測驗編製程序

➢ 蕭儒棠

測驗編製程序

蕭儒棠
國家教育研究院助理研究員

壹、簡介

評量 (assessment) 是教與學之間的橋樑，若運用得當，可有效聯結教與學，進而提升教與學的品質。每位教師有其獨特的教學方式，每位學生適合的學習方式也不相同，依據評量得到的回饋訊息，教師能了解學生作答狀況與觀念不清之處，進而對症下藥，及時補救，以提高學生的學習動機，並增進學生學習的信心與興趣。

根據評量得到的回饋訊息，教師可檢視學生的學習需求，了解學生是否具備學習某個單元應有的知識和技能，並據以調整課程內容，為教與學預作準備。評量也可審視學生學習的狀況，分析學生學習的優缺點，檢視教與學的進程，確定學生的學習進展，並針對學習困難之處，機動修訂教學內容，改進教學方法、進行個別輔導或補救教學，以提高學習效果。

評量同時關注教與學。教師可分析學生的解題過程，根據解題的錯誤類型，採取可行的教學策略，以導正學生的錯誤或迷思概念。此外，評量可針對課程的某個單元，檢驗教學目標完成的狀況，確定學生的學習成效，作為提升教學成效的參考。除了授課教師，學生同樣可藉由評量了解自己的學習需求、學習進展，和學習成效。根據評量得到的回饋資訊，學生可確定學習盲點，修正學習策略，進而提升學習品質。

評量的種類非常多元，除了常見的紙筆測驗，也有教室觀察、口頭詢問、小組討論、觀察個人、學習單、實驗紀錄本、作業、習題、論文和實驗操作等不同的型式。測驗 (test) 由代表各種知識、概念或技能的試題 (item) 組成，評分時根據學生的作答反應給予預設的固定分數或適當的部分分數，將作答反應轉換為量化的分數。得到每一道試題的分數後，考慮全部試題的分數則為學生的測驗分數。測驗分數是測驗的結果，它代表學生某種潛在特質的強弱程度，或在該學科習得的能力、技術或知識的程度。

將學生的作答反應轉換為測驗分數或能力值的方法，因測驗的特性及需求而有不同。大型測驗常以試題反應理論 (Item Response Theory) (Hambleton &

Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; 王寶墉，1995；余民寧，2009) 得到的能力值推估學生的學習狀況，而一般校園常見的測驗則採用古典測驗理論 (Classical Test Theory) (Allen & Yen, 2001; Crocker & Algina, 1986; Gulliksen, 1987; Lord & Novick, 1968; Nunnally & Bernstein, 1994; Suen, 1990; 余民寧，2009)，以所有試題得分的加總作為測驗分數。

為了提高測驗品質，發揮測驗應有的功能，扮演教與學的關鍵角色，測驗的試題必須經過質和量兩方面的分析，根據分析得到的結果不斷修審試題，以提高試題的品質。本文就測驗編製程序及試題編擬原則加以說明，作為教師或研究者編製測驗或編寫試題時的參考，並進一步運用測驗的回饋訊息，有效提升教與學的品質。

貳、測驗的類型

測驗因其目的、功能或使用時機而有不同的類型。依編製過程的嚴謹程度可分為標準化測驗 (standardized test) 和教師自編測驗 (teacher-made test) ；若以測驗結果解釋的方式區分，則有常模參照測驗 (norm-referenced test) 和標準參照測驗 (criterion-referenced test) ；若根據測驗使用的時機，可分成形成性測驗 (formative test) 和總結性測驗 (summative test) ；而根據測驗的功能區分時，可分為安置性測驗 (placement test) 與診斷性測驗 (diagnostic test)。

一、標準化測驗與教師自編測驗

測驗依標準化程度可分為標準化測驗和教師自編測驗。教師自編測驗的過程較為簡單彈性，可即時檢驗教與學的現況，而標準化測驗的編製過程以嚴謹著稱，標準化程度較高 (Berk, 1984; Koretz, 1988)。

(一)標準化測驗

標準化測驗是由測驗專家、課程專家和學科教師等共同編製的測驗。國內的台灣學生學習成就評量資料庫 (Taiwan Assessment of Student Achievement, TASA)、國民中學學生基本學力測驗、大學學科能力測驗、大學入學指定科目考試；美國的國家教育進展評量 (National Assessment of Educational Progress, NAEP)、國際間的學生能力國際評量計畫 (Programme for International Student Assessment, PISA)、國際數學與科學教育成就趨勢調查 (Trends in International Mathematics and Science Study, TIMSS) 和國際閱讀素養研究 (Progress of International Reading Literacy Study, PIRLS) 等，皆屬於標準化測驗。標準化測驗的涵蓋面較廣，目的是比較同年級或同年齡學生之間的學習成就，也可用於比較學校、學區甚至國家之間學生的學習成就差異。標準化測驗的編製、施測與評分等程序，必須經過嚴謹的規劃，確保每個環節都能遵守嚴格制訂的程序，選用的試題也必須經過預試及分析，以確保測驗的信度和效度，使

不同施測條件下接受測驗的考生，能得到公平的測驗分數，進而作有意義的比較。必須注意的是，TASA、NAEP、PISA、TIMSS、PIRLS等大型評量的設計目的是評估整體學生的能力、成就或素養，用於個人比較時，應謹慎考慮評量的設計是否適用，以及數據的解讀是否合理。

(二)教師自編測驗

教師自編測驗是教師根據具體的教學目標、課程內容和測驗目的，自行編製的測驗，用於測量學生的學習狀況。教師自編測驗的考生人數少，包含的內容範圍小，試題的題型多樣化，常用於檢驗階段性的學習成就與教學成效。常見的教師自編測驗包含隨堂測驗和定期測驗等，是教師根據教學需要，自行編製的測驗。教師自編測驗的目的在確定學生是否達成教學目標，作為教與學的改善依據，其試題編寫、施測、計分和結果的解釋與應用，完全由教師自行彈性決定。

教師自編測驗和標準化測驗在教學過程中具有互補的功能，教師自編測驗根據教師的實際教學需求彈性調整與編製，可獲得直接且及時的回饋資訊，是改進教學和提升學習的重要方式。而標準化測驗經過嚴謹的編製和施測程序，可針對學生的學習成就提供客觀的量化資訊，是決策者決定教育政策的重要參考資訊。

二、常模參照測驗與標準參照測驗

根據測驗結果的解釋方式，測驗可分為常模參照測驗和標準參照測驗 (Berk, 1984; Glaser, 1963; Gronlund, 1993; Linn, Miller, & Gronlund, 2009; 陳英豪、吳益裕，2001)。常模參照測驗的目的是區別考生彼此之間能力的差異，確定考生的相對排序位置，而標準參照測驗則是為了檢驗考生在特定的領域中，是否達到某些預設的精熟目標。

(一)常模參照測驗

常模參照測驗強調個別考生在群體中的相對位置或名次，為擴大測驗分數的分布範圍，以有效區分考生的學習表現，常模參照測驗通常選擇難度中等且鑑別度高的試題，捨棄所有考生都可能答對或答錯的試題。常模參照測驗注重考生與考生之間的比較，以相對比較的觀點呈現考生的測驗結果，適用於篩選或評定等第的測驗，可作為編班或入學等依據。

(二)標準參照測驗

標準參照測驗不考慮其他學生的測驗結果，而以某些預設的標準檢驗學生對特定知識和技能的掌握程度。標準參照測驗的試題應配合預計測量的學習結果，不需考慮試題的困難度和鑑別度，也不需刪除過於簡單或困難的試題。多數學生不能正確回答某些試題時，應進一步檢驗這些試題是否符合教學目標，教學方法是否恰

當。標準參照測驗以絕對比較的觀點看待個別學生的測驗結果。醫師、會計師、建築師、律師等專業證照考試，皆屬於標準參照測驗，通過與否不會因其他考生的表現而受到影響，只與某個預設的特定標準比較，若考生符合此特定標準，即可獲得證書。

三、形成性測驗與總結性測驗

測驗依施測的時機可分為形成性測驗和總結性測驗 (Linn, Miller, & Gronlund, 2009)。形成性測驗是教學過程中，配合教與學的需求，隨時彈性進行的小型測驗，目的是提供教與學的回饋訊息。總結性測驗是指某個單元或課程結束後進行的測驗，測驗的內容比較廣泛，通常用於評定成績。

(一)形成性測驗

形成性測驗檢視學生的學習進展，教師和學生可根據測驗結果對教與學採取滾動式的即時調整。形成性測驗著重教學中的「調查」，通常選在新的概念、技能或單元的教學結束後進行，針對與教學活動密切相關的小範圍內容進行測驗，不評定學生的等第或成績，而是確定學生是否掌握進入下一個概念、技能或單元的關鍵內容。相關的回饋資訊，除了可幫助學生掌握個別差異，學習尚未掌握的內容，也可協助教師檢討課程設計與教學策略的階段性成效，作為進行個別輔導或補救教學的依據。編製形成性測驗時應說明對應的教學策略和設計相關的學習建議，以發揮形成性測驗特有的功能，為下個概念或單元的教與學預做準備。

(二)總結性測驗

總結性測驗的內容範圍較廣，試題涵蓋課程內容中的基本知識和技能，測量學生對整體課程內容掌握的狀況，檢驗學生達到教學目標的程度。總結性測驗著重教學後的「回顧」，通常在完整的課程或教學活動結束後施測，對學生的學習成就評定成績，或檢驗某個教學方案是否有效，全面性檢視「教」與「學」的成效，以確定是否達成教學目標，期末考試或結業考試都屬於此類。

四、安置性測驗與診斷性測驗

測驗可依其功能分為安置性測驗與診斷性測驗二種 (Linn, Miller, & Gronlund, 2009)。安置性測驗於教學過程前實施，目的在檢驗學生是否具備某些先備知識，確定並安排適合的教學計畫。診斷性測驗於教學過程中實施，目的是確定學生學習困難的原因，以作為補救教學的依據。

(一)安置性測驗

安置性測驗著重教學前的準備，施測的時機選在教學開始之前，用於檢視學生

的學習背景，確定學生具備的基本能力及個別差異，並了解學生對新的學習任務的準備狀況。安置性測驗協助教師瞭解學生的特徵，教師則可依據測驗結果，評估學生的性向、能力與需求，並針對教學的內容、方式、型態與順序等，預作適當的調整與規劃，例如，決定教學起點與教學順序，教材教法的選擇，是否複習相關內容，是否進行分組教學等，將教學的重心集中於更深入的學習，或根據學生的分組，設計特定的教學策略或可行的學習互動。安置性測驗的結果只作為教師教學的參考，有時也作為教學前後學生學習成就與教師教學成效的比較，並不列入學生的成績報告。

(二)診斷性測驗

診斷性測驗於教學過程中施測，屬於心理測驗的一種，著重學習困難的分析。診斷性測驗運用精密的方式尋找學生在某個特定學習內容或技能上的問題，以確定學生學習困難的真正原因。藉由分析學生的作答反應，診斷性測驗可找出學生學習過程中的弱點，研判出現學習困難的可能原因，並進一步設計可行的補救措施或教學策略。

參、測驗編製的原則

測驗編製應遵循一定的程序，以確保測驗內容與測驗目的相符，降低其它因素對測驗結果的影響，使測驗結果儘可能反映考生所具備的知識和技能 (Gronlund, 1993; Haladyna, 1996; Haladyna, 2004)。測驗編製時應注意以下原則：

一、測驗應反映課程內容與教學目標

測驗是為了檢驗課程內容及教學目標中，學生對知識和技能的學習狀況。然而測驗並無法涵蓋課程內容中全部的知識和技能，因此，選擇的測驗內容應具有代表性，以充分代表學科的課程內容。測驗同時兼具考核教學成效的功能，因此，測驗應以教學目標為依據，藉由測驗審視教學目標的完成狀況 (林世華，2000)。若測驗結果顯示，多數學生無法掌握測驗涵蓋的課程內容及教學目標，則應考慮大幅修改或重新編製測驗內容。另一方面，若重新編製測驗內容後，多數學生仍無法通過測驗，則應考慮適度調整教學策略，以達成應有的教學目標。

二、測驗目的應能促進師生的教與學

測驗是結合教與學的重要環節，教師可利用測驗結果調整教學，並指導學生學習。對學生而言，測驗的回饋資訊能幫助學生釐清自己對課程內容的掌握狀況，找出學習狀況較薄弱的環節，進而調整學習方法和學習重點，將有限的時間和精力集中於需要加強的內容。測驗結束後，應儘快提供學生測驗的回饋資訊，導正學習的錯誤，並提供正確的答案及合理的解題思路。對教師而言，教學前的測驗，有助於

教師了解學生的起點行為，規劃適合的教學活動。教學過程中，教師可透過測驗的回饋資訊，隨時檢視學生對課程內容的理解狀況，瞭解影響學生學習的各種因素，進而調整教學目標、教學計畫、課程內容、教學方法和教學進度。測驗的回饋資訊也可以協助教師了解學生的學習類型及學習困難，進而採取適合的補救措施。

肆、測驗編製的步驟

為了避免與測驗目的無關的因素影響測驗結果，確保測驗內容與測驗目的相符，使測驗能如實反映學生具備的知識和能力，編製測驗時應遵循共同的標準作業程序，這套程序因測驗的特性而有不同的重點或嚴謹度，編製時可考量測驗的特性作適當的調整。無論測驗的類型為何，測驗始終是教與學的重要環節，而教師則永遠扮演試題編製的重要角色。為了確保測驗能發揮應有的功能，達成教學評量的目的，教師應熟悉測驗的編製原則和步驟，以編製適合的測驗。歸納國內外文獻 (余民寧，1993，2009，2010，2011；洪碧霞、邱上真、林素薇、葉千綺，1998；歐滄和，1993；劉湘川、蔡良庭，2005)，編製理想的測驗時，應考慮「確定測驗目的與編製計畫」、「試題的編寫與評分原則」和「試題和試卷的審查與分析」等三大面向。

一、確定測驗目的與編製計畫

測驗編製的首要工作是確定測驗目的，接著根據測驗目的擬定測驗編製計畫，以確定測驗類型、測驗題型、試題分析、評分方式及測驗報告等測驗中的每個步驟。此外，為了搭建教與學之間的橋樑，測驗編製計畫中關於試題內容的取樣，可參考教學指南中的課程內容與教學目標，以檢驗教學是否包含應有的課程內容，並達到預期的教學目標。一份周詳且具體可行的測驗編製計畫應考慮測驗目的及測驗類型、教學目標及課程內容及測驗藍圖及測驗題型，以下分別就上述三點進一步說明。

(一)確立測驗目的及測驗類型

測驗是為了授予專業證照，是為了比較考生之間的能力，是為了診斷學生學習困難的原因，抑或是為了獲得教與學的回饋資訊，不同的需求對應不同的測驗目的。而測驗目的決定測驗的方向、內容、方式甚至影響測驗結果的解釋方式。此外，測驗的類型相當多樣化，不同類型的測驗，具有不同的特性與功能。因此，測驗首先應依據教與學的需求確定測驗的目的，並依照測驗的目的，決定適當的編製的程序、施測的時機、測驗的特徵及測驗的解釋，如此才能充分且完整地反映教學目標，並提高回饋資訊的參考價值。

1. 根據編製的程序

若測驗對象為教師授課的學生，或是同年級的全部學生，這類較小規模的測

驗，可選擇教師自編測驗。教師自編測驗的編製過程較簡化且具有彈性，可依受測學生的特徵，彈性調整測驗的內容。若測驗的對象擴大為跨校、跨學區、跨地區甚至跨國界的學生時，應選擇標準化測驗，以維持測驗的公平性，確保測驗得到的回饋資訊能進行有意義的比較。標準化測驗的建置工作通常委由特別成立的專責機構負責，編製程序極為嚴謹，一般校園常見的測驗以教師自編測驗為主，編製程序較為彈性，二者各有所長，各司其職。

2. 根據施測的時機

安置性測驗通常於教學開始之前舉行，它檢驗學生是否具備課程的入門知識或技能，幫助教師了解學生對學習某一門課程的準備程度。形成性測驗在教學中施測，可確定學生對課程內容的熟悉程度，就課程內容、教師教學和學生學習三個環節提供調整與改進的資訊，也可作為教師是否進入下個單元的參考。診斷性測驗的功能在分析教學過程中學生反覆出現的學習困難，教師根據測驗結果的分析，設計可行的個別輔導或補救教學。教學後的總結性測驗則著重通盤的了解，測驗的結果可供學生及家長參考，或作為升學的依據。

3. 根據測驗的特徵

依課程階段劃分或聚焦方式的不同，有時安置性測驗可視為某個單元的總結性測驗，而總結性測驗也可能是下個學習階段的形成性測驗。若測驗結果顯示某位學生的學習狀況符合預設的通過標準，則該名學生可獲得學分及選修進階課程的資格，獲得學分屬於總結性測驗，而獲得選修進階課程的資格則為安置性測驗。此外，診斷性測驗和形成性測驗二者同樣以「發現」學生的學習困難為目的，同樣於教學過程中施測，但是形成性測驗著重於學習狀況的「發現」與「調查」，針對教學活動進行品質管制，隨時掌握學習是否達到預期成效，並盡可能對學習困難的部分進行補救。而診斷性測驗則更強調學習困難的「分析」，它針對形成性測驗無法立即處理的問題，進行更精密的診斷，作為進行補救的參考。

4. 根據測驗的解釋

常模參照測驗和標準參照測驗性質不同，適用於不同的測驗類型。相對於標準參照測驗，常模參照測驗中，考生的測驗分數變異性較大，得分範圍的分布較廣，能充分顯示學生的個別差異，尤其適合編組、編班或入學測驗等安置性測驗和總結性測驗。而標準參照測驗著重在瞭解個人的測驗表現是否達到事先所設定的標準，測驗結果提供考生在某個考科或領域的表現描述，回饋訊息較豐富，有助於學習診斷、補救教學或個別指導，因此，形成性測驗和診斷性測驗通常屬於標準參照測驗。至於標準化測驗和教師自編測驗究竟適合常模參照測驗抑或標準參照測驗，可參考二者的優缺點，依測驗的需求彈性選擇。此外，選用時也應特別留意，常模參照測驗提供學生測驗分數，對診斷學習的功能較弱，且同儕關係和學習情緒容易因

競爭產生負面的影響；而標準參照測驗說明學生當前的學習狀況，雖然降低了競爭的負面的影響，卻削弱了測驗追蹤或評估學生學習差異的功能。

(二)確定教學目標及課程內容

1. 教學目標

教學目標引導學生的學習方向、達成教師的教學任務，並說明學生完成指定的學習內容後，應具有的知識 (knowledge)、技巧 (skill)、能力 (ability) 或態度 (attitude)。清晰且具體的教學目標，不僅有助教與學雙方的溝通，提高教學成效與學習成就 (Kemp, 1985; Simpson, 1972)，根據教學目標編製的測驗，也更能準確地提供教與學所需的回饋資訊。關於教學目標的理論可參考Bloom的《教育目標分類》(taxonomy of educational objectives) (Bloom et al., 1956; Krathwohl, Bloom, & Masia, 1964)，文中將教育目標的認知領域 (cognitive domain) 分為認識與記憶 (simple recall or recognition of facts) 及能力與技巧 (intellectual ability and skill) 二部分，其中認識與記憶表現的是記憶能力，屬於知識 (knowledge) 層次，而能力與技巧則表現批判、反省或問題解決等較複雜的思考能力，可進一步區分為理解 (comprehension)、應用 (application)、分析 (analysis)、綜合 (synthesis) 與評鑑 (evaluation) 等五個層次。

為了因應教育理論的發展與演進，2001年Anderson等人考慮更廣泛因素，提出修訂版本的Bloom分類 (Anderson et al., 2001)。新版的 Bloom 分類修訂為名詞層面的知識向度 (knowledge domain) 和動詞層面的認知歷程向度 (cognitive process domain)，前者協助教師區分應該教什麼 (what to teach)，而後者旨在促進學習者保持 (retention) 和轉換 (transfer) 學得的知識。知識向度專指知識的分類，將知識分為事實 (factual)、概念 (conceptual)、程序 (procedural) 及後設認知 (meta-cognitive) 等四類知識；認知歷程向度由低層次至高層次依序為記憶 (remember)、理解 (understand)、應用 (apply)、分析 (analyze)、評鑑 (evaluate)、創造 (create) 等六項 (Anderson et al., 2001；葉連祺、林淑萍，2003；李坤崇，2004)。

教學目標引導且決定如何設計測驗，若教學和測驗的內容不一致，即使有高品質的教學，高成就的學生也無法在測驗中有優異的表現 (Airasian & Miranda, 2002)。因此，編製試題與施測最主要的依據是教學目標而不是教材內容，妥善運用Bloom分類，可加強測驗與教學的一致性 (Airasian & Miranda, 2002；葉連祺、林淑萍，2003；李坤崇，2004；李坤崇，2006)。一般認為知識屬於較基礎的層次，適合基礎入門的課程或年齡層較低的學生，而理解、應用和分析等較高層次的學習，則以進階課程或較年長的學生為主。因此，確定教學目標不僅是測驗編製計畫中的第一個步驟，也是最重要的步驟。

2. 課程內容

測驗是「教」與「學」雙方的橋樑，為了提供「教」與「學」所需的回饋資

訊，測驗必須忠實反映教學目標。教學目標描述教學的「結果」，而課程內容則是教學的「內容」，它考慮學科教學應有的內容範圍。內容範圍指的是教材範圍或能力指標，是學科的具體知識或技能，確定內容範圍是為了確保測驗內容來自應有的課程內容。測驗的目的在於測量學生在某一課程領域的學習成果，測驗的內容當然必須根據課程的內容範圍。

編寫試題之前必須清楚規劃測驗預計測量的範圍和層次。範圍是測驗所要測量的內容，包含學科教學的知識和概念，而層次則是測驗所測量的能力，包括某個教育領域中的某個層次。此外，應儘可能蒐集正確且可靠的資料，作為試題的取材依據，例如，課程標準、課程綱要、教科書、參考書、教師手冊、測驗理論、相關測驗題本以及心理學相關著作等。參考資料愈完整，試題的編寫工作愈順利，測驗內容愈有效，測驗結果的代表性愈高。

(三)確定測驗藍圖及測驗題型

1. 測驗藍圖

測驗藍圖描述重要的教學目標與評量目標之間的關係，避免試題編製時，命題者依自己的喜好隨意命題。此外，受限於測驗時間的長度，測驗內容無法涵蓋全部的教學目標與課程內容，為了使測驗內容具有較高的代表性，同時反映各種認知層次的相對比重，命題前應參考測驗目的、測驗類型、教學目標及課程內容，完成測驗藍圖的規劃，並將測驗藍圖轉化為具體的雙向細目表，以確保測驗品質，精確達成測驗目的 (余民寧，2009；余民寧，2011)。

雙向細目表中，每一橫向的表格代表一特定的課程內容，縱向的表格則代表知識、理解、應用、分析、綜合、評鑑等不同層次的教學目標。制定雙向細目表時，可以參考下列步驟：

(1) 於雙向細目表最左側的表格內，由上而下填入課程內容中每個單元的名稱。
(2) 於雙向細目表最上方的表格內，由左而右填入知識、理解、應用等教學目標。
(3) 依據課程內容的重要性，確定每個單元的試題數量。
(4) 依據課程內容的特性，確定每個單元的試題應有的教學目標類型。
(5) 確定每個單元、每個教學目標的試題數目，在教學目標一欄，填入數字。
(6) 重複步驟 (3) 至 (5)，合理分配每個單元、每個教學目標應有的試題數目。
(7) 計算並填入每個教學目標的試題總題數與分配比例。
(8) 計算並填入每個單元的試題總題數與分配比例。
(9) 重複步驟 (3) 至 (8)，檢視每個單元、每個教學目標的試題題數與分配比例是否合適。

2. 測驗題型

測驗可測量不同學習階段、不同學科的學習狀況，也可用於挖掘學習動機、自

我概念、創造力、……等，不同面向的潛在特質。測驗由試題組成，用於蒐集應試者學習狀況或潛在特質的相關訊息。試題是測驗的核心，也是影響測驗品質的重要因素。選擇測驗的題型時，應依據題型的特性，以發揮該題型特有的功能。測驗藍圖規劃測驗應包含的課程內容以及對應的能力層次，測驗藍圖轉化為具體的雙向細目表後，命題者可根據雙向細目表標示的課程內容及教學目標，選擇適合該層次的題型。試題的題型種類繁多，每種題型的測量功能均不相同，依照考生作答反應的方式可分為選擇反應試題 (selected-response item) 和建構反應試題 (constructed-response item) 二大類。

(1) 選擇反應試題

選擇反應試題包含單選選擇題 (multiple-choice item)、是非題 (true-false item) 和配合題 (matching item) 等類型，屬於評分較為客觀的題型。選擇反應試題對學生的作答反應限制較多，試題中提供幾個預設的選項，考生由其中挑選最適合的選項作為答案，它的特徵是作答內容簡短、具體且明確，評分結果準確、客觀且公平。另一方面，選擇反應試題不易察覺學生的答案是否經由猜測得到，且無法有效測量學生的表達能力或其它較高層次的能力。此外，選擇反應試題所需的作答時間較短，考生於相同的測驗時間內可回答更多試題，因此，測驗可涵蓋的範圍更大，可測量的內容更多。

編寫選擇反應試題時，可就程序 (procedural)、內容相關 (content concerns)、題幹結構 (stem construction)、一般選項發展 (general option development)、正確選項發展 (correct option development) 及誘答選項發展 (distractor development) 等面向，檢視是否符合優良試題的原則，以達成鑑別考生的測驗目的 (Haladyna & Downing, 1989a; Haladyna & Downing, 1989b; Haladyna, Downing, & Rodriguez, 2002)。

(2) 建構反應試題

建構反應試題包含簡答題 (short answer items)、限制反應題 (restricted response essay question)、和申論題 (extended response essay question) 等類型，屬於評分較為主觀的題型。建構反應試題的試題中不提供任何預設的選項，考生作答時，根據試題的要求，自行組織相關內容，並以適當的方式陳述答案，它的閱卷過程繁複冗長，評分過程及結果較為主觀。

建構反應試題並沒有明確的正確答案和評分標準，容易受到評分者的主觀因素影響，評分的公平性容易受到質疑。然而，建構反應試題可以觀察並有效測量考生對於知識和問題的概括、統整、分析與解決等多方面的能力，同時也能避免答案是經由猜測而來的可能性，是客觀性試題所無法取代的。此外，由於建構反應試題所需的作答時間較長，限制測驗包含的試題數量與涵蓋的內容，降低試題內容取樣的代表性。關於建構反應試題編寫，可參考美國教育測驗服務社 (Educational Testing

Service, ETS) 所出版的《建構反應及實作評量編寫指引》(Guidelines for Constructed-Response and Other Performance Assessments) (Baldwin, Fowles, & Livingston, 2005)，它提供了許多建設性的試題編寫原則，可作為編寫建構反應試題的指引。

測驗採用的題型影響甚至引導教師教學及學生學習的方式，若測驗以記憶性試題為主，為了通過測驗，教師的教學將重複講述記憶性知識，以加強學生的印象，而學生將大量背誦知識，忽略理解及運用等較高層次的學習目標。因此，如何選擇並活用各種試題題型，將影響測驗結果的品質，若測驗採用的題型強調問題解決的能力，鼓勵理解與應用，將引導教師與學生往更高層次的「教」、「學」目標邁進。編擬試題時，除了根據學科的內容知識、教師的教學經驗及考生的背景特徵，同時也可參閱測驗許多相關的研究成果與文獻，以提高測驗的品質 (余民寧，2011；簡茂發，2000；Brennan, 2006; Downing, & Haladyna, 2006; Haladyna, 1996; Haladyna, 2004; Hogan & Murphy, 2007; Roid & Haladyna, 1982)。

二、試題和試卷的審查與分析

編擬試題時應儘可能增加試題初稿的數量，最後再依據雙向細目表的預設數目，挑選部分審查通過的試題組合為測驗卷。欲使測驗試題臻於完善，所包含的試題必須經過嚴謹的審查程序，分別就試題的內容、形式、困難度和鑑別度等逐一檢驗，以反應試題的功能與特徵，進而發揮其測量的功能。審查方式可分為邏輯審查 (logical review) 和實證審查 (empirical review) (余民寧，2011；Haladyna, 1996; Lawshe, 1975; Roid & Haladyna, 1982)。邏輯審查針對試題的內容和形式，審查試題內容的取材是否符合課程內容及教學目標，又稱為形式審查 (facial review)。實證審查又稱為客觀審查 (objective review)，以預試結果分析試題的困難度 (difficulty)、鑑別度 (discrimination)，以及考生的作答反應組型 (response pattern)，審查各個選項 (option) 的反應情形是否符合預設的測驗目標。

(一)邏輯審查

測驗各有其特定的功能和適用範圍，因此，編製試題時應以測驗目的為依據。以成就測驗為例，為了測量學生於某一學科教學活動中的學習成效，了解學生於不同的層次的行為變化，編製測驗時應以教學目標和課程內容為依據。分析教學目標和課程內容後，將二者結合為雙向細目表，並以雙向細目表作為試題編製的依據。審查時應逐一確認試題的格式、敘述的品質等，確保試題符合編製的原理和要求，邏輯審查的內容包含：

1. 試題是否符合雙向細目表的規劃
2. 試題是否代表預期測量的教學目標
3. 試題是否依據試題命題原則編寫
4. 試題的敘述是否能清楚表達題意

5. 試題的呈現方式與作答說明是否適當
6. 試題的敘述是否提供暗示答案的線索

檢驗試題的測量目標與教學目標是否一致時，也可參考「試題與目標一致性」(item-objective consistency, IOC) 作為依據 (Rovinelli & Hambleton, 1977)。IOC指標的值域介於 -1.0 和 +1.0，IOC指標愈接近 +1.0，表示試題與目標的一致性愈高。計算時邀請學科專家檢視每道試題測量目標的程度，並依據下列定義評分：
+1分：很明確的斷定某個試題是測量某個目標
0分：無法確定某個試題是否能測量某個目標
-1分：很明確的斷定某個試題不是測量某個目標
評分後以下列公式計算試題與目標一致性指標 (IOC)

$$\text{IOC} = \frac{(N-1)S_1 - S_2 + S_2}{2(N-1)n}$$

N：目標個數
n：學科專家人數
S_1：所有專家在某個試題上的某個目標的評分總和
S_2：所有專家在某個試題上的所有目標的評分總和

(二)實證審查

實證審查用於分析試題功能，以統計方法獲得客觀的量化數據，作為判斷試題品質、挑選試題、完成組卷的參考。標準化成就測驗中，編擬完成的試題，通常透過預試 (pilot test) 結果進行實證審查。實證審查可確定試題的難度和鑑別度，並比較考生於各個選項的作答反應，以確保試題的品質，作為挑選試題的參考。實證審查分為試題分析與測驗分析二部分。

1. 試題分析

試題分析主要在於透過量化數據，分析每道試題的困難度與鑑別度，若試題為選擇反應試題，也可分析試題的選項誘答力 (distraction)。透過試題分析，可瞭解試題的品質，刪除或改寫品質不佳的試題，進而改善試題的品質。

(1) 困難度分析

理想的測驗應能有效地依測驗目的篩選考生，測驗的困難度過高或過低時，均無法發揮測驗應有的篩選功能。常模參照測驗以「相對」的觀點，比較每位考生在全體考生中的相對位置，測驗困難度過高或過低時，多數考生的測驗分數落於低分群或高分群，皆無法有效區分考生的差異。若為標準參照測驗，測驗標準過高或過低時，多數考生的測驗結果落於「未通過」或「通過」，同樣無法有效區分其程度

差異。

古典測驗理論對試題困難度定義與通過率有關：

$$P_i = \frac{R_i}{N}$$

其中，N 為應試總人數、R_i 為該題正確作答人數，P_i 為通過率，也可視為試題的困難度，通過率愈高，試題的困難度愈低。若再進一步依測驗的總分，將考生分為高分組 (全體受試者當中分數最高的27%至33%) 及低分組 (全體受試者當中分數最低的27%至33%)，則高、低分組考生通過率的平均，即為試題的「困難度指標」(difficulty index) ：

$$P_H = \frac{R_H}{N}$$

$$P_L = \frac{R_L}{N}$$

$$P_i = \frac{P_H + P_L}{2}$$

其中，R_H 為高分組該題正確作答的人數，R_L 為低分組該題正確作答的人數，P_H 為高分組該題正確作答的通過率，R_L 為低分組該題正確作答的通過率，P_i 為試題的困難度指標。困難度指標 P_i 最大值為1，最小值為 0，愈接近 1 代表答對人數愈多，試題愈簡單；愈接近 0 代表答對人數愈少，試題愈困難 (周文欽、歐滄和、許擇基、盧欽銘、金樹人、范德鑫，1995；郭生玉，2004；Ebel & Frisbie, 1991)。

(2) 鑑別度分析

鑑別度是試題區辨高能力考生與低能力考生的功能，透過算鑑別度，可顯示每道試題是否能讓高能力考生傾向答對，而低能力考生傾向答錯。優良的測驗除了應有難易適中的試題，也應儘可能提高試題的鑑別度。鑑別度值越高，表示試題越能區分出高能力考生與低能力考生；反之，則無法區分出高能力考生與低能力考生。測驗編製者希望能力高的考生在每道試題的答對率應高於能力低的考生，通常以測驗成績表示考生的能力高低，鑑別度指標即是呈現這樣的訊息。測驗的鑑別度與考生測驗成績的變異數有關，鑑別度愈高的測驗，測驗成績的變異數愈大。若試題太困難或太簡單，考生作答的情形趨於一致，測驗成績的變異數較小，鑑別度較低，因此，調整試題的鑑別度時，應同時考慮試題的困難度，若試題太困難或太簡單，測驗成績的變異數較小，鑑別度較低。若試題的變異數太小，表示考生作答的情形趨於一致，該試題的鑑別度低，屬於不良試題。

鑑別度分析時，將受考生分成高分組 P_H (全體受試者當中分數最高的27%至33%) 及低分組 P_L (全體受試者當中分數最低的27%至33%)，求高、低兩組考生通過率的差，即為試題的「鑑別度指標」(discrimination index)。鑑別度指標 D_i 的最大值為 +1，最小值為 -1，愈大代表試題鑑別程度愈好，愈小代表試題鑑別程度愈差 (周文欽等，1995；郭生玉，2004；Ebel & Frisbie, 1991)。高低能力組別的答對率計算步驟條列如下：

I. 將考生的原始作答反應比對標準答案後，答對編碼為「1」，答錯編碼為「0」，使其變成二元計分 (答對是1，答錯是0) 模式。

II. 將每一位受試者的每一題的分數加總變成原始總分。

III. 將所有受試者的原始總分由高到低排序，取前面27%至33%的受試者為高能力組，即 N_H，最後面27%至33%的受試者為低能力組，即 N_L。

IV. 針對第 i 題，分別計算高能力組在此題的答對人數，即 R_{iH}，和低能力組在此題的答對人數 R_{iL}。

V. 計算高能力組別在 $P_{iL} = \frac{R_{iL}}{N_L}$ 第 i 題之答對率，即 $P_{iH} = \frac{R_{iH}}{N_H}$；計算低能力組別在第 i 題之答對率，即。

計算高低能力組別在第 i 題之答對率差 $D_i = P_{iH} - P_{iL}$，就是鑑別度指標。

在第III步驟中，取多少比率的受試者作為高低能力組別受試者之人數並無定論（四分之一到三分之一皆可），只要能夠將受試者群分成三個區段，以區分出高、中、低能力組別之受試者均可 (余民寧，2011)。

此外，二系列相關 (biserial correlation) 和點二系列相關 (point-biserial correlation) 也可作為鑑別度指標，其意義是受試者在某一題的答對或答錯與測驗總分之相關，表示某道試題之作用與測驗總分作用之間的一致性程度。由於其計算方式較複雜，一般皆是透過統計軟體協助計算此指標。

當某變數是屬於二元變項 (dichotomous variables)，如試題答對以1表示，答錯以0表示，而另一變項是連續變項 (continuous variables) 如測驗總分，則可計算點二系列相關係數作為鑑別度指標，公式如下：

$$\gamma_{pb_i} = \left(\frac{\overline{X}_{ip} - \overline{X}_{iq}}{s_T}\right) \times \left(\sqrt{p_i q_i}\right)$$

其中，其中，$\overline{X}_{ip}$ 是第 i 題答對學生在校標（如測驗總分）的平均得分，$\overline{X}_{iq}$ 是第 i 題答錯學生在校標（如測驗總分）的平均得分，p_i 是第 i 題之答對率，q_i 是第 i 題之答錯率 ($q_i = 1 - p_i$)，S_T 全部受試者在測驗總分之標準差。

二系列相關的公式如下：

$$\gamma_{bi} = (\frac{\overline{X}_{ip} - \overline{X}_{iq}}{s_T}) \times (\frac{p_i q_i}{y_i})$$

其中，$\overline{X}_{ip}$ 是第 i 題答對學生在校標（如測驗總分）的平均得分，$\overline{X}_{iq}$ 是第 i 題答錯學生在校標（如測驗總分）的平均得分，p_i 是第 i 題之答對率，q_i 是第 i 題之答錯率($q_i = 1 - p_i$)，S_T 全部受試者在測驗總分之標準差，y_j 答對率 p_i 在常態分布下所在位置相對應之曲線高度。

(3) 誘答力分析

選擇題的選項有其篩選功能，學習狀況良好的考生選擇正確的選項，對課程內容的認識仍一知半解、有迷思概念、甚至學習狀況不佳的考生則可能選填錯誤選項。提高錯誤選項的誘答功能，可增加試題的鑑別能力，選擇題的選項誘答力分析，可進一步提供教師試題分析的指標，協助教師改進編擬試題的技巧，並藉由學生的作答反應組型了解整體學生的學習狀況 (Haladyna, 1996)，進而採取可能的教學策略。

進行選擇題的選項誘答力分析時，同樣將應試的考生分為高分組 (全體考生中分數最高的27%至33%) 及低分組 (全體考生中分數最低的27%至33%)，以考生的作答反應組型為依據，分別計算、比較高分組和低分組於每一個選項的選答率。若試題具有優良的選項的誘答功能，每個錯誤選項至少應有一位低分組考生選答，且選擇錯誤選項的考生中，高分組人數應少於低分組人數 (Ebel & Frisbie, 1991; 王文中、呂金燮、吳毓瑩、張郁雯、張淑慧，2004；余民寧，2011；郭生玉，2004；陳英豪、吳裕益，2001)。試題選項的誘答力分析，可作為測驗編製者評估試題品質的參考依據之一，誘答功能不佳的試題，應考慮修改或刪除。

上述困難度、鑑別度和誘答力等三種試題參數指標，已有學者出版相對應之測驗統計軟體計算，如Tester for Windows 程式3.0版，有興趣的讀者可參閱余民寧 (2011) 所著《教育測驗與評量：成就測驗與教學評量》。參考試題分析的結果，其數據可提供測驗編製者作為挑選優良試題之準則，然而，試題挑選的準則仍須視測驗目的而定，並非一成不變。郭生玉 (2004) 建議，先挑出鑑別度較高的試題，再從中挑選難度較為適中之題目，故以下將先從鑑別度說明，提供幾位學者之建議供讀者參考。

(1) 鑑別度

Noll、Scannell和Craig (1979) 建議，鑑別度指標至少需為0.25以上，如低於鑑別

度值之試題，應視為品質不佳之試題。Ebe 和 Frisbie (1991) 提出鑑別度值0.40以上，是優良試題；鑑別度值 0.30～0.39，是良好試題，但可能需修改；鑑別度值 0.20～0.29，是尚可試題，需作局部修改；鑑別度值0.19以下，是品質不佳試題，可考慮刪除或修改。

(2) 困難度

困難度指標方面，測驗學者均建議挑選難易適中，即困難度值接近0.50的試題，因當試題的難易度值適中，其鑑別度值是最大的，然而要同時符合鑑別度佳且困難度又接近0.50的試題是不多的，有其實務運作之困難；因此，Ahmanan 和 Glock (1981) 建議選擇題之難度值應介於0.40～0.80之間。

(3) 誘答力

余民寧 (2011) 建議：

(a) 在不正確的選項上的選答率，低能力組別受試者不可以為0；

(b) 在不正確的選項上的選答率，低能力組別受試者不可以低於高能力組別之受試者。

如果某些選項沒有受試者選答，或出現違反上述兩個判斷原則，則表示選項不具誘答力，宜考慮修改或重新設計。

2. 測驗分析

信度 (reliability) 與效度 (validity) 可反映測驗的優良程度 (周文欽等，1995；Gronlund, 1993; Huck, 2011)，信度表示測驗結果的一致性或穩定性，也就是測驗分數的可靠性，而效度則說明測驗確實能測量的特質，也就是測驗結果的正確性。信度是效度的必要條件，但非充分條件，有效度必定有信度，有信度不一定有效度。因此，測驗的效度，必須以信度為前提，因為不可信的測驗並無法保證測量的結果是有效的。

(1) 信度分析

信度是指測驗的可靠性、一致性、穩定性或準確性，與變異數 (variance) 和量度誤差 (measurement error) 二個統計量有關。它說明同一份測驗重覆測量某項特質時，得到相同結果的程度，或前後二次測驗分數一致的程度。一份具有鑑別度的測驗，其測驗成績的變異數較大，測驗的信度也較高。而測驗成績的可信度高時，表示測驗本身是準確的，量度誤差較小；若測驗的信度較差，則測驗的量度誤差較大。因此，測驗編製完成後，必須考慮測驗信度以確定測驗是否可信。考驗信度的方法有再測信度 (test-retest reliability)、複本信度 (alternative form reliability)、內部一致性信度 (internal consistency reliability) 和評分者信度 (inter-rater reliability) 等。

(a) 再測信度

再測信度指的是，對同一批考生，以同一份測驗卷，於不同時間重複施測二次。二次測驗得分的相關係數 (皮爾森積差相關係數)，該係數即為再測信度係數 (test-retest reliability coefficient)。隨著時間的流逝，再測信度可估計測驗結果是否保持穩定，因此又稱為穩定係數 (coefficient of stability)。

(b) 複本信度

複本測驗指的是，測量的潛在特質或能力相同，施測的時間長度相同，且試題的格式、數目、難度相當，但試題內容不同的二份測驗。二份複本測驗對同一批考生施測後，二次測驗得分的相關係數即為複本信度係數，簡稱複本信度。複本信度越高，測量考生的潛在特質或能力時，二份測驗的測量結果具有越高的一致性，且在測驗範圍內，二份複本測驗中的樣本試題具有越高的代表性。

(c) 內部一致性信度

分析再測信度或複本信度時，無論二次施測或二份測驗，目的都是為了取得二次測驗結果，進而計算二者間的相關係數；而內部一致性信度則是運用一次測驗的結果估計測驗的信度，目的在於簡化施測程序，同時能正確估計信度，這種方式估計得到的信度係數稱為內部一致性信度係數 (internal-consistency reliability coefficient)。內部一致性信度可反映測驗的同質性、一致性或穩定度，檢查個別試題與整份測驗的功能是否一致，同質性越高，代表測驗包含的試題是測量相同的特質。常見的估計方法有折半方法 (split-half method)、K-R 方法 (Kuder-Richardson method) 和 Cronbach α 方法 (Cronbach's alpha method) 三種。折半方法是將測驗的所有題目，平均分成兩部份，分別計分後，再根據兩個「半測驗」的分數，計算其相關係數。K-R方法適用於「對」或「錯」的二元計分測驗，並假設試題不受作答速度的影響。而Cronbach's α 則可用於多元計分的測驗，Cronbach's α 值大於 0.7 者表示具有高信度，小於 0.35 則為低信度 (Cuieford, 1965)。

(d) 評分者信度

若測驗屬於主觀測驗 (例如，論文題)，或採用觀察法、判斷法或評定量表法時，評分結果受到評分者的主觀意識影響，因而出現評分者誤差，此時有必要以「評分者信度」估計不同評分者之間的評分一致性，作為測驗使用者的參考。常用的評分者信度有評分者間 (inter-rater) 的評分者信度和評分者內 (intra-rater) 的評分者信度二種，前者為不同評分者對相同受試者的評分一致性估計，而後者為相同評分者對相同受試者的評分一致性估計。若同一位評分者嚴格依照評分標準給分，並且在評分過程中保持一致，這樣的評分結果具有較高的評分者內的評分者信度。若多位評分者對考生得分高低的排序是相近的，即一致認為某位考生應得到高分，而某位考生只能得到較低的分數時，這樣的評分結果具有較高的評分者間的評分者信度。

效標參照測驗中，測驗分數是決定或判斷考生是否達到精熟的重要依據，因此，效標參照測驗首重「決定」的正確性，其次才是「估計」的精確性。在效標參照測驗的目的是決定考生是否達到預設的精熟標準，在這個標準中，由於多數考生可以達到某個預設的精熟標準，因此，考生得分的變異數極小，甚至可能趨近於零。在這個情況下，常模參照測驗使用的信度係數估計法便不適用，效標參照測驗可採用百分比一致性指標 (percent agreement, PA) 計算測驗的信度係數。百分比一致性指標分析前後二次分類決定結果是否一致，並以其百分比比值的總和表示。假設100位考生接受二次測驗，由於測驗的試題並不相同，二次測驗的結果略有差異，若其中60位考生二次測驗均達到精熟標準，另20位考生二次測驗均未達到精熟標準，則百分比一致性指標 P_A 為

$$P_A = \frac{60}{100} + \frac{20}{100} = \frac{80}{100} = 0.80。$$

若二次測驗的結果差異相當大，其中只有6位考生二次測驗均達到精熟標準，2位考生二次測驗均未達到精熟標準，則百分比一致性指標 P_A 為

$$P_A = \frac{6}{100} + \frac{2}{100} = \frac{8}{100} = 0.08。$$

由上述二個例子，分類的決定越一致，百分比一致性指標 P_A 越接近1，所使用的效標參照測驗具有較高的信度係數，即表示所採用的分類標準 (即效標) 較為適當，區分精熟與未達精熟的能力具有一致性。

關於測驗信度的詳細內容可進一步參考相關的著作及文獻 (余民寧，2011；Carmines & Zeller, 1979; Cohen, 1960; Dick & Hagerty, 1971; Gronlund, 1993; Kaplan & Saccuzzo, 2008)。此外，影響信度的主要因素包含「試題數量」、「試題難度」、「施測對象」和「施測過程」等。

(a) 試題數量

測驗是測量的一個樣本，因此，取樣是否合理，必然影響測驗的信度。試題的數量太少，不足以代表完整的課程內容時，測驗的信度較低。增加試題數量是提高信度的有效方法。然而盲目增加試題的數量並不一定會提高測驗的信度，除了試題的取樣必須有代表性之外，增加試題數量的效果是遞減的，過多的試題考生無法確實作答，反而會降低測驗的信度。

一份測驗應有的試題數量並沒有絕對的標準，教師自編測驗時，應根據教學目標、教材內容、測驗目的等因素，決定雙向細目表中的題數和比重，但也可以依據教學現場的反應，適當調整雙向細目表中的教學目標、教材內容、試題數量，控制試題與試卷的品。教師可參考測驗目的、測驗題型、信度分析、學生年齡、學生程度、作答時間等，彈性調整測驗適合的試題數量。

(b) 試題難度

試題的困難度和信度並無直接的關係，然而試題對某些考生過於困難或簡單時，測驗分數的變異數較小，信度也將降低。因為試題過於困難時，考生可能會盲目猜測答案，作答反應接近隨機分佈，因此測驗結果的信度極低；若試題過於簡單，幾乎全體考生均能正確作答，則測驗分數的分佈集中，信度也隨之降低。

(c) 施測對象

接受測驗考生的心理素質各不相同，面對壓力時的反應不一，應試的心理壓力可能提高某些考生的專注力與反應速度，也可能使某些考生產生負面消極的應試心理。每個心理因素除了影響測驗成績，對測驗的信度也可能產生負面的影響，是所有因素中最難控制的部分。

(d) 施測過程

施測過程中，施測人員的素質和施測環境也可能影響測驗的穩定性。考場悶熱、座位擁擠、考試秩序混亂、試場周圍吵雜等，都會導致測驗信度下降。此外，施測人員未依規定執行試場規則，擅自提早或延後收卷，也是影響測驗信度的因素。

(2) 效度分析

效度指的是測驗的正確性，由測驗結果解釋與運用的觀點而言，測驗的效度說明運用測驗結果作出解釋與決策的正確程度 (余民寧，2011；Thorndike & Thorndike-Christ, 2010)。測驗編製的過程中，應儘可能的增加明確的證據與論證以支持並提升測驗的效度。若測驗效度無明確的佐證，測驗本身將不具意義，而使用測驗結果作為決策的適切性也將受到質疑。

測驗效度分析的目的在鑑定測驗是否能對預計測量的行為特質發揮測量的功能。若測驗的效度低，表示該測驗無法達到預期的功能。測驗的效度與測驗的目的息息相關，因此，鑑定測驗的效度時，必須以測驗的目的為基礎。效度可分為內容效度 (content validity)、構念效度 (construct validity) 和效標關聯效度 (criterion-related validity) (American Educational Research Association, 1999)。

(a) 內容效度

內容效度是指測驗內容適當的程度，成就測驗藉由內容效度判斷測驗的內容是否符合測驗的目標，考慮測驗試題的內容適當性及取樣代表性，內容適當性可釐清課程的內容範圍，確定測驗的全部的試題均在此範圍內；而取樣代表性則進一步確定出自內容範圍的試題是否具有代表性。選擇試題時應根據課程內容和教學目標的重要性而非隨機取樣，確保試題能涵蓋主要的內容範圍，並具有適當的分配比例，避免出現過於冷僻的試題，因此，擬題、選題時應根據雙向細目表的規劃，以提高測驗的內容效度。

(b) 構念效度

構念效度檢視測驗是否具有測量心理學理論中某個概念或特質的能力。構念 (construct) 是指心理學理論涉及的抽象的假設性概念、特質或變項，例如智力、焦慮和動機等。構念效度則說明測驗的意義，由心理學的理論觀點詮釋測驗的結果。構念效度的重點在於理論上的假設和對理論假設的考驗。構念效度必須由某個理論基礎出發，針對測驗相關的心理功能或行為，導出相關的基本假設並建立架構，據以設計和編製測驗，施測後根據結果檢視是否符合假設，若不符合，則修改測驗或重新檢討理論及假設的適當性，以得到具有良好構念效度的測驗。

(c) 效標關聯效度

效標 (criterion) 是指運用測驗預測的某種特質或行為的標準，而效標關聯效度則是以實證分析方法探討測驗結果與效標相關的程度，因此，又稱為實證效度 (empirical validity) 或統計效度 (statistical validity)。建立效標關聯效度時，困難之處在於不易取得適當的外在效標 (external criterion)。外在效標是測驗分數所預測的某些行為或表現的標準，例如，學業成就、特殊訓練的表現、實際工作表現、評定成績，以及現存的可用測驗等 (Anastasi, 1997)。若測驗分數與外在效標的相關越高，表示效標關聯效度越高，測驗分數越能有效解釋並預測外在效標行為或表現標準。

根據效標資料搜集的時間，效標關聯效度可分為同時效度 (concurrent validity) 和預測效度 (predictive validity)。同時效度指測驗結果與當前效標的相關程度，測驗分數與外在效標約在同一時間內連續取得，目的在以測驗分數估計個人目前於外在效標的表現。而預測效度與預測將來結果的測驗有關，測驗分數與外在效標約的取得有時間差，先取得測驗分數，相隔一段時間後再取得外在效標，目的在於以測驗分數預測個人未來於外在效標的表現。效標應該是有效的、可靠的、客觀的，檢驗測驗的效標關聯效度時，重點在於尋找合適的效標，不適合的效標仍然可求得效標關聯效度，但並不能顯示測驗是否達到可被接受的效度。

伍、結語

「測驗」在「教」與「學」的過程中，始終扮演重要的角色。雖然近年來已有許多提倡多元評量的觀點和技術，但是為了儘可能了解學生的能力、興趣和進步的情形，教師自編的紙筆測驗仍是最常採用的教學評量工具。編製符合需求且適合學生的測驗，也是每位教師不可或缺的基本能力。「測驗」最重要的優點在於客觀，不容易受特定的主觀因素影響且能在短時間內同時蒐集大量的資料，是蒐集學生能力與學習狀況的最便捷方法。

編製測驗是創造力的工作，編製時應就實際的教學需求，彈性運用相關的參考原則，以獲得實用的回饋資訊。學生可透過測驗了解自己對學習內容的熟悉程度；而教師也可藉由測驗檢視自己的教學活動是否達成預期的教學目標。編製測驗試

題，對經驗豐富的教師而言應相當駕輕就熟，但檢驗測驗試題是否適合施測的學生，卻不能以「經驗」作為唯一的依據。根據測驗及試題分析理論得到的數據，能了解測驗試題是否優良，並進一步探討測驗結果中隱含的意義。

了解並運用試題及測驗的分析方法及相關數據代表的意義，有助於改善教師的教學品質，提升學生的學習動機，進而提高教師的教學成效與學生的學習自信。教師根據量化數據評估測驗試題是否合適，若不合適應由何處著手修改。教師同時可根據量化數據了解學生的學習狀況，或本身的教學策略是否適合任教的學生。除了傳統的測驗分數，教師也可根據分析的結果，了解教學策略是否合適，分析學生的學習狀況，挖掘學生在學習過程中，可能遭遇的困難、問題及觀念不清之處，進而研擬補救對策以對症下藥。

「教學」、「學習」與「測驗」三者環環相扣，互相驗證，密不可分。本文於測驗的類型中說明各種測驗類型的功能、特徵及可能的施測時機；測驗編製的原則中簡述試題的編擬及評分原則以及優良試題的特徵；測驗編製的步驟則整理測驗編製可共同遵循的步驟，及提高測驗信度和效度的方法。測驗運用客觀的方法和技術，蒐集學生的學習行為及學習成效的相關資料，再根據教學目標，就學生的作答反應，進行分析、研究與評估。無論測驗採用的類型或題型為何，絕對不存在優或劣的比較，應就學科特性、教學情境及測驗目的等因素彈性運用，以發揮測驗應有的功能，縮短「教」與「學」的鴻溝，提升「教」與「學」的品質。

參考文獻：

王文中、呂金燮、吳毓瑩、張郁雯、張淑慧 (2004)。**教育測驗與評量——教室學習觀點（第二版）**。台北：五南。

王寶墉 (1995)。**當代測驗理論**。台北：心理。

余民寧 (1993)。測驗編製與分析技術在學習診斷上的應用。**教育研究**，**28**，44-60。

余民寧 (2009)。**試題反應理論(IRT)及其應用**。台北：心理。

余民寧 (2010)。測驗建置流程及新概念。**測驗及評量專論文集：題庫建置與測驗編製**。台北：國家教育研究院籌備處。

余民寧 (2011)。**教育測驗與評量：成就測驗與教學評量**。台北：心理。

李坤崇 (2004)。修訂Bloom認知分類及命題實例。**教育研究**，**122**，98-127。

李坤崇 (2006)。情意技能教學目標分類與評量。**教育研究**，**144**，123-133。

周文欽、歐滄和、許擇基、盧欽銘、金樹人、范德鑫 (1995)。**心理與教育測驗**。台北：心理出版社。

林世華 (2000)。由多元評量的觀念看傳統評量的角色與功能。**科學教育月刊**，**231**，

67-71。

洪碧霞、邱上真、林素薇、葉千綺 (1998)。國小中低年級國語文成就測驗題庫建立之研究。**測驗年刊**。**45** (2)，1-18。

郭生玉 (2004)。**教育測驗與評量**。台北：精華。

陳英豪、吳裕益 (2001)。**測驗與評量**。高雄：復文。

葉連祺、林淑萍 (2003)。布魯姆認知領域教育目標分類修訂版之探討。**教育研究**，**144**，94-106。

歐滄和 (1993)。標準化測驗的編製發展程序。**測驗統計年刊**，**1**，33-42。

劉湘川、蔡良庭 (2005)。成就測驗試題編製概要。**測驗統計簡訊**，**64**，1-12。

簡茂發 (2000)。**心理測驗與統計方法**。台北：心理。

Ahmanan, J.S., & Glock, M. D. (1981). *Evaluating student progress: Principles of tests and measurement* (6th ed.). Boston, MA:Allyn & Bacon.

Airasian, P.W., Miranda, H. (2002). *The role of assessment in the revised taxonomy. Theory into Practice*, 41 (4) , 249-254.

Allen, W. J., & Yen, W. M. (2001). *Introduction to measurement theory* (2nd ed.). Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author

Anastasi A. & Urbina S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (Eds.). (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman.

Baldwin, D., Fowles, M., & Livingston, S. (2005) , *Guidelines for constructed-response and other performance assessments.* (727534) Princeton, NJ:Educational Testing Service.

Berk, R.A. (1984). *A guide to criterion-referenced test constrcution*. Baltimore, MD:

The Johns Hopkins University Press.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I, cognitive domain*. New York: David McKay.

Brennan, R. L. (Ed.) (2006). *Educational measurement* (4th ed.). Washington, DC: National Council on Measurement in Education.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Dick, W., & Hagerty, N. (1971). *Topics in measurement: Reliability and validity*. New York: McGraw-Hill.

Cuieford, J. P. (1965). *Fundamental statistics in psychology and education.* New York: McGraw Hill.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall, 1991.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519-521.

Gronlund, N. E. (1993). *How to make achievement tests and assessments* (5th ed.). Boston: Allyn & Bacon

Gulliksen, H. (1987). *Theory of mental test.* Hillsdale, NJ: Lawrence Erlbaum Associates (Originally published in 1950 by New York: Johe Wiley & Sons).

Haladyna, T. M. (1996). *Writing test items to evaluate higher order thinking.* New York: Allyn & Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1989a) , A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 1, 37–50.

Haladyna, T. M., & Downing, S.M. (1989b) , The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 1, 51-78.

Haladyna,T. M., Downing, S.M.,& Rodriguez, M.C. (2002) , A review of multiple-choice item-writing guidelines for classroom assessment, *Applied Measurement in Education,* 15 (3) , 309-334.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory.* Newbury Park, CA: Sage.

Hogan, T. P., & Murphy, G. (2007). Recommendations for Preparing and Scoring Constructed-Response Items: What the Experts Say, *Applied Measurement in Education,* 20 (4) , 427-441.

Huck, S. W. (2011). *Reading statistics and research* (6th ed.). Boston, MA: Pearson.

Kaplan, R. M., & Saccuzzo, D. P. (2008). *Psychological testing principles applications and issues.* (7th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Kemp, J. E. (1985). *The instructional design process*. New York: Harper & Row.

Koretz, D.M. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator,* Summer, 12 (2) : 8-15, 46-52.

Krathwohl, D. R., Bloom, B. S. & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II, affective domain.* New York: David McKay.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Linn, R. L., Miller, M. D., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Noll, V. H., Scannell, D. P. & Craig, R. C. (1979). *Introduction to educational measurement* (4th ed.). Boston, MA: Houghton Mifflin.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (2rd ed.). New York: McGraw-Hill.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. Orlando, FL: Academic Press.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal for Educational Research*, 2, 49-60.

Simpson, E. J. (1972). *The classfication of educational objectives in the psychomotor domain. The Psychomotor Domain* (Vol. 3). Washington, DC: Gryphon House.

Suen, H. K. (1990). *Principles of test theories*, Hillsdale. NJ: Lawrence Erlbaum Associates.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Upper Saddle River, NJ: Pearson / Merrill Prentice Hall.