學習成就評量的多元功能及其相應研究設計

Multiple Functions of Student Assessment and Their Proper Study Design

彭森明 Samuel S. Peng
國立台灣師範大學教育評鑑及發展研究中心主任
Director, Center for Research on Educational Evaluation and Development, National Taiwan Normal University

摘　要

學習成就評量是教育過程中一項必要工作，不僅教師們在課堂教學中需要進行評量，瞭解學生學習情形，而且學校及各級教育主管單位，也都要推展評量，為教育成效把脈，以便掌握各級教育績效以及教育品質，為教育決策提供實證資料。目前世界主要國家無不積極推展學習成就評量以檢視國民在國際上的競爭能力。有鑑於此，本文旨在討論大型學習成就評量的各種功能以及設計評量時應考慮的事項，供國內負責推展大型學習成就評量工作以及有志參與此類評量的人員做參考，以便能更有效地推展國內教育評量工作。

關鍵詞：學習成就評量、教育評鑑、評鑑功能、評鑑設計

Abstract

Student assessment is an essential part of the education process. At the classroom level, teachers would assess students to find out whether they have learnt what has been taught, and then make adjustments or modifications in their instruction. Similarly, at the school, school district, or national level, student assessments are performed to monitor educational progress and evaluate the overall quality and efficiency of education. Moreover, in the modern knowledge-based world, the information resulting from student assessment is extremely important because it reflects a country's competitiveness. For these reasons, this paper synthesizes information about the multiple functions of student assessment and their respective design considerations, in the hope that such information would be helpful for the development of an efficient and effective student assessment program in Taiwan.

Keywords: student assessment, educational assessment, purpose of assessment, assessment design

一、引言

學習成就評量是教育過程中的一項必要工作，所以學校教師都會經常在課堂上測驗學生，或問學生問題並觀察學生行為表現，其目的即是在檢測學生學會什麼、學到什麼程度、能做什麼、以及診斷學習上的缺失或學習上困難的地方，以便決定教學進度或調整教學方法。

這種評量工作，也經常向上延伸到以學校為單位來實施，比如每所學校都會有統一試題的期中考試及期末考試，不僅用來瞭解學生個別學習成果，以判斷其努力的程度，而且也可以藉此瞭解班級間之差異，以提供教師改進教材、教法的參考。

同樣地，這種評量工作也可以延伸到各教育地區或全國，進行大規模的學習成就評量。以美國為例，早在六、七十年前，各地區學校即開始使用民間機構研製的全國性常模測驗，如Stanford Achievement Test和Iowa Tests of Basic Skills，以衡量學生各學科的學習程度。另外美國聯邦政府從1969年起，推展「全國教育成果評量」（National Assessment of Educational Progress，簡稱NAEP），由美國教育部的教育統計中心（U.S. Department of Education, National Center for Education Statistic）主導。NAEP有系統地檢測全國四、六、八、及十二年級學生學習成果（assessing what students know and what students can do），涵蓋11項學科（包括閱讀、寫作、數學、科學、社會、公民、美國歷史、地理、文學、音樂、電腦教育），並分析不同性別、社經背景和族裔以及不同地區學校之間的差距。NAEP亦引導各州進行符合各州特殊需求之評量，並帶動測驗理論與方法的基礎研究，包括試題反應理論（item response theory，簡稱 IRT）的應用、個人預估分數（plausible values）的計算、各學科內容及能力層次規範的制定、試題校準（item calibration）和測驗等化（test equating）的技術、檔案歷程評量（portfolio assessment）和實作評量（performance assessment）的理念及應用，以及增值評量（value-added assessment）和發展模型（growth model）等，有很多創見，對教育測驗學有很大的貢獻。目前No Child Left Behind

（2001）教育法案要求各州對四至八年級的學生每年舉行會考，以考核學校教學績效是否達到標準，並檢測不同族裔和不同社經背景之間的差距是否縮短（U.S., 2001）。

美國大型學習成就評量的理念與技術亦逐漸延伸到國際教育比較上，用來檢測各國學生之素質以及未來國家之競爭力，並藉此相互學習尋求改進之道。比較顯著的有IEA（International Association for the Evaluation of Educational Achievement）的「國際數理學科成就趨勢評量」（Trends in International Mathematics and Science Study，簡稱 TIMSS），以及OECD（Organization for Economic Co-operation and Development）的「國際學生學習評量」(Program for International Student Assessment，簡稱PISA）和「國際學生閱讀能力評量」（Progress in International Reading Literacy Study，簡稱PIRLS）。1988年起，NAEP亦將數理學科擴增至國際間的共同研究，進行國際數理教育評量研究（International Assessment of Educational Progress，簡稱IAEP）（IAEP, 1989, 1991），以便增進國際間的相互了解，做為考核教育成效的指標。上述這些學習成就評量都有很多國家參加，比如2006的PISA有56個國家參與，2007的TIMSS將有65個國家參與。無庸置疑的，學生學習成就評量在教育成就上確實是舉足輕重的工作，不僅先進國家重視，很多開發中的國家，甚至未開發的國家都重視此項工作，使得學習成果評量已成為國際趨勢。另外這些學習成就評量資料，加上一些相關教師、學校、家長及學生之資料，即成為非常有價值的教育研究資源，供研究學者探討許多教育議題。（註：有關上述各評量方案的性質與詳細內容，請參閱各方案網站，網址附於文後。）

近年來台灣也參加TMISS，PISA和PIRLS的計畫。在TIMSS的評鑑中，台灣的學生有良好的表現，在國際排行中，數理兩科都排在前面，與新加坡、日、韓、香港等不相上下。

不過話說回來，雖然大型學習成就評量在國際上已進行多年，但在國內這種評量工作卻起步很晚，尚未能真正發揮評量功能，以致多年來國內的教育改革工作以及教育政策的制定，往往缺乏實證資料做依據。歷年來雖然

有國中基本學科能力測驗（基測），高中學力測驗（學測）以及大專入學指定科目測驗（指考或科考），但這些測驗只能用來檢測個別學生在當年測驗的成果，在群體中的相對位置，做為區分與選擇學生的依據，而無法有效地檢測學生的真實程度以及整體學習成果的變化。原因很多，主要是每科考試的時間有限，試題無法涵蓋各學科全部內容以及能力規範，而且由於國內補習風氣太甚，很難有效地保存核心試題做試題等化之用。更可惜的是，這些寶貴的測驗資料，未能與學校、家庭以及其他有關學生的資料結合，成為教育研究資源，供探討一些重要教育議題之用，包括群組及地區成果差異和教學資源及社經背景之關係，未能發揮資料之最大功能。

兩年前國內開始推展「台灣學生學習成就評量」計畫（Taiwan Assessment of Student Achievement，簡稱TASA），由國立教育研究院籌備處主導，其主要目的在建立國民中小學及高中職學生學習成就資料庫，分別針對國小四年級、六年級、國中二年級及高中二年級、國、英、數、自然及社會科學的學習，進行成果評量，追蹤、分析學生在學習上之變遷趨勢，進而檢視目前國家教育體制與政策實施之成效。此計畫開始了國內大型學習成果評量的新里程，值得肯定與支持（http://tasa.naer.edu.tw）。另外國內近年來也開始評鑑各級學校、大學科系以及學程等，以確保教育品質。因此雖然國內全國性評量工作要比其他先進國家起步晚些，但相信能借鏡其他國家之經驗，省去許多摸索及嘗試改進之工作，可以迎頭趕上甚或超過他們，發揮評量在教育工作上的最大功能。

基於上述評量之重要性以及對國內評量工作的期許，本文將綜合一些國際經驗，供國內負責推展大型學習成就評量工作以及有志參與此類評量的人員做參考，以便有效地推展國內教育評量工作。本文將討論大型學習成就評量的各種功能以及設計評量時應考慮的事項，茲分別敘述如下：

二、大型學習成就評量的多元目的與功能

為什麼要做大型學習成就評量？大型學習成就評量能有什麼功能？大

型學習成就評量能用來做什麼？

大型學習成就評量與班級課堂測驗不同之處，在於關注的對象不同而已。前者評量的對象是群體，而後者評量的對象是個人。因此前者評量關注的是群體表現，如全國學生或某一個地區、某一種學校、某一群學生之整體表現，而非個別學生的表現。

不過無論是大型或是小型課堂評量，其功能都可以是多元的。一般人也許以為學習成就評量就是以量化的方式，如考試，或以質化方式，如觀察與訪談，來檢測教學及學習成果而已。其實學習成就評量包含評鑑與測量兩部份，不僅測量學生學到什麼和能做什麼（即是資料收集如量身高、體重、體溫），而且要評鑑此成就所代表的意義以及可能的相關因素（即是闡釋資料意義如過高、過重、發燒及其可能的原因），所以學習成就評量可以很簡單，也可以很複雜。學習成就評量可以用不同的設計以及不同的資料收集和分析方法，達成不同的目的或功能。因此規劃學習成就評量時，必先釐清目的與預期功能，然後才能採用適當的設計方案，包括抽樣、資料內容規範、資料收集方式以及統計分析方法和報告釋出等，有效地達成評量目的與功能。

學習成就評量的功能大致可以分成下列幾項:

1.瞭解教學與學習成效現況：評量最基本的目的與功能即是瞭解現況，描述目前學生的程度，學生知道什麼及能做什麼？實施大型學習成就評量即可在各基本學科領域上，瞭解全國或某一地區、某一類別學生的學習成果，比如有多少百分比的學生達到合格標準？學生分數的分配情形如何？是常態分配還是趨向高分或低分？前述美國NAEP的最原始的目的之一，即是在於瞭解美國學生的學習成果狀況，瞭解美國四、六、八及十二年級的學生知道什麼及能做什麼，其設計亦以達成此功能為主軸。

2.診斷教師教學及學生學習缺失：評量的第二種目的與功能是診斷學生在學習上的缺失與困難，包括在課程類別及項目（如分數的加、減、乘、除計算），以及能力層次（如演算、應用）上的缺失，以便做改進。前述美國NAEP的最初分析，除了計算總分之外，亦依評量內容

細目分析，其目的即是想找出學生在課業上的強項與弱點以及不足之處，以便做為改進課程和教學以及提供輔導之依據。

另外大型評量亦可用來檢測社會及學校投入教學的資源（input）以及教學措施（process）是否不同或有不足之處，如前述TIMSS的課本（課程)分析，比較課程內容（input）以及課堂教學錄影分析（video study），比較教師教學方法（process），找出不同與不足之處，做為改進教學之用，即是為此目的而設。

3.偵測趨勢與變化：評量學習成果常要回答的問題是學生的表現是進步還是退步？哪一方面進步或退步？進步或退步多少？美國NAEP的閱讀能力、數、理能力評量，每二至三年重複一次，其目的之一即是偵測整體學生的學習成就變化。No Child Left Behind Act（2001）所要求的評量，因每年都實施，所以亦有檢測各校、各學區及各州四至八年級學生學習成就變化的功能。

4.衡量學生學習成長：學生知道的及能做的事，是否依年齡或年級穩定成長？成長多少？這些也是評量可以回答的問題。不過要回答這些問題，必需要有能準確評量學習成長的測驗工具，持續對同一群學生在某一年齡或年級階段施測，因此需要長期追蹤之研究設計。美國1980年代的High School and Beyond 長期追蹤方案，有此設計，是其特色之一。目前美國在No Child Left Behind（2001）教育政策下，每年所有四至八年級學生都要接受各州舉行的會考，因此也可以檢測學生從四年級至八年級的學習成長（learning growth）。

5.檢視個別或組別差異：這項功能是上述目的的延伸，進一步了解不同學生之間的差異，比如性別差異、貧富差異、城鄉差異以及地區性或不同類別學校之間的差異。這也是美國NAEP的原始目的之一，至今依舊保存。國際評量方案，如前述之TIMSS、PISA、及PIRLS，也以檢測各參與國家間之差異為重要目的之一。

6.提供分級、分等與分組依據：評量可以依據分數標準將學生分等（如優、良、佳、尚待改進），分組（如通過，不通過），和分級（初級、

中級、高級）等，做為編班、取才、分流及其他教學措施之依據。美國NAEP最初分析只看全國學生達成標準的百分比（通過之比率），後來亦嘗試將IRT scale scores 依專家認定之標準，將學生分成basic、intermediate、和advanced等級，以闡釋測驗分數的意義，說明各等級的學生能有什麼不同的表現。

7.探究與學習相關之因素：評量之目的除了描述現況、診斷學習缺失、偵測變化、衡量成長、檢測組別差異，以及分等或分級之外，往往需要了解與這些現象的相關因素，包括投入之資源、課程、教學方法、師資素質、學校環境、家庭背景以及學生之個別差異，以及相關法案及措施等，以便進一步瞭解學習如何產生，做為設定改進教學方案之依據或參考。目前各國及國際學習成就評量，皆包含此項目的。比如TIMSS的附加問卷調查資料、課堂錄影資料（video study）、以及課本內容分析，都是為此目的而設。另外國內之「台灣教育追蹤資料庫」（http://www.teps.sinica.edu.tw）及「高等教育資料庫」（http://www.cher.ed.ntnu.edu.tw），其主要目的之一亦是探討與學生學習成就有關的因素。

在許多可能的因素中，教師與學校對學生學習的影響目前甚受重視。美國No Child Left Behind Act （2001）教育法規要求各州呈現出教師與學校教學成果，因此引起許多「成果歸因分析與研究」（value-added analysis），試圖了解不同教師與學校真正帶給不同背景與資質的學生的教學成果，成為很熱門的教育評量研究。

三、大型學習成就評量應儘量發揮多元功能

綜上所述，大型學習成就評量可以達成多種不同的功能。雖然不是每一評量方案都能達成所有這些功能，但是評量是費時費力的工作，因此為使評量發揮最大的效率，在設計每一評量方案時宜儘量涵蓋多重目標與功能。

以美國NAEP為例，開始時只著重整體學生達到學習標準的百分比以及

族群、性別及社經背景間之差異。簡言之，即只是檢測學生學習成果而已。但後來發現這項資訊不能滿足教育工作者以及決策者之需求，因為他們想要進一步了解除了族群、性別及社經背景間之外，還有什麼因素造成不同的學習成果，因此後來又用教師問卷、學生問卷以及家長問卷，增收有關教師教學方法、學習環境、學生學習態度、生活狀況以及家庭教育資源與輔助等資料，以便了解這些因素與學生學習成果的關係。後來又偶爾加收學生在校學籍記錄之資料（school transcripts information），以了解學生在校課程和學習活動與學習表現之關係。IEA主辦的TIMSS也是一樣，除了測驗成績之外，也有教師、學生以及家長問卷資料。此外還加了課本分析（textbook analysis）以及一些課堂錄影分析（video study）以便了解「教什麼」以及「如何教」對學習成果的影響。

NAEP and TIMSS評量方案是很好的典範，不僅可以直接測量整體學生學習狀況、檢測各組群學生之間的差異（TIMSS還包括國際間之差異），而且還可以利用此評量資料，探討影響學生學習的各種可能因素，包括課程、教學方法、學習歷程以及各種社經背景等，成為豐富的教育研究資源。

可惜的是NAEP及TIMSS兩項評量都是橫斷性的設計，沒法檢測學生學習成長（growth）以及與學習成長有關的因素。NAEP曾經考慮過長期追蹤方案，但長期追蹤的工程比較複雜，時間很長，比較困難即時完成檢測學生學習狀況，因此採取分開進行。長期追蹤研究不像NAEP那樣頻繁，大約每八年至十年一個方案，以補助NAEP之不足，回答一些NAEP不能回答的教育議題（有關長期追蹤研究方案請參閱http://nces.ed.gov）。不過目前No Child Left Behind Act（2001）要求全國學生在四至八年級時接受會考，所以有重複被測驗的情形，不僅可以檢測學習成長，亦可探討與學習成長相關的因素，包括教師素質、教材、教法以及學校的學習環境。

總而言之，評量資料的收集，宜審慎設計。實施每一方案時，應盡可能將相關資料一併收集，成為整合型的大型資料庫，一方面減低資料收集的成本，一方面增加資料的使用價值，使每一評量方案都能達成多元功能。

四、大型學習成就評量之設計考量

設計大型學習成就評量時，應考量的事項很多，主要的有下列幾項:

1.評量目的：前面說過，評量可以用來達成多種不同的目的，達成多元的功能，因此評量設計首先必須釐清目的，確切了解為什麼要做此評量，因為不同的評量目的，需要不同的設計方案來達成。比如要診斷缺失，必須要有詳細的內容規範以及學生背景資料;要偵測變化必須要有相等的評量工具以及持續執行的評量;要檢視組別差異，必須要有適當的抽樣或普查設計，確保各組別有足夠人數做分析;要探究相關因素必須要收集適當的相關資料。因此要做好評量工作，必先有明確的評量目的，然後再依目的設計實施方案。

2.評量內涵: 評量什麼，即內容規範（assessment framework），必須詳細規劃。內容規範平常都含課程項目與能力層次，以及試題數目分配。這是非常重要的工作，需要專家依課程網要製定，因為評量內容規範會影響教師教學與學生學習。為使測驗內容不受政治或特殊意識型態干涉或操作，美國的NAEP還特別設置顧問諮詢委員會來決定內容。委員會由各方代表組成，委員由教育部長聘任。目前國內的TASA也訂有規範，以國小四、六年級數學科為例，其規範如下：

表1 2006 TASA 國小四年級、六年級數學測驗各內容領域的題數分配百分比

	四年級			六年級		
	選擇題	應用題	總題數 %	選擇題	應用題	總題數 %
數與計算	42	5	40%	42	5	40%
量與實測	31	4	30%	21	3	21%
幾何	10	1	9%	21	3	21%
統計	10	2	10%	10	1	9%
代數	11	1	10%	10	1	9%
總題數	104	13	100%	104	13	100%

資料來源：國立教育研究院籌備處所（http://www.naer.edu.tw）

不過上述規範只顯示課程內容而無能力層次，不夠精緻。因此一些高層次的能力可能未被檢測或忽略。美國NAEP的內容規範比較詳細，值得借鏡。以1990-2003的數學為例，其規範包括下列五大項內容 （content area）:

- number sense, properties, and operations（數與計量）;
- measurement（數與實測）;
- geometry and spatial sense（幾何）;
- data analysis, statistics, and probability（統計）;
- algebra and functions（代數）。

另外，每一項內容包括下列三大層次能力（mathematical abilities）:

- conceptual understanding（瞭解概念）,
- procedural knowledge（通熟程序）,
- problem solving（解決問題）。

以及 三大實際運用能力（mathematical power）:

- reasoning（推理）,
- connections（連結）,
- communication（闡釋）。

因此其整體內容規範（assessment framework）是5 x 3 x 3 =45個項目, 然後從每項目中再決定試題方式與題數。

資料來源: http://nces.ed.gov/nationsreportcard/mathematics/whatmeasure.asp

3.測驗方式：測驗的方式影響學習，因為如何考試往往引導學生如何學習和學什麼，所以評量時採用什麼方式來測量必須慎重考慮。目前大型學習成就評量方式主要還是一般選擇式測驗。此種測驗不易考核出分析、綜合、應用等高層次能力，因此有需要考慮兼用一些其他試題方式包括演算、問答、應用題等。另外亦有檔案歷程評量（portfolio assessment）（Paulson, Paulson, & Meyer,1991）、實作評量（performance assessment) （Linn, 1993）與真實評量（authentic assessment）的構想，其評量的方式可以包含學習檔案(portfolio)、成品（product）、行為表現檢核、觀察、訪談、日記、自我報告、實際操作、模擬演練（simulated performance）等，俾使評量能兼

顧過程與結果。上述這些評量方式雖然有耗時、經費和設備上、評分、技術品質等幾項難處（盧雪梅，1998），卻比較能測出目前極受國際重視的學生思考過程以及理解分析問題、綜合應用知識和解決問題的高層次認知能力。這些方式可以採用許多小樣本方式去分別做少量不同的試題，以檢核整體學生的能力而非個別學生的能力。

4.相關資料之收集:　除了測驗之外，還要附加哪些資料? 這些資料可依來源分類，如家長問卷、學生問卷、學生學校之紀錄、教師問卷及對學生之評量、學校問卷、課程分析以及課堂施教狀況等。資料內容與性質，通常分成背景、資源、課程內容、個別差異（input）以及過程（process）如教學方法兩大類。這些資料的取擇，往往有賴於文獻之彙整以及教學模式之規劃。下圖為教學模式範例之一，可依模式中之主要結構項目來分別決定要收集那些資料。這些項目包括：課程內容與性質、教學措施與方法、學校環境與資源、家庭社經背景、家庭教育活動、社區教育資源、個別差異、以及主動學習行為與態度。這些資料的收集，可用問卷方式要求相關人員填答問題，或直接從既存檔案中摘取。

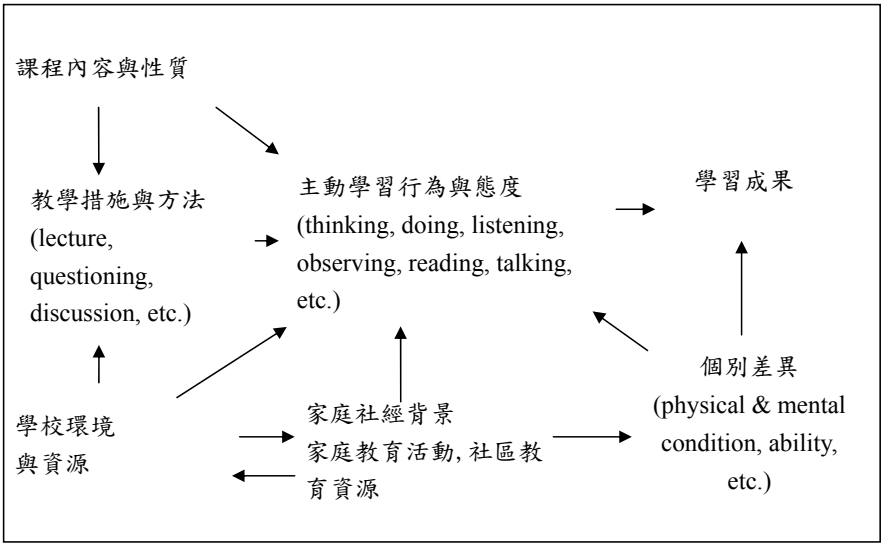

圖1 學習過程模式架構圖

5.普查或抽樣:評量是要全體學生還是抽樣學生參加，也是需要做考量。假如全體學生人數不是很多，則可採用普查方式，但假如人數很多，則可採用抽樣方式。但若採用抽樣，則需考慮分析需求以及誤差值之大小，以及作業上的方便等。另外抽樣設計往往需採用分層，以不同機率，隨機取樣方式。比如將學校依地區及學校大小分組，然後從各組中依比率隨機抽出適當數目的學校，之後再從學校抽學生。抽學生的方式，一種是將全體學生依個人背景分組，再從各組中依預定機率隨機抽出樣本。另外一種是以班級為單位抽樣，將被抽到的班級的學生全部納入評量。第二種方式比較容易執行，施測作業也比較方便，但因在同一班級中的學生同質性較高，在統計學上產生cluster effect的問題，造成分析上的困難，需要特別處理。目前兩種方式都有人用，TIMSS即是以班級為單位抽樣。

6.橫斷式或追蹤式評量:橫斷式的評量，係指在同一年度（或學年度）內評量不同年級或發展階段的學生，加以觀察或測量學生各項學習成就的情形，亦即參與的學生只做一次評量。而追蹤式的評量則需重複向同一群學生（全體或抽樣）做評量，連續觀察其在不同的年級或發展階段的學習情形。追蹤式評量若用適當同等質的測驗工具，則可檢測學習成長，而且也較合適做因果關係分析。不過追蹤式的評量在設計上及作業上都比較複雜，也容易發生樣本流失等情形，因此，這樣的長期計畫需要周詳的規劃以及政府的持續支持，以免有頭無尾，或事倍功半。

7.常模或效標參照方式計分（norm-referenced or criterion-referenced scoring）：評量分數的計算若採用常模參照方式（norm-referenced scoring），則標準是相對的性的，係根據某一個群體的表現所定的，學生的表現水準決定於該生與該群體比較的結果，因此，評量分數可以解釋個人在群體中的相對位置，一般係以百分等級或標準分數呈現結果。而效標參照（criterion-referenced scoring）方式計分，標準則是絕對的，著重於個人是否達到某種標準的程度（如是否學

會或沒學會某一課程項目），通常以傳統的百分制來呈現結果。目前評鑑都以常模參照方式計分，但NAEP最早計分方式則是採取效標方式，檢測達到標準的百分比。

8.每位學生考全部試題或部份試題:評量的內容很廣泛，所以需要很多試題來檢測，也可能需要很長時間來做答，所以為了考量學生作答時間，往往採用每位學生只做部份試題，只確切評估全體學生成效。此方式需要考慮如何分配試題以及如何採計個人分數。NAEP、TIMSS等都採用部份試題方式，所以若需採計個人分數，則需以統計方式估算個人分數，目前NAEP及TIMSS採用plausible value的方法，藉由每位學生既有測驗分數及背景資料來估算，過程相當複雜。

9.衡量個人或群體成績:目前國際上常見的評量，如NAEP、TIMSS、PISA等都非用來準確評量個人能力或學習成績，而是用來準確評估群體或主要組別之表現與程度。假如評量是用來準確評量個人能力，則需要求每位學生都做相同或等化的試題或透過複雜統計方式推估。國內國三的基測以及高三的學測和指定科目考試數於此種評量。

五、國內評量系統藍圖

由上可知，不同的評量目標需要不同的方式來達成。雖然每一評量方案可以有多元目標與功能，但是無法達成所有的功能，因此我們需要一完整的評量系統，以多項方案完成多元功能，提供完整豐富的資訊。此系統的構想分國內評量方案及國際評量方案兩部分，陳述如下。

國內評量方案含橫斷式及追蹤式兩種設計，各設計之細節，如目的、頻率、資料內容等，分列於表2。

表2 國內評量方案藍圖

評量方案	目的（註一）	普查 / 頻率 抽樣		資料內容（註二）	測驗方式
橫斷式					
1.基本學科能力（基測、學測、科考）	1,5,6	普查	每年	測驗+基本背景資料	標準測驗
2. 教育成果評量（四、六、八年級）	1, 2, 3, 5, 6, 7	抽樣	每2年	測驗+基本背景資料 +家長教授問卷資料 +學生問卷資料	標準測驗 +其他測驗
2. 大學畢業生（含碩博士）	1, 2, 3, 5, 6, 7	抽樣	每年	學歷+基本背景資料 +學生問卷資料	
追蹤式					
1.國小至高三	1, 2, 4, 5, 7	抽樣	每10年一新組	測驗+基本背景資料 +家長教授問卷資料 +學生問卷資料	標準測驗
2. 大一至大四	1, 2, 4, 5, 7	抽樣	每3年一組	基本背景資料 +學生問卷資料	核心能力測驗
2. 畢業生（含碩博士）	1, 2, 4, 5, 7	抽樣	每10年一組	基本背景資料 +學生問卷資料	

註一：1. 瞭解教學與學習成效現況; 2. 診斷教師教學及學生學習缺失; 3. 偵測趨勢與變化; 4. 衡量學生學習成長; 5. 檢視個別或組別差異; 6. 提供分級、分等與分組依據; 7. 探究與學習相關之因素。

註二：詳細內容原則請參閱上節四.4所述。

國際評量方案是為了與國際教育接軌並做國際比較，應繼續參與一些國際教育評量方案。主要方案包括：

1. Trends in International Mathematics and Science Study
2. Program for International Student Assessment
3. Progress in International Reading Literacy Study

可喜的是，上列所提方案中，除了國小至高三的長期追蹤評量之外，其餘都已推展。因此，若加上前述長期追蹤評量，國內之學生學習成就評量

系統即具備相當規模，未來此系統可向下延伸，涵蓋幼兒學習能力的評量，以便提供資訊，做為強化幼兒教育之依據。

六、結語

總而言之，評量是教學過程中一項必要工作。藉由評量可以考核學生學習成果以及探究相關因素，以便設計改進辦法，提升教學成效或品質。因此不僅教師們在課堂教學需要進行評量，使評量成為教學工作的一部份，而且學校及各級教育主管單位，都應推展評量，為教育成效把脈，掌握各級教育績效以及品質，將此評量結果做為教育決策珍貴的實證依據。美國的No Child Left Behind（2001）教育法案還特別要求各州執行評量，以評量結果來做為教育獎助與懲戒學校之依據。另外，由於評量之重要性，我們需要建立一套健全的評量機制來完成多元的目的與功能。本文所提之各項考量，希望能有助於此機制之建立。希望國內的未來的教育評量不僅都有明確的多元目標，發揮多元功能，而且也有完善的資料收集設計以及分析方式來配合達成多元評量目標。

目前國內的TASA已成形，在其設計過程中，也已將許多國際經驗納入，包括評量內容規範，評量實施程序以及附加資料之收集，都與國際接軌，其貢獻值得肯定。唯一值得商榷的是，各科評量應多久施測一次。美國的NAEP依不同科目性質，分二年、三年、五年甚至八年或十年才重覆一次，值得參考。（TASA原設計每年施測，目前已考慮隔年施測一次。）測驗方式亦宜加強高層次思考能力的檢測，如使用問答題、解答題、演算題及短文寫作等，以引導教學加強思考能力的訓練。另外有關技術性的報告，如試題內容規範、品質方析、測驗信度與效度等，亦應更加詳盡。

最近TASA亦考慮加上長期追蹤評量方案，進一步評量學生學習成長情況，以及探討歸因於教師及學校附加效益（added-value）和其它相關因素，包括各項教育政策及教育輔助方案之影響。假如這些構想能一一落實，則國內之教育評量工作，將能更上層樓，助長教育研究及教育改革工作，成為國

際一流典範。我們期待此項構想的早日實施。

作者註：本文為2006/9/24中華民國教材研究發展學會主辦之「學習成就評量與測驗」研討會演講稿，雖有少許更新，但原意不變。

參考文獻

國外學生學習成就評量方案網站：

1. Iowa Tests of Basic Skills （http://www.bjupress.com）
2. National Assessment of Educational Progress（http://nces.ed.gov/nationsreportcard/）
3. Program for International Student Assessment（http://nces.ed.gov/surveys/pisa）
4. Progress in International Reading Literacy Study（http://nces.ed.gov/surveys/pirls）
5. Stanford Achievement Test（http://harcourtassessment.com）
6. Trends in International Mathematics and Science Study（http://nces.ed.gov/surveys/timss）

國內學生學習成就評量方案網站

1. 台灣學生學習成就評量（http://www.naer.edu.tw）
2. 台灣教育追蹤研究資料庫（http://www.teps.sinica.edu.tw）
3. 台灣高等教育資料庫（http://www.cher.ed.ntnu.edu.tw）

盧雪梅（1998）實作評量的應許、難題與挑戰，**教育資料與研究**，20，1-5。

IAEP (1989). A World of Differences. Princeton, NJ: Educational Testing Service.

IAEP (1991). The 1991 IAEP Assessment-Objectives for Mathematics, Science, and Geography. Princeton, NJ: Educational Testing Service.

Linn, R. L. (1993). Educational Assessment: Expanded Expectations and Challenges. *Educational Evaluation and Policy Analysis*, *15*(1), 1-16.

Paulson, F. L., Paulson, P.L., & Meyer, C.A. (1991). What makes a portfolio a portfolio? *Educational Leadership, 48*(5), 60-64.

U.S. (2001). No Child Left Behind Act (2001) [Legislation], Pub. L. No.107-110, 115 Wtat.1425.