由馬里蘭州的學習成就評量與其在臺灣的試測結果看一實作評量的功能與運用

*陳文典 °陳義勳 +李虎雄 ++簡茂發
*國立臺灣師範大學物理系
°台北市立師範學院數理教育系
+國立臺灣師範大學數學系
++國立臺灣師範大學教育系

摘要：MSPAP是馬里蘭州所設計的一種中小學學生的學習成就評量。其特徵在它是一個實作評量，試題係沿著解決問題的過程；觀察、提出預想、實驗設計、執行、整理資料、研判結果等各項心智活動來設計，答卷採取文字敍述方式。因此，評量的項目遍及科學探討的各種能力以及閱讀書寫的表達能力。評分則採取學生在該學年應達成的學習成就指標來評定，不以學生間彼此相對性比較來當標準。

如此的評量方式，費時耗事，評分時更耗人力。但是它却比較能確實地評量到學生的科學能力。

MSPAP試題在臺灣作小規模試測，我們據此發現此種測驗的功能及運用上的困難和適合運用的教學時機。

壹、緒　言

(一) 馬里蘭州的學習成就評量簡介

美國幅員廣大，爲順應各地方的情況與要求，教育制度也和地方自治的體制相似，在實施上採地方分權制。因此，建立一個全國性的教育評量，以顯示全國教育的一般狀況和瞭解各地教育實施的成果，在發展整體教育時，成爲重要的參考指數。美國的全國性教育評量自一九六八年創始，迄今已有二十多年的歷史（彭森明，1994）。在這其中，對於教育評量功能的認知有了很大的改變；從原先只爲獲得學生的基本學習資料，以訂定學生學習成就的指標、藉各地評量的結果收集資料以發現各地的教育差異性等，其主要在於發揮評鑑性的功能。後來，逐漸地對評量結果做進一步分析、瞭解、詮釋與應用，收集有關學生背景、課程及教學方式、學習環境等各方資料，探討影響教育的各項因素、了解教學缺失，因而評量更具有「了解學習，解決學習困難」的積極性意義。這種由以「評鑑」爲目的，進而應用於改進教學，是一個很重要的改變，它的重點不在於競爭和比較，而在於瞭解與改進。

1989年秋，美國總統與各州州長舉行全國教育高峰會議，訂定全國六項教育目標，

做爲教育改革的指針。其中對於基本學科要求達到的程度、以及學生思考與分析能力等均列有要求的目標和標準。這些學習成效的指標，成爲各級學校教學的指針。依據這些指標，1990年後，陸續發展出數學、語文、科學、寫作、歷史、地理、公民等學科的評量，這些評量經各州相繼引用後，已成爲一種全國性的教育評量。關於這些學科的評量，均以學科爲單位，定期有專集出刊，報導各學科的教學指標，和各州之間學生程度的比較，並成爲教育人員制訂政策、教師教學等的重要參考資源。

這類教育評量的重要特質在於能瞭解學習成效的一般狀況，並藉此探究出影響教育的因素。其功用可以比較出過去與現在教育的發展、可以比較原來預期的教育目標與成就現況之間的差距、可以激發各地對教育的重視、激發教師努力去改進教學等等。

馬里蘭州學校學習成就評量計畫（The Maryland School Performance Assessment Program（MSPAP）），主要目的在於藉此評量來考核各校推展教學的行政責任，以提昇該州公立中、小學教學品質及辦學成效。

自1987年10月由馬里蘭州薛佛州長設置一專門委員會以探討州內各校教育成效現況以後，經過1990年的計劃及開始執行，到1993年做施測及分析等工作，整個計劃已落實地運作。而此計劃內容，包括對評量目標、評量內容、評量模式及施行細則，均有詳細的規劃，深具觀摩參考的價值。

本文僅提出該評量的幾項重要特質，以做爲往後像類似的評量在臺灣實施的參照。

1. 評量的目的

   評分採取標準參照（CRT）方式，用以檢視教學目標的達成情形。以作爲改進教學品質及行政上落實責任制的評鑑之重要依據。

2. 評量的內容

   對於中小學各級學生在數學、科學、閱讀、社會、寫作等方面的能力均定有學習成就（outcome）指標，評量的內容即據此指標做整體性的評量，而非單科目的能力測驗。

3. 評量的模式

   是一種實作評量。由一個設計的情境中，引發出「問題」，再由解決問題的整個過程中，各階段心智活動所需的能力中去設計問題；如引發問題、提出假設、提供策略、規劃工作、執行、處理調適的小問題，到對所獲結果的詮釋和評判。整個評量包括閱讀、寫作、數學、科學各種能力的整體運用。

4. 評量施行

在三、五、八及十一年級都設有評量測試。以五年級的測試而言，共有A、B、C三份試題，每份全程測試需兩個半天，每次約105分鐘，是一項冗長的測驗。但是因爲是實作評量，一面操作一面思考，一面獲得資料一面整理資料，所以受測學生並不感到沉悶，反而因爲是生活化的題目而感到興趣。

㈡ 實作評量的功能

自然科學是種依據所發現的現象加以推衍而領悟得到的知識系統，此系統即是由科學概念以及概念間的相關關係來組成，而能以完滿解釋發生的現象爲驗證準則。由於「科學」本身具有的此種特質，有關科學的教學上，均注重在研討問題的過程中，各種心智運作能力的培養，以及嚴謹細心、求眞求實的處事態度，科學概念自然是過程中所自然獲得的結果。整個教學注重的是歸納或推理的過程及作結論的嚴謹上。以此爲科學教學的基本原則，在實施教學及評量上均發生若干困難，而教學上的困難又大半源自評量上的偏頗。因爲在評量整個學習的成果中，處事態度以及解決問題過程中的各種思考和操作，是比較難以具體化而又煩瑣的。故一般教師常感覺到難以執行，而放棄對這方面的評量，乾脆以對科學概念體系的認知來作爲評量的全部。此一趨向的促成，已經對整個教學與學習模式產生扭曲的影響了。

長期以來，在我國的評量方式，不管是平時或升學考試，一直是絕大部份以科學概念的認知和理解爲主，以紙上問答的方式進行。影響所及，學生逐漸捨去「由工作中學習」的學習模式，改由教師講解、習題演練的方式去學習，以符合考試中的要求。教學中若有實驗，也流於爲理論解說，用來驗證理論、體會理論的補強活動，而產生於研究過程中的創造契機也就很少呈現。即使學生偶爾有一些好的想法，也因爲不能或缺乏此種實作習慣而流於空談，所學的科學概念也因爲缺乏實作的過程而不能運用來解決自己的問題，「學問」有逐漸成爲「清談」裝飾品的危險。這都是「實作評量」長期偏廢所引致的教學與學習的偏差，輕忽了科學求眞求實的基本特質。

其次，在長期聯考制度的導引下，「絕對標準化」的答案和迅速的評分過程成爲評量的追求目標，命題的模式也受到更大的限制，使原本在紙筆評量中尚倖存的問答題、作文題更形萎縮，塡充題也受到評分的人力限制而大加限制，最後，剩下的可能只限於「是非」、「選擇」題型。而各級學校也因爲因應聯招的命題方式，影響到整個平時的評量也相繼跟進，以「是非」、「選擇」爲一般命題的基本模式。其結果是：企圖以紙筆的「是非」、「選擇」題型要來評量科學的整個教學成果，不啻是一個極高難度的挑戰，甚或可說：不可能。此項評量趨勢的導引下，學習的模式也受到很大的扭曲，不只

實做的學習活動被輕忽，對於整個思考的流暢性和發表的完整性都付之闕如，學習成爲零碎的知識理解，或零星事實的記憶。由於評量方式的窄化，評量沒有去照顧到整個科學教育的教學目標，已經使整個教學活動受到極大的影響。

除此，評量也因其宗旨不同而不同。爲了升學而行聯招考試，其目的在甄選特殊的學生，注重在排名次，比出相對的成就程度，其評定的標準係採常模參照（NRT）的類型。在平時評量方面，也習慣以比名次及同儕競爭的方式來訂標準。這類標準的採取，對於以評量學習困難和教學目標是否達成的目的來說，是不合適的。至於顯示教學目標達成程度，以及呈現教學成效、教材適用情況、學校辦學績效等的評量，則評定的標準宜由課程標準及教學目標等標準參照（CRT）來訂定更合宜。只是這一方面在我國比較少被採用。

㈢ 實作評量試驗性試測

基於MSPAP評量計劃設計的週延性及其具有的特質，教育部自中華民國八十一年起藉舉辦兩次學科基本學習成就評量研討會（彭森明等，民81，民82）加以介紹，並於八十三年就MSPAP評量之中抽取數學及科學的部份，小學階段（以五年級代表）的試題，做小樣本的測試。

雖然在八十三年十二月所做的只是小規模的試測，若將其作爲評估臺灣區學生的程度，則代表性尚嫌不足。可是，此項試測至少達成幾項有意義的成果；⑴實作評量的設計與施測的經驗，⑵實作評量的特性與功能之瞭解，⑶闡述性與評判性的答案之評分標準化流程設計經驗。以下謹就實作評量實施時，獲得的學生回應及技術上困難的克服，特於本文加以報導。（詳情參閱研究報告，民83）

貳、實作評量的題目、施測與評分

由馬里蘭州開發的實作評量（參閱MSPAP，1994），具有幾項基本特質：1.題目係由解決一個問題出發，在解決問題的過程中，所遭遇到各種思考、研判等心智活動的考驗來設計小題目，形成一系列的問題，故原則上是對「解決問題」整個能力的評量。2.一個題目，是由一系列小問題組成，配合實際的操作。解決過程所需的時間，可能長達50分～60分。3.採取簡短的，敍述性的方式作答。4.評分採取標準參照的模式，依各年級學習成就要求的指標來評量，不由學生間能力之比較來定標準。

相應於以上的各項特質，MSPAP在研發及應用此一測驗時，也表現出一些措施，如1.對於各題目、各子題相對於學習成就指標而言，其難度均經過統計上的分析而標定。

2.各子題所評量的內容如科學概念、態度、方法等均有所標定，評量目的明確。3.由於敍述性的回答，評分標準力求確定；評分員的培訓及複檢均設有既定的程序，以確保評分標準不致滑動。

2.1　實作評量的題目：

本次試測著重在可行性的探討，及試探臺灣學生對此種評量模式的反應。題目則完全由馬里蘭學校實作評量計劃（MSPAP）提供。爲了不影響測驗性質，雖因應本地情況限制，變更少許器材或所安排的情境，但題目的內容應該可以認爲是相同的。

以國小五年級爲對象，共有A，B，C三份試卷，各卷內容包括數學、物理化學、生物的題目（見表一），受測學生在處理每個題目時，等於在解決一個問題。而在解決此問題所經歷的過程中，每一步驟所需的思考與行動，即對應某一探討科學的能力，設計相應的評量小問題。因此，一個題目涉及觀察、提出假設、執行實驗、整理資料、研判資料、推廣與應用時，各設計了一系列小問題。

試題設計包括測試題本、實驗裝置資料本及施測手册。

表一　試卷內容及測試時間

卷別	作業名稱	操作時間（分鐘）	題材屬性
A	一、平衡操作	75	物理
	二、評估一項遊戲	30	數學
	三、鹽度	45	化學
	四、動物物種的生存	55	生物
	五、替游泳池鋪地磚	35	數學
B	一、球的滾動	65	物理
	二、五年級的迷你課程	25	數學
	三、規劃一座兒童遊樂園	25	數學
	四、浮萍的難題	35	生物
	五、灰石研究	35	化學
	六、布置公佈欄	30	數數
C	一、改建運動場	35	數學
	二、雪車滑行	75	物理
	三、煤之鄉	70	化學
	四、幸運號碼	30	數學
	五、爺爺的湖	35	生物

2.2 施 測

一共有A，B，C三種卷子，每一位學生只測其中的一種卷子，要完全測完其中的一種需要兩個半天的時間。

樣本採自全國2445所小學中之15個學校，每校隨機選出一個班級，每一種卷子測五所學校共五個班級，故每種卷子的受測樣本約200名（A卷168名、B卷155名、C卷210名）。

監試教師爲該校導師，並在施測前接受測試講習，以瞭解施測的過程和執行細則。

2.3 評 分

由於一個題目所包括的各個子題，其難度和對應的學習成就評量項目均已確定，評分的標準也可確定。不過，由於問題的回答採取敍述性的方式，在評判上不容易評定，故針對評分上的不確定性困難，而有評分者的講習與訓練，務使評分標準化。

參、施測結果分析

由A，B，C三份測驗中，我們題出A卷「平衡操作」、B卷「球的滾動」及C卷「雪車滑行」等三個題目測試結果來分析。在學生對科學學習成就方面的表現作一報導，對實作評量的功能方面給予肯定，並對實作評量實施的可行性和策略提出建議。

3.1 學生在實作評量下，表現的學習成就情況：

(1) 傳達能力方面

・當題目要求對所操作的現象和結果作科學性的描述時（此類題目有「球的滾動」活動四步驟1，「雪車滑行」活動七步驟2），能夠針對觀測量與變因之間作因果關係的描述的僅40％。一般的說，即使是依因果關係來描述也都是很粗略，十幾二十個字就寫完了，表示運用文字表達的能力有待培養。

・設計圖表，用以表達資料的能力尚佳：運用雙向列表、長條圖等能力普遍具有（「平衡操作」活動三步驟4），但不習慣運用數線及統計上的莖葉圖。

(2) 對實驗上的認識普遍不眞確

・誤解「實驗前的預設」的功用，普遍認爲實驗結果若與預先的假設不合，就認爲實驗「失敗」。實際上預設只是一種先期經驗的判斷，做爲執行時方向的參考而已。施測結果發現，相信實驗結果若與預設不合，仍可能是一個好實驗的（「雪車滑行」活動七之2）只佔1.9％！比率相當低。

・當被詢問「和別人的實驗結果做比較有什麼價值」時，能持以「實驗狀況相同，

比較可增加信度」等科學性考量的（「雪車滑行」活動七之3）佔32.4％，其他同學（佔41.2％）大半以要謙虛、多學習、向成績好的同學學習等社會性的觀點來看待這種實驗時相互切磋的價值。

(3) 運用學過的理論或模型來解釋現象的思考方式並不普遍

・雖然在「雪車滑行」題目中，用「鋼珠由斜坡滾下衝出杯子，使杯子帶有動量」來模擬雪車下滑的動量，在「平衡操作」題目中，用「蹺蹺板的遊戲」來模擬槓桿，並且經過了實驗，獲得了許多的資料，也經過引導得出一些規則。可是，當被詢及「如何讓雪車不致滑太長」或「槓桿的平衡規則」時，大半的學生完全拋棄剛獲得的資料（引用資料，解釋到這些現象的，在「雪車」題目中只佔19.4％，在「平衡」題目中只佔22.4％），而運用自己過去生活上的經驗去作答，也就是在處理問題時又回歸到未學習前的素樸概念（naive concept）來處理問題。

3.2 實作評量的功能

(1) 實作評量藉由解決問題的實際歷程，經發現問題、觀察、形成假設、推理、控制變量、測量、歸納、解釋資料、研判、作決定，以及應用已獲得的規則、提出理論、建構模型等自然的活動，設計成一系列的問題。雖然每一個題目要經過30分鐘到70分鐘不等，才能完成，但是過程並不沉悶緊張，而是充滿了創意的、活潑的氣氛。

(2) 實作評量很眞實的把學生傳達的能力，運用時空關係、數學關係和科學過程技能、科學運用概念等，很自然的結合起來表現出來，而且評量出來。在情境而言，它本質上是整體的、連續的，和一般設計單題單一情境或是非、選擇題的模式之測驗完全不同。以整個科學教育目標來看，這種評量才算是不偏廢的全方位科學能力之評量。

3.3 實作評量實施的可行性（李虎雄等，民84）

・評量題目的設計不成問題

在平常教學的評量中，教師可依經驗來評估各子題的「難度」，設計「實作評量」的題目，並不比一般設計一份紙筆測驗要費事，主要的難處在於構想的情境要能引發出所要測試的題目，並且運用現有的設備，就足可用來解決此問題。

至於解決此問題的過程中，各操作及思考均可很自然的設計出一系列子問題來，而達到對整個科學學習的各目標作全方位的評量。

・評分工作是一大負擔

依照批閱的經驗，一個題目包括約15～20個子題。其中有些是由敍述方式回答的，有些是運用圖表表示資料的，有些研判，也有些是應用舉例，回答方式因各子問題性質

而設。像這樣的一個題目要完全批改完，以一班50個學生來估計，最少要5小時的時間。

而教學上的評量，不僅只是解決一個問題的能力而已，科學概念方面的領會和應用能力也很重要。此類考試因只限於對某一題目的解決，只能測驗到和這個題目有關的少許科學概念。若要全面的把學習過的概念都測試過，則必需包括數個題目，測試時間兩個半天，批改一星期，這就是完全不切實際的想法了。

肆、結論與建議

此次，以馬里蘭州實作評量（MSPAP）移植式地在臺灣試測，其結果的分析與檢討中，令我們獲致以下的一些結論與心得：

1. 此種實作評量的測驗模式，評量遍及各項科學能力：

將學生由如何察覺問題、提出對問題的看法、對可能的結果提出預設、規劃工作藍本、執行與操作、登錄整理及分析資料、研判資料意義、提出見解、解釋資料、到應用建構的概念於其他事例，藉著實際上進行工作的流程而一一的測試出來。一方面也是在強調科學教育完整的目標，另一方面也在導引學生如何去依照科學方法解決問題。

2. 我國學生在回答題目時，顯示傳達能力與獨立作業能力均不足：

由我國學生作答情形來看，在發覺問題、規劃工作流程、以及做獨立研判方面表現不如理想。在傳達方面，凡是涉及敍述方式回答的答案，均顯得簡略不全。例如，若有涉及數個重要變因的，回答中絕大部份只提一個。而且，在設計實驗時，絕大部份完全忽略其他變因的影響，只論及正在考慮的變因。即使敍述一個現象，也大部份寥寥數字，語意不流暢。

3. 實作評量及敍述方式的回答，其評分的客觀標準化，可經過評分者講習的培訓，達到目標：

MSPAP在培訓評分員方面，有一個較嚴格的程序：(1)先由專家試著批改測試卷，協調建立評分標準，並將各種回答類型列出，編成「評分者手册」。(2)評分員經過培訓後，試改數份題目，達一定標準（例如說八成）後及格，參加評分。(3)每批改一定份數（例如說500份），再檢定評分標準，及格後繼續批改，不及格者再度培訓。

因此，敍述性回答的評分標準化在執行上並無困難。只是，是項測驗係專對一科學問題的科學答案而言，可應用在聯考等大規模考試上。至於像國文的作文科評分，有時又涉及到文學性或意識型態方面的因素，要評分客觀標準化可能尚有困難。

4. 實作評量可適用於「平時作業」，或教育行政單位在「評鑑各校科學教育教學成效」，以及學生的「科學能力競賽」等場合。以強調科學能力涵蓋科學概念的認知和解決科學問題的各種能力而不偏廢。可矯正目前運用紙筆測驗，且以是非、選擇方式測試所引起的對教學及學習模式的扭曲情形。

實作評量固然在評分上相當的煩瑣，但是對科學學習的整個目標，可以做更切實更全面的評量。故宜在目前一般的評量方式上，互補的更換一些這種模式的評量，才能導引學生學習的方向，並對學習成效作確切的瞭解。

伍、補　記

本評量研究計劃由教育部主辦，國立臺灣師範大學執行。計劃由簡茂發主持，研究人員有李虎雄、陳昭地、林保平、王淑貞、陳文典、陳義勳、林秋麗、黃長司、黃萬居、鄭美雪、朱玲玲、曾文雄、吳美麗、卓娟秀、張武昌。

陸、參考資料

- 彭森明（民83），學生學習成果評量與分析—美國經驗。國立臺灣師範大學印。
- 彭森明等（民81），學科基本學習成就評量研討會。國立臺灣師範大學印。
- 彭森明等（民82），學科基本學習成就評量研討會。國立臺灣師範大學印。
- 研究報告（民83），國民教育階段學生學習成就評量研究。教育部主辦，國立臺灣師範大學印。
- 李虎雄、黃長司（民84），美國馬里蘭州學校實作評量工具在臺灣施測的可行性。科學教育月刊179期，pp. 41～49。
- MSPAP, 1994 Technical Report.
- MSPAP, 1994 Maryland Learning outcomes for Reading, Writing, Language usage, Mathematics, Science, Social studies.
- MSPAP, 1994 Score Interpretation Guide.

（收稿日期：84年10月18日，接受日期：84年11月1日）

The Function and Application of Performance Assessment

***Wen-Den Chen, °I-Shin Chen, †Hu-Hsiung Li, ‡Maw-Fa Chien**

*Department of Physics, National Taiwan Normal University
°Department of Science & Mathematics, Taipei Municipal Teachers College
†Department of Mathematics, National Taiwan Normal University
‡Department of Education, National Taiwan Normal University

Abstract

The MSPAP (Maryland School Performance Assessment Program) assessments are designed to assess the study achievement of primary and middle school students in Maryland. The assessments operate in manipulational activities. The items were designed along the problem-solving procedure: observation、assumption、experiment designation、practice、data organization and explanation、make decision and conclusion around a theme or problem solved. The students are required to respond in the form of short-paragraph writing. So that reading, writing and language usage assessment are tested simultaneously.

The specification of items covers all sorts of science abilities and communication abilities, as indicated in the Maryland Learning Outcomes.

The test is performed in about nine hours during two days and scoring work is clumsy. But the assessment is valid for testing the science ability.

The test has been performed in Taiwan only with a small sample (200 students). However, through this test we found the function of the assessment, the performance skill and the suitable opportunities for applications during class-instruction.

★