試題呈現與回饋模式對Angoff標準設定結果一致性提升效益之比較研究

吳宜芳
美國愛荷華大學教育測驗統計研究所博士生

鄒慧英
國立臺南大學測驗統計所教授

摘　要

在標準設定的眾多方法中，Angoff法及其相關變形、延伸與修正程序等，實為教育實景中相當普及的標準設定流程。然而，執行Angoff標準設定方法的設定者在概念化最低能力受試者、估計其答題概率時，面臨相當大的認知挑戰。試題特徵（如：試題難度）對設定者間或設定者內一致性的影響，可能影響最後產出標準的效度。基於此，本研究試圖以實徵P值排序回饋、Reckase表回饋與試題呈現分類與否等做法融入修正Angoff法的標準設定程序，以促進設定結果的一致性，並從中比較前述作法融入設定程序之優劣。

本研究係為測驗結束後所進行之標準設定研究，屬於事後做決定型，研究中探究不同回饋模式及試題是否分類呈現對標準設定結果之影響，藉以比較二種作法的優劣，此為本研究之獨特性所在。其次，透過這二種修正作法，期能使設定者對於試題難度有較佳的察覺，進而改善設定間或設定者內一致性，提高設定結果的一致性，並對標準之效度有所助益，是為本研究在功能性之貢獻。

關鍵詞：Angoff法、Reckase表、試題預先分類、標準設定

Evaluating the Utility of Different Item Presentation and Feedback Approaches with the Modified Angoff Method

Yi-Fang Wu
Ph.D Student, Educational Measurement and Statistics, University of Iowa

Hueying Tzou
Professor, Graduate Institute of Measurement and Statistics, National University of Tainan

Abstract

Numerous standard setting methods have been developed to assist panels in estimating the performance of the borderline examinees. Among them, the Angoff method is one of the most popular judgmental standard setting procedures. Its extensions, modifications, and variations are often applied in practice. In standard setting, panelists hold an important role, especially in the judgmental methods such as the Angoff method and its variations. The ability of panelists to accurately estimate the borderline examinees' performance is to some extent subjected to item difficulty. Once the accuracy is questioned, the validity of the performance standard would be damaged. Therefore, a variety of procedures and several types of feedback have been developed to reduce inconsistency among panelists or within a single panelist.

To compare different procedures embedded in the modified Angoff standard setting method for establishing cutoff scores on a large-scale achievement assessment, we designed two standard setting activities, integrating different procedures to help panelists make more accurate estimates.

Two sets of data from a national achievement assessment in mathematics in Taiwan were used in the standard setting activities. Each set contained 104 operational multiple-choice items used to measure students' grade-level math ability. Twelve panelists participated in the 4th grade standard setting activity and the 6th grade panel consisted of 14 panelists. They were all math educators and some had prior experiences in the modified Angoff standard setting procedures.

The standard setting procedures included two factors, each of which involved two conditions: test items with/without item-grouping in advance; different types of feedback, such as feedback with empirical p-values and feedback with IRT calibration/Reckase charts (Reckase, 1998, 2001).

We presented a generalizability analysis design to examine the improvement of consistency for different above mentioned procedures. Item effect, item difficulty effect

(both within difficulty level and between levels) and panelist effect were of interest.

First, the percentage of variance components of item effect increased consistently from Round 1 to Round 3, while the percentage of variance components of panelist effect decreased as the setting round passes. Panelists' consistency was raised; in addition, relatively more variability of panelists was eliminated in the procedure of feedback with Reckase charts. Secondly, with/without item-grouping, panelists could make similar estimates of item performance toward items with similar difficulty as the setting rounds passes. Finally, item-grouping integrated into feedback with Reckase charts having the best improvement of intra-judge consistency, since we observed that under this condition, the estimates of the root mean square error were the smallest and the estimates of generalizability coefficients and intraclass correlation coefficients (ICCs) were the highest.

Panelists are capable of distinguishing hard and easy items; however, with the help of item-group by difficulty and feedback with Reckase charts, the variability induced by item difficulty which has an impact on panelists' consistency, has been decreased as much as possible. This finding, undoubtedly, is beneficial in terms of defending the validity of standard.

Keywords: Angoff method, item-grouping, Reckase charts, standard setting.

壹、前言

現實生活場景中，唾手可得仰賴測驗結果或標準做出的相關決定，如大學入學考試的錄取校系、教師證照考試的通過與否、公務人員高等考試的錄取與否等。隨著標準參照測驗運動的興起，標準設定的施行及相關研究亦隨之開展（吳裕益，1988；Cizek, 2001）。所謂標準設定，係指為既定測驗建立標準的過程。文獻指出，已有許多專家、學者投入標準設定方法的相關研究，如Berk（1986）表示已發展、討論過的標準設定方法多達三十八種，國內吳裕益（1988）與鄭明長、余民寧（1994）亦指出，學者提議用來設定或調整通過分數的方法高達四十種；近年來，新的標準設定方法仍不斷在數量上有所突破（Raymond & Reid, 2001）。

建立標準或決斷分數的過程，一般稱之為「標準設定」（Cizek, 2006）。例如，藥劑師證照考試依表現結果將受試者區分為通過或失敗（pass/fail）；美國國家教育進步評量（National Assessment of Educational Progress, NAEP）依學生表現分為三個成就水準：基礎（basic）、精熟（proficient）、進階（advanced）。標準設定過程中，設定者須仰賴其個人專業知能、經驗及對各個表現標準的解讀與認知，將表現標準轉譯為分數量尺上的決斷分數（cutoff score）（Cizek, Bunch, & Koons, 2004）。

在百花齊放、各家爭鳴的方法中，Angoff法（1971）不啻為最廣泛、普遍的標準設定方法（Berk, 1986; Cizek & Bunch, 2007; Impara & Plake, 1997; Kane, 1994; MacCann & Stanley, 2006; Maurer, Alexander, Callahan, Bailey, & Dambrot, 1991; Reckase, 2006; Sireci & Bisken, 1992; Smith & Smith, 1988）。以此法為基礎衍生出的延伸方法、修正方法或變形（the extended and the modified Angoff methods,Angoff variations），更常見於許多標準設定的實務工作與相關研究。根據Berk（1986）觀察，Angoff法似乎是在技術適足性和實用性之間能取得平衡的最佳方法。依據Angoff的描述，設定小組必須對邊緣群組（borderline groups）的表現水準做出評估

（Angoff, 1971），然Angoff並未對此方法作出更微觀、更細節性的步驟說明，故後續在該法的使用上僅維持基本概念，實際作法多因人而異、因地制宜。簡言之，Angoff法為一判斷性的標準設定方法，具有相當程度的重要性與普遍性，設定者明顯處於此法的核心位階，其對受試者答題表現的評估與判斷，扮演著舉足輕重的角色，由於此法結合了人為判斷與心理計量方式，故相關議題之探究成為研究者關注的焦點所在。

本研究以國小四、六年級的數學成就評量進行標準設定活動，透過不同回饋模式及試題呈現分類與否，運用類推性理論探討對設定準確度的影響。具體而言，本研究之自變項有二，分別是不同的回饋模式與試題是否預先依難度分類呈現，依變項為標準設定結果的一致性，整個研究旨在比較二種回饋方式，及二種試題呈現方式對標準設定結果一致性的差異效果。

貳、文獻探討

本研究之標準設定採用Angoff法佐以修正流程，側重試題難度對於設定者判斷準確性之影響，探討不同回饋與試題呈現方式對提升設定結果一致性的增益效果；以下就相關重要概念進行探討或說明。

一、Angoff標準設定方法

Angoff法係William H. Angoff於1971年提出的一種標準設定方法。Angoff對其方法僅概略提出相關想法，諸如該法僅涉及一個階段或一回合的判斷過程，針對某一群最低能力受試（minimally competent candidates,MCC），設定者必須判斷其可能正確回答某試題的機率，之後將各試題的判斷機率值加總，便會產生一個最低可接受分數（minimally

acceptable score）。常見的Angoff法實施方式是要求設定者估計幾近精熟學生（barely proficient student，概念同於MCC）能正確回答某個二元計分問題的概率（Buckendahl, Smith, Impara, & Plake, 2002）；也或者，設定者必須估計MCC群體能正確回答測驗試題的概率，最後，設定者所產出的估計值將做為計算最低通過分數（minimum passing score）的基礎，用來區辨受試者可否通過測驗（Jaeger, 1995; Plake & Melican, 1989）。

目前對修正的Angoff法並未有一致的定義，Reckase（2000）曾提出修正的Angoff流程，大抵分為下述五個步驟進行之：

1. 選擇設定標準的人員；
2. 訓練設定者。
3. 要求設定者定義並描述受試者應當達到的表現水準。
4. 令設定者估計（最低能力）受試者在每個試題上的表現。
5. 引導設定者檢視實徵資料，如：試題的難度水準（P值）。

二、試題難度對設定者判斷準確度之影響

依「教育與心理測驗標準」所言，決斷分數不僅涵蓋技術面與實務面的考量，同時也包含了價值判斷（AERA, APA, & NCME, 1999）；Hambleton和Pitoniak（2006）便指出，設定表現標準時無可避免的會應用判斷，儘管判斷的本質隨方法而異，但總扮演著關鍵性角色。顯然，設定者的判斷對最終標準具有決定性的影響。過往研究陸續發現若干因素使得設定者的判斷失準，其中一個主要因素便是試題難度（Matter, 2000; McLaughlin, 1993; Schraw & Roedel, 1994; Shepard, 1995; Taube, 1997）。van der Linden在1986年談論設定者內的不一致性時，實意指單一設定者對試題難度的估計產生互相矛盾的情形（van der Linden, 1986）。如Angoff法的設定者對簡單試題分派較低的預期答對率，卻對困難試題分派較高的

答對率，表示出現上述設定者內不一致性的情形（Plake, Melican, & Mills, 1991）。事實上在判斷性的標準設定方法中，「設定者內一致性」是標準未能精準的來源（Plake et al., 1991）。Shepard、Glaser、Linn與Bohrnstedt（1993）特別強調，Angoff法的設定者必須估計一個隨機選取的MCC在某個測驗試題上的答對率，此乃一件極富認知挑戰的任務。以美國NAEP為例，該評量依據設定者產出的三個決斷分數，將學生分派至不同的精熟類目，便是一件相當複雜的認知工作，遠超出人類評定者所能及（Shepard, 1995; 引自Plake & Impara, 2001）。總括而言，當設定者對試題難度的估計值與試題實徵難度不一致時，其結果必會對最終標準的精確性有所影響，亦即標準設定過程出現設定者內不一致現象時，必會損及最終標準的精確性。

三、試題預先分類與Reckase表回饋模式

設定者判斷的失準顯然使設定標準的一致性與精確性受到質疑，標準設定過程中，試題難度無疑是影響設定者判斷準確度的重要因素。因此，若能經由合理、有效的作法來降低試題難度所帶來的負面影響，當能改善最終標準在效度所遭受之折損。

由於判斷MCC的試題答對率對設定者而言是一件極重的認知負荷，故降低設定者的認知負擔，或提供有助於評定一致性的支持、證據，或能有助於標準設定任務的完成。根據吳裕益（1986）所提出之評定量表法，在採用Angoff法的前提下，當試題數在30題以下時，可使用五點量表將所有試題依據難度分為五個類別，31至50題時可用七點量表將試題分為七個類別，51題以上則採用九點量表將試題區分為九個難度類別。據此，本研究為「提高相對難度的一致性」，將試題依難度預先分類，再根據修正的Angoff法步驟，要求設定者逐題估計各水準之最低能力受試者正確回答概率。藉此瞭解試題呈現有無預先依難度分類能否降低試題難度對設定者的

影響，或者提升設定者內的一致性。

由於Angoff法在設定者逐題評定的過程中，涉及大量的人為判斷，牽連複雜的認知、或認知負荷的問題，故為了促使設定者能產生一致的判斷，本研究除了以修正的Angoff法進行三回合的設定流程外，並將引入回饋實徵資料等作法於標準設定過程。van der Linden（1982）和Kane（1987）曾分別討論IRT模式和Angoff標準設定程序的相似性；Reckase（1998）更引入IRT方法設計NAEP的標準設定過程；MacCann與Stanley（2006）則實際採用Rasch模式做為實徵難度與回饋的資訊來源，以促進標準設定的結果。其中，Reckase（1998）所提出的Reckase表（Reckase charts），能提供設定者設定結果與IRT試題校準之一致性，以促進設定結果之內部一致性。此法原是Reckase為NAEP制定成就水準所撰寫之標準設定方法的部分內容，之後ACT的標準設定工作小組便以Reckase之名為此表命名（Reckase, 2001）。嚴格說來，此表並非完整的標準設定方法，然其設計初衷在輔助如Angoff法等以測驗為中心（test-centered）的標準設定方法，用以協助設定者進行邊緣群體答對概率的評定。

Reckase（1998）認為，NAEP的標準設定過程應視為轉譯National Assessment Governing Board（NAGB, 2006）政策性定義的過程，設定者參與標準設定過程的目的，是要提供試題答對率的不偏估計值。因此，Reckase利用IRT的三參數模式，以表格（如表1所示）呈現出各試題在不同精熟程度（通常以受試者θ值表示之，測驗總分或量尺分數亦可）之答對率，做為設定者調整其設定結果之依據。透過Reckase表，設定者可得知不同量尺分數（不同能力受試者）在各試題對應的預期答對率（依據IRT三參數模式而來）。設定者可藉此檢視自己在逐題評定最低能力受試群的預期答對率時，是否能對此一群體所應具備之能力秉持一致的看法，以做出穩定的判斷。本研究爰引Reckase表之設計初衷來呈現試題難度訊息、與評定一致性的回饋資訊，並將其設定結果與單純實徵P值進行回饋的設定結果加

以比較，進一步瞭解何種回饋模式對降低標準的變異性有較大的幫助。

參、研究方法與設計

本研究的標準設定程序採取修正的Angoff法，主要參考NAEP（Allen, Jenkins, Kulick, & Zelenak, 1997）、Reckase（1998, 2000）與Hambleton

表1 Reckase表舉隅——四題選擇題

量尺分數	試題一	試題二	試題三	試題四
…	…	…	…	…
188	.89	.96	.70	.92
185	.85	.95	.63	.91
182	.81	.93	.55	.89
179	.75	.91	.48	.86
176	.68	.89	.42	.83
173	.60	.86	.37	.80
170	<.53>	.83	.34	.77
167	.45	.79	.31	.73
164	.37	.74	.29	.69
161	.30	.69	.28	.66
158	.24	.63	.27	.62
155	.20	<.57>	.26	.58
152	.16	.52	.26	.55
149	.13	.46	.26	.52
146	.11	.41	.25	.49
143	.09	.36	.25	.46
140	.08	.32	.25	<.44>
137	.07	.29	.25	.43
134	.07	.26	.25	.41
131	.06	.24	.25	.40
128	.06	.22	.25	.39
125	.06	.20	<.25>	.38

註：1.<>表示評定者對MCC對該題的預期答對率。 2.資料來源：Reckase（2001:166）。

（2001）所建議之方式與準則來設計流程。研究架構如圖1所示。基本上，四、六年級標準設定小組在修正的Angoff法三回合設定流程大致相同，然為檢驗不同回饋模式（自變項一）對標準設定結果一致性（依變項）之增益情形，四年級設定小組在第三回合給予實徵P值排序結果做為回饋，六年級設定小組則以IRT三參數模式產生的Reckase表給予回饋，供二組設定者檢視其內部一致性。二個設定小組分別命名為「實徵P值排序回饋組」與「IRT模式回饋組」。另一方面，為降低試題難度對設定者判斷之影響、探討試題呈現方式（自變項二）之效益，四、六年級設定小組中各有半數

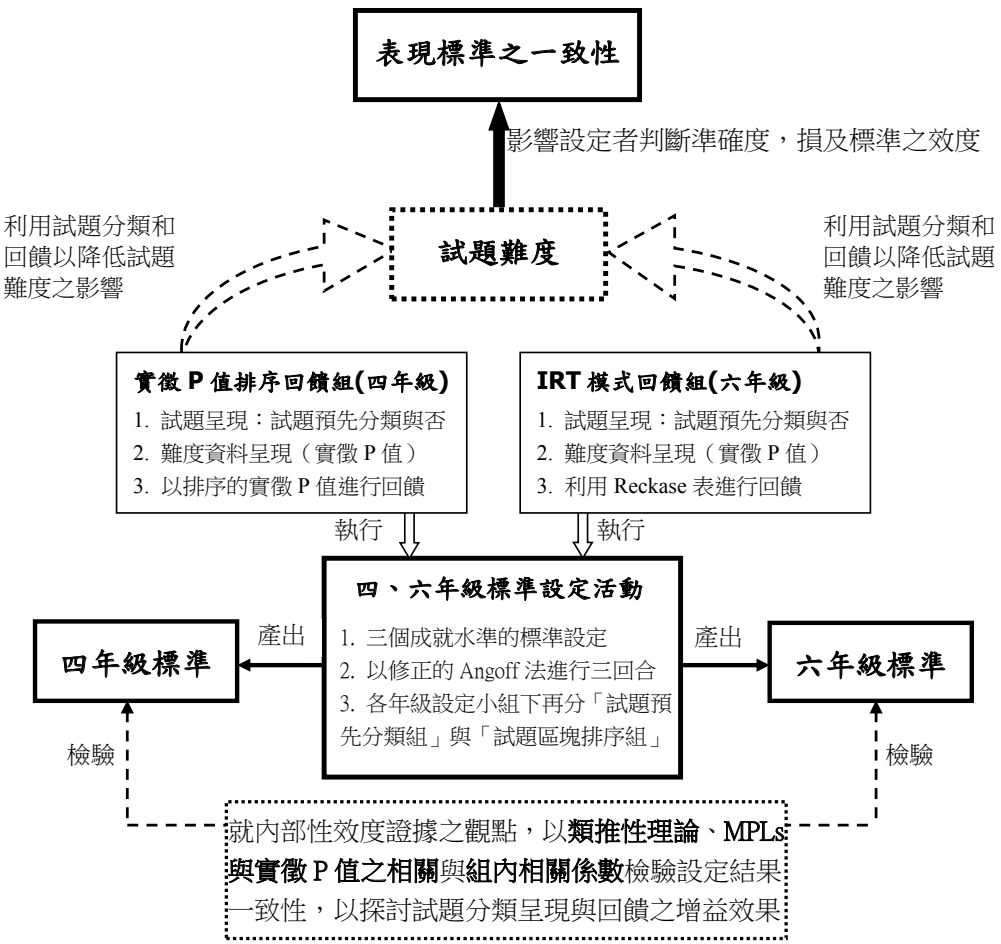

圖1 研究架構圖

設定者接受「試題依難度預先分類」之實驗處理，稱為「試題預先分類組」；另一半接受「試題依試題區塊排序」（試題排序與難度無關）之處理，稱之「試題區塊排序組」。最後，藉類推性理論設計、最低通過水準（minimum passing levels, MPLs）（Goodwin, 1999）與實徵P值之相關，以及組內相關係數（intraclass correlation coefficients, ICCs）來探究不同回合設定結果之演進變化情形，藉此驗證不同回饋模式與試題預先依難度分類對於標準設定結果一致性（依變項）之增益情形。

標準設定研究中包含三個互動的元素：一組設定者透過一系列標準設定的過程與一組試題進行交互作用（Plake et al., 1991）。以下分述本研究之設定材料、設定者與標準設定過程等三項元素。

一、標準設定材料（materials）

本研究所用之數學成就評量，四、六年級各有101與99個選擇題。傳統試題分析顯示，四年級測驗試題之難度範圍在0.12至0.91之間，平均難度為0.52、標準差為0.18；六年級測驗試題之難度範圍則在0.22至0.87之間，平均難度為0.59、標準差為0.17，此處所謂的試題難度係指受試者答對該試題的百分比。

二、設定小組（panel）與設定者（panelists）

本研究擔任標準設定的設定者主要為國小數學教師及教育局數學輔導團輔導員。根據Brandon（2004）的建議，執行修正的Angoff法至少需要10名設定者（理想上為15至20位），故本研究以16名設定者進行實際標準設定活動，除其中一名由國小教師轉任大學教職外，餘15名為國小教師（其中3名為數學輔導團輔導員），這15名國小教師均持續擔任四或六年級數學教學，多數曾參與各服務學校課程發展委員會數學領域小組。設定者平均教學年資為8.88年，任教數學領域的平均教學年資則為8.06年。最後，在考量各設定者可參與的時程下，四年級標準設定小組共計12位，六年級標準設定小組則有14位，同時參與二個年級標準設定者共10人。

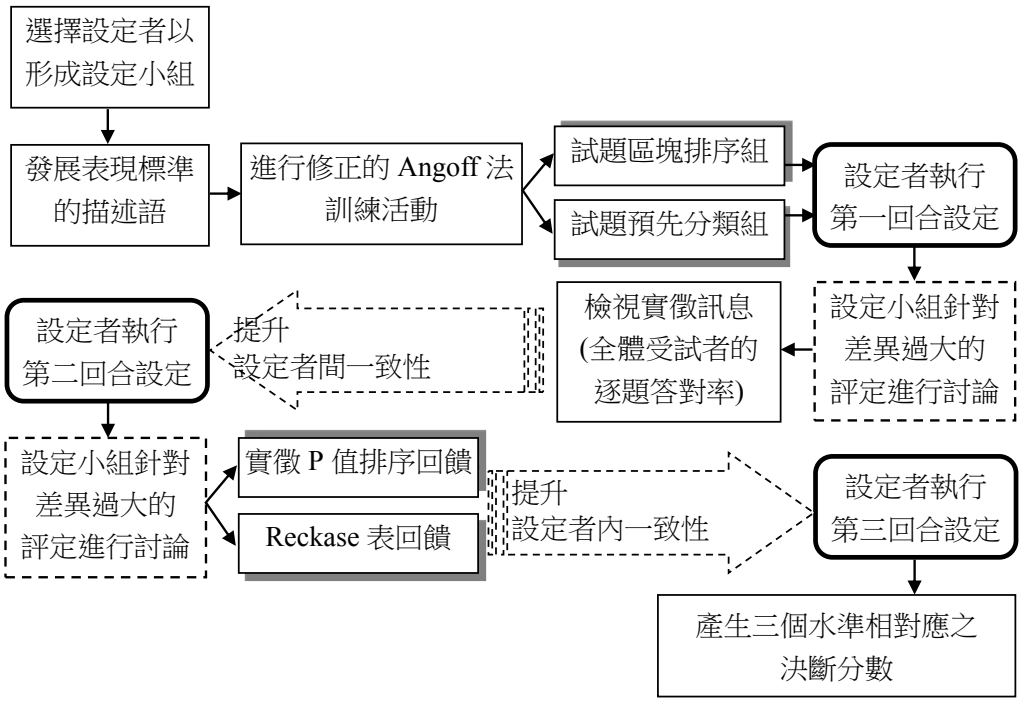

圖2 標準設定流程圖

三、標準設定過程

如前所述，本研究用以設定標準之方法為修正的Angoff法，圖2大致呈現本研究所採行之標準設定過程，各步驟簡要說明如後。

1.形成設定小組後，向設定者說明評量目的與進行標準設定之目的，帶領所有設定者概覽事先擬定之表現水準定義、檢視整體數學評量的架構並討論影響試題難度的因素。請個別設定小組獨立完成一份四／六年級的數學題本，目的在使設定者熟悉評量內容。之後，研究者帶領設定小組共同擴展並具體化表現水準定義，根據特定、明確的數學技能、知識與行為來判定四／六年級學生在三個成就水準的預期表現。

2.進行修正Angoff法之設定訓練。此時，研究者依設定者的教學背景將之分成二個次設定小組，分別是「試題預先分類組」與「試題區塊排序

組」。其差異處在於前者在進行標準設定時，試題呈現方式為依九個難度層級排序的試題卡（九種色卡），後者則為依試題區塊排序的試題卡（白色色卡）。而經由前述具體化表現水準定義的過程，設定小組在取得每個水準邊緣受試（亦即該水準當中最低能力受試）的特徵共識後，每位設定者依據評量簡介、難度來源分析、代表性水準描述語等資料，以樣本試題進行Angoff法之設定訓練。

3.正式設定過程。設定者須檢視並分別設定三個水準的邊緣受試在各個試題的預期答對率；共計三回合。

（1）第一回合設定開始，設定者首先針對水準一邊緣學生逐題估計其答對概率，其後對水準二邊緣學生執行相同步驟，最後則是水準三的邊緣學生。

（2）進行第二回合設定前，提供第一回合結果做為回饋。此步驟主要目的在提供「設定者間一致性」（interjudge consistency）訊息，包含個別設定者的平均設定結果、設定小組的總平均設定結果及兩兩設定者間的相關矩陣，做為調整設定者間一致性的參考依據。根據此回饋資料，設定小組成員必須覺察極端的設定結果，討論造成極端設定的原因，並鼓勵極端設定者（如：個人的平均設定遠高於或遠低於設定小組總平均者）說明設定的看法或理由，同時，探查是否有部份成員誤解試題設定的概念或方式。

（3）除提供前述（2）回饋資料外，於第二回合開始前提供設定小組試題難度訊息（即實徵P值），設定者可以此進行實際確認，用以提升設定者間的一致性。設定者對三個水準的預期答對率若與試題難度的差異過大，應當重新檢視該題，以覺察是否有誤解該題或誤判難度的可能。

（4）第二回合設定開始，設定小組檢視與第一回合相同的所有試題，此一回合中，設定者可利用「設定者間一致性訊息」、「試題難度訊息」、「第一回合提供的資訊」等相關訊息進行設定，設定者可再次確認第一回合的設定結果，也可以調整或重新設定各題的預期答對率。

（5）第三回合設定前，利用第二回合結果再次進行分析與回饋。除

了提供每位設定者的平均設定、設定小組的總設定及相關矩陣外，提供實徵P值排序表予四年級標準設定小組做為設定者內一致性之調整依據。作法上，以個別設定者為單位，呈現依實徵難度P值排序的試題，及各試題對應之設定者的三個水準預期答對率（根據第二回合設定結果而來），由於試題係依據實徵難度排序，因此設定者可以檢視具有相似難度的試題，其在各個水準的設定預期答對率是否出現不同的設定結果。六年級標準設定小組則是提供Reckase表（類似表1）做為設定者內一致性之調整依據；每位小組成員均有一份相同的Reckase表（根據第二回合設定結果而來），欄為試題號碼、列為量尺分數，表中數值則是利用IRT三參數模式計算所得之各題答對概率，並依照試題難度將題號由難而易進行排序。上述四、六年級不同回饋模式之主要目的在提供「設定者內一致性」（intrajudge consistency）的變異訊息。

（6）第三回合設定開始，設定小組檢視與第一、二回合相同的試題，利用新提供的「設定者內變異訊息」以及「第一、二回合提供的回饋」進行最終設定，此時，設定者可利用小組討論的方式，以決定他們單獨無法確定的預期答對率。最後，根據第三回合的設定結果計算決斷分數，產出對應三個成就水準的標準。

肆、結果與討論

本節探討不同回饋模式和試題呈現分類與否對於標準設定結果一致性的增益效果，透過類推性設計下設定者與試題變異隨設定回合消長的變化情形、誤差均方根（root mean squared error, RMSE）、類推性係數（G coefficients）、組內相關係數、試題的最低通過水準和受試實際表現（即試題難度）之相關程度等重要規準來依序呈現。以下分項說明之。

一、以類推性交叉設計檢視不同回饋模式對設定一致性的增益效果

首先在類推性理論設計中，以i表試題、p表示設定者，採$i\times p$交叉設計，探討設定者變異（σ_p^2）為標準設定結果主要誤差變異來源的情況下，觀察其是否能隨三個設定回合的演進而逐漸下降。由表2可知，實徵P值排序回饋組與IRT模式回饋組無論在任一水準下所產出的決斷分數，其試題變異百分比均「隨著各回合設定而漸次提升」，而設定者變異、與試題間交互作用變異則「隨著各回合設定而漸次降低」。三個水準的結果均顯示研究中二個設定小組透過修正的Angoff標準設定流程，能有效提升試題的變異並合理降低設定者的變異，可見整個設定小組能逐漸降低最終標準中所牽涉的個人主觀判斷，設定成員大多能在以試題為中心的考量下產出較為一致的評定。

整體而言，IRT模式回饋組之設定者變異百分比，幾乎都比實徵P值排序回饋組來得小；特別是在第三回合，IRT模式回饋組在三個水準的設定者變異均下降至5%以下（分別為4.2%、2.3%與4.5%），顯見設定間的不一致情形獲得相當程度的改善。然二個設定小組係在第二回合結束後始給予不同回饋，故比較第二與第三回合設定者的變異百分比降幅，實徵P值排序回饋組在三個水準的降幅依序為62.61%（11.5%降為4.3%）、41.21%（18.2%降為10.7%）、45.79%（32.1%降為17.4%），同理，IRT模式回饋

表2 不同回饋模式在各水準各回合之變異來源大小及變化情形

成就水準	變異來源與大小		實徵P值排序回饋組（N_P=12，N_I=101）			IRT模式回饋組（N_P=14，N_I=99）		
			回合一	回合二	回合三	回合一	回合二	回合三
一	σ_i^2	V.C.	0.010	0.019	0.019	0.012	0.017	0.021
		%	18.5	73.1	82.6	46.2	73.9	87.5
	σ_p^2	V.C.	0.030	0.003	0.001	0.005	0.001	0.001
		%	55.6	11.5	4.3	19.2	4.3	4.2
	$\sigma_{i\times p}^2$	V.C.	0.014	0.004	0.003	0.009	0.005	0.002
		%	25.9	15.4	13.0	34.6	21.7	8.3

表2 不同回饋模式在各水準各回合之變異來源大小及變化情形(續)

成就水準	變異來源與大小		實徵P值排序回饋組（N_P=12，N_i=101）			IRT模式回饋組（N_P=14，N_i=99）		
			回合一	回合二	回合三	回合一	回合二	回合三
二	σ_i^2	V.C.	0.008	0.021	0.021	0.011	0.018	0.041
		%	24.2	63.6	75.0	39.3	72.0	93.2
	σ_p^2	V.C.	0.012	0.006	0.003	0.008	0.002	0.001
		%	36.4	18.2	10.7	28.6	8.0	2.3
	$\sigma_{i\times p}^2$	V.C.	0.013	0.006	0.004	0.009	0.005	0.002
		%	39.4	18.2	14.3	32.1	20.0	4.5
三	σ_i^2	V.C.	0.003	0.014	0.015	0.008	0.010	0.019
		%	18.8	50.0	65.2	36.4	58.8	86.4
	σ_p^2	V.C.	0.009	0.009	0.004	0.007	0.003	0.001
		%	56.3	32.1	17.4	31.8	17.6	4.5
	$\sigma_{i\times p}^2$	V.C.	0.004	0.005	0.004	0.007	0.004	0.002
		%	25.0	17.9	17.4	31.8	23.5	9.1

註：1. N_P係指接受不同實驗處理的設定者人數，N_i係指經設定後採以分析的試題數。
2. σ_i^2表示試題變異；σ_p^2則為設定者變異；$\sigma_{i\times p}^2$則表試題與設定者之交互作用。

組的降幅依序為2.33%、71.25%、74.43%，結果顯示，除水準一外，IRT模式Reckase表的回饋效果較實徵P值的回饋效果為佳。換言之，利用Reckase表對於設定者間一致性具有良好的提升作用，故使設定者變異大幅降低。類似組型亦出現在試題與設定者交互作用的變異。綜此觀之，在第三回合前引入IRT模式之Reckase表，確能有效改善設定者間的一致性，進而獲得較穩定的最終設定結果。

二、以誤差均方根與類推性係數檢視不同回饋模式對設定一致性的增益效果

計算誤差均方根（RMSE）之目的在求取跨試題與設定者之平均設定的估計標準誤，藉以得知設定小組產出標準所隱含的誤差。本研究應用$i\times p$交叉設計，計算三個成就水準下，每回合決斷分數所對應之RMSE。不同

回饋模式設定小組其設定結果之RMSE與類推性係數（相對決策）如表3所示。RMSE與類推性係數（$E\hat{\rho}^2$）的計算公式分別為：

$$RMSE=\sqrt{\frac{\hat{\sigma}_p^2}{n_p}+\frac{\hat{\sigma}_{ip}^2}{n_i n_p}} \quad ; \quad E\rho^2=\frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)} \Rightarrow E\hat{\rho}^2=\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2+\hat{\sigma}_{ip}^2} \quad 。$$

符號說明如下：

n_i：試題數，n_p：設定者人數，$\hat{\sigma}^2$：對應效果之變異數成分估計值。

無論在任一水準當中，隨著各個回合的進展，RMSE有漸次降低或持平的趨勢。以水準一為例，實徵P值排序回饋組在第一回合的RMSE為0.050，第二回合則降為0.016，並在最終回合降至0.009；而IRT模式回饋組則依回合演進分別為0.019、0.009與0.009。簡言之，二個設定小組在所有水準產出之決斷分數，其RMSE均有隨回合演進下漸次降低的趨勢。根據Verhoeven等人（1999）的研究結果顯示，在常態分配下，10位設定者評定200題或12位設定者評定100題的條件下，如欲將通過分數（百分比）的精確性控制在1%之內，所對應之RMSE約在0.51，即 $0.51\times1.96=0.9996\approx1$，便可使其通過分數（百分比）的95%信賴區間是在±1%的誤差範圍。本研究四、六年級設定小組中的任一水準第三回合所產出最終標準的RMSE均

表3 不同回饋模式在三個水準各回合決斷分數之RMSE與類推性係數

	成就水準	實徵P值排序回饋組（N_p=12，N_i=101）			IRT模式回饋組（N_p=14，N_i=99）		
		回合一	回合二	回合三	回合一	回合二	回合三
RMSE	一	0.050	0.016	0.009	0.019	0.009	0.009
	二	0.032	0.022	0.016	0.024	0.012	0.009
	三	0.027	0.027	0.018	0.022	0.015	0.009
類推性係數（$E\hat{\rho}^2$）	一	0.737	0.971	0.983	0.919	0.974	0.988
	二	0.787	0.953	0.972	0.898	0.975	0.995
	三	0.705	0.923	0.956	0.886	0.957	0.988

註：1. N_p係指接受不同實驗處理的設定者人數；2. N_i係指經設定後採以分析的試題數。

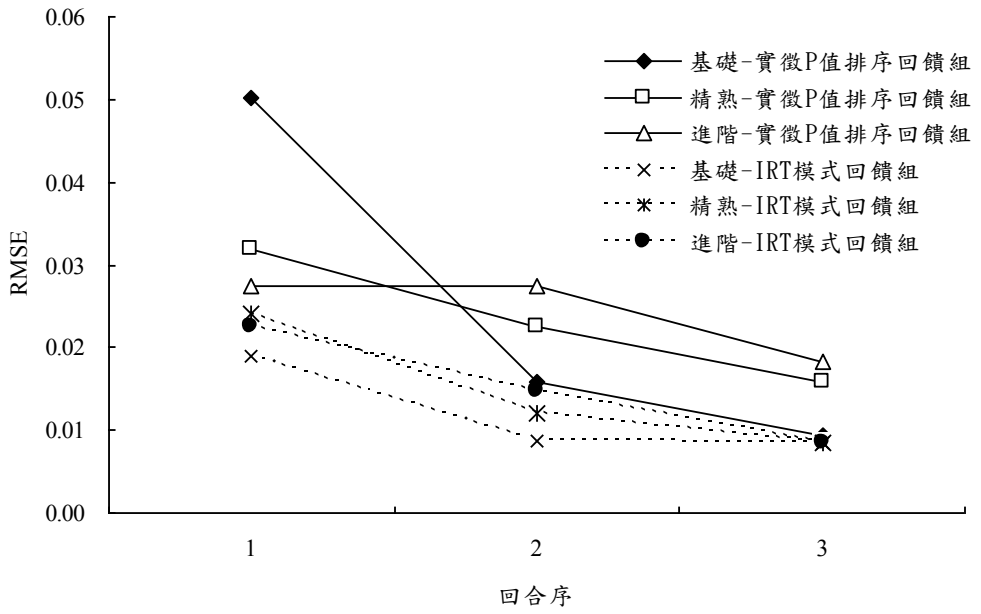

圖3 不同回饋模式在三個水準各回合設定結果之RMSE變化情形

為0.009（見回合三），以95%信賴區間而言， $0.009\times1.96=0.01764\approx0.02$ ，亦即本研究設定結果通過分數的95%信賴區間的誤差範圍僅0.02%，換言之，本研究二種回饋模式下各個水準所產出的最終標準，遠優於Verhoeven等人(1999)的研究結果。顯見，隨著標準設定流程的進行，二種回饋均能有效降低標準隱含的誤差或不穩定性，並在最終標準（第三回合結果）呈現出最佳的穩定狀態。上述變化亦可透過圖3觀察而得。

比較二組設定結果發現，實徵P值排序回饋組於各回合各水準所產出的結果，其RMSE估計值均較IRT模式回饋組來得高；若僅就第二、三回合的結果而言，實徵P值排序回饋組的RMSE降幅依序為43.75%、27.27%、33.33%，IRT模式回饋組的降幅為0%、25%、40%，二種回饋方式對降低RSME似無明顯孰優孰劣；然比較二組在各個水準的最終標準（第三回合結果）發現，IRT模式回饋組所產生的RMSE明顯較低（水準一 $0.009\leqq0.009$ 、水準二 $0.009<0.016$ 與水準三 $0.009<0.018$ ）；可見IRT模式

回饋所產出的決斷分數，其信度表現比實徵P值排序回饋組的結果來得好，換言之，利用IRT模式的回饋對設定者一致性有較佳的提升效果。

根據類推性交叉設計，以$i \times P$決策性研究結果求取各表現標準的類推性係數（$E\hat{\rho}^2$），發現上述RMSE所得到的結果，大致上也反映在各個決斷分數對應之類推性係數估計值。由表3可知，四年級設定小組在水準一三個回合的類推性係數估計值分別為0.737、0.971及0.983，顯示其估計值隨著回合演進的過程漸次提高；同樣趨勢亦得見於水準二、三；六年級設定

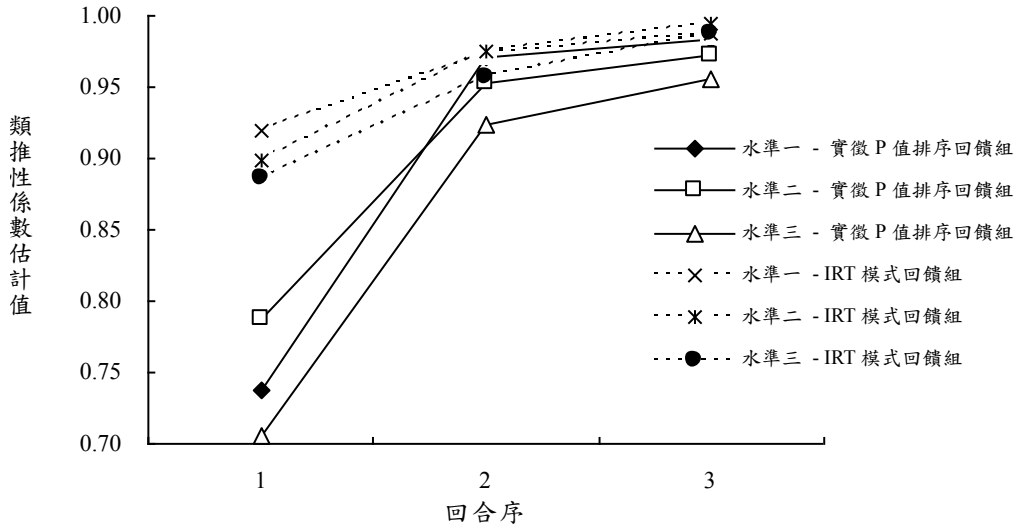

圖4 不同回饋模式在三個水準各回合設定結果之類推性係數變化情形

小組在各水準之類推性係數估計值亦呈現類似態勢。圖4呈現二設定小組在不同水準三回合設定結果的類推性係數變化情形。值得注意的是，透過IRT模式回饋組所產出的最終標準，其類推性係數估計值在三個水準中各為0.988、0.995與0.988，相較之下發現，此一回饋模式可獲致較佳的類推性係數估計值。IRT模式回饋組在類推性係數的表現優於實徵P值排序回饋組，與RMSE之現象相互呼應。

三、以試題的最低通過水準（MPLs）和受試實際表現之相關程度檢視不同回饋模式對設定一致性的增益效果

最低通過水準，意指標準設定者認為在某個測驗欲達到某個表現層次，最低能力受試者所應具備的成就水準。就單題而言，可以跨設定者對該題的所有評定加以平均，獲致該題的最低通過水準。一般而言，吾人預期各個試題的最低通過水準應能與所有受試的實際表現（試題難度P值）具有高相關。Pitoniak（2003）指出，試題評定與試題難度的相符程度可作為最終標準的內部性效度證據。基於此，本研究分別計算不同回饋模式設定小組在三個水準各回合下所有試題的「最低通過水準」（MPLs）與實徵試題難度之相關係數。

由表4可知，所有MPLs與實徵試題難度的相關係數均介於0.660至0.990

表4 不同回饋模式組之MPLs與實徵P值的相關係數變化情形

成就水準	實徵P值排序回饋組（N_p=12，N_i=101）			IRT模式回饋組（N_p=14，N_i=99）		
	回合一	回合二	回合三	回合一	回合二	回合三
水準一	0.660	0.982	0.991	0.819	0.923	0.941
水準二	0.618	0.978	0.993	0.818	0.983	0.990
水準三	0.557	0.976	0.937	0.897	0.984	0.861

註：1.以上係數均達0.05顯著水準。

2. N_p 係指接受不同實驗處理的設定者人數；N_i 係指經設定後採以分析的試題數。

之間，屬中、高度相關，可見設定者大多能判定試題的相對難度，呼應Shepard（1995）認為設定者能夠區辨簡單或困難的試題之看法。跨回合觀察發現，無論實徵P值排序回饋組或IRT模式回饋組，其第二回合結果與試題實徵難度的相關均明顯提高，係因在第一、二回合之間提供逐題的實徵P值讓設定者進行實際檢核，故第二回合結束時，設定者所決定的最低通過水準能在這類「真實訊息」（reality information, Cizek & Bunch, 2007, p.55）的輔助下，與試題難度間呈現高相關。

就實徵P值排序回饋組而言，其相關係數隨回合演進漸次提高。該組除

了在第二回合有試題難度做為參考，第三回合乃利用排序後的實徵P值對應自行判定的預期P值進行調整，因此在大多數情況下，本組設定者幾乎都與實徵難度值形影不離，由此結果可見實徵P值排序回饋組針對各水準邊緣學生進行逐題答對概率判定時，對難度的覺知較佳。

類似於實徵P值排序回饋組的結果，IRT模式回饋組在提供實徵P值後，第二回合的相關提高，第三回合給予Reckase表的輔助後，亦有助於設定者對試題難度的覺知，使相關提高。然Reckase表的輔助功能主要在改善設定者內與設定者間一致性，其內容乃經由三參數IRT模式產生，並非單純包含難度P值的訊息，故最後產出的相關係數不如實徵P值排序回饋組來得高；此外，在水準三第三回合的相關係數下降，或許與設定者長時間專注設定活動導致疲倦有關。

四、以組內相關係數檢視不同回饋模式與試題呈現方式對設定一致性的增益效果

Wuensch（2003）指出，欲瞭解兩個或兩個以上的評分者在測定相同事物的共識程度時，可利用組內相關係數（ICCs）的測量技巧；因此，本研究除了透過類推性交叉設計中變異來源之變化，呈現不同設定小組設定者間一致性的提升情形外，另計算各設定小組在三個水準各回合設定結果之組內相關係數，並分試題預先分類與區塊排序兩組討論之。

由表5可知，不同回饋模式的設定小組在三個成就水準之組內相關均

表5 不同回饋模式設定小組之組內相關係數在三個水準各回合之變化情形

	回饋程序	實徵P值排序回饋組（N_p=12，N_i=101）			IRT模式回饋組（N_p=14，N_i=99）		
分類方式	成就水準	回合一	回合二	回合三	回合一	回合二	回合三
試題區塊排序組（N_p=6 & 7）	一	0.348	0.775	0.821	**0.408**	**0.777**	**0.874**
	二	0.218	0.736	0.851	**0.392**	**0.778**	**0.913**
	三	0.150	0.592	0.707	**0.460**	**0.685**	**0.826**
試題預先分類組（N_p=6 & 7）	一	0.105	0.670	0.830	**0.551**	**0.691**	**0.956**
	二	0.222	0.499	0.618	**0.487**	**0.667**	**0.956**
	三	0.166	0.389	0.575	**0.428**	**0.533**	**0.887**

註：1.以上係數均達0.05顯著水準。

2. N_p係指接受不同實驗處理的設定者人數；N_i係指經設定後採以分析的試題數。

3.粗體表示IRT模式回饋組相關優於實徵P值排序回饋組。

隨著設定回合而漸趨提升。由於IRT模式回饋組的設定者在第三回合接受Reckase表的輔助產出設定結果，因此，若要確知IRT模式的回饋是否有助於提升設定者間一致性，需從第三回合結果驗證。觀察表5可知，IRT模式回饋中之試題預先分類組之組內相關係數為各小組最高者，三個水準的相關係數依序為0.956、0.956、0.887，由此可見，設定者在同時接受IRT模式回饋與試題預先分類處理二種實驗處理的情形下，其設定者間相關能達到較佳的情形。

五、以類推性套入設計檢視試題呈現方式對設定一致性的增益效果

由文獻探討可知，試題難度對設定者判斷所造成之影響，將使設定者在估計邊緣學生的最低通過概率時產生偏誤。理想上，設定者應能區分不同難度層次的試題，並對難度相同或相近的試題做出類似判斷，方能為合理、精準的最終標準背書。換言之，當設定者對試題判斷出現矛盾，例如：對簡單試題判定低的預期答對率、對困難試題判定高的預

期答對率，便會發生設定者內不一致的現象（van der Linden, 1986）。因此，本研究在標準設定過程中，除二種回饋方式外，亦提供二種試題呈現方式—試題預先分類及試題區塊排序—試題是否預先依實徵難度分類。分析時利用類推性理論（$i:d$）× p套入設計（d表試題難度層次，共分為九類），探討三回合最終設定標準之變異來源的變化情形，俾以得知試題預先依難度分類對提升設定者內一致性的助益。在考量設定者（p）與試題難度（d）二個測量面向的情形下，區分難度層次間（σ_d^2）與難度層次內的試題變異（$\sigma_{i:d}^2$），觀察二者在不同回合間之變化情形。理論上，難度層次間的試題變異應當大於難度層次內的試題變異，以支持設定者能合理區辨不同難度的試題，並對同一難度層內的試題做出相近的概率判斷；且試題預先分類組其難度層次間的試題變異（$\sigma_d^{2'}$）應當大於試題區塊排序組之變異（σ_d^2），試題預先分類組其難度層次內的試題變異（$\sigma_{i:d}^{2'}$）應當小於試題區塊排序組之難度層次內的試題變異（$\sigma_{i:d}^2$），如此方可驗證試題難度對設定者所帶來的影響是否可藉由試題預先分類加以調整。以下分就實徵P值排序回饋組與IRT模式回饋組探討試題呈現分類與否對降低試題難度的效果。

（一）實徵P值排序回饋組

實徵P值排序回饋組試題預先依難度分類與否，在三個水準三回合設定結果之各種變異來源的百分比如表6所示。以水準一為例，無論試題是否依難度預先分類，第一回合結果發現難度層次間變異百分比大於難度層次內試題變異百分比（$\sigma_d^2 = 21.5\% > \sigma_{i:d}^2 = 15.0\%$；$\sigma_d^{2'} = 8.3\% > \sigma_{i:d}^{2'} = 3.1\%$），足見設定者在沒有任何試題實徵難度資訊的情形下，確實能夠

表6 實徵P值排序回饋組在各水準各回合（$i:d$）×p套入設計之變異來源及變化情形

成就水準	變異來源與大小		試題區塊排序組（N_p=6，N_i=101）			變異來源與大小		試題預先分類組（N_p=6，N_i=101）		
			回合一	回合二	回合三			回合一	回合二	回合三
一	σ^2_d	V.C.	0.008	0.024	0.025	$\sigma^2_{d'}$	V.C.	0.006	0.019	0.019
		%	21.5	78.2	82.8		%	8.3	69.1	83.7
	$\sigma^2_{i:d}$	V.C.	0.006	0.0005	0.0004	$\sigma^2_{i:d'}$	V.C.	0.002	0.0002	0.0002
		%	15.0	1.5	1.4		%	3.1	0.8	1.0
	σ^2_p	V.C.	0.009	0.002	0.003	$\sigma^2_{p'}$	V.C.	0.052	0.003	0.000
		%	23.1	6.6	8.7		%	71.5	12.4	0.0
	$\sigma^2_{p\times d}$	V.C.	0.002	0.001	0.001	$\sigma^2_{p\times d'}$	V.C.	0.003	0.003	0.002
		%	5.1	3.4	4.0		%	4.2	10.7	10.8
	$\sigma^2_{p:d}$	V.C.	0.014	0.003	0.001	$\sigma^2_{p:d'}$	V.C.	0.009	0.002	0.001
		%	35.2	10.2	3.2		%	12.9	7.0	4.9
二	σ^2_d	V.C.	0.003	0.027	0.027	$\sigma^2_{d'}$	V.C.	0.005	0.019	0.021
		%	11.7	74.8	85.3		%	12.9	54.1	63.9
	$\sigma^2_{i:d}$	V.C.	0.003	0.0005	0.0005	$\sigma^2_{i:d'}$	V.C.	0.004	0.0002	0.0003
		%	11.2	1.3	1.5		%	10.5	0.5	1.0
	σ^2_p	V.C.	0.011	0.004	0.001	$\sigma^2_{p'}$	V.C.	0.014	0.009	0.005
		%	39.6	10.7	4.4		%	34.4	25.5	14.9
	$\sigma^2_{p\times d}$	V.C.	0.001	0.002	0.002	$\sigma^2_{p\times d'}$	V.C.	0.004	0.006	0.006
		%	4.1	4.9	4.9		%	10.8	16.2	17.9
	$\sigma^2_{p:d}$	V.C.	0.009	0.003	0.001	$\sigma^2_{p:d'}$	V.C.	0.013	0.001	0.001
		%	33.3	8.3	4.0		%	31.3	3.6	2.3
三	σ^2_d	V.C.	0.002	0.020	0.021	$\sigma^2_{d'}$	V.C.	0.001	0.011	0.013
		%	7.2	60.9	72.1		%	8.7	41.6	59.4
	$\sigma^2_{i:d}$	V.C.	0.002	0.0005	0.0003	$\sigma^2_{i:d'}$	V.C.	0.001	0.0001	0.0003
		%	8.6	1.4	1.3		%	8.8	0.5	1.3
	σ^2_p	V.C.	0.014	0.006	0.003	$\sigma^2_{p'}$	V.C.	0.006	0.011	0.005
		%	64.5	19.2	10.4		%	46.3	40.0	20.5
	$\sigma^2_{p\times d}$	V.C.	0.001	0.003	0.003	$\sigma^2_{p\times d'}$	V.C.	0.002	0.004	0.004
		%	3.0	9.2	9.6		%	15.2	15.7	16.1
	$\sigma^2_{p:d}$	V.C.	0.004	0.003	0.002	$\sigma^2_{p:d'}$	V.C.	0.003	0.001	0.001
		%	16.8	9.3	6.7		%	21.0	2.2	2.7

註：1. N_p係指接受不同實驗處理的設定者人數；N_i係指經設定後採以分析的試題數。

2. σ^2_d表示難度層次間試題變異；$\sigma^2_{i:d}$表示難度層次內的試題變異；σ^2_p則為設定者變異；$\sigma^2_{i\times p}$則表試題與設定者之交互作用；$\sigma^2_{p:d}$則表試題與設定者交互作用套入難度之變異。

有效區辨不同難度的試題，早期Lorge與 Kruglov（1953）之研究便說明了設定者能夠區分試題難易的事證。儘管差異縮小，上述同樣見於水準二；水準三則呈現難度層次內變異百分比稍大於難度層次間變異百分比的趨勢（σ_d^2＝7.2%＜$\sigma_{i:d}^2$＝8.6%；$\sigma_d^2{}'$＝8.7%＜$\sigma_{i:d}^2{}'$＝8.8%）。此結果顯示，設定者在水準一能辨識出不同難度層次的試題，到了水準三似乎不易區辨出不同難度層次間與相同難度層次內的試題難度。

再者，無論試題預先分類與否，難度層次間變異百分比（σ_d^2）隨著設定回合漸次升高。以水準一為例，區塊排序組在三回合的難度層次間變異百分比依序為21.5%＜78.2%＜82.8%，預先分類組則為8.3%＜69.1%＜83.7%。從一、二回合的明顯變化，不難看出在提供試題實徵難度資訊後，設定者愈能明確區分出不同難度層次的試題。此外，試題區塊排序組的難度層次內試題變異百分比（$\sigma_{i:d}^2$）亦在一、二回合間有明顯降幅。預先分類組的難度層次內試題變異（$\sigma_{i:d}^2{}'$）大致上也呈現下降的情形，尤其是一、二回合間更為明顯。由此顯示，無論試題預先分類與否，基本上設定者隨著回合演進過程中對同一難度層次內的試題越能做出相似近的概率判斷，設定者內一致性（呼應van der Linden之定義）隨之獲得改善。同樣的情形，亦見於水準二、三之設定結果。然而，進一步比較試題區塊排序與預先分類二組的難度層次內試題變異百分比（$\sigma_{i:d}^2$ vs. $\sigma_{i:d}^2{}'$）發現，後者在三個表現水準的每個回合均小於前者。由此可見，試題預先分類能使設定者對難度相近的試題有較好的掌控，有助於提升設定者內一致性。換言之，試題預先分類確實能使設定者在判斷的過程中，降低難度所帶來的不良影響。

（二）IRT模式回饋組

IRT模式回饋組試題預先依難度分類與否在三個水準三回合設定結果之各種變異來源的百分比如表7所示。以水準一為例，無論試題是否依難度預

表7 IRT模式回饋組在各水準各回合（$i:d$）× p套入設計之變異來源及變化情形

成就水準	變異來源與大小		試題區塊排序組（N_p=7，N_i=99）			變異來源與大小		試題預先分類組（N_p=7，N_i=99）		
			回合一	回合二	回合三			回合一	回合二	回合三
一	σ^2_d	V.C.	0.006	0.020	0.014	$\sigma^2_{d'}$	V.C.	0.015	0.017	0.029
		%	22.7	77.0	73.8		%	55.8	71.1	89.3
	$\sigma^2_{i:d}$	V.C.	0.005	0.001	0.003	$\sigma^2_{i:d'}$	V.C.	0.001	0.0001	0.002
		%	19.7	2.7	14.7		%	2.2	0.6	6.8
	σ^2_p	V.C.	0.007	0.001	0.0002	$\sigma^2_{p'}$	V.C.	0.002	0.000	0.0005
		%	25.2	4.4	0.9		%	9.0	0.0	1.4
	$\sigma^2_{p\times d}$	V.C.	0.001	0.001	0.0002	$\sigma^2_{p\times d'}$	V.C.	0.007	0.004	0.0003
		%	2.8	3.4	1.1		%	25.9	16.1	0.8
	$\sigma^2_{p:d}$	V.C.	0.008	0.003	0.002	$\sigma^2_{p:d'}$	V.C.	0.002	0.003	0.001
		%	29.6	12.5	9.6		%	7.1	12.6	1.7
二	σ^2_d	V.C.	0.005	0.022	0.043	$\sigma^2_{d'}$	V.C.	0.016	0.018	0.047
		%	20.0	77.3	89.1		%	49.8	68.5	93.6
	$\sigma^2_{i:d}$	V.C.	0.005	0.001	0.001	$\sigma^2_{i:d'}$	V.C.	0.001	0.0002	0.001
		%	20.4	2.4	3.1		%	1.8	0.9	2.6
	σ^2_p	V.C.	0.007	0.002	0.0003	$\sigma^2_{p'}$	V.C.	0.007	0.001	0.001
		%	27.7	7.3	0.7		%	21.9	3.2	1.6
	$\sigma^2_{p\times d}$	V.C.	0.0002	0.002	0.002	$\sigma^2_{p\times d'}$	V.C.	0.007	0.004	0.0005
		%	0.8	5.3	3.5		%	21.6	14.7	0.9
	$\sigma^2_{p:d}$	V.C.	0.008	0.002	0.002	$\sigma^2_{p:d'}$	V.C.	0.002	0.003	0.001
		%	31.1	7.6	3.6		%	4.8	12.7	1.3
三	σ^2_d	V.C.	0.003	0.011	0.014	$\sigma^2_{d'}$	V.C.	0.014	0.011	0.020
		%	21.5	69.4	66.1		%	43.7	55.3	74.8
	$\sigma^2_{i:d}$	V.C.	0.002	0.0002	0.004	$\sigma^2_{i:d'}$	V.C.	0.001	0.0002	0.004
		%	20.7	1.6	17.7		%	2.0	1.0	14.8
	σ^2_p	V.C.	0.001	0.0007	0.0003	$\sigma^2_{p'}$	V.C.	0.009	0.004	0.002
		%	11.4	4.3	1.6		%	28.7	21.7	5.8
	$\sigma^2_{p\times d}$	V.C.	0.0001	0.002	0.0003	$\sigma^2_{p\times d'}$	V.C.	0.006	0.003	0.001
		%	1.3	14.2	1.5		%	18.2	14.3	2.4
	$\sigma^2_{p:d}$	V.C.	0.005	0.001	0.003	$\sigma^2_{p:d'}$	V.C.	0.002	0.002	0.001
		%	45.2	10.6	13.1		%	7.4	7.6	2.2

註：1. N_p係指接受不同實驗處理的設定者人數；N_i係指經設定後採以分析的試題數。

2. σ^2_d表示難度層次間試題變異；$\sigma^2_{i:d}$表示難度層次內的試題變異；σ^2_p則為設定者變異；$\sigma^2_{i\times p}$則表試題與設定者之交互作用；$\sigma^2_{p:d}$則表試題與設定者交互作用套入難度之變異。

先分類，第一回合結果發現難度層次間變異百分比大於難度層次內試題變異百分比（σ_d^2＝22.7%＞$\sigma_{i:d}^2$＝19.7%；$\sigma_d^{2'}$＝55.8%＞$\sigma_{i:d}^{2'}$＝2.2%），此一現象與實徵P值排序回饋組之發現相呼應，顯示設定者確實能區辨不同難度的試題。此一情形同樣見於二組在水準二與水準三的第一回合結果，惟試題區塊排序組在水準二的第一回合，難度層次間與難度層次內的變異百分比無明顯差異。

其次，試題區塊排序組的難度層次間變異百分比（σ_d^2），在水準一、三的第三回合反而下降（前者依序為22.7%、77.0%、73.8%，後者依序為21.5%、69.4%、66.1%），並未如實徵P值排序回饋組般呈現隨回合上升的趨勢；然而在試題預先分類組，其難度層次間變異百分比（$\sigma_d^{2'}$）隨回合上升的趨勢則相當明顯（水準一55.8%＜71.1%＜89.3%、水準二49.8%＜68.5%＜93.6%、水準三43.7%＜55.3%＜74.8%）。此一結果顯示，設定者能在提供試題實徵難度資訊後，明確地區分不同層次的試題，並做出合乎邏輯的判斷，故一、二回合間呈現上升局面；然而試題區塊排序組在二、三回合對試題相對難度的掌握反而不佳，試題預先分類組在二、三回合反出現最佳效果，相較其他次小組在各水準的第三回合表現可知，IRT回饋模式的試題預先分類組在難度層次間變異百分比（$\sigma_d^{2'}$）是所有次小組最大者。

另外，試題區塊排序組的難度層次內試題變異百分比（$\sigma_{i:d}^2$）不但未隨著設定回合呈現一致下降的趨勢，其中水準一、三在第三回合反明顯上升（前者依序為19.7%、2.7%、14.7%，後者依序為20.7%、1.6%、17.7%）。似乎也反映IRT回饋組在二、三回合間在沒有試題預先依難度分

類的協助下，設定者無法持續降低難度層次內試題間的變異，或許也可能是因為Reckase表含有過多回饋訊息干擾所致。

最後，比較試題區塊排序與預先分類二組的難度層次內試題變異百分比（$\sigma^2_{i:d}$ vs. $\sigma^2_{i:d}$'）發現，前者在三個表現水準的每個回合均大於後者。舉例而言，二組在水準一中三個回合的變異關係依序為19.7%＞2.2%、2.7%＞0.6%、14.7%＞6.8%；水準二與水準三呈現相同趨勢。顯然，試題預先分類能使設定者對難度相近的試題有較好的掌控，能降低設定者在判斷過程中試題難度所帶來的負擔或影響。

伍、結論與建議

如前所述，為降低試題難度對設定者造成之影響，本研究於標準設定過程引入不同回饋模式與試題呈現方式，期望能提升設定者一致性，有助於最終標準之精確度。

一、結論

（一）IRT回饋模式對設定結果一致性具有提升效益

就時間點而言，利用Reckase表進行的IRT回饋模式發生在第三回合設定時，故僅自第三回合結果觀察其效益。就類推性交叉設計來看各成就水準的結果，此時IRT模式回饋組之試題變異百分比最大，而設定者變異則降至最小；就RMSE來說，IRT模式回饋組所產生的各個RMSE相對於實徵P值排序回饋組而言明顯較低，其所產生的類推性係數估計值則明顯較高；IRT模式回饋組的組內相關係數值亦大於實徵P值排序回饋組。由此可見，利用Reckase表的IRT回饋模式對於設定者間與設定者內一致性具有良好的提升作用，能將設定者的不一致性調整至最小，該組所產出的最終標準能使試題本身的效果發揮至極，並具有較完善的效度支持。

（二）試題呈現預先依難度分類對設定結果一致性具有提升效益

就類推性套入設計結果來看，試題呈現預先依難度分類的設定小組，其難度層次內試題變異百分比均小於試題區塊排序的設定小組，可見透過此一試題呈現方式，能使設定者對難度相近的試題有比較好的掌控，有助於提升設定者內一致性，可降低試題難度對設定者在判斷過程產生的不良影響。更重要的是，從套入設計結果和組內相關係數來看，IRT回饋模式在試題預先分類組能發揮最大效果，同時接受IRT回饋與試題預先分類二種實驗處理時，不僅能幫助設定者在概率判定的過程中維持良好的內部一致性，其設定者間相關亦能達到最佳情形。總括而言，二者交互作用之下對設定結果一致性有最佳的提升效益。

（三）多回合設定與回饋訊息對判斷性標準設定程序的必要性

最後，應用判斷性標準設定方法於選擇題時，建議以多回合方式進行，並且，加入例行性的設定者間與設定者內一致性回饋，皆是必要的作法，一如Brandon（2004）所言，試題估計值常因試題難度而有不一致的分散情形，需藉由標準設定當中各個回合的回饋與調整來加以彌補；Clauser、Swanson與Harik（2002）亦強調多回合的訓練能使設定者集中評定（centering ratings）、有益於設定者內一致性；透過本研究亦獲致同等結論。

綜上所述，經由本研究可知試題呈現依難度分類的作法明顯有助於一致性的提升；而另一方面，IRT回饋模式對於設定者內一致性的提升效果，顯然優於實徵P值排序回饋。因此，在必要的情況下，融合試題分類呈現與IRT模式回饋二種作法來進行標準設定，更能對設定者一致性發揮最佳的提升效益。

二、對未來建議

（一）探究不同IRT模式的回饋效果

依據前述結論，爾後的標準設定活動似可考慮納入IRT模式的Reckase表做為回饋機制，以提昇標準設定結果的信度與效度。然本研究所採行之

Reckase表係根據三參數IRT模式而來，其中的答題概率估計值受到鑑別度、難度與猜測率之影響，倘使利用不同IRT模式（如：Rasch單參數模式）所產出Reckase表來進行回饋，其設定結果或許有所不同。因此，可進一步探究何種IRT模式產出的Reckase表較為合適，或比較其設定結果之差異等。

（二）探究其他標準設定程序結果的一致性

本研究在考量評量涉及的試題類型與方法的實用性上，乃以修正的Angoff法進行標準設定活動，並根據其過程與所產出的決斷分數提出一致性的相關證據，建議未來應當採以其他標準設定方法，如書籤法，針對本研究所產出的決斷分數進行檢核，或者比較其設定結果之異同。

（三）操弄試題難度或探究設定者背景變項

為處理試題難度對設定者造成之影響，吾人或可設計一些難度相當的試題直接進行設定，是比較直接操控試題難度的作法；為提升設定結果之一致性，亦可引導設定者利用試題本身來進行成就水準的判斷，看哪些試題分屬哪一個水準，以此作法來取得設定者間一致性、減少評定的不一致；探究設定者背景變項與決斷分數之高低有何關聯……等，上述議題均是後續標準設定研究可進一步著墨、探討之處。

最終，本研究提供清楚例證，標準設定活動的設計經常融入不同回饋方式於各個階段／回合，然而，提供回饋的目的，必須清楚的呈現與說明，如此方能有助於提昇設定結果的一致性與準確性。此外，關於提昇設定結果的一致性是否能提昇標準設定的效度，從測驗信效度的觀點而言，信度只是效度的必要條件，而非充分條件，故提昇設定結果的一致性（信度提昇）不能保證效度的提昇；然依據學者Michael Kane（1994）對通過分數表現標準的效度驗證探討，他從程序、內部與外部三方面的證據與效標來說明表現標準的效度驗證。本文所探究的提昇設定結果的一致性正是Kane所指稱內在標準的效度檢核，據此而言，提昇設定結果的一致性，應能提昇設定標準之效度。由於本研究旨在探究設定結果一致性，故未論及效度方面的議題，該議題之探究可參見作者於2010年發表於測驗學刊

（57:1）「標準設定效度驗證之探究—以大型數學學習成就評量為例」一文。

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所博士論文(未出版)。

吳裕益(1988)。標準參照測驗通過分數設定方法之研究。**測驗年刊，35**，159-166。

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。

Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1997). *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: National Center for Education Statistics.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Measurement, 56*(1), 137-172.

Brandon, P. R. (2004). Conclusions about Frequently Studied Modified Angoff Standard-Setting Topics. *Applied Measurement in Education, 17*(1), 59-88.

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. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.1-17). Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (2006). Standard setting. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp.225-258). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Clauser, B. E., Swanson, D. B., & Harik, P. (2002). Multivariate generalizability analysis of the

impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39*(4), 269-290.

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18*(3), 257-267.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education, 12*, 13-28.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process1, 2. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement*, (4th ed., pp. 433-470). Washington, DC: American Council on Education.

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

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*(1), 15-40.

Kane, M. (1987). On the use of IRT models with judgmental standard setting procedures. *Journal of Educational Measurement, 24*(4), 333-345.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Lorge, I., & Kruglov, L. K. (1953). The improvement of the estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34-46.

MacCann, R. G., & Stanley, G. (2006, January). The use of Rasch modeling to improve standard setting. *Practical Assessment, Research & Evaluation, 11*(2). Retrieved from http://pareonline.net/pdf/v11n2.pdf

McLaughlin, D. H. (1993). Validity of the 1992 NAEP achievement-level setting process. In L. Shepard, R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Setting performance standards for student achievement tests: Background studies* (pp.81-122). Stanford, CA: National Academy of Education.

Matter, J. D. (2000). *Investigation of the validity of the Angoff standard setting procedure for multiple-choice items.* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA.

Maurer, T. J., Alexander, R. A., Callahan, C. M., Bailey, J. J., & Dambrot, F. H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel Psychology, 44*, 235-262.

National Assessment Governing Board (2006). *Writing framework and specifications for the 2007 National Assessment of Educational Progress.* Washington, DC: National Assessment

Governing Board.

Pitoniak, M. J. (2003). *Standard setting methods for complex licensure examinations* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA.

Plake, B. S., & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: an issue in judgmental standard setting. *Educational Assessment, 7*(3), 87-97.

Plake, B. S., & Melican, G. J. (1989). Effects of item context on intrajudge consistency of expert judgments via the Nedelsky standard setting method. *Educational and Psychological Measurement, 49*(1), 45-51.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issue and Practice, 10*(2), 15-25.

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

Reckase, M. D. (1998). *Setting standards to be consistent with an IRT item calibration.* Iowa City, IA: ACT.

Reckase, M. D. (2000). *The ACT/NAGB standard setting process: How "modified" does it have* to *be before it is no longer a modified-Angoff process?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, L. A. (ED442825)

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy and impact. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 159-173). Mahwah, NJ: Lawrance Erlbaum Associates.

Reckase, M. D. (2006). Some criteria for evaluating the functioning of standard-setting methods with application to bookmark and modified Angoff methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory and Cognition, 22*(1), 63-69.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of National Assessment of Educational Progress achievement levels. *Proceedings from the Joint Conference on Standard Setting for Large-Scale Assessments.* Washington, D.C.: National Assessment Governing Board and National Center for Education Statistics.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement tests.* Stanford, CA: National Academy of Education.

Sireci, S. G., & Biskin, B. H. (1992). A survey of national professional licensure examination programs. *CLEAR Exam Review, 3*, 21 25.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*(4), 259-274.

Taube, K. T. (1997). The incorporation of empirical item difficulty data into the Angoff standard-setting procedure. *Evaluation & Health Professions, 20*, 479-498.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19*(4), 295-308.

van der Linden, W. J. (1986). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting (Addendum). *Journal of Educational Measurement, 23*(3), 265-266.

Verhoeven, B. H., van der Steeg, A. F. W., Scherpbier, A. F. F. A., Muijtjens, A. M. M., Verwijnen, & van der Vleuten, C. P. M. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education, 33*, 832-837.

Wuensch, K. L. (2003). Inter-rater agreement. Retrieved from http://core.ecu.edu/psyc/wuenschk/docs30/InterRater.doc