第三節　題庫的試題分析

一、題庫試題分析概說

題庫乃是包含許多爲某些目標、技能或工作（task）而編製的題目，以爲測驗編製者爲某些需求（目的）編製測驗時的基礎。換句話說它是許多題目的匯集。當一個題庫包含有具備〝內容效度〞而且技術品質優良的試題時，測驗的編製就容易多了，其所產生的測驗通常比沒有題庫的情況下的品質要高（洪碧霞，民82）。

建立題庫最重要的一件工作就是如何來選擇高水準的題目。有了高水準的題目，有關教師命題時的困難就可迎刃而解，Hambleton(1982)及Popham(1981)曾分別提及判斷法(judgement methods)及經驗法(empirical methods)來進行題目特徵有關「量」的統計分析，並藉以選擇題目選擇題目。所謂判斷法乃是聘請學科專家們舉行專家會議(expert conference)，藉以評定所欲選擇的題目品質，並且判定是否與所定的目標之間和諧。經驗法是採用檢查受試者對題目反應的情形，以做爲修正題目的參考。

測驗編製有一個共同目標那就是希望以最少的題數（即最經濟），達到最大的功能（即最有效）。吳裕益（民81）認爲要達到以上的目標，通常是對一個大題庫（large pool of items）進行實地測試，然後從題庫中選取對信度和效度貢獻最大的一組題目來組成一份測驗。在編製一個新測驗（或縮短原有的測驗），最後所要採用的一組題目通常要經過「試題分析」（item analysis）過程來選取。

對測驗工具的評鑑（evaluation of instrument），通常也稱爲試題評鑑。王昭明（民75）認爲試題評鑑的目的不外乎下述五點：

1. 改進測量工具（improve the instrument）。
2. 增強學生學習（reinforce learning）。
3. 作爲補救教學（remedial instruction）的基礎。
4. 改進教學（improve instruction）。
5. 增加試題編製的技巧（increase skill in tset construction）。

Erickson 和 Wentling（1976）也認爲試題評鑑應包含下述兩個：

1. 對工具個別部份的評鑑。也就是對測驗工具進行試題分析（item analysis）。
2. 對工具做整體性的評鑑。也就是考慮工具的效度與信度。

優良的測驗必須具備相當水準的成的信度（reliability）與效度（validity），而測驗的信度與效度又取決於全部試題（test items）的性能。對於試題性能的檢驗，則有賴於邏輯的與統計的分析（logical and statistical analysis）（簡茂發，民76）。

因此，對於擬於收集進題庫中各個試題，事先予以進行試題分析的工作是相當重要的。

所謂試題分析就是對每一個測驗試題再做一次檢驗（Reexamining）發現試題的優點和缺點（Strengths and Flaws）即稱爲試題分析。所以，試題分析不但可以使我確定那些題目是不好的（Pour Items），同時也讓我們知道這些題目之所以不能達到預期目標的原因。是故，試題分析使我們變成更好的試題編製者（Chase，1978）。

Martuza（1977）認爲試題分析的目的有下述兩點：

1. 評估測驗試題的品質。
2. 發覺學生對教學內容不了解與誤解之處。

Erickson & Wentling（1976）則認爲試題分析的目的有下述四點：

1. 選出已被認爲適切的題目。
2. 檢查試題之結構有何缺點。
3. 協助教學者設計個別化之補助教學。
4. 增進教學的效果。

王昭明（民75）也提出試題分析的目的有如下諸端：

1. 鑑定試題的結構。
2. 改進測驗試題的品質。
3. 發現學生學習的障礙。
4. 作爲設計補救教學的依據。
5. 作爲改進教師教學的依據。

試題分析可以分爲質的分析（qualitative analysis）與量的分析（quantitative analysis）兩部份（簡茂發，民76）。前者就試題的內容和形式，從取材的適切性（relevaence）與編製試題的技術加以評鑑，就是對試題質的方面做邏輯的分析；後者則基於試題經過選題預試（try-out）的結果，對試題特徵做量的統計分析，這可從心理測驗理論中兩個不同學派對試題分析個別作法來加以說明。

心理測驗理論（或簡稱爲「測驗理論（test theory）」）是一種解釋測驗資料間實證關係（empirical relationships）的系統化理論學說。測驗理論一般劃分爲兩大學派，亦即：

1. 傳統測驗理論（classical test theory，簡稱CTT）

傳統測驗理論主要內涵是「眞實分數模式」（truescore model）（Gullikson, 1987; Lord & Norick,1968），亦即，觀察分數 X等於眞實分數 T與誤差分數 E之和，其數學表示式爲

：$X = T + E$。這個簡單而淺顯的弱勢假設（weak assumption）便是傳統測驗理論的理論基礎模式和發展骨幹。

2. 當代測驗理論（modern test theory，簡稱MTT）

當代測驗理論主要是以「試題反應理論」（item response theory，簡稱ＩＲＴ）（Lord, 1980； Hulin,Drasgow, & arsons, 1983; Hambleton, Swaminathan,& Rogers, 1991）爲理論架構。事實上，ＩＲＴ乃依據強勢假設（strong assumption）而來，也就是可使用一個「數學模式」，來表示可觀察到的受試者行爲（test performance）與其不可觀察到的潛在特質間之關係。

二、傳統測驗理論試題分析

(一).傳統測驗理論試題分析概述

傳統測驗理之試題分析除了前述所說有關試題「質」方面邏輯性的分析之外，而所謂「量」方面的分析乃是對試題逐一分析其難度（item difficulty）、鑑別度與受試者對各項配列答案（options）的反應情形，以作爲修改試題或選擇試題的依據。

而有關個別試題之題目特徵參數（item parameters）的檢驗通常分爲下述三類（Crocker & Algina,1986；吳裕益，民81）：

1. 描述受試於單一題目之反應的分佈情形之指數（如，題目反應的平均數和變異數）。

2. 描述受試於單一題目之反應和某些感興趣的效標之關聯程度的指數（如，題目與測驗總分之相關）。

數（如，題目與測驗總分之相關）。

3. 與「題目變異數」以及「題目與效標之關係」二者有函數關係的指數（如，題目信度和效度指數）。

以傳統測驗理論爲基礎所建立之題庫有其難以克服的缺點，亦即當題庫試題進行更新（update）時題庫將產生很大的限制。因爲以團體爲基礎的統計量數，會因不同的測驗或不同的預試樣本而使新舊題目間產生非線性關係，而導致新舊題目間之量尺的特質（scale property）不一致；也就是說，當題庫要進行更新時，不易將不同試題的試題組合之P值或D值重新量尺化（rescale），因此，每次建立的題庫只是用於單一的測驗（何榮桂，民80）。

洪碧霞（民81）也認爲傳統的題目參數像通過百分比（P）及題目與總分相關（r p.b.）對題庫中試題的描述並不適切，因爲這些參數是樣本依賴，換句話說，依考生特質的不同，參數就會跟著浮動。因此，對甲校很難的題目，對乙校可能並不難，如此，題庫中題庫參數的參考價值就會大打折扣，而題庫的功能也就會跟著劇降。

(二).試題分析的方法

常模參照測驗與標準參照測驗在目的上有很大的不同，題目分析可「質」的分析與「量」的分析，質的分析乃是分析試題的內容與形式，包括”內容效度”（content validity）的分析以及編擬試題技術之評鑑。量的分析是採用統計方法來分析題目的品質，其項目包括難度（difficulty）與鑑別度（discrimination）分析。一份測驗的信度與效度高低與否，則完全取決於個別試題之品質，所以經由題目分析的程序將可提高測驗的信度與效度。

1. 難度分析

(1).以通過百分比表示難度

試題的難度與測驗的效率（effectiveness）有關，難度適當的試題是構成優良測驗的必要條件（簡茂發，民75）。試題的難易程度通常以全體受試者答對或通過該題百分比（percentage passing）來表示，此種方法用於當一個題目是採用二元計分法時。其公式如下：

$$P = \frac{R}{N} \times 100\%$$

P：試題難度
R：答對該題的人數
N：全體受試者人數
P值介於 .00～1.00 之間

另一種求難度的方法是以高分組（測驗總分最高之27%）答對百分比加低分組（測驗總分最低之27%）答對百分比再除以二。當我們欲求得鑑別度時也需計算出高、低分組答對的百分比，所以很多測驗編製者較偏此種表示方法：

$$P = \frac{P_H + P_L}{2}$$

P ：試題難度

P_H：高分組答對該題的百分比

P_L：低分組答對該題的百分比

P值介於 .00～1.00 之間

當試題難度P值越接近 .50時，題目所具備的區別能力越高，越是接近1.00或 .00，就越無法區分受試者之間能力的差異。

在二元計分法當中，題目難度（P）與題目變異數之關係如下：

$$S_i^2 (或 \sigma_i^2) = p_i \times q_i$$

S_i^2 ： 某個題目之變異數

p_i ： 答對某題之百分比

q_i ： 答錯某題之百分比

當題目彼此之間之相關係數是一個常數，當 pi＝.50時，總測驗分數變異數爲最大。

(2).等距量表的難度指數

以通過百分比來表示難度，其試題難度P值是一種順序量尺（ordinal scale） 差距單位並不相等，它只能表示試題難易度的相對位置， 並無法指出各試題之間難易程度的大小。

針對此一缺點，美國教育測驗服務社（Education Testing Service） 另創一種具有等距尺度（interval scale）特性的難度指數，以Δ（delta） 表示。它是一種以13爲平均數、 4爲標準差、下限爲 1、上限爲25的標準分數。Δ值愈小，難度愈低；Δ值愈大，難度愈高。

此外它還有幾樣特點：

①.可表示試題難易的相對位置。
②.可指出不同難易之間的差異數值。
③.假設所測驗而得的題目難易數值成常態分配。
④.試題的難度可在常態分配曲線的橫軸上某一點以離差分數（deviation score） 表示之。

§求法§

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，使前者在又右，後者在左，找出兩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爲單位表示之，既爲 x ，其求法如下式：

$$\Delta = 13 + 4x$$

至於在實際的應用上，試題的Δ值可由范氏項目分析表（Fan，1952）查得。當P＞.50 時，Δ值小於13；當P＜.50 時，Δ值大於13，所有試題的Δ值介於 1～25之間。

2．鑑別度分析

許多測驗的目的是在提供有關受試者在該測驗所欲測量之構念或是該測驗分數所欲預測之外在效標的個別差異之訊息但有時候所測量的構念沒有辦法找到比該測驗總分更恰當的量數。（如課堂上的成就測驗，通常不容易找到適當的外在效標。）吳裕益（民81）認爲在此種情況下，就以測驗總分作爲受試者在所測量之構念相對地位之操作性定義（operational definition）。使用此種內在效標主要目的在於選擇那些得高分者其答對之機率也較高，得低分者其答對之機率也較低之試題。試題的鑑別力（discriminating power）之大小與測驗的信度與效度都有密切的關係。所以，如想增進測驗之預測與診斷的功能，都必須著重試題鑑別度的分析。試題的鑑別度可分爲題目效度（item validity）與內部一致性（internal consistancy）分析兩方面。前者在分析受試者在題目上的反應與效標上之表現的情形，後者是在分析個別試題與整個測驗總分的一致性。

(1).題目效度分析

①.題目特徵曲線法

題目特徵曲線（item-characteristic curves）可用來表示每個題目之效度。將受試者實際作答資料，繪製於一個以效標分數爲橫座標、以通過百分比爲縱座標的座標系上的試題特徵曲線。我們可直接由座標系上看出特別突出之題目表示效度越高，越平坦者效度越低。

②.相關係數法

以題目特徵曲線表示效度可以看出效度的高低，但要應用於電腦化題庫上，則最好採用單一數字指數來表示題目的效度。事實上用來求取題目效度指數的方法至少有五十種以上（陳英豪、吳裕益，民71）。以下便是幾種較常表示的方法：

a. 點二系列相關（point-bisorial correlation）

以點二系列相關來表示題目之效度，其最適用於效標爲連續分數，而題目爲二分變數。其計算式如下：

$$r_{pb} = \frac{\bar{X}_p - \bar{X}_q}{S_t}\sqrt{pq}$$

r_{pb}：某個題目之變異數

$\bar{X}_p$：答對者之受試者在效標上的平均得分

$\bar{X}_q$：答錯者之受試者在效標上的平均得分

S_t：全部受試者在效標得分的標準差

p：答對某題之百分比

q：答錯某題之百分比

b. 二系列相關（bisorial correlation））

此種效度的表示方法假定受試者在試題上的反應為常態分配，而且題目只有答對及答錯兩種情況。其計算式如下：

$$r_{bis} = \left(\frac{\bar{X}_p - \bar{X}_q}{S_t} \right) \left(\frac{pq}{y} \right)$$

r_{pb}：某個題目之變異數

$\bar{X}_p$：答對者之受試者在效標上的平均得分

$\bar{X}_p$：答錯者之受試者在效標上的平均得分

S_t：全部受試者在效標得分的標準差

p：答對某題之百分比

q：答錯某題之百分比

y：常態分配下答對百分比（P）所在位置之曲線的高度

c. ϕ 係數（phi coefficient）

以 ϕ 係數表示相關用於題目與效標均爲二分變數的情況。計算 ϕ 係數前先將資料化爲如下自由度爲 1，2 × 2 的列聯表。

	答對	答錯	
合格	A	B	（A＋B）
不合格	C	D	（C＋D）
	（A＋C）	（B＋D）	

其次將上述列聯表之資料帶入下式計算φ係數：

$$\phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

②.鑑別度指數

此種鑑別度指數是目前採用最普遍的題目分析法。一般以D來表示鑑別度指數D的值在-1.00～+1.00之間。其計算式如下：

$$D = P_H - P_L$$

D ：鑑別度指數
P_H：高分組通過人數百分比
P_L：低分組通過人數百分比

D值越高，表示該題的鑑別度越高。D的平均指數越高，表示測驗的信度越高。

表2-6　兩變項之性質與適用之相關係數

X / Y	二分類別變項	基於常態分配的二分變項	次序變項	等距或比例變項
二分類別變項	ϕ			
基於常態分配的二分變項	ϕ	r_{tet}		
次序變項	r_{rb}	r_{rb}	r_s、τ、W	
等距或比例變項	r_{pb}	r_{bis}	r_s、τ、W	r_{xy}

（資料來源：商茂發，民76，頁194）

ϕ ：ϕ係數（phi coefficient）

r_{tet}：四分相關係數（tetrachoric correlation coefficient）

r_{rb} ：等級二列相關係數（rank-biserial correlation coefficient）

r_{pb} ：點二相關係數（point-biserial correlation coefficient）

r_{bis}：二系列相關係數（biserial correlation coefficient）

r_s ：斯氏等級相關係列（Spearman rank-biserial correlation coefficient）

τ ：斯肯τ係數（Kendall's Tau）

W ：斯肯和諧度係數（Kendall's coefficient of concordance）

r_{xy} ：皮氏積差相關係數（Pearson product-moment correlation coefficient）

(2).內部一致性分析

內部一致性分析旨在瞭解各個試題的功能是否和整個測驗的功能相符一致。一般在進行題目分析時，經常以測驗的總分作爲分析的依據。此種分析可能改善內容效度和建構效度，但無法增進效標關連效度。

內部一致性分析的方法，與題目效度分析的方法很類似，主要是以測驗總分來取代外在效標。其主要方法有下列兩種：

①.求試題反應與測驗總分之關聯性

每一爲受試者之測驗總分屬於連續變數，而受試者的作答情形可分爲答對與答錯兩種，爲二元變數，故可用前述的點二系列相關（r_{pb}）或二系列相關法（r_{bis}）來求得其高低，而二系列相關法可由范氏題目分析表查得。

②.比較高低分組通過每一試題的百分比

$$D = P_H - P_L$$

D值越大，表示作答反應與測驗總分的一致性越高。

三、試題反應理論試題分析

(一)、試題反應理論試題分析概述

當我們擬予實施電腦化適性測驗（computerized adaptive testing）時，是以試題反應理論（item response theory；IRT）爲基礎來建立題庫：亦即在建立題庫之前，須選擇適當之試題反應模式（item response model）將試題加以校準（calibration），然後將校準過（calibrated）的試題參數（item parameters）挑選品質較佳之題目，以建立校準化的題庫（calibrated item bank）（何榮桂，民80）。

試題反應理論ＩＲＴ發跡於４０年代中葉，Ledyard Tucker（1946）就已經已提及「題目特徵曲線」（itemcharacteristic curve，簡稱ＩＣＣ）這個名詞與觀念，其實早在1916年Binet及 Simon編製智力量表時，就已經使用了ＩＲＴ的核心觀念－題目特徵曲線，但一直到60年代末，測驗領域仍以古典測驗理論中眞實分數模式爲主（亦即，觀察分數等於眞實分數與誤差分數之和，數學公式爲：Ｘ＝Ａ＋Ｅ。）。

表2-7　對ＩＲＴ發展有實際貢獻的代表學者與研究成果（余民寧，民81）

作者（年代）	代表作及其貢獻
Tucker (1946)	第一位提出試題特徵曲線概念的人。
Lord (1952)	第一位導出兩個參數常態肩形模式的參數估計公式，並考慮試題反應理論應用性的人。
Rasch (1960)	試題反應理論中 Rasch模式的創始者，影響深遠。
Lord & Novick (1968)	第一本介紹古典與當代測驗理論模式的經典作品，引發學者對「潛在特質」概念的重視與研究。
Wright & Panchapakesan (1969)	美國地區第一次介紹 Rasch模式的參數估計法，並發展有名的 BICAL電腦程式代表作品。
Samejima (1969)	她的一系列作品描述新的試題反應模式及其應用，其中包含處理多分法知連續性資料的模式，甚至擴展到多向度的試題反應模式，為一極具貢獻的重要著作。
Bock (1972)	提供許多估計模式參數的新概念。
Andersen (1973)	創造地區談論測量模式的重要著作。
1976年	Lord等創作第一版有名的電腦程式：LOGIST。
1977年	Journal of Education Measurement第四季出版一冊專門談討試題反應理論的專輯。
Baker (1977)	第一次評論試題反應模式參數估計法的文獻探討。
Wright & Stone (1979)	第一本描述各種 Rasch模式理論及其應用的專著。
Lord (1980)	第一本以試題反應理論為命名的專著，是當代測驗理論發展的里程碑。
Weiss (1980)	第一本編輯成的論文集，專談試題反應理論的實際應用課題——電腦適性化測驗。
Andersen (1980)	對測量模式參數估計法有貢獻的方法學專論。
Bock & Aitkin (1981)	提出邊際的最大近似值估計法——BM估計程序，對參數估計法之改進貢獻很大。
Masters (1982)	第一位發表部份知識計分模式，對改進Likert式評定量表的計分與次序反應資料的計分貢獻
Wright & Masters (1982)	陳述 Rasch模式的各種模式成員，說明各與部份計分模式相通，對Likert式評定量表與次序反應資料的計分方式改進不少。
Mislevy & Bock (1982)	發表另一有名的電腦程式：BILOG。
1983年	Applied Psychological Measurement 第四季出版一冊專門談討試題反應理論及其應用的專輯。
Weiss & Hessick (1983)	編輯成的論文集，以期表揚Lord一生對試題反應理論的貢獻，並談論該理論的應用與未來。
Weiss (1983)	編輯一本專談試題反應理論的應用與未來，並介紹它在電腦適性化測驗上應用的論文集。
Hulin,Drasgow, & Parsons (1983)	為一本試題反應理論的教科書，增加對「適合度測量」概念的說明與應用。
Hambleton (1983)	編輯一本試題反應理論模式與應用的論文。
Embretson (1985)	編輯一本試題反應理論的未來發展的論文集。
Baker (1985)	為一本導論性的試題反應理論教科書，專為沒有數學訓練基礎的讀者而作，並附有 CAI的電腦教學軟片。
Hambleton & Swaminathan (1985)	為一本通論的試題反應理論教科書。
Crocker & Algina (1986)	比較古典與當代測驗理論的導論性教科書。
Weiner & Braun (1989)	專談有關效度方面的論文集，也討論試題反應理論在效度上的應用。
Linn (1989)	負責主編第三版「教育評量」（Educational Measurement）其中一章專門介紹並評論試題反應理論。
Freedle (1990)	人工智慧及其在當代測驗理論上應用論文集。
Suen (1990)	介紹各種測驗理論方面的教科書。
Wainer等人 (1990)	專談電腦化適性測驗方面的入門書，也談試題反應理論在電腦化適性測驗上的應用。
Hambleton,Swaminathan, & Rogers (1991)	試題反應理論方面的入門書，適於非數學主修的初學者閱讀。

1. 試題反應理論的意義

所謂試題反應理論ＩＲＴ就是以一數學模式，來表示可觀察到的受試者行為（test performance）與其不可觀察到的潛在特質間之關係，也就是一種數學函數的形式（Hambleton & Cook，1977）。Hambleton（1985）也認為：試題反應理論是用來解釋測驗題目、受試者反應及個人特質等三者之間相互關係的一種理論架構（請參閱下圖）。倘依據上述所言，我們可以說：所謂試題反應理論就是以一個預先設定的數理統計學機率模式，將受試者看不見的潛力與他自己作答時的實際得分情形聯結在一起，當得分累積至一定程度時，受試者看不見的潛力便可藉由統計的方法推算出來。所以，試題反應理論的基礎實際上是奠定在預先設定的數學函數上，而函數的表示方法隨著實際應用情形而有調整，我們無法決定函數表示法是否會正確，但可依核對函數決定的得分情形和受試者實際的得分情形來看結果是否相符。

試題反應理論可以分為廣義和狹義的方式（許擇基、劉長宣，民80）；廣義的ＩＲＴ中，受試者的測驗成績是由一些看不見的能力特質來決定。考生的能力特質有高低之分，所以我們可以用數值來表達不同受試者在能力特質上的相對程度。通常受試者在第ｉ個特質上的程度以 θi 來表示，所以一個測驗若測量ｋ種能力特質，受試者在ｋ度空間中的位置是（ $\theta 1$ ，$\theta 2$ ，$\theta 3$ ．．．．θk ），函數和對應的假設加起來則組成為一個ＩＲＴ中的模式（model）。改變函數表示法或假設，則產生新的反應模式，所以在廣義的ＩＲＴ中，有無數的試題反應模式。但在

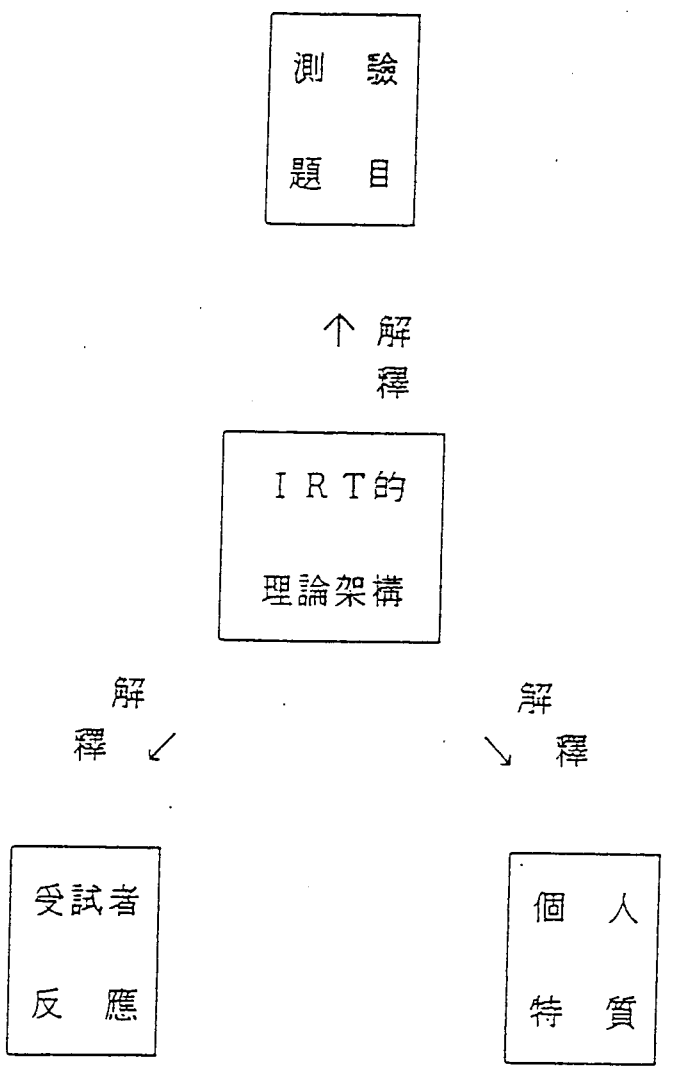

圖2-10 ＩＲＴ可用來解釋測驗題目、受試者反應與個人特質的理論架構

實際應用時，ＩＲＴ通常指的是狹義的定義；在狹義的ＩＲＴ中

試題測量的能力空間被縮小到一度。換句話說，將測量範圍，限制在一種能力或特質上，同時函數的表示法，採用的是logistic-family（數理邏輯家族）中的成員。基本上這些成員有一大家子人，但在狹義的ＩＲＴ中，經常被用來討論的函數表示法則是：⑴.單參數模式（one-parameter model）；⑵.雙參數模式（two-parameter model）；⑶.三參數模式（three-parameter model）。

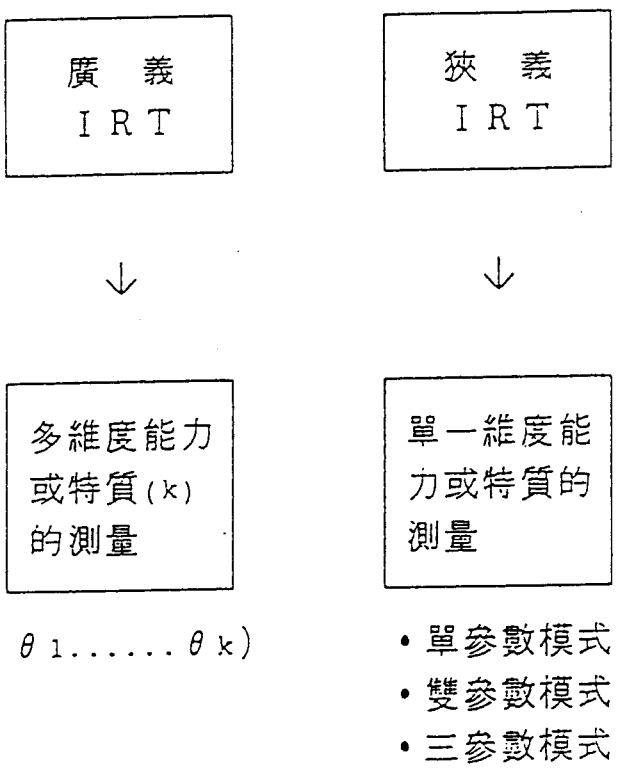

圖2-11　廣義與狹義ＩＲＴ

2. 試題反應理論的基本假設

前面我們曾經提及ＩＲＴ乃一個依據強勢假設（strong assumptions）而來，立論與假設均合理且嚴謹的學說。ＩＲＴ是利用函數來模擬受試者答題的心路歷程，而模擬的函數之實際應用是否能屬有效，則要視理論的假設與受試者實際的答題情形是否相符而定。有關ＩＲＴ的理論假設有下述諸端：

⑴「維度」（dimensionality）必須集中。
⑵ 各別試題「局部獨立」（local independence）。
⑶ 某個「題目特徵曲線」就是該題答對或然率對受試者能力的迴歸線。
⑷「答題速度」（speedness）沒有時間限制。

3. 試題反應理論的重要概念

基於ＩＲＴ應用的特質，Hulin（1983）曾提出了五項重要概念：（黃國彥、王以仁，民76）

⑴.在ＩＲＴ中，許多個人之間不同的特性（characteristics）均被加以考慮。在早期如 Lord（1952）建立ＩＲＴ模式時，所重視的個人特性爲能力（ability），如語文能力、數學能力等；Parsons（1979）則曾經將ＩＲＴ應用於職業滿意度之測量的研究。

⑵.在ＩＲＴ的研究中有一共同的現象：假如所觀察到的行爲或測驗反應，係映射出個人的某一潛在特性（latent

characteristic）。例如，某人在學業性向測驗語文部份（scholastic aptitude test-verbal section，SAT-v）中所獲得的分數，就代表了他的某種潛在特性的語文能力。由於潛在特質是無法直接觀察的，故利用ＩＲＴ將個人反應與潛在變項兩者加以連結；亦即以可觀察到的行爲反應（通常是對許多測驗項目的回答），來預測其個人的潛在特性。

⑶.在ＩＲＴ中，是以題目（item）來代表觀察的單位，以測驗（test）或量表（scale）來表示許多題目的集合，並以特質或能力來表示個人的潛在特性。同時採用希臘字母θ（theta）來表示「潛在特質」的數值。

⑷.潛在特質中θ被視爲呈一連續數值的分配，但並不需要假設它呈常態分配。事實上，某些母羣體特性的眞實分配中，也並非完全是常態的。

⑸.不同測量理論，對於界定特質與反應行爲及變項間關係的內容上，自是有所差異。因此，ＩＲＴ僅適用於對個人潛在特質能力θ之推估，而不需要對此特質提出建構效度。

余民寧（民82）則認爲ＩＲＴ建立在兩個基本概念上：

⑴.受試者（examinee）在某一個試題上的表現情形，可由一組因素來加以預測或解釋，這組因素叫作「潛在特質（latent traits）或能力。

⑵.受試者的表現情形與這組潛在特質之間的關係，可透過一條連續性遞增的函數來加以詮釋，這個函數便叫作「題目特徵曲線」（item characteristic curve，簡寫爲ＩＣＣ）。

(二).三參數題目參數估計

試題反應理論包含著一組數學模式（a family of mathematical models）（Hambleton & Cook，1977；Weiss，1983）。Hambleton & Cook（1977）指出：試題反應理論就是以一數學模式，來表示可觀察到的的受試者測驗行爲（test performance）與其不可觀察的潛在特質間之關係，亦即是一種數學函數的模式。所以試題反應理論也常爲試題反應模式（item response modles，簡稱ＩＲＭ）所代替。

所以在廣義的ＩＲＴ中，有無數的試題反應模式。但在實際應用時，ＩＲＴ通常指的是狹義的定義；在狹義的ＩＲＴ中試題測量的能力空間被縮小到一度。換句話說，將測量範圍，限制在一種能力或特質上，同時函數的表示法，採用的是 logistic family中的成員。基本上這些成員有很多，但在狹義的ＩＲＴ中，經常被用來討論的函數表示法則是：⑴.單參數模式（one-parameter model）；⑵.雙參數模式（two-parameter model）；⑶.三參數模式（three-parameter model）。

三參數模式之計算式如下：

$$P_i(\theta) = C_i + \frac{1 - C_i}{1 + exp\{-D_{ai}(\theta - b_i)}$$

$P_i(\theta)$ ： 能力值爲θ之考生答對試題的或然率

θ ： 考生之能力值

D_{ai} ： 爲一常數，三參數模式一般設爲1.7。

a_i ： 鑑別力

b_i ： 難度

c_i ： 猜測度

(三).測驗訊息

試題反應理論以試題及測驗訊息（information）取代傳統測驗理論的內部一致性信度（如α係數）。單一試題訊息是能力值（θ）的函數，它會隨著能力之不同而變化。而測驗訊息則是所有試題訊息之總和，它當然也隨著能力而變

化。表示訊息量時以能力值（-3～+3）爲橫軸，以訊息量爲縱軸。當能力值爲 0時之訊息量最高，而在+2及-2以外較低。

1. 試題訊息的計算方法

$$I_i(\theta) = \frac{D^2 \times a_i^2 \times Q_i(\theta) \left[P_i(\theta) - c_i\right]^2}{P_i(\theta) \times (1 - c_i)^2}$$

$I_i(\theta)$ ： 對 i個試題的訊息

$Q_i(\theta)$ ： 能力值爲θ之考生答錯試題的或然率

$P_i(\theta)$ ： 能力值爲θ之考生答對試題的或然率

θ ： 考生之能力值

Dai ： 爲一常數，三參數模式一般設爲1.7。

a_i ： 鑑別力

b_i ： 難度

c_i ： 猜測度

2．測驗訊息的計算方法

測驗訊息等於該測驗所包括之所有各別試題訊息之總和。

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

$$= \sum_{i=1}^{n} \frac{P_i(\theta)}{P_i(\theta) \times Q_i(\theta)}$$

(四).單元化分析

由題庫組成之測驗應該符合試題反應理論有關單元化（unidimensionality）的假設，亦即每次只測量受試者的一種能力。

有關單元化假設的檢視有下述三種：

1. 因素分析法

常用的方法爲四分相關矩陣而不是ϕ係數，它和二系列相關係數 r_{bis}有類似的意義。

2. 試題相關矩陣特徵法

試題之間相關矩陣的特徵值（eigenvalue）也可用來檢視單元化的假設，如果矩陣的特徵值中有一個特別大，則試題多只測一種能力。

3. 雙難度比照法

將試題分成幾個部份，並將每一部份試題的難度估計出來，最後再將幾個部份的測驗混合起來，混合起來的試題在作一次難度估計，因此每個測驗有兩個難度估計值。將兩組難度值相互比較其差異，我們就可決定各分測驗是否符合單元化的原則。

(五).適合度考驗

三參數模式的試題分析常採用χ 指數來考驗模式的適合度，並藉以考驗測驗中之試題受否有嚴重偏離三參數logistic模式。

四、傳統測驗理論與試題反應理論之比較

洪碧霞（民81）也認為傳統的題目參數像通過百分比（P）及題目與總分相關（r p.b.）對題庫中試題的描述並不適切，因為這些參數是樣本依賴，換句話說，依考生特質的不同，參數就會跟著浮動。因此，對甲校很難的題目，對乙校可能並不難，如此，題庫中題庫參數的參考價值就會大打折扣，而題庫的功能也就會跟著劇降。

當我們擬予實施電腦化適性測驗（computerized adaptive testin）時，是以試題反應理論（item response theory；IRT）為基礎來建題庫：亦即在建立題庫之前，須選擇適當之試題反應模式（item response model）將試題加以校準（calibration），然後將校準過（calibrated）的試題參數（item paraeters）挑選品質較佳之題目，以建立準化的題庫（calibrated item bank）（何榮桂，民80）。

(一).傳統測驗理論與試題反應理論之異同

綜合而言；傳統測驗理論ＣＴＴ與試題反應理論ＩＲＴ有下述幾種同：

1. 受試者能力之表示法不同

(1).ＣＴＴ—以觀察值X表示學生能力高低。
(2).ＩＲＴ—以能力值θ（+3～-3）表示學生能力高低。

2．理論基礎不同

(1)．ＣＴＴ—建立在簡單的函數假設（Ｘ＝Ａ＋Ｅ）上。
(2)．ＩＲＴ—奠立在一預先設定的數學函數模式上，將受試者看不見的潛力和他作答得實際情形聯結起來，當得分夠多時，看不見的潛力就可用統計方法推算出來。

3．參數值穩定性不同：

(1)．ＣＴＴ—樣本依賴。試題參數會依考生特質的不同而不同。
(2)．ＩＲＴ—樣本獨立。樣本不同的參數也可經由線性轉化而成爲樣本獨立的參數值。

4．主要功能不同：

(1)．ＣＴＴ— a.可提供測驗編製與評鑑的客觀標準及分析方法
b.協助正確的解釋和運用考試獲得的分數
c.協助發掘及解決現有的測驗應用問題

(2)．ＩＲＴ— a.測驗的編製
b.建立題庫
c.適性測驗
d.選項的不同加權
e.試題偏向
f.分數的等化

(二).試題反應理論ＩＲＴ尚未普遍使用的原因

國內外研究ＩＲＴ的學者已逐漸增加，但由於某些因素，因而大都停留在理論的探討階段，能將ＩＲＴ實際應用於眞實測驗情境者目前非常少見。其箇中原因有下列五點（Hambleton & Cook，1977）：

1. 試題反應理論係一複雜的測驗理論。同時研究者需要具備較佳的數學能力，以便閱讀許多與此理論有關之數學方面的報告與專書，而甚感困難。

2. 大多數研究ＩＲＴ的學者，在理論探討與介紹方面遠較實際應用者爲多。

3. 並無迅速而容易使用的有關試題反應模式的電腦套裝軟體程式，使得試題反應各參數模式在應用上相當不便（此一問題，迺來已經大爲改善）。

4. 部份學者對於ＩＲＴ的研究與發展，所能獲得的成效深表懷疑。

5. ＩＲＴ的各種測驗模式，往往基於被限制在較強的假設上，當實際應用於心理測驗資料時，卻不能完全符合這些假設，而產生相當大的問題。

當代測驗理論相較於傳統測驗理論有其優勢之處，但它被用來解決實際眞實資料者，則屬鳳毛鱗角，余民寧（民82）也認爲主要有下列五項原因：

1. 當代測驗理論係建立在理論假設嚴謹的數理統計學機率模式上，是一種複雜深奧、艱澀難懂的測驗理論，這對於在數學方面訓練有限的教育與心理學介學者而言，莫非是一大挑戰。閱讀有關此理論之數學方面的研究報告與專書，已感頗爲困難，實在更難以深入將之發揚光大

2. 多數當代測驗理論學者都出自於數學界或曾是學數學主修者，或至少在數理統計方面訓練有數者，他們偏愛對理論模式的探討，遠勝於對實際應用的推廣工作。

3. 過去，電腦科技的進步有限，沒有電腦套裝軟體的即時配合，當代測驗理論中對模式參數的估計，難以用手算或小型計算機順利進行。因此，在應用上更受限制。

4. 有些古典測驗理論的擁護者，對當代測驗理論的研究與發展，所能獲致的成效與應用性深表懷疑。爲了証明與解釋疑惑，當代測驗理論學派的支持者，便朝向理論模式的量化技術方面探討，致使當代測驗理論的發展愈趨數學化、數量化與電腦化。

5. 礙於嚴苛的基本假設，當代測驗理論所能適用的教育與心理測驗資料有限，並且需要大樣本的配合，因此使得它的應用性大打折扣，未獲一般測驗使用者的全力擁戴。

由上述之說明可知，古典測驗理論在立論上雖不夠嚴謹，但因其理論簡明易懂，比較有利於實際測驗情境，特別是小規模資料的實施，如班級、實習課的分組等。而當代測驗理論之立論雖然既明確且嚴謹，但因其包含一「數學模式」

在內，較爲艱澀難懂，僅適用於大樣本測驗資料的分析，如校際、地區、聯招或全國性技術士技能檢定學科筆試等。