「永續教育發展-創新與實踐論文集」

2010年國際學術研討會
—測驗及評量論文專輯

國家教育研究院

「永續教育發展-創新與實踐論文集」 2010年國際學術研討會—測驗及評量論文專輯
國家教育研究院

永續教育發展-創新與實踐論文集
測驗及評量論文專輯

林宜臻、張芳全、蔡孟憲、林原宏、盧思丞
涂柏原、謝名娟、謝進昌、藍珮君、譚克平、陳昭錦合著

國家教育研究院 編印

目錄

序

值此全球化、在地化、網路化、教育M型化、氣候變遷、環境永續等社會變遷與教育挑戰，教育各界須通力合作、有效因應，才能促進教育的永續發展。對此，《中華民國教育報告書》揭示「精緻、創新、公義、永續」的目標，以實現「新世紀、新教育、新承諾」的願景。

面對此多重挑戰，國家教育研究院於99年10月22日至23日舉辦「2010國際學術研討會：永續教育發展－創新與實踐」，希冀從創新與實踐出發，闡述永續教育發展的圖像與途徑，並期透過學術論述與實務對話，開創永續發展的教育新局。本論文專輯乃針對該次國際學術研討會中，評選為口頭發表之論文，以測驗及評量為主題範疇，徵求作者意願並依本院學術專書審查辦法進行外審，經審查通過後予以收錄之論文。

本書共收錄8篇文章，首先，林宜臻以2篇文章分別探討TASA2009小四、小六數學領域學習成就的標準設定執行過程之妥適性，以及組題與試題位置對試題特性之影響。張芳全則從20個參與TIMSS 2003國家的資料進行分析，以瞭解學生學習成就與學校規模之關係是否有一致趨勢。而蔡孟憲、林原宏則以應用多元計分概念詮釋結構模式（PCAISM）分析方法，並利用模糊集群分析將學生分群，探討各群大學生的微積分概念結構圖。

此外，盧思丞、涂柏原以總字數、認知需求層次與解題所需概念數三個成分，來描述認知成份與TIMSS八年級與國民中學基本學力測驗物理試題難度間的關係。謝名娟、謝進昌之文章，旨在比較書籤標定法與Yes/No Angoff標準設定方法在設定切斷分數上的差異，並使用TASA 2009年英文科來進行研究。藍珮君一文，則探討接受持續且長期的評分訓練後，評分者評分一致性的變化情形，包含與其他評分者之間的一致性，以及評分者自身的一致性。最後，譚克平、陳昭錦建議以概念圖為工具，在現行九年一貫國中自然與生活科技課程綱要，以及99學年度開始實施的高中基礎化學課程綱要之間，建立一個共同架構。

教育的主體為學生，有效的測驗與評量模式更是學生適性成長的重要環節，期盼本論文專輯的收錄與出版，能做為日後測驗與評量相關範疇，注入創新與實踐的新力量。最末，感謝各篇作者們對於各主題撰寫的投入、編輯團隊不辭辛勞的付出與努力，更感謝每篇兩位外審委員細心的評閱與指教，使得本論文專輯能以最佳的論述品質，呈現於國內外研究同好間。

國家教育研究院院長

TASA2009國小四、六年級數學領域學習成就標準設定

林宜臻

TASA2009國小四、六年級數學領域學習成就標準設定

林宜臻
國家教育研究院助理研究員

摘 要

本研究旨在執行TASA2009小四、小六數學領域學習成就的標準設定，並探討該學習成就標準設定過程的妥適性。本研究採Yes/No Angoff法執行設定，並搭配三輪的反覆遞迴操作，提供回饋訊息，以凝聚標準設定成員的共識。研究發現（1）標準設定過程具適切性；（2）PLD共識時間不足值得改善；（3）判斷基準的明確與否成為內部一致性的要素；（4）標準設定結果不足以反映層級表現。建議：（1）評量架構的認知要求宜與政策性定義一致；(2)評量目的宜與表現層級標籤一致；（3）評量架構與PLD宜置於命題前；（4）PLD與難度值的平衡點宜明確化；（5）基礎層級的之政策性定義宜明確化。

關鍵詞：標準設定法、表現層級描述、決斷分數

A research on the standard setting of TASA 2009 math for 4th and 6th grades

Yi-Jen Lin
Assistant Research Fellow, National Academy for Educational Research
jen@mail.naer.edu.tw

Abstract

This study aimed to (1) set cut-scores for mathematical achievement of 4th and 6th grades in Taiwan Assessment of Student Achievement 2009; (2) evaluate the procedure of standard setting; (3) understand the students' mathematical performance in 4th and 6th grades. Since the Angoff method is consistent with features of mathematics, therefore select Yes/No Angoff method for standard setting. Through evaluating the procedures of standard setting, we found: (1) The procedure is relevant；(2) The time of PLD is too short to form common consensus；(3) The criteria of judgment decide the consensus of standard setting. (4) The result of standard setting cannot truly reflect the performance level. The suggestions are as followed: (1) To match the assessment framework with policy definition; (2) To match the assessment purpose with the label of PLD levels; (3) To complete assessment framework and PLD before question designed; (4) To distinguish the role of PLD and P-value; (5) To clarify the definition of basic level.

Keywords: standard setting, performance level description, cut score or benchmark

壹、緒論

一、研究背景

臺灣學生學習成就評量資料庫設置的目的在於：（1）分析臺灣國小四年級、六年級、國中二年級、高中、高職二年級學生之學習成就表現及其關聯因素；（2）探討學生學習成就之表現差異與學習變遷之趨勢，進而檢視目前課程與教學實施成效。

基於當前國家教育體制與政策實施成效之檢視，評定或描述學生的學力表現是否達預期的水準有其必要。「臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement，TASA）」原委託國內五所大學執行標準設定（臺灣學生學習成就評量資料庫網站，2006），以致不同學科有不同的標準設定模式，層級劃分亦有所不同，例TASA2006數學領域採用修正的Angoff 法進行標準設定，分基礎、精熟、進階等三個成就層級（吳宜芳，2007），TASA2005與TASA2006的英文科標準設定，則採用書籤標定法（bookmark）進行通過與不通過之設定（陳彥名，2006）。由於TASA2005年與2006年的數學領域小四、小六的測驗內容，以教育部所公布之「國民中小學九年一貫課程暫行綱要」的能力指標為依據；2007年則以「國民中小學九年一貫課程綱要」的能力指標為依據；2009年則以同綱要的分年細目為依據，評量架構以分年細目替代能力指標。若繼續沿用先前之設定，已不符實務之運用，故引發重新修訂標準設定之需求。一般常以60分及格為標準（或通過分數），由於過於簡化，無法適用於TASA目的。本研究如圖1所示，設定基礎（basic）、精熟（proficient）、進階（advanced）三個層級的決斷分數，以便將學生劃分為基礎以下、基礎、精熟及進階等四個能力區塊，並確保通過某一層級表現的應試者，確實具備該層級的表現水準；而未能通過者，確實未達該表現水準。據此描述受試者能力以及達到該層級所具備的知識與技能，藉此瞭解受試者落於何種層級，得以示責、激勵與善誘。

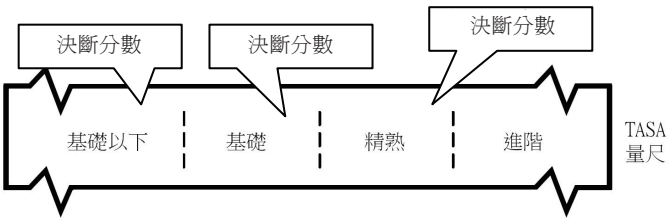


圖1 基礎以下、基礎、精熟及進階等四個能力區塊

二、研究目的

1. 執行TASA2009小四、小六數學領域學習成就的標準設定。
2. 探討標準設定過程的妥適性。

貳、文獻探討

標準設定（standard setting）係指標準設定成員經由一系列判斷過程，建立表現層級間合理界限的決斷點，以區別與鄰近層級的表現，並將此轉成分數量尺的位置（Hambleton, 2001；Cizek, Bunch, & Koons, 2004）。根據標準設定結果，明確描述受試者能力，及其所應具備的知識與技能。本章將探討標準設定的方法及其理論基礎與國際經驗，以為TASA標準設定流程規劃及檢視之參考。

一、標準設定方法

（一）Angoff標準設定法

Angoff法（1971）為較常見的標準設定方法之一，也是美國教育進展評量（National Assessment of Educational Progress, NAEP）大型資料庫所用的標準設定方法。Angoff法為William Angoff（1971）所提出，原始的Angoff法係要求標準設定成員針對每一層級的最低能力受試者（minimally competent examinee）對每一0-1量尺的試題，估計有多少比率的人可以答對此題，而把這個比率當成最低能力受試者答對此題的機率，再將每題可能答對之機率加總平均，便成為該設定者判斷的通過標準，最後將數位設定者判斷的通過標準加以平均，便成為測驗最後的通過標準。由於原始的Angoff 法，須每一位設定成員思索每一層級的最低能力受試者每一題的答對率，當題目眾多時，該種方式相對比較不合適。因此，產生出許多修訂版的Angoff法，其中以Impara 與 Plake （1997）提出的Yes/No Angoff法最被廣泛使用，該方法雖須逐題判斷，但不須逐題估計每一層級的最低能力受試者有多少比率的人可以答對該題，而只須逐題判斷每一層級的最低能力受試者「能否」答對該題，倘若可以答對則寫“Yes”，若不能答對則寫“No”，由於較為直觀判斷，減少原始Angoff法的執行困難度。

（二）bookmark標準設定法

源自於Angoff法的書籤標定法（bookmark）（Lewis, Mitzel, & Green, 1996）相對於Angoff法的逐題檢視，書籤標定法較為簡單、易懂與易執行。而且也能有效融入選擇題型與建構反應試題之標準設定，並適時連結測驗試題內容（item content）與表現層級描述（performance level description, PLD），書籤標定法同時納入試題反應理論（item response theory, IRT）與試題圖（item map）的概念。因此，稱之為修訂的IRT-Angoff法（IRT-Modified Angoff Procedure）（Lewis, Green, Mitzel, Baum, & Patz, 1998）。書籤標定法操作流程如下：

提供按簡單至困難排序的試題卷（ordered item booklets, OIB），標準設定成員逐一檢視OIB中所有試題後，搭配PLD的描述，推想各層級的邊緣學生（borderline students）應具備哪些知識而定，逐一對照OIB中的各個試題內容，挑選出的學生

應該有67%能正確作答的試題，書籤放置之際，兩群試題間應存在較大的知識區隔（Mitzel, Lewis, Patz, & Green, 2001）。將基礎、精熟、進階三個書籤逐一放置於分屬於不同層級的兩兩試題間，即完成三個層級通過分數的設定，將所有學生區分成基礎以下、基礎、精熟與進階等四個能力區塊。完成（1）-（3）第一輪的書籤標定法設定後，再以反覆遞迴的操作過程進行第二、三輪。

（三）Yes/No Angoff標準設定法之適用性

Angoff法（1971）為較常見的標準設定方法之一，採取逐題檢視，判斷過程較書籤標定法（bookmark）耗時，且判斷方式較為繁複，但書籤標定法的執行，係檢視由簡單至困難排序的試題卷，再將基礎、精熟、進階三個書籤逐一放置於分屬不同層級的兩兩試題間，基於數學內容雖具邏輯順序，其難度容易因受試者的解題方式與所備知能，以及命題者出題方式等之影響，又因相同的題目亦可能因組題方式的不同，造成難度值的不同（林宜臻，2010）等之因素，造成由簡單至困難排序的試題卷，未必能完全反映受試者的數學能力值之排序等之考量。由於標準設定方法適當與否甚於是否最佳（Loomis & Bourque, 2001；Reckase, 2000），因此，數學領域未採取較簡便的書籤標定法。此外，TASA數學領域小四與小六的主要試題類型為單一答案的選擇題，屬於0-1量尺類型，所以0-1量尺為前提的Angoff法適用於本研究。由於原始的Angoff 法，並須經由每一位設定成員思索每一層級的最低能力受試者每一題的答對率，再將每題可能答對之機率加總後平均，方成為該設定者判斷的通過標準，最後將數位設定者判斷的通過標準加以平均，才能成為測驗最後的通過標準（Buckendahl, Smith, Impara & Plake, 2002），其過程較為繁瑣，尤其題目眾多時，原始的Angoff 法比較不合適。相對於此，若以Yes/No Angoff法執行標準設定，標準設定者只須逐題判斷各表現層級中的邊緣受試者（marginally acceptable examinee，亦即該層級當中最低能力受試者）「能否」答對該題，由於執行過程較不繁瑣，而且符合數學領域特質。採取 Yes/No Angoff法執行2009年臺灣學生學習成就評量資料庫小四及小六數學領域學習成就的標準設定，較具妥適性。

二、 表現層級標籤之設定

（一）TIMSS 2007

1. TIMSS 2003與 2007小四學生數學各層級表現

國際數學與科學教育成就趨勢調查（The Trends in International Mathematics and Science Study, TIMSS）1995與1999的表現層級分成前10%（90th）、前¼（75th）、前½（50th）、後¼（25th）四大部份，並以正負5分為表現層級區間之範圍。TIMSS 2007 我國小四學生低標（low benchmark）、中標（intermediate benchmark）、高標（high benchmark）、頂標（advanced benchmark）的數學表現層級其分數範圍及各層級所佔比例如下（Olson, Martin, & Mullis, 2008）：

表1 TIMSS 2003與 2007我國小四學生數學各層級表現

層級	低標		中標		高標		頂標	
分數	395分-405分		470分-480分		545分-555分		620分-630分	
年次	%	S.E.	%	S.E.	%	S.E.	%	S.E.
2003年	7	0.2	31	0.7	45	1.1	16	0.9
2007年	7	0.2	26	0.5	42	1.2	24	1.2

2. 決斷分數的定錨題

數學表現層級分數範圍訂定後，選取各決斷分數的定錨題（anchoring items），選題之際，決定於以該層級受試者的答對率至少為.65，而且較低一個層級受試者的答對率至少低於.50，若符合該項標準，則可將該試題訂為該決斷點之定錨題。定錨題標準如下：

表2 TIMSS 2007 數學表現層級定錨題標準

低標	低標層級學生至少有65%的受試者答對該題 。
中標	中標層級學生至少有65%的受試者答對，而且低標層級學生低於50%答對該題。
高標	高標層級學生至少有65%的受試者答對，而且中標層級學生低於50%答對該題。
頂標	頂標層級學生至少有65%的受試者答對，而且高標層級學生低於50%答對該題。

（二）NAEP2009

1. NAEP的政策性定義

政策性定義的功能主要是編寫不同年級與各個學科學習表現層級描述（performance level description, PLD）的起始點，其作用在於概括界定成就層級的內涵。一般而言，政策性定義是由政策決策者制訂，以為學科專家編寫PLD時之參酌，以美國NAEP為例，其政策性定義是由美國的國家評量指導委員會（National Assessment Governing Board，NAGB）訂立，雖歷經多次修訂，但對基礎、精熟與進階之政策性定義，仍保有如下之特點：

基礎：學生學習表現在基礎層級，表示學生具備該年級學習之基本學力達部份精熟程度。

精熟：學生學習表現在精熟層級，表示學生具備紮實的學業表現，能展現學科相關的能力，包含該學科知識、該知識應用於真實情境的能力，並能適當分析該學科知識的能力。

進階：學生學習表現在進階層級，表示學生具有超越精熟層級更卓越的學習表現。

2. NAEP 2007-2009各層級數學表現描述

NAEP 2009針對將小四學生的基礎、精熟、進階層級的數學表現描述如下（National Center for Education Statistics, 2009：18）：

[基礎層級]

屬於基礎層級的四年級學生，應能部份掌握NAEP五大領域內容[1]的數學概念與過程：

能估算及進行簡單整數計算；能瞭解分數與小數；能解決NAEP各領域的一些簡單真實世界問題；能使用（雖未必準確）四種功能的計算機、直尺與幾何繪製用品；書寫的回應量少，而且無可支持的訊息。

[精熟層級]

屬於精熟層級的四年級學生，能統整過程知識及瞭解概念，將其應用於解決NAEP五大領域內容的問題：

能利用整數估算、計算及判斷結果是否合理；能瞭解分數與小數的概念；能解決NAEP領域內容真實世界的問題；能適當使用四種功能的計算機、直尺與幾何繪製用品；能利用如證明與適當訊息的策略以解決問題；能利用可支持的訊息組織與呈現解決方式，並解釋如何完成。

[進階層級]

屬於進階層級的四年級學生，能統整過程的知識與瞭解概念，將其應用於解決NAEP五大領域內容真實世界的非例行性問題：

能解決NAEP五大領域內容真實世界的非例行性複雜問題；能精確使用四種功能的計算機、直尺與幾何繪製用品；能提出邏輯的結論及證明答案，並能解釋如何完成解決過程；他們很明顯能清楚且簡潔的解釋以及溝通他們的思維。

3. NAEP 2005 2007 2009小四學生數學各層級表現

NAEP 2009將小四學生數學表現層級分為基礎（basic）、精熟（proficient）、進階（advanced），其決斷分數為214、249、282，NAEP 2007-2009各層級所佔比率如下：

表3 NAEP 2005 2007 2009小四學生數學各層級表現

層級	基礎以下		基礎		精熟		進階	
年次	%	S.E.	%	S.E.	%	S.E.	%	S.E.
2005	20	0.2	44	0.2	31	0.2	5	0.1
2007	18	0.2	43	0.3	34	0.3	6	0.1
2009	18	0.3	43	0.2	33	0.3	6	0.2

資料彙整來源：http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx

[1]數的屬性與運算、測量、幾何、數據分析/統計/機率、代數等五大領域內容。

參、研究方法

本研究採符合數學領域特質的Yes/No Angoff法執行設定。首先進行小四/小六數學領域 [基礎、精熟、進階等表現層級描述] 的妥適性討論及修訂，進而形成基礎、精熟、進階等各表現層級特徵之共識。搭配三輪的反覆遞迴操作，於第二、三輪標準設定時，提供回饋訊息，以凝聚標準設定成員設定的共識。並對不同背景團體成員間判定差異、極端判定值、輪次間成員判定結果等監控整個標準設定流程之穩定性。每輪設定結束後，施以評估問卷，以瞭解成員對於整個設定流程的想法。

一、執行標準設定

（一）第一輪設定

基於讓標準設定成員不受難度值影響，第一輪只提供每位標準設定成員一本只含有題目內容、選項、答案的小四/小六的TASA數學領域試題卷。每位設定者逐題判斷在該表現層級中的邊緣考生能否答對該題，亦即逐題判斷答對該試題至少需哪一層級程度方能答對，並在紀錄表欄位中打勾√，於第一輪標準設定後，填寫第一輪評估問卷。

（二）第二輪設定

第二輪設定方式與第一輪大致相同，差別只在於回饋訊息。為協助標準設定成員凝聚共識，成員於各輪結束後與新一輪開始之前，將收到回饋訊息。相較於第一輪只提供試題卷內容、選項、答案，第二輪的卷試題卷同時內含試題反應理論（item response theory, IRT）的a、b、c三參數值、古典測驗理論（classical test theory, CTT）的難度值、鑑別度值、各選項的百分比與通過率。此外，回饋訊息尚包括我國TIMSS及NAEP各層級通過百分比（詳見p.8表1及p.9表3），以為瞭解大型測驗的各層級能力的通過人數百分比；並提供含各題各層級選擇百分比，以及基礎、精熟、進階能力別的決斷值與通過百分比、各標準設定成員決斷值設定與平均值之差異分布圖[2]等第一輪設定結果，以為凝聚共識，增進成員內與成員間判定的一致性，於第二輪標準設定後，填寫第二輪評估問卷。

（三）第三輪設定

第三輪提供第二輪設定下結果的回饋訊息，其進行方式與第二輪的差別只在於各委員獨自完成，不再相互討論回饋訊息，於第三輪標準設定後，填寫第三輪評估問卷。

（四）將決斷分數轉換為量尺分數

經由如上三輪的判斷過程，最後建立區分不同的表現層級的決斷分數，再轉換為平均數250、標準差50的量尺分數（Cizek, Bunch,& Koons, 2004）。

[2]以平均值及正負1個標準差方式呈現，讓設定成員判斷其與平均結果的差異。

二、監控與評估標準設定流程

本數學領域標準設定採用如下的方式，進行監控整個標準設定流程的穩定性：

（一）不同背景團體成員間判定差異

針對不同性別、地區及身分類別的標準設定成員判定的通過分數，執行變異數分析，監控不同背景成員間判定結果的一致性。

（二）監控極端判定值的發生

各輪判定後，從分析的資料之中，檢視極端值可能發生的狀況，避免影響最後的通過分數，本研究以該成員設定結果超過該輪全部通過分數平均值以上或以下1個標準差者，視為極端值。

（三）監控第1輪-第2輪、第2輪-第3輪成員判定結果之改變

監控個別成員在第1輪與第2輪之間、第2輪與第3輪之間，其通過分數的變化形態，以成員改變試題判定結果之百分比，監控多少成員更動前一輪的設定結果，並以判定結果變化的平均及其1個標準差範圍，檢視設定結果過於劇烈者，以瞭解成員受到回饋訊息之影響及自身對於所訂立通過分數的信心程度。

（四）實施評估問卷

Cizek與Bunch（2007）建議編製評估問卷，調查成員對於訓練的瞭解程度。本研究請標準設定成員們於每輪設定結束後，填寫評估問卷，以瞭解成員對於整個設定流程的想法。評估問卷1檢視成員對於前導資料、PLD、Yes/No Angoff標準設定方法之解說等之意見，評估問卷2針對回饋訊息的適用性進行調查，評估問卷3則是調查成員對於整個結果的意見與想法。

肆、結果與討論

一、標準設定成員與訓練的適切性

（一）標準設定成員的選擇

標準設定涉及主觀判斷，所以設定成員的選擇與訓練成為標準設定的重要關鍵之一（謝進昌，2006；Hambleton, 2001）。NAGB（1990）是認為16-20人，就能達到心理計量中一定程度的精確性，但ACT（1994）認為在各年段、學科中，若有30名以上的標準設定成員，將提高運用的彈性與決斷分數估計的準確性，而此建議被多數後續研究採納或引用（ACT, 2005）。

本研究的小四及小六標準設定成員各以30名為原則，並以北、中、南、東與離島的教師、學者約20名，行政人員與家長代表約10名的比例，進行立意抽樣。基於標準設定成員組合方式，應多元化及具有代表性（Reckase, 2000）的考量，本研究以多元方式組合標準設定成員。

篩選之際，除了考量成員的人數及區域組成是否具有代表性外，成員對數學領域教學是否熟悉等，也列入條件。選定的學者代表，為現任教數學教育相關科系的教授；教師代表為（曾）擔任數學領域的中央輔導團或縣市輔導團或TASA數學特約命題教師等為主；行政人員代表為（曾）擔任數學領域輔導員；家長代表為三年內子女（曾）在該年段國小就讀，而且具有5年以上的教學經驗者。基於小四與小六教學經驗的多寡，將影響標準設定的共識，因此，小四與小六的標準設定人員，以任教年段的不同分開考量，具有輔導員身分者，則不在此限。基於設定結果信度的考量，如上所述，本研究將成員的人數、區域組成的代表性、任教年段、數學教學經驗等都列為篩選的必要條件，成員包含北、中、南、東與離島，同時兼顧身分別及性別的異質性。

表4 TASA2009數學領域小四與小六正式標準設定成員組成

地區	身分類別	性別				總人數	
		男		女			
		小四	小六	小四	小六	小四	小六
北	學者	0	0	3	2	3	2
	教師	1	2	7	5	8	7
	行政	3	3	1	0	4	3
	家長	1	0	0	0	1	0
中	學者	1	1	0	0	1	1
	教師	2	2	1	1	3	3
	行政	1	1	0	0	1	1
	家長	0	0	0	0	0	0
南	學者	0	0	2	0	2	0
	教師	0	1	2	3	2	4
	行政	0	0	1	1	1	1
	家長	2	2	0	1	2	3
東	學者	0	0	0	0	0	0
	教師	2	1	1	2	3	3
	行政	0	0	1	1	1	1
	家長	0	0	0	0	0	0
總人數		13	13	13	19	32	29

如表4所示，小四標準設定成員計32名，小六29名。小四標準設定成員總教學年資[3]：最低7年、最高36年，平均年資18年；地區分佈：北部16名（50%）、中部5名（16%）、南部7名（22%）、東部4名（13%）；類別分佈：學者6名（19%）、教師16名（50%）、行政人員7名（22%）、家長3名（9%）；性別分佈：男性13名（41%）、女性19名（59%）。

小六成員標準設定成員總教學年資：最低7年、最高32年，平均年資18.5年；地區分佈：北部12名（41%）、中部5名（17%）、南部8名（28%）、東部4名（14%）；類別分佈：學者3名（10%）、教師17名（59%）、行政人員6名（21%）、家長3名（10%）；性別分佈：男性13名（45%）、女性16名（55%）。

如上所示，本研究以北、中、南、東與離島的教師、學者約20名，以及行政人員與家長代表約10名的比例，組成標準設定團隊。成員的人數及區域組成具有代表性。

[3]含行政年資

（二）標準設定成員的訓練

Cizek與Bunch(2007) 認為參與成員資格並無一定的標準，重要的是必須與原先目的相互契合，提出成員訓練的關鍵元素：（1）提供先備資訊（例如：內容標準、評量架構、範例試題、標準設定過程經驗、標準設定目的等）；（2）清楚說明目的與工作任務；（3）熟悉表現標準與標準設定方法等。本研究除會議當日說明標準設定目的、任務及Yes/No Angoff標準設定方法的說明及流程等外，經由Q & A，進行雙方溝通與解釋，以使成員瞭解會議當日標準設定各流程的內涵，並藉由PLD修訂的討論，對PLD形成共識，並於會議舉行之前，經由預先寄送的前導資料，使成員事先瞭解標準設定目的、任務，以及如何設定等有關標準設定的概況。而前導資料內容包括：（1）數學領域標準設定目的及任務的說明；（2）Yes/No Angoff標準設定方法的說明；（3）數學領域小四/小六評量架構；（4）數學領域小四/小六的表現層級描述的修訂流程；（5）正式會議的議程。

84%及89%的小四與小六設定成員認為會議前寄送的前導資料，能幫助他們瞭解會議應扮演的角色；90%設定成員瞭解表現層級描述設定會議的目的；而PLD修訂的討論，有助於形成PLD共識，87%及83%的小四與小六設定成員認為PLD執行方式，有助於修訂表現層級描述。由此可以得知前導資料及標準設定當日的執行方式，對於標準設定成員的訓練，具有適切性。

二、時間分配的適切性

TASA數學領域學科專家建議小四、小六的基礎、精熟及進階等層級之表現層級描述委由標準設定成員處理之。因此，標準設定當日，於「開場說明與演練」之後，另設置檢視「表現層級描述」是否妥適的時段。87%及83%的小四與小六設定成員認為PLD執行方式有助於修訂基礎、精熟與進階表現層級的描述，但與標準設定於同天進行，不但壓縮標準設定時間，並造成PLD共識時間不足的現象，因此，只有37%小四標準設定委員認為「各階段任務的執行時間分配」長度適合。由此可以得知標準設定當日同時檢視「表現層級描述」，值得未來進一步改善。

有鑒於首日小四有些標準設定成員，未能及時完成，造成編碼與統計分析的延誤，而小六成員執行標準設定之際，則適時提醒小六設定成員對於時間的掌控。此外，在「開場說明與演練」之際，就先確立提供的回饋訊息只當成參考點，而由設定成員根據自身經驗判斷層級，以解決前一日小四評量設定時，發現試題未能契合「表現層級描述」的諸多衝突點。由於層級判斷準則於設定前確立，大幅縮短設定的時間。因此針對「各階段任務執行時間分配長度」的認同度，由前一日的37%提升至88%。

三、內部的效度

本節分析小四與小六數學領域全部標準設定成員在各輪的基礎、精熟、進階等表現層級所設定的決斷分數，以為瞭解內部的效度。

如表5所示，不論是在哪一輪或哪一個層級，被界定為極端值者相當少，顯示TASA小四數學領域所設定結果，受極端值的影響並不大。

表5 TASA小四數學領域全部標準設定成員設定之決斷分數一覽表

代號	匿名	身分	年資[4]	區域	性別	基礎			精熟			進階		
						輪次			輪次			輪次		
						1	2	3	1	2	3	1	2	3
1	**雝	行政	31	南	女	228.19	184.17	187.79	**387.65**	316.83	291.01	387.65	387.65	387.65
2	**琪	教師	15	南	女	226.54	183.12	112.8	272.58	287.59	228.28	387.65	387.65	387.65
3	**豐	家長	18	南	男	228.05	189.71	177.85	316.83	275.39	272.52	387.65	387.65	387.65
4	*基*	家長	19	北	男	118.91	172.71	110.62	272.41	316.08	228.19	387.65	387.65	387.65
5	**儒	教師	17	中	男	205.92	183.61	175.35	367.98	346.8	317.02	387.65	387.65	387.65
6	**男	行政	20	北	男	183.94	183.92	172.75	325.39	321.65	316.83	387.65	387.65	387.65
7	*慶*	教師	15	中	男	169.28	156.3	127.2	253.6	261.89	237.92	387.65	387.65	387.65
8	**智	教師	30	東	女	183.33	183.19	136.67	272.52	272.52	272.46	387.65	387.65	387.65
9	**葳	教師	17	北	女	131.65	165.58	111.25	273.31	306.15	268.96	387.65	387.65	387.65
10	*昔*	教師	26	北	女	227.83	234.73	228.02	334.45	**367.99**	**348.93**	387.65	387.65	387.65
11	*正*	學者	28	中	男	228.06	218.91	180.5	316.83	332.22	318.02	387.65	387.65	387.65
12	**葉	教師	9	南	女	173.38	188.4	168.07	272.46	316.83	306.41	387.65	387.65	387.65
13	**謙	教師	8	北	男	228.02	228.03	228.07	316.83	296.02	316.83	387.65	387.65	387.65
14	*如*	學者		南	女	138.97	170.89	179.95	316.83	316.83	317.45	387.65	387.65	387.65
15	**瑩	行政	19	東	女	228.05	227.92	143.55	316.54	317.37	272.54	387.65	387.65	387.65
16	**蘭	教師	21	北	女	171.69	182.94	*[5]	347.66	356.61	*	387.65	387.65	*
17	*煥*	行政	19	北	男	183.55	*[6]	124.21	316.83	*	272.52	387.65	*	387.65
18	**順	行政	12	北	男	118.07	136.81	133.26	272.53	286.36	272.52	387.65	387.65	387.65
19	*月*	行政	27	北	女	165.63	142.21	138.44	293.37	272.52	272.52	387.65	387.65	387.65
20	**詠	行政	13	中	男	189.72	191.88	182.94	275.57	303.49	272.58	387.65	387.65	387.65
21	*映*	教師	13	中	女	181.16	194.07	183.2	**231.71**	282.91	272.52	387.65	387.65	387.65
22	*敏*	教師	21	北	女	178.77	184.26	135.14	316.83	316.99	272.55	387.65	387.65	387.65
23	*昱*	教師	10	東	男	228.05	**272.24**	270.13	318.52	336.35	318.64	387.65	387.65	387.65
24	*幸*	學者		北	女	184.83	228.04	191.11	312.93	307.1	275.97	387.65	387.65	387.65
25	*旻*	學者		南	女	225.12	191.36	183.27	314.63	316.7	278.33	387.65	387.65	387.65
26	**梅	教師	20	北	女	134.59	175.19	146.82	272.52	272.53	272.51	387.65	387.65	387.65
27	*鳳*	教師	27	北	女	128.90	180.26	124.4	316.2	318.53	272.52	387.65	387.65	387.65
28	*偉*	家長	7	南	男	179.58	143.83	175.84	272.26	**239.41**	228.08	387.65	387.65	387.65
29	*寶*	教師	21	北	女	138.14	182.98	183.32	272.52	272.52	272.52	387.65	387.65	387.65
30	*寰*	教師	11	東	男	190.64	222.31	182.28	314.22	272.56	272.52	387.65	387.65	387.65
31	*靜	學者	36	北	女	227.52	182.7	*	316.83	316.54	*	387.65	387.65	*
32	*寧*	學者		北	女	226.87	228.12	227.99	340.92	359.27	**352.63**	387.65	387.65	387.65
平均數						186.03	190.66	167.43	303.82	305.89	283.08	387.65	387.65	387.65
中位數						183.75	183.92	175.59	315.42	316.08	272.53	387.65	387.65	387.65
標準誤						6.68	5.55	7.11	6.15	5.63	5.84	0	0	0
標準差						37.26	29.88	38.97	33.74	30.62	31.99	0	0	0
上2標準差						260.55	250.42	245.37	371.3	367.13	347.06	387.65	387.65	387.65
下2標準差						111.51	130.9	89.48	236.34	244.65	219.1	387.65	387.65	387.65
超過2標準差之總人數						0	1	0	2	2	2	0	0	0

*粗體數值代表極端值(決斷分數平均數上下2個標準差)

[4]教學年資含行政年資；學者身分者，若無小學教學經驗，則教學欄位以空白呈現。

[5]16及31號委員因故第三輪標準設定請假。

[6]17號委員因第二輪標準設定前，與其他委員討論試題，想法受影響，怕影響設定的信效度，故第二輪未設定。

如表6所示，不論是在哪一輪或哪一個層級，被界定為極端值者相當少，顯示TASA小六數學領域所設定結果，受極端值的影響也不大。

表6 TASA小六數學科全部標準設定成員設定之決斷分數一覽表

代號	匿名	身分	年資[7]	區域	性別	基礎			精熟			進階		
						輪次			輪次			輪次		
						1	2	3	1	2	3	1	2	3
1	**雝	行政	31	南	女	228.19	192.11	184.25	289.90	272.23	272.23	369.06	369.06	369.06
2	**豐	家長	18	南	男	183.64	183.58	178.88	272.22	272.22	272.23	369.06	369.06	369.06
3	*基*	行政	19	北	男	187.38	203.97	189.41	272.23	272.24	272.23	369.06	369.06	369.06
4	**儒	教師	17	中	男	183.30	183.71	183.78	316.61	316.56	316.47	369.06	369.06	369.06
5	**男	教師	21	北	男	146.25	151.05	183.33	272.23	272.23	272.23	369.06	369.06	369.06
6	*慶*	教師	15	中	男	170.81	164.29	157.09	272.21	272.21	272.23	369.06	369.06	369.06
7	**智	教師	30	東	女	213.09	182.00	183.31	272.23	272.23	272.28	369.06	369.06	369.06
8	*昔*	教師	26	北	女	228.58	**272.23**	228.19	335.35	**350.63**	320.37	369.06	369.06	369.06
9	**葉	教師	9	南	女	183.09	183.77	183.65	272.23	272.23	272.24	369.06	369.06	369.06
10	**謙	教師	8	北	男	219.59	183.95	180.80	316.60	302.22	316.61	369.06	369.06	369.06
11	**珍	學者		北	女	182.54	228.19	228.19	272.23	272.28	272.23	369.06	369.06	369.06
12	**瑩	行政	19	東	女	208.14	184.13	184.09	272.24	293.16	283.93	369.06	369.06	369.06
13	*雪*	教師	32	南	女	184.68	228.19	186.21	316.61	**359.32**	316.61	369.06	369.06	369.06
14	**泰	家長	11	南	男	186.81	186.59	183.55	278.06	285.87	314.09	369.06	369.06	369.06
15	**玲	教師	25	東	女	228.19	227.55	219.43	316.61	272.77	276.00	369.06	369.06	369.06
16	**順	行政	12	北	男	182.59	151.72	171.24	272.23	272.23	272.23	369.06	369.06	369.06
17	**詠	行政	13	中	男	185.59	183.79	*[8]	272.23	278.67	*	369.06	369.06	*
18	**永	教師	18	南	女	228.13	**272.23**	228.19	316.61	316.70	311.80	369.06	369.06	369.06
19	*映*	教師	13	中	女	184.21	183.83	183.83	259.34	272.23	272.23	369.06	369.06	369.06
20	*自*	學者	15	中	男	185.03	186.16	183.83	280.50	273.03	280.74	369.06	369.06	369.06
21	**華	教師	20	北	女	183.84	182.99	183.60	315.49	316.61	316.61	369.06	369.06	369.06
22	**梅	教師	20	北	女	183.64	183.49	183.84	272.23	272.23	272.23	369.06	369.06	369.06
23	**琴	教師	27	北	女	183.10	177.65	183.80	272.23	272.23	272.29	369.06	369.06	369.06
24	*寶*	教師	21	北	女	228.19	228.19	228.19	313.05	316.61	316.28	369.06	369.06	369.06
25	*寰*	教師	11	東	男	190.27	195.69	217.71	272.23	272.23	272.26	369.06	369.06	369.06
26	*偉*	教師	7	南	男	**140.39**	176.33	155.87	228.19	228.19	228.19	369.06	369.06	369.06
27	**雯	家長	20	南	女	**272.23**	184.34	182.79	**369.06**	319.35	316.61	369.06	369.06	369.06
28	*寧*	學者		北	女	228.12	210.85	224.37	317.99	315.52	316.61	369.06	369.06	369.06
29	**妙	行政	19	北	男	183.85	183.44	183.83	272.24	272.23	272.23	369.06	369.06	369.06
平均數						196.33	195.03	191.61	289.01	288.15	287.22	369.06	369.06	369.06
中位數						185.03	183.95	183.83	272.24	272.28	272.29	369.06	369.06	369.06
標準誤						5.13	5.41	3.98	5.34	5.19	4.49	0	0	0
標準差						27.63	29.12	21.05	28.78	27.94	23.76	0	0	0
上2標準差						251.6	253.28	233.71	346.57	344.04	334.75	369.06	369.06	369.06
下2標準差						141.06	136.79	149.52	231.44	232.27	239.7	369.06	369.06	369.06
超過2標準差人數						2	2	0	1	2	0	0	0	0

*粗體數值代表極端值(決斷分數平均數上下2個標準差)

[7]教學年資含行政年資；學者身分者，若無小學教學經驗，則教學欄位以空白呈現。
[8]17號委員因故，第三輪標準設定請假。

由圖2~圖4所示，小四設定成員執行第二輪設定時，針對第一輪設定，進行相當幅度的修訂，造成第一輪和第二輪的決斷分數差異性很大。原因在於第一輪設定時，只提供試題內容，第二輪設定時，試題卷同時內含試題反應理論（IRT）的a、b、c三參數值、古典測驗理論（CTT）的難度值、鑑別度值、各選項的百分比與通過率、每題各層級反應百分比等回饋訊息。由評估問卷結果，可以得知74%設定成員受回饋訊息影響，而且七成以上的成員，在第一輪設定時根據PLD設定，然而設定成員依Yes/No Angoff標準設定法逐題配合PLD進行判別之際，有如下衝突點：

(1) 試題內容含非單一的分年細目
(2) 同細目內含數項數學概念
(3) 試題與所列分年細目無法匹配[9]
(4) 同細目但解題所需概念及解題步驟計算等複雜程度不同
(5) 僅評量該年級指標內涵，忽略基礎層級題目[10]
(6) 題型變化不足，造成進階層級題數不足[11]

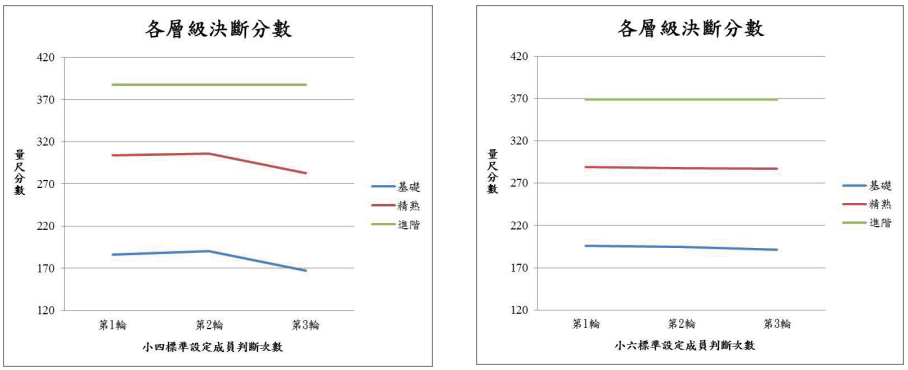


圖2 TASA2009數學科小四（左圖）、小六標準設定成員於三輪設定結果趨勢圖

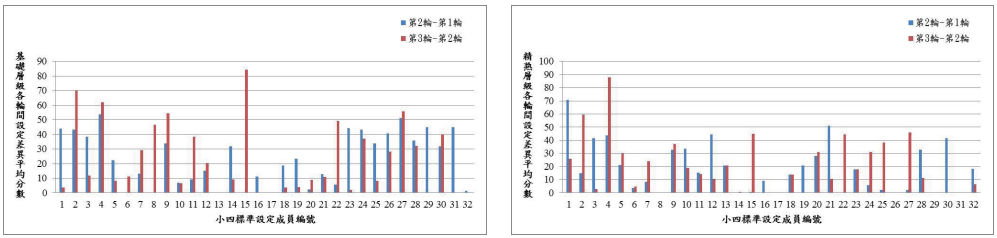


圖3 小四標準設定成員於基礎（左圖）、精熟層級各輪之間設定差異變化

[9] 24、25、47、52、54、55、56、64、77、88、92、96、98等題目的分年細目不對應。
[10] 導致面積與周長題數偏多，但數的位值與分解合成偏少。
[11] 造成以通過率為標準，將學生應了解的精熟，誤以為可以進階層級的挑戰題。

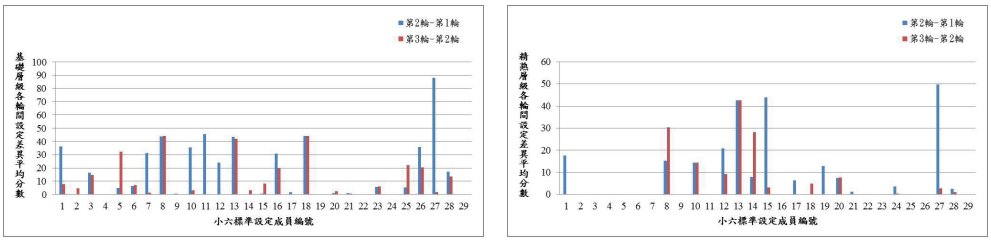


圖4 小六標準設定成員於基礎（左圖）、精熟層級各輪之間設定差異變化

如上所述，試題、分年細目與題型三者未能契合「表現層級描述」。有些題目依PLD敘述，是屬於進階層級試題，但由於試題設計方式未能與之匹配，試題複雜度及應用度偏低易解[12]，造成基礎、精熟、進階層級分類不易，影響成員們判斷區分時的不易。成員們在回饋評估問卷的書面意見提出：（1）如果以難度值與通過率判準，而降低標準，將無法藉此看到現場學生學習所呈現的問題，達成改善教學之成效；（2）判斷為基礎層級的試題，若學生表現通過率低[13]，正可反映教學現場的老師對此部分的忽略，或給學生操作理解的機會較少，而有些被判斷為精熟的試題，學生表現通過率高，也可反映出此部分老師教學成功；（3）有些基礎層級題目的通過率偏低是受題目資訊與描述的影響，因此將需要多一些的解讀步驟或涵蓋變化較大試題的決斷層級上移一層；（4）將回饋訊息視為參考點，據實呈現數據，以反映現場老師的解讀與轉化課程及其著力點如何影響學生的學習表現。如表7所示：每層級的各輪的波動大，顯示整體成員依據自我對於各層級判斷基準，進行不斷修正。

小六標準設定成員認為數據宜據實呈現，而視回饋訊息為參考點，以反映現場老師的解讀與轉化課程及其著力點如何影響學生學習表現，藉此看到現場學生學習所呈現的問題，以達改善教學之效。如表6以及圖2~圖4所示，相較於小四各輪的層級決斷分數的標準誤，小六整體決斷分數的標準誤較低，而且每層級的各輪之間波動小。由評估問卷的結果，可以發現小四設定成員七成以上，而小六五成多認為表現層級描述有助於層級判斷，係由於小六標準設定成員於設定前的共識降低成員對PLD的倚賴度。小四只有30%成員對於最後標準設定的結果具有信心，而小六設定成員達78%；小四只有45%對於最後的決斷分數感到滿意達，而小六成員達85%。由此，可以得知判斷基準的明確與否，成為內部一致性的要素之一。

四、不同背景之標準設定成員間評定差異分析

檢視不同性別、地區及身分別的標準設定成員，其決斷分數差異如下：

[12]試題反應理論計算的難度值及通過率也反應此現象。
[13]分數概念試題通過率普遍偏低。

(一) 不同性別成員設定之差異性

不同性別的標準設定成員所設立的決斷分數，其差異分析如表7所示：除小四基礎層級外，女性標準設定成員所設立的決斷分數都高於男性；小六不同性別成員間決斷分數的差距均較小四高。

表7 不同性別成員標準設定結果之差異分析

層次	性別	人數		決斷分數		均差絕對值	
		小四	小六	小四	小六	小四	小六
基礎	男	13	12	172.38	180.78	8.74	18.96
	女	17	16	163.64	199.74		
精熟	男	13	12	280.32	280.15	4.86	12.38
	女	17	16	285.18	292.53		
進階	男	13	12	387.65	369.06	0.00	0.00
	女	17	16	387.65	369.06		

(二) 不同地區別設定之差異性

不同地區別的標準設定成員所設立的決斷分數，其差異分析如表8所示：基礎層級以東部及離島的標準設定成員所設立的決斷分數最高；精熟層級小四以北部最高，小六以南部最高；小六不同地區別成員間決斷分數的平均數差距均小於小四。

表8 不同地區別成員標準設定結果之差異分析

層次	地區別	人數		決斷分數		均差絕對值	
		小四	小六	小四	小六	小四	小六
基礎	北部	13	13	166.33	190.71	23.26	13.79
	中部	5	5	169.73	187.34		
	南部	8	6	159.90	190.79		
	東部及離島	4	4	183.16	201.13		
精熟	北部	13	13	291.46	286.21	22.75	17.08
	中部	5	5	283.50	291.59		
	南部	8	6	268.71	293.20		
	東部及離島	4	4	284.04	276.12		
進階	北部	13	13	387.65	369.06	0	0
	中部	5	5	387.65	369.06		
	南部	8	6	387.65	369.06		
	東部及離島	4	4	387.65	369.06		

(三) 不同身分別設定之差異性

如表9所示：除精熟層級小六家長代表的決斷分數最高外，其他則以學者設定最高；小六不同身分別成員間決斷分數的差距均較小四小。

表9 不同身分別成員標準設定結果之差異分析

層次	身分別	人數		決斷分數		均差絕對值	
		小四	小六	小四	小六	小四	小六
基礎	學者	5	3	192.56	212.13	37.85	30.39
	教師	15	17	167.51	192.40		
	行政人員	7	5	154.71	182.56		
	家長代表	3	3	154.77	181.74		
精熟	學者	5	3	308.48	289.86	65.55	26.41
	教師	15	17	283.37	288.05		
	行政人員	7	5	281.5	274.57		
	家長代表	3	3	242.93	300.98		
進階	學者	5	3	387.65	369.06	0	0
	教師	15	17	387.65	369.06		
	行政人員	7	5	387.65	369.06		
	家長代表	3	3	387.65	369.06		

如表7~表9所示，不同背景別有如下設定的差異性：（1）除小四基礎層級外，女性標準設定成員所設立的決斷分數都高於男性；（2）基礎層級以東部及離島的標準設定成員所設立的決斷分數最高，小四精熟層級以北部最高，小六以南部最高；（3）精熟層級以小六家長代表的決斷分數最高外，其他則以學者設定最高。由於試題、分年細目與題型三者未能契合「表現層級描述」，造成不同背景變項影響內部設定的一致性。如何平衡既有之失衡現象，尚有努力的空間。

五、表現層級描述與評量架構間的平衡點

標準設定結果發現小四及小六進階層級比率有相當偏低現象。雖Reckase 與 Bay（1999）曾指出Yes/No Angoff法，容易把較低層級的決斷分數設得過低，而把較高層級的決斷分數設得過高。但相較TASA精熟層級與我國小四學生在TIMSS2007的數學頂標層級表現佔24%（Olson, Martin, & Mullis, 2008），其二者之差距甚大，而美國小四學生在其國內測驗NAEP 2009的數學進階層級也達6%，究其因在於TASA屬於進階層級的題數相當少。美國NAEP2009認為學生的數學學習表現若屬進階層級須「能統整過程的知識與瞭解概念，並將其應用於解決NAEP五大領域內容的真實世界的非例行性複雜問題。」，以及「能精確使用四種功能的計算機、直尺與幾何繪製用品；能提出邏輯的結論及證明答案，並能解釋如何完成解決的過程；能清楚且簡潔地解釋及溝通他們的思維。」（National Center for Education Statistics, 2009: 18）。換言之，NAEP2009對進階層級學生的要求是必須能應用該年級所學，而且解決的是非例行性問題，而非一般性問題。此外，NAEP 2009數學架構的主要評量向度除了內容領域外，另一向度則是對認知的要求（cognitive demands），要求的是低、中、高三種的數學複雜度（mathematical complexity），並與基礎、精熟、進階三個層級相互匹配（National Assessment Governing Board, 2008; National Center for Education Statistics, 2009, September 30）。

相較於NAEP要求複雜度，我國TASA2009評量架構中的認知要求是「概念理解」、「程序執行」、「解題與思考」。TASA依此評量架構設計評量題，政策性定義卻又參照NAEP設定的基礎、精熟、進階三個層級，但未結合數學複雜度。由於要求的向度不同，造成各個層級題數不均現象，所以屬於進階層級題數不足，造成拉高該層級決斷分數的現象。若又未能分散置於不同題本，將造成即便已達進階層級者，無相對題目可以作答的現象；而未具此能力者，亦因此導致總答對題數偏少現象。

六、評量目的與表現層級標籤的選擇

TASA與TIMSS的評量目的，主要是瞭解學校課程實施狀況，而NAEP的數學評量架構並不是課程架構（curriculum framework），其主要回答的是哪些數學技能（mathematics skills）應該列入評量，而非回答哪些或如何進行數學教學。因此，即便是學校課程重點的數學概念與技巧，NAEP並未將其納入（National Assessment Governing Board, 2008）。

TIMSS的表現層級標籤的頂標、高標、中標與低標，直接來自序位的前⅒、¼、½，以及後¼（Olson, Martin, & Mullis, 2008）。而NAEP則是根據表現層級描述，設定基礎、精熟、進階三個層級的決斷分數，將學生劃分為基礎以下、基礎、精熟及進階等四個能力區塊。我國數學評量目的與NAEP不同，而與TIMSS相同，但採用與NAEP相同的表現層級，我國評量目的宜與表現層級標籤一致。

七、PLD與難度值的平衡點

由於試題設計方式未能與之匹配，標準設定之際發現：依據PLD屬於進階試題，試題複雜度及應用度有偏低易解的現象；而屬於基礎層級試題，卻因資訊及描述方式，影響通過率偏低。小六標準設定成員認為：（1）若因難度值高與通過率低，而降低標準提高層級，將無法藉此看到現場學生學習所呈現的問題；（2）判斷為基礎層級的試題，若學生表現通過率低，正可反映教學現場的老師對此部分的忽略，或給學生操作理解的機會較少；（3）被判斷為精熟的試題，學生表現通過率高，可反映出此部分老師教學成功。因此，小六標準設定成員將IRT的參數值等回饋訊息視為參考點，將需要較多解讀步驟或涵蓋變化較大試題的決斷層級上移一層，以反映現場老師對課程的解讀、轉化及著力點，藉此看到學生學習上的問題。如上的層級判斷準則確認下，小六標準設定成員降低對PLD的倚賴度，而以教學經驗中學生對此問題的通過率，以及根據解題難易程度執行標準設定。小六標準設定成員的因應的模式，正反應NAEP以複雜度為認知要求的評量架構，所以小六整體標準設定的決斷分數其標準誤低於小四標準設定，而且每層級的輪與輪之間的波動小。

伍、結論與建議

本研究目的在於執行2009年TASA小四、小六數學領域學習成就的標準設定，以區分學生在基礎以下、基礎、精熟、進階等不同層級的表現，並探討該學習成就標準設定過程的妥適性。以下茲針對研究所得結論，進行說明，並提出建議，作為未來日後執行標準設定之參考。

一、結論

（一）標準設定過程具適切性

標準設定方法雖有差異，但具有如下相同的核心要素：（1）優質的標準設定成員；（2）合宜的訓練與逐題判斷的過程；（3）裨益於判斷的回饋訊息；（4）省思與修正等之程序（Reckase, 2000）。依此核心要素檢視，本研究具有相當的適切性。

（二）PLD共識時間不足值得改善

87%及83%的小四與小六設定成員認為PLD執行方式有助於修訂基礎、精熟與進階表現層級的描述，但與標準設定同一天進行，不但壓縮標準設定時間，並造成PLD共識時間不足的現象，值得未來進一步改善。

（三）判斷基準的明確與否成為內部一致性的要素

不同背景別有如下設定的差異性：（1）除小四基礎層級外，女性標準設定成員所設立的決斷分數都高於男性；（2）基礎層級以東部及離島的標準設定成員所設立的決斷分數最高，小四精熟層級以北部最高，小六以南部最高；（3）精熟層級以小六家長代表的決斷分數最高外，其他則以學者設定最高。係由於試題未能契合「表現層級描述」的諸多衝突點，尤其在基礎層級的認定上，產生最大的分歧，所以成員內設定的標準誤高達7.11。小六成員執行標準設定前，由於先行確立層級判斷準則，所以成員內設定的標準誤縮至3.98，而且第三輪的標準誤在三輪中最小，每層級各輪之間的波動小。小六設定成員78%對於最後標準設定的結果具有信心，而小四只有30%；小六成員85%對於最後的決斷分數感到滿意達，小四只有45%。由此可以得知判斷基準的明確與否，成為內部一致性的要素之一。

（四）標準設定結果不足以反映層級表現

整體而言，小四與小六都是屬於基礎層級者最多，分別佔全體的55.32%與58.34%；精熟層級者次之，佔全體的36.49%與27.05%；再其次為基礎以下層級者，佔全體的7.97%與14.42%；進階層級者最少，佔全體的0.22%與0.19%。標準設定結果有均差絕對值偏高現象，雖可呈現不同背景別設定的差異性，由於內部一致偏低，標準設定結果不足以反應不同層級的表現，但設定結果足以反應評量架構等尚有待努力。

二、建議

Loomis & Bourque（2001）認為標準設定方法適當與否甚於是否最佳，並提出判斷標準設定方法的適切性準則，允許參與者思考更複雜的各個面向。本研究主要藉由執行TASA2009小四、小六數學領域學習成就的標準設定，瞭解小四與小六學生在不同層級的表現。標準設定過程雖具妥適性，卻發現標準設定結果不足以反映學生在不同層級表現，其原因在於如建議所示，尚有若干值得改善的空間，倘若設定前條件具足，將更能反映學生在不同層級的表現，裨益於檢視數學課程實施成效。

（一）評量架構的認知要求宜與政策性定義一致

TASA標準設定之際，若「政策性定義」繼續參照NAEP設成基礎、精熟、進階三個層級基礎，評量架構除內含內容領域向度外，認知的要求宜納入低、中、高等不同複雜度。

（二）評量目的宜與表現層級標籤一致

我國數學評量目的與NAEP不同，而與TIMSS相同，但採與NAEP相同的表現層級，評量目的宜與表現層級標籤一致。

（三）評量架構與PLD宜置於命題前

評量工具建置的前後流程，宜「釐清測驗目的」→「訂立學科評量架構」→「撰寫表現層級描述（PLD）」→「編製測驗內容」，最後才執行標準設定。如此，除了能避免標準設定當日須另外設置檢視「表現層級描述」是否妥適的時段，造成標準設定時間的壓縮與PLD共識時間不足的現象，最重要的是，試題撰寫內容能明顯區隔不同的層級。

（四）PLD與難度值的平衡點宜明確化

上述「評量架構的認知要求與政策性定義一致」、「評量目的與表現層級標籤一致」與「評量架構與PLD宜置於命題前」未能付諸實施前，若需執行標準設定，宜將PLD與難度值各自明確定位，形成判斷準則，以提高成員內部設定的一致性。

（五）基礎層級的政策性定義宜明確化

「學生學習表現在基礎層級，表示學生具備該年級學習之基本學力達部分精熟程度」這是TASA工作推動委員會對基礎層級政策的定義。「該年級學習之基本學力」應該指的是該年級之前的學習內容，小六TASA數學試題尚包括小四的垂直定錨題[14]而非僅考該年級的內容尚可適用，但也未必是該年級學習的基本學力。若TASA日後試題仍維持僅考該年級的內容，宜更明確將基礎層級定位為「該年級較基礎的學習內容達部分精熟程度」。

[14]定錨題必須具備高鑑別度和適切的難度，例小四試題被選為和小六的定錨題，則須該定錨題的鑑別度高且較難的試題。

參考文獻

吳宜芳（2007）。**標準設定效度議題之探究：以數學學習成就評量為例。**國立台南大學測驗統計研究所碩士論文，未出版，台南。

林宜臻（2010）。**2006-2007年台灣學生學習成就評量資料庫（TASA）數學領域小六評量架構與試題分析之研究（未出版）。**台北縣：國家教育研究院籌備處。

國家教育研究院籌備處（2005）。**2005年臺灣學生學習成就評量資料庫數學領域評量結果報告（未出版）**。台北縣：國家教育研究院籌備處。

國家教育研究院籌備處（2006）。**2006 年臺灣學生學習成就評量資料庫數學領域評量結果報告（未出版）。**台北縣：國家教育研究院籌備處。

國家教育研究院籌備處（2007）。**2007年臺灣學生學習成就評量資料庫數學領域評量結果報告（未出版）。**台北縣：國家教育研究院籌備處。

陳彥名（2006）。**台灣學生學習成就資料庫（TASA）英語聽讀能力標準設定之效度探討**。國立台北教育大學教育心理與諮商學系碩士論文，未出版，台北。

臺灣學生學習成就評量資料庫網站（2006）。**臺灣學生學習成就評量資料庫建置計畫**。2009年12月20日，取自：http://tasa.naer.edu.tw/plan.htm

謝進昌（2006）。精熟標準設定方法的歷史演進與詮釋的新概念。**嘉義大學國民教育研究學報，16**，157-193。

American College Testing[ACT](1994). *Setting achievement levels on the 1994 National A*ssessment of Educational Progress in geography and in U.S. history and the 1996 *National Assessment of Educational Progress in science(Final version)(Design document).*Washington, DC: National Assessment Governing Board.

American College Testing[ACT](2005). *Developing achievement levels on the 2005 Nation*al Assessment of Educational Progress in grade twelve mathematics: Process *report.* Washington, DC: National Assessment Governing Board.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp.508-600). Washington, D.C.: American Council on Education.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253-263.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, California: Sage Publication Ltd.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary Methods. *Educational Measurement, 23*(4), 31-50.

Hambleton, R. K. (2001). *Setting performance standards on educational assessments and cri*teria for evaluating the process. In Gregory J. Cizek(Ed), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp.89-116). NJ: Lawrance Erlbaum Associates.

Impara, J. C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 355-368.

Lewis, D. M., Mitzel, H.C., & Green, D. R. (1996). *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Boulder, CO.

Lewis, D.M., Green, D. R., Mitzel, H. C., Baum. K., & Patz, R. J. (1998, April).*The bookmark standard setting procedure: Methodology and recent implementations.*Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J.Cizek (Ed.).*Standard setting: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark method: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance.standards :Concepts,methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

National Assessment Governing Board[NAGB](1990). *Setting appropriate achievement lev*els for the National Assessment of Educational Progress: Policy Framework and *Technical Procedures.* Washington, DC: Author.

National Center for Education Statistics (2009, September 30). *What Does the NAEP. Mathematics Assessment Measure?* Retrieved from http://nces.ed.gov nationsreportcard/mathematics/whatmeasure.asp

National Center for Education Statistics [NCES] (2009). *The Nation's Report Card: Mathematics 2009* (NCES 2010–451). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Olson, J., Martin, M., & Mullis, I. (2008). *TIMSS 2007 technical report.* Chestnut Hill, MA : TIMSS & PIRLS International Study Center, Boston College.

Reckase, M. D. (2000). *The evolution of the NAEP achievement level setting process: A summary of the research and development efforts conducted by ACT.* Iowa City, IA: ACT.

Reckase, M. D., & Bay, L. (1999). *Comparing two methods for collecting test-based judgments.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

【致謝辭】

感謝審查委員的肯定與鼓勵，專業的相挺是這篇論文得以順利產出的關鍵：感恩總計畫主持人林世華教授的領航，以及各子計畫主持人曾建銘、謝進昌、謝名娟等位好夥伴們對於宜臻的質疑，總是不厭其煩一一解答；感謝數學教育大老們的專業相挺，讓標準設定得以順利成功；謝謝張宛婷、林妲君、林哲慈、吳嘉峰、陳筱琦、蕭錡凌、李薇、童育緩、蔡佩儒、許思雯、李慧雯、葉善炫等位助理們之相助，使得標準設定流程得以有條不紊執行。

組題與試題位置對試題特性之影響

林宜臻

組題與試題位置對試題特性之影響

林宜臻
國家教育研究院助理研究員

摘要

本研究旨在探討組題與試題位置對試題特性之影響，提供日後組題者之參考。本研究採用古典測驗理論與試題反應理論三參數模式，以BILOG-MG、TEST2 等針對試題難度、鑑別度與誘答力等進行共同試題特性分析及比較。由於TASA 2006 之題本等化設計採取PBIB 設計，而TASA 2007 則採取NEAT 設計；又TASA 2007 為避免試題受順序的影響，各題本的定錨試題的題數與施測順序皆相同，並將定錨區塊置前，非定錨區塊置後，此相較於採取PBIB 設計的TASA 2006，其定錨題未必將全數置前，讓研究者得以藉此分析定錨題是否因試題擺放位置影響試題參數。

本研究發現：（1）試題置前的定錨試題的八成以上，其鑑別度值與CTT 通過率高於置後的定錨試題；（2）近七成置後定錨試題的IRT 難度值高於置前的定錨試題；（3）相同定錨題前的試題平均難度值為0.12 的相同試題其難度值為0.96，相同定錨題前的試題平均難度值為0.48 的相同試題其難度值為1.72，倘若定錨題前的試題平均難度值差距大時，即便是相同的題目，也會造成難度值相差高達0.76 的現象；（4）若相同題的前一區塊難度高，又位於容易作答區塊的最後，將造成IRT 鑑別度值僅0.59，而非此條件的相同題其鑑別度值達1.24，兩者IRT 鑑別度相差高達0.65 的現象。

為避免IRT 難度值，非完全來自於內容，而是受試題擺放位置的影響，造成通過率、鑑別度、難度，以及猜測度等參數的偏誤現象，本研究建議組題之際宜（1）區塊內試題由易至難排序：避免影響IRT 難度值；（2）區塊內與區塊間的難度值由易至難排序：避免因前面區塊過難，簡單試題置後，無時間作答，猜測代之，影響IRT 鑑別度值及CTT 通過率；（3）正確答案位置的序號分配應大致相等：避免某一特定選項序號過於集中，影響猜測度；（4）使用0-1 線性規劃方法組題：力求隨機選題，組題客觀；（5）分析預試與正式施測共同試題特質間之相關：以為瞭解組題參考用之預試試題特質資料之可信度。

關鍵字：試題參數、試題位置、組題

A Study on the Item Position How to Influences the Item Parameter Estimation

Yi-Jen Lin
Assistant Research Fellow, National Academy for Educational Research
jen@mail.naer.edu.tw

Abstract

This study aimed to explore the item position how to influence the item characteristic based on classical test theory (CTT) and item response theory (IRT). This study utilizes data from the 6th grade Taiwan Assessment of Student Achievement (TASA) 2006 and TASA2007 mathematics Assessment. Just because the TASA2006 equated with partially balanced incomplete block (PBIB) and the TASA2007 with non-equivalent groups with anchor test (NEAT); moreover the TASA2007 put all the anchor items in front of test. Results from this study indicate that the item position significantly affects the three parameter of IRT and the passing rate of CTT.

Keywords: item parameter estimation, item position, item assembling

壹、前言

一、研究背景

《教育基本法》第九條第一項第六款明定透過評量學生在各學習領域（科目）的表現，以評鑑學生的學習狀況，為中央政府之教育權限之一。基於我國長期缺乏量化指標和標準化測量工具以檢視學生學習成就的表現及其差異，因此建置「臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement，TASA）」，以確實瞭解課程實施的成效。TASA 的2005 年及2006 年試題由國家教育研究院委由國立台南大學負責試題研發、預試及正式施測，2007 年試題則由國家教育研究院主導與負責。

TASA 為能有效涵蓋所有測驗目標，以及準確比較受試者之間的表現，分配受試者接受部分試題，並透過等化設計，建立出可互相比較的共同量尺（國家教育研究院籌備處，2010 年2 月1 日）。林宜臻（2010）執行「2006-2007 年台灣學生學習成就評量資料庫（TASA）數學領域小六評量架構與試題分析之研究」之際，發現TASA2006 與TASA2007 等化定錨用的相同試題中，某些相同試題IRT 鑑別度值相差高達0.65，難度值也相差高達0.76 的現象。相對於古典測驗理論（classical test theory，CTT）的題目參數會受到受試者能力影響，試題反應理論（item response theory，IRT）的題目參數，理應具有估計不變性（余民寧，2009）。因此，激發研究者探討究竟什麼因素造成該一現象。國外研究發現試題擺放位置將影響單參的難度值，IRT 難度值非完全來自於內容，而是受到試題擺放位置的影響，造成參數的偏誤，少數則受疲勞因素，指出試題反應理論的題目參數不變性受到考驗（Adams & Wu, 2002; Davis & Ferdous, 2005; Eignor &Stocking, 1986; Haertel, 2004; Hohensinn, Kubinger, Reiff, Holocher-Ertl, Khorramdel &Frebort, 2008; Meyers, Miller &Way, 2009; Whitely & Dawis, 1976）。

由於TASA 2006 題本的等化設計採取PBIB 設計，而TASA 2007 採取NEAT 設計。採取NEAT 設計的TASA 2007 定錨區塊的16 題試題，正是TASA 2006 與TASA 2007的共同試題。又TASA 2007 為避免試題受順序的影響，各題本定錨試題的題數與施測順序皆相同，而且將定錨區塊置前，非定錨區塊置後，此相較於採取PBIB 設計的TASA2006，基於各試題區塊出現的次數相同，其共同試題未必將全數置前，讓研究者得以藉此分析定錨題是否因試題擺放位置影響試題參數。

二、研究目的

本研究旨在探討TASA2006與TASA2007小六數學試題的參數與組題之間的關係，以為提供組題者之參考。

貳、文獻探討

量化分析試題係利用試題測驗結果，難易度（difficulty）、鑑別度（discrimination）、猜測度（pseudo-chance）、選項分析（option analysis），以及選項誘答力分析（distraction analysis），以說明個別試題的特性。本節將探討古典測驗理論（classical test theory, CTT）與試題反應理論（item response theory, IRT），及其應用於測驗編製的優缺點；並探討題本等化設計與試題位置如何影響試題特性等。

一、 測驗理論與編製

(一) 測驗理論

1. 古典測驗理論

「古典測驗理論」又稱真分數理論，是以[實得分數=真實分數+誤差分數（X=T+E）]為其架構，假設個人在測驗上的實得分數（observed score）是由真實分數（true score）和誤差分數（error score）兩部份組成，該線性模式應用時，須滿足（1）實得分數的期望值等於真實分數；以及（2）真實分數與誤差分數互為獨立的基本假設。據此假設，古典測驗理論衍生難易度、鑑別度等試題分析時的重要指標。

（1）難度

難度指試題難易程度的指數，以答對某題的人數佔總人數的百分比表示，或以高分組（全體受試者分數前27%）與低分組（全體受試者分數後27%）答對該試題百分比總和的平均，代表試題難易度。其值介於0 到1 之間，數值愈小表示試題愈困難，反之則愈簡單。學者主張以0.4到0.8 之間的難度值範圍，作為個別試題難度的挑選標準，而整份測驗的平均難度值以接近0.5為佳（郭生玉，2004）。

（2）鑑別度

鑑別度指試題能否反應不同能力學生答題的差異，以高分組答對率減低分組答對率的相差百分比代表鑑別度，其值介於 -1 到1 之間，數值愈接近1表示鑑別度愈高。鑑別度判斷的標準，如表 1（郭生玉，2004）。如表1所示，鑑別度指數0.4以上的試題屬於非常優良；0.3至0.39屬於優良；0.2至0.29屬於尚可。鑑別度最低標準至少要0.25，低於此標準者，即可視為鑑別度不佳或品質不良的試題，0.19以下的劣質試題，建議大幅度修改或刪除。

表1 CTT鑑別度評鑑標準表

鑑別度	試題評鑑
0.40 以上	非常優良
0.30~0.39	優良，但需小幅度修改
0.20~0.29	尚可，需部分修改
0.19 以下	劣，需要大幅度修改或刪除

2. 試題反應理論

「試題反應理論」以函數表示受試者能力與試題難易度、鑑別度及猜測度等參數間的關係。試題反應理論的基本假設為：（1）單向度：測驗中的所有題目主要都是測量相同的某一項特質，或是受試者在測驗題目上的答題反應主要是受到單一項特質所影響；（2）局部獨立性：相同能力水準的受試者，在各個題目上的答對機率是互相獨立。依參數的多寡，可分單參、雙參與三參等（王寶墉，1995；余民寧，2009）。

（1）Rasch單參模式

$$P_{ij} = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}$$

θj為考生 j 的能力；

bi是試題 i 的難度；

e代表以底為2.718的指數；

Pij是受測者答對某個題目的機率

單參模式只有難易度參數，受試者時答對第i題的機率Pi（θ）是0.5時，表示第i題的試題難易度參數與受試者能力值θ相等；若能力值小於試題難易度，受試者答對第i題Pi（θ）低於0.5；反之，高能力的受試者方能答對Pi（θ）高於0.5的題目。

（2）二參數模式

$$P_{ij} = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

雙參數模式中有難易度與鑑別度兩個參數。鑑別度通常以a表示，指不同能力受試者答題反應的差異。

（3）三參數模式

$$P_{ij} = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

三參數模式，則除了難易度與鑑別度兩個參數，還有猜測度參數c。

一般而言，鑑別度值a介於0~2之間為多，而以.8~1.25之間最為有效；難度b大部分介於 -3~3之間，值愈高難度愈難；猜測度c則宜為$0 \leqq C < \frac{1}{\text{選項數}}$（王寶墉，1995）。

相對於古典測驗理論，試題反應理論應具有：（1）能力估計不變性；（2）具有題目參數估計不變性，以及（3）測量精準度較合理；（4）應用層面較廣等優點（余民寧，2009）。

(二) 測驗編製

古典測驗理論（CTT）的測驗編製主要以試題的內容和特徵（如：難度、鑑別度等）為選擇題型的組題依據（余民寧，2002，頁234-236）。古典測驗理論的測驗編製方法，由於：（1）估計值是樣本依賴，樣本能否代表測量的母群體，成為試題選擇成功與否的主要因素；（2）由於學生生理與心理的成熟與成長，期初所建立的測驗試題無法適當地應用到期末的測驗情境中；（3）不同族群受試者下所建立的試題指標，彼此間無法比較，造成從題庫中所建立起來的任何測驗，無法適用於某一特定的群體等試題特徵本身不具有不變性之缺失。相對於此，以試題反應理論為架構的測驗編製，由於建立在同一量尺的基礎上，可以在任何能力水準上，挑選出最能精確測量到該能力範圍，且滿足對測驗所需的訊息總量最有貢獻的試題，方便測驗的編製與測驗結果的解釋。然隨機方式所挑選出的測驗試題所組成的測驗，造成分類錯誤的機率幾乎是依最理想方式的兩倍，而且高估高鑑別度α值，會造成訊息函數偏差，宜使用大樣本，以確保試題參數估計值的正確性與穩定性（余民寧，2009，頁255-262）。

由於試題反應理論的測驗編製，著眼參數估計值而忽略試題內容，導致缺乏內容效度，而線性規劃（linear programming）可以兼顧兩者。我國的國中基本學力測驗使用0-1線性規劃方法來進行組題，依據題目的各種屬性（例如：題目的類型、認知層次、所屬的章節範圍…等），以及各項統計上的指標（例如：題目的難度、鑑別度、訊息量…等），設計出線性方程組的限制條件，再配合線性規劃求取目標函數的極大（小）值進行組題。基測採取該一模式組題較隨機選題或人工組題客觀（余民寧，2009，頁259；宋曜廷、許福元、曾芬蘭、蔣莉蘋、孫維民，2007，頁38）。

二、題本等化設計與試題位置

(一) 題本等化設計

測驗等化（test equating）設計係將X 測驗之得分轉換成Y 測驗之量尺分數，使兩測驗得以於同一量尺進行比較。可將欲銜接的兩份測驗，利用（1）單一組設計（Single-group design）：給予同一組考生施測；（2）相等組設計(Equivalent-group design)：給予隨機選出相似但不完全相同的兩組考生施測；（3）定錨測驗設計（Anchor-test design）：每組考生另接受一部份的共同試題；（4）共同考生設計（Common-person design）：部份考生重覆接受這兩份測驗等方式蒐集等化資料。等化必須滿足對稱性（symmetry property）、相等性（equity properties）、團體不變性（group invariance property）及測驗必須是單向度（unidimensionality of the tests）等性質，等化才能進行。而將所有測驗題本均內含定錨試題，再根據受試者於定錨試題之作答反應作為連結，利用線性轉換的銜接方式，將轉換所需的常數值加以量化、估算，再調整不同測驗之間的差異的定錨測驗設計較為常用，也較可行。若兩組考生於定錨試題的能力分配上，具有高度的相似性，而且定錨試題的數量約佔測驗試題數的20%至25%間，並能代表兩份銜接測驗的內容，效果將會最好（余民寧，2009，頁217-238）。

TASA為比較（1）同年級不同題本；（2）不同年級不同測驗；（3）同年級不同年度測驗等之不同年級與不同年度受試學生之變化，藉助測驗間連結的垂直與水平等化，建立共同量尺，以為比較不同年級與不同年度受試學生之變化。為滿足涵蓋不同的認知層次與評量內容的測驗目標，而且所有評量內容都能施測，並考量受試者同時接受大量試題的測試，會造成受試者精神上的虛耗，導致發生漏答、誤答或亂答情形，影響施測結果的準確度，TASA利用等化設計編製測驗題本。等化設計有PBIB設計、BIB設計、NEAT設計等方式，「平衡不完全區塊（balanced incomplete block, BIB）」設計與「部份平衡不完全區塊（partially balanced incomplete block, PBIB）設計」中之「不完全」意指任一題本無法涵蓋所有試題，此二種等化設計，首先將試題分成若干試題區塊，使用螺旋排列方式，配置試題區塊於題本中，而區塊間與區塊內的試題皆不重複，受試者只需接受若干試題區塊的試題，且不同受試者可能接受部分相同、完全相同、或完全不同的試題區塊。PBIB設計與BIB設計的差別在於：BIB設計的每一試題區塊在所有題本中出現的次數，以及成對試題區塊於題本中的出現次數相等，呈現「平衡」狀態；PBIB設計僅各試題區塊出現次數相等，而成對試題區塊出現次數未必相等，係「部份平衡」。而「定錨不等組（non-equivalent groups with anchor test, NEAT）」設計，則於不同測驗之中內含定錨試題，以為不同測驗間之連結。定錨試題的品質將會影響等化估計的效果，此外，僅須假設受試群體隨機抽取，不必假設受試群體彼此之間有相同的能力值。但為了避免試題順序的影響，不同測驗定錨試題的題數與施測順序須相同，而且定錨試題的內容和難度與各分測驗的測驗內容與難度也須相似（余民寧，2009，頁217-238；國家教育研究院籌備處，2010年8月15日）。

(二) 試題位置與難度值

研究證明試題的擺放位置會影響試題難度：Whitely 與 Dawis（1976）發現15題核心試題，其中6題因位置明顯影響難度的估計值；Eignor與Stocking (1986) 發現學術評估測驗（Scholastic Assessment Test, SAT）的預試與正式施測間的試題難度值有所差異；國際學生評估計畫（Programme for International Student Assessment, PISA）的2000年技術報告，指出題本間原彼此的難度值相當一致且穩定，然而這些難度值在不同國家時，卻呈現或易或難的現象，甚至題本間的差異高達0.52（Adams & Wu, 2002, p.157）；Haertel(2004)發現相同題的參數值因放置於1年級或2年級的測試中，而有所不同；Davis 與Ferdous（2005）發現無論是在3年級閱讀測驗或是五年級的數學與閱讀測驗，試題位置影響難度值；Hohensinn, Kubinger, Reiff, Holocher-Ertl, Khorramdel與Frebort（2008）發現利用等化設計編製測驗題本的大型測驗，其定錨題的IRT難度值，非完全來自於內容，而是受到試題擺放位置的影響，造成參數的偏誤（bias），少數則受疲勞因素（fatigue effect）影響；Meyers, Miller 與Way（2009, p.41）指出大型的州測驗的預試與正式施測的相同試題，因試題位置影響單參的難度值。這些研究都指出相同試題，因擺放於題本的前或後位置等會影響難度值，指出試題反應理論的題目參數不變性受到考驗。

參、研究方法

一、研究範疇

本研究以TASA2006及2007年六年級學童抽樣施測的答題結果為分析對象。如表2所示，TASA 2006之題本等化設計採取PBIB設計，而TASA 2007則採取NEAT設計。TASA 2007與TASA 2006的共同試題，則是TASA 2007定錨區塊的16題試題。採取NEAT設計的TASA 2007為避免受試題順序影響，各題本的定錨試題的題數與施測順序皆相同，而且將定錨區塊置前，非定錨區塊置後。此相較於採取PBIB設計的TASA 2006，共同題未必全數置前，讓研究者得以藉此觀察試題位置是否影響試題參數。

表2 TASA數學領域2006與2007小六之題本等化設計

年次	題本數	每題本區塊數	每區塊題數		總區塊數	等化設計
2006	13	3	8		13	PBIB設計
2007	10	定錨區塊與非定錨區塊各一	定錨區塊 16	非定錨區塊 8	11	NEAT設計

二、研究方法

本研究採用古典測驗理論與試題反應理論三參數模式，以BILOG-MG（Zimowski, Muraki, Mislevy, & Bock, 1996）、TEST2（余民寧，2002）等針對試題難度、鑑別度與誘答力等進行TASA2006及TASA2007小六數學評量共同試題的參數特性分析。

肆、研究結果與討論

一、試題位置與鑑別度值

由表3可以發現：TASA2007定錨題16題中有14題（約佔88%）的IRT鑑別度值高於TASA2006。TASA2007定錨區塊的16題，全數置放於各題本之前16題，之後才放置8題非定錨區塊的試題，所以較無時間不足未作答的現象，所以鑑別度相對高。

二、試題位置與通過率

由表3可以發現：TASA2007定錨題16題中有13題（約佔81%）的通過率高於TASA2006，而通過率相差6%以上者達27%，有否時間充分作答成為關鍵因素。

表3 TASA2006及TASA2007小六數學評量共同試題量化分析一覽表

區塊/題號		IRT試題分析						古典理論CTT選項分析					
		a 鑑別度		b 難度		c 猜測度		通過率		難 度		鑑別度	
2006	**2007**	**2006**	**2007**	**2006**	**2007**	**2006**	**2007**	**2006**	**2007**	**2006**	**2007**	**2006**	**2007**
8/4	1/01	2.025	**2.125**	0.689	0.541	0.200	0.200	43.478	**50.140**	0.500	0.551	0.822	0.760
4/4	1/02	2.315	2.017	1.411	**1.416**	0.147	**0.160**	24.063	**25.758**	0.293	0.351	0.497	0.490
2/8	1/03	1.670	**1.859**	0.520	0.480	0.213	**0.250**	***48.486***	***54.851***	0.527	0.568	0.768	0.700
1/1	1/04	1.237	**1.791**	-1.287	**-1.085**	0.207	0.200	85.032	85.015	0.763	0.779	0.465	0.426
11/3	1/05	1.463	**1.663**	-0.183	-0.251	0.142	0.140	***62.894***	***69.093***	0.576	0.612	0.823	0.755
6/7	1/06	1.239	**1.613**	-0.334	-0.404	0.217	0.210	***69.109***	***76.738***	0.625	0.687	0.726	0.602
13/8	1/07	1.143	**1.613**	***1.717***	***0.958***	0.274	0.250	34.852	33.609	0.375	**0.370**	0.483	0.365
11/4	1/08	1.458	**1.584**	1.181	1.100	0.122	0.130	27.923	**32.051**	0.356	0.404	0.588	0.537
8/5	1/09	1.431	**1.505**	0.883	0.897	0.203	0.240	40.562	**43.433**	0.470	0.508	0.716	**0.735**
10/6	1/10	1.057	**1.436**	0.304	0.216	0.155	0.190	***52.396***	***59.696***	0.512	0.569	0.772	0.742
5/1	1/11	1.158	**1.407**	-0.025	-0.160	0.216	0.190	62.975	**63.104**	0.583	0.595	0.757	0.713
13/5	1/12	1.360	**1.397**	0.443	0.531	0.151	0.220	46.133	**48.131**	0.477	0.514	0.699	**0.733**
6/5	1/13	0.998	**1.328**	-0.469	-0.445	0.131	0.140	67.580	**68.691**	0.615	0.619	0.719	0.684
2/2	1/14	1.362	1.269	0.754	0.674	0.197	0.190	42.645	**45.551**	0.482	0.499	0.743	0.668
2/1	1/15	1.023	**1.240**	1.404	1.426	0.277	0.310	39.724	**41.741**	0.445	0.452	0.553	0.399
11/8	1/16	***0.593***	***1.240***	-1.705	**-1.095**	0.203	**0.220**	83.396	80.097	0.777	**0.747**	0.421	0.405

*斜體粗體含底線的數值表示差距大者；粗體數值表2007年較2006年有提升現象

*由於TASA2006受委託單位未提供原始數據，而只提供統計後的數據，因此無法檢定TASA2006與TASA2007定錨題的三參數之差異值是否達到顯著差異。

三、試題位置與難度值

由上表3可以發現：（1）TASA2007定錨題16題中有5題（約佔31%）的難度值高於TASA2006；（2）TASA2007區塊1的第7題與TASA2006區塊13的第8題，雖屬相同題目，於TASA2007難度值為0.96，於TASA2006難度值高達1.72，兩者難度數值相差高達0.76。

由表4可以發現：TASA2006區塊13第8題的前7題，適合高成就學生的試題3題，適合中上成就學生的試題2題，適合中成就學生的試題1題，適合中低成就學生只有1題，前7題的平均難度值為0.48。相較於此，TASA2007的前6題中，適合高成就學生的試題1題，適合中上成就學生2題，適合中等成就學生2題，適合中低成就學生1題，前6題平均難度值為0.12。此外，猜測度c值理應小於 $\frac{1}{\text{選項數}}$（王寶墉，1995），但該相同題在TASA2006的猜測度c值0.26，在TASA2007猜測度c值0.27，兩者猜測度值都偏高，也造成難度值不穩的現象。

表4 TASA2006及2007相同題前試題難度值之比較

TASA2006			TASA 2007		
區塊/題號	難度（b）	平均	區塊/題號	難度（b）	平均
13/1	**-0.051**		1/1	0.541	
13/2	1.853		1/2	0.060	
13/3	-1.735		1/3	0.480	0.12
13/4	0.060	0.48	1/4	-1.085	
13/5	0.443		1/5	**-0.251**	
13/6	1.771		1/6	**-0.404**	
13/7	1.012		1/7	0.958	
13/8	1.717				

*雙底線數值，表示適合高成就學生；單底線數值，表示適合中上成就學生；粗體字表示適合中等成就學生

四、區塊位置與鑑別度數

TASA 2006數學領域小六採取PBIB 之題本等化設計，如表5所示，將試題分成13 個試題區塊（M1~M13），試題在區塊間與區塊內皆不重複，13 個試題區塊編製成13個題本（S1~S13），每個題本包含3 個試題區塊。TASA 2007採取NEAT設計，將試題分成11個試題區塊（M1~M11），編製成10 個題本（S1~S10）。每個題本包括1個16題組成的定錨試題區塊（M1），另有10個各8題組成的非定錨試題區塊（M2~M11），每一個題本24題，均包含一個定錨試題區塊和一個非定錨試題區塊。TASA2007相同試題隸屬於定錨區塊，定錨區塊全數放置於各題本之前。相較於此，相同試題在TASA2006未必置放於前。

表5 2006年及2007年數學科各年段試題區塊對應表

2006年				2007年		
題本序號	區塊位置			題本序號	區塊位置	
	K1	K2	K3		K1	K2
S1	M1	M2	M5	S1	M1	M2
S2	M2	M3	M6	S2	M1	M3
S3	M3	M4	M7	S3	M1	M4
S4	M4	M5	M8	S4	M1	M5
S5	M5	M6	M9	S5	M1	M6
S6	M6	M7	M10	S6	M1	M7
S7	M7	M8	M11	S7	M1	M8
S8	M8	M9	M12	S8	M1	M9
S9	M9	M10	M13	S9	M1	M10
S10	M10	M11	M1	S10	M1	M11
S11	M11	M12	M2			
S12	M12	M13	M3			
S13	M13	M1	M4			

※S：題本 M：區塊 K：區塊位置

由表3可以得知：TASA2006區塊11的第8題與TASA2007 區塊1的第16題，屬於定錨的相同試題，前者鑑別度值0.59，而後者鑑別度值高達1.24，兩者鑑別度數值相差高達0.65。如上表5及表6所示，TASA2006區塊11的第8題，是題本7的最後一題，TASA2006區塊11的IRT平均難度值只有-0.19，適合中低程度學生，是第4簡單的區塊；而位於該區塊的前一區塊8的平均難度值為0.63，卻是13區塊中第3難的區塊。若該試題的前一區塊難度高，又位於容易作答區塊的最後，會因時間不足無法作答，容易以猜測代之，造成鑑別度僅0.59；TASA2007定錨區塊的16題，全數置放於各題本之前16題，之後才放置8題非定錨區塊的試題，所以較無時間不足未作答的現象，所以鑑別度高達1.24。難度低的相同試題置放於題本最後，會因位置順序的不同，造成鑑別度數值相差高達0.65。

表6 TASA2006各區塊IRT難度平均值

區塊	-3~-2	-2~-1	-1~0	0~1	1~2	2~3	難度	難度	
	低成就	中低成就	中等成就	中上成就	高成就	極優	平均	易1	難1
1	0（0%）	4（50%）	1（12.5%）	2（25%）	0（0%）	1（12.5%）	**-0.50345**	2	12
2	1（12.5%）	0（0%）	3（37.5%）	3（37.5%）	1（12.5%）	0（0%）	**-0.06166**	6	8
3	0（0%）	2（25%）	1（12.5%）	3（37.5%）	2（25%）	0（0%）	**0.24399**	9	5
4	0（0%）	0（0%）	1（12.5%）	5（62.5%）	2（25%）	0（0%）	**0.73267**	13	1
5	1（12.5%）	1（12.5%）	2（25%）	3（37.5%）	1（12.5%）	0（0%）	**-0.19027**	3	11
6	0（0%）	0（0%）	5（62.5%）	3（37.5%）	0（0%）	0（0%）	**-0.00230**	7	7
7	1（12.5%）	4（50%）	3（37.5%）	0（0%）	0（0%）	0（0%）	**-1.14566**	1	13
8	0（0%）	0（0%）	1（12.5%）	6（75%）	1（12.5%）	0（0%）	**0.63048**	11	3
9	0（0%）	0（0%）	3（37.5%）	4（50%）	1（12.5%）	0（0%）	**0.12061**	8	6
10	0（0%）	0（0%）	2（25%）	3（37.5%）	2（25%）	1（12.5%）	**0.54867**	10	4
11	0（0%）	1（12.5%）	5（62.5%）	1（12.5%）	1（12.5%）	0（0%）	**-0.18810**	4	10
12	1（14.3%）	0（0%）	2（28.6%）	3（42.9%）	1（14.3%）	0（0%）	**-0.16058**	5	9
13	0（0%）	1（12.5%）	1（12.5%）	2（25%）	4（50%）	0（0%）	**0.63365**	12	2

*粗體數值，表適合中等程度學生之區塊；單底線粗體數值，表難度適合中上以上程度之區塊

五、正確答案選項位置

正確選項的號碼分配機率應大致相等，避免學生不知如何作答時，傾向選擇某一特定選項。

由表7可以得知：2006年M3、M12及M13區塊的正確答案位置，在有些選項序號一題都沒有，S12題本的正確答案位置都同時集中序號③高達50％。而題本S3、S8、S9、S12有些選項序號只佔13％。

表7 2006年區塊及題本正確答案位置

區塊	選項序號 題數（共8題）				題本	區塊組合			選項序號 題數（共24題）%			
	①	②	③	④		K1	K2	K3	①	②	③	④
M1	2	2	1	3	**S1**	M1	M2	M5	5（20.83%）	8（33.33%）	5（20.83%）	6（25.00%）
M2	2	3	2	1	**S2**	M1	M3	M6	7（29.17%）	4（16.67%）	8（33.33%）	5（20.83%）
M3	3	0	5	0	**S3**	M1	M4	M7	7（29.17%）	6（25.00%）	8（33.33%）	3（12.50%）
M4	2	4	1	1	**S4**	M1	M5	M8	5（20.83%）	8（33.33%）	5（20.83%）	6（25.00%）
M5	1	3	2	2	**S5**	M1	M6	M9	4（16.67%）	6（25.00%）	6（25.00%）	8（33.33%）
M6	2	1	1	4	**S6**	M1	M7	M10	6（25.00%）	5（20.83%）	6（25.00%）	7（29.17%）
M7	2	2	2	2	**S7**	M1	M8	M11	6（25.00%）	6（25.00%）	6（25.00%）	6（25.00%）
M8	2	1	2	3	**S8**	M1	M9	M12	**3（13.04%）**	5（21.74%）	8（34.78%）	7（30.43%）
M9	1	2	3	2	**S9**	M1	M10	M13	**3（12.50%）**	8（33.33%）	**9（37.50%）**	4（16.67%）
M10	2	2	3	1	**S10**	M1	M11	M1	6（25.00%）	7（29.17%）	6（25.00%）	5（20.83%）
M11	2	3	2	1	**S11**	M1	M2	M2	4（17.39%）	8（34.78%）	7（30.43%）	4（17.39%）
M12	**0**	2	4	2	**S12**	M1	M3	M3	**3（13.04%）**	6（26.09%）	**12（50.00%）**	**3（13.04%）**
M13	**0**	4	3	1	**S13**	M1	M4	M4	4（16.67%）	**10（41.67%）**	5（20.83%）	5（20.83%）

※單底線粗體數值表較不合理者

由表8可以得知，2007年M5、M6、M7、M9及M10區塊的正確答案位置，在有些選項序號一題都沒有。而M7與M9區塊的正確答案位置集中於選項序號③，分別佔8題中6題與5題。題本S1及S5的正確答案位置集中於同一序號達38%，而題本S5有的選項序號只佔8%。

表8 2007年區塊及題本正確答案位置

區塊	選項序號 題數（共8題）				題本	區塊組合		選項序號 題數（共24題）%			
	①	②	③	④		K1	K2	①	②	③	④
M1	6	3	2	5	**S1**	M1	M2	**9（37.50%）**	4（16.67%）	3（12.50%）	8（33.33%）
M2	3	1	1	3	**S2**	M1	M3	8（33.33%）	4（16.67%）	5（20.83%）	7（29.17%）
M3	2	**1**	3	**2**	**S3**	M1	M4	7（29.17%）	4（16.67%）	5（20.83%）	8（33.33%）
M4	1	1	3	3	**S4**	M1	M5	6（25.00%）	7（29.17%）	4（16.67%）	7（29.17%）
M5	**0**	4	2	2	**S5**	M1	M6	**9（37.50%）**	6（25.00%）	**2（8.33%）**	7（29.17%）
M6	3	3	**0**	2	**S6**	M1	M7	7（29.17%）	4（16.67%）	8（33.33%）	5（20.83%）
M7	1	1	6	**0**	**S7**	M1	M8	7（29.17%）	5（20.83%）	5（20.83%）	7（29.17%）
M8	1	2	3	2	**S8**	M1	M9	7（29.17%）	5（20.83%）	7（29.17%）	5（20.83%）
M9	1	2	5	**0**	**S9**	M1	M10	6（25.00%）	6（25.00%）	4（16.67%）	8（33.33%）
M10	**0**	3	2	3	**S10**	M1	M11	7（29.17%）	7（29.17%）	4（16.67%）	6（25.00%）
M11	1	4	2	1							

※單底線粗體數值表較不合理者

伍、結論與建議

本研究旨在探討TASA2006與TASA2007小六數學試題參數與組題之間的關係，獲得如下之結論，並提出建議。

一、結論

(一) 試題置前將提高鑑別度值與通過率及降低難度值

研究發現：定錨題置前的TASA2007相同試題的88%，其IRT鑑別度值高於TASA2006；TASA2007相同試題的81%，其CTT通過率高於TASA2006；TASA2007相同試題的31%其IRT難度值低於TASA2006。

(二) 低難度試題置於題本最後且前區塊難度高將大幅降低鑑別度值

研究發現：若試題的前一區塊難度高，又位於容易作答區塊的最後，會因時間不足無法作答，容易以猜測代之，造成該相同試題於置前的TASA2007其鑑別度值為1.24，置後的TASA2006鑑別度值僅0.59，兩者相差高達0.65。難度低試題若置於簡單題本的最後，若前區塊難度又高，將大幅降低IRT鑑別度值。

(三) 相同試題的前試題平均難度高將提高試題難度值

研究發現：共同試題前7題的平均難度值為0.48的TASA2006其難度值為1.72，相較於此，共同試題前6題的平均難度值為0.12的TASA2007其難度值為0.96，兩者IRT難度值相差高達0.76。前面試題的平均難度值高，將提高IRT難度值。

(四) 正確答案位置序號分配不均

TASA2006及TASA2007的試題區塊或題本，尚存有些正確答案位置過於集中於同一序號，或有些選項序號一題都沒有，或有些選項序號的試題甚少的現象。為避免學生不知如何作答時，傾向選擇某一特定選項序號影響猜測參數，所以正確答案的位置序號分配應大致相等。

二、 建議

本研究發現：（1）試題置前的定錨試題的八成以上，其鑑別度值與CTT 通過率高於置後的定錨試題；（2）近七成置後定錨試題的IRT 難度值高於置前的定錨試題；（3）相同定錨題前的試題平均難度值為0.12 的相同試題其難度值為0.96，相同定錨題前的試題平均難度值為0.48 的相同試題其難度值為1.72，倘若定錨題前的試題平均難度值差距大時，即便是相同的題目，也會造成難度值相差高達0.76 的現象；（4）若相同題的前一區塊難度高，又位於容易作答區塊的最後，將造成IRT 鑑別度值僅0.59，而非此條件的相同題其鑑別度值達1.24，兩者IRT 鑑別度相差高達0.65 的現象。

為避免IRT 難度值，非完全來自於內容，而是受試題擺放位置的影響，造成通過率、鑑別度、難度，以及猜測度等參數的偏誤現象，本研究建議組題之際宜（1）區塊內試題由易至難排序：避免影響IRT 難度值；（2）區塊內與區塊間的難度值由易至難排序：避免因前面區塊過難，簡單試題置後，無時間作答，猜測代之，影響IRT 鑑別度值及CTT 通過率；（3）正確答案位置的序號分配應大致相等：避免某一特定選項序號過於集中，影響猜測度；（4）使用0-1 線性規劃方法組題：力求隨機

選題，組題客觀；（5）分析預試與正式施測共同試題特質間之相關：以為瞭解組題參考用之預試試題特質資料之可信度；（6）受委託單位宜提供原始數據：委託單位若只提供統計後的數據，無法進行數據清理，尤其標準誤差不合理時，將無法利用統計檢定方法，檢定TASA2006 與TASA2007 定錨題的三參數之差異值是否達到顯著差異。

參考文獻

王寶墉（1995）。**現代測驗理論**。臺北：心理。

余民寧（2002）。**教育測驗與評量-成就測驗與教學評量（第二版）**。臺北：心理。

余民寧（2009）。**試題反應理論（IRT）及其應用**。臺北：心理。

宋曜廷、許福元、曾芬蘭、蔣莉蘋、孫維民（2007）。國民中學學生基本學力測驗的回顧與展望。**教育研究與發展期刊，3**（4），29-50。

林宜臻（2010）。**2006-2007年台灣學生學習成就評量資料庫（TASA）數學領域小六評量架構與試題分析之研究**。國家教育研究院籌備處自行研究計畫（編號NAER-97-24-B-1-02-03-1-08）。

國家教育研究院籌備處（2007）。**「臺灣學生學習成就評量資料庫」建置計劃－數學領域—（小四、小六、國二、高中二、高職二）96年命題研習手冊**。臺北：作者。

國家教育研究院籌備處（2010年2月1日）。臺灣學生學習成就評量資料庫試題等化**設計。「臺灣學生學習成就評量資料庫」電子報 第5 期**。取自http://tasa.naer.edu.tw/。

國家教育研究院籌備處（2010年8月15日）。BIB、PBIB 與NEAT 題本編排設計與等化介紹。**「臺灣學生學習成就評量資料庫」電子報 第12 期**。取自http://tasa.naer.edu.tw/。

教育基本法（2006年12月27日）。

郭生玉（2004）。**心理與教育測驗（修訂一版）**。臺北：精華。

Adams, R., & Wu, M. (2002). *PISA 2000 technical report.* Paris: Organisation for Economic Co-operation and Development.

Davis, J., & Ferdous, A.(2005). *Using item difficulty and item position to measure test fatigue.*Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.847&rep=rep1&type=pdf

Eignor, D. R., & Stocking, M. (1986). *An investigation of possible causes for the inadequacy of IRT pre-equating* (ETS RR-86-14). Princeton, NJ: Educational Testing Service.

Haertel, E. (2004). *The behavior of linking items in test equating.* CSE Report 630. LosAngeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California。

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*, 391-402.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Applied measurement in education, 22*(1), 38-60. doi: 10.1080/08957340802558342

Whitely, E., & Dawis, R. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36*, 329-337.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items (Version 3.3)[Computer software]. Chicago: Scientific Software International.

【致謝辭】

感恩審查委員的肯定、鼓勵與指正；感恩謝研究員進昌總是不厭其煩專業解答宜臻的疑惑；感恩曾建銘主任對於宜臻的協助，總是不遺餘力；感謝得力助手林妲君、張宛婷等的幫忙。

中等學校規模與學生數學成就之關係研究

張芳全

中等學校規模與學生數學成就之關係研究

張芳全
國立臺北教育大學教育經營與管理學系教授

摘要

各國的中等學校規模與學業成就之關係為何？本研究以20個參與TIMSS 2003國家的資料進行分析，研究中以二次式的迴歸方程式瞭解學校規模與學生學習成就之關係，以學校規模為投入變項，以學生學習成就為依變項。針對這20個國家(包括高度所得、開發中與亞洲四小龍)來瞭解學生學習成就與學校規模之關係是否有一致趨勢。研究發現，20個國家之中，學校規模與學習成就呈現四種關係，其中兩者之間具U型關係僅有比利時、匈牙利及印度，其餘的國家則否。臺灣則呈現線性的關係。依據結果，提供未來建議。

關鍵詞：學校規模、數學成就

Study on the Relationships between the School Size of Secondary Education and Student's Achievement

Fang-Chung Chang
Professor, Department of Educational Management,National Taipei University of Education
fcchang@tea.ntue.edu.tw

Abstract

Are the relationships between the school size of secondary education and student's achievement? Data from the Trend International Mathematics and Science Survey of 2003 (TIMSS 2003) were used to investigate the relationships between the school size and student's achievement. There were 20 countries in this study. It employed the multi-regression analysis to test the relationships between the school size and student's achievement. The school size was regarded as independent variable, and the student's achievement was regarded as dependent variable. We want to research the consistence on the relationships between the school size and student's achievement in 20 countries. The results showed that there were 4 types relationships between the school size and student's achievement, and only three countries (e.g. Belgium, Hungry, India) were inverse U-type relationships between the school size and student's achievement, and the other countries were not. Taiwan was linear-type relationships between the school size and student's achievement. From these findings, some suggestions were given.

Keywords: school size, mathematics achievement

壹、緒論

一、研究動機

學校規模與學生數學成就之關係頗為爭議，是否學校規模愈大，學生學習成就更好？或學校規模愈小，才讓學生學習成就表現比較好呢？是值得思考議題。然而，國內外在這方面研究相當少。過去在學校規模經營研究大多以學校所投入的經費多寡、學生單位成本或學校經營的總成本做為分析的依據（林文達，1977，1990；林淑貞，1979；吳炳銅，1994；郭添財，1991，1996；蓋浙生，1985；Edington & Martellaro, 1990；Thompson, 1994），較少重視一所學校規模，與學生學習成就之關聯的分析。其實，學校規模應與學生學習表現應有密切關聯。它可能是學校規模愈大，學校的師資素質較為齊全、教師平均任教科目較少，可以提供聯課活動及專業選修科目較多，因此可以提高學生的學習興趣及學習成效；相對的；學校規模小，學校教師的素質及對於學生教學的表現可能受到影響，學校規模小，教師任課科目相對較多，學生可以選擇的專業科目選修及課程較少，學生學習可能受到限制，因而學習表現可能受到影響。

本研究檢視國內自1980年之後，以學校規模與學習成就之關係的研究，僅有陶韻婷（2007）與張芳全（2010）的研究。這兩篇研究僅以臺灣為樣本進行分析，無法瞭解其他國家在這方面的表現情形，因此，這就令本研究嘗試以不同國家的資料來分析學校規模與學習成就之關聯性，本研究期待從其他國家在這兩者之間的關聯現象，歸納出學校規模與學習成就之關係，以提供未來研究及實務的參考。重要的是，本研究分析各國的學校規模與學習成就之關係，還有以下的研究動機：

首先，TIMSS 2003年的資料可以提供學校規模與學習成就之分析。國際教育成就調查委員會（The International Association for the Education Achievement, IEA）在1990年起推動進行的第三次國際數學與科學教育成就趨勢調查（The Trend International Mathematics and Science Survey, TIMSS, NCES, 2004）資料，這調查資料包括學生、教師及學校面向的資料，學校面向詢問了校長的學校發展情形，包括學校規模、學生就學人數及出缺席情形、學校教學資源及學生行為問題等。所以如能透過此資料來分析學校規模與數學成就之關係，可以瞭解學校規模與學生學業成就表現之重要性。IEA於1995年、1999年、2003年各進行一次成就調查，臺灣在1999、2003年與2007年都有參加。它對各國調查資料統一且較客觀，運用這筆資料分析此議題實為嘗試。

其次，雖然學習成就高低與學校規模宜考量國家教育制度及文化差異，但是學校規模大小是影響學生學習表現重要因素。然而各國中等學校的規模究竟與學生學習表現之關聯為何？很值得研究。2003年TIMSS調查近50個國家及地區八年級學生、校長及學校教師，因而若能以各國資料作為分析基礎，來瞭解學校規模與學習成就之關聯，可以獲得結論並提出具體建議。若以國民所得來區分，TIMSS2003調

查國家有高度經濟發展（2萬美元以上）、開發中國家（2萬元以下），究竟在不同經濟發展國家，學校規模與學習成就之關係為何？因此，本研究將選定8個先進國家與8個開發中國家來掌握它們之關係，究竟這些國家的學校規模愈大，是否學生學習成就愈好嗎？

第三，亞洲四小龍學習成就表現優異，過去以學校規模討論四小龍學生學習成就之關係研究仍缺乏。本研究運用TIMSS2003做為跨國分析，選定臺灣、南韓、香港及新加坡來分析是考量，四小龍國家學生學業成就高於歐美國家（Stigler, Lee & Stevenson, 1987），亞洲四小龍的學校規模與八年級學生的數學成就較高，在過去分析這些國家學生學習成就較高是受到學習抱負及家庭教育因素的影響很大（張芳全，2010a），然而究竟學習成就與學校規模之關聯為何？亞洲四小龍學生數學成就較高，除了受到重視升學之外，學校規模是否也是影響學習成就的因素之一呢？因此，本研究除了上述以高度國民所得與開發中國家（16國）之外，本研究還要掌握亞洲四小龍在這些變項之關係。

最後，余民寧（2006）曾試著企圖歸納整理出一個影響學習成就的統整性因素結構模型架構，他也陳述各種影響學習成就因素結構模式，以作為後續進行探索與建構理論模型的導引，他歸納出影響學生成就的五個模型，也就是：學生個人背景模型、家庭背景模型、教師背景模型、學校管理背景模型、政府教育政策因素模型等。本研究，就是要以學校經營管理模型，即運用20個國家在TIMSS 2003的學校規模來分析，它是否影響學生學業成就。

二、研究目的

本研究目的如下：從TIMSS2003的資料，來分析高度國民所得、開發中國家及亞洲四小龍等20個國家，在學校規模與學生學習成就之關係。針對結果，提供建議，供未來研究參考。

貳、文獻探討

一、規模經濟的意涵及其理論

(一)規模經濟的意涵

學校經營良窳可以從學生學習表現來掌握。良好學校氣氛、學校環境與經營方式，會讓學生學業成就有好表現，相對的，不良的學校經營，很可能會讓學生學習表現變得更差。然而，校長與教師在學校經營，與學校規模大小有關。學校規模大的學校所需要投入的人力及資源較多，教師的授課科目可以減少，教師可以發展專門的才能，也較容易有優質的教師。蓋浙生（1985）指出，學校規模過大，優點在於教師可以專業分工，教育資源相對於小型學校來得好，但是其限制在於師生的人際關係疏離，學生學業成就表現不一定比較好。相對的，學校規模過小，其優點在於師生關係較好，但是教師的專業分工較差，教師需要兼任較多的行政工作，但這

樣對於學生的學業成就也不一定比較好。上述可以看出，學校規模大小影響了學校經營之後的結果。

林文達（1990）指出，學校經營過程中應找出最適規模的學生人數，如此可以讓學校在經營的成本最低，但是學生的學習成果表現最好。林文達認為，最適規模的學校較適當，當學校規模逐次增加的學校有利於教師專業專業化，也能吸收優良的師資，同時每位教師平均任教科目較少，而每為學生平均的圖書數也比較多，這有益於學生的學習表現。相對的，如果學校規模太小，每位教師任教科目較多，還可能兼負行政工作，教師教學專業較無法施展，很多科目都要包辦教學，因而影響到學生學習。因此，學校的規模小或過大，對於學生的學業表現不一定好，相對的，學校規模適中才是學生成就表現最好的學校類型。Donna, Dennis與Shirley（2003）研究認為，降低學校區域規模與推動教育改革是同等重要。Paul（2000）認為，要提昇高度貧窮學生的學業成就，解決方法的重要方式之一是，從降低學校的規模，採用小班小校做起。這可以看出，學校的學生人數規模對於學生的學習成就的影響很重要。

郭添財（1991）研究指出，最適經營代表性成本曲線為L型曲線，即代表學校隨著規模愈大，單位學生成本下降；他研究發現，就教育規模經濟考量，臺灣省南部地區國民小學之「最佳經營規模值」，整體地區為每校501至800人；城市地區為801至1,600人；鄉鎮地區為501至800人；偏遠地區為201至350人；此外，在教育規模效應方面發現：臺灣省南部地區不同經營規模之國民小學學校組織結構、教師工作滿意、師生人際互動，以及學生學習安排具有明顯差異；且經營規模愈小，愈能得到有利的教育規模效應。

然而，學校規模大小是否會影響學生的學習成效，許多研究有不同的結果發現Chopin（2003）研究美國路易西安那公立學校的學校規模、社經地位與學業成就的等級組態發現，學校規模大小與學業成就之關係呈現U型曲線。同時，學校規模過大，會產生學校老師與學生互動疏離，在學生人數過多，學生可以使用的資源相對的減少，因而學生的學習表現將會受到不利影響。因此，學校的學生人數如果是在最適經營規模之下，學生學業表現會最好呢？是本研究所要分析的。

綜合上述，學校經營規模常以經濟來衡量，即學校投入經費、每位學生的單位學生成本，而在產出方面以教師工作滿意度、師生人際互動與學生學習安排，然而學校規模經濟還可能與學生的學習效果有關，也就是在學校規模經營下，學生的學習成果及其表現情形。學校規模過大或過小對於學生的學習成果都可能有不良的影響，因此本研究在分析學校規模是否與學業成就之間呈現了「U」字型的關係。

(二)規模經濟的理論

學校的最適規模是由經濟學的規模經濟（economics of scale）而來，陳正倉、林惠玲、陳忠榮與鄭秀玲（2006，頁245-246）指出，一個產業或廠商的規模擴大（產量的增加），而其長期平均成本呈現遞減的現象時，稱為規模經濟；當場商的規模

擴大，而其長期平均成本呈現遞增的現象時，稱為規模不經濟；當場商的規模擴大，而其長期平均成本固定時，稱為固定規模經濟。而產生規模經濟的原因主要是由於機器及設備的不可分割性，以及規模報酬遞增所致。

如果將規模經濟理論應用於學校經營來說，一方面應掌握學校規模的內涵，一方面應掌握規模經濟的條件。就前者來說，蓋浙生（1985，頁244-251）指出，學校規模有三項重要的內涵，一是學校資源運用的整體性與不可分割性，係指學校興辦必需同時投入資源與運用，無論規模大小，都應有必要的設施；而不可分割是指學校在經營過程中，資源運用至少是一個單位，不可以因為需求不及一個單位，而將其分開使用。二是學校人力的分工與專業化，學校規模小，教師編製有限，需要任教科目較多，專長不易發揮；而在行政方面，規模小易有負擔過重的情形。三是對學生多樣性的適應，它是指學校為適應不同的學生應有更多的課程提供學生選修，尤其規模擴大之後，應提供更多類型及更高成本的課程給學生學習。因此學校規模過大及過小對於學生學習都是不好的。

而教育規模經濟的形成是由於教育經營規模擴大，使得資源充份及適當的使用，同時在規模擴大之後，並沒有衍生不經濟的缺陷下才能成立（林文達，1990，頁191-193），換言之，規模經濟的條件必需要包括：1.資源充份的運用、2.資源獲得適當的使用、3.規模擴大並不衍生不經濟的缺陷。就第一項來說當學校規模擴大，整體教育資源減少，因而使得每一單位學生資源成本降低，如果因為每一單位學生資源成本下降，但是沒有減少學生應該獲得的教育功能，如知識學習、課外活動的參與以及教師的專業教學等，此時就是學校規模經濟的條件之一。就第二項來說，學校資源獲得適當的使用代表學校資源運用於所需要的領域之中，例如專用建築設備、教師及行政人員專才專用等，當學校規模較大，教師可以依其專長任教，學校規模小，教師所要任教的科目較多，就是無法專長專用的情形。第三項來說，學校規模如果持續擴大，可能造成了學校教育人員及師生的人際關係疏離與行政僵化，就前者來說，因為學校規模太大，行政體系需要階層劃分，相同層級的單位也因為規模較大，而需要有較多的單位存在，如此在行政溝通不良，容易產生人際關係疏離的情形。因為行政溝通困難，所以部門之間容易產生本位主義，因而行政人員願意服從的情形。

總之，學校規模太小，學校可以使用的資源有限，人力編製亦較不完整，學生學習課程無法多樣化；而規模擴大之後，學校增加資源、教師人力完整，可以提供更多的專業課程，而學生也可以獲得更多的選修課程，有助於學生的學習。然而，學校擴大到一定的規模之後，會因為學校規模過大，學校行政溝通困難，使得學生學習及教育效果減低。

二、國內外相關研究

(一)最適規模學校的研究

最適規模學校（optimal size school）研究主要集中在於學校投入的變項，如教育經費、每生單位成本、教師人數及土地空間與設備，來瞭解師生及學校營運的情形（產出），通常這些研究認為如果學校投入愈多的資源，對學生學習表現有正向的助益（Wendling & Cohen, 1981）；而這些研究並依相關的投入變項來估算出學校最適規模，找出學校投入與產出之學校最佳經營規模類型。林文達（1977）分析臺灣省及臺北市的國民中學最適經營規模發現，臺灣省為1,800名至2,200名，而臺北市則為2,200名至2,400名之間。林淑貞（1979）針對臺北市國民中學的經營規模分析發現，最適當的經營規模為4,006名學生。林嘉薇（2002）分析八十八、八十九和九十學年度臺北縣公立國民小學和產出、教育品質及生產要素價格相關的實際資料，臺北縣公立國民小學，學校經營代表性成本曲線為近似L型，學校經營存在規模經濟現象，其最適經濟規模不存在。江亞萍（1999）以學生人數、生師比、勞動及物料投入要素價格等變項，以問卷調查蒐集資料實證發現：台閩地區國民中學經營過程中存在規模經濟現象、國民中學最適代表性平均成本曲線為遞減狀態、國民中學學校規模至少在2,000人以上，方具規模經濟。吳炳銅（1994）採問卷調查法，對臺北縣195所公立國民小學進行普查，在學校規模的投入因素包括師生比、每生平均校地、每班平均學生數、教師平均年資、學士以上教師比例、合格教師比例、教師平均任教科目數、教職員工平均薪資、每生平均圖書冊數、每百生團隊數，單位學生平均經常成本為依變項發現，無論是全縣、偏遠地區、市型、鎮型或鄉型學校，經營規模之代表性成本曲線均為L型曲線模式，且在經營過程中均具有規模經濟現象存在，若以決定係數增量考量，U型曲線模式較合理；最適經營規模之範圍值分別為：全縣1,501-2,200人、偏遠地區101-150人、市型3,001-4,000人、鎮型1,501-2,200人、鄉型501-1,000人。

Riew（1966）分析了美國威斯康辛州的108高中的經營規模發現，最適當的學校經營規模為1,675名學生，這些學校的特性是每日學生的出席人數在160名至2,400名之間，其單位學生平均經常成本為405美元，而教師素質最好。Cohn（1968）研究美國愛華達州發現，學校最適規模經營的人數為1,500名至2,000名。Osburn（1970）分析密蘇里州高中的規模發現，最適當規模的學生人數為2,244名。Bowles和Bosworth（2002）研究美國的學校發現，如果學校每增加10%的學校規模，可以減少2%的每位學生單位成本，可見學校規模愈大，可以縮減學生成本支出。

(二)學校規模與學習成就

學校規模與學習成就之關係研究說明如下。陶韻婷（2007）研究採用TIMSS 2003資料庫進行分析，以皮爾森相關分析全台、城市及鄉村地區，學校規模、語言變項、學生的特質來瞭解是否在科學成就有差異，以迴歸分析學生科學成就企圖建立全台、城市及鄉村地區的模式發現，造成城市地區與鄉村地區之間的科學成就差異的相關變項是：學校規模、學生對科學評價。張芳全（2010b）以TIMSS 2007年的

資料，並採用多層次模型對學習成就進行分析發現，總體層次變項對數學成就的影響中，學校規模、學生家庭富裕比率、學校所在的城鄉、數學教學資源對各校平均數學成就有顯著影響，其中前三項對於學習成就具有正向關聯，代表學校規模大、學校學生家庭富裕比率愈高，以及愈在城市學校，該校學生數學成就愈高。

Fowler與Walberg （1991, p.191）歸納了許多文獻指出，國民小學階段中，學校規模愈小，對於學生學習成就、課外活動、學生滿意度及出席率有正向的顯著影響。Sande （1993）也分析美國伊利諾州的高中學生學習成就與學校規模發現，兩者有正向顯著的關係，但他並沒有分析究竟是否呈現倒U的關係。Chopin（2003）研究美國路易西安那公立學校的學校規模、社經地位與學業成就的等級組態發現，學校規模大小與學業成就之關係呈現U型曲線。簡言之，學校太小或太大都不是一個經營的合適標準。

Gentry（2000）對喬治亞公立高中的學校規模與學業成就之關係研究指出，較大學校的學生學業成就，高於在較小的學校學生，學校規模較大的學生在認知學習比較有利，但關於動作技能與情意未被進行調查，其建議若要說明學校規模較大，就是較佳效能的整體觀點前，應考慮到動作技能與情意向度調查。

Okpala（2002）分析美國國民小學的最適規模情形，其分析資料以1993-1994年、1994-1995年，以及1995-1996年的資料，在其研究中以學校發展的特性、教師特質、學校經費支出與學生的特性為自變項，而以學校品質為中介變項，以學習表現（數學及閱讀成就）為依變項，研究發現，1994-1995年的學校規模與數學及閱讀成就有正向顯著關聯，其他年度則否，而每生教育經費支出對於學習成果，在三個年度都有正向顯著影響。

Gardner, Ritblatt與Beatty （2000）研究指出，從學業成就、曠課、輟學率及父母參與因素，對中學學校規模進行分析，研究指出小型學校學生有較多課外參與、較佳滿意度及較低輟學率，就學業成就言，不一定學校規模愈大，學業成就愈好，較好的學校表現出較高的學業成就在「學術水平測驗考試」（Scholastic Assessment Tests, SAT）總分及口語與數學分數，與該研究原來假設相反，而在小規模學校的學生有較低曠課、低輟學率與高父母參與。

Melvin與Roy（2003）在小學規模對學生學業成就影響研究指出，先前對學校規模與學業成就之關係的研究發現也有矛盾現象，有些研究發現學校規模與學業成就呈現正相關；有些呈現負相關，也有些研究未將學生能力納入學校規模對學業成就影響的分析。兩人研究在於將學生能力與其它變項納入在學校規模與學業成就發現，學校規模與學業成就之間關係，並不是呈現線性關係，而是呈現不規則的現象。Donna, Dennis與 Shirley（2003）對學校區域規模與學生表現研究指出，降低在小學層級的區域、學校與班級大小重要性。Paul（2000）提出，要提昇貧窮學生學業成就應從降低學校規模，尤其應降低學校規模。

最後，還有一些研究結果發現，學校規模與學生學習並沒有顯著關聯。Michelson（1972）研究指出，學校規模並沒有顯著影響學習表現，而Brown和Saks（1975）也有相同的研究發現。Pritchard（1987）分析學業成就與學校效能概念，並分析學校規模對學習成就影響發現，並無法支持學校規模可以提高學生成就或對學生成就產生正面影響的發現。Lamdin（1995）分析美國的學區國民小學學生及學校為研究對象，依變項為加州成就測驗（California Achievement Test, CAT），包括閱讀及數學成就，而自變項納入學生的家庭背景、生師比等之後研究發現，學校規模與學生學習成就沒有顯著正向關係。

(三)學校規模與學生表現

然而，有一些研究分析學校規模與學生學習表現或感受情形。Francis（1992）以學生的感受進行分析，即以學生是否感到上學快樂，來分析國民小學經營規模與學習表現發現，國小規模愈小，學生感受到的快樂程度不一定高於規模較大的學校。Merritt（1993）分析美國康乃迪克州的學生發現，規模愈大的學校，學生的缺席率較高。Stevens與Peltier（1994）分析美國的國小小型學校，其學生參與聯課活動的情形比大型學校還要多。

綜合上述分析有幾項歸納。首先，學校的最適規模大小會因為學校所在的地區而有不同的規模，而這必需要考量所分析的變項，因為分析的結果變項不同，其學校規模也有不同。其次，學校規模的最適標準與學習成效不一定具有U形關係，也有可能是非線性關係。第三，以不同研究變項（如，從生師比、教育經費、學生的曠課與輟學率、學生感受）來分析學校的最適規模，其估算的學校人數有不同說法，研究發現較大及較小規模，各有其優劣。第四，雖然有幾篇是學校規模與學業成就之關係研究，但其關係也不明確，也就是說，學校規模與學業成就之間，並非線性觀點就能解釋，並不是規模較大的學校在認知學習成就，比規模較小的學校來得好，這說明，不能單一線性觀點詮釋。上述研究指出，學校規模與學習成就以U型曲線概念來代表，僅是一種方式，究竟兩者之關係為何？是本研究要分析的重點。

參、研究設計與實施

一、研究問題與研究方法

本研究的問題如下：
第一，高度國民所得的國家，其學校規模與學生學習成就之關係為何？
第二，亞洲四小龍的學校規模與學生學習成就之關係為何？
第三，開發中國家的學校規模與學生學習成就之關係為何？

本研究運用次級資料分析來掌此研究研究問題，本研究以TIMSS的資料，接著依據相關文獻，建構學校規模與學習成就之關係模式，接下來運用多元迴歸分析對資料處理，本研究的研究方法屬次級資料分析。

二、研究變項的定義

本研究納入學校規模及學習成就變項，研究內容也涉及幾個名詞，如「高度國民所得的國家」、「亞洲四小龍」、「開發中國家」等，茲將其意義說明如下。

學校規模是指一所學校的空間及人力與資源的大小，其中空間包括學校可以使用的面積，而人力則包括了學校的學生、教師、行政人員等人數多寡，而資源是指學校可以使用於提升學生學習效果的一切總稱，如教學資源、人力及財力等。而本研究的學校規模係指為各國的中學學校規模（以臺灣來說是國民中學，以各國來說是中等教育前段），它是指一所學校的總註冊學生人數，如果學校註冊學生人數愈多，代表學校規模愈大，反之則否，它以每位學生為單位（人）。在TIMSS 2003的校長問卷資料中可以獲得。

學生學習成就係指學生在學校的學習表現，包括學生認知學科的成就表現及其他非認知科目的學習表現，前者如學生在學校紙筆測驗的表現分數高低，而後者包括情意及寄能方面的表現。本研究的學習成就係以TIMSS 2003年各國參與調查的學生數學成就為主，該分數包括幾何、代數、資料處理等五個領域，而在本研究中的數學成就以數學總分為代表。TIMSS 2003對於數學成就有五個估計數，本研究以第一個估計數為主。因為TIMSS對於各國學生的數學成就估算，有其標準估計公式，雖然是在不同國家所進行的調查，但是各國的八年級學生可以進行跨國資料的分析。如果此項分數愈高，代表學生的學習成就愈好。

國民所得高低是衡量國家發展程度的指標之一。本研究的「高度國民所得的國家」係指以世界銀行（World Bank, 2004）的統計為主，如果該國在2003年的國民平均國民所得高於2萬美元以上的國家，而「開發中國家」也是以世界銀行（World Bank, 2004）的統計為主，在2003年該國的平均國民所得在2萬美元以下者。而「亞洲四小龍」則是指臺灣、南韓、香港及新加坡。

三、研究對象

本研究以20個國家做為分析的對象，其中有8個為高度國民所得國家，也是在TIMSS 2003的數學成就高於世界各國的平均值，同時這些國家也是國民所得較高者，它們是美國、比利時、挪威、日本、西班牙、澳洲、英格蘭、紐西蘭，而亞洲四小龍在TIMSS 2003的數學成就相當優異，即香港、新加坡、南韓及臺灣也在研究之列。最後，本研究納入8個開發中國家，在TIMSS 2003的數學成就傾向低於世界平均值者也納入分析，分別是埃及、匈牙利、馬來西亞、印度、突尼西亞、賽普勒斯、約旦、菲律賓。這些國家參與TIMSS 2003者，以他們的中等教育（國中）為樣本，而參與TIMSS 2003的學校，各國的學校數是不相同。

四、模型的設立

為了探究此問題，本研究以20個國家的學生數學成就與這些國家參與學校規模的學生人數進行分析。本研究以二次式的迴歸方程式，來瞭解學校規模與學生的學

業成就之關係。針對研究問題，以多元迴歸分析來檢定，迴歸分析在探討自變項及依變項之間的關係。本研究以二次式的迴歸分析，以學校規模為投入變項，學習成就為依變項，檢定模式如下：

$$學習成就 = a + \beta_1(學校規模) + \beta_2(學校規模)^2 + e$$

模式的依變項代表各個國家參與學校的學生學習成就之平均值；學校規模代表學校學生人數，後續表將以「規模」代表，$(學校規模)^2$代表學校學生人數平方，以「規模平方」代表。本研究透過對20個國家在學校規模與數學成就之關係，研究中將針對這20個國家找出兩者之關係，瞭解這些國家在學生學習成就與學校規模之關係是否有一致趨勢。

五、資料來源與限制

本研究研究的相關資料取自於TIMSS 2003年報告書（TIMSS 2003 user guide for the international database）（NCES, 2004）。本研究僅以數學成就為學業成就，未能更廣義的學習成就（如非認知成份）項目納入研究，實為大型資料庫在研究工具的限制。本研究經過描述統計，如表1可以瞭解20個國家的學校規模情形，先進國家參與的學校大致在1,000名以下，只有英格蘭的學校學生人數超過1,000名，而亞洲四小龍也都超過1,000人以上，可見這四個國家的學校都傾向規模較大的，開發中國家的學校規模人數不等，菲律賓超過2,000人，匈牙利僅有466人。

表1 各個國家的描述統計　　單位：分、校、人

變項	平均數	標準差	校數	變項	平均數	標準差	校數
臺灣				澳洲			
成績	582	50	150	成績	495	51	185
人數	1921	1090		人數	821	404	
香港				比利時			
成績	581	56	117	成績	535	60	141
人數	1070	175		人數	625	426	
南韓				英格蘭			
成績	586	28	149	成績	510	63	57
人數	1080	419		人數	1168	312	
新加坡				西班牙			
成績	600	49	141	成績	529	35	141
人數	1178	313		人數	665	348	
賽普勒斯				挪威			
成績	459	22	58	成績	461	25	135
人數	484	151		人數	318	141	
埃及				紐西蘭			
成績	440	64	211	成績	487	52	160
人數	1011	732		人數	708	496	
匈牙利				日本			
成績	527	48	147	成績	569	31	145
人數	466	230		人數	507	215	
印度				美國			
成績	416	67	148	成績	505	49	197
人數	736	705		人數	730	387	
約旦				馬來西亞			
成績	422	46	137	成績	507	55	150
人數	757	425		人數	1417	611	
突尼西亞				菲律賓			
成績	411	29	141	成績	383	63	129
人數	953	397		人數	2266	2192	

肆、研究結果與討論

一、高度國民所得國家的學校規模與學生學習成就之分析結果

經過迴歸分析後的結果如表2，表中看出，在這8個高度國民所得的國家，僅有比利時的學校規模與學生學習成就呈現倒U字型的關係，澳洲則僅有線性關性達到顯著水準，並沒有倒U字型的關係，而其他的六個國家不僅線性沒有關係，二次式也沒有達到統計的顯著關係，也就是說，這些高度國民所得的國家在中等學校的規模與學生學習成就，並沒有呈現倒U字型的關係。

表2 高度國民所得的學校規模與學生學習成就之迴歸分析

變項	*b*	標準誤	*β*	*t*	Adj-R^2
日本					
常數	552.246**	12.125		45.546	5.2%
規模	0.023	0.048	.1574	0.484	
規模平方	0.000	0.000	.0732	0.225	
澳洲					
常數	448.880**	13.906		32.281	7.5%
規模	0.083**	0.030	.6317	2.740	
規模平方	0.000	0.000	-.4078	-1.767	
挪威					
常數	451.919**	8.953		50.480	1.1%
規模	0.040	0.046	.2211	0.883	
規模平方	0.0000	0.001	-.1384	-0.552	
紐西蘭					
常數	473.509**	10.853		43.630	4.1%
規模	0.017	0.026	.1673	0.660	
規模平方	0.000	0.000	.0379	0.150	
美國					
常數	511.939**	11.25		45.525	0.8%
規模	-0.026	0.026	-.2040	-0.979	
規模平方	0.000	0.000	.2491	1.196	
比利時					
常數	502.506**	12.905		38.939	5.7%
規模	0.083**	0.029	.5959	2.867	
規模平方	0.001**	0.000	-.5657	-2.722	
英格蘭					
常數	610.703**	92.533		6.600	6.2%
規模	-0.134	0.169	-.6668	-0.792	
規模平方	0.000	0.001	.4324	0.514	
西班牙					
常數	507.911**	9.299		54.623	14.5%
規模	0.018	0.031	.1794	0.588	
規模平方	0.000	0.000	.2043	0.670	

***p*<.01

二、亞洲四小龍的學校規模與學生學習成就

經過迴歸分析後的結果如表3，表中看出，亞洲四小龍之中，臺灣與南韓僅有線性關性達到顯著水準，也就是一次的直線達到顯著關係，但是沒有呈現倒U字型的關係。很特殊的情形是，新加坡呈現的是學校規模與學生學習成就之間的關係為U字型關係，其意義是，學生人數過多或者太少，其學校的學生學習成就表現都比較好，適度的學校規模則否，而香港不僅線性沒有關係，二次式也沒有達到統計的顯著關係，也就是說，這些國家的學校規模與學生學習成就沒有呈現倒U字型的關係。

表3 亞洲四小龍的學校規模與學生學習成就之迴歸分析

變項	*b*	標準誤	*β*	*t*	Adj-R^2
臺灣					
常數	542.907**	15.096		35.964	6.3%
規模	0.0340*	0.015	.6673	2.265	
規模平方	0.0000	0.000	-.4695	-1.594	
新加坡					
常數	640.405**	25.673		24.945	31.6%
規模	-0.192**	0.049	-1.2187	-3.897	
規模平方	0.001**	0.000	1.6918	5.4100	
韓國					
常數	536.331**	9.201		58.291	29.1%
規模	0.061**	0.017	.9128	3.553	
規模平方	0.0000	0.000	-.3989	-1.553	
香港					
常數	357.356**	88.465		4.040	14.8%
規模	0.3044	0.163	.9436	1.869	
規模平方	-0.0001	0.001	-.5811	-1.151	

**p<.01

三、開發中國家的學校規模與學生學習成就

經過迴歸分析後的結果如表4，表中可以看出，在這8個開發中國家僅有匈牙利與印度的學校規模與學生學習成就呈現倒U字型的關係，而其他六個國家不僅線性沒有關係，二次式也沒有達到顯著關係，也就是說，這些國家的學校規模與學生學習成就沒有呈現倒U字型的關係。

表4 開發中國家的學校規模與學生學習成就之迴歸分析

變項	*b*	標準誤	*β*	*t*	Adj-R^2
埃及					
常數	425.305**	10.514		40.450	3.2%
規模	0.013	0.015	.1448	0.875	
規模平方	0.000	0.000	.0369	0.223	
賽普勒斯					
常數	427.672**	23.113		18.503	3.8%
規模	0.144	0.100	.9773	1.448	
規模平方	-0.001	0.001	-.9900	-1.467	
匈牙利					
常數	471.000**	15.678		30.043	9.2%
規模	0.207**	0.064	.9833	3.226	
規模平方	-0.001**	0.001	-.7868	-2.581	
約旦					
常數	401.406**	10.574		37.962	10.1%
規模	0.021	0.0200	.1907	1.035	
規模平方	0.000	0.000	.1350	0.733	
菲律賓					
常數	383.955**	11.283		34.031	0.2%
規模	-0.002	0.008	-.0558	-0.200	

規模平方	0.000	0.000	.0910	0.327	
馬來西亞					
常數	486.781**	21.891		22.236	4.5%
規模	0.009	0.029	.1020	0.315	
規模平方	0.000	0.000	.1117	0.346	
印度					
常數	373.169**	11.303		33.015	11.4%
規模	0.075**	0.018	.7928	4.316	
規模平方	0.000	0.000	-.7026	-3.826	
突尼西亞					
常數	407.819**	11.359		35.902	1.3%
規模	-0.001	0.021	-.0142	-0.048	
規模平方	0.000	0.000	.1284	0.435	

**p<.01

四、圖示呈現學校規模與學生數學成就

上述研究發現看出，學校規模與學業成就有四種關係。第一，兩者如倒U字型關係，如比利時、匈牙利與印度。第二種情形為線性關係，也就是學校規模大小與學業成就之間是線性關係，學校規模愈大，學生學習成就表現愈高，如臺灣、南韓、澳洲。第三種情形是U型關係，即學校規模愈大及愈小，學生的學習成就表現愈好，如新加坡。第四種情形是，不管學校規模大小都不會影響學生的學習成就，如美國、英格蘭、紐西蘭、挪威、約旦、菲律賓、馬來西亞、突尼西亞等。

為了讓讀者掌握四種關係情形，以下將上述發現說明，就U字型來說，圖1是比利時，圖中每個點代表一所學校，圖中曲線即為U字型線條。圖2為印度，圖3為匈牙利。新加坡U型關係呈現如圖4。臺灣的情形為直線關係，兩者呈現正向關係，如圖5；有很多國家是不管學校規模大小都不會影響學生的學習成就，就以美國為例，如圖6，其直線為平行X軸，就可以看出兩者沒有顯著的相關。

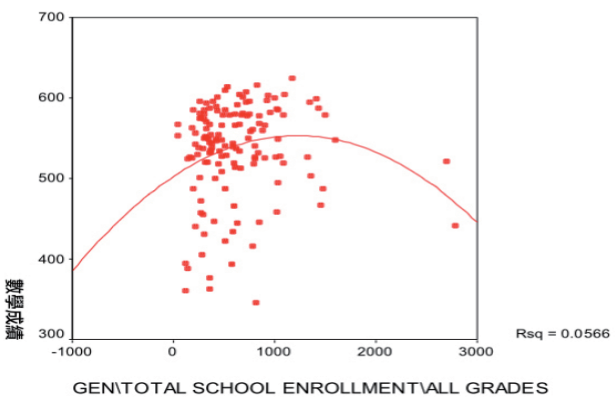


圖1 比利時的學校規模與學生數學成就之散布情形

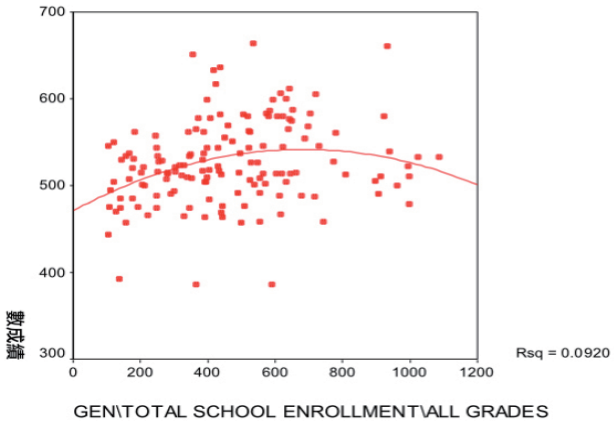


圖2 印度的學校規模與學生數學成就之散布情形

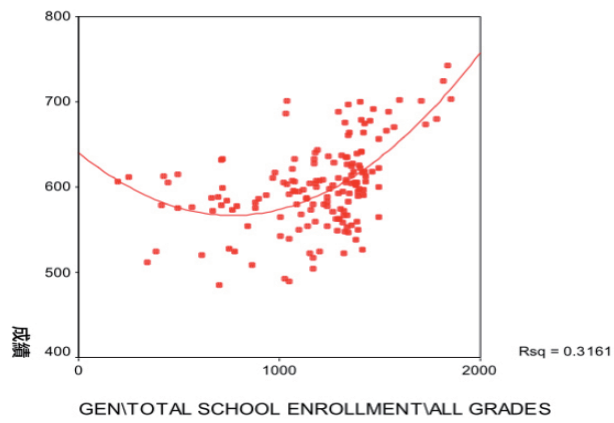


圖3 匈牙利的學校規模與學生數學成就之散布情形

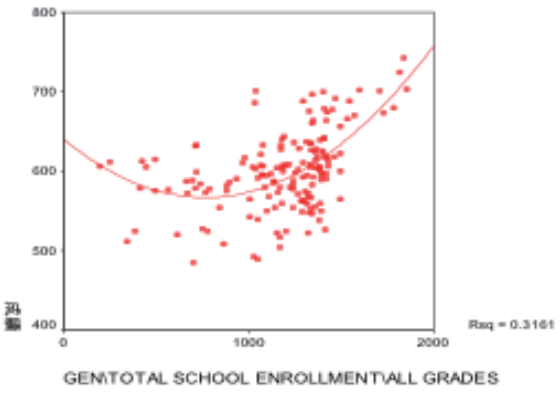


圖4 新加坡的學校規模與學生數學成就之散布情形

圖5 臺灣的學校規模與學生數學成就之散布情形

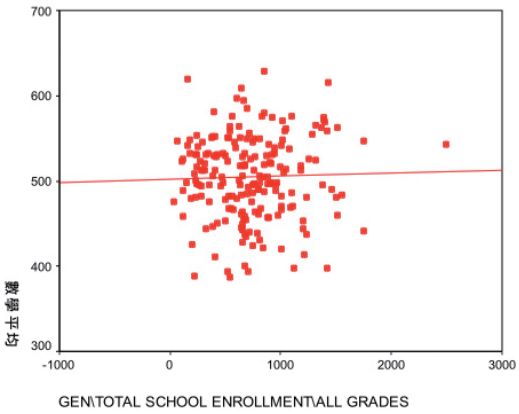


圖6 美國的學校規模與學生數學成就之散布情形

五、綜合討論

近年來，對於TIMSS的研究集中在影響學生學業成就的因素之分析，這方面因素包括家庭、學校、教師、學生個人及同儕的因素探討（余民寧，2006）。但是，也有一些研究則是以跨國方式對於影響學生數學及科學成就研究進行分析（張芳全，2007，2010a；鄭心怡，2004；House,2006；O’ Dwyer, 2005）。上述的研究均沒有掌握學校規模與學生學習成就之關聯。國內外的研究中，以學校投入與產出的教育投資觀點，來找出學校最適規模的研究卻是不少（林文達，1977，1990；林嘉薇，2002；林淑貞，1979；吳炳銅，1994；郭添財，1991，1996；蓋浙生，1985；Edington & Martellaro, 1990；Riew,1966；Thompson, 1994）。這些研究大抵是以學生單位成本及學校規模進行分析者多，而以學校規模與學習成就的實徵研究卻相當少，因此本研究探討此議題是一種新的嘗試。本研究以TIMSS 2003的大型資料庫對20個國家的學校規模與學生學習成就之關係探究，研究中以二次式的迴歸分析來檢

定兩者之間是否呈現倒U字型關係，此分析也是過去所無。本研究從20個國家的研究發現，中等學校規模與學業成就之關係，具有四種類型。茲就研究的結果，討論如下：

第一，本研究發現在各國的中等學校之中，學校規模與學習成就之間呈現倒U字型關係者，如比利時、匈牙利、印度是如此。這樣的研究結果發現呼應的本研究預期的情形，這也與過去許多研究結果發現，學校規模與學習成就呈倒U字型的關係一樣，這些研究如林文達（1990）、Chopin（2003）、Melvin與Roy（2003），然而，上述的研究結果並不是以學習成就為依變項，而是以教育經費投資多寡為依據，如學生單位成本為主。雖然教育成本與學校規模，以及學習成就與學校規模，在兩種議題所運用的研究變項屬性不相同，但是可以確定的是，學習成就、教育經費（學生教育成本），在某些國家之中，它們與學校規模呈現倒U字型關係。而本研究發現呈現倒U字型的情形，不僅是高度國民所得國家有這種情形，開發中國家也有，這可以說明學校規模與學習成就之倒U字型關係，可能跨文化上，也有相同者。

第二種情形是僅有線性的關係，也就是學校規模大小與學業成就之間是線性關係，即學校規模愈大，學生學習成就表現愈高，這個情形發現於臺灣、南韓、澳洲，它與張芳全（2010b）發現一樣，也與Fowler與Walberg（1991）、Riew（1986）、Sande（1993）的研究發現一致。

第三種情形是，學校規模愈大及愈小，學生數學成績愈好，這種情形發現在新加坡之中，它代表數學成就表現不是在學校的最適當規模，而是在較小及較大的學校中。這正呼應了Paul（2000）對學校與班級認為，它可採用較小規模能幫助學生提高學業成就，以及Gardner, Ritblatt與 Beatty（2000）的研究發現，較小規模學校有助於降低曠課、輟學率以及提高父母參與，其原因可能是教師與學生、父母互動機會增加、教師減輕教學負擔而更有能力關注學生，並給予支持，增進其求學興趣，以提高學生的學習成果等。

第四種情形是，不管學校規模大小都不會影響學生學習成就，這樣的研究發現與Brown和Saks（1975）、Michelson（1972）、Lamdin（1995）的研究發現一樣。而本研究在兩個變項所發現的國家包括美國、英格蘭、紐西蘭、挪威、約旦、菲律賓、馬來西亞、突尼西亞等都是有這樣的情形，也就是中等學校規模的大小並沒有影響學生的數學成就。若從這些國家的結果來看，不管是高度國民所得國家或開發中國家，學校規模不見得影響學生的學習成就，這代表還有其他重要的因素會影響學習成就。

上述研究發現，可以理解學校規模與學生學習成就屬於非線性關係與直線關係，因為新加坡的規模較大及較小的學校與學業成就有正相關，有些研究卻又發現較小規模的學校是幫助學生提高其學業成就，如臺灣、南韓、澳洲等。本研究發現，在20個國家之中，學校規模與學習成就之關係有不同的情形，這說明兩者之關係可能有跨文化因素，或有其他更重要因素影響學生數學成就，如學生因素（個人智商、學習態度、學習信念）、學校因素（教師教學、教育資源或學校氣氛）、學

生家庭因素（如家庭文化資本、財務資本、社會資本等），這說明兩者為非線性關係，若單就線性來進行兩者關係論述，可能流於武斷。就如學業成就除了智能表現之外，更應包含情意與技能層面，如單以學業成就來與學校規模掌握之間的關聯，可能太過於簡單。

伍、結論與建議

一、結論

經過上述的研究分析與討論，本研究結論：各國的學校規模與學生學習成就之間存在四種類型的關係。第一，兩者之間為倒U字型關係，代表的國家如比利時、匈牙利與印度，也就是學校規模過大與過小對於學生學習成就都不好，最適規模下，學生學習成就表現較好。而在高度所得國家與開發中國家都可能有第一種情形。第二種情形是僅有直線性關係，即中等學校的規模大小與學業成就之間為線性關係，學校規模愈大，學生數學成就表現愈高，這種情形發現於臺灣、南韓與澳洲。第三種情形是，學校規模愈大及愈小，學生數學成績愈好，這種情形發現於新加坡，它代表數學成就表現不是在學校的最適當規模，而是在較小及較大的學校之中。第四，不管學校規模大小都不會影響學生數學成就，這現象發現於美國、英格蘭、紐西蘭、挪威、約旦、菲律賓、馬來西亞、突尼西亞等，中等學校的規模不見得影會響學生的數學成就，這代表了，還有其他重要因素會影響學習成就。

二、建議

本研究依據結果，建議如下：

首先，學校規模小是否可以提昇學習成就還是很有爭議，要提高學生成績表現，仍應考量學校內外的其他因素（如教師教學投入、學生來自的社會階層、學校氣氛、家長參與、甚至學校所在地區的特性等）。本研究以TIMSS2003年的20個國家學校規模與學習成就發現，學校規模與學習成就之間存有四種關係，這四種關係中較多國家的現象是，學校規模愈大不一定會影響學習成就，也就是兩者不僅沒有線性關係，也沒有完全呈現為倒U字型的關係。換言之，各國的學校規模與學生學習成就之關係並非線性，也非倒U字型關係，在影響學生學習成就，學校規模可能僅是原因之一，但並不一定是主因。

其次，試著評估臺灣增加學校規模，提高學業成就的可行性。本研究結果發現，臺灣的中等學校規模與數學成就存有直線關係，沒有倒U字型的關係，代表中等學校規模愈大，其學生數學成就傾向較高。但是這種情形仍需後續的研究。近年來，臺灣面臨少子化，對偏遠地區採用併校增加學校規模，也許就是一種因應措施，但宜再試著評估其可行性。

針對未來研究有以下建議：

首先，未來如果需要討論學校規模與學習成就之關係，或許還應納入可能影響變項進行分析。學校規模大小強調將教育資源投入於學生後所產生成效，而學業成就只是一項而已，其他如輟學率、課外參與等也是影響因素之一。未來可以納入更多的變項，如以教育資源效果作為探討學業成就變項，將會有新的研究發現。尤其學校規模與教育成本關係密切，然而在各年度的TIMSS資料庫之中，並沒有教育成本或相關的教育經費資料，未來研究如果有教育成本、學生學習成就及學校規模人數或生師比，可以進一步分析，如此或許會有更完整分析。當然，學生的情意表現、社區資源、家庭背景等可能干擾因素，亦可以一併納入分析，這更能瞭解學校規模與學習成就的關係。此外，一個國國家的學校規模可能受到學區或城鄉之差異，因而對於學生學習成效，乃至於學校經營成效有不同，未來對於學校規模與學習成就之關係的研究，還可以考量學校所在的地理區域，如城市與鄉間的學校，這更可以瞭解不同城鄉的學校在學校規模與學習成就之關聯。當然，未來的研究在學生學業成就可以採取更廣的界定，例如納入一些非認知領域的學習表現情形，如學生的學習態度、滿意度。本研究是以TIMSS 2003的數學學習成就為主，但是這僅僅為認知的表現，學生的非認知部分，如情意、生活適應、學習態度等都能未納入分析，易言之，本研究將的學業成就僅限於數學成就，是否過於狹隘？僅以標準化的測驗為主，多元評量與多元智慧宜納入考量，以及非認知的表現亦應納入，這或許更可以瞭解學生的學習成就與學校規模之關聯性。

其次，採用比較研究法進行地區性分析，理解學業成就與學校規模關係。研究發現，不同研究採用樣本不相同，研究樣本所在的區域也有所差異，所以即使未來研究發現學校最佳規模模式，未來可以以每一個國家的學校人數或班級人數進行區域性比較，使研究盡可能周延。當然，城鄉教育的差距也影響學校規模，進而影響學生數學成就，例如Stewart（2009）以美國德州的五所不同經濟發展區域的高中分析小型學校與大型學校的學生學習表現，研究發現在鄉村的小型學校學生在閱讀、寫作、數學及科學成就的表現，比起在較都會區大型學校，學生的社會階層較低者還好。可見，都會區的大型學校不一定是學生學習成就表現較好，仍需有其他的因素納入分析，所以未來宜將城鄉差距的發展因素納入。學校規模大小反應教育資源投入差距，也反映每位學生平均可獲得的資源差異，未來更應朝影響學校規模進一步探究。同時，以規模經濟的效用，或效能為立論基礎並無不可，但是除了經濟方面的考量之外，各國的社會、文化、歷史因素之探究基礎是本研究較為不足，未來應將相關的變項納入分析。

此外，運用不同等級的學校（如國小及國中）探討學校規模與學業成就之關係，也是一種方式，因為本研究以國中為研究對象，若以職業與升學導向作為區分學校標準，學校規模與學業成就之間，又將呈現何種關係？或者以國民小學為樣本又會有哪些發現，也是值得思考。最後，本研究是以TIMSS 2003年的資料，未來可以運用1999及2007年，或者後續年度的資料來追蹤分析，或許更能瞭解各國的學校規模與學習成就之關係。

參考文獻

林文達（1977）。當前國民中學經營的策略。**人與社會，3**（5），34-40。

林文達（1990）。**教育經濟學**。臺北市：三民。

林淑貞（1979）。**臺北市國民中學經營規模之研究**。國立政治大學教育研究所碩士論文，未出版，臺北市。

江亞萍（1999）。**台閩地區國民中學教育規模之實證研究**。私立淡江大學產業經濟研究所碩士學位論文，未出版，臺北縣。

林嘉薇（2002）。**最適班級與學校經濟規模分析—以臺北縣公立國民小學為例**。東吳大學經濟學系碩士論文，未出版，臺北市。

余民寧（2006）。影響學習成就的因素探討。**教育資料與研究，73**，11-24。

吳炳銅（1994）。**臺北縣國民小學最適經營規模之研究**。國立臺北師範學院初等教育學系碩士論文，未出版，臺北市。

陳正倉、林惠玲、陳忠榮與鄭秀玲（2006）。**個體經濟學—理論與應用**。臺北市：雙葉。

陶韻婷（2007）。**國中生科學成就與學生背景、學校規模及城鄉之關聯性探討--以TIMSS 2003為例**。國立臺灣師範大學生命科學研究所碩士論文，未出版，臺北市。

張芳全（2007）。臺灣、美國及日本之國二學生家庭作業與數學成就關係之比較。**教育資料集刊，34**（3），285-316。

張芳全（2010a）。以SEM檢定影響數學成就因素：亞洲四小龍國二生參與TIMSS2003的資料為例。**教育行政論壇，2**（2），1-34。

張芳全（2010b）。**多層次模型在學習成就之研究**。臺北市：心理。

蓋浙生（1985）。**教育經濟學（再版）**。臺北市：三民。

郭添財（1991）。**臺灣省南部地區國民小學最適經營規模之研究。**國立高雄師範大學教育研究所碩士學位論文，未出版，高雄市。

郭添財（1996）。**臺灣省國民小學規模經濟之研究**。國立政治大學教育研究所博士學位論文，未出版，臺北市。

鄭心怡（2004）。**教育指標與經濟指標對學業成就影響之國際比較：以TIMSS為例**。臺北師範學院教育政策與管理研究所碩士論文，未出版，臺北市。

Bowles, T. J., & Bosworth, R. (2002). Scale economies in public education: Evidence from school level data. *Journal of Education Finance, 28*(2), 285-299.

Brown, B., & Saks, D. (1975). The production and distribution of cognitive skills withinschools. *Journal of Political Economy, 83*, 571-593.

Chopin, S. L. (2003). *The effect of school size, socioeconomic status, and grade-level configuration on academic achievement in Louisiana public schools.* Unpublished doctoral dissertation, Louisiana Tech University, Louisiana.

Cohn, E. (1968). Economies of scale in Iowa high school operations. *Journal of Human Resources, 3*(4), 422-434.

Donna, D., Dennis, H., & Shirley, S. (2003). School district size and student performance *Economics of Education Review, 22*, 193–201.

Edington, E. D., & Martellaro, H. C. (1990). Does school size have any relationship to academic achievement? *Rural Educator, 11*(2), 6-11.

Fowler, W. J., & Walberg, H. J. (1991). School size, characteristics and outcomes. *Educational Evaluation and Policy Analysis, 13*, 189-202.

Francis, L. J. (1992). Primary school size and pupil attitudes: Small is happy? *Educational Management and Administration, 20*(2), 100-104.

Gardner, P., Ritblatt, S., & Beatty, J. (2000). Academic achievement and parental school involvement as a function of high school size. *High School Journal, 83*(2), 21.

Gentry, K. J. (2000). *The relationship between school size and academic achievement in Georgia's public high schools.* Unpublished doctoral dissertation, University of Georgia, Georgia.

House, J. D. (2006). Mathematic beliefs and achievement of elementary school students in Japan and the United States: Results frrom the Third International Mathematics and Science Study. *The Journal of Genetic Psychology, 167*, 31-45.

Lamdin, D. J. (1995). Testing for the effect of school size on student achievement within a school district. *Education Economics, 3*(1), 33-42.

Melvin, V. B., & Roy M. H. (2003). An examination of the effect of elementary school. *International Review of Education, 49*(5), 463-474.

Merritt, R. (1993). The effect of enrollment and school organization on dropout rate. *Phi Delta Kappan, 65*(3), 224.

Michelson, S. (1972). Equal school resource allocation. *Journal of Human Resources, 7*, 283-306.

NCES(2004). *TIMSS 2003 user guide for the international database.* USA Department and Education: National Center for Education Statistics.

Okpala, C. O. (2002). Educational resources, student demographics and achievement scores. *Journal of Education Finance, 27*(3), 885-907.

O' Dwyer, L. M. (2005). Examining the variability of mathematics performance and its correlates using data from TIMSS' 95 and TIMSS' 99. *Educational Research and Evaluation, 11*(2), 155-177.

Osburn, D. D. (1970). Economies of size associated with public high schools. *The Review of Economics and Statistics, 52*(1), 113-115.

Paul, A. (2000). How school size affects academic achievements. *School Planning & Management, 39*(5), 86.

Pritchard, G. W. (1987). *Academic achievement and perceptions of school effectiveness and their relationship to school size.* Unpublished doctoral dissertation, South Carolina State University, South Carolina.

Riew, J. (1966). Economies of scale in high school operation. *The Review of Economics and Statistics, 48*(3), 280-288.

Riew, J. (1986). Scale economies, capacity utilization, and school costs: A comparative. analysis of secondary and elementary schools. *Journal of Education Finance, 11*, 433-446.

Sander, W. (1993). Expenditures and student achievement in Illinois. *Journal of Public Economics, 52*, 403-416.

Stevens, N. G.., & Peltier, G.. L. (1994). A review of research on small-school student. participation in extracurricular activities. *Journal of Research in Rural Education, 10* (2), 16-20.

Stewart, L. (2009). Achievement differences between large and small schools in Texas. *The Rural Educator, 30*(2), 20-29.

Stigler, J. W., Lee, S., & Stevenson, H. W. (1987). Mathematics classrooms in Japan,Taiwan, and the United States. *Child Development, 58*, 1272-1285.

Thompson, J. A. (1994). Scale economies and student performance in Hawaii *Journal of*

Education Finance, 19(3), 279-291.

Wendling, W., & Cohen, J. (1981). Education resources and student achievement: Good news for schools. *Journal of Education Finance, 7*, 44-65.

World Bank(2004). *World development report 2004.* Washington, D.C.: Oxford university press.

【致謝辭】

作者感謝三位匿名審查者的寶貴意見，供本文修改參考，使本文可讀性更高，文中若有不周，實為作者責任。同時並感謝行政院國科會專題研究補助計畫得以完成研究，計畫編號為NSC95-2413-H-152-012-

大學生微積分學習之分群化概念結構圖分析

蔡孟憲 / 林原宏

大學生微積分學習之分群化概念結構圖分析

蔡孟憲
臺中教育大學數學教育學系碩士班研究生

林原宏
臺中教育大學數學教育學系教授

摘要

「微積分」課程是普通大學、科技大學理工及商管學院必修的專業基礎科目，用來培養學生修習進階專業課程所需要的各種能力，奠定紮實微積分基礎觀念對於大學課程相關領域的學習有相當大的幫助。

本研究旨在應用多元計分概念詮釋結構模式（Polytomous Concept Advanced Interpretive Structural Modeling,簡稱為PCAISM）分析方法，並利用模糊集群分析將學生分群，探討各群大學生的微積分概念結構圖。本研究探討44位大學生，測驗試題共10題，測量6個微積分基本概念。

研究結果發現：1.運用概念結構圖可進行個別化的認知診斷，診斷學生概念精熟與否，藉由分析得到的結果，瞭解個別學生概念間的指向和連結性，以作為補救教學之參考依據；2.透過模糊集群有助於教師進行分組補救教學，經由有效的管理方法，以模糊集群進行分群，本研究受試者可分為低精熟組和高精熟組，發現各群學生有其共同相似特徵：（1）就概念階層而言，不同組別的概念階層數、層次會有所不同；（2）就概念連結而言，各群學生概念間的指向關係有所不同；（3）就概念精熟度而言，高精熟組各概念的精熟度都很高，優於低精熟組。教學者若能將認知診斷結果相似的學生集中並進行分組補救教學，即可有效地提昇學習者的學習成效。

本研究亦提出相關建議，可利用同儕互助學習（peer assisted learning）的理念進行補救教學。由於本研究發現高精熟組的學生概念精熟度近似專家，可將低精熟組的學生分組並指派高精熟組學生協助同學課後學習，亦可使高精熟組的學生概念更加精熟，可作為補救教學及未來研究之參考。

關鍵字：多元計分概念詮釋結構模式、微積分概念、補救教學、模糊集群、同儕學習

Concept Diagram on the Cognition Diagnosis of Statistics Learning and Clustering with Application for University Students

Meng-Xian Tsai
Master degree student, Department of Mathematics Education,
National Taichung University
exin0955@hotmail.com

Yuan-Horng Lin
Professor, Department of Mathematics Education,
National Taichung University

Abstract

Calculus is an important course for university students because it is the foundation of quantitative research. The purpose of this study is to analyze the concept diagram of calculus concepts for university students with clustering based on concept proficiency. Methodology in this study is PCAISM (polytomous concept advanced interpretive structural modeling). This method can not only present the individualized concept structure by hierarchical diagram, but also calculate the magnitude of mastery on each concept. Besides, fuzzy clustering on concept proficiency expresses the cognitive characteristics. Empirical data comes form paper-and-pencil assessment of calculus course. The results show that all students could be classified into two clusters. Proficiency and characteristics of concepts between these two clusters are quite different. According to the results, it shows PCAISM can provide useful information for cognition diagnosis. It is found that using peer assisted learning will be a potential way to help student learning calculus in university. Finally, some suggestions and recommendations for future investigation are discussed.

Keywords: polytomous concept advanced interpretive structural modeling, calculus concepts, remedial instruction, clustering, peer assisted learning

壹、緒論

「微積分」課程是普通大學、科技大學理工及商管學院必修的專業基礎科目，用來培養學生修習進階專業課程所需要的閱讀、分析、推理、計算及演算能力，奠定紮實穩固微積分基礎觀念對於大學課程相關領域的學習有相當大的幫助。傳統只看測驗總分、排名次的評量，只看到表面的現象，無法得知其背後的意涵，如成績相同並不代表學生能力相同，試題的作答反應組型亦不相同，且無法顯示每個學生對概念的理解程度，也無從得知個別學生對概念間的落差，造成老師無法針對學生所缺乏的概念進行補救。

本研究使用了透過模糊理論（fuzzy theory）的計算方法並運用詮釋結構模式（interpretive structural modeling）的階層結構運算法則，將呈現出個人化概念階層結構，藉以分析微積分考試的資料。從多元計分概念詮釋結構模式（polytomous concept advance interpretive structural modeling, 簡稱為 PCAISM）方法得之其結果比只單用成績更能有效的解讀個別學生的學習狀況與成效，並且提供較多的訊息供老師解讀學生的狀況，使測驗能確實發揮評量的功效，並輔助老師了解學生的學習狀況與學習困難之處。

在學校教室環境中，限於教師人力資源的限制，實際上很難進行個別補救教學。根據認知診斷訊息進行適當分群，使得「群內同質，群間異質」，是有效知識管理的一環。Zadeh（1965）提出的模糊理論，考慮隸屬度（membership）的非二元觀點（Kaufman & Rousseeuw, 1990），此觀點亦適合社會科學的資料分析（吳柏林，2005）。由於學生的認知狀態非二元分類所能解釋；因此，模糊集群（fuzzy clustering）所依據的隸屬度計算，適合應用於學習結果的分析（Lin, Yu, and Wu, 2006）。所以，本研究以微積分教學行動研究資料，應用模糊集群方法，依據學生的認知診斷訊息給予分群，以作為進行分組補救教學的依據。

貳、文獻探討

一、多元計分概念詮釋結構模式

多元計分概念詮釋結構模式是以 Lin, Hung, and Huang（2006）所提出的概念詮釋結構模式（concept advanced interpretive structural modeling, CAISM）為基礎，針對其計分法進行擴展與改良，並推導出適用於多元或混合的計分測驗資料之演算法則，增進概念詮釋結構模式理論之應用範疇（Lin and Lin, 2010; Warfield, 1976）。有關多元計分概念詮釋結構模式，其步驟如圖1所示：

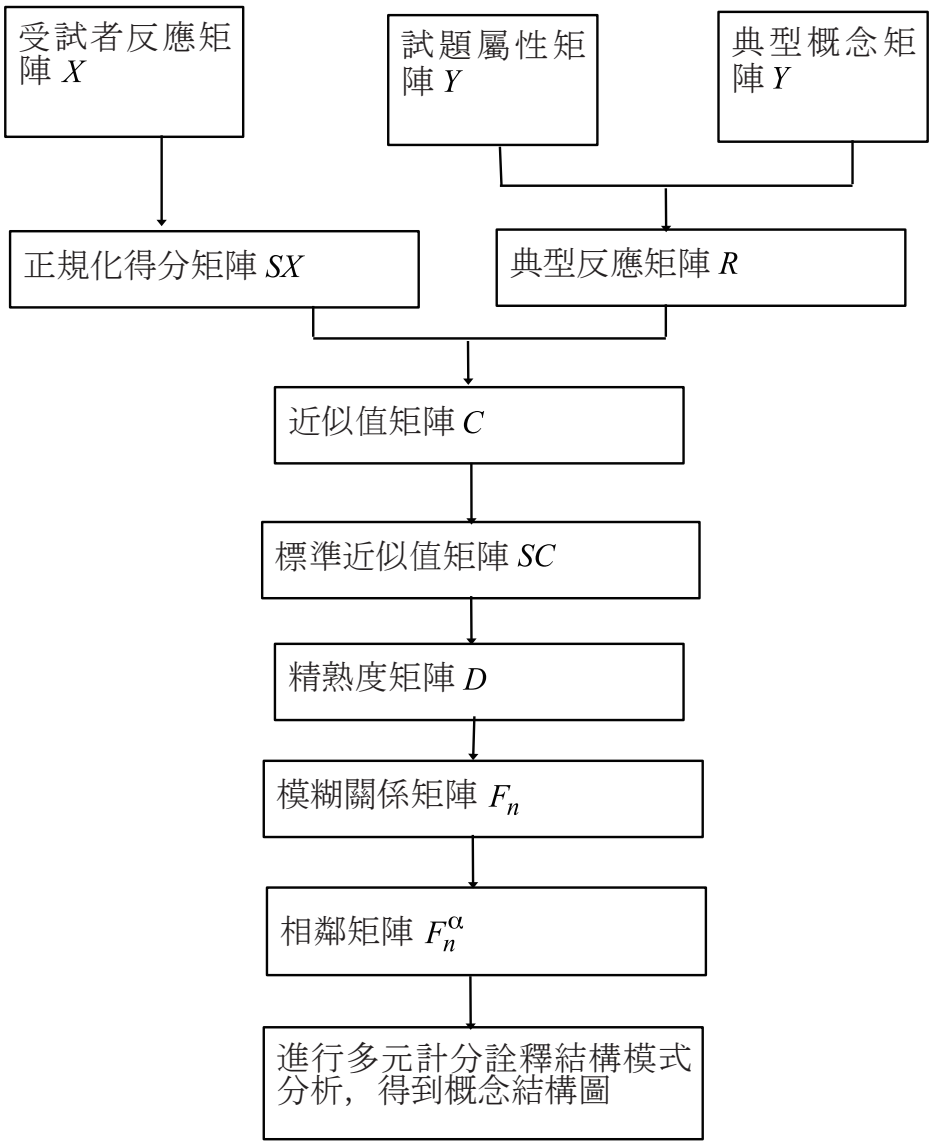


圖1 多元計分概念詮釋結構模式的演算法

二、模糊集群分析及相關研究

Zadeh（1965）提出模糊理論，將元素和集合之間的關係，以介於[0,1]之間的隸屬度描述（吳柏林，1996；林原宏，2005）。模糊集群融合隸屬度訊息，在相關的研究實例中，由於學習的量化結果具有不確定性（uncertainty），因此適宜以模糊集群來分析，依據學生的學習成果給予分群，以利進行補救教學（林原宏，2007）。本研究採目標函數法（objective function）之fuzzy c-means進行模糊集群分

析（Bezdek, 1981）。林原宏、黃國榮（2003）根據其理論撰寫模糊集群分析程式，軟體名稱為FCUT，研究者僅需輸入欲分析 N 位個數的 M 個變項，並設定預計想分群的群數，至於群數的決定，採用分割係數（partition coefficient）$F(U;C)$ 與分割亂度（partition entropy）$H(U;C)$ 兩個指標來決定（Bezdek, 1981）。其公式如下所示：

$$F(U;C)=\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}(u_{cn})^2 \qquad \text{（公式1）}$$

$$H(U;C)=\frac{-1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}u_{cn}\ln(u_{cn}),\forall u_{cn}\neq 0 \qquad \text{（公式2）}$$

在上述公式中，N 為受試總人數，u_{cn} 為受試者 n 隸屬於群組 c 的隸屬度。當 $F(U;C)$ 值較大時且 $H(U;C)$ 值較小時，為較佳的群數（Bezdek, 1981），研究者可利用不同群數下的分割係數與分割亂度大小比較，據以決定最佳群數。利用模糊集群應用在學習成效的評量上，獲致相當不錯的成效。黃馨瑩、林原宏、莊曜遠（2007），在分析國小學童知識結構特徵與容量概念中，指出以模糊集群方式可將學童有效分群，進而發現各分群學童的不同學習特性，提供補救教學之參考；Lin, Yu and Wu（2006）以模糊集群分析方法分析國小學童在機率概念的解題策略，將學童分成不同集群，發現學童的解題策略，會因年級不同而有所差異。

三、相關研究文獻

辛靜宜、林珊如、葉秋呈（2005）以五年制專科生為研究對象，利用「微積分學習動機導向策略問卷」，進行微積分學習動機與學習策略之初探研究，發現微積分學習動機可分為自我效能與控制信念、學科價值、解題動機、理論動機、考試焦慮等5個因素，學習策略也分為理解、習題演練、上課學習習慣、個人學習習慣、後設認知等5個因素。該研究針對學習動機與學習策略的情意面向，可提供學習輔導的參考，但若進一步進行知識結構的診斷，則可進行補救教學的實施。王財印、林坤霖、柯麗蓉、郭柏立（2010）探討技專校院統測學成績與大一微積分學習成就之相關性及其影響性，發現不同學院與系別之間的學習呈顯著差異，可進一步了透過適當的銜接教材教法，使得微積分學習有困難的學生，能有更高更有效的學習成就。本研究的後續研究，可以利用知識結構的診斷方法，提供微積分學習困難學生的補救策略。

Fennema and Sherman（1978）研究美國中學學生數學學習態度對於數學成就的影響發現，不同數學成就的學生對數學實用性看法不同，這樣的研究發現，可進一步分析不同微積分知識結構的學生，其對微積分應用性看法的有哪些不同。Ferrini-Mundy and Gaudard（1992）發現同樣一起剛進入大學學習微積分的學生，在高中有先接觸過微積分基本概念的學生，對於大學課程的學習較優於其他學生。此研究意涵著，有良好的微積分先備知識（prior knowledge），可以得到較佳的學習效果。Stylianou, Kenney, Silver, and Alacaci（2000）也指出類似的看法，建構良好基本概念的學習，對於學習微積分有很大的幫助。

綜合以上文獻所述，可以看出有關微積分的學習，無論在學習動機與學習策略方面，知識結構或概念結構的探討，對於微積分學習的情意與認知層面，將有重要幫助。而且結構化的微積分概念分析方法，可進一步做為補救教學或進階學習輔導的參考依據。因此，本研究以概念結構圖進行微積分學習的探討，有其重要與可行之處。

參、研究設計與實施

一、資料來源

本研究資料來源為研究者所就讀學校大學部一年級的必修課程，該課程的大學生必須接受一學年的基礎微積分課程。且本研究為下學期課程中平時考所實施的測驗資料，試題內容如附錄。

二、研究樣本與分析軟體

本研究採便利取樣，受試者為共同修習同一門課的大學生共44位。研究者以「認知診斷之測驗分析即時服務系統」中所提供的多元計分概念詮釋結構模式（PCAISM）。如圖2所示，經選取適當的閾值 ($\alpha = .55$) 後，可獲得受試者的概念知識結構圖及各概念之精熟度。所得之概念精熟度，以FCUT軟體進行模糊集群分析（林原宏，2005），據以獲得分群結果。根據上述流程分析結果，來探討不同集群的大學生的基礎微積分概念。

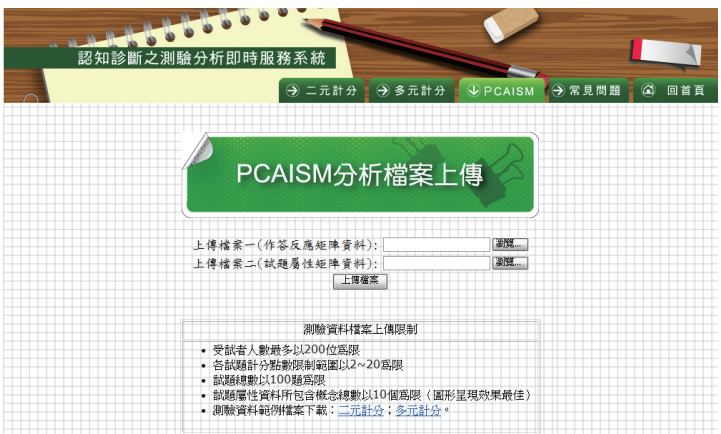


圖2 認知診斷之測驗分析即時服務系統

三、研究工具

本研究所測量的基礎微積分概念為教學者參閱微積分相關資料後，所認定對於修習微積分概念課程時，所應該了解的基礎微積分概念，且由研究者自行編製測驗，本測驗的 Cronbach's α 信度為 .59 ，以大學生為研究對象。測驗包含受試者44

人，6個概念（教學者認定的），總計10題，答對1題得10分，若受試者於作答時，有提出試題所應具有的概念，但未能將試題作答完全，則給予一半的分數即5分。試題為多元計分，試題概念屬性如表1所示。試題反應矩陣 $Y=(y_{ma})_{M\times A}$ 及試題答對率如表2所示，該表中，1代表有該題有測量到該概念；0代表該題沒有測量到該概念。

表1 試題概念屬性

概念編號	概念名稱
1	微積分基本定理
2	積分定理
3	定積分均值定理
4	定積分的性質及基本定理
5	偶奇函數在定積分上的應用
6	平面上曲線所圍的區域面積求法

表2 試題反應矩陣 $Y=(y_{ma})_{M\times A}$ 及試題答對率

試題	概念屬性						答對率
	1	2	3	4	5	6	
1	1	0	0	0	0	0	94%
2	0	1	0	0	0	0	36%
3	0	0	0	0	0	1	61%
4	0	0	1	0	0	0	70%
5	0	0	0	1	0	0	41%
6	0	0	0	1	0	0	45%
7	1	0	1	0	0	0	67%
8	0	0	0	0	0	1	88%
9	0	1	0	0	1	0	44%
10	0	0	0	0	0	1	41%

肆、研究結果與討論

一、微積分概念精熟度之模糊集群分析

以學理上由言，分群群數可從2至 $N-1$ 群（ N 為施測總學生數），但在實證研究上，以實際可行的不同群數間，選擇一個較佳的群數。本研究屬於微積分教學行動研究，因此，就現場教學資源而言，群數以2群至7群間選擇個較佳群數決定之。所以，施測資料之各概念的精熟度進行模糊集群分析，以2群至7群數下比較其最佳群數，其分割係數和分割亂度之值如表3所示。由於在2群的情形下，其分割係數最大且分割亂度最小，符合最佳群數的決定。因此，本研究將全體學生分成二群，各學生在二群的隸屬度與隸屬群組如表4所示。

各群人數與群中心與群中心之各概念精熟度如表5，其折線圖如圖3所示。由表5和圖3可知，顯示第一群的概念精熟度都介於.50到.57之間，第二群的概念精熟度都是1。因此，研究者將第一群命名為低精熟組，第二群為高精熟組。

全體學生中，低精熟組有41位，高精熟組有3位。限於篇幅，無法一一呈現每位學生的概念詮釋結構圖。因此，研究者在二群學生中，各抽取兩位受試者，低精熟組為受試者32與受試者40，高精熟組為受試者23與受試者43。研究者以此四位學生，比較說明各群學生的微積分概念階層知識結構圖之特徵。

表3 不同分群數之分割係數及分割亂度

	群數					
	2	3	4	5	6	7
分割係數	.97	.68	.54	.52	.52	.51
分割亂度	.07	.50	.80	.92	.99	.99

表4 隸屬度模糊矩陣及分群

受試者編號	第一群隸屬度	第二群隸屬度	隸屬群組	受試者編號	第一群隸屬度	第二群隸屬度	隸屬群組
1	0.997864	0.002136	1	23	0.000014	0.999986	2
2	0.990472	0.009528	1	24	0.972177	0.027823	1
3	0.991847	0.008153	1	25	0.982394	0.017606	1
4	0.974436	0.025564	1	26	0.990472	0.009528	1
5	0.979177	0.020823	1	27	0.997864	0.002136	1
6	0.979177	0.020823	1	28	0.993118	0.006882	1
7	0.989297	0.010703	1	29	0.986142	0.013858	1
8	0.991679	0.008321	1	30	0.989297	0.010703	1
9	0.983104	0.016896	1	31	0.980180	0.019820	1
10	0.992298	0.007702	1	32	0.982226	0.017774	1
11	0.988592	0.011408	1	33	0.983104	0.016896	1
12	0.970727	0.029273	1	34	0.987371	0.012629	1
13	0.973791	0.026209	1	35	0.993118	0.006882	1
14	0.976802	0.023198	1	36	0.989297	0.010703	1
15	0.985858	0.014142	1	37	0.984778	0.015222	1
16	0.976271	0.023729	1	38	0.982258	0.017742	1
17	0.989297	0.010703	1	39	0.974329	0.025671	1
18	0.980770	0.019230	1	40	0.980770	0.019230	1
19	0.984889	0.015111	1	41	0.000014	0.999986	2
20	0.991016	0.008984	1	42	0.991679	0.008321	1
21	0.994578	0.005422	1	43	0.000014	0.999986	2
22	0.987500	0.012500	1	44	0.979177	0.020823	1

表5 各群人數與群中心之各概念的精熟度

群組	人數	概念編號					
		1	2	3	4	5	6
第一群	41	.55	.50	.53	.57	.52	.54
第二群	3	1.00	1.00	1.00	1.00	1.00	1.00

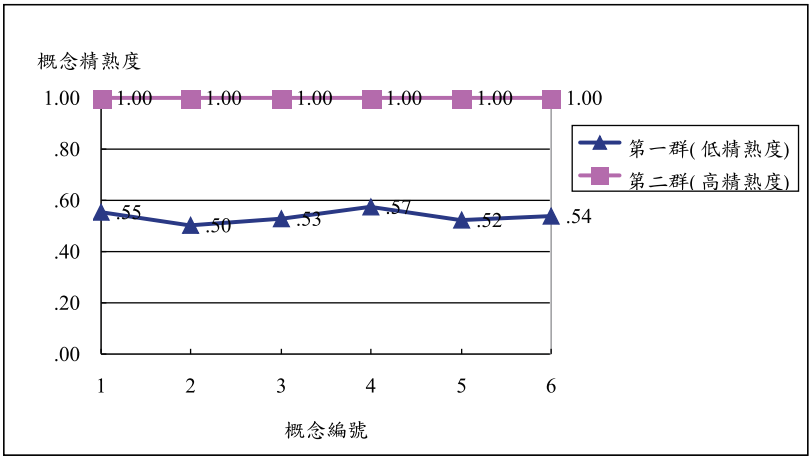


圖3 二群之群中心(概念精熟度)折線圖

二、各群學生概念階層知識結構圖特徵

以下分別就低精熟組和高精熟組所選取之學生，探討概念階層知識結構圖；圖中之圓圈內上方的數字分別代表概念1到概念6，圓圈內下方的小數則代表受試者在該概念之精熟度 d_{na} ，精熟度介於0和1之間，數值越高代表受試者在該概念越精熟。

（一）低精熟組學生的概念階層知識結構圖特徵

低精熟組之受試者32和受試者40的微積分概念階層知識結構圖分別如圖4、圖5所示，根據圖4、圖5，其歸納如下：

1. 就受試者32而言，其分佈在階層三的概念1（微積分基本定理）、概念3（定積分均值定理）和概念5（偶奇函數在定積分上的應用）均為等價關係（equivalent relation）的概念，精熟度為0.52；概念2（積分定理）和概念4（定積分的性質及基本定理）位於第二層，且其與第一層和第三層中的每個概念元素均有聯結關係，顯示概念2和概念4，對受試者32而言，為概念1、概念3和概念5的先備概念知識；概念6（平面上曲線所圍的區域面積求法）位於第一層，與第二層的概念均有聯結關係，顯示概念6對受試者32而言，為概念2和概念4的先備概念知識。

2. 就受試者40而言，其分佈在階層四的概念2和概念3與第三層的概念均有聯結關係，對受試者40而言，概念5為概念2和概念3的先備知識概念；概念5位於第三層，且其與第二層和第四層中的每個概念元素均有聯結關係，顯示概念5對受試者32而言，所需的先備概念知識為概念1和概念4；概念6位於第一層，與第二層的概念均有聯結關係，顯示概念6對受試者40而言，為概念1和概念4的先備概念知識。

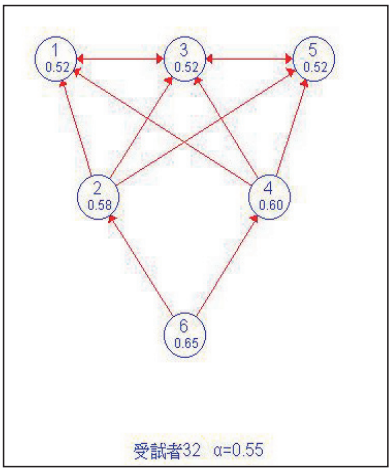


圖4 受試者32之概念階層知識結構圖

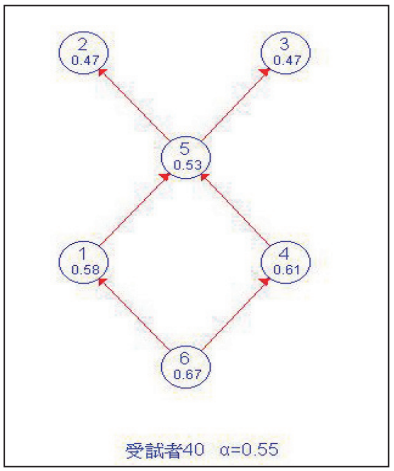


圖5 受試者40之概念階層知識結構圖

（二）高精熟組學生的概念階層知識結構圖特徵

高精熟組之受試者23和受試者43的微積分概念階層知識結構圖分別如圖7、圖8所示，根據圖7、圖8，其歸納如下：

1. 概念階層只有一層但兩兩間互相指向，表示在概念結構上，這些概念彼此間呈現等價關係。從認知心理學的觀點言之，專家的概念結點是關聯密切的。因此，受試者23和受試者43的概念結構和專家相似。
2. 所有概念間皆有互相指向關係，各概念的連結性相當緊密。
3. 各概念的精熟度為1。

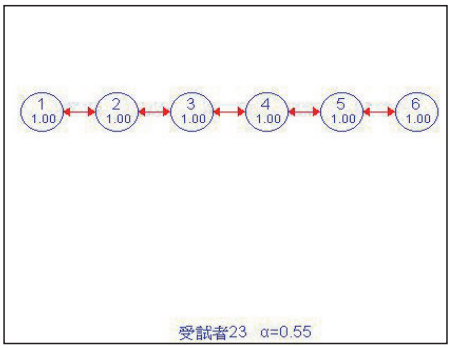


圖6 受試者23之概念階層知識結構圖

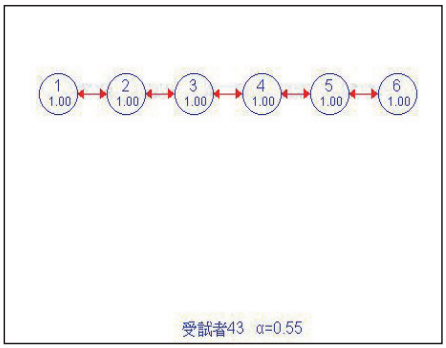


圖7 受試者43之概念階層知識結構圖

綜合以上說明，可發現本研究兩組的概念階層明顯，低精熟組為三個或四個階層，概念間的連結關係緊密；高精熟組為一個階層，概念之間關係密切，互有連結指向關係。教師可依據各群學生概念階層知識結構圖和概念精熟度訊息，有效管理對於瞭解學生學習微積分的知識，以進行有效率的補救教學。

伍、研究限制與結論建議

針對本研究限制和結論建議，具體臚列說明如下：

一、研究限制

1.本研究僅以單一班級的小樣本進行研究，雖然多元計分的概念詮釋結構模式不受大小樣本所限，但小樣本的施測資料，對於微積分精熟度和概念結構的診斷分析結果推論仍有限。

2.本研究測驗僅有10題，所評量的微積分概念亦僅有6個，此乃因為僅以平時考試為實證資料。所以，對於測驗工具的效度以及分析結果一般推論，仍是未來研究改進的方向。

二、結論

1.運用概念階層知識結構圖可進行個別化的認知診斷。本研究以一個班級的人數為樣本進行測驗分析，診斷學生概念精熟與否，藉由分析得到的概念階層知識結構圖，瞭解個別學生概念間的指向和連結性，以作為補救教學之參考依據。

2.透過模糊集群有助於教師進行分組補救教學。經由有效的管理方法，以模糊集群進行分群，本研究受試者可分為低精熟組和高精熟組，發現各群學生有其相似性的特徵：（1）就概念階層而言，不同組別的概念階層數、層次會有所不同；（2）就概念連結而言，各群學生概念間的指向關係有所不同；（3）就概念精熟度而言，高精熟組各概念的精熟度都很高，優於低精熟組。教學者若能將學習認知結果相似的學生集中並進行分組補救教學，即可有效地提昇學習者的學習成效。

三、建議

根據本研究的研究限制與結論，提出下列建議，以提供未來研究之參考

1.本研究以模糊集群的方法進行各群之概念階層知識結構圖的分析與比較，教師可依據受試者測驗結果和概念階層知識結構圖，針對精熟度不佳的概念，進行補救教學。教師亦可根據受試者個別的需求，補強其較不精熟的概念，提昇學生的學習能力，此方法可供教師進行課程設計和補救教學的參考依據。

2.O’Donnell and King（1999）學習發生在社會互動之中，因此老師不是唯一與學生互動、從而學習的人，就算是知識技能或地位相似的同學，也可以互助學習，

因「同儕學習」可說是同學們互相幫助、互相支持、互相需要的一種學習方式。Toping（2001）提出同儕協助學習（peer assisted learning），意指經由地位相似的同伴所提供的主動協助與支援，已習得知識及技能。因此研究者發現高精熟組的學生概念精熟度近似專家，可將低精熟組的學生分組並指派高精熟組學生協助同學課後學習，亦可使高精熟組的學生更加精熟。

3.本研究僅以單一班級的小樣本之平時考試進行研究，雖然多元計分概念詮釋結構模式亦適用於小樣本，但後續研究可進行較大樣本系統性探究，並嚴謹建立微積分評量工具的信度與效度，以獲得更多的認知診斷證據；或進行補救教學前後的效果分析，進行概念階層知識結構圖和概念精熟度的差異比較。

4.本研究以大學微積分進行實證研究，但多元計分概念詮釋結構模式亦可應用於其他具結構性知識的學習領域或科目，而且學校教學者亦可實際應用於教學現場，針對使用後的成效提出建議，作為未來研究改進的依據。

參考文獻

王財印、林坤霖、柯麗蓉、郭柏立（2010）。**技專校院統測數學成績與大一微積分學習成就之相關研究。第二屆科技與數學教育學術研討會**。臺中市：臺中教育大學。

吳柏林（1996）。社會科學研究中的模糊邏輯與模糊統計分析。**中國統計通訊，7**（11），14-30。

吳柏林（2005）。**模糊統計導論：方法與應用**。臺北市：五南圖書公司。

辛靜宜、林珊如、葉秋呈（2005）。五年制專科學生微積分學習動機與策略之初期研究。**南大學報：教育類，39**（2），65-82。

林原宏（2005）。模糊取向的詮釋結構模式之概念結構分析與應用。**教育與心理研究，28**，161-183。

林原宏（2007）。模糊理論在社會科學研究的方法論之回顧。**量化研究月刊，1**（1），53-84。

林原宏、黃國榮（2003）。**FCUT軟體[軟體和說明]**。臺中市：國立臺中教育大學。

林原宏、莊惠雯、易正明（2009）。教師對於學童數學概念之知識管理整合方法－概念詮釋結構模式與分群在時間概念之分析應用。**管理科學與統計決策，6**（3），46-58。

黃馨瑩、林原宏、莊曜遠（2007）。整合集群分析與多元計分次序理論於五年級兒童容量概念的知識結構。**2007第四屆測量統計方法學學術研討會暨臺灣統計方法學學會年會**。臺北市：東吳大學。

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms.* New York: Plenum Press.

Fennema, E., and Sherman, J. (1978). Fennema-Sherman Mathematics Attitude Scales: Instruments designed to measure attitude toward the learning of mathematics by females and males. *Journal for Research in Mathematic Education, 7*, 324-326.

Ferrini-Mundy, J., and Gaudard, M. (1992). Secondary school calculus: preparation or.pitfall in the study of college calculus? *Journal for Research in Mathematics Education, 23*, 57-69.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding groups in data.* NY: John Wiley & Sons.

Lin, Y. H., Hung, W. L., & Huang, K. J. (2006). *CAISM software* [manual and software for CAISM]. Taiwan, Taichung City: National Taichung University.

Lin, Y. H., Yu, M. N., and Wu, B. L. (2006). Fuzzy classification analysis of rules usage on probability reasoning test with multiple raw rule score. *Proceedings of the 2ndWSEAS/IASME International Conference on Educational Technologies* (pp.54-59). Bucharest, Romania.

Lin, Y. H., and Liu, M. H. (2010). Integration of polytomous IRS and S-P Chart in concept. diagnosis of fraction addition based on learning styles. *Proceedings of the 10th WSEAS International Conference on Systems Theory and.Scientific Computation* (pp. 48-53). Taipei, Taiwan.

O' Donnell, A. M. and King, A. (1999). *Cognitive perspectives on peer learning.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stylianou, D. A., Kenney, P. A., Silver, E. A., and Alacaci, C. (2000). Gaining insight into students' thinking through assessment tasks. *Mathematics Teaching in the Middle School, 6*, 136-144.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Journal of Educational and Psychological Consultation, 68*, 249-276.

Warfield, J. N. (1976). *Societal systems planning, policy and complexity.* NY: Wiley.
Zadeh, L. A. (1965). Fuzzy set. Information and Control, 8, 338-353.

附錄 微積分試題

題號	試題內容
1	請詳細寫出微積分基本定理
2	試利用定義求$\int_{-2}^{2}(x+3)dx$之值
3	試求曲線$y=x^3$與x軸所圍成區域在[-2，0]之間的面積
4	請詳細寫出微積分之積分均值定理並證明之
5	已知$\int_0^2 f(x)dx=4$，$\int_2^5 f(x)dx=7$，$\int_3^5 f(x)dx=3$，試算出$\int_0^3 f(x)dx$之值
6	已知$\int_0^2 f(x)dx=4$，$\int_2^5 f(x)dx=7$，$\int_3^5 f(x)dx=3$，試算出$\int_0^3 f(x)dx$之值
7	若$\int_1^3 x^2dx=2f(c)$，求$f(c)$值
8	試求曲線$y=x^2$與$y=\sqrt{x}$所圍成的區域面積
9	已知f是偶函數且$\int_{-a}^{a} f(x)dx=6$，試求$\int_0^a f(x)dx$之值
10	試求曲線$y=\sqrt{x}-1$，x軸，及$x=0$，$x=9$所圍成的區域面積

TIMSS八年級與國中基測物理試題認知成份之探討

盧思丞 / 冷柏原

TIMSS八年級與國中基測物理試題認知成份之探討

盧思丞
台南大學測驗統計所博士生

涂柏原
台南大學測驗統計所副教授

摘要

本文統整了TIMSS科學認知架構與文獻中對於認知成份研究所得到的結果，以總字數、認知需求層次與解題所需概念數三個成份做為認知成份來描述認知成份與TIMSS八年級與國民中學基本學力測驗物理試題難度之間的關係；同時藉由不同能力水準考生在不同試題難度水準上之答對率，說明各種能力水準考生與所精熟的認知需求層次之間的關係，以提供教學與評量的具體參考資料。研究使用TIMSS八年級物理試題共69題，與國民中學基本學力測驗物理試題75題試題，進行認知成份分析，發現三個認知成份能夠解釋TIMSS八年級及國中基測物理試題77%以上的難度變異。

關鍵字：認知成分、TIMSS、國中基測

The cognitive component analysis of TIMSS's and BCTest's physics items

Szu-Cheng Lu
Ph. D Student, Graduate Institute of Measurement & Statistics National University of Tainan
cheni2345ster@gmail.com

Bor-Yaun Twu
Associate Professor, Graduate Institute of Measurement & Statistics National University of Tainan

Abstract

Based on the TIMSS science framework and the literature perspectives, the study used three cognitive components, including number of words, cognitive level, and number of science concepts, to explain the variation of item difficulty. Sixty-nine items of TIMSS science assessment and 75 physics items from the BCTEST were used for this analysis. It was found that 77% of item difficulty variance was explained by the three cognitive components for both TIMSS and BCTEST. And for the students with different ability levels did perform differently on the items which called for different cognitive requirements. The more difficult items did require more advanced cognitive abilities in order to give correct responses.

Keywords: cognitive components, TIMSS, BCTEST

壹、緒論

一、研究動機

為了在國際比較的脈絡中瞭解科學教育的成效，台灣目前參加兩種國際評比：國際數學與科學教育成就趨勢調查（Trends in International Mathematics and Science Study, TIMSS）和學生能力國際評量計畫（The Programme for International Student Assessment, PISA）。其中，TIMSS偏重學科知識學習成就的評量，包含數學和科學兩個內容，調查對象為四年級和八年級兩個年級的學生（Olson, Martin, & Mullis, 2008）。

TIMSS成就測驗是在知識內容與認知能力二維架構下設計發展的，八年級科學的知識內容包含生命科學、化學、物理、地球科學等四個內容領域，四個內容領域各自又區分為三至六不等的若干主題；科學的認知能力則分為知識、應用與推理三個認知領域（Mullis, Martin, Ruddock, O'Sullivan, Arora, & Erberber, 2005）。其目的在於瞭解各國學生是否學會了學校科學與數學課程中預定的學習目標，因此其成就測驗基本上是以學校學科知識內容為架構設計的（Martin, Mullis, Gonzalez, Gregory, Smith, & Chrostowski, 2000）。對TIMSS八年級的試題來說，在49個國家中有80%的國家的課程涵蓋率高於70%，而台灣的涵蓋率更高達91%（Martin, Mullis, & Foy, 2008）。

國中生基本學力測驗（以下簡稱國中基測）是一種學科成就取向的測驗，測驗內容包含國文、數學、社會、英文與自然五個科目。其中自然科內容領域含括物質與能、生命世界、地球環境、生態保育、資訊科技等的學習、注重科學及科學研究知能等，主要在評量學生的科學知識與技能，評量對象為國內九年級的學生。

TIMSS與國中基測之評量目的基本上是相同的，都是評量科學學習成就。TIMSS的題型包含選擇與問答兩部分，比較能評量出學生較完整的科學知識，不被選擇題的選項所局限，而國中基測僅有選擇題。相較於TIMSS的問答題，國中基測係以題組的形式來評量學生對於同一主題的不同概念，雖受限於選擇題的題型，效能可能不如問答題，但仍可以測驗出學生的科學知識與閱讀理解能力。雖然TIMSS與國中基測仍有些差異存在，二者的試題之主軸與核心概念方面仍有許多相同之處，二者皆致力於以科學的方法發展更創新的測驗工具，以有效評量學生對科學概念的理解與應用，並幫助教師或研究者更加認識學生的科學學習成就，且施測的對象年齡相近，約為15到16歲的學生。

台灣學者有關TIMSS的研究，探討的主題涵蓋了：1.了解學生數學及科學學習成就，提供規劃數學及科學課程之參考（張殷榮，2001；林碧珍、蔡文煥，2003）；2.了解學生數學及科學學習成就與家庭背景、學習環境、教師因素等影響因子的關係，並作國際比較分析（張殷榮，2001）；3.與TIMSS 1999及TIMSS 2003之結果作比較；4.了解學生在數學及科學學習成就之趨勢（羅珮華，2000；林鳳女，2000）；

4.了解國際上評量學生學習成就的趨勢與新的評量方法，提供教師參考（李濟國，2001；陳竹村，2003）；5.提昇研究人員資料分析研究能力（劉春初，2004）。

國中基測的研究則大致上包含以下四類：1.實施成效以及計分方式（涂柏原，2007；余民寧、賴姿伶、劉育如，2005）；2.試題品質與造成試題差異功能之原因（盧雪梅、毛國楠，2008）；3.影響國中基測分數的原因，包括學校因素、同儕群體因素及個人和家庭背景特徵因素等（陳吉仲、郭曉怡、李佩倫，2007）4. 國中基測分數與未來表現之相關及預測情形，對教師教學之建議（宋曜廷、許福元、曾芬蘭、蔣莉蘋、孫維民，2007）。

以TIMSS或國中基測當前的研究趨勢來看，主要仍停留在學生、教師、或學校相關的背景變項上之研究，探討其對學生學習成就之影響，對於試題認知成份對試題難度之影響等主題，尚未有相關的研究產出。再者，在認知成份分析領域中，大多研究仍將焦點至於閱讀或數學，且偏重以題幹字數或選項字數為研究之核心，似乎過於強調閱讀負荷量對試題難度之影響。部分學者也曾研究科學認知成份之特徵（Enright, Allen, & Kim, 1993; Rosca, 2004; Yepes-Baraya, 1997），提出一些影響科學試題難度、或受試者成就表現的認知特徵，例如題幹或選項字數、內容領域、誘答項、以及認知歷程……等等，在在顯示當試題所使用的認知特徵不同時，會造成試題難度有所差異，Yepes-Baraya（1997）就將這些特徵稱為「認知成份」，因此認知成份與試題難度之間的關係是一個值得探討的議題。Leong（2006）認為控制試題難度不僅可以有效的測量到測驗本身所要測量的構念，還能減少因為試題較困難或較容易時，導致學生的分佈情形有偏態產生，並且可降低每個試題都進行預試的成本。

Enright、Allen和Kim（1993）進行難度成份分析，指出難度成份有文本與選項特徵、認知需求、知識層次、與科學教育者對試題難度的判斷四類，難度成份可解釋52%的難度變異量。Rosca（2004）使用17個難度成份對1999年TIMSS科學領域104題選擇題，進行難度成份分析，發現可讀性指標、Bloom的認知層次、誘答項的平均字數、答案字數與平均誘答項字數的比率等四個難度成份可解釋的變異量約28%，加上圖表的呈現，五個難度成份約可解釋30%的變異量。張銘秋等人（2010）分析PISA 2006的103題科學試題，結果顯示以知識類別的數量、知識層次、科學能力及字數四個成份預測科學素養試題難度，其認知模式可解釋52%的試題難度變異。

綜上所述可知認知成份與試題難度間的關係極為密切，且研究也顯示認知成份可以解釋難度的變異情形（Enright, Allen, & Kim, 1993; Rosca, 2004; 張銘秋、謝秀月、徐秋月，2010），因此對教育工者者而言是不容忽視的。若事先知道哪些認知成份會影響試題難度的變異，則可針對這些認知成份對學生進行教學，進而改善學生的成就表現。由於TIMSS與國中基測都屬成就測驗，且施測的對象年齡相近，因此研究者採用兩種大型測驗探討認知成份對試題難度之解釋變異情形，並比較兩種大型測驗的認知成份模式是否有差異，期待能歸納出科學試題的認知成份，以幫助教師瞭解學生所需具備的認知需求，提升學生的學習效果和成就。

二、研究目的與問題

本研究以TIMSS八年級與國中基測驗科學試題為主，因科學試題面向過於廣泛，故筆者只以物理的試題為研究對象，使用TIMSS八年級1999到2007年三個年度的物理試題，以及國中基測2001年到2003年三個年度的物理試題，進行認知成份分析研究。藉由試題的成份特徵和試題難度之間的關係，探討TIMSS與國中基測物理試題認知成份模式，並評估認知成份對試題難度變異的解釋力，針對由TIMSS和國中基測所得到的結果，進行比較。根據研究目的，主要的研究問題如下：

1. 依據TIMSS八年級物理測驗內容架構，國中基測物理試題的分配特徵為何？
2. 本研究所提之認知成份分析架構對TIMSS八年級與國中基測物理試題難度變異解釋力為何？
3. 依據認知成份架構，不同物理能力水準的學生其認知運作差異為何？

貳、文獻探討

一、認知成份分析之重要

依據認知取向編製測驗是目前評量趨勢，雖然現代測驗理論對測驗實務有許多影響，但心理計量和認知理論的連結仍然不大（Embretson, 1993）。認知心理學重視變異來源與理論，忽略個別差異，鮮少提及測量穩定性與精確性；而心理計量則重視個別差異，講求測量精確性和穩定性，缺乏效度驗證。因此結合兩取向之優點進行試題認知成份分析，將有很大的學術產出性（Dimitorv & Raykov, 2003; Embretson & Gorin, 2001）。

以往的認知心理學家強調題目的解題歷程，忽略題目內容對試題難度的變異來源，也未探討題目內容之變異與解題歷程的交互作用（丁振豐，1995）；林世華、葉嘉惠（1999）指出認知心理學強調解題歷程，題目是由其包含的認知成份所組成，因此測量的基本單位不是題目，而是所涉及的認知成份。當測量的單位愈詳細，試題所能提供的訊息量也愈豐富；藉由認知成份分析，可協助區辨測驗架構缺點，提供不同形式試題與不同測驗間比較的基礎，對系統性的發展測驗頗有助益（張銘秋、謝秀月、徐秋月，2010）。

洪碧霞、林素微和林娟如（2006）指出認知科學研究拓廣了學習評量內涵的視野，而測量方法學的進步提昇多元而複雜學生表現的可解釋性。美國評量基礎委員會（the Committee on the Foundations of Assessment）提出一項評量的三角架構，來統整學生在學習領域表徵知識和發展能力的認知模式、用以觀察學生表現的作業或情境設計、及針對學生表現資訊進行推論的解釋方法等三者之間的關係。認知與學習模式具有比以往更重的比重，藉由對學生成就目標與學習進展內涵性質之描述，以協助學生在學校進行成功的學習（Pellegrino, Chudowsky, & Graser, 2001）。Dimitorv和Raykov（2003）認為一個試題難度預測模式中認知與程序操作的知識，能夠允許

測驗發展者（1）在試題施測之前，編製已知難度的試題；（2）避免在研究團體進行個別試題的預試，以減少開支成本；（3）使試題難度與學生的能力值配合一致（match）；（4）發展以特定認知和處理特徵為目標的教學策略。

綜合上面的說法，可歸納出使用認知成份的七個優點：（1）藉由認知模式解釋試題難度來源，以便預測新試題之參數；（2）以認知模式預測試題特徵，所以不需對每個試題進行預試；（3）可說明試題的構念效度，且試題的意義與解題歷程的認知模式相互關聯；（4）藉由試題特徵（難度）的重新組合產生新試題於電腦化測驗中；（5）針對受試者能力配對產生的試題難度，促進電腦化適性測驗的發展；（6）不需考慮安全性問題，因為系統中只有認知因素是可知的；（7）可針對認知及處理的特徵，協助教師理解學生學習表現的認知過程，發展更有效的教學策略（Dimitrov & Raykov, 2003; Dimitorv, 2007; Embretson & Daniel, 2008; Embretson & Gorin, 2001）。

因此，倘若對於認知成份瞭解越多，對於測驗編製者而言，可省下繁複、費時的工作，即可操弄所編寫的試題之難易度；對於教學者而言，可更明確知道什麼對學生而言是困難的，進而發展更好的教學策略。故本研究希望能透過認知成份分析將影響測驗試題難度的成份分解出來，以提供教師或測驗編製者的參考。

二、科學難度成份的相關研究

Yepes-Baraya（1996）使用1993年NAEP科學評量研究學生接受測驗時所使用的認知程序。Yepes-Baraya將難度成份分為認知特徵與試題特徵兩種，其中認知特徵又可分為內容知識、推理和說明、以及假設公式和測試；而試題特徵可分為程序比喻資訊及試題形式和閱讀難度。在這些內容下，各含有4~11個試題特徵，共36個試題特徵。與考生面談進行放聲思考，分析答案藉以確定該測驗確實有測得其評量所欲測量的構念。

後來Yepes-Baraya（1997）使用1996年NAEP科學評量所有試題的資料進行分析，包含4年級、8年級及12年級等共三個年級45個試題區塊（block），總共超過500題試題，每個年級有將近2500名受試者作答。這個研究中所使用的難度成份與Yepes-Baraya（1996）的不同，包含了內容知識、含內容與解釋的推理、假設形成與考驗、處理圖表訊息、試題格式和閱讀難度、與實作任務的處理技巧等六類。在這六類之下，又分別含有4至11個特徵，共38個特徵。然而，作者並未使用統計方法來考驗難度成份的有效性，但是作者指出「含內容與解釋的推理」是最重要的特徵。

Enright、Allen和Kim（1993）使用1985年與1986年44題NAEP生命科學的選擇題來進行難度成份分析，他們將難度成份分為四類：文本與選項特徵、認知需求、知識層次、與科學教育者對試題難度的判斷。他們的研究結果顯示這四個難度成份可以解釋52%的變異情形，若結合試題特徵與科學教育者對試題難度的判斷，則可以提升7%至15%的解釋力。在個別難度成份的貢獻上，知識層次的解釋力高達38%。

Rosca（2004）使用1999年TIMSS科學領域的104題選擇題進行分析，其難度成份包含題幹的字數、題幹中的句子數、每一段的字數、每一句的字數、圖表的呈現、Bloom的認知層次、TIMSS的表現層次、TIMSS的科學內容分類、答案的字數、誘答項的平均字數……等共17種。Rosca採多元迴歸與線性邏輯斯模式（Linear Logistic Test Model, LLTM）進行分析，發現可讀性指標、Bloom的認知層次、誘答項的平均字數以及答案字數與平均誘答項字數的比率等四個難度成份可解釋的變異量約28%，加上圖表的呈現只可增加解釋約1%的變異量，五個難度成份約可解釋30%的變異量。

涂柏原、梁恩琪、翁大德與楊毅立（2004）針對2001到2003年國中基測自然科試題共345題，進行試題分析，提出影響自然成就表現的難度因素有五個，分別為試題資訊量、選項異質性、圖文推測程度、轉化程度與知識量。其中知識量表示試題本身所需具備的概念基礎多寡，當所需的概念基礎愈多，試題愈難。

由上面的描述可歸納得到影響試題難度的成份包含了題幹字數、各誘答項字數（或是誘答項平均字數）、可讀性指標、認知層次與圖表呈現、試題資訊量、選項異質性、圖文推測程度、轉化程度與知識量等。因為本研究使用的TIMSS八年級物理試題包含了選擇題和問答題兩種，而國中基測物理試題只有選擇題，故僅以各試題的總字數、認知需求層次、以及解題所需概念數三者，作為本研究進行分析時所用的認知成份。

參、研究方法

一、認知成份分析架構

本研究所用的認知成份係參考Rosca（2004）、Yepes-Baraya（1997）與Enright等人（1993）之難度成份架構、以及涂柏原等人（2004）對國中基測科學試題的分析結果而得到的，認知成份包含各試題的總字數、認知需求層次、以及解題所需概念數等三者。其中認知需求層次包含成七個層次，包含了解簡單訊息、了解複雜訊息、應用科學原則解決問題、科學公式解決問題、解釋現象、辨識熟悉的實驗控制、與設計調查等，整個架構如表1所示。

表1 本研究之認知成份分析架構表

成份	變項的內容	編碼的值
總字數	1-n	1-n
認知需求層次	1.了解簡單訊息	0
	2.了解複雜訊息	1
	3.應用科學原則解決問題	2
	4.科學公式解決問題	3
	5.解釋現象	4
	6.辨識熟悉的實驗控制	5
	7.設計調查	6
解題所需概念數	根據解題所需的概念數而來	0-k

註：研究自行整理歸納得到。n代表試題的總字數，k為所需要的概念數，若為常識則概念數編碼為0。

認知需求七個層次的內容說明如下：

1. 了解簡單訊息：知道科學事實、關係、程序、概念所代表的意義，對於科學名詞、所用單位、測量方法、科學儀器和設備有所了解。
2. 了解複雜訊息：了解物質之特質，科學方法間之相同或相異處，進而比較差異，並對於物質的一些特質和特徵進行區辨、排序。
3. 應用科學原則解決問題：運用圖表或模式來解釋科學概念、原則、結構、關係、程序或科學中的循環（如電子電路），使用基本的科學原則或相關知識，進行觀察與推論。
4. 科學公式解決問題：解釋文章、表格與圖片中與科學概念和原則有關的資訊，並使用科學公式找到量化的解決方法，此步驟涉及科學原則的直接應用或證明。
5. 解釋現象：利用一些科學原則、概念和理論，對觀察的或自然的現象提出合理的解釋、或科學驗證。
6. 辨識熟悉的實驗控制：分析問題以決定相關概念和問題解決步驟，發展與解釋解題策略，且考慮不同因素或概念所造成的影響；利用科學資訊、問題解決經驗、物理條件之改變，對科學實驗進行預測、並進一步檢驗研究假設，獲得證明或證據。
7. 設計調查：能自行設計合適的科學探究活動，藉以回答科學問題或檢驗假設，並自行操控影響實驗結果的變項或程序，探討在不同情境下，可能產生的實驗效果。

二、研究的資料

TIMSS的 1999、2003與2007年等三個年度八年級物理試題根據TIMSS技術報告（Martin, , Mullis, & Foy, 2008）將主要分成六項內容領域，分別為物理現象（包含物質分類與特性、物質的物理狀態與變化）、能量轉換（例如熱的傳播與溫度）、光與聲音、電與磁力、及力與運動。以台灣來說，TIMSS所公開釋出的試題共有69題，筆者根據上述的認知需求層次以及六項內容領域將所有69個試題歸納整理得到如表2的雙向細目表。這69個試題包括1999年的18題、2003年的25題和2007年的26題，題型包含選擇題與問答題兩種。而台灣在TIMSS 1999、2003與2007年三個年度的測驗中，參與測驗的學生人數分別為5889人、5379人與4046人。國中基測資料來自2007到2009年的自然試題，因研究之需要筆者只取物理試題共75題，三個年度分別為23題、25題與27題，題型均為選擇題。筆者亦依照與TIMSS相同的分類方式，得到表3的國中基測試題之雙向細目表。國中基測試題資料是由「國中基測委員會」所提供，每個年度每試均隨機抽取5000人，三年度總共為30000名考生。

表2 TIMSS物理試題之雙向細目表

	物理現象	能量轉換	光	聲	電與磁	力與運動	總題數
了解簡單訊息	1	2	2	0	1	4	10
了解複雜訊息	2	1	4	0	4	1	12
應用科學原則解決問題	2	4	3	0	0	4	13
科學公式解決問題	6	0	1	1	1	4	13
解釋現象	4	2	4	1	2	0	13
辨識熟悉的實驗控制	2	0	0	1	0	1	4
設計調查	3	0	0	0	1	0	4
總題數	20	9	14	3	9	14	69

註：研究者自行整理歸納得到，橫列是認知需求層次，縱行是物理內容領域。

表3 國中基測物理試題之雙向細目表

	物理現象	能量轉換	光	聲	電與磁	力與運動	總題數
了解簡單訊息	4	0	2	0	1	0	7
了解複雜訊息	2	1	1	3	1	2	10
應用科學原則解決問題	6	0	2	2	8	6	24
科學公式解決問題	2	3	0	0	3	8	16
解釋現象	2	2	0	0	2	3	9
辨識熟悉的實驗控制	1	1	2	0	5	0	9
設計調查	0	0	0	0	0	0	0
總題數	17	7	7	5	20	19	75

註：研究者自行整理歸納得，橫列是認知需求層次，縱行是物理內容領域。

三、認知成份編碼說明

研究中所用的三個認知成份分別為試題總字數、認知需求層次、與解題所需概念數等三者，其意義及說明如下：

1.字數：包含題幹與每個選項的總字數，若是有圖、表，則圖表內的字數也列入計算。

2.認知需求：認知需求層次共有七個層次，包括了解簡單訊息、了解複雜訊息、應用科學原則解決問題、科學公式解決問題、解釋現象、辨識熟悉的實驗控制與設計調查。

3.解題所需概念數：在解決科學問題時，學生會使用到一些科學概念，研究者根據試題的內容、敘述，歸納出學生需要的科學概念，包含0個科學概念（一般知識，不需科學概念解題）到4個以上的科學概念。

筆者根據表1所列編碼的值，對研究中所探討的每一個試題進行這三個認知成份的編碼。總字數這個成份直接以試題之總字數表示之；認知需求部分，因為是有次序性的，所以依照0到6的編碼方式將其進行編碼，故了解簡單訊息編碼為0，了解複雜訊息編碼為1，應用科學原則解決問題編碼為2，科學公式解決問題編碼為3，解釋象現編碼為4，辨識熟悉的實驗控制編碼為5，以及設計調查編碼為6；解題所需概念數中，分0個科學概念到4個以上的科學概念。另外，試題難度部分，則是經由單參數試題反應理論求得（即Rasch model），故試題難度之範圍大致上為-4.0到4.0之

間，當難度值為負值時表試題趨於簡單，難度值為正值時表試題趨於困難，而難度值接近0表試題難度適中。

實際編碼的方式，以底下四個TIMSS的物理試題作為編碼示例（見圖1至圖4）。圖1是難度值小於-1.0的範例試題，總字數50個字；難度值-1.206；主要是評量學生是否能根據題幹的說明，找出有作功的圖示，因為已經給作功的定義，所以學生不用再另外尋找相關的科學概念或知識，所以屬於認知需求層次的第一個層次「了解複雜訊息」，故編碼為0，其解題所需的概念數則編碼為0。

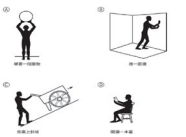
範例一（b＝-1.206， 50字，層次一-了解複雜訊息， 0個概念）	
當一個物體沿著作用力的方向移動時，這就是對它作了功。有一個人作了下圖所示的各種活動，在哪一個圖中的人有作功？	

圖1 難度值小於-1.0之試題範例

圖2是難度值介於-1.0~0.0的範例試題，該題的總字數為34個字，難度值為-.768；以題幹來看，是在評量學生是否能夠比較不同顏色之間對光的反射差異，因此在認知需求層次分類是屬於第二個層次「了解複雜訊息」，故編碼為1；而解題所需的概念則是色光的反射概念，只需要一個解題概念，所以編碼為1。

範例二（b＝-0.768， 34字，層次二-了解複雜訊息， 1個概念）
欲將建築物的牆壁塗上油漆以盡量反射光線，應該用什麼顏色的油漆？ A.白　B.紅　C.黑　D.粉紅

圖2 難度值介於-1.0~0.0之試題範例

圖3則是難度值介於0.0~1.0的範例試題，總字數為52個字；難度值為0.553；以題幹來看，是希望學生利用圖表資料、與科學概念或原則解決問題，並使用公式獲得一個量化的答案，所以研究者將之歸類為第四個層次「科學公式解決問題」，故編碼為3；而解題所需的概念包含兩個，一為距離和時間，二為速率，所以編碼為2。

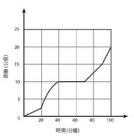
範例三、（b＝0.553，52字，層次四-科學公式解決問題， 2個概念）	
瑪莉騎腳踏車的途中發生爆胎，她修補後立刻繼續向前騎。下圖表示她的整個騎車過程。瑪莉花多少時間修補車胎呢？ A. 20分鐘　B. 30分鐘　C. 40分鐘　D. 70分鐘	

圖3 難度值介於0.0~1.0之試題範例

範例四（b = 1.854，308字，層次六-辨識熟悉的實驗控制，4個概念）

某個科學班級要找出裝有汽水的罐子的密度。全班分為四組來完成這項工作。每組各取得一罐汽水。當各組完成任務後，他們發表的實驗結果如下表所示。

	A組	B組	C組	D組
密度(g/mL)	1.04	0.04	2.77	1.05

全班同學很驚訝，每組對罐子的密度所測得的結果都不同。他們檢視每組所用來測汽水罐質量和體積的方法。表2表示每組如何測得汽水罐的體積。

表2：體積

組別	方法	體積(ml)
A	我們把燒杯加水到1400 mL 的刻度。我們把沒打開的罐子放入燒杯裡。罐子沈下去。然後水位到達1776 mL。	376.00
B	我們把燒杯加水到1400 mL 的刻度。我們把空罐開口朝下，垂直放入燒杯裡。我們用一枝鉛筆把罐子壓入水中。然後水位到達1776 mL。	376.00
C	我們把燒杯加水到1600 mL 的刻度。我們把空罐放到燒杯裡，開口朝上。我們把罐子壓入水中，看到有氣泡從罐子裡冒出來。當不再有氣泡冒出時，罐子沈到燒杯底，此時水位到達1605 mL。	5.00
D	我們把罐子打開，用量筒測量罐中汽水的體積。	371.00

B 和 C 兩組試著測量汽水空罐的體積。請解釋為什麼他們的結果不同。

圖4 難度值大於1.0之試題範例

圖4範例試題的總字數為308個字；難度值為1.854；根據試題說明，所要評量的是測量物體體積之方法，以及對於物體質量、密度和體積間關係之瞭解程度，學生必須先知道物體之質量、密度和體積三者之間的關係，並判斷何者是正確的測量方法，以及當實驗結果不同時，是何種因素造成的。B組和C組最大的差異是空罐放入燒杯的方式（一個瓶口朝上、另一個瓶口朝下），所以在計算體積時，B組的學生會將瓶子內的空氣體積也計算進去，而C組則只計算了鐵的部分；因此，是在分析問題以決定解決步驟時，考慮不同因素或概念所造成的影響，且利用問題解決經驗、條件之改變，對科學實驗進行預測、進一步獲得證明或證據，故屬於認知需求中的第六個層次「辨識熟悉的實驗控制」，故編碼為5；解決問題所需概念為四個，包含體積、質量、密度、以及三者間的關係（質量/體積＝密度），故編碼為4。

四、資料分析

本研究根據筆者所自行得到的TIMSS八年級與國中基測之物理試題的內容架構，並以多元迴歸方法進行分析，探討總字數、認知需求層次與解題所需概念數，對由Rasch模式所估計得到的試題難度（b值）之解釋力為何，並進一步說明各種能力水準的學生在各種難度水準試題上面的表現情形以及試題之特性。

肆、研究結果與討論

一、物理試題認知成份編碼一致性

由於認知需求層次與解題所需概念數的編碼會受主觀判斷影響，導致不同的人有不一樣的編碼，為確定本研究的編碼是合理的，因此筆者除自行編碼以外，另外

邀請了嘉義大學科學教育領域專家、與現任自然科教師一同進行編碼的工作，其中科教領域專家只對認知需求部分進行編碼。三人完成編碼之後，進行一致性分析分析以確定本研究的編碼是否合理，一致性分析的結果如表4所示。相較於Enright、Allen、與Kim（1993）的50%~79%，本研究的一致性百分比無論是TIMSS的物理試題，或是國中基測的物理試題，皆在80%以上，顯示認知成份編碼之一致性是可接受的（請參見表4）。因此，本研究後續根據各個試題的編碼所進行的難度分析，所得到的結果應是可以信賴的。

表4 TIMSS與國中基測物理試題認知成份編碼一致性百分比

編碼者	TIMSS		國中基測	
	認知需求層次（完全一致）	概念數（完全一致）	認知需求層次（完全一致）	概念數（完全一致）
1 vs. 2	56（81%）	61（88%）	60（80%）	64（85%）
1 vs. 3	57（83%）	--	--	--
2 vs. 3	59（86%）	--	--	--

註：表中的數字是編碼一致的題數，總題數TIMSS為69題，國中基測為75題；編碼者中的1是筆者，2為自然科教師，3為科教領域的專家。

二、物理試題內容架構的分配特徵

依據TIMSS的內容架構（參見表2），以物理現象的試題數最多，約30%，而聲的試題最少約4%；若以認知需求層次來看，層次六辨識熟悉的實驗控制與層次七設計調查所佔的試題較少為6%，其餘層次的試題分佈較平均，約佔總題數的14%~19%。

相較於TIMSS的內容架構，國中基測物理試題（參見表3）中，物理現象、電與磁、以及力與運動所佔之比率較高，介於23%~27%間，而能量轉換、光、及聲之試題所佔比率只有7%~9%，與TIMSS試題之分配有差異存在；在認知需求層次上，國中基測與TIMSS架構有些微差異，完全沒有第七個層次設計調查，且在第三層次應用科學原則解決問題、以及第四層次科學公式解決問題所佔的比率最多，分別為32%與21%，其餘所佔比率則介於9%~13%之間。

三、物理試題難度變異的解釋力

本研究所提出之認知成份包含總字數、認知需求層次與解題所需概念數三個，表5和6分別是TIMSS與國中基測物理試題難度與認知成份間之相關矩陣；不論是TIMSS或國中基測，物理試題難度與認知需求層次的相關都是最高的，均在.800以上，其次是解題所需概念數、以及總字數，且三個認知成份與試題難度之相關都達顯著。

不論TIMSS或國中基測，均可用總字數、認知需求層次與解題所需概念數三個變項來預測效標變項，且三者之間的相關介於.200~.560之間，預測變項間的共線性問題應該不顯著，因此對於預測變項對效標變項之解釋力不會有影響。對TIMSS物理試題而言，如表7所示，三個預測變項的解釋力為80.2%（調整後79.3%）；對國中

基測物理試題而言，三個預測變項的解釋力為78.5%（調整後77.6%），但如表8中顯示，總字數並未達到.05之顯著水準，因此，只有兩個預測變項時，其解釋力為78.2%（調整後為77.6%）。

表5 TIMSS八年級物理試題難度與認知成份間之相關矩陣（題數＝69）

	試題難度	總字數	認知需求	概念數
試題難度	1			
總字數	.398**	1		
認知需求	.800**	.295*	1	
概念數	.767**	.292**	.555**	1

*p < .05. **p < .01

表6 國中基測物理試題難度與認知成份間之相關矩陣（題數＝75）

	試題難度	總字數	認知需求	概念數
試題難度	1			
總字數	.344**	1		
認知需求	.859**	.337*	1	
概念數	.608**	.159	.497**	1

*p < .05. **p < .01

TIMSS八年級或國中基測物理試題之迴歸方程式如下：

$\hat{Y}_{TIMSS\text{試題難度}}$=.002×(總字數)+.214×(認知需求層次)+.298×(概念數)-.770

$\hat{Y}_{\text{國中基測試題難度}}$=.353×(認知需求層次)+.197×(概念數)-1.246

由迴歸方程式看出，TIMSS八年級或國中基測之預測方程式係數雖有不同，但仍可以發現除總字數以外，認知需求層次與解題所需概念數對試題難度的影響是很大的，當變動一個單位時，難度變動的幅度介於.197~.353之間。

表7 TIMSS八年級物理試題難度參數多元迴歸方程式係數摘要表

預測變項	非標準化係數	標準誤	標準化		
（常數）	-.770	.098	--	-7.831***	--
總字數	.002	.001	.119	2.013*	3.837
認知需求層次	.214	.029	.504	7.510***	4.756
解題所需概念數	.298	.044	.460	6.844***	5.415

註：整個模式的解釋力為R^2 = .802（調整後為.793）。

*p < .05. ***p < .001

表8 國中基測物理試題難度參數多元迴歸方程式係數摘要表

預測變項	非標準化係數	標準誤	標準化係數	t	條件指數
（常數）	-1.246	.116	--	-10.729***	--
總字數	.001	.001	.063	1.082	4.005
認知需求層次	.353	.033	.718	10.812***	5.306
解題所需概念數	.197	.052	.241	3.798***	7.414

註：整個模式的解釋力為R^2 = .785（調整後為.776）。

***p < .001

四、不同物理能力水準學生認知運作之差異

為進一步探討國中基測中不同物理能力水準的學生之認知運作差異情形，同上的作法，藉由Rasch模式得到基測物理試題難度（b值）、以及受試者的能力值（θ值）。當難度值為負表試題趨於簡單，難度值為正表試題趨於困難，而難度值接近0表試題難度適中；受試者能力部份，也是同樣的情形，當能力值為負表受試者能力較低，難度值為正表受試者能力較高。

依據本研究所得之認知成份架構，研究者根據試題難度分佈情形將試題依難度分成三個水準（表9），水準一為試題難度小於-.5的試題，共18題，此為較簡單的試題，平均難度為-.905，總字數平均為66.831，認知需求之平均為.998，而概念數之平均為.556；水準二則是試題難度介於-.5到.5的試題，有42題，這是屬於中等難度的試題，平均難度為.035，總字數平均為81.789，認知需求之平均為2.524，概念數之平均為1.293；水準三為試題難度大於.5的試題，共15題，是指難度較高的試題，平均難度為.989，總字數平均為89.071，認知需求之平均為4.203，而概念數之平均為2.134。

若由總字數來看，發現水準一試題平均字數只有66字，比水準二和三的試題來的少，而水準二和三之字數平均差異不大。認知需求部分，水準一試題的認知需求平均是.998，故該水準試題要求認知需求必須在層次一的了解簡單訊息或層次二的了解複雜訊息之間；而對水準二試題而言，其認知需求需在層次三的應用科學原則解決問題和層次四的科學公式解決問題；至於對水準三的試題來說，題認知需求必須在層次五解釋現象與層次六的辨識熟悉的實驗控制。以概念數來看，其各水準的平均分別為.556、1.293、以及2.134，因此難度水準一的試題最容易的，只要具備0或1個概念數，就可以答對該層次的試題；而難度水準三試題最難，概念數要求為2個以上才可。

表9 不同水準試題難度及認知成份平均數對照表

試題水準	試題數	難度	總字數	認知需求	概念數
一（b<-.5）	18	-.905	66.831	.998	.556
二（-.5<b<.5）	42	.035	81.789	2.524	1.293
三（b>.5）	15	.989	89.071	4.203	2.134

研究者進一步將受試者能力依據分配情形分成四個水準（表10），亦即將能力低於-1.0的學生歸類為低於基礎水準、-1.0~0.0之間為基礎水準為、0.0~1.0之間為精熟水準、而能力值在1.0以上者則屬於進階水準的學生。以低於基礎水準之考生來說，不論在哪一個難度水準其答對率均在.40以下，且在難度水準二和三答對率都不到.20；而基礎水準的考生，在試題水準一的答對率為.60，且在試題水準二和三之答對率均低於.40，與研究者預期相同；若以精熟水準考生來看，在試題水準一的答對率為.85，而試題水準二的答對率為.591，試題水準三的答對率為.321，也符合預期的結果；相較於前三個水準的考生，進階水準的考生是能力最好的考生，所以在試題水準一和水準二的試題以上的答對率都在.85以上，且在最困難的試題水準上答對率

也最高，其答對率為.70。

表10 不同能力水準考生在不同難度水準試題之答對率

	試題水準		
能力水準	一（b<-.5）	二（-.5<b<.5）	三（b>.5）
低於基礎水準	.315	.176	.109
基礎水準	.606	.372	.204
精熟水準	.855	.591	.321
進階水準	.965	.871	.703

綜合以上的分析（根據表9和表10），在基礎水準之考生，所具備的認知需求層次為一和二，亦即了解簡單訊息與了解複雜訊息；精熟水準之考生，其具備的認知需求層次為三和四，即應用科學原則解決問題與科學公式解決問題；而進階水準的考生則是具備認知需求層次五和六，解釋現象與辨識熟悉的實驗控制。此一結果可提供教師在教學上的參考，對於測驗編製者也具有一定的意義。

伍、結論與建議

本研究根據試題的內涵與認知成份分析的觀點，來對TIMSS八年級物理試題進行難度分析，發現總字數、認知需求、以及解題所需概念數等三項認知成份解釋了79.3%的難度變異。而在國中基測的試題方面，總字數、認知需求與解題所需概念數等三項認知成份對於試題難度的變異，解釋力為77.6%，但因總字數之迴歸係數不顯著，故將該變項刪除，其所得之解釋力仍為77.6%，因此最後只以認知需求及解題所需概念數來解釋國民中學基本學力測驗物理試題難度變異。

在考生能力水準與試題水準之比較，也可以發現基礎水準的考生只能了解簡單訊息與了解複雜訊息，精熟水準的考生則是具備應用科學原則解決問題與科學公式解決問題兩個認知需求，而進階水準之考生則是有解釋現象與辨識熟悉的實驗控制之認知需求。因此，在教學上，教師可以針對不同能力水準之考生，提供不同的教學輔助，以使學生的能力水準能夠有所提升。

參考文獻

丁振豐（1995）。幾何圖形類比推理題目內容成份分析之研究。**臺南師院學報，28**，83-114。

余民寧、賴姿伶、劉育如（2005）。國中基本學力測驗實施成效之初步調查：**學校的觀點。教育與心理研究，28**(2)，193-217。

宋曜廷、許福元、曾芬蘭、蔣莉蘋、孫維民（2007）。國民中學學生基本學力測驗的回顧與展望。**教育研究與發展期刊，3**(4)，29-50。

李濟國（2001）。影響高中學生物理學習成就因素之探討。**科學教育月刊，240**，21-30。

林世華，葉嘉惠（1999）。數字系列完成測驗試題認知成分分析之研究。**教育心理學報，31**，139-165。

林碧珍、蔡文煥（2003）。**我國國小四年級學生在國際教育成就2003試測的數學成就表現**。九十二學年度師範學院教育學術論文發表會論文集。台南：國立台南師範學院。

涂柏原、梁恩琪、翁大德、楊毅立（2004年3月）。**國中基測自然科試題分析研究**。「科技化測驗與能力指標評量國際研討會」發表之論文。國立台南師範學院。

涂柏原（2007）。國中生基本學力測驗量尺分數轉換之實徵研究。**教育研究學報，41**(1)，61-77。

洪碧霞、林素微、林娟如（2006）。認知複雜度分析架構對TASA-MAT六年級線上測驗試題難度的解釋力。**教育研究與發展期刊，2**(4)，69-86。

陳吉仲、郭曉怡、李佩倫（2007）。影響國中基本學力測驗分數的因素之分析。**教育政策論壇，10**(4)，119-142。

陳竹村（2003）。TIMSS 1999 台灣名列前茅極可能因素探討。**教育研究月刊，108**，133-146。

張殷榮（2001）。我國國中學生在國際測驗調查中科學學習成就影響因素之探討。**科學教育，244**，5-10。

張銘秋、謝秀月、徐秋月（2010）。PISA科學素養之試題認知成份分析。**課程與教學季刊，13**(1)，1-20。

劉春初（2010）。TIMSS-R架構與DEA分析法的運用—以台灣地區國民中學學校經營效率南北地區比較為例。**臺東大學教育學報，15** (1)，167-184。

盧雪梅、毛國楠（2008）。國中基本學力測驗自然科之性別差異和差別試題功能（DIF）分析。**測驗學刊，55**(4)，725-759。

羅珮華（2000）。第三次國際數學與科學教育成就研究後續調查之抽樣設計。**科學教育月刊，235**，14-20。

Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement, 31*(5), 367-387.

Dimitrov, D. M., & Raykov, T. (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavior Research, 38*(1), 1-23.

Embretson, S. E. (1993). Psychometric Models for Learning and Cognitive Processes.In Frederiksen, N., Mislevy, R. J. & Bejar, I. I. (Eds), *Test Theory for a New Generation of Tests* (pp. 120-150). Hillsdale, N J: Elbaum.

Embretson, S. E., & Daniel, R. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50*(3), 328-344.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

Enright, M. K., Allen, N., & Kim M. (1993). *A Complexity Analysis of Items from a Survey of Academic Achievement in the Life Sciences* (ETS Research Report 93-18). Princeton , N.J.: Educational Testing Service.

Leong, S. C. (2006, May). *On varying the difficulty of test items.* A paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). *TIMSS 2007 international science report: Findings from IEA's trends in international mathematics and science study at thefourth and eighth grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S.J., et al. (2000). *TIMSS 1999 international science report.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber,E. (2005). *TIMSS 2007 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Pellegrino, J. W., Chudowsky, N., & Glaser R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment.* Center for Education, National Research Council.

Rosca, C. V. (2004). *What makes a science item difficult? A study of TIMSS-R items using regression and the Linear Logistic Test Model.* Unpublished doctoral dissertation, Boston College, Boston.

Yepes-Baraya, M. (1996, April). *A cognitive study on the National Assessment of Educational Progress (NEAP) science assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York. Retrieved.from ERIC database. (ED404343)

Yepes-Baraya, M. (1997, March). *Lessons learned form the coding of item attributes for the 1996 NAEP science assessment G4 results.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago. Retrieved from ERIC database. (ED409356)

不同標準設定方法之比較研究

謝名娟 / 謝進昌

不同標準設定方法之比較研究

謝名娟
國家教育研究院助理研究員

謝進昌
國家教育研究院助理研究員

摘要

在本篇研究中，比較書籤標定法與Yes/No Angoff標準設定方法在設定切斷分數上的差異，並使用TASA 2009年英文科來進行研究。研究結果顯示，兩種標準設定方法所做出的決斷分數略有不同，而成員們覺得Yes/NoAngoff法執行起來較為簡單，而最終決定的切斷分數也較符合預期，書籤標定法對於成員而言，需要克服的技術層面較多，尤其是在面臨題本難度的順序，和心中期望的難度順序不一致時，往往會造成他們很難找到確切的書籤位置，但是，書籤標定法所需的執行時間較短，做出的切斷分數結果，反應在2009年的實徵數據上，發現中間部分的學生比例明顯低於基礎以下與進階的學生，許多成員覺得做出切斷分數呈現M型分布，較能反應台灣學生學習成就的真實情形。

關鍵字：Yes/No Angoff方法、書籤標定法、標準設定

The Comparison of Different Procedures in Standard Setting.

Ming-Chuan Hsieh
Assistant Research Fellow, Research Center for Testing and Assessment, National Academy for Educational Research
mhsieh@mail.naer.edu.tw

Jin-Chang Hsieh
Assistant Research Fellow, Research Center for Testing and Assessment, National Academy for Educational Research

Abstract

This study compared two standard setting methods- bookmark and yes/no Angoff method, using 2009 TASA 2009 English as research subject. It is found that the resulting cutoff points from these two methods are somewhat different. Judges regarded that yes/no Angoff method is easier to implement and the resulting cutoff points are more close to their expectation. Comparatively, there are more difficulties need to be solved for the bookmark procedure, especially when the item difficulty order does not follow judges' expectation. When this kind of situation happened, it is very hard for the judges to place the bookmarks. However, the needy time for implementing the bookmark procedure is much shorter. The resulting cutoff point from bookmark method shows that most Taiwanese students centered at the below proficient or advance level, and not many in the middle level. Many judges regarded this kind of M shape distribution of student performance more closely reflected the reality of education condition in Taiwan.

keywords: yes/no Angoff, bookmark, standard setting

壹、研究背景及目的

臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement，簡稱TASA）的建置目的，為了解國小四年級、六年級、國中二年級、高中、高職二年級學生之學習成就表現，並探討學生學習成就上之表現差異與學習上變遷之趨勢，進而檢視當前國家教育體制與政策實施之成效。在此目的下，其中最重要的工作項目之一，即是評定或描述學生目前的學力表現，是否達到政府決策者所預期的程度或標準。所謂標準設定，其意涵在於測驗研發者或使用者根據測驗結果，將不同學生的表現加以分類，並確保某些符合通過資格的應試者能通過檢定或達到某一成就水準。

本研究針對2009年臺灣學生學習成就評量資料庫國民小學英文科的學習成就標準設定，並設定基礎、精熟、進階三個水平，為了瞭解標準設定的結果是否會受到不同設定方法的影響，本研究在TASA的英文科的標準設定採用兩種設定方法-書籤標定法與Yes/No Angoff標定法，來比較不同設定方法間所產生之決斷分數的差異。

貳、研究方法

本研究針對2009年臺灣學生學習成就評量資料庫國民小學英語文，來進行學習成就標準設定，並設定基礎、精熟、進階三個水平，為了瞭解標準設定的結果是否會受到不同設定方法的影響，本研究採用兩種設定方法，來比較設定方法間所產生之決斷分數的差異。

會議的第一天，進行的方法為Yes/No Angoff 方法，在這個方法中，成員們必須針對題庫中的每一個題目，來判斷是哪一種水平的邊緣受試者（borderline students，即為此水平中，最低成就的學生）才能夠答對。如果覺得基礎水平邊緣受試者可以答對此題，則在此題的基礎水平空格上打勾，同樣的，也需考量是否精熟與進階的邊緣受試者是否可以答對此題。

會議的第二天的成員和第一天相同，而進行的方法為書籤標定法，在這個方法中，先把題庫裡所有的題目由簡單到難依序排列（ordered item booklets，簡稱OIB），每一位標準設定成員檢視完試題之後，則須在OIB中依序放置基礎、精熟、進階的書籤。而放置書籤的原則，則以各水平中，最低成就的學生，應具備的英文科知識為主。然後對照OIB中的各個試題內容，找出哪一題為該水平的邊緣受試者，可以有67%的機率能正確作答的題目，並把書籤放在那個試題位置。

在這兩種方法中，每一輪判別完，研究者會提供各個成員所判別的決斷分數位置，以供全體成員討論。這樣的過程會重複三次，而成員們也可以在每一次過程中，修正自己的決定。第三輪之後，依據成員們在每個水平中所判定分數的平均數，計算最終的決斷分數。

兩種方法所判別的決斷分數，將會進行比較。此外，每一輪標準設定後，會對成員們進行問卷調查，以瞭解哪一種方法最符合成員們心目中理想的決斷分數。以下，就這兩種標準設定方法，加以詳加描述：

一、Yes/No Angoff 方法

Angoff方法為William Angoff（1971）所提出，Angoff方法使用上很簡單易懂，而且能夠輕易為不同形式的題型設定決斷分數。在最原始的Angoff方法中，標準設定成員，必須對OIB 中的每一個題目進行判別，並決定邊緣受試者（minimally competent examinee），有多高的機率，可以達對這個題目。然而，若是使用原先的Angoff 方法，除了要對每一題進行判斷之外，成員們還須思索每一題的答對機率，當題本內的題目很多時，這種方式就變得較不合適。

Angoff方法有許多修定的版本，其中一個廣泛使用的版本為Impara 與 Plake（1997）所設計的Yes/No Angoff 方法。這個方法，和原先設計雷同，必需要對題本中所有的題目進行判斷，但不同的點是，此方法不用寫出邊緣受試者答對題目的機率，而是直接寫下邊緣受試者是否能夠答對此題。如果邊緣受試者可以答對此題，則在這個題目的表格上寫下「是」，如果不能答對，則寫下「否」。這種較為直觀的判斷，減少了原先Angoff方法的執行難度。

Yes/NoAngoff 標準設定法的實際操作概念及流程，大致是如下所述。首先，研究者會事先提供每位標準設定成員一本英語文試題卷，而每頁的試題內容大致如下表1的範例所示，包含有題目內容、選項、答案及評測項目等。

表1 英語文某一範例試題格式

科別	年級	編號	答案
英語文	六	1	2
題目	【聽到】：book, book, book		
	【題目】：① good ② book ③ put		
能力指標	1-1-3能聽辨課堂中所習得的詞彙。		

註：本題目為經修訂之範例試題，非正式試題。

而後針對題本中的每一個試題，判斷是否基礎水平的邊緣學生，可以答對該題。若是不能，則判斷是否精熟水平的邊緣學生，可以答對該題，若是不能，則判斷是否進階水平的學生，是否可以答對該題。若是成員覺得此題很難，連進階水平的邊緣學生均無法答對，則可判定該題為超出進階水平的能力的題目。

每一位標準設定成員，逐題判斷之後，研究者將每一位成員的填答數據，輸入電腦軟體BILOG-MG中，並依據試題在題庫中的a, b, c 參數，固定試題參數來計算每個成員在基礎、精熟、進階三個水平上所估的能力值。將BILOG-MG估算出每一位成員在各個水平上所估的能力值加以平均，即可算出各水平的決斷分數。

第二輪執行的任務大致是重覆第一輪的動作，但差別在於研究者會提供一些回饋訊息，以作為成員參考，包含第一輪標準設定後，其它成員（與自己）所對各題的判斷、決斷分數。成員即依據回饋訊息，分成小組來討論第一輪所設定通過分數的適切性與聆聽其它成員發表自己對題目的判別依據，進行第二輪的設定，並再次對各題進行判定。第二輪標準設定結束後，所提供的回饋訊息為其它成員（與自己）所對各題的判斷、決斷分數，與依此決斷分數，使用2009的學生真實成績，來

計算各水平的通過人數百分比。成員再次依據這些回饋訊息，重新進行每一個題目的判定。

第三輪執行的任務亦是雷同於前兩輪，但差別在於此輪任務是由成員獨自完成，不能相互討論，最後，研究者即是根據標準設定成員於第三輪所判定的成果，經換算後，以得基礎、精熟與進階水平的正式通過分數。

二、書籤標定法

書籤標定法（bookmark）（Lewis, Mitzel, & Green, 1996）的設計不僅能避免原始Angoff法逐題檢視、評定之疲勞、繁瑣干擾，同時，也能輕易的加入選擇題型與建構反應試題來進行標準設定，而在執行這個方法時，研究者會事先提供每位標準設定成員一本已經藉由IRT預先計算出的難度值，並加以由簡單至困難排序的英語文試題卷（OIB）。由於TASA的所有技術報告均以試題反應理論三參數進行分析，為了使標準設定的結果能應用於TASA，對於難度值的計算，是採用三參數試題反應模式來進行估計。

標準設定成員逐一檢視OIB中所有試題後，將基礎、精熟、進階三個書籤逐一放置於某一試題上，即完成三個水平通過分數的設定，然而，放置書籤位置時，須依照標準表現描述對於各水平學生的描述進行推理和想像，並推測該水平中，程度最差的那一位學生，具有67%的機率能該答對該題，並把書籤放置在那個試題的位置上，記錄在記錄表中。例如：若將基礎書籤放置在第16題、精熟書籤是放置在第40題、進階書籤是放置在第70題。即可將所有學生區分出基礎以下、基礎、精熟與進階等四個能力區塊。接續，研究者會將各水平書籤所放置對應的試題，搭配該題的已知的試題難度、鑑別度及猜測度參數，於反應機率（response probability）0.67（Huynh, 2006）下，進行能力值的轉換，如此，即是該水平的通過能力。和Yes/No Angoff方法相同，成員總共要進行三輪來放置書籤，第一輪放置後，研究者會提供一些回饋訊息，以作為成員參考，並進行討論，以了解設定通過分數的適切性，而後，成員再依據前述的原則，重新放置第二輪的基礎、精熟、進階三個書籤的位置。第二輪後提供回饋訊息，進行成員討論後與第三輪的標準設定，但提供的回饋訊息，可略有不同，以了解不同回饋訊息，對於成員設定分數的影響。

最後，研究者即是根據標準設定成員於第三輪所放置於基礎、精熟、進階三個書籤的位置，經換算後，以得基礎、精熟與進階水平的正式通過分數。

三、標準設定成員

本研究發出標準設定成員徵求訊息後，經為期二個多月的成員甄選，正式參與TASA英語文小六標準設定成員總人數為32名，而成員是分佈在北部（20名，62.5%）、中部（7名，21.9%）、南部（2名，6.3%）、東部（3名，9.4%），其中教師占有24名（75%），其次為行政人員4名（12.5%）、學者4名（12.5%），而性別的分佈是男性占有6人（18.8%）、女性占26人（81.2%）的比率。最後，小六成員總教

學年資或行政年資，最低是2年、最高是31年，平均年資是10.2年。整體而言，TASA英語文小六標準設定成員大致能含括北中南東四個區域的人員、同時能兼顧教師、行政人員及學者的代表組成，最後，教學或行政年資具高水平，平均皆超過10年。

四、會議流程

這32名選定的標準設定成員，應邀參加兩天的標準設定會議。 在會議進行前的一個禮拜，研究小組先寄送會議的前導資料，包括標準表現描述、評量架構、會議簡介、與會議流程說明。會議第一天，所邀請之32位成員全數出席，先由研究者簡要說明會議的目的、流程之後，並請所有成員，就標準表現描述內的細項內容，逐一檢視，並加以討論，並請TASA英文科召集成員協助釐清成員們的疑問。之後，成員檢視題本，並進行第一輪Yes/No Angoff方法的標準設定，第一輪結束後，研究小組進行統計分析，並繪製每一位成員給定分數的散佈圖、各試題傳統難度P值，與每一題，成員給定基礎、精熟、進階的比例。成員們就回饋訊息的內容，逐題進行討論，原先設計30分鐘進行小組討論，30分鐘全體討論，由於題目眾多，成員無法在30分鐘內完成小組討論，因此要求取消全體討論時間，以求能盡量能與小組成員討論題本內的題目。小組討論後，成員們修正彼此的看法與意見並填寫執行方法的評估問卷，內容包括對於標準設定方法的理解，或是對於會議流程進行的建議等。而後進行第二輪的判定，判定後則同樣提供成員相同的回饋訊息，並再次進行小組討論與填寫回饋訊息的評估問卷，問卷內容為對於回饋訊息的理解度等。最後進行第三輪的標準設定。 設定的結果公布給成員之後，進行成果問卷的填寫，問卷內容包括對於成員們對於自己所設定的分數信心強度、覺得最終結果是否合理等，會議結束後研究者將所有題本、回饋訊息回收。

會議第二天為同一批成員，除了有一位成員因病缺席之外，其它31位委員均參與第二天的會議。第二天所執行的標準設定方法為書籤標定法。流程與第一天雷同，第二天採用同樣的會議資料來進行書籤標定法的標準設定。但會議前與成員們強調題本是依據三參數試題反應理論中的難度，由易至難進行排序，由於三參數的難度估計，易受鑑別度和猜測度的影響，因此在題本難度的排序上，可能和成員們的心中的期望的難度排序不完全吻合。

第二會議流程，和第一天相同。但在回饋訊息的提供上略有不同，第一輪結束後，提供的回饋訊息為每一位成員所給定書籤位置的散布圖，而後進行30分鐘小組討論與30分鐘全體成員討論。第二輪結束後，除了提供每位成員給定書籤位置的散布圖之外，應某部份成員要求，呈現依據成員所給訂的切斷分數，在2009 年TASA的實徵數據下，有多少百分比會落在基礎以下、基礎、精熟和進階四個水平，研究者解釋完回饋訊息之後，則進行30分鐘小組討論與30分鐘全體討論。第二天會議成員評估問卷和第一天雷同，問卷中除了詢問關於書籤標定法所設出的結果評估之外，亦有幾題是詢問成員對於兩種標準設定法執行後的感想。

參、研究結果

一、切斷分數

表2與表3呈現在每一輪各水平的切斷能力值。對於書籤標定法而言，精熟水平在三輪中的書籤標定位置產生的波動較大，由第一輪的-0.34，第二輪的-0.43，到第三輪的-0.08，相對而言，對於基礎水平和進階水平而言，相對變動的幅度較小。對於Yes/No Angoff法而言，進階水平的切斷分數則於第二輪到第三輪的變動的幅度較大。相對而言，基礎與精熟的切斷分數在每一輪的變動幅度較為穩定。此外，書籤標定法在判定每一水平的切斷分數上，成員們評定分數的標準差較低，代表成員們在執行書籤標定法時較能達到共識，但唯一的例外為精熟水平的標準差，達到0.5左右，代表成員們在經過第二輪的討論之後，對於精熟的書籤標定位置有較大的分歧。

最終的切斷分數是以第三輪的結果來判定，就第三輪的切斷分數來看，Yes/No Angoff法在基礎與精熟水平的切斷分數略低於書籤標定法，但是進階水平的分數卻略高於書籤標定法有1個logit 之多。

表2 書籤標定法結果

輪	決斷分數（標準差）		
	基礎	精熟	進階
1	-0.90 (0.16)	-0.34(0.28)	0.21(0.22)
2	-0.94 (0.11)	-0.43(0.23)	0.30(0.17)
3	-0.96 (0.06)	-0.08(0.50)	0.22(0.10)

表3 Yes/No Angoff法結果

輪	決斷分數（標準差）		
	基礎	精熟	進階
1	-1.53 (0.42)	-0.41(0.28)	1.59(0.38)
2	-1.44 (0.22)	-0.35(0.27)	1.69(0.27)
3	-1.51 (0.28)	-0.44(0.00)	0.44(0.36)

表4為依據每一輪在各水平的切斷分數，並將此切斷分數放入2009年TASA的實徵數據中進行運算各階層的百分比人數。這兩個表中可以看出對於書籤標定法而言，精熟水平在每一輪中的變化較大，由第一輪的18%，第二輪的25%，到第三輪的11%。而對於Yes/No Angoff法而言，則在進階水平的變化較大。之所以產生變化的原因，在於進行Yes/No Angoff時，第一次的回饋訊息，單單只提供每個成員所給定判斷值得相對位置散布圖與每一題的傳統難度P值，而在第二次回饋訊息時，除了第一輪給定的回饋訊息之外，還呈現了各階層的百分比人數，許多成員在看到進階水平的人數過少，覺得如此的切斷分數會造成社會觀感不佳的問題，因此普遍降低原先所設定的進階水平切斷分數。

進行完兩天的標準設定之後，研究者要求成員寫下他們心目中理想的各階層人數的分配比率，所得的各階層分配如表五，由這此分配而言，較為接近Yes/No Angoff方法所設出的結果。 然而，在進行訪談時，多位成員表示，書籤標定法所設

出的結果反應出教學現場的真實情況，即為一個M型的學習生態，大多的學生集中在非常優秀或是極為落後的兩群，而中間程度的學生比率相當少，因此，成員們覺得雖然Yes/No Angoff法所設出的成果，較貼近表現標準描述與社會的期待，但是書籤標定法所呈現的結果，卻能反應教學現場的真實情況。

表4 各階層的百分比人數-書籤標定法、Yes/No Angoff法與成員心目中的理想比率。

水平	第一輪		第二輪		第三輪		成員的理想比率
	bookmark	Angoff	bookmark	Angoff	bookmark	Angoff	
基礎以下	20	8	19	9	19	8	13
基礎	16	26	14	27	25	25	24
精熟	18	65	25	64	11	34	34
進階	46	1	42	0	45	33	29

在標準設定的會議進行中，成員們就整體的會議說明、方法執行、回饋訊息的提供等各項層面對兩個標準設定方法進行評估，評分採李克氏量表， 1代表態度為正向的程度為最低（如 非常不同意、非常不清楚、非常沒信心等）、5為態度為正向的程度最高。而 1到 5中間的數字，則代表程度上的差別。例如1代表非常不同意、2代表不同意、3代表沒意見、4代表同意、5代表非常同意等。

以下就這幾個層面，來進行說明：

（一）會議說明

成員們認為收到前導資料能夠有助於了解會議中，自己所應扮演的角色，對於TASA的測驗目的與標準方法的執行，都呈現極高的評價與理解。然而，成員們對於Yes/No Angoff方法，給予的評價高於書籤標定法。例如在「我已經了解Yes/NoAngoff方法，並可以使用這個方法，進行試題的判別與歸類」，成員們給的平均分數為4.06，但對於「我已經了解Yes/No書籤標定法，並可以使用這個方法，進行試題的判別與歸類」，則給的分數為3.84。 然而、對於會議流程的進行來說，書籤標定法的分數高於Yes/No Angoff方法，例如「我瞭解會議接續的標準設定流程」，Yes/No Angoff方法為4.13，書籤標定法則為4.26。這是因為第一天執行的方法為Yes/No Angoff方法，而第二天執行的方法為書籤標定法，因兩天的流程都相同，成員對於自己要進行的任務，在第二天也有較清楚的認識。

（二）方法執行

兩個方法執行上來說，成員們覺得Yes/No Angoff方法，執行上比較容易，也較能夠和PLD作連結。例如在「基礎的表現標準描述（PLD）有助於我判別邊緣基礎受試者可以答對的題目」，成員給Yes/No Angoff方法的平均分數為3.78，而在「基礎的表現標準描述（PLD）有助於我置放介於邊緣低於基礎/基礎的書籤」，的平均分數為3.43，相似的，精熟與進階也呈現類似的情況。此外，成員們覺得Angoff方法比書籤標定法，更容易理解與應用自己的教學經驗。然而，由於Angoff方法必須逐題判斷，許多成員覺得討論的時間不夠充裕。相對而言，書籤標定法的給予成員們的討論時間就較為充裕。

（三）回饋訊息與討論

對於回饋訊息而言，就各成員給定分數的散佈圖、試題難度、各水平百分比而言，成員們都很了解個訊息的意涵，然而，成員對於Yes/NoAngoff 的回饋訊息理解程度，還是略高於書籤標定法，例如，成員對於其它成員（與自己）判定決斷分數的散布圖說明，Yes/No Angoff方法的清楚度，達到4.06，而對於書籤標定法，則為3.84。但成員們覺得書籤標定法的結果較容易討論，因為在只需要討論三個書籤的放置位置，但Yes/No Angoff法卻必須要逐題討論，很多成員都反應討論時間不夠。

（四）兩種設定方法的比較

成員普遍認為Yes/No Angoff標準設定法，是比較簡單、易懂的執行方法，在「哪一種標準設定方法較為簡單易懂」 與「哪一種標準設定方法較為易於執行」上，有27位的成員選擇Yes/No Angoff法， 3位選擇書籤標定法，而1位沒有意見。此外，有30位成員覺得Yes/No Angoff法，較能與標準表現描述作結合，最後，對於標準設定法所設出的結果，26位成員覺得Yes/No Angoff法所做出的結果較符合自己心中的預期，只有4位覺得書籤標定法所做出的結果比Yes/No Angoff 法較符合心中預期。

詳細的問卷內容的比較請詳見表5。

表5 Yes/NoAngoff法與書籤標定法評估問卷比較

題目		Angoff	書籤標定
會議說明	我認為先前收到的前導資料能充分幫助我瞭解本次會議應扮演的角色	4.34	4.19
	我瞭解TASA的測驗目的與施測對象	4.35	X
	我瞭解本次TASA標準設定會議的目的	4.41	4.16
	本次會議對於表現標準描述（PLD）的說明及其範例的陳述，我認為	4.00	X
	本次會議對於如何執行Yes/No Angoff／書籤標定法的說明，我認為	4.06	3.94
	我已經了解 A Yes/No Angoff方法，並可以使用這個方法，進行試題的判別與歸類。 B書籤標定法，並可以使用這個方法，進行標準設定的工作	4.03	3.84
	我瞭解邊緣受試者的涵義	3.97	4.13
	我瞭解會議接續的標準設定流程	4.13	4.26

方法執行方法執行	本次會議的解說與導引時間分配，我認為	3.16	3.06
	本次會議提供歸類試題／置放書籤的時間分配，我認為	2.94	3.23
	基礎的表現標準描述（PLD）有助於我 A判別邊緣基礎受試者可以答對的題目 B置放介於邊緣低於基礎/基礎的書籤	3.78	3.43
	精熟的表現標準描述（PLD）有助於我 A判別邊緣精熟受試者可以答對的題目 B置放介於邊緣基礎/精熟的書籤	3.56	3.45
	進階的表現標準描述（PLD）有助於我 A判別邊緣精熟/進階受試者可以答對的題目 B置放介於邊緣精熟/進階的書籤	3.78	3.47
	研究者所提供依難度排序試題本（OIB）符合我所知覺試題間的相對難度	X	2.68
	我對採用67%的正確作答標準，去界定書籤的位置，感到合適	X	3.39
	我先前的教學經驗，有助於我瞭解 A進行試題的判別與歸類 B該如何置放各階層的書籤位置	4.28	3.81
回饋說明與討論	對於其它成員（與自己）判定決斷分數的散布圖說明，我認為	4.06	3.84
	我瞭解其它成員（與自己）判定決斷分數的相對位置	4.13	4.00
	對於基礎、精熟、進階等水平通過人數百分比的說明，我認為	4.03	4.00
	我瞭解基礎、精熟、進階等水平通過人數百分比的意涵	4.25	4.00
	對於試題通過率的說明，我認為	4.31	X
	我瞭解試題通過率的意涵	4.45	X
	基礎、精熟、進階等水平通過人數百分比會影響我 A對試題的歸類 B放置書籤的位置	3.81	3.71
	其它成員所判定之決斷分數會影響我 A對試題的歸類 B放置書籤的位置	3.44	3.42
	試題通過率會影響我對試題的歸類	3.84	X
	試題品質的好壞（如誘答選項的設計，或是題目敘述不清）會影響我 A判別各水平受試者可以答對的題目 B放置書籤的位置	4.50	4.61
	透過小組的討論，我充分瞭解其他成員的想法	4.25	4.26
	透過與小組成員的討論，有助於我 A對試題的歸類 B放置書籤的位置	4.19	3.97
	本次會議提供小組討論標準設定結果適切性的時間分配，我認為 1.太短 2.略短 3剛剛好 4略長 5太長	2.59	3.32

結果評估與方法比較時間分配設定方法	我相信自己對每一個試題／水平的判別，是與表現標準描述（PLD）一致	3.78	3.23
	我對於自己所設置的決斷分數，深具信心	3.97	3.81
	我對於最後的決斷分數，感到	2.71	3.14
	本次會議各階段任務的執行時間，我認為 1.太短 2.略短 3剛剛好 4略長 5太長	3.03	3.28
	哪一種標準設定方法較為簡單易懂	27人	3人
	哪一種標準設定方法較為易於執行	27人	3人
	表現標準描述（PLD）對於執行哪一種標準設定方法較有幫助	30人	1人
	哪一種標準設定方法所設出的結果，較符合您的預期	26人	4人

註：X代表該天問卷中沒有這個題目。評分採李克氏量表，除非特別標明評分的尺度，否則1代表態度為正向的程度為最低（如非常不同意、非常不清楚、非常沒信心等）、5為態度為正向的程度最高。而1到5中間的數字，則代表程度上的差別。例如1代表非常不同意、2代表不同意、3代表沒意見、4代表同意、5代表非常同意等。

肆、結論與建議

在本篇研究中，比較書籤標定法與Yes/No Angoff標準設定方法在設定切斷分數上的差異，並使用TASA 2009年英文科來進行比較的研究。研究結果顯示，兩種標準設定方法所做出的決斷分數略有不同，而成員們覺得Yes/No Angoff方法對他們來說是執行起來較為簡單的方法，而最終決定的切斷分數也較符合預期，書籤標定法對於成員而言，需要克服的技術層面較多，尤其是在面臨題本難度的順序，和心中期望的難度順序不一致時，往往會造成他們很難找到確切的書籤位置，但是，書籤標定法所需的執行時間較短，對於成員的心理與體力負擔較輕，做出的切斷分數結果，反應在2009年的實徵數據上，發現中間部分的學生比例明顯低與基礎以下與進階的學生，許多成員覺得做出切斷分數呈現M型分布，較能反應台灣學生學習成就的真實情形。

此外，在標準設定中，標準表現描述（PLD）為成員們達成共識的主要依據，然而，就應用層面而言，成員們反應Yes/No Angoff方法較容易和PLD進行連結，因為Angoff方法為逐題判斷，成員們比較容易一題一題找出哪一題為基礎水平的邊緣學生可以達對的題目，哪一題為精熟水平的邊緣學生可以達對的題目。然而，對於書籤標定法就比較不容易了，因為要找出三個切斷點，基礎的切斷點代表基礎水平的邊緣學生，大概能答對所有的題目，可是對於OIB題本，是依造試題反應理論的難度來排的，並不是依據標準表現描述來排的，因此，成員反應雖然可能將切斷點到某一試題，但是並不代表此試題之前的所有題目，該水平的邊緣學生都可達對。例如，基礎的切斷點是放置在第20題，但是有可能第 5、 8題，是精熟學生才能答對的。因此，成員進行書籤標定法時，與到許多困難與衝突點，但透過討論後，的確能加速成員達成共識。

對於回饋訊息方面，成員們對於自己與其他成員的切斷分數散布圖感到十分重視，因為許多成員都不希望自己成為「異類」，因此，看到散布圖後，均會修正自己原先的判定，而逐漸和其他人判定的結果相同。但是，在兩天的標準設定會議中，卻看到明顯的城鄉差距。城市的學校老師，所設定的標準較高，而偏遠的學校老師，則設定的標準較低，即使設定時是依據PLD來進行推論，但是教師們還是會加入自己在教學現場經驗和看法，來進行標準設定。

就成員們評估的結果而言，Yes/No Angoff方法比較適合TASA英語科的標準設定，然而，進行此方法的前提為要有足夠的會議時間進行討論。就本研究原先設計，為每一輪成員進行標準設定的時間為50分鐘，討論為40分鐘，但是，由於Angoff的方法要逐題討論，因此，對於題本題目較多的測驗，是有執行上的困難的。例如TASA英文科的題本考題有一百多題，成員大多沒辦法在 40分鐘內完成討論，造成許多成員是邊進行標準設定，邊進行討論。此外，在成員的數據輸入和統計分析上也是一大挑戰，本研究總共有 32位成員，每一位成員進行103題的判斷，因此要進行三千多筆數據的輸入。本研究動員5-6位助理同時進行輸入與檢誤，花費約30-40分鐘的時間，因此，使用Yes/No Angoff方法雖然比較簡單，但是要考量到時間與人力的負荷。 相對而言，書籤標定法的討論就很簡單，成員們無須逐題討論，只需討論自己放置書籤的位置就可以了，同樣安排討論的時間為40分鐘，而成員們只需30分鐘左右就討論好了，而輸入更為簡單，只要為每一位成員輸入三筆數據，本研究的書籤標定法的成員為31人，總共需要輸入的數據不到 100筆，本研究只有動員兩位研究助理，花不到10分鐘的時間就完成輸入和檢誤。因此，就題目較多的OIB或是人力不足的研究小組，較適合使用書籤標定法。

隨著國際的趨勢，國內雖然漸漸知覺到長期學習成就評量資料庫建置的重要性，但對於標準設定的議題，瞭解可說是少之又少，而本研究的研究過程與結果，除了做為TASA實務的運用，也可作為其他國內建置大型學習成就評量資料庫標準設定之參考。

參考文獻

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp.508-600). Washington, DC: American Council on Education.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard.set tings based on Bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Impara, J. C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353-366.

Lewis, D. M., Mitzel, H.C., & Green, D. R. (1996). *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Boulder, CO.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark method: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

以多面向Rasch測量模式分析TOCFL口語測驗評分者訓練效果

藍珮君

以多面向Rasch測量模式分析TOCFL口語測驗評分者訓練效果

藍珮君
國家華語測驗推動工作委員會研究員

摘要

本研究目的為探討在接受持續且長期的評分訓練後，評分者評分一致性的變化情形，包含與其他評分者之間的一致性，以及評分者自身的一致性。研究方法為收集2009年2月至12月五次華語文口語能力測驗（簡稱TOCFL口語測驗）基礎級預試評分結果，採用多面向Rasch測量模式(many-facet Rasch measurement)分析軟體FACETS，針對評分者嚴格度（rater severity）、評分結果一致性（consistency）以及評分者面向之適配度（fit）進行分析。

研究結果顯示：1. 經過多次且密集的評分訓練後，評分者之間嚴格度差異逐漸縮小，但整體評分嚴格度達到顯著差異水準，顯示評分者之間的嚴格度仍有不同；2. 評分者本身的給分相當一致，適配度數值均在可接受之範圍內；3. 部分評分者間隔二個月後的評分嚴格度有所變化。

現今以華語為第二語言的口語測驗研究相當缺乏，尤其是採用多面向Rasch模式進行分析之研究。本研究結果提供華語文口語測驗的實徵資料，對於華語文口語能力測驗的評分訓練效果有初步的瞭解，也能做為未來進行華語文口語能力測驗評分訓練及測驗實施的參考。

關鍵字：多面向Rasch測量模式、口語測驗、評分者訓練

Using many-facet Rasch measurement to examining rater training effects of TOCFL Speaking

Pei-Jiun Lan
Steering Committee for the Test Of Proficiency-Huayu / Researcher
martinalan@sc-top.org.tw

Abstract

This research aims to investigate how rater consistency varies after a continuing, long-term rater training by observing the inter-rater consistency and the intra-rater consistency. We collect the rating results from the five pilot tests of the TOCFL Speaking (Test of Chinese as a Foreign Language- Speaking, holding from February, 2009 to December) at beginner level. This research applies the many-facet Rash measurement by adopting the FACETS software to analyze the rater severity, rater consistency, and the fit statistics of the raters.

Three major findings discussed in this research are that: 1) after frequent rater training sections, the variation of inter-rater severity has reduced, but the overall rating severity has reached the significant differences, which indicates that the discrepancy of severity still exists among the raters; 2) individual raters are consistent in their own rating as most of the fit statistics fall in the acceptable range; 3) several raters reveal variations in their severity when they rerated two months later.

So far, only little research on the speaking assessments which focus on using Chinese as a second language (CSL) has been done, and those which applied many-facet Rasch measurement are even less. Empirical data from the results of this study will be provided to help gain a preliminary understanding toward the effects of rater training in CSL speaking tests, and as a reference for the future rater training.

Keywords: many-facet Rasch measurement, speaking test, rater training

壹、前言

華語文口語能力測驗（前稱為Test of Proficiency-Huayu，現英文名稱更名為Test of Chinese as a Foreign Language -Speaking，以下簡稱TOCFL口語測驗）為一種表現測驗（performance assessment），考生需以口說的形式，完成考試的各種溝通任務。任務的類型均為建構反應題（constructed response item），考生在觀看一段影片或一至數張圖片後，依照題目的說明或要求，以自己的方式表達與建構出答案。此種考試型態，與一般常見的選擇題，除了答題方式外，計分方式也有明顯不同，選擇題有標準答案，計分客觀；而建構式題型沒有標準答案，考生的成績是由評分者根據評分規準，對考生的測驗表現進行評分而來。因此，與選擇題相較之下，影響考生在建構反應題表現的因素，除了考生自身能力以及試題難度以外，評分者也是一重要的影響因子。

正因如此，TOCFL口語測驗自研發以來，即相當注重評分者的培訓。目的在透過嚴謹的評分訓練，使評分者充分理解評分規準並依據此原則給分，進而讓參與測驗的考生得到公平可靠的成績。在每次預試的評分工作完成後，也針對評分結果進行評分者間信度分析，瞭解評分者給分的一致情形，以確保評分者的評分品質。即使如此，仍有一些疑問難以釐清，例如，當二位評分者給分一致性高時，究竟是表示二位評分者均依據評分原則進行評分，或其實評分者恰巧都是給分偏嚴格或偏寬鬆，所以有看似良好的一致性。而評分者一致性低時，會不會是其中一位評分者依循評分原則給分，但另一位評分過於嚴格或寬鬆所致？以往的分析方式，除非事先在評閱的樣本中置入標準卷（standard rating），否則很難看出端倪。

然而，多面向Rasch測量模式（many-facet Rasch measurement，以下簡稱MFRM）的出現，解決了上述難題。延伸自單面向Rasch模式的MFRM，能夠同時估計二個以上的面向（稱之為facet），除了試題難度以及考生的能力外，也能將評分者等其他相關因素納入估計，可以瞭解評分者評分的嚴格或寬鬆程度，甚至是評分面向的難易度，這對於測驗發展者來說不啻是一大福音。目前已有許多研究採用MFRM檢驗口語或寫作測驗評分者給分的品質，其中有研究發現，藉由評分訓練，能改善給分極端的評分者嚴格度，但無法使評分員的給分達到一致；此外，評分訓練可以提高評分員給分的信心，進而提升評分者內信度（Park, 2004；Weigle, 1998）。Lumley與McNamara（1995）的研究則指出，評分訓練的效果無法維持很久，評分者的嚴格度會隨時間產生變化。可惜這些研究多半是針對英語或德語等外語測驗，針對華語測驗所做的實徵研究相當缺乏，研究者透過搜尋引擎與各資料庫，最後僅在中國知識資源總庫查詢到二篇，均為中國大陸學者發表之期刊及論文（田清源，2007；羅丹，2008）。

綜上所述，本研究將採用MFRM對TOCFL基礎級口語測驗預試的評分結果進行分析，研究目的在探討接受持續且長期的評分訓練後，評分者評分一致性與嚴格度的變化，包括：

1.評分者間評分一致性的變化。
2.評分者內給分一致性的變化。
3.評分者在二次評分訓練中評分嚴格度（rater severity）的變化。

貳、文獻探討

一、評量口語能力的測量模式

Engelhard於1992年提出寫作評量計畫的測量模式，由於口語測驗的實施方式與計分流程與寫作測驗相似，本研究的口語能力測量模式參考Engelhard的模式加以修改後如圖1所示。測量模式包含三個主要的面向[1]：口語能力（speaking ability）、評分者嚴格度（rater severity）以及口語任務難度（difficulty of the speaking task）。此模式將評分者與口語任務視為中介變項（intervening variables）。

因此從圖1可知，考生口語能力會受到考生特質所影響，而最後得到的觀察分數，除了考生的口語能力外，還受到評分者給分嚴格與否，以及口語任務難易度的影響。此外，評分量尺的架構（structure of the rating scale）也會影響考生獲得的成績。

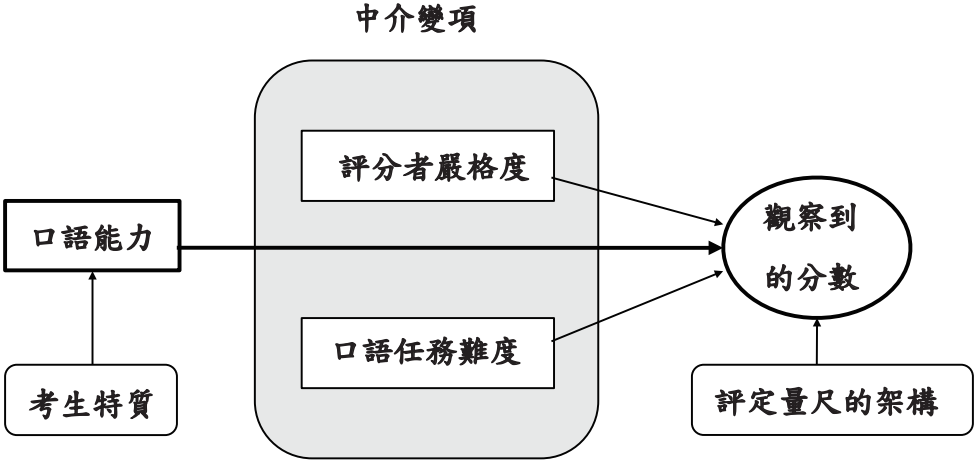


圖1 口語測驗的測量模式(修改自Engelhard, 1992)

除了Engelhard之外，亦有其他學者重視評分者因素對於考生測驗成績的影響。如Cason與Cason（1984）的研究即指出考生得分不僅代表其真實能力水平，而是考生能力和評分者個人特質的函數（引自羅丹，2008）；而Hoyt和Kerns（1999）對79篇概化（generalizability）研究進行後設分析（meta-analysis）發現，平均有37%的成績變異來自於評分者主要效果以及評分者與考生的交互作用（引自Eckes, 2005）。Engelhard（1994）採用Saal、Downey和Lahey（1980）提出的四種主要評分者誤差類型（評分嚴格度、月暈現象、趨中效應以及全距限制），來檢視15位評分者評閱寫作測驗的表現，結果顯示評分者的嚴格度達到顯著差異，其餘三種誤差類型也獲得證實。

[1]原模式中介變項中，尚納入評分面向難度(domain difficulty)，因本研究不對此進行探討，故省略。

上述研究結果在在顯示出評分者因素對於表現測驗成績的影響不容忽視，若一份測驗沒有良好的評分品質，所得到的成績將沒有信度及效度可言。有鑑於此，測驗機構及測驗發展者會藉由一些方式或步驟，盡可能讓評分過程和結果更為可靠，例如：在評分前舉行評分者訓練，讓評分者熟悉評分的規準；或對評分者進行評分測驗，挑選出合格的評分者；評分結束後對結果進行分析，作為之後評分之參考……等等。傳統上，舉辦評分者訓練目的就是在減少整體評分嚴格度以及隨機誤差兩者的變異（Lumley & McNamara, 1995）。

二、多面向Rasch測量模式

在過去，分析評分者間信度的方式，主要使用百分比一致性、皮爾森等級相關、肯德爾和諧係數、Kappa一致性係數等相關分析。百分比一致性在瞭解評分者對考生評定成績完全相同的比例；相關分析的目的在檢視評分者之間給分是否存在相對應的次序關係。當參與測驗的考生人數眾多時，測驗機構為減輕評分者負擔，往往將二位評分者分為一組進行評分，當二位評分者給分有落差時，再由第三位評分者加入評閱，並以上述方式分析評分者間信度。然而，高的評分者間信度，並不一定代表評分結果是正確的，也可能是二位評分者都是較為嚴格或是較為寬鬆的評分者。在沒有置入標準卷或其他介入機制的情形下，得到的評分者間信度可能反而提供錯誤的訊息，使研究者判斷失當，也連帶影響考生權益。

多面向Rasch測量模式（Many-Facet Rasch Measurement，以下簡稱MFRM）是Rasch測量模式的延伸，由Linacre於1989年提出，可以同時校準（calibrate）多個面向，並能分開呈現估計結果（引自Engelhard, 1992）。MFRM之下有幾種分析模式，一般常見的有評定量尺模式（rating scale model）和部分給分模式（partial credit model）。二者的差異在於前者假定每個試題的量尺架構相當；而後者每一個試題都有各自的量尺架構（Bonk & Ockey, 2003）。本研究由於每個題目均以相同的評分原則（rubric）進行給分，因此使用評定量尺模式，公式如下（Linacre, 2010）：

$$\log (P_{nmjk} / P_{nmj(k-1)}) = B_n - A_m - C_j - F_k$$

在此，

P_{nmjk} 是指考生n被評分者j在題目m評為k分的可能性

$P_{nmj(k-1)}$ 是指考生n被評分者j在題目m評為k-1分的可能性

B_n 是指考生n的能力

A_m 是指題目m的難度

C_j 是指評分者j的嚴格度

F_k 是指得到k分與k-1分之間的界線（barrier），也有人稱為難度階（threshold difficulty）

由於MFRM是針對評分者嚴格度、題目難度以及考生能力等面向的資料，同時進行估計及校準至一個共同的量尺上，因此各個面向得到的數值是可以互相進行比較的。估計考生能力時，因將評分者嚴格度差異納入考量進行調整，故獲得的考生能力比原始分數更能代表考生的真實能力。

本研究使用FACETS統計軟體（3.67版）進行分析，FACETS提供一些指標呈現各個面向的分析結果，在本研究中，使用到的指標與評分者嚴格度有關，包括：分散指標（separation index）、信度（reliability）、卡方考驗，以及評分者給分一致性（rater consistency）的infit統計值。以下分別說明各指標的定義與數值代表的含意。

（一）分散指標（separation index）

分散指標是測量值的校正標準差（Adj.SD）與均方根誤差（RMSE）的比值。數值越接近0，表示評分者間的嚴格度越相近，數值越大，表示評分者間的嚴格度差異較大（Weigle, 1998）。

（二）信度（reliability）

FACETS提供的信度數值是上述分散指標的可信度，也就是表示分析結果能區分不同能力水準考生的可信度。這裡所指的信度指標，就是眾所知悉的KR20信度指標的Rasch相似形（analogue），只是改以能力量尺而非原始分數進行計算（Pollitt & Hutchison, 1987）。一般來說，信度數值越高越好，表示越能區分考生的能力，如在考生面向，分散指標的信度越高，表示測驗越能將考生能力區分為不同程度。不過在評分者面向來說，較低的信度數值是好的，因為理想上希望不同的評分者有相同的評分嚴謹度（Park, 2004）。

（三）卡方考驗

為同質性考驗，目的在考驗所有評分者的嚴格度是否相等，虛無假設為所有評分者的嚴格度是相等的，若達到顯著水準，拒絕虛無假設，則表示至少有二位評分者的嚴格度是不同的。

（四）適配度（fit）

一般來說，適配度統計值是指MFRM模式中，觀察到的評分與預期評分結果的適配情形，分為outfit與infit二種。outfit值對於偶發性的極端非預期評分較為敏感；而infit值則對於累積的非預期評分結果較為敏感（Eckes, 2009）。因此，許多學者認為infit數值較為適合作為判斷適配度的指標（Pollitt & Hutchinson, 1987；Park, 2004）。

當測量模式與觀察資料適配時，infit的數值為1.0。Lunz、Wright和Linacre（1990）建議可接受的範圍為0.6至1.5（引自Engelhard, 1992）；Linacre（2002）建議用0.5為低標，1.5為高標，也有其他研究建議使用比較嚴格的標準（0.7-1.3）（McNamara, 1996；Bond & Fox, 2001；引自Eckes, 2005）。Lunz 與Stahl（1990）表示，若infit數值大於等於1.5，表示評分者有太多非預期的給分，稱為misfit；若小於等於0.5表示評分者給分的變異不足，稱為overfit（引自Weigle, 1998）。

Bonk與Ockey（2003）提到，比起考生的misfit，評分者的misfit對於測驗的效度可能造成更嚴重的威脅，因為這表示評分者的給分偏離自身的標準，這對於其他面向的估計影響很大，misfit也是Rasch分析無法進行校正的。

三、運用MFRM對於評分者及評分訓練的相關研究

Linacre（1989）提出MFRM後，許多學者開始採用此一測量模式對評分者因素及評分訓練的成果進行相關分析。Engelhard（1992）以接受了三天評分訓練，且通過評分測驗（評20篇作文，最少需達到62%給分完全相同，其餘篇數僅能相差一級分）的82名合格評分者，對1000名學生的英語寫作測驗進行評分。結果顯示，這82名評分者中，給分最嚴格及最寬鬆的評分者嚴格度差距為3.52 logits，嚴格度的卡方考驗達到顯著差異，而separation index的信度也達到0.87，顯示評分者即使已經受過訓練且通過嚴格的測試，但正式評分時的嚴格度仍是有差異的。

田清源（2007）以漢語水平考試（HSK）高等作文進行分析，結果顯示8位評分者嚴格度相差4.28 logits；羅丹（2008）對漢語水平考試（HSK）中級口語測驗的研究結果顯示，9位評分者嚴格度差異為2.72 logits，separation index為4.93，信度達到0.96，卡方考驗亦達到顯著，顯示評分者整體的給分嚴格度並不一致；評分者內一致性方面，有1位評分者給分變異過大。

Weigle（1998）則針對有評分經驗與無經驗的評分者，比較兩者接受評分訓練前後的評分嚴格度。結果發現比起有經驗的評分者，無經驗的評分者給分較為嚴格，本身的給分也較不一致；而經過評分訓練後，二個群體間的嚴格度差異減少，但仍達到顯著差異水準，此外，大多數評分者內一致性已經提高。

Bonk與Ockey（2003）、Park（2004）以及Eckes（2005）分別以英語口語測驗、英語寫作測驗（CEP writing test）與德語測驗（TestDaF）的寫作與口語測驗進行研究的結果也顯示，接受評分訓練後，評分者給分的嚴格度仍有顯著差異，然而，評分者內的一致性則較佳。McNamara（1996）從McIntyre（1993）、Weigle（1994）與Shohamy等人（1992）的研究結果發現：1.評分者訓練能讓評分者在評分時更有信心，訓練的成效在於降低評分者給分的隨機誤差（random error）；2.評分者訓練能降低評分者整體嚴格度的變異，但無法完全消除。特別是能降低非常極端嚴格或寬鬆的評分傾向，但是明顯的評分者差異依然存在。

其他研究則有不同的發現，Du、Brown與Rogers（1997）以學生能力、評分嚴格度、文本難度以及評分標準四個面向對寫作測驗的評分結果進行分析。30位評分者嚴格度的差異不到1 logits，顯示評分者嚴格度沒有非常極端嚴格或寬鬆，不過由於作者未提供其他指標結果，所以無法得知整體評分嚴格度有沒有差異。Liu與Wen（2007）邀請學生與教師對一批參加口語演說比賽的學生進行二次評分，前後間隔二個月。結果發現學生評分嚴格度的差異雖然縮小，但卡方考驗結果仍達到顯著差異；而教師的評分嚴格度差異極小，二次評分結果，卡方考驗均未達到顯著水準，顯示教師整體評分嚴格度沒有不同。唯該研究將教師分為男女二組進行分析，分別為3人和2人，評分者人數較少。

另外，也有研究更進一步針對評分者評分嚴格度的變化進行探討。Lumley和McNamara（1995）檢視13名評分者對同樣10名考生二次評分嚴格度的變化，評分

時間間隔18個月；其中有6位評分者，隔2個月後，再對另一群考生（73人）進行評分，持續觀察嚴格度的改變。結果發現，有些評分員的嚴格度會隨著時間改變，且這個改變並不穩定，有的評分者變得比較嚴，有的則變得較為寬鬆，顯示評分訓練的結果不必然能持久。Bonk與Ockey（2003）的研究也有類似的結果，13位評分者在連續二年的評分嚴格度變化很大，有8位評分者變得較為嚴格，3位評分比之前寬鬆，另2位變化的幅度較小。

綜合上述研究結果及McNamara（1996）的看法，無論是何種語言測驗，評分訓練對於提升評分者間信度（interrater reliability）的效果似乎不盡理想，但對於提升評分者內信度（intrarater reliability）的幫助較大。畢竟評分者有其個人的人格特質、專業知識、教學經驗等不同背景，即便接受評分訓練，共同理解了評分原則與規準，掌握了大致的評分方向，但對於一些細部的解讀上可能仍有差異，特別是一些比較難評閱的學生表現，很難透過培訓達到完全一致的評分結果。但是透過評分訓練，評分者在短時間練習評閱了為數不少的樣本，對於考生整體能力開始產生較為具體的認知，並逐漸歸納出不同級分考生的能力表現；可能因而使評分者在培訓後的正式評閱，更有依據進行給分，達到個人的評分一致性。

先前的研究，二次評分訓練的間隔期間多半為一年左右，次數最多為三次，若採取較為密集的評分訓練模式，是否能使評分者的嚴格度有更為明顯的改善，達到一致的評分嚴格度，是研究者感興趣的議題。故本研究採取多次且密集的評分訓練模式，希望藉此瞭解評分者的評分嚴格度與一致性是否與過去研究發現相同，亦或是因為密集的評分訓練而能達到更為一致的嚴格度。

參、研究設計

一、研究參與者

本研究於2009年2月至12月間共陸續舉辦五次基礎級口語測驗預試，在每次預試結束後，即緊接著開始進行評分訓練及評分工作，考試時間、樣本大小以及參與評分老師分佈如表1所示。五次評分訓練間隔約二至三個月不等，第一次評分訓練有3位評分者參加，第二次則有5位評分者參加，其餘三次評分者均為4人。若從評分者參加的次數來看，以評分者A和B共五次最多，評分者E參加二次，評分者C與D均參加四次。此五位評分者包含四位華語教師，以及華測會口語測驗研發專員一名(評分者A)，華語教師平均教學年資均超過十年。

考生人數方面，由於10月及12月預試使用同樣題目，研究者欲藉此進行評分者嚴格度的比較，故於12月評分時，隨機選取20位10月預試之考生錄音檔案，一併納入評分，為避免評分者憑先前一次評分的印象給分，事先未告知評分者。12月實際預試人數為63人，加上重複評閱之20名考生，共計為83人。

表1 TOCFL基礎級口語測驗預試時間、考生人數以及評分者參與次數一覽表

考試時間	人數	評分者A	評分者B	評分者C	評分者D	評分者E
2009/02	60	√	√	√		
2009/05	28	√	√	√	√	√
2009/07	49	√	√		√	√
2009/10	69	√	√	√	√	
2009/12	83	√	√	√	√	

二、測驗簡介

TOCFL口語測驗是專為母語非華語之人士研發的一種外語/第二語言口語能力測驗，目前規劃有基礎、進階、高階以及流利四個等級。測驗形式為電腦化測驗，口語考試題目透過電腦螢幕與耳機播放，題目播放完畢後，給予考生準備時間，準備時間結束後，考生以麥克風回答問題，錄音檔案將記錄在電腦裡。

在基礎級測驗，題目分為二大部分，第一部份為暖身題，共有3題，目的讓考生熟悉測驗介面，故考生在此部分的回答不納入計分；第二部分為正式題目，共有5題，命題方向著重於描述個人經驗、表達對事物的喜好，以及回答與日常生活有關的話題等，考試時間約30分鐘。

計分方式採用0至5級分之整體式評分法，評分者依照內容組織、表達能力以及語言運用三大方向對考生回答內容進行評分，考生在每一題的回答均會得到一個成績，若考生沈默未作答或回答離題，則評為0級分，3級分以上表示通過。

三、研究程序

TOCFL口語測驗研發人員於2008年底舉行第一次評分研習，邀請教學經驗豐富之華語教師進行培訓，培訓後初步挑選出數名評分一致性較高，且能配合進行後續評分訓練及評分工作之華語教師做為核心評分者，開始長期的合作。2009年五次評分訓練期間，研發人員亦陸續邀請曾於2008年參加過評分研習的教師擔任評分者，因有時在時間上難以配合，故部分評分者未能每次皆參加，評分者E因個人因素於2009年7月第三次評分訓練後，退出評分工作。

評分訓練流程可分為三個階段，第一階段測驗由研發人員簡介TOCFL口語測驗發展概況，以及說明基礎級考生的口語能力表現。第二階段測驗研發人員向評分者展示該次預試題目，以及各題各個級分的考生回答錄音檔範例（sample），同時提供評分細則，說明各範例給分依據。第三階段給予評分者每題各數個考生錄音檔案進行評分練習，再公佈練習檔案之得分，並與評分者針對給分較不一致的考生表現進行討論，以確認評分者對於評分原則的掌握程度相當。2009年5次評分訓練，皆依照上述流程進行。

三個階段評分訓練時間合計約為4至5小時，評分者於評分訓練後取得考生錄音檔案進行正式評閱，評分時間為一個月。待收齊評分者繳交之評分結果，研發人員隨即整理評分結果，挑選出給分歧異性較大的錄音檔案，加以反覆聆聽，再召集評分者針對這些檔案再作討論。如所有評分者給分相差2級分以上，或是分數介於通過與未通過臨界點的考生表現，進一步微調評分者給分的標準，目的在達成評分的一致性，整個討論約費時4至5小時。

肆、研究結果與討論

一、整體模式適配

資料與模式的整體適配度可以由非預期反應的次數得知，根據Linacre（2010），大約5%的標準化殘差等於或大於±2，大約1%的標準化殘差等於或大於±3，表示資料與模式適配。由於本研究五次評分訓練的試題和考生並非完全相同，故評分結果均採用獨立分析的方式。五次分析資料與模式的適配情形，標準化殘差等於或大於±2的比例介於4.7%至5.5%之間；等於或大於±3的比例介於0.3%至0.9%之間，大致上均符合適配的標準，顯示資料與模式的達到良好適配。

二、評分者嚴格度與一致性

表2至表6為五次評分訓練後評分者的嚴格度估計結果，在嚴格度欄位，數值為正表示偏嚴格，負值表示偏寬鬆，數值越大表示給分越嚴格，越小則表示越寬鬆。五次評分訓練後評分者嚴格度的差異分別為1.09 logits、0.93 logits、0.31 logits、0.32 logits以及0.31 logits，顯示經過幾次評分訓練後，評分者之間嚴格度的差異呈現出逐漸縮小的趨勢，然而由於五次分析分別為獨立的估計，每次估計得到的量尺（logit）未必相同，需再參考其他的指標。從separation index則可以得知，從第一次評分的6.08，逐步降低到3.30，甚至是1.15、1.57與1.28，顯示在長期且密集的評分訓練之下，評分者之間整體的嚴格度越來越接近；此外，信度數值也從剛開始的0.97、0.92，下降到0.57、0.71以及0.62。不過在評分者整體嚴格度的卡方考驗，除了2009年7月未達到顯著外（p=0.08），其餘四次的卡方考驗結果仍達到顯著水準，顯示整體來說，評分者的嚴格度雖然已趨於接近，評分者間一致性雖有所提升，但仍然未到達理想的水準。

在評分者內給分一致性方面，可以發現，即使採用較嚴格的標準（0.7-1.3）進行審視，五次評分結果，所有評分者的infit數值均落在適配的範圍之內（0.94-1.07、0.84-1.10、0.83-1.25、0.91-1.19、0.81-1.18），表示評分者內給分一致性良好，評分者參與評分訓練後，實際進行給分時能保持相當的穩定性，不會過於偏離自身的標準。

上述分析結果與Engelhard（1992）、Weigle（1998）、Bonk與Ockey（2003）、Park（2004）以及Eckes（2005）的研究發現相同，支持評分訓練對於給分極端的評分者能改善其嚴格度，但不能完全降低評分嚴格度的差異，評分訓練無法使評分者間的嚴格度達到一致，但是有助於提升評分者內給分一致性的論點。

表2 2009/02評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
C	0.38	0.10	0.98	0.99
B	0.33	0.10	0.94	0.92
A	-0.71	0.10	1.07	1.06

RMSE 0.10 Adj S.D. 0.61 Separation 6.08 Reliability 0.97

Fixed (all same) chi-sqaure: 76.8 d.f.: 2 sig: 0.00

表3 2009/05評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
C	0.46	0.12	0.92	0.90
B	0.33	0.12	1.10	1.06
A	0.12	0.12	0.84	0.84
E	-0.44	0.12	1.07	1.07
D	-0.47	0.12	1.09	1.09

RMSE 0.13 Adj S.D. 0.41 Separation 3.30 Reliability 0.92

Fixed (all same) chi-sqaure: 47.2 d.f.: 4 sig: 0.00

表4 2009/07評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
B	0.22	0.10	0.83	0.87
A	-0.04	0.10	0.86	0.84
D	-0.09	0.09	0.94	0.96
E	-0.09	0.09	1.25	1.30

RMSE 0.10 Adj S.D. 0.11 Separation 1.15 Reliability 0.57

Fixed (all same) chi-sqaure: 6.8 d.f.: 3 sig: 0.08

表5 2009/10評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
D	0.18	0.08	1.19	1.17
C	0.05	0.08	0.92	0.92
A	-0.09	0.08	0.96	0.95
B	-0.14	0.08	0.91	0.91

RMSE 0.08 Adj S.D. 0.12 Separation 1.57 Reliability 0.71

Fixed (all same) chi-sqaure: 10.4 d.f.: 3 sig: 0.02

表6 2009/12評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
D	0.14	0.08	1.14	1.14
B	0.03	0.08	0.82	0.82
A	0.00	0.09	0.81	0.81
C	-0.17	0.08	1.18	1.18

RMSE 0.08 Adj S.D. 0.10 Separation 1.28 Reliability 0.62

Fixed (all same) chi-sqaure: 8.6 d.f.: 3 sig: 0.04

三、評分者嚴格度的變化

由於2009年10月與12月的預試使用相同一套試題，研究者隨機挑選20名10月份考生口語錄音檔案，安插至12月份的評閱資料中，給評分者進行評分，藉由共同考生達到資料連結的目的。再將四位評分者二次的評分結果視為八位不同評分者的評分結果（如：評分者A分為A_10與A_12），對10月與12月二次評分結果進行同時估計，以比較四位評分者前後二次評分的嚴格度變化情形。評分者嚴格度的變化如表7及圖2所示，四位評分者中，評分者A與B評分嚴格度較先前一次嚴格，評分者A嚴格度由原先的-0.11提高為0.01，評分者B由-0.16提高為0.04；評分者C則是變得較為寬鬆，嚴格度由0.04降為-0.15；評分者D嚴格度的變化最小，僅降低0.04。

此結果和Lumley與McNamara（1995）以及Bonk與Ockey（2003）的發現一致，部分評分者嚴格度會隨時間改變，且變化的方向不同，有些評分者變得較嚴格，有些評分者則變得較寬鬆。本研究最後二次評分訓練與評分僅間隔二個月，評閱同一套試題，仍然有部分的評分者嚴格度產生變化，雖然變化幅度並非很大，但仍突顯出評分者很難維持相同的嚴格度。

表7 評分者二次評分嚴格度變化

評分者	2009/10	標準誤	2009/12	標準誤	變化情形
A	-0.11	0.08	0.01	0.08	0.12
B	-0.16	0.08	0.04	0.07	0.20
C	0.04	0.08	-0.15	0.07	-0.19
D	0.18	0.08	0.14	0.07	-0.04

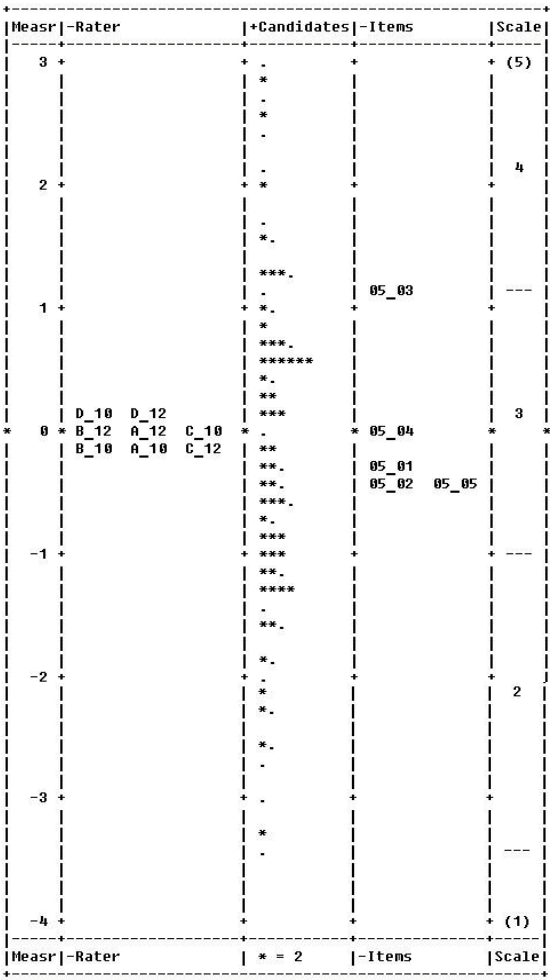


圖2 FACETS各面向對照圖

伍、結論

本研究結果顯示，密集且長期的評分訓練模式雖然有助於降低整體評分者嚴格度的變異，但無法完全消除評分者個別的差異，評分者間的一致性仍有待提升。但在評分訓練後，評分者內的一致性均符合0.7-1.3的範圍，顯示評分訓練對於評分者維持自身給分標準是有幫助的。此外，評分者在不同時期的嚴格度可能產生變化，不一定會維持同樣的嚴格度。上述研究發現顯示出評分者因素對於考生成績確實有影響，未來宜朝向以MFRM估計出考生能力參數後，再進行分數的轉換，取代傳統的原始分數，以提供更可靠的成績。因為謹慎挑選評分者以及密集的評分訓練仍然不能充分達到使評分者給分完全相等的目標，使用MFRM對於評分員給分的嚴格度進行統計上的調整，應是更為適當的作法。

本次研究參加的評分者人數較少，且因長期培訓，評分教師有時因個人因素無法每次皆參與，是較為可惜的地方。未來若能再增加評分者的人數，或許可以更明顯看出評分訓練對於評分者嚴格度的影響，當人數增加時，嚴格度的落差仍然與先前相當，或是變得更大。

此外，研究者觀察到9月第三次評分訓練之後接連三次的評分結果，無論是嚴格度差異、separation index或是reliability index的數值，均達到一個較為穩定的狀態，這也許表示評分者參加二到三次的評分訓練後，對於評分原則的掌握有了更清楚的瞭解，也更勝任評分工作。若是如此，未來新進評分者，可能需要先具備二到三次的評分訓練經驗，才較為適合擔任正式評閱的工作。而這些評分者若再繼續參與評分訓練，彼此之間的嚴格度是否會更接近，或是維持現狀，亦是研究者所好奇的，未來可持續加以觀察。

本研究結果提供華語文口語測驗的實徵資料，對於華語文口語能力測驗的評分訓練效果有初步的瞭解，也能做為未來進行華語文口語能力測驗評分訓練及測驗實施的參考。

參考文獻

田清源（2007）。HSK主觀考試評分的Rasch實驗分析。**心理學探新，第27卷**，第1期，65-69頁。民國99年7月5日，取自「中國期刊全文數據庫」(DOI: CNKI:ISSN: 1003-5184.0.2007-01-013)。

羅丹（2008）。**多面Rasch模型在HSK（中級）口語評分檢驗的應用**。北京語言大學課程與教學論碩士論文，未出版。民國99年7月5日，取自「中國優秀碩士學位論文全文數據庫」(DOI: CNKI:CDMD:2.2010.046217)。

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion. *Language Testing, 20*(1), 89-110.

Du, Y., Brown, W. L., & Rogers, C. (1997, March). *Raters and single prompt-to-prompt equating using the Facets model in a writing performance assessment.* Paper presented at the Ninth International Objective Measurement Conference,Chicago, IL.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages; Learning, teaching, assessment (Section H).Strasbourg, France: Council of Europe/Language Policy Division.

Engelhard, G. (1992). The measurement of writing ability with a Many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Linacre, J. M. (2010). *A user's guide to Facets: Rasch-model computer program (Version 3.67).* Chicago: Winsteps.com

Liu, Y. L., & Wen, S. M. (2007). *Rating Reliability on the Assessment of Speaking Performance.* Proceedings of English Education and Inter-Discipline Learning, Shih.Chien University, Taipei, 408-427, April, 28-29.

Lumely, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54-71.

Lynch, B. K., and McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL&Applied Linguistics, 4*(1), 1-21.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72-92.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15* (2), 263-287.

透過概念圖建立國高中科學課程之共同架構

譚克平 / 陳昭錦

透過概念圖建立國高中科學課程之共同架構

譚克平
國立臺灣師範大學科學教育研究所教授

陳昭錦
國立臺灣師範大學科學教育研究所博士生

摘要

國中九年一貫與高中課程綱要的銜接性是當前教育實務中的重要議題，如果課程設計能有妥適的銜接，將能使學生以先前的學習經驗為基礎，繼續發展對重要概念更加深及加廣的理解。就自然科學課程現況而言，國中的「自然與生活科技」領域是以能力指標的方式，藉以對學生應習得的能力作出要求，而高中則是將基礎科學區分為物理、化學、生物及地球科學四個科目，以內容綱要的形式呈現，兩個學習階段存在顯著的不對應情形。為改善現行國、高中科學課程的銜接性問題，本研究的目的是針對現行九年一貫國中自然與生活科技課程綱要，以及99學年度開始實施的高中基礎化學課程綱要，建議可以概念圖為工具，建立兩者之間的一個共同架構，使國中與高中科學課程的教與學，在銜接性與連貫性等方面能有妥適的參考依據。

本研究進一步以物質科學中兩個核心概念「物質組成」與「物質結構」為實例，說明如何透過概念圖建立國、高中科學課程的共同架構，並繪製不同學習階段的概念圖。透過概念圖可以呈現概念之間的階層與邏輯關係，一方面可協助國中教師掌握課程中哪些概念將會在高中的學習扮演重要角色，另一方面可協助高中教師在教學時，掌握學生應有的先備概念，並留意先備概念與待學概念之間的銜接關係，以便有效協助學生對於核心概念建構更深層且完整的理解。

關鍵字：科學課程、概念圖、課程銜接

Using concept map as a tool to align the junior and senior high school science curricula in Taiwan

Hak-Ping Tam
Professor, Graduate Institute of Science Education, National Taiwan Normal University
t45003@ntnu.edu.tw

Chao-Chin Chen
Doctoral student, Graduate Institute of Science Education, National Taiwan Normal University

Abstract

The vertical alignment between the Grade 1-9 Curriculum Guidelines and the Curriculum Standards for the Senior High Schools is among one of the core issues that demands attention by educational practitioners. A good alignment can greatly facilitate students' development of a deeper and broader understanding of important concepts based on their prior learning experiences. However, there is currently a big difference between the science curricular for the junior and the senior high schools, with the former organized in terms of integrated sciences and daily life technology while the latter in the form of four individual science subjects, namely, physics, chemistry, biology, and earth science. The purpose of this study is to introduce a powerful strategy regarding how the alignment framework can be constructed by way of a careful and thorough application of the concept mapping technique.

This study takes two core concepts from materials science, namely, composition of matters and the structure of matter, as an example to illustrate how a common framework can be established by means of concept mapping. Towards this purpose, concept maps for different learning stages will be drawn and compared. Since concept maps reflect the hierarchical and logical relationships among the relevant materials science concepts, it can help junior high school teachers to grasp which concepts will play an important role in the senior high school curriculum. On the other hand, it can help senior high school teachers to identify prerequisite concepts that should be acquired by students in junior high school. If teachers can be aware of the connection between the prior knowledge and the concepts to be learned, they can assist their students to attain a more effective and deeper understanding of the core concepts.

Keywords: concept map, curriculum alignment, science curriculum

壹、緒論

一、前言

人類的學習是累積漸進的過程，單一的學習經驗很難對學習者產生深遠的影響。無論是思考方式、基本習慣、概念、態度以及興趣等方面的改變，除了一些較特殊的經歷之外，通常都是逐漸發生的，需要在很多教育經驗不斷累積的情況下，才可能在學習者身上觀察到顯著的變化。因此課程設計者如果能妥善組織學習者的經驗，俾使該等經驗能夠透過彼此之間的關係相互增強，藉以有效提升學生的學習效能（Tyler, 1949; Ornstein & Hunkins, 2009）。相反地，如果不同階段的學習經驗缺乏妥適的銜接，可能導致某些重要概念在學生背景不足的情況下即開始教導，有些重要概念應該先複習以利加深加廣的學習但卻被忽略掉，而較不重要的概念卻重複出現的窘境，這不僅浪費寶貴的學習時間，也會有降低學生學習效果以及學習興趣的風險。

國民中小學九年一貫課程綱要自民國91年實施至今已邁入第十年，至於高中課程方面，自民國95年開始實施普通高級中學課程暫行綱要（教育部，2004），而99學年度則開始實施普通高級中學課程綱要（教育部，2008），這兩項高中的課程綱要中均明白指出，高中課綱應該延續九年一貫課程的精神。此外，關於高中教材之設計與編寫，亦應以中小學一貫課程體系參考指引為依據，並注意其與國民中小學九年一貫課程的銜接，組成可誘導學生學習興趣與發揮其潛能之科學教材。

儘管高中課綱中有強調與九年一貫課程的銜接，然而在現今的教科書審定制度下，各出版商的教科書編輯未必能認同或體現課程銜接的重要性。以科學課程現況來說，九年一貫自然與生活科技領域的課程綱要是以綜合科學（integrated sciences）的方式作整理，未採分科安排，並且以撰寫能力指標的方式來呈現，但在普通高級中學必修科目課程綱要中，卻將自然科學區分為基礎物理、基礎化學、基礎生物及基礎地球科學等四科，採主題方式呈現，並且未附能力指標。換言之，高中基礎科學的四個科目之課程綱要，並未直接與國中自然與生活科技領域互相銜接或對齊。從結構的角度而言，國、高中學習階段在課綱規定上已經是各自為政，在缺乏充分的溝通機制之下，要期待教科書的編寫者注重課程的銜接性，妥適地設計教材，在實務上存在著實質的困難。

現階段國、高中課程規劃，在課程設計的連貫性與銜接性等原則方面並不十分理想，導致高中生的學習經驗無法與國中時期有妥善的銜接，可能會引發一些學生及教師兩方面都需要面對的問題，茲以自然科學科目為例，分述如下。在學生方面，他們從國中修習單一的「自然與生活科技領域」科目，至升上高中之後，所要面對的卻是每學期至少修讀基礎科學四門學科中的兩門課程（各兩學分，每學期總計四學分，上下學期合計八學分），甚至有部分高中的安排是高一上、下學期均同時修讀四門基礎科學（各一學分，每學期總計四學分，上下學期合計八學分），因此學生面臨的挑戰，是要在最短的時間將國中三年的「自然與生活科技領域」學得

的經驗，分別對應到各種不同的基礎科學課程，並建立適當的聯結。例如：高一上修基礎化學時，學生應瞭解哪些概念是國中自然與生活科技曾經學過的，然而問題是學生能自行掌握這些概念嗎?另一方面，對高中自然科教師的挑戰則是，該如何掌握學生既有的、且未必有明確學科知識結構的先前習得概念？以及該如何引導學生把先前的概念，適當地應用到學科分界明確的高中學習體系當中？例如：任教高一基礎化學的教師，應知道哪些概念是學生在國中自然與生物科技曾有初步的涉獵，哪些概念是以國中的概念為基礎繼續延伸，哪些概念在國中完全未被提及，是在高中初次介紹等等。而這些分析，高中的自然科學課程綱要中並未詳細敘明，因此需要高中任課教師透過分析與比較國中自然與生活科技和高中課綱後，才能適切掌握。

由此觀之，為了協助教師以及學生克服目前面臨的挑戰，實有必要分析現行的國中與高中科學課程綱要內容，尋找適當的工具及方法，以建立可以兼納兩個學習階段的共同架構。

二、研究目的

本文的目的，是以現行九年一貫國中自然與生活科技課程綱要，以及99學年度開始實施的高中基礎化學課程綱要中某一概念範圍為例，介紹一套有效的方法，以建立這兩個學習階段的一個共同架構，使國中與高中科學課程的教與學，能有妥適的銜接性與連貫性，可作為日後教材編寫者參考之依據。

科學教育工作者Novak（1990）曾指出，概念圖能呈現眾多概念之間的連結以及有意義的階層與邏輯關係，它是學生學習科學概念的有效工具之一，教師應用概念構圖不僅能掌握學生的先前概念結構，也能作為規劃教學的有效媒介。為了達成前述國中與高中科學課程銜接之目的，本研究建議可使用概念圖為工具，發展國、高中科學課程的共同架構，協助教師、學生以及教科書編寫者們瞭解兩個階段課程共同的核心概念，彼此間之連結與銜接關係。

貳、文獻探討

由於篇幅所限，文獻探討將僅環繞與本文有關的課程組織原則與概念圖這兩方面做簡要介紹，分述如下。

一、課程組織的原則

泰勒（Tyler, 1949）曾以三項效標「連續性」（continuity）、「程序性」（sequence）以及「統整性」（integration），來檢視學習經驗設計的有效程度。這三個效標都是從學習者的觀點，以制定有效組織學習經驗的基本要素，其要旨簡述如下：

1.連續性：這是檢視課程是否有妥善的縱貫性組織當中的主要因素，意指對於課程中所包含的重要概念，予以「縱貫式」的重複敘述（vertical reiteration）。舉一個簡

化的例子，在國中階段可學習到物質組成的「粒子」概念，該概念在高中的基礎化學（一）與基礎化學（二）需要連續地出現，以使學生能充分掌握此一重要概念。

2.程序性：在學科概念的引介序列方面，應著重概念的內涵與應用層面的加深與加廣，透過適當的教學設計使概念的每一次處理，皆能協助學生對該概念建立更廣、更深的理解。如果高中的學習能適當地建立在國中所提供的經驗之上，那麼它們便能互相增強，學生在相關概念、技能等方面的發展，才能擴大其認知的深度與廣度。例如：國中階段對於「物質結構」的概念，主要著重在巨觀的觀察，對於微觀的原子結構只做概略性的介紹，而到了高中一年級，則深入介紹人類對於原子結構的探索歷程，並進一步指出原子結構在物質性質與交互作用當中扮演的重要角色。

3.統整性：這是指課程經驗的橫向聯繫（horizontal relationship）。學習經驗的組織務必做到能協助學生逐漸獲得統整的觀點，例如：高中階段在基礎物理及基礎化學課程中皆有探討「物質結構」的學習經驗，該等經驗能互相連結統整為更完整的認知理解。

二、概念圖簡介

概念圖始自上世紀六、七十年代，由美國Cornell大學的Novak教授首先提出。及至1984年，Novak教授和學者Gowin共同出版了名為“Learning how to learn” 的權威書籍，內文中他們根據Ausubel（1968）所提出的有意義學習（meaningful learning）的觀點，進而提出了概念構圖（concept mapping）學習策略，並深入探討如何運用概念圖改善科學概念的學習與教學，以達到有意義學習的目的。Gowin（1979）指出概念構圖是透過圖解的形式，呈現概念彼此間縱向與橫向的關係，並能展現學習者對特定主題相關概念的理解情形。文獻中有一些與概念圖相近似的觀念，例如知識圖，它們之間略有不同，本文以下討論將以Novak所提出的概念圖為依歸。

至於何謂概念，根據Novak和Gowin（1984）的意見，他們將概念視為是對事或物所能被感受到而且可被命名之常規性質（perceived regularity in events or objects designated by a label）。在概念圖中，每一個概念都以一個節點（node）的方式呈現。概念與概念之間，則透過連接詞（linking words）形成一個命題（proposition），也就是具意義並可以判斷真假的最小單位。此外，概念圖是具有階層結構的圖形，上層為一般化的概念，下層則為特殊化的概念，對兩個有連接線連結的概念而言，上層概念包含下層的概念。Novak和Gowin（1984）認為，當學習者繪製一個具階層性的概念圖時，他必須判斷概念間哪些具包含性，哪些具特殊性，這是一個學習者需要投入於思考之中的認知過程，從而增加學習相關概念的效果。透過分析學生繪製的概念圖，則可以瞭解學生對相關主題所習得的知識結構內容（例如參Edmondson, 2000）。

概念圖是一種有效的認知學習工具，概念圖之所以能被建構及繪製，其背後隱含一些基本假設，包括：學習者有先備知識的存在；人類會主動建構所經驗事物的意義；有意義的學習是指學習者將新概念與命題同化融合於既有的概念與命題網絡之內；知識是有組織及結構的；認知結構是可運用圖形工具來呈現其組織及內部關係（譚克平，2009）。

概念圖初期的用途主要用於展現學習者的起始概念、學習者經學習後的認知結構、評量學習者學習後的認知理解情形以及診斷迷思概念等等（Novak & Gowin, 1984; Novak, 1990）。有初步研究的結果顯示，概念圖對於學習者而言，具有減輕認知負荷、改善記憶並強化對知識結構的理解力（O'Donnell, Dansereau, & Hall, 2002），可運用在知識的提取與結構化等學習活動中。綜合來說，概念圖可以應用在課程規劃、教學活動設計以及學習成果的評量等層面。

概念圖在國際上的影響逐漸深遠，有學術團體每兩年舉辦一次國際研討會議，並先後在西班牙、哥斯大黎加、芬蘭及智利等地舉行。過去，概念圖的研究是以美國及加拿大的研究者為主，研究範圍集中於科學教育領域，及至目前，包括如英國、法國、荷蘭、希臘、義大利、墨西哥、南非及台灣等地，皆有研究者使用概念圖做為研究工具，研究範圍亦從科學教育擴展至其他領域。近年來更有協助繪製概念圖的軟體陸續被開發，其中，CmapTools軟體從網路上即可免費下載（參陳學志，2009）。

參、研究方法及過程

由於現行科學課程組織龐大，為方便聚焦，本研究首先以內容分析法來分析現行國中與高中的課程綱要，依據的是國中「自然與生活科技」學習領域之綱要及教材內容要項，以及普通高級中學必修科目「基礎科學」之課程綱要。然後選定主題，並以建構概念圖的技術，繪製國中及高中學習階段的概念圖，藉以彰顯該等概念的階層從屬關係，從中探討兩個學習階段相關概念的銜接性。

九年一貫「自然與生活科技」學習領域之綱要，是以「科學與科技素養」的分段能力指標來呈現，其中包括過程技能、科學技術與認知、科學態度等共計八個項目，其中「科學技術與認知」與教材內容密切相關，在課程綱要中有將其內容要項在附錄一陳列（教育部，2003），此內容要項是按課題、主題及次主題的架構呈現。以第一項課題之「自然界的組成與特性」為例，其中涵蓋地球的環境、地球上的生物、物質的組成與特性等三個主題，第一項主題可連結到高中基礎地球科學，第二項主題連結到高中基礎生物，第三項主題則是連結到高中基礎化學。在這種情況下，如要進行系統化的分析，從高中教學端的觀點，仍應從高中的學科角度出發。但受限於篇幅，本研究將只以基礎化學為示例，首先分析高中基礎化學課綱，以此為依據檢視九年一貫「自然與生活科技」學習領域之教材內容要項，擷取與高中基礎化學相關的課題、主題及次主題，篩選彙整如以下表1所示。

表1「自然與生活科技」學習領域之教材內容要項中與高中基礎化學相關的課題、主題及次主題篩選彙整

課題	主題	次主題
自然界的組成與特性	物質的組成與特性	＊物質的構造與功用
		＊物質的形態與性質
自然界的作用	改變與平衡	＊化學反應
		＊化學平衡
	交互作用	＊水與水溶液
		＊氧化與還原
		＊酸、鹼、鹽

另一方面，普通高級中學必修科目「基礎化學」的課程綱要是按主題、主題內容及應修內容的形式呈現，為能與前述九年一貫學習領域之主題與主題內容相對應，本研究將高中基礎化學綱要中應銜接自國中的主題篩選彙整如表2所示。結合表1及表2，可初步協助高中基礎化學任課教師瞭解哪些國中的教材內容要項與高中課程有關，在講授哪些單元時，需特別留意國中的銜接概念。另一方面也可使國中教師知道哪些內容要項是學生將來在高中會繼續學習的，可在教學過程中提醒學生在國中時打好學習基礎。

表2 普通高級中學「基礎化學」課程綱要中銜接「自然與生活科技」學習領域之主題篩選彙整

主題	主題內容	應修內容
物質基本組成	物質的組成	＊物質的分類　＊原子與分子　＊溶液
物質基本構造	原子構造	＊原子結構　＊原子中電子的排列
物質構造	物質的構造與特性	＊離子鍵與離子晶體
		＊共價鍵與分子化合物
		＊網狀固體　＊金屬固體
物質變化	化學反應	＊化學式　＊化學反應式與均衡
		＊化學計量　＊化學反應中的能量變化
	常見的化學反應	＊結合反應　＊分解反應
		＊酸鹼反應　＊氧化還原反應
	化學反應速率	＊反應速率定律　＊碰撞學說
		＊影響反應速率的因素
	化學平衡	＊化學平衡　＊平衡常數
		＊影響平衡的因素

比較表1及表2可發現，「物質組成」與「物質結構」為重要的核心概念之一，因此接下來將分別針對國中與高中的課程綱要，繪製出「物質組成」與「物質結構」兩個主題的概念圖，再進而分析兩學習階段結構上的異同。

至於概念圖的繪製，譚克平（2009）曾參考不同研究者的意見，歸納出一連串具體建構概念圖的步驟，本研究即依據其建議進行的步驟繪製概念圖（另外亦可參Cañas, & Novak, 2006），相關之步驟如下：

第一步：尋找相關文本
第二步：從文本篩選出重要的相關概念
第三步：將概念寫在卡片並置於紙上
第四步：選出最一般化概念作為最上層的概念
第五步：依從屬關係將比較特殊化的概念排在下面的層次
第六步：用線將上、下層相關的概念做聯結
第七步：填寫能表達出兩概念間聯結關係的連結詞
第八步：將有關係但在不同分枝上的概念做交叉聯結
第九步：填寫交叉聯結的連結詞
第十步：加入例子於最底下的層次，並視為最特殊化的概念
第十一步：不能納入的概念可置於一旁日後再作思考
第十二步：相同或非常相似的概念可放在一起，並加框線
第十三步：反思後再進行增修

本研究首先由兩位任教公立高中有17年化學教學經驗的教師，其中一位曾有5年的國中教學經驗，以及兩位有3年國中理化教學經驗的教師，依據上述第一步至第三步，將相關概念用國中或高中通用的科學用語，分別製作出國中及高中的概念卡片，接著四位教師共同檢視兩組概念的差異性，並針對不一致的部分進行討論後，才決定出要選擇哪些概念來繪製國中及高中的概念圖，接著即依前述其餘步驟共同完成概念圖。至於最後第十三步的反思，則由科教背景的學者進一步檢視概念的上下位階及連接詞的適切性，並與參與的教師共同討論後進行增修，最後完成國中及高中的概念圖。本研究接著依據繪製完成的概念圖分析比較國、高中不同階段的學習內涵差異，並據此比較之結果，建議日後相關單位進行課程與教學設計時，作為其應考慮項目的參考依據。

肆、研究結果

一、「物質組成」概念圖

1.國中與高中課程的「物質組成」概念圖如以下圖1與圖2所示。

2.比較圖1與圖2可知，關於「物質組成」的概念，國中階段著重於由巨觀角度出發的物質分類依據，而高中階段則深入到微觀層次，從物質的組成單位進一步細探不同物質類別之間的差異。學生在國中階段要能從巨觀的相態區別純物質與混合物，以及區別混合物中的均質混合物與非均質混合物。然而在國中階段，課程內容僅提及溶液，並未比較溶液與其他混合物的區別，透過圖1希望提醒國中教師，如能提示學生均質混合物（溶液）與非均質混合物的區別，並建立兩者均在混合物此一分支下的從屬關係，相信能使學生更清楚何謂溶液，也能有助於其日後在高中階段的學習。

3.有關純物質的分類，高中階段須進一步從微觀粒子世界的角度，建立原子與分子的概念，進而從化學鍵的差異區分化合物的種類，包括離子化合物與共價化合物等等，從概念圖的比較中，可清楚觀察到國中與高中階段共同的概念，高中新增的概念，以及概念之間如何銜接、延伸與加廣。

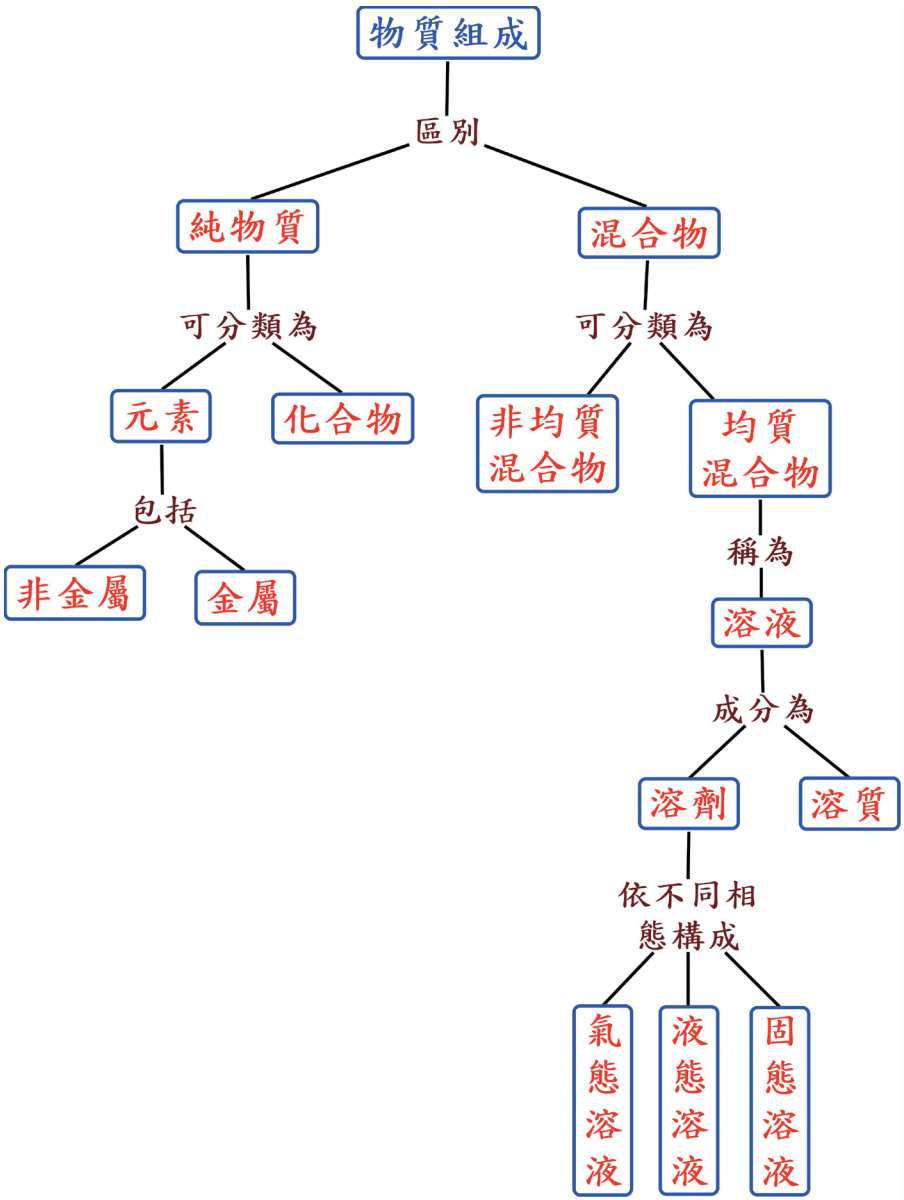


圖1 國中課程「物質組成」的概念圖

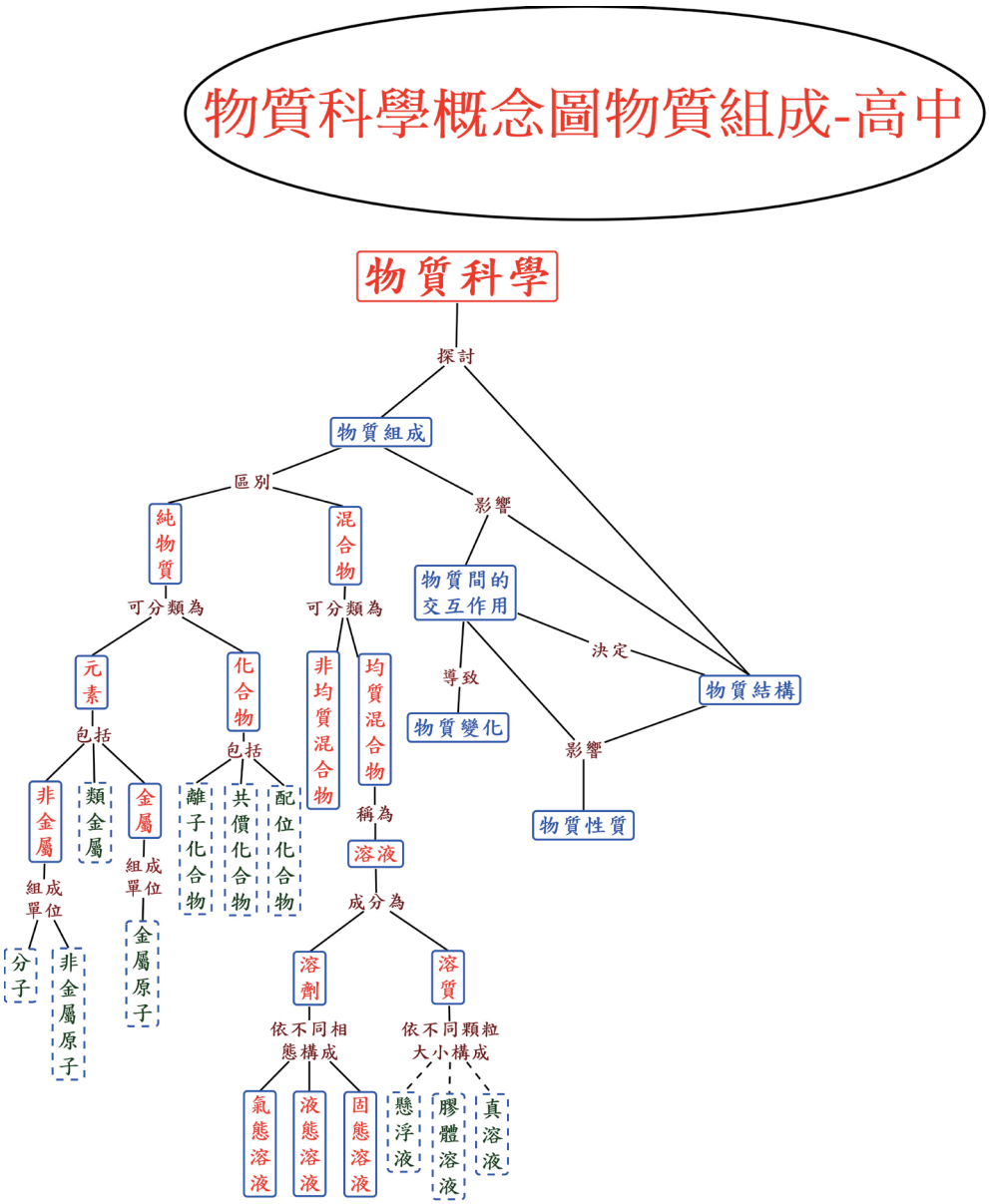


圖2 高中課程「物質組成」的概念圖（圖中實線框的概念為國中與高中重疊者，虛線框的概念則屬於高中課程特有者）

二、「物質結構」概念圖

1.國中與高中課程「物質結構」的概念圖如以下圖3與圖4所示。

2.比較圖3與圖4可知，關於「物質結構」的概念，國中階段分別呈現了巨觀及微觀層次的世界，關於巨觀的介紹較多，微觀的部分屬於基本概念的認識，然而現今國中教材中並未出現"巨觀"和"微觀"的科學詞彙。透過概念圖可提醒國中教師如果要協助學生日後在高中階段的學習，可視學生的接受程度引入這兩個重要概念，透過巨觀及微觀兩種尺度的比較，學生應可對物質結構有更清晰的認識。

3.在高中階段，有關物質結構概念的學習主要的延伸在於微觀層次，並增加了微觀層次與巨觀層次的聯繫。以圖3及圖4之國、高中的共同概念“固態”為例，國中只從巨觀層次介紹固態的特徵為“固定的形狀和體積”，然而如果要能與高中的微觀層次有適切的銜接，建議國中階段可稍作延伸至微觀層次，使國中學生能理解從粒子堆積是否散亂或有規律的觀點，可區別出晶形固體與非晶形固體。到了高中階段，學生學習了化學鍵的基本概念後，結合化學鍵的概念，可進一步探討晶體（巨觀）依結合方式（微觀化學鍵）之不同可分類為離子晶體、分子晶體等，並能據此解釋這些不同結合方式之晶體其物理性質的差異。此外高中延伸的概念所佔比例相當高，當中又涉及化學鍵及物質的交互作用等重要觀念，對高中生而言，物質結構是重要卻又不易理解的主題，透過國中與高中概念圖的呈現，提醒高中教師檢視學生在國中階段已習得的先備概念掌握情形，以協助學生能妥適地從巨觀世界轉換至微觀的觀點。

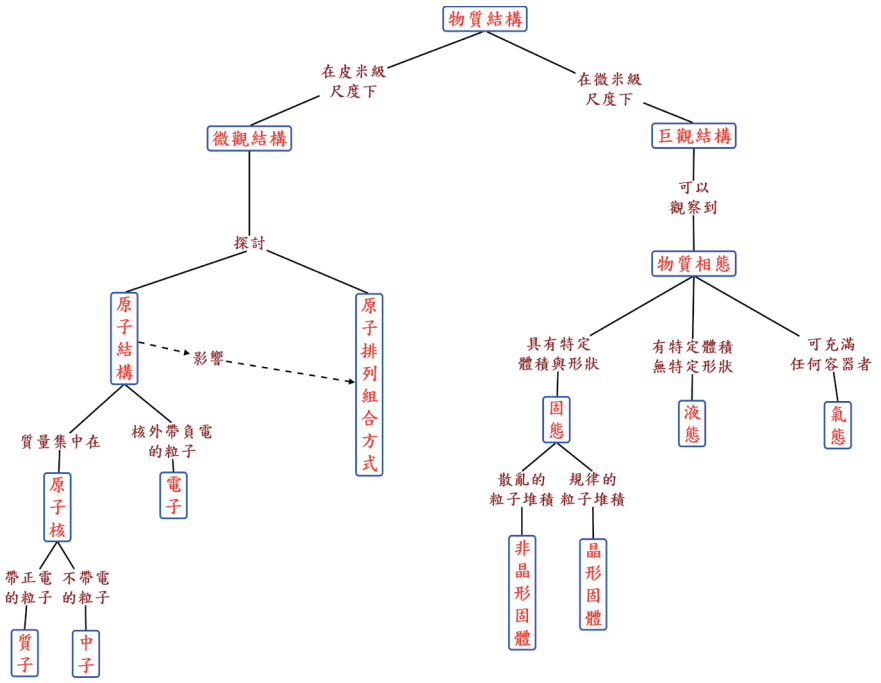


圖3 國中課程「物質結構」概念圖

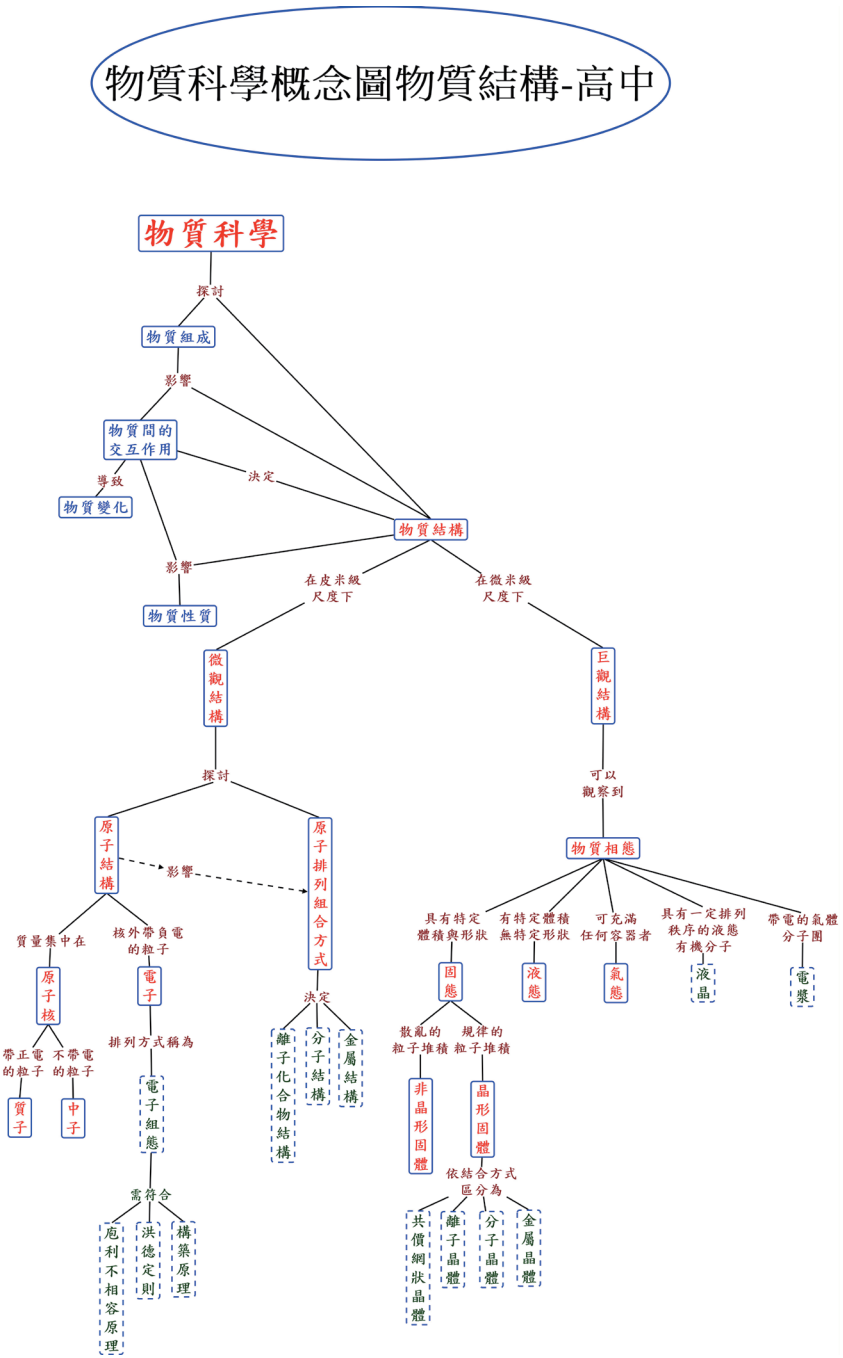


圖4 高中課程「物質結構」概念圖（圖中實線框的概念為國中與高中重疊者，虛線框的概念則屬於高中課程特有者）

伍、討論與建議

本研究以物質科學中的核心概念「物質組成」與「物質結構」為主題，說明如何透過概念圖以展現國、高中相關科學課程的架構。以下將討論以概念圖達成本研究目的之優點與缺點。

首先，透過概念圖可以清楚呈現概念與概念之間的階層與連結關係，有助國中教師掌握課綱中所列的概念對於學生日後升上高中繼續學習時，它們將會有何種地位、重要性以及延伸關係，如果國中老師能在教學過程中適時提醒學生，哪些重要概念是日後高中學習的基礎，相信此舉能讓學生在國中階段便能注重這些概念的學習，日後接觸高中課程時，也能以先前概念為基礎來學習新的延伸或擴充的概念。另外，高中教師可以透過概念圖的分析，瞭解學生在國中階段已習得哪些概念，進而思考如何協助學生將已習得概念與將要學習的概念銜接起來。高中教師一方面可先利用概念圖診斷學生先備概念的內涵，另一方面也能在教學中，提示學生過去所學與目前待學內容之間的聯繫情形。由是觀之，概念圖能發揮到連續性的功能，並起複習的作用（Tyler, 1949）。

再者，若能針對同一主題分別繪製國中及高中課程的概念圖，將可以方便比較不同學習階段概念涵蓋的範圍、難易程度以及學習的先後序列。對高中老師而言，透過概念圖的分析能適當地整合學科概念之間的關係，在備課的過程中，即能據此安排各概念引介的順序，發揮出Tyler（1949）所提之程序性的功能，能藉由加深加廣先備概念的方式來教導新概念。因此概念圖能有效檢視特定主題教學內容的完整性，以免遺漏重要概念，或者是因為忽略了先備概念而讓學生的學習效果受影響。

此外概念圖是有效的認知學習工具，教師若能適當教導學生以概念圖呈現所習得的內容知識，將有助於教師進行學習診斷，以及學習者進行自我檢視。在帶領學生繪製概念圖之前，如果教師本身有足夠的繪製與運用概念圖經驗，將更能把概念圖的優點及繪製要領適切地傳達給學生。因此，概念圖不僅可以扮演教師規劃課程及設計教學的利器，也能成為師生共同探究學習主題時的有效溝通工具。

然而，概念圖的運用並非毫無缺點。對於繪製概念圖的新手而言，他們或許會認為運用此工具繪圖相當費時，對概念的列出並不容易進行取捨，而且還有可能遺漏了重要的概念。除此之外，概念之間的階層關係亦不一定容易被界定。再者，如何選擇適當的連接詞以呈現概念之間的邏輯關係，亦有可能為一大挑戰。儘管如此，概念圖對老師的挑戰正是其功能所在。正因為概念圖能驅使教師更深入思考教學主題所涉及各概念之間應該如何適當連結，以及該如何將這些連結關係透過教學來體現，因此雖然教師在備課過程中要費心繪製概念圖，卻能藉此對其教學內涵有更深刻的理解，進而提升教學效能，反而能達到事半功倍的效果。

本研究以基礎化學為例，具體呈現高中教師可以如何透過概念圖這項工具，建立國、高中基礎化學相關主題的共同架構，期能藉此幫助教師掌握概念圖的應用要

領，提醒教師關注國、高中課程銜接的重要性，至於所提出具體落實的作法，希望能有助於教師的專業成長。

參考文獻

教育部（2003）。**國民教育九年一貫課程綱要**。台北市：教育部。

教育部（2004）。**普通高級中學課程暫行綱要**。台北市：教育部。

教育部（2008）。**普通高級中學課程綱要**。台北市：教育部。

陳學志（2009）。**概念圖的原理：語義網路與命題敘述**。各教育階段奈米及能源科技課程概念圖建構研習營研習資料。Available at http://140.112.65.252/home/project/100106/2.ppt

譚克平（2009）。**概念圖的製作及在科教領域中之應用**。各教育階段奈米及能源科技課程概念圖建構研習營研習資料。Available at http://140.112.65.252/home/project/100106/2.ppt

Ausubel, D. (1968). *Educational psychology: A cognitive view.* New York: Holt, Rinehart, and Winston.

Cañas, A. J., & Novak, J. D. (2006). Re-examining the foundations for effective use ofconcept maps. In A. J. Cañas & J. D. Novak (Eds.), *Concept maps: Theory,methodology, technology. Proceedings of the Second International Conference on Concept Mapping* (Vol. 1, pp. 494-502). San Jose, Costa Rica: Universidad de CostaRica.

Edmondson, K. M. (2000). Assessing science understanding through concept map. In J. J.Mintzes, J. H. Wandersee, & J. D. Novak, (Eds.) *Assessing science understanding: Ahuman constructivist view.* San Diego: Academic Press.

Novak, J. D. (1990). Concept maps and Vee diagrams: Two metacognitive tools for science and mathematics education. *Instructional Science, 19*, 29-52.

Novak, J., & Gowin, E. (1984). *Learning how to learn.* Cambridge: Cambridge University Press.

O'Donnell, A. M., Dansereau, D. F., & Hall, R. H. (2002). Knowledge maps as scaffolds for cognitive processing. Educational *Psychology Review, 14*(1), 71-86.

Ornstein, A.C., & Hunkins, F. P. (2009). *Curriculum: Foundations, principles, and issues* (5th ed.). Boston: Allyn and Bacon.

Tyler, R. (1949). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

【致謝辭】

本文感謝行政院國家科學委員會予以部分經費補助，計畫編號為NSC-98-2514-S-003-013-MY2。

國家圖書館出版品預行編目(CIP)資料

永續教育發展-創新與實踐論文集：國際學術研討會-測驗及評量論文專輯. 2010年 / 林宜臻等著. -- 初版. -- 新北市：國家教育研究院, 民101.02
面；　公分
ISBN 978-986-03-1573-8(平裝)

1.教育測驗 2.教育評量 3.文集

521.307　101000638

書　名：「永續教育發展-創新與實踐論文集」2010年國際學術研討會—測驗及評量論文專輯
著　者：林宜臻、張芳全、蔡孟憲、林原宏、盧思丞、涂柏原、謝名娟、謝進昌、藍珮君、譚克平、陳昭錦合著
出版機關：國家教育研究院
地　址：新北市三峽區三樹路2號
網　址：http://www.naer.edu.tw/
電　話：（02）8671-1111
出版年月：民國101年2月
版次：初版
其他類型版本說明：無
定　價：新臺幣170元
展　售：政府出版品展售中心
五南文化廣場：臺中市中山路6號
電話：04-22260330；傳真：04-22258234
網址：http://www.wunan.com.tw/
國家書店松江門市：臺北市松江路209號1樓
電話：02-25180207；傳真：02-25180778
網址：http://www.govbooks.com.tw

GPN:1010100108
ISBN：9789860315738

版權所有・翻印必究