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。

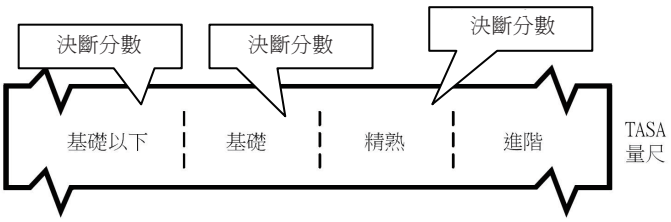

圖1 基礎以下、基礎、精熟及進階等四個能力區塊

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]

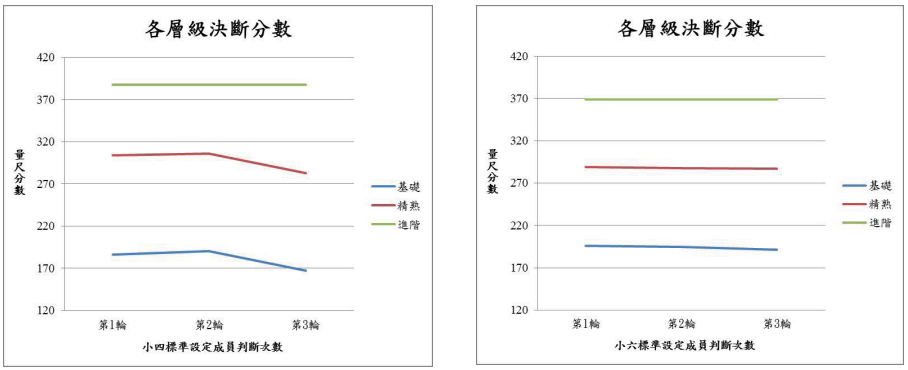

圖2 TASA2009數學科小四（左圖）、小六標準設定成員於三輪設定結果趨勢圖

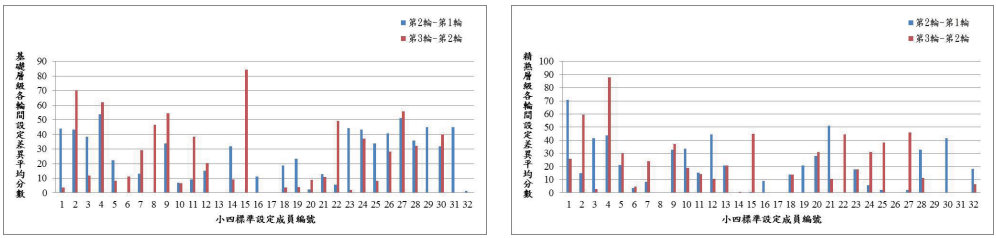

圖3 小四標準設定成員於基礎（左圖）、精熟層級各輪之間設定差異變化

[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。

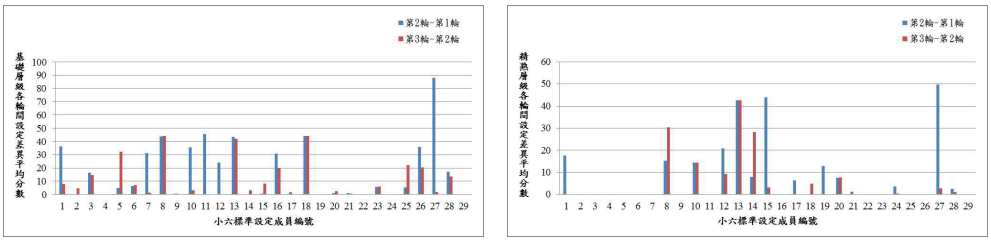

圖4 小六標準設定成員於基礎（左圖）、精熟層級各輪之間設定差異變化

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。