試題反應理論的介紹（十一）

——題庫的建立

余民寧

題庫（item bank 或 item pool）不光只是一堆試題的集合體而已，而是一堆經過校準（calibration）、分析、歸類、與評鑑後，貯存起來的測驗試題組合體。Millman & Arter（1984）便將題庫界定爲一群使用方便的試題彙編；他們的意思是說該群試題可資應用於各種測驗場合的數量非常龐大，並且都是經過分析、編碼、與結構分類處理後的試題，並且有逐漸走向電腦化的趨勢。

題庫（尤其是以試題反應模式參數估計值所建立起來的試題）具有下列的優勢：

1.可使測驗編製者（也許是教師或專業機構）隨心所欲地編製能夠符合各種目標的測驗。
2.可使測驗編製者就題庫的範圍內，編製出每個目標都有適當題數的試題來測量到它的測驗。
3.如果題庫能夠包含內容有效且編題技巧純熟的試題在內的話，則測驗品質通常會比測驗編製者自己編的測驗品質還好。

由此可見，題庫具有改進測驗品質的潛能，在可預期的將來，它對測驗編製者的重要性將日益增加，同時對節省編製測驗所花的時間，亦將無可限量（Hambleton & Swaminathan ,1985,pp.255-256）。

建立題庫的步驟

題庫的建立，當然是依據課程標準、教材大綱、雙向細目表的編寫而成，它的建立過程，可以分成下列十一個步驟（許擇基、劉長萱，民81）：

1.試題的編寫與修訂：首先，仿照傳統編製測驗的原則，撰寫大量的試題，並邀請學科專家（如任課教師）和測驗專家就試題進行形式審查，看看是否能符合內容效度的要求，否則加以修改或刪減。
2.選題預試：放在題庫裡的試題，都必須是建立在同一量尺上的才行，否則試題間無法比較或延用。因此，

選擇適當的試題和考生樣本，是一項很重要的步驟，幸好前二文所談的定錨測驗設計（anchor test design）可以提供本步驟的參考（Vale-,1986）：

(1)定錨試題的數目：若使用同時校準法，則至少必須使用兩題定錨試題；若使用等組法，則可以不用定錨試題。一般而言，若將來所編製的測驗，具有60到80個試題的話，則在題庫建立過程中，必須使用15至20個定錨試題才夠。

(2)每個定錨組別至少要包含30位考生。

至於考生樣本數要多大才算足夠？大致上可以這麼說：若使用二個參數或三個參數對數型模式來進行校準時，則至少必須使用1000位考生；若使用一個參數對數型模式的話，則可以減少到500位考生便行。樣本的能力範圍，最好是呈常態分配。

3.試題的校準與銜接：選擇適當的反應模式來分析資料，必須考慮試題的性質。就選擇題而言，當然是以三個參數對數型模式最適合。決定好適當模式後，便可採用適當的電腦程式（如：BILOG3或LOGIST5等），以進行試題參數與考生能力參數的估計與適合度分析，統稱爲校準（calibration）。經過校準後的試題，必須能夠通過適合度的考驗者，方可被保留在題庫裡，因爲它們可以被適當的反應模式所解釋。如果在校準時，所使用的是不同的考生樣本，則在將試題放入題庫之前，還必須做試題銜接的工作（請參考前二文的說明），如此才能將所有的試題參數都建立在同一個量尺上。

4.更新題庫：理想的題庫特色是，包含題數相當充份的試題、試題具有內容效度、鑑別度不低於0.8以下、難度分佈均勻、猜測度愈小愈好等。並且由於試題被選入不同的測驗裡，和不同的試題出現在同一份試卷中，在施測時會產生不同的背景影響（context effect）。因此，當題庫裡的試題被選用之後，都必須有詳實的施測記錄，甚至必須再重新校準一次，以確定該試題參數的眞正適合度。如此可以確保題庫之素質能夠不斷地更新，也可以保持題庫之安全，避免淪爲考古題而被眾多考生熟悉，因而喪失題庫的功能。另外，也可以視測驗目的，使用題庫的目的，和學科的性質，於每次施測前，重新組合與排列題庫中的試題，以方便未來的使用。

5.測驗編輯：如果題庫的素質很高，則從題庫中抽取試題來編製一份測驗，便會很容易。編輯測驗的方式很多，最主要是看測驗目的而定。往往是由專家先將試題按學科、單元、屬性、和概念等，先行予以電腦編碼，再按其他考慮事項（如：試題參數值、訊息函數值、估計標準誤等），撰寫在電腦程式裡，以便編輯時輸入幾個關鍵字，就可獲得想要的測驗。

因爲題庫的內容龐大，幾乎不太可能用人工選題的方式，來編印試卷。通常都是仰賴電

腦的幫助，因此在列印試卷上，也有幾種方法可供參考：

(1)分層隨機抽樣選取試題：按教材內容來分，將題庫予以分成幾個層次，然後就每個層次中隨機抽取適當的題數，以作爲列印試卷的內容。這種作法，無法保證被選出的試題品質就一定是最好的。

(2)依試題參數值隨機抽樣：測驗編製者可依據教材內容，決定具有所欲的難度、鑑別度、和猜測度的試題參數範圍，及擬使用的題數，再由電腦自合格的試題中隨機抽樣，以編成一份試卷。這種作法的最大優點便是，免除人爲的偏見，並確保試題具有一定的品質。

(3)由試題訊息量來選取試題：首先，由測驗編製者決定理想的目標訊息曲線（target information curve）（讀者可以參閱前面「訊息函數」與「測驗編製」二文），然後自已校準的試題中，選取訊息量能夠填滿此一曲線的候選試題，可中途更換較佳的候選試題，每選出試題便計算其訊息量是否已接近理想的曲線，若否，則一直繼續這種選題過程，直到理想的目標訊息曲線被填滿爲止。

(4)由測驗編製者主觀選題：測驗編製者依據試題的特性和統計分析的資料，再由本身的專業判斷，以便決定選取何種試題。

6.評估測驗品質：對於新編製的測驗，可用試題反應理論所適用的電腦程式（如：BILOG3）和LOGIST5）來預測其特性。例如，電腦程式可利用所選取試題的難度、鑑別度、和猜測度的估計值，來計算出試題參數估計值的平均值、信度估計值、測驗訊息期望值和平均值、和各種不同長度下的預期測驗訊息量等資料，以便讓測驗編製者來判斷所編測驗的優劣。如果所編的測驗不符理想，則可以依據前述步驟來重編。

7.測驗是否達預期的水準：根據第六步驟的資料，來判斷所編的測驗是否有達到預期的水準：如果達到；則進行第八步驟；如果尚未達成，則回到第四步驟，重新更新試題再來。

8.執行考試：如果前個步驟顯示測驗品質不錯，則可對考生進行施測。當然，施測應有的指導語、測驗情境的安排與佈置、和其他會影響考試的注意事項等，都應該事前的準備與策劃。

9.評分：在經過考試後的學生作答資料，可再被拿來進行試題校準，此時，學生的考試成績，可用下列二種方法之一來加以評分：

(1)直接以學生的能力估計值 θ 來代表學生的能力。唯這種作法，比較不容易被大眾所瞭解，因此解釋起來，頗費周章。

(2)以眞實分數（true score）來表示學生的能力。亦即將每位考生在每個試題上的答對機率，加總起來的和，即是他的眞實分數。眞

實分數的值域將分佈於全部試題的猜測度之和與試題總題數（n）之間。唯這種作法，仍有其解釋上的不便處，因此，可將眞實分數除以試題總題數，以轉換成正確答對試題的百分比分數，此分數則與一般學校慣用的百分制計分方式的意義相同：愈接近百分之百，表示其能力愈高；反之，愈接近零，則表示其能力愈低。

10.決策：此步驟旨在應用上述評分與試題評鑑的結果，作爲甄選學生，診斷命題技巧，與改進教學的參考。

11.研究與評鑑：題庫的應用，不僅是用於編製新測驗，以節省人力、物力、和時間，並可透過每次考試完畢後，針對試題與考生能力參數進行校準，以評鑑試題品質的好壞、試題內容有否偏差（如：有利於某種族群的考生，而不利於另一種族群的考生）、以及診斷學生的作答資料有否不尋常、或找出學習有誤差的部份等，這種不斷研究與評鑑的過程，正是題庫所提供的特色。

發展題庫的適當時機

已知上述建立題庫的步驟後，什麼情況下，才需要去建立並運用題庫？Millman & Arter（1984）建議在至少滿足下列條件之一的情況下，才需要著手建立題庫，並充份發揮題庫的價值。

(1)現存測驗無法廣被接受，並且客觀環境需求編製屬於自己的測驗時。

(2)經常需要進行測驗時。

(3)需求具有多份複本測驗時。

(4)實施個別化適性測驗（individualized adaptive testing）時。

(5)許多測驗使用者願一致建立滿足自己所需的題庫時。

(6)已具備題庫系統，如：電腦設備和可用之電腦軟體時。

建立題庫所面臨之重大課題

1.題庫應該包含多少試題？

基本上而言，題庫內的試題當然是愈多愈好，但是應該考慮所加入題庫的試題，是否具有內容效度和統計品質應達成的標準，以及考慮測驗的目的何在？Prosser（1974）建議每個概念至少要包含10個試題，每一單元課程內容至少要包含50題。Reckase（1981）則建議一百至二百個難度均勻分佈，且具有合理的鑑別度的試題，便可適用在電腦化適性測驗（computerized adaptive testing）裡。另外，測驗的目的如果是在對課程作一整體的評估，則不需針對每項學習細節編製太多試題；如果僅在作學習診斷，則許多學習細節部份，仍需要編製許多試題去測量它們。

2.題庫系統應該如何分類？

常見的分類系統是依據內容來分類，它有兩項作法：一是依主題或教學目標來檢索試題，另一是採關鍵字方式來檢索試題。一般而言，採用關鍵字方式比較富彈性，可以同時適用於目標、內容、年級、及思考歷程等；但是依主題或教學目標檢索方式，則比較可以顯現知識結構的層級分明。當課程修訂時，採用關鍵字系統者修訂比較容易、迅速；但如果電腦無法處理多重關鍵字時，或分類系統本身就具有明確的界定（如：生物學中的種、屬、科、目等層次的分類）時，則採用固定的分類方法就比較適合。專家們的建議，任何題庫系統都應兼具這兩種編碼檢索的方法。

3.題庫內試題是否必需具備量尺化的參數？

所謂量尺化的試題參數，是指將試題參數（如：難度值）經過校準後，都換算成同一量尺單位的指標。這種參數，正是試題反應理論具有試題參數不變性始然，也是古典測驗理論所採用的試題參數指標（但容易受樣本影響）所無法媲美的。因此，如果能對大樣本進行施測，則試題參數的量尺化就非常必要；但如果僅能對教師個別班級施測的話，則試題參數的量尺化問題便可予以忽略，因爲不僅是不可能，而且也沒有必要。至於，如果學校仍想運用試題參數的量尺化過程，來建立起屬於自己學校適用的題庫的話，則本文作者的建議：不妨採行幾個學校聯合命題的方式，以力求獲取大樣本（如：大於1000人以上），來建立起量尺化的參數試題。

4.題庫是否可以公開？

如果題庫可以公開，讓任課老師任意取用，則有人擔心：教師是否從此就僅教題庫內容，而使教學活動更形窄化。這點憂慮也是必然的現象。不過，有個觀點必須釐清：由於建立一個量尺化的題庫，不是一件容易的事，它必須投入大量的人力、物力、時間、和金錢，才能有所展獲，因此只要題庫夠大（理論上可以達到無窮大），教學是否會因此窄化的問題，倒可以不必擔心，因爲教師無法作到僅教題庫，而不教正常的課程內容；但是，如果題庫不夠大的話，公開題庫必然導致窄化教學活動，因此，是否要完全公開，則有待進一步的商榷。但是，基本上公開少數的樣本試題，以讓教師和考生明瞭運用題庫的評量方式和重點，則是正確、並且有其必要的作法。

5.題庫是否安全？

題庫的建立，故然可以使日後的測驗編製更加容易，也可以使評量問題更輕鬆地獲得解決。但是，題庫的重覆使用，是否會妨礙到試題的安全性（如：雷同或考古試題的出現）？這點憂慮，也許在題庫少時，會有此顧慮，但在題庫大時，這層顧慮也許就可以不必有。另外，隨時更新題庫的內容，確保試題的內容效度和統計品質，也是保障題庫安全的一項作法。

參考書目

許擇基、劉長萱（民81）。試題作答理論簡介。台北：中國行爲科學社。

Hambleton, R. K., & Swaminathan, H.（1985）. Item response theory：Principles and applications. Boston：Kluwer.

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

Prosser, F.（1974）. Item banking. In G. Lippey（Ed）, Computer-assisted test construction（pp.29-66）. Englewood Cliffs, NT：Educational Technlolgy.

Reckase, M. D.（1981）. Tailored testing, measurement prlblems and latent trait theory. Paper presented at the annual meeting of the National Council for Measurement in Education, Los Angeles.

Vale, C. D.（1986）. Linking item parameters onto a conmon scale. Applied Psychologioal Measurement,10,333-344.

（作者：國立政大教授兼附小校長）