試題反應理論的介紹(八)

——測驗的編製

余民寧

在古典測驗理論下，編製成就或性向測驗的方法，往往僅考慮試題的內容和特徵（如：難度和鑑別度），就當成是選擇試題的依據；例如：先挑選出鑑別度較高（如：大於0.25）的試題，再依據實施測驗的目的和考生的能力分配情況，挑選出難度較適中的試題，編成整份測驗（郭生玉，民79，頁269-272）。

然而，我們也在前幾篇文章裡評論過古典測驗理論所使用指標的缺失，例如：難度和鑑別度都不是不變值(invariant)，它們會隨著考生群體的能力分配的不同，而有不同的估計值出現，這些估計值都是樣本依賴(sample dependent)的估計值。因此，用來決定試題指標的樣本組能否適切代表測驗所要測量的母群體，便成爲決定某種審慎選擇試題的技術能否成功的主要因素。當這個代表性堪疑時，所獲得的試題指標（如：難度和鑑別度）便不太適用於將來所欲測量的母群體。

此外，由於學生生理與心理的成熟與成長，原本在學年度開始所建立的試題指標，到了學年度終了時，便不適用於原本的學生族群，因爲參與測驗的學生的能力分配，經過一學年的期間，已發生明顯的變化，因此導致期初所建立的測驗試題，無法適當地應用到期末的測驗情境中。

另一種情況也會使得古典的試題指標無法適用於未來所欲測量的母群體，那就是來自題庫(item bank)的測驗編製。在發展一套題庫之時，所有要被放入題庫的試題特徵，應該都已經事先被估算出來，並且事先決定好。實際上，這些被稱作「實驗性」的試題，是在被編入一份測驗卷，並對一群受試者施測後，才計算出試題指標估計值的。由於實驗性試題的數目遠比測驗卷數還多，我們只能把它們編成幾份測驗卷，每份均含有不同的實驗性試題和不同的題型，再拿來對不同族群的受試者施測。由於我們無法保證這些接受不同題型測驗的學生，都是能力相等的學生，因此，我們在不同族群受試者下所建立起來的試題指標，彼此間便無法比較。在這種情況下，題庫試題的指標若被假設成是可以比較的，則從該題庫中所建立起來的任何測驗，便無法適合用於某一特定的群體。

除了試題指標本身不具有不變性之外，即使在已有一個編製良好的現成題庫下，古典測驗理論的測驗編製方法，仍有一項很嚴重的缺

失，那就是被選入編成測驗的試題，無法滿足事前訂定的測量精確度的要求。試題對測驗信度的貢獻量，不僅受該試題特徵的影響，同時也受到該試題與其他試題間關聯性的影響。因此，我們無法單獨計算某個試題對測驗信度，甚至對測驗的測量標準誤的貢獻量，而不受其他試題的影響。

爲了彌補古典測驗理論在編製上所面臨的困難和缺失，試題反應理論提出一項比較強而有力的方法來克服這種窘境，那就是運用試題和測驗訊息函數來參與編製測驗的工作。運用試題與測驗訊息函數的最大好處是，它可以挑選出對滿足某份特殊測驗所需的訊息總量最有貢獻的試題，以編製成可以達成測量目標的測驗卷。因爲，訊息量和測驗的精確度息息相關，並且，試題難度指標和學生能力指標又定義在同一量尺上，所以，我們可以在任何能力水準上，挑選出最能精確測量（亦即該測量標準誤最小）到該能力範圍的試題，以編製成我們所需要的測驗。

測驗編製的基本方法

試題反應理論應用到測驗編製上，最常用的工具莫過於使用訊息函數(information function)。根據一般建立題庫的過程，在選定合適的試題反應模式來分析資料後，除了可以獲得試題參數和學生的能力參數估計值外，也可以獲得訊息函數值。利用試題訊息函數，以編製能夠滿足某種特殊需求的測驗編製過程，已由學者(Lord, 1977)提出綱要如下：

1.決定所要的測驗訊息函數的形狀，該形狀的曲線便叫作「目標訊息函數」(target information function)。
2.由題庫中先挑選一組試題，使得這些試題的試題訊息量累加起來的和，能夠填滿目標訊息函數下最難填的部分（通常是訊息函數曲線最突起的部份）。
3.每加入一個試題，便計算現有測驗試題所有的測驗訊息函數。
4.繼續上述的選題步驟，直到測驗訊息函數接近目標訊息函數到達某種令人滿意的程度爲止。

上述這些測驗編製的步驟，通常需要仰賴大電腦和測驗編製專家的共同合作，否則光靠筆算會費時、費力。

由已知（或現成）的試題反應模式下所建立起來的題庫中，我們可以根據Lord(1977)所提出的綱要，編製出可以在某個能力範圍內充分發揮鑑別功能的測驗來；也就是說，假設我們已知某組受試者們的能力水準，我們便可以挑選出能夠使該能力範圍內的測驗訊息量達到最大的測驗試題來，以作爲測量該等能力水準的工具。這種挑選測驗試題的作法，將能增進對能力參數估計值的精確性。

舉例來說，根據Lord(1977)所提的綱要，一個涵蓋範圍較廣的能力測驗(broad-range ability test)，其目標訊息函數應該是個相當平坦的曲線，它所表示的涵意是，在整個能力量尺上，該測驗希望能夠提供幾乎是同樣精確的能力估計值，以表明它所能適用的能力範圍較爲寬廣。而對一個設有切割分數(cut-off score)以區別精熟者(masters)和非精熟者(nonmasters)的效標參照測驗(criterion-referenced test)而言，其所期望獲得的目標訊息函數，應該是個對應於能力量尺上的切割分數附近，呈現極爲尖狹峯分配的曲線，這種情況顯示出，在切割分數附近，該測驗最能夠精確測量到區分精熟與非精熟二者的能力估計值。

透過試題訊息函數的使用，測驗編製者可

以編製出滿足各種特殊需求的測驗來。例如，Yen(1983)便曾舉例說明，如何運用試題訊息函數來編製一份大規模的測驗。van der Linden & Boekkooi-Timminga(1989)也已發展出一套程序，說明在測驗上加諸一些限制，以確保內容效度、適當的測驗長度、和其他特徵之後，可以自動挑選測驗試題以符合某種測驗訊息函數的作法。

爲了說明上述的過程起見，茲舉一份成就測驗爲例。就成就測驗而言，前測時的表現情形往往遠低於後測的表現情形，這是一種很常見的現象。有鑑於此，測驗編製者便可以挑選較爲簡單的試題作爲前測的內容，而挑選較爲困難的試題作爲後測的內涵。在每一個測驗情境裡，考生能力範圍所最常出現的地方，其測量的精確性往往會達到最大。甚至於，由於在這兩份測驗上的試題，都是在測量相同的能力，並且，能力估計值也不受特別挑選的試題群影響，因此，後測的能力估計值減去前測的能力估計值，其差值便可以用來測量成長(growth)量的大小。

de Gruijter & Hambleton(1983)和Hambleton & de Gruijter(1983)已著手研究，在測驗編製之前便先決定好切割分數或測驗的通過標準，看看最理想的試題挑選方法，會對一份測驗的決策正確性產生什麼樣的影響。爲了解釋這項結果，通常是以隨機的方式來挑選試題，以編製成所需要的測驗。在效標參照測驗的編製過程中，從一堆現成的候選測驗試題庫裡，以隨機方式挑選試題以組成測驗，是一種常用的作法。只不過是，依隨機方式所挑選出的測驗試題所組成的測驗，其錯誤率(error rates)（亦即是造成分類錯誤的可能機率）幾乎是依最理想方式來挑選測驗試題以組成測驗所造成的錯誤率的兩倍。因此，以試題反應理論爲架構，來挑選最理想的測驗試題的作法是有可能的，因爲試題、學生、和切割分數都是建立在同一量尺的基礎上，所以方便測驗的編製與測驗結果的解釋。

其實，設定目標訊息函數和挑選試題的程序，仍存在有許多值得商榷的問題。其中一個便是，單依靠統計學的效標來挑選試題的作法，並沒有辦法保證就可以編製出一個內容有效(content-valid)的測驗來。可惜，我們通常卻過度强調統計學的效標，而忽略試題內容在測驗編製上所扮演的重要角色。忽視內容的考慮事項，往往會導致編製出一個缺乏內容效度的測驗來。爲了解決這個難題，van der Linden & Boekkooi-Timminga(1989)使用線性規劃(linear programming)的技術，提出許多同時考慮試題內容和統計學的效標等有用的組合方法，以作爲挑選試題的參考依據。

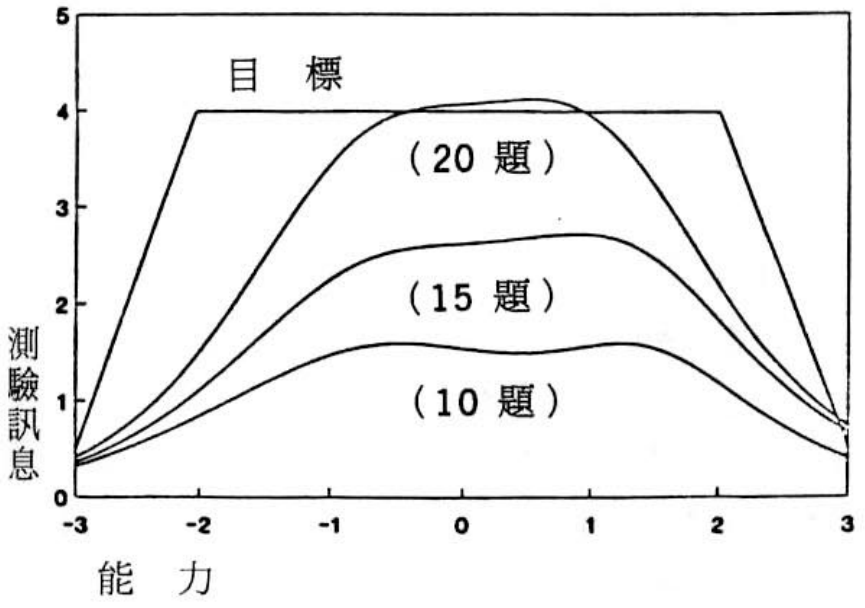
使用試題訊息函數來作爲測驗編製的依據，還有另一項問題會產生，那就是很可能高估高鑑別度（即a）值，以致於訊息函數也許會產生偏差。使用具有高鑑別度的試題所編製出的測驗，很可能會與期望中的測驗相去甚遠。由於測驗訊息函數將會被高估，所以增加額外的幾個試題到測驗裡，也許會緩和高估的情形。而最好的解決辦法，還是儘量使用大樣本，以確保試題參數的估計值都很正確、很穩定。

下列所舉的例子，是說明如何運用訊息函數來編製特殊測量目的的測驗。

廣泛能力測驗的編製

假設某位測驗專家想編製一份包含廣泛能力的測驗，他認爲該能力範圍應涵蓋(-2.00, 2.00)之間，並且只容許有0.50以下的估計標準誤存在，而在此能力範圍外者則允許

有少許較大的誤差存在。例如，假設選定SE(θ)＝0.50，則I(θ)＝4.0，典型的目標訊息函數可如圖一所示建立起來。爲了編製一份能夠滿足此目標，並且具有愈少試題愈好的測驗，我們就必須從具有難度值介於 -2.00 和 2.00 之間、高鑑別度和低猜測度的試題群中，去挑選符合要求的候選試題。圖一所示，即爲在既定的目標訊息函數（即θ值介於±2.0之間，且I(θ)＝4.0，呈現平坦的曲線）下，從題庫中挑選出最理想的 10、15和20 題測驗試題後，所計算出的測驗訊息函數。很明顯的可以從圖一看出，20題下的測驗，最爲接近我們想要編製的目標測驗。若增加難度值接近±2.0的試題，則所獲得的測驗訊息函數更加接近目標訊息函數。

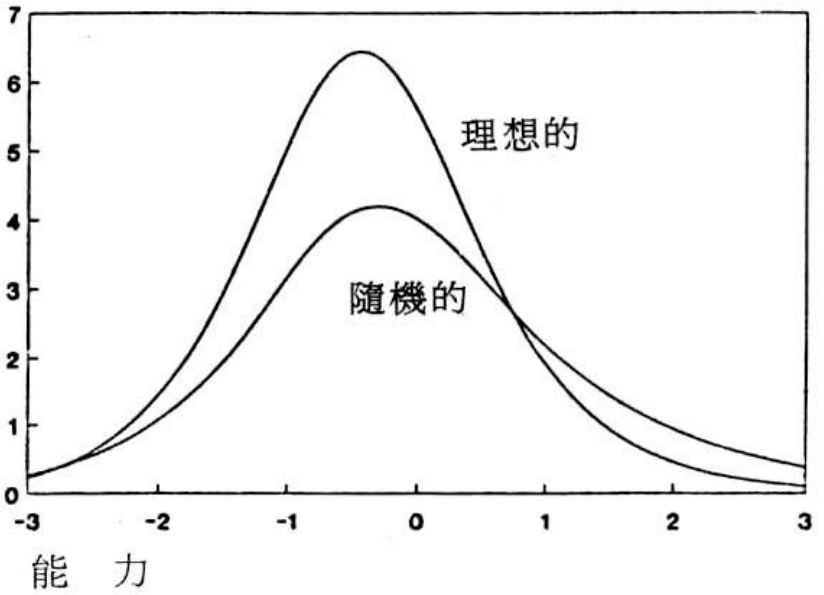


圖一　含有 10、15和20 題試題測驗的測驗訊息函數

效標參照測驗的編製

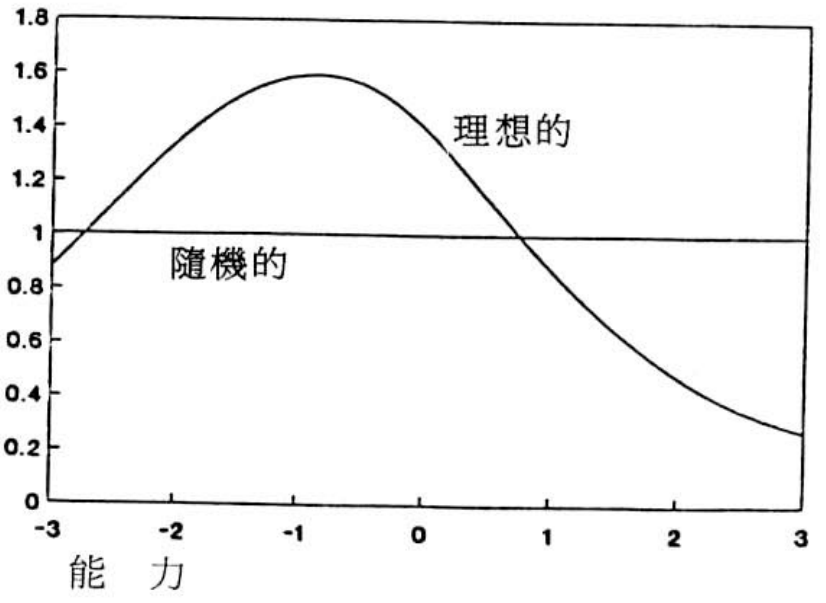
假設某位測驗專家想要編製一份含有15個試題的效標參照測驗，使得其測驗訊息數在切割分數θ＝-0.50處達到最大。爲了比較起見，也以一般常用的方法隨機抽取15個試題編成測驗（稱作標準測驗），並計算出其應有的測驗訊息函數。茲將這兩種不同方式挑選試題編製成的測驗訊息函數，畫於圖二裡，以資比較。

測 驗 訊 息



圖二　理想的和隨機的挑選方法下15個試題的測驗訊息函數

相 對 效 能

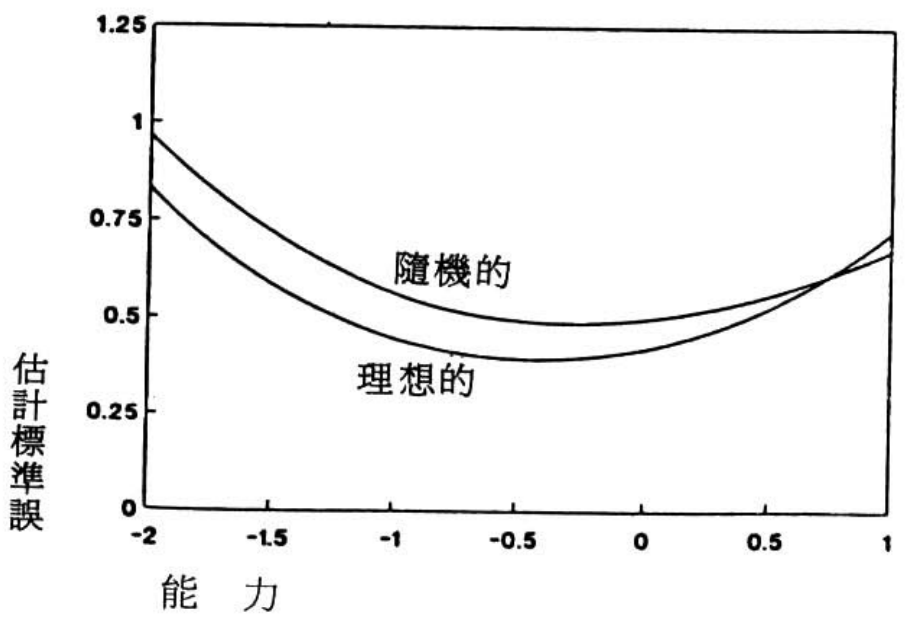


圖三　理想的對隨機的15題試題測驗的相對效能

圖三所示，爲理想的測驗對標準測驗之相對效能圖。很明顯的，理想的測驗在切割分數（即 θ = -0.50）處，提供較大的測量精確性；它比標準測驗在此處高出60%的相對效能，也就是說，標準測驗的長度必須從15題增加到24題，才能發揮與理想的測驗同等的效能。

由圖二和圖三可以看出，對高能力學生而言，理想的測驗表現得不如標準測驗表現的好。這是由於理想的測驗僅包含能夠在切割分數附近發揮鑑別功能的試題，而忽略許多適合於高能力學生的試題的緣故。由此可見，標準測驗包含比較多的異質試題在內。

實際說來，題庫中的試題愈異質化，或所欲編製之測驗長度佔題庫大小的比率愈小，則理想的試題挑選方法遠比隨機的試題挑選方法爲優。相對於這兩種挑選方法下的測驗訊息函數的估計標準誤，則如圖四所示。由圖四可知，理想的測驗的估計標準誤比隨機的測驗的估計標準誤還小。



圖四　理想的與隨機的挑選方法下15個試題的估計標準誤

參考書目

郭生玉（民79）。心理與教育測驗（五版）。台北：精華。

de Gruijter, D.N.M., & Hambleton, R.K.(1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), Applications of item response theory (pp.142-154). Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K., & de Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educcational Measurement, 20, 355-367.

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

van der Linden, W.J., & Boekkooi-Timminga, E.(1989). A maximum model for test design with practical constraints. Psychometrika, 54, 237-247.

Yen, M.W.(1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton(Ed.), Applications of item response theory (pp.123-141). Vancouver, BC: Educational Research Institute of British Columbia.

（作者：政大教授兼附小校長）