教師專業成長與實踐智慧
2004 年 12 月，頁 93-104

測驗品質考驗與 *TestGraf 98* 的應用

楊志強
國民教育學系副教授
國立台北師範學院

壹、前言

在筆者初任國小教師時，記得每位老師前輩們總練就一手刻鋼板的好功夫，才能油印出一份工整的月考試卷。近年來，電腦科技的普及與書商的激烈競爭，各種版本的教科書皆提供相當便捷的題庫光碟系統，老師想要出一份隨堂考或是段考試卷，只要經電腦隨機選題、排版及列印，不消幾分鐘的光景，便能完成一份試卷，節省了許多時間。

筆者擔心過度依賴書商的題庫系統，而忽略了教學評量的目的與價值，特別是一份優良試題品質的試卷是經過質化分析與量化分析的判斷過程。在質化分析方面，出題者在試題施測前，可藉由試題的內容審查，有效命題原則與教學目標評鑑工作來進行，確保試題具有教學內容的代表性。在量化分析方面，則在試題施測之後，分析試題的難度（difficulty）、鑑別度（discrimination）與誘答力（distraction），以考驗測驗品質的優劣。藉由試題分析的過程，篩選試題，建立題庫的過程，才能真實地達到教學評量的目的與價值。

本文的主要目的是在介紹一套新的試題分析軟體，稱為 *TestGraf 98*，提供給教育領域的相關研究人員與現職老師們在量化試題分析使用。*TestGraf 98* 並不如傳統的試題分析軟體，計算試題難度指數（item difficulty index）、鑑別度指數（item discrimination index）或各試題選項的選答率，而是提供一選項特徵曲線（option characteristic curve，OCC），來解釋試題的難度、鑑別度與選項誘答力。以下章節即先介紹傳統的試題分析方法，再導入 *TestGraf 98* 軟體簡介與選項特徵曲線的解讀，最後再介紹 *TestGraf 98* 的實例操作。

貳、傳統的試題分析方法

傳統的試題分析方法主要是在分析每個試題所具備的三大統計特徵，即難度、鑑別度與誘答力（Ebel 與 Frisbie，1991）：

一、難度分析

難度分析即在計算試題難度指數或稱為答對百分比（number correct ratio），最便捷的一種計算方式就是某一試題全體受試者答對的比例，以 *N* 為全體受試者人數，*R* 則為答對人數，公式如下：

$$P_i = \frac{R_i}{N} \times 100\%$$

舉例而言，假設班上有 30 位學生。某次數學月考，計有 20 人答對第一題試題。則其試題難度指數為：

$$P_1 = \frac{20}{30} \times 100\% = 67\%$$

難度指數最大值為 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。根據 Ebel 與 Frisbie（1991）、王文中等（民 88）、余民寧（民 86）、郭生玉（民 74）、陳英豪與吳裕益（民 81）…等提出之試題評鑑原則，個別試題的難度指數應介於 0.4 至 0.8 之間，整體試題的難度指數應在 0.5 左右。

二、鑑別度分析

鑑別度分析即在計算試題鑑別度指數，將受測者的評量總分，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）。並依照高、低兩組通過測驗試題百分比的差，即為試題鑑別度指數 D，公式如下：

$$D_i = P_H - P_L$$

舉例而言，假設某次數學月考的第五題試題，高分組的難度指數為 0.8；低分組的難度指數為 0.2。則其鑑別度指數為：

$$D_5 = 0.8 - 0.2 = 0.6$$

鑑別度指數的最大值為+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。根據 Ebel 與 Frisbie（1991）、王文中等（民 88）、余民寧（民 86）、郭生玉（民 74）、陳英豪與吳裕益（民 81）…等提出之試題評鑑原則，鑑別度指數 0.4 以上是屬於非常優良試題；0.3 至 0.39 屬於優良試題，但可能需要修改；0.2 至 0.29 屬於尚可試題，但需做局部修改；0.2 以下屬於劣質試題，建議刪除。

三、選項誘答力分析

以選擇題試題題型為例，除了有一正確選項之外，尚須提供 3 至 4 個不正確的錯誤選項，來誘使那些知識不夠完整或僅具不份知識的受測者選擇，以發揮試題誘答功能，增加試題的鑑別力。也就是說，當學生的知識不夠完整，無法做出正確作答時，僅能猜題，一個具有良好誘答力的不正確選項，能夠吸引該類學生選擇。選項誘答力分析的分析方式是將受測者的評量總分，分成高分組（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），再分別計算出每一個選項的選答率即可，如表 1。

表 1　選項誘答力分析

試題	分組	選項			
		A	B	C	D
1	高分組	2	6*	1	1
	低分組	3	2*	3	2
2	高分組	8*	0	0	2
	低分組	7*	0	3	0

註：*為正確選項

根據 Ebel 與 Frisbie（1991）、王文中等（民 88）、余民寧（民 86）、郭生玉（民 74）、陳英豪與吳裕益（民 81）…等提出之選項誘答力評鑑標準，試題的選項誘答力應至少有一位低分組受試者選擇任何一個不正確選項，並且選擇不正確選項的低分組受試者應比高分組多。舉例而言，表 1 的試題 1 誘答力遠優於試題 2。

參、*TestGraf 98* 軟體介紹

一、理論基礎

有別於傳統的試題分析方法，另一分析方法為項目反應理論（item response theory，IRT），自 1950 年代 Lord 發表雙參數常態肩型模式及潛在特質理論，發展已有五十年的歷史（王寶墉，民 84）。特別是，最近加拿大心理計量學者 Ramsay 結合高低試題鑑別指數與核平滑無參數估算法，發展出正確選項與誘答選項均可分析之核平滑法無參數試題特徵曲線估算法（kernel smoothing approaches to nonparametric item characteristic curve estimation）於 1991 年發表於 Psychometrika 期刊上。核平滑（kernel smoothing）就是被估計的受試者加以排序後的函數和試題選項是否被選（選則指示值為 1，否則為 0）而成為二元變數（binary variable）之間的關係。此方法並無假設任何適當的模式，完全根據受試者實際作答資料來進行分析，是一種無參數（nonparameter）的試題反應理論（Ramsay，1991）。

Ramsay 根據上述理論，發展出 *TestGraf98* 軟體，可估計選項特徵曲線（option characteristic curve，OCC）。軟體為一分享軟體（shareware），可在 http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html 網頁免費下載（在楊志強教學資訊網 http://tea.ntptc.edu.tw/~cyang 亦有連結），參考文獻暨操作手冊 *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*（Ramsay，2000）亦可同時下載。

二、選項特徵曲線

選項特徵曲線的特性是利用圖形化的方式來記錄或比較資料及數據，將比繁冗的文字敘述或單純的數字表現來得詳盡且清楚。它以受試者的能力為橫軸，而以受試者者在某一試題之選答率為縱軸，事先並無假設其服從某一特定之試題反應模式，完全根據受試者之作答資料，配合上述之核平滑化法，得一平滑之曲線圖。選項特徵曲線對於診斷試題的問題能提供有效的幫助，它能決定是否重新命題以消除曖昧不明的試題選項或是提供較合理的誘答選項。以下圖例，正確選項以綠線呈現，錯誤選項則以紅線呈現，說明選項特徵曲線的意涵與解釋方式（楊志強與楊志堅，民 92）：

如圖 1 顯示，正確選項是 3，正確選項具有良好的鑑別度，能力愈高的受試者通過率愈高，能力愈低的受試者通過率則愈低。其他錯誤選項具有誘答力，能力愈高的受試者愈不會選擇該錯誤選項，能力愈低的受試者選擇該錯誤選項的比例就愈高。整體而言，這個試題是鑑別度與誘答力兼具的優良試題。

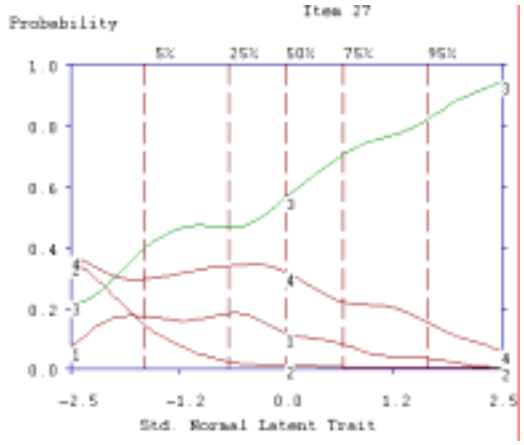

圖 1　選項特徵曲線（一）

如圖 2 顯示，正確選項是 3，正確選項在能力介於 5%至 25%之間的受試者鑑別度不佳。錯誤選項 2 對於能力低於 50%的受試者具有誘答力，其餘錯誤選項不具誘答力。

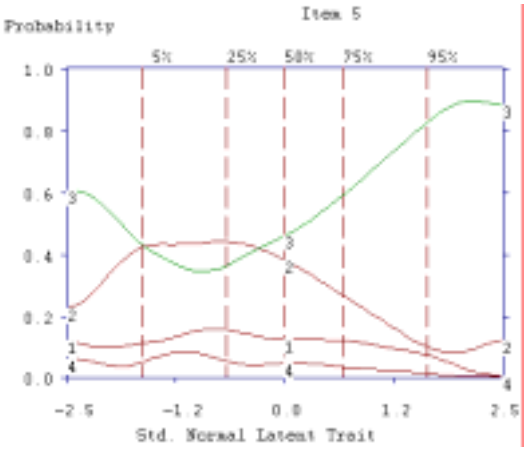

圖 2　選項特徵曲線（二）

如圖 3 顯示，正確選項是 4，正確選項在能力低於 25%以下的受試者稍具鑑別力，對於能力高於 25%以上的受試者不具鑑別力。所有錯誤選項僅對能力低於 25%以下的受試者稍具誘答力。整體而言，這個試題過於簡單，對於中高能力的受試者不具鑑別力。

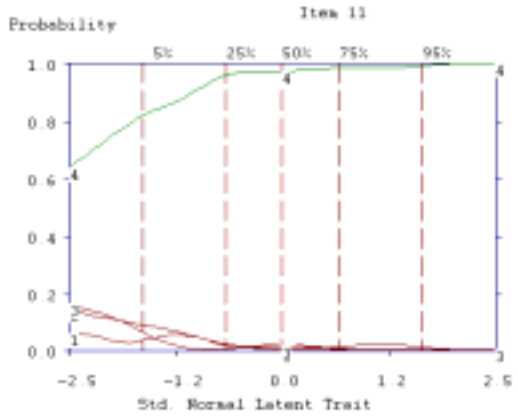

圖 3　選項特徵曲線（三）

如圖 4 顯示，正確選項是 2，能力愈高的受試者通過率愈高，能力愈低的受試者通過率則愈低，但高能力受試者與低能力受試者的答對率差異不大；另外，錯誤選項 1 與正確選項是 2 的曲線相似。整體而言，這個試題鑑別力不佳，而且，推斷選項 1 與選項 2 的概念可能重疊，選項 1 與選項 2 可建議修改，鑑別度亦可能因此提升。

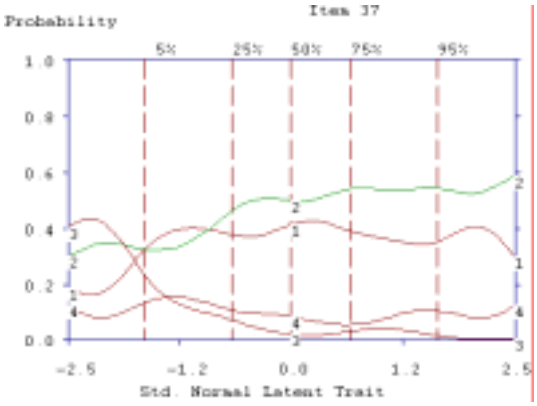

圖 4　選項特徵曲線（四）

更多實例詳見於楊志強與楊志堅（民 92）、楊志強（民 93）、吳慧民（民 90）及林清芳（民 92）…等之相關研究論文。

肆、*TestGraf 98* 軟體操作

一、資料輸入

在未執行 *TestGraf 98* 之前，事先輸入原始測驗資料，方可啟動程式進行試題分析。舉例而言，有一份 5 題的四選一選擇題試卷，共有 10 個受測者，資料如表 2：

表 2　受測者的原始測驗資料

座號	第 1 題	第 2 題	第 3 題	第 4 題	第 5 題
01	1	2	1	2	1
02	1	1	2	1	4
03	1	2	4	2	1
04	1	2	4	2	1
05	1	2	3	4	2
06	1	2	1	3	3
07	1	1	2	2	1
08	2	4	3	2	2
09	1	4	4	2	1
10	3	3	1	2	2

TestGraf 98 所需的測驗資料，須以記事本、Wordpad、Word 或其他文書處理軟體事先輸入原始資料，以文字檔（text file）形式儲存，再供 *TestGraf 98* 讀取。以圖 5 為例，開啟記事本，輸入資料，第一列資料須宣告有 5 題試題，座號代碼為 2 個字元；第二列資料為正確答案，00 為座號代碼，1、2、4、2、1 分別為第一題至第五題的正確答案；第三列起，即為表 2 內容的測驗資料。輸入完畢，以文字檔形式儲存，將該筆資料命名為 exam.dat（可自行命名）。

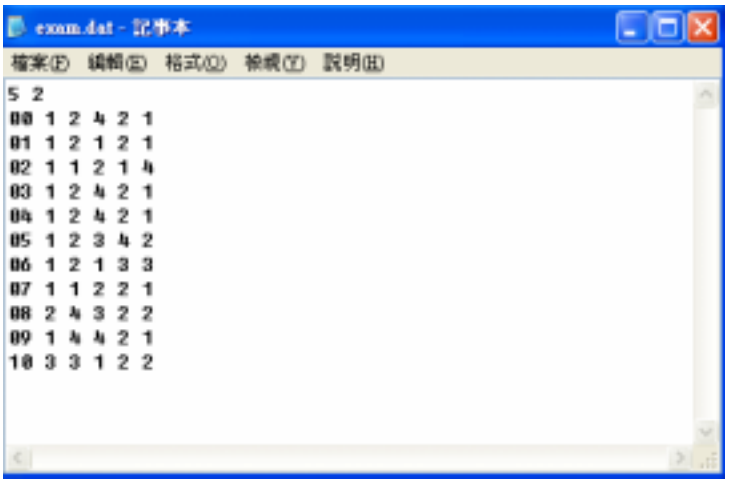

圖 5　以記事本輸入測驗資料

二、操作步驟

點選執行檔 Testgraf98.exe 啟動 *TestGraf 98* 程式，開啟畫面如圖 6。

圖 6　*TestGraf 98* 的啟動畫面

第一步驟首先點選 New 選項，開啟 exam.dat 之後，會先後出現圖 7 與圖 8 的對話框，詢問試題型式，逕行點選 OK 即可，宣告試題為 5 題的四選一選擇題。

圖 7　試題型式對話框（一）

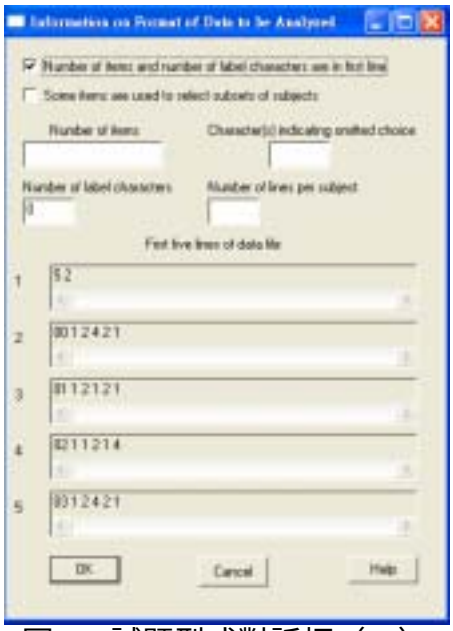

圖 8　試題型式對話框（二）

第二步驟接著再點選 Edit 選項，開啟 exam.tg（該檔會自動產生），出現圖 9 的編輯對話框，進行編輯，我們以最簡單的方式，取程式預設值即可，無須再行編輯，即選擇 Quit，再點選 OK。

圖 9　編輯對話框

第三步驟點選 Analyze 選項，再次開啟 exam.tg，出現圖 10 的分析選項對話框，進行分析，我們仍以最簡單的方式，取程式預設值，點選 OK 即可。

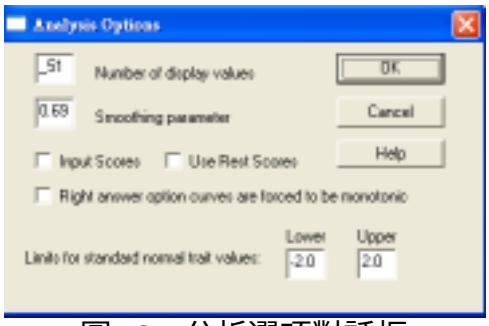

圖 10　分析選項對話框

第四步驟，點選 Display 選項，再次開啟 exam.tg，出現圖 11 的展示對話框，進行分析結果展示，選擇 Options，點選 OK，則輸出選項特徵曲線。在圖 11 的展示對話框尚可選擇各項試題反應理論的圖形或數值，本文省略，若有興趣，可詳見操作手冊（Ramsay，2000），實例可詳閱楊志強（民 93）之研究論文。

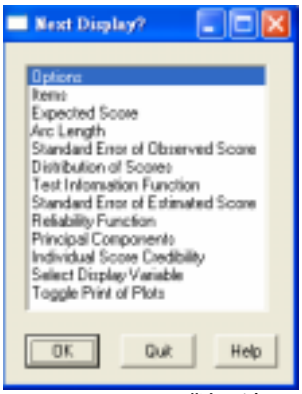

圖 11　展示對話框

在出現圖 11 的同時，也會出現圖 12 的選題對話框，在圖 12 中可點選 Next，

觀看下一題的選項特徵曲線；點選 Previous，可觀看上一題的選項特徵曲線；點選 Display Variable，可選擇選項特徵曲線的 *Y* 軸資料是原始分數（score）或轉換成潛藏能力（latent trait）；點選 Print，會進行列印功能。

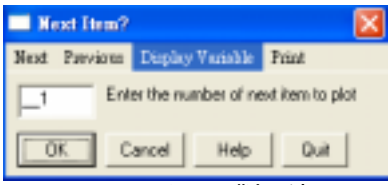

圖 12　選題對話框

伍、結語

透過客觀量化的試題分析，分析現有學生測驗資料，提升教師自編成就測驗試題的品質，並能藉以蒐集優良試題，建立真正具有科學數據的題庫系統，達到教學評量的目的與價值。同時，*TestGraf 98* 軟體僅是在網際網路中無數分享軟體中的其中一種，兼具學術研究及實務功能，期望教師也都能廣泛使用資訊科技與網路資源分享，降低教育投資的成本，提升教育服務的品質。

參考文獻

王文中等（民 88）。**教育測驗與評量－教室學習觀點**。台北：五南。

王寶墉（民 84）。**現代測驗理論**。台北：心理。

余民寧（民 86）。**成就測驗的編製原理**。台北：心理。

吳慧民（民 90）。**Kernel smoothing 估計選項特徵曲線理論之改進與應用**。國立台中師範學院教育測驗與統計研究所碩士論文。

林清芳（民 92）。**九年一貫課程自然與科技領域評量工具之發展**。國立台中師範學院教育測驗與統計研究所碩士論文。

郭生玉（民 74）。**心理與教育測驗**。台北：精華。

陳英豪、吳裕益（民 81）。**測驗與評量，第四版**。高雄：復文。

楊志強、楊志堅（民 92）。**選項特徵曲線在科學教育評量之應用**。應用教學科技於科學教育學術研討會，國立嘉義大學。

楊志強（民 93）。**技職校院學生科技專業能力指標縱貫研究（2/3）**。九十二年度國科會專題研究結案報告，NSC 92-2516-S-240-001。

American psychological association (1997). *Standards for educational and psychological testing, revised ed*. Washington, DC: American Psychological Association.

Beyer, A. (1994). Assessing students' performance using observations, reflections, and other methods. In N.L. Webb and A.F. Coxford (Eds.), *Assessment in the mathematics classroom* (pp. 111-120). Reston, VA: NCTM, Inc.

Clark, G.M. (1998). *Assessment for transitions planning*. Austin, TX: PRO-ED, Inc.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. CA: Sage Publications, Inc.

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

Fowler, F. J. (1995). *Improving survey question*. CA: Sage Publications, Inc.

Gonzalez,E.J.& Smith,T.A.(Ed.) (1997). *User Guide for the TIMSS International Database*. Chestnut Hill, MA:Boston College.

Mondy, R. W., Noe, R. M. & Premeaux, S. R. (2002). *Human resource management, 8th ed*. New Jersey: Prentice Hall Inc.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, pp. 611-630.

Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. Quebec, Canada: McGill University.

Robertson, I.T. & Iles, P.A. (1988). Approaches to managerial selection. In C.L. Cooper & I.T. Robertson (Eds.), *International review of industrial and organizational psychology*, (pp.159-211). New York: John Wiley.

Santor, D. A. & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, 10, pp. 345-359.

Santor, D. A., Zuroff, D. C. Cervates, P. Palacios, J, & Ramsay, J. O. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment*, 7, pp. 131-139.

Santor D. A., Ramsay, J. O. & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, 6, pp. 255-270.

Schielack, J.F. & Chancellor, D. (1994). From multiple choice to action and voice: teachers designing a change in assessment for mathematics in Grade 1. In N.L. Webb and A.F. Coxford (Eds.), *Assessment in the mathematics classroom* (pp. 121-129). Reston, VA: NCTM, Inc.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education, 6th ed*. New Jersey: Prentice Hall Inc.