科學學習的評量理念

張惠博*　黃文吟**
*國立彰化師範大學　物理系
**國立彰化高級中學

一、考試與教學的關係，敵乎？友乎？

「考試領導教學」雖是老生常談，可是，這樣的現象究竟是好是壞，是值得教育工作者重新審慎思考的問題。其實，評量對於教學的影響，可能是正面的，亦可能是負面的。再好的課程、教材、教材內容，倘若缺乏合適的評量方式與內容，即無法引導學生學到他（她）們所應該知道的。例如：教師過度使用選擇題，則學生對於實作或動手解決問題的學習，即易於失去興趣與動機。有鑑於傳統評量不足以測出學生真正的學習成果，採用另類的評量方式，可以彌補傳統評量方式之不足，亦可提昇所有與評量有關之關係人之間的良性互動，這是提昇教學與學習成效的重要課題。本文即引介幾種另類的評量方式，包括：實作評量 (performance-based assessment)、真實性評量 (authentic assessment)、歷程檔案 (portfolios) 等等，介紹其基本理念與特性，並藉以反省科學課程改革之時，評量的觀念與作法亦必須相攜並進之處。

二、評量觀念的演變

近年來，由於建構主義派典的興起，有關成就測驗/評量的理念，已從測驗文化 (testing culture) 轉移爲評量文化 (assessment culture) (Birenbaum & Dochy, 1996)。這兩種取向的主要差異，在於前者較重視概念學習的成果／產出 (product)，測驗的規劃與施行過程中，受試者總是處於被動以及無選擇權力的情勢下接受測驗。而近代的評量取向則重視於評量與學習過程密切結合，在進行評量的過程中，被評量者可扮演主動參與的角色，並在過程中分享權力。

根據 Birenbaum (1996) 的說法，成就測驗/評量，可依其取向之不同，分爲傳統的標準測驗 (traditional standardized test) 與另類評量 (alternative assessment)，二者又各包括許多不同的測驗/評量形式，如傳統的標準測驗有標準參照測驗與常模參照測驗，而另類評量則是許多評量方式的總稱，其中較爲常見的如：檔案評量、實作評量、真確評量、直接評量 (direct assessment)、與建構式的評量 (constructive assessment) 等。

傳統的標準化測驗受到兩個基本理念的影響頗深，其一是把教學視爲概念傳遞的活動，

也因此認爲知識可經由重覆練習或背誦而得。另一方面，又特別崇尚心理測量 (psychometric) 傳統下對客觀、公平與推廣性的要求。例如，以色列學者 Tamir (1990) 即曾指出，選擇題的問題，雖遭受許多負面的批評，然而，其仍被廣泛的使用，特別是在美國，理由如下：

1.在短的測驗時間裡，容許涵蓋較廣泛範圍的題材。

2.它可用來評量不同階層的學習。

3.在計分上而言，是比較客觀的，也因此，信度較高。

4.它比較容易與快速評分，也可使用機器評分。

5.它可避免學生雖然懂得學科內容，卻因不擅書寫，而遭受評分上的不利。

6.它較可用來進行項目分析，以了解那些題目太容易，或過於艱難，甚或是題意不清。

然而，傳統式（以選擇題或測驗題）測驗，固然有省時、省錢的優點，但也因此使學習容易流於記憶，學生缺乏面對真實情境的解決問題之經驗與學習等缺點。而且，這種重視概念學習的結果／產出，以及爲達客觀與公平要求而予標準化的測量模式，其測驗結果往往順理成章的成爲據以排序學生學習或教師教學成效的重要（或唯一）資料來源。然而，這樣的測量模式雖可達到高信度水準，但因其標準化的結果，往往忽略了學生在種族、語言、認知偏好等方面的差異 (Birenbaum, 1996)，而且，無法反映教室裡真實的教學實況。

三、另類評量的意涵與種類

另類評量講求的則是多元的、真確的反映學生的學習實況。在智慧與價值多元化的資訊時代，教育的目標也趨向多元發展。另類評量的特性在於與教學過程密切結合，評量基準與方式的多元化，尊重學生爲一能自我調節 (self-regulation) 的學習者，因此，評量過程中賦予學生選擇的權力，並鼓勵討論協商的合作方式以評估學生的表現。資料蒐集的方法，則除了傳統的紙筆測驗外，亦可使用觀察 (observation)、晤談 (interview)、問卷 (questionnaire)、使用日記或記錄 (diary/record)、實地測試 (field or pilot test) 或放聲想 (think-aloud) 等等方法。評量時，以學生爲中心，持續且自然地觀察其是否能建構出良好的解決問題的方案？能否提出多元的觀點？以及是否能合理化他們的構思？教師亦可以知道學生在學習過程中所付出的努力，以及，進步的情形和達成多少學習目標等等。教師倘有意進行有別於傳統的評量，最爲有效的方法是從其教學改變做起。根據 Shavelson 與 Baxter (1992) 的說法，一位以動手操作爲主，或是探究教學爲主的教師，其進行實作評量的經驗與能力亦較佳。這樣的教師比起以講述法爲主的教師，對於實作評量的命題與實施較能得心應手。然而，不同於傳統測驗甚至可以機器代勞的計分方式，另類評量需仰賴人力決定學生的學習成果究竟爲何，因此，另類評量方式通常是耗廢時間與人力的。此外，

有關的幾種不同另類評量形式之內涵與特性，分別說明如後：

檔案評量：足以展現學生學習過程與進步的任何形式的資料，例如：書面或是利用錄音、錄影等視聽媒體的形式等等，都可成爲 portfolios 的內容物 (Popham, 1995)。在建立檔案的過程中，學生有選擇其所欲包含的內容物之自由，甚至可以是他人的作品，亦或跨學科的。但重要的是，學生必須展現出對每個內容物的自我反省與統整能力。檔案評量因此具有真確的、機動的、持續的、多元的、互動的與豐富的等特性。評斷學生學習檔案的優劣時，除了各項資料個別的評鑑之外，也需要特別留意整體協調連貫性的展現，Birenbaum 並提供評量整體面向的判準 (參考 Birenbaum, 1996)。

實作評量：是藉由學生完成某特定作業 (task)，來評量其學習狀況的一種途徑。這樣的評量方式，能帶動學生視學習爲自己的責任，而不是老師的，因此，學生能投入更多的心力來學習。但是，當指定的課業超出學生能力所及，也將使學生放棄完成任務。另一方面需注意的是，實作評量通常涵蓋較少的作業，容易造成爲測驗而訓練 (coach to the test) 的情況，因而最易影響其結果效度 (consequential validity)。所謂結果效度是指評量結果的推論與使用之有效程度 (Messick, 1989；Popham, 1995)，例如，教師倘利用大多數的時間教授測驗的內容與概念，而甚少提及測驗之外的問題，這就涉及結果效度的問題 (Linn, Baker, & Dunbar, 1991)。易言之，有些教學，因著重於學生對於測驗內容的背誦與記憶，對於學生的思考與解決問題能力的發展，並沒有助益。因此，即使學生考得高分，也不能說學生獲得學習，這樣的評量，就缺乏結果效度。所以，實作評量也有可能遭致傳統測驗的命運，學生是因爲記憶或熟練題目而獲得高分，並不是獲得學習的成就，即所謂的測驗污染 (test pollution)。爲避免此一情況發生，評量者要特別注意過程的評量，以及學生對整個學習進程的自我反省。爲了解決此一問題，Shavelson 與 Baxter (1992) 即主張應進行課程嵌入的評量 (Curriculum-Embeded Assessments)，亦即，評量是與課程的進行搭配的，評量不是偶一爲之。爲了獲得對於學生較真確的了解，甚至可進行長達一年期的評量，以獲得較大的樣本的評量。如此，應可獲得較有效度的評量。

事實上，Shavelson, Carey, 與 Webb (1990) 即曾指出，不少科學家都會同意，倘意欲測量學生真正的了解，最好的方式，應是將學生置於實驗情境，提供問題，再由學生利用實驗室的資源去解決問題。科學家會認爲，在教學中，應教導學生有關探究的技能，這是在傳統紙筆測驗中所無法測得的。在探究歷程中，學生將會獲得科學事實知識與科學方法的練習。因此，對於傳統方法的測驗不足之處，即成爲科學教育工作者，所應思考與解決的問題了。

此外，Shavelson 與 Baxter (1992) 亦曾指出欲編擬一套與教學活動相互對應匹配的實作評量，是耗時、也須要投入相當多的科學、技術的智慧與方法。換言之，發展一套有品質的實作測驗工具，須要先通過學生的試驗，並獲得學生的想法及改進意見，才能益臻完全。反之，抄捷徑的結果，將會造成構想不佳及結構不良的評量工作，惡性循環的結果，將很可能導致不良的教學，卒而，影響及以探究、活動爲取向的課室教學。Shavelson 與 Baxter (1992) 也指出，當一旦完成實作評量的工具編擬時，有四個問題就會被提出：

1.這份評量是否能提供有信度的評量？（事實上，評量的效度應重於信度）。

2.這份評量是否能與學生的學習經驗相匹配？亦即，評量的內容是否爲學生所被教過的？

3.經由這樣的評量，所獲得的成就之訊息，是否有別於傳統方式的評量？（實作評量是否真能測出學生的學習？應是最重要的考慮）。

5.實作評量所欲評量的，是否可經由其它方式來測得的？

前述這四個問題，確實是在設計實作評量所應注意的，倘能合宜的達到這些要求，實作評量工具的編擬，才有其意義。

真確評量：Wiggins (1989) 首先引介此一概念，它具有下列四項特性，一、是真正與相關現場發生關聯，或說是在相關現場取得資訊而完成的實作展現。二、特別投注心力在思考如何對教與學作判準。三、相較於傳統測驗，自我評量在此扮演著更爲吃重的角色。四、過程中，學生常被要求公開展示他們的作品，並爲自己的作品辯護，以確認他們專精的程度。

實作與真確評量有頗多共通之處，有時，甚至很難將它們區分開來。也有學者會以真實的實作評量 (authentic performance assessment) 稱之。Hill 與 Ruptic (1994) 指出，若著眼於作業的選擇與實際生活有關聯，而不僅是以學校的世界爲範圍時，我們會以真確評量來稱呼。Black (1998) 也指出實作性評量的特徵如下：

1.你(妳)不僅能了解學生學到什麼，還能知道他們是如何學習。

2.學生必需嘗試藉由綜合性的知識去評估複雜的作業。

3.學生們爲形成新知識與產出而接受挑戰。

4.作業反映了在課程裡進行實地研究的探索之真實狀況。

5.作業完成的時間可長可短。

6.作業與課程同步進行，而評量則是學生真實學習經驗的反應，並非虛擬營造的。

7.作業情境的安排，將學校活動與真實世界的經驗聯結在一起。

前述六、七兩種特性，是屬於真確評量所獨有，其他的特徵，亦爲實作性評量所有。

四、「全美科學教育標準」中的教學與評量基準

一九九六年，全美研究基金會出版了「全美科學教育標準」一書，對於美國中小學科學教育，建立了目標、基準與評鑑的標準，對於科學教育的實施，應有引導、提昇的作用。雖然，美國國情與傳統和我國有很大的不同，可是，美國「全美科學教育標準」(NRC, 1996)應可作爲國內進行課程改革與評量時的參考。茲引述其「科學教學標準」與「評量標準」，並藉此反思國內科學教學與評量應注意之處。

(一)科學教學標準 A（科學探究取向的課程）

1.爲學生發展一整年的學程架構與短程目標。

2.選擇科學教材，改編與設計課程，以符合學生的興趣、知識、理解、能力和經驗。

3.選擇教學與評量的策略，支持學生理解的發展，以及培養科學學習者的社群。

4.與跨學科和跨年級的教師同工。

面對九年一貫新課程，科學教師亟須重新架構其課程，所以，此一向度，可以用來評鑑教師在這方面的表現。

(二)科學教學標準 B（幫助學生學習的行動）

1.與學生互動時，著重並支持探究活動。

2.對於科學的想法引起學生熱烈的討論。

3.讓學生接受與面對學習的責任。

4.認知與回應學生的異質性，並鼓勵所有的學生全力投入科學的學習。

5.鼓勵與建立科學探究的技能及科學的特質：好奇心，能坦然接受新的想法與數據，及懷疑。

在幫助學生面對新課程時，學生也需要學習如何學，上述的標準可用來做爲教師教學的參考。

(三)科學教學標準 C（教學與學習的評量）

1.利用多重方法以及有系統的對學生之理解與能力等收集資料。

2.分析評量所得的資料，以引導教學。

3.指導學生進行自我評量。

4.藉由學生資料、教學觀察、以及與同事的互動，進行反省，並據以改進教學。

5.使用學生資料、教學觀察、以及與同事的互動，對學生、教師、家長、決策者及社會大眾，報導學生的成就與其學習機會。

前述評量的內涵顯然已超過傳統的評量觀點與方法，教師宜努力練習，才能達到評量的目標。

(四)科學教學標準D（學習環境）

1.安排可運用的時間，使學生能夠從事深入的研究。

2.爲學生開創一個有變通及有助於科學探究的環境。

3.確保安全的工作環境。

4.提供可資運用的工具、材料、媒體，以及學生容易取得的技術資源。

5.確認及利用校外的資源。

6.讓學生參與學習環境的建立。

以往我們的教育，側重在個人，或競爭式的學習，事實上，著重環境與社會情境，應是科學學習的合宜途徑。

(五)科學教學標準E（學習社群）

1.表達對所有學生的不同想法、技能、和經驗之尊重。

2.讓學生參與有關工作內容情境的決定，並要求學生爲其學習社群的所有人員負責。

3.培養學生間的合作。

4.針對科學對話規則的了解之共識，安排與促進正式或非正式的討論。

5.形成與建立科學探究的技能、態度和價值。

科學知識的形成，常須要經過科學社群的同意，同理，倘能建立學習社群，對於學生的學習，應有莫大的裨益。

(六)科學教學標準F（計畫與發展）

1.計劃與發展學校科學學程。

2.參與有關科學學程的時間與其他資源分配的決定。

3.爲教師自己及其同事，全力參與計劃與實踐專業成長和發展的策略。

至於，有關評量的論點列述如下：

(一)評量標準A：評量必須與其所欲得知的決定一致。

1.評量是精心設計的。

2.清楚的敘述評量之目的。

3.對於評量的決定與數據間的關係是清楚的。

4.評量的步驟是具有內部一致性的。

一般而言，紙筆測驗常是評量的代名詞，其實，這是很不對的，因爲其是否能真確的

評量學生，是很有疑問的。所以，九年一貫的到來，評量亦應改弦易轍。

(二)評量標準 B：學習的成就與學習機會是必須評量的。

1.所收集的成就之數據，是聚焦於對學生的學習是最爲重要的科學內容。

2.對於有機會學習的數據收集，應聚焦於最有效率的指標上。

3.對於學生是否有機會學習與學生成就的評量應是同等重視的。

我國的評量，常以考試代替。即使是考試，也常令學生有不知所措之驚恐。因此，評量內容應與教學內容有關聯，甚至，應如何選擇重點，以進行評量，應是最爲重要的課題。

(三)評量標準 C：所收集的數據之技術品質與基於其詮釋所欲採取的決定與行動具有良好的匹配。

1.所宣稱要測量的特徵，能真確的測得（效度）。

2.評量的作業是真確的（Authentic）。

3.個別學生在二個或更多的作業，顯示其有相近的表現，這樣即可說明測到了學生的成就面相（信度）。

4.學生有充分的機會去展現他們的成就。

5.評量作業，以及呈現這些作業的方法所提供的數據，具有足夠的穩定性，以致，即使在不同時機使用，也能得到相同的結果。

評量工具或評量結果要具有信度與效度，才能較真確地測得學生的學習成就。

(四)評量標準 D：評量實務必須是公平的。

1.評量作業必須受到一些審視，類如：是否刻板印象？亦即，是否只反映某特定團體的經驗或觀點？所使用的語言，是否會冒犯到某特定團體？或者其他一些可能使學生自其喜愛的作業上分心。

2.大規模的評鑑，須要使用統計方法以辨認可能存在於不同族群間之偏差。

3.爲了順應身體不便、學習能力較遲緩、或語言能力不足的學生之需要，評量的作業必須做合宜的修改。

4.評量作業必須建立在各種不同的情境之中，俾具有不同興趣與經驗的學生，皆能參與學習，不能假設爲僅是針對某一特定的性別、種族等。

在目前教育環境之下，仍有太多青少年的次文化不被了解與重視，尤其，在評量問題上，更是如此。舉例來說，學校常因作業方便，舉行統一考試，以致不同資質、興趣的學生，皆須使用同一試卷，評量內容與其學習能力常不相配，評量意義盡失。

(五)評量標準 E：對於學生成就與學習機會的評量之推論必須是合理的。

在對於學生成就與學習科學的機會之評量數據作推論時，必須考慮形成推論時所依據的假設。

五、評量的挑戰與革新

綜合而言，有關另類評量，特別是實作評量，及以作業 (tasks) 爲主的評量，國外的文獻已相豐富，諸如：TIMSS (The Third International Mathematics and Science Study)，New standards (1995)，或是 Shavelson 等人 (1990)，及 Ruiz-Primo 與 Shavelson Study (1996) 等研究或報告，皆有可資參考的研究結果。然而，這樣的評量方式，將無可避免地必須作進一步的研究，特別是對於學校、教師、學生等的影響。此外，這種新方式的評量，是否亦能夠類如心理測驗那般地擁有信度與效度，以測量學生對於科學的了解？接著，更重要的，即在於由這種新式的評量，是否能提昇學生在學科學習方面的成就，進而，提昇整體教育的成果，這是全體教育界所應共同努力的課題。當然，在國內也有許多學者/教育工作者嘗試進行另類評量的研究，且已獲致一些成果。在中小學課程改革之際，評量方式的改革亦是非常重要的課題，至於，如何落實於科學的課室教學之中，則有待科學教育研究者與實務工作者共同努力，且應審慎的研究其實施方式與可行性，進而提高另類評量方式在教學過程，甚至學力評鑑或認證時的機會與地位，使學生與教師的知能與權力都得以獲得更適度的發展與尊重。

謝辭

本文的部分內容，曾獲安排於八十九年元月二十七、八日，由台灣師範大學理學院舉辦的「新世紀中小學自然科學課程與師資培育研討會」發表。其次，本文的撰寫，亦獲得國科會 NSC 89-2511-S-018-004 的支持，均此致謝。

參考資料

1.Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum, & Filip J. R. C. Dochy (Eds), Alternatives in assessment of achievements, learning processes, and prior knowledge, pp3-29. Boston: Kluwer Academic Publishers.

2.Black, P. J. (1998). Testing: Friend of foe? The theory and practice of assessment and testing. London: The Falmer Press.

3.Hill, B. C., & Ruptic, C. (1994).Practical aspects of authentic assessment: Putting the pieces together. Norwood: Christopher-Gordon Publishers.

4.Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment:

Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

5.Messick, S. (1989). Validity. In Robert L. Linn (Ed). Educational Measurement (3rd ed., pp.13-104). New York: Macmillan.

6.National Center on Education and the Economy (1995). Performance standards: English language arts, mathematics, science and applied learning (II). Arington: Kirby Lithographic.

7.National Research Council (1996). National science education standards. Washington, D. C.: National Academy Press.

8.Popham, W. J. (1995). Classroom assessment: What teachers need to know. Boston: Allyn and Bacon.

9.Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. Journal of Research in Science Teaching, 33, (10), 1045-1063.

10.Shavelson, R. J. & Baxter, G. P. (1992). What we've learned about assessing hands-on science. Educational Leadership, 49, 20-25.

11.Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. Phi Delta Kappan, 71, 692-697.

12.Tamir, P. (1997). Justifying the selection of answers in multiple choice items. International Journal of Science Education, 12(5), 563-573.

13.Terwilliger, J. (1997). Semantics, psychometrics, and assessment reform: A close look at "authentic" assessment. Educational Researcher, 26(8), 24-27.