試題反應理論的介紹(三)

——試題反應模式及其特性

余民寧　著

根據 Thissen & Steinberg (1986)的分法，所有的試題反應模式(item response models)依其基本假設與參數估計時的設限不同，可以歸納爲下列三大類：

(一)差異模式(difference models)：適用於次序反應的資料；

(二)除總模式(divide-by-total models)：適用於次序和名義反應的資料；

(三)左加模式(left-side added models)：適用於有猜題(guessing)可能的單選題反應資料。

雖然歸類方式不盡相同，到目前爲止，大多數已發展出來並且已在使用中的試題反應模式，還是以適用於二元化計分(binary or dichotomous scoring)的性向或成就測驗資料爲主。本文的目的，即在介紹試題反應理論中最常用的基本模式及其具有的特性。

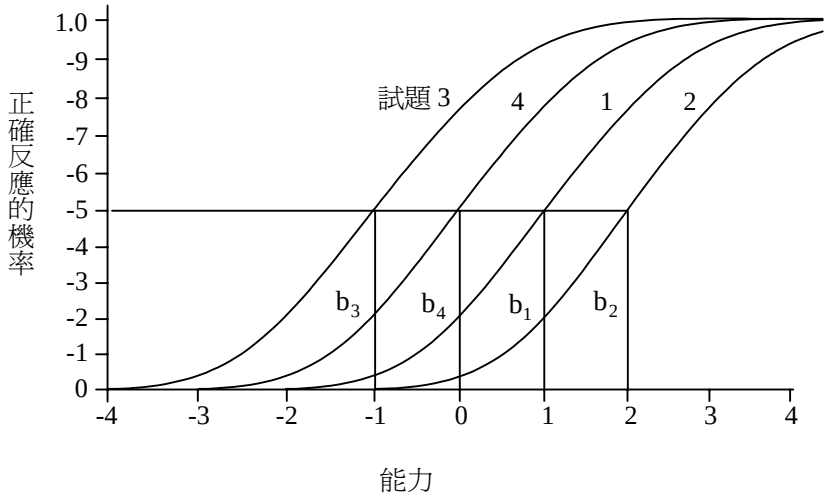
基本的試題反應模式

前文已經說過，試題特徵函數或試題特徵曲線是用來描述測驗所欲測量的潛在特質，與其在試題上正確反應之機率間的一種數學關係；因此，每一種關係就有其相對應的一條試題特徵曲線存在，亦即是每一種試題反應模式都是用來描述特質與正確反應機率間的關係。常用的試題反應模式，有下列三種，每一種模式都依其採用的試題參數的數目多寡來命名，都僅適用於二元化的反應資料（亦即，正確反應者登錄爲 1，錯誤反應者爲 0 的資料）。

(一)一個參數對數形模式(one-parameter logistic model)：這個模式的數學公式如下所示：

$$P_i(\theta)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}} \qquad i=1,2,\cdots,n$$（公式一）

其中，$P_i(\theta)$ 表示任何一位能力爲 θ 的考生答對試題 i 或在試題 i 上正確反應的機率；b_i 表示試題難度(difficulty)參數；n 是該測驗的試題總數；e 代表以底爲 2.718 的指數；且 $P_i(\theta)$ 是一種 S 形曲線，其值介於 0 與 1 之間。一個參數的試題特徵曲線如圖一所示。



圖一　四條典型的一個參數試題特徵曲線

根據公式一的定義，試題難度參數 b 的位置正好座落在正確反應機率爲 0.5 時能力量尺(ability scale)上的點；換言之，當能力和試題難度相等時（即 $\theta - b_i = 0$ ），考生答對某試題的機會只有百分之五十。當能力小於試題難度時（即 $\theta - b_i < 0$ ），考生答對某試題的機會便低於百分之五十；反之，當能力大於試題難度時（即 $\theta - b_i > 0$ ），考生答對某試題的機會便高於百分之五十。b_i 值愈大，考生要想有百分之五十答對某試題的機會，他／她便需要有較高的能力才能辦到，亦即該試題是屬於較困難的題目。愈困難的試題，其試題特徵曲線愈是座落在能力量尺的右方；反之，愈簡單的試題，其試題特徵曲線愈是座落在能力量尺的左方。圖一所示，四條試題特徵曲線的試題難度參數分別爲 $b_1 = 1.0, b_2 = 2.0, b_3 = -1.0, b_4 = 0.0$ ，其值的大小，分別決定該四條曲線在能力量尺上的相對應位置，因此，試題難度參數有時又叫作位置參數(location parameter)。

理論上，b 值的大小介於 $\pm\infty$ 之間，但實際應用上，通常只取 ± 2 之間的範圍。由圖一所示，b 值愈大表示試題愈困難，b 值愈小表示試題愈簡單。b 值的概念符合常理的想法，但不同於古典測驗理論中對難度 P 值的概念定義： P 值愈大表示試題愈簡單， P 值愈小表示試題愈困難，其概念正好與常理的想法相反。這正是試題反應理論在解釋試題特性上的一大優點。

由圖一所示，四條曲線的形狀是一致的，但在能力量尺上的位置各有不同，這點顯示出：在一個參數模式下，影響考生在試題上表現好壞的試題特性只有一個，那就是該試題的難度。一個參數對數形模式並不把試題鑑別度(discrimination)指數考慮在內，其實，這種作法等於是假設所有試題的鑑別度都是相等的（通常設定爲 1）。同時，它亦假設試題特徵曲線的下限(lower asymptote)爲零，亦即對於能力非常低的考生而言，他／她答對某試題的機會是零；換言之，一個參數對數形模式假設能力低的學生沒有猜題猜中的可能，雖然考生們在單選題試題上往往會猜題。

很明顯的，一個參數模式的假設非常地嚴格。這些假設的適當與否，端視資料

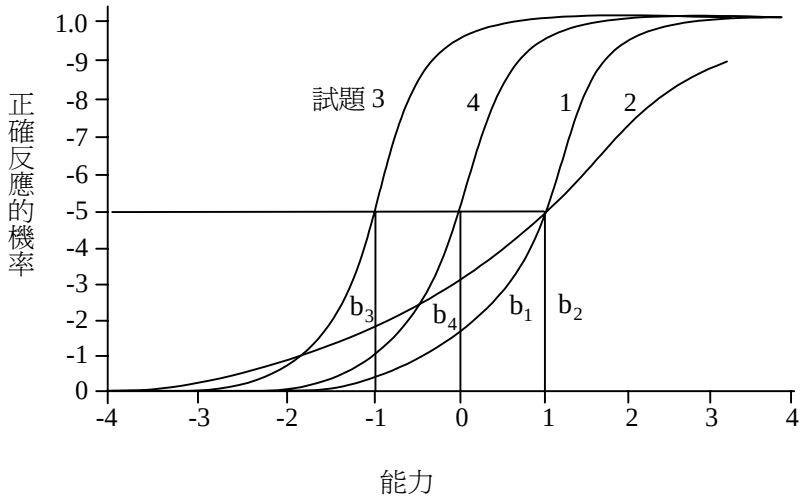
本身的特性和所欲應用該模式的重要性而定。例如，從一堆同質性頗高的題庫(item bank)中選取相當容易的試題編製而成的測驗，便非常符合這些假設的要求，這類情境常見於在有良好施測指導語下的效標參照測驗(criterion-referenced tests)中。

一個參數對數形模式相通於 George Rasch (1960)的模式，因此又有 Rasch 模式之稱，以紀念這位丹麥的數學家在測驗理論上所作的貢獻。Rasch 模式通行於歐洲地區的心理計量學界，以及美國芝加哥大學等大學，有關 Rasch 模式的發展詳情可參閱 Rasch (1960)、Wright & Stone (1979)、和 Wright & Masters (1982)。

(二)兩個參數對數形模式(two-parameter logistic model)：這個模式的數學公式如下所示：

$$P_i(\theta)=\frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \qquad i=1,2,\cdots,n$$（公式二）

其中，各符號的定義與公式一相同，唯多了一個參數：試題鑑別度(item discrimination) a_i，它的涵義與在古典測驗理論中的涵義相同，同是用來描述試題 i 所具有鑑別力大小的特性。典型的二個參數的試題特徵曲線，可參見圖二所示。



圖二　四條典型的二個參數試題特徵曲線

試題鑑別度參數 a 的值，剛好與在 b 點的試題特徵曲線的斜率(slope)成某種比例。試題特徵曲線愈陡(steeper)的試題比稍平滑的試題，具有較大的鑑別度參數值；換句話說，鑑別度愈大的試題，其區別出不同能力水準考生的功能愈好，亦即分辨的效果愈好。事實上，該試題能否區別出以能力水準爲 θ，上下兩組（即高於 θ 和小於等於 θ）不同能力考生的有效性，是與對應於 θ 量尺的試題特徵曲線的斜率成某種比例。

理論上，a 值的範圍在 $\pm\infty$ 之間，但學者們通常捨棄負的 a 值不用，因爲該試題反向區別不同能力水準的考生，此外，帶有負值 a 的試題特徵曲線代表著：能力愈高的考生答對某試題的機率愈低，這似乎與學理相違背，所以負的 a 值不用。通常，a 值也不可能太大，常用的 a 值範圍介於 0 與 2 之間；a 值愈大，代表試題特

徵曲線愈陡，試題愈有良好的分辨能力；a 值愈小，代表試題特徵曲線愈平坦，正確反應的機率與能力間成一種緩慢增加的函數關係，亦即試題愈無法明顯有效地分辨出考生的能力水準。

很明顯的，二個參數對數形模式是由一個參數對數形模式延伸演變而來，亦即把試題鑑別度參數考慮進一個參數對數形模式裡，便成爲二個參數對數形模式。圖二所示，四條試題特徵曲線的試題參數分別爲 $a_1=1.0, b_1=1.0, a_2=0.5, b_2=1.0, a_3=1.5, b_3=-1.0, a_4=1.2, b_4=0.0$，這些參數決定試題特徵曲線的形狀不會是平行的，因爲有不同大小的試題鑑別度值存在的關係。當這四條試題特徵曲線的 a 值都相等時，這些曲線便成平行的 S 形曲線，如圖一所示；因此，我們可以這麼說：一個參數對數形模式是二個參數對數形模式的一種特例，亦即把試題鑑別度參數都設定成一致時（通常設定 $a_i=1, i=1, 2, \cdots, n$），公式二的數學式子便簡化成公式一的數學式子，這種說法於是成立。

由圖二亦可知，這些曲線的下限值都是零，亦即二個參數對數形模式並不把考生的猜題因素考慮在內，這點假設與一個參數對數形模式雷同。猜題因素不存在的假設，往往使二個參數對數形模式適用於自由反應(free-response)的試題分析，或試題不太困難的單選題測驗分析，對於有良好施測指導語的能力測驗資料亦可適用。

二個參數對數形模式是由 Birnbaum (1968)修改自 Lord (1952)的原始二個參數常態肩形模式(normal ogive model)而來，由於它比常態肩形模式易於計算和解釋，目前已取代常態肩形模式，而成爲主要的試題反應模式。如果我們把公式二的分母與分子同時除以 $e^{a_i(\theta-b_i)}$，公式二也可以寫成下列的橫等式：

$$P_i(\theta)=[1+e^{-a_i(\theta-b_i)}]^{-1}$$

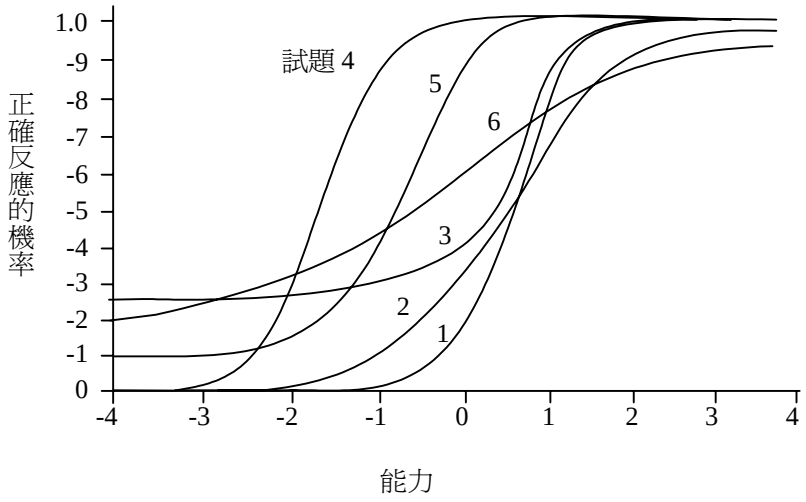
這個公式是二個參數對數形模式的另一種慣用表示方法。

(三)三個參數對數形模式(three-parameter logistic model)：這個模式的數學公式如下所示：

$$P_i(\theta)=C_i+(1-C_i)\frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \qquad i=1,2,\cdots,n$$（公式三）

其中，各符號的定義與公式二相同，唯多出一個參數：機運參數(pseudo-chance parameter) C_i。這個參數提供試題特徵曲線一個大於零的下限，它代表著能力很低的考生答對某試題的機率。

三個參數對數形模式是由二個參數對數形模式延伸演變而來，它多增加一個參數 C，即是把低能力考生的表現好壞因素也考慮在模式裡，當然，猜題可能是這些考生在某些測驗試題（如：選擇題）上唯一的表現行爲。通常，C 參數的值比考生在完全隨機猜測下猜答的機率值稍小，亦即 $C_i \le 1/A_i$，A_i 代表試題 i 的選項數目。Lord (1974)認爲，這是由於出題者通常會在試題中佈置誘答選項的緣故，基於這項理由，C 不應該完全被視同「猜題參數」。三個參數的試題特徵曲線如圖三所示。



圖三　六條典型的三個參數試題特徵曲線

圖三所示，六條試題特徵曲線的試題參數分別爲 $a_1 = 1.8, b_1 = 1.0, c_1 = 0.0, a_2 = 0.8, b_2 = 1.0, c_2 = 0.0$, $a_3 = 1.8, b_3 = 1.0, c_3 = 0.25, a_4 = 1.8, b_4 = -1.5, c_4 = 0.0, a_5 = 1.2, b_5 = -0.5, c_5 = 0.1, a_6 = 0.4, b_6 = 0.5, c_6 = 0.15$ ，這些參數決定這六條試題特徵曲線的形狀各不相同。其中，由第一條與第四條曲線的比較，可以顯現出試題難度參數在試題特徵曲線上的位置的重要性來：較困難的試題（如第 1，2，3 題）大多偏向能力量尺的高能力部份，而較簡單的試題（如第 4，5，6 題）則多偏向能力量尺的低能力部份。由第 1，3，4 條與第 2，5，6 條曲線的比較，可以看出試題鑑別度參數對試題特徵曲線的陡度(steepness)的影響力。最後，由第 1 條與第 3 條曲線的比較，*C* 參數對試題特徵曲線的形狀也扮演著決定性的角色；同樣的，試題 3、5 和 6 的下限的比較，也提供我們不少有關 *C* 參數的訊息。

其他常用的模式

除了上述三種基本的試題反應模式外，還有其他適用於非二元化資料的模式。例如：Bock (1972)的名義反應模式(nominal response model)是適用於名義反應資料的試題反應模式。Bock 的模式可用來分析單選題中每個選項被選中之機率；假設每個試題有 *m* 個選項，對每個 θ 而言，選擇這 *m* 個選項之機率和爲 1，這點是本模式的基本假設之一，另一個則是假設每個試題的 *m* 個選項間沒有任何次序大小(ordering)的關係存在。當試題選項只有兩個時，Bock 的模式便簡化成二個參數對數形模式，所以 Bock 的模式是一種通用的模式(general model)。

另一類資料是多元化計分(polytomous scoring)的資料，一如 Bock 的模式所適用的資料，但資料本身多了一項特性：就是試題的選項（或反應）間具有次序大小的關係。適用於這類次序反應(ordered response)資料的模式有 Samejima (1969)的等級反應模式(graded response model)，Andrich (1978a, 1978b, 1978c, 1978d, 1982)的二項式嘗試模式(binomial trials model)和評定量表模式(rating scale model)，以及經 Masters(1982)歸納各種適用於次序反應資料的模式而提出的部份計分模式(partial

credit model)，和經本文作者(Yu, 1991)擴充 Masters 的模式而成的「二個參數部份計分模式」。這類模式可用來作爲分析李克氏量表(Likert scale)所屬各種資料的工具，以改進社會科學研究的測量精確度，對社會及行爲科學，甚至教育研究的量化方法學，具有著實的貢獻潛能。

上述這些模式都是由基本的對數形模式延伸演變而來，由於新的模式還在層出不窮地誕生，本文無法一一詳述，僅挑選基本的三種對數形模式作介紹，其餘可參見 Thissen & Steinberg (1986)的分類說明。

參考書目

Andrich, D. (1978). A binomial latent trait model for the study of Likert-style attitude questionnaires. British Journal of Mathematical and Statistical Psychology, 31, 84-98.

Andrich, D. (1982). An extention of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika, 47, 105-113.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories Psychometrika, 37, 29-51.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Rasch, G. (1980). Probability models for some intelligence and attainment tests. Chicago: The University of Chicago Press (Original edition publised in 1960).

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Yu, M. (1991). A two-parameter partial credit model. Doctoral dissertation of University of Illinois at Urbana-Champaign (unpublished).