試題反應理論的介紹(七)

——訊息函數

英譯名：(Information functions)

余民寧　著

英譯名：(Min-Ning Yu)

我們曾在前幾篇文章裡談到過，試題反應理論與古典測驗理論有兩點不同：一爲參數具有不變性(invariance)，另一爲訊息函數(information function)概念的提出。不變性已經在前幾篇文章裡談論過了，本文即集中在訊息函數的討論上。

基本概念

試題反應理論提出一個能夠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，該方法即需要使用試題訊息函數(item information function)，作爲建立、分析、與診斷測驗的主要參考依據。

試題訊息函數的定義如下：

$$I_i(\theta)=\frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad i=1,\cdots,n$$ （公式一）

其中的符號，$I_i(\theta)$ 代表試題 i 在能力爲 θ 上所提供的訊息，$P_i'(\theta)$ 爲在 θ 點上的 $P_i(\theta)$ 值的導數，而 $P_i(\theta)$ 爲能力 θ 在試題 i 上的試題反應函數，$Q_i(\theta)=1-P_i(\theta)$。試題訊息函數可以應用到前面所談到的一個、二個、與三個參數對數形試題反應模式，這些模式都適合用於二分法計分(dichotomously scored)的測驗資料。例如，以三個參數對數形模式爲例，公式一可以化簡爲(Birnbaum, 1968; Lord, 1980)：

$$I_i(\theta)=\frac{a_i^2(1-C_j)}{[C_j+e^{a_i(\theta-b_i)}][1+e^{-a_i(\theta-b_i)}]^2}$$ （公式二）

從公式二裡，我們很容易便可推知 a，b，和 c 參數在試題訊息函數中所扮演的角色：(1)當 b 值愈接近 θ 時，訊息量較大；反之，b 值愈遠離 θ 時，訊息量則較小；(2)當 a 參數較高時，訊息量也會較大；(3)當 c 參數接近 0 時，訊息量則會增加。

試題訊息函數在測驗的發展與編製上，以及試題好壞的診斷上，扮演著舉足輕重的角色，因爲它能表示出試題對能力估計正確性的貢獻量大小。該貢獻量的大小，端受兩個主要因素的決定：一爲試題的鑑別度參數的大小（亦即，a 值愈大，試題特徵曲線便愈陡，$P_i(\theta)$ 的斜率便愈大，所以訊息量便愈高）；另一爲試題的難度參數，它的位置會決定訊息量的高低。Birnbaum(1968)指出，某個試題所提供的最大訊息量，剛好出現在能力參數爲 $\theta_{\max}$ 的點上，$\theta_{\max}$ 的值爲：

$$\theta_{\max}=b_i+\frac{1}{a_i}\ln[0.5(1+\sqrt{1+8C_i})]$$ （公式三）

如果猜測機率爲最小時（亦即，當 $C_i=0$ 時），則 $\theta_{\max}=b_i$。一般而言，當 $C_i>0$ 時，某個試題在能力水準比其難度值稍高的位置上，所提供的訊息量會達到最大。訊息量最大值所對應的能力水準，即代表該試題所能最精確測量或估計到的能力參數估計值。因此，算出試題的最大訊息量，便可知道該試題所精確測量到的潛在特

質大概是多少，或者是說該試題適合何種潛在特質程度的測量。

在發展測驗或評鑑試題上若使用試題訊息函數的協助，尚需有個基本前提必須先成立，那就是假設我們所選用的試題特徵曲線(ICC)能夠適用於測驗資料。如果這種資料與試題特徵曲線間的適合度很差的話，則我們所計算得到的試題參數估計值和試題訊息函數，將會產生誤導的作用；甚至，當這個適合度尚屬良好時，如果 a 參數很低，且 c 參數很高，則試題的有用性亦會受到限制，它無法通用於所有的測驗中。此外，測驗試題的有用性有時也受到測驗編製者在編製某種具有特殊用途測驗的需求的限制。因此，某個試題在某種能力量尺上也許可以提供相當可觀的訊息量，但在另一種用途的能力量尺上，則無法提供絲毫有價值的訊息量。

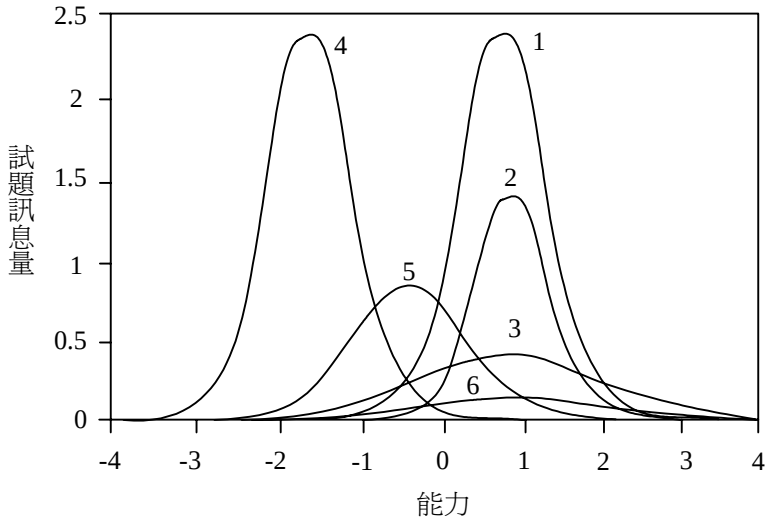
有了上述基本概念後，我們舉六個不同訊息量的試題爲例，說明訊息量所具有的應用涵義。

訊息函數的例圖

試題訊息函數與能力水準二者是組成訊息函數圖的兩個主軸，所畫出的訊息函數就如圖一所示，它是根據表一的試題參數值所畫成的。

表一　六個試題的試題參數值

測驗試題	試題參數		
	b_i	a_i	c_i
1	1.00	1.80	0.00
2	1.00	0.80	0.00
3	1.00	1.80	0.25
4	-1.50	1.80	0.00
5	-0.50	1.20	0.10
6	0.50	0.40	0.15



圖一　表一中六個試題的試題訊息函數

由圖一所示，我們可知它們提供許多寶貴的見解：

1.當 $C > 0$ 時，試題所提供的最大訊息量，大約出現在它的難度水準或比其難度水準稍大的位置（我們只要比較最大訊息量所對應的能力量尺上的位置和表一中的相對應的 b 值便知）。

2.試題的鑑別度參數很顯然地影響試題所提供的訊息量（這點可由比較試題一和試題二的試題訊息函數中得知）。

3.在其他條件均相等的情況下，具有 $C > 0$ 的試題比較不適於用來評定能力水準（這點可由比較試題一和試題三的試題訊息函數中得知）。

4.具有較低鑑別力的試題，在整份測驗中則幾乎不具有任何統計學的用處（如試題六一般）。

5.在評定某些能力水準的範圍內，即使是最具有鑑別力的試題（如試題一和試題四），也會比某些鑑別力較差的試題（如試題五），提供較少的訊息量。例如：在評定具有中等能力的考生能力（即能力水準約在-.50 左右者）時，試題五比試題一和試題四提供較爲有用的訊息；換句話說，對中等能力的考生而言，試題五比試題一和試題四較爲適當且有用。

由上述的五個見解可知，試題訊息函數可以提供我們判斷測驗試題和編製測驗的有效性的一個新方向。

一般而言，$C > 0$ 的試題訊息函數都會比 $C = 0$ 的試題訊息函數還小，在這種情況下，研究者也許會考慮使用一個或二個參數模式，以求合適所使用的測驗資料。結果所得到的試題訊息函數將會比較高些；因此，也唯有在試題特徵曲線能夠適用於所分析的資料時，一個和二個參數的試題訊息曲線才能發揮用處。若試題特徵曲線並無法很適當地適用於所分析的測驗資料，且其相對應的試題訊息曲線也偏離理想的形狀很遠，而我們仍然使用它們時，則我們會獲得具有誤導作用的結果。de Gruijter(1986)便曾舉例說明，在某些情況下，樣本太少時而仍然使用 Rasch 模式，便會產生偏差的結果。

測驗訊息函數

根據 Birnbaum(1968)的推導，一份測驗的訊息函數(test information function)是指它在某一個 θ 值上所提供的訊息量，該訊息量剛好是在 θ 值上的試題訊息函數之總和，記作 $I(\theta)$ ：

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$ （公式四）

由於在 θ 值上的測驗訊息函數是其試題訊息函數之總和，從公式四裡可以看出：每個試題都單獨地對測驗訊息函數作貢獻，因此，每個試題所作的貢獻量大小，並不受在該測驗中其他試題的影響。這個特性是古典測驗理論所沒有的，也正是試題反應理論所具有兩項特點之一。然而，測驗試題對測驗信度和試題鑑別度指標（如：點二系列相關係數）的貢獻，卻受在該測驗中其他試題特性的影響，而無法單獨地決定；因爲在計算這些指標時必須用到測驗分數，而測驗分數卻依所選擇

的測驗試題的不同而不同。甚至，只要改變一個試題，便會對測驗分數產生影響，緊接著，古典的試題和測驗指標也會隨著改變。

在θ值上的測驗訊息量與該能力的估計值的精確性成平方根反比，其符號記作：

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$（公式五）

其中，$SE(\hat{\theta})$稱作估計標準誤(standard error of estimation)。該項指標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。有關這點說明，我們可以由公式五中的定義得知，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。

在試題反應理論的架構裡，$SE(\hat{\theta})$所扮演的角色和古典測驗理論中的測量標準誤(standard error of measurement)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，但古典的測量標準誤對所有能力水準的考生而言，卻都是一致的；換句話說，古典的測量標準誤的意義是認爲每位考生能力估計值的誤差都是一致的，而試題反應理論的估計標準誤則認爲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)。

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的測驗的標準誤還小）。標準誤的大小很快地會趨近於穩定，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的影響，僅會發生小小的作用，典型的例子可以參見 Green, Yen, & Burket(1989)的論文。

<u>相對的效能</u>

有了測驗訊息函數之後，測驗編製學家們往往感興趣的是：比較兩份或多份測量到同樣能力的測驗訊息函數。比較兩份或多份測驗的訊息函數，可以提供測驗專家作測驗評鑑和選擇的參考（參見 Lord (1977)的例子）。所以在發展一份全國性的成就測驗時，往往就需要比較不同測驗的訊息函數，以幫助選擇優良試題來組成所

需的測驗；或者，在編製一份標準化成就測驗時，可參考過去有關學生表現的訊息函數概況，再優先挑選在某段能力範圍內能產生最大訊息量的試題，彙編成我們所需的標準化成就測驗（至於其他因素，如：效度、成本、內容、和測驗長度等，當然也必須在考慮之內）。

比較兩份測驗的訊息函數是這樣進行的：把兩份同樣測得能力估計值爲θ的測驗訊息函數相除，該商值便定義爲某個測驗的相對效能(relative efficiency)：

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}$$ （公式六）

其中，$RE(\theta)$便是相對效能，而$I_A(\theta)$和$I_B(\theta)$則爲定義在一個共同能力量尺θ上的A測驗和B測驗的訊息函數。相對效能的涵義可由下列例子的說明得知：假設$I_A(\theta) = 25.0$，$I_B(\theta) = 20.0$，則代入公式六，得$RE(\theta) = 1.25$，我們可以解釋爲：「在能力水準爲θ時，測驗A所發揮的效能比測驗B所發揮的效能要多（或長）25%，因此，測驗B必須要加長 25%（即把訊息函數相當的試題加入原有的測驗試題中），才能產生與測驗A對θ值一樣的精確測量；或者是，測驗A可以縮短 20%的長度，就可以產生與測驗B對θ值一樣精確的能力估計值。」當然，上述解釋中的加長或縮短測驗長度的作法，都是假設所增減的試題都和原有測驗中的試題，一樣具有可資比較的統計品質（如：類似的難度、鑑別度，產生大約一致的訊息量，都適用於同一範圍程度內的θ值的測量等）。

由上述的舉例說明，訊息函數的應用性非常的廣，我們將在後續文章裡逐一介紹試題和測驗函數，以及相對效能的應用實例，尤其是應用在測驗編製裡。

參考書目

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (chapters 17-20). Reading, MA: Addison-Wesley.

de Gruijter, D. N. M. (1986). Small N does not always justify the Rasch model. Applied Psychological Measurement, 10, 187-194.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2(4), 297-312.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Jaurnal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.