國立臺南大學
「教育研究學報」第 42 卷第 2 期(民國 97 年):67~82

國中基測量尺分數轉換議題探討

涂柏原
國立台南大學測驗統計研究所

摘　要

本文目的在於探討基本學力測驗各科量尺分數對照表經過微調時，對於考生總分名次排序的影響以及對於目前考生全錯或僅答對數題時，皆對應到量尺分數 1 分的情形，是否能夠加以改善。在所模擬的 5000 考生資料中，當原始分數與量尺分數的對照表經過微調之後，至少有 14.3%的考生之總分排名受到影響，排序名次降低的遠多於名次升高。在維持目前的將原始分數轉換成量尺分數以及量持分數維持在 1-60 分的架構之下，利用將 IRT 中的 1PL 模式所估計得到的考生能力值加以直線轉換以得到量尺分數的作法，似乎是可行的，雖然可能需要進一步的調整；而若以 3PL 模式所估計得到的考生能力值加以直線轉換以得到量尺分數的作法，有許多困難需要克服。

關鍵字：量尺分數、試題反應理論、基本學力測驗

壹、緒論

從 2001 年國中生基本學力測驗開始啓用至今，除了 2007 年五月份中央研究院的林妙香研究員指稱國中基測的量尺及等化程序有一些問題外（參見林妙香〔2007〕的研究報告），一般說來，社會大眾對於基本學力測驗並沒有特別的意見的。而歸納這幾年來社會大眾對於國中基測的量尺分數之主要評論或想法，主要有下列二者：（1）單科答錯一題時，考生所得到的量尺分數可爲是 55 分，與全對的考生所得到的 60 分，是否差異過大，以致於影響錯一題的學生之排名；（2）答對 10 題的學生可能與完全答錯的學生得到相同的量尺分數，都是 1 分。針對這兩個議題，本文將利用模擬資料來加以探討，期待獲致有價值的結果，可以作爲基測改良其量尺分數時之參考。在詳細說明本研究之目的前，筆者將先簡單地回顧基本學力測驗現行的量尺分數之產生方法。

一、基本學力測驗量尺分數產生的方法

假定一個測驗有 K 個試題，如果答對的題目得到 1 分，答錯是 0 分的話，一個考生的原始分數（raw score）是這 K 個二元計分（dichotomously scored）試題得分的和。假定隨機變項 X 代表原始分數，那麼在一個母群體裏 X 這個隨機變項（random variable）的值等於 i（$i = 0,1,...,K$）的機率函數是

$$\Pr(X = i) = \int_0^1 \Pr(X = i \mid \tau) g(\tau) d\tau , \qquad (1)$$

其中 τ 是母群的答對率真分數（proportion-correct true score），$0 \le \tau \le 1$，其機率密度函數爲 $g(\tau)$，而 $\Pr(X = i \mid \tau)$ 則是以真分數爲 τ 的考生所形成的原始分數之條件分配（conditional distribution）。根據 Kolen、Hanson 和 Brennan（1992）的描述，$g(\tau)$ 與 $\Pr(X = i \mid \tau)$ 的分配可以用強真分數模式（strong true score model）來加以估計。

國中生基本學力測驗的量尺分數是根據考生的答對題數原始分數加以轉換得到的，因此 $\Pr(X = i)$ 的取得，就成了計算「原始分數與量尺分數轉換表」時關鍵的重點。基本學力測驗量尺分數的建立方式是利用公式（1）得到 $\Pr(X = i)$ 之後，接下來利用 Kolen、Hanson 和 Brennan（1992）所描述的程序，透過利用正弦反函數轉換（arcsin transformation; Freeman & Tukey, 1950）來將各個分數點上之條件分配測量標準誤（conditional standard error of measurement, CSEM）加以近似化或穩定化（stablize），亦即使得各個分數點上的 CSEM 都大致相等；然後將答對題數原始分數轉換成 1-60 分的量尺分數。詳細的程序請參見 Kolen、Hanson 和 Brennan（1992）的文章，或是涂柏原（2005, 2007）的示範。

Kolen、Zeng、和 Hanson（1996）提到 $\Pr(X = i)$ 也可以用底下的公式表示

$$\Pr(X = i) = \int_{-\infty}^{\infty} \Pr(X = i \mid \theta) \varphi(\theta) d\theta , \qquad (2)$$

其中 θ 是母群的能力參數（ability parameter），其機率密度函數爲 $\varphi(\theta)$，而 $\Pr(X = i \mid \theta)$ 則是以能力爲 θ 的考生所形成的原始分數之條件分配。用這個公式時，$\varphi(\theta)$ 與 $\Pr(X = i \mid \theta)$ 的機率函數則可利用試題反應理論（item response theory, IRT）的模式來加以估計。其中，以能力爲 θ 的考生所形成的原始分數之條件分配 $\Pr(X = i \mid \theta)$ 可用 Lord 與 Wingersky（1984）的遞迴公式（recursion formula）來計算。涂柏原（2005, 2007）的研究結論是以公式（1）和公式（2）所得到的原始分數與量尺分數之對照表結果類似，但是如果原始分數的分配如果偏離單峰分配的情形比較嚴重

時，則以公式（2）所得到的 $\Pr(X=i)$ 會比較接近觀察得到的原始分數的分配。因此，他建議利用公式（2）來建立基本學力測驗的量尺分數是比較合適的方法。

目前利用公式（1）所得到的量尺分數對照表所招致的批評主要有兩點：（1）許多人認爲當考生只答錯一題時，所得到的量尺分數常常是 55 或 54 分，他們認爲比起全部答對的 60 分，少了太多了。如果量尺分數是 56 分甚或更高時，也許考生會有比較好的排名。（2）目前的量尺分數對照表，在低分的部分，常見的現象是全部答錯或是僅答錯少數幾題的考生，所對應到的量尺分數皆爲 1 分。部分人士建議應該讓答對題數少的這部分之量尺分數也有不同，以有效地的將考生的能力區別開來。

二、研究目的

因此，本文有兩個目的，第一個目的在於探討基本學力測驗各科量尺分數對照表經過微調時，對於考生總分名次排序的影響；具體的說是，若單科答錯一題時量尺分數從 55 分改變成 56 分時，對於考生最後的總分排序的影響爲何。第二個目的是在維持現有的原始分數轉換成量尺分數以及 1-60 分的架構不變之下，探討利用試題反應理論（IRT）中的三參數對數模式（three-parameter logistic model, 3PL）和單參數對數模式（one-parameter logistic model, 1PL）所估計得到的考生能力值，直接線性轉換成 1-60 的量尺分數時，對於上述第二種批評的現象，是否能夠有所改善。

針對上述的目的，本研究計分成兩大部分：（1）先以實際參加 91 年第 1 次基本學力測驗的考生資料中所抽取出來的 5000 人樣本之資料所估計得到的各科試題參數來模擬考生在各科的作答反應，以探討各科原始分數與量尺分數對照表若進行人工調整時，對考生量尺分數總分排序名次的影響。（2）以前一個步驟所產生的模擬資料進一步探討利用 1PL 和 3PL 模式所估計得到的能力參數值，加以線性轉換成爲量尺分數，來取代目前的方法之可行性。

貳、研究方法

一、資料產生的方法

要模擬每個考生在國文、英文、數學、社會和自然等五個科目上的答題反應，需要每個考生在各個科目上面的能力參數、以及各科的試題參數。在本研究中，各科的試題參數是由參加 91 年第 1 次測驗由學力測驗小組所提供的 5000 人樣本中得到的。而能力參數的產生則利用 Morgan（1984, p. 86 & p. 88）所提到的方法。

Morgan 的方法基本上如底下的描述。假如相同考生在五個科目上得分的相關矩陣爲 $\underline{R}$，則對 $\underline{R}$ 進行 Cholesky Decomposition，可得到一個下三角形（或上三角形）的矩陣 $\underline{A}$。若隨機從 $N(0,1)$ 抽樣得到 5000 個能力值作爲考生的國文能力，得到一個 5000×1 的矩陣或向量 $\underline{\theta_1}$，則依照此方式，我們可以得到英文能力 $\underline{\theta_2}$、數學能力 $\underline{\theta_3}$、社會能力 $\underline{\theta_4}$ 和自然能力 $\underline{\theta_5}$；其中 $\underline{\theta_1}$、$\underline{\theta_2}$、$\underline{\theta_3}$、$\underline{\theta_4}$ 和 $\underline{\theta_5}$ 之間是獨立的。若 $\underline{\Theta}=[\underline{\theta_1},\underline{\theta_2},\underline{\theta_3},\underline{\theta_4},\underline{\theta_5}]$ 爲 5000 名考生在五個科目上面的能力值矩陣，則 $\underline{A}\underline{\Theta}'$ 的轉置矩陣就是用來產生 5000 名考生在五個科目的答題反應所需的能力參數（視爲模擬資料考生之真實能力）。有了用來產生答題反應資料的能力值之後，可利用各科的試題參數和這些能力值，以 3PL 模式來產生各科的答題反應。

而各科之間的相關矩陣 $\underline{R}$ 的值是多少？因爲筆者手上所有的 91 年第 1 次基測資料不是同一個人作答所有科目的，因此特別請學力測驗小組再次提供實際參加 95 年第 1 次測驗的考生中 5000 人樣本在五個科目上之答題反應，根據這筆資料，得到五個學科的原始分數之間的相關以及五科

能力之間的相關，分別呈現在表 1 與表 2 之中，二者十分類似。在本研究中，以表 2 中所列的各科能力之間的相關爲 $\underline{R}$ 。

表1. **95年第一次基測各科原始分數之間的相關矩陣**

	國文	英文	數學	社會	自然
國文	1				
英文	.782**	1			
數學	.771**	.773**	1		
社會	.865**	.784**	.806**	1	
自然	.833**	.807**	.873**	.877**	1

** p<.01.

表2. **95年第一次基測各科能力估計值之間的相關矩陣**

	國文	英文	數學	社會	自然
國文	1				
英文	.778**	1			
數學	.782**	.792**	1		
社會	.847**	.767**	.794**	1	
自然	.834**	.800**	.871**	.859**	1

** p<.01.

二、資料模擬及後續分析的程序

根據本研究第一個目的所進行的資料模擬與後續的分析處理之步驟如下：

（1）利用前一段所描述的方法，得到用來產生答題反應的考生在各科之能力值，這五科之能力值之相關如表 3 所示，與表 2 的數據相比較，顯示模擬所得到的各科能力值之間的相關結構與實徵資料十分相似。

表3. **模擬產生的各科能力估計值之間的相關**

	國文	英文	數學	社會	自然
國文	1				
英文	.779**	1			
數學	.780**	.788**	1		
社會	.843**	.763**	.783**	1	
自然	.834**	.808**	.873**	.850**	1

** p<.01.

（2）以所得到的5000名考生在各個科目上的能力值，以及各科的試題參數，利用3PL模式來模擬產生5000名考生在5個學科上之答題反應資料。

（3）利用Bilog-MG來對所產生的資料進行試題參數以及能力參數的估計，並取得各人的答對題數原始分數、能力估計值、以及考生能力的實徵分配。

（4）因爲用廣義的beta-binomial的方法（即現行的基本學力測驗量尺分數計算方法）和用IRT的方法來將原始分數轉換成量尺分數所得到的結果基本上是一樣的。因此，本研究利用IRT的方法（參見Kolen , Zeng & Hanson, 1996; 涂柏原，2005，2007）來將各科答對題數原始分數轉換成量尺分數之對照表建立起來。

（5）根據原始分數與量尺分數對照表來計算5000名考生的各科量尺分數以及五科量尺分數之總

分。

（6）爲了回答本研究第一個目的所欲探討的問題，也就是量尺分數對照表微調對於考生總分排名的影響，因此對國文與數學兩科做了一點點的改變，如表4所示。國文科的原始分數48分原來對應到量尺分數52分，現在將之改成53分，而原始分數49分之量尺分數由55改成56分；而數學科的原始分數30分的量尺分數由55改爲56分。這裡所操弄的基本上就是若該科僅答錯一題時，所對應的量尺分數應該爲多少的問題。其餘的三科維持不變，然後重新計算每一個考生這兩科的量尺分數以及量尺總分。

表4. **量尺分數的變更情形**

	原始分數	量尺分數	微調後之量尺分數
國文科	48	52	53
	49	55	56
數學科	30	55	56

（7）計算每一個考生這二種排名的差異（未更動之總分排名減去更動後總分之排名），然後計算此排名差異之次數分配。

而根據本研究第二個目的所進行的資料分析處理之步驟如下：

（1）將本研究第一個部分所產生的5000人考生在五個科目上面的答題反應資料利用IRT的3PL模式來估算試題參數以及考生能力參數。

（2）直接將前一個步驟所得到的能力值乘上10，加上30，來進行線性轉換，經過如此的轉換，理論上所得到的量尺分數應該是在1-60之間。

（3）將本研究第一個部分所產生的5000人考生在五個科目上面的答題反應資料利用IRT的1PL模式來估算試題參數以及考生能力參數。

（4）直接將前一個步驟所得到的能力值乘上10，加上30，來進行線性轉換，經過這樣子的轉換，理論上所得到的量尺分數也應該是在1-60之間。

（5）探討這兩種直接將考生能力估計值加以線性轉換放大的方式，是否能夠成爲基本學力測驗量尺分數之替代計算方法。

參、結果與討論

一、分數微調對排名的影響

表 5 是同一個考生在兩種量尺分數總分排名次序變動情形的次數分配表，從中可以發現在模擬的 5000 人次中，有 3529 人的名次沒有變動，因此名次變動的有 1471 人次，佔全部的 29.4%，這個比率是有些驚人。因爲表 5 中的差異是指第一種量尺分數的總分排名來減去第二種排名，且分數越高的名次越少（或越小、越低），因此負數表示在第一種量尺分數之計算方式下，名次排在比較前面。由表 5 的結果發現，對於現行的量尺分數來說，將量尺分數對照表微調之後，因此而造成名次往前的僅有 5.72%，而名次退後的卻有 23.7%。

表5. **兩種量尺總分排名差異的次數分配表（調整國文、數學兩科）**

差異	次數	百分比
-7.50 ~ -5.01	102	2.04
-5.00 ~ -.01	1083	21.66
.00	3529	70.58
.01 ~ 5.00	102	2.04
5.01 ~ 10.00	79	1.58
10.01 ~ 15.00	48	0.96
15..01 ~ 20.00	29	0.58
20.01 ~25.00	18	0.36
25.01 ~ 30.00	7	0.14
30.01 ~ 38.50	3	0.06
小計	5000	100.0

表6. **兩種量尺總分排名差異的次數分配表（僅調整數學）**

差異	次數	百分比
-3.50 ~ 0.01	582	11.64
.00	4333	86.66
0.01 ~ 5.00	37	0.74
5.01 ~ 10.00	24	0.48
10.01 ~ 15.00	16	0.32
15.01 ~ 20.00	6	0.12
20.01 ~ 23.00	2	0.04
小計	5000	100.0

爲了對此現象能有更清楚的瞭解，因此筆者再試一次僅改變一個科目的情形，因此底下所嘗試的是維持其他四科的量尺分數對照表不變，只改變數學，而且只是像表4所呈現的，只調整答錯一題時的量尺分數（由55變成56）。再重新計算量尺分數總分之後，排名差異的次數分配如表6所呈現的。

只調整數學一科時（將答錯一題時之量尺分數由55分改成56分），發現5000人中有4333人的量尺總分排名沒有改變的，改變的有667人（佔14.34%）；如同前一段所看到的，因爲量尺分數的改變而排名往上提升的遠少於往下降的人數。當然，對於這種現象，並未表示答錯一題的量尺分數由55分調整成56分是不好的，而是說明了量尺分數之決定的確是一件很重要、需要非常慎重的事情。因此該如何處理比較好，需要進一步更詳細的研究。

二、尋求新的量尺分數產生方法

如前面所提到的，現行的量尺分數對於僅答對少數題目的考生所給予的量尺分數都是1分的問題，例如，國中基測小組所公告的96年的量尺分數對照表中，社會科答對題數爲17題或17題以下的量尺分數皆爲1分（參見表7），這種情形招致了一些批評，的確是有改善的空間。爲了探討這種現象改善的可能性，因此進行了這個部分的研究。然而，筆者進行此一部分的研究之立場是希望在維持以原始分數轉換成量尺分數以及1-60分的架構之下，找出比現行的方法更好的量尺分數產生方法。

表7. **九十六年國中基測各科量尺分數爲1的最高答對題數**

	國文	英文	數學	社會	自然
最高題數	12	14	6	17	6

註：整理自http://www.bctest.ntnu.edu.tw/9602score.htm，2007年8月23日。

汪慧瑜、余民寧（2006）曾建議利用IRT中的3PL模式來估計考生的某一科能力值（θ），直接用IRT所估計出來的能力參數加以線性轉換至1-60分，以得到每一個考生的量尺分數。此一建議，從學理上來看是十分有道理的，但是在實務上面之可應用性如何，是需要進一步加以探討的。因此，本文的第二個部分，將一併探討直接利用IRT所估計得到的能力值，加以線性轉換以得到基本學力測驗之量尺分數之可行性。

（A）利用3PL模式估出考生能力值，直接線性轉換

由本研究的模擬資料，各個考生的答對題數之原始分數和能力估計值（利用3PL模式估計得到）的描述統計資料分別呈現於表6和表7之中。

表8. **五個科目的原始分數之描述統計**

	最小值	最大值	平均數	標準差	題數
國文	5	50	31.49	11.175	50
英文	4	45	22.43	11.132	45
數學	1	31	15.46	7.413	31
社會	2	63	37.96	13.341	63
自然	4	58	33.23	13.265	58

表 9. **五個科目的能力值（3PL）之描述統計**

	最小值	最大值	平均數	標準差
國文	-2.7280	2.1678	-.0144	.9649
英文	-1.8254	2.3334	.0006	.9584
數學	-1.7752	2.3514	.0084	.9413
社會	-2.7534	2.6433	.0066	.9663
自然	-2.7612	2.3631	.0005	.9595

由表8的數據可以得知，每一個科目都有全部答對的考生，但是從表9當中可以看到各科考生的能力估計值最大只到2.6左右，所以如果用 $SS = 10 \times \theta + 30$ 的方式來得到量尺分數 SS，則即使理論上的 θ 是具有標準常態分配，轉換後的量尺分數也應在1-60分之間；但是，實際上經過轉換之後*SS*的最大值卻不會等於60，除非轉換之後再利用人工加以調整，這樣的情形會與一般的期待有落差。根據現行的作法，大家會期待全部答對的學生會得到60分的量尺分數。

（B）利用1PL模式估出考生能力值，直接線性轉換

因爲目前基本學力測驗的等化應用了1PL模式，因此筆者接著就嘗試了用1PL模式所提供的能力估計值來進行上述的線性轉換，以得到量尺分數。前述模擬資料經過以Bilog-MG程式來進行1PL模式的試題和能力參數估計之後，得到5000名考生能力估計值之描述統計資料如表10所示。

表 10. **五個科目的能力值（1PL）之描述統計**

	最小值	最大值	平均數	標準差
國文	-1.7982	2.4016	.00016	.9618
英文	-1.2182	2.7620	.00018	.9694
數學	-1.3943	2.6870	.00025	.9497
社會	-2.7999	2.6934	.00011	.9680
自然	-1.8895	2.6134	.00020	.9695

從表 10 的資料中，可以發現若直接將能力值乘上 10，然後加上 30 的話，最大值將只有 57.6（英文科），也就是完全答對的考生之量尺分數爲 57.6，不是目前所用的系統中之 60 分。所以，若是想要用估算出來的能力值加以線性轉換的方式來計算考生的量尺分數，則必須再進一步處理。利用 1PL 模式所估計得到的能力值乘上 10 再加上 30 的作法，所得到的分數之描述統計資料如表 11 所示。

表 11. **利用 1PL 所估出之能力值直接線性轉換得到的量尺分數之描述統計資料（未調整）**

	最小值	最大值	平均數	標準差
國文	12.02	54.02	30.0016	9.6184
英文	17.82	57.62	30.0018	9.6940
數學	16.06	56.87	30.0025	9.4966
社會	2.00	56.93	30.0011	9.6801
自然	11.10	56.13	30.0020	9.6949

筆者決定先嘗試將各科的平均數固定在表 11 所呈現的平均值上面，然後依比率將最高分放大至 60 分，因此最低分也依比率往下修，經過這樣的處理，各科量尺分數的最大值就變成了 60 分，也就是全部答對的考生之量尺分數是 60 分了。經過這樣的調整之後，各科量尺分數之描述統計資料如表 12 所示，而各科答對題數原始分數與量尺分數之對照表則呈現於表 13 之中。可以輕易的看到這種方式所建立的量尺分數對照表與現行的對照表之差異處，各科最低幾分的原始分數所對應到的量尺分數不是 1 分，而是其他的分數。從表 10 各科能力估計值的最小值可以看出各科用這種方法所得到的最低量尺分數不是 1 分的原因，尤其是英文科經過轉換後的量尺分數是 17 分（16.77 四捨五入之後變成 17）。就某些角度來說，這樣的量尺分數換算方式也許會比目前的好，不會讓許多低能力的考生無論答對三題或五題，皆得到相同是 1 分的量尺分數。

表12. **利用1PL所估出之能力值直接線性轉換得到的量尺分數之描述統計資料（調整後）**

	最小值	最大值	平均數	標準差
國文	7.54	60.00	30.0016	12.0131
英文	16.77	60.00	30.0018	10.5294
數學	14.43	60.00	30.0025	10.6030
社會	1[a]	60.00	30.0011	10.7836
自然	8.31	60.00	30.0020	11.1309

註：[a]經上、下修之後，原是-1.19 分，當然得調整爲 1 分。

表 13. **利用 1PL 所估出之能力值直接線性轉換得到的量尺分數對照表（調整後）**

原始分數	國文	英文	數學	社會	自然
0					
1			14		
2			15	1	
3			16		
4		17	17		8
5	8	17	17	5	9
6	8	18	18	6	10
7	9	18	19	7	11
8	9	19	20	8	12
9	10	19	21	8	12
10	11	20	22	9	13
11	11	20	23	10	14
12	12	21	25	11	14
13	13	22	26	11	15
14	14	22	27	12	16
15	14	23	28	13	16
16	15	24	30	13	17
17	16	24	31	14	18
18	17	25	32	15	18
19	18	26	33	16	19
20	19	27	35	16	20
21	19	28	36	17	20
22	20	28	37	18	21
23	21	29	39	19	22
24	22	30	40	19	22
25	23	31	4	20	23
26	24	32	45	21	24
27	24	33	47	22	24
28	25	33	50	22	25
29	26	34	53	23	26
30	27	35	56	24	26
31	28	36	60	24	27

（續下頁）

表 13. （續）

原始分數	國文	英文	數學	社會	自然
32	28	37		25	28
33	29	38		26	29
34	30	39		26	29
35	31	40		27	30
36	32	41		28	31
37	33	42		28	31
38	35	43		29	32
39	36	45		30	33
40	37	46		31	34
41	38	48		31	34
42	40	51		32	35
43	41	53		33	36
44	43	57		34	37
45	44	60		34	38
46	47			35	39
47	49			36	40
48	52			37	41
49	56			38	42
50	60			39	43
51				40	45
52				41	46
53				42	48
54				43	50
55				44	52
56				45	54
57				47	57
58				49	60
59				50	
60				53	
61				55	
62				57	
63				60	

註：1. 1PL 模式的特性是相同原始分數的考生具有相同的能力估計值，表中出現不同原始分數對應到相同量尺分數純粹是四捨五入的緣故。2. 低分部份的空白處是初步分數轉換後，需要利用外插法或內插法才能得到的。

因爲表 12 及表 13 的結果是以 5000 人的樣本得到的，筆者手上恰好有 91 年第 1 次英文科所有考生的答題反應資料（共計 299714 人），因此以這份資料再一次試試以 IRT 的 3PL 與 1PL 模式所估計得到的能力值直接線性轉換成量尺分數之可行性。

經過次數分配分析之後（未列表呈現），發現從 0 至滿分共 46 種分數（91 年第 1 次的英文科有 45 道試題）的次數皆不等於 0，而用原始分數、3PL 和 1PL 模式所估計得到的能力值之描述統計資料呈現於表 14 之中。91 年第 1 次的英文科共有 1159 人答對所有的試題，這些人的 3PL 模式之能力估計估計值 $\hat{\theta}_{3PL}=1.9333$，而 1PL 能力之估計值爲 $\hat{\theta}_{1PL}=2.2198$，所以如果直接用 $SS=10\times\hat{\theta}+30$ 的方式來將這些結果轉換至 1-60 的話，則這些滿分的考生之量尺分數將無法得到 60 分。此結果與前面用 5000 人樣本所得到的類似，如果要用能力值直接轉換到 1-60 分之量尺分數，可能需要再人爲調整。

表14. 九十一年第一次英文科考生得分之描述統計

	人數	最小值	最大值	平均數	標準差
原始分數	299714	0	45	23.22	11.708
3PL能力值	299714	-1.6281	1.9333	-.00184	.97602
1PL能力值	299714	-3.4773	2.2198	-.00045	.97566

從理論可以得知，3PL模式的能力估計值受到試題的鑑別力參數（*a*參數）和難度參數（*b*參數）的影響，因此具有相同答對題數的考生之能力估計值未必會一樣，因此答對題數原始分數與用能力值轉換得到的量尺分數會呈現多對多的現象，這種情形可以從表15當中得到印證，表15中的數字是那種原始分數與量尺分數組合的次數（或人數）。

表 15. 以 **3PL** 的能力值線性轉換時之原始分數與量尺分數對照表（部分）

	量尺分數								
原始分數	5	6	7	8	9	10	11	12	13
0	3382	0	0	0	0	0	0	0	0
1	8	3	0	0	0	0	0	0	0
2	15	2	6	3	0	3	0	0	0
3	46	10	34	15	7	5	4	1	0
4	102	38	74	72	33	45	16	18	4
5	174	82	169	225	100	117	90	54	26
6	286	165	294	393	266	291	255	197	113
7	361	279	420	585	541	516	524	460	334
8	409	355	620	796	833	807	903	825	654
9	416	409	608	813	1081	980	1206	1164	1040
10	1351	404	518	755	1105	1119	1380	1458	1420
11	199	377	429	730	930	1010	1323	1423	1514
12	191	240	280	495	2081	820	1076	1363	1461
13	129	144	162	313	473	658	760	949	1242
14	48	144	70	137	245	347	447	646	801
15	4	43	60	96	120	194	252	338	477
16	1	15	18	22	52	71	120	177	256
17	0	2	11	9	26	27	39	66	102
18	0	2	1	1	7	9	27	26	32
19	0	0	0	0	0	3	5	6	20
20	0	0	0	0	1	0	2	1	3
21	0	0	0	0	0	1	0	0	1
22	0	0	0	0	0	0	0	0	1
23	0	0	0	0	0	1	0	0	0
24	0	0	0	0	0	0	0	0	0

對基本學力測驗來說，如果要維持目前容易與家長溝通的答對題數原始分數與量尺分數對照的型態，利用 3PL 模式所得到的能力估計值來加以線性轉換的作法，可能會面臨到不易製作原始分數與量尺分數對照表的問題。因此，筆者還是回到 1PL 模式的情況。利用 1PL 模式中，答對題數是能力參數的充分統計數這個特性，無論答對的試題是否相同，只要具有相同的答對題數，則必定有相同的能力估計值的特點，要製作原始分數與量尺分數之對照表，就比較容易。

表 16 呈現了原始分數與用 1PL 模式所估計得到的能力值和將能力值加以線性轉換所得到的量尺分數。表中的「量尺分數 1」是能力估計值線性轉換之後的結果（直接乘上 10，然後加上 30），因爲最高分只有 52.2 分，要再加上 7.8 才會等於 60，因此對於此處小於 0 的部分就不加以處理。爲了讓滿分等於 60，因此全部加上 7.8 分，得到「量尺分數 2」（也就是，在此例中只有平移，沒有改變標準差），而「量尺分數 3」則是「量尺分數 2」四捨五入之後的結果。由「量尺分數 2」那一欄，可以清楚看到不同的原始分數對應到不同的量尺分數，但是因爲四捨五入的原因，使得「量尺分數 3」那一欄，產生許多不同原始分數對應到相同量尺分數之情況。然而，至少「量尺分數 3」看起來比現行的量尺分數，應當會讓更多的人接受。

表16. **九十一年第一次英文科的原始分數與量尺分數對照表（1PL）**

原始分數	能力估計值	量尺分數 1	量尺分數 2	量尺分數 3
0	-3.47731	-4.77	3.03	3
1	-3.32226	-3.22	4.58	5
2	-2.68658	3.13	10.93	11
3	-1.42521	15.75	23.55	24
4	-1.15912	18.41	26.21	26
5	-1.10492	18.95	26.75	27
6	-1.06595	19.34	27.14	27
7	-1.02878	19.71	27.51	28
8	-0.99218	20.08	27.88	28
9	-0.9557	20.44	28.24	28
10	-0.9189	20.81	28.61	29
11	-0.88125	21.19	28.99	29
12	-0.84209	21.58	29.38	29
13	-0.80049	22	29.8	30
14	-0.75521	22.45	30.25	30
15	-0.70462	22.95	30.75	31
16	-0.64669	23.53	31.33	31
17	-0.57938	24.21	32.01	32
18	-0.5015	24.98	32.78	33
19	-0.41402	25.86	33.66	34
20	-0.32074	26.79	34.59	35
21	-0.22728	27.73	35.53	36
22	-0.13846	28.62	36.42	36
23	-0.05645	29.44	37.24	37
24	0.019144	30.19	37.99	38
25	0.090146	30.9	38.7	39
26	0.158649	31.59	39.39	39
27	0.226482	32.26	40.06	40

（續下頁）

表16. **（續）**

原始分數	能力估計值	量尺分數 1	量尺分數 2	量尺分數 3
28	0.29507	32.95	40.75	41
29	0.365472	33.65	41.45	41
30	0.438493	34.38	42.18	42
31	0.514778	35.15	42.95	43
32	0.594885	35.95	43.75	44
33	0.679308	36.79	44.59	45
34	0.768463	37.68	45.48	45
35	0.862639	38.63	46.43	46
36	0.961956	39.62	47.42	47
37	1.066383	40.66	48.46	48
38	1.175844	41.76	49.56	50
39	1.290421	42.9	50.7	51
40	1.410646	44.11	51.91	52
41	1.537876	45.38	53.18	53
42	1.674832	46.75	54.55	55
43	1.826566	48.27	56.07	56
44	2.00242	50.02	57.82	58
45	2.219805	52.2	60	60

註：利用299714人資料得到的。

至此，筆者的想法是，在維持現有的1-60分以及原始分數與量尺分數相對應的架構之下，用IRT的$\hat{\theta}$來轉換成爲量尺分數是件不容易的事。用1PL模式是稍微容易一些，但是，也無法簡單的轉換即可得到，還需要進一步的人爲調整。因爲各科題數不同，所以各科不大可能有相同的原始分數與量尺分數對照表，且各科的量尺分數雖然平均數是可以調整成爲30分，各科的標準差應該仍是不同的。

肆、結論與建議

本文的第一個目的在於探討基本學力測驗各科量尺分數對照表經過微調時，對於考生總分名次排序的影響。由第一個部分所獲致的結果，我們初步有的瞭解是當原始分數與量尺分數的對照表被微調時，的確有一些考生的總分排名會受到影響。以本研究爲例，當只微調數學一科時，答錯一題時之量尺分數從 55 改成 56 分，結果有 14.34%的考生之總分排名受到影響。若同時改變國文與數學兩科時，則 29.4%的考生受到影響。然而，要對此現象有更完全的了解，需要進一步的研究。

而就直接利用從 IRT 模式中所估計得到的能力值，加以線性轉換以得到 1-60 分的量尺分數這個部分，在維持目前 1-60 分以及原始分數與量尺分數對照表的架構下，用 3PL 模式來估計能力參

數，然後加以線性轉換之作法，是有其困難議題需要解決。因爲受到鑑別力參數（*a* 參數）的影響，雖然答對的題數相同，但是答對的題目不同的考生，其能力估計值是不同的；而且要得到一個合理的原始分數與量尺分數的對照表，在 3PL 模式的情況中是不容易的。而從表 11 和表 14 來看，用 1PL 模式所得到的能力值加以線性轉換之後，再加以微調，是可以得到一組看起來不差的量尺分數對照表，且原來的答對題數不多時都對應到量尺分數 1 分的問題也可以得到解決。再加上目前學力測驗小組在進行一年兩次測驗成績等化的工作時，也是利用 Rasch 模式，因此利用從 1PL 模式所估計得到的能力值來轉換成量尺分數應是可以考慮的。至於表 13 中低分部分的空白處，在使用全部參加第一次測驗的資料時，應該就可以全部估計得到（如表 16 那樣），若是仍然有缺的，也許可以考慮用差補的方法解決。另外，不同的原始分數對應到相同的量尺分數的問題，由於這是四捨五入所造成的，因此如果真的介意的話，可以利用非整數的量尺分數加以解決。當然，如果國中基測真的想要將量尺分數之產生方式改誠如本文所建議的，那麼在正式實施之前，是需要更多的研究的。

參考文獻

林妙香（2007）。**國中基測量尺及等化程序缺失**。中央研究院技術報告 C-2007-01。2007 年 6 月 20 日，取自 http://ww3.stat.sinica.edu.tw/library/c_tec_rep/2007-1.pdf。

汪慧瑜、余民寧（2006）。國中基本學力測驗量尺分數的另類表示方法。**測驗學刊**，**53**(2)，205-238。

涂柏原（2005）。如何將原始分數轉換成量尺分數─以國中生基本學力測驗爲例。**測驗學刊**，**52**(2)，1-28。

涂柏原（2007）。國中生基本學力測驗量尺分數轉換之實徵研究。**教育研究學報**，**41**(1)，61~77。

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics*, *21*, 607-611.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*, 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*, 129-140.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, *8*, 453-461.

Morgan, B. J. T. (1984). *Elements of simulation*. New York: Chapman and Hall.

投稿日期：96 年 9 月 12 日
修正日期：97 年 4 月 1 日
接受日期：97 年 4 月 1 日

The Scale Score Transformation Issues of the BCTEST

Bor-Yaun Twu
National University of Tainan Graduate Institute of Measurement and Statistics

Abstract

Two issues of BCTEST scale scores were investigated in this study. The first of them is that the scale score an examinee will receive when he/she correctly answers all but one items is five or six points less than the examinees who get all items correctly. Some suggest that the rank order of examinees would be different if scale score had been one or two points higher for those who nearly get all items correctly. The second issue is that the scale score for the examinees who answered only a few items correctly tend to be the same, always 1. The purposes of this study are to investigate the effect of adjusting the scale score conversion table on the rank order of examinees' total score and to explore if the ability parameter obtained from the IRT calibration could be used to construct a meaningful score scale. The results indicated that at least 14.3% of examinees would be ranked differently when the scale score conversion table for one of the five subjects was modified. And it seems that the ability estimate given by 1PL model can be linearly transformed into 1-60 scale scores, although further adjustment is needed. For using the ability estimate obtained by 3PL model to form the scale scores, there still are some difficulties needed to overcome.

Key word: scale score, item response theory, BCTEST