



初任理化教師之教學成效評量

林煥祥

國立高雄師範大學 化學系

(投稿日期：84年9月25日，接受日期：85年5月20日)

摘要：本研究之目的，係探討利用短時間群集評量方式 (Time series design)，評估初任理化教師教學成效之可行性及有效性。四位初任理化教師及其任教之四班國中二年級學生參與本研究。評量方式乃採用短時間群集設計，追蹤這些學生在理化第八章化學反應的學習成就。研究結果顯示，在職前訓練中專業學科成績及教學表現優異，且具現代科學哲學觀的初任理化教師教學之下，學生的學科成就進步型態，如預期般顯著的優於那些教學表現較差，且具傳統科學哲學觀的初任理化教師所教之學生，顯示出該評量方式在評估初任理化教師之教學成效的可行性。另外有關該評量方式的效度之考驗，由其對高成就與低成就學生的鑑別能力，以及研究工具本身的效標關連效度，進一步顯示其有效性。未來若經進一步與其它評量方式比較結果，確具實用性，則在師資檢定授證系統，似乎可將該評量方式當成另外一種評估初任理化教師教學成效的方法，以便提供較為客觀的量化數據給相關人員參考。

關鍵詞：教學成效、短時間群集評量、學習成就。

壹、緒論

一、理論依據及文獻探討

1980年代美國的科學教育之主要目標為推動全民科學素養，讓受教者能了解科學、科技、及社會之間的互動與密切關係，並能利用學校裏所學的科學知識於日常生活中的社會事件，做出明智的決定 (National Science Teachers Association-

tion, 1982)。因此科學教育的重點擺在課程的發展及教學的改進。然而任何課程的改革若沒有伴隨著評量的同步改革，則其改革成功的機率不大。因此，緊接而來的 1990 年代，將會因課程及教學之改革而在評量方面有所變遷。誠如 Tamir (1993) 所言，1990 年代科學教育的焦點將集中在評量 (Assessment) 的發展與改進。不幸的是，有些評量方式甚具創意卻鮮為一般教師所熟知或使用。例如短時間群集評量設計 (Time series design) 即是一例。

大部份的評量都只是在長時間的學習期間內進行一次或二次的資料收集，雖然可以呈現該時段內之學習效果，卻無法顯示出學習期間內進步的型態。有鑑於此，評量專家們早在 60 年代就發展出具有創意的短時間群集評量，(Campbell & Stanley, 1966)。為了呈現一群學生學習進步的型態，首先在未教學前即多次對該群學生進行前測。然後進行教學，在教學過程中並多次測驗。教學完成後再進行多次後測。由於測驗次數頻繁，因此每次測驗時間均甚短。從這些密集的前測、施教中測驗、及後測結果中，研究者可以清楚的看到學生的進步情形是否隨著教學日數有所差異。比起單一前測、後測之數據，更能詳細而有效的呈現教學效果。其評量過程詳如圖 1。

在圖 1 中，橫座標為評量次數，縱座標為全班學生各該次之平均得分數。由 A、B、C 之三曲線比較中可以發現，C 曲線之進步型態起落不定、不明顯。

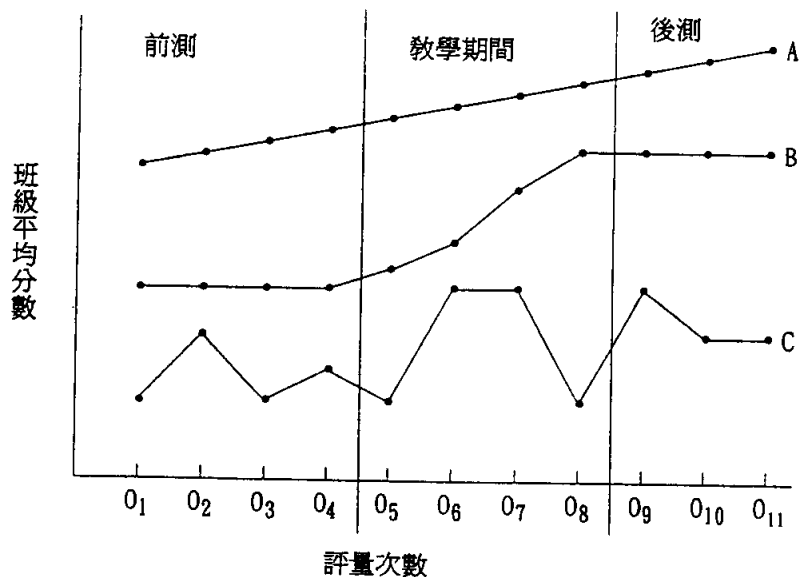


圖 1：短時間群集評量之曲線圖。

A 曲線之進步型態雖然明顯，但在未教學前，該群學生即已有進步之傾向，顯然此進步情形非完全來自教學之處理。而 B 曲線之進步很顯然的是來自教學之效果。因為前測的 4 次平均及後測的 3 次平均，都沒有進步情形發生，唯有在教學期間才呈現出進步的型態，因此研究者可以很肯定的指出，此進步來自教學之效果。同樣的由圖 1 中來看，如果不是採用短時間群集評量，而只採用前測及後測，則 A，B，C 三曲線之平均分數差異幾乎相同，表面上看來，似乎這三群學生的學習進步情形相同，然而經由短時間群集評量的曲線呈現，才能更進一步了解到只有 B 曲線那一群的學生，進步是真的由教學效果而產生的。由這個例子的說明也更能看出短時間群集評量在顯示學生學習狀況之長處。

大多數人很自然的會對這樣的評量方式產生一些疑問。諸如：既然每一次測驗的時間都很短（5 分鐘以內），那麼所收集到的數據是否可靠或有效？學生接受如此多次的測驗，是否會因厭倦而隨便作答，產生不可靠的測驗結果？評量工具被重覆或多次使用，其數據是否仍然有效？數據該如何分析？結果是否能確實反應出學生之學習狀況？由於這些疑問甚受關心，有些學者即對此展開研究並陸續發表於學術期刊中，以下即針對這些研究結果加以探討以便釐清疑慮。

Mayer 及 Lewis (1979) 爲了探討短時間群集評量是否適用在科學教育的領域，選定了一班高中暑期生物課的學生共 21 人，以對生物科學習態度爲依變項，展開爲期八週的評量。學生利用每節下課前的 3 分鐘作答該態度問卷，內容包括今天上課是否有趣、刺激、值得、氣氛和諧等。該研究之自變項則爲各種不同的教學方式，如：討論、實驗、幻燈片、考試、作業習作、學生演示、以及戶外活動等。每日問卷回收後，該節課之全班平均態度分數即行求出。最後分析發現，每次戶外活動教學的平均分數（共有 6 次），都是一星期內各種教學方式中最高分者，而全班態度平均分數最低者，則爲每次 2 小時的小考（共有 4 次）。當以全班態度之平均分數爲縱座標，以日數爲橫座標，所得之圖可以看出其型態之一致性。由於其他研究中也證明戶外教學活動之效果，因此可證明該研究工具及評量方式，的確能描述出教學活動受學生喜愛的程度。再者檢查生物科成績高成就群（A 等第）及低成就群（C 及 D 等第）的態度平均分數，可以顯示出其差異性，也能證明該研究之效度。進一步的相關係數分析（教學方法與態度分數），更證明了以上座標圖的型態分析之有效性。由這些收集數據的分析結果，證明了短時間群集評量用在科學態度評量的可行性及有效性。也證明了雖然學生

不斷重覆接受測驗，不致因而產生厭倦致使測驗結果失效。該研究者在每日收集數據的同時，發現同學們的合作態度及施測時之良好氣氛，也能支持“測驗疲倦症候群”現象並未發生。不過維持良好的測驗氣氛，使學生願意合作，也是研究者收集數據時應該注意的技巧之一。

短時間群集評量除了適合於診斷學習態度外，也曾被成功的用來評量概念的學習成就。Mayer 及 Kozlow (1980) 以 2 班 8 年級的學生為研究對象，班級人數各為 24 人及 25 人。選定地球科學中地殼變化這個學習單元，自行發展了一套包含 54 題選擇題的研究工具，進行為期 26 天的短時間群集評量。第 1-8 天為前測對照期，第 9-18 天為教學期，第 19-26 天為後測對照期。兩個班級都是以同一研究工具施測，但第一班每日每人只考 1 題選擇，而且每位學生每一天都做答不同的題目，而另一班則每日每人考 3 題。資料分析結果顯示，每天每位學生抽考 1 題的方式較為有效。作者在結論中指出短時間群集評量的方式，若用來評量學習成就，似乎單題施測的效果比 3 題施測效果好些，而且每日施測所需之時間亦可較短。另外作者也指出學生每次考完後，有可能利用下課時間互相討論，如此則會影響該評量方法之效度，值得注意預防。

繼 Mayer 及 Kozlow (1980) 成功的利用短時間群集評量來解釋學生的學習成就之後，Fransworth 及 Mayer (1984) 進一步以此方法來評量不同認知能力的學生之學習成就。由於評量結果中，形式推理期 (Formal reasoning stage) 的學生之學習成就，與具體操作期 (Concrete operational stage) 的學生之學習成就，具有顯著差異性，他們更進一步肯定了短時間群集評量用在評量學習成就的適切性及有效性。

以上的幾個例子說明了短時間群集評量成功的被用在科學態度及科學成就的評量上。然而該評量方式採用次數頻繁的測驗，是否會對所收集到的數據有任何不良的影響或作用？這個問題倒是引起不少關注。針對重覆測驗所可能引發導致受試者厭惡而隨便作答，或由重覆考試中熟悉考題，使得接下來的測驗都得高分，讓研究者誤以為是教學效果所造成的學習進展，有許多研究者曾做過研究。例如 Mann, Taylor, Proger, Dungan 及 Tidey 等人 (1970) 曾對中學生在一星期內重覆施測 4 次，每次都是相同的考題，內容包含了數學及語文，且每次考完後並未檢討試題內容。結果發現進步最大的是在第二次的測驗，可見在第三次及第四次的測驗時，學生可能有了厭煩的效應發生。另外 Catanzano 及 Wilson (1977)

將一群 7 年級修習普通科學的學生，分成 3 組並分別給予不同的測驗方式。第一組不施行重覆考試，第二組間歇性的重覆考試，第三組則施行持續性的重覆考試，每次考完試後都同時複習測驗內容。資料分析結果發現，間歇性重覆考試及持續性重覆考試兩組的考生，其學習成就顯著的優於未重覆考試的考生。此結果顯示，重覆考試的確可能會影響往後的學習成就。另外的結果也顯示了這三組的學生在科學態度的表現上，以間歇性重覆測驗組的學生最為優異，顯示重覆測驗可能會影響到學生的學習態度。

雖然重覆施測可能對學生的學習成就造成影響，而且使學生因感厭煩而隨便作答 (Resentful demoralization)，而短時間群集評量也是採用密集式的重覆施測，但是它在設計上考慮到這些因素而採用每次只施測 1 題，且每位考生每次皆作答不同的試題。所以也就是因為施測時間短及試題的變化性，且每次考後不複習的設計方式，應該可以避免前述研究中所發現重覆施測的不利效應 (Testing effect)。Mayer 及 Rojas (1982) 爲了澄清這些疑慮，乃選擇俄亥俄州郊區的一所中學內 4 個班級修習地球科學的學生，隨機分配成三組不同的測驗方式。A 組的學生每日接受施測（內容爲 1 題成就測驗的選擇題，1 題態度評量），B 組的學生每 4 日受測一次，C 組的學生每隔 10 日受測 1 次。A、B、C 組學生若按日程安排同時施測，則該日各組皆採用相同的題目。經過 30 天的評量結果，發現 A 組的學生並未因每日的重覆施測而從考試經驗中學習到考試內容，以致學習成就優於 B、C 組的學生。另一方面，在學生對該科之學習態度的評量上，結果發現 A 組的學生在學習態度上與 B 或 C 組的學生並沒有顯著的差異，證明 "resentful demoralization" 並未發生在該研究之學生。因此 Mayer 及 Rojas 的研究更進一步的支持了短時間群集評量設計的有效性及實用性。也值得在科學教育的其他學科之教學更進一步推廣。

教師的專業學科知識及對科學本質 (The nature of science) 的觀點，是影響其教學表現 (Teaching practice) 的兩大重要因素。有關教師的專業學科知識與教學表現的研究雖然不多，但研究結果一致指出該二者是有密切的關連 (Carlsen, 1987; Ferguson & Womack, 1993; Hashweh, 1987)。Druva 及 Anderson (1983) 也曾利用後設分析 (Meta-analysis) 的方式證明其高度相關性。Lantz 及 Kass (1987) 的研究則更進一步指出化學教師的化學學科知識如何塑造出化學教師的教學表現。另外有關教師對科學本質的觀點與教學表現的研究結果則較不一

致。有的研究發現該二者是有深切而直接的關連 (Brickhouse, 1990; Gallagher, 1991) 而有的研究則發現較低的相關性 (Duschl & Wright, 1989; Lederman & Zeidler, 1987)。但是 Lederman (1992) 進一步指出，這些低相關性的結果可能是受教育政策或課程限制的影響，以致在教室教學中沒有表現出來。近幾年來科學教育學者也一致強調教師及學生對科學本質瞭解的重要性 (American Association for the Advancement of Science, 1989)。然而對科學本質的觀點不同，如何影響教師的教學方式呢？Brickhouse (1990) 爲了回答此一問題，分別觀察了對科學的本質具有不同觀點的科學教師之教學行爲，他在該研究中指出，當教師具有現代科學哲學觀時，上課中比較會鼓勵學生，運用創意想出不同的解題方法，或設計不同的實驗過程得到實驗結果。反之，當教師的科學哲學觀屬於實證觀者（例如把科學學說看成是自然界已經存在的真理），在其教學過程當中，如果學生的實驗結果與教科書答案不合，該教師則要求學生重新檢查實驗步驟以便得到“正確答案”。由這些研究的結論中，不難發現教師的專業學科知識及對科學本質的觀點，對其教學表現的確具有決定性的影響力。在瞭解了教師的專業學科知識及對科學本質的觀點可能影響其教學表現後，一般社會大眾更感興趣與關心的是，那麼這些具有不同背景的教師，其教學成效又會如何不同呢？是不是有具體的方法可以評量出來？尤其負責師資檢定的相關人員，對於具有創意且可靠又有效的評量方式，能夠評量出教師的教學成效，可能更感興趣。有鑑於此，利用短時間群集評量去評估同樣不具教學經驗，而卻各自具有不同的專業學科知識，及不同科學哲學觀的初任理化老師，的確值得嘗試。然而教學成效的評量並不容易，評量的方式各有不同。大部份的方式，偏重於上課的觀察，再依據各種量表上的設定標準給分，例如，Ferguson 及 Womack (1993) 的研究就是一例。Cook (1985) 則利用學生在標準測驗的成績當成是教師教學成效的指標。後繼者更有 Mwamwenda 及 Mwamwenda (1989) 及 Purser (1987) 也利用學生的考試成績當成是教師的教學成效。綜觀這些評量方式，觀察法的費時及評審者的主觀因素是其弊病。而利用一次或二次學生的評量成績爲教學成效，雖然量化的數據可以較客觀，但卻較不易代表教學過程中學生的改變情形。短時間群集評量則不但兼具了省時及客觀的評量優點，而且更能呈現出學生成績進步的型態。因此它是否適用於教學成效的評量值得加以探討。

二、研究目的

具體而言，本研究之主要目的為：

- (一)發展短時間群集評量的工具並分析其評量結果的效度。
- (二)探討由此評量方式所得學生學習曲線，是否真能反映出不同背景、能力的初任理化教師之教學成效？

貳、研究方法及過程

一、研究對象

四位初任理化教師及該四位理化教師所任教的國二班級共四班學生參與本研究。該四位初任理化教師的選取，係由一所師範大學化學系應屆結業生中，針對下列兩項評量結果，選出適當的研究樣本四位：

- (一)化學學科、科學教育、教學實習、教材教法之平均成績。
- (二)對科學本質的看法 (Views on the nature of science)。

上述(一)(二)兩項評量中，由(一)項結果將結業生分成高成就組及低成就組，亦即百分比排序 (Percentile rank) 前三分之一者屬高成就組，後三分之一者屬低成就組。而(二)項的評量係根據中文版 Aikenhead (1989) 所發展的 Epistemology of Science Questionnaire 將全班學生作答結果，分成對科學本質較具傳統觀和較具現代觀者兩組。最後由高成就且對科學本質具現代觀之結業生中，隨機選取兩位（以 A 組簡稱之），而低成就且具傳統觀之結業生中，也隨機選取兩位（以 B 組簡稱之）。

以上 4 位初任理化教師選定後，乃就該 4 位教師所任教之國中二年級班級中，隨機各選取一班為研究對象，因此共計 4 位理化教師 4 班國中二年級學生參與本研究。A 組兩位老師任教的學校中，一校地處市區，另一校則為郊區。而 B 組兩位老師任教的學校，一校為郊區學校，另一校為較偏遠之國中。

二、研究工具 — 學生學習成就測驗卷

該成就測驗卷係研究者根據國中理化第八章化學反應之課程內容自行發展成一份包含 45 題選擇題之評量工具。採用選擇題的原因係由於文獻探討中，

Mayer 及 Kozlow (1980)，以及 Fransworth 及 Mayer (1984) 的研究結果發現短時間群集評量若採用單題選擇題，其適用性及有效性都足以發揮最大的效果。而且因為施測時間短，加上每位學生每次受測都得到不同的題目，因此學生不會有厭倦，也不會從重覆考試中，得知下次將考的題目內容而產生所謂的考試效應，以致使收集到的學生得分失去準確性。有鑑於此，本研究採用先前研究者的建議，以單題選擇施測，故依課程內容發展此一研究工具，其發展過程如下：

首先分析理化第八章的內容，找出一些重要的主題，如原子量、分子量、莫耳、莫耳濃度、化學反應方程式等。為了確保命題內容分佈與教學所花時間一致，研究者分別徵詢了三位國中教師，有關他們在各主題之授課時間及對各主題之教學目標。在考量該三位教師的意見之後，乃建立內容雙向分析表 (Two-way specification table) 如表 1。

利用表 1 為命題之藍圖，並參考 Gay (1985) 及 Gronlund (1985) 對選擇題命題之建議，進行命題工作。包含 48 題的初稿命題完成後，商請一位大學化學教授，一位科學教育專家，一位中學教師審核其適切性，再將他們的建議列入修改之參考。修改後以此試題對二班常態分班的國三學生（已學過該單元）共 74 名進行試測，再以 VAX 主機上的 SAS 套裝軟體做試題分析。剔除鑑別度低於 0.25 的試題後，該測驗之難度值均勻分佈於 0.15 ~ 0.90 之間。各選項之誘導度

表 1：「化學反應」教學目標與內容分析表

內容主題	教 學 目 標			合計
	知識性	理解性	應用性	
莫耳概念與亞佛加厥數	4*	10	2	16
原子量、分子量	2	2	2	6
莫耳濃度	1	4	3	8
化學反應方程式	5	7	3	15
	12	23	10	45

*：表內數字表示該主題在試卷內包含之題數。

也均達 2% 以上。最後，定稿的 45 題，其信度為 Cronbach α 0.94。另外，各學生該次測驗得分，與其上學期理化成績之相關係數為 0.86，這個結果也進一步支持了本工具之效度。

三、研究過程

利用上述之研究工具，針對研究對象進行評量，評量過程分成三個階段：第一階段—前測 (Pre-test)：即教師尚未進行各該單元之教學前一星期之評量 4 次。第二階段「教學中測驗 (During-treatment test)：自開始教學各該單元至教完本單元為止，共評量 (11 次)，評量次數同該單元之授課節數。第三階段—後測 (Post-test)：教完各該單元後之評量 4 次。以上三階段之評量，每次評量都使用同一工具 (45 題選擇分印成 45 張考卷，每張考卷只印一題)，利用每節課下課前之 5 分鐘，由研究助理協助任課教師，隨機分發每位學生一題作答，每個學生在各單元之評量中，不可重覆上次已作過之試題，因此每份試卷只有一題且以不同顏色分類之，使分發試題較為容易。由於隨機分配試題，因此程度好或差的學生都有可能抽到難度高的試題。當學生抽到的試題內容屬教過的範圍時，若其了解則應該答對。因此隨著授課日數增加，逐漸涵蓋試題內容，則全班平均分數也會逐漸進步。若教師之教學成效良好，該班之全班平均分數也會進步的較明顯。不過某一天的考試中，若程度好的學生都抽到低難度且已教過之試題，而程度差的學生都抽到高難度且未教過之試題，則該次之全班平均分數可能會很高。這也是造成學習曲線起伏的因素之一。然而在隨機且多次施測的情況下，整個學習曲線之進步形態還是可以分析出來。

四、資料分析

學習成就評量所得資料，按日計算全班的平均分數。全班各該次之平均分數計算公式為：平均分數 = 答對人數 / 全班人數。例如教學前的測驗中，由於考試內容老師尚未教到，因此全班答對人數一定很少。若全班 47 人中只有 3 人答對，則該日全班平均分數為： $3/47 = 0.06$ 。如此依序求出每次之全班平均分數，再以日數為橫座標，全班平均分數為縱座標作圖，以便看出該班之學習進步型態。利用目視方法可以找出這些座標點的最佳直線及該直線之斜率。斜率愈大，表示學生進步情形愈顯著，在前測、教學中測驗、及後測的三階段評量中，

我們期待施教中測驗結果的斜率會最陡，而且教學愈成功的班級，其斜率愈陡。

除了以上座標圖表示法之外，將日數與全班平均分數進行相關係數的分析，也可進一步檢視學生之學習狀況。照理預測，每日之全班平均得分應該隨著教學日數增加，全班同學逐漸了解教學內容之科學概念，而使得分提高，亦即教學階段的第三天之全班平均分數，理應高於第二天或第一天之平均數。故相關係數在教學期間之期待值應為正數，且愈高愈好。

至於有關本研究工具的效度，可由下列方法分析求得。首先將每一位學生在教學期間 11 次測驗的答對次數加起來做為其得分，進而將全部學生依其在校之理化成績分為高低兩個成就群，再利用獨立性 t 考驗，比較該兩群學生在 11 次測驗的得分數。若高分群顯著的優於低分群，則根據 Gronlund (1985) 所指，該結果可當成是本研究工具的效度指標。也就是說該工具能有效的將學生的成就鑑別出來。

參、研究結果

參與本研究的四位初任理化教師中，其中一位 B 組的教師任教於苗栗偏遠山區，交通不便，研究助理不克前往協助施測。且該教師向研究者表示其班級經營困難，很多學生沒有專心聽課而在評量時隨意作答，因此其評量次數曾中斷幾次。由於其評量結果不完整，且學生並未盡心作答，根據 Mayer 及 Lewis (1979) 的建議，該班資料在分析時須予刪除。因此計有 A 組兩位教師的任教班級（人數分別為 47 人及 46 人）及 B 組一位教師的任教班級（人數為 47 人）所得資料可供分析。

為了探討本研究評量結果的效度，乃自 A 及 B 組中各選出一個學校，並向學校當局收集這些國中學生的月考成績。這些月考題目都是校際間聯合命題，大部份取材自事先經過試測修正的題庫，因此其鑑別度及效度都達一定水準，而且學生在三次月考的成績表現上也頗具一致性。根據 Gay (1985) 的建議，這些月考成績可以用來當成是效標關連效度 (Criterion-related Validity) 指標之參考依據。如果學生在本研究之評量中，得分與其月考成績高度相關，則代表本研究之評量結果具有良好的效度。因此首先將每位學生在本研究的三個評量階段中答對次數求出，再與其第一次月考的成績（因為第一次月考的內容與本研究的評量內容相同。）進行相關係數之分析。A 組與 B 組學生在本研究之評量得分與月考

得分之相關係數如表 2。

由表 2 中可以看出不管是 A 組或 B 組學生，其月考成績皆與本研究教學階段，利用短時間群集評量所得之成績具有顯著之相關 ($P < 0.001$)。另外，正如我們預期之結果，前測評量階段的分數，由於尚未教學，未具任何學習效果，所以與月考分數沒有顯著相關。而在後測階段，僅 A 組的學生得分具有顯著相關 ($P < 0.001$)。

另外，Gronlund (1985) 曾經建議，如果一個評量能有效的鑑別出高成就及低成就群，則具有良好效度。因此在資料分析之時，遂將各班學生依其第一月考分數分類，凡其月考分數高於該班中數者為高成就群，低於全班中數者為低成就群。進而比較該二群學生在本研究中，短時間群集評量之得分。由表 3 中可以看出 A 組的班級中，高成就群學生在教學階段的 11 次評量中，平均答對分數高達

表 2：短時間群集評量成績與月考成績之相關性

組別	前 測		教 學 期 間		後 測	
	r	p	r	p	r	p
A	-0.01	0.9606	0.85	0.0001***	0.58	0.0001***
B	-0.05	0.7613	0.50	0.0005***	0.14	0.3476

***: $p < 0.001$

表 3：高成就群與低成就群在教學階段得分比較

		N	Mean	SD	t
A 組	高成就群	21	5.28	2.49	5.83***
	低成就群	22	1.86	1.04	
B 組	高成就群	22	5.00	2.22	3.08**
	低成就群	23	3.30	1.33	

** : $P < 0.01$ *** : $P < 0.001$

5.28 分，顯著的高於低成就群的平均答對分數 1.86 分 ($p < 0.001$)。而在 B 組的班級中，高成就群的分數 5.00 分也顯著的高於低成就群的 3.30 分 ($p < 0.01$)。

除了以上的方法可以檢驗本研究評量之效度外，如果這些學生的平均得分，在前測未教學階段，及教學階段的得分上具有顯著差異，則亦表示該評量方式能夠有效的反映出學生的學習成就。因此，首先將每位學生在三階段的答對題數統計出來，再以此題數除以該階段之評量次數為其得分。例如某生在教學階段的 11 次評量中答對 4 次，則其得分為 4/11 分。經分析結果，A 組與 B 組學生之平均分數、標準差、及相依性 t 考驗（檢查前測與教學階段之得分差異）結果如表 4。在表中的平均分數比較中可以發現，兩組學生在教學階段都顯著的高於前測未教學階段 ($P < 0.001$)。這一項分析結果，更進一步支持了短時間群集評量的有效性。

短時間群集評量的另一項優點就是它多次的評量結果作圖後，可當成學生學習進步情況之指標。為了進一步檢驗這些資料是否真能反映出學生的學習成效，首先乃以日數為橫座標，全班學生平均分數為縱座標作圖，得圖 2，3，及 4。由圖 2 及圖 3 中可以看出 A 組的兩位初任教師（在職前訓練上成績優良且在科學本質上具有現代觀者），其所教出的學生在學科學習成就上，進步的型態似乎較為明顯，也就是說，圖中教學期間的曲線圖之斜率較大。尤其以圖 2 更為明顯。由目視可以看出圖 2 中，在前測階段（第 1 天至第 4 天）看不出有任何分數上的進步情形。而在教學期間（第 5 天至第 15 天），全班平均分數似乎隨著教學日數增加而進步。在後測階段（第 16 天至第 19 天），則又看不出這個進步的

表 4：短時間群集評量的平均分數、標準差、及 t 值

組別	前 測		教 學 期 間		difference	t
	mean	sd	mean	sd		
A	0.18	0.14	0.32	0.23	0.14	3.71***
B	0.16	0.15	0.37	0.18	0.21	5.55***

***: $p < 0.001$

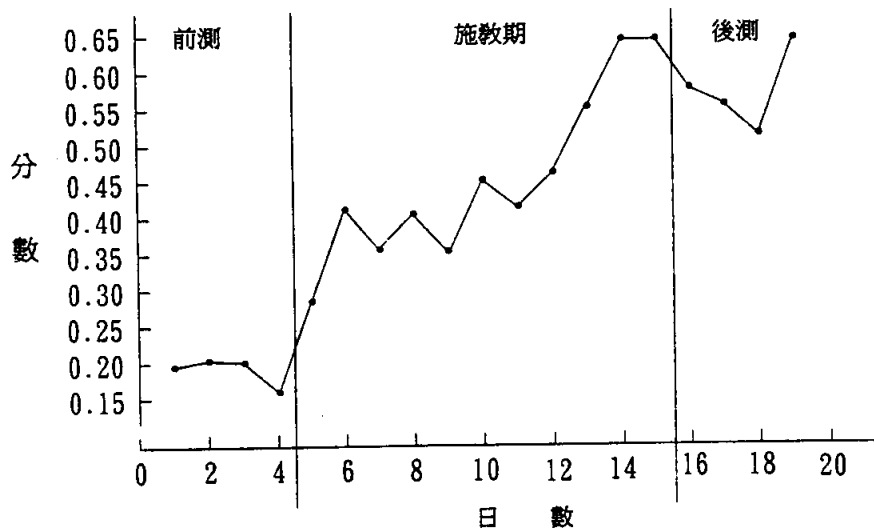


圖 2：A 組初任教師任教班級（市區）之學生成績進步曲線(一)。

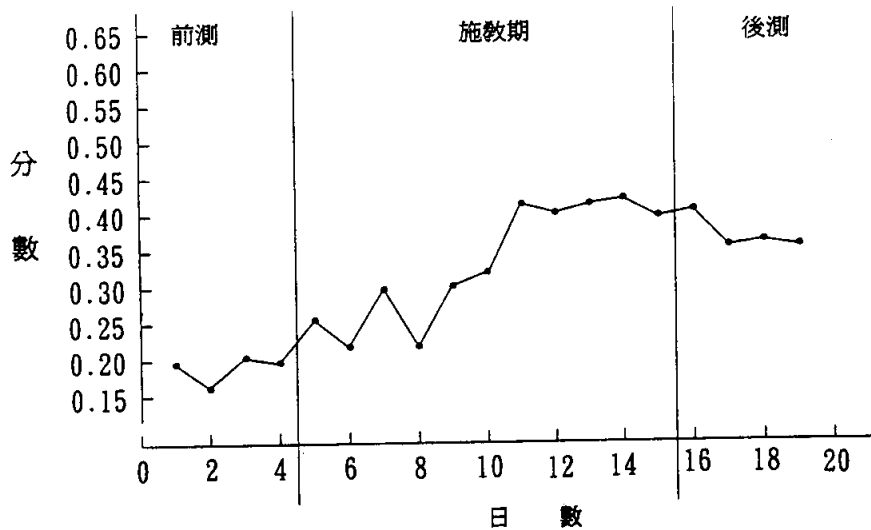


圖 3：A 組初任教師任教班級（郊區）之學生成績進步曲線(二)。

趨勢。同為 A 組的另一位老師所教的班級，學生之學習成就進步情形由圖 3 中可以看出，不若圖 2 那麼明顯，但若是和 B 組的老師所任教的班級之學生學習成就進步曲線圖（圖 4）比較之下，仍然可以比較出差異之處。在圖 4 中，不管是那一階段（前測、教學期間、後測），似乎都看不出全班的平均分數有顯著上揚的趨勢。

爲了要更進一步呈現出以上三個學習成就進步曲線圖的不同，遂以日數及每

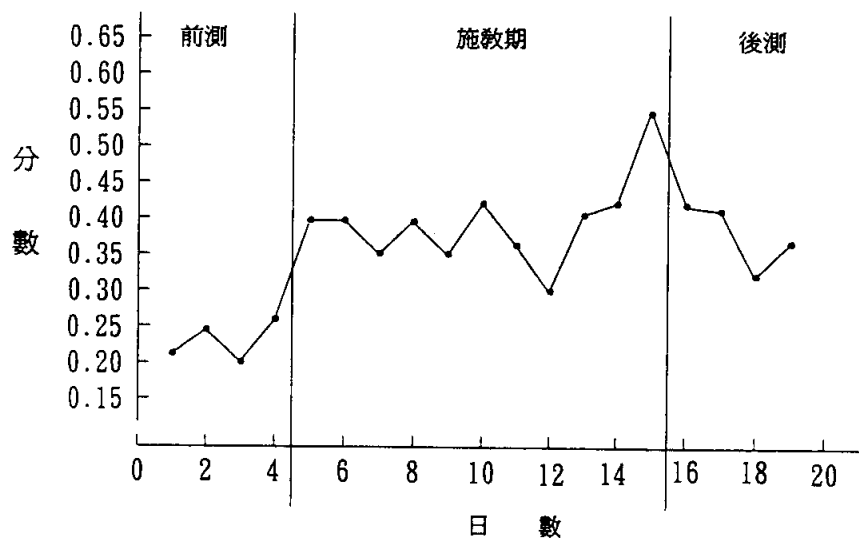


圖 4：B 組初任教師任教班級（郊區）之學生成績進步曲線。

表 5：各學習階段日數與學習成就相關係數

組別	前 測		教 學 期 間		後 測	
	r	p	r	p	r	p
A1	-0.60	0.4028	0.92	0.0001***	0.47	0.5275
A2	0.22	0.7835	0.88	0.0007***	-0.89	0.0400*
B2	0.54	0.4626	0.47	0.1415	-0.74	0.2558

*: $p < 0.05$ ***: $p < 0.001$

日之全班平均分數二個因子，進行相關係數的分析。我們可以預期的是，若是全班平均分數隨著施教日數而增加，則其相關係數值愈大，也表示進步情形愈明顯。當全班平均分數值與日數相關係數值分成前測、教學期間、後測三個階段分析時，其結果如表 5。表中的比較可以發現，A 組的二個班級在教學期間的該相關係數值，分別高達 0.92 ($P < 0.001$) 及 0.88 ($P < 0.001$)。換句話說，這兩個班級學生的學習成就，很顯然的隨著老師的教學而逐漸提高。反之，B 組的班

級，雖然在教學期間 11 天的班級平均分數比前測時來得進步，但其進步型態起伏變化甚大，我們較無法確定這些進步，是來自老師的教學而對課程內容及科學概念了解有所增進。

表 5 中，另外值得注意的一點是 A2 這個班級的學生，在後測的 4 次評量中，全班平均分數有逐漸退步的趨勢 ($r = -0.89, P < 0.05$)。這個情形在其他的兩個班中，倒是沒有出現。這個情形顯示，學生對較早學過的內容有遺忘的趨勢。Mayer 及 Kozlow (1980) 的研究中也有此相同情形發生。

肆、討論

Mayer 及 Lewis (1979) 認為，短時間群集評量的特色在於不做「不同班級間實驗組與控制組的比較」，而以同一班級為實驗組兼控制組，只檢驗一個班級在教學前、教學期間、教學後三階段的學習進步情形。因此該評量方式的基本假設是，一個具有教學成效的教師，不管教到市區程度好的或偏遠地區程度差的班級，他（她）都能使該班級在教學期間的『每次評量之全班平均分數』逐漸進步，且優於未教學期間的『每次評量之全班平均分數』。如果老師的教學成效愈好，則該班的進步也愈顯著。由於不進行不同班級間成績的比較，只進行各班間進步型態 (Pattern) 的比較，因此在「學生的相關知識背景」的考量上，不若傳統的實驗研究法中，必須特別強調。不過，為了進一步探討該評量方式的效度及客觀性，也就是兼顧到學生的相關背景，林煥祥 (1995) 也進一步利用相同的設計，在常態編班的同一學校內選出一位初任及一位資深優秀教師（由該校校長、教務主任、及同校之理化教師所推薦）加以探討。結果顯示資深優秀教師任教班級之學習曲線，正如預期般，呈現較為顯著的進步型態。其結果也進一步支持了該評量方式的適用性。

在選取 A 組與 B 組教師時，雖然學科知識及科學哲學觀是兩個重要的條件，但另一個更重要的條件是教學表現（包括大四教學實習的模擬試教以及校外三個星期的實際教學表現），由大學教授及國中資深教師評定其教學表現，作為本研究分組之參考。所以在研究者已知 A 組老師的教學成效優於 B 組老師的前提下，如果有一種評量能忠實的反應出此結果，則該評量方法即具良好效度（詳見 Gronlund, 1985, p73-75）。Mayer 及 Lewis (1979)，Mayer 及 Kozlow (1980) 等研究者利用短時間群集方式評量學生的學習態度及不同認知能力的學生

之學習成就，也都是依據與本研究類似的理論邏輯（即研究者根據文獻預知形式操作期的學生學習成就會高於具體操作期的學生，而短時間群集評量能反應此一結果，故具良好效度。）

本研究試圖利用短時間群集評量方式評估初任理化教師的教學成效，所獲致的初步結果令人鼓舞。進而對這種評量方式，未來在師資培育中的應用深具信心。首先在評量結果的效度考驗上，Fransworth 及 Mayer (1984) 由其研究結果中，發現短時間群集評量的結果，確實能區別出形式推理期與具體操作期學生間的學習成就；本研究的結果證實，該評量方式也確實能夠鑑別出高成就群及低成就群學生間的差異。這項結果不但證實了本研究工具之效度，也支持了Fransworth 及 Mayer 的結論，短時間群集評量應用於學習成就評量的可行性。另外參與本研究的學生，在短時間群集評量的得分與月考的得分具有高度顯著的相關，顯示本評量方式能適當的反應出學生的理化學習成就，更進一步支持了該評量的效標關連效度。

根據 Cook (1985) 的結論，學生的學習成就分數可以當成是教師的教學成效，那麼本研究利用短時間群集評量方式所得之學生學習成就，不但與學生之月考成績顯著相關，而且有更多的證據支持其效度，因此應當可以當成是教師的教學成效指標。尤其評量時間都選擇在上課結束前 5 分鐘，更避免了許多影響學生成績的外在因素（如：自修時數、寫作業時數等），比月考分數更真實的反映了教學成效。再者由於各校月考題目難度不同，學生程度參差，若以月考分數直接當成教學成效有失公平。而短時間群集評量方式，係以同一班級為實驗兼控制組，不與其它班級進行成績比較。不管任何一位老師教到多差的班級，只要他/她的教學具有成效，則該班之教學曲線必能呈現出顯著的斜率。而在日數與班級平均分數的相關係數分析上，也必能呈現出顯著的正相關。由此可見利用短時間群集評量的方式，所收集到的成就分數來代表教師的教學成效，不但可以避免不客觀的實驗組與控制組間的比較，它利用多次的評量數據，更能夠方便且詳細的呈現出一個班級的學習狀態。

接下來我們所關心的就是這些評量所得的學習曲線圖，是否真能反映出學生的學習狀態？由 Carlsen (1987)，Ferguson 及 Womack (1993)，以及 Hashweh (1987) 等人的研究中發現，教師的專業學科知識、教育專業知識與其教學表現有著密切的關係；另外 Brickhouse (1990) 及 Gallagher (1991) 等人的研究也證

實，教師的科學哲學觀與其教學表現有關。那麼根據以上這些研究，如果選出專業學科知識及教學表現良好，且具備現代科學哲學觀的理化教師，其教學成效必然優於專業學科知識及教學表現較差，且具備傳統科學哲學觀的理化教師。所得的研究結果若與此一致，則代表其評量資料忠實的反映出其教學成效。由本研究的结果中可以發現，A組的理化老師之授課班級，不但學習進步曲線較為陡峭，日數與全班平均得分的相關係數值也較為顯著。這些結果也支持了利用短時間群集評量來評估教學成效，確實具有可行性。另外，在前面緒論中曾經提及，一般的前後測評量方式，由前後測得分差異推斷學生的學習成就，無法得知學生的進步過程及型態。而短時間群集評量方式的多次評量，則能補此不足。本研究中的B組班級平均分數，教學階段亦顯著的高於前測階段，若非多次評量，則無法分析出其進步型態不如A組教師任教班級那般顯著。因此這項優點確非其他評量方式所能及。

伍、在科學教育上的應用

由Mayer及Lewis(1979)，Mayer及Kozlow(1980)，Mayer及Rojas(1982)，以及Fransworth及Mayer(1984)等一系列對短時間群集評量方式的探討，已然確立了這種評量方式在科學教育上的應用價值。本研究更進一步探討了這種評量方式的效度、信度及其在教學成效評量的應用，所得結果更增強了這種評量方式的可行性。國內目前正值師資培育法細則制定階段，對於合格師資的鑑定方法及證照發給條件之訂定等方面，尚在發展階段，或許這種短時間群集評量方式，可以考慮為評量教師教學成效的方法之一。如此一來，在證照的發給或教學成效的評比，也比較有一些量化的數據可供參考。

陸、研究限制及未來研究建議

為了符合短時間群集評量的原則，本研究中的學習成就評量工具內容全部為選擇題，因此在學習成就的評量層面上較受限制。例如實驗技能則無法評量。學生在這方面的成就則受限於研究工具，無法加以探討。另外要提醒讀者的是，科學教育的評量應是多元性的，很多評量專家所建議的方式諸如：學習歷程檔案(Portfolio)(Collins, 1990)，實作評量(Performance assessment)(Baron, 1990)，群體評量(Group assessment)(Johnson & Johnson, 1990)，以及李克氏量尺評量

(Likert-scale)(Zollar, 1992) 等，都是評量學生學習成就或教師教學成效的方式。未來若欲利用短時間群集評量方式在師資檢定上，做為教學成效之考核，宜擴大研究對象，並將其評量結果與其他檢定工具或評量方式之考核結果互相比照。再者未來的研究設計，若能採用另一班「非群集評量」以為對照組，或許會更週延，值得在往後進一步之研究中探討，以確立短時間群集評量的有效性及實用性。本研究屬探索性質，實施之細節可能尚有若干須改進之處，但期望此研究結果能有拋磚引玉的效果，並且做為未來推廣研究之參考。

柒、誌謝

本研究承蒙國家科學委員會補助經費(計畫編號：NSC83-0111-S-017-014F)，乃得完成，謹誌謝忱。

捌、參考文獻

1. 林煥祥(1995)：《由短時間群集評量方式評估新任理化教師之教學》。國科會專題研究計畫成果報告，計畫編號：NSC 83-0111-S-017-014F。
2. American Association for the Advancement of Science (1989). *Science for All Americans*. Washington DC: Author.
3. Aikenhead, G. S., & Ryan, A. G. (1989). *The development of a multiple choice instrument for monitoring views on Science-Technology-Society topics*. Ottawa: Social Science and Humanities Research Council of Canada.
4. Baron, J. B. (1990). Performance assessment: Blurring the edges among assessment, curriculum, and instruction. In Champagne, A. B., Lovitts, B. E. & Calinger, B. J. (Eds.). *Assessment in the Service of Instruction*, pp 127-148. Washington DC: American Association for the Advancement of Science.
5. Brickhouse, N. (1990). Teachers' beliefs about the nature of science and their relation to classroom practice. *Journal of Teacher Education*, 41(3), 53-62.
6. Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experi-*

- mental designs for research. Chicago: Rand McNally and Company.*
7. Carlsen, W. S. (1987, April). Why do you ask? The effect of science teacher subject matter knowledge on teacher questioning and discussion. Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 293 181).
 8. Catanzano, R., & Wilson, M. S. (1977). The effect of retesting contingencies on achievement, anxiety, and attitude in seventh grade science. *Science Education*, 61, 173-180.
 9. Collins, A. (1990). Portfolios for assessing student learning in science: A new name for a familiar idea? In Champagne, A. B., Lovitts, B. E. & Calinger, B. J. (Eds.). *Assessment in the Service of Instruction*, pp 157-166. Washington DC: American Association for the Advancement of Science.
 10. Cook, D. (1985). An objective component for the evaluation of teaching. (Abstract from Dissertation Abstract International, 46, 562).
 11. Druva, C., & Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching*, 20, 467-479.
 12. Duschl, R. A., & Wright, E. (1989). A case study of high school teachers' decision making models for planning and teaching science. *Journal of Research in Science Teaching*, 26(6), 467-501.
 13. Ferguson, P., & Womack, S. T. (1993). The impact of subject matter knowledge and education course work on teaching performance. *Journal of Teacher Education*, 44(1), 55-63.
 14. Fransworth, C. H., & Mayer, V. J. (1984). An assessment of the validity and discrimination of the intensive time series design by monitoring learning differences between students with different cognitive tendencies. *Journal of Research in Science Teaching*, 21(4), 345-355.
 15. Gallagher, J. (1991). Prospective and practicing secondary school science

- teachers' knowledge and beliefs about the philosophy of science. *Science Education*, 75, 121-133.
16. Gay, L. R. (1985). *Educational Evaluation and Measurement*. Merrill: Columbus, Ohio.
 17. Gronlund, N. E. (1985). *Measurement and Evaluation in Teaching*. MacMillan: New York.
 18. Hashweh, M. (1987). Effects of subject matter knowledge in teaching of biology and physics. *Teaching and Teacher Education*, 3, 109-120.
 19. Johnson, D. W. & Johnson, R. T. (1990). Group assessment as an aid to science instruction. In Champagne, A. B., Lovitts, B. E. & Calinger, B. J. (Eds.). *Assessment in the Service of Instruction*, pp 149-156. Washington DC: American Association for the Advancement of Science.
 20. Lantz, O., & Kass, H. (1987). Chemistry teachers' functional paradigms. *Science Education*, 71(1), 117-134.
 21. Lederman, N. G. (1992). Students' and teachers' conception of the nature of science: A review of the research. *Journal of Research in Science Teaching*, 29(4), 331-359.
 22. Lederman, N. G., & Zeidler, D. L. (1987). Science teachers' conceptions of the nature of science: Do they really influence teacher behavior? *Science Education*, 71(5), 721-734.
 23. Mann, L., Taylor, R. S., Proger, B. B., Dungan, R. H. & Tidey, W. (1970). The effects of serial retesting on the relative performance of high and low test anxious seventh grade students. *Journal of Educational Measurement*, 7, 97-104.
 24. Mayer, V. J., & Kozlow, M. J. (1980). An evaluation of a time series single subject design used in an intensive study of concept understanding. *Journal of Research in Science Teaching*, 17(5), 455-461.
 25. Mayer, V. J., & Lewis, D. K. (1979). An evaluation of the use of a time series single subject design. *Journal of Research in Science Teaching*, 16 (2), 137-144.

26. Mayer, V. J., & Rojas, C. A. (1982). The effect of frequency of testing upon the measurement of achievement in an intensive time series design. *Journal of Research in Science Teaching*, 19(7), 543-551.
27. Mwamwenda, T. S., & Mwamwenda, B. B. (1989). Teacher characteristics and pupils' academic achievement in Botswana primary education. *International Journal of Educational Development*, 9(1), 31-42.
28. National Science Teachers Association (1982). *Science-Technology-Society: Science Education for the 1980s*. Position paper. Washington DC: Author.
29. Purser, S. R. (1987). The relationship between teacher effectiveness and teacher evaluation and selected teacher demographic variables. Paper presented at the annual meeting of the American Association of School administrators, New Orleans, LA, February 20-23.
30. Tamir, P. (1993). A focus on student assessment. *Journal of Research in Science Teaching*, 30(6), 535-536.
31. Zollar, U. (1992). Faculty teaching performance evaluation in higher science education: Issues and implication (A "cross-cultural" case study). *Science Education*, 76(6), 673-684.

The Assessment of Beginning Science Teachers' Teaching Effectiveness

Huann Shyang Lin

Department of Chemistry National Kao-Hsiung Normal University

Abstract

The purpose of this study was to explore the validity and feasibility of using time series design in the assessment of beginning science teachers' teaching effectiveness. Four beginning physical science teachers and four classes of 8th grade students who were taught by the four teachers participated in this study. The students' learning achievement on the chapter of chemical reaction was assessed by time series design. The results indicate as predicted that students taught by teachers with high GPAs, higher rated teaching performance, and contemporary views about the nature of science made more significant progress on learning achievement than their counterpart who were taught by teachers with low GPAs, lower rated teaching performance, and more traditional view about the nature of science. The convincing criterion-related validity of the instrument and the procedure's high capability of discriminating high and low achieving students suggest that the time series design is effective and feasible in assessing beginning science teachers' teaching effectiveness. Our results suggest that time series design can play a role in the teacher certification system by serving as an alternative of providing quantitative evidence of practicing teachers' teaching performance. More effort are encouraged to verify the validity and feasibility of time series design with other assessment methods.

Key words: Learning Achievement, Teaching Effectiveness, Time Series Design.