第二章　文獻探討

本章旨在說明研究所需的理論基礎進行文獻探討及回顧，共分為三節敘述，分別為：第一節為測驗分析理論，旨在探討古典測驗理論（CTT）及試題反應理論(IRT)兩者的基本假設及分析方法。第二節為測驗題本分析，探討整份試題分析方法，包括信度及效度。第三節為試題題型分析，採討針對測驗個別試題的分析方法。

第一節 測驗分析理論

自 1905 年法國學者 Binet-Simon 投入智力測驗研究後，漸發展出客觀評量人類內在的心理測驗的學問－「心理計量學」，包括量化心理學、個別差異和心理測驗理論的範圍。測驗理論發展至今大致可分為兩大主流；以 H.Gulliksen(1950)為代表的古典測驗理論（classical test theory，簡稱 CTT）以及 Georg Rasch(1960)為代表的試題反應理論（item response theory，簡稱 IRT)。

一、 古典測驗理論

發展時間較早，至今仍是實用上最多的測驗理論。主要是利用某個測驗去估計實得分數(observed score)的信度，企圖估計出實得分數和真實分數(true score)間的關聯程度，又稱為真實分數理論(true score theory)。真實分數模式是以 X = T + E 為架構，其中 X 表示測驗實得分數，T 表示真實分數，E 表示誤差分數。當受試者在某個特定的情境下，接受一個測驗施測後，會得到該測驗上的實得分數（X)，也就是在這樣試題上所表現的能力(ability)。或許在不同情境或是同範圍而不同試題樣本的情況下，受試者的表現也可能不一樣。若在所有可能的情境下或使用不同試題樣本來針對同一位受試者進行多次（近乎無限）的測驗，可以獲得許多該受試者的實得分數，這些實得分數的平均數，可定義為受試者的真實分數（T)。而真實分數無法在少次測驗中測量到，為受試者本身的潛在特質；另外在實得分數和真實分數之間有不可測得非受試者潛在特質的隨機誤差量，即為誤差分數（E)。

CTT 具有三個基本假設：

（一）實得分數的期望值等於真實分數。

（二）實得分數、真實分數、誤差分數互不相關。

（三）假設有兩個測驗符合上列二個假設，實得分數分別為 X_1 及 X_2，且對每一群考生而言，亦滿足 $T_1 = T_2$ 和 E 的變異數 $\sigma_{E1}^2 = \sigma_{E2}^2$ 時，這兩個測驗稱為「複本測驗」，測驗結果能相互比較及解釋。

CTT 理論根據上述的假設，可以做出下列詮釋：

（一）假設具有潛在特質在。
（二）為多次測量的推論結果。
（三）單獨一次的測量必有誤差存在。
（四）假設潛在特質與誤差之間是獨立的。
（五）複本測驗假設是嚴格的。

從 CTT 理論衍生出試題分析幾個重要指標，如：難度、鑑別度、信度等，由於 CTT 所採用的公式簡單明瞭、淺顯易懂，適用於大多數的教育與心理測驗情境等多項優點，為目前應用及流通最為廣泛；但因為模式過於簡單，產生下列幾項缺點：

（一）CTT 所採用的指標，如：難度、鑑別度、信度等，都是屬於樣本依賴的指標；也就是隨著受試者樣本不同而改變，同一份測驗難獲得一致的指標。
（二）對某一份測驗的所有受試者的潛在能力估計值的測量誤差，CTT 只以一個共同的測量標準誤來計算，忽略受試者反應的個別差異，對於具有高、低兩極端潛在特質的受試者而言，不合理也不精確，致使 CTT 模式的適當性受到質疑。
（三）CTT 只對複本測驗所獲得的受試者分數，才能提供有意義的比較；對於非複本的測驗則無法對受試者分數進行有意義的比較。
（四）CTT 對信度的假設是建立在複本測驗的假設上，但此假設在實際測驗情境是不存在的。因為在實際測驗情境下，不可能要求每個受試者接受同一份測驗無數次後，仍然保持每次反應結果彼此獨立、互相不影響。而且並不是每個測驗在編製測驗時就能同時製作複本。
（五）CTT 忽略受試者作答的試題反應組型所代表的意義，對於原始得分相同的受試者或正確反應總和相同的試題，即當作受試者的潛在能力（如：能力）或試題參數（如：難度）的估計值相同。此觀點是不正論的，因為總分相同的受試者或總和相同的試題，其試題反應組型不見得相同，因此試題反應組型所顯示的意義也不會相同，所估算出的潛在能力和試題參數的估計值，應該也會不一樣。

二、 試題反應理論

在 CTT 假設下的試題分析具有上列缺點，測驗學者就針對以上缺點提出試題反應理論（IRT），IRT 是將受試者的潛在能力和實際得分聯結在一起，每個受試者在接受測驗後，會有不同的潛力表現，通常以數值來表達出不同受試者潛在能力的相對程度，也就是受試者的能力參數。IRT 模式是用機率函數表示受試者能力參數與試題特性參數間的關係，以受試者的答對機率為縱軸，受試者的能力為橫軸所畫出來的曲線即為試題特徵曲線（Item characteristic curve，以下簡稱 ICC）。ICC 必須要符合二個條件才能符合其假設，第一個是能力越高者，答對機率越大，因此 ICC 必須是單調遞增（monotonic increasing）。第二個條件是對

任何能力水平的人而言，簡單的題目的答對機率比困難的題目大。可舉一個例子說明上列條件：若有二個題目施測，對能力水平低的人第一題答對機率較第二題大，但對能力水平高的人而言，第二題的答對機率反而比第一題大時，那麼第一題和第二題何者比較簡單，就看不出來，因此該題目的難度在常理上就無法理解其意義。ICC 能表示出試題參數與能力的關係變化，藉由模式求出受試者在試題上的表現與對能力的估計量關係。不同的 ICC 代表不同試題參數與能力的變化關係，每一個關係就有一條 ICC，也就是說每一種試題反應模式都是用來描述受試者能力與答對機率的關係。

常用 ICC 函數的參數有難易度、鑑別度、猜測度三種試題參數，也因此發展出單參數、雙參數及三參數此三類 IRT 模式。

（一）單參數 IRT 模式

Georg Rasch(1960)提出，此模式的試題特性只有一個試題參數，即難易度；其假設受試者（N）在面臨試題（I）的作答反應，是由受試者的潛在特質（Θ'_n）及作答時該題的試題特性（b'_i）兩者的影響；當潛在能力值Θ'_n相對於試題特性 b'_i比值高時，其勝率就越大。

由上列假設可以做出以下數學推導，

令答對機率為 Po，而答錯機率為 Pf，

其勝率

$$O_n = Po / Pf = \Theta'_n / b'_i$$

若有兩位受試者Θ'_1及Θ'_2作答同一題（b'_1）時，

則

$O_1 = \Theta'_1 / b'_1$

$O_2 = \Theta'_2 / b'_1$

兩者能力的比值

$O_1 / O_2 = \Theta'_1 / \Theta'_2$

令

$\Theta'_n = \exp(\Theta_n)$

$b'_i = \exp(bi)$

$O_1 / O_2 = \exp(\Theta_1 - \Theta_2)$

兩者的能力比值和題目特性無關，是客觀測驗，為比率量尺。

若取自然對數 log，可得

$\text{logit}_i = \log(O_n) = \log(\Theta'_n / b'_i) = \log(\Theta'_n) - \log(b'_i)$

$\text{logit}_i = \Theta_n - bi$

若二位受試者作答同一題時，則

$\text{logit}_1 = \Theta_1 - b_i$　　$\text{logit}_2 = \Theta_2 - b_i$

二位受試者差距則為

$\text{logit}_1 - \text{logit}_2 = \Theta_1 - \Theta_2$

若以 logit 單位時，兩者的能力差異和題目特性無關，是客觀測驗，為等距量尺。

由於 $\text{logit}_i = \log(0_n) = \log(Po / Pf)$ 且 $Po + Pf = 1$

因此可以用下列公式來表示說明，上述參數的關係：

$$P_i(\theta) = \frac{e^{D\tilde{a}(\theta - b_i)}}{1 + e^{D\tilde{a}(\theta - b_i)}}$$ （公式 2-1）

$P_i(\theta)$：答對第i題的機率函數，其值介於0與1之間。

b_i：表示題目難度(difficulty)參數。

e：自然對數約為2.71828；Dã＝1.7。

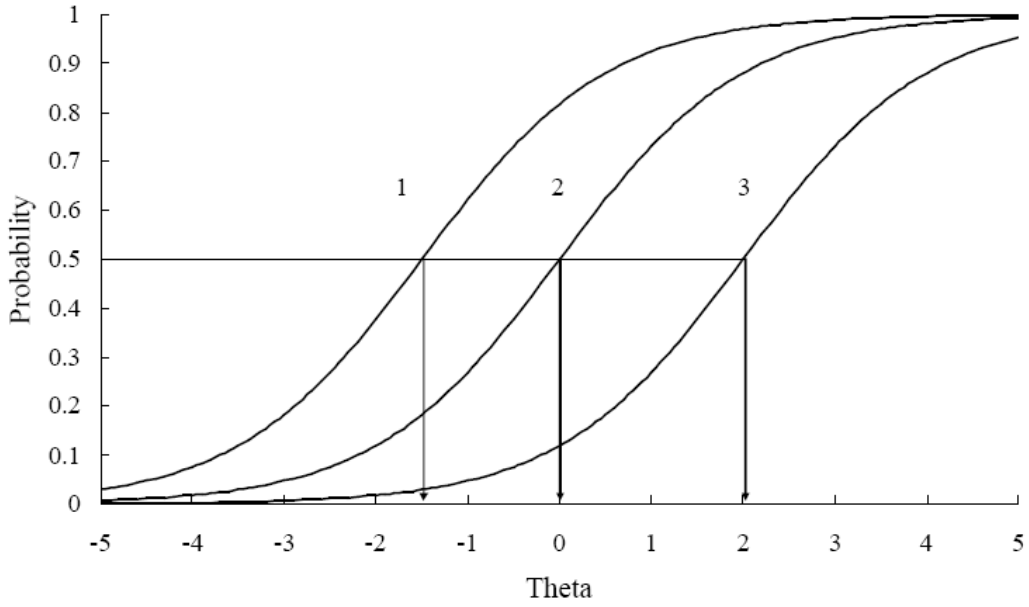

圖 2-1 單參數 IRT 模式的試題特徵曲線圖

公式中 ã 是所有題目的共同鑑別力，b_i是題目的難度，可因題目而不同。

由於 ā 是所有題目均相同的共同鑑別力，是一個常數，故所有題目特徵曲線的斜率均相同；但由於題目的難度不同，故題目特徵曲線的位置不同，如圖　中三條曲線的分布。此式為單參數IRT模式的ICC曲線函數，又稱為Rasch模式。具有客觀測驗及等距量尺的優點。

（二）多參數 IRT 模式

Lord(1952)提出後經 Birnbaum(1968)改用 logistic 分配所得到的模式，針對題目增加鑑別度（a_i）的差異來修正 Rasch 模式，即為二參數 IRT 模式；其公式如下：

$$P_i(\theta)=\frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

（公式 2-2）

除了難易度參數和上述單參數模式相同外，鑑別度 a_i 是指試題對不同受試者是否能反應出其答題的差異，也就是鑑別度大的試題，能力高者答對率高，而能力低者答對率低。由圖 2-2 的曲線 2 及曲線 3 比較，ICC 愈陡的試題，鑑別度愈好；而 ICC 愈緩的試題，鑑別度愈差。

Birnbaum 再增加受試者猜測度（c_i），可得三參數 IRT 模式；其公式如下：

$$P_i(\theta)=c_i+(1-c_i)\left[\frac{1}{1+e^{-Da_i(\theta-b_i)}}\right]$$

（公式 2-3）

除了難易度參數與鑑別度和上述單參數模式相同外，猜測度 c_i 是將能力極低的受試者考慮在模式裡，計算此類的受試者答對試題的機率。c_i 值愈小，表示猜測的因素愈小，試題愈有效。由圖 2-2 的曲線 1 及曲線 2 比較，曲線 1 在能力較低時的左下方出現 P=0.2 的漸近線。

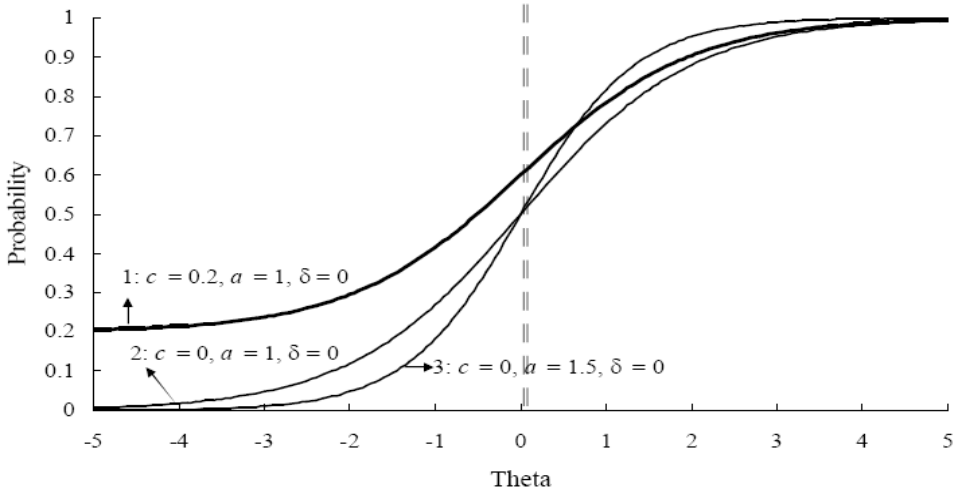

圖 2-2 多參數 IRT 模式的試題特徵曲線圖

（三）IRT 的假設

由於 IRT 模式的嚴謹性，其所適用的測驗資料必需符合 IRT 的假設下才能被用來分析受試者的作答資料，其假設如下：

1. 單向性

指測驗只測量一個特質或能力，其代表的是測量同一特質或能力的試題組合，Hambleton & Swaminathan（1985）認為只要測驗資料有一個主控的因素或成分，就可算符合單向性的假設，而這個主控因素便是特質及能力。受試者答題的好壞，基本上由一種能力來支配，至於其他非能力因素的發生皆為隨機狀態，這現象在理論上可以歸納為隨機誤差。

Crocker & Algina (1986)認為單向性就是試題間統計依賴，也就是對整體受試而言，試題間為相互關聯，故應只有一條迴歸線，以表示只有一個特質存在。單向度的檢定方法通常為因素分析法，透過因素分析中的主成份分析，可尋找到一份測驗中影響受試者得分的因素有哪些，當主成份因素特徵明顯大於第二個因素特徵值（Reackase，1979），測驗大致符合單向性的假設。

2. 局部獨立性

對某特定能力的受試者而言，試題間無相關性的存在，這意味著在 IRT 模式的能力因素，才是唯一影響受試者在測驗試題上反應的因素。受試者對某一試題作答的結果，並不受到其他試題的影響，因此個別試題在編寫時必須沒有連帶關係。局部獨立性的假設實際上是指當能力的影響力除去時，試題和試題的相關就不存在。局部獨立在下列情況下無法成立，一是影響測驗的表現能力向度不只一種時，二是連鎖性試題以及試題本身提供答案線索時，在這些情況下 IRT 模式無法適用於該筆測驗資料。局部獨立性假設的檢定通常與單向性假設的檢定相同，都是利用因素分析法來檢定，因為當單向性假設成立時，局部獨立性假設也會成立。局部獨立性假設是單向性假設成立下的必然結果（Lord，1980）。

3. 非速度性

測驗的實施不是在速度限制下完成，受試者在答題時沒有時間限制，其成績不理想是因為能力不足而不會答，並非時間不夠所致。如果受試時間不足，受試者答題的好壞，普遍受到答題速度的影響，測驗的成績會受到能力以外的因素所影響，單向性的假設就會受到質疑，非速度性的假設常隱含在單向性假設中。非速度性的檢定，除了單向度檢定外，可檢查有多少受試者沒有答完整份測驗，以及有多少試題沒有被受試者作答。

4. 知道－正確假設

受試者只要知道某試題的正確選項，便會答對該試題；也就是說，若受試者答錯某試題，就表示受試者不知道該試題的正確選項。因此有些受試者把正確選項填錯，就不在本假設的考慮範圍，因為人為疏忽是任何測驗理論所無法顧及的。此外，省略不答和未答完的試題有所不同，前者是能力所影響，也就是本假設所提的；後者則是施測時間所影響，為違反前述非速度性所造成的，此二項亦隱含在單向性假設中。

5. 等鑑別度

為單參數 IRT 模式的假設，研究者的重點在難易度參數，各個試題鑑別度需大致相等，或選擇鑑別度大致在一個合理的範圍內的試題。等鑑別度的假設檢定時，若鑑別度參數落於 0.65-1.34 之間(Baker，2001)或在模式符合度分析上其指標能符合模式，便算符合等鑑別度，等鑑別度的假設雖不實際，但因 Rasch 模式的強韌性，能允許某程度範圍的違反假設。

6. 最小猜測度

此為單參數及雙參數模式的假設，即模式不能受測驗試題猜測因素的影響。可以使用試題誘答分析來讓猜測因素減到最小，或刪除能力特低的受試者來防止猜測行為。最小猜測度的檢定，可以用 Wright(1982)所提供的公式，刪除能力特低的受試者，以四選項的選擇題為例，假設總題數為 K，假設 S 為刪除的臨界分數（critical scores)，則受試者得分低於 S = 0.25 K + 2 ((0.25)(0.8)K)^0.5 ，即可視為低能力而予以刪除。

7. 模式符合度檢定

除了上述六項 IRT 的基本假設外，測驗資料能否符合模式，還要看受試者在試題上實際的表現，和由模式預測出來的期望表現是否一致，適合度的計算方式是根據殘差而來，殘差是指觀察值和模式期望值之間的差距，可用試題的殘差分析來檢定測驗資料與每個試題符合性（goodness-of-fit)。

（四）IRT 的優勢

IRT 模式是建立在數理統計學機率模式的基礎上，為一立論與假設均合理與嚴謹的學說，深受測驗學者的青睞，其地位逐漸凌駕在 CTT 之上，連目前國中基本學力測驗也採用單參數 IRT 模式來計算，以下是 IRT 所具有的優點：

1. IRT 的試題參數估計值，並不會受樣本依賴的影響，即試題參數不會因受試者樣本不同而有所差異。

2. IRT 以試題訊息函數針對每位受試者提供個別差異的測量誤差，不像 CTT 只提供整份測驗相同的測量標準誤，能精確估計受試者的能力。
3. IRT 可利用試題的設計，進行測驗分數或受試者能力的等化，同時對不同受試者間的分數或能力進行有意義的比較。
4. IRT 以試題訊息量(item information)及測驗訊息量(test information)的概念來估計某試題或整份試題的測量誤差，並以測量誤差來取代 CTT 的信度指標。
5. IRT 所採用的適合度統計量 (FIT)，可以檢測模式與資料間的適合度。

上述 IRT 的各項特點中，多參數 IRT 模式看似考慮周詳，能符合大多數的測驗；單參數 IRT 模式似乎為多參數 IRT 模式中的一個特解，但事實上，多參數 IRT 模式有二個爭議之處，第一點是其 ICC 無法符合「對任何能力水平的人而言，簡單的題目答對機率比在難的題目大。」的條件。第二點是在多參數模式下，其量尺是否為具有測量意義的等距量尺？

若有兩位受試者能力為Θ'_1及Θ'_2，在二參數模式下，在作答第 i 題的勝率分別為

$$0_1 = \exp\ a_i(\Theta_1 - b'_1) \qquad 0_2 = \exp\ a_i(\Theta_2 - b'_1)$$

若取自然對數 log，可得兩者的受試者作答同一題時

$$\text{logit}_1 = \log(0_n) = a_i(\Theta_1 - b'_1)$$
$$\text{logit}_2 = \log(0_n) = a_i(\Theta_2 - b'_1)$$

二位受試者差距則為

$$\text{logit}_1 - \text{logit}_2 = a_i(\Theta_1 - \Theta_2)$$

發現兩者的能力差異取決為題目特性，可證明其量尺無法成為客觀的等距量尺，三參數亦同。

在單參數 Rasch 模式所得到的量尺Θ具有測量的意義，並具有等距量尺的特性，所以可以直接進行統計分析。而多參數模式的量尺並沒有測量品質，缺乏等距特性，無法滿足參數統計的要求。因此本研究採取以 Rasch 模式來進行 IRT 的分析。

（五）能力量尺分數

就測量特性而言，原始分數並不能達到有意義測量的基本要求：單向度、線性、客觀性。客觀測量必須滿足單向度、線性、 客觀性。單向度是指測量必須只描述測量物的單一屬性。線性指的是測量單位必須均等，也就是必須達到至少

等距量尺以上的要求。客觀性是測驗必須以達到樣本獨立（sample-free）及試題獨立（test-free），也就是說試題的校準(item calibrations)不會因為受試者不同而有不同的結果；受試者能力的估計也不會因為試題的不同就產生不同的結果。

Θ量尺最主要的轉換用途是將它轉換成真實分數量尺(true-score scale)；因為真實分數量尺的範圍是由 0 到n，n為測驗的題數，而Θ量尺的範圍卻是介於正負無窮大之間（亦即$-\infty < \theta < \infty$），若將Θ量尺轉換成真實分數量尺，不僅有助於我們陳報考生的能力高低，更有助於我們解釋測驗分數和作為對換測驗(test equating)之用。

真實分數 T 是答對題數和分數之期望值，也是能力為Θ的考生在一堆試題上的試題特徵曲線(item characteristic curves)之和。由此看來，真實分數其實就是考生在某一測驗上的測驗特徵曲線(test characteristic curves)，當然，這種說法也僅有在試題反應模式適用於該資料的條件下才成立。

將Θ轉換成真實分數或內容範圍分數（domain score）有許多好處：第一，負的分數可以被消除，便利於大眾的理解能力；第二，新量尺的範圍介於 0 與n之間（或 0%到 100%之間），分數本身即具有解釋涵義在裡頭；第三，內容範圍分數比Θ量尺更好決定區別精熟與否的切割分數(cut-off score)，便利於精熟測驗(mastery testing)的實施；第四，將真實分數對照其相對應的Θ值，畫成一個雙向度的分布圖，有助於判定切割分數的位置（Hambleton & deGruitjer, 1983）。真實分數可以被看成是Θ的一種非直線轉換，因為Θ與 T 間具有一種依序遞增的函數關係。

第二節 測驗題本分析

測驗分析是分析整份測驗的種種特性，在此針對兩種最主要的測驗特性來作文獻探討，信度及效度。其中在信度方面，將從 CTT 的信度分析及 IRT 信度指標來討論。

一、CTT 信度：

一般而言，信度也可以說是可靠度，從字面上的意義來看，就是一份測驗的結果是否「可靠」，而信度可以由兩個方向來加以說明。

（一）信度並非「全有或全無」的概念，而只是程度的問題：

分析一份測驗的信度一定會有一個數值，然而沒有一份測驗是「完全」可靠的，也就是說沒有完全的信度（黃元齡，民 86），所以我們不能簡單地說一份測驗沒有信度或有信度。

（二）信度也就是一致性：

理論上指出「同一份測驗施測於同一組受試者，得出來的結果會一模一

樣」，然而實際測驗的狀況卻非如此。這也就是我們要求一份測驗信度時，則必須要看「這一份測驗的一致性高不高」的主要原因。

(三) 有關於信度的求法：

一般較常見的有以下三種，我們將逐一介紹其理論基礎與實際的算法（余民寧，民 86)。

1. 重測信度：

同一份測驗針對相同的受試者，在不同的時間前後重複施測兩次，如果第一次施測的結果不會影響到第二次施測的結果，則根據這兩次施測的測驗分數求相關係數，此相關係數就是我們所所謂的重測信度係數（test-retest reliability coefficient)。

2. 複本信度：

以兩份類似的測驗，亦即兩份測驗在試題格式、題數難易度、指導語說明、施測時間與例題舉例..等方面均相當接近或相似，並且都是用來測驗同一潛在特質或課程內容，但是試題的內容卻不盡相同的測驗，施測於同一組受試者，再根據兩次施測的測驗分數求相關係數，此相關係數即為複本信度係數（parallel-forms reliability coefficient)。

3. 內部一致性信度：

上述兩種信度的估計方法，均需對相同受試者進行兩次施測，才能求得兩次測驗分數的相關係數，用以作為信度係數。這樣的做法在實務上是有困難的，因為會增加測驗編製的負擔，而在受試者方面，更可能因為在短時間內重複作兩次相同的測驗或在同一時間做兩份複本測驗，而造成受試者合作意願低落、動機減低和疲勞增加..等現象，而直接或間接影響到施測的結果。因此測驗統計學者就積極嘗試尋找一種做法，希望只要根據一份測驗結果，就可以估算出此份測驗的信度。用這種方式所估計出來的信度係數，就是內部一致性信度係數（internal consistency reliability coefficient）。這種信度係數也是目前最常用的信度係數，雖然以這種方式的估計方法有很多種，但我們在此會提到三種方法，其中兩種我們只是提及他們的做法而已，而一筆帶過，但本論文所採用第三種做法的信度係數，詳細做法與理論基礎討論如下。

(1) 折半方法：

本方法是將一份測驗，以隨機的方式分成兩半，分別將這兩半的總分算出，最後求出這兩個總分間的相關係數，這種相關稱為折半相關（split -half correlation)，又稱為折半信度（split-half reliability)，通常折半信度愈高表示兩半測驗的內容愈一致或愈相等，亦即測驗內容的適當性愈高。將測驗分成兩半主要的目的是在得到兩個儘可能接近平行（parallel）的兩半測驗。Crocker & Algina （1986）認為

要將一個測驗分成兩半測驗，其方式有很多種，其中較常用的有下列四種（引自吳裕益，民 93）：

i. 依奇偶數題號來分成兩個測驗，奇數題組成一式測驗，偶數題組成另一式測驗。

ii. 先依據受試者答題狀況計算每個題目的難度（P 值），依其大小順序排列，然後再按奇偶數題分成兩半測驗。

iii. 隨機將所有題目分成兩半測驗。

iv. 依題目內容配對，組成兩半測驗。

每位受試在兩半測驗之得分分別加以計分，然後計算此兩半測驗分數之各項統計數，並代入下列適當的公式，就可以得到全測驗信度之估計值。其中以Rulon法來計算測驗的折半信度最為簡捷，其公式：

$$\rho = 1 - (\sigma_E^2 / \sigma_X^2) \quad \text{.....（公式 2-4）}$$

式中 σ_E^2 是全體受試者差異分數的變異數，σ_X^2 為全體受試者答對題數的變異數

（2）K-R 方法：

此方法是由測驗統計學者 Kuder 和 Richardson 於 1937 年所創，一般簡稱為 K-R 方法，此法適用於二元化計分（dichotomously scoring）的測驗資料，主要依據受試者對所有試題的反應，分析試題間的一致性，以確定測驗中的試題，是否都測量到相同特質的一種信度估計方法。一般而言，有兩個常見的 K-R 方法所導出來的信度公式，分別為 KR_{20} 與 KR_{21} ，其公式如下：

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{S_x^2}\right) \quad \text{.....（公式 2-5）}$$

$$KR_{21} = \frac{n}{n-1}\left[\frac{1 - \overline{X}(n - \overline{X})}{nS_x^2}\right] \quad \text{.....（公式 2-6）}$$

其中 n 為整份測驗的題數；p_i 為答對第 i 個試題的機率， q_i 為答錯第 i 個試題的機率，且 $q_i = 1 - p_i$，所以 $\sum_{i=1}^{n} p_i q_i$ 為每一試題得分的變異數總和， S_x^2 為測驗總分的變異數；$\overline{X}$ 為測驗總分的平均數。由 KR_{21} 公式所求出來的信度，

通常傾向產生較小（比較不精確）的係數（Cronbach,1990），但因為它較容易計算，所以也常被使用，如果整份測驗所有的試題難易度指數都一樣，或平均難易度指數接近 0.50 時，根據 KR_{20} 公式或 KR_{21} 公式，所估算出來的信度值都將相等。K-R 方法的測量誤差，主要來自於測驗內容抽樣的誤差，尤其是受到抽樣內容同質性或異質性程度的影響最大。

3. α係數方法：

針對 K-R 方法只能處理二元化計分測驗的缺失，測驗統計學者 Cronbach 於 1951 年提出α係數方法，此方法可以處理多元計分的測驗，例如：在情意方面的測驗常使用到「李克氏五點評定量表」(Likert’s five-point rating scale)。α係數公式是由 KR_{20} 公式所發展出來的它的計算方式如下：

$$\alpha = \frac{n}{n-1}\left(1-\frac{\sum_{i=1}^{n} S_i^2}{S_x^2}\right) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 2-7)}$$

其中 Si 為第 i 個試題得分的變異數。從α係數公式可以清楚發現，它只是將 $KR2_{20}$ 公式中的 $\sum_{i=1}^{n} p_i q_i$ 之值改成 $\sum_{i=1}^{n} S_i^2$，也就是說 KR_{20} 公式只是α係數公式的一個特例，亦即α係數是信度估計值的一個通式。α係數所估算出來的信度，在測驗試題呈現同質性時，會接近於用其他方法所估算出來的信度；在測驗試題呈現異質性時，所估算出來的信度就會低於用其他方法所估算出來的信度，因此 Novick & Lewis（1968）認為我們可以說α係數為信度估計的下限（lower bound）。α係數的測量誤差和 K-R 方法的測量誤差一樣，主要來自於測驗內容抽樣的誤差，尤其是受到抽樣內容同質性或異質性誤差的影響較大。由於α係數是所有信度估計值的下限，所以α係數的值很高時，表示真正的信度值比它還高，由此我們可以斷定所分析的測驗，是一份值得信賴的測驗；如果α係數偏低時，則無法提供有關測驗較準確的訊息，也就是說，我們不能判斷該測驗是否真的值得信賴，這是我們在解釋α係數信度時應特別注意的。正常狀況下，α 係數受到題數多少的影響，試題間相關係數平均數愈低，則其影響愈大，題數愈大，相對的 α 係數也會提高。在一般社會科學領域研究中，α 係數受到測驗或量表中的題項、試題間的相關係數之平均數與向度數目等三個因素所影響：

(1)當題數多時，即使試題間相關低和量表具有多向度的特性，該量表仍有個高的α係數。

(2)試題間相關係數之平均數增加，α係數亦會增加。

(3)當量表之向度愈多，則α係數會變小。（傳粹馨，民91）

內部一致性係數要多大，才表示測驗的分數是可靠的，根據Henson(2001)的觀點，認為這與研究目的與測驗分數的運用有關，如研究者目的在於編製預測問卷或測驗(predictor tests)或測量某概念之先導性研究，信度係數在.50至.60已足夠。當以基礎研究為目的時，信度係數最好在.80以上。當測驗分數是用來作為截斷分數(cutoff score)之用而扮演重要的角色，如篩選、分組、接受特殊教育等，則信度係數最好在.90以上，而.95是最適宜的標準。如果以發展測量工具為目的時，信度係數應在.70以上。

二、IRT信度指標

題目反應理論用來評鑑測驗之優劣的指標不是信度係數，是測驗訊息(test information)。CTT的信度是一個常數，從信度得到的測量標準誤也是一個常數，無論受試能力高低均用同一個信度或標準誤來評鑑和解釋測驗分數。IRT的測驗訊息以及能力估計值的標準誤並非常數，而是能力的函數，稱為訊息函數（information function）。例如，很簡單的題目或測驗，對於能力較高的受試自然不易精確估計其能力， 因此其訊息量必然較小，反之，其估計誤差也就較大。訊息越大的題目或測驗對受試能力的估計越精確，估計誤差越小。

(一) 題目訊息函數(item information function)

題目訊息為用來描述試題或測驗、挑選測驗試題、以及比較測驗的相對效能的實用方法，該方法需使用「題目訊息函數」，作為建立、分析、與診斷測驗的主要參考依據。

題目訊息函數的定義如公式：

$$I_i(\theta)=\frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \quad \text{.....(公式 2-8)}$$

$I_i(\theta)$是第i題能力為θ者的訊息。$P_i'(\theta)$是θ點ICC的第一階導數，即曲線的斜率，$P_i(\theta)$、$Q_i(\theta)$分別是θ點的答對率及答錯率，二者相乘就是θ點的變異數。此公式說明某一能力點的題目訊息就是該點ICC之斜率的平方除以該點的變異數。

表2-1是三種模式的P_i'、$I_i(\theta)$及最大訊息量(θ_{max})所在之計算公式。單、雙兩種logistic模式的θ_{max}均在難度值所在，如b=0的題目在θ=0處的訊息量最大，三參數模式受c之影響，最大訊息量所在略有偏離，比難度值所在處稍高。

表2-1　三種logistic模式的題目訊息函數之計算公式

模　式	P_i'（斜率）	$I_i(\theta)=P_i'^2/P_iQ_i$	θ_{max}
單參數	DP_iQ_i	$D^2P_iQ_i$	b_i
雙參數	$Da_iP_iQ_i$	$D^2a_i^2P_iQ_i$	b_i
三參數	$Da_iQ_i(P_i-c_i)/(1-c_i)$	$D^2a_i^2(P_i-c_i)^2/(1-c_i)^2$	$b_i+\frac{1}{Da_i}\{1n\frac{1+(1+8c_i)^{1/2}}{2}\}$

（引自　黃國清，民93）

對不同能力水準的受試者而言，相同的試題提供了不同的訊息量，試題訊息量可使用於適性施測過程中，依據受試者目前的能力估計，選擇最適合目前估計能力水準的題目施測。

（二） 測驗訊息函數(test information function)

根據Birnbaum(1968)的推導，一份測驗訊息函數(test information function)是指它在某一個能力值（θ）上所提供的訊息量，該訊息量剛好是在能力值（θ）上題目訊息函數值之總和，亦即測驗訊息量等於所有組成該測驗的題目個別訊息量之總和，也就是說

$$I(\theta)=I_1(\theta)+I_2(\theta)+.....+I_n(\theta)$$

$$=\sum_{i=1}^{n}\frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \qquad \text{（公式 2-9）}$$

其中n 為測驗之總題數。

（三） 測驗訊息之特徵

依Hambleton和Swaminathan(1985)之看法，測驗訊息具有下列特徵：

1. 測驗訊息之定義是針對能力量尺之某一點而下的，換言之，同一測驗之訊息在能力量尺不同點之訊息可能是不一樣的。
2. 測驗訊息之高低是受試題的質(鑑別力和猜對率)和題數之影響。 TCC曲線的斜率愈大，訊息愈高。測驗在某一能力點的變異數(variance)愈小，訊息也愈大。
3. 測驗訊息函數I(θ)並不依靠某些試題的特別組合。每一試題對測驗訊

息的貢獻是獨立的，亦即每個試題所做的貢獻量大小並不受在該測驗中其他試題的影響。換言之，試題之間不會有交互作用，測驗的訊息只是各試題訊息之總和。

4. 測驗訊息與該能力估計值的標準誤(standard error on θ, SE(θ))有下列關係：

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 2-10)}$$

其中，$SE(\hat{\theta})$只要在能力參數的最大近似估計值求出後，便可計算得出。有了能力參數的最大近似估計值，並且也求出在θ值上的測驗訊息之後，我們便可以估計信賴區間的方式來解釋能力估計值的涵義。一般而言，最大的測驗訊息量所對應的能力估計值θ，便是該份測驗所精確測量到的能力參數，也可以說是該份測驗適用於該能力估計值範圍內的測量。

當$I(\theta)$值達到最大時，$SE(\hat{\theta})$值便達到最小，也就是說該θ值的最大近似估計值的估計誤差達到最小，亦即此時的θ的最大近似估計值最精確。

在IRT架構裡，$SE(\hat{\theta})$所扮演的角色和CTT中的測量標準誤的角色相同，然而有一點需要注意者，$SE(\hat{\theta})$的值隨著能力水準的不同而不同，但CTT測量標準誤的意義是認為每位考生能力估計值的誤差都是一致的，而IRT的估計標準誤則認為每位具有不同能力水準的考生，皆應有不相同的估計誤差(或估計的精確性)。

其實，θ的最大近似估計值$\hat{\theta}$的標準誤，$SE(\hat{\theta})$，是這個特定θ值的最大近似估計值所構成的漸近性常態分配的標準差。當測驗的長度夠長時，該分配是呈常態的；即使是測驗長度僅有10至20個試題，這種以常態分配的估計方法，也可以滿足多數測驗目的的要求(Samejima, 1977)。可知當測驗對某能力水準的受試者提供的訊息愈多時，則測驗對該水準受試者的能力估計愈準確；反之則愈不準確。測驗訊息依受試者能力水準的不同而異，可取代傳統測驗之信度指標及測量標準誤的概念。

一般而言，估計標準誤的大小受三個因素的影響：(1)測驗試題的數目（例如：較長的測驗會有較小的標準誤)；(2)測驗試題的品質（例如：鑑別度較高的試題往往讓能力低的考生沒有僥倖猜對的機會，所以它的標準誤便較小)；(3)試題難度與考生能力之間的配合程度(例如：組成測驗的試題難度參數若與考生的能力參數相近，則其標準誤會比具有相當困難或相當簡單試題測驗的標準誤還小)。標準誤的大小很快地趨近穩定，因此，當訊息量增加到超過25時，訊息函

數對能力估計值的估計誤差，僅會發極小的影響(Burket，1989)。

(四) 分離統計參數分析(Separation)

IRT為了與CTT中的信度概念相對照，乃使用根均方誤來求得一個類似CTT固定標準誤的值，並據以求得與CTT相互對照的信度指標，包含分離指標與分離信度指標。而根均方誤即是個別標準誤的平方和再取平均後的開根號值，以此固定的標準誤來代表原先個別不同的標準誤，並據以計算在IRT中的信度指標。

分離統計參數是將上述根均方誤所轉換成統計參數(Wright，1982)，其包括兩組：受試者分離參數(Person Separation)及試題分離參數(Item Separation)：

1. 受試者分離參數，用來區分受試者可被區分能力級段(strata)差異的統計參數，以Gp來表示

   Gp = SAp / Sep ……(公式 2-11)

   其中SAp是受試者得分的標準偏差，Sep是受試者得分的根均方誤(Root Mean Square Error，簡稱 RMSE)。

2. 試題分離參數，用來區分試題可被區分難度級段差異的統計參數，以Gi來表示

   Gi = SAi / Sei ……(公式 2-12)

   其中SAi是試題的標準偏差，Sei是試題的根均方誤。

一般而言，Gp越大，表示受試者間的區分越明顯，評分量標也越有效，測驗信度越佳。Gi越大，表示試題評分量標分類級段越佳，試題難度差異明顯，能清楚分辨受試者的差異。

(五) 信度指標

分離參數可以計算得到測驗信度指標(Prieto，1997)，包括受試者信度指標及試題信度指標兩部分：

1. 受試者測驗信度：其指標和α係數方法類似可以用下式來計算

   $Rp = (Gp)^2/(1+Gp^2)$ ……(公式 2-13)

2. 試題測驗信度：可以用下式來計算

   $Ri = (Gi)^2/(1+Gi^2)$……(公式 2-14)

三、效度

效度就是「正確性」，也就是「測驗所得的測驗分數」代表「施測者原本想測試的潛在特質」的程度，或者是測驗能夠達到其編製目的的程度，例如：命題者絕不會出一篇作文來測量受試者的數理能力。因為就算是受試者的作文得滿分，我們也不能確定他的數理能力是否和其作文能力是一樣好？所以就測驗的目

的而言，效度（正確性）是比信度（可靠性）來得重要。

根據1985 年美國教育研究學會（American Education Research association）、美國心理學會（American Psychological Association）和國立教育測驗委員會（National Council on Measurement in Education）等三個專業團體所組成的聯席委員會，所出版的一本有關編製與使用方面的規範標準－「教育與心理測驗標準」一書中提到，「針對測驗使用目的不同，規定再推論和解釋測驗分數時，應該報告三種不同的測驗效度」。所以效度分析主要有三個類型（余民寧，民92）：

（一）內容效度：

內容效度（content-related）是指「測驗試題的抽樣樣本」內容是否具有教學目標、教材代表性或適當性程度的一種指標，例如：要測驗學生是否具有「瞭解氣象要素的概念」，那所使用的測驗試題必須都為與氣象要素的概念相關的試題，而不能有其他方面的試題，如此這份測驗才會有較高的內容效度，一般而言，測驗的試題若能涵蓋所有教學目標和教材內容，並且根據雙向細目表來命題，且具有足夠的代表性試題，即能確立該測驗的內容效度的適當性（陳英豪、吳裕益，民79）。也就是說，教學目標與教學內容是內容效度評鑑的兩項重要因素。對成就測驗而言，內容效度是最重要的（王文科，民91）。

（二）關聯效度：

能夠預測受試者未來的表現或評估受試者在某些效標表現上的未知狀況，稱為關聯效度（criterion-related）。例如：我們也可以從學生大學聯考的分數來預測學生在大學時的表現。而在某些科目加重計分，使得在某些科目表現較好的同學，能進入大學相關科系就讀，也就是考量到關聯效度的因素。本研究雖然是可用在升學方面的成績考量，但還是以能測出學生在地球科學的能力為主，因此不做關聯效度的分析。

(三) 建構效度

推論受試者是否具備某種理論上特質的程度，稱為建構效度（con-struct-related)(周文欽等人，民84)。例如：給受試者做一份邏輯推理測驗，即可用來推論受試者的數學性向。通常，建構效度的建立過程都需經過下列步驟（Gronlund, 1993； Hopkins, Stanley & Hopkins,1990）：

1. 先對理論建構進行分析，以發展出一套完整的評量工具；也就是說先提出有關理論建構的說明，再根據此說明設計評量用的測驗。
2. 提出可以評鑑該理論建構是否存在的預測或假設說明。
3. 根據步驟1 所設計出的測驗進行施測，以取得實際的資料進行分析，以驗證上述的預測或假設是否屬實。
4. 收集其他相關的資料，淘汰與理論建構相反的試題或是修正理論，並重複

步驟2 和3，直到上述的預測或假設得到驗證，而測驗的建構效度獲得支持為止。

由此可見，建構效度的建立過程即是一種教育研究的過程，需要先有理論的建構、形成假設、蒐集資料去驗證、反覆修正及檢討建構過程，直到理論建構獲得令人滿意的驗證結果為止。

第三節 試題題型分析

構成一份測驗的基本單位即為試題，有良好的試題才會有良好的測驗，透過個別試題性能的檢驗，以避免測驗中出現過難或鑑別度過低的試題，所有試題均須經過正式的施測，並根據施測的結果進行測驗統計分析，以確定各試題之各項指標數值分析。以下就一般比較常見的幾項量的分析做一番探討：

一、難易度：

試題的難易度與測驗的效率有關，難易度適當的試題是構成優良測驗的必要條件。通常以P 代表試題難易度，N 表示全體受試者人數，R 為答對該題的人數，P_H表示高分組（全體受試者當中分數最高的27% ）答對該題的百分比， P_L表示低分組（全體受試者當中分數最低的27% ）答對該題的百分比。難易度的計算方式有三種方式（簡茂發，民80）：

（一） $$P=\frac{R}{N}$$ ……(公式 2-15)

例如：有一個測驗總共有100 名受試者，其中某一題答對的人數為50 人，則此題的難易度為 P = 0.25

（二） $$P=\frac{P_H+P_L}{2}$$ ……(公式 2-16)

例如：有一個測驗總共有100 名受試者，其中某一題高分組答對的百分比為75% ，低分組答對的百分比為25% ，因此可算得難易度為 P = 0.5

P 值為介於0 與1 之間的一個值，P 值愈大表示試題的愈容易， 相反的P 值愈靠近0 表示試題愈難，例如：三個試題其難易度分別為P =0.25、P =0.5、P =0.75，表示第一題比第二題難，第二題又比第三題難，但難易度為一次序量尺（ordinal scale），差距單位並不相等，其值僅代表試題難、易程度的相對位置，兩個難易度的差不具任何意義，由上面的例子說明，不能說第一題和第二

題的難易度差別與第二題和第三題的難易度差別是一樣的。

（三） 等距量尺（interval scale）分析

美國教育測驗服務社（Educational Testing Service；簡稱ETS）另創一種具有等距量尺特性的難易度指數，以B（delta）表示之。它是一種以13為平均數、4 為標準差、下限為1、上限為25 的標準分數。B值愈小，困難程度愈低；B值愈大，困難程度愈高。它不但可以表示試題難、易程度的相對位置，而且可以指出不同難易度之間的差異數值。此種難易度是基於試題所測量的特質呈常態分配的假設，認為試題的難易度可以在常態分配曲線的橫軸上某一點，以離差分數（deviation score）表示之。其求法係根據答對某一試題的人數百分比與答錯該題的人數（包括未作答者）百分比，使前者在右，後者在左，找出二者在常態分配曲線橫軸上的分界點，此點的相對位置以標準差為單位表示之，即為X，再按下列公式求出B值：B=13＋4X。例如：某一試題的通過人數為84% ，亦即P=0.84，則可知其相當的X 值為-1，代入上述公式，則其B值：B=13＋4（-1）= 9。其次，在實際應用上，試題的B值可由范氏項目分析表查得。當P＞0.5 時，B值＜13；當P＜0.5 時，B值＞13。且所有試題的B值均介於1 和25之間。

關於如何利用難度值來挑選試題，美國的測驗學者Ebel & Frisbie（1991）將試題的難度分為五個等級，如表2-2所示：

表2-2 試題難易度等級表

難 易 度（P）	難 易 度 等 級
P≧0.80	極容易
0.80＞P≧0.60	容易
0.60＞P≧0.40	難易適中
0.40＞P≧0.20	困難
0.20＞P	極困難

（引用Ebel，1979）

一般測驗專家均建議挑選難易度約為 0.5 的試題，也就是難易適中的試題，因為這樣的試題鑑別度可以達到最大，不過在實際的選題上，要使每一題的難易度都接近 0.5 是有些困難的。因此有學者便主張以 0.4 到 0.8 之間的難易度範圍作為選擇題的挑選標準（Chase, 1978），但平均而言，整份測驗的平均難度值還是以接近 0.5 為佳。

基於本研究的受試者是由各校推薦，整體平均能力強，答對率偏高，在難度等級表中，極困難的難易度值需要調整，採用表 2-3 標準作區分。

表2-3試題難易度等級表（調整後）

難 易 度（P）	難 易 度 等 級
P≧0.80	極容易
0.30＞P	極困難

二、鑑別度：

構成測驗的試題必須具有鑑別某種心理特質的作用，才能使測驗成為可靠又正確的測量工具，也就是說試題的鑑別度高低與測驗的信度和效度有著密切的關係，欲增進測驗的預測與診斷的功能，必須要很仔細的分析試題的鑑別度，主要的目的在檢查個別試題與整份測驗之間的一致性分析，分析的方法主要有兩種（余民寧，民86）：

(一) 探求試題反應與測驗總分之間的關聯性：

受試者對某一試題的作答反應可分為答錯或答對兩種情形， 屬於二分變項（dichotomous variable）；而對整份測驗有一個總分，屬於連續變數，兩者之間的關聯性可以使用二系列相關係數（γ_{bis}）或點二系列相關係數（γ_{pb}）來表示其內部一致性的高低，亦可具有此一測驗題目鑑別能力高低的功用（林清山，民82）。

(二) 鑑別指數（index of discrimination）：

D 為鑑別指數，其公式為：

$$D=P_H-P_L \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 2-17)}$$

其中P_H、P_L的定義如同難易度所述。例如：某一試題的$P_H=0.90$ ， $P_L=0.35$，則此試題的鑑別指數D = 0.90－0.35 =0.55，鑑別指數為介於 -1 與1 之間的數值，愈靠近1 表示個別試題反應與測驗總分之間的一致性愈高。

由鑑別度的定義，可以知道鑑別度高的試題，應該可以清楚地將能力高與能力低的受試者分辨得很仔細，但到底鑑別度要到什麼程度才能算是一個好的試題呢？根據Noll、 Scannell 及Craig（1976）等人的看法，可接受的最低標準至少要0.25 以上，低於此標準者，即可視為鑑別度不佳或品質不良的試題。美國的測驗學者Ebel（1979）亦曾提出一套鑑別度的判斷標準，供試題命題者作為選題的參考，如表2-4（引自郭生玉，民79）。

表2-4　Ebel鑑別度評鑑標準表

鑑別指數	試題評鑑
0.40 以上	非常優良
0.30 － 0.40	優良，但需小幅度修改
0.20 － 0.30	尚可，但需部分修改
0.19 以下	劣，需要大幅度修改或刪除

基於本研究的的受試者整體平均能力強，同質性高，在鑑別度等級表中，鑑別度劣的試題其鑑別指數需要調整，採用表2-5標準作區分。

表2-5 鑑別度評鑑標準表（調整後）

鑑別指數	試題評鑑
0.30 以上	非常優良
0.10 以下	劣，需要大幅度修改或刪除

三、選項分析

選擇題的選項包括正確選項與誘答選項，正確選項必須明確且不會引起任何爭議，而誘答選項則必須具有誘答的功能，要知道這些性質是否成立則需透過選項分析。選項分析可以讓施測者清楚知道每一試題的所有選項是否符合命題的原則，進行選項的誘答力分析，可以藉整體學生的答題情形，來診斷整體學生的學習狀況，其主要特色於提供一個主要的正確答案供判斷選擇外，另提供兩個以上的不正確項選項，用來吸引或迷惑知識不夠完整或只有片段觀念的學生來選擇它們，以發揮其特有的「誘答」功能，增加試題的鑑別度；意即具有良好誘答能力的不正確選項，能讓程度不夠學生以「隨機猜題」(Randomly Guess)的方式來選擇它，所以選項誘答力的分析從另一角度來看亦可做為判斷試題好壞的參考。

選項分析是透過比較高分組與低分組對正確與誘答選項的選答率，如果分析的結果符合下面兩項要求，則表示該試題的所有選項是合理有效的（郭生玉，民79）：

（一）、正確選項的選答率，高分組必須高於低分組。

（二）、每一個誘答選項均有低分組的受試者選答，且低分組的選答率高於高分組。

如果不符合第一個要求，表示此試題具有負向的鑑別度，不能清楚區別高分組與低分組；至於第二個要求，又有兩個方面需要討論，首先是如果一個選項沒有任何低分組或高分組受試者選答，表示該選項不具任何誘答率，應該在修改題目時將此選項更換；而如果是該誘答選項高分組的選答率高於低分組，則表示該誘答選向的敘述可能有不清楚或錯誤誘導的地方，使得高分組的受試者有較多誤選的情形，因此在修改試題時應特別注意這些選項。