第四章 結果與討論

本研究旨在尋找出能有效融合精熟標準與測驗建構流程的設定方法，乃採用 2002 年國中基本學力測驗自然科的兩次測驗資料，事先於第三章研究方法中以文獻探討方式佐證其符合測驗編製流程中的先前假設，而後輔以實徵分析以提供方法可行性的證據。本章則旨在探討實徵分析的結果，在此共分為三節，第一節分析 2002 年國中基測的精熟標準設定結果，並檢測換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益，同時採行定錨點以描述精熟/未精熟者差異能力；第二節探究測驗長度因素與最大測驗訊息量法、換算古典測驗分數法、測驗特徵曲線構圖法與定錨點間的互動效果；第三節則探究測驗異質性因素對最大測驗訊息量法、換算古典測驗分數法、測驗特徵曲線構圖法與定錨點應用上可能產生的影響。另，本研究為簡化表格形式，乃以英文代號或中文縮寫呈現，在此預先將各自相對應含意做簡述，以利讀者釐清各表格意義。

表 4-1 本章英文代號與中文縮寫含意對照表

TEST1 / TEST2	2002 年國中基測第一次測驗 / 2002 年國中基測第二次測驗
$\theta_{\max}$ / $\theta^{i}_{\max}$	最大測驗訊息量法求得之精熟標準 / 最大試題訊息量對應之能力值
IRT 測驗難度	代表 IRT 求得之平均測驗難度值
古典偏態 / IRT 偏態	由受試者古典測驗分數(答對試題數) / IRT 能力值求得之偏態係數
TCTS	由換算古典測驗分數法求得之相對應通過分數
TCCM	由測驗特徵曲線構圖法求得之相對應通過分數
TEST1-TEST2	代表以第一次測驗求得精熟標準，而後以第二次測驗作為效標，驗證精熟/未精熟者之分類結果。
TEST2-TEST1	代表以第二次測驗求得精熟標準，而後以第一次測驗作為效標，驗證精熟/未精熟者之分類結果。
TEST1-TEST1	於驗證轉換效益時，在相同第一次測驗下，以不同方法轉換之結果比較
TEST2-TEST2	於驗證轉換效益時，在相同第二次測驗下，以不同方法轉換之結果比較
10(TEST1)20(TEST1) 30(TEST1)40(TEST1) 50(TEST1)	在測驗長度因素議題操弄中，代表第一次測驗中長度分別為 10 題、20 題、30 題、40 題與 50 題的測驗
10(TEST2) 20(TEST2) 30(TEST2) 40(TEST2) 50(TEST2)	在測驗長度因素議題操弄中，代表第二次測驗中長度分別為 10 題、20 題、30 題、40 題與 50 題的測驗
Easy(TEST1) Hard(TEST1)	在測驗異質性因素操弄中，代表第一次測驗中分屬於簡易與困難的測驗
Easy(TEST2) Hard(TEST2)	在測驗異質性因素操弄中，代表第二次測驗中分屬於簡易與困難的測驗
Normal((20)TEST1) Normal((20)TEST2)	在測驗異質性因素操弄中，乃以本研究中測驗長度為 20 題之測驗作為簡易與困難測驗的對照，在此稱為常態測驗

第一節 2002 年國中基測精熟標準設定結果之分析

本節乃分三大部份，首要分析以最大測驗訊息量法求得之精熟標準設定結果，其次探討換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益，最後，運用定錨點以解釋精熟/未精熟者間的差異能力，茲將結果陳述如下：

壹、精熟標準設定結果分析

以下茲就精熟標準設定方法求得之精熟標準與其相對應分類一致性效果分析之：

一、精熟標準設定方法求得之精熟標準結果

表 4-2 呈現的是 2002 年自然科兩次測驗(TEST1、TEST2)中，由最大測驗訊息量法求得之對應能力值$\theta_{\max}$、及根據換算古典測驗分數法(TCTS)與測驗特徵曲線構圖法(TCCM)將此$\theta_{\max}$轉換為古典測驗答對題數的結果。在第一次測驗上，$\theta_{\max}$值為 0.2500，在本研究中即視為第一次最大測驗訊息量法的精熟標準，而第二次測驗求得之精熟標準同樣為 0.2500。這表示若採以第一次測驗的精熟標準進行分類時，受試者在測驗上的能力值大於或等於 0.2500 者，即視為精熟者，反之則視為未精熟者；另一方面，IRT 精熟標準$\theta_{\max}$能力值在經轉換後(詳細轉換結果請見附錄二)，恰不能等於某一古典測驗分數，只是介於某二古典測驗分數間，因此使得 TEST1 與 TEST2 的$\theta_{\max}$值所對應的分數分別介於 36(IRT 能力值：0.22762)與 37(IRT 能力值：0.36162)分、35(IRT 能力值：0.05662)與 36(IRT 能力值：0.25336)分之間，這表示若採用換算古典測驗分數法於第一次測驗中求得之通過分數為分類準則時，則受試者在測驗上的總分若大於或等於 37 分者，即視為精熟者，反之則視為未精熟者。同理，於 TEST1、TEST2 中，經由測驗特徵曲線構圖法求得之相對應古典測驗分數分別為 37.8929、38.4678，相對於精熟分類中，鑑於古典測驗答對題數中並未有小數值，因而，如 37.8929 分亦等同於通過分數是介於 37-38 分間，即是受試的測驗總分若大於或等於 38 分者，即視為精熟者，反之則視為未精熟者。

表 4-2 自然科兩次測驗最大測驗訊息量對應能力值、各方法之轉換古典測驗分數與相關結果一覽表

測驗別	$\theta_{\max}$	IRT 測驗難度	古典偏態	IRT 偏態	TCTS	TCCM
TEST1	0.2500	0.0046	-0.023	-0.036	36-37	37.8929(37-38)
TEST2	0.2500	-0.1165	-0.239	-0.222	35-36	38.4678(38-39)

二、精熟標準設定方法分類一致性結果

本研究乃以最大測驗訊息量法求得精熟標準，接續以換算古典測驗分數法與

測驗特徵曲線構圖法求得相對應古典測驗答對題數，而後，採交叉驗證方式進行分類一致性的檢定。實際操作上，以 TEST1 計算出之精熟標準為兩次測驗(TEST1-TEST2)的決斷分數，再計算其百分比一致性與κ係數值；而第二次分類時，則改採以 TEST2 之精熟標準為兩次測驗的決斷分數，以求得各分類信度指標。

在盡量保持兩次測驗長度(58 題)一致、受試能力分配(如表 4-2 偏態值)、測驗難度(如表 4-2IRT 平均測驗難度值所示)一致下，分類結果如表 4-3 所示：在運用最大測驗訊息量法時，採用 TEST1 計算出的$\theta_{\max}$為兩次測驗的精熟標準，進行首次分類驗證時，其百分比一致性高達 0.9117(即 91.17%)，κ係數為 0.823；而反過來採用 TEST2 的$\theta_{\max}$為兩次測驗精熟標準，進行第二次分類驗證時，由於精熟標準即等同於首次分類的結果，因而，其分類百分比一致性同樣為 0.9117 (即 91.17%)，κ係數為 0.823。這顯示不論採用 TEST1 或 TEST2 的$\theta_{\max}$為精熟標準時，在 TEST2 或 TEST1 為分類效標下，其分類百分比一致性都有高達九成以上的水準，分類成效實屬良好；而在轉換為古典測驗答對題數的分類表現上，換算古典測驗分數法與測驗特徵曲線構圖法求得之百分比一致性與κ係數值，則與使用最大測驗訊息量法進行分類之結果十分雷同，都擁有不錯的分類精熟/未精熟者表現。

表 4-3 自然科兩次測驗各精熟標準設定方法之分類結果

精熟設定方法	精熟標準	比較測驗	百分比一致性	κ係數
最大測驗訊息量法	TEST1	TEST1-TEST2	0.9117	0.823
最大測驗訊息量法	TEST2	TEST2-TEST1	0.9117	0.823
TCTS	TEST1	TEST1-TEST2	0.9079	0.815
TCTS	TEST2	TEST2-TEST1	0.9060	0.812
TCCM	TEST1	TEST1-TEST2	0.9075	0.812
TCCM	TEST2	TEST2-TEST1	0.9058	0.806

在本研究中，最大測驗訊息量法於兩次測驗中，皆獲得相同$\theta_{\max}$(即 0.2500)值，對此，採交叉驗證時，不僅具有實務應用的公平性優勢(決策者當然期望兩次測驗有相同精熟標準，若不一致則易引發爭議)，且就理論而言，同時符合最大測驗訊息量法穩定估計此精熟標準附近受試者能力的特性(即隱含穩定分類精熟/未精熟者)。若以區間觀點視之，如圖 4-1 所示，測驗最大訊息量所對應之能力值雖為最佳之標準，但其鄰近區間亦算是具備不錯的穩定能力估計效果，因此，於精熟/未精熟分類上仍能具備水準以上之表現。若將此觀念，推導至最大測驗訊息量法的延伸應用中，本研究強調藉由此法所求得之精熟標準可作為實務參照之用途，專家學者們可根據對社會、經濟影響的考量，對此精熟標準基點上下加以調整，則調整後之精熟標準區間，以能力估計穩定面向而言，雖非最佳者，但以上述區間觀點思考之，仍應具備不錯之估計表現。若相較於參照其它方法所求得之基點，此法則具備由測驗透露出客觀訊息的優勢。

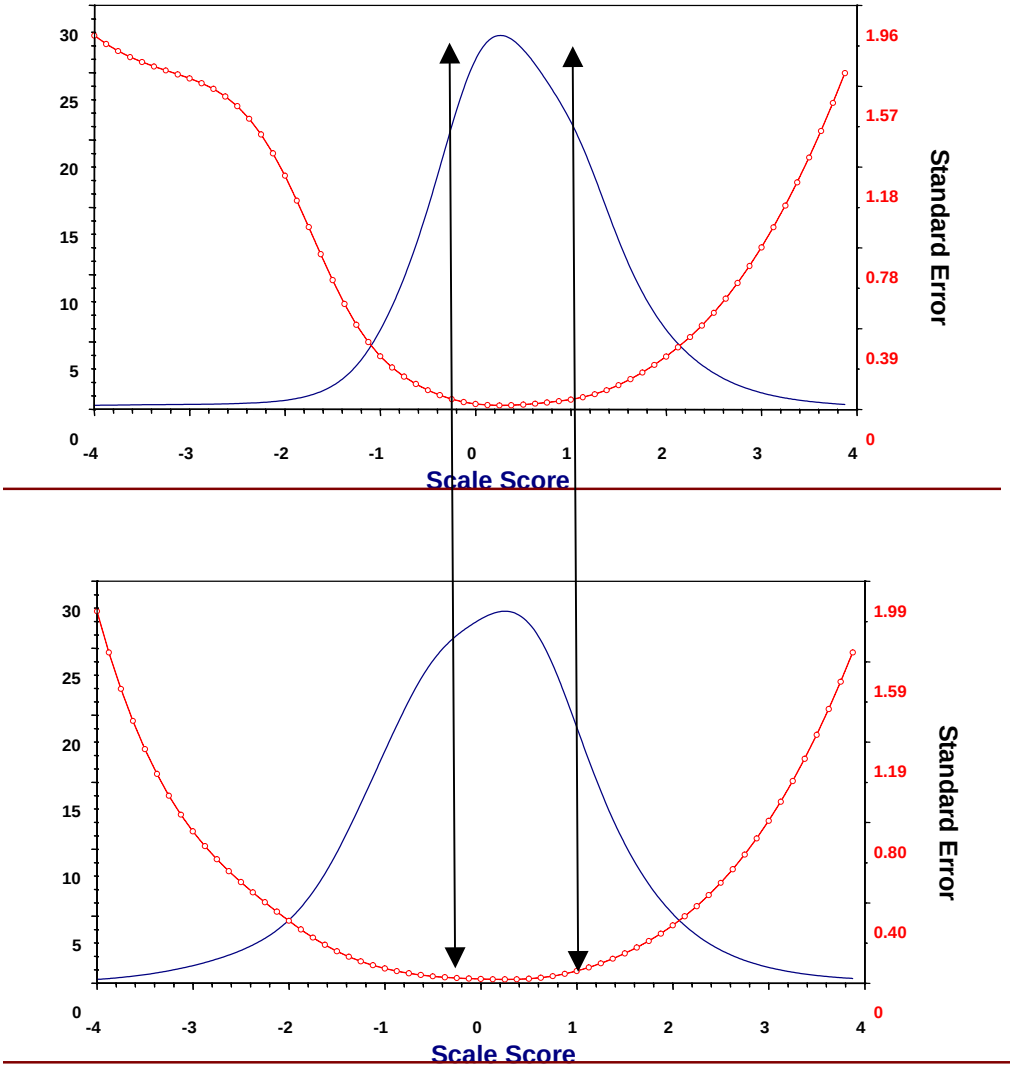


圖 4-1 自然科兩次測驗最大測驗訊息量圖

註：上圖為第一次測驗之結果；下圖為第二次測驗之結果；實線為訊息量之變化，採左側之刻度；圓點線為標準誤之變化，採右側之刻度。

貳、換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益分析

以下茲就換算古典測驗分數法與測驗特徵曲線構圖法之分類一致性效果與分類差異結果兩方面探討之，陳述如下：

一、換算古典測驗分數法與測驗特徵曲線構圖法之分類一致性效果分析

對於評估換算古典測驗分數法與測驗特徵曲線構圖法於轉換 IRT 精熟標準 θ 能力值與古典測驗答對題數間之效益時，乃以求轉換分數上之精熟/未精熟者分類的一致性效果為主，概念上如表 4-4 所示，顯示在同一份測驗下(都是 TEST1，以表示完全控制測驗、受試者表現)，約有 324 人(287+37)在此被分類錯

誤，其百分比一致性如表 4-5 所示，高達 0.9676 (96.76%)，κ 係數值為 0.935，而在 TEST2 上，亦有良好表現。同理，最大測驗訊息量法與測驗特徵曲線構圖法之間的分類一致性結果，如表 4-5 所示，在相同 TEST1 或 TEST2 下亦有高達 94.82% 與 93.29%的正確分類比率。但若從細部百分比一致性分析，在兩次比較中，皆顯示由換算古典測驗分數法轉換自最大測驗訊息量法求得之精熟標準是較測驗特徵曲線構圖法為佳(0.9676 大於 0.9482；0.9797 大於 0.9329)。

表 4-4 自然科第一次測驗於轉換分數上之分類一致性效果評估方式

TEST1		TCTS		合計
		精熟	未精熟	
最大測驗訊息量法	精熟	4586	287	4873
	未精熟	37	5090	5127
合計		4623	5377	N=10000

表 4-5 自然科兩次測驗各轉換方法間分類效果一覽表

精熟標準設定方法	比較測驗	錯誤分類數	百分比一致性	κ 係數
最大測驗訊息量法 vs．TCTS	TEST1-TEST1	324	0.9676	0.935
最大測驗訊息量法 vs．TCTS	TEST2-TEST2	203	0.9797	0.959
最大測驗訊息量法 vs．TCCM	TEST1-TEST1	518	0.9482	0.896
最大測驗訊息量法 vs．TCCM	TEST2-TEST2	671	0.9329	0.865

註：表 4-4 之(4586+5090)/10000 等於本表第一欄之百分比一致性 0.9676；而表 4-4 之 287+37 即為本表第一欄之錯誤分類數 324 人。

二、換算古典測驗分數法與測驗特徵曲線構圖法之分類差異分析

在比較換算古典測驗分數法與測驗特徵曲線構圖法間分類的差異上，如表 4-6 所示，在相同 TEST1 下，由於換算古典測驗分數法之轉換結果為 36-37 分，與測驗特徵曲線構圖法之轉換結果為 37-38 分(如表 4-2 所示)相較時，兩者相差 1 分，因而，使得兩者間會出現 250 名分類不一致人數(此即原始古典總分為 37 分的人數)。另一方面，在相同 TEST2 下，兩者轉換後之差異分數為 3 分(35-36 vs．38-39)，至使出現 660 名分類不一致人數(即原始古典總分為 36(197 人)、37(216 人)、38(247 人)。註：相關分數點人數可參照附錄二)，但整體而言，如表 4-7 所示，在 TEST1、TEST2 表現上兩者皆具頗高的分類一致性。

同理，若將此推論至上述最大測驗訊息量法與換算古典測驗分數法、測驗特

徵曲線構圖法之比較時，將由最大測驗訊息量法求得之θ_{max}想像為某一原始古典分數，與其它轉換方法相較時，則可發現其中分類不一致人數(即是百分比一致性的對照)多為彼此間轉換的差異分數，而當該差異分數點人數較多或具較大差異分數時，則百分比一致性多表現較差。相反的，若θ_{max}若能正確無誤的經由換算古典測驗分數法或測驗特徵曲線構圖法轉換至原始測驗分數上(即同一通過分數)，則在相同測驗、同一受試者下，錯誤分類人數應為 0。

進一步探討實質影響分數轉換間的差異分數大小與差異分數點人數多寡的因素。就前者而言，在換算古典測驗分數法下，可能因本研究乃採用受試者實際表現資料進行分析，某些反應組型並未呈現出來，或者出現某些極端的反應組型，而使得分數轉換間有些許落差；另一方面，對於測驗特徵曲線構圖法，乃因轉換後之 TCC-Y 軸值與古典測驗答對題數間的落差，而出現分類不一致現象。而就後者影響差異分數點人數多寡因素而言，在本研究接續探討之測驗長度與測驗異質性因素中，將有更深入分析，故不在此加以描述。

但綜觀此兩種方法，根據資料分析結果顯示，其誤差皆非常小，最大不超過 7%，這顯示最大測驗訊息量法於精熟標準設定上，不論是搭配運用換算古典測驗分數法或者測驗特徵曲線構圖法皆是值得參照的方法。

表 4-6 自然科第一次測驗於轉換分數方法間分類一致性差異

TEST1		TCCM		合計
		精熟	未精熟	
TCTS	精熟	4373	250	4623
	未精熟	0	5377	5373
合計		4373	5627	N=10000

表 4-7 自然科兩次測驗轉換方法間分類差異效果一覽表

精熟標準設定方法	比較測驗	錯誤分類數	百分比一致性	κ係數
TCTS vs. TCCM	TEST1-TEST1	250	0.9750	0.950
TCTS vs. TCCM	TEST2-TEST2	660	0.9340	0.867

註：表 4-6 之(4373+5377)/10000 等於本表第一欄之百分比一致性 0.9750；而表 4-6 之 250+0 即為本表第一欄之錯誤分類數 250 人

參、精熟/未精熟者間差異能力分析

對於第三章研究方法中曾討論到有關運用換算古典測驗分數法與測驗特徵曲線構圖法轉換效益上之解釋的效果(以古典測驗分數作為解釋時所收效益)，對此，乃僅限於提供如某次資格考試，及格標準為“60 分“之訊息，而並未能更深入剖析精熟/未精熟者間真實差異能力，於廣泛解釋上實屬有限。對此，定錨

點則能發揮其功效，茲就以下幾方面探討之：

一、定錨點的選擇

在定錨點的選擇上，為有效針對精熟/未精熟者間差異能力作廣泛描述，本研究乃採用 4 個定錨點，分別以 2 個定錨點解釋精熟/未精熟者的能力，分析結果如表 4-8 所示，在 TEST1 中乃以 1.7296 與 0.7432 分別代表精熟者的定錨能力值，而以-0.2432 與-1.2296 代表未精熟者的定錨能力值，對此，本研究為方便後續分析，乃由左而右分別將之稱為第 1、2、3 與第 4 定錨點。同理，在 TEST2 中，由於標準差的不同使得雖然兩次測驗雖擁有相同精熟標準，但定錨點仍會有些微差異，而此對於篩選定錨點題目上應不致於產生太大的影響。

表 4-8 自然科兩次測驗定錨點一覽表

測驗別	$\theta_{\max}$	標準差	精熟定錨點		未精熟定錨點	
TEST1	0.2500	0.9864	1.7296	0.7432	-0.2432	-1.2296
TEST2	0.2500	1.0093	1.76395	0.75465	-0.25105	-1.26035

二、定錨點題目的選擇

藉由定錨點，可求得相對於每一試題之答對機率值，接續，再採納 Huynh(1998)建議之三個準則：一、位於第i個定錨點上受試者正確反應機率值大於或等於.75；二、位於下一個定錨點受試者正確反應機率小於.56；三、兩定錨點間試題正確反應機率差異大於或等於.19。進行定錨點題目之篩選。表 4-9 為第一次測驗下採用 TEST1 定錨點求得之定錨點題目與其相對之正確反應機率值，在此，以第 5 題為例，其於定錨點-0.2432 下，正確反應機率值為.82453，大於.75 標準，且其下一定錨點-1.2296 其正確反應機率值為.52896，小於.56，且兩者差距達.29557，亦符合.19 準則，因此，得以做為在該定錨點(-0.2432)下進一步篩選具內容代表性的候選試題。同理，其餘測驗完整之定錨點篩選結果請詳見附錄三。

表 4-9 第一次測驗下根據 TEST1 定錨點挑選之定錨點題目與其相對之正確反應機率值

	定錨點			
	1	2	3	4
定錨點題目	(1.7296)	(0.7432)	(-0.2432)	(-1.2296)
1	.99484	.98750	.97013	.93094**
2	.97572	.95471	.91755	.85599**
5	.99715	.97444	.82453**	.52896
7	.99419	.95558	.76532**	.50891
8	.99254	.95298	.78540**	.54849
9	.99968	.99351	.88893**	.45880

12	.99142	.94551	.75156**	.48339
13	.97395	.82003**	.41282	.20053
16	.95290	.78353**	.48335	.33419
22	.98222	.88057**	.55176	.30702
23	.95115	.79568**	.49361	.29951
24	.97304	.80304**	.43014	.29077
25	.96087	.75945**	.40780	.28765
29	.98207	.86049**	.49373	.29593
31	.98534	.82869**	.39202	.26626
32	.99584	.90992**	.35820	.13894
35	.97940	.82860**	.36396	.12814
36	.99564	.93306**	.50468	.19323
37	.97906	.77561**	.28625	.15640
43	.98537	.85342**	.41084	.21291
45	.90337**	.44243	.25815	.24516
46	.99279	.98391	.96459	.92448**
48	.82998**	.55559	.28670	.16931
49	.92466**	.33450	.17712	.17234
51	.86585**	.43750	.15715	.11099
52	.97279	.78659**	.36006	.20534
55	.78072**	.40633	.21448	.17641

註 1：**代表該試題符合該定錨點之篩選準則，乃為區別於測驗長度議題中以*代表另一種寬鬆篩選準則。

2：左側之定錨點題目為相對於原始測驗之題號，而未達標準者，為簡化呈現方式，故不列入表中，完整資料請見附錄 3。

3、各定錨點由左而右分別為第 1 至第 4 定錨點，而括號內為其對應能力值。

表 4-10 與表 4-11 為自然科兩次測驗各定錨點之試題篩選結果。表 4-10 代表著在 TEST1 與 TEST2 中，以 TEST1 求得之定錨點進行試題篩選之結果，例如，在 TEST1 第 1 定錨點 1.7296 下，TEST1 中有 5 題試題(原始題號為 45、48、49、51、55)符合準則，而 TEST2 為效標下，則有 7 題符合準則；同理，表 4-11 則為以 TEST2 定錨點下，TEST2 與其效標 TEST1 之定錨點題目篩選結果。相較於兩次篩選結果，如上所述，雖然兩次測驗之標準差有些微差異，但篩選出之定錨點題目皆相仿，例如在 TEST1 下，採用 TEST1 與 TEST2 之第 1 定錨點結果，皆為 45、48、49、51、55(總題數 5 題)。同理，於其它測驗結果上，僅只於第 2 與第 3 定錨點上有些許差異，但此對於往後選擇具內容代表性試題時，應不致於會產生太大的影響，因而，在進一步描述精熟/未精熟差異能力時，在本研究則不刻意區分是以 TEST1 或 TEST2 求得之定錨點下的結果，而是全部皆加以採納（如表

4-11 在 TEST1 下由第 2 定錨點求得之 14、38、50 試題，並未於表 4-10 的 TEST1 之第 2 定錨點題目中出現，但描述上仍將其列入)，以描述 TEST1 精熟者/未精熟者間差異能力為何？而在 TEST2 精熟/未精熟者差異能力亦為何？探討兩者是否具備相仿的精熟/未精熟能力？為核心。

另，由表 4-10、表 4-11 可發現於極端定錨點上(如第 4 定錨點)相較於精熟標準附近(如第 2 定錨點)上，求得之總定錨點題目數，具有某程度的差異，是否存在某種影響因素？對此，於測驗長度與測驗異質性時，將有更深入剖析。

表 4-10 TEST1 定錨點下自然科兩次測驗定錨點題目篩選結果

成就水平 / 測驗	精熟		未精熟	
定錨點	1(1.7296)	2(0.7432)	3(-0.2432)	4(-1.2296)
TEST1	45、48、49、51、55	13、16、22、23、24、25、29、31、32、35、36、37、43、52	5、7、8、9、12	1、2、46
總定錨點題目數	5	14	5	3
TEST2	39、41、46、48、49、50、56	4、13、15、18、19、23、24、25、26、28、31、37、40、45、52、53、55、58	3、5、6、8、11、35	1
總定錨點題目數	7	18	6	1

註：定錨點由左而右為第 1 至第 4 定錨點，而括號內為其對應能力值。

表 4-11 TEST2 定錨點下自然科兩次測驗定錨點題目篩選結果

成就水平 / 測驗	精熟		未精熟	
定錨點	1(1.76395)	2(0.75465)	3(-0.25105)	4(-1.26035)
TEST1	45、48、49、51、55	13、14、16、22、23、24、25、29、31、32、35、36、37、38、43、50、52	5、7、8、9	1、2、46
總定錨點題目數	5	17	4	3
TEST2	39、41、46、48、49、50、56	4、13、15、18、19、23、24、25、26、28、31、37、40、42、45、52、53、55、58	3、5、6、8、11、35	1
總定錨點題目數	7	19	6	1

註：定錨點由左而右為第 1 至第 4 定錨點，而括號內為其對應能力值。

三、精熟/未精熟者間差異能力描述

藉由上述初步的篩選，得以選定各測驗下適當之定錨點題目，而後，則需就此進一步挑選出足以描述該定錨點能力之內容代表性試題，並加以突顯精熟/未精熟者間差異能力。對此，如 Koretz & Deibert (1995)所認為，對於描述受試者於該定錨內容領域上之表現，可以呈現其所具備知識與技能為核心，參照此概念，如表 4-12 所示，首先，統整出各定錨點題目下，相對檢測之知識與技能。在此，如以第一次測驗第 1 題為例，該試題題目為：

1．有關使用瓦斯的安全問題，下列敘述何者正確？

(A)瓦斯熱水器置於室內較為安全

(B)洗澡水溫度要適中以防瓦斯外洩

(C)瓦斯燃燒不完全會導致二氧化碳中毒

(D)定期檢查瓦斯開關與接頭，以策安全

該試題乃屬於第 4 定錨點，以測量受試者是否具備日常正確使用瓦斯的安全性知識(註：為呈現由基本往進階的概念，因此，呈現上乃是依續由第 4 定錨點(低能力)往第 1 定錨點(高能力)描述)，如此，第 2、46 題亦屬第 4 定錨點的範疇，加以綜整後，可歸納出足以代表該定錨點能力之概念，乃顯示該定錨點受試者僅具備生活中基本的安全與救護知識。同理，對於第 3 定錨點受試者而言，則不僅具備日常生活的安全知識，更需認識：1、學科中基礎生物器官特徵與物理現象知識(如第 7 題認識人體的生殖器官及功能；第 9 題認識細胞核的構造與功能等)；2、理解簡單的圖示意義(如第 8 題光於各介質中不同折射角度圖)。

表 4-12 自然科第一次測驗未精熟者能力描述

<table>
<tr><td colspan="4">第 4 定錨點</td></tr>
<tr><td>定錨點題目</td><td>試題能力指標</td><td>定錨點能力指標</td><td>未精熟者能力指標</td></tr>
<tr><td>1</td><td>具備日常正確使用瓦斯的安全性知識</td><td rowspan="3">生活中基本的安全與救護知識</td><td rowspan="11">1、具備學科基礎知識

2、簡易圖示理解</td></tr>
<tr><td>2</td><td>具備處理骨折情境的救護策略</td></tr>
<tr><td>46</td><td>具備本土腸病毒的相關資訊</td></tr>
<tr><td colspan="3">第 3 定錨點</td></tr>
<tr><td>題數</td><td>試題能力指標</td><td>定錨點能力指標</td></tr>
<tr><td>5</td><td>認識生活中常見肥皂組成與應用</td><td rowspan="5">1、具備學科中基礎生物器官特徵與物理現象知識
2、理解簡單的圖示意義</td></tr>
<tr><td>7</td><td>認識人體的生殖器官及功能</td></tr>
<tr><td>8</td><td>認識光的於不同介質中的特性</td></tr>
<tr><td>9</td><td>認識植物中細胞核的構造與功能</td></tr>
<tr><td>12</td><td>認識淨水器中活性炭的功能</td></tr>
</table>

註：各定錨點題目原始試題內容請參見附錄四。

但具備學科基礎知識與簡易圖示理解能力，仍屬於未精熟者的能力範圍。對於精熟者，仍需進一步具備如表 4-13 所示之知識與技能，擁有如第 2 定錨點受試者對於廣泛學科知識的了解(如第 16 題具備現行生物分類系統知識；第 22 題認識聲音中振幅、頻率、速率、波長的特性等)、複雜問題、資料與圖表詮釋(如第 38 題認識天氣圖及其表現的天氣現象)、邏輯推理與分析實驗結果以獲得相關論點(如第 22 題認識以加熱是否易變焦黑的方法分辨有機物質與無機物質的實驗)等能力，或者更進階如第 1 定錨點受試者般具備進階學科知識與綜合、評鑑資料、情境傳遞之訊息。

表 4-13 自然科第一次測驗精熟者能力描述

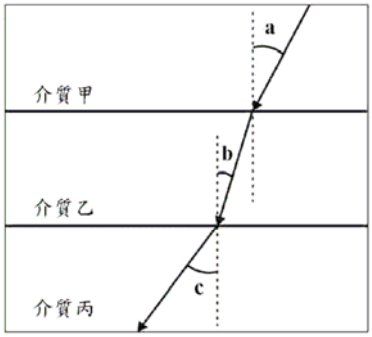
第 2 定錨點			
定錨點題目	試題能力指標	定錨點能力指標	精熟能力
13	認識水的密度與溫度的關聯	1、具備廣泛學科知識，並了解彼此關聯性 2、具備複雜問題、資料與圖表詮釋能力 3、應用學科知識以邏輯推理出適當答案 4、具備分析實驗結果的能力，並以此獲得研判的論點	1、具備廣泛且進階的學科知識 2、合理邏輯推導能力 3、綜合、評鑑資訊以獲得正確結論 4、具備複雜圖表、實驗的分析能力
14	具備原子相關知識，並依此推測屬性間關係		
16	具備現行生物分類系統知識		
22	認識聲音中振幅、頻率、速率、波長的特性		
23	認識大氣的重要組成成分與演變過程		
24	認識力的作用與傳動現象，以推導木塊重量		
25	認識以加熱是否易變焦黑的方法分辨有機物質與無機物質		
29	認識以適當的方式登錄長度測量結果		
31	認識生物遺傳的特性，並以此推導染色體數量		
32	認識基因可控制性狀的遺傳特性，並以此推測適點的結論		
35	認識平衡的化學反應中各化學式的關係		
36	認識植物細胞中遺傳基因的特性		
37	認識溫室效應對地球環境的影響		
38	認識天氣圖及其表現的天氣現象		

43	由實驗的結果，推測浮力與液體體積關係		
50	知道指北針的磁針與磁場(導線通以電流)交互作用的方式		
52	認識氫的物理與化學性質，並以此獲得研判的論點		
第 1 定錨點			
題數	試題能力指標	定錨點能力指標	
45	理解演化規則，以推論不同的蝗蟲體色的成因	1、具備進階學科知識 2、理解資料、情境傳來之訊息，綜合後以形成適當概念	
48	由實驗數據推論溫度與時間變化關係		
49	理解電功率與歐姆定律，運用附圖及所提供的公式作推理的能力		
51	認識各營養成份所含熱量以推測結論		
55	理解圖形及符號之間的關係，能以適當化學式正確表達化合物名稱		

對此，進一步的舉例詳細描述精熟/未精熟者之差異能力。採以第 2 定錨點之第 8 題與第 3 定錨點之第 50 題分別代表未精熟者與精熟者之範例試題，各自試題內容為：

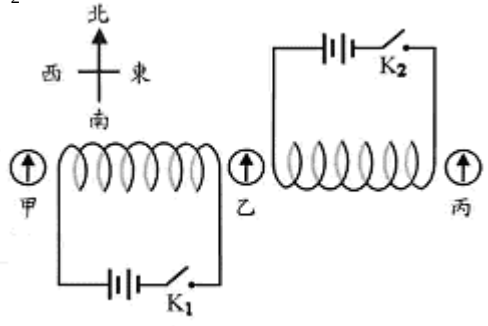
8．如圖，光線經過甲、乙、丙三層介質時發生折射，且角度 c>a>b，則光線在三介質中的速率大小關係，下列何者正確？

(A)甲>乙>丙
(B)甲>丙>乙
(C)丙>甲>乙
(D)丙>乙>甲



50．將二個完全相同的線圈放在桌面上，另有甲、乙、丙三羅盤，乙羅盤在兩線圈的正中間，如圖(二十二)。當開關 K_1、K_2 按下接通電流後，下列何者正確？

(A)甲羅盤磁針的 N 極向東偏轉
(B)乙羅盤磁針的 N 極向西偏轉
(C)丙羅盤磁針的 N 極向東偏轉
(D)乙羅盤所在位置的磁場最強



就第 8 題而言，受試者僅需認識光於不同介質中折射角度愈大，速率愈大的特性(基礎學科知識)，即可輕易回答問題，而圖示乃僅重複題意以作為輔助用途，並非屬複雜，因而，隸屬未精熟者能力範疇；而就第 50 題而言，受試者不僅需了解指北針的偏轉是磁針與磁場(地磁或導線通以電流)交互作用的結果，更應清楚知道其作用的方式(廣泛學科知識)，且受試者更需加以詮釋圖示，由甲、乙、丙不同羅盤的位置推測出正確結論，因而，隸屬精熟者能力範疇。

同理，第二次測驗精熟/未精熟者間之差異能力，分析結果如附錄五所示，與此所強調之具備廣泛且進階的學科知識、合理邏輯推導能力、綜合、評鑑資訊以獲得正確結論、具備複雜圖表、實驗的分析能力等結果一致。綜合上述，對照於國民中學學生基本學力測驗推動工作委員會(2002c)所認為，基本學力測驗著重評量學生在知識、理解、應用、分析、綜合、評鑑等層次的能力。學生除了需要具備各學科的基礎知識外，尚需應用這些知識來詮釋問題、資料和圖表，甚至進一步地應用學科知識及邏輯推理來整理出問題的答案。相對於本研究結論，似乎印證前者乃隸屬未精熟者能力，而若欲達精熟狀態，則需具備後者邏輯推理與國民中學學生基本學力測驗推動工作委員會(2002b)表示之高層次思考及統整學科知識等能力。

第二節 精熟標準設定方法間測驗長度因素結果之分析

本節主要在探討測驗長度因素與最大測驗訊息量法、換算古典測驗分數法、測驗特徵曲線構圖法、定錨點間的互動效果，對此，可從三方面探討之。首先，分析在不同測驗長度下，最大測驗訊息量法、換算古典測驗分數法與測驗特徵曲線構圖法求得之精熟標準及其一致性分類效果；其次，探討測驗長度因素對於換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益可能的影響層面；最後，欲確立測驗長度因素如何影響定錨點題目之篩選。茲將結果陳述如下：

壹、測驗長度因素下精熟標準設定結果分析

以下茲就不同測驗長度類型下精熟標準設定方法求得之精熟標準與其相對應分類一致性效果分析之，陳述如下：

一、不同測驗長度類型下精熟標準設定方法求得之精熟標準結果

有關本節探討測驗長度對分類結果的影響，如第三章研究方法中實施流程與研究設計一節所述，期望在控制測驗異質性、受試者能力分配因素下，分別以10、20、30、40、50 題試題，代表五種不同的測驗長度類型(所篩選之試題如附錄六所示)，以探究不同測驗長度與精熟標準設定方法間可能的變化情形。表 4-14 顯示在盡量控制無關因素下(各測驗長度類型最大平均難度差異值為 30 題的 0.0973；不論古典測驗分數或 IRT 能力值的偏態係數，其最大偏態差異在 IRT 能力值為 10 題的 0.289)，不同測驗長度類型，各$\theta_{\max}$值、換算古典測驗分數法與測驗特徵曲線構圖法之分析結果，如第一次測驗長度為 10 題時，求得之$\theta_{\max}$為 0.1250，相對於由換算古典測驗分數法與測驗特徵曲線構圖法求得之轉換分數分別為 6-7 與 6.2262(即相當 6-7 分)(各長度類型下詳細轉換結果請見附錄七)，在此，即以此作為各長度類型下之精熟標準。

表 4-14 自然科兩次測驗不同測驗長度類型下最大測驗訊息量對應能力值、各方法之轉換古典測驗答對題數與相關結果一覽表

測驗長度	$\theta_{\max}$	IRT 測驗難度	古典偏態	IRT 偏態	TCTS	TCCM
10(TEST1)	0.1250	-0.1063	-0.114	0.214	6-7	6.2262(6-7)
10(TEST2)	0.0000	-0.1276	-0.255	-0.075	5-6	5.8346(5-6)
20(TEST1)	0.2500	-0.0432	-0.070	0.085	13-14	13.1039(13-14)
20(TEST2)	0.3750	-0.0234	-0.138	0.007	13-14	13.4266(13-14)
30(TEST1)	0.2500	-0.1243	-0.140	0.003	19-20	20.2280(20-21)
30(TEST2)	0.5000	-0.0270	-0.191	-0.084	22-23	21.2503(21-22)
40(TEST1)	0.2500	-0.1135	-0.126	-0.026	26-27	27.1760(27-28)
40(TEST2)	0.0000	-0.1975	-0.392	-0.239	24-25	24.9030(24-25)

50(TEST1)	0.3750	-0.0121	0.012	-0.006	32-33	34.0707(34-35)
50(TEST2)	0.2500	-0.0587	-0.162	-0.152	30-31	32.4440(32-33)

在本章第一節原始測驗長度下(58 題)，如表 4-2 所示，曾提到最大測驗訊息量法於兩次測驗中，皆獲得相同$\theta_{\max}$ (0.2500)，對此，採交叉驗證時，不僅具有實務應用的公平性優勢，且就理論而言，同時符合最大測驗訊息量法穩定估計此精熟標準附近受試者能力的特性(即隱含穩定分類精熟/未精熟者)。相對於此，如表 4-14 所示，在不同測驗長度類型議題研究中，兩次測驗所求得之$\theta_{\max}$，皆有微幅的差距，若以差異最大的測驗長度 30 與 40 題為例，兩者相差皆達 0.2500 (0.2500 vs. 0.5000；0.2500 vs. 0.0000)。在為確保實務應用公平性下，仍以第一次測驗求得之精熟標準為兩次測驗之分類準則時，雖然就第一次測驗而言，該精熟標準為該測驗最適合之能力值，但就第二次測驗而言，該精熟標準並非是最佳的。

在此，以測驗長度 20 題為例，詳細說明之。在進行首次精熟/未精熟者分類時，若採用 20(TEST1)之$\theta_{\max}$ 0.2500 為 20(TEST1)與 20(TEST2)之分類準則時，此精熟標準對於 20(TEST1)而言的確是最佳的分類準則，但對於 20(TEST2)而言，並非是最佳的分類準則(該測驗最佳的分類準則應是 0.3750)。同理，交叉驗證的另一面向，以 20(TEST2)求得之精熟標準亦非是 20(TEST1)的最佳精熟標準能力值，因而，理論上而言，會使得分類上出現較不一致情況。雖是如此，同樣以先前曾討論之區間觀點視之，測驗最大訊息量所對應之能力值雖為最佳之標準，但其鄰近區間亦是具備不錯穩定能力估計效果，因而，於精熟/未精熟分類上理應仍能具備水準以上之表現。而詳細結果，接續分析如下。

二、不同測驗長度類型下精熟標準設定方法分類一致性效果

表 4-15 呈現在不同測驗長度類型下，自然科兩次測驗最大測驗訊息量法分類結果，整體而言，皆可約高達 8 成附近(或以上)之百分比一致性。而就不同測驗長度間比較上，可明顯發現隨著測驗長度增加，百分比一致性信度值也隨之增加，這與 Hambleton(1983)、Hambleton, Mills, & Simon(1983)等人以事先設定精熟標準方式下分析的結果相仿。其中當測驗長度由 10 題增加至 20 題時，信度值增加幅度較大，以採第一次測驗精熟標準為例，百分比一致性值由 0.8001(即 80.01%)增至 0.8631(即 86.31%)，增幅達約 6.3%，錯誤分類人數減少了 630 人(1999-1369)，接續再增加測驗長度，分類信度增幅就較無如此明顯現象。而在換算古典測驗分數法與測驗特徵曲線構圖法下，其分類結果分別如表 4-16 與表 4-17 所示，亦有同樣的類似現象產生，測驗長度增加亦帶動分類一致性信度值上昇，測驗長度由 10 題增至 20 題時，信度值亦有較大增幅。

另一方面，從縱觀的角度出發，由最大測驗訊息量法求得之百分比一致性信度值，不論在何種測驗長度類型下，皆一致較換算古典測驗分數法與測驗特徵曲線構圖法為佳，但是差距皆非常微小，最大差異出現於測驗長度 10 題時，以 TEST2

為精熟標準之 3.03% (79.98% vs. 76.95%、76.95%)。此外，當測驗長度增加時，差距更是微小，至長度 40 題時，差異已縮小至 1%內，就此，顯示出測驗長度能有效弭平分數轉換間的不一致效果，但是增進幅度有限，並不足以稱之有顯著效果。

綜合上述，研究者若對分類精確性的要求是很一般化(約 80%分類一致性)，且在考量金錢、時間成本效益下，希望測驗的題數能儘量少，則建議於測驗長度 20 題時，會是必備的基本值。

表 4-15　自然科兩次測驗不同測驗長度類型下最大測驗訊息量法分類結果

測驗長度	精熟標準	比較測驗	錯誤分類數	百分比一致性	κ 係數
10	TEST1	TEST1-TEST2	1999	0.8001	0.596
10	TEST2	TEST2-TEST1	2002	0.7998	0.600
20	TEST1	TEST1-TEST2	1369	0.8631	0.723
20	TEST2	TEST2-TEST1	1373	0.8627	0.713
30	TEST1	TEST1-TEST2	1220	0.8780	0.745
30	TEST2	TEST2-TEST1	1192	0.8808	0.697
40	TEST1	TEST1-TEST2	1078	0.8922	0.784
40	TEST2	TEST2-TEST1	1085	0.8915	0.783
50	TEST1	TEST1-TEST2	931	0.9069	0.812
50	TEST2	TEST2-TEST1	951	0.9049	0.809

表 4-16　自然科兩次測驗不同測驗長度類型下換算古典測驗分數法分類結果

測驗長度	精熟標準	比較測驗	錯誤分類數	百分比一致性	κ 係數
10	TEST1	TEST1-TEST2	2116	0.7884	0.570
10	TEST2	TEST2-TEST1	2305	0.7695	0.530
20	TEST1	TEST1-TEST2	1470	0.8530	0.690
20	TEST2	TEST2-TEST1	1470	0.8530	0.690
30	TEST1	TEST1-TEST2	1396	0.8604	0.717
30	TEST2	TEST2-TEST1	1300	0.8700	0.693
40	TEST1	TEST1-TEST2	1157	0.8843	0.767
40	TEST2	TEST2-TEST1	1174	0.8826	0.765
50	TEST1	TEST1-TEST2	983	0.9017	0.798
50	TEST2	TEST2-TEST1	988	0.9012	0.802

表 4-17　自然科兩次測驗不同測驗長度類型下測驗特徵曲線構圖法分類結果

測驗長度	精熟標準	比較測驗	錯誤分類數	百分比一致性	κ係數
10	TEST1	TEST1-TEST2	2116	0.7884	0.570
10	TEST2	TEST2-TEST1	2305	0.7695	0.530
20	TEST1	TEST1-TEST2	1470	0.8530	0.690
20	TEST2	TEST2-TEST1	1470	0.8530	0.690
30	TEST1	TEST1-TEST2	1366	0.8634	0.714
30	TEST2	TEST2-TEST1	1333	0.8667	0.707
40	TEST1	TEST1-TEST2	1150	0.8850	0.764
40	TEST2	TEST2-TEST1	1174	0.8826	0.765
50	TEST1	TEST1-TEST2	969	0.9031	0.792
50	TEST2	TEST2-TEST1	983	0.9017	0.798

貳、測驗長度因素下換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益分析

以下茲就不同測驗長度類型下，換算古典測驗分數法與測驗特徵曲線構圖法之轉換一致性效果與轉換差異兩方面探討之，陳述如下：

一、不同測驗長度類型下換算古典測驗分數法與測驗特徵曲線構圖法之轉換一致性效果分析

對於測驗長度因素與換算古典測驗分數法、測驗特徵曲線構圖法間之互動影響上，在同樣完全控制測驗、受試者表現時，探討不同測驗長度類型下，轉換分數上之精熟/未精熟者分類的一致性效果。資料分析結果如表 4-18 所示，隨著測驗長度的增加，由換算古典測驗分數法轉換自最大測驗訊息量法求得之精熟標準 $\theta_{\max}$ 時，百分比一致性信度值增加幅度皆非常的微小，於測驗長度 30 題時，即有 97.10%、97.26%的百分比一致性，相對之錯誤分類數亦僅有 290 與 274 人，而後隨測驗長度增加，受此影響則漸不明顯，並未能發現有效減低錯誤分類數。同理，如表 4-19 所示，由測驗特徵曲線構圖法轉換自最大測驗訊息量法求得之精熟標準 $\theta_{\max}$ 時，受測驗長度影響依舊不明顯。

對此，若從較細微差異角度來比較換算古典測驗分數法與測驗特徵曲線構圖法轉換之結果時，顯示不論在何種測驗長度下，表 4-18 之換算古典測驗分數法皆擁有較佳的百分比一致性表現，尤其於 30-50 題時，錯誤分類數皆可控制於 400 人(4%)以下，相較於表 4-19 測驗特徵曲線構圖法，則有少部份錯誤分類數仍大於 500 人(5%)，但差異仍屬微小。整體而言，兩者皆可達九成以上的一致性水準，顯示不論在何種測驗長度下，採用何種轉換方式，其轉換一致性仍具一定的水平。

表 4-18 自然科兩次測驗不同測驗長度類型下最大測驗訊息量法與換算古典測驗分數法間轉換結果一覽表

測驗長度	比較測驗	錯誤分類數	百分比一致性	κ係數
10	TEST1-TEST1	707	0.9293	0.857
10	TEST2-TEST2	530	0.9470	0.894
20	TEST1-TEST1	540	0.9460	0.890
20	TEST2-TEST2	412	0.9588	0.913
30	TEST1-TEST1	290	0.9710	0.942
30	TEST2-TEST2	274	0.9726	0.930
40	TEST1-TEST1	381	0.9619	0.924
40	TEST2-TEST2	287	0.9713	0.943
50	TEST1-TEST1	337	0.9663	0.932
50	TEST2-TEST2	247	0.9753	0.950

表 4-19 自然科兩次測驗不同測驗長度類型下最大測驗訊息量法與測驗特徵曲線構圖法間轉換結果一覽表

測驗長度	比較測驗	錯誤分類數	百分比一致性	κ係數
10	TEST1-TEST1	707	0.9293	0.857
10	TEST2-TEST2	530	0.9470	0.894
20	TEST1-TEST1	540	0.9460	0.890
20	TEST2-TEST2	412	0.9588	0.913
30	TEST1-TEST1	471	0.9529	0.905
30	TEST2-TEST2	573	0.9427	0.861
40	TEST1-TEST1	641	0.9359	0.871
40	TEST2-TEST2	287	0.9713	0.943
50	TEST1-TEST1	825	0.9175	0.831
50	TEST2-TEST2	620	0.9380	0.875

二、不同測驗長度類型下換算古典測驗分數法與測驗特徵曲線構圖法之轉換差異分析

對於測驗長度因素影響換算古典測驗分數法與測驗特徵曲線構圖法間分類差異上，影響的層面亦不明顯，由此二法轉換之差異結果，如表 4-20 所示，兩者間轉換後之古典測驗差異分數，皆非常微小，並未明顯受測驗長度的增加而有所變化。其中，最大的差異分數更出現於測驗長度為 50 題時的 2 分差距，對照表 4-21 的分類結果時，在 50 題試題，以 TEST1 為比較測驗下，錯誤分類的受試者達 532 人，對照表 4-20，可發現此錯誤分類人數乃由換算古典測驗分數法求得之轉換結果(32-33 分)，與測驗特徵曲線構圖法轉換之結果(34-35 分)，兩者間差異所組成，即是為古典原始總分 33(262 人)與 34 分(270 人)的受試者。

對此，進一步詳細描述表 4-20 以便於解說接續之細部分析，該表乃為各測驗長度類型下，由換算古典測驗分數法與測驗特徵曲線構圖法轉換後之古典測驗答對題數的差異分數與其鄰近分數點之相對人數對照表(在此為簡化呈現，僅列少數分數點之人數對照表，詳細請參考附錄七)，以上述 50(TEST1)為例，差異分數為 2 分(即 33 與 34 分於測驗特徵曲線構圖法中仍視為未精熟者)，其鄰近分數點分別為 32 至 35 分，而該分數點相對人數則為 244 至 265 人。

表 4-20 不同測驗長度類型下轉換方法間之差異分數與通過分數鄰近分數點人數

測驗長度	差異分數	通過分數鄰近古典測驗分數點（人數）			
10(TEST1)	0	5 (1469)	6 (1467)	7 (1402)	8 (1323)
10(TEST2)	0	4 (1085)	5 (1167)	6 (1252)	7 (1253)
20(TEST1)	0	12 (682)	13 (677)	14 (698)	15 (667)
20(TEST2)	0	12 (643)	13 (622)	14 (564)	15 (630)
30(TEST1)	1	19 (476)	20 (455)	21 (411)	22 (476)
30(TEST2)	1	21 (426)	22 (439)	23 (414)	24 (426)
40(TEST1)	1	26 (336)	27 (338)	28 (336)	29 (346)
40(TEST2)	0	23 (324)	24 (331)	25 (363)	26 (357)
50(TEST1)	2	32 (244)	33 (262)	34 (270)	35 (265)
50(TEST2)	2	30 (247)	31 (257)	32 (228)	33 (259)

表 4-21 自然科兩次測驗不同測驗長度類型下換算古典測驗分數法與測驗特徵曲線構圖法間轉換結果一覽表

測驗長度	比較測驗	錯誤分類數	百分比一致性	κ 係數
10	TEST1-TEST1	0	1.000	1.000
10	TEST2-TEST2	0	1.000	1.000
20	TEST1-TEST1	0	1.000	1.000
20	TEST2-TEST2	0	1.000	1.000
30	TEST1-TEST1	455	0.9545	0.908
30	TEST2-TEST2	439	0.9561	0.895
40	TEST1-TEST1	338	0.9662	0.931
40	TEST2-TEST2	0	1.000	1.000
50	TEST1-TEST1	532	0.9468	0.890
50	TEST2-TEST2	485	0.9515	0.902

從細部精確錯誤分類人數角度視之，相較於表 4-21 測驗長度為 30 題時，以 TEST1 為比較測驗求得之結果，雖然二法轉換之通過分數差異僅為 1 分(分別為 19-20 與 20-21 分)，但錯誤分類的受試者即達 455 人(對照表 4-20 可看出即為

古典原始總分 20 分的受試者)，與上述 50 題之錯誤分類人數 532 人相比，兩者已非常接近，顯示測驗長度愈長，雖然方法轉換後之差異分數較大，但亦未產生較大錯誤分類人數。對此，可同時搭配表 4-20 與圖 4-2 作解說，該圖 X 軸代表著原始古典測驗總分(如測驗長度為 10 題時，其範圍則自 0-10 分，其餘以此類推)，Y 軸為各原始分數點相對應之受試者人數，圖中十條線自左上自右下，分別代表 TEST1 與 TEST2 測驗長度為 10 至 50 題之結果。當測驗長度為 10 題時，分數點於 5-8 分時皆有高的受試者人數(如表 4-20 皆超過 1000 人)，而漸漸從 20 至 50 題時，分數點人數分佈亦隨之分散，至 50 題時各分數點人數約莫不超過 350 人，由此觀點視之，顯示當測驗長度增加時，每個原始總分分數點人數亦隨之分散，因而，由各方法求得之轉換分數，彼此間即使有較大的差異，但影響錯誤分數的層面亦會較測驗長度短時輕微。

由此觀之，若將其推論至上述最大測驗訊息量法與換算古典測驗分數法、測驗特徵曲線構圖法之轉換效果時，同樣將由最大測驗訊息量法求得之 θ_{max} 想像為某一原始古典分數，與其它轉換方法相較時，即不難理解隨著測驗長度的增加，即使兩者於轉換分數間有較大的差異表現，應同樣能有效減少錯誤的分類人數。對照實務用途時，可發現對於影響轉換時差異分數的因素，決策者並不容易掌握與控制，但卻可藉由增加測驗長度，分散分數點人數，以彌平錯誤分類的影響。

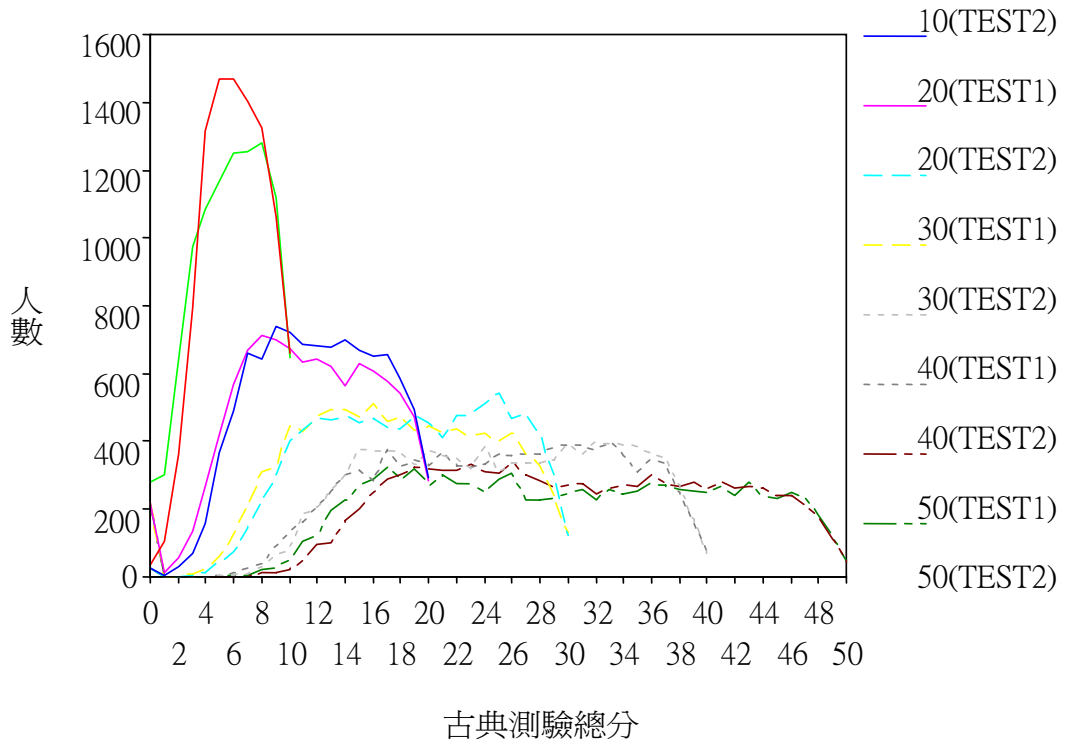


圖 4-2 不同測驗長度類型下各分數點人數變化圖

參、定錨點題目與測驗長度因素之互動分析

以下茲就不同測驗長度類型下各定錨點與定錨點題目篩選結果進行分析，陳

述如下：

一、不同測驗長度類型下各定錨點篩選之結果

實務應用上，為能有效的深入剖析精熟/未精熟者間真實差異能力，首要條件即是在各定錨點上須具備充足的定錨點題目，以便選擇具內容代表性之範例試題。對此，測驗長度因素所扮演的角色則不容忽視。在此，本研究即探討在固定 4 個定錨點下，定錨點題目與測驗長度間的互動結果。而各測驗長度類型下定錨點篩選結果如表 4-22 所示，在同樣是採以精熟標準為中心，上下各 0.5 個標準差、1.5 個標準差能力值作為定錨點，例如，在 10(TEST1)中乃以 1.37555 與 0.54185 分別代表精熟者的定錨能力值，而以-0.29185 與-1.12555 代表未精熟者的定錨能力值。另由於各測驗長度類型間的 $\theta_{\max}$ 與標準差各有差異，使得定錨點亦隨之不同，但是彼此間差距並非過於極端，仍能保持於相似定錨點範圍內。

表 4-22 不同測驗長度類型下各定錨點篩選結果一覽表

測驗長度	$\theta_{\max}$	標準差	精熟定錨點		未精熟定錨點	
10(TEST1)	0.1250	0.8337	1.37555	0.54185	-0.29185	-1.12555
10(TEST2)	0.0000	0.8689	1.30335	0.43445	-0.43445	-1.30335
20(TEST1)	0.2500	0.9196	1.6294	0.7098	-0.2098	-1.1294
20(TEST2)	0.3750	0.9430	1.7895	0.8465	-0.0965	-1.0395
30(TEST1)	0.2500	0.9483	1.67245	0.72415	-0.22415	-1.17245
30(TEST2)	0.5000	0.9736	1.9604	0.9868	0.0132	-0.9604
40(TEST1)	0.2500	0.9666	1.6999	0.7333	-0.2333	-1.1999
40(TEST2)	0.0000	0.9857	1.47855	0.49285	-0.49285	-1.47855
50(TEST1)	0.3750	0.9777	1.84155	0.86385	-0.11385	-1.09155
50(TEST2)	0.2500	1.0021	1.75315	0.75105	-0.25105	-1.25315

二、不同測驗長度類型下定錨點題目篩選結果分析

在相仿的定錨點下，不同測驗長度因素與定錨點題目間的互動關係，可從兩方面探討：首先，從直觀的角度，測驗長度愈長相對能擷取到定錨點題目的機會理應會較多，對照表 4-23 所示，在不同測驗長度類型中第 2 個定錨點下，有較顯著隨著測驗長度增長，定錨點題目亦隨之增加(從 10(TEST1) 第 2 定錨點下，兩次測驗的 2 題、1 題至 50(TEST1)第 2 定錨點下，兩次測驗的 21 題、23 題)，可證實此項觀點。但另一方面，若參照其它定錨點(如第 1 或第 4 個定錨點)，測驗長度的效果相較之下似乎並不顯著，尤其是處於較極端的定錨點更是如此，例如，第 4 個定錨點並無明顯因測驗長度較長，因而能獲取較充足的定錨點題目，至多亦只篩選到 3 題定錨點題目。

表 4-23 不同測驗長度類型下各定錨點題目篩選結果一覽表

測驗長度	精熟		未精熟	
	第 1 定錨點	第 2 定錨點	第 3 定錨點	第 4 定錨點
定錨點 10(TEST1)	1.37555	0.54185	-0.29185	-1.12555
10(TEST1)	1 (10%)	2 (20%)	0 (0%)	1 (10%)
10(TEST2)	0 (0%)	1 (10%)	1 (10%)	0 (0%)
定錨點 10(TEST2)	1.30335	0.43445	-0.43445	-1.30335
10(TEST1)	1 (10%)	1 (10%)	0 (0%)	1 (10%)
10(TEST2)	2 (20%)	1 (10%)	1 (10%)	0 (0%)
定錨點 20(TEST1)	1.6294	0.7098	-0.2098	-1.1294
20(TEST1)	5 (25%)	6 (30%)	1 (5%)	1 (5%)
20(TEST2)	4 (20%)	4 (20%)	3 (15%)	0 (0%)
定錨點 20(TEST2)	1.7895	0.8465	-0.0965	-1.0395
20(TEST1)	5 (25%)	7 (35%)	2 (10%)	1 (5%)
20(TEST2)	5 (25%)	6 (30%)	5 (25%)	0 (0%)
定錨點 30(TEST1)	1.67245	0.72415	-0.22415	-1.17245
30(TEST1)	6 (20%)	10 (33.33%)	5 (16.67%)	2 (6.67%)
30(TEST2)	5 (16.67%)	8 (26.67%)	4 (13.33%)	0 (0%)
定錨點 30(TEST2)	1.9604	0.9868	0.0132	-0.9604
30(TEST1)	2 (6.67%)	12 (40%)	5 (16.67%)	2 (6.67%)
30(TEST2)	5 (16.67%)	14 (46.67%)	5 (16.67%)	0 (0%)
定錨點 40(TEST1)	1.6999	0.7333	-0.2333	-1.1999
40(TEST1)	4 (10%)	16 (40%)	5 (12.5%)	3 (7.5%)
40(TEST2)	5 (12.5%)	13 (32.5%)	7 (17.5%)	1 (2.5%)
定錨點 40(TEST2)	1.47855	0.49285	-0.49285	-1.47855
40(TEST1)	7 (17.5%)	8 (20%)	2 (5%)	3 (7.5%)
40(TEST2)	3 (7.5%)	12 (30%)	4 (10%)	1 (2.5%)
定錨點 50(TEST1)	1.84155	0.86385	-0.11385	-1.09155
50(TEST1)	5 (10%)	21 (42%)	5 (10%)	3 (6%)
50(TEST2)	8 (16%)	23 (46%)	8 (16%)	1 (2%)
定錨點 50(TEST2)	1.75315	0.75105	-0.25105	-1.25315
50(TEST1)	7 (14%)	22 (44%)	3 (6%)	3 (6%)
50(TEST2)	9 (18%)	19 (38%)	5 (10%)	1 (2%)

對此，可從第二個面向，以最大試題訊息量觀點來解釋。現以 50 題下採用 50(TEST1)定錨點為例，分析結果如表 4-24 所示，試題 1 與 2，其最大試題訊息量對應之能力值 $\theta^{i}_{\max}$ 分別為-3.3597、-2.9702，顯示題目較為簡易，乃屬較適合

低定錨點能力之受試者，因此，對於第 3 定錨點-0.11385 而言，能夠輕易符合 Huynh(1998)認為第i個定錨點上受試者，正確反應機率值大於或等於.75 的準則，但對下一個定錨點-1.09155 而言，該試題仍非常的簡單，因而，無法符合 Huynh 認為的下一個定錨點受試者，正確反應機率小於.66(較寬鬆標準)的標準，使得最後分析結果乃歸屬於第 4 個定錨點。同理，試題 4、6、7、10 的最大試題訊息量對應之能力值約介於-.2295 至-.6632 之間，對於第 3 定錨點而言，仍屬相當簡易，可符合上述正確反應機率大於或等於.75 的準則，再進一步檢查第 2 個準則時，可發現就第 4 定錨點而言，該試題則頗具難度，正確反應機率會因而下降，因此可輕易符合該準則的要求，使得最後分析結果乃歸屬於第 3 定錨點。

以此推論，如表 4-24 所示，試題 11、12、13、17…35 乃歸屬於第 3 定錨點 0.86385，而 37、40、41…47 等較難試題則較適合於較高的第 4 定錨點 1.84155。由此觀之，最大試題訊息量所對應之適合能力值，乃與作為各定錨點之定錨點題目習習相關(其餘結果如附錄八所示，亦可發現此性質)。而相對於本研究採用之最大測驗訊息量法，精熟標準所在位置，即為整份測驗多數試題最適於施測之能力值，由此，即不難理解為何於精熟標準附近之定錨點(如表 4-23 之第 2 與第 3 定錨點)能獲取較多定錨點題目(如 50(TEST1)定錨點 0.86385 於兩次測驗即有 21(42%)、23(46%))，而於較極端之定錨點(即表之第 1 與第 4 定錨點)，即使測驗長度較長亦無法具備充足的定錨點題目(如 50(TEST1)定錨點-1.09155 於兩次測驗僅有 3(6%)、1(2%))。

綜合上述，可發現測驗長度，並非是絕對影響定錨點題目篩選的因素，更重要的在於最大試題訊息量所對應之最適能力值是否能與定錨點相搭配，就此觀點，對照本研究於第二章文獻探討第三節通過分數轉換與能力標準描述時，提出圖 2-11，以篩選各定錨點下之最大試題訊息量以填滿整份測驗之目標訊息量的概念，即可獲得支持的證據。但實際進一步的佐證，仍有待後續之研究。

表 4-24 第一次測驗長度 50 題下採用 TEST1 定錨點之定錨點題目篩選結果

定錨點題目	定錨點				$\theta^{i}_{\max}$
	1 (1.84155)	2 (0.86385)	3 (-0.11385)	4 (-1.09155)	
1	.99534	.98877	.97332	.93839*	-3.3597
2	.97741	.95800	.92363	.86636*	-2.9702
4	.99779	.98035	.85841*	.56147	-.4428
6	.99542	.96505	.80262*	.53316	-.2295
7	.99977	.99549	.92092*	.51033	-.6632
10	.99309	.95615	.78848*	.51124	-.2729
11	.97944	.85432*	.46503	.21374	.2351
12	.99170	.91134*	.60726	.45875	.3746

13	.96122	.81547*	.51812	.34478	.4745
17	.92306	.80057*	.59914	.41154	.2516
18	.98585	.90371*	.60163	.32507	.0764
19	.95916	.82490*	.53273	.31549	.3041
20	.97897	.84028*	.47031	.29819	.4365
21	.96901	.80003*	.44259	.29424	.5493
23	.91453	.75502*	.49104	.27343	.2824
24	.98601	.88885*	.54274	.30782	.2276
26	.98915	.86835*	.43603	.27125	.4012
27	.99710	.93686*	.43577	.14604	.1173
29	.98406	.86443*	.42493	.14157	.1669
30	.98440	.82504*	.33222	.16160	.4206
31	.95757	.78628*	.45976	.29954	.5489
32	.98266	.75043*	.37466	.32934	.7953
35	.98892	.88647*	.46716	.22278	.2298
37	.92829*	.49953	.26447	.24553	1.1689
38	.99342	.98541	.96803	.93188*	-3.4859
39	.96529	.76769*	.37547	.23254	.5808
40	.85146*	.59442	.31330	.17861	.8734
41	.94897*	.40189	.18016	.17242	1.1595
42	.94741	.78203*	.47449	.28737	.4692
43	.89257*	.49702	.17368	.11309	1.0233
44	.97894	.82836*	.40584	.21325	.3942
47	.81444*	.44853	.22660	.17849	1.3065
49	.99598	.96522	.77015*	.39054	-.3699
50	.98942	.92289*	.65852	.40917	.0864
總定錨點題目數	5	21	5	3	

註 1：測驗長度議題中，乃有別於上述原始測驗長度(58 題)篩選定錨點題目的準則，在此，是以較寬鬆的標準(下一個定錨點受試者，正確反應機率小於.66 即可)，代表著在最低標準下的結果，因此，為作區別，以*代表該試題符合該定錨點之篩選準則。

2：左側之定錨點題目為相對於原始測驗之題號，而未達標準者，則不列入表中。

3：各定錨點由左至右分別為第 1 至第 4 定錨點，而括號內為其對應能力值。

第三節 精熟標準設定方法間測驗異質性因素結果之分析

本節乃以探討測驗異質性因素對於最大測驗訊息量法、換算古典測驗分數法、測驗特徵曲線構圖法與定錨點的影響為主軸。對此，可就三個面向描述之，首先，分析在測驗異質性下，最大測驗訊息量法、換算古典測驗分數法與測驗特徵曲線構圖法求得之精熟標準及其一致性分類結果；其次，探討測驗異質因素對於換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益可能的影響層面；最後，深究測驗異質性因素於定錨點題目選擇上可能產生的影響力。分析結果茲陳述如下：

壹、測驗異質因素下精熟標準設定結果分析

以下茲就測驗異質下精熟標準設定方法求得之精熟標準與其相對應分類一致性效果，陳述如下：

一、測驗異質下精熟標準設定方法求得之精熟標準結果

對於測驗異質性議題，研究設計上乃需具備簡易與困難兩類型測驗試題，相對於自然科兩次測驗上，即如第三章研究方法中測驗異質性研究設計圖 3-2 所示，乃將其分為四類(詳細試題篩選結果請參見附錄九)。期望在控制測驗長度下(在此皆設定為 20 題)，針對各測驗難度類型，求得最大測驗訊息量法、換算古典測驗分數法與測驗特徵曲線構圖法之精熟標準。分析結果如表 4-25 所示，在第一次測驗中簡易(Easy(TEST1))與第二次測驗中困難(Hard(TEST2))的類型，其 IRT 測驗難度值分別為-0.9151 與 0.6724，相差達 1.5875，可謂符合測驗異質效果，而最大測驗訊息量法求得之 $\theta_{\max}$ 值分別為-0.1250、0.6250，亦符合於簡易測驗中精熟標準較低，而於困難測驗中精熟標準相對較高的實務應用概念。

在此，進一步對此概念詳細描述之。決策者會施行較簡易的測驗，通常代表其欲甄選較多的受試者(視其目的而定)，因而必需降低門檻，促使多數人選皆能符合資格；相反的，若決策者欲甄選較優秀的人才，當然必須將測驗難度提昇，以排除多數人選，門檻相對需較高。而最大測驗訊息量法即具備此隨測驗難易程度來決定門檻的特性，測驗困難時，精熟標準自然較高(表示測驗適合該高能力者)，而測驗較簡易時，精熟標準相對會較低。對照於國中基本學力測驗而言，試題編製的出發點，乃用以評量學生的基本能力，若以精熟測驗的角度視之，隱含著學生若具備此基本能力，即符合上高中資格；反之，則不具資格。因而，在僅要求學生能通過基本門檻下，難度應屬簡易，以使多數學生能符合準則。

另一方面，在第一次測驗中困難(Hard(TEST1))與第二次測驗中簡易(Easy(TEST2))的類型，測驗難度差異(達 1.7206)亦十分充足，而求得之相對 $\theta_{\max}$ 分別為 1.1250、-0.6250，更符合上述準則。此外，為作對照用途，乃引用測驗長度議題中 20 題的資料，將其視為非異質測驗類型(在此視為常態 Normal(Test1)與 Normal(Test2))，其測驗難度差異僅達 0.0198，符合本研究對常態標準的要

求。

在轉換分數方面，雖然由最大測驗訊息量法求得之θ_{max}，符合上述所稱實務應用的概念，但經轉換後之古典測驗答對題數，卻有些許差異，於 Easy(TEST1)和 Hard(TEST2)中，並未能如上述顯示困難測驗(13-14 分)會較簡易測驗(16-17 分)求得較高之標準，此乃因，在相同 θ 能力值下，受試者於簡易測驗上需較困難測驗答對較多的試題數，才足以達到相同能力標準，如以 Easy(TEST1)之θ_{max}值 -0.1250 而言，在 Hard(TEST2)中，則僅需獲得 8-9 分即可，就此可參照附錄十之詳細轉換結果。若對照實務用途亦符合常理，舉例而言，若某次資格考試，決策者因人為疏失，而將測驗編製過於簡易，若在同樣能力標準要求下，理當應答對較多試題，才足以證明符合資格；反之，亦是如此。

另一方面，由 Hard(TEST1) 與 Easy(TEST2)間求得之結果，概念是亦是如此，雖然，Hard(TEST1)轉換後之分數 14-15 分，雖較 Easy(TEST2)之 11-12 分為大，此乃因兩者求得之θ_{max}差異較上述為大，但是參照附錄十，同樣可發現 Hard(TEST1)之 θ_{max} 1.1250 在 Easy(TEST2)中，則需獲得 19-20 分，才足以達到此能力標準。

表 4-25 測驗異質性下最大測驗訊息量對應能力值、各方法之轉換古典測驗答對題數與相關結果一覽表

測驗類型	θ_{max}	IRT 測驗難度	古典偏態	IRT 偏態	TCTS	TCCM
Easy(TEST1)	-0.1250	-0.9151	-0.876	-0.288	16-17	15.6290 (15-16)
Hard(TEST2)	0.6250	0.6724	0.416	0.441	13-14	11.7484 (11-12)
Hard(TEST1)	1.1250	0.7579	0.455	0.505	14-15	14.3532 (14-15)
Easy(TEST2)	-0.6250	-0.9627	-1.163	-0.588	11-12	12.7981 (12-13)
Normal(Test1)	0.2500	-0.0432	-0.070	0.085	13-14	13.1039 (13-14)
Normal(Test2)	0.3750	-0.0234	-0.138	0.007	13-14	13.4266 (13-14)

二、測驗異質性下精熟標準設定方法分類一致性效果

對於測驗異質性議題，相對於實務應用的概念，本研究乃假設當測驗編製者因人為疏失，而使得兩次測驗間難度有所偏差，如先前所討論的，在避免爭議的情況下，仍以相同精熟標準為兩次測驗之分類準則，而非採用上述最大測驗訊息量法以變動的θ_{max}為各自測驗精熟標準的特性，以探討在此情況下可能的影響。

相對於本研究設計，乃以最大測驗訊息量法於 Easy(TEST1)求得之標準同時為 Easy(TEST1)、Hard(TEST2)的分類準則，反過來，再由 Hard (TEST2)中求得之標準，同時為 Easy(TEST1)、Hard(TEST2)的分類準則，如此，交叉驗證下進行分析；另一方面，對於 Hard(TEST1)、Easy(TEST2)的設計亦是如此。

資料分析結果如表 4-26 所示，顯示在不同測驗難度類型下，由兩次最大測驗訊息量法求得之分類結果，可發現以 Easy(不論 TEST1 或 TEST2)或 Hard (不論 TEST1 或 TEST2)為精熟標準時，其錯誤分類數皆較常態難度(Normal)測驗為高，但彼此差異不大，整體的百分比一致性都能維持在 80%以上，而 κ 係數亦符合 Berk(1984)、Hambleton (1990)、Subkoviak (1988)所述，精熟標準離平均表現較遠時，有明顯低估現象。

另呈現 0-1 指數，代表第一次測驗時被分類為未精熟者(以 0 代表)，卻於第二次測驗時被分類為精熟者(以 1 代表)的人數，而 1-0 指數代表意義正好相反。若以 Easy(TEST1) vs. Hard(TEST2)為例，0-1 指數則代表於 Easy(TEST1)被視為未精熟者，卻於 Hard(TEST2)被分類分精熟者；反之，1-0 指數亦是如此。

由表 4-26 可看出，0-1 與 1-0 的人數會隨著精熟標準位置的不同而有所變化，以表中 Easy(TEST1) vs. Hard (TEST2)採 Easy(TEST1)為精熟標準為例，若與 Normal(不論是 TEST1 或 TEST2)相比，其 1-0 人數(1057 人)明顯是大於 0-1 人數(611 人)，代表著精熟者於第二次測驗上被分類為未精熟者人數較多；而若是比較 Hard(TEST1) vs. Easy (TEST2)採 Easy(TEST2)為精熟標準時，則未精熟者反於第二次測驗中被分類為精熟者人數較多(0-1：1449 人大於 1-0：613 人)。就此，相較於表 4-27 與表 4-28 之換算古典測驗分數法與測驗特徵曲線構圖法的分類結果，明顯改善許多，可能原因接續分析如下。

表 4-26 測驗異質性下最大測驗訊息量法分類結果一覽表

比較測驗	精熟標準	0 - 1	1 - 0	錯誤分類數	百分比一致性	κ 係數
Easy(TEST1) Hard(TEST2)	Easy (TEST1)	611	1057	1668	0.8332	0.667
Easy(TEST1) Hard(TEST2)	Hard (TEST2)	427	1159	1586	0.8414	0.593
Hard(TEST1) Easy(TEST2)	Hard (TEST1)	579	871	1450	0.8550	0.380
Hard(TEST1) Easy(TEST2)	Easy (TEST2)	1449	613	2062	0.7938	0.457
Normal(TEST1) Normal(TEST2)	Normal (TEST1)	673	696	1369	0.8631	0.723
Normal(TEST1) Normal(TEST2)	Normal (TEST2)	696	677	1373	0.8627	0.713

對此，進一步輔以圖 4-3 作說明，該圖分別為 Hard(TEST1)與 Easy(TEST1)求得之最大測驗訊息量圖，若由 Hard(TEST1) 求得之θ_{max} 1.1250 為兩次測驗之分類準則時，如上所述，此對於 Easy(TEST1)而言，並非是最佳的精熟標準，因而，在該精熟點附近能力估計會較不穩定，致使分類時較易產生不一致結果；若對照於圖 4-1 測驗長度 20 題時之結果(即為此 Normal 測驗之結果)時，精熟標準間的差距較其為大，因而，穩定分類上相對會較差，但有藉於 IRT 具有因試題參數估計受試者能力的特性，使得分類上，相較於換算古典測驗分數法與測驗特徵曲線構圖法之分類結果，仍不致於產生太大的影響。

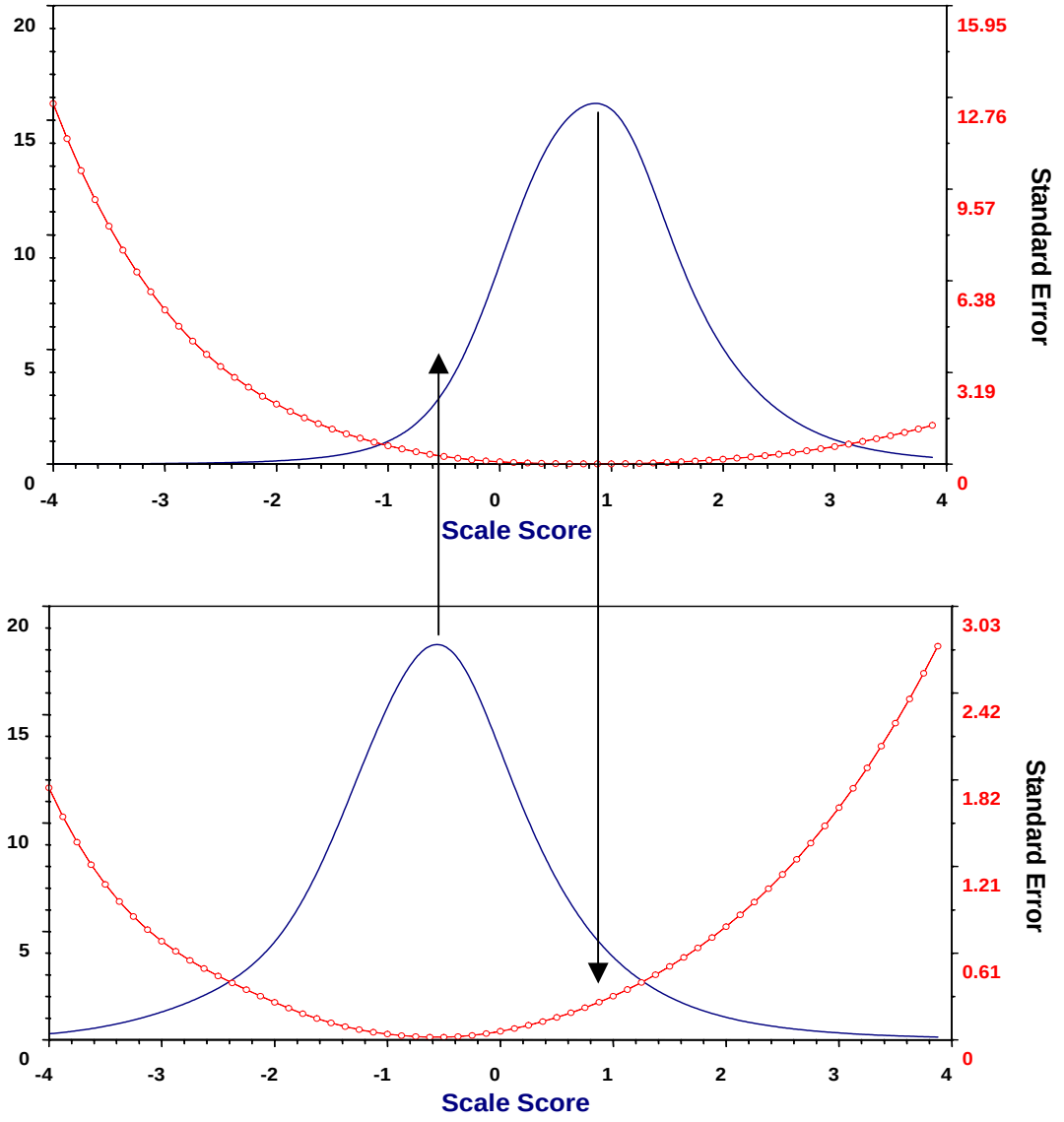


圖 4-3 Hard(TEST1)與 Easy(TEST1)之最大測驗訊息量圖

註：上圖為 Hard(TEST1)之結果；下圖為 Easy(TEST1)之結果；實線為訊息量之變化，採左側之刻度；圓點線為標準誤之變化，採右側之刻度。

表 4-27 顯示在不同測驗難度類型下，換算古典測驗分數法於自然科兩次測

驗之分類結果，由此即可明顯看出在兩份異質測驗上，同時採用某一固定通過分數(如同時採用第一或第二次測驗之通過分數)為兩次測驗之精熟標準時，相較於兩次皆為同質測驗(皆為 Normal 難度)時，分類錯誤明顯大了許多，最大錯誤分類數更達 5050 人；另一方面，表 4-28 為運用測驗特徵曲線構圖法下之分類結果，同樣亦出現龐大錯誤分類數，最高亦達 5224 人。若由此觀點視之，明顯可看出現行考選部採行 60 分(或 70 分)固定精熟標準的缺失，其分類一致性易深受測驗異質性因素的影響。相較於最大測驗訊息量法之分析結果，則能有效彌平此干擾。

表 4-27 測驗異質性下換算古典測驗分數法分類結果一覽表

比較測驗	精熟標準	0 - 1	1 - 0	錯誤分類數	百分比一致性	κ 係數
Easy(TEST1) Hard(TEST2)	Easy (TEST1)	4	4112	4116	0.5884	0.190
Easy(TEST1) Hard(TEST2)	Hard (TEST2)	2	5048	5050	0.4950	0.192
Hard(TEST1) Easy(TEST2)	Hard (TEST1)	4943	16	4959	0.5041	0.170
Hard(TEST1) Easy(TEST2)	Easy (TEST2)	4935	35	4970	0.5030	0.195
Normal(TEST1) Normal(TEST2)	Normal (TEST1)	548	922	1470	0.8530	0.690
Normal(TEST1) Normal(TEST2)	Normal (TEST2)	548	922	1470	0.8530	0.690

表 4-28 測驗異質性下測驗特徵曲線構圖法分類結果一覽表

比較測驗	精熟標準	0 - 1	1 - 0	錯誤分類數	百分比一致性	κ 係數
Easy(TEST1) Hard(TEST2)	Easy (TEST1)	1	4486	4487	0.5513	0.201
Easy(TEST1) Hard(TEST2)	Hard (TEST2)	5	5219	5224	0.4776	0.168
Hard(TEST1) Easy(TEST2)	Hard (TEST1)	4943	16	4959	0.5041	0.170
Hard(TEST1) Easy(TEST2)	Easy (TEST2)	5078	28	5106	0.4894	0.186
Normal(TEST1) Normal(TEST2)	Normal (TEST1)	548	922	1470	0.8530	0.690
Normal(TEST1) Normal(TEST2)	Normal (TEST2)	548	922	1470	0.8530	0.690

對此，可能的原因，在此藉由 Easy (TEST1) vs. Hard(TEST2)中採 Easy(TEST1)為精熟標準為例，搭配參照圖 4-4 作說明。在古典測驗理論下，於第一次測驗時，因題目較為簡易，受試者得分會明顯偏高，因而呈負偏態得分分配；但在第二次測驗時，因題目較困難，受試者得分明顯偏低，而呈正偏態得分分配，這使得兩次測驗受試者得分分配差異趨大(由表 4-25 古典偏態值可看出，其差異達 1.292)，加之古典測驗理論計分較無彈性，若採同一固定精熟標準下(本例採 Easy(TEST1))，將會使得精熟人數於第二次測驗時明顯被分類至未精熟狀態，但在最大測驗訊息量法中，其偏態值差異則較古典測驗理論下分析結果為小(僅達 0.729)，乃因具有考量試題參數進行能力估計的特性，能力估計較穩定，顯示受測驗異質因素的影響較不明顯。

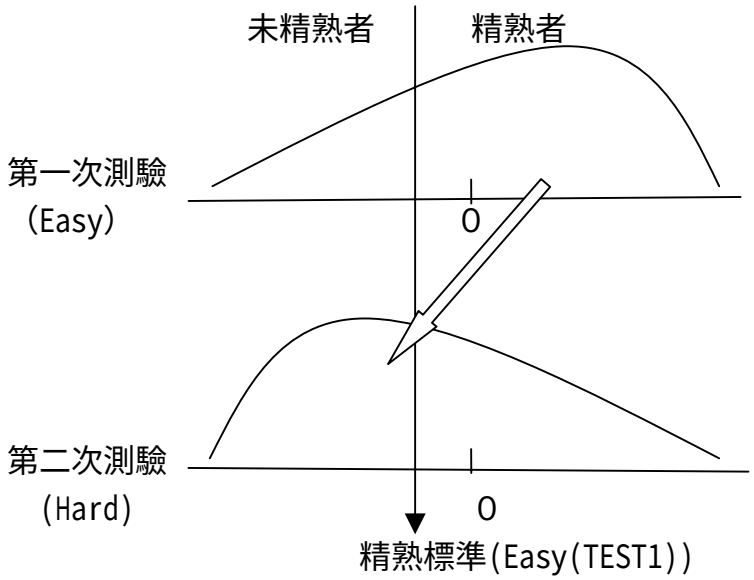


圖 4-4 異質性測驗下採行古典固定精熟標準之缺失解說圖

貳、測驗異質因素下換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益分析

以下茲就測驗異質下換算古典測驗分數法與測驗特徵曲線構圖法之轉換一致性效果與轉換差異兩方面探討之，陳述如下：

一、測驗異質性下換算古典測驗分數法與測驗特徵曲線構圖法之轉換一致性效果分析

對於測驗異質因素於換算古典測驗分數法與測驗特徵曲線構圖法之轉換效益影響上。在此，仍集中於比較分類精熟/未精熟者百分比一致性的效果，但檢視的面向則有所差異。本研究有關測驗異質性的設計，乃強調同一次比較中兩次測驗間難度的差異，而測驗分數的轉換效益則是強調完全控制測驗、受試者表現(即同一份測驗，採不同精熟標準設定方法)之結果，兩者具有不同意函。因而，研究上乃稍作調整，轉而探討不同比較中，採用簡易測驗、困難測驗或常態測驗

間於轉換效益上是否會產生不同的效果？

表 4-29 為最大測驗訊息量法與換算古典測驗分數法間轉換結果，資料分析結果顯示，不同比較測驗中，是否為 Easy(不論是 TEST1 或 TEST2)或 Hard(不論是 TEST1 或 TEST2)的測驗，百分比一致性皆較 Normal 下之結果為佳，最小錯誤分類數乃出現於 Hard(TEST1)下之 238 人，而最大錯誤分類數則出現於 Normal(TEST1)之 540 人，但彼此差異皆十分微小；另一方面，表 4-30 為最大測驗訊息量法與測驗特徵曲線構圖法之轉換結果，亦有類似的結果，並未發現測驗異質性明顯影響轉換分數上分類一致性效果。

若從細部角度，比較表 4-29 與表 4-30 之分析結果，可發現換算古典測驗分數法不論於何種測驗難度類型下，其表現皆較測驗特徵曲線構圖法為佳，但同樣彼此間的差異均不大。整體而言，兩者皆具備良好百分比一致性與κ係數，顯示不論測驗比較中，於何種測驗難度類型中，採用何種轉換方式，並不致於影響轉換分數間一致性分類的效果。

表 4-29 測驗異質性下最大測驗訊息量法與換算古典測驗分數法間轉換結果一覽表

比較測驗	錯誤分類數	百分比一致性	κ係數
Easy(TEST1)-Easy(TEST1)	410	0.9590	0.918
Hard(TEST2)-Hard(TEST2)	269	0.9731	0.922
Hard(TEST1)-Hard(TEST1)	238	0.9762	0.905
Easy(TEST2)-Easy(TEST2)	256	0.9744	0.924
Normal(TEST1)-Normal(TEST1)	540	0.9460	0.890
Normal(TEST2)-Normal(TEST2)	412	0.9588	0.913

表 4-30 測驗異質性下最大測驗訊息量法與測驗特徵曲線構圖法間轉換結果一覽表

比較測驗	錯誤分類數	百分比一致性	κ係數
Easy(TEST1)-Easy(TEST1)	508	0.9492	0.897
Hard(TEST2)-Hard(TEST2)	777	0.9223	0.802
Hard(TEST1)-Hard(TEST1)	238	0.9762	0.905
Easy(TEST2)-Easy(TEST2)	523	0.9477	0.855
Normal(TEST1)-Normal(TEST1)	540	0.9460	0.890
Normal(TEST2)-Normal(TEST2)	412	0.9588	0.913

二、測驗異質性下換算古典測驗分數法與測驗特徵曲線構圖法之轉換差異分析

對於測驗異質性因素影響換算古典測驗分數法與測驗特徵曲線構圖法間分類差異上，同樣以探討不同測驗比較中，採用簡易測驗、困難測驗或常態測驗間

於分類上是否會產生不同的效果？資料分析結果如表 4-31 所示，於不同測驗難度類型中，由換算古典測驗分數法與測驗特徵曲線構圖法求得之差異通過分數，顯示並未明顯出現較特殊變化，兩者最大差距為 Hard(TEST2)中的 2 分(分別為換算古典測驗分數法之 13-14 分與測驗特徵曲線構圖法之 11-12 分的差異)，對照於表 4-32 的分析結果，有最大的錯誤分類人數(870 人)，此即由古典原始總分 12(467 人)與 13 分(403 人)的受試者所組成，而 Easy(TEST1)與 Easy(TEST2)中錯誤分類數亦可如此推演，乃分別由 16 分(748 人)與 12 分(445 人)所組成。

表 4-31 測驗異質性下轉換方法間之差異分數與通過分數鄰近分數點人數

測驗難度	差異分數	各古典測驗分數點下相對應人數											
		7	8	9	10	11	12	13	14	15	16	17	18
Easy(TEST1)	1	158	234	305	408	427	501	537	649	711	748	936	1076
Hard(TEST2)	2	907	762	582	537	476	467	403	422	377	371	356	328
Hard(TEST1)	0	1029	970	816	691	604	595	474	423	381	340	307	212
Easy(TEST2)	1	229	264	346	344	416	445	498	522	606	750	962	1316
Normal(Test1)	0	661	642	738	722	685	682	677	698	667	650	657	586
Normal(Test2)	0	668	712	700	672	636	643	622	564	630	606	576	542

表 4-32 測驗異質性下換算古典測驗分數法與測驗特徵曲線構圖法間轉換結果一覽表

比較測驗	錯誤分類數	百分比一致性	κ係數
Easy(TEST1)-Easy(TEST1)	748	0.9252	0.850
Hard(TEST2)-Hard(TEST2)	870	0.9130	0.776
Hard(TEST1)-Hard(TEST1)	0	1.000	1.000
Easy(TEST2)-Easy(TEST2)	445	0.9555	0.878
Normal(TEST1)-Normal(TEST1)	0	1.000	1.000
Normal(TEST2)-Normal(TEST2)	0	1.000	1.000

進一步搭配表 4-31 與圖 4-5 從細部精確錯誤分類人數角度分析之，在簡易的測驗中受試者的得分通常會較偏高，因而人數多集中於高分數點上，而困難的測驗則是相反，人數會多集中於低分數點上，對照表 4-31 分析結果，顯示在 Easy 上(不論 TEST1 或 TEST2)，得分點自 7 至 18 分(為簡化篇幅以免過於複雜，在此僅列簡要結果，完整結果請見附錄十)，人數呈現一致上昇趨勢，即如圖 4-5 之負偏態表現，乃於高分數點上有較多受試者；而在 Hard 上(不論 TEST1 或 TEST2)，人數則隨分數點呈現下降趨勢，即如圖之正偏態表現，乃於低分點上有較多受試者。若對照於表 4-31Hard(TEST2)分析結果，雖然兩種轉換方法出現 2 分差異，但其差異的分數點 12 與 13 分，皆出現於較中偏高的位置上，因而未造

成過大的錯誤分類數(即 467+403 = 870 人)，而 Easy(TEST2)亦是如此，其差異分數點為 12 分，仍屬於中間位置，因而錯誤分類仍可控制在一定水準(445 人)。另，就 Easy(TEST1)而言，雖然二法轉換之差異分數僅相差 1 分，但其差異分數點乃為 16 分，較屬偏高位置，因而錯誤分類數即達 748 人。

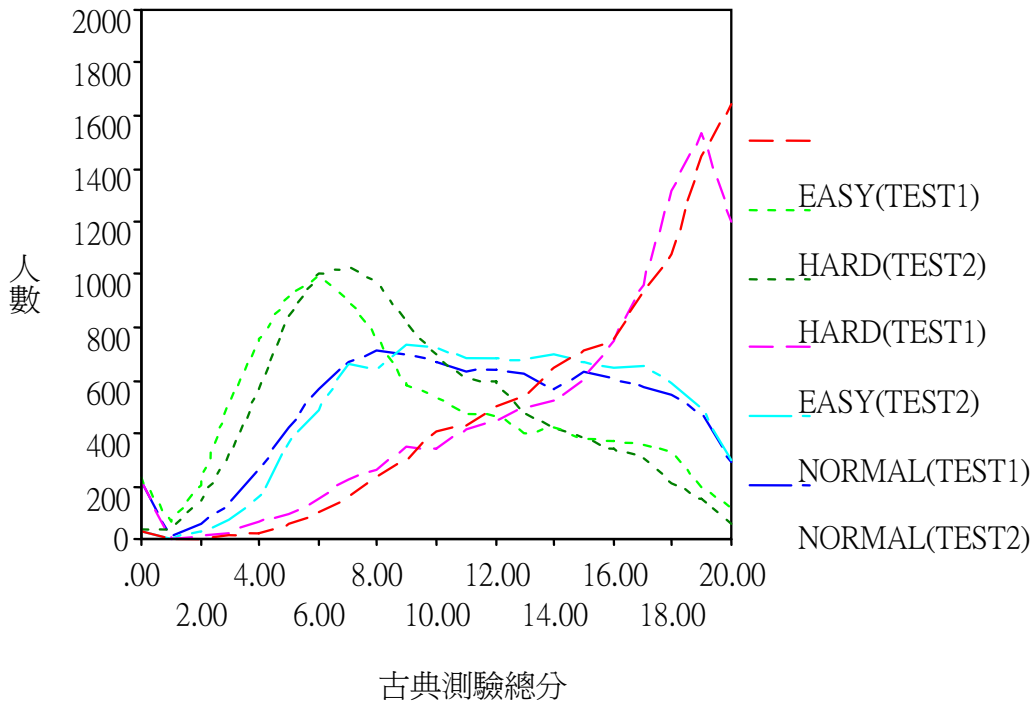


圖 4-5 測驗異質類型下各測驗分數點人數變化圖

由此觀之，若將其推論至上述最大測驗訊息量法與換算古典測驗分數法、測驗特徵曲線構圖法之轉換效果時，如前述，同樣將由最大測驗訊息量法求得之 θ_{max} 想像為某一原始古典分數，與其它轉換方法相較時，即不難理解精熟標準位置與測驗異質性因素間的關係。對照於本研究所採用之最大測驗訊息量法，如前述，此法乃具備隨測驗難易程度來決定門檻的特性，於簡易測驗中求得之精熟標準較低，而於困難測驗中求得之精熟標準相對較高，因而，對照於上述概念，由此法求得之精熟標準附近人數並不會太多，相對於轉換概念上，代表著即使有較大的差異分數，亦不會造成嚴重的錯誤分類人數。

以此論點，即可合理解釋為何在上述表 4-29 中 Easy(不論是 TEST1 或 TEST2)或 Hard(不論是 TEST1 或 TEST2)百分比一致性皆較 Normal 下之結果為佳。假設由最大測驗訊息量法求得之精熟標準與換算古典測驗分數法轉換之結果，兩者相差 1 分(將最大測驗訊息量法求得之 IRT 精熟標準 θ 能力值，想像為某一古典測驗分數)，由上述特性可知，該精熟標準不論於簡易或困難測驗上，相對人數皆會較常態下之分數點人數為少，若採以圖 4-5 解說，可視為如同於 8 分(簡易測驗)或 16 分(困難測驗)下，其附近變動範圍的相對人數皆會少於常態下同等變動範圍之人數。綜合上述，本研究所採最大測驗訊息量法具有因測驗難易程度來決

定精熟標準的特性，恰好能與測驗異質性互相搭配，有效減低錯誤分類人數，而不致造成嚴重錯誤的分類。

參、定錨點題目與測驗異質因素之互動分析

以下茲就測驗異質下各定錨點與定錨點題目篩選結果分析之，陳述如下：

一、測驗異質性下各定錨點篩選之結果

除了測驗長度因素會影響定錨點題目的選擇外，在此，本研究另探討測驗異質性因素的影響，依舊探討於固定 4 個定錨點下，分析其與定錨點題目間的互動結果。對於各異質測驗類型下定錨點的篩選，同樣是以精熟標準為中心，上下各 0.5、1.5 個標準差能力值作為定錨點，資料分析結果如表 4-33 所示，可發現 Easy(不論 TEST1 或 TEST2)測驗求得之定錨點多偏向於低能力點位置(如 Easy(TEST1)求得定錨點分別為：1.23955、0.32985、-0.57985、-1.48955)，而 Hard(不論 TEST1 或 TEST2)測驗求得之定錨點則多偏向於高能力值位置(如 Hard(TEST2)求得定錨點分別為：1.9756、1.0752、0.1748、-0.7256)。對此，由上述最大測驗訊息量法於簡易測驗中求得精熟標準較低，而於困難測驗中求得精熟標準相對較高的特性，則不難理解此一結果。而此差異取向是否會影響定錨點題目的篩選，接續分析如下。

表 4-33 測驗異質類型下各定錨點篩選結果一覽表

測驗別	$\theta_{\max}$	標準差	精熟定錨點		未精熟定錨點	
Easy(TEST1)	-0.1250	0.9097	1.23955	0.32985	-0.57985	-1.48955
Hard(TEST2)	0.6250	0.9004	1.9756	1.0752	0.1748	-0.7256
Hard(TEST1)	1.1250	0.8526	2.4039	1.5513	0.6987	-0.1539
Easy(TEST2)	-0.6250	0.9328	0.7742	-0.1586	-1.0914	-2.0242

二、測驗異質性下定錨點題目篩選結果分析

對於異質測驗與定錨點題目篩選間的相關，資料分析結果如表 4-34 所示，可初步發現於極端定錨點上(第 1 或第 4 定錨點)，多出現無法獲得定錨點題目的現象。在此，同樣可就兩方面分析之。首先，就直觀角度而言，於 Easy(TEST1)下求得之定錨點能力值多偏低，若同樣採用 Easy(TEST1)以選擇定錨點題目時，如表 4-34 中所示，則易造成高能力之定錨點(如第 1 定錨點：1.23955)無法獲取相對的定錨點題目，此乃因於簡易測驗中，即使對於較低定錨點而言，答對機率值仍高，因此，對於高能力定錨點而言，則不易符合 Huynh(1998)認為下一個定錨點受試者，正確反應機率小於.66 的準則，因此易產生沒有定錨點題目的情況；另一方面，若採用 Hard(TEST2)以選擇定錨點題目時，對於 Easy(TEST1)下求得之低定錨能力點而言，在困難測驗中，答對機率則易偏低，因此多無法符合 Huynh 認為第i個定錨點上受試者，正確反應機率值大於或等於.75 的標準，因而造成如表 4-34 中第 2 至 4 定錨點(分別為 0.32985、-0.57985、-1.48955)同樣

無法獲得充足之定錨點題目。

同理，於 Hard(TEST2)下求得之定錨點，在選用 Easy(TEST1)以挑擇定錨點題目時，則會產生於高能力定錨點下無法獲得相對之定錨點題目，而選用 Hard(TEST2)以挑選定錨點題目時，則會產生於低能力定錨點無法獲得適當定錨點題目。綜合上述，即可發現於簡易測驗中，對於高能力定錨點而言，多無法獲取適當的定錨點題目；相對的，於困難測驗中，對於低能力定錨點而言，亦無法產生適當定錨點題目，在此情況下，若搭配由 Easy 或 Hard 測驗求得之偏低或偏高定錨點，則情況更加明顯。對此，乃是經由直觀廣義角度分析之，讀者亦可發現於表 4-34 中最後一列，藉由 Easy(TEST2)求得之定錨點，於 Easy(TEST2)下決定定錨點題目時，反而於高能力定錨點上出現了較多的定錨點題目(1 題與 11 題)。對此，同樣可從上述測驗長度議題中曾介紹過之最大試題訊息量觀點解釋之。

表 4-34 測驗異質類型下各定錨點題目篩選結果一覽表

測驗別	精熟		未精熟	
	第 1 定錨點	第 2 定錨點	第 3 定錨點	第 4 定錨點
定錨點 Easy(TEST1)	1.23955	0.32985	-0.57985	-1.48955
Easy(TEST1)	0 (0%)	9 (45%)	2 (10%)	3 (15%)
Hard(TEST2)	11 (55%)	0 (0%)	0 (0%)	0 (0%)
定錨點 Hard(TEST2)	1.9756	1.0752	0.1748	-0.7256
Easy(TEST1)	0 (0%)	0 (0%)	8 (40%)	5 (25%)
Hard(TEST2)	6 (30%)	9 (45%)	0 (0%)	0 (0%)
定錨點 Hard(TEST1)	2.4039	1.5513	0.6987	-0.1539
Hard(TEST1)	1 (5%)	7 (35%)	0 (0%)	0 (0%)
Easy(TEST2)	0 (0%)	0 (0%)	1 (5%)	13 (65%)
定錨點 Easy(TEST2)	0.7742	-0.1586	-1.0914	-2.0242
Hard(TEST1)	1 (5%)	0 (0%)	0 (0%)	0 (0%)
Easy(TEST2)	1 (5%)	11 (55%)	0 (0%)	0 (0%)

以試題的最大訊息量觀點進一步深入分析時，在此乃以 Easy(TEST1)求得之定錨點，採用 Easy(TEST1)測驗以挑選定錨點題目為例作示範說明，詳細篩選結果可參照附錄十一所示。表 4-35 即為第一次簡易測驗(Easy(TEST1))下採用 Easy(TEST1)定錨點所篩之定錨點題目結果，可發現各定錨點題目最大訊息量對應之能力值 $\theta^{i}_{\max}$ 多屬低能力範圍(皆是負能力值)，因而，對於能力值較高的第 1 定錨點 1.23955 而言，各試題多非常簡易，可輕易符合正確反應機率值大於或等於.75 的標準，但就第 2 定錨點 0.32985 而言，多數試題亦屬簡易，因而對第 1 定錨點而言，即無法符合下一個定錨點受試者，正確反應機率小於.66 的準則。

相對的，分析 Easy(TEST2)求得之定錨點於 Easy(TEST2)下挑選定錨點題目之結果，如表 4-36 所示，最大試題訊息量對應之能力值亦多屬低能力範圍，但相對的定錨點卻較上述為低(就第 1 定錨點 0.7742 與接續之第 2 定錨點-0.1586 而言，遠小於上述之 1.23955 與 0.32985)，因而，會出現於第 12 題時，該最大試題訊息量對應能力值-.4946，就第 1 定錨點 0.7742 而言，正確反應機率如預期的大(為.83793，符合大於或等於.75 準則)，但就第 2 定錨點-0.1586 而言，因與該試題訊息量對應能力值-.4946 頗為相近，致使出現趨於臨界正確反應率標準附近(為.62294)，而勉強符合下一個定錨點受試者，正確反應機率小於.66 的寬鬆準則;相反的, 若該試題乃是藉用上述 1.23955 與 0.32985 定錨點衡量時, 則多半無法符合標準。

另一方面，同理可由此推論為何表 4-36 簡易測驗中於低能力定錨點上，並未如預期出現較多的定錨點題目。就第 3 與第 4 定錨點而言，其能力值過低(分別為-1.0914、-2.0242)，因此，相較於該測驗中各最大試題訊息量對應之能力值時，可發現多半的試題對此二定錨點而言，多過於困難，致使正確反對機率值偏低，無法符合.75 準則，即使，有少數試題乃偏屬簡易(如 3、6、7、10 與 16 題)，但多差臨門一腳，僅是符合臨界準則附近。

綜合上述，可發現測驗異質性與定錨點題目的篩選關聯，若以廣義觀點視之，多呈現簡易測驗中，對於高能力定錨點而言，較難獲取適當的定錨點題目，相反的，於困難測驗中，對於低能力定錨點而言，則無法產生適當定錨點題目的情況，但此非確切影響的因素，更重要的如同測驗長度議題中所探討的，在於最大試題訊息量所對應之最適能力值是否能與定錨點相搭配。

表 4-35 第一次簡易測驗下採用 Easy(TEST1)定錨點之定錨點題目篩選結果

	定錨點				
	1	2	3	4	$\theta^{i}_{\max}$
定錨點題目	1.23955	0.32985	-0.57985	-1.48955	
1	.99198	.98195	.96003	.91469*	-3.3900
2	.96684	.94157	.89969	.83482*	-2.9986
3	.99663	.97122	.82246*	.58385	-.6512
9	.99856	.97764	.76028*	.39189	-.6795
10	.97876	.85775*	.51118	.29844	-.1904
11	.97815	.88969*	.64875	.44250	-.2872
12	.99533	.88534*	.40187	.30057	-.0428
14	.92905	.78407*	.54903	.37572	-.0339
15	.95267	.76783*	.43679	.28294	.0640
16	.97979	.84420*	.47380	.29885	-.0889
17	.98266	.81578*	.33668	.17583	-.0623
18	.98925	.97755	.95393	.90858*	-3.5170

19	.98466	.89547*	.57836	.30586	-.3846
20	.96313	.80641*	.50941	.37225	.0740
總定錨點題目數	0	9	2	3	

註：定錨點由左而右乃為第 1 至第 4 定錨點，括號內為其對應能力值，另在此仍以*代表.66 的定錨點題目篩選準則；左側之定錨點題目為相對於原始測驗之題號，而未達標準者，仍不列入表中。

表 4-36 第二次簡易測驗下採用 Easy(TEST2)定錨點之定錨點題目篩選結果

	定錨點				
	1	2	3	4	$\theta^{i}_{\max}$
定錨點題目	(0.7742)	(-0.1586)	(-1.0914)	(-2.0242)	
3	.99094	.90955*	.48605	.09931	-1.0058
4	.99266	.91957*	.50369	.12133	-.9940
5	.96622	.81686*	.46405	.24845	-.6160
6	.88722	.75769*	.55798	.34591	-1.1144
7	.99480	.94464*	.60864	.14879	-1.1820
8	.96738	.75948*	.35498	.22709	-.3671
9	.98662	.77979*	.22803	.12691	-.4216
10	.94019	.81558*	.55866	.27732	-1.1300
12	.83793*	.62294	.34950	.15543	-.4946
16	.93873	.83892*	.64124	.38665	-1.4478
17	.97207	.84154*	.45484	.13419	-.9131
19	.97679	.78346*	.26719	.08209	-.5590
總定錨點題目數	1	11	0	0	

註：定錨點由左而右乃為第 1 至第 4 定錨點，括號內為其對應能力值，另在此仍以*代表.66 的定錨點題目篩選準則；左側之定錨點題目為相對於原始測驗之題號，而未達標準者，仍不列入表中。