
# 測驗之編製

## 命題技巧與測驗資料之分析

蕭儒棠/曾建銘/吳慧珉/林世華/謝佩蓉/謝名娟
合 著

國家教育研究院
National Academy for Educational Research
www.naer.edu.tw

ISBN 978-986043496-5 定價 170元
GPN 1010302916

測驗之編製—命題技巧與測驗資料之分析

國家教育研究院


# 測驗之編製

## 命題技巧與測驗資料之分析

蕭儒棠/曾建銘/吳慧珉/林世華/謝佩蓉/謝名娟
合 著

# 序

不管是國際學生能力評量計劃（PISA）或是促進國際閱讀素養研究（PIRLS）的調查顯示，臺灣學生在數學與閱讀的素養，明顯地提升。這代表過去數年來，政府對於教師進修、基礎教學研究以及推廣的成效。

十二年國民基本教育強調因材施教、適性揚才理念，如何透過測驗、評量、問卷瞭解學生的興趣、性向、潛能、現有知識狀態，面對教學時知識狀態的轉變，無疑是重要的事情。如何讓教師具備編制、使用測驗，與分析及判讀測驗結果的能力，是實踐適性揚才的重要手段。

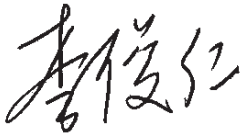
本書的產生，正是基於此目標：讓現場教師熟悉測驗評量的基本概念，能解讀測驗結果，甚至具備動手編制測驗以及分析測驗的能力。

本書共收錄5篇文章，首先，蕭儒棠老師介紹測驗的類型、編製基本原則與步驟，強調因應學科特性、教學情境與測驗目的整體考量。曾建銘老師介紹選擇反應試題之編寫原則，並分科說明不良試題、修審方向，與修訂試題範例。吳慧珉老師說明新一代的測驗分析方式，試題反應理論，以及如何解釋測驗分析的結果。林世華老師、謝佩蓉老師以試題反應理論的Rasch模式為主軸，並介紹等化設計的概念，除理論外，並輔以實例，希冀提供教師們簡易操作以及解讀的程序。最後，謝名娟老師針對實作評量提供實施的步驟與準則，作為教師進行實作評量的參考。

教師能增能，學生則受益。期盼本專書的出版，能協助教師提升編制以及運用測驗、評量、問卷的專業知能。

感謝各篇作者們對於撰寫各篇章的投入與付出，更感謝審查委員細心的評閱與指教，使本專書得以順利呈現。

國家教育研究院測驗及評量研究中心 主任



# 目錄

# 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，$p_j$ 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 $\alpha$ 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 $\alpha$ 則可用於多元計分的測驗，Cronbach's $\alpha$ 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.

# 選擇題命題原則與不良題範例

➢ 曾建銘

# 選擇題命題原則與不良題範例

**曾建銘**
國家教育研究院副研究員

測驗最重要是要具有高的效度，即測驗的結果能符合其目的。根據測驗編製的步驟，在確定測驗目的與編製計畫、教學目標及課程內容後，接下來就要訂定測驗藍圖及測驗題型 (Linn & Gronlund, 2000; 余民寧，2012)，但在有限的測試時間下，測驗內容不能涵蓋所有的教學目標和課程內容，為了使測驗內容能具有較高的代表和反映各種認知層次的相對權重，命題應該就測驗目的、測驗題型、教學目標和課程內容，完成規劃測驗藍圖，將測驗藍圖變成雙向細目表，以保證測驗的品質，準確的達到測驗的目的，使測驗具有高效度。如何根據雙向細目表來命題、修審為一份測驗良窳的關鍵。

## 壹、試題編寫的一般原則

試題由「題幹」(stem) 和「作答反應」(response) 組成。「題幹」描述試題的問題情境，是一個完整的敘述或問句，有時也包含相關的圖表、圖形或照片，是用於詢問學生的一組「刺激材料」(stimulus material)。「作答反應」是學生根據題幹的敘述或提問，所完成的回答紀錄，例如，「是」或「否」、選項代碼、字詞、文句、短文、計算過程、……等。

編擬試題可參閱許多相關的文獻 (余民寧，2011；簡茂發，2000；Brennan, 2006; Downing, & Haladyna, 2006; Haladyna, 1996; Haladyna, 2004; Roid & Haladyna, 1982)，上述文獻內容包含相關的研究成果與經驗，是提高測驗品質的參考資料。雖然測驗類型眾多，性質與功能也不相同，且各種題型有其特殊的編製要領，但編製時仍應遵守某些「試題編寫的一般原則」，以提高測驗的信度和效度。

### 一、試題應依循測驗目的

測驗的目的有許多種，是作為選拔、診斷之用，抑或作為分類之用。測驗目的不同，編製測驗的取材範圍、試題題型及試題難度也隨之改變。

### 二、試題內容應具代表性

由於測驗內容無法涵蓋全部教學目標與課程內容，只能測量其中的某些樣本，因此，試題內容的取樣應有代表性，不能偏重或忽視任何內容領域，以提高測驗的效度。

### 三、試題的題型應多樣化

「選擇反應試題」所需的作答時間較短，測驗可涵蓋的範圍較大，可測量的內容較多，而「建構反應試題」則可有效測量考生對於理解與應用等較高層次的能力。一份優良的測驗，應依據測驗的目的與要求，配合題型的特徵與功能，選用各種不同的題型。

### 四、試題文句應簡潔明確

試題的敘述應以淺顯、簡潔為原則，避免艱深的字詞和文句。編擬試題時除了應避免與課程內容和教學目標無關的資訊，也必須確保試題的敘述包含解題需要的資訊，避免無法作答的無效試題。

### 五、試題答案應明確可靠

試題的正確答案必須明確，並避免任何可能的疑慮與爭論，以提高測驗的信度與效度。編擬試題時除了應提供試題的正確答案，也應說明試題編擬的理念，協助其他人理解試題的測驗目標，作為未來試題修審、組卷或評分的參考。

### 六、測驗試題應彼此獨立

試題不應含有任何回答該題或其它試題的線索。若某一道試題的敘述或答案提供另一道試題的解題線索，則試題作答的正確與否，將無法準確反應考生的學習狀況，即學生是真的學習後了解做對，還是因其它試題提供解題線索而答對。

## 貳、選擇反應試題的編寫原則

編寫試題應考慮測驗目的，以達到鑑別考生程度的目的，過程中應遵守某些編寫原則。1989 年 Haladyna 與 Downing 分別就「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) 等六大面向，列舉優良單選選擇題的編寫原則，作為編寫試題時遵循的參考 (Haladyna & Downing, 1989a; Haladyna & Downing, 1989b)。

隨著教育研究及理論的發展與演進，研究者對優良單選選擇題的編寫原則，

提出進一步的修正，並補充了更多的實證證據 (Haladyna, Downing, & Rodriguez, 2002)。修正後的編寫原則包含「內容」(content)、「格式」(formatting)、「風格」(style)、「題幹」(stem) 和「選項」(option) 等面向。

## 一、內容編寫原則 (以下內容參考 (Haladyna, Downing, & Rodriguez, 2002))

「內容」是試題編寫中最重要的面向，學科專家的專業知識是編寫優良試題的重要因素，關於試題「內容」的編寫原則包括：

(一) 試題應根據測驗說明 (雙向細目表、測驗藍圖) 的規劃，且試題內容應反映特定的內容及某項具體的心理行為。
(二) 試題應建立於重要的學習內容基礎，避免測驗瑣碎的內容。
(三) 運用新的材料測驗較高層次的學習成果。試題應避免教科書或課堂上使用過的詞句，避免考生僅憑記憶作答。
(四) 確保測驗中每道試題的內容彼此獨立。
(五) 編擬單選選擇題時，應避免過於侷限性或過於一般性的內容。
(六) 避免以意見為主的試題 (opinion-based items)。
(七) 避免陷阱題 (trick items)。
(八) 確保試題使用的詞彙對考生而言是簡單的。

## 二、格式和風格相關的原則

(一) 使用傳統單選選擇題、二選一題 (alternate choice)、判斷正誤題 (true-false, TF)、多項判斷正誤題 (multiple true-false, MTF)、配合題 (matching)、依據上下文回答題以及試題組形式，但是避免使用複雜單選選擇題型 (K型)。
(二) 試題應縱向排列，而非橫向排列。
(三) 確定試題的文法、標點、字母大小寫以及拼寫正確無誤。
(四) 將每道試題的閱讀量降至最低。

## 三、題幹編寫原則

(一) 確保題幹提供的指引明確。
(二) 試題的核心概念應於題幹中敘述，而非選項。
(三) 試題的敘述應簡潔明確，避免非必要的詞句 (過度繁瑣)。
(四) 題幹應使用肯定語句，避免否定語句，如「不是」或「除了……」等詞句。使用否定語句應謹慎，運用時應以「大寫」或「黑體」標出。

## 四、選項編寫原則

(一) 提供的有效選項越多越好 (研究顯示3個選項應已足夠)。

(二) 確定選項中只有一個是正確答案。
(三) 根據正確選項數目調整正確答案的選項位置，避免正確過度集中於某一或二選項。
(四) 選項若是數字或有長短不一，位置需依邏輯或數值的順序排列。
(五) 確定選項的獨立性，如選項間的範圍不應相互重疊，以避免有兩個以上之正確答案，而產生疑義。
(六) 選項的內容和文法結構應同質 (homogeneous)，以避免學生很容易排除非同質性之選項，降低誘答選項之功能。
(七) 選項的文句長度應相當。
(八) 謹慎使用「以上皆非」(none-of-the-above) 的選項。
(九) 避免使用「以上皆是」(all-of-the-above) 的選項。
(十) 選項儘量用肯定詞，避免否定詞。
(十一) 避免提供正確答案的線索，例如：
(1) 特定限定詞，包含「總是」(always)、「從不」(never)、「完全」(completely) 和「絕對」(absolutely)；
(2) 避免題幹中出現與正確選項有關或相近的詞彙；
(3) 避免題幹及選項敘述中「不一致的文法」提供考生選答的線索；
(4) 避免顯而易見正確的選項；
(5) 避免某二個或三個選項提供考生正確選項的線索，即非正確選項要具誘答；
(6) 避免荒謬、可笑的選項。
(十二) 誘答項必須是合理的。
(十三) 根據學生的典型錯誤撰寫誘答項。
(十四) 若教師以及學習環境允許，試題中可加入幽默元素。

## 參、不良題範例

測驗題型可分為選擇反應題與建構反應題兩大類，選擇反應題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；但比較不適合測量評鑑與綜合的能力 (余民寧, 2012)，而其中之選擇題 (multiple-choice item) 仍為中等以下各級學校教師所廣泛運用。因此，以下採分科就選擇題命題時需注意之原則提出不良試題範例，包含不良題之分析、修審方向及修訂試題範例等，供一般教師命題時參考。

### 一、國語文

**(一) 題幹應為完整句並以簡單而清晰的用詞來陳述問題。**

不良試題範例

1. 「張飛穿針」之於「大眼瞪小眼」，猶如「諸葛亮借箭」之於

①有借有還
②不自量力
③有借無還
④小題大作

【說明】：
題幹不完整，因此於最後加上「何者」，使題幹語意完整。

修訂試題範例
「張飛穿針」之於「大眼瞪小眼」，猶如「諸葛亮借箭」之於下列何者？
①有借有還
②不自量力
③有借無還
④小題大作

**（二）選項不宜插在題幹中間，以免題幹分裂為二；尤其不宜將選項置於題幹之前。**

不良試題範例
1. 「老闆和店員」之間的關係，就像
①同學和朋友
②丈夫和妻子
③將軍和士兵
的關係？

修訂試題範例
「老闆和店員」之間的關係，就如同下列哪一組選項的關係？
①同學和朋友
②丈夫和妻子
③將軍和士兵

【說明】：
原題幹被選項分裂為二，因此將題幹重新敘述為完整句。

**（三）題幹應避免包括一個以上的問題，否則將減少試題的診斷價值。**

不良試題範例
閱讀下文，哪一個選項是這段文字的涵義？句中將人與樹並列，其理由最不可能是下列何者？
**人和樹，樹和人，我們都是大地上的風景。**
①都是大自然的一份子
②最終都要回歸大自然

③最漂亮的風景是人和樹
④都需要水和空氣的滋潤

【說明】：
題幹問了兩個問題，究竟是要回答前者或後者，容易產生試題疑義。

修訂試題範例
閱讀下文，哪一個選項不是句中將人與樹並列的理由？
**人和樹，樹和人，我們都是大地上的風景。**
①都是大自然的一份子
②最終都要回歸大自然
③最漂亮的風景是人和樹
④都需要水和空氣的滋潤

【說明】：
根據命題原則將第一個問題刪除，就單一概念進行測驗。

**（四）避免在題幹上堆砌一些不切題且毫無作用的材料。**

不良試題範例
下列□□□□中，宜填入哪個成語最適當？
**審查國小課本，發現抒情、敘事的文章多半問題不大，而只要一遇到議論文，無不讓審查者傷透腦筋，究其原因，恐怕也是因為道理真的不容易用寫的。所以，這類文章大多說辭夾纏，每段都大同小異地重複上段同樣的論點，寫那樣的文章讓學生讀，也難怪學生無法心領神會！這部份和真實的人生頗有相互映照的趣味。大多數喜歡講道理的人，都潛藏囉唆的特質，無法□□□□，常常一發不可收拾，讓人聽了備感焦慮，焉能指望這樣的道德教訓會有效果！尤其大人總是喜歡在生氣時才對孩子說道理，這時候，再有理的話也總夾帶著幾分的憤怒和威權，孩子哪裡聽得進去！**
①言簡意賅
②舉一反三
③移樽就教
④心領神會

修訂試題範例
下列□□□□中，宜填入哪個成語最適當？
**大多數喜歡講道理的人，都潛藏囉唆的特質，無法□□□□，常常一發不可收拾，讓人聽了備感焦慮。**
①言簡意賅

②舉一反三
③移樽就教
④心領神會

**（五）題幹陳述盡可能簡明，避免不必要的複雜用字或句型結構，否則將變成閱讀、理解能力測驗。**

不良試題範例

世棒賽正如火如荼在臺灣舉行，球迷絞盡腦汁，搬出個人的國文造詣，創造出許多令人拍案叫絕的創意標語。有些標語使用了修辭學上的「雙關」用法，試問下列有幾個標語和「韓冤莫白」使用相同的修辭方式？
甲、終日之戰　乙、荷苦來哉　　丙、后羿射日
丁、美終不足　戊、吞下高麗蔘　己、血流成荷
庚、義想天開
①三個
②四個
③五個
④六個

修訂試題範例

有些標語使用了修辭學上的「雙關」用法，試問下列有幾個標語和「韓冤莫白」使用相同的修辭方式？
甲、終日之戰　乙、荷苦來哉　　丙、后羿射日
丁、美終不足　戊、吞下高麗蔘　己、血流成荷
庚、義想天開
①三個
②四個
③五個
④六個

**（六）如果由於需要，必須在題幹中採用反面敘述時，則要特別強調反面或否定的字詞，或在這些字詞底下畫線。**

不良試題範例

下列句子的修辭用法，那一個與其他三者不同？【句型和修辭】
①就算整個世界被寂寞綁票，我也不會奔跑＜蘇打綠・小情歌＞
②從明天到永遠，你我不停的犯錯＜劉若英・冰點＞
③我要我的世界口味最特別，感動加上調味就會很完美＜張紹涵・C大調＞
④當你的笑容撞進心中，心跳突然定格＜曹格・愛到底＞

修訂試題範例

應在不同兩字畫底線，如不同，餘皆相同。

【說明】：

1. 在設計題目時，應該盡可能以正面陳述的方式設計題幹和問題，如果迫不得已 (例如正向且誘答性高的答案很難設計) 必須在題幹中採用反面敘述時，則要特別標註強調反面或否定的字詞，使學生專注在問題上。
2. 一份試卷和題本內，這種負面問法的題目數量不宜過多

**（七）在各選項中共同文字應置於題幹內**

不良試題範例

閱讀下文，並判斷其主要在說明什麼？

**既相齊，食不重肉，妾不衣帛**

①描寫晏子力行的節操

②描寫晏子節儉的美德

③描寫晏子持家的規矩

④描寫晏子重義的精神

修訂試題範例

閱讀下文，並判斷其主要在描寫晏子哪一點？

**既相齊，食不重肉，妾不衣帛**

①力行的節操

②節儉的美德

③持家的規矩

④重義的精神

**（八）題幹是否避免使用雙重否定**

不良試題範例

下列選項中的成語，何者沒有錯字？

①蠶食京吞

②裹足不前

③以蠡測海

④力網狂瀾

修訂試題範例

下列選項中的成語，何者用字完全正確？

①蠶食京吞

②裹足不前

③以蠡測海
④力網狂瀾

**（九）答案必須明確，對單一選擇題而言，僅可供一個正確選項。**

不良試題範例

「如何棕黑色的泥土竟長出灰褐色的枝子，如何灰褐色的枝子會溢出深綠色的葉子」，關於這段文字，主要傳達什麼寓意？
①讚頌大自然的生命力
②讚頌造物者的神祕力
③讚頌春天獨一無二的神蹟
④讚頌春天無以名之的美麗

【說明】：
此題原應要考原作者對春季的讚美，但只從原文中截取出兩句要判斷寓意，其實每一個答案皆可，因此會產生疑義。

**（十）避免在題幹與正確答案中，使用相同的字詞。**

不良試題範例

曾國藩說：「**弟之廉人人料之，其不儉，則阿兄所不及料也**」，主要在勸勉澄弟：
①修養「廉」潔
②培養「儉」德
③加強「忍」德
④戒除「貪」心

【說明】：
1. 題幹宜改為疑問句，讓語意敘述完整。
2. 題幹與正確選項皆出現儉，容易被學生猜對

**（十一）避免正確答案敘述較為詳細，或特別突出。**

不良試題範例

1.下列哪一個語詞所表示的時間最「長」？
①「片時」之間，功課就寫完了
②一「霎時」，天空下起傾盆大雨
③他按下快門，捕捉「剎那」美景
④在「悠悠」的歲月裡，媽媽的黑髮刷上了白

不良試題範例

閱讀下文，哪一個選項可以說明這段文字的涵義？

**我覺得發明補習的人應該是馴獸師，他可能是從訓練動物得到的靈感。因為補習班就是讓你一直練習，一直練習，直到寫對為止。不過，被訓練的動物如果做對了，會有糖吃；而補習卻得自己繳費練習。這是兩者間最大的不同。**
①補習會讓成績更優秀
②鼓勵大家成為馴獸師
③補習班應該提供糖果
④補習班的大量練習像是把學生當作動物來訓練

修訂試題範例

閱讀下文，哪一個選項可以說明這段文字的涵義？

**我覺得發明補習的人應該是馴獸師，他可能是從訓練動物得到的靈感。因為補習班就是讓你一直練習，一直練習，直到寫對為止。不過，被訓練的動物如果做對了，會有糖吃；而補習卻得自己繳費練習。這是兩者間最大的不同。**
①補習會使學生成績變好
②鼓勵大家成為動物馴獸師
③補習班應該提供糖果給學生
④大量的練習讓人失去學習興趣

**（十二）避免意義相同的選項。〔指誘答選項〕**

不良試題範例

下列哪一個選項填入□中最恰當？
**這條山路□□□□，你可要小心駕駛，注意安全！**
①蜿蜒曲折
②迂迴曲折
③曲折離奇
④百折不撓

修訂試題範例

下列哪一個選項填入□中最恰當？【能運用詞語】
**這條山路□□□□，你可要小心駕駛，注意安全！**
①蜿蜒曲折
②拐彎抹角
③曲折離奇
④百折不撓

【說明】：
第一與二的答案意義相同，如此將造成兩個答案皆可的疑義，因此將選項二改為拐彎抹角。

**（十三）避免錯誤的選項缺乏似真性以及選項應具誘答力。**

不良試題範例

下列哪一個選項填入□中最恰當？

**冷冽的風無情的吹來，大地一片□□□□，卻阻礙不了我們攀登山頂的決心！**

①春暖花開
②風調雨順
③陽光普照
④白雪皚皚

【說明】：
根據題意選項應為描述氣候不佳之選項，而原答案123皆不具誘答，因此進行修改。

修訂試題範例

下列哪一個選項填入□中最恰當？【能運用詞語】

**冷冽的風無情的吹來，大地一片□□□□，卻阻礙不了我們攀登山頂的決心！**

①寒風刺骨
②天寒地凍
③雪上加霜
④白雪皚皚

**（十四）避免使用「以上皆是」或「以上皆非」的選項。**

不良試題範例

(甲) 外祖不二日人問「遺」(乙) 遂營目前之務，而「遺」千載之功 (丙) 攀條折其榮，將以「遺」所思。上面句中的三個「遺」字的意思：
①甲乙相同
②乙丙相同
③甲丙相同
④三者皆相同

修訂試題範例

下列文句中「遺」字的解釋，何者與「齊桓公飲酒醉，遺其冠，恥之，三日不朝。」的「遺」字相同？
①外祖不二日人問「遺」
②攀條折其榮，將以「遺」所思

③遂營目前之務，而「遺」千載之功
④漢王亦因令良厚「遺」項伯，使請漢中地

## 二、英語文

### （一）避免在題幹上堆砌一些不切題且毫無作用的材料。

不良試題範例

看圖辨義，每題播出兩遍。請聽錄音機播出題目和三個選項，選出與所看到的圖畫最相符的答案。

Which is true about the picture?
①The office has two employees.
②The office is in a mess.
③The office has only one door.

【說明】：
1. 建議圖片應主題明確，圖片內容混亂且顏色不清楚，不必要的擺設及人物過多。
2. 選項②牽涉個人主觀判斷，容易引起爭論。
3. 選項③雖然是正確答案，但圖片中的門很小且不清楚，粗心的學生很容易沒有注意到。
4. 選項①過於簡單沒有誘答力。

修訂試題範例

Where are these people?
①They are in an office.
②They are on the street.
③They are at the beach.

【說明】：
經過試題的修審後，辦公室的主題在圖片中很明確，圖片變得簡單易懂，且答案非常明確。

**（二）如果採用「最佳答案」的形式，必須指出「在下述選項之中，那一項最適當」。**

不良試題範例

看圖辨義，每題播出兩遍。請聽錄音機播出題目和三個選項，選出與所看到的圖畫最相符的答案。


What is the girl doing?
①She is taking a picture of the boy.
②She is talking to the boy.
③She is helping the boy with his homework.

【說明】：
選項②和③也可以是正確答案，容易引起爭論。建議修改選項②和③。

修訂試題範例


What is the girl doing?
①She's helping the boy write his homework.
②She's taking a picture of the boy.
③She's jogging with the boy.

【說明】：
經過試題的修審後，答案非常明確，不至於誤導學生。

**（三）選項需具同質性。**

不良試題範例

小朋友，下列每題的三個選項中，只有兩個選項在字義上屬於同一種類群。請將不屬於該類群的選項挑出來，並塗在答案卡上。
① Friday　② night　③ Monday

【說明】：
選項②為正答，也是選項中唯一一個選項第一個字母為小寫的字，學生容易因此增加猜中正答的機會。建議將選項②修改為第一個字母大寫的字，以增加測驗學生能力的準確性。

修訂試題範例

小朋友，下列每題的三個選項中，只有兩個選項在字義上屬於同一種類群。請將不屬於該類群的選項挑出來，並塗在答案卡上。
① Friday　②Taiwan　③ Monday

【說明】：
經過試題的修審後，學生所看見的選項為三個同為大寫的專有名詞，作答時較

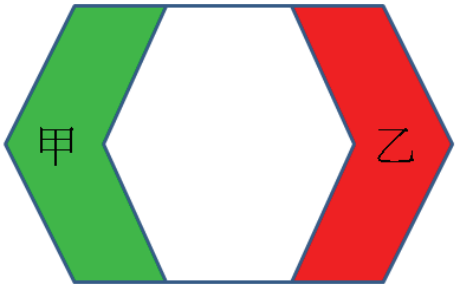
易針對試題的內容去做答案的判斷，而非似懂非懂的由同中求異的方式去選出答案。

## 三、數學

### (一) 題幹應為完整句並以簡單而清晰的用詞來陳述問題。

不良試題範例

將二個六邊形的紙部分重疊成如下圖，已知六邊形的面積為48平方公分，甲的面積是六邊形面積的$\frac{1}{3}$，甲和乙的面積共是多少平方公分？




① 16
② 32
③ 48
④ 144

【說明】：
題幹敘述不清，兩個六邊形是否全等，六邊形的面積是指一個六邊型的紙或中間空白的部分，皆未敘明清楚。

修訂試題範例

將二個全等六邊形的紙部分重疊成如下圖，已知每一個六邊形的面積為48平方公分，甲的面積是六邊形面積的$\frac{1}{3}$，甲和乙的面積共是多少平方公分？

### (二) 選項不宜包括專為粗心學生而設的陷阱。

不良試題範例

小華有一台電動玩具車，在啟動10秒後，才會以每秒鐘5公分的速度等速前進，則這輛車從啟動到行進了$a$公尺，共花多少秒？

① $\frac{a}{5}+10$　② $20a+100$　③ $10(1+2a)$　④ $\frac{a+10}{5}$

【說明】：

題幹中先說明以每秒鐘5公分的速度等速前進，但接下來是說“行進了 $a$ 公尺”，很明顯題幹中前後長度單位做了改變，若學生沒注意都用公分來作答，而選項①又是求出的結果，如此，粗心的學生將會選到答案①而不自覺。學生若不是猜而是粗心選①，以測驗目的而言學生已具備答對此題的能力，卻是因為粗心而錯，此非測驗的目的，建議將 $a$ 公尺修為 $a$ 公分，選項②議做調整；或者題幹不改，將選項①做調整，以測出學生是否具備解出該題的能力。

修訂試題範例

小華有一台電動玩具車，在啟動10秒後，才會以每秒鐘5公分的速度等速前進，則這輛車從啟動到行進了$a$公分，共花多少秒？

① $\frac{a}{5}+10$　② $50+a$　③ $10(1+2a)$ ④ $\frac{a+10}{5}$ 。

## 四、社會

### （一）不符合測驗目標

不良試題範例

這群原住民多半分布在臺東、屏東、高雄一帶，以華麗的衣物著稱，配戴百合花是婦女純潔高貴的象徵，族人以雲豹的後裔自居。請問：上述是在描述臺灣原住民哪一族的特色？

①賽夏族
②排灣族
③魯凱族
④泰雅族

【說明】：

1. 此題測驗目標為「探討臺灣文化的淵源，並欣賞其內涵」，旨在測量學生是否了解是魯凱族的文化特色。
2. 命題原意佳，但因題幹對於魯凱族特色的描述流於瑣碎記憶，可能影響教師要求學生記憶各原住民文化的瑣碎知識，致使學生發展出錯誤的學習方式-死背，故此題不適宜。
3. 測驗內容應凸顯各原住民文化的主要特色，強調其主要分布、祭典活動，以及為因應時代變遷的轉型過程，使學生學會欣賞傳統文化、尊重多元文化，不需過度著墨於細節瑣碎的部份。

修訂試題範例

這群原住民多半分布在臺東、屏東、高雄一帶，每年國曆八月十五所舉行的豐

年祭是其最重要的農耕禮儀祭典。請問：上述是在描述臺灣原住民哪一族的特色？

①賽夏族
②排灣族
③魯凱族
④泰雅族

【說明】：

1. 此題測驗目標為「探討臺灣文化的淵源，並欣賞其內涵」，旨在測量學生是否了解是魯凱族的文化特色。
2. 修訂後之試題強調魯凱族原住民文化的主要分布與最重要的祭典活動。學生能學會欣賞傳統文化、尊重多元文化

### （二）取材來源不適當

不良試題範例

爸爸、媽媽趁著暑假帶著曉美一起前往巴西拜訪移民的親戚，炎熱的天氣和廣大的熱帶雨林區，讓曉美留下深刻的印象。曉美心想：「熱帶雨林對世界生物那麼重要，對於維護熱帶雨林，我們可以做什麼呢？」請問：下列哪一個組合是正確的？

甲、重複使用信封
乙、種植綠色植物
丙、減少使用塑膠袋
丁、做紙類的資源回收

①甲乙丙
②甲乙丁
③甲丙丁
④乙丙丁

【說明】：

1. 此題測驗目標為「列舉地方或區域環境變遷所引發的環境破壞，並提出可能的解決方法」，旨在測量學生是否了解維護熱帶雨林的方式。
2. 型態為多重選擇題，會提高測驗複雜性，造成學生的作答壓力，對國小六年級的學生並不適切，應盡量避免使用。
3. 維護熱帶雨林不能僅單純由樹木的角度直觀思考，「甲、重複使用信封」、「乙、種植綠色植物」、「丙、減少使用塑膠袋」、「丁、做紙類的資源回收」四個選項皆為保護環境直接或間接的作法，選項的適切性具有爭議性，故此題不適宜。

**（三）資料轉化為選文時，內容詮釋須精簡並符合測驗概念：**

不良試題範例

自2007年下半年受全球性金融風暴影響，國內景氣衰退，失業率節節上升，政府為了減輕人民負擔及刺激景氣，於2008年初以來陸續修法減稅，如大幅降低遺產稅與贈與稅稅率，卻因此造成中央稅收嚴重短徵；由於減稅項目均屬國稅，地方稅稅收並未受中央減稅影響，反而受惠於政府為刺激景氣，房地產交易明顯回升，與房地產景氣相關的稅收表現都不錯。請問：中央政府與地方政府依據事務性質所分配到的稅收來源是屬於哪一種制度設計？

①權能區分
②權利分立
③均權制度
④直接民主

【說明】：

1. 題文資料太冗長，須擷取有用的資料，聚焦於問題核心。
2. 題幹的概念有誤，中央與地方的稅收分配非依據事務性質，用稅收考均權制度有問題，甚至可能倒果為因。

**（四）注意單一因果的危險**

不良試題範例

民國初年的「二次革命」和「護法運動」均以失敗告終，孫中山苦於國民黨軍事力量不足，無法與軍閥抗衡，因此意欲引進外來勢力建立黨軍。但是當時的西方列強均拒絕施予援手，因此孫中山先生引進別股勢力，企圖擴大國民黨的軍事實力。請問：上文敘述是在描述下列哪一事件產生的背景？

①護國戰爭
②聯俄容共
③北伐統一
④寧漢分裂

【說明】：

1. 在設計因果的試題時，應在題幹內有完整且有邏輯的敘述，如此才可將其中的因果關係完整呈現。
2. 如果某一事件是多重因果，那麼可考慮在選項內都將之呈現，並設計一個錯誤選項，以負面方式問答，讓學生排除掉與該事件無關的選項，如此可讓學生知道，一個事件的成因往往不是因為單一因素。

**（五）問「為什麼」，比問「是什麼」更有意義**

不良試題範例

「唐朝的崩潰，表示關中地區優勢從此喪失，長安作為國都的時代也從此告終，此後政治的核心逐漸東移。……大運河與黃河交會口的都市□□，在唐代就已經躍升為華北的經濟樞紐。朱全忠就是因為出任此地的節度使，而奠定其中原霸主的地位。……五代諸朝 (除了後唐) 以及宋，皆定都於□□。」請問：空格中應填下列哪一個城市？

①洛陽
②汴京
③南京
④北京

【說明】：

本題的問題在問學生一個歷史事實，即首都南遷，此題可以從兩個方向修改，一是問中國重心遷移的方向，另一可以描述歷史事實，然後問造成這樣現象的理由為何？

修訂試題範例

唐朝的崩潰，長安作為國都的時代也從此告終，此後政治的核心逐漸東移。大運河與黃河交會口的都市汴京，在唐代中期其重要性逐漸上升，到後來的五代諸朝 (除了後唐) 以及宋，皆定都於此。請問：造成汴京取代長安成為都城的主要原因為何？

①北方因戰亂而殘破，汴京因有運河之利而乘勢興起
②北方因為外患眾多，因此五代及宋往南遷以避戰禍
③北方因為疫病流行，人口大量往南遷徙以躲避瘟疫
④南方因為開發程度較高，精緻的文化吸引北人南渡

【說明】：

修正過後的試題，將問題著眼於都城為什麼會由北往南遷，而不是問一個既定的歷史事實。歷史上發生過的事情太多太多，如果要求學生都要知道，會造成很大的負擔，不妨把一些現象寫入題幹描述，並給予足夠的線索，讓學生自行判斷造成這種現象的可能原因為何，或許是比較能鼓勵學生思考的一種方式。

**（六）題幹應為完整句並以簡單而清晰的用詞來陳述問題。**

不良試題範例

以下關於舊石器時代的敘述何者正確？

①北京人距今約2萬年，已知用火，以漁獵、採集維生
②山頂洞人使用骨針縫製獸皮為衣，但比北京人腦容量小
③稱為舊石器時代主要依據是因為未發明文字
④臺灣舊石器時代的代表為距今約5萬年前的長濱文化

【說明】：

1.題幹沒有提供回答的線索，學生回答問題的線索，應該在題幹內，否則就會變成判斷四道選項敘述正確與否的是非題。

2. 應該盡量避免讓學生判斷各自獨立且沒有太多關聯的資訊。

修訂試題範例

下段文字是描寫一個臺灣舊石器時代人的生活，請問哪一段敘述是正確的？

魯夫生存在據今約五萬年前的東臺灣的海岸附近 (甲)，他靠漁獵和種植農作物維生 (乙)，並將獵捕到的野獸用骨針縫製獸皮衣 (丙)，日常生活中發生的大小事，他都用文字記錄在房屋牆壁上 (丁)。

①甲

②乙

③丙

④丁

【說明】：

原本題目的題幹很簡略，學生回答的依據只能由選項內判斷，因此會變成學生在判斷四道選項正確與否的是非題。現將題幹改寫成完整且相關的敘述，修改後的試題完整度較高。

**（七）題幹應避免包括一個以上的問題，否則將減少試題的診斷價值。**

不良試題範例

某一資料曾記載：在廣東一帶有許多民眾信奉外來的天主教，這些人無論其身分地位的差別，皆以「兄弟」稱之，唯有上天可稱為「父」，而親生的父親也稱「兄弟」；親生的母親則稱「姐妹」。請問：這段資料是在記載哪一個組織被反對的原因？

①耶穌會的西方異俗

②義和團的怪力亂神

③太平天國的悖離傳統

④八國聯軍的殘忍暴虐

【說明】：

1. 此題的問題有兩個層面，一要學生判斷這是哪一個組織，其二學生還需要判斷這個組織被反對的理由。學生回答此題，我們不曉得他到底知道哪一個層面的問題，是知道太平天國呢，還是知道太平天國因為背離傳統而不被人接受？學生是否只要知道其中一個問題的答案，另外一個問題的答案就自動浮現，那麼第二個問題的設計是否具有意義？

2. 一道試題有二個問題，將會降低試題的診斷價值。

修訂試題範例

某一資料曾記載：在廣東一帶有許多民眾信奉外來的天主教，這些人無論其身分地位的差別，皆以「兄弟」稱之，唯有上天可稱為「父」，而親生的父親也稱「兄弟」；親生的母親則稱「姐妹」。請問：這段資料是在描述哪一個組織？

①耶穌會
②義和團
③太平天國
④八國聯軍

【說明】：

1. 修改過後的試題，將問題專注於資料是在描述哪一個組織。
2. 在設計試題時，問題的設計要精確，測驗的概念可以複雜，但是題目的問題必須要明確且單一。

**（八）避免在題幹上堆砌一些不切題且毫無作用的材料。**

不良試題範例

胡小花最愛看的電視節目「世界真奇妙」正要開播了，她最喜歡和大明星一起猜題拿大獎。以下是節目中出現的題目，讓我們一起來幫她找出正確的答案。世界各國「吃」的習慣都不相同，下列哪一種敘述最有可能是印度人的飲食習慣？

①吃拉麵的時候，發出很大的吸吮聲
②手拿筷子，夾取盤中的美食，再放到碗裡
③一手掌刀，一手持叉，慢慢切割盤中的牛排
④將右手洗乾淨，然後以手直接抓取盤中的食物

修訂試題範例

世界各國「吃」的習慣都不相同，下列哪一種敘述最有可能是印度人的飲食習慣？

①品嚐拉麵的時候，發出很大的吸吮聲
②手拿筷子夾取盤中的美食，再放到碗裡
③一手掌刀，一手持叉，慢慢切割盤中牛排
④將右手洗乾，然後以手直接抓取盤中的食物

**（九）題幹陳述盡可能簡明，避免不必要的複雜用字或句型結構，否則將變成閱讀、理解能力測驗。**

不良試題範例

以下是對於某個政治運動的描述：**此一運動是以黨員為對象。由於黨員數量快數增加，缺乏以往那股內聚力。於是領導者決定以下放的方式來整治，將黨員調到鄉下去，以更接近實際問題的地方去。接受改造的人，首先必須受調查，繼而被迫坦白，把自己的出身與經歷都交代清楚，讓別人可以挑出批評之處。之後的鬥爭大會裡，被批者是孤立的，在一大群人中被冷嘲熱諷，公開的被指控和羞辱。最後被批者才可以進入重生、與黨修好的階段。大家接受他的認錯與坦白，歡迎他重新回到黨的懷抱。這個時後他會欣喜若狂，甘願接受黨的指導。不論如何，其結果都是使人人順從黨的路線。**請問，最有可能是下列哪一運動？

①大躍進
②白色恐怖
③整風運動
④文化大革命

【說明】：

本題的文字量很大，學生在閱讀上有很大的負擔，在設計題目時，無助於作答的文字可以刪除或是以刪節號替代，惟需注意，修正過後的資料，就不可再當做原文引用。

修訂試題範例

以下是對於某個政治運動的描述：**此一運動是以黨員為對象。由於黨員數量快數增加，缺乏以往那股內聚力。於是領導者決定以下放的方式來整治。接受改造的人，首先必須受調查，繼而被迫坦白，把自己的出身與經歷都交代清楚，讓別人可以挑出批評之處。之後的鬥爭大會裡，被批者是孤立的，公開的被指控和羞辱。最後被批者才可以進入重生、與黨修好的階段。**請問，最有可能是下列哪一運動？

①大躍進
②白色恐怖
③整風運動
④文化大革命

【說明】：

本題的修改方式是直接刪除掉多餘的敘述，但也可透過改寫的方式減少文字量，如此可幫助學生仔細閱讀題目，避免學生因為不耐煩而倉卒作答，增加了粗心答錯的機率，如此將不可得知學生是否了解題目欲測驗的概念。

**（十）不宜為粗心學生，而故意在題目中設計陷阱。**

不良試題範例

小明去參觀一處考古遺址，內有磨製的石器，還有陶器，以及種植作物和大量貝殼和魚骨的遺留痕跡，請問小明參觀的可能是哪一文化的遺跡？
①長濱文化
②卑南文化
③圓山文化
④十三行文化

【說明】：

此題對學生而言，他們判斷的依據可能就只有磨製石器和陶器，而排除掉選項一，三個文化都屬於新石器時代，而其中有貝塚的只有圓山文化，但學生必須很細心注意到此一細節或是知識很廣，才有辦法作答，否則就會變成在三個選項中猜答的情形。

修訂試題範例

小明去參觀一處考古遺址，內有磨製的石器，還有陶器，以及種植作物和大量貝殼和魚骨的遺留痕跡，請問小明參觀的可能是哪一時期的文化遺跡？
①舊石器時代中期
②舊石器時代晚期
③新石器時代中期
④新石器時代晚期

【說明】：

將問題改成題幹中敘述的文明約處於哪一個時期，如此學生就不需要去記得各種文化的細節，只需要大概知道文明演變的進程有什麼樣的特徵就可以了。學生可從磨製石器，排除掉一、二選項，而新石器時代晚期已經有金屬製品出現，題幹中沒有提及，因此也可以排除掉第四選項。

**(十一)選項需具同質性。**

不良試題範例

主辦三天兩夜專業研習的機構，決定將第一天的研習課程提早結束，讓學員可以進行自由聯誼活動，認識彼此；研習主持人說：「來參加研習，不只是學習知識，透過聯誼活動，也是一種學習，因為可以增進 (甲) 」請問：(甲) 應該是下列何者？
①風險貼水
②社會資本

③機會成本
④多元智慧

【說明】：

四個選項亦非同一知識概念下的範疇。此種現象可見選項與試題之間的關連性不強，喪失選擇題的邏輯，且誘答選項較缺乏誘答力。

**(十二)試題之文句必須重新組織，無論題幹或選項的文辭用句，均應避免直接抄課本或原有材料。**

不良試題範例

討伐袁世凱失敗後，孫中山流亡日本，於民國三年，將國民黨改為下列哪一個組織？
①中興會
②同盟會
③中華革命黨
④中國共產黨

【說明】：

上述的題幹是直接抄自某一出版社的教科書在談及「二次革命」的課文內容。直接抄教課書或參考書的文句而產生的試題，學生可能僅是依照記憶作答，並非按照自己的認知理解能力解題。因此命題者在設計試題時，應注意要將試題改寫，切勿直接引用課文或參考書之內容。

修訂試題範例

建國之初，袁世凱嚴重破壞憲法體制，孫中山因而發起「二次革命」討袁，在二次革命失敗後，孫中山遠走日本，成立新的組織，其宗旨訴求「以實行民權、民生主義為宗旨」、「以掃除專制統制、建設完全民國為目的」。請問，上述提及的組織為下列何者？
①國民黨
②中國共產黨
③中國國民黨
④中華革命黨

【說明】：

將題目的敘述改寫，使學生必須理解題幹敘述後，方能作答，如此可避免學生僅是靠記憶作答，實質上卻不知道歷史事件的始末。

**（十三）選項宜按邏輯順序排列。**

不良試題範例

某時期移民臺灣的條件如下：

**一、定居意念堅定。**

**二、身體強壯，沒有傳染病。**

**三、素行端正，無前科和酗酒等惡習。**

**四、已成家者，必須帶妻子、家人一同前來。**

請問：這樣的移民條件，可能是出自於哪一時期的規定？

①鄭氏時期

②荷據時期

③日治時期

④清領初期

【說明】：

選項的設計要有邏輯性，而本題的選項沒有按照時間先後順序排列。

修訂試題範例

選項重新排列如下：

①荷據時期

②鄭氏時期

③清領初期

④日治時期

【說明】：

1. 選項的設計要有邏輯性，譬如：時間的先後順序、人物的年代順序、字數的長短排列、數字的大小排列……等。
2. 在排列選項時，應考慮題目的問題，例如有的題目會問，下列哪一個事件發生的年代最早，諸如此類的試題，如果選項還是按照時間先後順序排列，精明一點的學生往往可以利用小技巧而答對試題，就失去了原先測驗的意義了。

## 五、自然

**(一) 題幹應為完整句並以簡單而清晰的用詞來陳述問題。**

不良試題範例

下列有關臺灣紅樹林的敘述，何者正確？
①多位於東部各河口附近
②所有樹種皆具有胎生苗
③生物種類少但數量龐大
④樹木根系短淺以利呼吸

修訂試題範例

紅樹林生態系的植物為了適應河口沼澤的環境，具備許多特殊的形態與構造，下列何者不屬於紅樹科植物的特徵描述？
①多位於東部各河口附近
②所有樹種皆具有胎生苗
③生物種類少但數量龐大
④樹木根系短淺以利呼吸

【說明】：
此問題主要在問紅樹科植物的特徵描述，但原題幹無法表達清楚。

**(二) 選項不宜插在題幹中間，以免題幹分裂為二；尤其不宜將選項置於題幹之前。**

不良試題範例

對空的試管口吹氣會發出聲音，是什麼原因?
①試管中的水
②試管中的空氣
③嘴唇
④舌頭
的振動。

修訂試題範例

對空的試管口吹氣會發出聲音，是利用振動。請問是下列哪一項原因?
①試管中的水
②試管中的空氣
③嘴唇
④舌頭

**（三）題幹應避免包括一個以上的問題，否則將減少試題的診斷價值。**

不良試題範例

關於日、地、月的敘述，下列何者正確？
①月球繞地球公轉與其自轉的週期相同，故月球自轉一周約需1個月
②地球繞太陽公轉，故地球是太陽的衛星
③月球因自行發光又離地球極近之故，故月圓時顯的特別亮
④太陽誕生後6億年地球才形成

【說明】：
同時考了日、地、月三個概念，且長短不一。

**（四）題幹中避免使用否定的陳述，盡可能以正面陳述的方式強調敘述試題的題幹。**

不良試題範例

民國98年8月8日莫納克颱風帶來超大豪雨，重創南臺灣，山區土石流肆虐，造成重大傷亡。一定會發生土石流的條件不包括：
①位於山坳地區
②有大量降雨的時期
③在颱風侵襲的時間
④有大量鬆軟土石分布的地區

【說明】：
1. 題幹之敘述不完整，沒有以完整句子呈現，且強調“一定會”又要“不包括”，本身的敘述問法就不佳。
2. 題幹之敘述包含兩個敘述，一為災害發生，但又要問“發生土石流的條件”，如此會造成試題疑義。
3. 建議試題題幹敘述要完整，陳述四種情況 (二對二錯或三對一錯)，再設計選項包含二或三個情況，讓學生選取正確答案。

修訂試題範例

民國98年8月8日莫納克颱風帶來超大豪雨，重創南臺灣，山區土石流肆虐，造成重大傷亡。發生土石流的條件不包括下列何者？

**（五）答案必須明確，對單一選擇題而言，僅可供一個正確選項。**

不良試題範例

雨水滲入地下的速率常受地表坡度、岩層透水性、地面植被等因素影響。當降雨無法及時滲入地下，而大量往低窪處匯集常易導致水災。坡地災害的發生也

常與降雨有關，此外，植被的對土石、水分的抓附，也會對坡地的穩定性造成影響。下列何種情形引發水患的可能性較小？
①地表岩層的透水、排水不良
②颱風帶來大量而集中的降雨
③坡地被植被完整披覆
④排水系統老舊失修並且淤塞

【說明】：
題幹與選項中皆含有多個概念，從不同方向解讀將會導致不同的答案，易引起爭議。建議修定如下，擬考之知識概念題幹敘述表達要清楚呈現與選項間能連貫。

修訂試題範例
坡地災害的發生常與降雨有關，而植被的對土石、水分的抓附，也會對坡地的穩定性造成影響。下列何者是大雨後容易發生坡地災害的原因？
①水分增加了坡地土石與岩層間的摩擦
②雨水沖刷經常導致坡地植被嚴重破壞
③降雨容易引發地震進而導致土石鬆動
地表的土石吸收了水分因而增加重量

## 肆、結語

本章提供選擇題命題原則與不良題範例，是以國家教育研究院測驗及評量研究中心，臺灣學習成就評量題庫建置時各科研發修審過程中之一些值得注意範例為主，擬提供給現場教師作為參考。因基於保密，各科所釋放之例題數不一，因此造成例題比例不同。臺灣學習成就評量題庫之建置目的為「建立國民中小學、高中及高職學生學習成就長期資料庫，以追蹤、分析學生在學習上變遷之趨勢，進而檢視目前課程與教學實施成效。」(臺灣學生學習成就評量資料庫網站，2014)，其評量架構之內容係以九年一貫能力指標及高中課程剛要為主，本章之重點為命題原則，因此試題範例說明中若沒有提及測驗目標，其前提就假設已符合測驗目標前提下做建議。

在本章內容所提及之試題編寫原則，一般會採用表單的方式，讓命題者進行勾選，以提醒與修正。修審時，再由修審委員就題幹、選項的表單進行核檢，以確保試題之內容效度。而試題的內容修審常涉及主觀因素，因此本章所提供之試題範例與建議修審，亦未臻完善、見仁見智，在此僅提供給命題者與修審者一個參考方向。

## 參考文獻


余民寧 (2012)。**教育測驗與評量：成就測驗與教學評量**。臺北：心理出版社。

簡茂發 (2000)。**心理測驗與統計方法**。台北：心理。

臺灣學生學習成就評量資料庫網站 (2014)。**TASA建置計畫**。2014年4月2日，取自 http://tasa.naer.edu.tw/1about-1.asp?id=2

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

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

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.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.) . Upper Saddle River, NJ: Prentice-Hall.

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

# 測驗理論與測驗分析技術

➢ 吳慧珉

# 測驗理論與測驗分析技術

**吳慧珉**
國家教育研究院助理研究員

測驗編製者編製完成一份測驗後，透過預試或正式施測可收集到受試者作答反應資料，此時需搭配測驗理論，以解釋測驗資料間的實證關係。測驗理論學者通常把測驗理論劃分成二大學派：一為古典測驗理論 (classical test theory，簡稱CTT)，主要是以真實分數模式 (true score model) (Gullikson, 1987; Lord & Novick, 1968) 為核心；一為現代測驗理論 (modern test theory)，主要是以試題反應理論 (item response theory，簡稱IRT) (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hulin, Drasgow, & Parsons, 1983; Lord, 1980) 為架構，IRT是目前測驗理論之主流，許多國際大型測驗，如學生能力國際評量計畫 (the Programme for International Student Assessment，簡稱 PISA)、國際數學與科學教育成就趨勢調查 (Trends for International Mathematics and Science Study，簡稱 TIMSS) 等，均使用IRT分析測驗資料，本章節首先將簡介古典測驗理論，而後著重於介紹試題反應理論與其相對應之測驗分析技術，讓讀者深入了解IRT之理論與軟體分析技術。

## 壹、古典測驗理論簡介

### 一、古典測驗理論

古典測驗理論也被稱為是真實分數理論 (true score theory)，可說是最早之測驗理論，主要的理論基礎是真實分數模式 (true score model) (Lord & Novick, 1968）。所謂的真實分數模式，是建立於直線關係之數學模式，假設任何可以觀察到或測量到的觀察分數皆包含兩個部分所構成的數學函數關係：一是觀察不到但卻是研究者想獲得的潛在特質 (latent trait)，也就是「真實分數」；另一個部分是觀察不到，卻不代表潛在特質且是研究者想要極力避免或設法降低的部分，叫作「誤差分數」。這兩個部分合起來構成一個實際的測量值，也就是觀察分數，且彼此之間延伸多種基本假設，符合這些基本假設的測量問題，就是古典測驗理論探究的範圍。古典測驗理論主要有幾項項基本假設，分述如下：（余民寧，2011）

1. $x = t + e$，其中，$x$ 是觀察分數，$t$ 是真實分數，$e$ 是誤差分數。即觀察分數＝真實分數＋誤差分數。

受試者的真實分數無法直接觀察而得，只能由測量的方式得到觀察分數，這種觀察分數含有誤差，而此誤差被假設為一個隨機變數，其分配是以零為集中趨勢指標的常態分配。這種誤差有時大於真實值也有時小於真實值，但總平均起來誤差為零。此外，觀察分數的分配亦為常態分配。

2. $E(x) = E(t + e) = t$，觀察分數的期望值＝真實分數。

用相同的測量方式，去重覆測驗同一個人很多次，所得的觀察分數的期望值，也就是觀察分數的平均值，就是受試者的真實分數，此時誤差分數的期望值等於零。因此雖然測量有誤差，但如果我們能收集到足夠多次的觀察分數，則這些觀察分數的平均值可被視為是真實分數。

3. $\rho_{te} = 0$，$\rho_{te}$ 表示誤差分數與真實分數的相關，誤差分數與真實分數的相關是零。

真實分數的高低和誤差分數的高低是沒有關係的。

4. $\rho_{e_1e_2} = 0$，$e_1$ 假設是測驗 1 的誤差分數，$e_2$ 假設是測驗 2 的誤差分數。不同測驗之間的誤差分數的相關是零。

不同測驗之間，其誤差分數是沒有關係的，也表示受試者在一測驗上有較高的誤差，不一定在另一測驗上也有較高的誤差。

5. $\rho_{e_1t_2} = 0$，$e_1$ 假設是測驗1的誤差分數，$t_2$ 假設是測驗 2 的真實分數。不同測驗的誤差分數與真實分數是零相關。

一個測驗的誤差與另一個測驗的潛在特質或真實分數不相關，因此測某種特質的測驗並不受另一種測驗的誤差影響。

## 二、古典測驗理論之限制

古典測驗理論理論基礎是建立於簡單的函數假設，也就是 $x = t + e$，由於計算公式簡單明瞭、淺顯易懂，適用大多數的教育與心理測驗情境，以及社會科學研究資料之分析，在測驗編製、評估及使用上被廣為流通與應用。但隨著測驗需求量的日益增加及形式的多樣化，簡單的假設造成其應用上的許多限制（Lord & Novick, 1968）：

1.試題的難度指數是受試者通過某題的百分比，而鑑別度指數是受試者在某題的得分和其總分之相關，不論是百分比或相關都會有樣本依賴的現象。一份測驗給優秀的受試者施測，難度指數值會偏高；反之，一份測驗給能力較差的受試者施測，難度指數值會偏低。試題的特性，如試題是簡單的或是難的，取決於受試樣本的特

性，即樣本依賴的現象 (王文中，2004)。

2.受試者的成績會受到測驗難度的影響，同一位受試者，在較難的測驗容易答錯，得到60分，在較簡單的測驗得到90分，也就是受試者程度是好是差，取決於測驗的特性，使用簡單的測驗能力變好，使用較難的測驗，則能力變差，即所謂的測驗依賴 (王文中，2004)。

3.理論模式中只假設一個測量標準誤 (standard error of measurement)，即所有的受試者均有相同的測驗誤差，誤差指數大，代表測驗測量的每一個受試者都不準確，沒有例外。但在實務應用上，我們常發現測驗測量的能力對於某些受試者是十分精確，對某些受試者卻相反，如有些題目對於能力較高的受試者能提供較豐富的訊息，因此能精確地測量出這些能力較高的學生能力，相對的，這些題目對於高能力的學生，測量誤差就會較低；然而這些題目，對於中低能力的人，可能都會答錯而都是 0 分，無法準確估計出中低能力學生的能力，而有較高的測量標準誤，故用一個測量誤差指數代表所有受試者的測驗誤差是不合理的。

4.古典測驗理論沒有將受試者的作答反應組型 (response pattern) 和試題特性相連結，認為原始總分相同的受試者，其能力必定一樣，但在現實測驗情境中，即使總分相同的受試者，其作答反應組型亦有可能不一樣。如一份五題的測驗，第五題是比較難的題目，答對得 1 分，答錯得 0 分，考生甲的作答反應組型是 11100，考生乙的作答反應組型是 11001，同樣都得到 3 分，古典測驗理論會認為考生甲乙的能力是一樣的，依據作答反應組型，乙考生答對的是較難的三題，其能力估計值應該不一樣才合理。

5.在測驗應用方面，古典測驗理論是假設相同的測驗或複本測驗的分數方能提供有意義的比較，當不同的人接受不同的測驗，如非複本測驗，其分數是無法進行比較的。然在真實情境中，編製一份符合複本測驗的假設是有其難度的，故當不同的人接受不同的測驗，在古典測驗理論架構下，測驗分數是不能比較的。如果測驗的涵蓋範圍廣泛，如想要完整評估六年級學生在十二年國教課程架構下的數學表現，可能須有 200 題試題方能完整涵蓋六年級數學課程架構，囿限於考試時間與疲勞因素，可能須將 200 題試題拆成四份各 50 題的測驗分給不同班的學生作答，此種情境即是「不同的人接受不同的測驗」，即所謂的等化 (equating) 議題，古典測驗理論在這一方面未能提供較完善的解決之道。

## 貳、試題反應理論簡介

### 一、試題反應理論之特點

由於古典測驗理論的諸多限制，因此才有現代測驗理論的興起，現代測驗理論

主要是以試題反應理論為主要架構，試題反應理論由於理論架構較嚴謹，考慮層面廣泛，不僅可以延續古典測驗理論的功能，並藉由電腦科技發達之協助，得以突破古典測驗理論在應用上的瓶頸。目前試題作答理論已取代古典測驗理論，成為當前的主流測驗理論之一 (Embretson & Yang, 2006; Hambleton, 1989; Yen& Fitzpatrick, 2006; 許澤基和劉長萱，1992；余民寧，2009，2011)。

Lord與Novick (1968) 提出以模式為基礎的測量開啟試題反應理論之發展。試題反應理論主要是建立在兩個基本的假設：(一) 受試者在測驗試題的表現，可由一組因素加以預測或解釋，這組因素稱為潛在特質或能力 (ability)，因此試題反應理論是一種潛在特質理論，能力的估計是用來解釋測驗的結果。(二) 潛在特質無法直接測量，於是將受試者的測驗表現與這組潛在特質的關係，透過一條連續遞增的函數，詮釋測驗的表現與不可直接測量的潛在特質之間的關係，這個函數稱為試題反應函數 (Hambleton, Swaminathan & Rogers, 1991)。

試題反應理論具備幾項功能，正好可以補足上述古典測驗理論之限制：

1.古典測驗理論中，試題的難度指數和鑑別度指數，會因為受試者的能力分配不同而得到不一樣的估計結果，受試者能力之估計後受到測驗難度之影響；試題反應理論具有參數不變性 (parameter invariance) 之特色，即某試題的試題參數之估計，不受考生能力分配之影響，而某位受試者能力值之估計，亦不受到使用哪一組測驗試題之影響 (Hambleton & Swaminathan, 1985)。在線性轉換的條件下，也就是說考生能力不同時，可透過線性轉換達到參數不變的特性。

2.試題反應理論中，一份測驗中的每一題試題皆有其對應的試題訊息函數 (item information function)，將所有題目的試題訊息函數加總，就可以得到測驗訊息函數，其意義相似於古典測驗理論中的信度，但針對不同的能力有不同的訊息函數值。試題訊息函數是試題的特性，透過一個量化的統計數字，指示題目區辨出不同能力考生的能力。測量標準誤的定義是與訊息函數的平方根成反比，每位具有不同能力水準的受試者，對應不相同的測量標準誤。如一群考生的能力是介於 -3 到 3 之間，第 1 題對於能力是 2 的考生，試題訊息量是 0.8，對於能力是 -1 的考生，試題訊息量是 0.1；對於能力是 2 的考生，測量標準誤約是 1.12，對於能力是 -1 的考生，測量標準誤約是 3.13，表示第 1 題適合區辨能力高的考生，測量標準誤較低，對於能力低的考生，測量標準誤較高，不同能力水準的受試者，對應不相同的測量標準誤。

3.受試者能力的估計主要依賴於受試者的作答反應組型和所施測的試題特性，原始總分相同的受試者，若是其作答反應組型不一樣，使用特定的試題反應理論亦有可能得到不同的能力估計值。如考生甲的作答反應組型是 11100，考生乙的作答反應組型是 11001，考生甲能力估計值是 1.8，考生乙能力估計值是 2.0，因考生乙答對的是較難的第 5 題，而考生甲答對的是較簡單的第 3 題，考生乙能力應高於考生甲。

4.對於等化、差異試題功能、題庫的建立、電腦化適性測驗之實施提供良好之解決之道，此部分由於內容較深奧，將不再本章節中探討。

## 二、試題反應理論模式

試題反應理論中，包含多種的試題反應函數。依據所測量之能力維度數，可分為單向度試題反應理論 (unidimensional IRT, UIRT) 和多向度試題反應理論 (multidimensional IRT, MIRT) ；依據計分型態，可分為二元計分，如答對是 1 分，答錯是 0 分，和多點計分模式，如計分方式是 0 分、1 分、2 分、3 分。以下將從不同能力維度數的觀點，介紹幾種常見的單向度試題反應理論模式和多向度試題反應理論模式。

### (一)單向度試題反應理論

$$E(X) = f(I, A) \qquad \text{(公式 1)}$$

其中，$X$ 為試題反應的正確性，如 $X=1$，表示答對此題，$X=0$，表示答錯此題；$I$ 為試題參數向量；$A$ 為能力參數向量。（公式 1）指 $X$ 的期望值是由試題參數和能力參數所成之函數所決定的；然而，單向度試題反應理論模式必須符合單向度（unidimensionality）、局部獨立（local independence）、受試者對於試題的反應可透過數學模式表徵等基本的假設（Weiss & Yoes, 1991）。

1.單向度：是指試題之間是否具有同質性，均測量同一特質，讀者可透過因素分析檢驗是否所有試題均依附於同一因素，符合單向度之假設。

2.局部獨立：在給予受試者能力的條件下，受試者在作答所有題目，各題之間是獨立的，不會互相影響，如受試者答對第一題的機率和答對第二題的機率是彼此獨立的。局部獨立的假設可簡化IRT參數估計的複雜度，簡化數學式。

3.受試者對於試題的反應可透過數學模式表徵：將受試者能力與試題之間的關係透過數學式模式化，即所謂的試題反應函數，許多不同的試題反應函數被學者提出，以描述試題的特性和受試者能力的特性以及彼此之間的關係，建立IRT的理論核心架構。

基於試題反應理論的單向性假設，一般使用之試題反應理論為單向度試題反應理論。以下介紹二元計分對數型模式及多點計分模式，二元計分對數型模式依所採用的參數個數來命名，分別為單參數羅吉斯模式（one-parameter logistic model，簡稱1PL）、二參數羅吉斯模式（two-parameter logistic model， 簡稱2PL）及三參數羅吉斯模式 (three-parameter logistic model，簡稱3PL)；多點計分模式則介紹部分給分

模式 (partial credit model，簡稱 PCM) 和廣義部分給分模式 (generalized partial credit model，簡稱GPCM)。

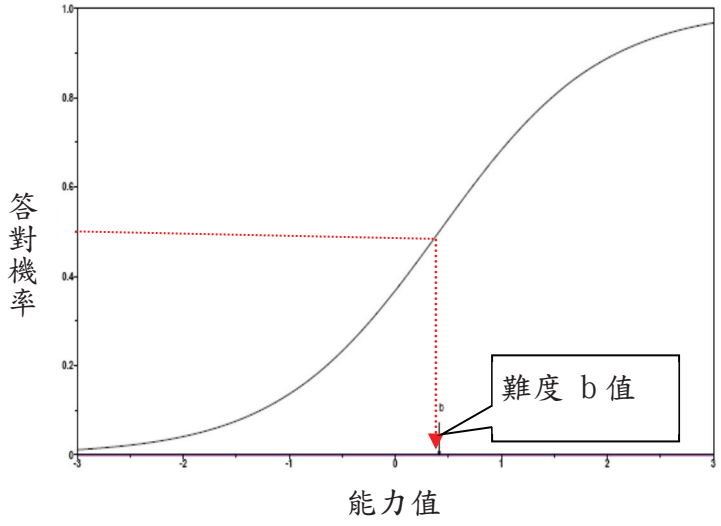
1.二元計分羅吉斯模式

（1）單參數羅吉斯模式

在試題反應理論的1PL模式下，假設受試者 $j$ 之能力為 $\theta_j$，其作答試題 $i$ 通過的機率如下：

$$P(X_{ij}=1\,|\,\theta_j,b_i)=\frac{1}{1+\exp[-(\theta_j-b_i)]} \qquad \text{(公式2)}$$

其中，$X_{ij}$ 為受試者 $j$ 在試題 $i$ 的作答反應，答對記為 1，答錯記為 0；$b_i$ 為試題 $i$ 之試題難度參數，$-\infty<b_i<\infty$，是當答對機率是0.5時，相對應的能力值位置。在單參數羅吉斯模式中，總分是能力值之充分統計量，即總分相同的受試者，其能力估計值適一樣的。從（公式2）可算出當受試者能力與難度參數相等時，在試題的答對機率是 0.5，當受試者能力高於難度參數時，在試題的答對機率會高於 0.5，即受試者有很高的機率答對此題。單參數羅吉斯模式之圖示如圖 1。




註：畫面擷取自BILOG-MG軟體

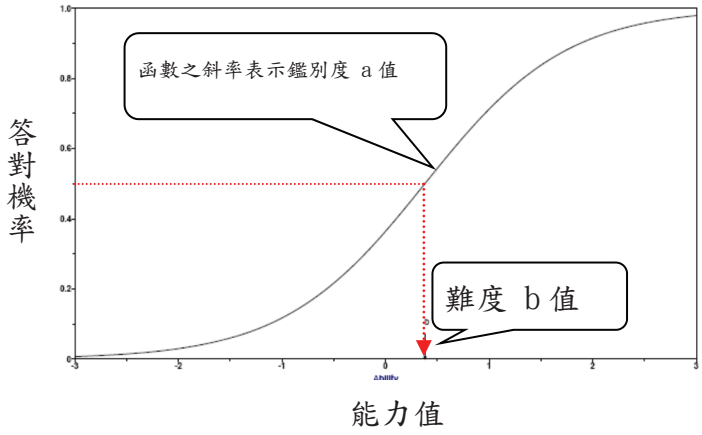
圖 1 單參數羅吉斯模式之試題特徵曲線 ($b$ = 0.416)

（2）二參數羅吉斯模式

在試題反應理論的2PL模式下，假設受試者 $j$ 之能力為 $\theta_j$，其作答試題 $i$ 通過的機率如下（Birnbaum, 1968）：

$$P(X_{ij}=1\mid\theta_j,b_i,a_i)=\frac{1}{1+\exp[-a_i(\theta_j-b_i)]} \quad \text{(公式 3)}$$

其中，$X_{ij}$ 為受試者 $j$ 在試題 $i$ 的作答反應，答對記為 1，答錯記為 0；$a_i$ 為試題 $i$ 之試題鑑別度參數（discrimination parameter），$-\infty<a_i<\infty$，是試題反應函數之斜率，斜率越陡，鑑別度越高，表示題目越能區分不同能力的受試者；$b_i$ 為試題 $i$ 之試題難度參數，$-\infty<b_i<\infty$。二參數羅吉斯模式之圖示如圖 2。




註：畫面擷取自BILOG-MG軟體

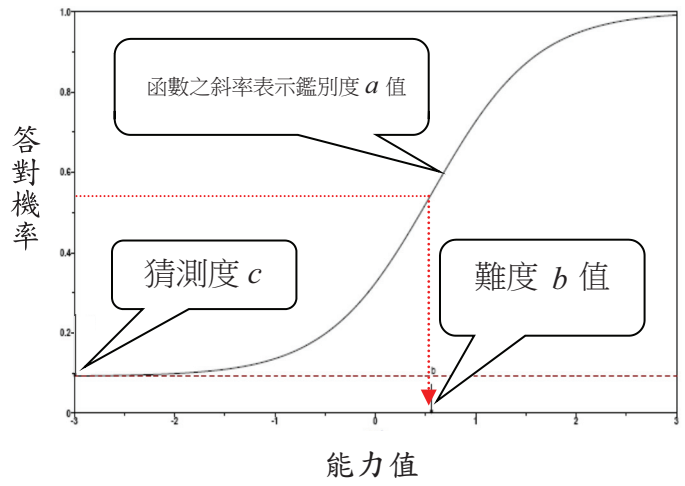
圖 2　二參數羅吉斯模式之試題特徵曲線 ( $a=0.860$ , $b=0.382$)

（3）三參數羅吉斯模式

在試題反應理論的 3PL 模式下，假定測驗會發生猜題之現象，表示受試者即使能力很低，仍然有機會透過猜測答對此題。假設受試者 $j$ 之能力為 $\theta_j$，其作答試題 $i$ 通過的機率如下（Birnbaum, 1968；Lord, 1980）：

$$P(X_{ij}=1\mid\theta_j,b_i,a_i,c_i)=c_i+\frac{(1-c_i)}{1+\exp[-a_i(\theta_j-b_i)]} \quad \text{(公式 4)}$$

其中，$X_{ij}$ 為受試者 $j$ 在試題 $i$ 的作答反應，答對記為 1，答錯記為 0；$a_i$ 為試題 $i$ 之試題鑑別度參數，$-\infty<a_i<\infty$；$b_i$ 為試題 $i$ 之試題難度參數，$-\infty<b_i<\infty$；$c_i$ 為試題 $i$ 之試題猜測度參數（guessing parameter），$0\le c_i<1$。三參數羅吉斯模式之圖示如圖 3。




註：畫面擷取自BILOG-MG軟體

圖 3 三參數羅吉斯模式之試題特徵曲線 ( $a = 1.124$ , $b = 0.56$, $c = 0.092$)

2.多點計分對數模式

（1）部分給分模式

部分給分模式是由Masters (1982) 所提出，當計分模式不再只有對或錯兩個類別，而是有更多類別時，如數學應用問題，教師常按照學生的解題步驟給分；問答題時，根據答案的完整性給予不同的分數等級；問卷設計時，需填答非常同意、同意、尚可、不同意、非常不同意等，只要題目是需依據知識程度的不同或依據回答的重要性不同而給予不同的分數，均適用此模式，其類別反應公式如下：

$$P_{ik}(\theta_j)=\frac{\exp\left[\sum_{v=1}^{k}(\theta_j-b_{iv})\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c}(\theta-b_{iv})\right]}=\frac{\exp\left[\sum_{v=1}^{k}(\theta_j-b_i+d_v)\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c}(\theta_j-b_i+d_v)\right]} \qquad \text{(公式 5)}$$

且 $b_{i1}=0$ 且 $\sum_{v=1}^{1}(\theta_j-b_{iv})\equiv 0$

$\theta_j$ ：表示受試者 $j$ 的能力。

$k$ ：為受試者的回答所屬類別，從$1\cdots m_i$。如計分方式是0分、1分、2分，則有3個類別 $m_i=3$。受試者如果得0分，則 $k=1$ ；受試者如果得1分，則 $k=2$ ；受試者如

果得2分，則 $k=3$。

$c$：題目之類別數，$c=1\cdots m_i$ ，$m_i$ 第 $i$ 題所有的類別數。

$P_{ik}(\theta_j)$：能力為 $\theta_j$ 的受試者 $j$ 在第 $i$ 題得 $k$ 類的機率（$0<P_{ik}(\theta_j)<1$）。

$b_{iv}$：指第 $i$ 題第 $v$ 個的試題階難度參數（step parameter）或類別閾參數（category intersection parameter），隨著類別界線（category boundary）而變，相鄰在兩類別間，就有一個 $b_{iv}$ 參數 $(-\infty<b_{iv}<\infty)$，即 $b_{ik}$ 為 $P_{i,k\text{-}1}(\theta_j)$ 和 $P_{ik}(\theta_j)$ 的交點。

$b_{iv}=b_i-d_v$ ：$b_i$ 為試題 $i$ 位置參數、$d_v$ 為閾參數，$v=1\cdots k, d_k$ 為同一試題內的第 $k$ 類和其他類別的相對難度（Andrich, 1982）。$b_{iv}$ 為第 $i$ 題第 $v$ 個的試題階難度參數或類別閾參數，隨著類別界線而變，相鄰在兩類別間，就有一個 $b_{iv}$ 參數 $(-\infty<b_{iv}<\infty)$，即 $b_{ik}$ 為 $P_{i,k-1}(\theta_j)$ 和 $P_{ik}(\theta_j)$ 的交點。

如解數學題目： $\sqrt{2}\times\sqrt{4}$

$$\sqrt{2}\times\sqrt{4}$$
$$=\sqrt{8}\cdots\text{(步驟 1)}$$
$$=2\sqrt{2}\cdots\text{(步驟 2)}$$

教師如果認為受試者需完成步驟（1）和（2）才算是完整解答此題，未完成任何步驟給0分，完成步驟（1）給1分，完成步驟（2）給2分，則計分類別是0、1、2 三個類別，則 $m_i=3$，此題有兩個階難度參數，假設是 $b_{i1}=-0.38$, $b_{i2}=1.88$，則三個類別（0分、1分、2分）的類別反應函數如下，部分給分模式之圖示如圖 4。圖 4 顯示：0分類別的反應函數隨著能力值增加而降低；1分類別的反應函數呈現一個鐘形曲線；2分類別的反應函數隨著能力值增加而增加。第一個階難度參數是 $b_{i1}=-0.38$，是0分類別和1分類別的交點；第二個階難度參數是 $b_{i2}=1.88$，是1分類別和2分類別的交點。部分給分模式的公式中，我們可以發現如果試題為二元計分，則用部分給分模式模式分析，試題會與使用單參數模式分析得到相同的結果。

$$P_{i0}(\theta_j)=\frac{1}{1+\exp(\theta_j-b_{i1})+\exp(2\theta_j-b_{i1}-b_{i2})}\cdots 0\text{分的機率}$$

$$P_{i1}(\theta_j)=\frac{\exp(\theta_j-b_{i1})}{1+\exp(\theta_j-b_{i1})+\exp(2\theta_j-b_{i1}-b_{i2})}\cdots 1\text{分的機率}$$

$$P_{i2}(\theta_j)=\frac{\exp(2\theta_j-b_{i1}-b_{i2})}{1+\exp(\theta_j-b_{i1})+\exp(2\theta_j-b_{i1}-b_{i2})}\cdots 2\text{分的機率}$$

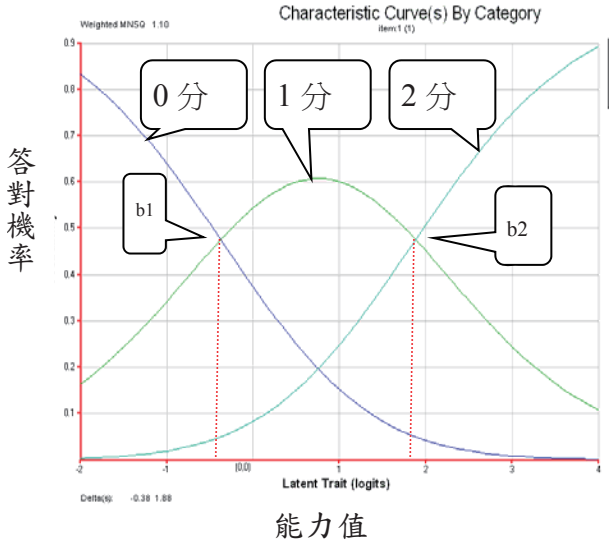



圖 4 部分給分模式之試題類別特徵曲線 ($b_{i1} = -0.38$, $b_{i2} = 1.88$)

（2）廣義部分給分模式

Muraki (1992) 提出廣義部分給分模式是部分給分模式的延伸，故大部分參數的解釋均和部分給分模式相同，但多了一個鑑別度參數，各試題之間有不同的鑑別度參數，其意義同單向度二參數羅吉斯模式之鑑別度，廣義部分給分模式為：

$$P_{ik}(\theta_j)=\frac{\exp\left[\sum_{v=1}^{k} a_i(\theta_j - b_{iv})\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c} a_i(\theta_j - b_{iv})\right]}=\frac{\exp\left[\sum_{v=1}^{k} a_i(\theta - b_i + d_v)\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c} a_i(\theta - b_i + d_v)\right]} \quad \text{(公式 6)}$$

其中 $d_1 \equiv 0$，為了在進行參數估計時，使其有一個相對原點，$b_{iv} = b_i - d_v$。

$\theta_j$：表示受試者 $j$ 的能力。

$k$：為受試者的回答所屬類別，從 $1\cdots m_i$。如計分方式是0分、1分、2分，則有3個類別，$k = 1,2,3$。

$c$：題目之類別數，$c = 1\cdots m_i$，$m_i$ 第 $i$ 題所有的類別數。

$P_{ik}(\theta_j)$：能力為 $\theta_j$ 的受試者 $j$ 在第 $i$ 題得 $k$ 類的機率（$0 < P_{ik}(\theta_j) < 1$）。

$b_{iv}$：$b_{iv} = b_i - d_v$。$b_{iv}$ 為第 $i$ 題第 $v$ 個的試題階難度參數或類別閾參數，隨著類別

界線而變，相鄰在兩類別間，就有一個 $b_{iv}$ 參數 $(-\infty < b_{iv} < \infty)$，即 $b_{ik}$ 為 $P_{i,k-1}(\theta_j)$ 和 $P_{ik}(\theta_j)$ 的交點。$b_i$ 為試題 $i$ 位置參數、$d_v$ 為閾參數，$v = 1 \cdots k, d_k$ 為同一試題內的第 $k$ 類和其他類別的相對難度（Andrich, 1982）。

$a_i$：試題 $i$ 的斜率參數，同一試題在各類別選項有相同的斜率參數，但不同的試題有不同斜率。

**(二)多向度試題反應理論**

隨著測驗的評量架構設計日趨複雜，單向度試題反應理論逐漸被擴展為多向度試題反應理論。所謂的多向度是指所測量的能力向度是多維度，以PISA為例，根據PISA2012年技術報告中所呈現的數學科評量架構，包括四個主題領域 (subject domain) 量尺分數，包括數量 (quantity)、空間與形體 (space and shape)、改變與關係(change and relationships) 及不確定性 (uncertainty) 四個向度之數學能力 (OECD, 2014)。PISA是使用多向度試題反應理論分析此四種數學能力。目前常見的多向度試題反應理論大多是單向度試題反應模式的衍生模式。

多向度測驗可以分為題間多向度測驗 (between-item multidimensional test) 與題內多向度測驗 (within-item multidimensional test) 兩種 (Adams, Wilson & Wang, 1997)。若在測驗裡的每個試題只測量一種能力，即單向度的試題，若整份測驗包含多個測量不同能力的單向度試題，則稱此測驗為題間多向度測驗 (如圖5)；若在測驗裡的每個試題不只測量單一種能力，也就是試題內就包含多向度，稱此測驗為題內多向度測驗 (如圖 6)。

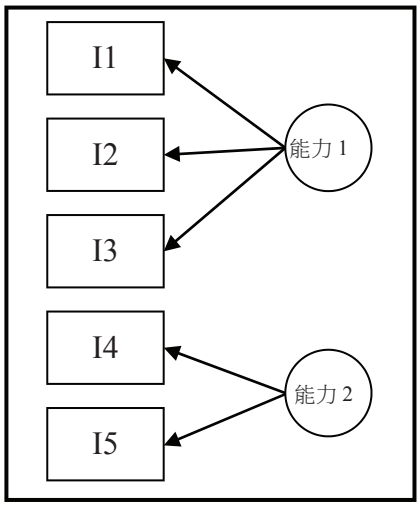



圖 5　題間多向度測驗

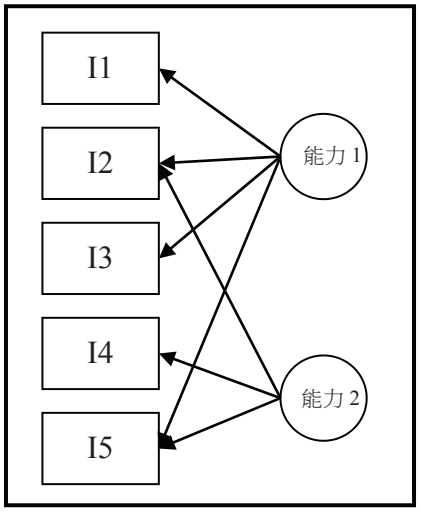



圖 6　題內多向度測驗

常見的多向度試題反應理論模式，分別為多向度隨機係數多項 logit 模式 (Multidimensional Random - Coefficients Multinomaial Logit Model, MRCMLM)、多向度二參數模式 (multidimensional two parameters model, M2PL)、多向度三參數模式 (multidimensional three parameters model, M3PL)，然囿限於欠缺相對應之套裝分析軟體，多向度試題反應理論模式之應用並不普及，多數是學者之理論研究，目前廣為人知的是 PISA 使用的分析軟體 Acer ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) 可以分析 MRCMLM 之理論模式，故應用較廣泛，再者，MRCMLM可套用至不同的評量架構，如題間多向度與題內多向度、二元與多點計分、單向度與多向度等評量架構，故本章節著重於介紹MRCMLM，對於多向度二參數模式有興趣的讀者，可參閱Reckase & Mckinley (1991) 的文章，而對於多向度三參數模式有興趣的讀者，可閱讀Reckase (1997) 的書籍。

1.多向度隨機係數多項 logit 模式 (MRCMLM)

多向度隨機係數多項 logit 模式是由 Adams、Wilson 與 Wang (1997) 等人所提出，MRCMLM 為 Rasch 模式的衍生模式，是一個混合的 coefficients模型，試題參數是由未知的參數 $\xi$ 所描述，同部分計分模式的試題的階難度參數，差別在於在部分計分模式試題階難度參數只與單一向度能力有關係，然在MRCMLM的試題難度階參數不僅可對應至單向度能力，亦可對應至多向度能力。受試者的潛在變數$\theta$，是一個隨機變項，以向量表示受試者之單向度能力或多向度能力。故若讀者的評量架構是屬於多向度的評量架構，就可以使用此模式，MRCMLM之試題反應函式如下：

$$P(\mathbf{X}_{ik}=1;\mathbf{A},\mathbf{B},\boldsymbol{\xi}|\boldsymbol{\theta})=\frac{\exp(\mathbf{b}'_{ik}\boldsymbol{\theta}+\mathbf{a}'_{ik}\boldsymbol{\xi})}{\sum_{k=1}^{K_i}\exp(\mathbf{b}'_{ik}\boldsymbol{\theta}+\mathbf{a}'_{ik}\boldsymbol{\xi})} \quad \text{(公式 7)}$$

$\mathbf{X}_{ik}=(X_{i1},X_{i2},...,X_{iK_i})'$，$(k=0,1,...,K_i+1)$ ：受試者反應類別，如計分是0分、1分、2分，則有三個類別，0分是第1個類別，1分是第2個類別，2分是第3個類別等。

$$X_{ik}=\begin{cases}1 & \text{表第}\boldsymbol{i}\text{題作答第}\boldsymbol{k}\text{個反應類別}\\ 0 & \text{表其他}\end{cases}$$

$\boldsymbol{\xi}'=(\xi_1,\xi_2,...,\xi_p)$：試題參數向量（$p$ 個參數），如三個計分類別，則 $p=2$。

$\boldsymbol{\theta}'=(\theta_1,\theta_2,...,\theta_D)$：受試者的能力向量（$D$個向度），如PISA的評量架構，則 $D=4$ (數量、空間與形體、改變與關係、不確定性)

$\mathbf{A}'=(\mathbf{a}_{11},\mathbf{a}_{12},...\mathbf{a}_{1K_1},\mathbf{a}_{21},\mathbf{a}_{22},...\mathbf{a}_{2K_2},...,\mathbf{a}_{nK_n})$ ：整份測驗的設計矩陣，控制試題所對應

的反應類別。

$\mathbf{a}_{ik}(i=1,...,n \text{ and } k=1,...,K_i)$：第$i$題中第$k$個反應類別的設計向量，每個向量長度為$p$

$\mathbf{B}=(\mathbf{B}_1',\mathbf{B}_2',...,\mathbf{B}_n')'$：整份測驗的計分矩陣

$\mathbf{B}_i=(\mathbf{b}_{i1},\mathbf{b}_{i2},...,\mathbf{b}_D)'$：第 $i$ 題的計分子矩陣

$\mathbf{b}_{ik}=(b_{k1},b_{k2},...,b_{ikD})'$：在 $D$ 個向度中，第 $i$ 題回答第 $k$ 個反應類別的計分向量

MRCMLM 模式較複雜，為了讓讀者易於瞭解，下面將以實例說明模式之應用：

| | |
|---|---|
| 題目示例 | 請計算$\sqrt{2}\times\sqrt{4}$ |
| 能　　力 | 解決平方根的計算問題 |
| 計分說明 | 作答需列計算過程，完成部分步驟得1分，完成全部步驟得2分。 |

此題是單向度多點計分之試題，計分是0分、1分、2分，共有三個類別，二個難度階參數，如受試者得2分，則 $\mathbf{X}_i=(X_{i1},X_{i2},X_{i3})'=(0,0,1)'$　$\xi'=[\xi_{11},\xi_{12}]$

$\theta'=[\theta]=$ [解決平方根的計算問題]

$$\overset{\quad\xi_{11}\ \ \xi_{12}}{\mathrm{A}=\begin{bmatrix}\mathrm{a}_{11}\\ \mathrm{a}_{12}\\ \mathrm{a}_{13}\end{bmatrix}=\begin{bmatrix}0 & 0\\ -1 & 0\\ -1 & -1\end{bmatrix}}$$

，此例題有三個類別，二個難度階參數。使用設計矩陣A矩陣控制二個難度階參數之開關，第一列表示受試者反應於第一個類別 (0分類別) 需跨越的難度階參數，故元素是 [0,0]；第二列表示受試者反應於第二個類別 (1分類別) 需跨越第一個難度階參數 $\xi_{11}$，故元素是[-1,0]；第三列表示受試者反應於第三個類別 (2分類別) 需跨越第一個難度階參數和第一個難度階參數，$\xi_{11}$和$\xi_{12}$，故元素是 [-1,-1]。A矩陣控制試題所對應的反應類別，受試者要得到高分，需跨過越多難度階參數。

$$\mathrm{B}=[\mathrm{B}_1']=\begin{bmatrix}\mathrm{b}_{11}\\ \mathrm{b}_{12}\\ \mathrm{b}_{13}\end{bmatrix}=\begin{bmatrix}b_{111}\\ b_{121}\\ b_{131}\end{bmatrix}=\begin{bmatrix}0\\ 1\\ 2\end{bmatrix}$$

，計分模式，此題是0分、1分、2分。

為了清楚說明，我們先計算公式3-11的分子，受試者反應在第1個類別：

$$\mathrm{b}_{11}'\theta+\mathrm{a}_{11}'\xi=[0][\theta]+[0,0]\begin{bmatrix}\xi_{11}\\ \xi_{12}\end{bmatrix}=0$$

反應在第2個類別：

$$b'_{12}\theta + a'_{12}\xi = [1][\theta] + [-1,0]\begin{bmatrix}\xi_{11}\\ \xi_{12}\\ \end{bmatrix} = \theta - \xi_{11}$$

反應在第3個類別：

$$b'_{13}\theta + a'_{13}\xi = [2][\theta] + [-1,-1]\begin{bmatrix}\xi_{11}\\ \xi_{12}\\ \end{bmatrix} = 2\theta - (\xi_{11} + \xi_{12})$$

透過公式3-11，可計算受試者在不同類別的得分機率，讀者可將下面式子與PCM的例子作一比較，可發現在單向度多點計分時，MRCMLM可簡化為PCM模式。

受試者得0分的機率：

$$P(X_{11} = 1;\xi \mid \theta)$$
$$= \frac{\exp(b'_{11}\theta + a'_{11}\xi)}{\exp(b'_{11}\theta + a'_{11}\xi) + \exp(b'_{12}\theta + a'_{12}\xi) + \exp(b'_{13}\theta + a'_{13}\xi)}$$

$$= \frac{1}{1 + \exp(\theta - \xi_{11}) + \exp[2\theta - (\xi_{11} + \xi_{12})]}$$

受試者得1分的機率：

$$P(X_{12} = 1;\xi \mid \theta)$$
$$= \frac{\exp(b'_{12}\theta + a'_{12}\xi)}{\exp(b'_{11}\theta + a'_{11}\xi) + \exp(b'_{12}\theta + a'_{12}\xi) + \exp(b'_{13}\theta + a'_{13}\xi)}$$
$$= \frac{\exp(\theta - \xi_{11})}{1 + \exp(\theta - \xi_{11}) + \exp[2\theta - (\xi_{11} + \xi_{12})]}$$

受試者得2分的機率：

$$P(X_{13} = 1;\xi \mid \theta)$$
$$= \frac{\exp(b'_{13}\theta + a'_{13}\xi)}{\exp(b'_{11}\theta + a'_{11}\xi) + \exp(b'_{12}\theta + a'_{12}\xi) + \exp(b'_{13}\theta + a'_{13}\xi)}$$
$$= \frac{\exp[2\theta - (\xi_{11} + \xi_{12})]}{1 + \exp(\theta - \xi_{11}) + \exp[2\theta - (\xi_{11} + \xi_{12})]}$$

## 參、試題反應理論之測驗資料分析

應用試題反應理論分析測驗資料時，必須估計所選用的試題反應函數的參數，參數的估計常涉及艱深難懂的數學公式及繁瑣的計算過程，若沒有電腦套裝程式的即時配合，則在應用上會受到限制，但值得慶幸的是，電腦科技突飛猛進，各種適

用於試題反應理論的電腦軟體程式相繼問世，只要使用者學會這些程式，便能有效率的獲取所需要的參數估計值，進一步對測驗資料進行分析與解釋，相當方便。

目前比較廣泛使用的單向度試題反應理論軟體程式主要是由 Scientific Software International, Inc (簡稱SSI) 所發行的測驗套裝軟體，包含 BILOG-MG、MULTILOG、PARSCALE和TESTFACT四種軟體，這四種軟體所適用的理論模式是分析單向度試題反應理論之資料。若以本章所提及之理論模式而言，BILOG-MG、MULTILOG和TESTFACT主要適用於二元計分單參數羅吉斯模式、二參數羅吉斯模式以及三參數羅吉斯模式； PARSCALE除了分析二元計分的資料單參數羅吉斯模式、二參數羅吉斯模式以及三參數羅吉斯模式外，亦可以分析適用於部分給分模式和廣義部分給分模式之資料，讀者若需要更詳細的資料，可參閱http://www.ssicentral.com/irt/index.html網站，SSI發行之使用者參考手冊，IRT from SSI (Du Toit, 2003)，中詳細說明BILOG-MG、MULTILOG、PARSCALE和TESTFACT四種軟體之使用方法。多向度試題反應理論軟體程式則有Acer ConQuest 2.0 (Wu, et al., 2007)，下面將以例子按步驟示範BILOG-MG和Acer ConQuest2.0的使用程序和指令。

## 一、BILOG-MG程式應用

### (一)資料檔

受試者的原始作答反應以文書處理軟體輸入後，存成dat副檔名，舉例如表 1。為了資料分析的正確性，分析前先對受試者作一次資料篩檢：

1. 受試者於整份測驗均未作答，則刪除。
2. 檢視答對率低於0.25之受試者作答反應，若發現作答反應疑似亂答（有規則性的亂答），則刪除。

使用Bilog-MG進行試題分析所需準備資料，分別有三個所需要的檔案：

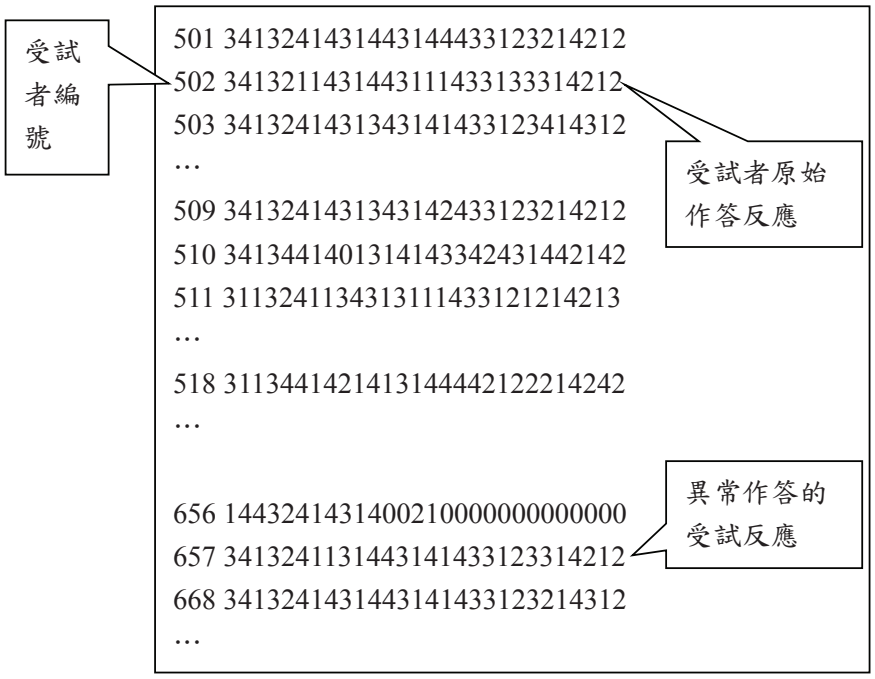
1. **DATAT1.dat**：受試者反應資料，為每位學生在每題上的作答反應，可以為「01」資料也可以為選項的類型「01234」，其中「01」資料為受試者於每題上的答對答錯情形，「0」為表示答錯該題，「1」為答對該題；「01234」則為受試者於每題上的選項反應類型，其中「0」表示受試者未作答或遺失資料，「1」為選擇選項「1」、「2」為選擇選項「2」，…以此類推。
2. **DATAT1.KEY**：各試題的正確答案檔案，當受試者反應資料為1234時，則需要此檔案。
3. **DATAT1.BLM**：資料分析程式檔。當安裝完成後，Bilog-MG會提供一個「EXAMPL」的資料夾，通常與 Bilog-MG 位於同一個目錄，裡面有許多範例檔案，建議使用EXAMPL01.BLM進行修改。

請將DATAT1.dat、DATAT1.KEY與DATAT1.BLM存在同一資料夾中，以利BILOG-MG執行程式。以下針對三個檔案分別進行說明：

1. DATAT1.dat：

本範例是333位受試者參加一28題的四選一形式的數學科選擇題測驗。第一列開始是受試者原始作答反應，前三欄是受試者編號，第五欄開始是受試者在28題之原始作答反應，「0」表示受試者未作答或遺失資料，如編號656的受試者未作答的題數過多，建議資料分析前應先刪除，以免影響分析之正確性，經資料篩檢後，剩下325位受試者，存成DATA1.dat。

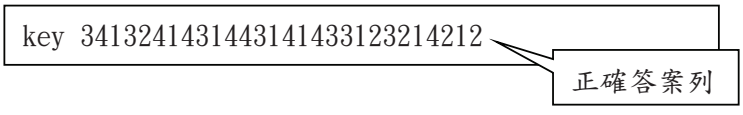
表 1　BILOG-MG程式所需要之資料檔格式




2. DATAT1.KEY：

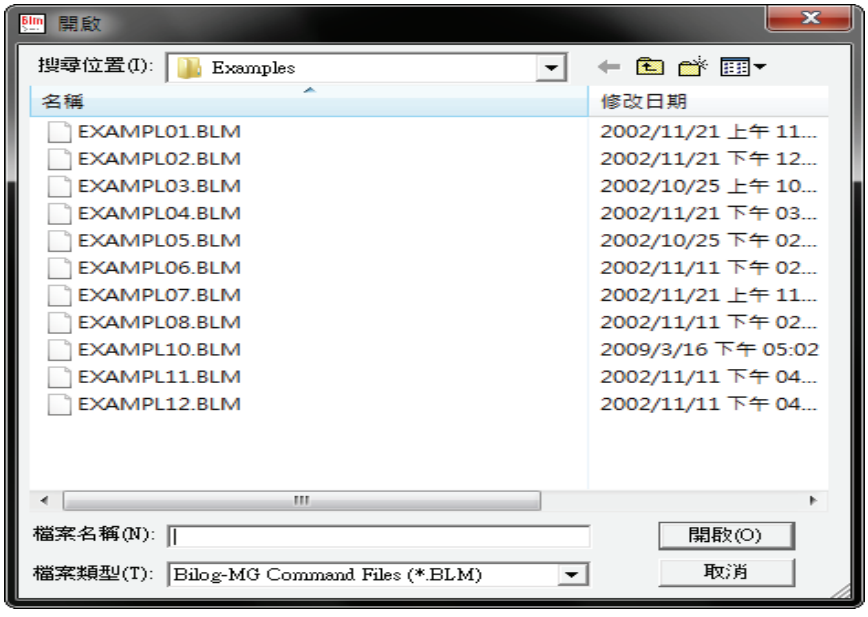
答案儲存模式，存為key副檔名型式，須注意正確答案的起始位置須與DATA1.dat的試題起始位置相同，前三欄是「key」指令，第五欄開始是28題之正確答案，其格式如下：

表 2　BILOG-MG程式所需要之答案檔案格式



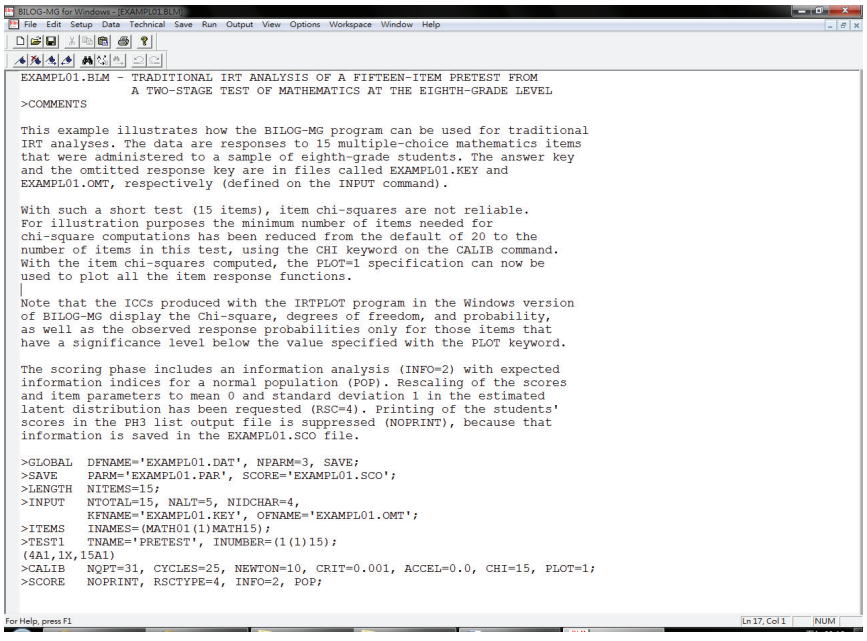
3. DATAT1.BLM：

為了讀者使用方便，建議開啟BILOG-MG之範例檔案EXAMPL01.BLM檔案修改，如圖 7，開啟後會看見範例檔案如圖 8，各程式說明如下：




註：畫面擷取自BILOG-MG軟體

圖 7　BILOG-MG之範例檔案




註：畫面擷取自BILOG-MG軟體

圖 8　BILOG-MG範例檔案之程式畫面

範例檔最上方為說明此範例檔所適合的模式與測驗。前二列總是標題列，若標題列只有一列則第二列要空白下個指令由第三列開始。每一列長度為80字元。結尾必須要特殊符號如＞或；

EXAMPL01.BLM - TRADITIONAL IRT ANALYSIS OF A FIFTEEN-ITEM PRETEST FROM A TWO-STAGE TEST OF MATHEMATICS AT THE EIGHTH-GRADE LEVEL

註：畫面擷取自BILOG-MG軟體

COMMENTS下方為說明此份測驗的試題，以及所使用的模式，分別的使用的檔案名稱，使用者可以在此敘述該測驗的內容，解釋程式執行過程。可輸入一列或更多列，此指令要在GLOBAL之上，每一列長度為80 字元，不需要以分號做為指令的結束，一般為句號。

>COMMENTS

This example illustrates how the BILOG-MG program can be used for traditional IRT analyses. The data are responses to 15 multiple-choice mathematics items that were administered to a sample of eighth-grade students. The answer key and the omtitted response key are in files called EXAMPL01.KEY and EXAMPL01.OMT, respectively (defined on the INPUT command).

註：畫面擷取自BILOG-MG軟體

自GLOBAL到SCORE為利用Bilog-MG進行分析的程式主體部分，各指令說明如下：

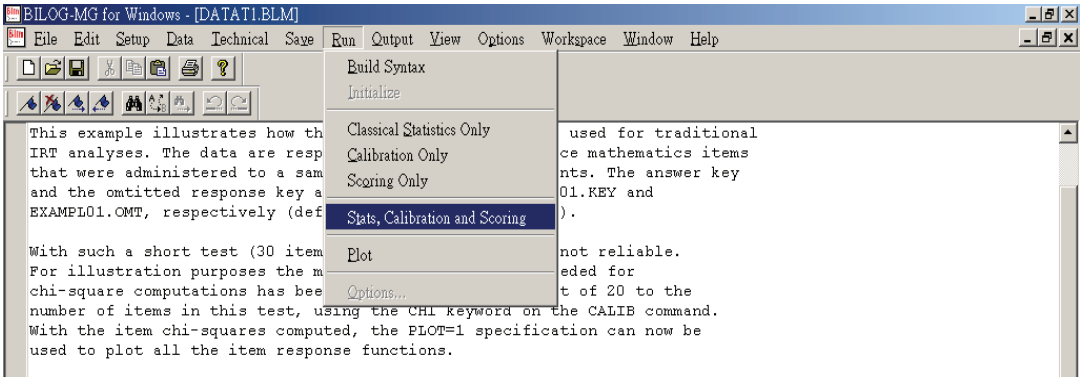
```
>GLOBAL   DFNAME='EXAMPL01.DAT', NPARM=3, SAVE;
>SAVE     PARM='EXAMPL01.PAR', SCORE='EXAMPL01.SCO';
>LENGTH   NITEMS=30;
>INPUT    NTOTAL=30, NALT=5, NIDCHAR=4,
          KFNAME='EXAMPL01.KEY', OFNAME='EXAMPL01.OMT';
>ITEMS    INAMES=(MATH01(1)MATH30);
>TEST1    TNAME='PRETEST', INUMBER=(1(1)30);
          (4A1,1X,30A1)
>CALIB    NQPT=31, CYCLES=25, NEWTON=10, CRIT=0.001,
          ACCEL=0.0, CHI=30, PLOT=1;
>SCORE    ;
```

註：畫面擷取自BILOG-MG軟體

- 指令>GLOBAL中，為提供程式執行所需要輸入檔案名稱及資訊，設定資料檔名，以及選擇所欲分析之模式，並決定是否儲存分析後資料。
- DFNAME為受試者作答反應檔案，若檔案儲存為 EXAMPL01.DAT，則指令可以撰寫為DFNAME='EXAMPL01.DAT'。
- NPARM為所欲使用之模式，其中NPARM=1為使用單參數模式進行分析NPARM=2為使用二參數模式進行分析、NPARM=3為使用三參數模式進行分析。
- 指令>SAVE為告知是否儲存分析結果，分別定義儲存試題、能力值檔案的檔案名稱；指令>SAVE中提供輸出檔案的名稱，檔名不得超過128 字元。而為了使用SAVE指令，GLOBAL 指令中必須有save選項。PARM='DATAT1.PAR'為表示試題參數檔案名稱為DATAT1.PAR、SCORE='DATAT1.SCO'則表示能力值參數檔案名稱。
- 指令>LENGTH，NITEMS=30為表示測驗題目數為30題，若資料分析使用者測驗題數不相同者，可以自行修改。
- 指令>INPUT為分別定義測驗相關訊息資料，其中NTOTAL=30表示測驗題目數為30題。
- NALT=5表示該份測驗試題的選項數目。
- NIDCHAR=4表示學生的編碼位元數，如學生編號是0001，表示4個位元數。
- KFNAME則為答案檔案名稱，當學生作答反應為01資料時則無需此檔案。
- 指令>ITEMS中為說明所有試題的名稱及其相對應的編號INAMES= (MATH01 (1) MATH30)，此份測驗為定義第一道試題名稱為MATH01，以此類推至最後一道試題MATH02、…、MATH30。
- >TEST1指令中，TNAME為該測驗之名稱，若該測驗名稱為PRETEST，則指令可以寫為TNAME='PRETEST'。
- INUMBER= (1 (1) 30) 則表示試題題號從1到30中間間隔1題。
- (4A1,1X,30A1) 則分別表示了資料的格式，其中4A1表示了總共有4位ID碼、1X則表示了在ID碼以及作答反應間有一位空格、30A1表示了共有30道試題的作答反應。
- 指令>CALIB中有控制試題參數估計的程序和說明試題參數的分配函數等多項的相關設定，NQPT=31是近似分配函數的切割點，值越大估計越準確，但估計時間越久； CYCLES=25 是最大期望演算法（expectation-maximization algorithm）迭代次數設定, NEWTON=10, 是牛頓迭代法的次數設定；CRIT=0.001是最大期望演算法和牛頓迭代法的收斂指標；需要依據自行測驗更改的有CHI=30，其中的30表示施測的題數，必須依據該份測驗而有所變動，其餘的相關設定則可依據EXAMPLE01中設定即可。

■ 指令>SCORE個別受試者或反應模式的起始計分；計算試題與測驗資訊並畫訊息曲線；在樣本或潛在的機率分配中，依特定的平均值和標準差重新計分，如原本軟體預設值是受試者平均值是0，標準差是1，想要重新計分為平均值是250，標準差是50。

當所有檔案皆準備完成時，即可執行BILOG-MG分析，步驟為-->Run--> Stats, Calibration and Scoring




註：畫面擷取自BILOG-MG軟體

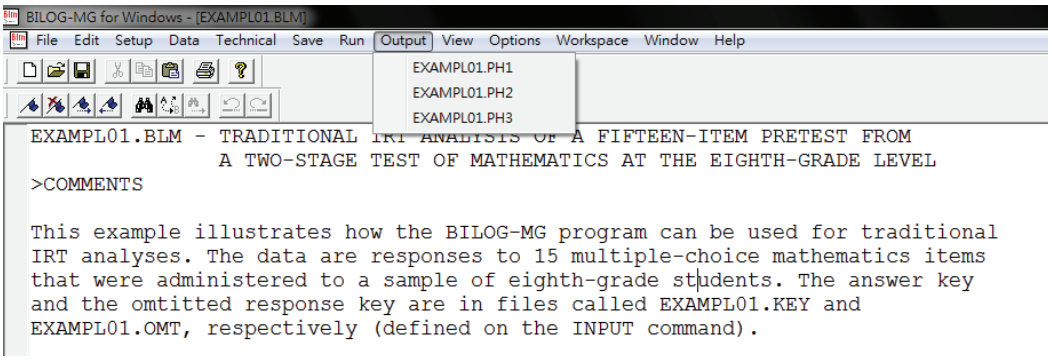
圖 9　執行BILOG-MG之程式畫面

資料分析完成之後，會出現以下的訊息，訊息視窗如下：


註：畫面擷取自BILOG-MG軟體

圖 10　執行成功BILOG-MG之程式畫面

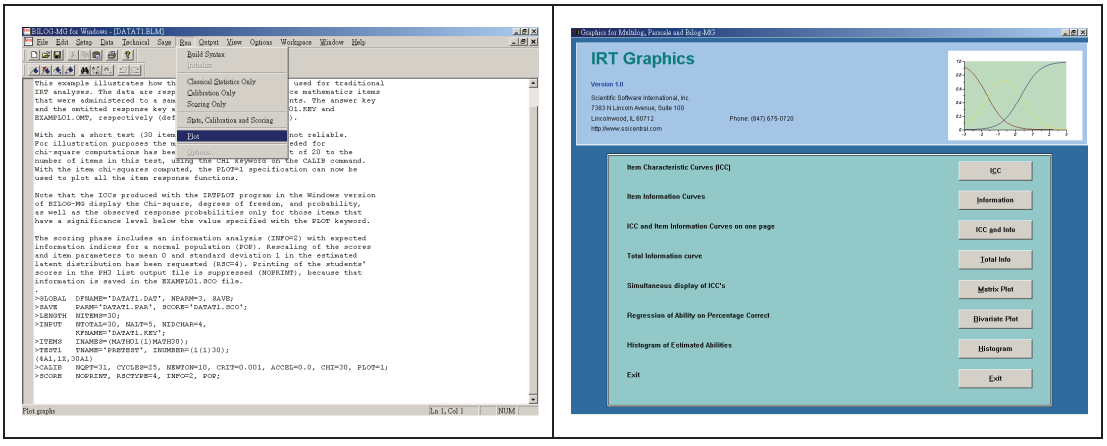
按「確定」後，可以回到程式命令視窗，此時點下「Output」可以觀看分析結果，分別有EXAMPLE01.PH1（如圖11）、EXAMPLE01.PH2、EXAMPLE01.PH3三個檔案，三個檔案分別呈現不同的結果，結果說明於下節呈現：




註：畫面擷取自BILOG-MG軟體

圖 11　BILOG-MG輸出檔案之程式畫面

此外，Bilog-MG提供了試題特徵曲線，測驗特徵曲線，以及每道試題的訊息函數，可以透過Run→Plot看執行結果，如圖 12。



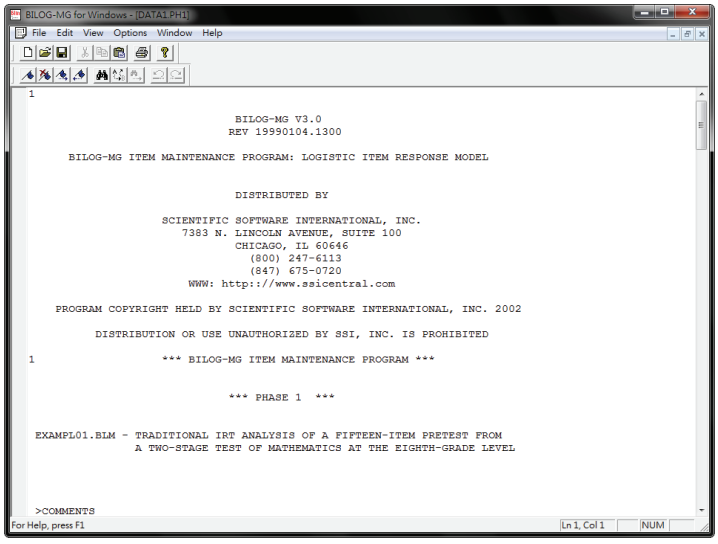

註：畫面擷取自BILOG-MG軟體

圖 12　BILOG-MG每一題試題圖形介面之程式畫面

## (二)分析結果

資料分析完成後，回到程式命令視窗，此時點下「Output」可以觀看分析結果，分別有EXAMPLE01.PH1、EXAMPLE01.PH2、EXAMPLE01.PH3三個檔案，三個檔案分別呈現不同階段的估計結果，其中EXAMPLE01.PH1呈現古典測驗理論為基礎之難度與鑑別度資料，如圖 13。在PH2這個檔案，讀者可藉由-2 LOG LIKELIHOOD檢查參數估計是否收斂，如果-2 LOG LIKELIHOOD的值越來越小，表示估計值有收斂。此外在資料夾中另外可以增加 EXAMPL01.PAR、EXAMPL01.SCO兩個檔案，則是分別儲存了試題參數值以及受試者能力參數值，各檔案解釋說明如下：

## 1. EXAMPLE01.PH1




ITEM STATISTICS FOR SUBTEST PRETEST

| ITEM | | #TRIED | #RIGHT | | LOGIT/1.7 | ITEM*TEST PEARSON | CORRELATION BISERIAL |
|---|---|---|---|---|---|---|---|
| 1 | MATH01 | 325.0 | 305.0 | 93.8 | -1.60 | 0.280 | 0.553 |
| 2 | MATH02 | 325.0 | 277.0 | 85.2 | -1.03 | 0.163 | 0.250 |
| 3 | MATH03 | 325.0 | 302.0 | 92.9 | -1.51 | 0.256 | 0.485 |
| 4 | MATH04 | 325.0 | 228.0 | 70.2 | -0.50 | 0.544 | 0.717 |
| 5 | MATH05 | 325.0 | 309.0 | 95.1 | -1.74 | 0.278 | 0.588 |
| 6 | MATH06 | 325.0 | 291.0 | 89.5 | -1.26 | 0.212 | 0.359 |
| 7 | MATH07 | 325.0 | 251.0 | 77.2 | -0.72 | 0.206 | 0.286 |
| 8 | MATH08 | 325.0 | 243.0 | 74.8 | -0.64 | 0.246 | 0.335 |
| 9 | MATH09 | 325.0 | 225.0 | 69.2 | -0.48 | 0.549 | 0.720 |
| 10 | MATH10 | 325.0 | 252.0 | 77.5 | -0.73 | 0.470 | 0.655 |
| 11 | MATH11 | 325.0 | 263.0 | 80.9 | -0.85 | 0.032 | 0.046 |
| 12 | MATH12 | 325.0 | 241.0 | 74.2 | -0.62 | 0.422 | 0.571 |
| 13 | MATH13 | 325.0 | 240.0 | 73.8 | -0.61 | 0.488 | 0.660 |
| 14 | MATH14 | 325.0 | 304.0 | 93.5 | -1.57 | 0.323 | 0.628 |
| 15 | MATH15 | 325.0 | 222.0 | 68.3 | -0.45 | 0.361 | 0.471 |
| 16 | MATH16 | 325.0 | 250.0 | 76.9 | -0.71 | 0.203 | 0.281 |
| 17 | MATH17 | 325.0 | 224.0 | 68.9 | -0.47 | 0.319 | 0.418 |
| 18 | MATH18 | 325.0 | 250.0 | 76.9 | -0.71 | 0.536 | 0.743 |
| 19 | MATH19 | 325.0 | 290.0 | 89.2 | -1.24 | 0.412 | 0.689 |
| 20 | MATH20 | 325.0 | 301.0 | 92.6 | -1.49 | 0.400 | 0.747 |
| 21 | MATH21 | 325.0 | 216.0 | 66.5 | -0.40 | 0.594 | 0.769 |
| 22 | MATH22 | 325.0 | 262.0 | 80.6 | -0.84 | 0.408 | 0.587 |
| 23 | MATH23 | 325.0 | 172.0 | 52.9 | -0.07 | 0.423 | 0.531 |
| 24 | MATH24 | 325.0 | 297.0 | 91.4 | -1.39 | 0.443 | 0.790 |
| 25 | MATH25 | 325.0 | 271.0 | 83.4 | -0.95 | 0.358 | 0.534 |
| 26 | MATH26 | 325.0 | 237.0 | 72.9 | -0.58 | 0.426 | 0.572 |
| 27 | MATH27 | 325.0 | 256.0 | 78.8 | -0.77 | 0.439 | 0.619 |
| 28 | MATH28 | 325.0 | 200.0 | 61.5 | -0.28 | 0.418 | 0.532 |

試題名稱
試題通過率，古典難度
試題編號
樣本數
答對人數
古典鑑別度

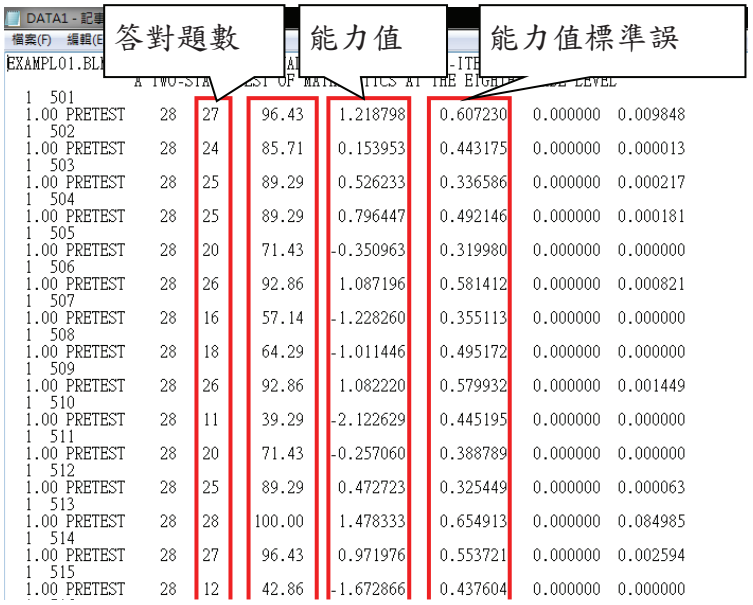
註：畫面擷取自BILOG-MG軟體

圖 13　BILOG-MG 之ph1畫面

## 2. EXAMPL01.PAR


圖 14　BILOG-MG中.PAR檔案之內容

## 3. EXAMPL01.SCO



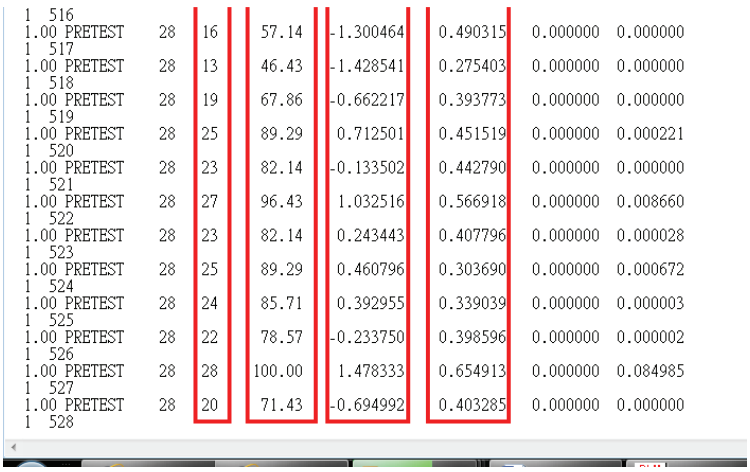



圖 15　BILOG-MG中.SCO檔案之內容

## 二、ConQuest程式應用

Acer ConQuest 2.0（Wu, et.al, 2007）的估計是以多向度隨機係數多項洛基模式（MRCMLM）為基礎，可估計單向度單參數邏輯斯模式和多向度單參數模式，目前PISA則使用Acer ConQuest 2.0進行參數估計（OECD, 2014），下面將以兩個能力向度為例，說明如何使用ConQuest 2.0進行多向度單參數IRT之參數估計。

### (一)資料檔

受試者的原始作答反應以文書處理軟體輸入後，存成dat副檔名。為了資料分析的正確性，分析前先對受試者作一次資料篩檢：

1. 受試者於整份測驗均未作答，則刪除。
2. 受試者於整份測驗中，連續五題試題未作答（含以上），則刪除。
3. 檢視答對率低於0.25之受試者作答反應，若發現作答反應疑似亂答（有規則性的亂答），則刪除。

使用ConQuest進行試題分析所需準備資料，分別有三個所需要的檔案：

1. **ex7a.cqc**：資料分析程式檔，建議使用軟體提供之Example檔案進行修改。
2. **ex1.dat**：受試者反應資料，為每位學生在每題上的作答反應，可以為「01」資料也可以為選項的類型「01234」或「ABCD」，其中「01」資料為受試者於每題上的答對答錯情形，「0」為表示答錯該題，「1」為答對該題；「01234」則為受試者於每題上的選項反應類型，其中「0」表示受試者未作答或遺失資料，「1」為選擇選項「1」、「2」為選擇選項「2」、…以此類推。
3. **ex1.lab**：此檔案為選擇題試題名稱檔案，可以使用亦可以不使用，使用者可以依據資型需求修改此檔案。

請將ex7a.cqc、ex1.dat與ex1.lab存在同一資料夾中，以執行ConQuest程式。以下針對三個檔案分別進行說明：

1. **ex7a.cqc**：

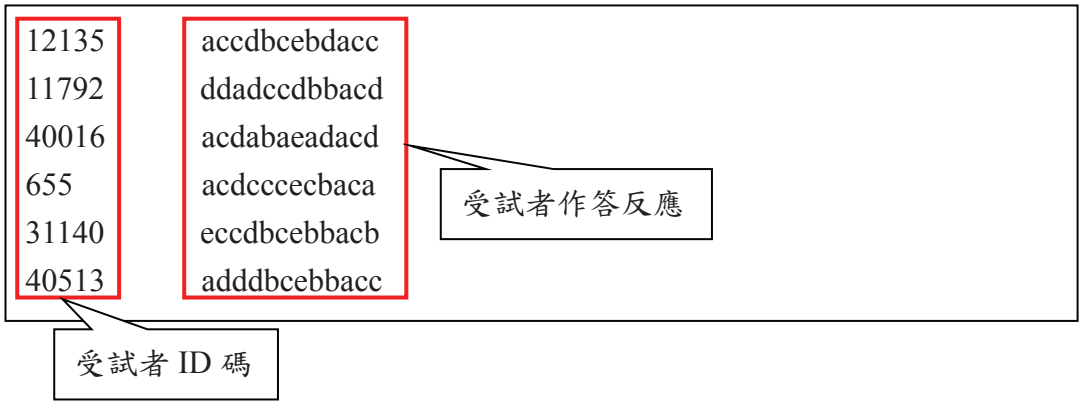
```
datafile ex1.dat;
format id 1-5    responses 12-23;
labels << ex1.lab;
key acddbcebbacc ! 1;
score (0,1) (0,1) ()! items(1-6);
score (0,1) () (0,1)! items(7-12);
model item;
estimate ;
show !estimates=latent,tables=1:2:3:9>> ex7a.shw;
itanal >> ex7a.itn;
show cases !estimates=eap >> ex7a.eap;
show cases !estimates=mle >> ex7a.mle;
```

註：畫面擷取自ConQuest軟體

- 程式碼第一行datafile為定義作答反應類型的檔案名稱為ex1.dat，每行結尾需以;結束。
- format指令下為定義反應類型的檔案資料，其中id 1-5為表示受試者的ID碼為1-5個欄位，responses 12-23為表示受試者的作答反應為第12-23的欄位。
- labels << ex1.lab則為輸入自行定義各題試題的檔案名稱。
- key acddbcebbacc！1為定義各題試題的正確答案，以此檔為例第一題的正確答案為a、第二題正確答案為c，最後!1為定義每一題的得分類型。
- 接著兩行的score為分別定義試題所屬的向度，第一行score (0,1) (0,1) ()！items (1-6) ;說明了資料形式為01的方式，其中 (0,1) 為屬於第一個向度，試題1-6為第一向度的試題，第二行score (0,1) () (0,1)！items (7-12) ;說明了資料形式為01的方式，其中 (0,1) 為屬於第二個向度，試題7-12為第二向度的試題。
- model item是定義反應資料的變數名稱，本例子是分析試題所產生的作答反應，故為item。
- estimate為使用軟體中的預設的設定進行估計。能力值的估計預設方法是Gauss-Hermite法，node預設的設定數是15；預設計算每一題模式適配度值。
- show cases !estimates=eap >> ex7a.eap為估計個體能力值時使用期望後驗法 (expected

a posteriori，EAP) 的方法，並將其儲存成ex7a.eap；show cases !estimates=mle >> ex7a.mle為估計個體能力值時使用最大概似估計法 (maximum likelihood，ML) 的方法，並將其儲存成ex7a.mle。此兩種估計法之理論，有興趣之讀者可參閱Acer ConQuest 2.0（Wu, et.al, 2007）之使用手冊。

2. **ex1.dat**：




註：畫面擷取自ConQuest軟體

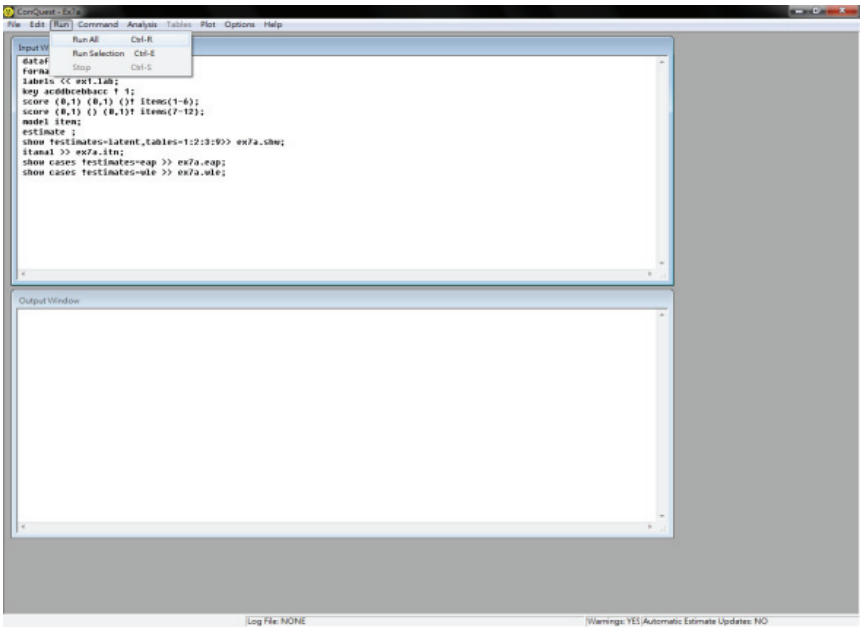
ex1.dat檔案為儲存受試者的作答反應檔案，以ex1.dat為例，前五碼為受試者的ID編碼，第12-23為受試者的作答反應，在此依據受試者的原始作答選項進行紀錄。

3. **ex1.lab**：


註：畫面擷取自ConQuest軟體

ex1.lab為定義各試題的名稱，其中1234為題號的部分，題號後方為該題號的試題名稱，以題號為1的試題為例，其試題名稱為BSMMA01。

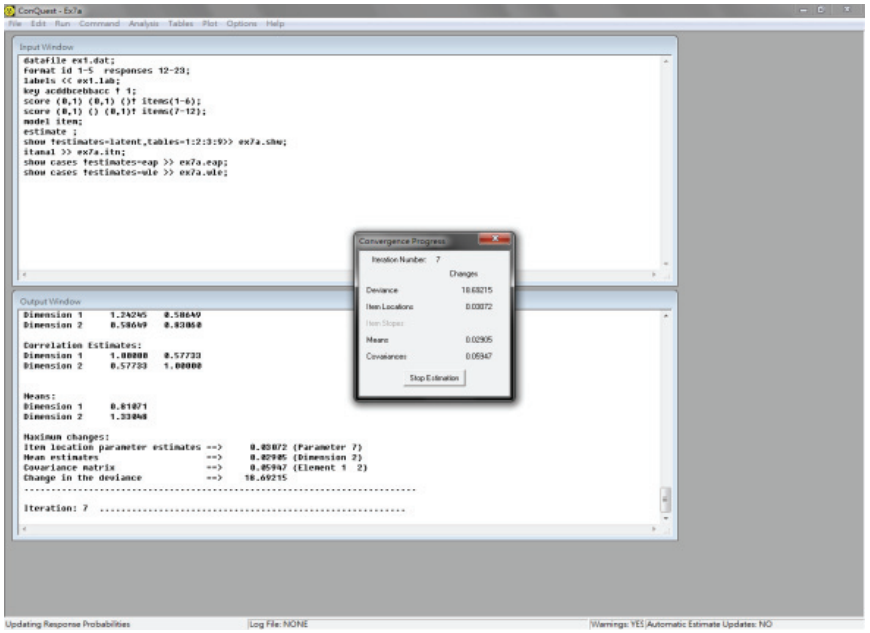
將上述的三個資料分別準備好之後，開啟ex7a.cqc的檔案，File →Open，選擇三個檔案儲存路徑，需特別注意的是儲存路徑中，不能有中文資料夾名稱或檔名，打開ex7a.cqc，確定無誤後，Run →Run All即可進行分析（圖16）。



註：畫面擷取自ConQuest軟體

圖 16　ConQuest 之執行畫面一

程式執行之後，會出現以下畫面（圖17），程式執行結束後，資料夾檔案會出現ex7a.shw、ex7a.eap、ex7a.mle，各檔案說明如下節所示。

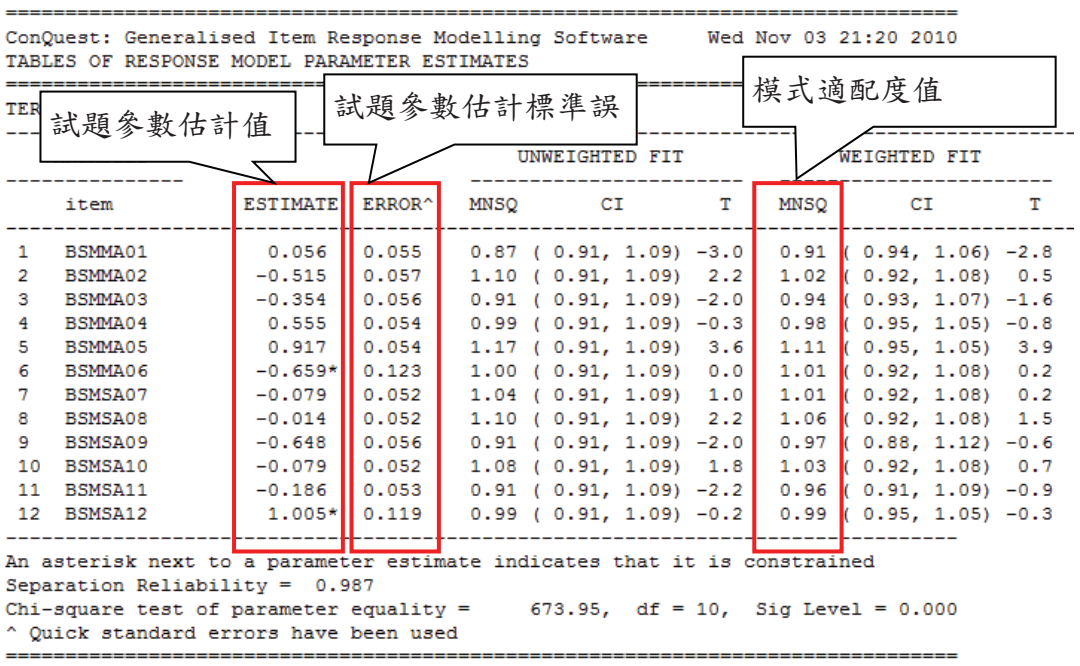


註：畫面擷取自ConQuest軟體

圖 17　ConQuest 之執行畫面二

## (二)分析結果

### 1. ex7a.shw



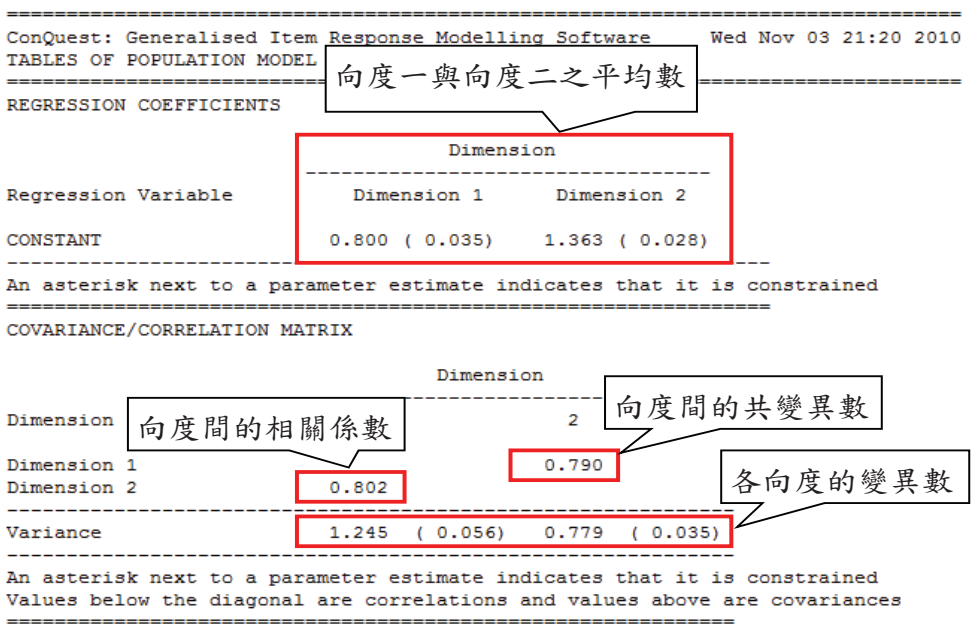

```
============================================================================
ConQuest: Generalised Item Response Modelling Software     Wed Nov 03 21:20 2010
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
============================================================================
TER
---------------------------------------------------------------------------
                                    UNWEIGHTED FIT                WEIGHTED FIT
---------------
     item          ESTIMATE  ERROR^  MNSQ      CI        T     MNSQ      CI        T
---------------------------------------------------------------------------
 1   BSMMA01          0.056   0.055  0.87 ( 0.91, 1.09) -3.0   0.91 ( 0.94, 1.06) -2.8
 2   BSMMA02         -0.515   0.057  1.10 ( 0.91, 1.09)  2.2   1.02 ( 0.92, 1.08)  0.5
 3   BSMMA03         -0.354   0.056  0.91 ( 0.91, 1.09) -2.0   0.94 ( 0.93, 1.07) -1.6
 4   BSMMA04          0.555   0.054  0.99 ( 0.91, 1.09) -0.3   0.98 ( 0.95, 1.05) -0.8
 5   BSMMA05          0.917   0.054  1.17 ( 0.91, 1.09)  3.6   1.11 ( 0.95, 1.05)  3.9
 6   BSMMA06         -0.659*  0.123  1.00 ( 0.91, 1.09)  0.0   1.01 ( 0.92, 1.08)  0.2
 7   BSMSA07         -0.079   0.052  1.04 ( 0.91, 1.09)  1.0   1.01 ( 0.92, 1.08)  0.2
 8   BSMSA08         -0.014   0.052  1.10 ( 0.91, 1.09)  2.2   1.06 ( 0.92, 1.08)  1.5
 9   BSMSA09         -0.648   0.056  0.91 ( 0.91, 1.09) -2.0   0.97 ( 0.88, 1.12) -0.6
10   BSMSA10         -0.079   0.052  1.08 ( 0.91, 1.09)  1.8   1.03 ( 0.92, 1.08)  0.7
11   BSMSA11         -0.186   0.053  0.91 ( 0.91, 1.09) -2.2   0.96 ( 0.91, 1.09) -0.9
12   BSMSA12          1.005*  0.119  0.99 ( 0.91, 1.09) -0.2   0.99 ( 0.95, 1.05) -0.3
---------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =   0.987
Chi-square test of parameter equality =     673.95,  df = 10,  Sig Level = 0.000
^ Quick standard errors have been used
============================================================================
```

註：畫面擷取自ConQuest軟體

圖 18　ConQuest 中.SHW之部分內容(一)

圖18為ex7a.shw的檔案中所呈現的試題參數估計值，以及其估計標準誤，其中試題參數估計值的第6以及第12題的部分，在估計值旁有*表示為估計值為限制(constraints) 模式，用來使得整體平均難度值估計值為零的情況。右邊欄位則呈現加權和未加權的模式適配度的值，讀者可以加權的模式適配度的值為主，代表理論模式與實證資料之適配情形，如第一題的值是0.91，位於信賴區間[0.94,1.06]之外，表示模式適配稍差一些；第二題的值是1.02，位於信賴區間[0.92,1.08]之間，表示模式適配良好。

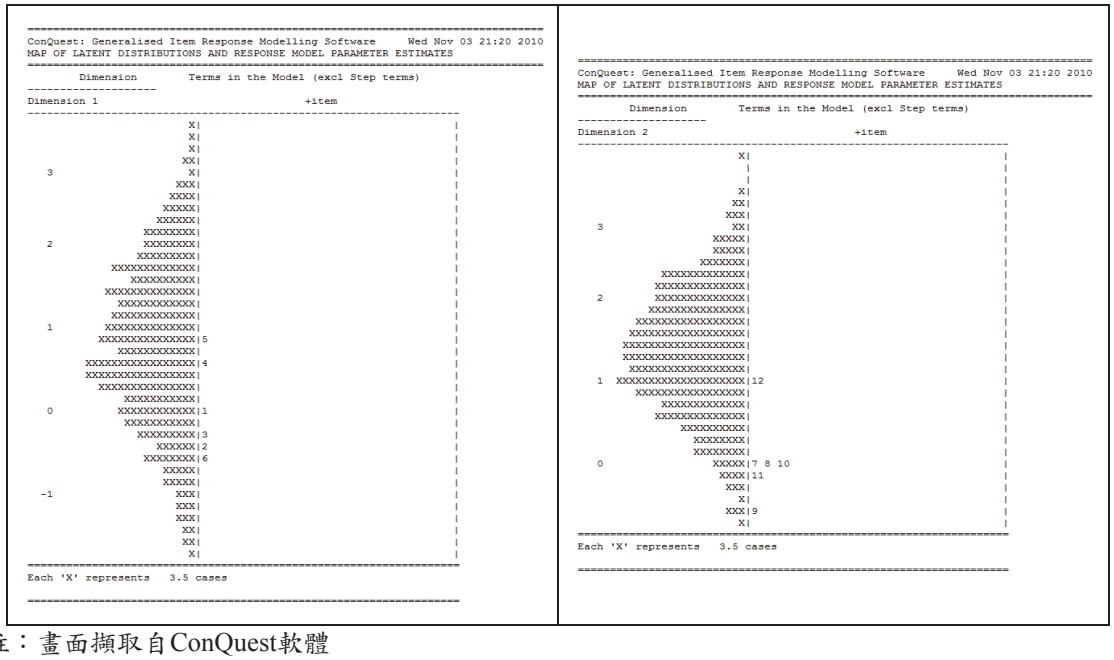
圖19為ex7a.shw的檔案中所呈現的能力參數估計值之群體平均數，受試者在向度一的能力平均值是0.800；在向度二的能力平均值是1.363，由於向度一和向度二是各自校正於自身的量尺，此兩個平均數不能直接拿來比較，讀者不能說向度二的能力平均高於向度一，這樣的比較是沒有意義的。兩個能力向度之相關係數與共變異數和各能力向度之變異數，左下角是向度一和向度二的相關係數0.802，右上角是向度一和向度二的共變異數0.790，向度一的變異數是 1.245，向度二的變異數是 0.779。




註：畫面擷取自ConQuest軟體

圖 19　ConQuest 中.SHW之部分內容(二)

圖20為各向度上受試者以及試題的表現圖，縱軸為-3~3的能力值以及難度值，其中中央軸的右方，為各試題所處的難度位置，並用數字標記其題號，中央軸的左方為受試者的能力值位置，每個‘X’表示了3.5位受試者。

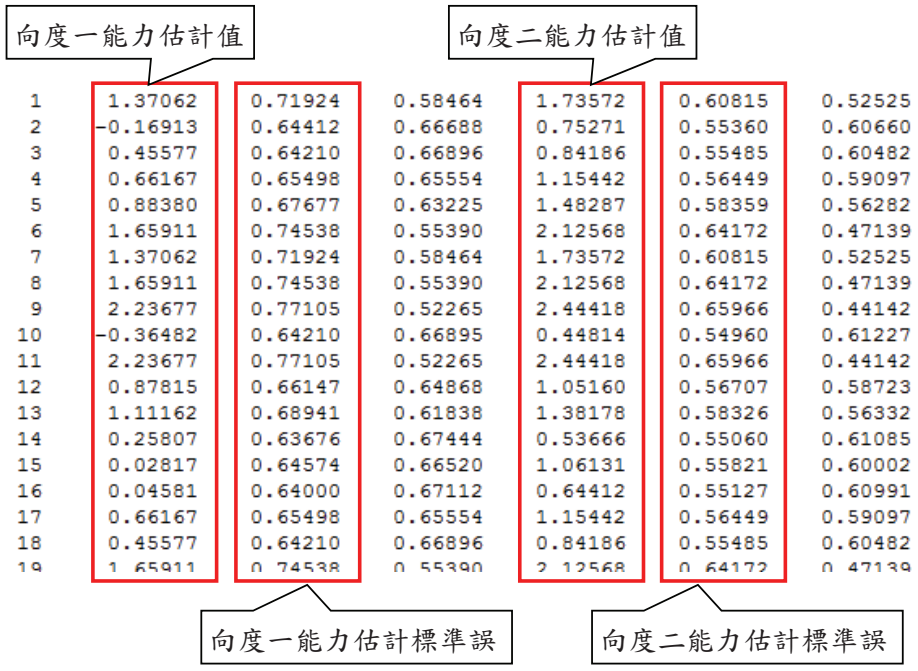



註：畫面擷取自ConQuest軟體

圖 20　ConQuest 中.SHW之部分內容（三）

圖21是使用期望後驗法估計而得每位受試者在兩個向度之能力值與標準誤。圖22是使用最大概似估計法估計每位受試者在兩個向度之能力值與標準誤。

2. **ex7a.eap**




| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1.37062 | 0.71924 | 0.58464 | 1.73572 | 0.60815 | 0.52525 |
| 2 | -0.16913 | 0.64412 | 0.66688 | 0.75271 | 0.55360 | 0.60660 |
| 3 | 0.45577 | 0.64210 | 0.66896 | 0.84186 | 0.55485 | 0.60482 |
| 4 | 0.66167 | 0.65498 | 0.65554 | 1.15442 | 0.56449 | 0.59097 |
| 5 | 0.88380 | 0.67677 | 0.63225 | 1.48287 | 0.58359 | 0.56282 |
| 6 | 1.65911 | 0.74538 | 0.55390 | 2.12568 | 0.64172 | 0.47139 |
| 7 | 1.37062 | 0.71924 | 0.58464 | 1.73572 | 0.60815 | 0.52525 |
| 8 | 1.65911 | 0.74538 | 0.55390 | 2.12568 | 0.64172 | 0.47139 |
| 9 | 2.23677 | 0.77105 | 0.52265 | 2.44418 | 0.65966 | 0.44142 |
| 10 | -0.36482 | 0.64210 | 0.66895 | 0.44814 | 0.54960 | 0.61227 |
| 11 | 2.23677 | 0.77105 | 0.52265 | 2.44418 | 0.65966 | 0.44142 |
| 12 | 0.87815 | 0.66147 | 0.64868 | 1.05160 | 0.56707 | 0.58723 |
| 13 | 1.11162 | 0.68941 | 0.61838 | 1.38178 | 0.58326 | 0.56332 |
| 14 | 0.25807 | 0.63676 | 0.67444 | 0.53666 | 0.55060 | 0.61085 |
| 15 | 0.02817 | 0.64574 | 0.66520 | 1.06131 | 0.55821 | 0.60002 |
| 16 | 0.04581 | 0.64000 | 0.67112 | 0.64412 | 0.55127 | 0.60991 |
| 17 | 0.66167 | 0.65498 | 0.65554 | 1.15442 | 0.56449 | 0.59097 |
| 18 | 0.45577 | 0.64210 | 0.66896 | 0.84186 | 0.55485 | 0.60482 |
| 19 | 1.65911 | 0.74538 | 0.55390 | 2.12568 | 0.64172 | 0.47139 |

圖 21　ConQuest 中.EAP之內容

3. **ex7a.mle**


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.00 | 6.00 | 5.00 | 6.00 | 1.42012 | 1.03235 | 1.38735 | 1.02665 |
| 2 | 2.00 | 6.00 | 4.00 | 6.00 | -0.65412 | 0.88259 | 0.61942 | 0.87834 |
| 3 | 4.00 | 6.00 | 3.00 | 6.00 | 0.64143 | 0.88691 | -0.01349 | 0.83947 |
| 4 | 4.00 | 6.00 | 4.00 | 6.00 | 0.64168 | 0.88671 | 0.61942 | 0.87834 |
| 5 | 4.00 | 6.00 | 5.00 | 6.00 | 0.64143 | 0.88691 | 1.38735 | 1.02665 |
| 6 | 5.00 | 6.00 | 6.00 | 6.00 | 1.42018 | 1.03225 | 2.72314 | 1.63175 |
| 7 | 5.00 | 6.00 | 5.00 | 6.00 | 1.42012 | 1.03235 | 1.38735 | 1.02665 |
| 8 | 5.00 | 6.00 | 6.00 | 6.00 | 1.42018 | 1.03225 | 2.72314 | 1.63175 |
| 9 | 6.00 | 6.00 | 6.00 | 6.00 | 2.75387 | 1.63290 | 2.72314 | 1.63175 |
| 10 | 2.00 | 6.00 | 3.00 | 6.00 | -0.65412 | 0.88259 | -0.01350 | 0.83947 |
| 11 | 6.00 | 6.00 | 6.00 | 6.00 | 2.75387 | 1.63290 | 2.72314 | 1.63175 |
| 12 | 5.00 | 6.00 | 3.00 | 6.00 | 1.42012 | 1.03235 | -0.01349 | 0.83947 |
| 13 | 5.00 | 6.00 | 4.00 | 6.00 | 1.42012 | 1.03235 | 0.61915 | 0.87856 |
| 14 | 4.00 | 6.00 | 2.00 | 6.00 | 0.64168 | 0.88671 | -0.63862 | 0.87257 |
| 15 | 2.00 | 6.00 | 5.00 | 6.00 | -0.65412 | 0.88259 | 1.38738 | 1.02660 |
| 16 | 3.00 | 6.00 | 3.00 | 6.00 | -0.00924 | 0.84919 | -0.01354 | 0.83947 |
| 17 | 4.00 | 6.00 | 4.00 | 6.00 | 0.64168 | 0.88671 | 0.61942 | 0.87834 |
| 18 | 4.00 | 6.00 | 3.00 | 6.00 | 0.64143 | 0.88691 | -0.01349 | 0.83947 |
| 19 | 5.00 | 6.00 | 6.00 | 6.00 | 1.42018 | 1.03225 | 2.72314 | 1.63175 |
| 20 | 1.00 | 6.00 | 4.00 | 6.00 | -1.41597 | 1.02406 | 0.61942 | 0.87834 |
| 21 | 6.00 | 6.00 | 6.00 | 6.00 | 2.75387 | 1.63290 | 2.72314 | 1.63175 |
| 22 | 2.00 | 6.00 | 6.00 | 6.00 | -0.65412 | 0.88259 | 2.72314 | 1.63175 |

圖 22　ConQuest 中.MLE之內容

## 三、挑選試題之準則

現代測驗理論是以模式為基礎的測驗理論，若要估計準確，需考慮樣本數大小之問題，教師自行編製測驗時，建議採用跨校或跨班級的方式收集約300人以上之樣本，進行分析會較準確。另外上述的測驗分析軟體，同一份測驗必須使用同一個理論模式，如整份測驗都是使用單向度單參數羅吉斯模式；有的軟體，如 BILOG-MG，整份測驗都必須是二元計分，讀者須清楚測驗分析軟體之限制，方能應用得宜。

試題特性分析在測驗資料處理上是相當重要的一環，藉由測量模式或一般性描述統計分析試題各項參數的穩定性可以提高測驗的品質。試題反應理論模式之鑑別度參數越大表示試題越能區辨不同能力的受試者，難度參數越大表示試題難度越高，猜測度參數介於 0 與 1 之間，猜測度參數越大表示受試者越容易猜對此題。

測驗中所使用的試題分析完後如何進行檢測與篩選，以下介紹臺灣學生學習成就資料庫所使用之篩選準則，提供讀者於試題分析之參考依據 (郭伯臣、曾建銘、吳慧珉，2012)。

1.試題鑑別度參數介於0 ~ 0.4之間，$0 < a < 0.4$；

2.試題難度參數小於或等於 –3，$b \leq -3$；

3.試題難度參數大於或等於3，$b \geq 3$；

4.試題猜測度參數大於或等於 0.3，$c \geq 0.3$。

符合上述條件之試題需特別留意，可考慮修改題目。若是試題參數估計時無法收斂或產生異常的情況（例如 $a = 8.88$、$b = 13.27$ 等）可考慮刪題。

另外，有些軟體 (如ConQuest) 有提供每一題的模式適配度的值，讀者亦可以考慮模式適配度的值，將嚴重不適配的題目，即遠離於信賴區間的題目刪除重新分析。

## 參考文獻

王文中 (2004)。Rasch測量理論與其在教育與心理之應用。**教育與心理研究**，27 (4)，637-694。

余民寧（2009）。**試題反應理論 (IRT) 及其應用**。台北：心理出版社。

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

郭伯臣、曾建銘、吳慧珉主編（2012）。**大型標準化測驗建置流程應用於TASA之研究**。國家教育研究院。

許擇基、劉長萱（1991）。**試題作答理論簡介**。台北：中國行為科學社。

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, 47 (1), 105-113.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Du Toit, M. ( 2003). *IRT from SSI.Scientific Software International,* Inc.

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

Embretson, S., & Yang, X. (2006). Item Response Theory. In J. L. Green, G.. Camilli, & P. B. Elmore (Eds.) (2006). *Handbook of complementary methods in education research.* (pp. 385-409). Mahwah, NJ: Lawrence Erlbaum Associates.

Gullikson, H. (1987). *Theory of mental tests. Hillsdale*. NJ: Lawrence Erlbaum Associates.

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

Hambleton, R. K. (1989). Principles and applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

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

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement,* 16 (2),159-176.

OECD (2014). PISA 2012 results in focus: what 15-year-olds know and what they can do with what they know. OECD

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement,* 15, 361-373.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton and J. Zaal (eds.), *Advances in educational and psychological testing.* Boston: Kluwer Academic Publishers.

Wu, M. L., Adams, R. J., Wilson, M. R.,& Haldane, S. A. (2007). *Acer ConQuest Version 2.0 :generalised item response modelling software.* Australian Council for Educational Reesearch.

Yen, W. M., & Fitzpatrick, A.R. (2006). Item response theory. In R. T. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-154). Westport, CT: Praeger.

# 教室中的成績等化食譜

➢ 林世華 / 謝佩蓉

# 教室中的成績等化食譜

**林世華**
國立臺灣師範大學副教授

**謝佩蓉**
國家教育研究院助理研究員

## 壹、緒論

常言道，總結性評量和形成性評量兩者不可偏廢。形成性評量（formative evaluation）意指評估學生於教學過程的學習進展，可視為教學過程的一部分，讓教師能調整其教學；其目的，乃在於協助教師與學生達成學習目標（Miller, Linn, & Gronlund, 2013）。而若要能適當地評估學生的學習進展，適切地運用某些統計方法是不可或缺的。本文透過淺顯的文字，簡單地介紹Rasch模式與等化設計的重要概念，接著便以實例示範，帶領讀者一步步操作測驗成績等化，使1至12年級的教師們能按圖索驥、掌握要訣，輕鬆連結學科測驗成績，記錄學生的學習進展。

## 一、具備物理學常用的測量量尺

量尺的議題，常常被簡化為「單位」的議題，二者也常被混用。不同的量尺，在物理學上就是不同的單位，例如：公分、吋，是長度的單位，公斤、磅，則是重量的單位。因為採用的量尺不同，使得公制和英制測量結果的數值不同，然而因為關係明確，彼此之間的關係是可以換算的，進行比較是沒有問題的。此外，由於量尺確立，我們對於體重數值所呈現的意義，很快就能有概念。譬如，四年級的小孩80公斤，就是太胖啦。

我們在對學生施測時，量尺會出現問題。期中考國語文和期末考國語文都是國語文測驗，施測後都產生一個數量，似乎隱含「分」這個單位，例如：期中考卷面分數88分，期末考卷面分數87分，然而此「分」非彼「分」，「分」這個單位並不具備物理學的特性。

$$\Pr(X_{ij}=1;\theta_i,b_j)=\frac{e^{\theta_i-b_j}}{1+e^{\theta_i-b_j}} \qquad \text{（公式1）}$$

丹麥數學家Rasch於1960年提出一個數學模式（公式1），而利用這個公式所發展出來的分析模式，稱為「Rasch模式」（Andersen & Olsen, 2001）。它具有很多良好的數學特性，也因此可以應用的情境很廣，其中一項便是適合用來作為發展性的測量。

暫時不看指數的話，分子和分母分別是 $\frac{e}{1+e}$；公式操弄的重點在於指數部分，也就是希臘字 $\theta$（讀音：theta）與英文字 $b$。

$\theta$　就如同學生的分數，$b$ 是題目的難度，就好像是這題有百分之多少的人答對它。

$\theta_i$　表示第 $i$ 個學生的能力，每位學生的能力都不一樣。

$b_j$　表示第 $j$ 個題目的難度，各題有各題的難度。

$e$　是自然底數，它是一個常數，大約等於 2.71828。

當學生完成一份測驗，我們會關心兩個向度：孩子考幾分？，另一面就是這一題被多少學生答對？我們會關心學生考幾分，是因為分數的背後代表了某種能力。這一題被多少學生答對，則是想知道這些測驗題項的難易程度。傳統的做法是人的能力與題目難度，單位分開建造，人的能力是「採取百分制的幾分」，題的難度則是「答對人數百分比」。既然公式1將學生能力值與題目難度放在同一個數學式中，而且能力值如果要能減難度值，表示能力與難度的單位相同（公斤和公分不能相減），意即「人的能力」與「題目難度」是同一量尺。

經由公式1所算出來的值，是個介於0至1的數值，用來表示：某個人帶著他的某項能力（例如：數學能力），作答一份測驗之中的某一題，答對的機率有多大？如果 $\theta_i - b_j > 0$，則答對這一題的機率大於 0.5；如果 $\theta_i - b_j = 0$，表示答對這一題的機率等於 0.5。用這個方式來描寫，當某位學生的能力越來越高，他答對某一題的機率是越來越大的；若是某位學生的能力非常低，則他答對某一題的機率就非常小。如要將學生區隔高下，最有用的試題是答對機率 0.5 的情況。

要注意的是，這個公式只適用於作答結果為「對」或「錯」兩種狀態的試題（例如：選擇題、是非題、填充題），其他給分方式試題的答對機率，要用其他的公式計算。關於 Rasch 模式更詳盡的介紹，可參閱本書第三章測驗理論與測驗分析技術。

## 二、生活中的等化情境

「等化」這個名詞，聽來陌生。事實上，你可能沒想過，「等化」和我們的生活息息相關。在正式進入等化議題之前，讓我們先來看兩個例子。

「物價高漲、薪水不漲」是近年社會上的熱門議題。根據行政院主計總處（2013）的統計指出，2000年工業及服務業的平均薪資為41,861元，2013年為47,881元，增加幅度很小。然而，2000年你可以用542元買到一桶家用液化石油氣，2013年

卻得花871元才能買到一桶（經濟部能源局，2013），是1.6倍的價格。臺北市信義區的房價變動就更大了，2000年的時候公寓每坪23.2萬元，2013年要價61.08萬元（內政部地政司，2013），幾乎是三倍價。這些現象告訴我們，數字的大小，其背後意義各不相同；2000年拿100元可以買到的東西，比2013年多！這個意思是說，2000年的物品價格和2013年的物品價格價是不能直接比較的。因此，經濟學家會透過公式，算出「消費者物價指數」來比較一般消費者在兩個不同時期購買相同商品組合所需付出的成本。

第二個例子和貨幣有關。出國旅遊前，我們常到金融行庫兌換當地貨幣。究竟手中的新臺幣可以換到多少當地貨幣呢？倚賴「匯率」決定。例如，要到香港旅遊，在不考慮手續費的情況下，每3.85元新臺幣可以換到1元港幣（表1）。但若是要到澳門旅遊，澳門幣並非國際貨幣，無法在臺灣直接兌換。這時我們就可以透過港幣作為新臺幣和澳門幣的共同量尺，將換得的1元港幣再換為1.0306元澳門幣（表1）。也就是說，有了港幣這個量尺基礎，我們便可知道澳門幣比新臺幣的幣值大。這樣的轉換也可以應用在測驗分數上，如果我們知道乙校期中考80分的學生能力，相當於甲校期中考100分的學生能力；又知道丙校期中考90分的學生能力，相當於甲校期中考100分的學生能力，那我們就可以知道甲校期中考最簡單，丙校次之，乙校最難。

值得注意的是，消費者物價指數的誤差甚大，而貨幣轉換是受經濟供需的法則所支配，亦受眾多因素所影響。同樣地，我們在進行測驗分數等化之際，也得留意其中所包含的誤差和可能的影響因素。

**表1　匯率牌價範例**

| 港幣 | 新臺幣 | 澳門幣 |
|---|---|---|
| 1 | 3.85 | |
| 1 | | 1.0306 |

## 三、等化的意義、設計以及限制

等化（equating）的定義是：對同一群學生而言，一份新測驗的分數和一份舊測驗的分數，用來代表其中某位學生的相對位置時，兩者是等值的（Livingston, 2004）。其目的是，透過統計對於測驗分數進行轉換，以校準不同測驗間的難度（Kolen & Brennan, 2004）。等化並沒有辦法校準內容，純粹就「難度」這個議題做校準。等化可依其應用情境分為水平等化（horizontal equating）、垂直等化（vertical equating）以及分數連結（score linking）（van der Linden, 2000）。托福網路測驗（TOEFL iBT）讓世界各地的考生在不同時間點作答不同的測驗卷，成績仍然可以相互比較，便是水平等化的一例。Rasch當年在發展模式時，曾經檢驗某個世代學生閱讀能力的進展，該種設計便需採用垂直等化，連結學生在不同年齡所測量之相同構

念（閱讀表現）的分數（Peter, Cieza, & Geyh, 2013）。分數連結的典型例子為ACT和SAT兩種美國大學常採用的入學考試成績之間的連結，它們是兩種不同的測量工具，但彼此需要有分數之間的對應以便用於入學申請（Dorans, 1999）。

等化的設計和方法有很多種，本文介紹的是「共同題等化設計」（common-item equating），也就是新編測驗（new form）的題項之中，包含了一組參照測驗（reference form）之中的試題，而這些重複使用的試題便被稱為「共同題」。通常，我們會將第一份卷作為參照測驗，例如：期初考、期中考，並從中搬一些題目到下

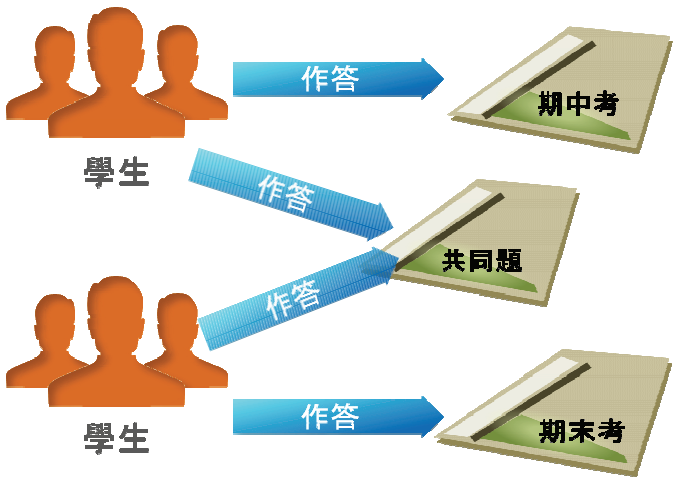



圖1　共同題等化設計示意

一次的考試，以檢視學生於教學之後的進展（圖1）。

共同題等化是國內外大型評量採用的設計，像國際數學與科學教育成就趨勢調查（Trends for International Mathematics and Science Study, TIMSS）、臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement, TASA）等，也被應用於校園之中（Baghaei & Amrahi, 2011; Meyer & Zhu, 2013; Stewart & Gibson, 2010）。共同題等化共同題等化設計的好處是，較其他等化設計更有彈性，可適用於各種情境。有的等化設計，同一位學生需要於施測日作答兩份試卷；而共同題等化設計，同一位學生只需於施測日作答一份試卷。最重要的優點是，除了共同題內容不能曝光，其餘的非共同題都可以公告，讓老師和學生能於考試結束後一同檢討試題內容，符合教室中的實務需求。

可想而知，共同題等化設計成敗的關鍵點，在於「共同題」。共同題的試題品質要好，不能太難也不要太簡單。共同題要有內容代表性，最好是該份測驗所涵蓋內容的迷你版。一般建議，一份40題左右的測驗，共同題的題數至少應該占測驗題數的20%（Kolen & Brennan, 2004），亦即40題之中至少應含8題為共同題，但可以更

多。

凡有測量必有誤差，凡有等化當然也有誤差。等化技術存在著許多限制，是我們在使用這項技術時需要格外謹記在心的。首先，等化無法精準地校正每一位受測學生的個別差異。假如A學生恰好對於參照測驗的題項格外精熟，而B生對於新編測驗的題項格外精熟，那麼對這兩位學生而言，兩份測驗的難度很可能會恰好相同。因此，等化較為適用於校正群體的測驗分數。此外，若我們以卷面分數進行等化，再轉換為量尺分數，數值常常不是整數，便衍生「進位誤差」，而此類誤差會隨著學生的卷面分數離散程度上升而增加（Livingston, 2004）。

垂直等化的誤差來源就更多了。垂直等化意欲比較學生在不同時間點的成長進展，題項難度的設計是否真能和學生成長相符，是一項很大的挑戰。再者，同一群學生於不同年級接受測驗時，所經歷的題項本質和評量程序都可能會產生變化，使得垂直等化可能混雜了題項內容與評量方式的變化，準確估計試題難度的挑戰更大（Lissitz & Huynh, 2003）。

## 貳、實例操作示範

以下透過實例操作過程，說明如何設計期中考與期末考卷，並以固定試題參數量尺化（fixed common item parameter calibration, FCIP calibration）方法，進行兩次考試分數之連結校準。

### 一、成績等化設計

某科目修課學生數為58人。期中考題數35題，題型為選擇題，期末考題數50題，亦為選擇題。每題均含4至5個選項，其中只有一個選項是正確答案。測驗後，並未將試卷發回給學生，也沒有公告試題。為連結期中考與期末考測驗分數，兩次測驗之間有7題共同題，由期中考試題之中挑選而來。鑒於共同題挑選原則「試題品質要好，不能太難也不要太簡單。內容要具代表性，最好是該份測驗所涵蓋內容的迷你版」，因而此例7題共同題分別來自期中考所涵蓋的7個章節。教師於學期初便告知學生，期末考的測驗範圍部分涵蓋期中考的範圍，讓學生能事先有所準備。

### 二、成績等化材料

1. 期中考逐題作答反應[1] 與試題正確答案
2. 期末考逐題作答反應與試題正確答案
3. 期中考與期末考共同題對照表與共同題參數檔

[1] 本文實例所使用之期中考與期末考逐題作答反應下載網址：https://www.dropbox.com/s/nnnlsdyany2bb1z/1011.zip

4. 試題分析軟體 ACER ConQuest[2]
5. 試算分析軟體 Microsoft Office Excel

## 三、成績等化做法

### （一）期中考學生能力值與試題難度分析

1. 整理期中考逐題作答反應

將每位學生期中考的逐題作答反應鍵入電腦，並整理存檔為ACER ConQuest所需的檔案格式；也就是資料與資料之間緊密相連、沒有空白。檔案內容由左至右分別是：學生座號（共四碼）、第1題至第35題的逐題作答反應（圖2）。整理好之後，

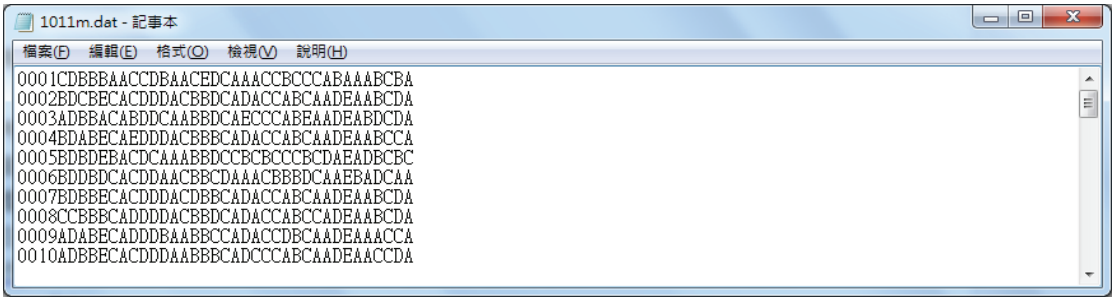
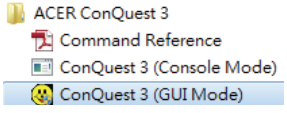



圖2　期中考資料格式實例

將檔案存檔並命名；此例，我們將之命名為「1011m.dat」。

2. 準備試題分析指令檔

於程式集中點選執行 

。點選「開新檔案」（圖3）後，於左側 Input Window 輸入 ACER ConQuest所使用的指令檔（圖4），並存檔為「D:\1011m.CQC」。

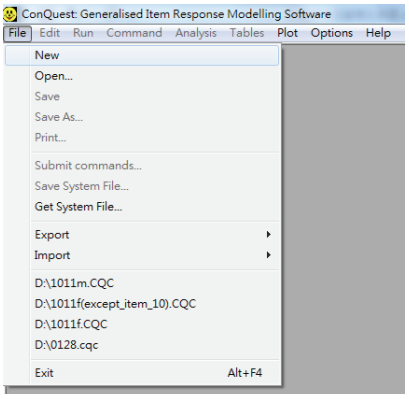



圖3　開新檔案實例

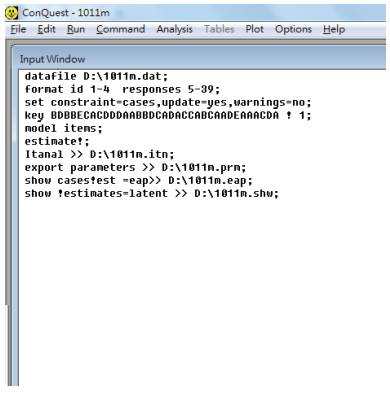



圖4　輸入指令實例

[2] ACER ConQuest 試用版和操作手冊下載網址：http://conquest-sales.acer.edu.au/index.php?cmd=collect&e=g05DzYDarjM%3D

ACER ConQuest所使用的指令檔非常容易上手，只需10列即可完成所需要的分析，逐列詳細說明如表2。

表2　ACER ConQuest 指令與說明

| 指令 | 說明 |
|---|---|
| datafile **D:\1011m.dat;** | 讀取資料檔「1011m.dat」，檔案存放在「D:\」。 |
| format id **1-4**  responses **5-39**; | 資料檔的第1格至第4格資料為學生座號，第5格至第39格為作答反應。 |
| set constraint=cases,update=yes,warnings=no; | 以人為定位，設定能力值的總和為0；估計出來能力值為0的學生，表示站在正中間。 |
| key **BDBBECACDDDAABBDCADACCABCAADEAAACDA**！1; | 依照題號順序輸入標準答案，共35個正確答案。 |
| model items; | |
| estimate!; | |
| Itanal >> **D:\1011m.itn**; | 輸出傳統的試題分析報表於「D:\」，檔名為「1011m.itn」。 |
| export parameters >> **D:\1011m.prm**; | 輸出試題難度於「D:\」，檔名為「1011m.prm」。 |
| show cases!est =eap>> **D:\1011m.eap**; | 輸出學生能力值於「D:\」，檔名為「1011m.eap」。 |
| show !estimates=latent >> **D:\1011m.shw**; | 輸出總表於「D:\」，檔名為「1011m.shw」。 |

註：Acer ConQuest所使用的指令檔非常容易上手，只需依照實例，將粗體字部分依照資料真實情況進行修改，便能應用於教室中的真實情境。

3. 執行試題分析

按下「執行全部」（圖5），程式便開始進行分析。分析結束後，便可以在「D:\」找到「1011m.eap」、「1011m.itn」、「1011m.prm」以及「1011m.shw」四個檔案。

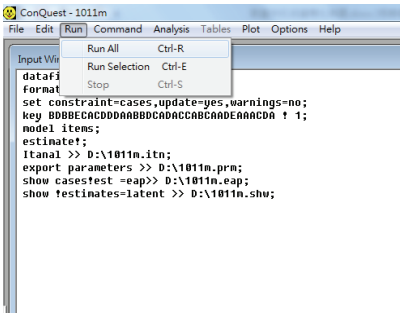



圖5　執行試題分析實例

4. 詮釋試題分析結果

（1）「1011m.itn」報表重點說明

這個檔案呈現的是傳統的試題分析結果，可用來和學校既有的試題分析報表相互印證。如果出現全部學生都答對與全部學生都答錯的題目，則是沒有功能的題目，並沒辦法將學生區分高下。

以第五題為例（圖6），作答此題的學生共有58人（Cases for this item）。此題得分與測驗總分之間的相關是0.37（Item-Total Cor.），為鑑別度的指標；當試題的鑑別度越佳，越能區隔學生的能力。此題的正確答案為E（選擇E者，Score為1，其餘為0），58位學生中有37位答對，通過率63.79%，屬於中間偏易的試題。

```
Item 5
------
item:5 (5)
Cases for this item      58   Item-Rest Cor.  0.26   Item-Total Cor.  0.37
Item Threshold(s):    -0.64   Weighted MNSQ   1.00
Item Delta(s):        -0.64
-------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1
-------------------------------------------------------------------------------
   A       0.00         6      10.34   -0.16    -1.25(.216) -0.23     0.63
   B       0.00        11      18.97   -0.16    -1.24(.221) -0.55     0.57
   C       0.00         1       1.72    0.03     0.20(.844)  0.04     0.00
   D       0.00         3       5.17   -0.07    -0.55(.584) -0.35     0.80
   E       1.00        37      63.79    0.26     2.05(.045)  0.23     0.71
===============================================================================
```

圖6　古典測驗理論試題分析實例

（2）「1011m.prm」報表重點說明

這個檔案呈現的是以Rasch模式進行分析後，所得到的試題難度，也就是該題答錯（得0分）跨到答對（得1分）的閾值。就定義而言，試題難度指的是擁有50%機率可以答對該題的學生能力值（Verhelst, 2004）。如果學生的能力值呈現常態分配，那麼68%學生能力值介於-1至+1之間，95%學生能力值介於-2至+2之間；試題難度達到2以上為高難度的題目，試題難度落於-2以下為極簡單的題目。

同樣以第五題為例（圖7），試題難度為-0.63852，表示能力值為-0.63852的學生，有50%的機率可以答對此題。由於我們已經在指令界定「估計出來能力值為0的學生，表示站在正中間」，能力值為-0.63852的學生能力值未達整體學生能力值一半，反應出此題屬於中間偏易的試題，和此題通過率63.79%相互呼應。再看第10題，試題難度為-4.32683，表示能力值為-4.32683的學生即有50%的機率可以答對此題，因而此題顯得非常非常容易。

```
 1  -0.89911   /* item 1 */
 2  -1.50364   /* item 2 */
 3  -0.55523   /* item 3 */
 4  -3.60308   /* item 4 */
 5  -0.63852   /* item 5 */
 6  -0.99060   /* item 6 */
 7  -2.59618   /* item 7 */
 8  -0.31258   /* item 8 */
 9  -2.59618   /* item 9 */
10  -4.32638   /* item 10 */
```

圖7　Rasch模式分析所得試題難度實例

（3）「1011m.eap」報表重點說明

這個檔案呈現的是以Rasch模式進行分析後，所得到的學生能力期望值。第一欄為識別碼，依照期中考資料順序排列，識別碼1代表學生座號0001、識別碼2代表學生座號0002、識別碼3代表學生座號0003，可供辨識出是哪位學生的資料。

第二欄即為學生能力值，如果學生的能力值呈現常態分配，那麼68%學生能力值介於-1至+1之間，95%學生能力值介於-2至+2之間。第三欄為學生能力值變異數（Wu, Adams, Wilson, & Haldane, 2007）。圖8所呈現的10位學生之中，識別碼2（座號0002）能力值0.69792最高、識別碼5（座號0005）能力值-1.54555最低。

```
 1    -1.17822    0.35749    0.78445
 2     0.69792    0.46151    0.64076
 3    -0.34681    0.38430    0.75090
 4     0.12747    0.41385    0.71113
 5    -1.54555    0.34141    0.80340
 6    -1.17822    0.35749    0.78445
 7     0.49339    0.44317    0.66875
 8    -0.03889    0.40210    0.72730
 9    -0.03889    0.40210    0.72730
10     0.49339    0.44317    0.66875
```

圖8　Rasch模式分析所得學生能力值實例

（4）「1011m.shw」報表重點說明

這個檔案呈現的是以Rasch模式進行分析後，所得到的總表，內容包含數個部分。首先，呈現的是試題的適配指數MNSQ及其95%信賴區間，用來檢視試題是否符合Rasch模式的前提假設。理論上，MNSQ數值的虛無假設等於1，若分析所得之MNSQ數值超出其95%信賴區間，表示該題並不符合Rasch模式「鑑別度 = 1」的前提假設，此時，該試題之T值的絕對值也會大於 2（Wu, Adams, Wilson, & Haldane, 2007）。當原始 T 值大於2，表示和理論模式相較，試題之鑑別度較差；當原始 T 值

小於 -2，表示和理論模式相較，試題之鑑別度更高。

此外，「未加權MNSQ」（UNWEIGHTED FIT）較容易受到極端值的影響，Bond與Fox（2007）建議使用者，首要以「加權MNSQ」（WEIGHTED FIT）作為判斷試題是否適配的指標。

同樣以第五題為例（圖9），試題難度為-0.639（ESTIMATE），加權MNSQ為1.00並未超出其95%信賴區間，T 值的絕對值為0，表示此題符合Rasch模式「鑑別度=1」的前提假設。再以第八題為例（圖9），試題難度為-0.313（ESTIMATE），加權MNSQ為1.21略超出其95%信賴區間，而T值的絕對值為2；顯示此題鑑別度較低，但還不致於非常差，在沒有更合適試題的情況下，可以考慮保留。在試題適配度皆為可接受的情況下，亦顯示此份測驗符合Rasch模式單向度（unidimensionality）的前提假設，可視為測量同一特質「統計學能力」。

| VARIABLES | | | | UNWEIGHTED FIT | | | WEIGHTED FIT | | |
|---|---|---|---|---|---|---|---|---|---|
| | item | ESTIMATE | ERROR^ | MNSQ | CI | T | MNSQ | CI | T |
| 1 | 1 | -0.899 | 0.318 | 1.22 | ( 0.64, 1.36) | 1.1 | 1.18 | ( 0.74, 1.26) | 1.3 |
| 2 | 2 | -1.504 | 0.356 | 1.23 | ( 0.64, 1.36) | 1.2 | 1.16 | ( 0.63, 1.37) | 0.9 |
| 3 | 3 | -0.555 | 0.305 | 1.12 | ( 0.64, 1.36) | 0.7 | 1.13 | ( 0.79, 1.21) | 1.1 |
| 4 | 4 | -3.603 | 0.737 | 0.30 | ( 0.64, 1.36) | -5.2 | 0.84 | ( 0.00, 2.23) | -0.1 |
| 5 | 5 | -0.639 | 0.308 | 0.99 | ( 0.64, 1.36) | 0.0 | 1.00 | ( 0.78, 1.22) | -0.0 |
| 6 | 6 | -0.991 | 0.322 | 0.87 | ( 0.64, 1.36) | -0.7 | 0.91 | ( 0.73, 1.27) | -0.7 |
| 7 | 7 | -2.596 | 0.493 | 1.38 | ( 0.64, 1.36) | 1.9 | 1.11 | ( 0.31, 1.69) | 0.4 |
| 8 | 8 | -0.313 | 0.300 | 1.24 | ( 0.64, 1.36) | 1.3 | 1.21 | ( 0.81, 1.19) | 2.0 |
| 9 | 9 | -2.596 | 0.493 | 0.52 | ( 0.64, 1.36) | -3.1 | 0.86 | ( 0.31, 1.69) | -0.3 |
| 10 | 10 | -4.326 | 1.021 | 0.30 | ( 0.64, 1.36) | -5.3 | 0.89 | ( 0.00, 2.82) | 0.2 |

圖9　Rasch模式分析所得試題適配指標實例

其次，解讀「試題與受試者關係圖」（圖10）所呈現的訊息。垂直的虛線將圖一分為二，虛線左側「×××」圖示，表示受試者；虛線右側「數字」表示試題題號，最左側「下至 -3 上至 +2」表示能力值與試題共用的刻度。傳統的試題分析結果是以百分比呈現，學生與試題有各自的計算基準，兩者無法畫在同一張圖；Rasch 模式使得學生與試題使用相同量尺，便能發揮優勢，以一張圖呈現兩者之間的相互關係。

整體而言，受試者能力值分佈介於 -2 至 +2 之間，試題難度分佈介於 -3 至 0 之間，唯獨第 32 題難度最高，介於 1 至 2 之間；顯示這 35 題試題相對於受試者而言，乃如指諸掌。

細部觀察可知，對於能力值介於 -2 至 0 的受試者而言，尚有足夠的試題可以區辨其能力；反觀能力值介於 0 至 2 的受試者而言，幾乎沒有試題能夠區辨他們。此

外，為數不少的試題難度介於 -3 至 -2 之間，卻沒有能力值相對應的受試者。顯示期中考的命題為中間偏易，大多數學生均駕輕就熟、輕鬆作答。

另一方面，這個圖隱含了「適性」概念在其中，意即困難的題目應該給能力比較高的學生作答、簡單的題目給能力比較低的學生，這樣對兩者都有挑戰性。試想，若是簡單的題目給能力比較高的學生、而困難的題目給能力比較低的學生，那麼前者輕鬆答完、後者無法理解題意，亦非評量所欲達成之目的。

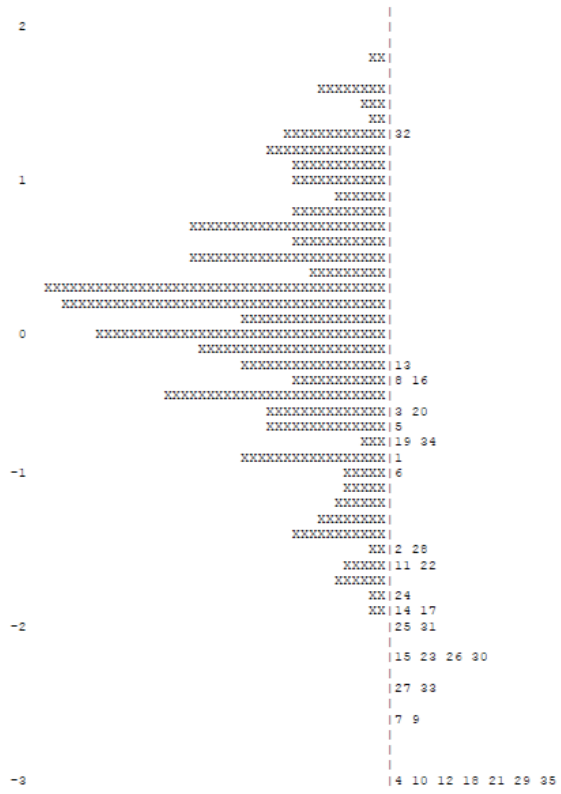



圖10　試題與受試者關係圖實例

## （二）期末考學生能力值與試題難度分析

1. 整理期末考逐題作答反應

和期中考資料檔作法相同，將每位學生期末考的逐題作答反應讀進電腦，並整理存檔為ACER ConQuest所需的檔案格式。檔案內容由左至右分別是：學生座號（共四碼）、第1題至第50題的逐題作答反應。整理好之後，將檔案存檔並命名；此例，我們將之命名為「1011f.dat」。

2. 備妥共同題對照表與共同題參數檔

為了將期末考試題難度和期中考試題難度置於相同量尺，採用固定試題參數量尺化（fixed common item parameter, FCIP）方法，校準期末考試題參數；也就是估計期末考試題難度時，匯入已知之共同題參數而不再估計共同題難度。

透過共同題對照表（表2）得知，期中考的第1題為期末考的第1題、期中考的第7題為期末考的第3題、期中考的第14題為期末考的第6題，以此類推。考量共同題品質對於期末考試題難度與學生能力值估計影響甚鉅，在分析期末考資料之前，再次檢查「1011m.shw」中，共同題的MNSQ值，確認它們皆符合Rasch模式前提假設，才進行下一個步驟。

表3　共同題對照表實例

| 期中考題號 | 期末考題號 |
|---|---|
| 1 | 1 |
| 7 | 3 |
| 14 | 6 |
| 19 | 9 |
| 24 | 12 |
| 30 | 15 |
| 34 | 18 |

開啟「1011m.prm」，刪除非共同題，僅保留第1、7、14、19、24、30、34題（圖11），並將最左方「題號」變更為1、3、6、9、12、15、18（圖12），試題難度不變，另存新檔為「D:\1011a.prm」即完成。

```
 1  -0.89911   /* item 1 */
 7  -2.59618   /* item 7 */
14  -1.88579   /* item 14 */
19  -0.81016   /* item 19 */
24  -1.74923   /* item 24 */
30  -2.19878   /* item 30 */
34  -0.81016   /* item 34 */
```

圖11　共同題於期中考之題號與參數

```
 1  -0.89911
 3  -2.59618
 6  -1.88579
 9  -0.81016
12  -1.74923
15  -2.19878
18  -0.81016
```

圖12　共同題於期末考之題號與參數

3. 準備試題分析指令檔

於程式集中點選執行ConQuest 3，開啟「D:\1011m.CQC」（圖13）小幅度修改指令，也就是小部分修改期中考試題分析指令檔，即可完成期末考試題分析，並達成測驗成績等化之目的。修改完成後，另存新檔為「D:\1011f.CQC」。

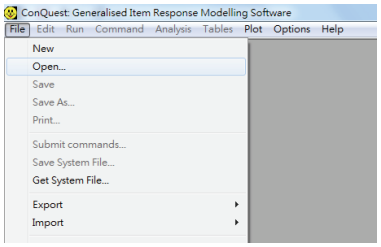



圖13　開啟舊檔實例圖

測驗分數等化所使用的指令檔非常容易上手，只需11列即可完成所需要的分析，逐列詳細說明如表。和表1的主要差異在於「以共同題為定位」和「匯入共同試題參數」兩列指令。

表4　ACER ConQuest測驗分數等化指令與說明

| 指令 | 說明 |
|---|---|
| datafile **D:\1011f.dat**; | 讀取資料檔「1011f.dat」，檔案存放在「D:\」。 |
| format id **1-4**  responses **5-54**; | 資料檔的第1格至第4格資料為學生座號，第5格至第54格為作答反應。 |
| set constraint=**none**,update=yes,warnings=no; | 以共同題為定位，需將set constraint=改為「none」。 |
| key **BAADABAADCBBCAADBDDCBABABBEBCDCBDDCDBDBDACAAAACBCB!** l; | 依照題號順序輸入標準答案，共50個正確答案。 |
| model items; | |
| import anchor_parameters << **D:\1011a.prm**; | 新增一列指令，匯入共同題試題參數。 |
| estimate!; | |
| Itanal >> **D:\1011f.itn**; | 輸出傳統的試題分析報表於「D:\」，檔名為「1011f.itn」。 |
| export parameters >> **D:\1011f.prm**; | 輸出試題難度於「D:\」，檔名為「1011f.prm」。 |
| show cases!est =eap>> **D:\1011f.eap**; | 輸出學生能力值於「D:\」，檔名為「1011f.eap」。 |
| show !estimates=latent >> **D:\1011f.shw**; | 輸出總表於「D:\」，檔名為「1011f.shw」。 |

註：Acer ConQuest所使用的指令檔非常容易上手，只需依照實例，將粗體字部分依照資料真實情況進行修改，便能應用於教室中的真實情境。

4. 執行試題分析

按下「執行全部」，程式便開始進行分析。分析結束後，便可以在「D:\」找到「1011f.eap」、「1011f.itn」、「1011f.prm」以及「1011f.shw」四個檔案。

5. 詮釋試題分析結果

（1）「1011f.itn」、「1011f.prm」以及「1011f.eap」報表

「1011f.itn」、「1011f.prm」以及「1011f.eap」三個報表的詮釋方式，和期中考試題分析產生的輸出檔相同，不再贅述。

（2）「1011f.shw」報表重點說明

試題適配指標的部分，和「1011f.shw」報表略有出入：7題共同題的試題難度旁邊，出現「*」註記，並且沒有顯示標準誤！（圖14）乃因這個例子採用FCIP進行試題參數校準，7題共同題的試題難度已經被固定不再估計，也就不會有估計標準誤。謹慎起見，再次核對7題共同題的試題難度，確認和「1011a.prm」之中的數值相同，顯示程式讀取資料正確。

| VARIABLES | | | UNWEIGHTED FIT | | | WEIGHTED FIT | | |
|---|---|---|---|---|---|---|---|---|
| item | ESTIMATE | ERROR^ | MNSQ | CI | T | MNSQ | CI | T |
| 1 1 | -0.899* | | 0.60 | ( 0.64, 1.36) | -2.5 | 0.68 | ( 0.74, 1.26) | -2.8 |
| 2 2 | -3.262 | 0.622 | 0.87 | ( 0.64, 1.36) | -0.7 | 0.98 | ( 0.02, 1.98) | 0.1 |
| 3 3 | -2.596* | | 1.67 | ( 0.64, 1.36) | 3.1 | 1.32 | ( 0.32, 1.68) | 1.0 |
| 4 4 | -3.704 | 0.745 | 1.18 | ( 0.64, 1.36) | 1.0 | 1.04 | ( 0.00, 2.25) | 0.3 |
| 5 5 | -0.259 | 0.313 | 1.32 | ( 0.64, 1.36) | 1.6 | 1.24 | ( 0.81, 1.19) | 2.3 |
| 6 6 | -1.886* | | 0.98 | ( 0.64, 1.36) | -0.0 | 1.10 | ( 0.54, 1.46) | 0.5 |
| 7 7 | -1.034 | 0.335 | 1.08 | ( 0.64, 1.36) | 0.5 | 1.03 | ( 0.72, 1.28) | 0.3 |
| 8 8 | -1.034 | 0.335 | 0.89 | ( 0.64, 1.36) | -0.5 | 0.93 | ( 0.72, 1.28) | -0.5 |
| 9 9 | -0.810* | | 0.90 | ( 0.64, 1.36) | -0.5 | 0.93 | ( 0.75, 1.25) | -0.5 |
| 10 10 | -4.434 | 1.028 | 0.25 | ( 0.64, 1.36) | -5.9 | 0.92 | ( 0.00, 2.85) | 0.2 |
| 11 11 | 0.466 | 0.317 | 1.08 | ( 0.64, 1.36) | 0.5 | 1.08 | ( 0.80, 1.20) | 0.8 |
| 12 12 | -1.749* | | 1.41 | ( 0.64, 1.36) | 2.0 | 1.40 | ( 0.58, 1.42) | 1.7 |
| 13 13 | -2.939 | 0.551 | 0.85 | ( 0.64, 1.36) | -0.8 | 1.05 | ( 0.18, 1.82) | 0.3 |
| 14 14 | -2.939 | 0.551 | 0.78 | ( 0.64, 1.36) | -1.2 | 0.97 | ( 0.18, 1.82) | 0.1 |
| 15 15 | -2.199* | | 0.96 | ( 0.64, 1.36) | -0.1 | 0.88 | ( 0.46, 1.54) | -0.4 |
| 16 16 | -2.463 | 0.468 | 0.61 | ( 0.64, 1.36) | -2.4 | 0.81 | ( 0.37, 1.63) | -0.5 |
| 17 17 | 0.302 | 0.314 | 1.02 | ( 0.64, 1.36) | 0.2 | 1.05 | ( 0.81, 1.19) | 0.6 |
| 18 18 | -0.810* | | 1.11 | ( 0.64, 1.36) | 0.6 | 1.16 | ( 0.75, 1.25) | 1.3 |
| 19 19 | -2.274 | 0.442 | 1.01 | ( 0.64, 1.36) | 0.1 | 1.06 | ( 0.43, 1.57) | 0.3 |
| 20 20 | -2.680 | 0.503 | 0.69 | ( 0.64, 1.36) | -1.8 | 0.89 | ( 0.29, 1.71) | -0.2 |

圖14　FCIP所得試題適配指標實例

至於「試題與受試者關係圖」的詮釋方式，亦和期中考試題分析產生的輸出檔相同，不再贅述。

（三）期中考與期末考學生能力值差異分析

1. 彙整期中考與期末考能力值

分別開啟「1011m.eap」和「1011f.eap」，將第二欄學生能力值複製貼上至Microsoft Excel（圖15），可以看出，和期中考相較，座號0001、0004、0006、0008以及0010五位學生，期末考能力值是增加的；而且可以確定的是，期中考的能力值與期末考的能力值是建造在同一量尺，兩者可以相互比較。

| | A | B | C |
|---|---|---|---|
| 1 | 座號 | 期中考能力值 | 期末考能力值 |
| 2 | 0001 | -1.17822 | -1.10513 |
| 3 | 0002 | 0.69792 | 0.65959 |
| 4 | 0003 | -0.34681 | -0.91384 |
| 5 | 0004 | 0.12747 | 0.2475 |
| 6 | 0005 | -1.54555 | -1.83439 |
| 7 | 0006 | -1.17822 | -0.81529 |
| 8 | 0007 | 0.49339 | 0.12379 |
| 9 | 0008 | -0.03889 | 0.2475 |
| 10 | 0009 | -0.03889 | -0.11013 |
| 11 | 0010 | 0.49339 | 0.99052 |

圖15　期中考與期末考能力值彙整實例

2.全班學生期中考與期末考能力值差異分析

除了可以看出個別學生能力值的發展，也可以透過「成對母體平均數差異檢定」（paired-samples T Test）分析全班學生能力值的發展。

首先，點選Microsoft Excel「資料分析」功能（圖16）。

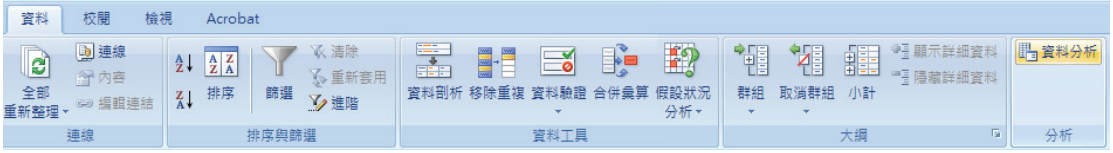



圖16　Microsoft Excel資料模組介面

接著，選擇「t檢定：成對母體平均數差異檢定」，按「確定」（圖17）。

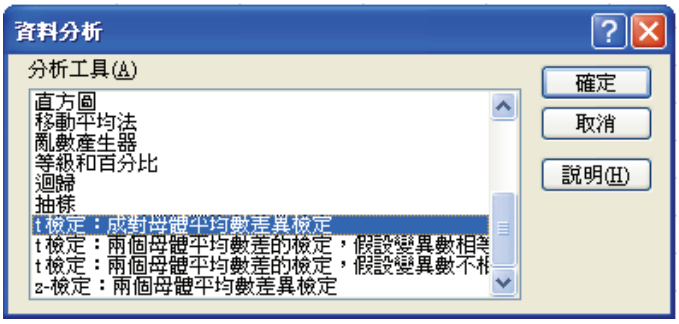



圖17　Microsoft Excel資料分析模組介面

利用滑鼠選取期中考能力值資料範圍，讓Excel讀入「變數1的範圍」，本例為「$B$2:$B$59」；再利用滑鼠選取期末考能力值資料範圍，讓Excel讀入「變數2的範圍」，本例為「$C$2:$C$59」；最後，按「確定」（圖18）。


圖18　Microsoft Excel成對母體平均數差異檢定介面

從分析結果報表可知（圖19），期中考全班學生能力值平均數為-0.0008、期末考全班學生能力值平均數為-0.02548，似乎微幅退步？然而，再進一步檢視「P(T<=t)雙尾」為0.739071，大於社會科學常用的判斷標準「0.05」，表示期中考全班學生能力值平均數與期末考全班學生能力值平均數差異，並未達到統計上顯著水準，也就是和期中考相較，期末考全班學生能力值平均數，沒有進步也沒有退步，維持平盤。由於一般國小學生數學基本能力的年進步量0.5至0.7（任宗浩、譚克平、張立民，2011），大學一年級學生統計能力的年進步量約為0.04（Lin & Hsieh, 2013），顯示學習年段越高，一年能進步的能力值越少。合理推論，此例期中考和期末考僅相距二個月，致使能力值的改變不明顯。

t 檢定：成對母體平均數差異檢定

| | 變數 1 | 變數 2 |
|---|---|---|
| 平均數 | -0.0008 | -0.02548 |
| 變異數 | 0.429169 | 0.63463 |
| 觀察值個數 | 58 | 58 |
| 皮耳森相關係數 | 0.717059 | |
| 假設的均數差 | 0 | |
| 自由度 | 57 | |
| t 統計 | 0.334712 | |
| P(T<=t) 單尾 | 0.369535 | |
| 臨界值：單尾 | 1.672029 | |
| P(T<=t) 雙尾 | 0.739071 | |
| 臨界值：雙尾 | 2.002465 | |

圖19　Microsoft Excel成對母體平均數差異檢定實例

## 參、結語與展望

測驗成績等化，當然不是只侷限於選擇題型，是非題、填充題、甚至問答題，也都可以進行遵循這一套操作程序，略加修改執行指令即可達成。教師們一旦學會操作成績等化，教務處便可以建立新版校內成績冊。未來，期中考試過後，學生會有兩欄成績，第一欄是原有的卷面分數，就是滿分一百分而學生得幾分。第二欄則是將作答反應轉化成能力值，期中考呈現的是直接估計所得的能力值，期末考呈現的則是和期中考連結等化後的能力值，供老師判斷學生學習的進展和演變。

等化技術的應用層面是很廣的，不但能了解學生於期中考至期末考的能力變化，也可以了解補救教學、差異化教學等各式教學方案實施前後，學生能力的變化。如果設計得宜，甚至可以用於了解學生從1年級升至12年級的能力變化。要注意的是，校準後的能力值仍有其使用上的限制，不適合作為繁星計畫等升學方案之參考指標。

本文所演示之實例，連結等化的重要設計在於「測驗後，並未將試卷發回給學生，也沒有公告試題」，學生不會因為期中考試後反覆練習舊題目，而使得期末考時表現得更好。然而，這個做法卻不符合現今1至12年級教學實務所需。

建議未來由各學科中心主導建構「各學習單元的共同題題庫」，確保試題品質優良，沒有試題參數漂移（item parameter drift, IPD）的現象（IPD意指共同題的參數因為時間的不同而產生變化致使測驗分數效度受到威脅）（Goldstein, 1983），且內容亦具備該單元的代表性。這樣不但能於試前獲悉共同題參數並確保試題功能無

虞，也能於試後回收共同題，避免學生背誦共同試題干擾評量結果。意即期中考與期末考的試題，一部份來自學校教師，一部份來自學科中心。學校教師命題的部分仍維持傳統做法，試後發回並檢討考卷；但共同題的部分則收回，不公開試題。

期中考之前，各校教師自共同題題庫中，挑選合適試題融入其中；期末考的時候，共同題仍然融入題目中再測驗一次。教師需要及早準備學生，告知期末考的測驗範圍將會有部分涵蓋期中考的範圍，讓他們在心理上和時間上，都有足夠的機會能事先準備。

在更長遠的未來，可由教育部建置「雲端等化系統」，教師們僅需依照格式匯入學生作答反應，並輸入共同題題號與難度參數，系統即能自動完成學生能力值估計並輸出報表，供教師與學生們了解學習趨勢，是不是相當便利又有效率呢！

## 參考文獻


內政部地政司（2013）。**房地產交易價格**。檢索自http://www.land.moi.gov.tw/chhtml/new_quehl.asp

任宗浩、譚克平、張立民（2011）。二階段分層叢集抽樣的設計效應估計。**教育科學研究期刊，56**（1），33-65。

行政院主計總處（2013）。**受僱員工薪資調查統計**。檢索自http://www.dgbas.gov.tw/ct.asp?xItem=1135&ctNode=3253&mp=1

經濟部能源局（2013）。**家用液化石油氣大母體區平均價格年報表（零售價）**。檢索自http://web3.moeaboe.gov.tw/oil102/

Andersen, E. B., & Olsen, L. W. (2001). The life of georg rasch as a mathematician and as a statistician. In A. Boomsma, M. A. J. van. Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 3-24). New York, NY: Springer-Verlag.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*, 192-211.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores (College Board Rep. No. 99-1)*. New York, NY: College Entrance Board.

Goldstein, (1983). Measuring changes in educational attainment over time: Problems and

possibilities. *Journal of Educational Measurement, 20*(4), 369-377.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and practices (2nd ed.). New York, NY: Springer-Verlag.

Lin, S.-H., & Hsieh, P.-J. (2013, August). *Longitudinal study of undergraduate students' use of learning strategies in introductory statistics.* Paper presented at the Pacific Rim Objective Measurement Symposium 2013, Kaohsiung, Taiwan.

Lissitz, R.W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability *Practical Assessment, Research & Evaluation, 8*(10). Retrived from http://pareonline.net/getvn.asp?v=8&n=10

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Service.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research and Practice Assessment, 8,* 26-39.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2004). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson.

Peter, C., Cieza, A., & Geyh, S. (2013). Rasch analysis of the general self-efficacy scale (GSES) in spinal cord injury (SCI). *Journal of Health Psychology.* Advance online publication. doi:10.1177/1359105313475897

Stewart, J. & Gibson, A. (2010). Equating classroom pre and post tests under item response theory. *JALT Testing & Evaluation SIG Newsletter, 14*(2), 11-18.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65*(4), 437-456.

Verhelst, N. D. (2004). *Reference supplement to the preliminary pilot version of the manual for "relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment" sectiong: Item response theory.* Strasbourg, France: Council of Europe.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software.* Victoria, Australia: ACER.

# 活化測驗方式的另一個選擇—實作評量

➢ 謝名娟

# 活化測驗方式的另一個選擇—實作評量

**謝名娟**
國家教育研究院副研究員

## 壹、前言

十二年國教即將上路，教育部長蔣偉寧曾指出-十二年國教的重點不在考試，而在學生是否能主動學習。因此，老師除了發展自己的特色課程、特色教學之外，也應思考如何使用有別於傳統的紙筆考試，以刺激學生主動學習的能力。尤其是重視操作能力的世代，有些能力是紙筆測驗較不容易進行施測的。例如，報導（曾薏蘋，2012）指出，透過高考三級進用的電機工程人員，去鄉公所任職時，完全不知道應該如何修檢發電機；也有醫事技術類別的公務員被派去醫院進行業務督導，卻因為沒有證照，被當地醫院認為是外行領導內行。因此，目前的公務人員考試制度偏重於紙筆考試，而非實作的能力，即使有些考生很會寫題目，但不見得會動手操作。在現今的教育現場，考甚麼人們就重視甚麼，由於我們的考試並不重視實作的能力，只重視傳統紙筆測驗的能力，因此電機工程人員不會修電機，也是意料之中的事。

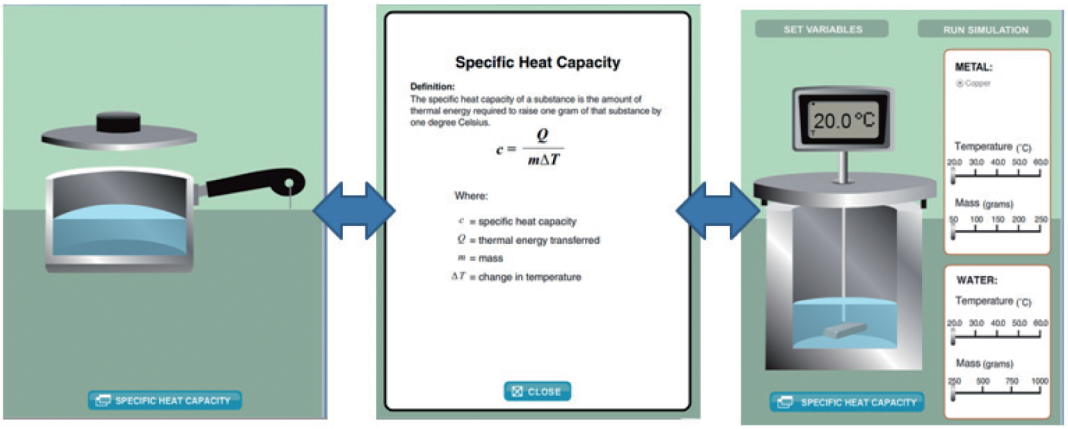
在1980年代初期，實作評量被視為具有價值的教育改革方式 (Linn, 1993; Resnick & Resnick, 1992; Wiggins, 1989)，而其被重視的主要原因則由於現行考試著重在受試者的高層次思考和問題解決的能力，且希望所學得知識技能可應用在現實生活中。例如，美國在國家教育進展評量中 (National Assessment of Educational Progress，簡稱NAEP)，將其評量重點擺在評估學生所學習的知識，是否能運用在日常生活中，而其高層次技能的評量，藉由開放式的問題，允許學生使用不同的策略來回答，甚至透過電腦模擬互動，讓學生能連結不同的知識與能力。例如在科學的測驗中，相當著重學生動手做實驗的能力，受限於空間經費，很難有真實的器材供給每個學生進行實驗。因此會用模擬的情境來測試學生操做實驗的能力，如圖1的範例所示，要請學生回答溫度、水分子與銅原子變化的一些問題。在操作過程中，須設定銅塊的質量（圖中銅塊的大小隨著質量的改變而增大或縮小）、銅塊的溫度、水的質量（水的質量改變，圖中的水位隨之上升或下降）、設定水的溫度（水的溫度改變，圖中溫度計顯示數值同時配合改變）、模擬實驗儀器的數據輸出說明，來執行並找出實驗數據。而後，透過相關問題，如：

*根據你模擬過程得到的數據，銅塊和水的比熱何者較大？*

*A. 銅塊*

*B. 水**

*解釋你選擇的選項，並針對你模擬得到的數據作為解釋的依據。*




這個任務中，你將研究製作平底鍋時，鍋底部分最適合的金屬材料爲何。爲了充分加熱食物，烹調過程中鍋底部分將加熱至高溫。最適合的金屬材料是外界輸入熱能時，金屬可到達的溫度較高。

比熱(Specific heat capacity)爲挑選鍋底金屬材質時考慮的特性之一。這個任務中，你將研究二種可能用於製造鍋底的金屬材料的比熱。

任何時間點選 SPECIFIC HEAT CAPACITY 可顯示比熱的定義及方程式

圖爲熱量計模擬實驗。二物質接觸時的熱能轉換可用熱量計量測。熱量計經過特殊設計以避免系統內部與外在環境間的熱能轉換。

將不同溫度的放入熱量計內部時，可以嘗試操作模擬的熱量計，研究熱量計中水的溫度變化量。

圖 1　NAEP 科學測驗範例試題

實作評量將學生從被動的、被給予的角色，轉換成主動的、積極的學習者。傳統的選擇題試題具有標準答案，選到正確的答案才能得分，選到不正確的答案則失分，而在實作評量中，許多答案都可以是正確的，完全是看學習者要如何建構自己的答案。實作評量鼓勵學生積極的展現成就，進行高層次思考與問題解決技巧；並運用他們的問題解決能力。學生在進行實作評量所需要的時間會依據問題情境有所不同，可能從幾分鐘到幾天甚至更久的時間 (Aschbacher, 1991; Baron, 1991; Herman, Aschbacher, & Winters, 1992; Madaus & O'Dwyer, 1999; Stiggins, 1987)，而進行的方式，除了個人完成之外，也可以透過小組合作。

實作評量在台灣教育領域中已逐漸受到重視，然而，真正應用在教學中的現場老師還是相當有限，其主要原因並非設計上的困難，而是對於要如何進行實作評量

的步驟與準則不夠熟悉，本文的目的，則希望能提供一份關於實作評量的簡單食譜，讀者只要能依樣畫葫蘆，照著食譜的步驟依序進行，即使菜餚不甚美味，但至少能呈盤上桌，而未來只要透過不斷的練習與修正，則可以設計一份不錯的實作評量教案。

本文包括幾個部份，首先簡述實作評量的內涵與定義、實施步驟、評量規準等，而後則依據測驗理論，提出應該如何評估實作評量的信度與效度，當然，對於現場老師而言，信效度的評估可能過於困難，然而以測驗評量領域而言，信效度的評估就像是做菜要試味道，不是只把菜做完就結束了，最重要的透過顧客的品嘗，給與菜餚的回饋，以做為改進的依據。信度的評估可看出評分者的一致性，並顯示評分規準的潛在問題，效度評估則是評估測驗內容和評量目的之一致性，本文提出簡單的信效度評估方法，供讀者參考。最後，提出一個實際教案來展示應如何進行實作評量的完整過程。

## 貳、文獻探討

教育及心理測驗的標準（全美教育研究協會、美國心理協會及教育研究協會[AERA, APA, & NCME], 1999）宣稱實作評量的測驗情境和現實生活中是很類似的，透過學生在這些從特定領域表現中取出的樣本，用以解釋了這領域典型或預期表現。 實作評量的主要特徵，就是表現能直接被觀察到，而其表現要能將目標及分數解釋進行連結。例如，在藝術理論課程中，實作評量可以是要求學生寫一篇關於繪畫理論的論文，然而，若是在一個繪畫教室，要來評估學生的繪畫能力，則不可能要求寫一篇論文，而是應該要學生真正的動筆，來畫出一幅畫來。

Larkin, McDermott, Simon, & Simon (1980) 指出實作評量應與測驗目的緊密連繫。測驗目的可以透過反思教學目標，與學生最後應該表現出來的能力做緊密連結。舉例來說，寫作過程在寫作中是一個相當重要的部分；因此，寫作實作評量最好應包含寫作過程的各方面因素，如何構思、如何校對與編輯修改的草稿等，另一個例子，若是數學的實作評量，則內容可包含問題解決、理由及證明，在解答中，也應指出需要落實和調整各種不同的策略以解決問題、發展和評估數學證據和論點，組織、摘要和傳達他們數學思考過程。所有應具備的表現能力應該都定義清楚，並將具體行為羅列出來。

### 一、實作評量的任務發展

在1990年代初期，研究者提出對於設計實作評量任務的一般準則至今仍適用(Baron, 1991; Herman et al 1992; Linn et al., 1991)。Baron (1991) 主張表現任務設計的內容必須很豐富，學生進行評量的過程中，老師必須引導學生參與教學活動，且使他們深入的了解，並在學年度中就讓學生清楚知道評量的準則，使他們能有機會對

自己的學習成果進行自我檢視。在討論有效的任務需具備哪些特性中，Baron認為表現任務包含以下幾個面向：需要學生替自己的問題想出對策、允許多樣化的策略和解決方法、使學生能夠使用他們先備知識及互助合作。再者一些任務可能需要持續工作好幾個星期甚至好幾個月，讓學生可控制如何解決問題及調查，也需要學生設計和與他們的調查連結，且需要自評及自我檢視（Baron, 1991）。最後，此任務必須對學生是有意義的，有挑戰性的，並且在真實社會背景下是切合的，使學生能將他們的理解和技巧轉換到相關的任務上。

Marzano et al.（1993）認為，實作評量是讓學生透過在不同的情境下，來完成某些工作任務，以展現學生對知識的理解、技巧的運用及思維的習慣。實作評量一般包括動手量度、圖形製作等。實作評量具有一個特定的模式。這個模式的第一步是確認內容的標準，標準可能是建構在教學目標或是評量架構中，希望學生達成具體、且可測量的評量目的。舉例來說在第一步驟中，想讓學生討論一個關於西安事變的觀點：「張學良是否應該脅持蔣介石以達到與日本宣戰的目的」以這個內容而言，許多議題可被拿來討論，學生可以討論並決定，「到底是不是應該與日本宣戰?或是有沒有其他不用脅持蔣介石，就可以讓蔣介石對日本宣戰的替代方案」。

第二步是建構相關的知識，以完成任務，在第二個步驟，內容變得更明確，任務是找尋關於此議題的相關內容，包括去圖書館找文獻、上網找資料，或是訪談耆老等。

第三步則建構前兩步驟的資訊內容。這個過程可能會發展出以下任務：你是張學良的幕僚，九一八事變後，日軍慘無人道屠殺東北人民。然而，由於蔣介石為當今時局最高領導人，面對共產黨日益壯大，蔣介石決定要與日本保持和平，提出先安內、後抗日的政策，因此不願出兵與日本對抗。張學良已多次和蔣介石建議應與日宣戰，然而均無法達成目的。你是張學良的幕僚，你必須權衡是否應該要發動西安事變，迫使蔣介石與日軍宣戰。你必須解釋你的決定，並解釋是否有其他方式，無須脅持蔣介石，就可以讓蔣介石對日本宣戰的替代方案。

第四步驟為呈現成果並使用標準進行評鑑，對於學生可以使用不同的方式來呈現成果，舉例來說，學生以兩種以上的不同方式發表他們的發現，例如根據當情境，進行歷史文獻蒐集，並撰寫一篇報告等。作者也提醒為了使評量能夠被我們掌握，在一個特定的任務中，不要使用超過三到四個評量指標（亦即你會從那些角度去看學生的作品），各評量指標應寫下如何評分，例如用三點量尺評定，「3-優異」、「2-普通」和「1-要加油」，以描述作品的特徵或展現的品質。

除了使用單一實作評量，來評鑑學生的學科能力之外，也可以將數個實作評量串起來評鑑，Shavelson & Ruiz-Primo （1998） 描述一個對於科學實作評量的框架，在這計劃中，他們以各種不同的評分標準跨越了各種任務。任務的類別有：調查較

有用的方法、識別主要構成要素、分類調查及觀測調查。分析評分計劃由以下四方面組成：基於過程、基於證據、基於分類合理性，以及基於數據準確性。表格1為上述四個要件，提供了關於實作類型、任務、回答格式及評分要素各種不同類型的例子。

表 1　科學實作評量類型

| 實作類型 | 任務 | 回答格式 | 評分要素 |
|---|---|---|---|
| 比較性的探索：溶解實驗 | 給與學生三種粉末，學生須判定哪一種粉末最容易溶解於水中。 | 學生必須寫下如何進行實驗，與發現的結果。 | 歷程導向: 學生進行實驗的過程是否合理，最後解答的正確性。 |
| 成份辨識：神秘粉末 | 一個紙袋裝了幾種粉末，學生須判定紙袋中有哪些粉末。 | 學生須寫下如何進行實驗來決定袋子中所具有的粉末，與最後的辨識結果。 | 證據導向:如何判斷粉末存在的證據，與最後結果的正確性。 |
| 分類：石頭圖鑑 | 提供數種石頭，學生須將這些石頭進行分類，並指出各種石頭的特性。 | 學生須展示如何進行各種石頭的分類，並指出各種石頭的不同特性。 | 分類合理性:分類特質的正確性與精準性。 |
| 觀察：地質調查 | 依據石頭的出產地進行實地訪查，並描述當地地質、地形、氣候等。 | 學生需提供觀察的歷程與結果。 | 數據準確性:如何蒐集證據與描述之準確性。 |

思考及推論解決的任務過程也被用來設計評量，舉例來說，分析認知型的任務可使用放聲思考（Ericsson & Smith, 1991），目前已被運用在醫學領域上（Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999）。透過受試者思考、知識、過程及提問的特徵，都可用來評估此領域技術的專業程度（Glaser, Lesgold, & Lajoie, 1987），這些特徵更可以用來嵌入不同的評分規準。

## 二、評分規準與原則

設計評分項目是一個不斷反覆的過程，教師設計完之後，可先進行測試，看看評分項目是否學生所有的行為表現，而後透過幾次的修改，才能讓評分的項目較為完整。設計評分項目需要明確的標準，不管是對於評論表現品質或選擇一個評分步驟（例如：分析型的或整體型的）。基本上來說，評分標準是由一群專家學者，藉由他們在不同領域的知識，以及身為教育家的經驗過程發展而來，這些專家也同樣參與設計表現任務，且擁有關於學生在不同階段所表現的不同精熟程度的知識。Clauser (2000) 指出有幾種可能的方法用來確認標準是否合宜，例如請專家進行放聲思考，或解析對任務應有的可能回應。

評分標準說明在各個分數階段和被測量的架構是有關的，事實取決於包括是否

為一個成果或過程的測驗、測驗的任務需求、參與測驗人數及測驗目的和分數解釋。得分水平的數量取決區分多少個不同層次的表現。然而，得分的水平設定不宜太多，三到四個層次即可。

評分規準（rubric）有三種主要類型：總結性、分析性及檢核表（Huot,1990; Miller & Crocker, 1990; Mullis, 1984）。

選擇一個特定測驗的評分程序取決於測驗的目的與分數的詮釋，使用總結性評分，測驗者會根據寫作的品質做出單一且全面性的評論並打一個分數，例如表2所陳述的為國中基本學力測驗中的寫作測驗的評分規準摘錄，其採用規準為總結性評分，受試者在收到成績時，只會拿到一個來代表自己的成績。這種評分對於受試者人數多，如國中基測的作文考試，動輒數萬考生，較為適合。而分析式的評分則如表3，評分者評量寫作是根據數個面向，例如：立意取材、結構組織、遣詞造句、錯別字、格式與標點符號。而評分要呈現出每個項目的品質。根據 Mullis （1984）的摘要指出「總結性的評分是設計來描述全盤性的；或是各部份的總結，而分析式的評分則設計為描述單一特徵或部份主題，並以加總方式達成全面性的評分。」分析性各部份的評分高低，取決於內容的相對重要性，若是這項指標，是評量的重點，則給予這個分向度的配分可以較高。然而，分析性評分對於教師而言較為費時費力，但可提供受試者表現強項與弱項資訊的回饋訊息。

表 2　總結性評分範例：國民中學學生基本學力測驗寫作測驗評分規準摘要表

| 級　分 | 國民中學學生基本學力測驗寫作測驗評分規準 |
|---|---|
| 六級分 | 六級分的文章是優秀的，這種文章明顯具有下列特徵：<br>※立意取材：能依據題目及主旨選取適切材料，並能進一步闡述說明，以凸顯文章的主旨。<br>※結構組織：文章結構完整，脈絡分明，內容前後連貫。<br>※遣詞造句：能精確使用語詞，並有效運用各種句型使文句流暢。<br>※錯別字、格式與標點符號：幾乎沒有錯別字，及格式、標點符號運用上的錯誤。 |

註：內容摘取自國中基測網站http://www.bctest.ntnu.edu.tw/writing.htm

表 3　分析性評分範例: 改編自國民中學學生基本學力測驗寫作測驗評分規準

| 項　　目 | 分項得分 | 教師評分 |
|---|---|---|
| 立意取材 | 25% | |
| 選取適切材料 | 15 | |
| 能闡述說明，以凸顯文章的主旨 | 10 | |
| 小計 | 25 | |
| 結構組織 | 25% | |
| 文章結構完整 | 10 | |
| 內容前後連貫 | 15 | |

| | |
|---|---|
| 小計 | 25 |
| 遣詞造句 | 25% |
| 能精確使用語詞 | 15 |
| 有效運用各種句型使文句流暢 | 10 |
| 小計 | 25 |
| 錯別字、格式與標點符號 | 25% |
| 沒有錯別字 | 20 |
| 標點符號運用正確 | 5 |
| 小計 | 25 |

檢核表則列出表現或成果的測驗向度、在適當空格中做記號 ，教師可將所有需評鑑的要點寫下來，而後再來檢核是否學生的表現符合標準（如表4）。檢核表適合用在只想檢視學生達成某項任務與否，而不想用分數來區分學生成績的高低情況。

表 4　檢核表範例: 改編自國民中學學生基本學力測驗寫作測驗評分規準

| 項　　目 | 是否達成 |
|---|---|
| 選取適切材料 | |
| 能闡述說明，以凸顯文章的主旨 | |
| 文章結構完整 | |
| 內容前後連貫 | |
| 能精確使用語詞 | |
| 有效運用各種句型使文句流暢 | |
| 沒有錯別字 | |
| 標點符號運用正確 | |

總體而言，製作評分規準的步驟如下:

1. 參考其他Rubric範例，選擇適合的類型。
2. 依據需求來選擇總結性、分析性或檢核表的評分規準。
3. 配合教學目標，來思索受試者應該要表現出的行為能力。
4. 根據行為能力來定義評分規準。
5. 羅列學生在每個規準上的不同表現方式或是程度。
6. 訂出不同的程度等級、或界定每個等級的分數範圍。
7. 制定適合的表格。
8. 依據表格評分，檢討是否有不足之處並進行相關修正。

制定評分方式的嚴謹度需依考試的風險度而定。若是班級課堂評量使用，則可由任課老師，依據現有的評分表格進行修改。若是風險較高，重要的國家考試，則須由一群專家學者，共同制定出評量要點，透過不斷的修正與改進，才能制定出一個較為完善的評分方式。

國外有一些現成的網站可以使用，新手老師可以參考這些網站的內容，來進行評分規準的設計。以下大致介紹這些網站，有興趣的讀者可以深入研究網站的內容。

1. Rubistar (網站為http://rubistar.4teachers.org)

這個網站可以下載別人已經制訂好的歸準，也可以用裡面的程式來寫制定自己要用的規準。其中可以直接下載的評分規準內容涵蓋口頭報告、作品、多媒體、科學作業、寫作、工作技能、數學、藝術、音樂、閱讀等領域。然而，別人寫好的評分規準不見得可以直接拿來用，所以網站將各類別的評分規準予以統整，並將所有主要的子項目進行歸類，使用者在使用其系統時，不用重新去想評分項目，而是可以點選系統內已經設好的向度進行修改即可，對於老師而言，應該可以節省不少時間。

2. Authentic Assessment Toolbox (網站為http://jfmueller.faculty.noctrl.edu/toolbox/)

這個網站為美國North Central College in Naperville大學教授Dr. Jon Mueller所架設，除了有豐富的理論介紹外，例子多都搭配作業的描述，所以讀者可以清楚看到作者原先所設計的實作評量，與其搭配的規準為何，除此之外，還提供檔案評量的作業範例。其範例涵蓋國小、國中、高中與大學階段各式各樣的實作評量，不僅評量者可以使用其評量規準，教學者亦可以參考其實作評量的範例融入課程使用。

3. Rubric Library (網站為http://www.fresnostate.edu/academics/oie/assessment/rubric.html)

此網站提供許多評分規準的範例，尤其是非學科的部分，包括政策性的寫作、領導才能、口試、計畫案、批判性思考、戲劇寫作等。

## 三、信效度

教室評量中，很少能對實作評量的信效度進行評估，然而，評量設計者可借由信效度的評估來檢視評量的成效。信效度亦可讓教師對於自己所設計的評量品質更加了解。

信度是指評量結果的穩定性（stability）及一致性（equivalence; consistency）。評量結果的穩定性可由再測信度來評估，然而，由於實施實作評量費時費力，大多都只能施測一次。因此，大多使用評分者的一致性來檢視實作評量的信度。

大多數的教室評量，評分者侷限為教師一人，這種情況無法計算評分者一致性。然而，若是能選出一些學生當作評分者，或是在進行合作教學時，和搭擋的教師一起評分，此時即可以使用評分者信度來進行實作評量的信度評估。評分者在進行評分時，常常會有盲點，例如對某些學生平常的印象很好，即使在此測驗的表現

不好，也會因為印象分數而給高分。這些干擾的因素，都可能會影響到測驗的結果，透過其他評分者的評分，可進一步檢視評分的客觀性。

肯德爾和諧係數常用評估評分者的信度，此係數用來評估 K 位評分者，針對 N 位受試者表現評比時的評分一致性，也可以視同一個評分者先後 K 次評 N 個對象。其計算公式如下，其中 K 代表評審者的數目，N 代表受試者的數目，∑Ri 為每個被評對象所評等級之和。$\sum Ri^2$為每個被評對象所評等級之平方和。

$$W = \frac{\sum Ri^2 - \frac{(\sum Ri)^2}{N}}{\frac{1}{12}K^2(N^3 - N)}$$

計算時，除了可以使用現成的統計軟體之外，也可以用excel甚至計算機來進行計算，其計算步驟如下。

步驟1：陳列每位評分者對每位選手的評分總分。假設有五位評分者ABCDE，他們針對三位選手歌唱表現進行評分，而下面則是每位評分者所給的總分，第一位評分者A，給選手甲75分，給乙選手73分，給丙選手63分…以此類推。

| | 甲選手 | 乙選手 | 丙選手 |
|---|---|---|---|
| 評分A | 75 | 73 | 63 |
| 評分B | 67 | 73 | 65 |
| 評分C | 86 | 82 | 69 |
| 評分D | 54 | 70 | 66 |
| 評分E | 83 | 77 | 87 |

步驟2：把分數排序。每位評分者的成績進行排序，例如，對於評分者A來說，給甲選手的成績最高，所以排序為一，乙選手的成績次之，所以排序為二，丙選手為第三，排序為三。而 ∑Ri 則是將排序的成績加總起來，而 $\sum Ri^2$ 則是將Ri先平方再進行加總。

| | 甲選手 | 乙選手 | 丙選手 |
|---|---|---|---|
| 評分A | 1 | 2 | 3 |
| 評分B | 2 | 1 | 3 |
| 評分C | 1 | 2 | 3 |
| 評分D | 3 | 1 | 2 |
| 評分E | 2 | 3 | 1 |
| Ri | 9 | 9 | 12 |
| $Ri^2$ | 81 | 81 | 144 |

步驟3：帶入公式

K = 幾位評審 = 5
N = 幾位受試者 = 3

$$W = \frac{\sum Ri^2 - \frac{(\sum Ri)^2}{N}}{\frac{1}{12}K^2(N^3 - N)}$$

$$= \frac{(81+81+144) - \frac{(9+9+12)^2}{3}}{\frac{1}{12}5^2(27-3)}$$

$$= 0.12$$

評分一致性為0.12，和諧係數w越大則一致性越高。一般來說，W值介於0.9~1.0代表評分者之間的評分非常高相關， 0.7~0.9代表高相關，0.5~0.7代表中等相關，0.3~0.5代表低相關，0.0~0.3代表微相關。因此在本例中，評分者的信度較低，代表評分者評分分數之間的相關性很弱。

效度是用來評鑑評量結果的解釋與使用的合適性， 在許多考試中，效度的評估是極為重要的。若是可以的話，效度的證據可多方蒐集，用以評估評量的成效。效度有分為很多種，包括內容效度、效標關聯效度、構念效度等。內容效度指的是指測驗的內容是否符合測驗的目的，若是題目的內容符合教學目標、所選的教材也有代表性, 則我們稱測驗的內容效度是良好的。教師在設計實作評量時，可以將設計好的教案，請其他老師檢視看看，甚至可以請相同領域的學者專家來看看是否所設計的實作評量任務，能夠與教材內容所涵蓋的範圍與教學目標相符。

效標關聯效度指測驗的分數與其他相關測驗或指標的相關性。如果設計的為數學的實作評量測驗，那麼可選用的效標為學校老師給定的成績、學生課堂時的表現、其他相關的數學測驗（像是學生的期中考數學成績）、或是學生是否在其他數學競賽中的得到優異的成績。如果測驗分數和外在效標之間的相關越高，則代表效標關聯效度越高，也就是越能用測驗分數來有效解釋及預測外在的效標。

建構效度則是建立在構念之上， 而構念是在心理學或社會學上所存在的一個理論上的構想特質，不容易觀察，也很難被測量，但是我們卻可以假想這是存在的。建構效度的建立，必須由研究者先提出假說，並蒐集資料去驗證 並反覆檢討、修正整個建構的過程， 直到建構效度可以成立為止。而內容效度與效標關聯效度的建構方法與結果，都可用來當作建構效度的證據。建構效度的驗證方法種類繁多, 有內部一致性的分析法、外在效標分析法、因素分析法、結構方程式模式、多特質-多方法分析法等。

每種效度的證據若能都蒐集是最好的，可一般教師在課堂上要能找到這些證據較為困難，比較可行的應該是蒐集內容效度與效標關聯效度的相關證據，內容效度的獲得可透過其他教師的相互討論，而效標關聯效度則是依據學生在班上的其他表現，如學習成績等。而建構效度，需要一些較為複雜的統計方法與統計軟體來執行，有興趣的讀者可參閱余民寧（2011）專著。

此外，針對教室評量其相關更深入的信、效度議題可參考Brookhart（2003）。

## 四、實作評量應用

### (一)大型測驗

實作評量在操作上較為費時費力，但還是可能以大規模的方式進行施測。若將數個實作評量集結起來，並進行系統性的歸類，則可變成檔案評量。LeMahieu et.al（2005）所進行檔案評量，即是將數個實作評量的內容集結起來。實驗中將美國匹茲堡的學區內六年級到十二年級的學生要求進行寫作的檔案製作，並對其中部份的檔案進行隨機抽樣，取出了1250份的檔案，針對寫作的三個面向，來進行學生寫作能力的評估。每份檔案須包含以下幾份作品

(1) 依據自己的標準，選擇一份最重要的作業
(2) 一份自己最滿意的作業
(3) 一份自己最不滿意的作業
(4) 一份自選作業，但須寫出選這份作業的理由
(5) 若是班級教師覺得學生所選擇的作業不夠具有代表性，則可以再為學生選一份作業。

除了這五份作品之外，檔案中還需提供一份目次表，來描述檔案中作品的內容與製作日期，一份寫作的問卷來描寫個人成為一個寫作者的經驗，及一個反省回顧的描述，來記錄學生過去一年來的寫作能力的變化。從這份檔案的內容的描述可以看出，學生對於自己檔案的內容有充分的決定權，某學生可以選擇寫新詩，而另一位學生可以選擇寫短文，在所有作品選擇中，並沒有硬性的規定學生應選取何種作業放進檔案中。為了避免學生會對所選取的文章進行過多的修飾，抽到的學生在一個禮拜前才會進行通知。

此檔案的評分從三個面向來進行探討（見表5），而每一個面向都有六分，得0分代表學生的表現不足（inadequate）而6分代表學生的表現卓越，若是評分者覺得檔案內容不足，無法對某個面向進行評分，也可以評證據不足（no evidence）。第一個面向為寫作的成就。內容包括學生的寫作品質，對於寫作能力、技巧、架構、文章標題的了解程度與語言的表達能力都與以評價，這個面向與一般對寫作的要求相同。第二個面向為評估學生對於寫作過程及策略的運用能力，內容包括有效的使用

預寫的策略，使用草稿來形成自己的想法，並利用外界的資源（例如同儕討論、讀者或是其他成人的回饋）來進行文章的修正。第三個面向為寫作者的成長與發展的能力評估。學生必須展現對於寫作的熱誠與態度、如何看出個人寫作的優缺點並給予評鑑，並能夠對不同的目的、題材與對象來寫作。

表5　匹茲堡寫作檔案評分向度與內容

面向一:寫作成就
(1) 達到具有價值性的挑戰
(2) 建構及維持目的
(3) 使用技巧及選擇題材
(4) 控制慣用語、字彙與語句結構
(5) 瞭解讀者的需求（組織、發展與使用細節）
(6) 使用語言、聲音、圖片與語態
(7) 幽默感、比喻、有趣性

面向二:寫作過程及策略運用
(1) 有效的使用預寫的策略（prewriting）
(2) 使用草稿來發現並修正想法
(3) 使用討論的機會來修正寫作（同儕、成人、或讀者）
(4) 有效的進行修改（改造、重新聚焦與修改）

面向三:身為一個作者的成長、發展與專注
(1) 對寫作任務投入性的證據
(2) 增進對寫作的投入性
(3) 發展身為作家的感覺
(4) 個人寫作標準的演進
(5) 能看出某人寫作的優點及需要
(6) 能在進行寫作作業上展現冒險與創新
(7) 可以使用不同目的、題材或是對象來寫作
(8) 最早的作業和最近的作業之間的進步、成長與發展

註: 翻譯自LeMahieu et al.（1995）

此研究共有25位評分者參與檔案的評鑑。其中，12位進行國中的檔案評鑑，而13位進行高中的檔案評鑑。評分者的組成份子為教師與閱讀或是寫作領域的專家，國中組的專家每人須評鑑99份的檔案，而高中的專家則須評鑑78份的檔案。進行正式評分前，每位評分者都須經過訓練並使用幾個範例檔案進行評分練習。透過不斷對於各面向的討論與實例探討，評分者對於規準的了解程度與實例檔案的各面向評分達到一致性之後，才正式對學生的檔案進行評分。

每份檔案須有兩位評分者對三個面向進行評分，若是兩位評分者所評的分數差異性在一分之下，則以兩位評分者所給分的加總作為這份檔案的分數。如果差異性在一分以上，則有第三位評分者進行仲裁評分。

評分過程經歷整整一周的時間，得到相當高的信度，三個向度的信度約為0.74到0.87之間，尤其以寫作成就的信度最高。而對於評分者之間的信度（inter-rater reliability），也達到0.80到0.84之間。從這個研究可以看出，只要能夠對實作評量進行充分的規劃，對於評量的內容詳細加以說明，並設定嚴謹的評分程序與對評分者的訓練，即使讓受試者自由選擇檔案中的內容，亦可以得到良好的信度。

### （二）醫學臨床技能測驗

實作的能力在醫學界的需求已相當風行，且考選部在2013年已經將臨床技能測驗直接納入醫生職照的先備考試中，現在的醫學生必須先通過這個測驗，才能參加國考(曹以會，2013)。

這項測驗藉由情境模擬實作的歷程，來評估考生應具備的能力，測驗分為12站，其中前8站是透過標準化病人演出的試題，考生依序到不同的測驗站接受測試，每個測驗站都設定一個情境，病人會有不同的身體狀況來”演出”某種疾病的症狀，考生必須在15分鐘中內，來進行問診、身體檢查、溝通衛教等。而後4站則是臨床技能的操作題，包括操作醫療器材的準確度與精確度等。

在這個測驗中，其主要的評量向度包括與病人溝通、為病人看診的態度，以及面對病人時能否表現出良好的態度與互動能力，透過這些向度，來當作評估醫學生是否合適擔任醫生的標準。 在所有的12站中，受試者必須通過七站才算合格。

這項測驗動員了龐大的人力，包括768位主治醫師擔任考官， 512位標準化病人配合測驗，還有眾多的試務工作人員。雖然實施的成效還需評估，但可看出醫學界已相當重視使用實作評量，來進行評選適合的人才。

## 參、實作評量範例

在生活課程中，著重讓學生體驗各種姿態、表情動作的美感，並表達出自己的感受。因此，在這位老師的實作評量設計中，讓學生親自體驗當模特兒，透過走台步和產品代言拍攝來展現自己。

### 一、設計理念

這是一份讓同學們親自體驗當模特兒的實作評量，評量內容包含走台步和產品代言拍攝兩大部分，這兩項測驗都是以測驗學生是否具有模特兒必備的專業能力為目的，畢竟模特兒的專業不能只靠書本知識的吸收，能否將吸收的知識展現在這兩個測驗項目，才是能否成為專業模特兒的關鍵，透過這樣的實作評量，同學們才能知道自己不足之處，進而能透過本次的評量自行調整改善。

## 二、指導語和作業說明

### (一)指導語

通常在實作評量的開始進行前，或有簡單的指導語，來告知同學接下來要做甚麼樣的活動。說明活動的目的與正確的回答方法，讓受訪者能認真的據實回答，有助於增加評量的效度。若活動中有特殊需求及回答方法，也應事先說明。

指導語須包含以下幾個要點：

1.敘述實作評量的題目與認真作答的重要性
2.告訴受試者評分的重點
3.評量活動進行的方式與可能需要花費時間
4.如果有疑問，應如何尋求幫助

在此處所提供的範例，所提供的指導語如下:

*給未來的模特兒們：*

*在過去專業訓練下，準備好要show出你們努力的成果了嗎？記住，唯有跨過重重難關的模特兒才能站在群眾的面前發光發亮！準備好要接受考驗了嗎?這次的成果驗收將從由兩個不同層面分開檢視評分，分別是走台步和產品拍攝代言的部分。走台步方面，透過不斷的練習，希望學生能在走秀時態度自信，在肢體動作上展現模特兒的水準。另外，在產品代言攝影的部份，也期待學生能夠發揮所學，構想出符合產品功能及風格，且與肢體能相搭配的和諧畫面，以下將有更詳細的說明。*

*台步的展示將會是第一個施測的項目，同學需沿著地上的直線行走，且在標記處擺出一兩個姿勢後轉身回到出發時的標記處再次擺出姿勢後離開。在這樣的過程當中，眼神除了要透露出自信之外，也要盡量保持平視，避免東張西望或是看地板，在最後的姿勢方面則要記得自己身體的優勢，擺出能夠展現優勢、掩蓋缺點的姿勢。轉身的時候也要記得保持平衡，抬頭挺胸，這些都會列入評分的項目之中。至於攝影的部分，受試者需要在限制的四樣產品當中（分別是雨傘、包包、水壺和手機），選擇自己喜愛且能發揮的產品進行拍攝，一個產品將會拍攝三張照片，在拍攝過程中，將由同一位攝影師進行拍攝，但模特兒可有自己的想法與攝影師溝通拍攝與取景角度。在拍完兩個商品共六張照片之後，評審們將會根據攝影的過程和拍攝出來的照片進行評分。除了能夠明確表達出該產品的核心概念外（例如：雨傘的功能是遮雨），肢體和表情是否和產品有互相搭配也是這個測驗項目的評分重點之一。另外，照片當中是否有使用不同的姿勢和表情來傳達商品也會在評分考量內。這是我們蒐集的資料，僅當做研究使用，且會對你的個人絕對保密。如果你有甚麼問題，可以問身邊協助的同學喔!*

## (二)評分規準

採用評定量表作為判斷模特兒表現的評量工具，量尺類型屬於數字型評定量表。評分者依據情境角色所表現出的程度圈選適當表現該特質程度的數字；「1」代表不佳、「2」代表有待加強、「3」代表尚可、「4」代表優良、「5」代表非常優秀。

| 級分 | 評分規準 |
|---|---|
| 5 | 模特兒表現優異，符合以下特徵：<br>(1) 肢體動作能自由駕馭，態度自信大方，不扭捏。<br>(2) 代言充分呈現產品特色與核心概念，模特兒充滿個人魅力與特質，創意度極高。<br>(3) 照片具有吸引觀眾購買之潛力。 |
| 4 | 模特兒表現已在水準之上，符合以下特徵：<br>(1) 肢體動作表現得宜，偶有不流暢之處，但不致影響評分。<br>(2) 代言已能掌控產品特質，經提示後，能發揮創意表演。<br>(3) 照片可激發大部分觀眾購買意願。 |
| 3 | 模特兒表現已達一般水準，符合以下特徵：<br>(1) 肢體動作表現尚稱完整，偶有銜接不順之處。<br>(2) 代言尚可，偶需旁人提點產品核心概念，創意度尚可。<br>(3) 照片效果尚可，觀眾購買意願持平。 |
| 2 | 模特兒表現未達基本水準，符合以下特徵：<br>(1) 肢體動作斷斷續續，銜接不順暢。<br>(2) 代言表現度待加強，大部分需旁人提點產品核心概念，無法自行發揮創意表現產品特質。<br>(3) 照片效果待加強，觀眾購買意願低。 |
| 1 | 模特兒表現不佳，符合以下特徵：<br>(1) 肢體動作不協調，神情渙散沒有自信。<br>(2) 代言表現度不佳，經提示及示範後，仍無法呈現產品特質。<br>(3) 照片效果不好，觀眾無購買意願，非模特兒人選。 |

共5位評分者 (k = 5) 及10位被評者 (N = 10)，經過總分計算，每位評分者對 10 位受試者的評分排序如下：

| 評分者（k=5） | 學生作品（N=10） | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 甲 | 乙 | 丙 | 丁 | 戊 | 己 | 庚 | 辛 | 壬 | 癸 |
| A | 5 | 6 | 4 | 1 | 2 | 3 | 5 | 9 | 8 | 10 |
| B | 6 | 7 | 3 | 2 | 8 | 1 | 9 | 4 | 5 | 10 |
| C | 7 | 9 | 2 | 1 | 5 | 3 | 4 | 8 | 6 | 10 |
| D | 2 | 9 | 6 | 1 | 4 | 7 | 5 | 8 | 3 | 10 |
| E | 7 | 8 | 2 | 1 | 6 | 3 | 9 | 5 | 4 | 10 |
| Ri | 27 | 39 | 17 | 6 | 25 | 17 | 32 | 34 | 26 | 50 |
| Ri2 | 729 | 1521 | 289 | 36 | 625 | 289 | 1024 | 1156 | 676 | 2500 |
| $\Sigma Ri=27+39+17+6+25+17+32+34+26+50=273$ | | | | | | | | | | |
| $\Sigma Ri2=272+392+172+62+252+172+322+342+262+502=8845$ | | | | | | | | | | |

代入和諧係數公式可得和諧係數為為=0.674

以和諧係數來計算評分者間的信度係數，由這五位評分者對十位參賽者的評分結果以等第分數評定後，計算其評分者間信度係數為0.674。

## 肆、結語

在本文中提出的內容，希冀能提供相關單位參考，最後，提出幾點建議。

### 一、評量帶動教學的改變

考試領導教學是大多數學者所反對的一個方向，然而，不諱言的在台灣升學主義下，往往是升學考試要考甚麼，老師就教甚麼，在傳統選擇題的紙筆考試下，著重的往往是某部份、片面性的知識，而且要能夠廣泛的考到所有的內容非常困難，而這些部份性的、片面性的知識卻要用以代表學生所有的學習成效，因此老師在教學時，只教要考的重點，對於考試不考的內容就跳過，然而，這些片段式的知識往往是見樹不見林，學生記幾個重點公式，會看關鍵字套公式，卻不了解整個內容來龍去脈，因此可能考完就忘記，沒辦法學習到整體性的知識。

在實作評量是鼓勵考試來帶動教學的改變，先把評量規準告訴學生，告訴他們這學期要評鑑的重點就是依照規準來評，因此老師按照規準的重點來教，學生也要依重點來學，如寫作測驗中，評量歸準可能包含大綱的訂定、文獻搜集的深廣度、段落的清晰、文句排列的邏輯性等，老師在教學時，依據這些重點來教，而學生也照這些重點來學，一整個學期下來，學生大概就會知道寫作的重點有甚麼了，也能夠寫出一篇像樣的文章。因此實作評量上而言，學生需要如何去操作一個實際的任務，而且事先可以先看到評分規準，知道甚麼是好的表現並依據這樣的要求去完成，並在整個學習過程中，可以培養完整的知識與技巧。

### 二、實作評量的任務設計

實作評量搭上多元評量的列車，許多學校都在推行，然而，在任務執行中，須隨時檢視評量最終的目的，是為了解學生學習的成效，應以學生的學習為中心，且評量應與教學目的、課程內容緊密結合。因此在設計實作評量任務時，若能與相同領域的老師共同討論，甚至能跨不同學科共同設計，方能達其效果。而目前對於實作評量的理解與推行上仍有努力的空間，希冀相關單位能藉由辦理研習課程、來推廣這方面的知能。

### 三、標準本位評量的趨勢

過去的教室評量結果重試排名，只知道學生的名次，卻不知學生在學習上是哪

裡不懂，也不知理解程度為何，十二年國教推行後，標準本位評量將為推行重點所在，重視的是學生了解學科內容的程度，依據學科習得知識的了解與應用程度，將學生的表現分成幾個等級，而每個等級對應出甚麼樣的表現水準，都有明確的指出方向。例如在每個學習階段、學科領域都有各自的課程綱要與對應出的基本能力指標，而透過這些能力指標將其轉換成學習內容方向、並設計相關的實作評量任務，發展評分規準，而據此評量與了解學生的表現。這樣的標準本位評量模式，將是未來評量的新趨勢。

根據美國視導與課程發展學會（The Association for Supervision and Curriculum Development, ASCD）研究指出，在各種考試形式中，以學生檔案評量效果最佳，其次為實作評量、上台報告、期末考試與州立大型標準化測驗（ASCD Smartbrief, 2012），而檔案評量和實作評量相當類似，都是透過真實性評量來評估學生的學習結果，然而，這樣的考試進行方式，需要相關教育單位、教師、家長與學生共同的努力與配合，方能推廣與執行。

## 伍、參考文獻


余民寧 (2011)。**教育測驗與評量：成就測驗與教學評量**（第三版）。台北：心理出版社。

曾薏蘋 (2012) 。奶嘴公務員鬧笑話公職考試尋專才。**中國時報**。取自：http://blog.udn.com/baogon/6339187

曹以會 (2013) 。OSCE醫學臨床測驗今年正式舉辦 近99%及格。**聯合報**。取自:http://mag.udn.com/mag/edu/storypage.jsp?f_ART_ID=459982Power By udn.com

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

ASCD Smartbrief. (2012). Which do you think provides the most accurate summative assessment of student learning? Retrieved February, 4, 2013 from http://www.smartbrief.com/news/ascd/poll_result.jsp?pollName=89A63917-5867-47C5-8F56-1B50832D90D3&issueid=1318C04D-9D20-4609-9608-FA30F3D9363A

Aschbacher, P. R. (1991). Performance assessment: State activity, interest and concerns. *Applied Measurement in Education,* 4 (4) , 275-288.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education,* 4 (4) , 305-318.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment

purposes and uses. *Educational Measurement: Issues and Practice, 22* (4) , 5-12.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement,* 24 (4) , 310-324.

Ericsson, L. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An instruction. In LA. Ericsson & J. Smith (Eds.) , *Toward a general theory of expertise: Prospects and limits,* 1-38, Cambridge, MA: MIT Press.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning J. Glover, J.C. Conoley, & J. Witt (Eds.) , *The influence of cognitive psychology on testing and measurement.* The Buros-Nebraska Symposium on measurement and testing, 3, 41-875, Hillsdale, NJ: Lawrence Erlbaum.

Herman, K. L., Aschbacher, P. R., & Winters, L (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Huot, B. (1990). The literature of direct writing assessment major concerns and prevailing trends. Review of Educational Research, 40 (2) , 237-263.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science,* 208, 1335-1342.

LeMahieu P. G. Gitomer, D. H., Eresh, J. T. (1995). Portfolios in large scale assessment difficult but not impossible. *Educational Measurement: Issues and Practice.* 14 (3) , 11-28.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher,* 20 (8) , 15-21.

Linn, R. L. (1993). Educational Assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis,* 15, 1-16.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessments. *Phi Delta Kappan,* 80 (9) , 688-696.

Marzano, J., Schmitt, A., Bleistein, C. (1993). *Sex-related performance differences on constructed-response and multiple choice sections of the Advanced Placement Examinations* (RR-93-5). Princeton, NJ: Educational Testing Service.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education,* 3 (3) , 285-296.

Misley, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, LA. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior,* 15-335-374.

Mullis, I.V.S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement :Issues and Practice,* 3 (1) , 16-18.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.G. Gifford & M.C.O' s Conner (Eds.). *Changing assessment; Alternative views of aptitude, achievement and instruction,* 37-55, Boston: Kluwer Academic.

Shavelson, R. J., & Ruiz-Primo, M. A. (1998).On the assessment of science achievement conceptual underpinnings for the design of performance assessments: Report of year 2 activities (CSE Technical Report 481). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices,* 6 (1) , 33-42.

Wiggins, G. (1989). A true test：toward more authentic and equitable assessment. *Phi Delta Kappan,* 20, 703-713.

國家圖書館出版品預行編目資料

測驗之編製 : 命題技巧與測驗資料之分析 / 蕭儒棠等合著. -- 初版. -- 新北市 : 國家教育研究院, 民103.12
面 ; 公分
ISBN 978-986-04-3496-5(平裝)

1.教育測驗 2.命題 3.文集

521.307 103025269

書　　名：測驗之編製－命題技巧與測驗資料之分析
著　　者：蕭儒棠、曾建銘、吳慧珉、林世華、謝佩蓉、謝名娟
出版機關：國家教育研究院
地　　址：新北市三峽區三樹路 2 號
網　　址：http://www.naer.edu.tw
電　　話：(02) 8671-1111
出版年月：民國 103 年 12 月
版　　次：初版
其他類型版本說明：本書另有電子版本，網址為：http:// teric.naer.edu.tw
定　　價：新臺幣 170 元
展　　售：政府出版品展售中心
五南文化廣場：臺中市中山路 6 號
電話：04-22260330　傳真：04-22258234
網址：http://www.wunan.com.tw/
國家書店松江門市：臺北市松江路209號1樓
電話：02-25180207　傳真：02-25180778
網址：http://www.govbooks.com.tw/

GPN：1010302916
ISBN：9789860434965