第二章 文獻探討

文獻探討部份，共分為四節，第一節係從效標參照測驗的起源與定義為開端，陳述精熟標準方法與研究議題的演進，進而引導出詮釋本研究精熟標準設定的三項主要核心概念；第二節則說明本研究所運用之最大測驗訊息量法，闡述其於精熟標準設定上之學理基礎；第三節則就換算古典測驗分數法與測驗特徵曲線構圖法於轉換 IRT-θ 能力值與古典測驗答對題數的概念及運用定錨點於解釋精熟/未精熟者差異能力進行討論；第四節則主要是將焦點置於探討上述方法與測驗長度、測驗異質性議題之關聯。

第一節 精熟標準設定

壹、效標參照測驗的起源與定義

測驗、評量的起源來自於考試制度，原本興起於中國，但加以量化、採科學方式進行探討則肇始於國外，其目的在探究人的能力，以藉此拔擢人才。歷史發展中，對於測驗以不同角度的觀點加以分類的，實不在少數，其中，常模參照(norm-referenced)與效標參照(criterion-referenced)的評量模式即是一例。早自 Thorndike(1918)(引自吳裕益(1986))即曾對此概念有明確區分：認為常模參照測驗是奠基在心理學個別差異(individual difference)概念上，而效標參照測驗則是以自然科學模式為依據。長久以來，教育評量方式為常模參照模式所把持，以致競爭、排等第的觀念深植於心，但強調回歸個體本質能力的聲音亦不斷湧現，此概念即是效標參照的理念。

Glaser 於 1963 年在美國教育研究協會(American Educational Research Association)上，就曾針對效標參照測驗提出詳盡說明，他認為效標參照測驗的概念是「從未熟練到完美能力(perfect performance)的範圍間一種連續知識的獲得，而個體的成就水平乃落於藉由測驗中所展現出的行為來表達此連續知識中的某一點…(Glaser, 1963, p.519)」之後，對此加以延伸、重新詮釋的定義方式十分多元。Glaser & Nitko (1971, p.653)認為「效標參照測驗是一個經過仔細編製，並依特定能力標準來直接解釋結果的測驗，而此能力標準(performance standards)通常是由個體必須學會的學習行為領域(domain of tasks)來加以界定，並由代表此領域抽樣而得的行為所組成的測驗」;Popham & Husek(1969, p.2)則由精熟/未精熟的角度認為「效標參照評量是用以確立個體相對於某些效標(criterion)(如，能力標準)中的地位，因而個體是與某些已建立的效標作比較，而非與其它個體」; Ivens(1970, p.2)則以較廣義觀點認為「效標參照測驗是由一組行為目標的許多試題所構成的測驗」。此外，由此延伸出的同義詞彙，如領域參照測驗(domain-referenced tests)、目標導向測驗(objectives-referenced tests)、能力本位測驗(competency-based tests)、精熟測驗(mastery tests)都顯示出不同專家、學者對此議題重視與見解，Gray 於 1978

年時，就曾對此概念加以分類，並確認出有多達 57 種不同定義存在。雖然存在如此眾說紛紜的定義，但相對於目前較能被一致接受的是：「效標參照測驗乃指被用來確定個人在某個界定清楚的行為領域(well-defined behavioral domain)中表現程度的測驗」(Popham, 1978, p.93)。

雖然效標參照測驗定義十分多元，但仍可從中萃取出幾項主要組成元素，分別為界定清楚的行為領域(well-defined behavioral domain)、目標(objectives)、能力標準(performance standards)、能力解釋(performance interpretation)等四項，若加以組合，即可表達成在界定清楚的行為領域中，根據學習或行為目標訂定能力標準，以檢定個人是否能達到預設的標準，並用此作為表現程度解釋的測驗。

在檢視過效標參照測驗內涵後，搭配 Hambleton, & Zaal(1991, pp.10-11)、Hambleton(1990)對於效標參照測驗編製流程建議與 Hambleton(2001)、Kane(1994)對於「標準」的定義，可將上述融合為如圖 2-1 所示，以能力標準設定(performance standard setting)為核心的測驗建構流程，簡述如下：

一、設定核心目標

此階段乃屬於政策考量(policy decision)，測驗主辦者需會同專家、學者或相關人員，根據測驗目的，考量如經費限制、受試者特性、命題委員遴選、測驗格式等因素，以設定該次測驗或評量的中心目標。此時，亦初步決定該次測驗對於最低能力表現者的要求水平(是屬於高水平能力亦或是中等水平能力)。

二、確立測驗內容與具體目標

根據測驗目的，確定評量的內容範圍，並將概念式的核心目標轉換為具體測驗指標，而相對於「標準」定義，此階段正如 Hambleton(2001)所指稱，全部受試者在課程中被期望應該知道與能達到的內容標準(content standard)，例如，能執行基本數學運算、閱讀一篇文章等(p.91)。

三、題庫建立

施測者根據具體目標，將其轉換為測驗試題以檢測受試者能力，此階段乃依循標準化測驗的編製程序，從雙向細目表的擬定、試題的編製與修訂、預試、試題分析以及信、效度的評估等，以建立龐大測驗題庫(item bank)。

四、判定能力標準以選取試題正式施測

根據內容標準，施測者需進一步判定精熟/未精熟者間所擁有的差異能力標準(performance standard)，以描繪兩者間的成就水平(achievement level descriptions)。而此正如 Kane(1994)認為能力標準是介於可接受與不可接受的成就水平間的概念性界限(conceptual boundary)(p.433)，而定義之能力標準乃屬於決策領域(domain of policymakers)(Kane, 2001)，即代表著決策者期望精熟者應具有的最低能力水平，例如，在數學乘法運算中，決策者認為精熟者至少

需懂得二位數乘法運算，若受試者僅能正確作答一位數乘法運算，則將之視為未精熟者。據此，再選擇適合的試題(同時包含一位數與二位數乘法試題)，以正式進行施測。

五、決定通過分數標準

此階段則根據測驗結果，選擇欲採行的能力標準設定方法，並綜合政策考量因素，以決定測驗的通過分數(passing score)，此正如 Kane(1994)認為在量尺分數上所設定的某一特定分數點，並用以對受試者作決策(p.433)。相較於上述屬於質性描述的能力標準，通過分數則是將之轉換為實際的分數(如 60 分)，將測驗得分高於此分數者，即視為精熟者，反之則視為未精熟者。

六、解釋測驗結果

本階段首重測驗結果的解釋，參照試題編製的具體目標，提供精熟/未精熟者能力差異標準，同時解說整個決策過程、能力標準設定方法、試題編製、測驗信、效度等證據，以取信於社會大眾，並建立評量的權威形象。

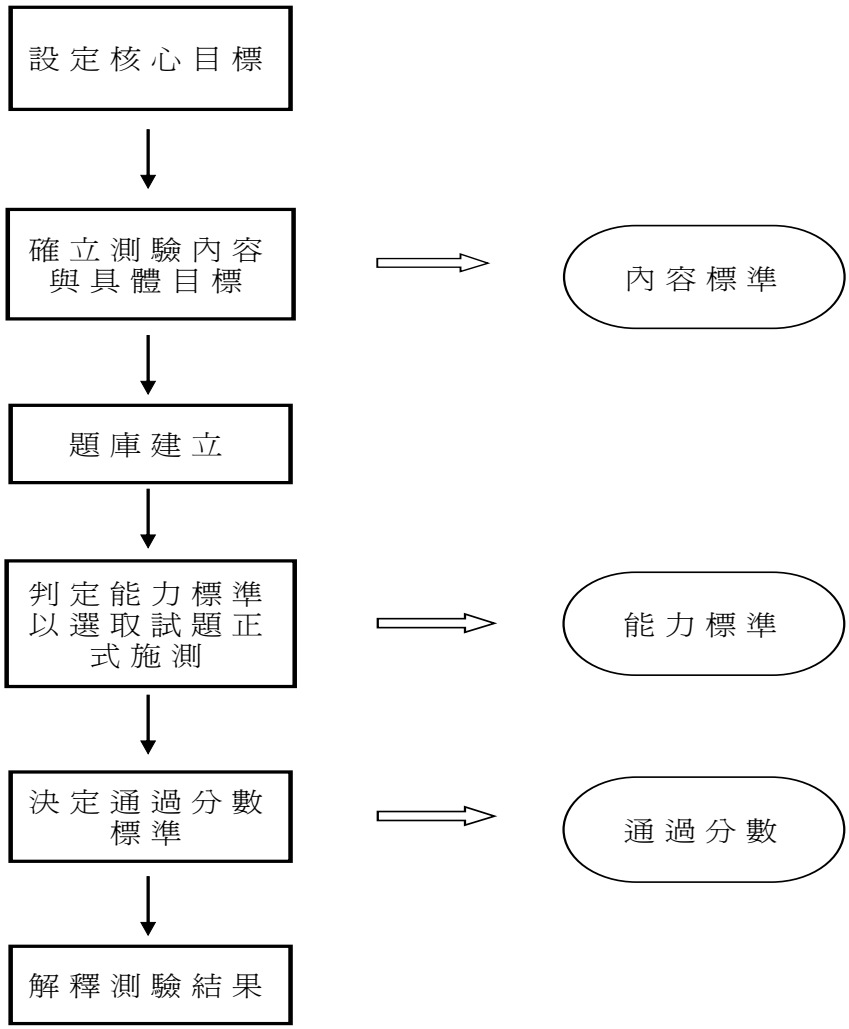

圖 2-1 以能力標準設定為核心的測驗建構流程圖

此測驗建構流程觀點即是描繪本研究精熟標準設定方法的首要核心概念(後續將有更詳細原因解說)，接續，再從設定方法的創新與議題演進中粹取出另兩項詮釋概念，茲陳述如下。

貳、精熟標準設定

有藉於本文研究基礎乃建立在試題反應理論上，因此，乃援用 Lord(1980)的說法，以「精熟測驗」一詞顯現於此方面應用，以區別古典測驗理論所稱「效標參照測驗」的不同。同時，本文所探討標準設定方法主要用以區分精熟/未精熟者，不加以延伸至多層次(如基本、精熟、進階)，因而多援用「精熟標準」一詞以表示由 IRT 相關方法求得之標準，而藉由轉換分數方法求得之古典測驗答對題數則稱之為「通過分數」。

自一九七〇年代以來，精熟測驗的相關議題即不斷被提出探討。茲將其探討的七大議題要點，整理歸納如下：

一、信度：關注於精熟/未精熟者正確分類決定的信度、測驗分數的信度、真實分數估計的信度等；

二、效度：探討精熟測驗的內容效度(content validity)、建構效度(construct validity)、效標關聯效度(criterion-related validity)相關概念，如：試題命題技巧及代表性問題、所作精熟/未精熟決策的有效性等；

三、試題選題方式：比較不同試題選題方式，如：隨機選題法、一致法(agreement approach)、Phi 係數法、IRT 法等的特點，並探討何者方法具較高精熟/未精熟者分類一致性等；

四、測驗長度：探討測驗長度與正確分類精熟者/未精熟者、試題選題方式、精熟標準設定等變項間的關聯性等；

五、精熟標準設定：探討不同精熟標準設定法間的特性及一致性分類表現等；

六、測驗計分與結果解釋方式：探討如何能有效改善測驗分數的解釋，與使用簡易、清晰的分數報告格式等；

七、電腦化適性精熟測驗(computerized adaptive mastery testing)：探討電腦精熟判定的準確性、試題選擇方式、終止條件及相關題庫建立與系統的研發等。

在上述精熟測驗各項議題中，關鍵的一環，即在於精熟標準設定(standard setting)。不論是教育或是證照發放用途，我們所關心的是哪一個精熟標準，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。在此領域上，已有許多專家、學者先後投入相關的研究，Berk(1986)確認過有 38 種方式被發展出；鄭明長、余民寧(1994)認為有超過 40 種方法，至今 2005 年，陸續也有許多精進方式被研發。面對如此龐雜的方法，分類上亦各家雜陳，Meskauskas(1976)曾將通過標準分為：一、狀態模式(state models)：即將受試者能力視為「全有」或「全無」兩種情形，精熟者被視為擁有測驗所要求的基本

能力，而未精熟者則無，相關方法可參考 Emrick(1971)、Roudabush(1974)等人作品；二、連續模式(continuum models)：即將受試者能力視為連續累進的情形，認為受試者在一連續能力分布量尺上，若通過某設定標準，即視為精熟，反之，則視為未精熟。

Hambleton(1980)、Hambleton & Eignor(1980)、Jaeger(1980)則進一步將連續模式分為三類：一、判定模式(judgmental models)：針對測驗的內容如試題難度，以此作為判定精熟標準的依據；二、實徵模式(empirical models)：依據受試者實際作答表現來設定精熟標準；三、組合模式(combination models)：融合上述兩者，以判定受試者能力為核心，並參照受試者於此測驗內容上的表現，以決定精熟標準。但不論何種模式，都牽涉到判斷或判定(judgment)，如判斷測驗試題、受試者能力等，差別只在於判斷的主要核心焦點、過程與程度的不同。因此，Jaeger(1989)、Kane(1994)、鄭明長、余民寧(1994)企圖將三類加以整合，只分為測驗中心模式(test-centered model)(以測驗試題內容為判斷依據)與受試者中心模式(examinee-centered model)(以受試者能力或實際表現為判斷依據)兩類。然而，當精熟標準設定方法不斷擴增(如延伸至實作評量或多元計分試題時)，如此的分類機制則易自限於描述不同的方法，因而 Hambleton, Jaeger, Plake, & Mills(in press, 引自 Pitoniak(2003))刻畫六個著重於「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(評審個別或團體判定、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(內在證據、外在證據等)；六、評量的本質(試題類型：選擇題或建構反應試題、計分方式等)。

同時，隨著電腦科技的精進，使得電腦化精熟判定方式又成為另一探討主題，主要分為二個領域來進行：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)。

而在「標準設定」外，另有一類是屬於 Berk(1986)所稱「調整通過分數」的方法，如 Millman(1972)、Kriewall(1972)、Novick & Lewis(1974)、Wilcox(1979)等人提出有關二項式模式(binomial based model)、貝氏理論(Bayesian method)與決策理論(decision method)的方法，則是在真實分數量尺上的通過分數之標準已設定後，再選擇一個適當的觀察分數量尺上的通過分數，以期能將分類錯誤降到最低。但 Berk(1986)將其視為只是前者的輔助，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，只偏重於依分類錯誤的損失來調整求適當通過分數，作法適切性頗值得懷疑。

在參照上述各學者、專家的分類後，本文作者企圖對精熟標準設定方法提出

一較完整架構，如圖 2-2 所示，以利讀者釐清彼此關係。

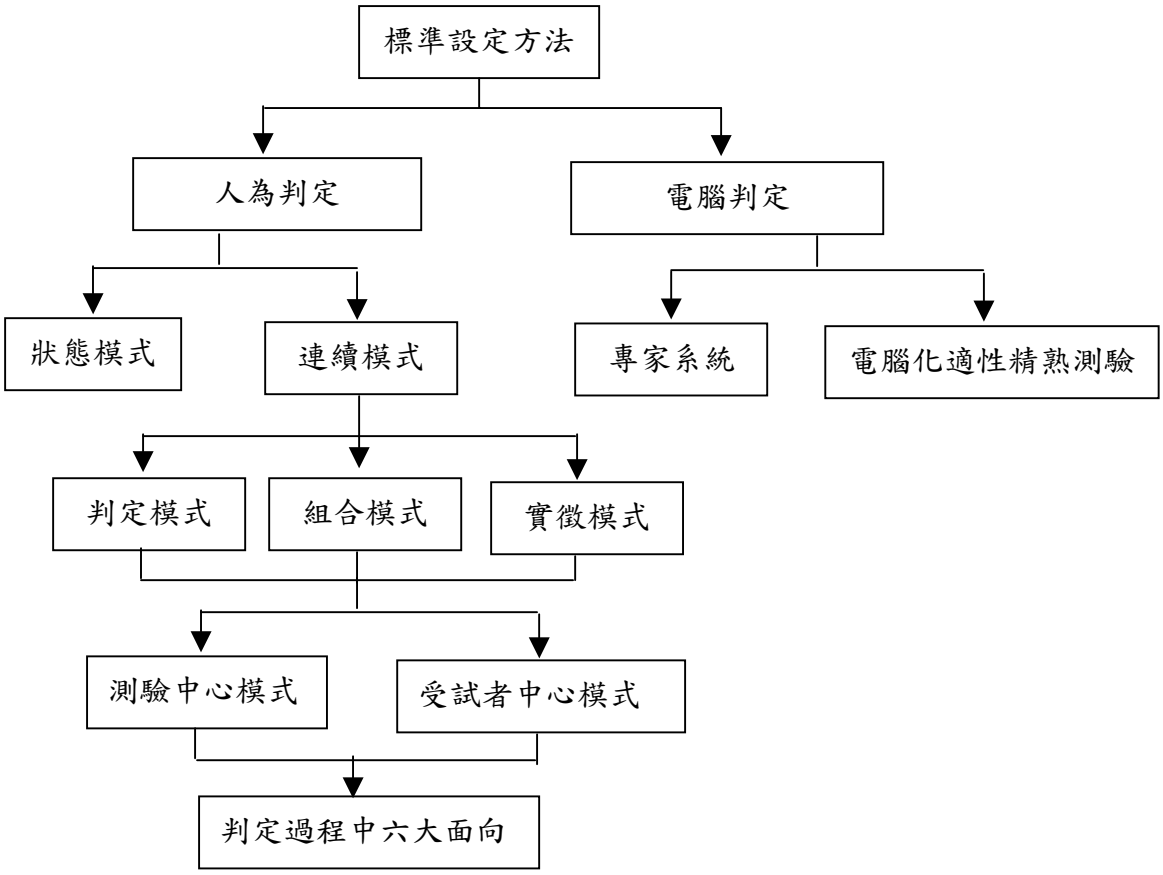

圖 2-2　精熟標準設定方法分類圖

面對如此龐雜方法，本研究欲藉由精熟標準設定方法與其相關探討議題於近幾十年演進過程中，從中萃取可能的訊息，以形成詮釋與支持本研究所運用方法的證據(其實，於上述方法分類的演變上，即可發現晚近研究漸漸重視所謂「判定過程」的概念」)。在此，本文於方法介紹上，為解說與理解的方便性，仍採用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，主要以 Nedelsky(1954) 、Angoff(1971)、Ebel(1972)這三種方法運用最廣、研究討論也最多，至今各種延伸的方法，概念上亦是建基於此，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 2-1 所示之 Nedelsky 法之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找出那些「最低能力表現」(minimally competency)學生或 Nedelsky 所稱 F-D 學生(代表介於 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之機率值加總，再進一步求所有評審平均值，以此作為精熟標準。其中，Nedelsky 乃假定評審是具有能力區辨哪些選項對最低能力表現學生而言，是具有吸引力的。此外，更認為學生對於不會作答的題目，會先挑出被認為絕對錯誤的選項，而後就剩下的選項中，採用隨機猜題方式作答。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，代表測驗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 2-1 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	
9	(0) 1 2 3	.25	SUM
10	0 (1) 2 3	.33	5.07

註：取自 Zieky & Livingston(1977, p.5)

二、Angoff 家族

(一)Angoff 法(Angoff' s method)

Angoff 乃於 1971 年在其文章上概略提出相關理念，乃要求評審針對一群最低能力表現者，判斷其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 2-2 之 Angoff 法之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的機率值，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。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 2-2 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)

(二)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)，為了有效凝聚評定結果，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。

(三)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)。

(四)反覆二選 Angoff 法(iterative two choice Angoff)又稱 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:每位畢業學生是否都能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，其它評審所建議之標準的分配、評審本身先前的評定結果、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，並藉由實際問題具體化對於最低能力表現的界定。而於第三輪中，所有評審判定的最小中位數值(minimum median standard)，即為精熟標準。

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

Bernknopf, Curry, & Bashaw(1979)乃針對改良式選擇型 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，然後，以此求得各試題平均百分率之標準誤來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，所得到的試題機率值再依據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。

(六)反覆 Angoff 法(iterative Angoff)

由 Saunders & 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。

(七)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2-3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2-3 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。

(八)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 & 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。

(九)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 & 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)。

(十)認知元素法(cognitive component approach)

相較於 Angoff 法於整個試題作判定過程,McGinty & 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成份的相關描述,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成份的機率值,而後,將試題中所有成份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。

(十一)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, & 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 2-3 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, & Patz(1998)又將其稱為修正的 IRT-Angoff 法(IRT-Modified Angoff Procedure)。

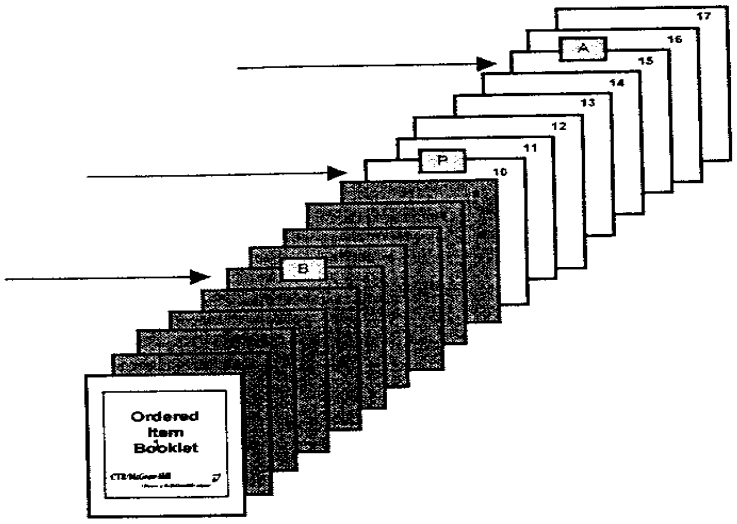

圖 2-3 書簽技術中排序的試題卷(取自 Mitzel, Lewis, Patz, & Green(2001, p.256)

(十二)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 2-4 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。

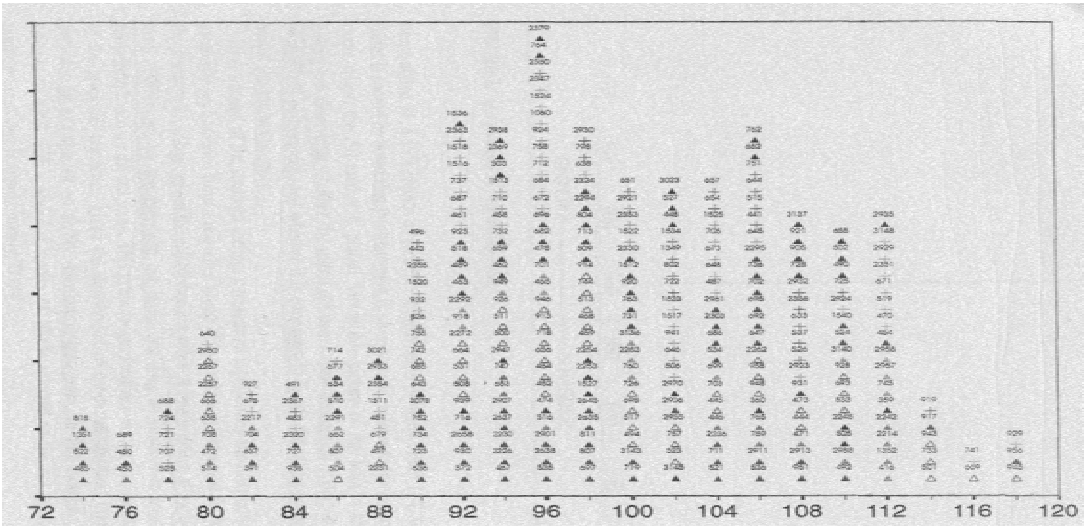

圖 2-4 試題構圖法之直方圖範例(取自 Wang(2003, p.234))

(十三)IDEA 法(interdependent evaluation of alternatives method)

Angoff 法在概念上乃著重於試題本身的判定，而 Nedelsky 法則著重於試題的選項，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, & 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

(一)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(questionable 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 2-4 乃 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 2-4 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)

(二)難度-目標分類 Ebel 法(difficulty-taxonomy Ebel)

Skakun & 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。

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

在概念上，Skakun & 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，傳統上，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，且概念亦漸漸難於測驗中心模式區分，而為解說方便，在此仍依其主要概念加以分類，茲介紹如下：

一、臨界組法(the borderline group method)

Zieky & 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)(乃因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)

(一)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 2-5 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，強調著執行測驗的受試者(people take the examination)，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 & 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。

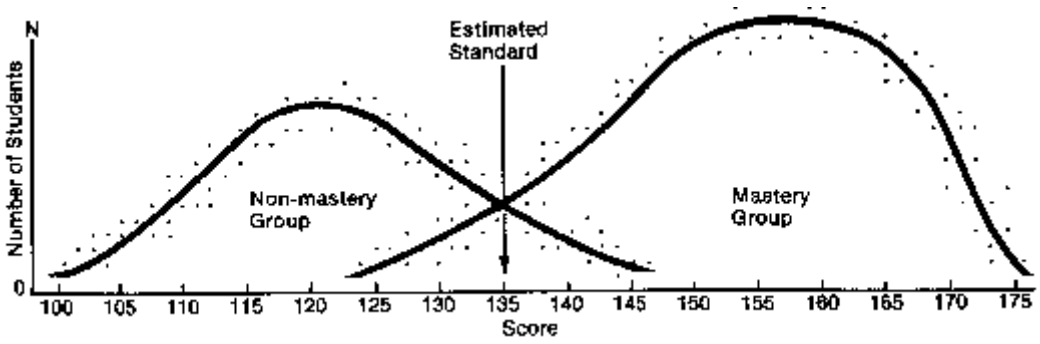

圖 2-5 對照組圖形法之範例圖(取自 Zieky & Livingston(1977, p.9))

(二)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：

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

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 2-1 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。

(三)對照組 QDF 法

相較於對照組 LDF 法，其公式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，其公式為：

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

(公式 2-2)

其符號意函與公式 2-1 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 2-2 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。

(四)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 & 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 2-1，以求得精熟標準。

(五)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：

$$M_1 - \left[S_1(M_1 - M_2)/(S_1 + S_2)\right] \qquad \text{（公式 2-3）}$$

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 & 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, & 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, & 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, Robin, & Patelis(1999)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(適用於一個集群解)或對照組法概念(適用於兩個集群解)決定精熟標準。

六、課程參與法(course enrollment method)

Giraud, Impara, & Buckendahl(2000)提出一項牽涉到課程安置(course placement)的精熟標準設定方法，事前，乃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, Impara, & Buckendahl(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(1996)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，但不同專家、學者所提方法在概念上皆十分雷同，乃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，可將其分為二個面向，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 & 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 & 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, & Jaeger(1995)、Plake, Hambleton, 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, Pence, & Jaeger (1995)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、連結(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、轉折(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，而此概念即形成本研究第二項詮釋核心。對照上述架構圖，多元方法的演變就如同 Hambleton, Jaeger, Plake, & Mills(in press, 引自 Pitoniak(2003))將分類面向由單純測驗與受試者中心模式擴展至判定過程中各要素的分類，已漸漸強調所謂精熟標準設定的「過程」。

精熟標準設定方法接續的提出，圍繞此所探討的相關議題亦是不斷湧現，討論的面向大致可分為三類：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。首先，在方法的比較研究上，Berk(1986)、Jaeger(1989)、Bontempo, Marks, & Karabatsos (1998)皆曾綜整過去此類比較研究的文章(相關研究簡要整理如表 2-5)，其結果就如多數研究者所發現的，不同的方法所產生結果多不一致，理由上，Hambleton(1978)認為對於評審的指導不同，過程的不同，自然不會產生類似結果；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；Berk(1986)認為可能是各種方法及評審對於精熟有不同的概念，以及對學習目標或測驗內容的行為領域有不同的解釋；Jaeger(1989)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 2-5 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、迴歸法、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

資料來源：作者整理。

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。

其次，對於精熟標準設定過程的探討，Hurtz & 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(含相關調整法)於設定過程中的相關議題，綜整之相關研究可參考表 2-6，所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 2-6 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, Small, & Impara(1985)	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

資料來源：作者整理。

對於信度與效度議題方面，多相對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，如在方法的比較研究上，多採用 Hambleton, Swaminathan, Algina, & Coulson (1978)、Berk(1980)有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)所指稱的外部效度證據(external validity evidence)與內部與過程效度證據(internal and

procedural validity evidence)(詳細描述於效度議題上會有更詳盡解說)，強調著提供多元效度的證據，而此即為本研究第三項詮釋概念。

綜合上述，可發現歷史的演進中，方法的創新乃強調著元素的組合搭配與調整，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 2-6 詳細描述之，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，相對於上述本章初曾提出之以能力標準設定為核心的測驗建構流程，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。

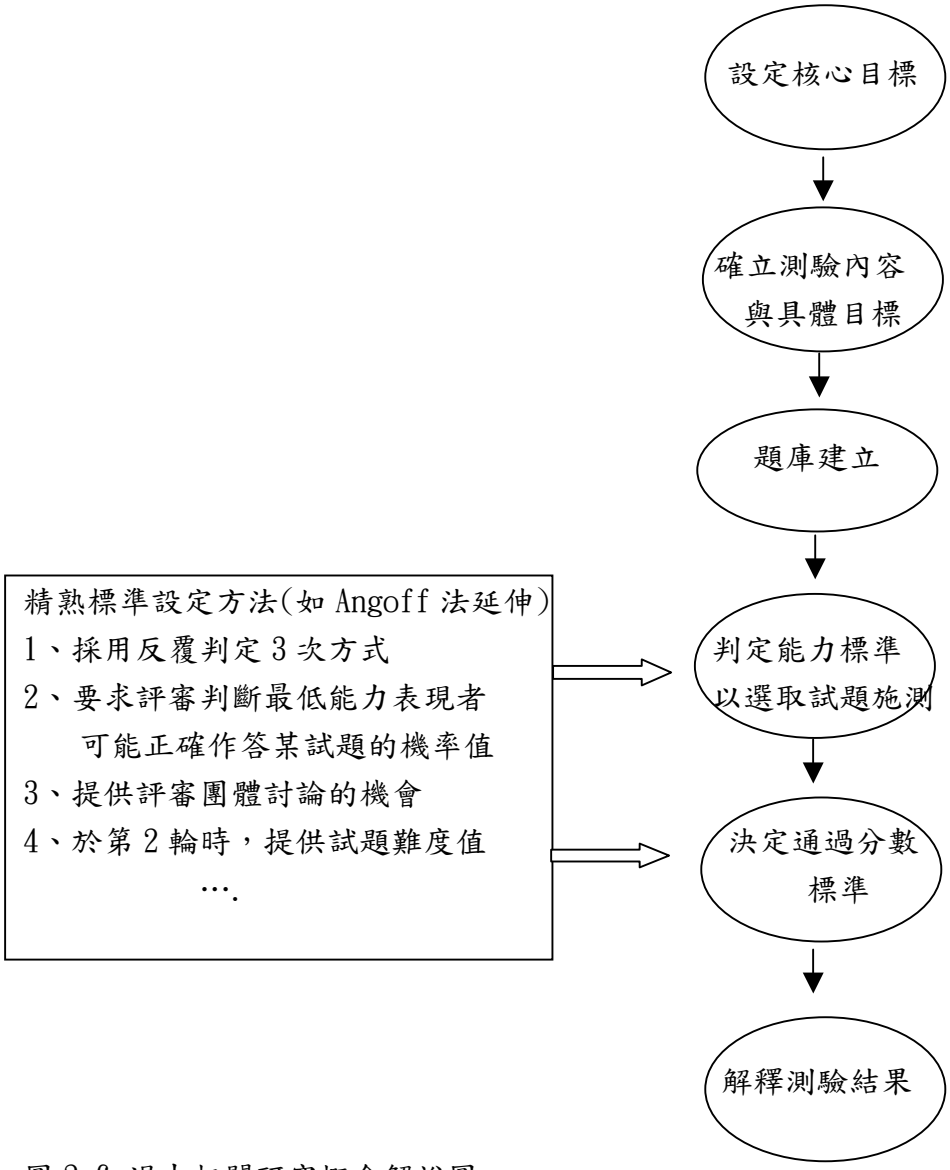

圖 2-6 過去相關研究概念解說圖

有藉於此，本研究乃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的，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)曾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，本研究乃採以圖 2-7 廣義測驗建構流程概念檢視本研究所運用之精熟標準方法：最大測驗訊息量法，除探討本身理論性外，更輔以文獻探討方式檢驗其運用工具是否具完善核心目標、具體目標、優良試題等，再提出合理方法以解釋測驗結果，企圖提供多元的效度證據。

簡言之，本研究即是期望從上述歷史的演進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(註：此概念將有助於後續解釋最大測驗訊息量法與精熟標準設定融合的合理性)。

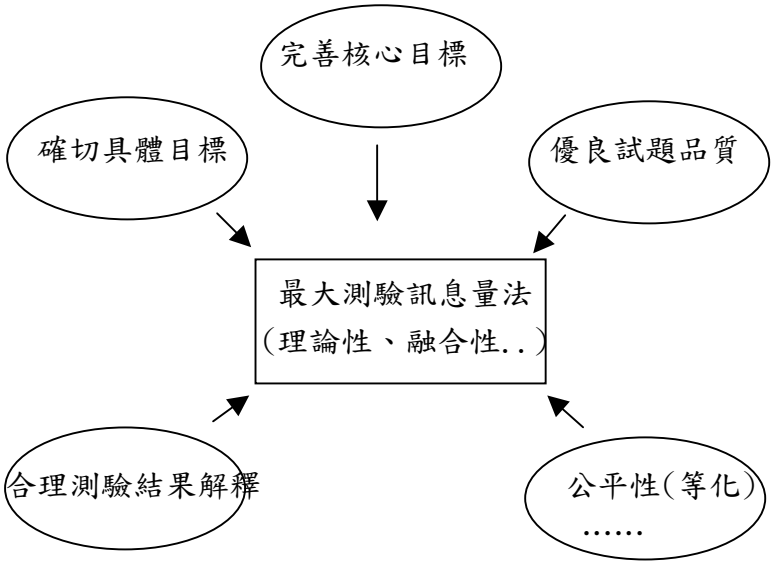

圖 2-7 本研究所運用精熟標準設定方法之廣義流程概念圖

第二節 最大測驗訊息量法

在精熟標準設定方法中，絕大部份是建立於古典測驗理論之下，不僅易受樣本限制、能力估計偏誤較大，且有平行測驗建構困難等難題存在，使得建諸於此的方法，在理論、應用上都有所限制。對於此，試題反應理論則能有效克服與解決，本研究即應用試題反應理論中最大測驗訊息量的概念於精熟標準的設定，於此範疇上相關理論基礎，茲探討如下。

自從 Lord(1980)發表第一本以「試題反應理論」為名的專書後，當代測驗理論即正式以 IRT 為中心架構，於是開始有許多學者嘗試應用其理論來解決測驗上相關議題，而於試題訊息函數(item information function)運用上，則多傾向於試題選擇(item selection)面向(Hambleton & de Gruijter, 1983；Harwell, 1983)或信度的探討(Samejima, 1994)，而於精熟標準設定上的探討，雖然其相關概念早先即有學者提出，但實徵研究則極為少見。因而，本研究在此嚐試就其公式意涵、試題選擇、統計考驗力等面向說明其於精熟標準設定上之理論基礎。

壹、公式意涵面向

在精熟標準的設定中，我們所關心的是該標準能從受試者的反應中，準確(即分類誤差最小)的將精熟/未精熟者加以分類，而最有效將受試者分為精熟/未精熟的測驗，如 Lord & Novick(1968)所說是包含能有效鑑別精熟標準附近能力的試題。加以延伸，即代表著其試題反應函數(item response function)在精熟標準能力值的第一階導數(first derivative)期望是最大的，除此之外，van der Linden(1981, p.393)認為測驗中受試者在精熟標準附近的試題反應分佈(scatter of the item responses)亦是影響鑑別力的重要因素，若愈分散，則其鑑別力就愈小。若將上述概念對照於 Birnbaum(1968)所提試題訊息函數，其定義為：

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \qquad i=1，2\cdots n \qquad （公式 2-4）$$

$I_i(\theta)$代表著試題i在能力θ值上提供的訊息，$P_i'(\theta)$為在θ點上$P_i(\theta)$值的第一階導數，$P_i(\theta)$為能力值θ的受試者在試題i上的試題反應函數(item response function)(即正確反應機率)，$Q_i(\theta)=1-P_i(\theta)$則代表錯誤反應機率；此即代表著試題在精熟標準能力值的訊息量期望是最大的(分子期望愈大愈好/分母期望愈小愈好)。

若將此應用到三參數對數型模式(three-parameter logistic model)時(即本研究所採用的模式)，則公式 2-4 可化簡為(Birnbaum, 1968；Lord, 1980)：

$$I_i(\theta) = \frac{a_i^2(1-c_i)}{[c_i + e^{a_i(\theta-b_i)}][1+e^{-a_i(\theta-b_i)}]^2} \quad \text{（公式 2-5）}$$

由公式 2-5 可推知，當b_i值愈接近θ時(即試題難度符合某考生能力時)，訊息量較大；當a_i參數(即試題鑑別度)較高時，訊息量也會較大；當c_i參數(即能力低的考生答對試題的機率)接近 0 時，訊息量則會增加。在綜合考量各試題參數影響後，Birnbaum(1968)指出某個試題所能提供的最大訊息量(maximum information)，剛好會出現在能力參數為θ_{max}的點上，其值為：

$$\theta_{max} = b_i + \frac{1}{a_i}\ln\left[0.5(1+\sqrt{1+8c_i})\right] \quad \text{（公式 2-6）}$$

根據 Birnbaum(1968)的推演結果，每一試題對整份測驗訊息量所作的貢獻，並不會受到測驗中其它試題的影響，這代表著一份測驗在某一特定θ值上所提供的訊息量，是等於在此θ值上的所有試題訊息量的總和，稱為「測驗訊息函數」(test information function)，記作$I(\theta)$：

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \quad \text{（公式 2-7）}$$

在精熟標準點的試題訊息量期望是最大的，而其所組成的測驗，同樣期望測驗訊息量是最大的。根據公式 2-6，乃隱含著只要給定某試題的a_i、b_i、c_i即可求得此試題最大訊息量對應的能力值θ_{max}，而推演到整份測驗上，同樣可以求得整份測驗的最大訊息量所對應的能力值θ_{max}；綜合上述，此即代表整份測驗最佳的精熟標準。

貳、試題選擇面向

試題訊息函數最廣泛的運用，當屬於試題選擇範疇，其具有表示「試題對能力估計正確性貢獻量大小的功能」，可用於診斷試題的好壞上，以便再進一步挑選合適的測驗試題(Hambleton & de Gruijter, 1983；Harwell, 1983)。具體而言，試題的最大訊息量所對應的能力水準θ_{max}，即代表某試題所能精確測量或估計的能力參數值。因此，只要求出θ_{max}，便可推估該試題所精確測量到的潛在特質大概是多少，或者說，該試題較適合於測量何種程度的潛在特質或能力(余民寧，1992)。基於此概念，電腦化適性測驗(computerized adaptive testing)從而蘊育而生，乃根據受試者先前反應，選擇適合其能力標準的試題給與施測，若將此延伸至與精熟測驗相結合，即如同 Reckase(1983)、Kingsbury & Weiss(1983)、Weiss & Kingsbury(1984)所稱之適性精熟測驗(adaptive mastery testing)，強調著根據受試者現行能力水平，選擇最適合試題給與施測，直至受試者估計出之能力區間未包含精熟標準時即停止，最後，以估計之區間能力值位於精熟標準以上或以下，以決定受試者精熟/未精熟狀態。但對此，Spray &

Reckase(1994, 1996)在其研究中顯示，選擇在精熟標準點上產生最大訊息量的試題，是比依據現行受試者能力選題方式為佳，其概念就如同 Wiberg(2003)以另一同義辭彙所稱效標參照電腦化測驗(criterion-referenced computerized testing)，認為我們所關心的並非是估計受試者現行的能力，而是以檢測受試者是否具備至少某水平的能力為主。對於此概念，進一步輔以圖 2-8 的能力區間穩定估計概念作說明，假設有二名受試者 A、B，其得分較為遠離精熟標準θ_m，另一名受試者 C，得分則較靠近精熟標準，若選擇在精熟標準點上產生最大訊息量的試題給與施測，則隨著試題的增加，受試者 C 能力估計會愈趨穩定(能力區間漸小)，但對受試者 A 與 B 而言，因試題較不適合其能力水平，能力區間範圍始終較 C 為大，但其能力的方向性已漸趨穩定(亦即是朝精熟或未精熟方向)。就精熟測驗的角度而言，雖然受試者 A 與 B 能力估計仍不穩定，但我們所關心的是判別受試者精熟/未精熟的狀態即可，並不需精確估計其能力，因此，是否應以關切精熟標準附近能力的受試者為優先，以期能夠儘速達到穩定，減少分類錯誤，而較遠離精熟標準之受試者，只需具穩定精熟/未精熟方向性即可。

本研究即運用上述相同的概念於精熟標準設定上，我們所關心的是一份精熟測驗中何者才是最佳的精熟標準？是否代表整份測驗試題所形成的最大訊息量所對應的能力值即是最佳的精熟標準？才真正具有精確測量該能力值與區分精熟/未精熟者的特性。

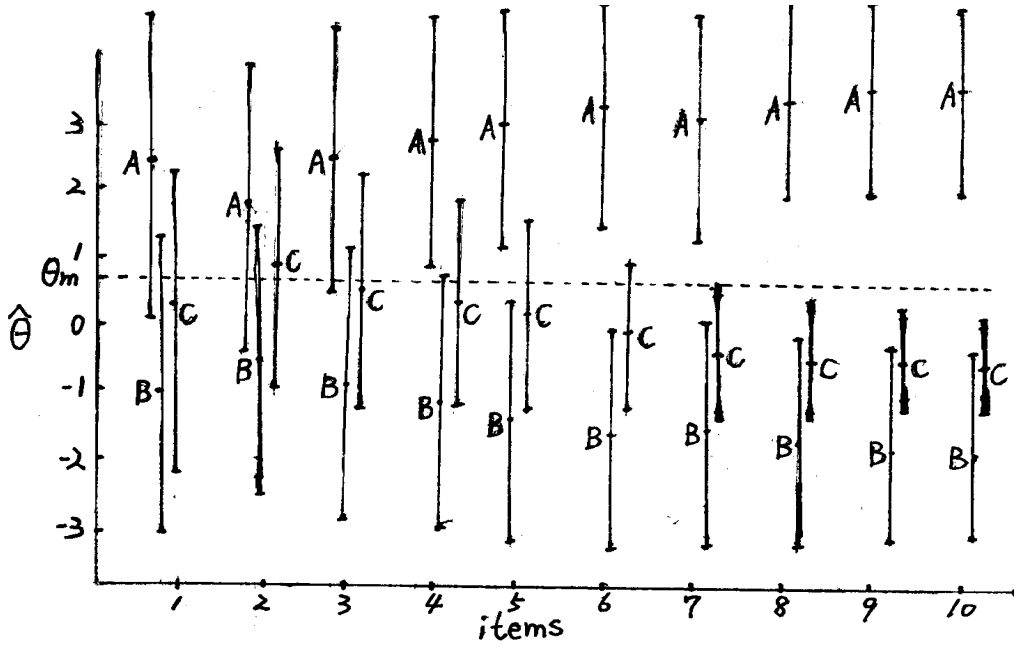

圖 2-8 能力區間穩定估計概念圖

參、統計考驗力面向

如同上述，我們所關心的問題仍然是一份精熟測驗中，何者才是最佳的精熟

標準？Wiberg(2003)認為此乃屬於統計假設範疇，意在考驗受試者是否具備至少某水平能力，而其假設可為：

$$H_0：\theta \le \theta_0$$
$$H_1：\theta > \theta_0$$

其中，θ代表著受試者的能力值，而θ_0代表著我們所設立的精熟標準，第一類型錯誤即為受試者事實上缺乏能力但我們卻將他視為精熟者，第二類型錯誤則為受試者事實上有能力但我們卻將他視為非精熟者。而根據 Wiberg(2003)的推導，認為能使上述統計考驗力達最大的試題難度值，即如下公式：

$$b_i = \theta_0 - \frac{1}{a_i}\ln\left[0.5(1+\sqrt{1+8c_i})\right] \qquad \text{（公式 2-8）}$$

經轉換後

$$\theta_0 = b_i + \frac{1}{a_i}\ln\left[0.5(1+\sqrt{1+8c_i})\right] \qquad \text{（公式 2-9）}$$

會發現精熟標準θ_0正好等於最大訊息量所對應能力值$\theta_{\max}$，符合 Birnbaum (1968)也曾基於將錯誤分類機率減低到最小程度的考量，認為最大訊息量所在點的鑑別度與難度值最具有區別的功能。

基於上述理論基礎，本研究乃以測驗的最大訊息量所對應求得之$\theta_{\max}$，代表著整份測驗的精熟標準，並將此稱為「最大測驗訊息量法」(maximum test information approach)。另，就此亦可延伸出與精熟標準設定融合的議題，在本研究接續探討之轉換通過分數與差異能力描述一節中，將會有更詳細探討。

第三節 轉換通過分數與差異能力描述

相較於傳統古典測驗理論下之精熟標準設定方法，建基於 IRT 的最大測驗訊息量法雖然在理論背景、運用彈性上具有較佳優勢，但其仍有缺點存在，即是該理論多建立在艱深的數理統計學基礎上，一般民眾難以理解，因此在關乎自身重大權益的測驗時，都難以被接受。因而，本研究企圖探討幾種有助於解釋測驗結果的方法與概念，以下茲就轉換通過分數(屬於量的範疇)與差異能力描述(屬於質的範疇)，分別討論之。

壹、轉換通過分數的方法

由最大測驗訊息量法求得的精熟標準 $\theta_{\max}$，為了與社會大眾傳統答對題數或百分比率(percent correct)概念作連結，本研究在此提出換算古典測驗分數法與測驗特徵曲線構圖法於轉換 IRT-θ 能力值成古典測驗答對題數，兩者相互比較，企圖挑選出較佳方式，以作為解釋測驗結果時運用。以下茲就二法的理論、概念說明之。

一、換算古典測驗分數法(transformed classical test scores approach，簡稱 TCTS)

古典測驗理論下，客觀測驗的評分方式，乃於每一試題之得分，答對者給予 1 分，答錯得 0 分；或是以不同配分表示，如答對得 2 分…等，但其本質上都不變，皆是以每位受試者於整份測驗之答對題數總和，即視為其能力估計值，傳統評分方式可表示為(Crocker & Algina, 1986)：

$$X_j = \sum_{i=1}^{n} X_i \qquad \text{（公式 2-10）}$$

X_j 代表著第 j 位受試者之總分，即是整份測驗試題(1 至 n)答對題數的總合。而在試題反應理論中，三參數對數型模式下，欲計算某位考生答對某試題正確反應機率或能力值時，則需考量到該試題難度參數、鑑別度參數、和機運參數三者。因此，在古典測驗理論下，一份測驗之答對題數（即總分(total raw scores)）相同的考生，其 IRT-θ 能力估計值便會隨著其反應組型(response pattern)的不同而有所不同。舉例來說，假設有五題測驗試題，其可能的總分及其對應的反應組型如表 2-7 所示：

表 2-7　五個測驗試題總分及其對應的反應組型

總分	可能的反應組型				
0	00000				
1	10000	01000	00100	00010	00001
2	11000 01010	10100 01001	10010 00110	10001 00101	01100 00011
3	11100 10011	11010 01110	11001 01101	10110 01011	10101 00111
4	11110	11101	11011	10111	01111
5	11111				

註：反應組型中，1 代表答對；0 代表答錯；題數乃由左而右，如 10000 即代表五題試題中只答對第一題

就得分 1 分的考生而言，其可能的理論反應組型即有 5 種，在 IRT 下，每一種反應組型都會有一種相對應的能力估計值被估算出來，如下所示：

10000 = θ_1

01000 = θ_2

00100 = θ_3

00010 = θ_4

00001 = θ_5

這代表對古典測驗分數總分為 1 分的考生而言，其可能的 IRT-θ 能力估計值就會有 5 種之多，因而在余民寧、汪慧瑜(2005)所發展的「答對題數與量尺化能力分數」轉換方法中，乃建議採用各反應組型下所有能力估計值的平均數，作為某種相同測驗分數的對應代表，並將其視為該古典測驗分數（即答對題數）所對應之能力值，在本研究即將此稱之為換算古典測驗分數法(transformed classical test scores approach，簡稱 TCTS)。茲以總分為 1 分考生而言，其 θ 能力值則為：

$$x_1 = \overline{\theta_1} = \sum_{j=1}^{5} \theta_{1j} \Big/ 5 \qquad \text{（公式 2-11）}$$

其中，x_1 代表答對題數為 1 下的所有 θ 值的平均數，並簡化成以 $\overline{\theta_1}$ 來表示。由此，我們可以延伸計算出答對題數總分為 2 分、3 分、…、一直到 n 分的考生與其相對應的各個平均能力估計值，其通用公式為：

$$x_i = \overline{\theta_i} = \sum_{j=1}^{m} \theta_{ij} \Big/ m \qquad \text{i=1, 2, …, n} \qquad \text{（公式 2-12）}$$

其中，x_i 代表答對題數為 i 下的所有 θ 值的平均數，n 代表某一測驗中的試

題總數，m 代表得分為 i 分者的所有可能反應組型的總數，理論上最多為 m=$\binom{n}{i}$ 的組合數(combination)。但於實際資料上，理論上的反應組型並非每一種都會出現，通常，只會侷限出現在其中的少數幾類上。對此，本研究乃應用相同的概念，採用實徵資料上受試者的實際表現，將各得分所對應的實際能力 θ 估計值平均數，視為古典測驗答對題數與 IRT-θ 能力值的對照值，進一步求得精熟標準 $\theta_{\max}$ 所對應的古典測驗答對題數，以作為另一種精熟標準，並評估其可行性。

二、測驗特徵曲線構圖法(test characteristic curve mapping method，簡稱 TCCM)

試題反應理論下，受試者的能力表現與潛在特質間的關係，代表著一種數學函數關係，對此，可透過一條連續性遞增的函數來加以詮釋，這個函數便叫作試題特徵曲線（item characteristic curve，簡稱 ICC)，如圖 2-9 所示，其 X 軸即為受試者能力估計值 θ，Y 軸即代表著在第 i 題答對機率值 $P_i(\theta)$，實際上，即是將不同受試者在單一試題(item)上各能力估計點連結而構成的曲線。

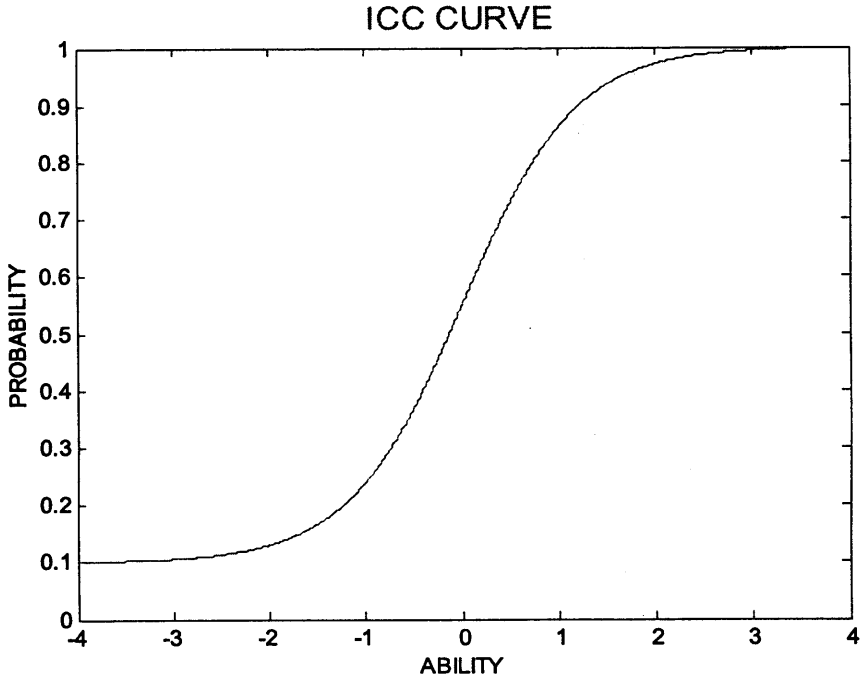

圖 2-9 試題特徵曲線圖

若進一步將各試題的試題特徵曲線加總起來，即可形成整份測驗的測驗特徵曲線(test characteristic curve, 簡稱 TCC)，亦即將不同受試者在某一份測驗各試題上能力估計值加總後，所得之整份測驗的能力估計點連結而構成的曲線，其 X 軸仍為受試者能力估計值 θ，Y 軸則代表著受試者於全部測驗試題答對機率值的總和，其 Y 軸刻度，以公式表示可為：

TCC-Y 軸=$\sum_{i=1}^{n} P_i(\theta)$　　i=1，2…n　　（公式 2-13）

其值乃介於 0 至 n 之間，而較常見的 TCC 曲線，其 Y 軸刻度乃更進一步將其轉換為領域分數π(domain score)，以公式表示可為：

$$\pi = \frac{1}{n}\sum_{i=1}^{n} P_i(\theta)$$　　（公式 2-14）

乃將受試者於全部測驗試題答對機率值的總和除以測驗題數，使其值介於 0 至 1 間，因是屬於直線轉換，所以仍不會改變其相對關係。

測驗特徵曲線構圖法(TCC mapping method)即是利用上述的概念，採用圖形對應方式進行能力值的轉換，如圖 2-10 所示。首先，將 X 軸上求得之 IRT-θ 能力值，畫一垂直於 X 軸的直線，交於 TCC 曲線上的一點，再經由此點，畫一垂直於 Y 軸的直線，交於 Y 軸上的一點，此點即代表著換算後之古典答對題數或機率值(視其 Y 軸為 TCC-Y 軸或π值)。

利用 TCC 於轉換 IRT-θ 能力值與古典測驗答對題數或比率值，其相關的文章早在 van der Linden(1981, p.396)、Hambleton & de Gruijter(1983, P.358)、Kane(1987, p.341)等人的研究中，皆曾於概念上提出，而 Hambleton(1998)、Reckase(1998)更進一步對其具體描述，並實際運用此法於轉換 NAEP(National Assessment of Educational Progress)測驗中 IRT 量尺分數與專家評定古典精熟答對題數，足以顯見其於實務用途的重要性，但於轉換效果探討方面，卻較少見諸於研究中，因此，本研究乃企圖評估比較上述換算古典測驗分數法與測驗特徵曲線構圖法於轉換時的效益，以供未來應用時參考。

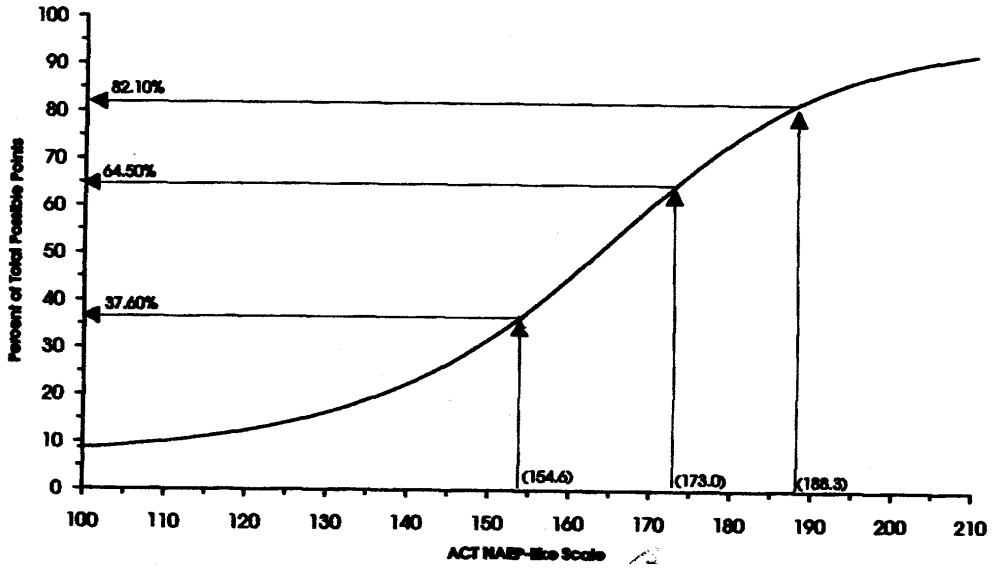

圖 2-10 測驗特徵曲線構圖法（取自 Reckase(1998, p.18)）

貳、能力標準描述

本研究於文獻探討初，即曾於以能力標準設定為核心的測驗建構流程中說明，測驗結果的解釋需參照試題編製的具體目標，提供精熟/未精熟者能力差異標準，此屬於「質」的解釋範疇(如精熟者需懂得三位數加法)，乃有別於上述轉換分數上「量」的解釋面向(如精熟者得分需高於 60 分)。因而，本研究在此進一步探討如何描述精熟者在預設內容領域中較未精熟者所具備較突出或精熟的知識或技能？

對於隸屬實質層面的測驗結果報告，NAEP 亦發展出多種的解釋方法，如定錨點(anchor points)(Beaton & Allen (1992))、成就水平法(achievement levels approach)或稱之為能力標準(performance standards) (Koretz & Deibert (1995))等，概念上乃需於量尺分數上任意選擇數個定錨分數點(如200、250、300 分)或採用各能力標準的分數點(端視研究者所採方法而定)，而後,根據此定錨點(或能力標準分數點)選擇一個或數個具內容代表性的範例試題(exemplary items)，對此將之稱為定錨點題目(anchor items)(為有別於等化概念中之定錨試題，乃以此名詞稱之)，以此為基礎作為受試者於該分數點的能力表現描述。對此，加以延伸，可進一步從幾方面作深入探討：

一、定錨點的選擇

有關於定錨點個數、類型的選擇等，端視研究者自身的考量，若採用的個數過多時，易無法配合挑選具代表性試題，個數過少時，整體上較缺乏廣泛的能力描述，而傳統上，如 NAEP 於 0 至 500 分量尺上，多採取五個定錨點(150、200、250、300、350)以作為測驗結果的解釋(Beaton & Allen, 1992)。而本研究為有效區辨精熟/未精熟者間差異能力，乃分別以 2 個定錨點代表精熟/未精熟者能力，以期收廣泛解釋之效。

二、定錨點題目的選擇

關於定錨點題目的選擇，除需具內容代表性意義外，更期望試題對於大部份擁有較多知識或技能的高能力定錨點受試者而言，能較次低能力定錨點的受試者，於試題表現上具備一致性高的正確反應機率，因而，於 NAEP 中另有二種常見準則(Beaton & Allen, 1992)：

1、位於第i個定錨點上受試者的正確反應機率值，至少需大於或等於.80；相對的，位於下一個定錨點受試者的正確反應機率值，則需小於.50。

或

2、位於第i個定錨點上受試者的正確反應機率值，至少需大於或等於.65；相對的，位於下一個定錨點受試者的正確反應機率值，則需小於.50，此外，兩定錨點間試題答對機率差異需大於或等於.30。

有關上述機率值的標準，就如 Huynh(1998)所認為，多半建立於實務經驗上，並不具理論基礎，因而，其乃將試題的表現分離為正確反應(X=1，代表在此試題下被期望具有此知識的意函)與不正確反應(X=0，代表不具此知識)，運用最大訊息量的概念，確立各自對應能力值，以作為上述選題準則。具體運用上，曾建議於僅有四個選項的選擇題中，其選題準則可為：位於第i個定錨點上受試者的正確反應機率值，至少需大於或等於.75，而位於下一個定錨點受試者的正確反應機率，則需小於.56(或為避免過於嚴苛，採取兩者中數.66)，此外，兩定錨點間試題正確反應機率差異需大於或等於.19。而本研究即採 Huynh 較符合理論性之建議，以作為篩選定錨點題目。

三、定錨點描述

受試者於各定錨點上所具備的能力表現，Koretz & Deibert (1995)認為可分三種方式加以呈現，第一、技能為基礎(skill based)，主要以描述受試者在內容領域上所具備知識與技能為目的；第二、年級為基礎(grade based)，乃刻劃著不同定錨點以反應哪一年級的能力表現為主；第三、預測用途(predictive)，則描述受試者於未來活動中，可能成功的表現。而本研究乃認為實務運用上，描述與表達受試者能力的方式可以是多元的，決策者可根據本身所擁有資料、測驗目的等等，加以調整，研擬一份具代表性能力描述報告，因而，本研究乃欲以層次、能力累進的方式以描述精熟/未精熟間差異能力。

四、與最大測驗訊息量法融合議題

由最大測驗訊息量本身的概念出發時，也會引發某些疑義，亦即有人會認為尋求測驗最佳的精熟標準乃屬工作分析(task analysis)的範疇，主要由專家依據任務需求及測驗難度判定通過標準，但由最大測驗訊息量法求得之 IRT-θ 能力值，主要是由組成測驗的試題所決定，與受試者是否適任工作似乎無關聯。針對此項難題，本研究企圖運用先前所強調「元素的搭配組合與調整」、「廣義測驗建構流程」概念來作補充解說。

首先，需明瞭測驗試題的組成，乃係由更高階層之具體目標與核心目標所決定，由此檢視初次(暗指尚未能完全確立特定精熟標準)設定某資格檢定測驗之精熟標準時，本研究提出如圖 2-11 之流程概念圖作說明。在確立的核心目標主導下，於決定精熟標準前，專家、學者乃會針對欲檢測的內容範圍、最低能力標準要求、人才甄選取向與目的達成共識，進一步，會具體化該測驗適合的能力範圍(如約 50%-75%左右考生能答對的題目)，以利試題的編撰與接續之選題組卷，而後，不僅需考量配合先前所設定之能力範圍以挑選題目，並且需納入各定錨點之具內容代表性題目，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，經由最大測驗訊息量法以求得的暫時θ_{max}，即可代表決策者經由測驗表達對精熟標準的要求。但此時，可進一步搭配

另一項元素：專家判定，並考量此θ_{max}對社會、經濟可能造成的影響等因素，再加以作調整，以決定出最佳的精熟標準，之後，再進行後續之分數轉換與測驗結果的解釋即可。

簡言之，本研究即是強調以廣義測驗建構流程來檢視「專家+最大測驗訊息量法+專家」等三項元素搭配，預先經由專家、學者確立大致的能力範圍，再經由最大測驗訊息量法以確立該測驗中，何者為最佳且特定的精熟標準以供參照用途，後續，再經由專家考量其它因素以作調整。對照上述疑義而言，此流程即顯示，本研究雖然以集中探討最大測驗訊息量法為核心，但並非強調精熟標準需完全由最大測驗訊息量法所求得之θ_{max}所決定，而是期望實務運用時，能藉由最大測驗訊息量法具備定位出測驗中最佳精熟標準點的能力，再搭配專家意見作調整，以尋求於該資格檢定中最適切之精熟標準。而相對於本研究設計上，即是在此概念下，專注於提供多元效度證據，以佐證最大測驗訊息量作為精熟標準設定之輔助工具的可行性。

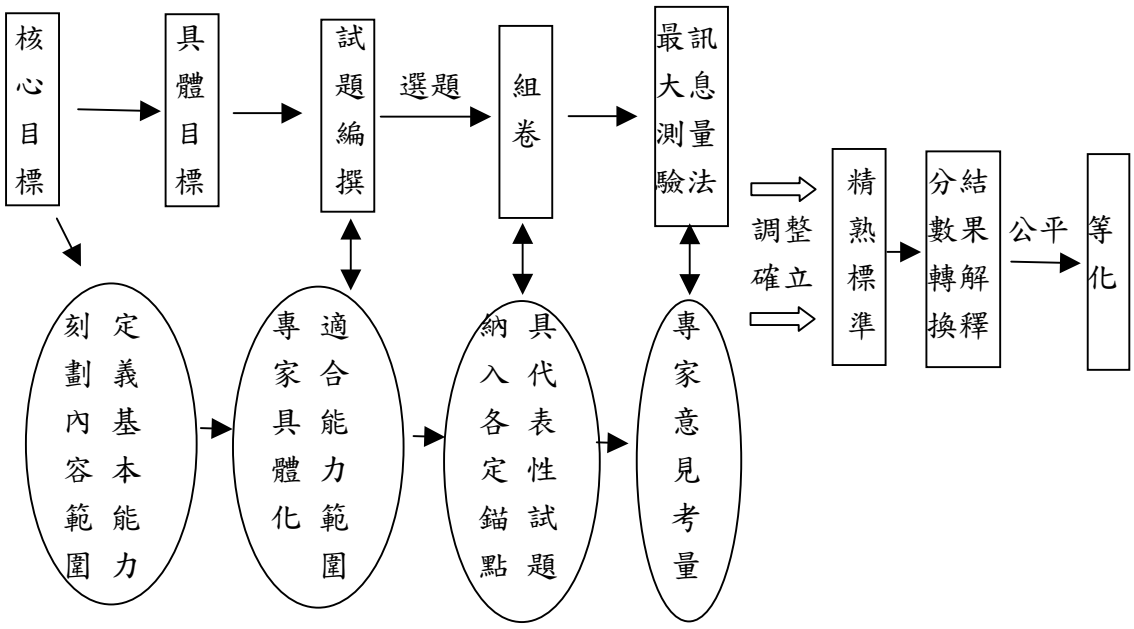

圖 2-11 最大測驗訊息量法延伸議題解說圖

另一方面，於文獻探討第二節，筆者曾說明試題訊息量的概念最常用以作為試題挑選，而組成的測驗，其最大測驗訊息量對應之 θ 能力值，乃具最大區別力，可為最佳精熟標準，若將此與上述定錨點概念相結合，會產生某種解釋上的困難，即是利用試題訊息量作為挑選試題依據時，組成的測驗皆以準確測量精熟標準的能力值為目的，但其是否足以符合整個測驗內容代表性，則值得懷疑。因而，採用定錨點解釋測驗結果時，可能產生於某些定錨點上較難以挑選具內容代表性的定錨點題目，對此概念，就如 Haladyna & Roid(1983)所認為，必須於解釋與測量準確性上作抉擇，而相對上述即是在探討如何有效納入各定錨點之具代表性試題。在此，進一步輔以圖 2-12 作詳細說明，若 10、30、50、70、90 分分

別代表某份測驗的定錨點，以 60 分為通過分數時，因而，若採用試題訊息量概念(以挑選最符合精熟標準的試題)組成的測驗，會產生多數的試題皆為符合 50、60、70 分等通過分數附近定錨點之代表性試題為主，而較遠離通過分數的定錨點，則易發生定錨點題目數不足的情況，而形成解釋上的難點。

在此，同樣以廣義測驗建構流程檢視之，並以最大測驗訊息量選題的角度出發，亦可導引出有效解決定錨點題目數不足的難題。而此解決的策略，乃結合選題、定錨點與目標訊息量(target information)的概念。就此，同樣輔以圖 2-12 作為解釋範例，首先，假設在某大型題庫中，先依據 10、30、50、70、90 分等預設的定錨點，將試題依其內容代表性(如一位數、二位數或三位數加法概念)分別歸類於適合各定錨點描述的類別中，而後，依照比例，分別從各定錨點中挑選能填滿預設的精熟標準 θ_{max} 所對應之測驗訊息量(即目標訊息量)的試題，如此，即可避免傳統僅根據試題最大訊息量指標挑選試題，而忽略內容代表性問題。但從較細部角度而言，此法雖可增加解釋上的方便性，同時亦可能犧牲某些具準確性但不具內容代表性的試題，因此而增加測驗長度，對此，研究者可自行加以調整，如僅按照某比例試題(如總測驗的 50%)依據上述方式挑選，其餘試題則仍依照試題訊息量作選擇，就此乃端視研究者自身需求，在測驗長度與內容代表性間作得失衡量。

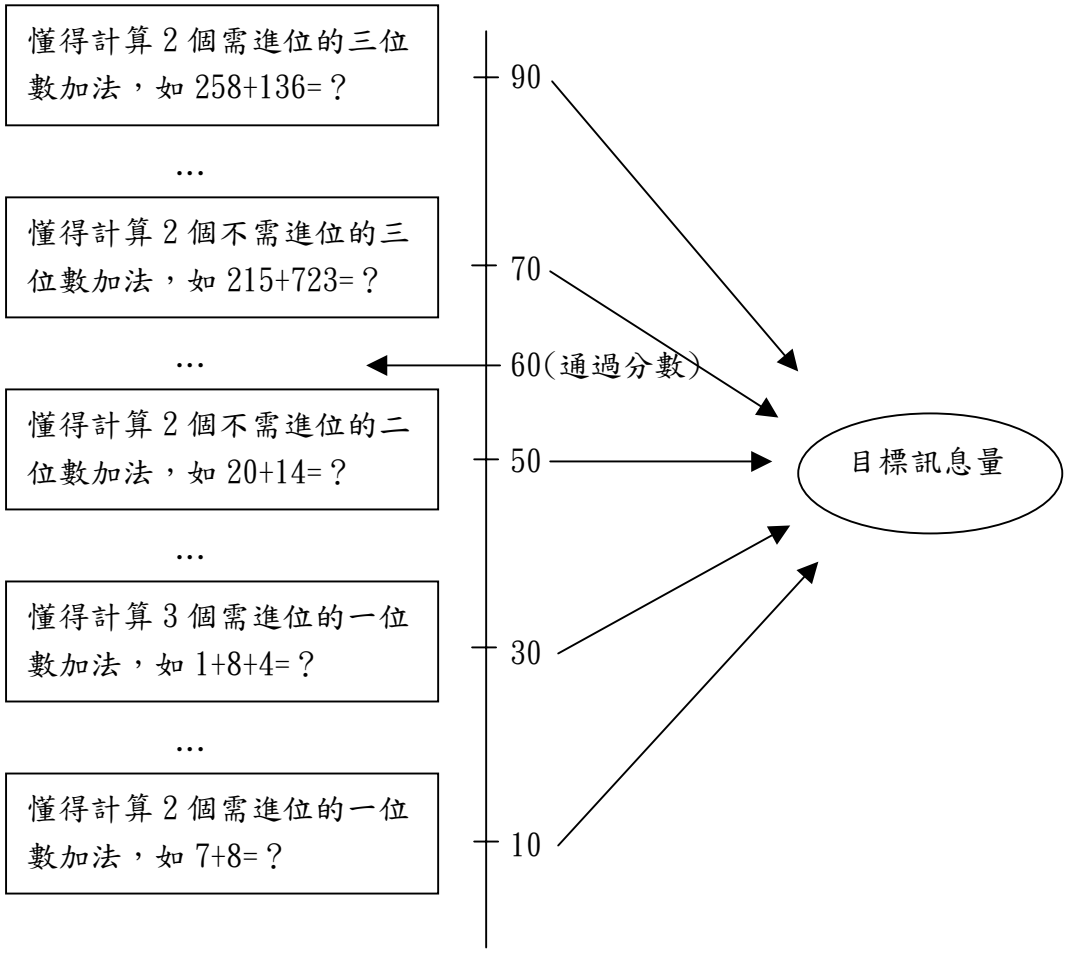

圖 2-12 解決最大測驗訊息量法疑義與定錨點題目數不足示意圖

第四節 精熟標準設定之相關議題探討

在精熟標準的設定上，我們除了關心此標準或決斷分數為何之外，仍希望這個分數或標準是最佳的，最能有效區別精熟/未精熟者。因而，在相關議題方面，本節將先就精熟標準的效度、信度議題進行探討，介紹判定精熟標準優、劣的準則與如何尋求支持效度的相關證據，而後針對測驗長度、測驗異質性等主要影響信度的因素，作統整描述。

壹、精熟標準效度議題

長久以來，效度(validity)的概念即是心理計量領域中最主要、最基本的核心，其概念演變、推展亦十分廣泛多元，Garrett(1937, p.324)認為此代表著測驗能測量到它所欲測量的程度；Guildford(1942)則認為此隱含測驗能有效預測未來行為的特質(引自 Angoff(1988))。不同學者對效度的描述，概念上皆有些微差異，延伸的分類類別亦不同，如內容效度(content validity)、效標關聯效度(criterion-related validity)、建構效度(construct validity)，乃強調因應不同測驗目的及情境而所需不同的證據(Angoff, 1988)或者更進一步演變融合成 Messick(1988, 1989)強調測驗分數的意義及結果詮釋的單一統整的建構效度，考量分數價值意涵與分數使用、決策的社會後果所呈現的證據。

對照於精熟標準設定，效度上則強調通過分數的意義與建基於此所作決策影響，就如同 Kane(1994)認為精熟標準所重視的兩個面向，第一，強調通過分數需具備相對應的特定能力標準，以瞭解所代表的實質意義，將此稱為描述性假設(descriptive assumption)，以探討高於能力標準的受試者較低於能力標準受試者所具備更多符合預設的知識或技能，此乃延伸 Messick(1989)所定義測驗解釋的證據基礎(evidential basis of test interpretation)建立於建構效度的概念，以連結通過分數的解釋乃建基於能力標準；第二，強調著能力標準的適切性，是否足以符合決策過程中的目標，乃牽涉政策採用、結果期望與研究者價值觀等，將此稱為政策假設(policy assumption)，即是延伸 Messick 所稱測驗解釋的影響基礎(consequential basis of test interpretation)，強調決策目的下，能力標準的適當性及其社會影響。

關於效度並非是直接證明其有或無的問題，而是如連續量尺上的不同程度，可採用 Messick(1988)認為累積不同證據以支持精熟標準有效性，或者如 Kane(1994)所認為檢視不同準則以確立精熟標準無效性(invalid)，而如何去評定某精熟標準的有效性，專家、學者所提準則則十分多元，但大致可依 Kane 所認為的分為三類：

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

過程的證據強調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：

(一)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)；

(二)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)；

(三)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)；

(四)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)；

(五)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)；

(六)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)。

由此觀之，過程證據乃強調精熟標準設定過程中各個面向的特徵，皆可提供相關效度的證據，此同時亦呼應著本文文獻探討第一節中，曾簡述過去相關研究議題的演進，乃漸強調著判定過程的嚴謹性、合理性。對照於本研究設計，乃從更廣義的「測驗發展」角度視之，以探討各個面向可行性(如核心目標、具體目標是否完備，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)，以提供有關精熟標準設定中多元效度的過程證據。

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」為核心所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，僅強調方法內運用的穩定與一致性，而根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：

(一)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)。

(二)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。

(三)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，藉以連結與所設定精熟標準間的相關，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：

(一)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)。

(二)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)。

相對於本研究設計，乃企圖提供多元的效度證據，除運用上述「以能力標準設定為核心的測驗建構流程」為基準，提供廣義效度過程證據外，另，採用受試者於兩份測驗上表現，以一份建立精熟標準，另一則視為外部效標資料，建立推論效果，而判定指標上採用過去研究中常用的百分比一致性與κ係數，相關概念茲於分類一致性信度的議題中探討。

貳、分類一致性信度的議題

本研究對於信度指標，如同 Kane(1994, 2001)主張乃將信度歸屬於上述效度證據中，但在此，為解說方便仍以傳統信度稱之。對於正確區分精熟者/未精熟者的分類一致性議題上，Hambleton & Novick (1973)曾建議對於精熟分類決策的信度，應定義為在同一份或複本測驗上兩次作答表現所作分類決策的一致性。而為了確定本研究所設立的精熟標準是可信的，在此採用此二位研究者所提供的檢視指標作為依據，其公式可以表示如下：

$$P_o = \sum_{J=1}^{m} P_{JJ} \qquad \text{(公式 2-14)}$$

P_o代表受試者被正確分類到j類精熟狀態的比率。實際上，若m等於 2 時，即是只有精熟/未精熟兩種狀態，此公式會與百分比一致性(percent agreement)公式完全一樣，皆代表著在兩次測驗上正確分類為精熟與未精熟者的比率總和。而百分比一致性公式為：

$$P_a = \frac{a}{N} + \frac{d}{N} \qquad \text{（公式 2-15）}$$

其中，如表 2-8 所示，a 代表兩次測驗皆被分類為精熟者人數，d 代表兩次測驗皆被分類為未精熟者人數，N 代表總人數。

表 2-8 受試者於兩次測驗上精熟分類的摘要表

測驗/精熟狀態		第二次測驗		合計
		精熟	未精熟	
第一次測驗	精熟	a	b	$a+b$
	未精熟	c	d	$c+d$
合計		$a+c$	$b+d$	N=$a+b+c+d$

雖然 P_a 值所代表的是兩次測驗結果分類完全一致的百分比，但Novick, Lewis, & Jackson (1973)認為它包含了一些隨機的成份在內，會對測驗的真實信度有高估的傾向；Swaminathan, Hambleton, & Algina(1974)亦認為此並無法將因隨機而造成的一致性分類比率納入考量。因此，Swaminathan, Hambleton, & Algina(1974)建議可採用 κ 係數(Kappa coefficient of agreement) (Cohen, 1960)來控制隨機因素所造成的影響，它是一種代表在排除隨機因素後，分類一致性的信度指標，其數值愈大，代表兩次測驗複本的信度或兩種判定模式的一致性愈高。其公式表示如下：

$$\kappa = \frac{P_a - P_c}{1 - P_c} \qquad \text{（公式 2-16）}$$

其中，$P_c = (\frac{a+c}{N})(\frac{a+b}{N}) + (\frac{b+d}{N})(\frac{c+d}{N})$，代表著當兩份測驗均無信度時，而將受試者歸為精熟機率為 $(\frac{a+c}{N})(\frac{a+b}{N})$，歸為未精熟的機率為 $(\frac{b+d}{N})(\frac{c+d}{N})$，將二者加結合，以表示在理論上由於隨機因素而評定為一致的百分比期望值。

以上這幾類指標，各有其優缺點，余民寧（2002）和Nitko (1983)曾建議如果強調重點在於全體一致性的分類，而不考慮過程的隨機因素，則使用百分比一致性指標即可；若研究重點關心測驗過程結果對分類一致性的貢獻程度，則選用 K 係數。此外，Berk(1984)、Hambleton(1990)、Subkoviak(1988)認為這兩個信度值會因決斷分數設定的高低而有所改變，不過兩者是呈相反之關係。有藉於此，本研乃同時採用百分比一致性(本研究只分精熟/未精熟兩種狀態，所以與P。值完全一樣)與K 係數信度指標，作為判定方法間分類正確性高低的指標。

但在僅提供信度指標下，即使分類的第一類型與第二類型錯誤率是最低的，並未能充分證明精熟標準設定方法是較佳的，據此輔以圖 2-13 為範例說明，假設有五名受試者其得分分別為 10-90 分，而當受試者於兩次測驗上表現完全一致時(同時隱含兩次測驗為嚴格複本)，精熟標準不論是 1、2 或 4 不等，皆具完美分類一致性信度，若就此準則判定，是否代表任何一個精熟標準皆適合呢？答案

明顯是否定的，仍需搭配效度訊息(如測驗結果解釋)才足以勝任，因而，對照於本研究設計，除提供信度指標外，仍搭配多元效度證據以支持本研究所採精熟標準設定方法。

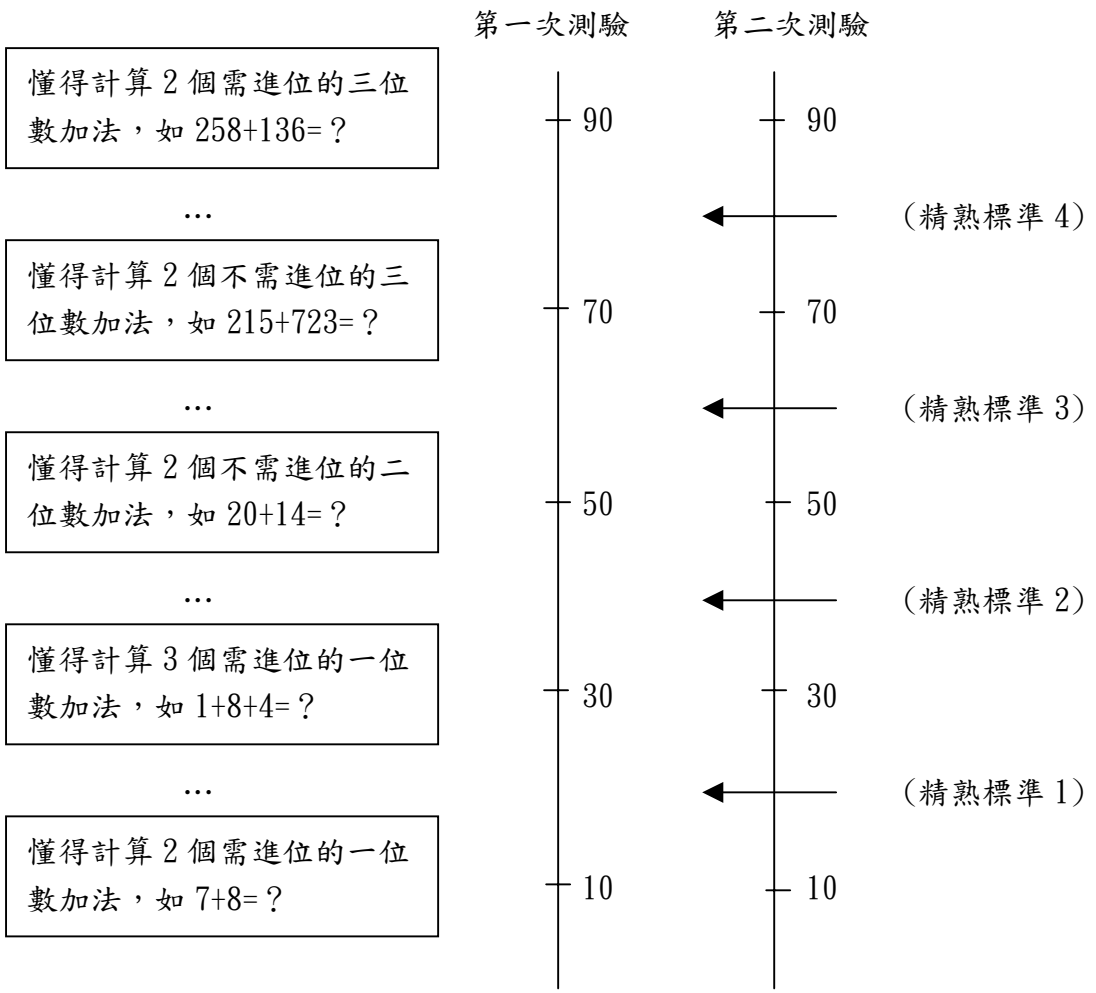

圖 2-13 完美百分比一致性信度示意圖

參、影響信度因素議題

在企圖提供多元的效度證據下，信度指標亦是重要的一環，但影響分類的一致性因素，除了上述設定精熟標準方法外，仍存在著許多干擾因子。在參考了 Hambleton, Mills & Simon(1983)、Hambleton(1983)、Hambleton(1990)、Eignor & Hambleton(1979)、Hambleton & de Gruijter(1983)等人的研究和分析之後，可歸納出如下幾項影響分類一致性的因素：

一、精熟標準設定方法的選擇；

二、測驗長度；

三、測驗異質性：考量兩次測驗上試題難度不一致時，其分類正確性；

四、精熟標準的位置：探討精熟標準為兩極端能力值或是屬於平均能力值時，其

分類正確性為何，此與第五項因素概念雷同；

五、受試者能力分配：探討受試者得分分配在不同偏態情形下，可能的分類情形，同時比較分配人數在較遠離或聚集於決斷分數時，可能的分類一致性結果；

六、試題特徵的差異：考量測驗試題的鑑別度值、難度值等對分類正確性的影響；

七、試題選題方式；

八、試題反應模式選擇；

九、測驗試題編製優劣程度等。

由上述可發現，對於影響正確分類精熟/未精熟者問題上，同時存在著太多干擾因素，且彼此又相互重疊、交互影響。因此，為了能夠較完整探討本研究所採精熟標準設定方法對分類一致性的影響，在考量本研究是採實徵資料且上述因素中某些乃屬研究者實務運用時無法探制的因素(如受試者能力分配)情況下，在此，儘量將如試題特徵差異、受試者能力分配、及試題反應模式選擇等干擾因素控制保持一致，暫時不在此納入分析探討，而僅針對測驗長度及測驗異質性等主要干擾因素進行研究。所以，本研究擬進行探討的主要相關議題，茲歸納敘述如下：

一、測驗長度的議題

測驗長度對於正確分類的影響，自早即被廣泛進行討論與研究，但通常都伴隨著其它因素，如 Hambleton(1983)曾採用電腦模擬資料，比較不同參數模型、測驗長度(分成 10、15、20、40 題)、不同精熟標準等對正確分類機率的影響；Hambleton, Mills, & Simon(1983)探討試題選擇方式、測驗長度(分成 2 至 20 題)、測驗試題異質性對分類一致性影響；Millman (1973)、Wilcox (1976)、Eignor & Hambleton(1979)則提供許多對照表，用以幫助決定適當的測驗長度與領域分數(domain score)、精熟標準、受試者能力分配狀況等因素之於分類正確率的對照。上述的研究，大多一致肯定隨著測驗長度增加，均能有效增進受試者被正確分類的機率。

但進一步深究，可發現上述的研究多是探討在事先設定之某個固定精熟標準下，測驗長度與其它相關因素間於分類準確性上之互動影響，而並非從精熟標準設定方法角度出發，運用某種方法，就其產生的精熟標準，再檢定其可能受到測驗長度干擾或影響程度，就如同 Plake & Melican(1989)即曾探討運用 Nedelsky 法時，測驗長度與難度對於評審判定時可能的影響一般。同時，隨著方法間特性的不同，其影響層面亦有所差異，如採用 Angoff 法時，測驗長度(長或短)對於評審判定時的倦待感或熟練程度，以致影響判定精熟標準結果，乃不同於採用集群分析法所產生的效果，就此概念，本研究乃以方法角度出發，探討最大測驗訊息量法、轉換通過分數之方法、定錨點題目篩選與測驗長度間之關聯，描述如下：

(一)測驗長度與最大測驗訊息量法之關聯

相對於本研究所運用之最大測驗訊息量法，不論是由 Weiss & Kingsbury(1984)從適性精熟測驗的選題角度，認為隨著試題的增加，受試者能力會愈趨穩定(同等於愈容易確立精熟/未精熟方向)或 Wiberg(2003)認為選擇愈佳的試題(具最大訊息量的試題)，則可以較少測驗長度即達相同區別考驗力，皆隱含測驗長度是影響精熟/未精熟者分類的要素，且測驗愈長，分類愈趨穩定，但何種測驗長度才是較基本、適切的要求值呢？若能加以定位，則採用傳統紙筆精熟測驗亦能具有電腦化精熟測驗之以較少測驗長度即能有效分類精熟/未精熟的優點。

對此，再加以詳細說明之。電腦化精熟測驗如本章第二節中曾說明，於篩選試題時，僅需挑選適合精熟標準能力者之題目即可，因而，只會組卷而形成單一的測驗，乃不似電腦化適性測驗般，需藉由電腦協助以檢測受試者能力來挑選不同試題，而形成不同的受試者皆有屬於自我的測驗(為傳統紙筆測驗較難執行部份)，對此，是否代表在傳統紙筆測驗方式下只需以精熟標準為核心加以組卷，亦能具有相同測驗長度下達穩定分類的效果，而不需借用電腦選題的功能(在此，不隱含同等具有電腦本身的優點)，有藉於此，本研究乃極欲探討何種測驗長度才是較適切以達穩定分類精熟/未精熟者的要求值。

(二)測驗長度與轉換通過分數、定錨點題目篩選間關聯

在探討測驗長度與轉換通過分數的關聯上，首先，就換算古典測驗分數法而言，其轉換的穩定性較易因受試者作答反應的不同，而隨之變化，尤其是對於出現異常作答反應情況下，採用平均能力值以替代古典測驗答對題數，更是如此。相對於測驗長度，較長的測驗亦隱含著在同一古典分數下會出現較多的能力對應值(包含異常能力值)，這是否代表測驗較長較易產生轉換不穩定的狀況呢？但另一方面思考時，似乎又並非是絕對的，得視實際異常反應出現情況而定。若就能力估計穩定性而言，測驗較長時，能力估計較為穩定，似乎又能減低異常能力估計值影響？此外，就百分比一致性而言，測驗長短與錯誤分類數似乎有某種關聯，亦是如當測驗長度愈長時，各分數點人數會相對較測驗長度短時為少，因而，於相同轉換分數差異時(如兩種方法經轉換為古典測驗答對題數後之差異為 2 分時)，測驗長度愈長時，該差異分數點人數會較少，相對的所造成分類誤差人數，理論上而言，亦會較少(即百分比一致性會較高)。另一方面，就測驗特徵曲線構圖法而言，與測驗長度的關聯似乎亦存在著此一互動關係，但實質情況為何？是本研究所欲探討之重點。

測驗長度與篩選定錨點題目間關聯上，若由直觀角度視之，當測驗題目少時，相對於可提供挑選合乎準則且具內容代表性的範例試題，自然會較少，尤其當所設定的定錨點個數增加時，更是如此，但其是否為真正影響的主因，亦或是有其它成因存在，由於過去研究顯少提供這方面議題探討，因而，本研究乃企圖以探索角度分析兩者之互動關係。

二、測驗異質性議題

測驗異質性的議題係指在兩次測驗上，若試題是由較異質題庫(即試題難度、鑑別度分歧較大)中採隨機方式抽取出時，而組成之測驗卷，試題難易程度度較易產生不一致現象，致使對於一致性分類精熟/未精熟者產生影響。對此議題，Hambleton, Mills, & Simon(1983)與 Hambleton & de Gruijter(1983)皆從選題面向著手，分析古典測驗理論下選題方式的不適當性。對此，Hambleton & de Gruijter(1983)曾舉了類似表 2-9 例子(為符合本研究主題而稍經修正)，剖析古典測驗理論選題下，難度值符合通過分數並非是一較佳選題方式，而本研究乃延用其概念，嘗試從精熟標準設定方法的面向切入，探討兩次測驗難度差異大時，其分類的正確率變化情況，以下茲就測驗異質性與最大測驗訊息量法、轉換通過分數與定錨點題目篩選間關聯作陳述。

(一)測驗異質性與最大測驗訊息量法之關聯

相對於古典測驗理論，在 Hambleton, Mills, & Simon(1983)運用 IRT 概念以選題的研究中，則顯示經選題後組成之嚴格複本測驗(難度較一致)是會較隨機複本測驗(難度較不一致)於某固定精熟標準下，具有較佳的分類效果，但兩者差異甚小。此概念若相對於由精熟標準設定方法求得之精熟標準(係指由方法求得精熟標準，而非如上述，任意指定某一固定精熟標準)，在兩次異質測驗進行精熟/未精熟者分類時，是否代表著，愈異質測驗，分類效果愈差呢？但兩者差異亦不大呢？若參照本研究所探討之最大測驗訊息量法時，乃建立於 IRT 理論基礎下，具有因試題參數特徵以計算 θ 能力值的特性，在對照於異質測驗中產生之精熟標準，以進行精熟/未精熟者分類時，是否受影響程度亦會較小呢？本研究即在探討運用最大測驗訊息量法時，受測驗異質因素影響的程度，同時檢視是否具有改善接續將介紹之古典測驗理論中固定通過分數問題的特性。

(二)測驗異質性與轉換通過分數、定錨點題目篩選間關聯

在探討測驗異質性與轉換通過分數的關聯上，就換算古典測驗分數法與測驗特徵曲線構圖法而言，皆欲透過最大測驗訊息量法求得之 IRT 精熟標準 θ 能力值，加以轉換為古典測驗答對題數，因此，若將此分數作為兩次測驗之通過分數時，同樣具備古典測驗理論計分不適當之特性。對此，可進一步藉由表 2-9 來說明：假設現有受試者 100 人，分為 5 群，分別接受兩次測驗，其中，精熟標準設定為 0.80，在第一次測驗全部 60 題試題中，若分別在三種難易度上各 20 題，則如表 2-9 所示其分類結果將會有 60 名受試者(GroupA、B、C)被視為精熟者；但若在第二次測驗中，將高難度試題(Hard(一))刪除，改加上簡易試題(Easy(二))，則在同樣固定 0.80 精熟標準下，則會有 80 名受試者被視為精熟者，在比較兩次測驗結果時，會發現 GroupD 則成為分類不一致對象，即多出了 20 名受試者被錯誤分類。此即突顯出古典測驗理論下計分的不適當性(即答對較難試題與答對較易試題所給與之能力評等皆相同)，以致測驗異質性影響分類精熟/

未精熟者的現象。反觀國內考試院現行採用固定 60 分或 70 分作為精熟標準的決策下，是否亦會面臨因測驗異質性議題而產生錯誤分類的問題。

測驗異質性與篩選定錨點題目關聯上，如上述，在過去有關於定錨點之探討上，多以實務應用為主，顯少提及與其它議題之互動關聯，因此，本研究對此議題，乃同以探索角度出發，欲深究測驗異質性對於運用定錨點以挑選合乎準則且具內容代表性的範例試題之影響程度。

表 2-9 兩次測驗試題難易度與作答表現

Group	Sample size	Item difficulty				Domain score	
		Easy	Medium	Hard(一)	Easy(二)	(一)	(二)
A	20	20	20	20	20	1.00	1.00
B	20	20	20	20	20	1.00	1.00
C	20	20	20	20	20	1.00	1.00
D	20	20	20	0	20	0.67	1.00
E	20	20	0	0	20	0.33	0.67
Item p-value		1.00	0.80	0.60	1.00		

註：修改自 Hambleton & de Gruijter (1983, p.357)。