

教 育 論 壇

主題：實作評量與案卷評量

實作評量的應許、難題和挑戰

引言人：盧雪梅

台北市立師範學院國民教育研究所副教授

壹、緒論

這次教育論壇的主題—「實作評量和案卷評量」(performance assessment and portfolio assessment)，又稱做「另類評量」(alternative assessment)，正是美國九〇年代教育界和教育測量界的熱門議題。當我還是博士班學生的時候，從近幾年來美國教育測量或心理計量界就業市場的資訊中，就很強烈感受到這股另類評量潮流的強度和熱度。事實上，實作評量稱不上是新的評量方式，近來之所以受到注意的主要原因是：美國教育界開始廣泛將這種評量方法運用在學校學科領域中，並且做為評估學校教學績效和瞭解學生成就水準一般狀況的策略，例如，和我國有過合作經驗的馬里蘭州即為一例。另類評量的方法受到歡迎的主要原因之一是：一般人對於選擇式的測驗題感到不滿意，例如，選擇式的測驗題只能測量學生「知道」什麼但無法測量學生「能做」什麼；此外，認為以選擇題為主的標準化測驗對於教師的教學和學生的學習造成一些偏差的影響。標準化測驗經常是各學區(school districts)或各州用來做為評估學校教學績效的方式，在績效的壓力之下，造成部分教師教學側重於測驗的內容，而扭曲了教學的面貌和窄化學生的學習結果。在教育改革的推動上，美國教育界期望藉著評量的革新來提升教師教學的品質和學生學習的成就。此外，一些入學考試和證照考試也在原有的選擇題之外，增加建構反應題或實作測驗部分，實作評量在美國教育界和測驗界已是一個非常顯明的趨勢。

評量的革新也是國內教改的重點之一，實作評量(含案卷評量)的精神和方法具有多元化的特色，相信實作評量在國內也將會推展開來。事實上，目前國民中小學學生成績考查辦法中所列定的方法，絕大多數都是屬於實作評量的方式，如何落實多元化的評量，才是當務之急。

貳、甚麼是「實作評量」和「案卷評量」？

何謂「實作評量」？簡而言之，以觀察和專業判斷來評量學生學習成就的評量方式都可以稱為實作評量（Stiggins, 1987），其型式非常的多元化，例如建構反應題、書面報告、作文、演說、操作、實驗、資料蒐集、作品表示等，都是實作評量的例子，案卷評量也是實作評量的一種型式。實作評量具有下列幾點特徵（Herman, Aschbacher, & Winters, 1990）：(1)要求學生執行或製作一些需要高層思考或問題解決技能的事或物；(2)評量的作業（tasks）是有意義性、挑戰性且與教學活動相結合；(3)評量的作業能與真實生活產生關聯；(4)歷程（process）和作品（product）通常是評量的重點；(5)表現的規準（criteria）和標準（standards）——也就是評量的重要層面與給分標準，要事先確定。實作評量有時也被稱為真實性評量（authentic assessment）。

案卷評量通常也被視為實作評量的一種型式，由於它的實施方式具有獨自的特色，所以「案卷評量」一詞常被單獨使用。案卷（portfolios）簡單的說，就是學生的作品集，學生每一個人都有一個屬於自己的資料夾，有目的地蒐集個人在某個（些）領域的作品，這作品集以說故事的方式傳達出學生的努力、進步和成就（Arter & Spandel, 1992）。案卷夾中蒐集的項目除了學生的作品外，還包括：(1)案卷夾的內容一先經過師生共同討論與決定的；(2)作品挑選的指導說明書；(3)作品判斷的要點；(4)學生的自我評鑑。案卷評量結合教學和評量，重視學生學習的歷程和學生對學習的自我反省，是一種相當能適應學生個別差異的評量方式。

參、實作評量的優點

實作評量有甚麼優點呢？理論上，實作評量可以讓教師瞭解學生對問題瞭解程度、投入程度、解決的技能和表達自我的能力，能夠較完整的反映出學生的學習結果；因為實作評量與真實生活較為相近，其支持者認為實作評量能夠增進學生學習的動機、提高學生參與和投入的程度、幫助學生建構有意義的學習情境、發展問題解決能力、批判性思考和表達自我的能力。有時候實作評量也可以做為一種教學策略，提高學生的學習興趣和學習結果。總之，實作評量的精神和方式對於教和學均能提供較完整的回饋訊息，有助於促成教師教學品質和學生學習成就的提升。然而，實作評量所帶來的應許並不是自動地產生，而是要經過嚴謹的計畫、實施和運用，實作評量的功能才能發揮出來。

肆、實作評量的難題

實作評量雖然有很多的優點，實施起來也有不少困難，下面幾點是最常見的難題：

- (一)時間上的難題：實作評量在實施上和記分上所需的時間比較多。
- (二)經費和設備上的難題：實作評量的花費通常比紙筆測驗來的多；有時需要購置一些器材或儀器，在空間的需求和器材的保管維護上也可能遇到問題。
- (三)評分上的難題：實作評量除了需要時間和人力去計分外，評量和觀察重點的掌握和評分標準的訂定有時也是一個難題，尤其是對非結構性的作業項目進行評量。例如，最近ACT對二百多名中學教師進行案卷評量實施困難的調查報告中指出，大多數教師認為時間和評分是他們進行案卷評量感到最困擾的兩個問題（Wolfe & Miller, 1997）。
- (四)技術品質上的難題：最主要的是評量結果的信度和效度。例如，在信度方面

，評分者間評分的一致性通常不高；在效度方面，由於實作評量的實施通常需較多的時間，因此作業項目通常很少，甚至只有一項，以極少數的行為樣本是否能適當推論學生的學習結果的全貌，令人疑慮。這方面的考量常是實作評量做為評估學校績效的手段的可行性和適切性，爭議最多的地方。

伍、實作評量的設計

實作評量適用的學科範圍非常的廣，例如，聽、說、讀、寫、語文、數學、自然科和社會科等，都可以採用實作評量；評量的方法也非常多樣化，教師可以根據教學目標和評量的目的，編擬適用的評量計劃。下列是擬定實作評量計劃時需考慮到的幾個重要層面：

(一)評量的目的是什麼？必先確定目的後，才可以開始進行評量。

(二)要評量的行為表現是什麼？從下列三方面加以界定：

- 1.重要的學習內容和技能。
- 2.行為表現的性質：觀察重點是著重於「歷程」(process)或是「作品」(product)，或是此二者兼顧？
- 3.判斷的規準(criteria)和標準(standard)：明確列出行為表現的重要層面和各層面表現的評分標準，例如怎樣的表現是優異的、普通的或是不佳的。此項是實作評量品質關鍵的所在，教師要事先和學生就這一項進行溝通，給學生回饋時也是從這點切入。

(三)設計作業—如何蒐集資料或學生表現的憑據？

- 1.選擇資料蒐集的型式：要設計特定的作業來引發學生的表現行為，或者以觀察教室中自然發生的事件來做為評量的依據？如果能夠有系統和客觀地觀察和記錄學生的表現，教室觀察是一種經濟且有效的評量方法。
- 2.決定要資料蒐集的數量：一次蒐集一個行為樣本、一次蒐集多個行為樣本、或者多次蒐集多個樣本？

(四)設計評分計劃

- 1.計分的型式：採整體性計分或分析性計分？
- 2.評分者人選：由教師或專業人士來評分，或由學生自評或同儕互評？無論評分者人選是誰，都應先接受過評分訓練。
- 3.記錄方式：檢核表、評定量表、軼事紀錄、作品集。

Popham (1995) 根據實作評量的特性和限制，提出七點在設計或選擇實作評量時應考慮的要項：

- (一)類推性 (Generalizability)：學生在這個作業項目上的表現可以類推到其他相似項目上的程度如何？
- (二)真實性 (Authenticity)：這個作業項目是否和學生在真實生活中遇到的情境相類似？
- (三)多元焦點 (Multiple foci)：這個作業項目測量的是多元的而非單一的教學結果嗎？
- (四)可授性 (Teachability)：學生在這個作業項目的能力的提升是否是教學所致的？
- (五)公平性 (Fairness)：這個作業項目對所有學生都公平的嗎？
- (六)可行性 (Feasibility)：在費用、空間、時間和設備的考量上，這個作業項目

的可行性如何？

(七)可評性 (Scorability)：學生在這個作業項目上的反應之評定結果的可信度和準確度如何？

陸、啓示和挑戰

實作評量的浪潮在美國各地、各州快速流行，許多學者對於這種景象，提出了警告和呼籲，提醒實作評量的推展並不能單建立在信念上，而是要建立在實徵的證據上 (e.g., Dunbar et al., 1991)。美國測驗界和教育界在實作評量的運用和相關的研究投入了非常多的努力，無論在實務推展、理念和經驗的交流、心理計量技術的研究和發展、以及實徵研究結果的累積等方面都有相當的成果。實作評量是否如預期中促成教師教學品質和學生學習結果提升，有些調查指出教師和學生認為實作評量對教和學都有積極的影響 (e.g., Koretz et al., 1993)，然實作評量長期的效果和影響還有待更多的觀察。雖然國內測驗評量的背景環境與美國不同，然而對於評量的革新的需求是相似的，推展實作評量面臨的挑戰是非常多的。美國在這方面的經驗有一些可以做為我們的借鏡和參考。

(一)建立正確的評量觀，落實多元化的評量

美國實作評量的興起，雖然是針對客觀式測驗題，但並沒有完全否定客觀式測驗題的功能和特性，Reckase (1997) 在一篇比較標準化測驗和案卷集測量的建構 (constructs) 的研究中，結論提到「...很難說那一種方法優於那一種，只能它們測量到的層面不同，要使用那一種方法，端視評量的目的和需要的訊息的類型而定」，這也是許多研究者一致的結論。教學目標是多元化的，評量目的和方法基本上也是多元的，評量者應熟悉各種評量方法的優點和限制，配合教學目標，妥善靈活運用各種評量方式，才是正確的評量觀。和客觀式試題相比，實作評量雖是非常不經濟的評量方法，但不能為了運用上的方便或簡單的理由，把評量的層面給窄化或偏狹，評量者應配合教學目標選擇和其目的相合的評量方式和工具，以落實多元化的評量。

(二)加強教師評量的專業知能

教師是班級教學和評量的主導者，在評量多元化的時代裡，教師評量的專業知能越顯重要。美國推展實作評量的經驗也強調教師訓練的重要性，有研究者指出要教師運用一種新的評量方式之前，必須裝備他們在新評量方法的技能和信心，假如教師沒有技能，班級評量可能被誤導或沒有功能，假如教師沒有自信心，甚麼也不可能發生。八十三學年的第二學期初台北市教育局曾以行政命令規定，全市小學採用紙筆測驗以外之方式辦理第一次定期考查。結果引起一股評量風波，後來，教育局取消這項規定。這種結局並不令人感到意外，因為老師們在知能上和心理上都沒有裝備好。事實上，國民中小學學生成績考查辦法中所列定的方法，絕大多數都是屬於實作評量的方法，我們要落實多元化的評量之前，首要之務，除了建立教師正確的評量觀，還要提升教師的專業知能和信心，同樣的，也要加強職前教師在這方面的知能和訓練。此外，建立教學評量的諮詢和支援的系統，提供教師相關的服務和協助，也是推動多元化評量的重要一環。

(三)發揮團隊合作的精神

美國推展實作評量的經驗也非常強調團隊合作精神的重要性，例如，鼓勵教

師、課程專家、心理學家和測驗專家共同合作來研發實作評量的作業；鼓勵教師之間、學校之間做經驗的分享與交流；強調學校與社區觀念溝通與合作的重要性，以及行政人員對教師支持的重要性。這些觀念都非常寶貴，值得我們借鏡。

(四)呼籲更多的評量實徵研究

雖然評量活動是教學歷程中很重要的一環，然而在教育方面的研究中，教學評量的相關實徵研究並不多見。希望有更多的學者、研究生和第一線的教師能投入這一個重要且實際的研究領域，期盼透過研究來瞭解目前教學評量的現況、各種評量方面的功能、教學評量重要課題和研究改進的方法，並累積評量相關研究的成果。另外，教育部預定從八十七學年度起將建立我國中小學生的基本學力指標並進行測驗，是否會以實作評量的型式進行，尚未得而知。雖然籌備的時間非常匆促，但這將成為國內有史以來的第一次規模評量（largescale assessment），也將為國內教育測驗與評量帶入一個新的紀元，願審慎走出第一步。

參考文獻

- 陳文典、陳義勳、李虎雄、簡茂發（民84）。由馬里蘭州的學習成就評量與其在臺灣的試測結果看實作評量的功能和運用，科學教育，185, 2-11。
- Arter, J.A., & Spandel, V.(1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, 11(1), 36-44.
- Dunbar, S.B., Koretz, D.M., & Hoover, H.D.(1991). Quality control control in the development and use performance assessment. *Applied Measurement in Educational*, 4(4), 289-303.
- Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stechetr, B.(1993). The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program. Los Angeles: RAND Institute on Educational and Training/CRESS.
- Herman, J.L., Aschbacher, P.R., & Winters, L.(1990, November). *Issues in developing alternative assessments*. Paper presented at the annual meeting of the California Educational Research Association, Chicago.
- Mehrens, W.A.(1992). Using performance assessment for accountability purpose. *Educational Measurement: Issues and Practice*, 11(1), 3-9.
- Popham, W.J.(1995). *Classroom Assessment-What Teachers Need to Know*. Needham Heights, MA: Allyn & Bacon.
- Reckase, M.D.(1997, March). *Constructs assessed by portfolios: how do they differ from those assessed by educational test*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Stiggins, R.(1987). Design and development of performance assessment.*Educational Measurement: Issues and Practice*, 6(3), 33-42.
- Wolfe, E.W., & Miller, T.R.(1997). Barriers to the implementation of the portfolio assessment in secondary education. *Applied Measurement in Educational*, 10(3), 235-251.