

第二版

教育測驗與評量

— 教室學習觀點

王文中 呂金燮 吳毓瑩 張郁雯 張淑慧 ◎合著

Educational assessment :
A classroom learning perspective



第二版

教育測驗與評量

——教室學習觀點

王文中、呂金燮、吳毓瑩、張郁雯、張淑慧 合著

（依姓氏筆劃排序）

五南圖書出版公司 印行

國家圖書館出版品預行編目資料

教育測驗與評量：教室學習觀點/王文中等
合著.--二版.--臺北市：五南，2004 [民93]

面：公分

含索引

ISBN 978-957-11-3715-5 (平裝)

1.教學法

521.3

93015033



11BB

教育測驗與評量—教室學習觀點

作 者 — 王文中 呂金燮 吳毓瑩
張郁雯 張淑慧(5.2)

發行人 — 楊榮川

總編輯 — 龐君豪

主 編 — 陳念祖

責任編輯 — 蔣和平

出版者 — 五南圖書出版股份有限公司

地 址：106台北市大安區和平東路二段339號4樓

電 話：(02)2705-5066 傳 真：(02)2706-6100

網 址：<http://www.wunan.com.tw>

電子郵件：wunan@wunan.com.tw

劃撥帳號：01068953

戶 名：五南圖書出版股份有限公司

台中市駐區辦公室/台中市市區中山路6號

電 話：(04)2223-0891 傳 真：(04)2223-3549

高雄市駐區辦公室/高雄市新興區中山一路290號

電 話：(07)2358-702 傳 真：(07)2350-236

法律顧問 元貞聯合法律事務所 張澤平律師

出版日期 1999年9月初版一刷

2004年3月初版十一刷

2004年9月二版一刷

2011年10月二版七刷

定 價 新臺幣600元

二版序

20 世紀最大的教育革命莫過於教育測驗的改革。20 世紀初對於心智遲緩兒童的研究啟開了現代心理測驗的序幕。隨著家長與一般社會人士對辦學績效日漸關心，教育測驗與評量就成為校長、老師必須面對的問題。透過測驗與評量，向大眾交代辦學效果。為此，必須對大量的學童進行施測。選擇題也就因應而生，並成為最好的題型。電腦與讀卡機的問世，更令選擇題變為大規模教育測驗的主流，例如性向測驗、成就測驗、基本學力測驗、聯考、國家考試等。

到了 20 世紀末，逐漸有學者對於選擇題提出批判。他們認為考題形式會影響學習，間接影響教學。選擇題會鼓勵考生死記零碎的事實，而輕忽觀念的理解和問題解決。這個反思，帶動了建構反應式的評量，如寫作、申論題、證明題、計算題、以及其他實作題，如操作儀器、製作企畫等。這類評量雖然表面上遠比選擇題必須付出更大的經濟成本，例如評分的花費，不過如能因而激發學生深度與真正的學習，那麼所得到的報酬應該遠比這些花費大得多了。

在 21 世紀的今天，教育測驗與評量的目的不再侷限於早期的辦學績效評鑑，而變為增進學生個人的學習效果，達成教育目的。簡單的說，就是從機構本位變為學生個人本位。同樣的，評量的方式也是由原先的單一的紙筆測驗（尤其是選擇題），變為多元評量，例如建構反應評量方式、實作評量，甚至利用非正式的口語溝通進行評量。

為了幫助學生學習，教師必須隨時掌握其學習缺失，據以調整教學方法，因此評量必須持續進行，而不是在學習告一段落後，才進行綜合性的評量（如期末考）。透過這種動態評量，可以追蹤學習的軌

跡，調整教學，達到最大的學習效果。換言之，評量不再是靜態的，而是動態評量。

本書建立在這樣的評量思潮上，期能使讀者具體理解各種評量的方法，並靈活使用於課堂的教學情境中。我們先介紹測驗與評量的理論基礎，接著說明各種評量方式的適用時機與具體實施步驟，最後則是評量結果的解釋與應用。

自從 1999 年第一版上市以來，本書廣受歡迎，也得到許多批評和建議，繼而在 5 年後第二版更新出書。新版中的每個章節後面都新增思考問題，以利讀者反思運用各章的重要觀念，新知補給站則增添一些熱門的測驗與評量議題的討論。此外，為了讓讀者更容易將書本的概念運用於實務，新增附錄說明如何以 EXCEL 分析本書相關的統計問題。本書可供大學生修習教育測驗與評量、教育與心理測驗、心理測驗等課程之用，也可當作各級學校教師進行評量，以及一般工商業界進行人力資源管理之參考。

一版序

20 世紀最大的教育革命莫過於教育測驗的改革。二十世紀初對於心智遲緩兒童的研究啟開了現代心理測驗的序幕。隨著家長與一般社會人士對辦學績效日漸關心，教育測驗與評量就成為校長、老師必須面對的問題。透過測驗與評量，向大眾交代辦學效果。為此，必須對大量的學童進行施測。選擇題也就因應而生，並成為最好的題型。電腦與讀卡機的問世，更令選擇題變為大規模教育測驗的主流，例如性向測驗、成就測驗、基本學力測驗、聯考、國家考試等。

到了 20 世紀末，逐漸有學者對於選擇題提出批判。他們認為考題形式會影響學習，間接影響教學。選擇題會鼓勵考生死記零碎的事實，而輕忽觀念的理解和問題解決。這個反思，帶動了建構反應式的評量，如申論題、證明題、計算題，以及其他實作題，如操作儀器、製作企劃等。這類評量雖然表面上遠比選擇題必須付出更大的經濟成本，例如評分的花費，不過如能因而激發學生深度與真正的學習，那麼所得到的報酬應該遠比這些花費大得多了。

至此，教育測驗與評量的目的不再侷限於早期的辦學績評鑑效，而變為增進學生個人的學習效果，達成教育目的。簡單的說，就是從機構本位變為學生個人本位。同樣的，評量的方式也是由原先的單一的紙筆測驗（尤其是選擇題），變為多元評量，例如建構反應評量方式、實作評量，甚至利用非正式的口語溝通進行評量。

為了幫助學生學習，教師必須隨時掌握其學習缺失，據以調整教學方法，因此評量必須持續進行，而不是在學習告一段落後，才進行綜合性的評量（如期末考）。透過這種動態評量，可以追蹤學習的軌

跡，調整教學，達到最大的學習效果。換言之，評量不再是靜態的，而是動態評量。

本書建立在這樣的評量思潮上，期能使讀者具體理解各種評量的方法，並靈活使用於課堂的教學情境中。我們先介紹測驗與評量的理論基礎，接著說明各種評量方式的適用時機與具體實施步驟，最後則是評量結果的解釋與應用。本書可供大學生修習教育測驗與評量、教育與心理測驗、心理測驗等課程之用，也可當作各級學校教師進行評量，以及一般工商業界進行人力資源管理之參考。



二版序

一版序

第一章 測驗與評量的意義與趨勢 001

第一節 測驗與評量的昨日與今日 / 003

第二節 測驗與評量的關係 / 006

第三節 測量與量尺 / 008

第四節 測驗與評量的道德議題 / 012

第五節 評量的趨勢 / 016

第六節 本書的架構 / 020

第二章 評量與教學目標的連結 027

第一節 教學評量的傳統包袱 / 031

第二節 評量與教學連結的基礎 / 036

第三節 連結評量與教學的階段 / 048

第四節 結語 / 053

第三章 效度 057

第一節 效度的涵義 / 059

第二節 效度的性質 / 063

第三節 效度在內容代表性上的證據 / 068

第四節 效度在效標關聯上的證據 / 072

第五節 建構效度作為效度概念的核心 / 078

第六節 效度概念的綜合考量 / 085

第七節 影響效度的因素 / 088

第八節 結 語／089

第四章 信 度 095

第一節 信度的意義／097

第三節 信度評鑑／119

第四節 效標參照評量的信度估計／123

第五節 測量標準誤／125

第六節 信度概念在教學評量的應用／127

第七節 結 語／128

第五章 發展評量計畫 131

第一節 學習成果的分類／133

第二節 評量的方式／136

第三節 試題的類型／138

第四節 評量方式與學習成果的結合／140

第五節 建立課堂測驗細目表／145

第六節 選擇評量方式／147

第七節 結 語／150

第六章 選擇性題目與題組的編製 157

第一節 選擇題的性質／159

第二節 選擇題的優點／163

第三節 選擇題的缺點／167

第四節 編寫選擇題的注意事項／170

第五節 題組的性質／182

第六節 題組的優點、限制、與編製建議／188

第七節 其他議題／191

第八節 結 語／196

第七章 建構性題目的編製	201
第一節 建構性題目的適用時機／	204
第二節 建構性題目的題型／	204
第三節 建構性題目的優點與限制／	216
第四節 建構性題目的出題建議／	217
第五節 建構性題目的計分建議／	217
第六節 結語／	218
第八章 實作評量——理論	221
第一節 發展背景／	225
第二節 實作評量的本質／	229
第三節 實作評量的類別／	235
第四節 國內實作評量之應用研究／	246
第五節 實作評量應用上的優點／	257
第六節 實作評量應用上的限制／	260
第七節 結語／	264
第九章 實作評量——應用	269
第一節 決定評量的標的／	271
第二節 設計題目／	274
第三節 評分規準的設計與記錄的方式／	278
第四節 評量結果的分析／	291
第五節 優良實作評量的評鑑標準／	292
第十章 檔案評量	297
第一節 檔案評量是什麼／	299
第二節 檔案評量的理論基礎——情境／社會歷史觀點的 認知與學習理論／	307

第三節	如何實施檔案評量／	309
第四節	對檔案評量的評鑑／	317
第五節	結語／	319
第十一章	情意評量.....	327
第一節	情意評量的原則與範疇／	331
第二節	自我陳述式的評量／	334
第三節	同儕評量／	337
第四節	軼事錄／	345
第五節	心智習性的評量／	349
第六節	結語／	353
第十二章	題目分析.....	357
第一節	題目分析的重要性／	359
第二節	主觀的題目評鑑／	359
第三節	客觀的題目分析／	361
第四節	測驗總分分配形式／	373
第五節	結語／	376
第十三章	評量結果解釋	379
第一節	效標參照與常模參照的分數解釋／	382
第二節	常模／	384
第三節	發展性常模／	385
第四節	組內常模／	388
第五節	常模的評鑑／	400
第六節	解釋測驗分數的守則／	402
第七節	結語／	403

第十四章	學期成績.....	405
第一節	學期成績的意義與功能／	407
第二節	學期成績所包含的內容／	409
第三節	學習成果評定如何考慮學習過程／	411
第四節	分數形式的期末成績／	414
第五節	等第形式的學期成績／	417
第六節	學期總評量的效度想法／	421
第七節	結語／	426
附錄一	自然科定期成績考查試題案例	429
附錄二	黃敏老師之「閱讀及寫作」學習檔案評分規準 ...	433
附錄三	利用 Excel 進行資料分析.....	435
人名索引		447
漢英索引		449
英漢索引		456

第一章

測驗與評量的意義與趨勢

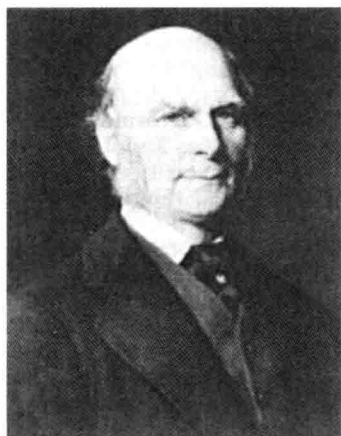
王文中

假設你是老師，一定想知道你的教學是否成功？教法或教材是否需要修改？學生有哪些學習困難？在哪些地方表現良好？如果你是學生，你會想知道自己在哪些地方已經精熟？哪些地方還需加強？比起其他同學你是否領先還是落後？和過去相比是進步還是退步？如果你是家長，你會關心兒女的表現。如果你是教育行政當局（如校長或其他教育行政人員），你會關心辦學的效果。如果你是社會的一般人士，你可能也會關心今天的學生是否已經習得必備的能力和正確的價值觀？和過去相比是進步還是退步？要回答這些問題，必須有系統、周延、效率的蒐集學生多方面的資料，客觀分析，作成結論，這必須仰賴測驗與評量。

第一節 測驗與評量的昨日與今日

測驗與評量有著悠久的歷史，中國人早在漢朝就發展出非常複雜和精緻的步驟來選取公職人士，隋唐時代的科舉制度更是一套完整的取士制度。西方則早在 13 世紀大學畢業考試就開始使用口試，不過廣泛實施筆試卻要等到紙張便宜後才開始。19 世紀起，人們對於測驗與評量的興趣蓬勃發展。英國和美國開始公職人員考試。精神醫學對於心理測驗的發展有著莫大的影響。法國醫生 J. E. D. Esquirol 研究心智遲緩和瘋子的差別，指出心智遲緩仍有程度之分，而且語言表達最能判別遲緩程度。另一位法國醫生 Seguin 也同樣的致力於心智遲緩的研究，並於 1837 年設立學校教育心智遲緩的兒童。後來，Alfred Binet 促使法國公共教育部改善心智遲緩兒童的處置，並促使智力測驗的問世。

英國生物學家 Francis Galton 爵士（1822～1911）可以算是現代心理測驗運動之父。他於 1884 年在國際博覽會中設置人體測量實驗室，測量視覺和聽覺的敏銳度、肌肉強度、反應時間和其他感覺動作的功能。展示結束後，該中心移至倫敦。Galton 爵士認為人類在這些感官測量上的差異，反映出他的智力程度。例如心智遲緩的人，比較無法



Francis Galton

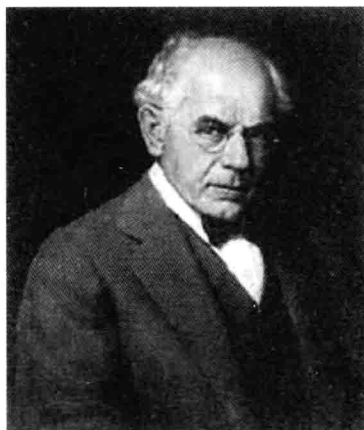


Karl Pearson

區分冷熱和疼痛。他還用評等表（rating scale）、問卷、自由聯想等技術的應用，也發展統計方法來分析個別差異，後來 Karl Pearson（1857～1936）延伸其方法，分析變項間的關聯，稱為 Pearson 積差相關。

1890 年，美國心理學家 Jams McKeen Cattell（1860～1944）在其論文裡，首次使用「心理測驗」（mental tests）這個名詞。這篇論文介紹用以評估大學生智力的測驗，內容包括肌肉強度、行動速度、痛覺敏感度、視覺和聽覺的敏銳度、重量區分感覺、反應時間、記憶力等。Cattell 於 1888 年在劍橋大學講學，期間與 Galton 爵士接觸並受其影響，因此其所謂的心理測驗，仍有著濃厚的 Galton 色彩：用感覺動作與反應時間來測量智力。後來這類的測驗被使用到學童、大學生和成人中。不過這類的測驗的成效卻大受質疑，因為各種測驗間的結果並不一致，而且和教師對學童智力狀況的評量，學童的學業表現的關聯性也不高。

對這類測驗的質疑引發 Alfred Binet（1857～1911）等法國心理學家對智力測驗的反思與革命。他們認為當時的心理測驗只強調感官知



James McKeen Cattell



Alfred Binet

覺，以及一些單一和專門的能力，忽略高階的心理功能。所以提出各樣的測驗，用以評估記憶力、想像力、注意力、理解力、美感等。1905年，Alfred Binet 和 Theodore Simon（1873～1961）發表著名的比西量表（Binet & Simon Scale）。該測驗涵蓋判斷、理解、推理等能力，這才是他們所認為的智力表現（詳見 Anastasi & Urbina, 1997）。

早期的心理測驗都採個別施測。到了第一次世界大戰，由於大量徵召軍人，因此需要快速的分派這些軍人到適當的職務，也就促使團體測驗的盛行。現今的選擇題測驗就是當時的產物。在第一次世界大戰後，測驗運動開始轉型。首先是技術的突破，尤其是電腦的發明，配合掃描器，計分可以變得既快速又正確。電腦也使得施測方式有了改變，例如增加三度空間、色彩與音效。電腦還可以隨著受試者作答情形，選擇適當難度的試題給受試者作答。例如受試者這一題答錯的話，下一題就會變得比較簡單。反之，如果這一題答對的話，下一題就變得比較難。這就是所謂的電腦適性測驗（computerized adaptive testing）。現今一些大型考試，如 TOEFL（Test of English as a Foreign Language）、GRE（Graduate Record Examinations）、GMAT（Graduate