國立臺灣師範大學教育心理與輔導學系
教育心理學報，2013，44 卷，4 期，793-811 頁

以多層面 Rasch 分析的角度來評估標準設定之變異性*

謝名娟
國家教育研究院
測驗及評量研究中心

執行標準設定時，研究者常需檢視標準設定成員之間的變異性。研究者希望標準設定成員在設定切斷分數時，彼此之間對於決斷分數的判斷能夠達到一致性，也就是成員之間的變異性越小越好。此外，在舉行標準設定時，成員之間爲了要能夠達到共識，必須經過好幾輪的討論。因此，如何提供有用的訊息回饋，供標準設定成員參考，以節省成員的討論時間，也是研究的重點。在本文中，將多層面 Rasch 分析應用於標準設定上，並使用 Yes/No Angoff 的方法來進行操作，研究發現透過多層面 Rasch 之分析，除了能檢視標準設定成員之間的變異性，也可提供有用的訊息供標準設定成員參考。除此之外，對於標準設定成員自身內的衝突決定亦能提供有效率的檢視。

關鍵詞：多層面 Rasch、標準設定、Yes/No Angoff

國中基測的概念源自常模參照測驗，其主要內涵爲著重個人與團體內其他成員的比較，藉由與其他競爭者的分數，來瞭解個人表現的優劣程度。而十二年國教推出的國中會考，則是源自於標準參照測驗，其主要目的是將測驗分數與事先設定好的標準做比較，學生透過表現標準的描述來瞭解自身的學習成就，不需要與他人比較，國中會考將學生分成三個等級-精熟、基礎、待加強。然而，要怎麼把考試的成績切成三段呢？這就要透過標準設定了。

標準設定主要的目的爲找出幾個切斷分數將學生分成數個等級。依據測驗的目的，透過標準設定所找出的決斷分數會對個人、學校或社會造成不同層度的影響。例如國中會考，若將精熟的分數設的越高，則代表學生能達到精熟水平的比率會越小。而對於高風險考試而言，決斷分數的正確設定與否更會造成嚴重的後果。例如醫師執照考試，若是通過的切斷分數設定過高，會造成一些具有能力的醫學系學生無法通過考試。但若切斷分數過低，則會造成能力不佳的醫學系學生拿到醫生執照，間接危害病人權益，因此是否能夠在標準設定的過程中，找出合適的切斷分數是相當重要的。

* 本篇作者之通訊方式 hm7523@hotmail.com

標準設定主要是將表現的標準和測驗分數做連結，透過標準設定成員專業的判斷，找出符合標準要求的切斷分數。雖然這個過程看似簡單，但相當費時、費力，由於過程中牽涉許多主觀性的判斷，因此研究者常面臨外界對於所設定出之標準分數合理性之質疑。首先，就標準設定成員的決定上，須先找出具有代表性的一群人（Cizek & Bunch, 2007）。若是設定的分數攸關學生成績，則標準設定的成員可以是學生家長、學校老師、校長、領域相關的大學教授等，透過不同的成員表達出不一樣的聲音，使得所設計出的判斷分數能更加周延。

在實務應用上，要能得到標準設定成員間的共識相當困難（Raymond & Reid, 2001），每個成員都來自不同的背景，要能夠透過標準設定的議程來說服別人自己的看法，或是妥協自己原先的意見，須經過不斷的討論，整個過程費時費力。而且在整個過程中，須不斷的檢視標準設定成員在設定決斷分數時的歧異性，以了解是否透過討論、成員能逐漸達成共識。

因此，對於經費、時間有限的機關團體，如何能提供標準設定成員有用的回饋訊息、以提高在標準設定時討論的效率是非常重要的。一般來說，提供給成員的回饋訊息可歸爲二類（Cizek & Bunch, 2007），第一類爲常模（normative）的回饋數據，包括每一位標準設定成員所設定的切斷分數、切斷分數的極端值分布、平均數、中位數、標準差等。第二類的回饋訊息則是影響（impact）層面的回饋數據，這類的回饋訊息，可以提供成員設定判斷分數的可能造成的影響訊息。例如就過去的研究數據來找出，若是設定這樣的切斷分數，在每一個層面受試者通過的比率是多少，成員們可以就這方面的訊息來討論這樣的切斷分數對於社會大眾的接受程度與觀感爲何。第三類訊息則爲真實性（reality）的回饋訊息。例如測驗試題中，每一個題目的難易度、鑑別度，或是其他類似測驗的切斷分數等。這二種回饋訊息爲最常在標準設定中使用的回饋訊息。

在實務應用上，這些回饋訊息並無法對於標準設定的過程有立即性的幫助，也就是說，即使知道某些成員所設定的切斷分數是極端值，但是在計算最終決定的判斷分數時，是否應該將極端值刪除？此外，對於某些標準設定方法，如 Yes/No Angoff 方法，即使能檢視題目的難易度，但是成員並不知道該如何修正原先的判決，而只能逐題討論，而由於題目眾多，成員意見要能達成共識需要時間，因此常常受限於時間，成員無法討論太多題目。因此對研究者而言，如何能夠提供成員一個比較直觀的、立即性的回饋訊息，是相當重要的。

在本文中，將應用多層面 Rasch 分析 TASA2009 國小六年級英語科標準設定上，並採用 Yes/No Angoff 方法。本研究主要研究目的有二，第一，使用多層面的 Rasch 分析，來評估標準設定過程中成員彼此之間與個體內判斷的歧異性。在本研究中，將檢視成員若來自不同的職業背景，所判斷出的決斷分數是否有歧異性與個體成員在進行標準設定時，是否有發生衝突判定的現象，例如，判定同一層級的受試者能答對困難的題目卻無法答對簡單的題目。第二，如何使用多層面 Rasch 分析來提供標準設定成員可能有用的回饋訊息，進行判斷決定的修正，或藉此協助研究者，評估標準設定成員設定決斷分數時過程的品質？

根據研究目的，主要的研究問題如下：

1. 標準設定成員來自不同的職業類別與性別，彼此之間背景的不同，是否會造成標準設定判斷時的歧異性，這種差異是否達到顯著？

2. 標準設定成員在做判斷時，是否發生內部衝突的現象？要如何使用多層面 Rasch 分析將此現象偵測出來，形成回饋訊息以供成員與研究者參考？研究者如何能藉此訊息，來評估標準設定成員在執行判斷時的品質?

以下先就本研究所使用 Yes/No Angoff 標準設定方法之文獻進行探討，並對多層面 Rasch 模組之理論背景進行說明，最後則呈現研究方法、數據分析與結論建議。

文獻探討

一、Yes/No Angoff 方法

現今常用的標準設定的方法有五十幾種（Berk, 1996），在這些方法中，基本上的流程都是要求標準設定成員先參加設定方法的訓練，等到成員已對整個標準設定的過程熟稔，再進行正式的標準設定會議，在會議中，成員們除了需要自行判斷之外，也需要不斷的和其他成員討論，而在整個過程中，會議主席需提供相關回饋訊息，供成員在討論中參考，當作修正判定的依據。當最後的成員間的彼此歧異性降到可接受的範圍，或已經達到最終討論的輪次時，則以成員的判斷分數之平均，或是中位數，當作最終的決斷分數。

國內對於標準設定的相關研究眾多，大多的研究著重於理論方面的研究（吳裕益，1986；林惠芬，1993；鄭明長、余民寧，1994；謝進昌，2005），或是信效度的研究（吳毓瑩等人，2009；陳彥名，2006；吳宜芳、鄒慧英、林娟如，2010；謝進昌等人，2011）。過去研究的標準設定執行方法中，Angoff 方法爲最常使用的標準設定法之一，此方法爲 Angoff（1971）所提出，Angoff 方法使用上很簡單易懂，而且能夠輕易爲不同形式的題型設定決斷分數。這種方法廣泛應用在各類測驗上，如 NAEP、美國各州政府評量（Council of Chief State School Officers [CCSSO], 2001）、瑞典的數學考試（Näsström & Nyström, 2008），醫學等證照考試。

在原先的方法中，標準設定委員必須對題本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，可以答對這個題目。也可以想成在一群人中，有多少比率的人可以答對此題，而把這個比率，做爲邊緣受試者答對此題的機率。若是使用原先的 Angoff 方法，除了要對每一題進行判斷之外，委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。

因此，Angoff 方法就產生出了許多修定的版本，其中一個廣泛使用的版本爲 Impara 與 Plake（1997）所設計的 Yes/No Angoff 方法。此方法和 Angoff 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，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能夠或是不能夠答對此題。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"Yes"，如果不能答對，則寫下"No"。這種較爲直觀的判斷，減少了原先 Angoff 方法的執行困難度。

Yes/No Angoff 標準設定法的實際操作流程，大致如下，首先，研究者會事先提供每位標準設定成員一本試題卷，而每頁的試題內容包含有題目、選項、答案及評測項目等。而後則需逐題判斷是否在該程度的邊緣受試者，能否答對該題，逐題判斷後，研究者會將每一位委員的填答數據，輸入程式軟體中算出全部成員對於該水平的平均數。 算出的平均數，則可代表在此水平的邊緣學生，應該可以答對整份題本中的題目比例，並可以此做爲此水平的決斷分數的依據。

同樣的步驟會重複幾輪，而每一輪大致重覆第一輪的動作，但差別在於研究者會提供不同的回饋訊息，以作爲成員參考，例如在第一輪的回饋訊息中提供常模訊息，如其它成員（與自己）對各題的判斷之散布圖、決斷分數之分配圖等，而在第二輪回饋訊息中則提供影響訊息，即就過去的研究數據而言，若是設定這樣的切斷分數，在每一個層面受試者通過的比率是多少。成員即依據回饋訊息，分成小組來討論上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，進行下一輪的設定，並再次對各題重新判定。最後，研究者根據標準設定成員於最後一輪所判定的成果，來決定正式通過分數。

二、多層面 Rasch 模式

多層面 Rasch 模式延伸自單參數的試題反應模式，由 Linacre 在 1989 年提出，可以同時分析測驗中所存在的多個面向，並可分開呈現所估計出的結果。多層面 Rasch 模式具有試題反應理論

的優點，是古典測驗理論所無法具備的（余民寧，2009），例如，試題反應理論所計算出的試題參數（如難度、鑑別度等），不受樣本的影響，即使用不同的受試樣本，所計算出的試題參數也爲會相當穩定。且對於受試者的能力估計，不受測驗的影響，只要是同質性試題組成的測驗也能對不同受試者之間的分數進行有意義的比較。再者，試題反應理論所採用的適配度考驗值（goodness of fit index），可以提供考驗模式與資料間的適配程度、受試者的反應是否出現非尋常（aberrant）等參考指標。

多層面 Rasch 模式還有幾項特色可應用於標準設定上。第一，標準設定過程中最爲研究者困擾的議題就是判斷分數易受到標準設定成員主觀性的影響。然而，即使成員事先受過訓練，但還是只能盡量在主觀中求客觀，但是很難避免某些成員在判定時，使用嚴苛的角度，而某些成員，則採寬鬆的角度，傳統的單向度的試題反應理論無法將這種評分者的嚴苛程度進行考量。然而，多層面 Rasch 分析，將成員的嚴苛度放進模組中，並透過參數估計，使得最終的決斷分數，能夠考量成員間彼此不同的嚴苛程度（Stone, Beltyukova, & Fox, 2008）。第二，Rasch 適配度值可以用來偵測非尋常的判斷反應。例如學生程度可以分成基礎和精熟兩個層級。對於某些簡單的題目，標準設定成員判定精熟程度的學生無法答對，但是卻判定同一個題目，基礎程度的學生可以答對。或是判定基礎學生無法答對簡單的題目，卻可以答對另一題較爲困難的題目。這種不合常理的判斷，可使用適配度指標偵測出來。第三，多層面 Rasch 分析將成員的嚴苛度、試題難度、成員個人的背景因素等，都可放入同一個模組中。而將這些面向的資料同時進行估計校準至同一個的量尺中，因此，各個面向得到的參數值可以互相比較（Linacre, 1999）。

Kozaki（2010）曾嘗試探究將多層面 Rasch 分析，應用在標準設定上。在其研究中，需要爲日文與英文的翻譯執照員設定通過分數，由於這是屬於一種低風險的考試，且受限於時間、經費，Kozaki 無法召集所有的成員同時到某個場地執行標準設定，所以他透過郵寄的方式，將相關材料寄給成員，並請成員將設定好的決斷分數用郵寄的分式寄回，而後則使用多層面 Rasch 分析，來分析出成員在做決定時的歧異性，並計算出最後的決斷分數。然而，由於可能造成試題外洩，這樣的方式在一般高風險考試的應用並不可行。Engelhard（2009）近年來亦推行多層面 Rasch 分析於標準設定上的應用，並發展了客觀性標準設定法（Objective Standard Setting），在其文章中可看見多層面 Rasch 分析在標準設定上的可行性與優勢。然將多層面 Rasch 分析應用在常用的標準設定法中，例如 Yes/No Angoff 方法，則在過去文獻中較無著墨。

研究方法

一、標準設定成員

本研究中，標準設定成員的選擇來源以大學英語系教授、中央輔導團、縣市輔導團、行政人員（如校長，但具備英語科的教學經驗）並輔以學科團隊的推薦國小英語教師爲主。

標準設定成員總人數爲 32 名，其中教師占有 24 名（75%），行政人員 4 名（12.5%）、學者 4 名（12.5%），而性別分佈爲男性 6 人（18.8%）、女性 26 人（81.2%）。女性占大多數的原因是因爲大多數在國中的英語老師，還是以女性居多。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，標準設定成員職業類別具有代表性，包含教師、行政人員及學科學者，此外性別比例與整體教師分布類似。標準設定成員背景分布如表 1 所示。

表 1 標準設定成員背景資料分佈

身分類別	性別		總人數
	男	女	
教師	5	19	24
行政人員	1	4	5
學者	0	3	3
總人數	6	26	32

二、標準設定材料

臺灣學生學習成就評量資料庫(The Taiwanese Assessment of Student Achievement;簡稱 TASA)主要用來評估小四、小六、國二、高中職二的學生學業成就，TASA 包含五個考科: 國語文、英語文、數學、自然科學與社會科學。本標準設定所採用的材料，爲小六英語文。測驗的內容制定乃依據教育部公布之九年一貫課程綱要，和英語學習領域的能力指標爲依據。TASA 英語文建置目的，旨在評估學生英文學習成就表現，除了可用來檢視國內現階段英語課程實施效益外，也可以檢視學生英文學習成就。而隨著 TASA 英語文試題研發團隊的更替與評量架構之調整，再加諸過去沿用的標準已多不符合實務運用，引發修訂之需求。而研究團隊在此背景下，藉由標準設定，重新檢視臺灣學生學習成就評量資料庫之國民小學六年級英語文在每個表現層級的切斷分數。

TASA 小六英語文測驗題型主要分爲兩大類，第一大類爲聽力，其中包含三選一的單選題型式，主要內容爲測試學生聽辯單字、語音、句子及生活日常用語的意義。第二大類爲閱讀，亦爲三選一的單選題型，主要爲測驗學生辨識單字、句子、簡易英文標示，以及瞭解短文、圖表的能力(臺灣學生學習成就評量資料庫網站，2012)。標準設定所使用的題目共 103 題。

正式施測時，無法讓受試者在短暫時間內施測全部試題，且對於小六學生而言，若是接受所有試題之測驗，不僅耗時，且容易造成心理上的負擔，而產生漏答或亂答的情形增加。因此 TASA 測驗採平衡不完全區塊的題本設計(Balanced Incomplete Block design, BIB)，並透過 IRT 的等化技術(Equating)，將每個題本的分數連結起來。對於 BIB 設計與其受試者能力參數計算有興趣的讀者，可參閱 Yates(1936)；郭伯臣與王暄博(2008)；郭伯臣、楊思偉、白曉珊與張鈺卿(2008)之專著。每一份英文題本內含聽力題 20 題與閱讀題 12 題。

TASA 小六英語科的試題難度，爲中間偏容易，依據曾建銘和陳清溪(2009)的研究發現，學生在聽力與閱讀兩方面的表現水準相似，答對率各爲 76% 與 78%，在每一項能力指標答對率介於 74%~93%。

三、標準設定會議流程

確認 32 名標準設定成員後，於會議進行前的一個禮拜，研究小組先寄送會議的前導資料，讓成員能事先瞭解本研究的進程與目的，其內容包括標準表現描述、英文科評量架構、會議簡介、與會議流程說明等。

會議開始時，研究者先進行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(Performance Level Description, PLD)內的細項內容，逐一檢視，並加以討論，並請 TASA 英文科召集委員協助釐清成員們的疑問。接續，成員們逐一檢視題本，並經過討論、練習後，嘗試進行第一輪的 Yes/No Angoff 的標準設定。

依據 PLD 的描述，請標準設定成員融入自己的專業經驗，分別對基礎、精熟、進階水平的邊緣學生程度進行想像，並逐題判斷該水平的邊緣學生是否能夠答對該題。而 Yes/No Angoff 標準

設定法的實際操作概念及流程，大致如下。首先，研究者提供試題卷，而每頁的試題內容包含有題目內容、選項、及評測項目等。而後要求成員對題本中每一個試題，做以下判斷：

（一）程度列爲基礎最低能力的學生，是否可以答對這一題，如果可以，則在「基礎最低能力者可以答對此題」的欄位內打勾；

（二）程度列爲精熟最低能力的學生，是否可以答對這一題，如果可以，則在「精熟最低能力者可以答對此題」的欄位內打勾；

（三）程度列爲進階最低能力的學生，是否可以答對這一題，如果可以，則在「進階最低能力者可以答對此題」的欄位內打勾；

（四）若進階最低能力的學生，無法答對此題，則在「進階較高能力者才可以答對此題」的欄位內打勾。

例如，如果標準設定成員認爲此題只有程度爲精熟的最低能力學生才可以答對，則於下表 2 之「精熟最低能力者可以答對此題」的欄位內打勾。其中，前述基礎、精熟及進階最低能力學生即爲界於兩層級間臨界點之邊緣學生（borderline student），成員對於這群學生的能力無法簡單歸類。例如精熟的最低能力者，他們的能力則是介於精熟與基礎層級的能力之間。

表 2　Yes/No Angoff 記錄表範例

題號	1. 基礎最低能力者可以答對此題	2. 精熟最低能力者可以答對此題	3. 進階最低能力者可以答對此題	4. 進階較高能力者才可以答對此題
1		V		

第一輪的標準設定總共花費時間約爲一個半小時，結束後由研究者回收成員記錄表，並交由助理進行統計分析。

依據成員記錄表，研究者繪製每一位成員給定分數的散佈圖、並提供各試題傳統難度 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，原先設計一小時的討論時間，因爲題目眾多且成員反映討論時間不足，延長爲 1 個半小時的討論。

討論後，成員們修正原先的判定並進行第二輪的標準設定。執行情形如第一輪，只是花費的時間較少，成員約花一個小時完成第二輪的設定。

助理完成第二輪的分析之後，研究者除提供和第一輪相同的回饋訊息之外，亦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、精熟和進階四個等級，並再次進行成員之間的討論。

討論之後則執行第三輪的標準設定。第三輪的標準設定結束完畢後。研究者公布決斷分數的設定結果並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。本研究主要使用第三輪的實徵數據。

四、多層面 Rasch 模式

由於本研究共有 32 位標準設定成員，每位成員須依據每一個題目判別是否基礎、精熟、進階的邊緣學生可以答對此題，總共有 103 題，所以使用的數據爲 32×103×3 的矩陣，此數據使用軟體 FACETS（Linacre，2007），進行分析。FACETS 程式主要應用在多層 Rasch 模式中並進行參數估計，將觀測的數據轉換成對數型尺度（logit scale）。

本研究考量的層面爲試題難度、成員者的嚴厲度與評分者性別與職業。則第 n 位評分者，其性別爲 m，職業爲 i，在表現層級 ▌，試題難度爲 b 時，針對這個題目，被此成員者評定爲 Y/N 分數之對數勝算比可表示爲：

其中 P_{nmijlY} 爲第 n 位成員者，性別爲 m，職業爲 i，表現層級爲 l，且試題難度爲 b 時評 Y 的可能性。

R_n 爲成員 n 的嚴厲度；

G_m 爲成員性別 m 評分時的嚴厲度；

S_i 爲當成員背景來自職業 i 評分時的嚴厲度；

E_l 爲表現層級 l 的難度；

D_b 爲題目 b 的難度；

I_h 指評定 Y 或是 N 之間的難度界線，也稱爲難度階（threshold difficulty）。

由此模型可見，標準設定成員本身、性別、職業、表現層級、題目難度都是要考量的層面，而各個層面之間的關係緊緊相扣，評分者在試題的判斷爲 Y 或 N 和本身的嚴苛度，試題的難度、本身的性別、職業都有關聯性。此模組將原本單純的評分方式，進而分離各種可能影響評分的因素，使標準設定的判斷模式能夠更爲精準。

在每一個層面，FACET 都會計算參數平均值，參數標準差、不同的適配度參數，例如 infit 均方值、outfit 均方值等。在模組中所估算的參數值，轉換成對數型的等距尺度（logit scale），而其學理上的範圍爲正無限大到負無限大之間。此外，多層面 Rasch 模型將所有的變數關係呈現在變數分布圖上，變數間的大小關係，可以互相進行意義性的比較。一般而言，越高的參數，代表成員進行判斷時越嚴厲或題目越難（Linacre, 2007）。

在多層面 Rasch 模型中，若評分者的給分不穩定，則數據結構會偏離模型的假設，代表 Rasch 分析不適合分析此數據。爲了要檢驗是否可以使用多層面 Rasch 分析來分析資料，可由兩個面向來觀察，第一，檢視是否資料本身之適配度合乎要求，只有適配度合乎要求，才能對後續的參數估計有意義的解讀。一般來說，常以 Infit 均方值與 outfit 均方值常用來檢視適配度，de Ayala（2009）建議這兩種均方值的理想範圍爲 0.5 到 1.5 之間，越接近 1 代表適配度越好。若均方值低於 0.5 或高於 1.5，則代表適配度有問題，需要進一步審核試題與受試者能力之間的關聯性。

第二，須檢視單向度的假設是否成立，Tennant 與 Pallant（2006）認爲 Outfit 或 Infit 的均方值仍有不足，應使用 Rasch 殘差主成份分析來佐證單向度的證據，只要解釋變異量大於 60%，第一殘差特徵值小於 3.0 或第一因素殘值變量佔殘差總量 5%內，符合任何一項條件時，則可提供 Rasch 模式符合單向度的假設的證據（Linacre, 2006）。

進行多層面 Rasch 分析時，亦需考量 Rasch 參數估計的分離度係數（The separation coefficients）。分離度係數爲一種依據假設所算出統計指標，而此假設爲所有的觀測值是從一個常態分布的母群所隨機抽取出來的，而此母群中的統計特徵和觀測值完全相同，在這種特性之常態分布母群中，能夠辨識出幾種具有統計顯著差異性的群組。分離度越高，則代表越能將層面的類別區隔出來，例如若評分者的分離性信度爲 10，則代表若是有一個與研究者所探究的評分者樣本分布相似的常態母群，其評分者之間的差異性至少可以被分成 10 個層級，在標準設定中，研究者會希望評分者的分離度越低越好，因爲接近 0 反倒是代表評分者的評分具有相當的一致性、共識很強。此外，透過卡方檢定，可以進一步檢視觀測值的差異性有沒有達到顯著。例如，研究者想要知道標準設定成員之間的評分是否一致，則可查看卡方檢定的結果，若卡方檢定不顯著，代表評分者之間的判斷是相當一致的，而顯著的卡方值則隱含評分者之間的判斷存在差異。

結果

多層面 Rasch 模式衍伸自試題反應理論，因此數據須符合單向度之假設才能進行接續的分析，首先將使用 Winsteps 軟體來進行主成分分析，來檢視數據是否呈現單向度。依殘差主成份分析報表顯示，解釋變異量爲 96%，第一及第二殘差特徵值（eigenvalue）爲 10.1 與 8.8，第一及第二因素分別解釋 10.5%及 9.2%的殘值變異量，由於解釋變異量大於 60%，表示資料符合 Rasch 模式單向度的測量。

圖 1 爲變數分布圖（variable map），由這個分布圖中可以看出本研究中所考慮各個層面中變數分布的狀況。

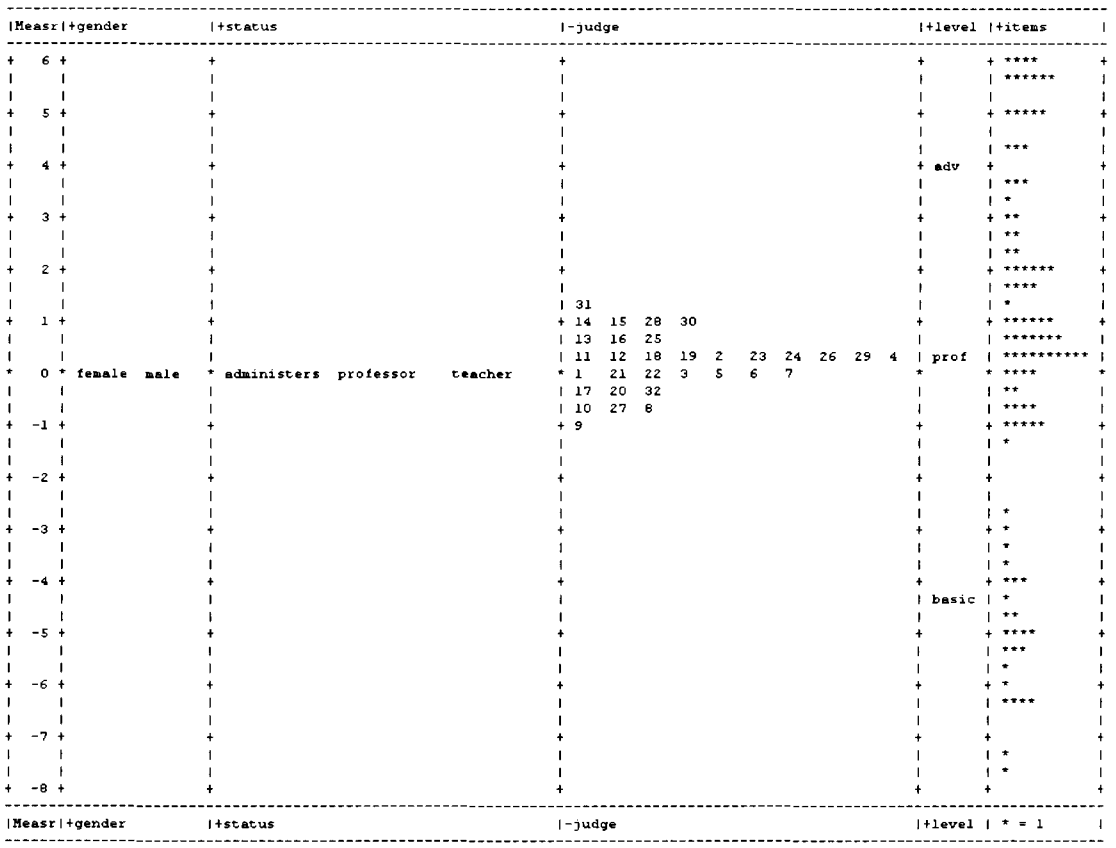


圖 1 變數分布圖

圖 1 最左邊的欄位是刻度，爲對數單位，對數值越高，代表評分者越嚴厲、或是題目越難，對數的單位已轉成爲等距尺度，爲一具有連續性、單位可進行大小比較的數值、且數值之間具有相等的距離。第二個欄位是評分者的性別差異，由這個欄位的變數分布來看，男性和女性的評分者進行試題判斷時，並沒有差異性存在。而第三個欄位則爲評分者的所在地區性差異，由此欄位的變數分布，可看出來自東部地區的成員，進行判別較爲嚴格，而來自南部地區的成員，判別較爲寬鬆。第四個欄位爲評分者的職業類別，不管成員的職業爲教授、學校教師或是行政人員，評分都很類似。第五個欄位則呈現所有 32 位標準設定成員的嚴厲度分布，由此可看出第 31 位成員是進行試題判別時，標準最爲嚴苛，而第 9 位成員最爲寬鬆。第六個欄位爲表現層級的分布，基礎、精熟、進階三個層級之間的難易差異相當顯著，代表成員進行判別時，認爲這三個層級之間是有很大的差別的。最後一個欄位爲試題，由此欄位可看出試題難度分布的相當廣泛，有些題目很簡單、有些題目則很難。

表 3 爲各層面分析參數整理，總共探討六個層面，就適配度 infit 與 outfit 來看，每個層面的適配度都在 0.5~1.5 的範圍內，代表使用的 Rasch 模式來進行估計應該是適合的。就分離度來評估變數本身的異質性，可看出不同的性別和職業之分離度的卡方檢定未達顯著，代表這兩種背景因素，並不會造成成員在判斷上的差異性。

表 3 各層面估計概況整理

	成員	性別	職業	表現層級	試題
Rasch 參數					
平均	0	0	0	0	0.27
標準差	0.23	0.07	0.09	0.07	0.50
N	32	2	3	3	103
Infit					
平均	0.98	1	0.92	0.99	0.98
標準差	0.2	0.1	1.2	0.3	0.1
Outfit					
平均	1.25	1.33	0.79	1.25	0.86
標準差	0.7	0.7	0.2	0.4	1
分離度	2.22	0	0.26	55.07	7.3
卡方檢定	0.00*	0.89	0.89	0.00*	0.00*

*代表 $p < 0.05$

分離度爲一種依據假設所算出統計指標，而此假設爲所有的觀測值是從一個常態分布的母群所隨機抽取出來的，而此母群中的統計特徵和觀測值完全相同。分離度指在這種特性之常態分布母群中，可以分辨出幾種具有統計顯著差異性的類群（Strata），Linacre（2012）指出，當分離度大於 2，即代表層面中的類別有顯著性的不同，此外，分離度（Separation）越高，則代表越能將層面的類別區隔出來，如表 3 所示，成員之間的分離度爲 2.22，代表若是有一個與本研究的組成成員相似的常態分布母群，其評判的嚴厲度差異至少可以被分成 2 個層級，這也隱含受過訓練與討論的標準設定成員，即使盡量訓練其判斷的客觀性，在標準設定時，還是很難脫離原本自身特質的主觀性。在進行決斷分數的評判時，有些成員較嚴格、有些成員較寬鬆。而表現層級的分離度爲 55.07，代表這個標準設定中所設定的基礎、精熟、進階三個等級有明顯的不同。而題目的分離度爲 7.3，則意涵題目難易範圍分布很廣，試題之間的難易度有顯著性的不同。

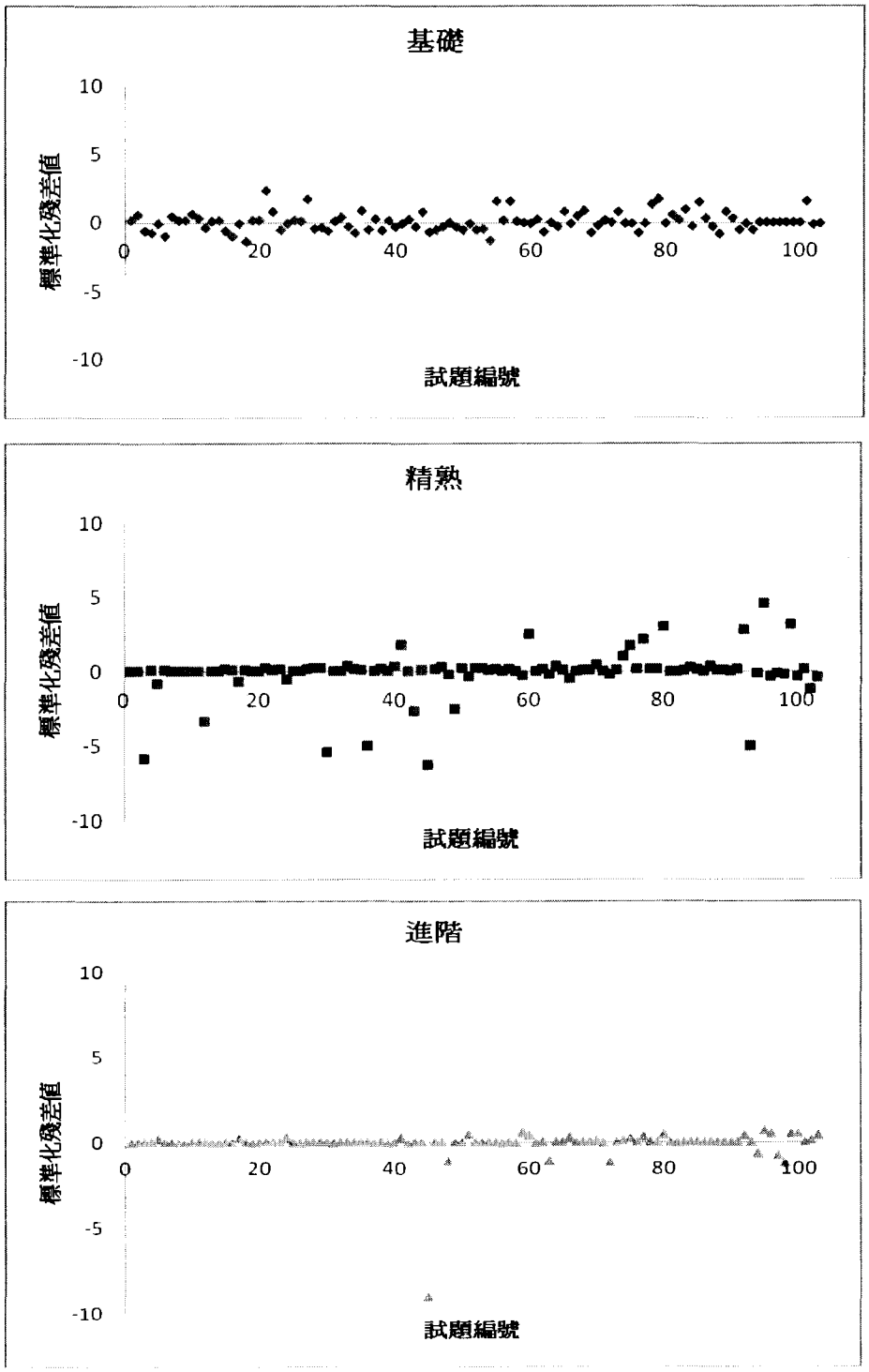


圖 2 成員 17 之標準化殘差值分布圖（infit = 3.55）

每一位標準設定成員，都能製作出標準化殘差分布圖，這個分布圖可做爲評估標準設定成員在執行試題判斷時的內部一致性參考。如圖 2，爲第 17 位標準設定成員之殘差分布圖，由這位成員的 infit 指標高達 3.55，代表這位成員的判斷與 Rasch 模組的估計有很大的出入。在圖 2 中，共有三個子圖，分別呈現基礎、精熟和進階三個階段的標準化殘差分布。一般的判斷原則是，標準化殘差須介於正負 2 個單位，若是大於 2，則代表成員的判斷高於模組預期，例如，題目很難，依據 Rasch 模型預估，學生應該不容易答對，然而，成員在進行題目判斷時，卻覺得學生可以答對。相對的，若是標準化殘差值低於-2，則是代表題目簡單，依據 Rasch 模型預估學生應該可以答對，然而成員在進行判斷時，卻覺得學生無法答對。模組預期值是經由其他成員的判斷綜合推估而成，因此，若是標準化殘差值越大，也代表這位成員和其他成員的判斷越不同。

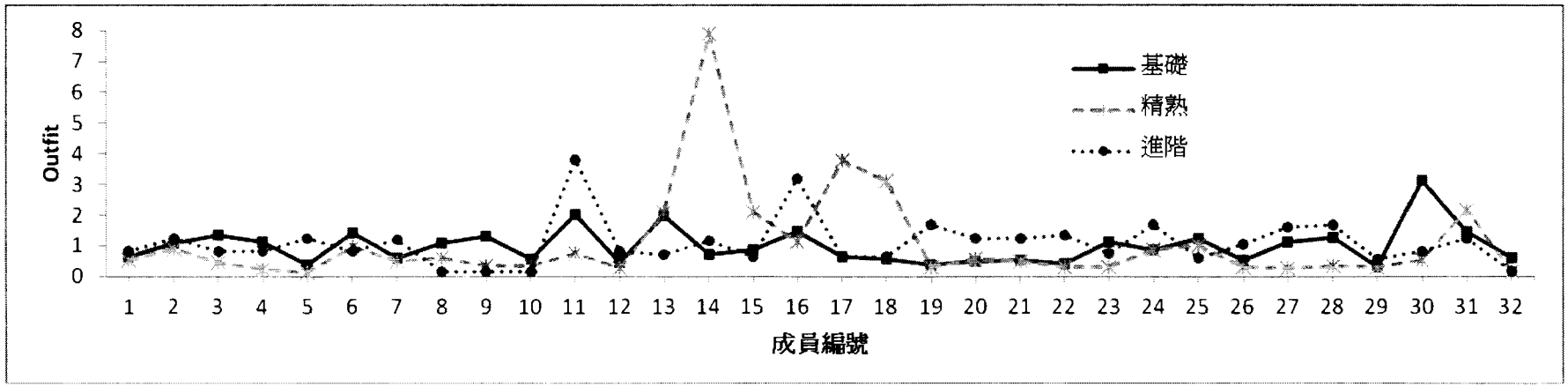
第一個子圖顯示在基礎層級中，這位成員的判斷多符合模組預期，因爲大多數的殘差值都落在±2 中，第二個子圖則爲精熟階級的判斷，成員 17 在精熟階級中，有好幾個試題的標準化殘差值超出±2，有些甚至低於-5 以上，代表這樣的判斷，非常不符合模組預期，例如，在精熟階段的第 3 題，這題應該是一個非常簡單的題目，然而，成員 17 卻認爲這一題很難，難道精熟程度的邊緣學生也無法答對這一題。這樣大的反差，造成標準化殘差值高達-5.68。第三個子圖爲進階階層的殘差分布圖，在這個圖中，顯示除了第 45 題之外，大多數的判決符合預期，因爲殘差非常接近 0。與先前一樣，這一題是非常簡單的題目，大多數的其他成員都覺得這一題是進階程度的邊緣學生絕對可以答對的題目，然而，這位成員卻覺得這一題很難，認爲進階邊緣學生無法答對。這樣的反差造成標準化殘差值爲-9。這樣的訊息，能提供研究者與標準設定成員當作回饋訊息的參考。若成員進行討論前，能先就個人的標準化殘差圖進行檢視並進行判斷修正，可節省不少討論時間。

整體而言，此成員進行基礎與精熟階段的判定時，較多有不一致的現象產生，即成員的判定與模組的預估不同。而這位成員，進行進階階段的評估，其判斷則與模組的預期相對而言較爲一致，所以殘差分布較接近於 0。

圖 3 則呈現標準設定成員適配度分布圖。若是 infit 與 outfit 的值越大，代表標準設定成員的判斷，和 Rasch 理論模組估計下的預期表現，差距越大，一般的期望是能維持在 0.5 到 1.5 之間。而其中 outfit 的統計值容易受到極端值的影響，因此研究者較爲著重 infit 穩定性。由此可看出第 30 位成員在基礎層級、與第 14、17 號成員在精熟階段的評判不符合模組預期，而在進階階段成員的評判均相當符合預期。

而圖 4 呈現的爲試題的適配度分布圖，由此圖可看出進階層級的第 80、97 題，與基礎階段的第 41 題較不符合模組預期，也代表這些題目是成員在做決斷時衝突性比較大的題目，可就這些題目進行優先討論。

研究者可藉由圖 3 和 4 檢視標準設定成員是否達成共識，或是更進一步找出是哪些成員還沒有達成共識，或是哪些是阻礙共識達成的試題，若會議時間不足，可優先進行那些題目的討論。



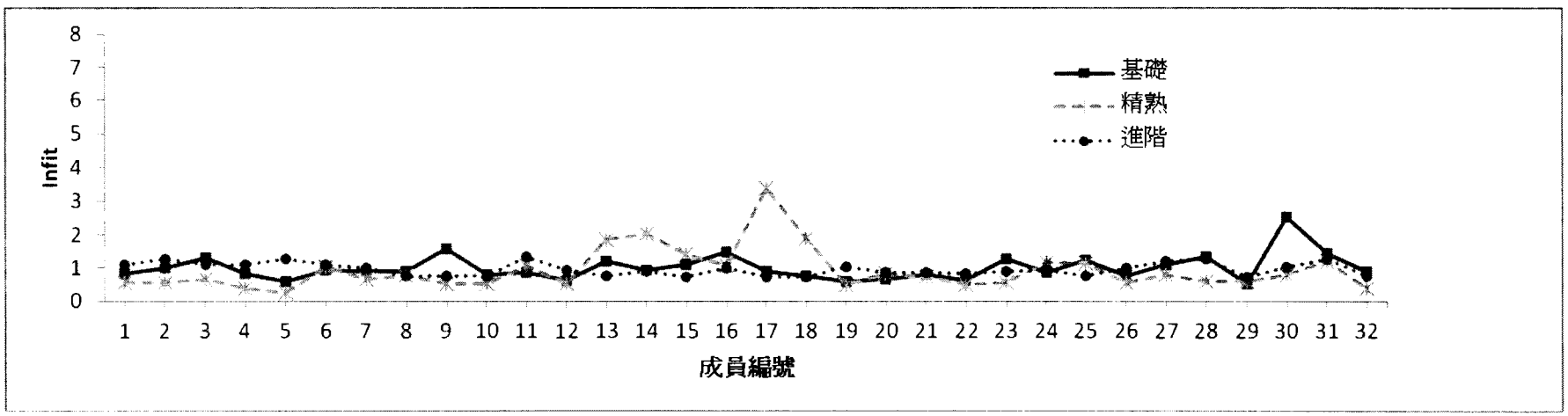
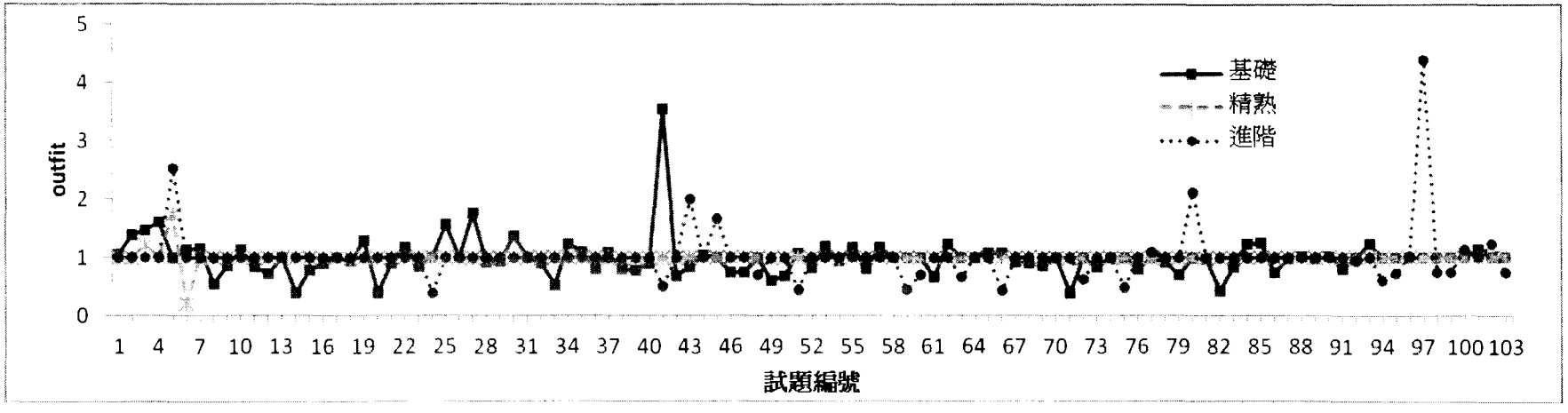


圖 3　標準設定成員 Rasch 適配度分布度



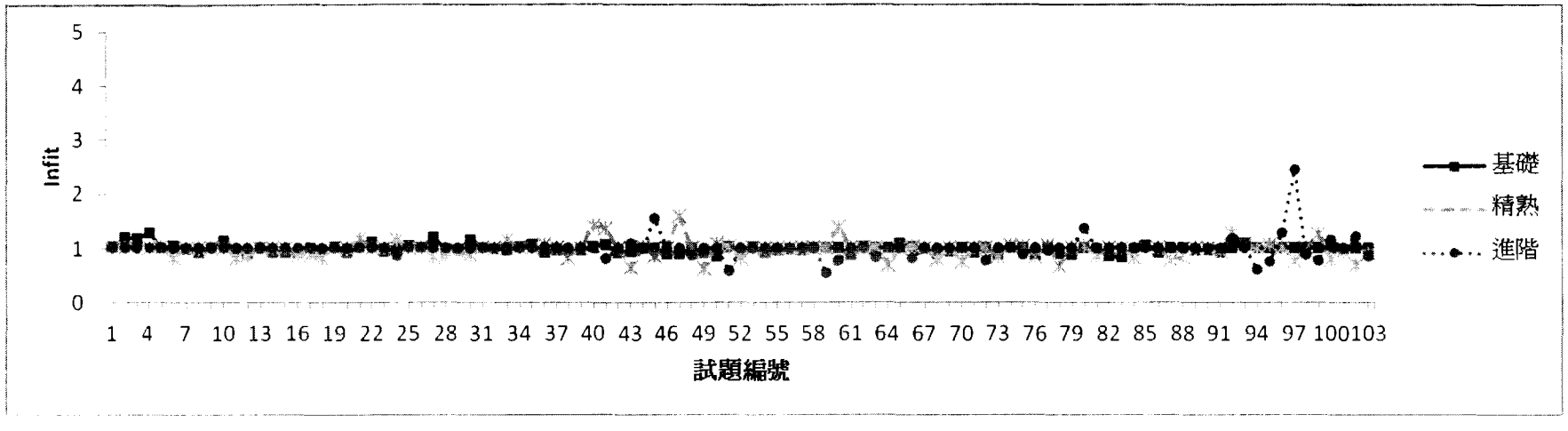


圖 4　試題 Rasch 適配度分布度

結論與建議

在標準設定的執行下，研究者最著重會議進行的效率與標準設定成員共識的達成與否，標準設定的過程是成員的判斷、心理計量的統計與實務的結合。在整個過程中，很難避免成員的主觀判斷，然而，成員判斷的品質，卻直接影響了決斷分數設定的適切性與否。研究者在進行標準設定時，除了提供成員充分的訊息與相關的訓練，也應盡量提供相關、且有用的回饋訊息，讓成員更了解自己所做的判斷。

本研究使用了多層面 Rasch 分析，展現在標準設定上應用的可行性。其中有許多統計指標與圖表，除了能夠提供研究者與成員做爲回饋訊息的參考之外，也能提供內部的效度證據，即標準設定成員評估結果是否具穩定性及一致性。表 4 則整理了在標準設定中，可用來評估與檢視成員與試題層面的相關訊息與可以檢視的問題。對於標準設定成員來說，變數分布圖可用來檢視成員嚴厲度的相對分布圖，除了可以檢視每一位成員的相對嚴厲度分布之外，亦可用此分布圖來檢視不同組群，如職業類別、性別與來自不同區域的成員的意見是否一致。若是在變數圖中，不同類別在變數圖中呈現在同一個水平，則可以推論他們之間的差異性是不顯著的，例如性別與職業類別，由於在本研究中的變數分布圖中都分布在同一個水平，因此研究者可以清楚的看出成員的性別或職業，在標準設定的判斷上並不會有歧異性。然而，來自不同地區的成員，則會有不同的意見。本研究的變數圖中，可看出東部地區的成員，判斷的標準最爲嚴厲，而南部地區的則最爲寬鬆。這樣的論點也可以從分離度和卡方檢定中看出來。分離度越大，則代表成員彼此判斷的歧異性越大，而卡方檢定則進一步提供這樣的歧異性是否達統計顯著水準。

在多層面 Rasch 分析中的標準化殘差分布圖和 Infit、outfit 均方值可提供研究者與成員值得參考的回饋訊息。對於標準設定成員而言，可透過個人的標準化殘差分布圖，找出自己判斷時，是哪一個試題、或哪一個層級，和別人的判斷差異最大，或是哪一項判斷，有可能是不小心出錯而造成的，成員可以透過這樣的回饋圖形，在討論前就先自我檢視與修正。此外，Infit、outfit 的均方值也可協助研究者找出哪些成員的判斷不符合預期，或哪些試題是有很大的歧異性，需要成員優先討論。

表 4 多層面 Rasch 分析在標準設定上的對於成員和試題檢視

統計指標或圖表	成員	試題
變數分布圖	成員間彼此的嚴厲度分布爲何?	試題難度的相對分布爲何?
分離度	成員間的判斷歧異性有多大?	試題難度分布的歧異性有多大?
卡方檢定值	成員間的判斷歧異性是否達到顯著水準?	試題之間的難度差異是否達顯著水準?
Outfit/Infit 均方值（分布圖）	成員對於試題的判斷是否與模組預期一致?	試題難度是否與模組預期一致?
標準化殘差分布圖	成員本身哪些的判斷不符合 Rasch 模組的預期? 比預期的判斷高還是低?	哪些試題的難度不符合 Rasch 模組的預期，比模組預期高還是低?

透過多層面 Rasch 分析所提供的訊息，可以進一步檢視那些適配度較差的成員，並針對這些成員對於標準設定的執行方式、PLD 的理解等，探究其評判不一致的原因。而適配度較差的試題，可針對題目設計、題幹選項、與 PLD 的連結性等方面進行深入討論，因此，本文所提供的統計指標，可作爲標準設定流程中輔助的回饋資訊，並可藉由這些資訊找出問題的癥結點，讓標準設定的執行達到更好的效果。

最後，研究者對未來的研究方向及應用提出幾點建議：

（一）本研究呈現多層面 Rasch 分析在分析標準設定的優勢，然而，本文僅使用現成的數據，並進行事後分析，未來研究者值得深入探討並利用本文所採用的回饋訊息，在標準設定會議的實務運作上是否可行，或是對於標準設定成員進行相關訪談、問卷調查，檢視這種新式的回饋訊息，是否比傳統的回饋訊息，更能提供標準設定成員有效的訊息。

（二）本研究著重多層面 Rasch 分析在 Yes/No Angoff 方法的應用，然而，這樣的分析是否能應用在其他常用的標準設定方法，則需要更深入的探究。例如，書籤標定法也是研究者廣爲使用的一種標準設定方法，然而此方法卻不是逐題檢視，而是先把所有的題目由簡單到難依序排列起來，要求每一位標準設定成員檢視完試題之後，在這些排序的題目中放置各水平的書籤。這樣的概念和 Yes/No Angoff 截然不同，然而，同樣涉及到成員的主觀嚴厲度判斷，如何使用多層面 Rasch 分析，來檢視在這種情境下成員個體內與彼此間的判斷一致性，並提供成員與研究者相關的回饋資訊，也是值得未來深入研究的議題。

參考文獻

吳裕益（1986）：**標準參照測驗通過分數設定方法之研究**。國立政治大學教育研究博士論文。[Wu, Y. Y. (1986). The study of standard setting methods in the criterion-referenced tests (Doctoral dissertation). National Chengchi University.]

吳宜芳、鄒慧英、林娟如（2010）：標準設定效度驗證之探究：以大型數學學習成就評量爲例。**測驗學刊，57**（1），1-27。[Wu, Y. F., Tzou, H. Y., & Lin, C. J. (2010). Validating the performance standards for cut scores in a large-scale mathematics assessment. *Psychological Testing, 57*(1), 1-27.]

吳毓瑩、陳彥名、張郁雯、陳淑惠、何東憲、林俊吉（2009）：以常態混組模型討論書籤標準設定法對英語聽讀基本能力標準設定有效性之輻合證據。**教育心理學報，41**（1），69-90。[Wu, Y. Y., Chen, Y. M., Chang, Y. W., Chen, S. H., He, T. H., & Lin, J. J.(2009). Normal mixture model as convergent validity evidence to bookmark standard setting of english reading and listening ability. *Bulletin of Educational Psychology, 41*(1), 69-90.]

余民寧（2009）：**試題反應理論及其應用**。台北：心理。[Yu, M. N. (2009). *Item response theory*. Taipei: psychological publishing.]

林惠芬（1993）：通過分數設定方法在護理人員檢覈筆試測驗之研究。**測驗年刊，40**，253-262。[Lin, H. F. (1993). Standard setting approaches in the nursing personals paper-pencil tests. *Psychological Testing, 40*, 253-262.]

陳彥名（2006）：**臺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究**。國立台北教育大學教育心理與諮商學系碩士論文。[Chen, Y. M. (2006). *Investigating the standard setting validity of english assessment in TASA (Master's thesis)*. national taipei university of education.]

郭伯臣、王暄博（2008）：大型測驗中同時進行垂直與水平等化效果之探討。**教育研究與發展期刊，4**(4)，87-120。[Kuo, B. C., & Wang, H. P. (2008). A simultaneous vertical and horizontal equating of large-scale assessments. *Journal of Educational Research and Development, 4*(4), 87-120.]

郭伯臣、楊思偉、白曉珊、張鈺卿（2008）：BIB 與 NEAT 設計在不同年度測驗連結效果之比較。**測驗統計年刊第 16 輯下期**，125-154。台中：國立台中教育大學。[Kuo, B. C., Yang, S. W., Pai, H.-S., & Chang, Y. C. (2008). Compared with the linking performance of examination by BIB and NEAT in different years. *Journal of Educational Measurement and Statistics, 16*, 125-154., Taichung, Taiwan：National Taichung University of Education]

曾建銘、陳清溪（2009）：2007 年臺灣學生學習成就評量結果之分析。**教育研究與發展期刊，5**(4)，1-38。[Cheng, C. M., & Chen, C. H. (2009). They analysis of taiwan assessment of student achievement 2007. *Journal of Educational Research and Development, 5*(4), 1-38.]

臺灣學生學習成就評量資料庫網站（2012）：**臺灣學生學習成就評量資料庫建置計畫**。取自 TASA 網站：http://tasa.naer.edu.tw/1about-1.asp?id=2.，2012 年 5 月 22 日。[Taiwan Assessment of Student Achievement (2012). About TASA. Retrieved May 22, 2012, from http://tasa.naer.edu.tw/1about-1.asp?id=2]

鄭明長、余民寧（1994）：各種通過分數設定方法之比較。**測驗年刊，41**，19-40。[Zheng, M. C., & Yu, M. N. (1994). The comparison of different standard setting methods. *Psychological Testing, 41*, 19-40.]

謝進昌（2005）：**以最大測驗訊息量決定通過分數之研究**。國立政治大學教育學 系教育與心理輔導組碩士論文。[Shieh, J. C. (2005). *Study of the Standard Setting by the Maximum Test Information* (Master's thesis). National Chengchi University.]

謝進昌、謝名娟、林世華、林陳涌、陳清溪、謝佩蓉（2011）：大型資料庫國小四年級自然科學習成就評量標準設定結果之效度評估 。**教育科學研究期刊，56**（1），1-32 [Hsieh, J. C., Hsieh, M. C., Lin, S. H., Lin, C. Y., Chen, C. H., & Hsieh, P. J. (2011). Validation of the standard setting procedure for a large scale 4th grade science assessment. *Journal of Research in Education Science, 56*(1), 1-32.]

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.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*(3), 215-235.

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

Council of Chief State School Officers (2001). *State student assessment programs annual survey.* Washington, DC: Author.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Engelhard, G. J. (2009). Evaluating the judgments of standard setting panelists using Rasch measurement theory. In Smith, Jr. E.V., & Stone, G. E. (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 312-346). Maple Grove, MN: JAM Press.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366.

Kozaki, Y. (2010). An alternative decision making procedure for performance assessments: Using the multifaceted Rasch model to generate cut estimates. *Language Assessment Quarterly, 7*, 75-95. doi: 10.1080/15434300903464400

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, J. M. (2006). *Winsteps: Rasch model statistical software*. Chicago, IL: MESA.

Linacre, J. M. (2007). *Facets Rasch measurement computer program* [Computer software]. Chicago, IL: Winsteps.

Linacre, J. M. (2012). *A User's Guide* to *FACETS*. Retrieved July, 1, 2012, from http://www.winsteps.com

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment Research and Evaluation, 13*(9). Retrieved July, 1, 2009 from: http://pareonline.net/getvn.asp?v=13&n=9

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard-setting. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, G. E., Beltyukova, S., & Fox, C. M. (2008). Objective standard setting for judge-mediated examinations. *International Journal of Testing, 8*, 180-196. doi: 10.1080/15305050802007083

Tennant, A., Pallant, J. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048-1051. .

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *Journal of Agricultural Science, 26*, 424-455.

收 稿 日 期：2012 年 05 月 29 日
一稿修訂日期：2012 年 09 月 11 日
接受刊登日期：2012 年 09 月 12 日

Bulletin of Educational Psychology, 2013, 44(4),793-811
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Evaluating the Variability in Standard Setting Using Many Faceted Rasch Model

Ming-Chuan Hsieh
National Academy for Educational Research
Research Center for Testing and Assessment

When conducting the standard setting, the variability of judgments between standard setting panelists is always an issue needed to be addressed. The researcher has to examine whether the variability between panelists is under the accepted range. In addition, standard setting is a time-consuming process. It usually takes several rounds to discuss the judgments. How to provide the useful feedback for panelists to review their judgments is crucial. In this study, the many facet Rasch model was applied on the Yes/No Angoff standard setting procedure. The result shows that the many facet Rasch model has advantages on examining the variability between panelists. It also provides useful feedback to review the internal conflict decisions within each panelist.

KEY WORDS: many facet Rasch, standard setting, Yes/No Angoff