試題反應理論的介紹（十二）
電腦化適性測驗

余民寧

前文談到測驗題庫的建立，題庫的另一個優點，就是爲電腦化適性測驗（Computerized Adaptive Testing, CAT）作準備，不僅可以節省施測時間，更可以達到精確估計考生能力或某種潛在特質的目的。

從前面幾篇文章裡，我們也大致可以知道：當測驗的難度能夠適合考生的能力程度時，這時測驗所測量到的考生能力最爲精確。因此，我們可以知道任何一次施測的結果，都無法針對每位考生提供最精確的能力測量，因爲該測驗的難度無法適合每位考生能力程度的需求。最理想的施測狀況是：能夠針對每位考生不同的能力程度，來提供適合個別情況的測驗方式，這也就是電腦化適性測驗所欲探討的內容。

最早應用適性測驗（即因才施測式的測驗方式）的例子，是1908年Binet所作的智力測驗的研究（Weiss，1985）。後來一度中斷好久，直到1960年代末期，才由在教育測驗服務社（Educational Testing Service）的F.Lord從事較爲完整的通盤研究（參見 Lord（1980）的作品）。由於Lord感覺到，對於低能力與高能力的考生而言，固定長度的測驗無法有效的滿足這些考生能力估計的需求，因此才極力投入適性化測驗的研究。Lord認爲如果被挑選用來施測的試題都能針對每位考生能力提供最大的參考訊息的話，則縮短測驗的長度（即減少施測的題數），應該不會降低對每位考生能力的精確測量。理論上而言，每位考生所接受的施測試題，應該都是不同的試題組，因此，適性測驗的實施是可能的。

但是要實施適性測驗，也唯有在電腦誕生發明後，才有可能施行。電腦科技的發達，日新月異，它的超大容量可以貯存測驗訊息（如：測驗試題及其特徵指標）、編製、施測和記錄測驗分數，因此使得推行適性測驗變得愈來愈可行（Bunderson , Inouye, & Olsen, 1989; Wainer, 1990）。在1960年代末，美國陸軍總署、人事管理局、及其他聯邦機構，均大力支持贊助有關適性測驗的研究，舉辦特殊的專題研討會，並且有數百篇的有關研究論文發表出來，後來都被收集出版成冊（例如：Wainer, 1990; Weiss, 1983）。

在電腦化適性測驗（CAT）裡，呈現給考生的試題順序，是依據考生在前一個試題上的表現好壞來作決定的。根據考生先前的表現好壞，下一個要呈現給考生作答的試題，便是對考生能力估計精確性最有貢獻的最大訊息量的試題。如此一來，測驗的長度便可以縮短，並且也不會犧牲任何的測量精確性；因爲對於高能力的學生，可以不必給他相當容易的試題作答，對於低能力的學生，也可以不必給他極度困難的試題作答，因爲這些試題對他們的能力水準的估計而言，只能提供極爲有限或絲毫沒有幫助的訊息。因此，實施電腦化適性測驗，不僅可以做到因才施測的精確估計考生能力的地步，也可以節省許多施測時間和成本，可說是至少一舉兩得。

在開始進行電腦化適性測驗之時，先由電腦終端機隨機呈現一組測驗試題（也許是兩題或三題），在考生作出反應後，電腦便根據這些反應資料，估計出考生的初步能力估計值（initial ability estimate）；然後，電腦會根據這些初步能力估計值，從現有的題庫（儲存在電腦的內部）中挑選出最能對能力水準的估計發揮最大貢獻力量的試題（通常這些試題的訊息量也是最大），再呈現這些試題給考生作答；這種施測過程一直繼續下去，直到事先預定的施測題數已測完，或某科預定的能力估計值的測量精確性（即標準誤）已獲得爲止。

試題反應理論的新天地

我們已知難易適中的試題，對估計考生能力的精確性最爲有效。而通常的一份測驗卷試題難度，很難滿足或適合每位考生的能力水準，因此要能做到試題難度隨考生能力不同（即個別差異）而調整的測驗方式，唯有採行適性測驗。而最適合在適性測驗中作應用的，便是試題反應理論（IRT）。由於試題反應理論中的試題反應模式，可以獲得不受不同施測試題影響的能力估計值（即具有試題獨立（參見本系列文章之二和之三）的估計特性），也就是說不同的考生考不同的試題，只要試題性質相同，不同能力考生的能力估計值可以被精確的估計出來，因此可以互相比較。事實上，也唯有試題反應理論才適合應用在適性測驗裡。

在應用試題反應理論到實際的測量情境時，必須先滿足該測驗只具有單一主要的因素的基本假設（參見本系列文章之二），這個基本假設在目前所使用的適性測驗裡，一般都能夠獲得滿足。目前，最適合應用到適性測驗上的試題反應模式，是三個參數對數型模式（即3PL）（Green, Bock, Humphreys, Linn, & Rechase, 1984; Weiss, 1983），最主要的原因是它比一個與二個參數對數型模式，更適合用在選擇題的試題資料上。

在適性測驗裡，測驗訊息函數也扮演著很重要的角色（參見本系列文章之七）。其中，能對測量精確性發揮最大貢獻力量的試題，會被優先挑選做施測的試題，呈現給考生施測。一般而言，能讓考生大約有50%或60%答對機率的試題，通常都是屬於能夠提供最大訊息量的試題。

適性測驗的基本方法

以試題反應理論爲架構下的適性測驗，有個基本目的，那就是要撮合測驗試題的難度和待測量的考生能力水準。爲了達成這項目的，我們必須擁有已知每個試題特徵的龐大試題庫，以便從中挑選出適當的試題（Millman & Arter, 1984）。根據Lord（1980）的看法，我們必須設計電腦程式，以便完成下列的目的，達到適性測驗的目標：

(1)根據考生先前的反應表現，預測他在尚未接受測驗的試題上的種種可能反應。

(2)根據上述的理解，有效地挑選試題，接著呈現給考生作答。

(3)最後在測驗完畢時，能夠計分，以分數來表示考生能力的大小。

適性測驗的基本方法，包括下列六個步驟。茲簡述各個步驟如下：

1.試題反應模式的挑選：針對不同類型資料和研究問題的瞭解，挑選適當的一個、二個、或三個參數對數型模式，作爲進行適性測驗的最基本模式根據。當然，三個參數對數型模式是最常被選用的模式。

2.題庫的準備：參考本系列文章之十一的說明。

3.測驗的起點：應該先考那一個試題，是適性測驗所需面臨的一件重要抉擇問題。

從理論上來看，試題的難度必須要能夠配合考生的能力水準。但是，除非我們已知考生過去的表現好壞，否則無法在施測之前就知道考生的能力。所以，常用的起點方法有：(1)自難度適中的試題中隨機抽取一個試題；(2)完全隨機抽取一個試題；(3)先調查學生的背景，然後再決定出那一類的試題。Lord（1977）認爲，只要測驗的題數不少於25題的話，以那一個試題做爲起點的影響不大。

4.選擇方式：使用試題反應理論作爲適性測驗的理論基礎，必須有根據某種理論建立的題庫存在，以方便經過校準過後的試題參數特徵，也都能儲存在題庫裡。校準時所選用的模式不同，也會影響到計分方法的選擇和能力的估計。一般而言，常用的試題選擇的方法有三種：(1)挑選能對考生能力估計提供最大訊息量的試題，爲了避免同樣的試題一再地被重覆選用，Green等人（1984）建議可從一堆產生最大訊息量的試題中，隨機抽取一個試題來進行就可以。(2)利用貝氏試題選擇法，將考生能力分配看成是某種事先分配（prior distribution），計算考生答對或答錯未用到的試題之事後變異數，再挑選能夠使這種考生能力事後分配之變異數變爲最小的試題，作爲施測的試題。使用貝氏的選題方法，頗受事前分配之假設的影響很大，但是只要施測的試題很多的話，這種影響是可以被排除的。(3)挑選難度最接進考生現階段能力估計之試題。

5.計分方法：其實也就是學生的能力估計方法，唯一不同的是，在適性測驗裡，考生每答對一個試題，就得重新估計一次考生的能力估計值。最大近似值估計法和貝氏估計法是適性測驗裡常用的兩種能力估計方法。最大近似值估計法的估計效能很好，但遇到題數少或估計值無法收斂時，都會產生很大的問題；貝氏估計法雖能克服這些困難，但對事前分配的假設如果不當的話，卻會產生有所偏差的能力估計值。

6.終止的標準：終止適性測驗的方法，和前述的選題與計分方法間有很密切的關聯。若以試題最大訊息量作爲選題標準的話，只要累積已測過之試題的訊息量，到達某種事先預定的標準後，便可終止。若以貝氏估計法來選題的話，則可以估計能力之變異數小到某個預定的標準時，便可終止施測。此外，如果前述兩種標準均很慢才達到的話，也可以預設施測試題數的上限，只要題數一測完，即使尚未達到預定的標準，也可以終止施測，以避免漫無止境地繼續下去，浪費考生的許多寶貴時間。

適性測驗的例子

假設從一已知的題庫（如：表1所示）中，進行適性測驗（事實上的題庫應有數百題，在此所列舉者，僅作爲例子說明用），則下列的步驟是電腦化適性測驗中會出現的事件：

(1)假設先挑選試題 3 ；因爲它具有平均難度值和最高的鑑別度值。又假設某考生答對，但此時的最大近似值估計法無法進行能力估計，必須等到至少有一題答對和一題答錯才行（全錯或全對的得分，會導致－∞和＋∞的能力估計值）。

表 1　假想的題庫試題

試　題	b	a	c
1	0.09	1.11	0.22
2	0.47	1.21	0.24
3	- 0.55	1.78	0.22

4	1.01	1.39	0.08
5	- 1.88	1.22	0.07
6	- 0.82	1.52	0.09
7	1.77	1.49	0.02
8	1.92	0.71	0.19
9	0.69	1.41	0.13
10	- 0.28	0.98	0.01
11	1.47	1.59	0.04
12	0.23	0.72	0.02
13	1.21	0.58	0.17

(2)假設其次挑選試題12，因爲它比前一個試題較難。又假設該考生答對。至此，最大近似值估計法仍無法進行能力估計。

(3)其次選中試題 7 ，因爲它比前二題較難。假設該考生答錯本題。該考生在三個試題上的反應祖型爲（1, 1, 0），利用這三個試題的已知特徵和最大近似值估計法，估計出該考生能力估計值爲 $\hat{\theta}=1.03$；這三個試題的測驗訊息量爲 $I(\hat{\theta})=0.97$，估計標準誤爲 $SE(\hat{\theta})=1.02$ ，如表 2 所示。

表 2　每一階段電腦化適性測驗後的能力估計值和估計標準誤

階段	試題編號	試題反應	$\hat{\theta}$	$I(\hat{\theta})$	$SE(\hat{\theta})$
1	3	1	—	—	—
2	12	1	—	—	—
3	7	0	1.03	0.97	1.02
4	4	1	1.46	2.35	0.65
5	11	0	1.13	3.55	0.55
6	9	1	1.24	4.61	0.47
7	2	1	1.29	5.05	0.45
8	1	1	1.31	5.27	0.44
9	8	0	1.25	5.47	0.43

(4)接著，計算出當 $\hat{\theta}=1.03$時，題庫中剩餘試題所提供的訊息量，如表 3 所示。由於試題 4 在 $\hat{\theta}=1.03$時所提供的訊息量最大，所以，它是下一個被挑選中的試題。假設該考生答對本題，接著估計出反應組型爲（1, 1, 0, 1），此時新能力估計值 $\hat{\theta}=1.46$，估計標準誤 $SE(\hat{\theta})=0.65$，如表 3 所示。

表 3　每一電腦化適性測驗階段中剩餘試題所提供的訊息量

$\hat{\theta}$	試題所提供之訊息量												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1.03	.034	.547	—	1.192	.010	.051	—	.143	1.008	.251	1.101	—	.166
1.46	.179	.319	—	—	.044	.017	—	.205	.579	.136	1.683	—	.175
1.13	.292	.494	—	—	.008	.039	—	.159	.917	.219	—	—	.170
1.24	.249	.433	—	—	.006	.029	—	.175	—	.187	—	—	.173
1.29	.232	—	—	—	.006	.026	—	.182	—	.175	—	—	.174
1.31	—	—	—	—	.005	.024	—	.186	—	.168	—	—	.174
1.25	—	—	—	—	.006	.028	—	—	—	.184	—	—	.173

(5)接下來，計算出$\hat{\theta}$ =1.46時，剩餘試題所提供的訊息量。然後，挑選出最大訊息量的試題。之後，再施測，重新估計能力，計算剩餘試題所提供之訊息量，再挑選最大訊息量的試題，如此繼續下去（如表 3 所示，接下來被選中的試題，依序爲試題11、9、2、1、和最後的試題8 ），一直到考生能力估計值的估計標準誤之遞減值小於事先預定的標準（如：小於 .01）。從表 2 可知，第九個階段施測試題8之後，它從第八個階段所遞減的標準誤值爲0.01，因此，施測的過程到此爲止。此時，該考生的能力估計值爲$\hat{\theta}$ =1.25。這個估計值便是我們從題庫中挑選 9 個試題進行適性測驗後，所精確估計出該考生的能力水準。

由本例可知，實施適性測驗將具有下列的幾項優點：

(1)加强測驗的安全性；
(2)依據需求來進行施測；
(3)無需使用答案紙；
(4)適合每位考生的作答速度；
(5)立即的計分和報告成績；
(6)降低某些考生的考試挫折感；
(7)加强施測的標準化過程；
(8)容易從題庫中找出並刪除不良的試題；
(9)對於試題類型的選擇更具彈性；
(10)減少監試的時間。

參考書目

1.Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed.)(pp. 367-407). New York: Macmillan.

2.Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. Journal of Educational Measurement, 21, 347-360.

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

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

5.Millman, J., & Arter, J. A.(1984). Issues in item banking. Journal of Educational Measurement, 21, 315-330.

6.Wainer, J. et al. (Eds.) (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

7.Weiss, D. J. (Ed.) (1983). New horizons in testing. New York: Academic Press.

8.Weiss, D. J. (1985). Adaptive testing by computer. Journal of Consulting and clinical Psychology, 53, 774-789.

(作者：國立政治大學教授並附小校長)