國民教育研究學報，第 2 期：1～26頁
民85年，國立嘉義師院國民教育研究所

從傳統到變通：教學評量的省思

黃　秀　文

國立嘉義師範學院

摘　要

本文是採文獻分析的方法，探討美國教學評量領域的變革。本文首先探討傳統評量的特色及衍生的問題，再述及傳統評量如何改進其缺失以及何種新式的變通性評量方法應運而生，本文接著介紹變通性評量的特色及文獻上對它的批評，最後再藉由對傳統評量及變通性評量基本假定的釐清，來評估兩種評量方式的適用範圍及可能的影響力。本文的結論是，我們應根據評量的目的－做比較或了解學生達到標準的程度－來選擇適當的評量方式；嘗試以單一評量方式來達到所有的評量目的，其結果極可能產生偏差。根據教育目的及學習理論，本文也建議變通性評量實有重視的必要。

壹、緒論

在美國教學評量的領域中，計量取向的正式標準化測驗一直是評量的主流，計分容易與客觀公平是此類測驗能取得優勢地位而且歷久不衰的主要原因。然而，近幾年來此類傳統評量的價值受到越來越多的質疑，此不僅促成了傳統評量技術的推陳出新，許多研究者及教師也紛紛尋求另類評量方式來評鑑學生的學習情形。美國此一評量變革的趨勢，即爲本文探討的重點。本文希冀此一他國經驗的探討，對於國內的教學評量，能有拋磚引玉之效。

美國的教學評量領域爲什麼會有這種趨勢？傳統評量的負面效應、新近學習理論的提出、政治社會力量的推動、以及種種新式教學法的出籠，應是主要的動力來源。

其實，傳統式評量對於教學與學習的負面影響早已引起甚多批評(黃政傑，民79; Shepard, 1989)。傳統評量基於行政及計分簡便的考慮，偏向封閉性考題，爲了配合此類考題，整個課程與教學因而窄化，學校及教師們也深受現實蹉絆之苦。許多教師及研究者認爲，要掙脫傳統評量的桎梏，有必要改進現行評量系統或建立一新的評量系統，來導引符合教育理念的教學。另外，Garcia和Pearson (1994) 也指出，傳統評量對弱勢文化及社會低階學生所產生的負面影響也是促進評量改革的趨力之一。他們指出，傳統測驗的題目常代表著主流文化及社會中高階層的價值觀，而此不利於來自弱勢文化及社會低階團體的學生；甚者，較之主流文化或中高社經背景的學生，來自於弱勢文化或低階社經背景的學生更常接受到低品質的教學：一種完全配合封閉式考題的僵化教學，於是惡性循環的結果，文化及社會階級間的差距日益加深。在目前強調多元文化的時代，此種文化上的不公平情形，常成爲撻伐的對象，傳統評量的改革也因此得到另一助力。

新近學習理論的提出，爲舊式評量的改進及新式評量的建立提供了理論架構基礎，也因而加速了教學評量的革命。新近學習理論強調，知識並非如物品可單方面地由一方傳達給另一方，學習也不僅是知識的累積。依據新近學習理論，當學習者與書本、教師、同儕、及周遭學習環境互動時，學習者會主動詮釋其在此互動過程中所得的訊息；知識即爲學習者主動建構所得訊息的結果，學習則爲學習者建構意義的過程，而學習脈絡中的人(包括學習者本身、教師、同儕等等)、事、物皆會影響學習者意義的建構(Rosenblatt, 1985; Shuell, 1986)。基於此學習理論，個體主動認知被視爲學習的核心，其背景知識及策略運用對學習的影響力開始受到重視，而學習環境的安排、活動設計、教師引導、同儕互動等等層面也受到關注。這些新的認知理念，不僅推動了各各學科的教學改革，也改變了長久以來偏向計量傳統的評量觀點，研究者開始嘗試找出能評鑑學生建構意義的過程與結果的其它評量方式。

而各種學科領域上的教學改革，也進一步促進了新的評量方式的發展(Glaser & Silver, 1994; Goodman, Goodman, & Hood, 1989; Simmons & Resnick, 1993)。許多改革者意識到，評量與教學息息相關，如果我們仍沿用傳統的封閉式測驗，教學上無可避免地會導向迎合封閉式測驗的教學型態，此時任何進步式的教學都可能在這種現實情境下，流於慘澹經營，甚至於無疾而終。相反地，如果我們採用較符合學習理論觀點的評量方式，評量與教學相互配合下，進步式的教育改革將得有更

多發揮的空間。因而嘗試進步式教學法的教師及研究者，也常根據其教學理念發展新的評量方法，期盼藉由教學與評量的相互爲用，使其進步式教學法發揮最大的功效。

另外，在體認到美國經濟競爭能力愈趨薄弱以及省視其它先進國家培育人才的教育體系之後，美國政府提出「America 2000」的教育目標，揭示教育應建立全國性的標準及評量系統，以普遍提昇人力素質，讓所有的學生在學習上皆能達到高標準(Gandal, 1995; Resnick & Nolan, 1995)。順應此政治社會的需求，許多教育人士及相關機構紛紛致力於擬定各學科應有的目標，及各年級應達到的水準，並且尋求建立一種高標準的評量系統來配合此教育政策上的革新。此評量系統強調提供較複雜的評量工作，以激發學生展現較高層次的思考及問題解決能力。

要之，由於新近學習理論及政治社會力量的影響，也基於推行新式教學法的需要及傳統評量本身的缺失，使得傳統評量面臨嚴峻挑戰，而此不僅促進了傳統評量本身的改進，也爲新式評量方法奠下了優勢利基。種種新式評量方法在此環境下應運而生，其名稱甚多，包括有實作評量(performance assessment)、眞實評量(authentic assessment)、卷宗（或作品、或檔案）評量(portfolio assessment)、動態評量(dynamic assessment)等等，這些新式的評量方法通稱爲變通性評量(alternative assessment)(莊明貞，民84)。從傳統標準測驗在教學評量領域的擅場到今日變通性評量的出現，使得評量領域上呈現傳統與變通對峙的局面，許多爭論的話題也因而產生。到底傳統式評量與變通性評量有何差異？他們個別的優勢及問題何在？他們眞的是「勢不兩立」嗎？這些即爲本文擬探討的重點。以下將分別敘述傳統評量的特色與問題、評量的變革、及變通性評量的特色與問題，並對傳統與變通兩種評量方式加以評估，而若干爭論的話題也將穿插其間予以討論。本文的目的不在爲若干爭議的論點提供解答，只是希冀文中的敘述及討論能夠爲教學評量提供更深一層次的思考。

貳、傳統評量的特色與問題

自五、六〇年代心理計量學開始興盛起，美國的教學評量也漸漸走向計量取向。

在本文中的傳統評量，指涉的即是此計量取向的標準化測驗，而本文將偏重有關成就測驗的探討。無可諱言地，此類傳統測驗仍佔目前美國教學評量的主流。究竟傳統評量的特徵爲何？何以它能成爲教學評量的主流？文獻對它又有何批評？以下將針對這些問題進行討論。

一、傳統評量的特色

以系統化的正式成就測驗而言，大部份都是屬於常模參照的選擇題式測驗。此類成就測驗通常是針對某個學科領域而設計，而且測驗內容傾向某些固定的型式(Garcia & Pearson, 1994)。如以語文學科爲例，測驗的內容大多分爲字彙、閱讀、文法句型等類目(有的測驗也有作文)；而在閱讀測驗上，大多設計幾篇短文，每篇短文之後再設計一些有關閱讀理解的單一選擇題，學生的閱讀能力即根據他們答案的對錯來評定。正式成就測驗強調「標準化」的歷程，在建立「標準化」的過程當中，測驗設計者首先根據雙向細目表及各種命題技術來編擬試題，然後再藉由預測的實施來考驗試題的信度與效度，設計者根據預測的結果來修正內容或增刪題項，以使測驗能達到一定水準的信度與效度；最後實施測驗時也有一定的標準程序，施測的情境也要加以適當的控制(郭生玉，民76)。時間上的限制也是標準測驗的特色之一，以美國研究所入學考試的GRE(Graduate Record Examination)爲例，其考試內容有七大項，每項測驗時間均以三十分鐘爲限。測驗時間的長短是依據預測時一定比例的受試者能完成測驗項目所需的時間來決定，此比例是由設計者依測驗的目的而擬定的。有些測驗也將時間列入區別能力的要素之一，受試者能力的評定不僅決定於答案的正確性，也決定於作答的速度；在此類測驗中，受試者的作答必須又快又正確，才能得到好的分數，受試者承受的緊張與壓力不言而喻。至於測驗結果的報告，其通常列出分項的分數、總分、以及代表受測者在團體中等級的百分位數。在常模參照的成就測驗中，其實分數本身並沒有任何意義，它是在與同群體的其他受測者分數比較後才有意義的，也因此大家關注的焦點是放在有「比較」意義的百分位數上面，而非分數本身。

在資料的量化及統計方法的運用下，傳統評量能有效率地呈現學生在團體中的相對地位。所謂「有效率」是指它施行容易，節省時間，而且有明確的數字提供清楚明瞭的評比訊息。此傳統評量特有的「效率」深受行政人員及決策者的歡迎，因爲它使

繁複的行政作業及決策作業變得較簡便，而且不會涉及是否達到標準的爭辯以及評分上主觀判斷的困擾，尤其用在大規模的評量時，它更是能節省相當大的成本。正由於傳統評量的「好用」，所以它在教學評量的領域上一直立於不敗之地，縱然變通性評量的來勢洶洶，傳統評量實仍爲主流。

二、傳統評量的問題

傳統評量雖仍爲現今教學評量的主流，然其對於教學及學習上的負面影響也常見於文獻之中，如不符合新近學習理論即受到許多研究者(如Resnick & Resnick, 1992; Valencia, Hiebert, & Kapinus, 1992)的批評。如前言所述，新近的學習理論強調知識是個體主動建構的，要幫助學生提昇學習的層次，有必要回歸到學習者本身，了解其建構歷程。傳統評量的比較性功能雖強，可以客觀顯示個體在團體中的等級地位，但對於學生解題、思考、策略運用等較過程性的學習則無法提供深入的訊息。學生的理解程度爲何、使用何種學習策略、如何與他人互動、學習失敗的原因爲何等等，這些其實是在教學第一線的教師們最需要的評量訊息，這些資訊有助於老師因勢利導、更適切地幫助學生提昇能力，然而傳統評量在這方面資訊的提供則付之闕如。換句話說，傳統評量的評比訊息對實際教學的改進並沒有太多實質的幫助。

此外，爲了計分的方便，傳統式評量常捨開放性的題目，而以封閉式的題目爲主，當中又以選擇題爲最主要的題目類型。封閉式的題目類型固然有其見長之處，但所能測驗出的知識及能力是有限的，而且易流於枝節技巧的評量，測量低層次及孤立的知識與簡單的推理，學科的眞正精髓反倒被忽視，也因此學生在測驗分數上的表現未必能反映他在此學科上的眞正能力。更嚴重的是，傳統評量常被用來作爲決定學生未來前途的關鍵性工具，基於學生未來發展的現實考量，教師在教學上不得不迎合傳統評量的內容。而由於傳統評量本身的種種缺失，教學爲了迎合評量也很容易產生偏頗的現象。當教學導向這些測驗型式與內容的灌輸和訓練時，教學無非是侷限自己於某一學習的層面，而忽視其它的學習面向；偏向學科枝節的介紹，忽略了學科眞正的宗旨。這也就是爲何傳統式評量一直被批評爲窄化學習，僵化教學，不能反映學生在某學科上的眞實能力(Shepard, 1989; Glaser & Silver, 1994)。此種考試引導教學的現象，不僅牽絆了教師的教學理念，誤導了學生對學習的看法，也產生了所謂的測驗分數污染(test score pollution)，也就是說老師將教學的重心置於測驗題式易出

的考題，測驗分數顯然會受到教學因素的影響，而不能公正地代表學生的眞正能力(Darling-Hammond & Wise, 1985; Haladyna, Nolan, & Haas, 1991; Shepard, 1989)。教學過度強調內容，強調灌輸，學生也會漸漸地視知識爲外在的，不是自己可以建構思考的，於是慢慢養成被動的學習習慣，坐待別人給予他「知識」。

再者，傳統評量的評比色彩強烈。測驗後將分數統計出來，對錯好壞分清楚的評比等級之後，評量工作即功成身退。學生從評量結果中得到的訊息，只有代表其學習的數字，以及告知其等級的百分位數，學生無法更進一步了解其學習上的優勢以及什麼能力仍有待提昇。同樣地，教師也只能從這表面的分數知道學生「分好壞」，無法進一步得知學生各個層面學習的情形。其實，分等級是一種無可避免的社會選擇性功能運作，它仍有其積極的一面，並非完全不好，而其正負面影響端視它是如何被使用的。善用之，它可以作爲個人求進步或抉擇方向的指標；誤用之，則不但會導引學生對學習產生錯誤的定義，也容易起標幟作用、傷害心理、扼殺學習動機。很不幸地，它被誤用的成份居多，而主要原因是在於它的功能被誇大了。傳統評量的評比結果常被單獨用來作爲對學生未來有重大影響的安置或選擇性的決定，因此造成傳統評量方式獨大，評比功能被過度強調，測驗分數和等級被誤用爲代表學生全部的能力(Taylor, 1994)。學生常因測驗評比的結果被安置於高低能力組別或一般及特殊教育的班級，而許多研究(如Allington, 1989; Delpit, 1988; Stuetzel, Shake, & Lamarche, 1986)顯示，被安置於低能力組別或特殊教育班級的學生接受到的常常是低品質的教學；他們被問低層次的問題，被教導簡單的知識與技巧，接受教師爲中心的教學。換言之，學校沒有提供這些學生有意義的、能激發思考的教學，而這也無形中剝奪了這些學生的教育機會，造成能力差距愈來愈大的惡性循環。更嚴重的是，此種安置與低品質的教學，常會產生負面的標記作用，對學生的心理造成很大的傷害。

另外，有些研究者(如 Garcia & Pearson, 1994; Mercer, 1989)也批評傳統式評量反映出語言、階級、及性別上的偏見。如 Garcia和Pearson(1994) 即以早期智力測驗的發展爲例，他們指出當 Terman 轉譯 Binet 的法文智力測驗爲英文時，Terman 將問題加以改變以符合美國的知識觀與價值觀，而其常模的建立是以中產階級的白人學生爲對象；其後由於施測結果發現女生的分數高於男生，測驗設計者基

於女生不會比男生聰明的假定，將那些女生得較高分數的題項加以刪除，但奇怪的是，測驗設計者並沒有修改或刪除城市學生比鄉下學生得較高分數的題項，很顯然地，設計者也預設了城市學生比鄉下學生聰明的假定。在這一段智力測驗發展的歷史過程中，從其樣本的選擇以及內容題項刪改的依據原因，我們不難看出智力測驗從一開始的發展上即透露出種族、性別、和階級的歧視，而依循智力測驗測量模式的學校成就測驗也遭受到同樣的質疑。其實，標準化測驗的常模化過程以及對極端分數的忽視，基本上不利於來自次級文化團體的學生。而事實也顯示，經過正式測驗，來自於非主流文化的學生大部份被安置於特殊教育課程或低能力組別，似乎學校特殊教育課程是專門爲非主流文化學生而設置的。這種現象不僅對這些學生不公平，也加深了種族與文化之間的鴻溝，而傳統測驗在這方面也難辭其咎。

參、評量的變革

由於傳統評量引發了前述的種種問題，對教學及學習的影響有負衆望，再加上緒論所提及的各種勢力的配合，要求改革評量的聲音因此受到重視。有的教育工作者從事體制內的改革，從改進既有的傳統評量著手，來呼應要求改革評量的聲音；有的教育工作者則向外尋求新式的評量方法，以取代傳統的評量。以下即從這兩方面來加以探討。

一、傳統評量的改進

面對種種的批評，測驗設計者也不斷地嘗試各種方法來改進傳統的評量。例如爲了提供更深入、更精確的解釋訊息，測驗設計者發展出試題反應理論(item response theory)、試題關聯結構(item relational structure)等等的分析方法來分析試題，企圖改進傳統測驗理論的一些缺失，如樣本依賴、誤差指標忽視個別差異、忽略相同的總分下可能有不同的反應組型等等(劉湘川、簡茂發、林原宏，民83)。這些分析方法的確修正了傳統測驗理論解釋分數時的一些偏失，不過，這些量的方法仍有其不足之處，其模式和統計方法也尚在改良發展之中。

針對不符合學習理論的批評，測驗設計者也著手從事測驗內容及測驗型式的改

革，以符應學習理論。以語文科爲例，密西根州及伊利諾州即根據學習理論發展出與傳統閱讀測驗迥異的新式閱讀成就測驗(Valencia, Pearson, Peters, & Wixson, 1989)。該測驗強調意義的建構，高層次的思考、並且重視學習者在思考過程中所運用的背景知識及策略方法。依此目的，測驗選擇了較長、較眞實有趣的文章，以便提供充分的脈絡來幫助受測者建構意義；而測驗題目的設計不僅重視文章內容的了解，也強調超越文章內容的高層次思考與推理。此外，測驗也設計了一些問題來了解學生對於文章內容的背景知識、閱讀時的策略運用、以及閱讀習慣和態度，這些問題不是用來評分，而是用來輔助說明學生閱讀理解的分數。伊州及密州改良舊式測驗以反映學習理論的作法實在值得喝采，不過Garcia和Pearson (1994) 指出一現實的問題。傳統閱讀測驗的內容涵蓋難易考題(如三年級的學生可閱讀到一、二年級程度的簡單文章及四、五年級程度難度較高的文章)，縱然是低成就學生，仍能回答一些簡單文章的問題。伊州及密州設計的新式閱讀測驗則以較長的文章取代傳統閱讀測驗斷章取義的短文，以較開放性的題目取代單一選擇題。其結果是，文章長雖能提供給學生較完整的文章脈絡，然而也因爲文章長，因此只能包含兩三篇文章，文章的篇數有限，選取的文章通常只配合欲評量的年級程度，所以三年級的學生所閱讀的很可能即爲兩三篇相當具有挑戰性、屬於三年級程度的文章，再加上需要思考的複選題及申論題，新式閱讀測驗的難度無疑地大於傳統的閱讀測驗，而此造成了低成就學生抗拒及非正常的表現。

此外，測驗設計者也針對最受批評的文化偏見問題謀求改進。例如爲了避免常模建立過程中產生的文化偏見，有些測驗專家致力於發展以某一文化團體或階級爲常模的測驗，不過，此類測驗由於與其它團體的比較有困難，再加上無法明確地區別學生屬於何種文化團體或階級，應用上並不普遍(Mercer, 1989)。另外，許多測驗學家也嘗試從測驗的主題(如涵蓋各式各樣的主題來減低文化背景知識的影響)、題目設計(如問題必須是閱讀文章後才會作答，以避免文化因素的干擾)、語言(如設計不同語言的測驗版本)、統計方法等等，來控制文化因素的干擾，不過，這些改進的方式仍無法獲得普遍的認同(Garcia & Pearson, 1994)。其實，包容各文化族群的歧異性，原本就不是件容易的事，更何況是提供他們一毫無文化偏見、客觀公正的評量？

二、變通性評量的興起

如上所述，由於種種勢力因素的交織配合，不同於傳統評量的新式評量方法相繼而出。這些變通性評量方法的名稱甚多，文獻中常見的包括有眞實評量(authentic assessment)、實作評量(performance assessment)、卷宗(或作品、或檔案)評量(portfolio assessment)、動態評量(dynamic assessment)等等。變通性評量的內容及方式均不同於傳統評量，其規模有的是以學校及學區爲單位，也有及於全州的大規模評量(如亞利桑那州、加州、馬里蘭州、肯塔基州等等，皆採用變通性評量於閱讀、寫作、自然、或數學科目上)。變通性評量通常先依循新近學習理論建立學科的學習架構，再根據此架構設計教學及評量活動，評量活動中也設計有評分的標準和等級來評量學生進步及達到標準的情形(Greenwood, 1993; Comfort, 1994)。以下將簡介幾種常見的變通性評量：眞實評量、實作評量、以及卷宗評量。

1. 眞實評量

眞實評量是在實際的教學活動中進行，教學即評量，評量即教學，兩者密切配合；教師在教學活動中是透過觀察、與學生的談話、以及學生的作品，蒐集各個學生學習情形的資料(Atwell, 1987; Calfee & Hiebert, 1991; Cambourne & Turbill, 1990)。由此我們也可以看出，眞實評量與教學同步，是一個不斷的歷程；而教學活動也不是固定不變的，它會依評量的訊息不斷地做調整。從 Garcia 和 Pearson (1994) 區分眞實評量與實作評量的敘述中，我們更可以看出眞實評量的特性。Garcia 和 Pearson 指出，眞實評量的工具和活動與實際教學息息相關，它是完全由教學的老師所設計，評量的是實際班級教學脈絡下學生的眞實表現；而實作評量可嵌於實際教學脈絡中，由授課的老師所設計，也可以施行於教學外的地點，由外面的專家來設計與評量(如考試)，但評量活動皆是與實際教學配合的實作活動。

2. 實作評量

實作評量是根據學生的實際表現所做的評量，其方式可藉由直接的現場觀察與判斷，或間接地從學生的作品去評判(Wiggins, 1989)。Resnick and Tucker (1990) 指出實作評量有三種型式：實作任務 (performance tasks)、實作考試 (performance axams)、和卷宗評量。實作任務及實作考試的評量方式是給予學生一項任務或工作，例如科學實驗、數學解題、寫作、或口頭報告等等，然後再根據學

生的實作過程及成果加以評量。實作任務及實作考試兩者的差異在於，實作任務即爲日常教學活動的一部份，其評量是在實際的教學脈絡中進行；實作考試則類似聯考，是在一段特定時間與特定地點進行評量。至於卷宗評量則是學生平日代表性實作成品的收集，其細節容後再詳述。以自然科爲例，目前加州正在試行以實作評量爲主的全州性評量，評量幼稚園至12年級學生對於自然科四十餘個核心概念(big ideas)的學習情形。其評量方式包括實作考試、改良式的選擇題測驗、開放式的問答題、和卷宗評量；在實作考試中，評量者提供給學生相關材料來進行若干小的實驗(可安排一天或數天進行，也可安插小組討論的活動)，學生將其實驗的程序、觀察到的資料、以及分析的結果記錄於一測驗小冊子(test booklet)上，評分者再根據若干評分標準來評分(Comfort, 1994)。

3. 卷宗評量

卷宗評量爲最常使用的變通性評量，通常其它類型的變通性評量，諸如眞實評量、實作評量、動態評量等等，也會用到卷宗評量。在卷宗評量中，學生不斷地收集其作品(如閱讀心得、研究報告、實驗報告、作文、日誌等等)於文件筴(portfolio)中，然後在學期中定期(約二至四次不等)整理反思這些作品，最後學生再選擇自認爲最具代表性的作品編輯成最後的portfolio，作爲評分的依據；最後的portfolio除了學生的作品外，教師通常也會要求學生加上自己對於學習過程的看法以及爲何選擇這些作品的反思(Arter & Spandel, 1992；Farr & Tone, 1994)。在學生整理反思及選擇作品的過程當中，教師會提供學生一些關於作品選擇、編輯格式、及評分標準的參考，並提供師生討論、同儕討論、及自我評鑑的機會，以引導學生自我提昇，提高作品的品質(Arter & Spandel, 1992; Farr & Tone, 1994)。至於卷宗評量的實施情形，在美國除了許多學校及學區(如匹茲堡學區的五年試行計畫)個別試行外，佛蒙特(Vermont)州更試行全州性的卷宗評量於寫作及數學課(Gomez, Graue, & Bloch, 1991)。關於評分的過程，通常是由學校、學區、或全州組成一評分委員會負責評分，評分者或爲學校教師或爲校外專家，或兩者皆包括。

肆、變通性評量的特色與問題

變通性評量爲新興的評量方式，其評量的方式甚多，但大抵皆依循新近學習理論的架構而設計。以下將介紹文獻對於變通性評量的正面評價，並探討其在教學與學習上可能產生的問題。

一、變通性評量的特色

變通性評量強調教學與評量應密切配合，並以思考及問題解決能力的提昇作爲教學與評量的目標。變通性評量的倡導者認爲，對學生日後的學習最有幫助的，不在於給予學生一大堆的知識或概念，而在於思考及問題解決能力的培養；不應強調分類分等級，而應強調個體主動認知的重要性，引導學生朝向標準，不斷地自我提昇。歸納 Baron (1990), Farr 和 Tone(1994), Glaser 和 Silver(1994), Wiggins(1989) 的觀點，變通性評量有以下幾個特色：

1.銜接教學與評量

變通性評量以學生在實際教學活動中的各種表現爲評量依據，也就是說，其評量是過程導向、持續性的歷程。此教學與評量的密切配合可以對學生的學習情形提供較全面性的、完整的、深入的訊息。此訊息可以幫助老師更了解學生的學習優勢及問題，掌握學生眞正的能力及進步情形，使老師能在教學上作適當的調整來幫助學生解決困難、提升其學習水準。而也惟有重視過程的評量，學生才有機會去反思自己學習上的問題，省察如何在學習上求進步，而這些也才是眞正的學習。考試及標準測驗在這方面資訊的提供則顯得較爲薄弱，測驗題目的對或錯可顯示出學生那裡會或那裡不會，但是會的部份的理解程度到底是多少，不會部份的癥結問題到底是出在那裡，則不得而知，此類訊息無論是對於老師的教學或學生的學習都沒有太多實質上的幫助。

2.使學習更有意義、更深入

變通性評量強調教學與評量的內容應爲重要的、完整的概念，而非瑣碎知識的累積；應重視思考與問題解決能力的培養，而非低層次的記憶與歸納。它的目的在幫助學生獲得完整、有意義的概念，增進表達技巧及運用策略的能力，並激發學生從事較

複雜的深層思考。所以變通性評量著重脈絡下有意義的學習，在教學與評量的過程中，它鼓勵學生主動探索、深入思考、並學習表達。此種評量方式有助於提昇學生的思考及問題解決能力，使學生的學習更有意義、更爲深入。

3. 強調學生知道什麼、能做什麼

變通性評量的重心，不在於偵測學生那裡做錯了，而在於強調學生知道什麼、能做什麼、及如何再進一步知道得更多、做得更好，簡而言之，其精神是「你會做很多事，你還可以學會更多事」。變通性評量鼓勵老師及學校避免強調「分類」及「標誌」色彩強烈的評比與競爭，對於學生嘗試去做好某一件事的努力(縱然尚未達到預期的標準)應給予正面的回饋。以學習理論而言，變通性評量似乎較符合學習理論中的公正性或正當性，亦即努力是有收穫的。Case (1995) 也指出，如果學生的主動探索及自我反思能受到鼓勵和肯定，學生才能得到學習的樂趣，才會對學習這件事抱有希望，產生信心，藉此學生對於知識的興趣與探索也才得以持續，不斷地求自我提昇。

上述部份是從變通性評量的理念及作法來推知它對於教學與學習上的正面影響，另外一些研究的結果也支持變通性評量，尤其與傳統的封閉式測驗比較時，變通性評量對於教學及學習的潛在影響更是受到肯定。以教學而言，許多教師表示變通性評量促使他們改變了課程內容與教學方式(Gomez et al, 1991; Lambdin & Walker, 1994)。教師們指出，他們將更多的教學時間用於指導思考或問題解決的策略，用於探討重要概念及其彼此之關係，而不再是零碎知識的解說與記憶。他們比以前花更多的心思去觀察與分析學生的學習情形，去反思自己的教學成效。他們也安排更多小組討論的活動，並提供給學生更多思考及表達的機會。以學習而言，研究(Comfort, 1994; Jones, 1994)發現，當學生被賦予較多自主空間，並學習著對自己的學習負責任時，學生對學習變得較積極，較有自信心，敢於探索，在學習上也表現得更好。而學生也表示，變通性評量方式較「友善」，較有趣，而且較具挑戰性。

不過有一點要提及的是，變通性評量尚是一新的領域，探討其對於教學與學習之影響力的相關文獻並不多，而這些相關文獻的資料大部份是來自於教師與學生的訪談，量的實證性研究相當缺乏。因此，變通性評量對於教學與學習的影響力究爲如何，仍待更充份的研究證據來確定。

二、變通性評量的問題

雖然變通性評量能顧及許多傳統的考試或測驗所無法評量的學習面向，但是變通式評量也並非毫無瑕玷。如上所述，支持變通性評量的文獻，大部分都是根據新近學習理論提出一些教學模式，再依據此模式發展出各種變通性的評量方法，然後再藉由訪談教師或學生的資料來證明變通性評量的潛在優點。至於變通性評量如何建立其信賴度的技術層面，它影響學習的深廣度到底如何，以及它與傳統式評量在學習結果上究竟有何根本差異，這些方面的相關研究至今仍不多見。因此，變通性評量常被視爲一種主觀、不夠嚴謹的評量方式，尤其當變通性評量應用於較大規模(如全縣市或全國性)的評量時，其信度、效度、可行性等等常受到質疑。

以信度方面而言，評分者的評分是否一致爲變通性評量最基本的信度。評分者是否同意評分的共同標準？他們是否對同一學生的作品或表現評相同或相近的分數和等級？此類評分者的信度有建立的需要，否則無法提供學生或學校正確的訊息，因爲學生得到的分數或等級可能只是代表評分者個人主觀的印象，而非學生眞實能力的表現。那變通性評量在這方面的信度爲何？Herman 和 Winters (1994)在探討 46 篇有關卷宗評量的實證性與質的研究報告後指出，46 篇中只有 13 篇文獻敘述有關評分者的評分一致性，而在此 13 篇中所得的評分者信度也不盡相同。以佛蒙特州(Vermont)、匹茲堡(Pittsburgh)、及一小學的卷宗評量爲例，佛州的評分者一致性不一，隨著分數合計方式的不同而不同，因此被質疑其資料無法正確分析出學生作品的水準到底屬於那一層次，也無法做學區之間和學校之間的比較，然而匹茲堡及一小學的評分者信度皆達高水準。Herman 和 Winters 發現，評分者信度的高低決定於卷宗的內容類目是否一致、評分標準是否具體明確、以及評分者所受的訓練及合作情形。他們也同時指出，很少文獻探討評分者評分一致性以外的其它重要信度指標，如分數在不同的時間及教學情境下的穩定性、分數在不同的評分團體之間的穩定性等等。

以效度方面而言，我們要問的是變通性評量的結果是否能眞正透露出所要評量的能力或潛能。探討變通性評量效度的文獻通常將變通性評量結果與它類評量結果相比較。如果變通性評量的目標是在於激發與評量學生的思考能力與問題解決能力，以計量學的角度來看，考驗效度的方式可將變通性評量的結果與另一同樣評量思考與問題

解決能力且具公信力的評量結果相比較，或將變通式評量結果與測量非思考能力與問題解決能力的其它評量相比較，若前者的相關度高，後者的相關度低或無相關度，則表示變通性評量結果具有某層面的意義(Herman and Winters, 1994; Shavelson, Baxter, & Pine, 1992)。Herman 和 Winters (1994)探討與此方面相關的兩個研究，在一研究中，數學卷宗評量的結果與其近似的數學寫作教學評量的結果有中度的相關；當與其較不相同的數學選擇題式測驗結果相比較時，竟然也具中度相關，而非無相關或低相關。在另一研究中，研究者比較卷宗評量的結果與標準測驗評量的結果，研究發現三分之二在卷宗評量中被列爲優等的學生，在標準測驗中將不會被評爲優等，換句話說，卷宗評量與標準測驗的評量結果並無相關存在，而且相對之下，學生在卷宗評量中的表現似乎較佳，有較多的學生列於較優的等級，而在標準測驗中，學生的表現似乎較差。那麼，到底變通性評量的效度爲何？從極有限而且結果不一致的文獻中，我們實在無法一覽端倪。

另外，文獻(Gomez et al, 1991; Herman & Winters, 1994)也指出，變通性評量強調的同儕討論與師生互動，常導致學生能力的高估(如前段提及的第二個研究即發現到此問題)，而這對於變通性評量的效度不無影響。變通性評量強調建立支持性的教學情境，希冀藉由同儕及老師的支持與協助，提高學生的學習興致，並誘發出學生最好的表現。但有些研究者質疑，額外的幫助可能造成學生能力的高估，而且那些獲得較多同儕或老師協助的學生，豈不是在評量過程中較佔優勢嗎？他們也指出，此類問題在作重要擇才決定的大規模評量中，將會益發明顯，因爲評分者所作的評量決定可能會因學生受幫助程度的不同而產生偏頗，而無法代表學生個人眞實能力的表現。

上述是若干研究者根據計量學的觀點所提出的批評，除了信度效度之外，變通性評量的可行性也備受質疑。許多教師表示，變通性評量雖然對教師的教學以及學生的學習多所裨益，但也相當費時費力(Gomez et al, 1991; Lambdin & Walker, 1994)。老師們指出，比起以往，他們要花更多時間去思考如何設計能提昇思考與問題解決能力的教學與評量活動，如何擬出評量的標準並應用於學生作品或表現的評分上，及如何引導學生了解及達到教學與評量的標準。他們也必須時時反思，如何因應學生新的需要而調整其教學與評量。教師們更表示，在學生作品及表現的評分上，他們尤其須要花費更多的時間。

另外，專業師資的配合也爲變通性評量的可行性投下變數。以佛蒙特州爲例，在其第一年的試探性實施中，其數學卷宗評量所得的結果無法加以評分，因爲教師們的教學中並沒有包含有關問題解決的教學活動(Herman & Winters, 1994)。Au(1993)的研究也有類似的情況發生。Au 試行卷宗評量於語文科的作文教學中，一學期後學生學習的評量結果普遍低落，研究者發現此現象的原因主要是來自於教師不熟諳變通性評量的理念，教師們雖採用卷宗評量，但其教學仍屬於傳統教學。由這些例子可知，要發展出好的變通性評量需要教師們高度的專業知能，因爲這其間牽涉的不只是評量方式的改變，而是整個教學內容、教學方法、以及師生互動關係的改變。變通性評量需要具備有教學與評量方面專業知能的教師，才能導引學生進入更深層次的認知活動及評量活動；缺乏專業知能的教師，變通性評量的可行性及實施效果皆會大打折扣。

此外，大規模變通性評量所需的成本也是其可行性的阻礙之一(Cole, 1988)。龐大師資及評分者的訓練，大規模複雜性評量活動的設計與時間的安排，多種評量資料的紀錄、儲存、與評分的行政管理...等等，這些行政、人事、專業發展、與技術改進所須的成本相當龐雜，是爲變通性評量未來施行發展的一大挑戰。

很明顯地，以計量學的角度來看，變通性評量若要應用於有關重要擇才決定的大規模評量時，其信效度建立的技術層面和可行性的確有許多爭議之處，而其對教學與課程的影響，也因缺乏充裕的計量方面資料而被批評爲證據薄弱。Herman 和 Winters (1994)建議，如果變通性評量的評量項目及方式能制度化與標準化，也許是解決這些技術性嚴謹問題的最好方式。然而，諷喻的是，當變通性評量朝向大規模與制度化發展時，它的結果無可避免地會關係到學生未來的前途，而此關聯常使變通性評量喪失其「最具價值」的優點之一：培養學生主動探索、自我反思的學習態度(Darling-Hammond, 1994)。爲何如此？Case (1995)指出，要幫助學生能自主支配其學習，需要老師與學生共同合作，時常溝通協商(negotiation)，協助學生自訂目標並提供機會使學生不斷地提昇其目標的標準。如果爲了建立客觀性或爲了適用於較大規模的評量，而將變通性評量植基於制度化的評量系統中，其評量標準將會由外在的權威所決定，此時老師的教學及學生的學習無可避免地會以這些外在標準爲依歸，而這無疑會削減學生自主探索的空間，「窒息」了變通的美意，標準也很可能在不正常教學下，失卻其眞實意義。此時老師與學生之間的溝通協商也顯得多餘，因爲

目標早已爲他人所決定，溝通協商的眞正意義已不復存在。此對教學的影響是，老師可能教孤立的實作工作，而非引導學生了解學科領域的完整概念；老師可能會支配學生的學習，而非引導學生自主支配其學習(Darling- Hammond, 1994)。對學習的影響是，當學生鑽研於外在的標準或遊戲規則，並根據這些標準或規則來從事學習時，其對自己的學習將慢慢地喪失主控權，也不會學習到或意識到自主學習的重要性，而此現象與傳統評量中學生爲考試而讀書所產生的負面學習情形並無二致(Darling-Hammond, 1994)。而學生非但在學習過程中不自主，對於學習結果也無法自己控制，交了考卷或作品，或執行完一件學習任務，其結果就聽由老師或評分者安排了，這種學習情形就如同賭注一般，是那麼地不可控，學習也因此顯得遙遠不貼近，枯燥無趣味。

以Freedman(1995)的研究爲例，Freedman透過問卷、觀察、和訪問，探討英美兩國6-9年級各四個班級的寫作交流活動。兩國皆採用卷宗評量，但在英國卷宗作品的優劣與9年級學生能否進高中有決定性的影響。結果發現9年級的英國學生在寫作上不似美國學生及其它年級的英國學生那麼投入，文章種類不似其它班級來得多，也顯得較沒有信心自如地發抒自己的見解。受訪的9年級英國教師表示，爲了幫助學生達到標準，他們無形中取代了學生的責任，他們選擇了題目並提供明確的作文指導；學生也表示，寫作是爲評分者而寫，爲某些標準而寫，是索然無趣的。換言之，當變通性評量朝向大規模與系統化發展時，其評量結果與評比功能無形中被誇大，學生可能會過度倚賴外在的標準，學習上傾向以老師或其它權威的判斷爲依歸，而此可能與變通性評量的精神－培養自己摸索、自己判斷、自我評量的獨立學習態度－背道而馳。

那難道變通式評量只限用於教室教學的評量，而無法提供大規模評量所需的評比訊息？其實，變通性評量與傳統評量的典範不同，我們以實證典範中建立信效度的技術層面以及評比的價值觀來評斷變通性評量是不公的。我們應以其它方式來建立變通性評量的可靠性及可行性，例如省視其預設的學習目標是否代表學科的中心概念、設計的教學及評量活動是否能反映這些學習目標、評量標準是否恰當、評分是否一致等等(Lynn et al, 1991; Trevisan, 1991)。況且，變通性評量受歡迎的原因之一，正是因爲大家越來越質疑那些計量技術的價值，而贊同變通性評量在課程、教學、及學習上的理念。目前雖無充裕的計量證據支持變通性評量對教學與學習的正面影響，但

質的資料所顯示出的結果，以新近學習理論去衡量實深具說服力。但是，正如Herman和Winters (1994)所提出的，如果我們要以評量的結果作爲某方面選擇的重大決定時，變通性評量如何在技術上建立其信賴度將會成爲關鍵性的論點。因爲面對攸關重大的社會性擇才決定時(如大學選擇人才的決定)，社會大衆想知道的是提供作決定的評量訊息是否正確、可以信賴。那到底變通性評量可行不可行？本質上，這些問題牽涉到傳統評量模式與變通性評量模式的基本假定，以及不同的團體對評量訊息的不同需求，以下部份將針對這些層面進行探索。

伍、傳統與變通性評量之評估

一般人咸認爲，倚重標準化測驗的傳統評量才是眞正公平客觀的評量方式，雖然其測驗分數不能絕對準確地代表個人的能力，但也不會太離譜，而其客觀性、公平性、及效率，則是其它評量方式不可及之處。不過，荷蘭及美國的例子可以說明此類觀念有欠完善(Resnick & Nolan, 1995)。在美國，標準化測驗現仍爲教育上安置及擇才的主要依據，然而其學生的素質卻有每況愈下的趨勢。荷蘭的教育一向斐譽國際，在國際性的學業競試上，該國學生的表現也一向名列前矛，然該國並沒有標準成就測驗，荷蘭是以考試作爲區分人才的準據，但其考試評量的方式傾向變通性評量。事實上，傳統評量有其價值，但若以其爲唯一「合法」性的評量方式，則無異將傳統評量導入歧途。美國目前要求另類評量的呼聲高漲，常模參照成就測驗的不當使用實爲關鍵因素。同樣地，變通性評量若實施不當，極可能落入傳統評量的窠臼中，同樣承襲舊式評量在教學與學習上的負面作用，而遑論能支持新式教學法的改革。要避免誤用傳統及變通性評量，我們有必先了解傳統評量和變通性評量理論上的基本假定及目的。惟有釐清兩者間不同的基本假定與目的，我們才能評估其適用範圍及可能的影響力。

傳統評量是屬於測量模式，Taylor (1994)指出測量模式的基本假定來自於特質理論(trait theory)，而此理論的基本假定有三:(1)人類在許多特質(如身高、體重)上具有一致性及穩定性的個別差異。計量學家們更以統計方法發展出鐘形曲線(nor-

mal curve)來代表人類在某些特質上的穩定性差異，並以之作爲區分人類能力的基礎。(2)這些特質是可以測量的，我們可以發展工具來客觀地測量人類在這些特質上的個別差異現象。要達到客觀評量的目的，測量工具都要經過信效度的考驗。影響測量工具的信度包括有測驗項目的鑑別度(使受測者不會傾向得低分、高分、或相同分數)、測驗項目的獨立性(使某一項目上的作答不會影響到另一項目上的作答)、測驗的長度(愈長，可靠性愈高)、施測的程序等等。至於考驗測量工具效度的方法則有內容效度、效標效度和構念效度(郭生玉，民76)。(3)個體在某一特質上的表現，可藉由其類似的群體在相同特質測量上分配的比較，而呈現出來。爲了使某特質的比較具有可靠性，計量學家發展出種種統計方法(如集中量數、變異數、信賴水準等等)來增進測量的準確性。獲得這些統計數字的程序，成爲測量過程的必備要素，也是考量測量是否嚴謹的決定因素。

承上所述，測量模式的評量方法是植基於人類特質有個別差異的信念，以及對於測量特質差異之種種統計方法的了解與信賴。此測量模式後來爲教育學者所採用，並用於學業成就測驗的發展。換言之，學業成就被視爲一種穩定性的人類特質(trait)，而且假定了個體在此特質上有一致性的個別差異現象(呈常態分配)。Taylor (1994)對此提出批評，她認爲學業成就是目標導向的、學習的、會變的，並非是一種相當穩定的、呈一致性分配的特質。她也指出，雖然測驗題目原先是根據學科目標設計的，但在後來題目的取捨上，主要是根據此題目是否具有鑑別度，而非其是否代表某學科領域的重要標準或目標。換句話說，許多很好的題目，很可能在建立鑑別度的過程中被淘汰了。而採用測量模式的學習成就測驗也透露出，測驗的主要目的是在於分高低等級；受測者的分數本身是沒有意義的，必須透過與他人比較後才有意義。在此種評量模式中，建立學科的重要學習標準顯得不重要了，因爲「優秀」或「成功」是決定於學習者的等級是否勝過其他學習者，而非學習者是否達到標準或期望。

至於變通性評量，其基本上是屬於標準模式。Taylor (1994) 指出標準模式的基本假定有四：(1) 我們可以建立共同的教育標準作爲努力的目標。在標準的建立上，過去的標準參照測驗強調學科的內容，目前標準模式的提倡者則重視學生的實作表現。標準是對於學生實作表現的要求，也就是對於學生在作品或實際執行某一工作時所應展現的知識技巧及過程方法上的要求。(2) 大部份的學生可以達到這些標準。不似過去的標準參照測驗刻意將評量獨立於教學之外，現今的標準模式提倡者強調評量

應與教學合一，而且允許較長的時間從事學生的作品或實作表現資料的收集，而不是在某一關鍵時刻實施評量。教學與評量合一以及較長時間的評量過程，提供機會給學生去內化(internalize)標準、追求標準，也給老師機會去運用不同的教學策略以幫助有個別差異的學習者達到相同的標準。換言之，依據標準模式的信念，只要我們提供成功的教學策略，大部份的學習者都可以達到期望的標準，其成就的分佈也將大不同於鐘形曲線的常態分配。(3) 相同的標準可以透過不同的學習表現方式呈現出來。相同的標準並不意味著要求學生都要有相同的表現，學生選擇的作品或實作表現的方式會依其興趣或背景而不同，但我們仍可根據標準分辨出何者的表現有較高的水準。Taylor (1994) 舉音樂競試爲例，雖然每位表演者選擇的曲子及樂器不同，但評判者仍可根據標準分出高下。(4) 經由訓練，教育者可以內化(internalize)標準，而且可以對不同的學生表現做出公正、一致性的判斷。要獲得深入細致的訊息，有賴人的省察判斷，而非客觀的工具(Glaser and Silver, 1994)。但在信賴人的判斷之餘，也要承認人有主觀偏見，因此標準模式的提倡者強調評分者的訓練，以增進評分者在學科及評量上的專業知能，裨益於公正的判斷。

很顯然地，植基於標準模式之變通性評量的目的是在於了解學生有否達到標準。雖然變通性評量基於評分方便及公正性的考量，也根據標準設立了一些評分的規準及等級，但此等級不是與同年級的學生常模一系列的「典型」表現比較後所得的等級，換言之，此等級並非常模參照。學生的等級是用來說明他在某某標準下的表現水準，而不代表能力的高低；它可幫助學生自己及教師了解學生是否達到標準、那裏需要改進，但無法用它來做一群體中各成員能力高下的比較。要求變通性評量以種種技術來達到客觀評比的功能，其結果極可能使變通性評量「失眞」，提倡者不能不審愼。而事實上，「比較」在變通性評量中並不是那麼重要，因爲變通性評量基本上認定大部份的學生皆可達到標準，雖然有些人較快，有些人較慢。如果變通性評量強調比較、分等級，等於違背了其基本假定，走向了測量模式，屆時同樣會重蹈傳統評量問題的覆轍，縱然其評量型式及評分方式不同於傳統測驗。

由上可知，兩種評量模式各有其獨特的基本假定與目的，單一評量模式無法滿足所有的評量目的；應用測量模式的評量結果來代表學生是否達到某些標準，或應用標準模式的評量結果來建立精確客觀的評比訊息，基本上皆是一種失當的使用。教育工作者可依據其教育目的選用適當的評量模式，如果我們視教育的目的在於分類、分等

級，在於區分學生能力的高下，則傳統評量當能提供這方面較好的服務；如果我們視教育的目的是在幫助所有的學生發揮潛力，達到高水準的學習標準，那麼變通性評量應該較能符合這方面的期望。以上那一種教育目的的解讀最能凸顯教育的理想，相信大家心理都明白。然而，實際上這不是簡單的選擇題，兩種評量方式的取捨輕重之間一直存在著許多衝突與爭辯。Farr (1992) 指出，兩者緊張的關係根源於不同的群體有不同的評量需求。政策擬定者、擇才決定者、及一般大眾所需的評量訊息是學生或學校之間的學業成就高低比較，而教師所需要的評量訊息是學生有否達到教學目標、那裏沒有達到標準、什麼學習問題需要解決。這是一個很現實的問題，也透露出社會選擇性功能與個人自我完成之間潛在的矛盾衝突。如何平衡兩者不同的評量需求，如何使學生不僅具備自我反思、自我評量的眞正學習能力，也知道外在社會的評比標準或期望，這些仍是有待深省的問題，不是簡單的「排列組合」就可以解決的。

陸、結語

傳統評量及變通性評量各有所長。傳統評量能經濟客觀地提供大規模的評比訊息，但常使教學及學習訴諸於成果導向，所以其公平性雖有餘，但人本精神不足；雖能迅速提供學習結果，卻忽視了能反映眞實能力的學習過程。過度強調評比色彩一向是傳統式評量最爲人詬病的地方，它讓絕大多數的中下程度學生自覺是失敗者，此一標記無論對於學生的心理上或知識的追求上皆有相當負面的影響。較之傳統的考試測驗，變通性評量重視過程，對學生的學習情形能提供較完整、較深入的訊息，更重要的是它強調自我反思，不斷地追求進步、追求高標準，而這種自主探索的態度，對學生日後的學習最有助益。以此等層面去考量，變通性評量實有重視的必要。

所以今後教學評量的理想走向，應該是相當清楚，但是我們也不容鄉愿，視傳統評量的強大威力及變通性評量的實施困難如無物。雖然傳統評量備受批評，但其基本信念已根深蒂固於我們的社會之中，因此大家雖然不滿意傳統評量，但在觀念上也不太能接受新式的變通性評量。所以要從傳統走向變通的決定性關鍵，在於大眾是否能接受變通性評量的基本信念，也就是能否相信大部份的學生都能學習、都能達到標

準，能否接受相同的標準下可以有不同的表現方式，能否信賴評分者能做公允的判斷。這種觀念上的改變並非易事，需要時間，也需要不斷的討論與溝通。而縱然大家能接受變通性評量的理念，變通性評量本身實施上的困難又是另一挑戰。當變通性評量跳離簡單的事實記憶和推理，進入較複雜、較具挑戰性的評量內容時，其評量必然會超越數字，要求較多的文字書寫，而此也連帶地會涉及較多老師主觀的判斷，所以「如何公正評分」無疑地將會是一最受關注的焦點。除了評分問題之外，變通性評量的實施，也牽涉整個教學結構及重心的改變、教師的專業水準、以及成本等等問題，這些都有待我們步步爲營，從長計議。

總之，變通性評量目前仍在起步階段，除了須要各種行政資源給予不斷的支持外，更須要給予老師們機會與時間去從事專業發展。當然，其對課程、教學、與學習所產生的改變品質和功效究爲如何，仍須不斷地探索與改進。不過，我們必須要了解變通性評量的效應在時間上有其惰性，無法立即呈現預期的結果，實施者不要因爲其效果緩慢就全盤否定或轉向傳統測量模式的評量。另外，變通性評量倡導者在尋求建立系統化以適用於大規模評量的同時，應避免評比與外在標準反客爲主，以免失卻了眞實評量的特質。要再強調的是，欲求變通性評量發揮其預期目標，提高教學與評量品質，單靠學者的高登一呼是不夠的，它需要在教學前線的教師們領會、實行、倡導，也更需要行政人員及決策者的參與和支持。

參考文獻

莊明貞(民84)。變通性評量。康橋教研學會雜誌，19期，頁4。

郭生玉(民76)。心理與教育研究法。台北：精華書局。

黃政傑(民79)。評鑑與控制。師友月刊，五月，頁7-9。

劉湘川、簡茂發、林原宏(民83)。試題關聯結構與試題反應理論之聯合分析研究－以乘除概念之「暗隱模式」爲探討基礎。測驗統計年刊，第二輯，頁53-143。

Allington, R. L., Stuetzel, H., Shake, M., & Lamarche, S. (1986). What is remedial reading? A descriptive study. Reading Research and In-

struction, 26, 15-30.

Arter, J. A. & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. Educational Measurement: Issues and Practice, 11, 36-44.

Atwell, N.(1987). In the middle: Writing, reading, and learning with adolescents. Portsmouth, NH: Heinemann.

Au, K. H.(1993). Portfolio assessment: Experience at the Kamehamaha Elementary Education Program. In E. Hiebert, P. Afflerbach, & S. Valencia (Eds.), Authentic reading assessment: Practices and possibilities. Newark, DE: International Reading Association.

Baron, J. B.(1991). Strategies for the development of effective performance exercises. Applied Measurement in Education, 4, 305-318.

Calfee, R. & Hiebert, E. (1991). Classroom assessment of reading. In R. Barr, M. L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), The handbook of reading research(Vol.2, pp. 281-309). New York: Longman.

Cambourne, B. & Turbill, J.(1990). Assessment in whole-language classrooms: Theory into practice. Elementary School Journal, 90, 337-349.

Case, S.(1994). Will mandating portfolios undermine their value? Educational Leadership, 52, 46-47.

Cole, N. S.(1988).A realist's appraisal of the prospects for unifying instruction and assessment. In C. V. Bunderson (Ed.), Assessment in the service of learning (pp. 103-117). Princeton, NJ: Educational Testing Service.

Comfort, K. (1994). Authentic Assessment: A systemic approach in California. Science and Children, 42-66.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. Harvard Educational Review, 64, 5-30.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. Elementary School Journal, 85, 315-336.

Delpit, L.D. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. Harvard Educational Review, 58, 280-298.

Farr, R. (1992). Putting it all together: Solving the reading assessment puzzle. The Reading Teacher, 46, 26-37.

Farr, R. & Tone, B. (1994). Portfolio and performance assessment. Orlando, FL: Harcourt Brace & Company.

Freedman, S. (1995). Exam-based reform stifles student writing in the U.K. Educational Leadership, 52, 26-29.

Gandal, M.(1995). Not all standards are created equal. Educational Leadership, 52, 16-21.

Garcia, G. & Pearson, D. (1994). Assessment and diversity. Review of Research in Education, 20, 337-391.

Glaser, R. & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. Review of Research in Education, 20, 393-419.

Gomez, M., Graue, E., & Bloch, M. (1991). Reassessing portfolio assessment: Rhetoric and reality. Language Arts, 68, 620-628.

Goodman, K. S., Goodman, Y. M., & Hood, W. (1989). The whole language evaluation book. Portsmouth, NH: Heinemann.

Greenwood, J. (1993). On the nature of teaching and assessing. Arithmetic Teacher, 11, 144-152.

Haladyna, T.M., Nolan, S.B., & Haas, N.S. (1991). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20, 2-7.

Herman, J. & Winters, L. (1994). Portfolio research: A slim collection. Educational Leadership, 52, 48-55.

Jones, G. (1994). Assessment potpourri. Science and Children, 14-17.

Lambdin, D. & Walker, V. (1994). Planning for classroom portfolio assessment. Arithmetic Teacher, 318-324.

Madaus, G. (1993). A national testing system: Manna from above? An historical/technological perspective. Educational Assessment, 1, 9-26.

Mercer, J. (1989). Alternative paradigms for assessment in a pluralistic society. In J.A. Banks & A.M. Banks(Eds.), Multicultural education: Issues and perspectives(pp. 289-304). Boston: Allyn & Bacon.

Resnick, L. & Nolan, K. (1995). Where in the world are world-class standards? Educational Leadership, 52, 6-10.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction (pp.37-75). Boston: Kluwer Academic Publishers.

Rosenblatt, L. (1985). Viewpoints: Transaction versus interaction— A terminological rescue operation. Research in the Teaching of English, 19, 96-107.

Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21, 22-27.

Shepard, L. (1989). Why we need better tests. Educational Leadership, 46, 4-9.

Shepard, L. (1992). What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B.R. Gifford & M. C. O'Connor(Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 301-328). Boston: Kluwer.

Shuell, T. (1986). Cognitive conceptions of learning. Review of Educational Research, 56, 411-436.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. Educational Leadership, 50, 11-15.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. American Educational Research Journal, 31, 231-262.

Valencia, S. W., Hiebert, E., & Kapinus, B. (1992). National Assessment of Educational Progress: What do we know and what lies ahead? Reading Teacher, 45, 730-734.

Valencia, S., Pearson, D., Peters, C., & Wixson, K. (1989). Theory and practice in statewide reading assessment: Closing the gap. Educational Leadership, 21, 57-63.

Wiggins, G. (1989). Teaching to the authentic test. Educational Leadership, 46, 41-47.

From Tradition to Alternative : The Reflection of Classroom Assessment

Hsiu-Wen Huang

National Chiayi Teachers College

Abstract

This article reviews literature about the reform of classroom assessment in America. The paper starts with the exploration of the characteristics and problems of traditional assessment, and then introduces how traditional assessment is improved and what alternative assessment methods come with the tide of reform fashion. Next, the charactristics and criticisms of altenative assessment are investigated. Finally, the use and potential impact of both traditional and alternative assessment are examined by clarfying the basic assumptions of them.

The article concludes that we should select appropriate assessment methods according to the purposes of assessment—to make comparison or to understand the extent of standards students reach. Using single assessment to fulfill both purposes will result in many drawbacks in all probability. In terms of learning theories and educational purposes, the article also suggests that alternative assessment should be emphasized.