力是某些領域所獨有的，欲在特定的工作中獲得成就，需要具備此種智力。在固定習慣性的工作中，使用較多此種智力。流體智力通常是採用文化公平測驗測量，結晶智力則採用語文測驗測量。Cattell 研究發現， 流體智力發展到 14 歲或 15 歲爲最高峰，但結晶智力則持續發展到 25 歲或 30 歲。

另外，Jensen（1969，1973）提出二層次理論（two-level theory）。第一個層次爲聯想智力（associative intelligence），包括機械學習能力和記憶能力； 第二個層次爲抽象智力（abstract intelligence），包括思考推理與問題解決能力，類似於 Spearman 的 g 因素或 Cattell 的流體智力。

四、題庫擴充時所需之等化與連結

題庫擴充時，新題目的量尺必須與原有題庫的題目參數之量尺相同，因此就需要用到等化與連結方法。

㈠等化

等化（equating）是將一式測驗上所得之分數，以統計公式轉換調整至另一式測驗分數量尺上的過程（Dorans、1990）。基本上等化乃是決定二個或二個以上不同式測驗（或量尺）分數之間關係的過程。換句話說，也就是要將不同式測驗結果的分數放到同一量尺單位上，其過程乃在於調整題目及測驗特徵的差異（Hung、1990）。

當測驗結果所做的決定具有很大影響時，基於公平、保密等之考慮，通常必需編製許多平行測驗，如果二式測驗完全合於平行的定義，等化就不需要了，然而編製嚴格平行測驗幾乎是不可能的，故只有在測驗分數經過適切的等化後，才能有意義的、公平的比較測驗結果。而且在發展標準化測驗的過程上，如修訂、增加題目……等，等化均是一項非常重要的工作，藉此新、舊版測驗間才能相互比較，題庫才能加以擴充，有關文獻的連貫也才有可能。等化有下列兩種型式：

1.平行等化

有某些考試（如托福、GRE 等）需要有多式複本測驗，才能在一年內實施多次測驗。但是，如果不同式測驗分數的分佈並不相同，那就需要決定不同式測驗相互等值的分數。此種爲多式複本測驗建立相互等值分數的系統，稱之爲平行等化（horizontal equating）。平行等化的目的在於編製含有多式，但在難度上相平行的測驗。

2.垂直等化

有某些測驗組（test battery）包括不同難度水準的測驗，如「修訂 Otis- Lennon 心理能力測驗」的水準三適用於小學一年級下學期至小學三年級下學期，水準四適用於小學四年級至六年級。在這些心理測驗中，題目難度的設計原本就有顯著的不同，要比較學生在這兩個測驗上的表現，就需決定兩個水準相互等值的分數。此種過程稱爲垂直等化（vertical equating）。垂直等化的目的在於編製同一特質而含有不同程度的測驗，並且這些不同水準的量表須使用相同的計分量尺，通常用於對個體長時間學習發展的縱剖評估。無論在理論或實施方面，垂直等化皆比平行等化複雜多了。

㈡連結

連結（linking）是 IRT 等化中一個必要的步驟，其目的是將所有題目的參數放到同一單位的量尺上，使題目的參數經連結後可互相比較。以下提出一些學者對於 IRT 連結的研究報告：Vale、Maurelli、Gialluca、Weiss和Ree（1981）對共同試題題數及其訊息曲線的形狀與等化結果的關係進行探討，結果發現 15 題至 25 題共同試題是必要的，共同試題的訊息曲線以長方形或常態爲佳。尖峰式的訊息效果較差。

Mckinlex 和 Reckase（1981）認爲 5 題共同試題是不夠的，25 題的結果較 15 題好，但如採用同時估計法則 15 題共同試題即已足夠。Raju，Edward 和 Osberg（1983）於垂直等化的研究中發現6至 8 題的共同試題就已足夠。Wingersky 和 Lord（1984）發現當採用同時估計法時，共同試題甚至於可少至 2 題，只要估計誤差小，其效果與 25 題不相上下。

Cook 和 Eignor（1981）對等化設計、共同試題、其估計標準誤和考生能力分配形狀進行研究，發現同時估計法與特徵曲線轉化法並無差異，但同時估計法較不需要太多的共同試題。共同試題估計標準誤儘可能接近原測驗的估計標準誤，以長方形分配爲佳，尖峰式的分配較差。Wingersky，Cook和 Eignor（1986）以三參數模式和模擬資料進行研究，發現以較長的共同試題進行連結時，能提供較穩定的結果。Hung（1990）以模擬資料進行研究，結果發現就難度而言，連結對於眞值與估計值及預定值與估計值間誤差的減小，是有幫助的；但對鑑別度則不然，連結會造成更大誤差、偏差、甚至系統性的偏差。在四種連方法中，以特徵曲線轉化法和平均數、標準差法二者對難度的連結結果最好關於考生特質方面，當樣本中所缺少的考生的能力位於該測驗能提供較豐富訊息的能力點上，則連結對難度量尺可有較大的助益；但如缺少位於測驗訊息較少的考生，則連結是不必要的 。

另外，採用 IRT 作爲等化之依據時，還需假定所要等化或連結的測驗是單一向度的。

本研究同一年級甲、乙兩式測驗之等化屬於平行等化，五、六年級之等化屬於垂直等化。由於本研究的學術性向測驗包括語文、數量及圖形三個分測驗，此三個分測驗之內容性質不同，因此採用的共同試題題數較多，預試時，兩式測驗有 24 題共同試題（每一分測驗 8 題），正式題本有 20 題共同試題，其中語文有 6 題，數量及圖形各有 7 題。在各種題目參數連結方法中，本研究採用「平均數、標準法」及「同時估計法」（或稱聯合估計法）。同一年級「高雄市與台南市」及「台灣省與台北市」四區同一測驗題目參數之轉換採用「平均數、標準差法」，同一年級男、女生題目參數之相互轉換（題目偏失研究時）也採用「平均數、標準差法」；

五、六年級題目參數之相互轉換採同時估計法，所得題目參數量尺屬於高年級，而非五年級或六年級。

㈢題目參數連結方法

根據原始分數的等化無法滿足公平、對稱和不受樣本變動影響等條件。如果受試反應的資料適合某種題目反應模式，那根據題目反應理論的等化可以解決原始分數等化的那些問題 。

依據題目反應理論，受試能力 θ 不受施測的試題子集之影響，而題目參數則不受受試群體之影響。當題目參數是已知的，由於 θ 的估計值是一致的，因此，θ 的估計值不會受到不同的試題子集影響。所以，不管受試接受較難或較易的測驗測試，所得的結果並沒有什麼不同。就題目反應理論來說，不管是水平或垂直等化的情境，只要將試題參數轉換至同一量尺即可，並不需要進行測驗分數的等化。但是，對不同受試群體施測，所得的題目參數表面上看來是不同的。所以會有這種差異是因爲我們任意固定 θ （或 b) 的零點。 不過從兩組受試所得的題目參數和能力參數存在著一種直線關係。此直線方程式的常數可用下列方式決定：

a. 迴歸法（regression method）；

b. 平均數、標準差法（mean and sigma procedure）；

c. 強韌平均數、標準差法（robust mean and sigma procedure)；

d. 重複強韌平均數、標準差法（iterative robust mean and sigma procedure）；

e. 特徵曲線法（characteristic curve method）；

本研究僅使用「平均數、標準差法」及同時估計法，故僅就此二法略作介紹，其餘請參閱吳裕益（測驗的等化，民 80）。

1.平均數、標準差法

此種方法是採用估計的題目難度值分配中的二級動差，即找到斜率 B 及截距 A 將測驗二的難度參數估計值的平均數和標準差轉化爲與測驗一的難度參數估計值的平均數和標準差。

$$\text{斜率 } B = \frac{S_{b_1}}{S_{b_2}}$$

$$\text{截距 } A = \bar{b}_1 - B\,\bar{b}_2$$

$$b_2^* = B\,b_2 + A$$

其中S_{b_1}，S_{b_2}分別爲測驗一及二之難度(b)的標準差。

$\bar{b}_1$，$\bar{b}_2$分別爲測驗一、二難度(b)的平均數。

b_2^*爲轉化到測驗一量尺後的新難度值。

b_2 爲測驗二量尺上的難度值。

此種線性轉化方法優於迴歸法，因爲它是對稱的，而其邏輯在於對共同試題組而言，二份測驗的估計值應是完全相關的，其差異來源只在於原點和單位不同（Hung，1990b）。

2.同時估計法

本研究所採用同時估計法是以 MicroCAT 測驗系統中的 ASCAL 程式同時估計兩式測驗所有題目的三參數。其使用的步驟與一般用法無異，只是在資料檔上的編排有些不一樣。由於甲、乙式有 20 題共同試題，因此進行同時估計法時，測驗的題數就以 100 題來估計，其資料檔格式如圖 2.2 所示。

圖2.2　同時估計法資料格式

<table>
<tr><td colspan="4">控　制　列</td></tr>
<tr><td colspan="4">正　確　答　案</td></tr>
<tr><td colspan="4">選　項　數</td></tr>
<tr><td colspan="4">測驗所包含的題目之代碼</td></tr>
<tr><td rowspan="2">受試者之代碼</td><td>甲(乙)非共同試題受試者之資料</td><td>甲(乙)共同試題受試者之資料</td><td>未答完之代碼</td></tr>
<tr><td>未答完之代碼</td><td>乙(甲)共同試題受試者之資料</td><td>乙(甲)非共同試題受試者之資料</td></tr>
</table>

五、國內外常用學術性向測驗的內容

國內外較常用的學術性向測驗，名稱有多種，但其性質盲似。

㈠比西量表（Binet-Simon Scale）

比西量表試題分成下列7大類。

1.語言
2.記憶
3.概念的思考
4.推理
5.數字推理
6.視覺動作
7.社會智慧

㈡魏氏兒童智力量表（WISC-R）

(1)語文

1.常識
2.類同
3.算術
4.詞彙
5.理解
6.記憶廣度

(2)作業

7.圖形補充
8.連環圖系
9.圖形設計
10.物形配置
11.符號替代
12.迷津

㈢陸軍普通分類測驗（Army General Classification Test）

1.語文理解
2.算術推理
3.方塊計算

㈣中學智慧測驗

1.類推測驗
2.刪字測驗
3.算術測驗

㈤國中智力測驗

1.語文類推
2.語文歸納

3.算術計算

4.算術推理

(六)加州心理成熟測驗（California Test of Maturity (level)）

(1)非語文（圖形）

1.空間關係

2.相似

3.類推

(2)算術

1.數字測驗（即數系測驗）

2.算術推理（即數字應用）

3.語文測驗

①語文理解測驗

②延宕回憶測驗

(七)羅桑智力測驗（The Multi-level Edition of the Lorge-Thorndike Intelligence Tests)

(1)文字測驗

1.字彙測驗

2.語句完成測驗

3.算術推理測驗

4.歸類測驗　5.語文類推測驗

(2)非文字測驗

1.圖形分類測驗

2.數系測驗

3.圖形推理測驗

(八)學校能力測驗（包括 School and College Ability Test 及 Army Classicfication Battery）

1.語文類推測驗

2.數學理解測驗

3.空間關係測驗

(九)圖形式智力測驗

1.歸納測驗（具體及抽象兩種圖形）

2.類推測驗（具體及抽象兩種圖形）

3.填充測驗（即圖形的矩陣推理）

(十)奧雷學校能力測驗 (Otis-Lennon School Ability Tests)

1.語文測驗

2.數字測驗

3.圖形測驗

㈣學術性向測驗（Scholastic Aptitrde Test (SAT)）

1.語文（反義字、語句完成、語文類推、閱讀理解等）

2.數學

㈤美國大學入學測驗（American College Testing Program (ACT)）

1.英語用法

2.數學

3.社會科閱讀測驗

4.自然科閱讀測驗

包含傳統的性向和成就測驗

㈥研究所入學測驗（Graduate Record Examination (GRE)）

1.語文（語文推理和閱讀理解）

2.數量（算術和代數推理、圖形推理、資料描述）

歸納上述 13 種測驗， 紙筆式學術性向測驗內容宜包括下列三大類（不包括操作測驗）

1.語文測驗（含語文理解及語文推理）

2.數字測驗（含數字應用及數字系列測驗）

3.圖形測驗（含圖形類推及圖形歸類，每種再分成具體與抽象兩類形）

六、題目偏失（item bias）研究方法

㈠傳統測驗理論題目偏失研究方法

1.兩個群體的平均數有差異，那題目或測驗就是有偏失。此定義顯然不妥，因爲影響兩個群體平均數差異的因素很多，並不是只有題目偏失這個因素。依照這個定義，那體重計與身高計都是有偏失的測量工具，因爲男、女的身高和體重平均數均有明顯差異。

2.兩個群體所得到的題目難度指數（或 P 值）相同，就是沒有題目偏失。 此定義與上述第一種定義有相同的問題。兩個群體在所有

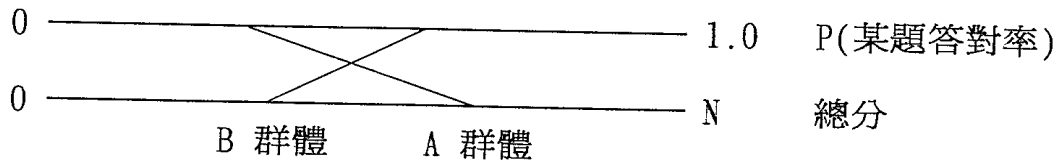
題目的 P 值均相同，那兩個群體的平均得分也必然相同，但是，兩個群體的平均數相同，其各題的 P 值卻不一定相同。

3.題目與受試群體有交互作用，就是有偏失。

圖2.3的(1)及(2)，題目與群體有交互作用，是有偏失的題目；圖2.3的(3)是沒有偏失的題目。

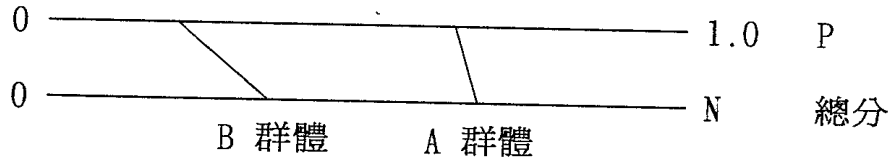
圖 2.3 題目與群體有交互作用及無交互作用的例子

(1)無次序性交互作用



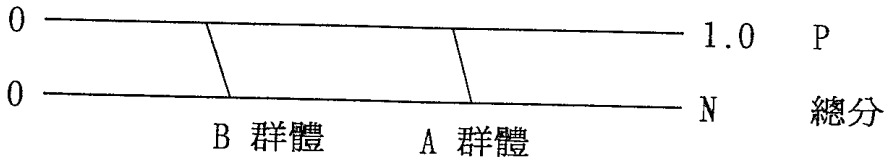
A 群體的總分比B群體高，但是某一題的P反而比B群體低。該題對B群體有利。

(2)有次序性交互作用



A 群體的總分及某題之答對率雖然均高於B群體，但是P的差距比總分的差距大或小，也就是不平行。該題對A或B群體有利。

(3)沒有偏失



A、B兩個群體總分之差距與P之差距相同，兩條線相互平行，此爲沒有偏失的題目。

4.兩個群體P之相關爲 1 ， 那題目就沒有偏失（或是有相同程度之偏失）。題目難度散佈圖上偏離最適合線的題目就是有偏失的題目。此種方法主要的問題是 P 值之相關不是直線的，即使是沒有偏失，只要兩個群體的能力不等，仍然得不到完全相關。

5.△值 (P 值之反轉常態轉換其平均數爲13，標準差爲4。如p＝.84，△＝9；P＝.50，△＝13；p＝.16，△＝17偏離雙變項散佈圖主軸之垂直距離，即爲該題之偏失值。Lord（1980）和 Angoff（1982）發現猜測、題目鑑別力不同以及兩個群體能力之差異，均會使△值偏離主軸。

6.具有相同潛在能力的兩組受試，其答對題目的概率相等，就是沒有偏失。

具有相同總分的不同群體受試，如果答對某題之概率不同，即爲有偏失。 此類方法均採用χ^2法，其中充足χ^2法 （ full chisq-uare method）是將總分分成 J 組（j=1 ，、、、， J）。

圖2.4　　用來計算chi-square的交叉分類表

題目 i		A 群體	B 群體	C 群體
	答對	N_{11j}	N_{12j}	$N_{1.j}$
	答錯	N_{21j}	N_{22j}	$N_{2.j}$
	全體	$N_{.1j}$	N_{2j}	N_{j}

第j 組，第i題的chi-square值爲

$\chi^2_{ij} = N_j (N_{11j} N_{22j} - N_{21j} N_{12j})^2 \quad (N_{1.j} N_{2.j} N_{.1j} N_{.2j})$

所有各能力組第i題的chi-square值爲

$$\chi^2_i = \sum_{j=1}^{J} \chi^2_{ij} ,$$

只有兩個群體時自由度爲J。如果有K個群體 ，則自由度爲 J(K－1)。此種方法主要的問題有二：

(1)分成多少組以及如何分組均相當主觀，會影響所得結果。

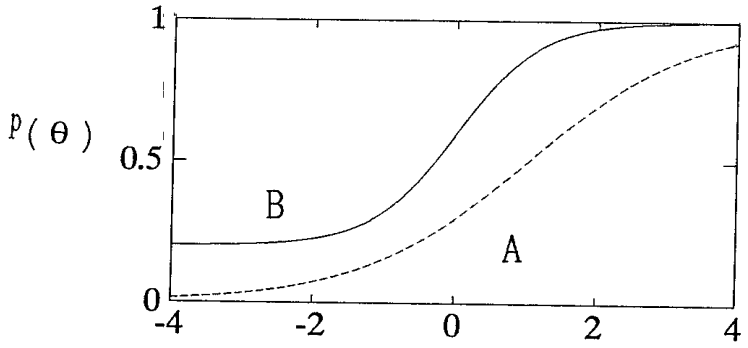
(2)chi-square對樣本數及細格數很敏感，只要人數夠多，χ^2值均會達顯著水準。

㈡題目反應理論 (IRT) 題目偏失研究方法

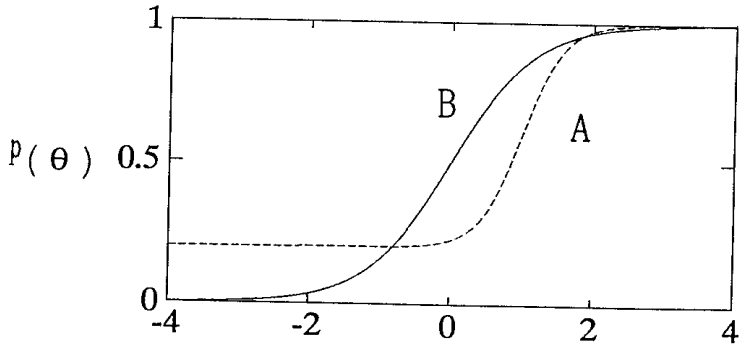
能力相同的不同群體受試，如果其答對某一題目的概率不同，那該題就是有偏失。也就是說，不同群體之題目特徵曲線 (ICC)或是題目參數如果完全相同，那就沒有偏失。

當二個群體某一題目之 ICC 相同時，其題目參數也相同。

圖 2.5 兩個群體題目特徵曲線差異情形



所有能力水準均對A群體不利的題目



低能力水準不利於B群體，高能力水準不利於A群體的題目

以 IRT 爲基礎所進行的題目偏失研究，通常採用下列二種方法：

1.題目特徵曲線比較法

此種方法是在計算兩條題目特徵曲線所涵蓋的面積之差異情形，面積差異越大，代表題目的偏失越大。完全沒有偏失的題目，其 ICC 完全相同，因此，面積之差異也等於 0。計算面積的方法有下列三種：

(1)絕對值法

$$A_{1i}=\sum_{\theta=-5}^{+5}\left|P_{i1}(\theta_k)-P_{i2}(\theta_k)\right|\triangle\theta$$

$$\triangle\theta=.005$$

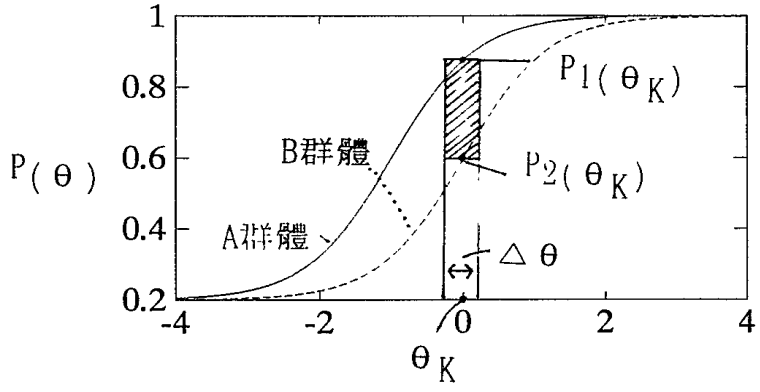
(2)平方根法

$$A_{2i}=\sum_{\theta=-5}^{5}\{[P_{i1}(\theta_k)-P_{i2}(\theta_K)]^2\triangle\theta\}^{1/2}$$

(3)偏向法

$$A_{3i}=\sum_{\theta=-5}^{5}\{[P_{i1}(\theta_k)-P_{i2}(\theta_K)]\triangle\theta\}$$

圖 2.6 兩個群體的ICC差異面積計算法



第(1)及第(2)兩種面積可找出有偏失的題目，第(2)種先平方再開二次方，加重極端差異值的重要性，此兩種均能找出有偏差的題目，二者所得的面積也均永爲正值，無法知道對那個群體較不利（或有利）。第(3)種面積值可提供題目偏失方向之訊息（從正負號判斷），但是較無法有效偵測有偏失之試題，像上圖有一段對 B 群體有利，也有一段對 A 群體有利，正負相抵，所得到的面積值可能很小，但實際上是有明顯偏失。

2.題目參數比較法

此種方法類似多變項變異數分析法（MANOVA），考驗每一個題目的兩組題目參數（a，b，c）有無顯著差異，其公式爲

$$Q=(X_1-X_2)'(V_1+V_2)^{-1}\ (X_1-X_2)$$

X_1 及 X_2 分別爲兩個群體的題目參數之向量，V_1 及 V_2 是兩組題目參數的估計誤差變異數 - 共變數矩陣。此項考驗的假設爲

H_0：$\tau_{1i}=\tau_{2i}$（i＝1，・・・n）

H_1：$\tau_{1i}\neq\tau_{2i}$

$$\tau_{1i}=\begin{bmatrix} a_{i1} \\ b_{i1} \\ c_{i1} \end{bmatrix}\quad ,\quad \tau_{2i}=\begin{bmatrix} a_{i2} \\ b_{i2} \\ c_{i2} \end{bmatrix}$$

題目參數估計誤差的變異數 - 共變數矩陣等於訊息矩陣（I）的反矩陣，

即

$$V_{1i}=[\,I_1(X_i)\,]^{-1}\equiv I_{1i}^{-1}$$

$$V_{2i}=[\,I_2(X_i)\,]^{-1}\equiv I_{2i}^{-1}$$

因此，

$$Q_i=(X_{1i}-X_{2i})'(I_{1i}^{-1}+I_{2i}^{-1})^{-1}(X_{1i}-X_{2i})$$

此項考驗所得Q值近似 χ^2 分配，其自由度等於所要比較的參數數目。如採單參數模式，那就只有一個b參數，因此考驗的公式可簡化爲

$$z_i=(b_{1i}-b_{2i})/(V_{1i}+V_{2i})^{1/2}$$

㈢IRT題目偏失研究之步驟

任何IRT模式題目參數比較法之步驟如下:

1.選擇一個適當模式。

2.分別爲每一個群體估計題目和能力參數。

3.由於兩個群體的參數是分開估計的，因此，必須安置在共同量尺上。參數估計時，如選擇將 b_i 標準化，那參數就在已同一量尺上。否則就要採用前面介紹的題目參數連結法來將題目參數安置於同一量尺上。

表2.1　三參數logistic模式題目參數訊息矩陣

參數	訊　息　矩　陣		
	a_i	b_i	c_i
a_i	$\frac{D^2}{(1-c_i)^2}\sum_{a=1}^{N}(\theta_a-b_i)^2(P_{ia}-c_i)^2Q_{ia}/P_{ia}$		
b_i	$\frac{D^2a_i^2}{(1-c_i)^2}\sum_{a=1}^{N}(\theta_a-b_i)(P_{ia}-c_i)^2Q_{ia}/P_{ia}$	$\frac{D_2a_i^2}{(1-c_i)^2}\sum_{a=1}^{N}(P_{ia}-c_i)^2Q_{ia}/P_{ia}$	
c_i	$\frac{D}{(1-c_i)^2}\sum_{a=1}^{N}(\theta_a-b_i)(P_{ia}-c_i)Q_{ia}/P_{ia}$	$\frac{-D_{ai}}{(1-c_i)^2}\sum_{a=1}^{N}(P_{ia}-c_i)Q_{ia}/P_{ia}$	$\frac{1}{(1-c_i)^2}\sum_{a=1}^{N}Q_{ia}/P_{ia}$

4.一旦題目參數估計值已量尺化後，接下來就要計算兩個群體的訊息矩陣（information matrices），其計算公式見表2.1。

5.使用下式計算每一個題目的 Q_i值

$$Q_i=(X_{1i}-X_{2i})'\ (I_{1i}^{-1}+I_{2i}^{-1})^{-1}\ (X_{1i}-X_{2i})$$

6.依據考驗統計數（Q_i）決定題目是否有偏失。

上述方法雖然直接和簡易，但使用此種方法時需考慮下列二點：

1.此種顯著性考驗只是漸近χ^2分配。到底樣本需多大才能使χ^2考驗精確並不清楚。

2.只有在θ_a已知的情況下，題目參數估計值的漸近分佈才有效。當同時估計θ_a，a_i，b_i，c_i 時，漸近理論可能無效。

上述題目特徵曲線比較法和題目參數比較法，本質上均在比較題目特徵曲線是否一致。雖然學者對此兩種方法之相對優點的看法不太相同，但此兩種方法顯然均合邏輯。Lord（1980a，p.217）推薦題目參數比較法。Linn等人（1981）認爲題目參數比較可能會導致錯誤的結論。他舉出下列例子來說明此種誤導的可能性。表2.2是兩個群體的題目參數：

表2.2　兩個群体某一題目參數

群體	題目參數		
	a	b	c
1	1.8	3.5	.2
2	.5	5.0	.2

雖然這兩個群體鑑別力和難度參數的差異很大，但是這兩個群體在能力介於－3 與＋3 之間的題目特徵曲線之差異則極微（任何 θ 值答對概率之差異均不超過 .05）。由於題目偏失之定義是指 θ 相同的不同群體受試，其答對概率有明顯差異，因此，Linn 等人認爲題目特徵曲線比較法優於題目參數比較法。

㈣適合度比較法

與 IRT 題目偏失有關的研究，還有不同群體題目參數之適合度比較法，也就是要比較以不同群體爲樣本，所得到的適合度統計數是否有差異，其步驟如下（Linn & Harnisch, 1981）：

1.將兩個群體合併來估計題目及能力參數。

2.計算每位受試的答對概率 P_{iag}（g＝1，2）。

3.分別計算每個群體的平均答對概率 $P_{i.g}$。

4.計算每一組的觀察答對率 $p_{i.g}$（即古典題目難度指數）。

5.比較 $P_{i.g}$與 $p_{i.g}$（g＝1，2）。

6.此外，還要計算每個人的標準殘差，其公式爲

$$z_{iag}=(U_{iag}-P_{iag})/[P_{iag}(1-P_{iag})]^{1/2}$$

(U_{iag}是第g組，第a個受試，在第i題的得分。答對，U＝1，答錯，U＝0。)

將每一群體的標準殘差平均，並比較兩個群體的平均標準殘差。

適合度統計數可能無法提供有意義的比較。就單參數模式來說 Shepard 等人（1981）發現適合度統計數的聚斂效度（convergent validity）不佳，也就是說適合度統計數與其他偏失指數的相關很低。事實上適合度不同，不一定是題目有偏失，像沒有考慮到猜測（guessing）和題目鑑別力之不同，均有可能造成適合度之不同。

三參數模式的適合度統計數或許能合理評估題目之偏失 (Linn & Harnisch, 1981）。不過，適合度比較之意義並不完全清楚。在採用此種方法之前，需先對此種方法進行效度考驗研究。

綜括上述有關題目偏失偵測方法之討論，題目反應模式的題目偏失研究似乎要比傳統方法有意義。偵測題目偏失最適切的模式是三參數模式（Ironson, 1982, 1983; shepard et al., 1981）。不過，除非所要比較的幾個群體均相當大，且能力散佈範圍也足夠廣，否則參數的估計會有問題。

Hambleton 和 Swaminathan（1985）推薦使用「面積法」和參數比較法。這兩種方法本質上是一樣的都是在比較特徵曲線。不過，此兩種方法之操作性定義不同，因此，對題目是否有偏失所作的決定也可能不一致。Hambleton 和 Swaminathan 認爲這兩種方法均有需要多累積一些有關偏失之效度證據。

題目偏失偵測方法效度研究的方法有下列方式:

1.比較各種方法所認定的偏失題目與題目內容分析法所認定有偏失的題目是否一致。

2.故意在原測驗加入部分有明顯偏失的題目，然後分析各種方法所得到的偏失指數與效標間之相關。

3.採用電腦模擬研究，方法同 2.。

雖然 Linn 等人的看法似乎言之成理，但是 Hambleton 和 Swaminathan（1985）則認爲也可以從反面來看，也就是說參數比較法要優於特徵曲線比較法。就「眞正偏失的題目」而言，兩個群體的題目參數

雖然有很大的差異，但是兩個群體的答對概率可能沒有實質的差異（在某能力範圍），可見題目參數比較法比較敏感，也比較適切。此兩種方法實質上是相同的，題目參數相同，特徵曲線必然相同，反之亦然。上述例子的 b 太極端，因此在計算特徵曲線之差異面積時，能力範圍不可限制在± 3 之間，如果能力範圍加大，那兩個群體的特徵曲線仍有很大的差異。筆者認爲採用特徵曲線差異面積法時，所含括的能力（ θ ）範圍至少要在兩個群體所有題目難度（ b ）的範圍上下各加 2 單位。例如，題目難度介於± 3 時，能力範圍宜介於± 5 ，這樣才能有效偵測出有偏失的題目。

特徵曲線差異及題目參數比較法有一共同的問題，那就是所要研究的兩個群體如果能力分佈分別集中在高和低能力的那二端，在此種情況下，題目參數 c 的估計會有問題。高能力組由於沒有足夠的低能力受試，c 的估計誤差必然很大。同樣地，如果兩個群體的能力分佈分別集中在能力量尺的兩端，a 和 b 參數的估計也會有問題。此問題並非是題目反應理論所獨有，即使是直線迴歸模式的估計也有同樣的問題。另一個更常見的問題是每一群體所能使用的樣本數可能太小，特別是少數群體更不容易得到足夠的樣本數。

Lord（1977, 1980, p.217）建議以下列方式來估計參數，或可解決參數估計的問題 ：

1.將兩個群體合併來估計題目和能力參數，選擇將bi標準化（這樣可將所有題目參數安置於同一量尺上）。

2.將 c_i 固定爲以第一步驟所得到的 c_i 值。

3.在 c_i 值固定的情況下，分別爲兩個群體估計能力、難度和鑑別力參數，選擇將 b_i 標準化（這樣就不必再進行等化研究，所得到的題目參數本來就在同一量尺上）。

使用上述方法時，由於兩個群體的c_i設定爲相等，因此，比較題目參數時，只比較兩個群體的a_i和b_i參數是否有差異。

shepard 等人（1981）發現此種方法有些問題。他們以合併的樣本（1593人）進行分析， 發現約有 40% 的 c 參數無法收斂。另一個更嚴重的問題，是將得自於合併組的 c_i 估計值固定後， 低能力組的難度和鑑別度參數的估計很不理想，可能的解釋是 c_i 值對低能力組受試來說太大，

以致於影響該組其他參數的估計（shepard et al., 1981）。本研究的題目偏失研究，採用題目特徵曲線比較法的「絕對值法」及「偏向法」。

七、解釋IQ分數之注意事項

有很多人誤用或錯誤解釋IQ分數，其主要原因是錯認一個人的IQ是固定的（不會改變的）數量。由於IQ分數得自於包括幾類心理能力的測驗，這些測驗的信度和效度也並非很完美，因此，不同時間或不同測驗所測得的IQ也就不太可能完全相同。要正確使用IQ分數，就要考慮到IQ分數是會變的，不能忽略會變的事實。

假定能滿足測驗的基本條件（如，發展的機會相同以及均具有最大動機，我們仍可以預期IQ分數會有下列變異來源和程度：

1.單就測驗標準誤，同一測驗所得IQ會有 5 至10分的差異。因此，IQ 105最安全的解釋是視爲100至110之間的一般IQ分數。

2.如果要比較不同測驗所得之IQ分數，我們可預期同一受試在不同測驗所得之IQ會有某種程度的差異，就目前經常使用的團體測驗而言，不同測驗所得的語文IQ，可能差到 8 分。非語文IQ的差異則較小（Hieronymus Stroud, 1969）。由於每一測驗所測量的心理能力總有某些程度的差異，而且每一測驗用來建立常模所使用的標準化樣本也不同，因此，不同測驗所得到的IQ不能直接比較，在解釋IQ分數時，必須知道所使用的測驗是那一個。

3.小學生的IQ分數之變異程度要大於中學生。這主要是因爲能力在形成過程較不穩定，如果使用團體測驗，則變異程度更大，這是因爲有很多會影響測驗結果的因素很難完全控制。

上述這些變異是在理想測驗條件下，我們可以預期IQ分數會有的正常變異。雖然對那些預期IQ分數完全不變的人來說，IQ會有變異是件惱人的事，但對於實際應用IQ分數的教師而言，則不會有太大不變。一個老師只要知道某生的IQ在80至90之間，或95至105之間，或110至120之間即能適當估計該生的學習能力。

解釋和使用IQ分數時最大的問題不是正常情況所造成的IQ變異，因爲這些變異可以相當精確的估計而且也可以事先加以考慮，如果性向測驗

的基本條件無法滿足那會造成一些無法適當估計的誤差，這才是解釋和使用IQ分數的主要問題。一般而言，下列學生的IQ分數較不可靠（Gronlund，1985）：

1.家庭環境沒有提供足夠的機會去學習測驗問題所要求的作業的學生（如，文化不利兒童）。

2.對學校功課沒有足夠動機的學生。

3.閱讀能力較差或有語言缺陷的學生。

4.情緒適應較差的學生。

參、研究方法與步驟

一、預試樣本：

本研究的預試樣本取自高雄市及台南市各國小六年級學生。爲使所得樣本具有相當程度的代表性。首先決定所需取樣班級數，共需九十六班學生數，五、六年級各一半，接著根據各區所包含學校的學校數比例及市區、市郊的校數比率，分別選取各區代表性學校作爲本次研究之樣本。抽樣學校如表 3.1 及表 3.2。全部樣本共有 4334 人，其中台南市有 2347 人，高雄市有 1987 人。