國民教育研究學報，第 16 期：157-193
2006 年國立嘉義大學國民教育研究所

精熟標準設定方法的歷史演進與詮釋的新概念[1]

謝 進 昌

國立政治大學教育學系博士班研究生

摘　　要

本文旨在介紹近幾十年間，精熟標準設定方法與議題的演進與發展，並從中粹取出有效的訊息，衍生出三項詮釋的新概念，分別為：元素的搭配組合與調整、廣義測驗建構流程、多元效度等，期望能以此重新檢視精熟標準設定的方法，以提供未來研究者於實務運用時不同的啟發與思考。

關鍵詞：精熟標準設定、元素的搭配組合與調整、廣義測驗建構流程、多元效度

[1] 本文與測驗學刊第 52 輯第 2 期「以最大測驗訊息量決定通過分數之研究」一文，皆精進修改自謝進昌(2005)碩士論文，但本文以評閱精熟標準設定方法與新概念提出為核心，該文則特以此新概念為基礎，提出實證研究證據為主體，兩者取向乃不盡相同，但於文獻評閱上會出現少部份雷同之處，在此特以引註說明以釐清。另，作者需特別感謝指導老師余民寧教授與兩位匿名評審的批評與指正，才得以讓本文順利完成。

壹、前言

「及格」對我們而言，是學生時代再熟悉不過的字眼，而對其最直接的聯想即是60 分，但是否曾思考過這個標準是如何訂定？是否存在著合理性呢？而在傳統學校評量外，於國家級證照考試上，其通過的標準又是如何設定？是否同樣具有堅強理論支持呢?若從學術研究角度而言，此類問題即是在探討效標參照測驗中，該效標或精熟標準的設定方式。關於精熟標準設定方法，至今仍不斷的演變與精進，在國外，依舊是屬於熱門的話題，所提出的方法十分多元與龐雜，尤其是伴隨著試題反應理論（item response theory, 簡稱 IRT）與相關統計、評量、電腦科技的發展，更使得此議題愈趨活絡。Berk 在 1986 年時，就曾確認過有 38 種方式被發展出，而至 1996 年時，更表示有近 50 種的方法被發表。面對精熟標準設定方法不斷更新的訊息與技術，但國內多數民眾所保留的概念，卻仍是停留於傳統的 60 分，即使是攸關你我權益的國家級證照檢定，其設定的標準亦是如此（考選部，2005），而此僅是一個長久以來約定成俗的習慣標準，乃屬於武斷、主觀決定的方法。

有藉於此，本文之目的即是希望能將國內、外過去幾十年間，主要發展出的設定方法，從歷史演進的角度作簡要描述，以加深國人對於精熟標準設定方法的瞭解，另期望從方法與相關議題探討中，粹取出新的元素，以重新詮釋目前所發展中的精熟標準設定概念，希望能帶給未來研究者與實務工作者不同的想法與啟發，以利後續的運用。

貳、精熟標準設定

在精熟標準設定（standard setting）中，不論是教育或證照發放用途，所提出的方法可謂非常多元，所關心者皆是欲確認出何種精熟標準，才能真正有效區別出所謂的精熟／未精熟者、或者說有資格／沒資格獲得證照的個體。同時，在分類上亦是各家雜陳，自早先 Meskauskas（1976）從能力觀點將其分為狀態模式（state models）與連續模式（continuum models），經 Jaeger（1989）、Kane（1994）、鄭明長、余民寧（1994）的整合與簡化為以測驗試題內容為判斷依據的「測驗中心模式」（test-centered model）與以受試者能力或實際表現為判斷依據的「受試者中心模式」（examinee-centered model）兩類。

演變至今，隨著精熟標準設定方法因實務應用的需求而不斷擴增（如於實作評量或多元計分試題的運用），分類機制上為避免描述的自限，因而 Hambleton、 Jaeger、Plake 與 Mills（in press, 引自 Pitoniak, 2003）乃建議以六個著重於「判定過程中要素」的分類面向：一、評審判定焦點：著重於試題、受試者亦或是受試者對試題反應等；二、評審判定任務：著重於評估最低能力者在試題上表現、將受試者反應分類等；三、判定過程：著重於評審是採個別或團體判定方式、提供評審於判定時回饋的訊息類型等；四、評審人數與組合方式：評審成員類型、同質或異質程度等；五、精熟標準效度形成方式：著重於是採內在效度證據、外在效度證據等；六、評量的本質：著重於檢視試題類型：選擇題或建構反應試題、計分方式等。

隨著電腦科技的精進，使得電腦化精熟判定方式成為另一項探討主軸，主要可分兩個領域：第一，探討從人工智慧（artificial intelligence）和認知科學所發展出的專家系統（expert system），另一則從心理計量和教育領域發展出的電腦化適性精熟測驗（computerized adaptive mastery testing）（Frick, 1992）。此外，另有一類是屬於 Berk（1986）所稱「調整通過分數」的方法，但 Berk 只將其視為前者的輔助，本身並非設定通過分數的方法，而鄭明長、余民寧（1994）認為此類方法未考量學生能力和試題內容，只偏重依分類錯誤的損失來調整適當通過分數，作法適切性頗值得懷疑，有藉於此，本文在此並不將其視為精熟標準設定方法。

在綜整上述各學者、專家的分類後，大致可形成如圖 1 所示之精熟標準設定分類架構。而此，可代表著近幾十年內精熟標準設定方法歷史演進架構，本文則欲從中粹取出相關訊息，以研擬新的詮釋概念。首先，乃從精熟標準設定方法的介紹，延伸出第一項新概念--元素的搭配組合與調整；而後，從設定方法的相關議題探討中輔以效標參照測驗建構的觀念引導出第二與第三項概念--廣義測驗建構流程與多元效度證據。最後，說明此三項概念於實務上的用途，茲依序陳述如下。

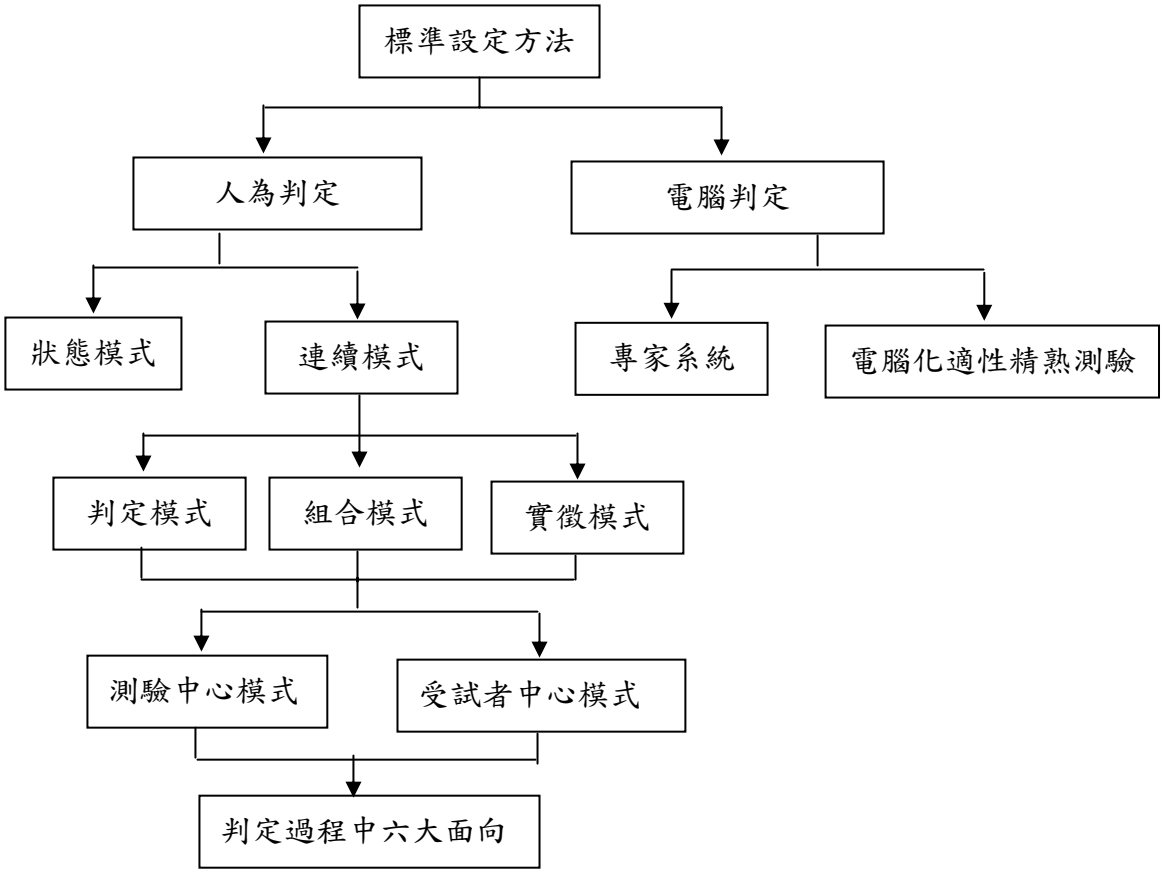

圖1 精熟標準設定方法分類圖（引自謝進昌，2005）

一、元素的搭配組合與調整

在精熟標準設定方法的介紹方面，本文僅以探討人為判定方法為主，同時，為理解的方便性，仍採目前較廣為通用的測驗中心與受試者中心模式分別描述之。在測驗中心模式下，在此主要以 Nedelsky（1954）、Angoff（1971）、Ebel（1972）三種方法為核心，依次就其延伸方法描述之，茲簡述如下：

（一）Nedelsky 法（Nedelsky's method）

Leo Nedelsky 於 1954 年提出，其核心理念乃要求評審針對試題中的作答選項（response options）作判斷。實際執行時，首先，會提供每位評審一份如表 1 之 Nedelsky 法之評判者記錄表與測驗題本，進而要求評審審視第一個試題並找出「最低能力表現」（minimally competency）學生或 Nedelsky 所稱 F-D 學生（意指 F：Failing 與 D：barely

passing 的受試者）能指出錯誤的作答選項，再將該題剩餘選項，取其倒數，此機率值即為該題的最小通過水平（minimum pass level）（如，第五題 4 個選項中，評審認為最低能力表現學生至少能辨識出 1 個誘答選項，則其最小通過水平為 1 除以 3 = 0.33），如此，將所有試題按此方式進行，最後，將每位評審於每一試題機率值加總，再進一步求所有評審平均值，以此作為精熟標準。其中，Nedelsky 乃假定評審能區辨哪些選項對最低能力表現學生而言，是具有吸引力，且認為學生對於不會作答題目，會先挑出被認為絕對錯誤的選項，而後就剩下的選項，採隨機猜題方式作答。因而，評審若認為最低能力表現學生能正確回應多數試題中選項時，則整體機率值則會偏高，代表對於評審心目中最低能力學生而言，試題是偏簡單的，則精熟標準自然會較高，反之亦然。

表 1

Nedelsky 法之評判者記錄表範例

Judge's Recording Form
NEDELSKY METHOD

Question Number	Circle Numbers of Choices identified	P	
1	0 1 2 **(3)**	1.00	
2	0 1 **(2)** 3	.50	
3	0 1 **(2)** 3	.50	
4	0 1 2 **(3)**	1.00	
5	0 **(1)** 2 3	.33	
6	0 **(1)** 2 3	.33	
7	0 1 **(2)** 3	.50	
8	**(0)** **(1)** 2 3	.33	SUM
9	0 1 2 3	.25	5.07
10	0 **(1)** 2 3	.33	

註：取自 Zieky & Livingston（1977, p.5）

（二）Angoff 家族

1、Angoff 法（Angoff's method）

Angoff（1971）曾概略提出相關理念，乃要求評審針對一群最低能力表現者，判斷可能正確作答某試題的機率值，再進一步將各試題機率值加總，即代表最低可接受分

數（minimally acceptable score）。實際執行時，會先給與每位評審一份如表 2 之 Angoff 法評判者記錄表與測驗題本，進而要求評審開始審視第一個試題並評定最低能力表現學生於每一試題可能答對機率值，如此，將所有試題按此方式進行，最後，將每位評審於每一試題判定機率值加總，再進一步求所有評審平均值，以作為精熟標準。流程上 Angoff 法與上述 Nedelsky 法類似，主要差別在於兩者對於整個試題中判定的焦點核心不同。而相較於 Nedelsky 法，此法則屬於較直接方式，若評審心目中的最低能力表現學生能正確回答測驗中多數試題時，則機率值相對較高，表示試題是偏簡易，其精熟標準自然較高，反之亦然。

表 2

Angoff 法之評判者記錄表範例

Judge's Recording Form ANGOFF METHOD		
Question Number	Estimated Probability	
1	1.00	
2	.90	
3	.80	
4	.70	
5	.35	
6	.45	
7	.25	
8	.30	
9	.25	SUM
10	.25	5.25

註：取自 Zieky & Livingston(1977, p.6)

2、改良式選擇型 Angoff 法（modified multiple choice Angoff）

由美國教育測驗服務社（Educational Testing Service, ETS）於 1976 年提出（Berk, 1986），為有效凝聚評定結果，此法進一步將判定的機率值具體化，直接給予固定的七個百分率（5%、20%、40%、60%、75%、90%、95%），要求評審選擇最接近自我主觀判斷的標準，如果評審無法在上述七個百分率中決定哪一個，則可以選擇不知道（Do not know）。最後，將每位評審於每一試題上決定之機率值加總，再進一步求所有評審平均值，以作為精熟標準。

3、Yes／No 的 Angoff 法(two choice Angoff)或稱修正的 Nedelsky 法(modified Nedelsky）

Nassif（1978）的想法乃要求評審判定對最低能力表現受試者而言，是否能正確回答某試題，若認為受試者可以正確回答，則評定為「yes」，不能正確回答時則評定為「no」或選擇不知道，最後，再根據評審所判定的「yes」題目於整份測驗中所佔百分比，以此作為精熟標準。在概念上，相較於此，Nedelsky 法則是屬於要求評審對於試題中選項作 yes／no 判定（如，此選項對於 F-D 學生而言，是否能正確辨識呢？）同時，對照改良式選擇型 Angoff 法，更是簡化判定時認知上的複雜性，減少評審間評定的變異（variability）。

4、反覆二選 Angoff 法（iterative two choice Angoff）（或稱 Jaeger 法（Jaeger's method））

概念上如同 Yes／No 的 Angoff 法，Jaeger（1978）將可能判定的機率值具體化為兩種選擇，但差別在於加入需反覆執行過程（iterative process），即是給與評審討論先前所評定結果的機會，以供調整時參考。Jaeger 於 1982 年時，將此進一步發展為結合整體判斷、考生表現、試題判斷的方法，稱之為反覆性結構的試題判斷過程（iterative structured item judgment process）或 Berk（1986）所稱反覆二選 Angoff 法。實際執行時，評審需反覆三次，每次都詢問自我以下問題：每位畢業學生能否正確回答這個題目呢？如果一個學生無法答對這個題目，是否就不應給與文憑？諸如此類的問題，概念上同樣是針對試題作 yes／no 判斷，但相較於傳統 Angoff 或 Nedelsky 法，此法強調對所有學生或受試者作判定，評審則不需在心目中概念化所謂最低能力表現者。此外，在反覆過程中額外提供三類參照訊息（normative information）：首次評定後，其它評審建議之標準分配、評審本身先前評定結果、依學生真實表現所得之試題難度值。理念上乃希望藉由反覆過程與提供參照資料以減少評審內（intrajudge）與評審間（interjudge）判定的變異性，並藉由實際問題以具體化界定最低能力表現。於第三輪中，以所有評審判定的最小中位數值，即為精熟標準。

5、修正的／改良式選擇型 Angoff 法（adjusted/modified multiple choice Angoff）

Bernknopf、Curry 與 Bashaw（1979）針對改良式選擇型 Angoff 法再加以修正，評審需就每一試題，分別從 9 個百分比中（15%、25%、35%、45%、55%、65%、75%、85%、95%），選擇 1 個他認為最低能力表現學生應答對之百分比，而後，以此求得各試題平均百分率之標準誤調整所評定試題的機率值，以使假精熟或假未精熟之分類誤差達最小。最後，得到之試題機率值再依隨機猜測誤差加以調整，而每位評審校正後的所有試題機率值的總和，求其評審平均值，即為精熟標準。

6、反覆 Angoff 法（iterative Angoff）

由 Saunders 與 Mappus 於 1984 年提出，此法類似於結合 Angoff 法與 Jaeger 法，在反覆三輪的過程中，同樣要求評審評定最低能力表現學生可能正確作答試題的機率值，而在評審整體性考量過自我設定的標準並檢視全部學生測驗分數分配與第 2 輪建議的標準所決策之相關描述統計結果後，精熟標準則在所有評審的共識（consensus of the judges）下決定。

7、評定量表法（rating scale method）

如同 Angoff 法對於試題作機率值判定，吳裕益（1986）起初乃要求評審主觀判定試題難度值，而後，再將各試題依判定結果，依續分派至各難度等第（依判定測驗的總題數不同，可分為 5、7 或 9 個等第），之後，再決定各難度等第對於最低能力表現學生而言，其通過機率為何？最後，求各等第下題數與其相對通過機率乘積，加總再求評審平均，即為測驗精熟標準，相關評定量表評審記錄表如表 3 所示。概念上則將「各試題間難度的比較」納入考量，避免執行 Angoff 法逐題審視時，而忽略試題間相對關係，藉此以提高評審間評定結果的一致性。

表 3

吳裕益評定量表法評審記錄表

難度	1(最易)	2	3	4	5(最難)
理論百分比	7	24	38	24	7
題號					
題數					
評定之通過機率					
題數×評定之通過機率					

註：引自吳裕益（1986, p.216），為五點量表，30 題以下適用。

8、Angoff 衍生法（Angoff derivative method）

美國全國教育進步評量（National Assessment of Educational Progress, NAEP）自 1994 年起，即應用多種 Angoff 法延伸的精熟標準設定方法於多元計分試題上，Loomis 與 Bourque（2001）將此四種方法分別稱之為：正確百分比法（percent correct method）、比率法（proportional method）、平均估計法（mean estimation method）、ISSE 法（the item score string estimation method）。

正確百分比法概念上乃要求評審評估有多少百分比的最低能力表現受試者於試題上至少能部分正確反應（partially correct response），例如，在某多元計分試題上，其計

分準則（scoring rubrics）乃以 1 分代表不正確、2 分代表部份正確、3 分表示完全正確作答等，如同二元計分般，執行時同樣將分數判定區分為 2 個區塊：2 分（含以上）與 1 分（含以下），即要求評審判定最低能力表現受試者可能得分 2 分（含以上）的百分比值。

比率法則要求評審評估最低能力表現受試者於試題中每個計分準則分數點上（each rubric score point）可能反應的機率，例如，判定最低能力表現受試者可能得 1 分、2 分、3 分等的百分率，相較於正確百分比法，此法則納入了部分計分的考量（乃考量各個分數點，而非如上述將其分成 2 個區塊）。

平均估計法在概念上則非常直接，乃要求評審判定最低能力表現受試者在每個多元計分試題上可能獲得的平均分數，以此為基準再求整份測驗之精熟標準。

ISSE 法如同 Yes／No 的 Angoff 法般，僅是加以延伸至多元計分試題，乃要求評審判定最低能力表現受試者在每個多元計分試題上，是或否能得到 1 分、2 分、3 分。

9、延伸的 Angoff 法（extended Angoff approach）

伴隨著實作評量的發展，使得精熟標準設定不僅需建立於紙筆測驗上二元計分的選擇反應（selected response）試題或多元計分的建構反應（constructed response）試題，更應將其延伸至實際作品評量。如同傳統的 Angoff 法判定試題的正確反應機率值，Hambleton 與 Plake（1995）乃要求評審評定最低能力表現者在每個多元計分的實際表現上可能獲得的期望分數。同時允許評審依據實作內容中不同計分重點與以加權，此外，並提供團體討論的機會、各階段中標準設定影響結果、精確未精熟的作品類型等。整體而言，此法在概念上只是延伸 Angoff 法於實作評量上，及同時融合如 Jaeger 法中幾項元素（如反覆執行、提供參照資料等）。

10、認知元素法（cognitive component approach）

相較於 Angoff 法於整個試題作判定過程，McGinty 與 Neel（1996）則要求評審將試題切割為數個互相獨立的認知元素（cognitive component）。舉例而言，某數學試題為：516+193+232 等於多少？受試者要正確回答此問題時，則需先具備幾項認知能力：（1）瞭解“等於“的意函；（2）懂得“+“代表加總的意思；（3）知道如何列出三位數加法的式子；（4）知道三位數加法運算方式；（5）懂得應用基本數學運算。執行時，評審則被提供有關此類認知元素的相關描述，並詢問“為了通過這份測驗，受試者必須有能力正確應用此類技巧至少多少百分比的次數。“換句話說，評審乃以判定最低需正確應用此技巧於需要它的情境中的比率（註：但並非詢問有多少百分比的試題需要此類正確作答的技巧）。此比率值則稱之為最小成功比率（minimum success rate），即代表著

最低能力受試者能正確應用此認知元素的機率值，而後，將試題中所有元素評審判定的平均機率值加以相乘，求試題判定結果的總和，即為整份測驗之精熟標準。

11、書簽技術（bookmark method）

此法乃結合同時檢視試題內容與真實受試者反應的技術，Lewis、Mitzel 與 Green（1996）要求評審逐一檢視經由 IRT 事先計算出的難度值加以由易至難排序的試題卷（ordered item booklets），同時提供評審一份條例著試題在排序後與排序前於測驗卷中所在位置與各試題所欲測量的內容領域或知識等資訊的試題圖（item map）以供參照，之後，評審被要求放置一個書簽（bookmark）於檢視的試題圖中，其認為最低能力表現受試者應有 3 分之 2 機率（約 67%）知道或能正確作答的 2 個試題間，此外，若加以延伸則可如圖 2 所示，判別多個不同能力標準（B：Basic；P：Proficient；A：Advanced），而精熟標準則根據評審選擇的兩個試題所代表的 IRT 難度值加以計算。由於此法乃融合 IRT 技術與 Angoff 法概念，因而 Lewis、Green、Mitzel、Baum,與 Patz（1998）又將其稱為修正的 IRT-Angoff 法（IRT-Modified Angoff Procedure）。

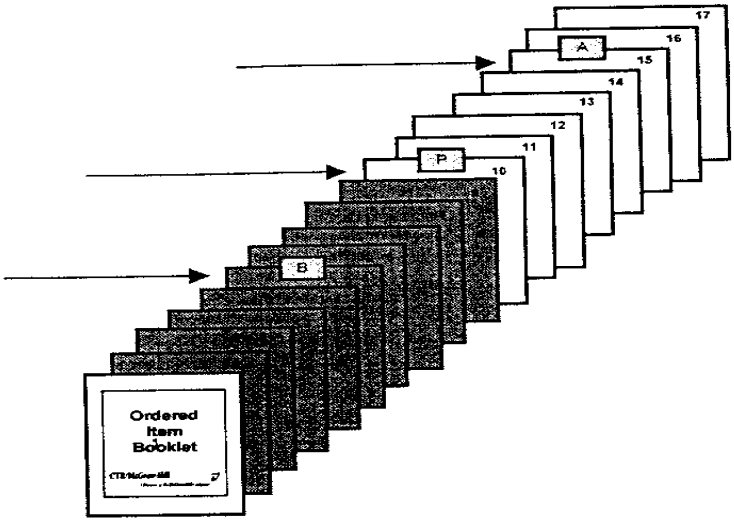

圖 2 書簽技術中排序的試題卷（取自 Mitzel, Lewis, Patz, & Green, 2001, p.256）

12、試題構圖法（item mapping method）

試題構圖法的概念與書簽技術可說雷同，同樣要求評審在依難度排序的試題中，尋找出其心目中最低能力表現受試者應有在特定機率值下知道或正確作答的 2 個試題，但 Wang（2003）認為兩者間在判定過程上仍有幾項相異點：（1）認為最低能力表

現者應至少具有 50%的機率（非上述 67%，乃因 Rasch 模式在此機率值下具有較大試題訊息）答對該試題；（2）額外提供一份如圖 3 所示之各試題難度計算結果的直方圖（histogram chart），使得評審能以更宏觀角度檢視所有試題相對位置（圖中三角黑點係為各試題相對於 X 軸之難度值位置）；（3）目的上，書簽技術主要應用於教育評量，因而傾向設定多個精熟標準（multiple levels），試題構圖法多應用於證照考試為主，傾向只設定精熟／未精熟標準；（4）書簽技術仍要求評審仍需逐題檢視判定，而試題構圖則僅需選擇具難度水平與測驗內容的代表性的試題作判定。而最後精熟標準則根據評審凝聚的共識試題決定。

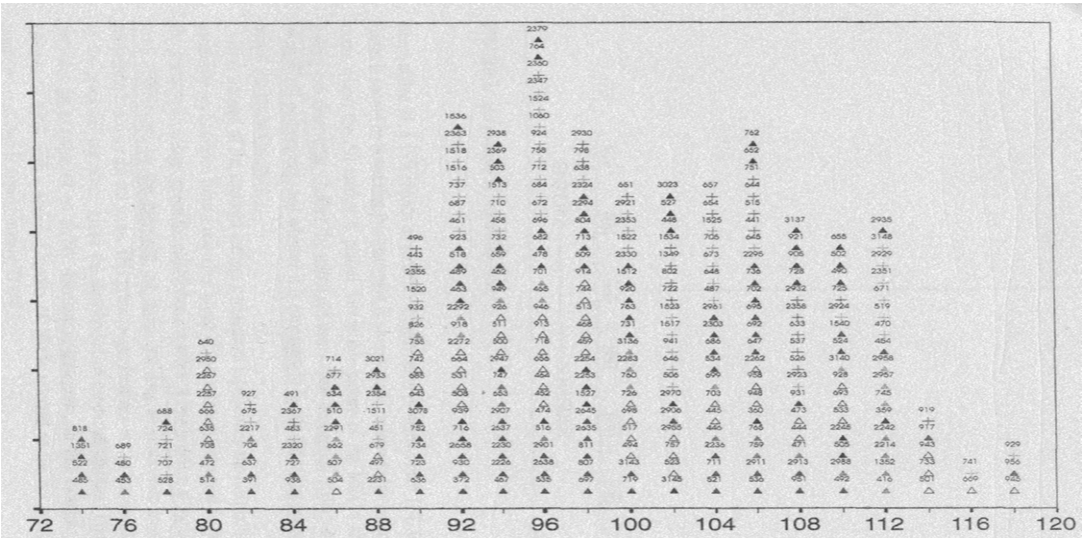

圖 3 試題構圖法之直方圖範例（取自 Wang, 2003, p.234）

13、IDEA 法（interdependent evaluation of alternatives method）

傳統 Angoff 法概念上著重於試題本身的判定，而 Nedelsky 法則強調試題的選項，雖然這兩種方法在執行時，皆要求評審需對試題本身（題幹與選項）作完整檢視以決定評定結果，但實際上評審判定時，仍易受方法著重核心不同的影響，如 Angoff 法的評審則易將焦點集中於檢視試題題幹與正確選項上，較不易受誘答選項吸引，因而多低估試題難度，而 Nedelsky 法評審則易受誘答題項的吸引，忽略了檢視正確選項，因而會喪失檢視潛藏在正確選項中的訊息，而高估試題難度。有藉於此，Chang、van der Linder 與 Vos（2004）企圖融合兩者，要求評審需考量到整個試題（題幹、正確選項、誘答選項）透露的訊息，並判定在現行題幹描述下，相對於其它選項，最低能力表現

受試者於每個選項上可能正確作答的機率值（前提：各選項判定結果總和需為 1）。其中，作相對選項的檢視，乃認為受試者作答時，並非針對個別選項上作絕對判斷，而會考量到選項間彼此的關係，而試題精熟標準則由所有評審於正確選項上（其它選項機率值則可加以忽略）所判定的平均機率值決定。

（三）Ebel 家族

1、Ebel 法（Ebel's method）

Ebel 於 1972 年提出，概念上乃藉由試題的特性來決定最低通過分數。評審首先根據試題的四種適切性（relevance）或重要性（importance）：基本必備的（essential）、重要（important）、尚可（acceptable）、存疑（questionabl）及三種難度：容易（easy）、適中（medium）、艱深（hard）形成一個 4×3 雙向細目表，然後，依據各試題的特性，經判定後分別將其置入各細格內，而後，再針對每一細格的重要性給予不同的權數（如表 4 為 Ebel 所建議之加權係數，認為簡易且基本必備的題目是最低可接受能力者應100%正確回答的，則給與此權數，其餘概念可以此類推，而此權數多寡乃評審可自由調整）。最後將各試題與權數相乘、加總，再求其平均試題權數，即為精熟標準。

表 4

Ebel 法中測驗試題的適切性、難度與期望成功機率值

Relevance Categories	Difficulty Levels		
	Easy	Medium	Hard
Essential	100%	---	---
Important	90	70%	---
Acceptable	80	60	40%
Questionable	70	50	30

註：取自 Ebel（1972, p.493）

2、難度-目標分類 Ebel 法（difficulty-taxonomy Ebel）

Skakun 與 Kling（1980）乃將針對傳統 Ebel 法稍作調整，將試題分類的特性區分為難度：容易、適中、艱深與目標分類：事實（factual）、理解（comprehension）、問題解決（problem solving），形成一個 3×3 雙向細目表，並由評審預先將題目歸到各細格中，而後，經由評審判定最低能力表現者應能正確回答不同特性細格之試題機率值（即是上述權數），再將此機率值乘上該細格試題數，加總，求其平均試題機率值，即為精熟標準。

3、適切性-目標分類 Ebel 法（relevance-taxonomy Ebel）

在概念上，Skakun 與 Kling（1980）同樣只是將原先 Ebel 法試題分類特性中的難度改以試題適切性取代，以形成一個 4×3 雙向細目表，而後，其評定歷程與傳統 Ebel 法皆相同。不論是難度-目標分類 Ebel 法或適切性-目標分類 Ebel 法，概念上皆等同於傳統 Ebel 法，差異點只在於試題著重的分類面向不同。

反觀在受試者中心模式下，若以測驗編製者的角度視之，由於受試者的表現，乃屬於無法控制的因素（uncontrollable factor）（如在大型證照測驗下，是無法完整掌握來應試者的特質），因而，使得此模式方法不僅在實用或學理上，皆不似以測驗中心模式的方法堅強，但隨著實作評量發展，此類方法亦日驅多元，且概念亦漸漸難與測驗中心模式區分，但為解說方便，在此仍依其主要概念將其分類為受試者中心模式，茲介紹如下：

（一）臨界組法（the borderline group method）

Zieky 與 Livingston（1977）於理念上乃要求評審事先找出一組被判定為未達精熟，但也非未精熟的學生，亦即處於精熟／未精熟的模糊狀態，對此將之稱為「臨界組」（borderline group），然後求此組學生於測驗上表現分數的中位數（median）（此統計量較平均數不受極端值影響），即將此視為精熟標準。

（二）對照組法（the contrasting groups method）

1、對照組圖形法

相較於臨界組法，對照組圖形法（Zieky & Livingston, 1977）恰可與之作一個對比，此法並非以界定精熟／未精熟模糊狀態之臨界組為目的，而是希望尋找出能明確界定為精熟與未精熟的學生，再將此二群人之測驗得分分配曲線畫出，而如圖 4 所示取其兩曲線的交叉點，即視為精熟標準（亦可針對錯誤分類的重要性加以調整），其機制在於認為此交叉點所形成之分類錯誤是最小的，同時於界定學生精熟表現上，是較上述判定模糊狀態為簡易。而 Brandon（2002）進一步從過去相關研究中加以歸納，認為可從兩方面觀點來檢視此法：受試者為中心觀點（person-focused version）、受試者反應為中心觀點（response-focused version），前者乃屬傳統觀念，強調著以接受測驗的受試者（people take the examination）為主體，評審是以選擇精熟／未精熟的受試者為主要任務，後者則強調受試者已完成的測驗（examinees' completed examinations）反應，評審以分類精熟／未精熟的作答反應為首要重點。Webb 與 Miller（1995）即曾以 Brandon（2002）所稱受試者反應為中心的對照組法於實作評量上，乃要求評審針對學生實際作品加以分類，精熟標準則依分類為最低具競爭力作品與分類為不具競爭力作品的分

數決定。

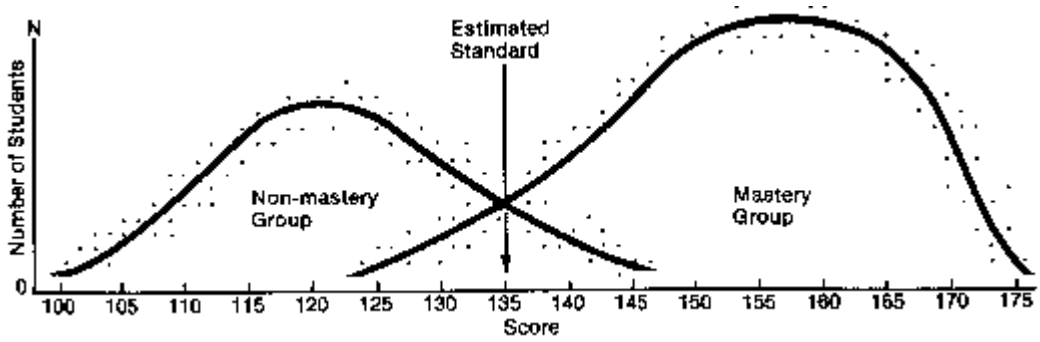

圖 4. 對照組圖形法之範例圖（取自 Zieky & Livingston, 1977, p.9）

2、對照組 LDF 法

傳統對照組法乃採用圖形方式以決定精熟標準，而 Koffler（1980）則認為在判定精熟／未精熟者後，亦可使用「明確的統計方法」獲得精熟標準，因而採用直線區別函數（linear discriminant function, LDF）（Fisher, 1936）的概念。此法需預先假定精熟與未精熟者之測驗分數呈常態分配、變異數亦相等，在母群體參數未知情況下，使用樣本估計數來替代，認為其最佳分類方式可為：

$$\left[\left(\overline{X_1}-\overline{X_2}\right)/S^2\right]\left[\left(T-\left(\overline{X_1}-\overline{X_2}\right)/2\right)\right] \qquad \text{（公式 1）}$$

其中，$\overline{X_1}$、$\overline{X_2}$ 分別代表判定後精熟組與未精熟組之平均測驗得分；S^2 代表合併後樣本變異數；T 代表整體學生之測驗分數。如此，將整體學生之測驗得分逐一代入公式 1 中，其結果再與 $\log(q_2/q_1)$ 求得之結果相比（其中，q_1、q_2 分別代表被判定為精熟者與未精熟者所佔比率），最後則以公式求得之結果，能大於 $\log(q_2/q_1)$ 中之最小測驗分數，即視為本次測驗精熟標準。

3、對照組 QDF 法

相較於對照組 LDF 法，其乃假定兩群人變異數是同質的情況下，若違反時，其韌性（robustness）則會顯得較差（Gessaman & Gessaman, 1972），認為可改採以二次區別函數（quadratic discriminant function, QDF）解決，公式為：

$$T\left(\overline{X_1}/S_1^2-\overline{X_2}/S_2^2\right)-\frac{T^2}{2}\left(1/S_1^2-1/S_2^2\right)-\frac{1}{2}(\overline{X_1}^2/S_1^2-\overline{X_2}^2/S_2^2)+\frac{1}{2}\log(S_2^2/S_1^2)$$

（公式 2）

其符號意函與公式 1 相同，另外，S_1^2 、S_2^2 分別代表精熟與未精熟者測驗得分變異數。執行上，同樣以公式 2 求得之結果，能大於 $\log(q_2/q_1)$ 中之最小測驗分數，即視為本次測驗精熟標準。

4、對照組等級 QDF 法

若違反常態分配假設時，在採用上述方式時，其韌性亦是顯得較差，因而 Conover 與 Iman（1978）乃建議採用等級轉換（rank transformation）方式，概念上乃事先將整體分數，由小至大排序，再依序自 1 至 n 給與等級，之後，將此等級分數視為各學生之測驗得分，代入公式 1，以求得精熟標準。

5、對照組 M-SD 法

吳裕益於 1986 年提出，理念上除為解決以圖形方式決定精熟標準所產生的缺點（如，資料分佈均勻時易造成誤差、人數少時分佈曲線易有不規則情況），而改以統計方法求得精熟標準點外，另一方面，則期望較 LDF、QDF 法潛顯易懂，因而提出如下之公式：

$$M_1-\left[S_1(M_1-M_2)/(S_1+S_2)\right]$$

（公式 3）

其中，M_1 、M_2 分別代表精熟、未精熟組測驗平均數；S_1 、S_2 分別代表精熟、未精熟組測驗標準差。將測驗而得的各數值代入公式中，所得結果即為對照組 M-SD 法求得的精熟標準。

（三）Berk 效標組法（Berk criterion group validation model）

早先於 Zieky 與 Livingston（1977）提出對照組法時，Berk 於 1976 年時即提出類似概念，其中乃將對照組法中精熟組與未精組的定義予以具體化，因而假定一個操作型定義：以接受講習者則視為為精熟者，未接受者則視為未精熟者，對此將之稱為效標分類（criterion classification），接著當測驗實施後，任選一個分數將學生分為通過與不通過，此分類稱為預測分類（predictor classification）。因此，在每個分數下皆可得一個 2×2 的細目表，再針對各種不同預測分類的分數逐一求其正確分類概率，而以正確

分類概率最大者，該分數即視為精熟標準。

上述乃以所選分數能達最大正確決定概率者即視為精熟標準，除此之外，亦可採用效度係數（validity coefficient）（即兩個二分變項間ϕ相關係數）最大者，即視為最佳精熟標準。對此，亦可延伸出計算效用分析（utility analysis）：評估分類錯誤時需付出的相對代價與損失（如醫師等執照發放，若將受試者誤判為精熟時，會造成較大損失，則評審可主觀判定應提高精熟標準，以降低此錯誤），來作為精熟標準的調整；增量效度（incremental validity）：用以比較使用此測驗所得訊息，和其它方式所獲得訊息間的相對大小。

（四）傳統直觀方法

Cascio、Alexander 與 Barrett（1988）乃提出幾項傳統直觀方法，基準法（base rate method）與迴歸法（regression based method）乃分別利用效標資料以決定基本最低能力表現者的比率值與預測的迴歸公式以求得精熟標準；標準差法（standard deviation method）則依據受試者在測驗分數上的表現所得之平均數與標準差，相互搭配以求得精熟標準或另加以延伸，如鄭明長、余民寧（1994）採行 IRT 計算出平均 θ 能力值與標準差取代此古典測驗理論的計分方式；測驗分數百分比法（percentage of test score method）亦是單純依據過去經驗或其它考量因素直接決定以多少百分比值決定精熟標準；考試院直接以固定 60 或 70 分為精熟標準等。

（五）集群分析法（cluster analysis method）

為了避免測驗中心模式下，評審需主觀的評定最低能力表現受試者的影響，Sireci、Robin 與 Patelis（1999）乃以較客觀的統計方法--集群分析，應用於精熟標準設定上，概念上乃將受試者依據各種分類變項加以分成幾個集群。分析時，團體中每位受試者會依其距離集群重心（cluster centroid）的距離值，加以分配至表現最相近的集群中，而欲使得相同集群內受試者表現差異最小，不同集群間受試者表現差異最大。此外，在分類變項的選擇上，Sireci, et. al.認為可採用受試者於個別測驗試題上表現、試題因素分析後所得直交因素分數、各次量表總分等，最後，就研究者所採用集群數，可運用臨界組法概念（一個集群解）或對照組法概念（二個集群解）決定精熟標準。

（六）課程參與法（course enrollment method）

Giraud、Impara 與 Buckendahl（2000）提出一項牽涉到課程安置（course placement）的精熟標準設定方法，其概念上需事前將受試者分配至不同難度水平的課程（course）上，而後，求得其於各課程內平均的測驗分數。最後，再選擇最適合或最符合受試者

能力水平的課程，即以受試者於此課程內平均測驗分數為精熟標準。

（七）評審的期望法（expert expectation method）

Giraud 等人（2000）提出一項與 Dillon（1996）類似的精熟標準設定方法，兩者在概念上，皆要求評審判定有多少百分比的學生。評審的期望法乃詢問評審學區內有多少百分比學生是屬於低於最低能力表現者，而 Dillon 則是詢問評審有多少百分比的學生已準備畢業或具有能力進階至下一年級。前者主要以判定最低能力表現者“以下“百分比數，後者則將焦點集中在判定“以上“的百分比數。

（八）分析或整體判定法（analytic or integrated judgment method）

實作評量的發展，使得多元的精熟標準技術日漸成熟，但專家、學者所提方法在概念上十分雷同，皆要求評審在檢視受試者於作品或試題上的反應後，根據能力標準的描述，將其分類至各能力水平類別內（如挑選符合臨界組、精熟、未精熟的代表性作品）。但此類的方法，若依評審判定焦點的差異，可將其分為兩個面向，第一種著重於要求評審於分類時，將焦點集中於實作內容中個別元素的判定（如以分類個別建構反應試題為核心），此類方法以 PS 法（paper selection method）（Loomis & Bourque, 2001）或 Plake 與 Hambleton（2001）所稱分析判定法（analytic judgment method）為主；第二種則著重於對學生作品或表現作整體判斷（holistic judgment）（如以分類完整測驗卷為核心），此類方法以 BC 法（booklet classification method）（Loomis & Bourque, 2001）、BOW 法（body of work method）（Kahl, Crockett, DePascale, & Rindfleisch, 1994, 1995）或 Jaeger 與 Mill（2001）所稱的整體判定法（integrated judgment method）。

（九）JPC 與 DPM 法（judgmental policy capturing method & dominant profile method）

Jaeger（1995）所提 JPC 法與 Putnam、Pence 與 Jaeger（1995）、Plake、Hambleton 與 Jaeger（1997）發展的 DPM 法，皆以評定實作評量精熟標準為目的，操作概念上，與分析或整體判定法類似，乃要求評審將每份實作檔案、作品分類至某個特定的能力標準（如優良、普通、差等），藉此以形成評審自我的分類決策方針（classification decision policy）。而 Putnam, et. al.認為有幾種決策方針類型：1、補償（compensatory）：精熟標準多以所有試題或實作表現上平均得分或總和決定，因而在某項表現較差時，仍可藉由其它方面彌補；2、合取的（conjunctive）：實作內容中精熟標準，乃各自有其底線標準（bottom line）（如認為寫作分數不得低於 3 分等），因而無法藉由其它方面表現彌補；3、析取的（disjunctive）：精熟標準決策方針可給與某些領域的實作表現有加權效果，因而，某些具優勢的實作表現（dominant profile）具有決定受試者是否精熟的效果。而 JPC 法過程中乃應用迴歸分析技術於適配評審最後精熟標準方針，較屬間接方式，且結

果多屬補償類型，相對的，DPM 法則改採以較直接方式，要求評審直接確認他們期望的精熟標準方針，再反覆執行討論後，以整體共識決定精熟標準，因而較能產生同時融合上述 3 種類型的決策方針。

Berk 在 1986 年的作品中，曾認為 1970 年代是精熟標準設定方法發起的初端，1980 年代初期則是方法間的比較研究與標準設定過程的探討，而 1990 年代的發展，Berk（1996）則認為深受實作評量的影響。此類特徵對照於上述相關方法歷史的演進脈絡，則頗有相互印證意味。雖然精熟標準設定方法是不斷的推陳出新，調整式的方法亦是多元，但仍能從中看出其核心理念依舊是不變的，皆期望能真正區分精熟／未精熟者，而在此之下，方法的創新亦多屬「元素間相互搭配或調整」，如調整評審檢視試題時的判定方式（判定試題機率值、可刪除選項、試題重要性亦或是受試者反應等）、考量是否提供參照資料、是否反覆執行、如何計算通過分數方式等等，而此概念即形成本文強調之第一項詮釋核心。

二、廣義測驗建構流程與多元效度證據

隨著精熟標準設定方法持續的發展與精進，圍繞此所探討的議題亦是不斷湧現，而討論的面向大致可分為：（一）方法的比較研究；（二）精熟標準設定過程議題探討；（三）信、效度議題等三方面。首先，在方法的比較研究上，Berk（1986）、Jaeger（1989）、Bontempo、Marks 與 Karabatsos（1998）皆曾綜整此類比較研究的文章（簡要整理如表 5），其結果就如多數研究者所發現，不同的方法所產生結果多不一致，理由上，van der Linder（1982）認為不一致的來源，可能是對於精熟概念理解差異、評審間學習目標的見解不一與評審內判定的不一致；吳裕益（1986）認為各種方法間的差異與標準設定之過程有關；Jaeger 認為即使是相同的判定者使用相同的方法，都不易產生相同的精熟標準，乃因精熟標準方法設定皆涉及主觀的看法、主觀的判定。在上述理由中多指向主觀的精熟標準設定過程差異會致使產生不同的結果，因而，有許多研究開始轉向探討設定過程中所引發的議題。

表 5

歷年不同精熟標準設定方法比較研究一覽表

相關研究	精熟標準設定方法通過分數比較
Andrew & Hecht （1976）	Nedelsky 法、Ebel 法

Schoon, Gullion, & Ferrara （1979）	Nedelsky 法、Ebel 法
Brennan & Lockwood（1980）	Nedelsky 法、Angoff 法
Koffler（1980)	Nedelsky 法、對照組等級 QDF 法
Skakun & Kling, （1980）	Nedelsky 法、難度-目標分類 Ebel 法、適切性-目標分類 Ebel 法、平均數下一個標準差法
Harasym（1981）	Nedelsky 法、Yes／No 的 Angoff 法
Behuniak, Archambault, & Gable（1982）	Nedelsky 法、Angoff 法
Mills（1983）	Angoff 法、對照組圖形法、對照組 QDF 法、臨界組法
Halpin, Sigmon, & Halpin（1983）	Nedelsky 法、Angoff 法、Ebel 法
Reilly, Zink, & Israelski（1984）	Nedelsky 法、Angoff 法
Cross, Impara, Frary, & Jaeger（1984）	Nedelsky 法、Angoff 法、Jaeger 法
吳裕益（1988）	Nedelsky 法、Angoff 法、評定量表法、Ebel 法、臨界組法、對照組圖形法、對照組 LDF 法、對照組 QDF 法、對照組 M-SD 法
Livingston & Zieky（1989）	Nedelsky 法、Angoff 法、臨界組法、對照組圖形法
Woehr, Arthur, & Fehrmann（1991）	Angoff 法、對照組圖形法、基準法、廻歸法、平均數法、平均數下一個標準差法
林惠芬（1993）	臨界組法、對照組 M-SD 法、Berk 效標組法、迴歸法、基準法、平均數法、平均數下一個標準差法、考選部 60 分
鄭明長、余民寧（1994）	平均數法、平均數下一個標準差法、教師判斷法、考選部 60 分、最大測驗訊息量法、最大試題訊息量法、IRT 能力平均數法、IRT 能力平均數下一個標準差法
Impara & Plake （1997）	Angoff 法、Yes／No 的 Angoff 法
Chang（1999）	Nedelsky 法、Angoff 法
Stephenson, Elmore, & Evans（2000）	Angoff 法、Jaeger 法、臨界組法
Giraud, Impara, & Buckendahl（2000）	Yes／No 的 Angoff 法、臨界組法、對照組圖形法、課程參與法、評審期望法
鄭清泉（2001）	Angoff 法、電腦化適性精熟測驗系統
Buckendahl, Smith, Impara, & Plake（2002）	Yes／No 的 Angoff 法、書簽技術
Green, Trimble, & Lewis（2003）	書簽技術、對照組圖形法、整體判定法

註：各研究所採方法，因某些於文中並無詳細註明確切應用方式或是有作其它調整，

因而，在此多以其隸屬的傳統方法或概念上較接近方法來註記。

對於精熟標準設定過程的探討上，Hurtz 與 Auerbach（2003）曾運用後設分析（meta-analysis）方法加以統整相關議題結果、Brandon（2002, 2004）則分別以文獻評閱方式，探討受試者中心模式最常用的對照組法與測驗中心模式中的 Angoff 法（含相關的調整方法）於設定過程中的相關議題，綜整之研究可參考表 6，涵蓋範圍可歸納為下列幾種：

（一）評審相關

以討論精熟標準設定過程中與評審相關之因素，例如，允許評審團體討論或僅限個別判定、可否重新考量修訂其判定結果、提供參照資料的時機（團體討論前、中、後）、評審訓練、評審專業程度、評審人數、評審背景、允許反覆判定、定義最低能力表現受試者（個別定義或者團體共識）、判定地點、判定前接受測驗等，以探討上述諸因素是否會影響判定結果。

（二）試題相關

探討試題中正確選項位置、題幹長度、誘答選項的誘答效用、測驗長度、試題難度、試題描述類型等試的相關因素，對於判定結果的影響。

（三）提供參照資料

探討提供評審參照資料，如試題難度（P）值、測驗試題的答案、前輪或現階段判定結果描述、實際受試者作答分佈、其它評審或本身判定結果與影響、各精熟水平下受試者應具備的特定能力或行為描述、現實層面（教育、財政）上影響等，是否會有助於判定的表現。

表 6

歷年精熟標準設定過程議題一覽表

相關研究	討論議題
Harasym（1981）	探討 Nedelsky 法與 Yes／No 的 Angoff 法在不同試題類型下判定的結果
Behuniak, Archambault, & Gable（1982）	探討在 Angoff 法與 Nedelsky 法下，不同評審群與評審背景（教學年資、年級、職位）等對於判定結果的影響
Halpin, Sigmon, & Halpin（1983）	探討在 Nedelsky 法、Angoff 法、Ebel 法下，不同評審群（研究生、中學教師、大學教師）於不同精熟標準設定方法間互動影響
Cross, Frary, Kelly,	採用臨界組法的概念於作文評分上，探討評審為判定內容的專家

Small, & Impara（1985）	與非專家、提供參照資料對判定結果效益比較
Norcini, Lipner, Langdon, & Strecker（1987）	探討在 Angoff 法判定過程中，團體討論前、中、後提供參照資料之效益
Norcini, Shea, & Kanya（1988）	探討運用 Angoff 法於醫學判定上，評審的專業、提供參照資料對於判定結果的效用
Smith & Smith（1988）	探討運用 Angoff 法與 Nedelsky 法時，判定試題的特徵（正確答案位置、題幹長度、誘答選項的誘答效用等）對判定結果的影響
Plake & Melican（1989）	探討運用 Nedelsky 法時，測驗長度與難度對於評審判定時可能影響
Busch & Jaeger（1990）	探討在調整的 Angoff-Jaeger 法下，不同評審類型（判定內容的專家與非專家）、提供參照資料、允許評審重新考量起初判定結果、允許評審討論起初判定結果對設定精熟標準之效益
Fehrman, Woehr, & Arthur（1991）	探討 Angoff 法下，評審接受不同訓練方法於判定結果之效益
Norcini, Shea, & Grosso（1991）	探討調整的 Angoff 法下，評審的人數、共同試題數於連結兩份測驗的精熟標準時可能的影響
Maurer, Alexander, Callahan, Bailey, & Dambrot（1991）	探討 Angoff 法下，評審的專業程度、評審人數對判定結果影響
Plake, Impara, & Potenza（1994）	探討 Angoff 法下，評審為評定內容的專家與非專家、提供參照資料對判定結果的效益
Hudson & Campion（1994）	探討 Angoff 法下，提供評審參照資料與判定試題難度間可能存在的互動關係以影響判定結果
Chang, Dziuban, & Hynes（1996）	探討在調整 Angoff 法下，評審具有判定的試題相關知識，對其判定結果的影響
Hurtz & Hurtz（1999）	運用概化理論探討在 Angoff 法下，需多少評審數才能達到較穩定的判定結果
Chinn & Hertz（2002）	探討 Angoff 法與 Yes／No 的 Angoff 法下，提供評審各精熟水平下受試者應擁有特定行為描述，對其判定結果的影響
Clauser, Swanson, & Harik（2002）	探討 Angoff 法下，評審經訓練與提供參照資料後於判定過程的穩定效果

在精熟標準設定相關研究發展中，可發現方法的創新乃強調著元素的搭配組合與調整，而研究的議題則由方法間的比較研究，轉至評審判定過程中相關議題的探討，顯示著判定過程中的嚴謹性、合理性，即相對暗示著該精熟標準設定方法的良窳或稱之為是否具有效度。但過去的研究，對於判定過程的概念仍僅限於「方法」內，而非

以較廣概念呈現，對此，輔以圖 5 詳述之，上述有關精熟標準設定方法的研究（如 Angoff 法相關延伸），對於判定過程的探討，如採反覆判定、判斷試題正確作答機率、提供參照資料等等，相對於以能力標準設定為核心的效標參照測驗建構流程，至多僅隸屬第四與第五階段，即使該判定過程具備相當嚴僅、完美表現，仍僅是提供少部份認定此精熟標準所具備優良特性的證據，但為何不從更廣角度視之，更能容納廣範效度證據。

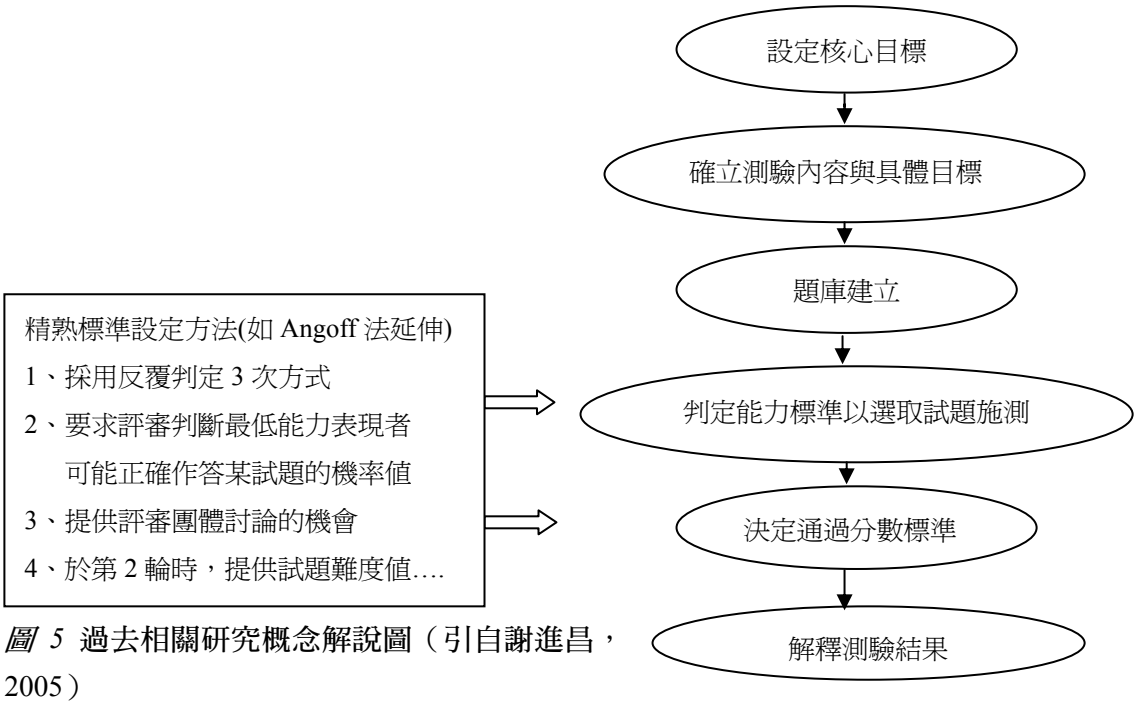

圖 5 過去相關研究概念解說圖（引自謝進昌，2005）

有藉於此，本文企圖以更廣的角度檢視精熟標準設定方法，此不僅考量到各階段間是習習相關，且更能顧慮到所採用方法是否能有效融入整個流程，此概念就如同 Kane（1998）表示選擇方法的準則，不僅在於此方法於設定結果中是否具有準確性（一致性分類精熟／未精熟者），更需關注研究者是如何依據精熟標準分數作決策，即代表著從測驗編製至計分、精熟標準設定、解釋報告應與欲解釋的測驗結果相一致。因而，本文第二項詮釋概念乃提出以「廣義測驗建構流程」檢視運用之精熟標準設定方法，此乃有利於研究者提供接續探討之多元效度證據。

對於信、效度議題方面，多代表著某種精熟標準設定方法是否具備一定分類精熟／未精熟者水準的同義詞，因而，於各種研究中多伴隨此類議題的探討，如在方法比

較研究上，多採 Hambleton、Swaminathan、Algina 與 Coulson（1978）、Berk（1980）關於精熟分類決策的信度，如百分比一致性（percent agreement）、κ 係數（Kappa coefficient of agreement）（Cohen, 1960），或僅比較通過分數、判定試題機率值與實徵難度 P 值間相關或差異程度等。但就如同上述，方法比較結果多呈現不一致情況，因而，單純比較其百分比一致性或 κ 係數並非能充份佐證該方法具備良好特性或特質，有鑑於此，判定方法良窳取向乃漸漸強化效度重要性，強調著精熟標準設定結果是否具有其合理性、實務應用性等等，此即如同 Kane（1994, 1998）所指稱的三類效度證據，檢視的準則茲簡述如下：

（一）效度的過程證據（procedural evidence for validity）

過程的證據強調著精熟標準設定過程的適當性及其執行時各階段的品質，而判定的準則包含：

1、方法的選擇（selection of methods）：確立其理論基礎、是否具備執行簡易、具可靠度且易於解釋結果的實用性質（practicability）（Berk, 1986）；

2、過程的執行（implementation of procedures）：確立包含決策目標、能力標準定義的明確性（explicitness）與評審挑選、訓練、資料搜集過程等，是否具系統性與嚴謹性（Berk, 1986; Kane, 1994, 2001）；

3、評審回饋（feedback from judges）：評審對判定過程與決策結果的知覺、意見與滿意程度（Kane, 1994, 2001）；

4、心理計量的合法程序（psychometric due process）：強調精熟標準需關切目標合法性，是否充份告知受試者精熟標準用途與確立基本公平性存在（Cizek, 1993）；

5、設定結果發表（documentation）：精熟標準設定過程中各面向可供檢視與發表程度（Cizek, 1996）；

6、社會影響與財政支出：確立由此所作決策，可能對如教育、心理或其它層面可能造成的影響，及考量連動的財政支出（Millman, 1973）。

（二）效度的內部證據（internal evidence for validity）

有別於效度的過程證據將焦點集中於以「方法」為核心延伸的各階段執行品質，在此，效度的內部證據則將範圍縮小，僅強調方法內運用穩定與一致性，而據研究者方法運用概念的不同，提供內部證據亦有所差異，大致可分為三類：

1、方法內的一致性（consistency within method）：強調方法在不斷的重覆過程中，所得精熟標準估計的準確性程度，即使橫跨不同試題或內容，皆能提供適當的精熟／未精熟者分類訊息（Berk, 1986; Kane, 1994, 2001），同時在運用其它相似的方法時，

兩者能產生一致的結果（Kane, 1998）。

2、評審內的一致性（consistency within judge）：強調判定過程中，評審於各階段內與各階段間，本身評定結果的穩定程度（Berk, 1996; Kane, 1994），對照過去研究，如文獻探討初所述，多提供回饋資料、加強評審訓練以增加此方面成效，多期望評審內判定變異較小且與實徵資料（如試題難度 P 值）間具有高相關。

3、評審間的一致性（consistency between judges）：強調判定過程中，評審間評定結果的一致性，多期望其判定變異較小，以利於匯整結果（Berk, 1996; Kane, 1994, 2001）。

（三）效度的外部證據（external evidence for validity）

單就方法內的探討，無法將結果作有效推論，因而，效度的外部證據強調運用有關受試者能力或其它方法的效標資料，藉以連結與設定的精熟標準間相關，以提昇精熟／未精熟者分類的預測效果，而判定準則主要可分二大類：

1、方法間的一致性（consistency between method）：有別於方法內的一致性，乃強調在同一研究中，運用不同的精熟標準設定方法，期望能產生相似的結果（Kane, 1994, 2001）。

2、對照其它外部資訊（comparisons to other information）：方法間的比較僅能提供不同精熟標準設定方法間適當或不適當的結論，無法說明產生不同結果時，何者具有較佳效度，因而，對照其它外部資訊（如受試者於其它相似測驗表現、相關成就資訊、有無接受教學或其它群體受試者表現等），以提供有效效標資料，以確立精熟標準正確分類的外在推論、預測效果（Berk, 1986, 1996; Kane, 1994, 2001）。

由此觀之，提供精熟標準設定具有效度或稱之為良窳的證據，十分多元，不單只限於傳統判定精熟／未精熟者一致性信度證據，其它如強調精熟標準設定過程中各個面向特徵的過程證據，皆可作為提供有效度的代表，若將此搭配上述所提廣義測驗建構流程觀點，則如下圖 6 所示，強調著是否具完善核心目標、具體目標、由題庫組卷而成的測驗品質是否建全，精熟標準設定方法是否具理論性且能有效融入其中，並且具有清晰、易懂測驗解釋結果等等，皆是提供精熟標準設定方法良窳的佐證，此即本文所強調之第三項詮釋概念：「多元效度證據」。

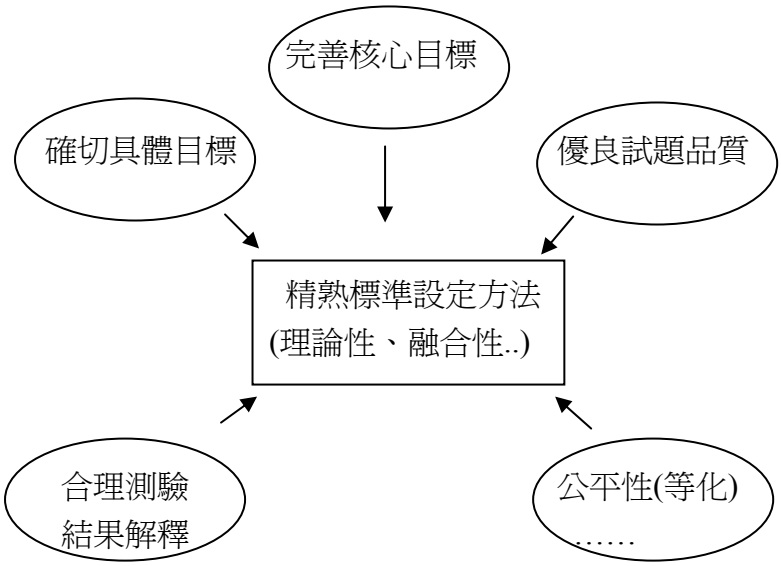

圖 6 本文所運用精熟標準設定方法之廣義流程概念圖（引自謝進昌，2005）

簡言之，本文乃期望從精熟標準設定方法與議題歷史的發展中，所萃取出的三個主要面向概念，「元素的搭配組合與調整」、「廣義測驗建構流程」、「多元效度」，重新詮釋精熟標準設定的方法，以期望能提供未來研究者於實務運用時不同的啟發，相關可能運用面向，茲於結論與建議中簡述之。

參、結論與建議

本研究旨在從精熟標準設定方法與議題的演進、發展中，衍生出三項詮釋的新概念，而此於實務上能促使精熟標準設定方法運用更具彈性，且能回應對多數受試者與測驗中心模式方法的評論，即是認為此類方法決定之精熟標準，多依靠受試者的表現（或評判者對於測驗試題答對機率的判斷）作為判定精熟標準的準則，並無考量到諸如：此精熟標準是否符合本測驗目的要求或者該標準設定後對於政治、財政面的影響等。有關上述論點，乃是對方法的觀點過於主觀與自限，若以本文強調之概念詮釋時，則可獲得不同思考面向，以下茲舉範例說明：

在初次設定某資格檢定測驗之精熟標準時，於測驗前，往往未能確切知道所欲設

定的通過分數為何？因此，多半需藉由首次測驗資料尋求出可供參照的標準，在傳統觀念下，測驗資料本身僅能提供受試者的反應概況或測驗試題（視研究者所採為何種模式之方法）的訊息以決定精熟標準，若採以本文強調之詮釋概念時，則能將此資料所透露之訊息作彈性運用。在圖 7 流程中，若以「廣義測驗建構流程」觀點檢視時，可發現精熟標準設定的開端從測驗編製初即已融入，在核心目標主導下，決策者需針對人才甄選取向、最低能力要求等達成共識，以作為接續篩選人才的概念，進而，在確切的測驗內容範圍內，訂立具體的能力指標，同時具體化該測驗適合的能力範圍（如多少百比分的考生能正確作答），並以此挑選具內容代表性試題，組卷施測（而題庫的建立可根據歷次的測驗逐步累積），此時，組成的測驗試題即隱含著決策者目的與要求（如，目的上欲甄選優秀人才，則能力範圍即會限定於高能力區域，於編製與挑選試題，則會相對編取適合該能力區者），施測後，藉由研究者所欲採行之精熟標準設定方法求得暫時的精熟標準，此即可代表決策者經由測驗表達對精熟標準的要求。而後再以「元素的搭配組合與調整」檢視時，可進一步搭配另一項元素：專家判定，考量此精熟標準對社會、經濟可能造成的影響等因素作調整，以決定出最適切之精熟標準，接續，再進行測驗結果的解釋與報告撰寫。

綜整之，即是強調以廣義測驗建構流程觀點來檢視「專家+精熟標準設定方法+專家觀點」等元素搭配組合，以決定適切之精熟標準，對照上述疑義而言，此流程範例不但考量到精熟標準設定方法對於社會、經濟的影響層面，且能有效融合該精熟標準與決策者之核心理念，同時，流程中各面向，皆是作為提供該精熟標準設定方法具備有效性或可信性之多元效度證據。

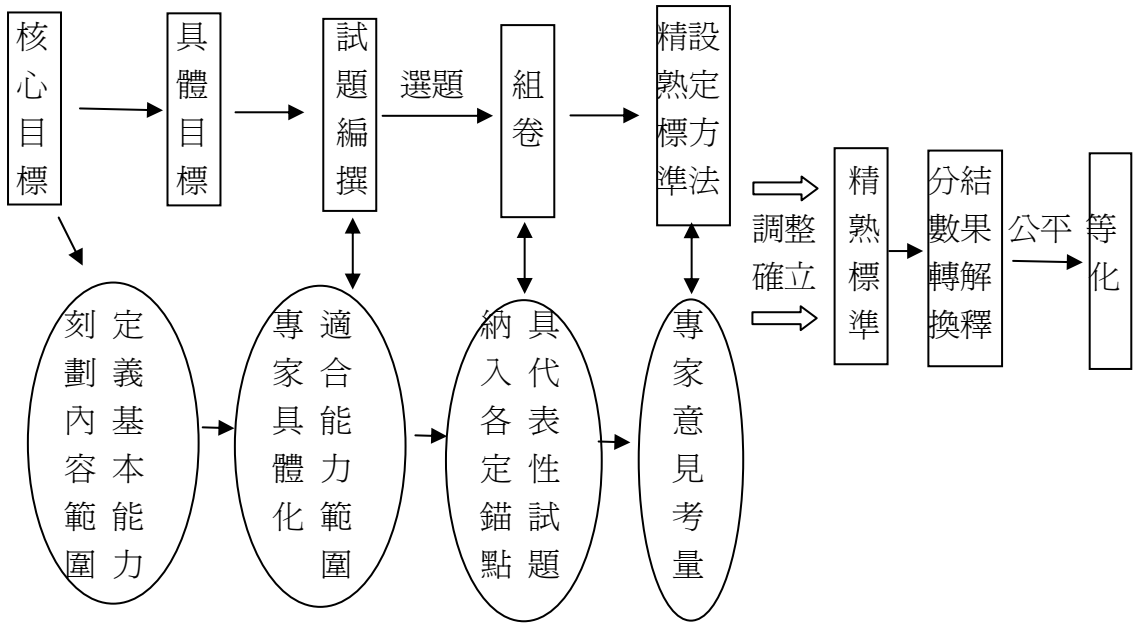

圖 7 以三項詮釋概念為核心建構之精熟標準設定模式範例

在過去研究中，對於精熟標準設定方法詮釋的觀念多過於自限（僅特定於方法本身），並未能以測驗建構觀點視之，考慮到與上（測驗編製目的、試題來源）及下（測驗解釋結果）的融合，致使運用上多所限制。因而，本文建議研究者於運用精熟標準設定方法時，可將其視為一種「元素的搭配組合與調整」，乃自行根據測驗目的、實務考量等因素，搭配各類元素，建立一套符合決策者目的要求之精熟標準設定流程。同時，搭配廣義測驗建構流程觀點，提供各面向的多元效度證據，以佐證該方法所得之精熟標準的可信度，以建立資格檢定考試本身的權威。此外，有鑑於篇幅限制，本文無法對實徵運用與研究作詳細佐證，有興趣研究者可參考謝進昌、余民寧（2005）作品，該文則詳列如何運用此新概念，以詮釋最大測驗訊息量（maximum test information）於精熟標準設定的合理性與適切性及其實證研究結果。

參考文獻

中文部份

考選部（2005）。考選部全球資訊網。2005 年 7 月 8 日，取自 *http://wwwc.moex.gov.tw/lp.asp?CtNode=970&CtUnit=31&BaseDSD=7*。

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。

吳裕益（1986）。*標準參照測驗通過分數設定方法之研究*。未出版之博士論文，國立政治大學教育研究所，台北。

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。

鄭清泉（2001）。*人工化與電腦化適性精熟能力判定在國小學童數學精熟分類一致性之比較研究*。未出版碩士論文，國立嘉義大學國民教育研究所，嘉義。

謝進昌（2005）。*以最大測驗訊息量決定通過分數之研究*。未出版之碩士論文，國立政治大學教育研究所，台北。

謝進昌、余民寧（2005）。以最大測驗訊息量決定通過分數之研究，*測驗學刊，52*（2），149-176。

外文部份

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*, 35-50.

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.

Berk, R.A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45*, 4-9.

Berk, R. A. (1980). A consumers' guide to criterion-referenced test reliability. reliability. *Journal of Educational Measurement, 17*(4), 323-349.

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

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

Bernknopf, S., Curry, A., & Bashaw, W. L.(1979). *A defensible model for determining a minimal cutoff score for criterion referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Behuniak, P., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implication for the validity of proficiency test score interpretation, *Educational and Psychological Measurement, 42*, 247-255.

Bontempo, B. D., Marks, C. M., & Karabatsos, G. (1998). *A meta-analytic assessment of empirical differences in standard setting procedures*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Brennan, R. L. & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory, *Applied Psychological Measurement, 4*, 219-240.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development, 35*(3), 167-181.

Brandon, P. R. (2004). Conclusion about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education, 17*(1), 59-88.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the national teacher examinations. *Journal of Educational Measurement, 27*(2), 145-163.

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

Cascio, W. F., Alexander, R.A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41*, 1-24.

Chang, L., Dziuban, C. D., & Hynes, M. C.(1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education, 9*(2), 161-173.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education, 12*(2), 151-165.

Chang, L., van der Linder, W. J., & Vos, H. J.(2004). Setting standards and detecting intrajudge inconsistency using interdependent evaluation of response alternatives. *Educational and Psychological Measurement, 64*(5), 781-801.

Chinn, R. N., & Hertz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education, 15*(1), 1-14.

Cizek, G. J. (1993). Reconsidering standard and criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 13-21.

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.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Conover, J. W., & Iman, R. L. (1978). *The rank transformation as a method of discrimination with some examples*, Albquerque, NM: Sandia Laboratories.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teachers Examinations, *Journal of Educational Measurement, 21*, 113-129.

Cross, L. H., Frary, R. B., Kelly, P. P., Small, R. C., & Impara, J. C. (1985). Establishing minimum standards for essays: Blind versus informed review. *Journal of Educational Measurement, 22*, 137-146.

Dillon, G. F. (1996). The expectations of standard setting judges. *CLEAR Exam Review, 2*, 22-26.

Ebel, R. L. (1972). *Essentials of educational measurement* (2rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Educational Testing Service (1976). *Report on a study of the use of the National Teachers' Examination by the state of South Carolina*. Princeton, NJ: Author.

Fehrman, M. L., Woehr, D. J., & Arthur, W. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and Psychological*

Measurement, 51(4), 857-872.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Edgenics, 7*, 179-188.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187-213.

Gessaman, M. P., & Gessaman, P. H. (1972). A comparison of some multivariate discrimination procedures. *Journal of the American Statistical Association, 67*, 468-472.

Giraud, G., Impara, J. C., & Buckendahl, C. (2000). Making the cut in school districts: Alternative methods for setting cut-scores. *Educational Assessment, 6*, 291-304.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice, 22*(1), 22-32.

Halpin, G., Sigmon, G., & Halpin, G.(1983). Minimum competency standards set by three divergent groups of raters using three judgemental procedures: Implication for validity, *Educational and Psychological Measurement* , *43*, 185-196.

Hambleton, R. N., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*, 1-47.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41-55.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N.(in press). *Handbook for setting performance standards*. Washington, DC: Council of Chief State School Officers.

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. *Educational and Psychological Measurement, 41*(3), 725-734.

Hudson, J. P. Jr., & Campion, J. E. (1994). Hindsight bias in an application of the Angoff method for setting cutoff scores. *Journal of Applied Psychology, 79*(6), 860-865.

Hurtz, G. M., & Hurtz, N. R.(1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and*

Psychological Measurement, 59(6), 885-897.

Hurtz, M. G., & Auerbach, M. A. (2003). A meta analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*(4), 584-601.

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

Jaeger, R. M. (1978). *A proposal for setting a standard on the North Carolina High School Competency Test*. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-476.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn(Eds.), *Educational Measurement* (3rd ed., pp.485-514). New York: Macmillan.

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

Jaeger, R. M., & Mills, C. N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.). *Standard setting: Concepts, methods, and perspectives* (pp.313-338). Mahwah, NJ: Erlbaum.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1994). *Using actual student work to determine cut-scores for proficiency levels: New methods for new tests.* Paper presented at the National Conference on Large-Scale Assessment, Albuquerque, NM.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1995). *Setting standards for performance levels using the student-based constructed response method.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Kane, M. (1998). Choosing between examinee-centered and test-centered standard setting methods. *Educational Assessment, 5*(3), 129-145.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.). *Standard setting: Concepts, methods, and perspectives* (pp.53-88). Mahwah, NJ: Erlbaum.

Koffler , S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement, 17*, 167-178.

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.

Livingston, S. A., & Zieky, M. J.(1989). A comparison study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121-141.

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.

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.

McGinty, D., & Neel, J. H. (1996). *Judgmental standard setting using a cognitive components model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. *Review of Educational Research, 46*, 133-158.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research, 43*, 205-216.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement, 20*(3), 283-292.

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.

Nassif, P. M. (1978). *Standard setting for criterion referenced teacher licensing tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25*, 57-65.

Norcini, J., Lipner, R., Langdon, L., & Strecker, C. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24*, 56-64.

Norcini, J., Shea, J. A., & Grosso, L. (1991). The effect of numbers of experts and common items on cutting score equivalents based on expert judgment. *Applied Psychological Measurement, 15*(3), 241-246.

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

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., Impara, J. C., & Potenza, M. T.. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement, 31*(4), 339-347.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57*(3), 400-411.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.). *Standard setting: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 57-83.

Reilly, R. R., Zink, D. L., & Israelski, E. W. (1984). Comparison of direct and indirect methods for setting minimum passing scores. *Applied Psychological Measurement, 8*,

421-429.

Saunders, J. C, & Mappus, L. L. (1984). *Accuracy and consistency of expert judges in setting passing scores on criterion-referenced tests: The South Carolina experience.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Schoon, C. G., Gullion, C. M., & Ferrara, P. (1979). Bayesian statistics, credentialing examinations, and the determination of passing points. *Evaluation and the Health Professions, 2*, 181-201.

Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education, 12*(3), 301-325.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. *Journal of Educational Measurement, 17*, 229-235.

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.

Stephenson, A. S., Elmore, P. B., & Evans, J. A. (2000). Standard-setting techniques: An application for counseling programs. *Measurement and Evaluation in Counseling and Development, 32*(4), 229-244.

van der Linden, W. J. (1982). A latent trait method for determining intra-judge inconsistency in the Angoff and Nedelsky techniques of standard setting , *Journal of Educational Measurement, 19*, 25-308.

Wang, N. (2003). Use of the Rasch model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*(3), 231-253.

Webb, M. W. I., & Miller, E. R. (1995). *A comparison of the paper selection method and the contrasting groups method for setting standards on constructed- response items*. U.S.; Pennsylvania: December 31, 2004, from ERIC database.

Woehr, D. J., Arthur, W., & Fehrman, M. L. (1991). An empirical comparison of cutoff score methods for content-related and criterion-related validity settings. *Educational and Psychological Measurement, 51*(4), 1029-1039.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.

文稿收件：2005 年 07 月 15 日
文稿修改：2006 年 01 月 02 日
接受刊登：2006 年 02 月 15 日

The Historical Movement of Standard Setting and New Concept of Interpretation

Jin-Chang Shieh

Graduate student, Department of Education
National Chengchi University

Abstract

The purpose of this paper is to introduce the evolvement and development of standard setting in recent ten years and we are trying to discover efficient information to lead to three new interpretation facets through it. They are respectively component combination and adjustment, generalized test construction processes and multiple validities. Depending on these three concepts, we hope to provide researchers different inspiration and thoughts in reviewing the methods of standard setting for future usage.

Key words: standard setting, component combination and adjustment, generalized test construction processes, multiple validities.