試題反應理論的介紹(四)

——精熟測驗

余民寧

近二十年來，在心裡與教育測驗領域裡有個重大的改變，那就是效標參照測驗 (criterion referenced tests) 逐漸受到重視，並廣爲流傳使用。至今，效標參照測驗的用途很廣，(1)在軍隊裡，它可被用來評量軍人的基本能力；(2)在工業界，它可被用來評定員工工作技能的純熟度，或評鑑在職訓練課程的好壞；(3)在證照考試上，它可被用在各行各業中，以區分出誰是「精熟者」、誰是「非精熟者」；(4)在學校教育中，它可被用來評量學生在某種知識技能上的表現程度。

由於效標參標測驗的名辭定義很多 (如 Gray(1978) 說有 57 種之多)，很難予以統一，不過，目前比較被一致接受的定義是：「效標參照測驗是指被用來確定個人在某個界定清楚的行爲領域中表現程度的測驗」(Popham,1978,p.93)。它又有幾個常見的同義詞，如：精熟測驗 (mastery test)、領域參照測驗 (domain-referenced test)、能力測驗 (competency test)、或基礎技能測驗 (basic skills test) 等，本文援用 Lord(1980) 的說法，以「精熟測驗」一詞來顯現試題反應理論 (IRT) 在這方面的應用，並用以區別與古典測驗理論所指稱的「效標參照測驗」的不同，以及介紹精熟測驗的整個過程和應用。

精熟測驗的設計和編製的步驟

精熟測驗的內涵，主要可以分成幾個重點：(1)測驗試題的設計與選擇，(2)測驗的計分與報告方式，(3)測驗的長度與精熟標準的決定。茲將這整個設計和編製的步驟，條述如下 (Hambleton & Zaal,1991,pp.10-11)：

1. 初步的考慮事項

a. 說明測驗的目的。

b. 說明該測驗所欲測量的目標。

c. 說明受試者的特性及特殊的施測設備。

d. 初步決定試題格式 (如：客觀測驗的試題或實作導向的試題)。

e. 決定編製測驗所需的時間和成本。

f. 慎選合格且適當的命題委員 (如：考慮他

個人的專長或發展測驗中所扮演的角色的重要性)。

g.說明初步的測驗長度(如：要測量那些目標、需要多少題數、及施測時間多長)。

2.審視測量的目標

a.審視測量目標的陳述是否明確清楚，目標的適當性可否被接受。

b.選擇測驗所欲測量到的一組目標。

c.針對每項目標描述所需試題的特徵，並且審視這些特徵的完整性、正確性、明確性、和實用性。

3.撰寫試題

a.撰寫大量試題，以作爲預試之用。

b.輸入電腦化題庫，以便利修改和存取(參見本系列論文之(十)——題庫的建立)。

c.進行試題初步編輯工作。

4.評量內容效度

a.延聘一批課程、學科、與測量專家。

b.請這批專家評閱這些試題是否符合它們所欲測量的目標、是否具有教材內容的代表性、以及是否不受刻板印象的影響。

c.審視這些試題,以判定其技術上的適切性。

5.修改試題

a.有必要時，根據上述 4b 到 4c 的步驟，修改試題或刪除不適當的試題。

b.如有需要，撰寫補充的試題，並重覆上述第 4 個步驟。

6.初步預試

a.編輯試題成試卷的形式，以便進行預試。

b.針對一群適當的考生施測。

c.進行試題校準和試題偏差的診斷(參見本系列論文之(十)——試題偏差的診斷)。

7.測驗試題再修改

a.根據 6c 的步驟，如有必要，需對試題加以修改或刪除。

8.組合成正式的測驗

a.決定測驗的長度。所需的題型數目、及每個目標需多少試題數。

b.從上述有效的候試題中(多半是由題庫中)，挑選所需要的適當試題。

c.準備測驗指導語、練習用的試題、測驗題本、計分卡、答案紙……等。

d.補充說明指導語不清楚的地方、考生作答有那些注意事項、特殊考生(如：殘障考生)所需的作答時間等。

9.設計精熟的標準

a.決定考生表現程度的描述或精熟程度的設定，是否能夠符合測驗的目的(如果該描述是主要的用途的話，則跳到第 10 個步驟)。

b.說明設定區分爲「精熟」與「非精熟」之標準的挑選過程；如果必要的話，設定一個以上的標準(如：分成「卓越」、「優良」、「尚可」等)。

c.說明殘障考生所適用的特殊標準。

d.說明需要重測的考生的另一種計分方式。

10.正式預試

a.設計施測的方式，以便收集測驗分數的信度與效度等方面的訊息。

b.對挑選出的一群適當的考生進行施測。

c.評估那些爲了符合特殊需求而改變之施測過程所可能造成對測驗的信度和效度估計的影響。

d.評量施測程序、測驗試題、及測驗分數的

信度和效度。

e.根據上述所獲得的技術性資料，進行最後的修改。

11.準備使用手冊。

a.準備一份施測者或監考者須知手冊。

b.準備一份技術性使用手冊。

12.收集額外技術性資料

a.進行信度和效度的分析研究。

測驗的長度

有關精熟測驗應該具備多少試題才算適當的研究，一直是個很數量化的研究課題，所累積的研究文獻也很多 (Hambleton,1984)。本文僅列舉較具實用性的一種如下。

當我們確定測驗分數的用途，也對其用法加以描述後，有關考生得分的專精分數估計值 (又稱作「領域分數」(domain score) 的測量精確度，可用公式表示如下：

$$(\text{精確度})^2 = \frac{\hat{\pi}\ (1-\hat{\pi}\ )}{n} \quad (\text{公式 1})$$

其中，專精分數估計值 π 可以表示如下：

$$\pi = \frac{1}{n} \sum_{i=1}^{8} P_i(\theta) \quad (\text{公式 2})$$

$P_i(\theta)$ 爲具有能力估計值爲 θ 的考生在試題 i 上答對的機率；所有測驗試題答對機率之和，即爲該考生的眞實分數 (True score)。因此，所謂的專精分數即是眞實分數的平均數；亦即是考生答對某種目標領域的測驗內所有試題的機率。該值爲一比率分數，其值域介於 0 與 1 之間；其值愈接近 1，表示該考生的精熟程度愈大，反之，該值愈接近 0，則表示該考生在測驗上的表現反應愈不精熟。由公式 1 可以推論出適當的測驗題數應該是：

$$n = \frac{\pi\ (1-\pi\ )}{(\text{精確度})^2} \quad (\text{公式 3})$$

例如，假設已知某群考生的專精分數爲 .80，且我們希望該專精分數估計值的精確度至少能夠達到 .10 的話，則將此二數值代入公式 3，可以獲得：

$$n = \frac{.80(1-.80)}{(.10)^2} = 16$$

換句話說，我們若想要使某群考生在某個測驗上的專精分數達到 .80，且其估計值的精確度也達 .10 的水準的話，則我們必需要編製出一份至少含有 16 個試題的測驗，才能符合我們所需要的測驗目的的要求。由此可見，要編製出一份達到某種測驗目的的精熟測驗，其題數的多寡完全取決於專精分數和估計精確度兩個因素：專精分數愈接近 50，所需之題數則愈多。專精分數愈接近於兩極端 (即 0 或 1)，則所需題數愈少。若所要求的精確度值愈大，而所需的題數愈少。若所要求的精確度值愈小，則所需的題數便需要愈多。如果精熟測驗分數的目的是用來區分「精熟者」與「非精熟者」的話，則可用來幫助決定題數多寡的參考依據就更多。Millman(1973) 和 Wilcox(1976) 提供了許多對照表，可用來幫助決定適當的測驗長度、專精分數、通過分數 (passing score) 或標準設定 (standard setting) 等問題。

精熟標準的決定

在精熟測驗中，有關通過分數等標準設定問題之研究文獻，可說是已經到了汗牛充棟的地步。根據數位學者 (Berk, 1984, 1986; Ham-

bleton, 1990; Hambleton & Zaal, 1991)的歸納，有關標準設定之研究方法，大致可以歸納成三大類，大類內各有數種較有名的方法：

一、判斷的方法 (judgmental methods)

這種方法主要是聘請專家評審每一個試題，以判別出最低能力考生所可能表現到什麼樣的最佳程度。這類設定通過標準的方法，有三種較爲常用的方法較有名，分別是：

㈠ Nedelsky 法：首先，請個別的專家找出最低能力考生在選擇題的誘答選項中，能夠刪除（或以消去法消除）的選項數目。因此，該試題的最低通過標準即訂定爲剩餘未被刪除之誘答選項的數目之倒數。即爲最低能力考生在該試題上的「猜測分數」(chance score)。再將每個試題之最低通過標準（即猜測分數）加總起來，便得此一測驗的通過標準。若有數位專家進行判斷，則以個別之通過標準之和的平均數，作爲該測驗之通過標準。

㈡ Ebel 法：根據試題的相關性和難度兩個向度，請專家進行評定。其中，相關性分成四個水準，難度分成三個水準，共形成 4 × 3 的列聯表，再請專家就每一細格中，最低能力考生所可能答對之百分比進行評定。再將數位專家評定一致之細格題數加總除以總題數，便得此一測驗之通過標準。

㈢ Angoff 法：請專家就每一個試題中，最低能力考生所可能答對之機率，進行評定。將每題可能答對之機率加總，便成爲該專家所判斷的通過標準。再將數位專家之判斷的通過標準加以平均，便成爲該測驗之最後的通過標準。

二、實徵的方法 (empirical methods)

這種方法是以考生實際的作答資料的分析結果，作爲設定通過標準之依據。又可分成：

㈠ Livingston 法：從外在選擇一個效標 (criterion)，並建立一條直線的效用函數 (linear utility function)，以決策理論的方法找出能夠使該效用函數達到極大的分數切割點 (cutoff score)，便是該測驗的通過標準。

㈡ Linden & Mellenbergh 法：找出能夠使「期望的損失」(expected losses)（即分類錯誤的代價）達到最小的分數切割點，便是該測驗所需之通過標準。若此點找出後，將使效標分數大於此點以上者（即被判定爲精熟者）能夠通過測驗；效標分數小於此點者（即非精熟者）無法通過該測驗。

三、混合的方法 (combination methods)

這種方法乃揉和上述兩個方法，用來設定通過標準的一種過程。又分爲：

㈠邊緣組法 (borderline-group method)：首先要求專家對每一教材內容的最低可接受的表現程度作一定義，,再列舉一批表現水準接近此一劃分爲精熟與非精熟的邊緣線的考生，然後編輯測驗對此批考生施測，取其得分之中位數 (median)，作爲該測驗之通過標準。

㈡對照組法 (contrasting-group method)：首先要求一批專家定義精熟某教材內容的最低可接受的表現程度，再找出他們已確知某些已達精熟和未精熟的考生。針對此二組考生施測，並將此二組考生的得分分配曲線，一一畫在每個目標範疇圖上，取其兩線的交叉點作爲起始的通過標準 (initial standard)。然後，再漸次調整該交叉點，使分類錯誤率達到最小的位置爲止，此時的決定點即爲最後的通過標準。

總之，標準的設定終究還是屬於判斷的歷程，需要參與者(1)熟悉教材內容和各種設定標準的方法，(2)有評定試題表現和測驗分數分配曲線的能力和經驗，(3)以及瞭解使用該測驗的社會與政治背景。如此才能有個良好、公正的標準誕生 (Hambleton & Powell, 1983)。

精熟測驗的未來發展方向

精熟測驗的編製技術、應用、與改進，均已日臻成熟的地步。目前它仍在研發的領域有：

(1)標準設定的方法，(2)改進測驗分數使用效果的分數報告格式，(3)以及描述目標的方法。未來，尚可改進精熟測驗的實用性和提高未來的運用潛力的方向，計有(1)配合微電腦的使用，研究如何存取、施測、和計分，(2)配合試題反應模式，研究發展目標、測驗試題、和考生可以隨時參考使用的各種繼續成長或發展的量表。有關精熟測驗的編製、應用、和發展方面的教科書和使用手冊，讀者可參閱 Berk(1984)、 Hambleton(1990)、 和 Popham(1987) 等人的專著。

參書書目

Berk, R. A. (1984). A guide to criterion-referenced test constrution. Baltimore, MD: The Johns Hopkins Univerity Press.

Berk, R. A. (1986). A criterion-referenced tests. Review of Educational Research, 56,137-172.

Gray, W. M. (1978) A comparison of Piagetian theory and criterion referenced measurement. Review of Educational Research, 48, 223-249.

Hambleton, R. K. (1984) Determining test lengths. In R. A. Berk(Ed.),A guide to criterion-referenced test construction (pp.144-168) Baltimore, MD: The Johns Hopkins Univerity Press.

Hambleton, R. K. (1990). A practical guide to criterion-referenced testing Boston, MA: Kluwer.

Hambleton, R. K., & Zaal, J. N. (Eds). (1991). Advances in educational and psychological testing Boston,MA:Kluwer.

Hambleton,R.K., & Powell, S.(1983). A framework for viewing the process of standard-setting. Evaluation and the Health Professions, 6, 3-24.

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

Millman, J. (1973). Passing scores and test lengths for domain referenced measures. Review of Educational Ressarch, 43, 205-216.

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall

Wilcox, R. (1976). A note on the lenght and passing score of a mastery test. Journal of Educational Statistics, 1, 359-364.

（作者：政大教授兼附小校長）