2-PL 模式比 1-PL 模式多考慮一個試題參數 a，即鑑別度參數，其試題特徵函數表示為：

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

（三）3-PL 模式：

此模式又比 2-PL 模式多使用了一個猜測參數 c，來描述試題，其試題特徵函數表示為：

$$P_i(\theta)=c_i+(1-c_i)\frac{\exp Da_i(\theta-b_i)}{1+\exp Da_i(\theta-b_i)}$$

c 參數為能力極低時仍有答對該題的機率。

第二節　測驗等化

由於測驗等化牽涉到不同測驗或不同受試者之間的關係，為了使得不同的估計數值可轉變成相同，或容易解釋與應用，測驗學家發展各種測驗等化的理論與技術。

一、建立評量量尺的重要性

若要申請美國研究所(除了商學院外)，大部份的學校都會要求 GRE 的成績，作為他(她)們申請入學之必要條件。假如甲生在今年 1 月份在計量部份考 700 分，語文考 400 分。乙生在今年 3 月份在計量部份考 600 分，語文考 500 分。雖然兩者的考試時間與試卷皆不同，但對於以上兩位考生的成績，我們將會有

一致的看法：甲生的計量部份比乙生好；乙生的語文部份比甲生好。此外，從甲生的計量得分，我們可以推論甲生的計量成績是屬於上，但語文成績部份則為中下，而甲生的數學能力可能比語文好。從乙生的得分，我們可以推論其數學能力是中上，語文能力是中等，同樣地，乙生的數學能力可能比語文好。且間隔一段期間後(如二個月)，甲生與乙生再重考一次，他們的得分應不會有太大的改變，除非在間隔的這一段時間，他們非常用功地準備與復習，或在兩次考試時之身心狀況有極大的差異。

二、等化的設計(王寶墉，民 84；余民寧，民 82；Hambleton & Swaminathan, 1985; Kolen & Brennan, 1995)

1. 單團體設計(single group design)：讓同一組受試者，接受兩個或兩個以上的測驗，是最簡單的一種設計。但易受重覆練習和疲倦因素的影響。
2. 等團體設計(equivalent group design)：以隨機選取或配合抽取能力分配相等的不同受試團體，接受兩個不同的測驗而進行等化。
3. 等測驗設計(equivalent test design)：隨機分派題庫或很多要校準的題目成不同的測驗組合，但有如傳統測驗的平行複本試題，將此平行測驗施測於受試能力分配不必相同的團體，但嚴格的平行測驗實際上不易達成。
4. 校準團體設計(anchor group design)：此設計是以不同測驗施測於不同團體，但校準團體接受每個測驗的施測，而其他團體則只施測一種測驗。此設計的關鍵在校準團體的樣本數決定估計數的穩定性。

5. 校準測驗設計(anchor test design)：此設計是將校準試題放入不同的測驗中，而對不同的團體施測。但將校準試題放入不同的測驗，有下列幾種方式：

甲、內在校準測驗(internal anchor test)：將校準試題放入不同的測驗內。

乙、外在校準測驗(external anchor test)：除了已有的測驗外，另外加測校準試題。

丙、分段校準測驗(cascading anchor test)：不同校準試題放入不同的測驗內。

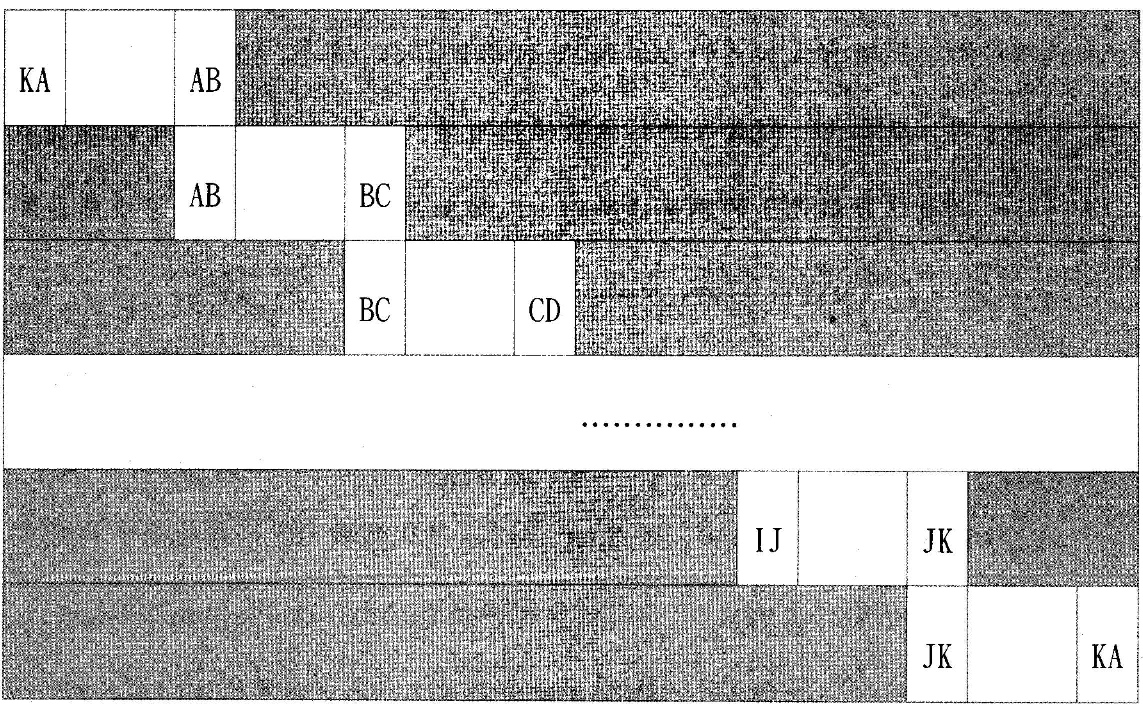
丁、一致校準測驗(uniform anchor test)：相同校準試題放入不同的測驗內。

6. 校準試題矩陣抽樣設計(matrix sampling design with overlapping items)：此設計主要有兩種型態，一是區塊共同試題多重矩陣抽樣設計，另一種是分散型共同試題多重矩陣抽樣設計。兩種將校準試題、非共同試題及不同的受試團體，靈活交互應用，是建立題庫時有效的設計。

7. 同時校準法:藉由測驗資料的收集設計與 IRT 電腦軟體之應用結合，將所收集之數筆測驗資料同時執行試題校準，經過校準之後，所有題目的參數估計值，皆有相同的量尺單位，此乃同時校準法之涵意。

將所收集到的試題，經專家學審題後，組成數份不同之試卷，且試卷裡之

部份題目(定錨試題)是相同的(比率約 15%)(Vale, 1986)，將其施測於預先選好其能力分佈與學測相似的不同學校中，在蒐集並整理測驗的資料中，將幾組樣本對幾份試卷反應資料做調整，並整理成一筆測驗資料矩陣，如下圖。然後以 IRT 電腦軟體，如 BILOG-MG(Zimowski, Muraki, Mislevy & Bock, 1995)，對此一資料矩陣進行試題校準工作。假設此資料矩陣含有 300 道題目，則此 300 題之參數估計值將對應於相同量尺上。經過校準後的試題，必須能夠通過適合度的考驗者，方可被保留在題庫裡，因為它們可以被適當的反應模式所解釋。



灰色部份通常假設其為考生未完成部份。

一般測驗機構所採用的量尺分數根據其產生的方式，大概可以分為兩種：(1) 將原始分數常態轉換後所得 (normalizing raw scores) 之量尺分數。例

如美國教育測驗服務社（ETS）的 GRE 或 TOEFL 測驗分數，就是將原始分數常態化轉換後的量尺分數。智力測驗中的比西量表分數（Stanford-Binet Intelligence Scale）及魏氏兒童智力量表分數（Wechsler Intelligence Scale for Children-Revised Form; WISC-R）等也是一種將原始分數常態化轉換後的量尺分數。(2) 均等測量標準誤（equalizing measurement error variability）之量尺分數。這是 E. L. Lindquist 在發展「愛荷華教育發展測驗」(Iowa Tests of Educational Development, ITED）時所提出來建立量尺分數的方法，也是美國知名測驗機構 ACT 公司的 ACT Assessment Test（ACT, 1997）所採用的量尺分數型態。基本上，這是在原始分數轉換成量尺分數的同時，利用數學的方法將每一個量尺分數點上的測量標準誤（或稱測量誤差）調整成相等或是非常接近（Kolen & Hanson, 1989; Kolen, Hanson, & Brennan, 1992)。

測驗等化的方法可分為古典測驗理論與現代測驗 IRT 的方法。古典測驗理論等化法，是利用原始總分進行等化，又可分為線性等化(linear equating)和等百分位數等化(equipercentile equating)。現代測驗 IRT 的等化法常用的有迴歸法(regression method)、平均數與標準差法(mean and sigma method)、強韌平均數與標準差法(mean and sigma method)、真分數等化法(true score equating)及特徵曲線法(characteristic curve method)(王寶墉，民 84；Angoff, 1971 ; Kolen, 1988；Hambleton & Swaminathan, 1985)。

一、古典測驗理論等化的方法

1. 線性等化法：

兩個測驗的分數分配相同，於各測驗中找出對應相同標準分數之觀察分數，即兩測驗之觀察分數可置於相同量尺比較，亦即除了平均數和標準差不同外，X 和 Y 分數的分佈是相同的。如果這個假設能夠成立，那就能從 X 和 Y 找出具有相同的 z 分數配對，這些分數即為等化分數。亦即如果

$$(X - \mu_X)/\sigma_X = (Y - \mu_Y)/\sigma_Y$$

則 X 與 Y 等值。其中 X 與 Y 分別代表測驗 X 和 Y 的原始分數，μ_X 、μ_Y 分別代表兩測驗的平均數，σ_X 與 σ_Y 分別代表兩測驗的標準差。其直線轉換為：

$$Y = aX + b$$

$$\mu_Y = a\mu_X + b$$

$$\sigma_Y = a\sigma_X$$

$$Y = (\sigma_Y / \sigma_X)X + [\mu_Y - (\sigma_Y / \sigma_X)\mu_X]$$

2. 等百分位數等化法

等百分位數等化法是在決定兩個測驗的哪些分數具有相同的百分等級。第一步是決定兩個測驗分數分佈上的百分等級。然後將兩個測驗的百分等級對原始分數的散佈圖畫出來。

在連結各資料點時必須利用直線內插法。如假設在 A 測驗卷的原始分數是 a，其對應的百分等級為 P，而在 B 測驗卷，從百分等級 P 所對應的原始分數是

b。則我們可以說一個人在A測驗的得分a與在B測驗的得分b是具有相同能力的。

Lord(1980)認為測驗分數要能公平的等化，需滿足下列需求：

A. 測量不同特質或能力的測驗不能等化。

B. 信度不相等的測驗之原始分數不能等化。

C. 難度不相等的測驗之原始分數不能等化。

D. 除非兩測驗是完全複本測驗，否則兩測驗不能等化。

E. 具有完全信度的測驗分數才可以等化。

此外，對稱性(symmetry)與不變性(invariance)等二條件亦是進行測驗分數等化所必備的。對稱性是指無論是X轉化為Y或Y轉化為X，等化結果必須相同。不變性則指等化的程序是獨立的，不受所選用樣本的影響。

由於古典測驗理論是依據弱勢假設而來，理論淺顯易懂、便於計算，但卻有下列的缺失與限制：

A. 難度、鑑別度與樣本相依，即難度、鑑別度隨樣本的不同而變化。

B. 不同測驗不易比較不同受試者的能力。

C. 理論中假設所有受試者等測量標準誤與實際不符。

亦即以原始分數來進行等化有其限制存在，如公平性、對稱性與不變性等要求常無法獲得滿足。

二、試題反應理論的等化方法

常用的試題反應理論的等化方法有下列幾種：迴歸法(regression method)、平均數與標準差法(mean and sigma method)、強韌平均數與標準差法(robust mean and sigma method)、真分數等化法(true score equating)及特徵曲線法(characteristic curve method)（王寶墉，民 84；Angoff, 1971；Kolen, 1988；Hambleton & Swaminathan, 1985）。

1. 迴歸法(regression method)

兩測驗分數的直線關係，可由下列迴歸式來表示：

$$y = \alpha x + \beta + e，\alpha = r_{xy}\frac{s_y}{s_x}，\beta = \bar{y} - \alpha\bar{x}$$

y， x 可以是能力，即 $y = \theta_y, x = \theta_x$，也可以是難度，如 $y = b_y, x = b_x$。誤差 e 則視為獨立而相等的隨機變項，迴歸係數α與β是斜率與截距。r_{xy}是兩測驗相關係數，s_x, s_y是標準差，$\bar{x}$與$\bar{y}$是平均數。迴歸法的缺點是不對稱，只有 Rasch 模式例外。

2. 平均數與標準差法(mean and sigma method)

平均數與標準差法是假如 $y = \alpha x + \beta$，則 $\bar{y} = \alpha\bar{x} + \beta$ 且 $s_y = \alpha s_x$，

所以得到$\alpha = s_y / s_x$，$\beta = \bar{y} - \alpha\bar{x}$，此關係是對稱的。

3. 強韌平均數與標準差法(mean and sigma method)

如果連結是包含兩個以上的參數，使用平均數與標準差法的估計誤差變化很

大，而且離異值(outlier)會影響係數的計算。因此，Linn, Levine, Hastings, & Wardrop(1981)提出強韌平均數與標準差法。

強韌平均數與標準差法認為應給每一對(x, y)的能力或參數加權，此加權得自兩個難度參數估計值的變異數中較大一個的倒數，其步驟為：

a. 決定每個配對(x_i, y_i)的加權植w_i，

$w_i = \max\{Var(x_i), Var(y_i)\}$， j=1, 2, …, k,

b. 加權值量尺化，新加權值： $w_i' = w_i / (\sum_{i=1}^{k} w_i)$

c. 計算新的 x 與 y 值

$x'_i = w'_i x_i$ ， $y'_i = w'_i x_i$ j=1, 2, …, k,

d. 求出加權後的平均數與標準差：

$\bar{x}', \bar{y}', s'_x, s'_y$

e. 以加權後的平均數與標準差求出α，β ：

$\alpha = s'_y / s'_x$ $\beta = \bar{y}' - \alpha\bar{x}'$

4. 特徵曲線法(characteristic curve method)

上述所提的各種等化方法，都只考慮能力與難度指數，其餘之鑑別度及猜測度都未列入考慮，因此，Haebara(1980)及 Stocking & Lord(1983)提出特徵曲線法，方法如下：

具有 θ_a 能力的受試者在測驗 X 的真分數為 ξ_{xa}

$$\xi_{xa} = \sum_{i=1}^{n} P(\theta_a, a_{xi}, b_{xi}, c_{xi})$$

具有 θ_a 能力的受試者在測驗 Y 的真分數為 ξ_{ya}

$$\xi_{ya} = \sum_{i=1}^{n} P(\theta_a, a_{yi}, b_{yi}, c_{yi})$$

試題參數是：

$$b_{yi} = \alpha b_{xi} + \beta,$$

$$a_{yi} = a_{xi} / \alpha,$$

$$c_{yi} = c_{xi}$$

為求兩真分數間差異的最小值，其函數為

$$F = \frac{1}{N} \sum_{a=1}^{N} (\xi_{xa} - \xi_{ya})$$

用 Newton-Raphson 解出以下偏導數方程式，即可得兩真分數的最小值：

$$\frac{\partial F}{\partial \alpha} = \frac{\partial F}{\partial \beta} = 0$$

因特徵曲線法涉及三個試題參數，等化的效果最佳。

5. 真分數等化法(true score equating)

IRT 能力分數之評量單位不易為一般人所接受，而且長久以來大眾皆習慣採用原始分數來計分。有鑑於此，IRT 原始分數(IRT observer score equating)與真實分數等化法(IRT true score equating)乃被發展運用來克服此問題。IRT 真實分數為考生在 n 題題目答對機率的總和，公式如下(Lord, 1980)：

$T \mid \theta = \sum_{i=1}^{n} P_i(\theta)$

真實分數等化方法包括兩個主要步驟(Kolen & Brennan, 1995)：

A. 將兩份試卷的試題參數估計值對應在相同之量尺上。

B. 使用 IRT 分數為橋樑，製做兩份試卷真實分數的對應表。

第二步驟之原理為：假定兩組(分為基礎組與等化組)樣本之能力參數之單位相同，且兩份試卷的試題參數值之單位亦同，將基礎組每一位受試者的 IRT 能力分數轉化為真實分數；同樣地，也等化組每一位受試者的 IRT 能力分數轉化為真實分數。然後以 IRT 分數為橋樑，製造出兩份試卷真實分數的對應表。IRT 真實分數等化方法之最大優點為：當我們從題庫裡選定試題樣本後，即可開始製造試卷間之分數轉化表，而不需要等到考完試取得考生作答資料再來做。但先決條件為題庫裡試題的參數估計值要準確。

6. 觀察分數等化法

真實分數ξ和觀察原始分數 r 有相同的量尺，而

$$r = \sum_{i=1}^{n} U_i$$

若項目反應理論是有效的，則

$$E(r) = \xi$$

一個測驗理論上的觀察分數分佈 $f(r \mid \theta)$ 可以由下列等式中獲得

$$\sum_{r=0}^{n} f(r \mid \theta)\, t^n = \prod_{i=1}^{n} [Q_i(\theta) + tP_i(\theta)]$$

若θ的值已知，如N個考生的能力為$\theta_1,\theta_2,...,\theta_a,...,\theta_N$，則$f(r)$的邊際分配為

$$f(r)=\sum_{a=1}^{N}f(r\mid\theta_a)$$

我們可以利用下列例子來做說明：

假設有一個2參數的logistic模式，有三個題目，其

難度 $b=[b_1,b_2,b_3]=[1.0,0.0,-1.0]$

鑑別度 $a=[a_1,a_2,a_3]=[1.5,1.0,0.5]$

給定下列5個能力 $\theta=[\theta_1,\theta_2,\theta_3,\theta_4,\theta_5]=[-2,-1,0,1,2]$

根據上面的條件，我們可以獲得在不同能力的理論上觀察分數分佈如下：

	條件相對次數 $f(r\mid\theta)$				
原始分數	$\theta=-2$	$\theta=-1$	$\theta=0$	$\theta=1$	$\theta=2$
0	.678	.420	.139	.012	.000
1	.313	.500	.475	.143	.009
2	.010	.080	.361	.488	.158
3	.000	.000	.025	.357	.833

假設總共的考生有100人，而且具有能力θ=-2, -1, 0, 1, 2的人分別有5, 15, 30, 40, 及10人，則原始分數的理論邊際次數分配可計算得如下表：

	條件相對次數 $f(r \mid \theta)$					邊際次數分配 $f(r)$
原始分數	$\theta=-2$	$\theta=-1$	$\theta=0$	$\theta=1$	$\theta=2$	
0	3	6	4	0	0	13
1	2	8	14	6	0	30
2	0	1	11	20	2	34
3	0	0	1	14	8	23
考生人數	5	15	30	40	10	100

根據以上，觀察分數等化法的步驟如下：

a. 根據 $\sum_{r=0}^{n} f(r \mid \theta) t^{n} = \prod_{i=1}^{n} [Q_i(\theta) + tP_i(\theta)]$ 及考生的能力及測驗 X 的項目參數，可以得到條件次數分配

b. 再從 $f(r) = \sum_{a=1}^{N} f(r \mid \theta_a)$ 得到邊際次數分配

c. 對於測驗 Y 重覆 a, b 的步驟

d. 用等百分位數來等化此兩個測驗的原始分數

與觀察分數等化法比較，IRT 真實分數等化方法之計算過程較為簡易，此方法無須依賴考生能力之分佈狀態，而且經過 Lord & Wingersky (1983)的研究發現，此兩種方法的等化結果相似，況且 ETS 測驗公司亦採用 IRT 真實分數等化方法。因此，本研究擬採用 IRT 真實分數法來對兩次學測的數學科做等化，然後與國中基本學力測驗中心的結果比較做探討。