國中基測量尺及等化程序缺失

林妙香研究員

中央研究院統計科學研究所

日期: 96 年 5 月 22 日

摘 要

基測實施六年來，公眾隱約可感受到基測計算方式可模糊前段及後段學生差距；一般大眾在艱澀學理如同黑箱作業下，無法判斷中段學生差距是否也被模糊了。由分發結果而言，中段學生差距之模糊可造成公、私立學校分發落差，連帶地使家長承擔不同對等的經濟負擔。此負擔年限短則三年長則七年。眾所周知，三年的私立高中職、五專學雜費至少20-30萬,而四年私立大學學雜費、生活費更是一筆可觀數目。試想台灣已成M型的社經結構，就讀私立學校所須費用對中、低收入家庭而言，無疑是雪上加霜。

鑑於一般大眾無法針砭基測量尺分數計算方式是否影響考生及家長權益，本文檢驗及分析基測量尺分數是否滿足程序正義的計算方式。此程序正義包括三個參照基準點:一為正確的學理應用；二為內部計分流程應與公佈於大眾的「計分遊戲規則」相符相合；三為擇優政策隱含的名次排擠效應的資訊應該透明化。

本文主要的發現:心測中心在計算量尺分數的過程中，存在著一些不尋常的地方：一是各個考科在計分過程中的最高分並非設定在 60 分，且各個科目在各個年度都不相同；二是第二次測驗量尺並無等化(equating)步驟，只是單純進行分數連結(linking)。

關鍵詞: 量尺分數，量化，IRT-等化，連結，強真分數理論，均等測量標準誤量尺分數型態。

一、序言

日前部分縣市教育局長指出國民中學學生基本學力測驗(簡稱基測)的量尺分數打擊弱勢學生，呼籲教育部全面檢討基測計分方式，教育部長杜正勝也表示檢討基測方案的時候了。

國中基測雖然已經實施六年，大眾及教育人員無法理解量尺分數的計算方法。例如，為何在某些基測科目答對一題或10多題，量尺分數都是1分。另一方面，雖然有許多學者及家長質疑**量尺分數的算法、意義、及公平性**，但在無法驗證其計算方式情況下，只能將質疑埋沒心中。

基測自90年度實施以來，師大心理與測驗發展中心(以下簡稱心測中心)只公佈答對題數與量尺分數之對照表，而將考生的試卷**原始分數分佈列為國家機密，不對外公佈。換言之，當**考生的**原始分數分佈**「能見度」為零時，驗證工作是很難進行的。

回溯90年度基測開始實施時，基測推動工作委員會以一序列專文及答客問闡釋基測分數的意義及學理基礎，要求大眾相信基測量尺分數轉換及等化過程之公信力。 基測成績作為高中、高職及五專分發依據，是可影響30幾萬國中考生的未來教育環境品質及經濟負擔，因而基測成績計算過程應該透明化的，且可受驗證的。

基測實施六年來，公眾隱約可感受到基測計算方式可模糊前段及後段學生差距；一般大眾在艱澀學理如同黑箱作業下，無法判斷中段學生差距是否也被模糊了。由分發結果而言，中段學生差距之模糊可造成公、私立學校分發落差，連帶地使家長承擔不同對等的經濟負擔。此負擔年限短則三年長則七年。眾所周知，三年的私立高中職、五專學雜費至少20-30萬, 而四年私立大學學雜費、生活費更是一筆可觀數目。試想台灣已成M型的社經結構，就讀私立學校所須費用對中、低收入家庭而言，無疑是雪上加霜。

鑑於一般大眾無法針砭基測量尺分數計算方式是否影響考生及家長權益，筆者檢驗及分析基測量尺分數是否滿足程序正義的計算方式。此程序正義包括三個參照基準點:一為正確的學理應用；二為內部計分流程應與公佈於大眾的「計分遊戲規則」相符相合；三為擇優政策隱含的名次排擠效應的資訊應該透明化。

二、研究背景導言

研究心路。筆者在92年度中央研院統計所科普演講，曾指出三項論點。其一為：心測中心對量尺分數及等化學理的宣導文是過於『誇大的』；其二為：1-60分量尺分數作為基測子學科共同量尺的學理限制；其三為：基測題庫品質優劣及等化程序正確與否連帶地使擇優政策美意淪為名次排擠效應。

筆者因而向國科會科教處申請計畫(林妙香)先行發展心測中心原本想用的四參數或五參數的beta-binomial理論模式(林妙香及謝育泰)，繼而檢驗其他方面質疑之處。筆者亦向基測學力小組申請研究資料；抽樣資料為90-93年度考生之二次基測成

績，樣本人數各年度約三萬多筆，是依據全國18個考區按比率隨機抽樣的。此外，筆者亦向基測學力小組申請90-93年度第一次及第二次基測整體統計資料。

在從事這項研究的過程中，最讓筆者感到困惑的部份是：一直無法依循心測中心對外公佈的量尺計分方式(請閱參考文獻:8)，來重新建立各年度基測中心所公佈的量尺分數。一開始，筆者花了許多的時間來重新檢視所有與基測量尺理論以及計算公式的相關公開文件，也研讀了其理論建構的所有參考論文，都無法找出問題的所在。直到交叉比對了抽樣資料的統計特性與申請的整體資料的統計特性後，赫然發現的是，並非筆者對於心測中心的計分方式的理解有誤，而是在其內部計算量尺分數的過程中，存在著一些不尋常的地方：一是各個考科在計分過程中的最高分並非設定在 60 分，且各個科目在各個年度都不相同；二是第二次測驗量尺並無等化(equating)步驟，只是單純進行分數連結(linking)。而這些不尋常的地方，都沒有看到心測中心有任何的公開說明。筆者並不願意就此認定這其中存在著「弊端」，但其事實上是影響到了所有考生的權益，筆者也會在本文中對此「影響」的部份做一個說明。

<u>與本文相關的基測特色導引</u>。基測一年施測兩次，考生可報考兩次或一次；考生兩次測驗分數可擇優使用，作為登記分發入學高中、高職、五專之依據。在此一提，基測自實施以來，其宣稱門檻功用已呈「死體化」，無法運作；基測一如「往昔聯招巨獸」，其成績仍是用來區分考生相對能力的表現，仍是扮演分發排序考生的功用。

基測為包含五個科目的一套測驗(test battery)，科目內容包含國文、英文、數學、自然、社會等學科。各科目測驗題數不等，少則 31 題（數學)多則 66 題(社會)。心測中心以 1-60 分的**量尺分數取代**學科試卷的原始分數作為考生測驗成績，因此分發入學依據是依照五學科**量尺分數總分(5-300)作**排序的。心測中心為配合基測一年兩次的施測作業，對五學科建立學科題庫，以編組該年度所需的二本試卷題本(two forms of the same subject test)。心測中心只對第一次基測建立「原始分數-**量尺分數**」對照表。第二次基測的原始分數透過測驗等化技術對應至第一次基測的**量尺分數**對照表。(詳閱參考文獻 1-7)

<u>本文探究方向</u>。心測中心宣導其 1-60 分的**量尺分數**是反映**均等測量標準誤量尺分數型態**，是仿照 The ACT Assessment 建立量尺的方法，**具有分數精準**特性。心測中心宣導其題庫裡是存放着優良試題及正確試題參數資訊(試題難度)，因此年度所需的前後二次學科試卷題本的平均難度是相近的。心測中心宣導其 IRT-等化程序使兩次基測結果仍然是客觀、公平、可以互相比較。本研究針對此三方面的宣導文探究下列主題:

1)考生的基測成績通知單的量尺分數的計算是正確的嗎?
2)各考科的 1--60 分的**量尺分數代表對**分發總分影響影響力是一樣的嗎?
3)心測中心的等化程序是真正的 IRT-等化程序嗎?
4)擇優政策是否不利於只報考第一次基測的考生?
5)心測中心真的有建構題庫嗎?

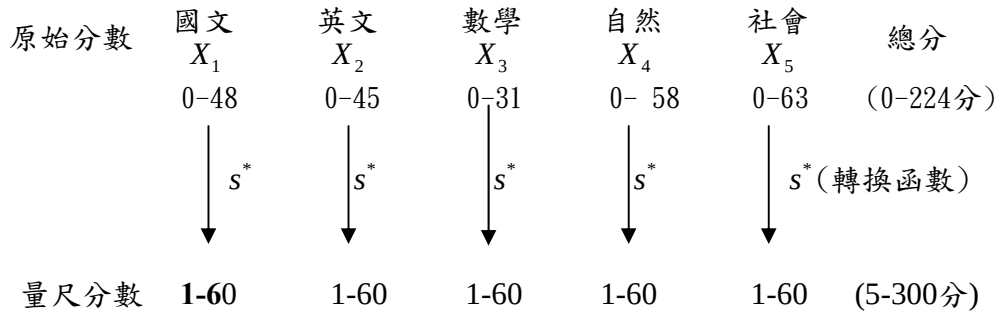
筆者重述探究主題的資料來源有二：A) 心測中心所公佈的 90-93 年度試題研究報告及整體統計資料，亦即全國考生第一次、第二次基測的統計資料，包括計算量尺分數所需的參數值，考生的量尺分數(公告的與內部作業的)；以及 B) 90-93 年度抽樣資料，各年度約三萬多筆。

三、 基測量尺分數計算方式的灰色地帶及其效應

3.1 基測量尺分數建立示意圖

基測各子學科試卷題數不等，少則 31 題 (數學)多則 66 題(社會)；各子學科試卷題本的題型皆為四選一的選擇題，試題給分方式為:答對 1 分，答錯 0 分。因此考生的試卷答對題數即為考生的原始分數。若以五學科原始分數加總作為分發入學依據，則題數最少的數學學科相對地對分發總分影響力最小，這對數學成績好的考生是不公平的。

心測中心在第一次基測考完後，採用量化(scaling)方法將子學科的原始分數轉換為 1-60 的量尺分數，如圖一所示。令 X_1、X_2、X_3、X_4、X_5 分別代表基測子學科的原始分數(92 年度基測為例)：



圖一、基測學科量尺建置示意圖

由示意圖所示，子學科的原始分數透過 s^* 函數皆轉換為1-60分的量尺分數。由示意圖所示，讀者會認為子學科對量尺分數總分(5-300)的影響力是一致的，是一樣的，因為五學科對量尺總分相對的比重是一樣的。**其實不然**，因為考生的基測成績通知單的量尺分數不是規規矩矩地按照心測中心所宣稱的公式計算的，更嚴重的是 s^* 函數不會使子學科量尺分數的標準差是一樣的。

3.2 心測中心的均等條件測量標準誤量尺分數

心測中心引經據典宣稱1-60分的**量**尺分數是均等測量標準誤的量尺分數型態，是採用Kolen (1992)的計算程序，將考生的基測原始分數轉換為量尺分數。我們先了解均等條件測量標準誤量尺分數的統計內涵及The ACT (1989)作法。

The ACT 的均等測量標準誤的量尺分數型態。均等條件測量標準誤量尺分數轉換統計內涵有三，一為以的強真分數理論(Lord, 1965；1968)的 compound binomial–beta 模式(1)描述考生原始分數(X)分配

$$\Pr(X=x)=\int_a^b \Pr(X=x|\zeta)g(\zeta)d\zeta\,, \qquad (1)$$

$\Pr(X=x|\zeta)$: compound binomial distribution

$g(\zeta)$: four-parameter beta distribution

ζ ：proportional true score，$0\le\zeta\le 1$

其二，透過正弦反函數(2)將原始分數轉換為 arcsin分數（$c(x)$）：

$$c(x)=c(x|n)=.5\left\{\text{SIN}^{-1}\left[(x/(n+1))^{1/2}\right]+\text{SIN}^{-1}\left[((x+1)/(n+1))^{1/2}\right]\right\} \quad (n\text{:題數}) \qquad (2)$$

其三，**是**非線性方法建立原始分數至量尺分數之轉換:

$$x\xrightarrow{\text{非線性}}c(x)\xrightarrow{\text{線性:}}s\Rightarrow x\xrightarrow{\text{非線性}}s\ ,$$

而$c(x)$至量尺分數的線性轉換公式如下：

$$s\ \equiv s^*[c(x)]=\frac{\sigma_s^*}{\sigma(E_c)}\{c(x)-\mu[c(X)]\}+\mu_s^*\ ,\quad \{\mu_s^*,\sigma_s^*\}\text{:指定值（給定值）} \qquad (3)$$

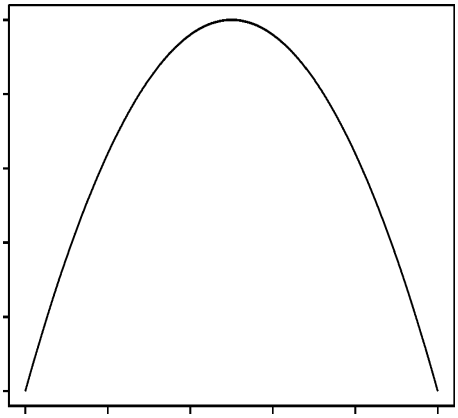
鑑於基測中心對公式(3)作了違反測量概念的『調整』，筆者特在此就公式(3)作進一步的說明。

公式(3)雖是標準化形式，卻是平均值與**測量標準誤**之相對轉換: $\{\mu_s^*,\sigma_s^*\}$相對$\{\mu[c(X)],\sigma(E_c)\}$，其中$\sigma_s^*$與$\sigma(E_c)$分別為量尺分數($s$)與 arcsin 分數（$c(x)$）的**測量標準誤**(SEM: standard error of measurement)，而不是**標準差**(SD: standard deviation)。

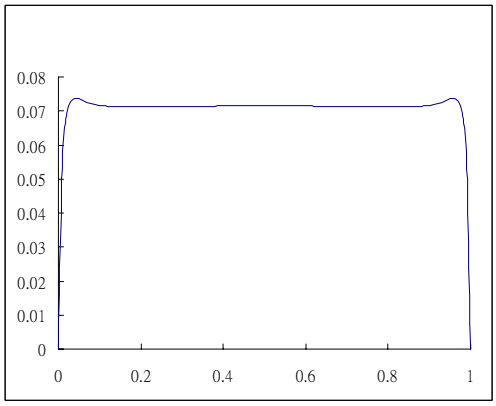
在強真分數理論線性模式下($X=n\zeta+E$)，**測量標準誤是測驗誤差分**數(E)的統計量，**標準差是測驗**分數(X)的統計量。因而$\sigma(E_c)$是指的$c(x)$的**誤差分**數的(平均)**測量標準誤**，$\sigma(E_c)$=sqrt[$\sigma^2(E_c)$]，$\sigma^2(E_c)=\int_a^b\sigma^2[c(X)|\zeta]g(\zeta)d\zeta$ 。

正弦反函數(2)具有穩定二項式變異數作用(Freeman and Tukey，1950)，使量尺分數之條件測量標準誤在整個量尺上幾乎有固定的均等值。(在compound binomial或binomial的條件機率分配下，原始分數(x)的條件測量標準誤($\sigma(x|\zeta)$)因與ζ值有關，不為常數，不是固定的)。(見下圖對照所示)。

原始分數(x)的條件測量標準誤
($\sigma(x|\zeta)$)



arcsin分數$c(x)$的條件測量標準誤
($\sigma(c(x)|\zeta)$)



正弦反函數之轉換能將靠近中間的量尺分數加以壓縮（compressing）而將量尺的兩端分數拉長開來(stretching)。換言之，在正弦反函數作用下，原本差距皆為1的原始分數以不同比例被轉換：量尺上，分數點的差距不是相同的，兩端分數拉長的比例與題數有關。

公式(3)**的量尺分數的**$\{\mu_s^*, \sigma_s^*\}$是指定值。Kolen對The ACT Assessment所建立的**量尺分數作如下**指定值:全距:1-36、平均值(μ_s^*)=18，平均測量標準誤(σ_s^*)=2。測量標準誤σ_s^*=2 賦予量尺之條件測量標準誤皆為2，使每一分數具有$s \pm 2$之測量誤差範圍(interval)。Kolen (1988)就分數精準量尺型態的意義，指定量尺全距(1-36)小於測驗題數(n) (The ACT Assessment的四個測驗題數介於40−70題)。 The ACT未考慮以**量尺分數加**總作為分發之用，不必考量s^*轉換(公式(3))未必使子學科的**量尺分數具有相同的標準差，對**分發**總**分產生不同的影響力。

心測中心的均等條件測量標準誤量尺分數。基測成績作為分發入學依據，心測中心理應建立具有常模意義的量尺分數型態(scale with property of normative meaning)，卻偏好套用ACT作法。在套用ACT作法時，對**量尺分數指定值**(測量標準誤及全距)的**觀念及推導出現即極大的偏差**，使致在計算考生量尺分數時，**必須進行『**調整**』**。

表一摘錄基測中心公佈的90-93年度第一次基測量尺分數指定值，及其三參數beta-binomial模式(Carlin及Rubin, 1991)之beta分配的三參數估計值。

表一 基測中心公佈的 90-93年度第一次基測量尺分數指定值 及beta -binomial模式參數值

			3參數beta -binomial模式				量尺分數指定值		
			$\hat{\Psi}=(\hat{a},\hat{\alpha},\hat{\beta})$					人為調整	
			下限	Alpha	Beta	平均數	測量標準誤	測量標準誤	標準差
年度	學科/題數	有效人數	$\hat{a}$	$\hat{\alpha}$	$\hat{\beta}$	μ_s^*	σ_s^*	σ_s^*	SD(s)
90	國文科(n=46)	300032	0.06173	2.76526	1.45796	30	7.5	4.17982	13.5
	英文科(n=44)	299983	0.17474	0.54148	0.48492	30	7.5	3.77680	18.5
	數學科(n=32)	299368	0.19351	1.22484	1.35433	30	7.5	4.28931	12.3
	自然科(n=56)	300016	0.27881	0.83148	1.08218	30	7.5	3.36531	12.5
	社會科(n=66)	299446	0.21944	1.39800	1.21219	30	7.5	3.46173	13.7
91	國文科(n=50)	296611	0.16537	1.92805	1.37447	30	7.5	3.99273	13.5
	英文科(n=45)	296315	0.17269	0.72929	0.99827	30	7.5	3.78014	14.7
	數學科(n=31)	295917	0.14926	1.11804	1.49486	30	7.5	4.15696	12.0
	自然科(n=58)	296593	0.26978	0.84412	1.15943	30	7.5	3.28283	12.2
	社會科(n=63)	295988	0.19365	1.82323	1.69067	30	7.5	3.64522	12.8
92	國文科(n=48)	308249	0.04363	2.89442	1.43802	30	7.5	4.40586	14.5
	英文科(n=45)	307601	0.19214	0.49407	0.44799	30	7.5	3.89570	19.5
	數學科(n=31)	307194	0.17095	0.96520	0.77167	30	7.5	5.26824	18.2
	自然科(n=58)	308095	0.24673	0.95790	1.10712	30	7.5	3.36947	13.0
	社會科(n=63)	307144	0.19938	1.66332	1.26534	30	7.5	3.55028	13.5
93	國文科(n=48)	312173	0.10920	2.11311	1.19738	30	7.5	4.29414	15.0
	英文科(n=45)	311796	0.18234	0.49888	0.43601	30	7.5	3.83078	19.5
	數學科(n=32)	311481	0.17377	0.78613	0.71828	30	7.5	4.65462	17.0
	自然科(n=58)	312156	0.30434	0.82194	0.97217	30	7.5	3.53505	13.0
	社會科(n=63)	311544	0.21138	1.54051	1.15506	30	7.5	3.46810	13.5

$$(\sigma_S^*)^2=\sigma^2(E_s)=\int_a^1\sigma^2(s\,|\,\zeta)g(\zeta)d\zeta$$

表一資料顯示心測中心指定『$\mu_s^*=30,\sigma_s^*=7.5$』作為學科共同量尺分數的特性。基測中心匪夷所思地將測量標準誤指定為$\sigma_s^*=7.5$，依此指定值之下，量尺分數何等地不精準。 例如，二位考生的單一考科的量尺分數的差距要達到15分以上($s\pm7.5$)才有統計上的意義。心測中心此一錯誤的解讀自陷於以違反**測量概念**的『調整』方式來計算量尺分數。

依$\sigma_s^*=7.5$指定下，心測中心必定發現:1)各年度五學科量尺分數最小全距落入[1-80]，最大全距則落入[1-100]；2) 在相同的測量標準誤下，五學科量尺分數最大值皆不一樣。(筆者發現$\sigma_s^*=2$或3指定下，量尺分數最大值達不到60分)。

依據固守全距=1-60及 $\mu_s^* = 30$ 作法下，因測驗題數的差異，心測中心勢必調整子學科測量標準誤(σ_s^*)的指定值，亦即以不等量的測量標準誤(σ_s^*)分別計算子學科的量尺分數。如此一來，導致二層的纍加負作用。 第一層負作用發生在子學科內，當(σ_s^*)的指定值改變，影響量尺分數點之間差距的斜率($\sigma_s^*/\sigma(E_c)$)也被改變；第二層負作用發生在總分，當斜率改變，子學科量尺分數的標準差會改變，子學科對總分影響力的權重亦被改變。更嚴重的是，**若斜率($\sigma_s^*/\sigma(E_c)$)的調整並非依據當年度考生實際作答反應資料計算而得，這就**違反了『公平計分遊戲規則』，考生所得的量尺分數的公平性就有問題。

心測中心未整數化量尺分數原貌。 筆者由心測中心所公佈的整體統計資料發現如表二的未整數化及整數化的**量尺分數原貌**：在內部作業的計分過程中的最高分並非設定在60分，且各個科目在各個年度都不相同。這可能是導因於以人為調整測量標準誤(σ_s^*)之故，而非依正確算法: $60 = A[c(n) - \mu[c(X)] + 30$ 來計算測量標準誤($\sigma_{s|60}^*$)及**量尺分數**。在正確的算法下($\sigma_{s|60}^*$)，未整數化**量尺分數**最高分皆會等於60分(參考表三)，不會產生低於或超過60分之狀況，也不會答對題數不同而有相同的60分最高分(針對答對n題數與答對n-1題數而言)。($60 = A[c(n) - \mu[c(X)] + 30$ 代表最高分設定在60分)。

表二 90-93年度基測中心統計資輸出的未整數化的量尺分數原貌--最高前二答對題數

年度		答對題數	公告量尺分	統計資輸出 量尺分數	統計資輸出 未整數化	年度		答對題數	公告量尺分	統計資輸出 量尺分數	統計資輸出 未整數化
90	國文科	45	54	54	53.61	92	國文科	47	55	55	55.17
	(n=46)	46	60	60	59.56		(n=48)	48	60	60	61.45
	英文科	43	54	54	54.00		英文科	44	55	55	54.85
	(n=44)	44	60	60	59.56		(n=45)	45	60	60	60.64
	數學科	31	54	54	53.96		數學科	30	55	55	54.92
	(n=32)	32	60	60	60.09		(n=31)	31	60	60	62.55
	自然科	55	56	56	56.33		自然科	57	57	57	56.73
	(n=56)	56	60	60	61.12		(n=58)	58	60	60	61.53
	社會科	65	57	57	57.29		社會科	62	56	56	56.14
	(n=66)	66	60	60	62.22		(n=63)	63	60	60	61.20
91	國文科	49	55	55	55.36	93	國文科	47	55	55	54.71
	(n=50)	50	60	60	61.05		(n=48)	48	60	60	60.84
	英文科	44	60	60	59.53		英文科	44	54	54	54.28
	(n=45)	45	60	60	64.93		(n=45)	45	60	60	59.98
	數學科	30	56	56	55.94		數學科	31	54	54	54.07
	(n=31)	31	60	60	61.87		(n=32)	32	60	60	60.82
	自然科	57	57	57	57.11		自然科	57	56	56	56.40
	(n=58)	58	60	60	61.79		(n=58)	58	60	60	61.45
	社會科	62	60	60	59.75		社會科	62	55	55	54.97
	(n=63)	63	60	60	64.94		(n=63)	63	60	60	59.92

3.3 測量標準誤調整後的影響層面

3.3-1 各考科量尺分數差距改變

以表二的91年度基測為例，得滿分的考生與錯一題的考生之五考科未整數化的量尺分數差距原本為：國文科 6分，英語科5分，數學科6分，自然科5分，及社會科5分；而公告的(整數化及截頭至60分)量尺分數差距降低為：國文科5分，英語科0分，數學科4分，自然科3分，及社會科0分。這對照下，心測中心所公告的量尺分數顯然對得滿分的考生不公平。茲將得滿分的考生人數列舉於下，讀者就不會和筆者一樣，以為只犧牲少數得滿分的考生。

90-93年度第一、第二次基測量尺分數滿分人數

	國文科	英語科	數學科	社會科	自然科
901	2180	14825	4495	1082	1760
902	1132	14379	4393	496	91
911	2050	4560	1402	818	1640
912	57	4508	5165	909	773
921	2569	16993	15651	1057	1913
922					
931	3819	20926	12041	2460	2864
932	1816	9073	6952	1358	1535

心測中心的**隨意調整**算法不是只對得滿分的考生不公平，對所有考生都不公平。這是因為當測量標準誤(σ_s^*)的值改變，連帶地影響量尺分數點之間差距的斜率($\sigma_s^*/\sigma(E_c)$)也被改變了。茲以表三數據說明此一事實。

表三數據顯示:數學32個分數點就有19個分數點的量尺分數不同於正確算法所得的量尺分數。答對題數落在14-19題的考生，其量尺分數不變；答對題數落在20-31題的考生，有的量尺分數被加1分，有的量尺分數被加2分，有的量尺分數沒加分；答對題數落在0-14題的考生，有的量尺分數被減1分，有的量尺分數被減2分，有的量尺分數沒被扣分。表三數據說明在心測中心**隨意調整**測量標準誤的算法之下，影響的層面不是僅限於得滿分的考生，而是絕大多數的考生。

表三 91年度數學量尺分數在正確算法與人為調整算法之差值

答對	基測公告		正確算法			反推及驗證基測內部計算			
題數	量尺分數	未整數化	量尺分數	未整數化	差值	$\hat{\Pr}(X=x)$	$C(x)$	量尺分數	未整數化
0	1	-3.61	1	-1.64	0	0.00022	0.08886	1	-3.61
1	2	2.33	4	3.95	-2	0.00149	0.2152	2	2.33
2	5	5.46	7	6.91	-2	0.00502	0.28193	5	5.46
3	8	8.02	9	9.31	-1	0.0114	0.33628	8	8.02
4	10	10.26	11	11.42	-1	0.0198	0.38387	10	10.26
5	12	12.29	13	13.33	-1	0.02844	0.42711	12	12.29
6	14	14.18	15	15.11	-1	0.03564	0.46726	14	14.18
7	16	15.96	17	16.78	-1	0.04068	0.50515	16	15.96
8	18	17.66	18	18.38	0	0.04369	0.54129	18	17.66
9	19	19.29	20	19.92	-1	0.04522	0.57609	19	19.29
10	21	20.88	21	21.41	0	0.04583	0.60984	21	20.88
11	22	22.43	23	22.87	-1	0.0459	0.64277	22	22.43
12	24	23.94	24	24.30	0	0.04565	0.67507	24	23.94
13	25	25.44	26	25.71	-1	0.0452	0.70691	25	25.44
14	27	26.92	27	27.10	0	0.04459	0.73843	27	26.92
15	28	28.40	28	28.49	0	0.04386	0.76976	28	28.40
16	30	29.87	30	29.87	0	0.04301	0.80103	30	29.87
17	31	31.34	31	31.26	0	0.04206	0.83237	31	31.34
18	33	32.82	33	32.65	0	0.04099	0.86388	33	32.82
19	34	34.32	34	34.06	0	0.03983	0.89572	34	34.32
20	36	35.84	35	35.49	1	0.03855	0.92802	36	35.84
21	37	37.38	37	36.95	0	0.03717	0.96095	37	37.38
22	39	38.97	38	38.44	1	0.03566	0.9947	39	38.97
23	41	40.61	40	39.98	1	0.03403	1.0295	41	40.61
24	42	42.31	42	41.58	0	0.03224	1.06565	42	42.31
25	44	44.09	43	43.26	1	0.03029	1.10353	44	44.09
26	46	45.97	45	45.03	1	0.02814	1.14369	46	45.97
27	48	48.01	47	46.95	1	0.02575	1.18692	48	48.01
28	50	50.24	49	49.05	1	0.02303	1.23452	50	50.24
29	53	52.80	51	51.46	2	0.01987	1.28886	53	52.80
30	56	55.94	54	54.41	2	0.016	1.3556	56	55.94
31	60	61.87	60	60	0	0.01075	1.48194	60	61.87

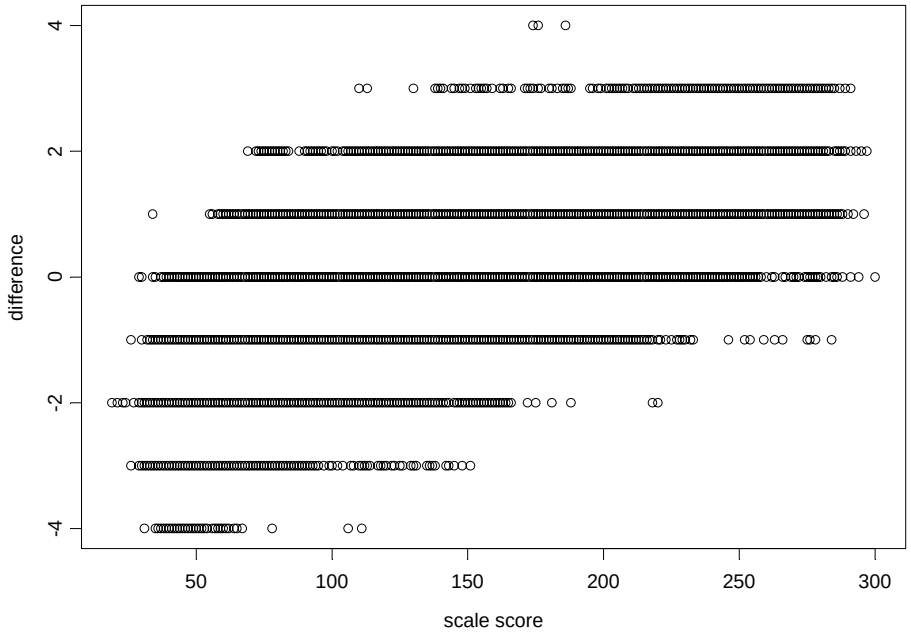
3.3-2 對五考科量尺總分變化之影響

量尺總分是五考科量尺分數以分非線性組合加總(nonlinear combination)。**筆者進一步比較**心測中心**隨意調整**算法的**量尺總分**與正確算法的**量尺總分之差異，二**算法之**總分皆**以整數化**量尺分數為相加基準。筆者也計**算**差異值**的**考生人**數**分佈。表**四摘錄這二方面的統計數據。

表四數據顯示90-93年度量尺總分變化差距各年度皆不一樣。90-93年度總分變化差距範圍依序分別為[減4分至加4分]，[減11分至加14分]，[減9分至加7分]，及[減4分至加3分]。亦即當以正確算法的量尺總分為參照基準，心測中心隨意調整的算法讓有些考生分發總分被扣11分之多，讓有些考生分發總分增加達到14分之多；此一增一減的範圍足足達到25分之多(91年度為例)。92年度達到16分之多，90年度達到8分之多，而93年度達到7分之多。

總分變化差距=0代表量尺總分沒有變化。表四數據顯示量尺總分沒有受到影響的考生人數是相對的少。90-93年度量尺總分沒有受到影響的考生人數比率依序分別約為29%、6%、10%、22%。換言之，90-93年度量尺總分受到影響的考生人數比率依序分別約為71%、94%、90%、78%。心測中心的隨意調整測量標準誤作法之下的影響層面真是不小。

圖三的縱座標為考生的量尺總分差異值，橫座標為考生基測成績通知單的量尺總分；圖三可進一步瞭解考生成績通知單的量尺總分是加分或扣分之後的成績(90年度為例)。以量尺總分=200之考生群為例，構成相同量尺總分=200來源不同，有的加分作用，有的是減分作用，且加減分數不一致，如+3，+2，+1，+0，-1，及-2分等皆構成量尺總分=200之來源。(加分代表基測通知單成績高於正確算法的成績，減分代表正確算法的成績高於基測通知單成績)。



圖三 90年度心測中心調整算法與正確算法的二量尺總分之差異分佈圖，橫座標為考生基測成績通知單的量尺總分。

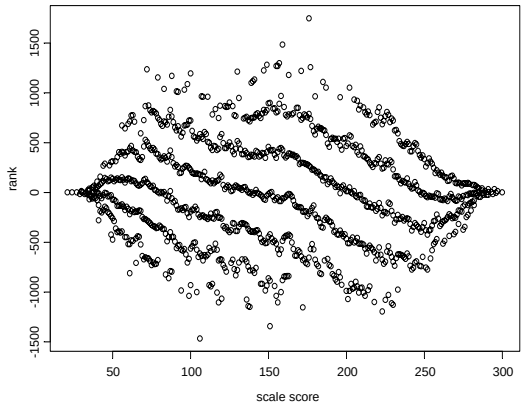
表四、90-93年度第一基測量尺總分變化差距及樣本考生、整體考生人數之分佈

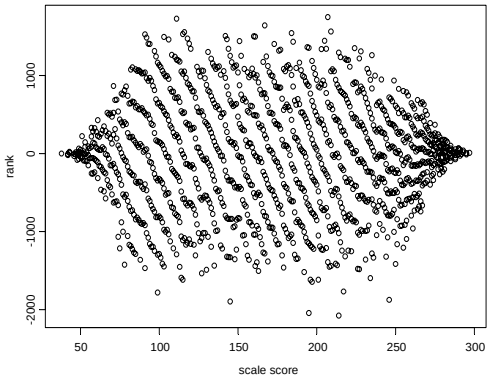
90年度			91年度			92年度			93年度		
總分變化差距	抽樣人數	整體人數	總分變化差距	抽樣人數	整體人數	總分變化差距	抽樣人數	整體人數	總分變化差距	抽樣人數	整體人數
-4	70	566	-11	1	8	-9	2	16	-4	35	355
-3	907	7339	-10	20	162	-8	25	204	-3	1312	13310
-2	3779	30578	-9	278	2256	-7	346	2832	-2	3845	39009
-1	7065	57168	-8	1085	8806	-6	1156	9465	-1	5400	54785
0	10797	87366	-7	1951	15834	-5	2112	17292	0	6729	68268
1	8386	67857	-6	2602	21118	-4	2965	24276	1	5693	57757
2	4784	38711	-5	2870	23293	-3	3459	28321	2	5709	57920
3	1580	12785	-4	2601	21110	-2	3712	30392	3	2293	23263
4	3	24	-3	2379	19308	-1	3894	31883	總人數	31016	314667
總人數	37371	302394	-2	2230	18099	0	3867	31662	差距=0之考生人數比		
差距=0之考生人數比率			-1	2113	17149	1	3663	29991		0.22	0.22
	0.29	0.29	0	2057	16694	2	3606	29525			
			1	1915	15542	3	3510	28739			
			2	1867	15152	4	2998	24546			
			3	1690	13716	5	2131	17448			
			4	1592	12920	6	761	6230			
			5	1671	13562	7	50	409			
			6	1929	15656	總人數	38257	313231			
			7	2033	16500	差距=0之考生人數比率					
			8	1667	13529		0.10	0.10			
			9	1174	9528						
			10	735	5965						
			11	333	2702						
			12	109	884						
			13	24	194						
			14	2	16						
			總人數	36928	299703						
			差距=0之考生人數比率								
				0.06	0.06						

3.3-3 名次分發排序之影響

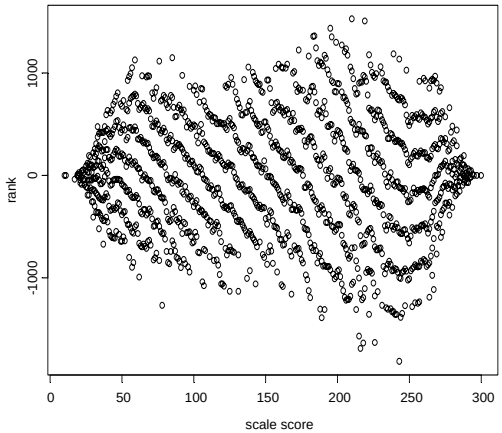
就台北市考區而言，考生量尺總分僅相差1至2分之內，足以影響分發至不同學校就學。又且，量尺總分200分為分發至公、私立學校之決勝點，筆者進一步比較量尺總分名次排序之差異。換言之，筆者將抽樣考生的心測中心調整算法的量尺總分(5-300)作排序，亦將考生的正確算法的量尺總分(5-300)作排序，然後計算二算法等第之差異。等第(ranking)差距範圍和考生人數有關，筆者進一步依抽樣比率估計台北分發區整體考生之可能的等第差異範圍。圖四(a)-(d)分別為90-93年度台北分發區考生名次差異範圍，縱座標為名次差異值(心測中心調整算法-正確算法)，橫座標為考生基測成績通知單的量尺總分。(台北分發區包括北一區、北二區及基隆區考生；90-93年度考生人數分別約為88206、87525、91714、及94223)。

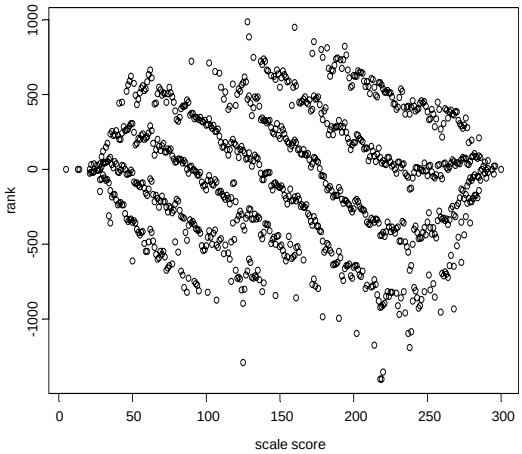
(a): 90年度　　(b): 91年度





(c): 92年度　　(d): 93年度





圖四、90-93年度台北分發區考生名次變化範圍(估計等第差值)， 縱座標為名次差異值(心測中心調整算法-正確算法)，橫座標為考生基測成績通知單的量尺總分。

圖四圖形所示，考生基測成績高於280分或低於50分，心測中心調整算法的名次與正確算法的名次一致，名次差異不大。除此之外，基測成績落在280分與50分之間的考生，名次變化皆很大。90、91年度基測成績落在250分與150分之間的考生，名次變化最大範圍約3000名；92年度基測成績落在280分與180分之間的考生，名次變化最大範圍約2500名；93年度基測成績落在220分與150分之間的考生，名次變化最大範圍約2500名。這些大範圍名次變化在在顯示足以影響考生分發至不同的學校。更嚴重的是，就基測成績落在200分附近的考生而言，名次變化足足影響其公、私立學校入學之改變。

圖四圖形的名次變化分佈皆呈現有規律的形態 (pattern)。有形態的分佈代表心測中心調整算法是會影響名次的，是不公平的。心測中心調整算法若是公平的，名次應該是不變的，圖內所有點皆會落在0的水平線。90年度與93年度具有相似形態的名次變化分佈，因二年度的量尺總分差異範圍相近，分別為8分和7分。91年度與92年度具有相似形態的名次變化分佈，因二年度的量尺總分差異範圍較相近，分別為25分和16分。(總分300分限制下，名次變化形態與量尺總分差異範圍有關)。

3.4 1-60分量尺分數正確算法

依藉 $\hat{\boldsymbol{\Psi}}=(\hat{a},\hat{\alpha},\hat{\beta})$ 是充分統計量(Maritz and Lwin, 1989)，由**心測中心所公告的** beta-binomial參數值估計值(見表一)，筆者可依照 $60=A[c(n)-\mu[c(X)]+30$ 求得**測量標準誤**($\sigma^{*}_{s|60}$)來計算**量尺分數，此為正確**算法的考生**量尺分數。依藉** $\hat{\boldsymbol{\Psi}}=(\hat{a},\hat{\alpha},\hat{\beta})$ 是充分統計量，及表一的『調整』**測量標準誤**(σ^{*}_{s})可驗證或計算**基測中心內部作業**的**量尺分數。茲說明正確**算法的**測量標準誤**($\sigma^{*}_{s|60}$) **之計算步驟於下。**

第一步： 依三參數beta-binomial模式及 $\hat{\boldsymbol{\Psi}}=(a,\alpha,\beta)$ 可得90-93年度第一次基測整體考生真分數分佈及原始分數分佈。

模式參數　　真分數分佈　　原始分數分佈

$$\hat{\boldsymbol{\Psi}}=(a,\alpha,\beta)\xrightarrow{\text{估計}}\hat{g}(\zeta)\xrightarrow{\text{估計}}\hat{\Pr}(X=x)$$

$$\Pr(X=x)=\int_a^1 p_n(x|\zeta)\,g(\zeta)\,d\zeta\ ,\quad \text{其中 } g(\zeta,\boldsymbol{\Psi})=\frac{(\zeta-a)^{\alpha-1}(1-\zeta)^{\beta-1}}{(1-a)^{\alpha+\beta-1}B(\alpha,\beta)}$$

$$p_n(x|\zeta)=\binom{n}{x}\zeta^{x}(1-\zeta)^{n-x}$$

第二步：將各年度五學科的原始分數轉換為反正弦分數 $x\to c(x)$，學科題數如表一所示。

$$c(x)=\frac{1}{2}\left\{\sin^{-1}\sqrt{\frac{x}{n+1}}+\sin^{-1}\sqrt{\frac{x+1}{n+1}}\right\},\ x=0,1,2,....,n\quad (n\text{:試卷題數})$$

第三步:依據**第一步之相關資訊**計算 $c(x)$ 的平均數及 $\hat{\sigma}^2(E_c)$

$$\hat{\mu}[c(X)] = \sum_{x=0}^{n} c(x)\hat{\Pr}(X = x) \qquad (c(x) \text{ 分數之平均數})$$

$$\hat{\sigma}^2(E_c) = \int_a^1 \sigma^2[c(X)|\zeta]\hat{g}(\zeta)d\zeta \qquad (c(x) \text{ 誤差分數之變異量})$$

第四步: 計算正確算法**的**測量標準誤 $\sigma^*_{s|60}$:

$$60 = A[c(n) - \mu[c(X)]] + 30$$

$$60 = \frac{\sigma^*_s}{\hat{\sigma}(E_c)}\{c(n) - \hat{\mu}[c(X)]\} + 30,$$

$\because \hat{\sigma}(E_c), c(n), \hat{\mu}[c(X)]$ 皆已知之下，**正確**算法的 $\sigma^*_{s|60}$ 值可求得

第五步: 反正弦分數轉換為量尺分數 $c(x) \to s = s^*[c(x)]$

$$s \equiv s^*[c(x)] = \frac{\sigma^*_{s|60}}{\hat{\sigma}(E_c)}\{c(x) - \hat{\mu}[c(X)]\} + \mu^*_s,$$

$$= \frac{\sigma^*_{s|60}}{\hat{\sigma}(E_c)}c(x) + \left\{\mu^*_s - \frac{\sigma^*_{s|60}}{\hat{\sigma}(E_c)}\hat{\mu}[c(X)]\right\}$$

（斜率） **（截距）**

步驟1-5說明正確算法與心測中心『調整算法』的量尺分數的計算差異之處，僅在於使用不同的測量標準誤: $\sigma^*_{s|60}$ 與 σ^*_s，其他統計量諸如 $c(x)$、$\hat{\sigma}(E_c)$、$\hat{\mu}[c(X)]$、及 μ^*_s 皆相同。

表五摘錄**正確**算法與基測中心『調整』算法的 σ^*_s 與 $\sigma^*_{s|60}$，斜率與截距，及量尺分數的標準差，可說明**不同測量標準誤之下，計**算量尺分數的斜率 $(\sigma^*_s/\sigma(E_c))$ 就不相等，量尺分數標準差也就不等，對量尺總分的影響力也就改變。下表是五學科對量尺總分的相對比重。 由百分比所示，考科對量尺總分的比重是不一樣的(見直行)；各考科在各年度對量尺總分的影響也是不一樣的(見橫行)。如果這些比重是考後看到考生成績內容後再去調整的，容易產生弊端。

基測中心	90年度	91年度	92年度	93年度
國文科	19%	21%	18%	19%
英文科	26%	23%	25%	25%
數學科	17%	18%	23%	22%
自然科	18%	19%	17%	17%
社會科	19%	20%	17%	17%
	100%	100%	100%	100%

表五 90-93年度基測量尺公式之斜率、截距及量尺分數標準差之比較

		固定值		正確算法				基測中心			
					量尺公式		量尺分數	人為調整	量尺公式		量尺分數
年度	學科/題數	$\hat{\mu}[c(X)]$	$\hat{\sigma}(E_c)$	$\sigma^*_{s\|60}$	斜率	截距	標準差	σ^*_s	斜率	截距	標準差
90	國文科(n=46)	0.98206	0.07289	4.2414	57.51	-26.48	13.65	4.1798	56.93	-25.91	13.45
	英文科(n=44)	0.93168	0.07210	3.8328	52.81	-19.20	18.72	3.7768	52.33	-18.75	18.45
	數學科(n=32)	0.87293	0.08702	4.2768	49.12	-12.87	12.26	4.2893	49.26	-13.00	12.30
	自然科(n=56)	0.89315	0.06609	3.2439	48.81	-13.60	12.05	3.3653	51.21	-15.74	12.50
	社會科(n=66)	0.94198	0.06099	3.2236	52.86	-19.79	12.76	3.4617	56.31	-23.04	13.70
91	國文科(n=50)	0.95656	0.06995	3.8576	55.09	-22.70	12.99	3.9927	56.80	-24.34	13.45
	英文科(n=45)	0.81792	0.07350	3.2479	44.29	-6.22	12.62	3.7801	50.69	-11.46	14.69
	數學科(n=31)	0.80389	0.08843	3.9125	44.30	-5.61	11.29	4.1570	46.94	-7.74	12.00
	自然科(n=58)	0.8762	0.06500	3.0988	47.53	-11.65	11.51	3.2828	50.21	-13.99	12.20
	社會科(n=63)	0.90931	0.06248	3.1301	50.00	-15.47	10.95	3.6452	58.16	-22.88	12.75
92	國文科(n=48)	0.98973	0.07137	4.2033	57.92	-27.32	13.83	4.4059	60.33	-29.71	14.50
	英文科(n=45)	0.94045	0.07094	3.8254	53.51	-20.32	19.10	3.8957	54.36	-21.12	19.45
	數學科(n=31)	0.94349	0.08714	4.8552	55.17	-22.05	16.77	5.2682	58.34	-25.04	18.20
	自然科(n=58)	0.89759	0.06497	3.2059	49.51	-14.44	12.37	3.3695	51.76	-16.46	13.00
	社會科(n=63)	0.95973	0.06241	3.4141	54.69	-22.49	12.98	3.5503	56.44	-24.17	13.50
93	國文科(n=48)	0.98749	0.07125	4.1775	57.76	-27.04	14.59	4.2941	59.33	-28.59	15.00
	英文科(n=45)	0.94454	0.07075	3.8432	53.90	-20.91	19.51	3.8308	53.81	-20.82	19.45
	數學科(n=32)	0.91585	0.08572	4.5315	52.34	-17.93	16.55	4.6546	53.75	-19.23	17.00
	自然科(n=58)	0.92859	0.06487	3.3729	51.99	-18.28	12.88	3.5351	54.21	-20.34	13.50
	社會科(n=63)	0.97024	0.06236	3.4779	55.59	-23.94	13.54	3.4681	55.36	-23.71	13.50

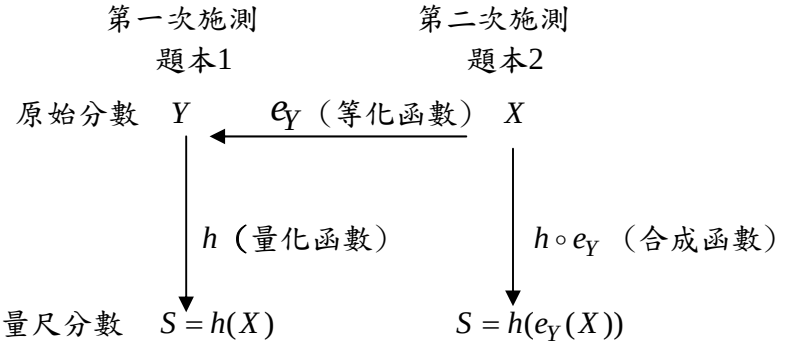
註：心測中心的統計資料將$c(x)$、平均數$\hat{\mu}[c(X)]$、$\hat{\sigma}^2(E_c)$皆冠上Kolen

註：心測中心以測量標準誤標示$\hat{\sigma}^2(E_c)$，$\hat{\sigma}^2(E_s)$，$\hat{\sigma}^2(E_X)$

四、 第二次基測等化程序的灰色地帶及其效應

4.1 簡介測驗等化

依據測量觀點，測驗施測情境若是一年多試且使用不同難度的試卷題本，得考量同時運作測驗量化及等化程序(Petersen, Kolen and Hoover,1989)，使考生不因受測題本難度的差異而吃虧或佔便宜。茲以圖五說明量化及等化運作的目的、對象及過程。令 X 和 Y 分別代表子學科第一次及第二次試卷題本的原始分數:

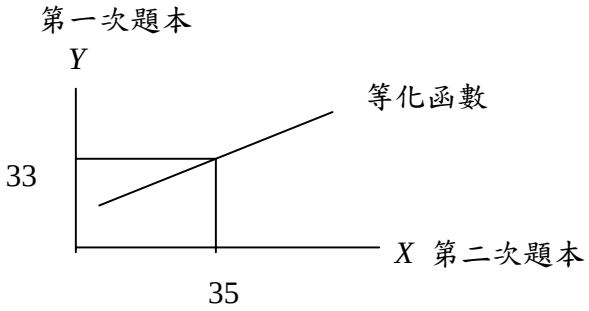


圖五、子學科試卷題本原始分數等化及量化程序示意圖

由示意圖路徑所示，第一次施測的題本作為建立共同量尺依據，題本1的原始分數只透過 h 函數轉換為量尺分數。 題本2的原始分數須先透過等化函數(e_Y)轉換為題本1的原始分數對等值，再透過 h 函數轉換為量尺分數。

A)等化統計方法皆認定題本的差異只在難度面向；因此要進行等化的題本皆須滿足內容同質試卷假設(homogenenous forms of a test)。所謂內容同質的二份題本是指所測的概念、認知能力等內容結構要盡可能地平行(parallel)，而題本的試題型態(type)及格式(format)要一致。在此一題，內容同質試卷不可僅由試題難度範圍相近定義的。

B)等化函數須是對稱的(symmetric)、可逆的(inversible)；意指Y轉換至X的等化函數($e_X(y)$)與X轉換至Y的等化函數($e_X(x)$)互為反函數關係。對稱函數的意涵如下圖所示:



此圖是說第一次題本的33分若轉換為第二次題本的35分，則第二次題本的35分只能對應至第一次題本的33分。等化函數可為線性或非線性函數(Holland and Rubin，1982)： $Lin_Y(x) = \mu_Y + (\sigma_Y / \sigma_X)(x - \mu_X)$ 或 $Equip_Y(x) = G^{-1}(F(x))$。（註:迴歸函數不具備對稱特質）。

C) 等化函數推導只容許二題本在難度面向的差異，而不容許受測考生群組在能力面向的差異。因此須透過資料收集設計(data collection design)來控制考生群組在能力面向的差異。共同考生群設計(common or equivalent group(s) design)及共同試題設計(common item design)為主要二大類設計。等化的統計架構 (Kolen and Brennan, 1995)共有三大類:測驗分數等化(observed score equating)，真分數等化(true score equating)，及IRT等化。此三類的等化的統計架構各有其優點及限制。

上述三要點在提醒正確等化程序的重要性。正確等化程序才能達成等化核心目標:**使不同難度題本的原始分數俱備可交換性**、等同**性**(interchangeability/ Equivalence)。可交換性是說第二次題本的35分是等同於第一次題本的33分。

4.2 基測題庫建立及等化方法

依據飛揚專刊及『國民中學學生基本學力測驗答客問』的宣導文，基測中心採用IRT 方法建立學科題庫(資料庫)，其「題庫」是存放著優良試題及試題參數(b_j)； 藉由試題參數資訊組合難度相近的二份基測學科題本、估計考生 IRT-能力值、及進行二次基測分數等化。心測中心強調其 IRT 等化程序是客觀的，因而考生第二次基測所得的量尺分數『**絕對是合理、公平、公正的**』(詳閱參考文獻 1-6)。

心測中心宣稱其所建立題庫是IRT-Rasch模式校準題庫(calibrated item pool)，是透過共同試題等化設計(見下表)，將所有預試題本試題的難度參數估計值($\hat{b}_j$)依公式(5)、公式(6)轉換至IRTθ–scale。

	共同試題	非共同試題	預試學生
預試題本 甲	1…. j	J' 題	組群 1
預試題本 乙	1…. j	J' 題	組群 2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
預試題本 M	1…. j	J' 題	組群 M

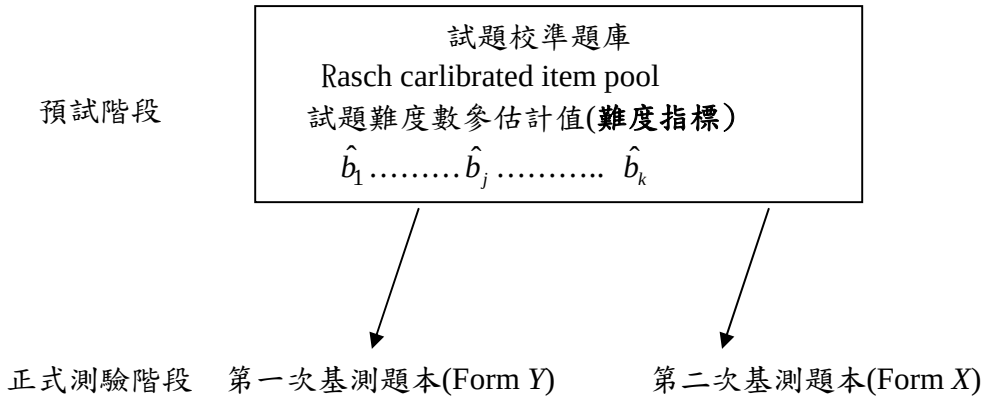
甲、乙二預試題本為例: $b_{甲j} = Ab_{乙j} + B$ ， (5)

$$\text{其中} \begin{cases} A = \sigma(b_{甲}) / \sigma(b_{乙}) \\ B = \mu(b_{甲}) - A\mu(b_{乙}) \end{cases} \quad (6)$$

$\sigma(b_{甲})$、$\mu(b_{甲})$: 甲題本之共同試題難度參數標準差及平均數

$\sigma(b_{乙})$、$\mu(b_{乙})$: 乙題本之共同試題難度參數標準差及平均數

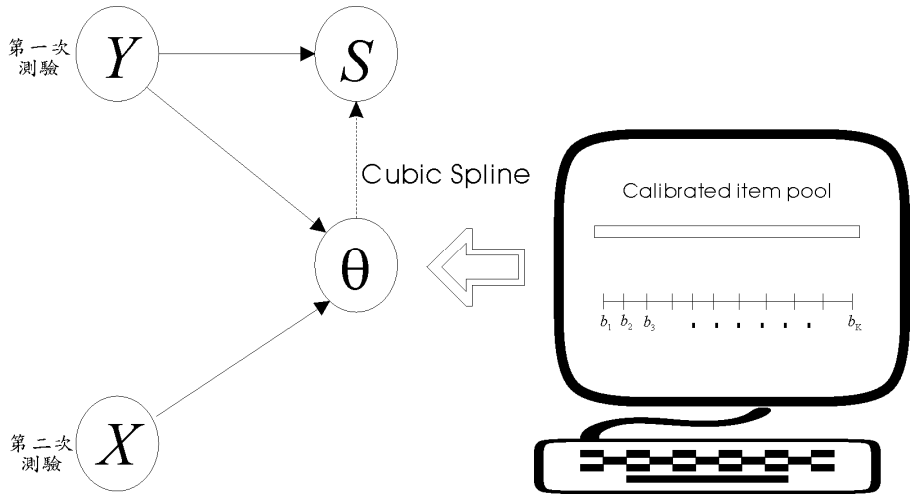
所有M預試題本若經過公式（5）、公式（6）的連結，即可建立如下圖所示的IRT-Rasch模式校準題庫。



（全國考生）　　（重複考生）

心測中心宣稱依據此題庫裡的試題難度指標$\hat{b}_j$組合前後二次正式學科測驗題本：Form Y and Form X，並使二測驗題本的平均難度相近(約75%答對率)。第一次基測實施(考完)後，由**考生**的實際**答題反應資料(0或1)及已知的**試題難度指標$\hat{b}_j$估計**考生的$\hat{\theta}$能力值**。同理，第二次基測實施後，參與第二次基測**考生的$\hat{\theta}$能力值亦可**估計出。心測中心宣稱前後二次$\hat{\theta}$**能力**估計值可合併在一起。

第一次題本(Form Y)的原始分數建立量尺分數轉換表(S)。考生第二次測驗的量尺分數是透過$\hat{\theta}$**能力**值連結至第一次測驗的量尺分數。例如，第一次測驗原始分數Y=60所對應的$S=52$及$\hat{\theta}=2.25$，若第二次測驗原始分數X=58所對應的$\hat{\theta}=2.25$，則$S=52$是X=58的量尺分數。換言之，二題本的原始分數**因**對應至**相同能力值而認定**具有相同的量尺分數。這種對應程序如下示意圖，即為心測中心宣稱的二次量尺分數等化程序。



圖六 心測中心宣稱的二次量尺分數等化程序示意圖

4.3 基測題庫與等化學理缺失

筆者比照相關文獻（Lord, 1980; Lord, 1984; Kolen and Brennan,1995），發現心測中心其題庫及等化內涵有嚴重的學理缺失。茲列舉於下。

<u>IRT Rasch-模式的缺失</u>。心測中心捨棄ACT或ETS的IRT 3-參數模式(包含試題鑑

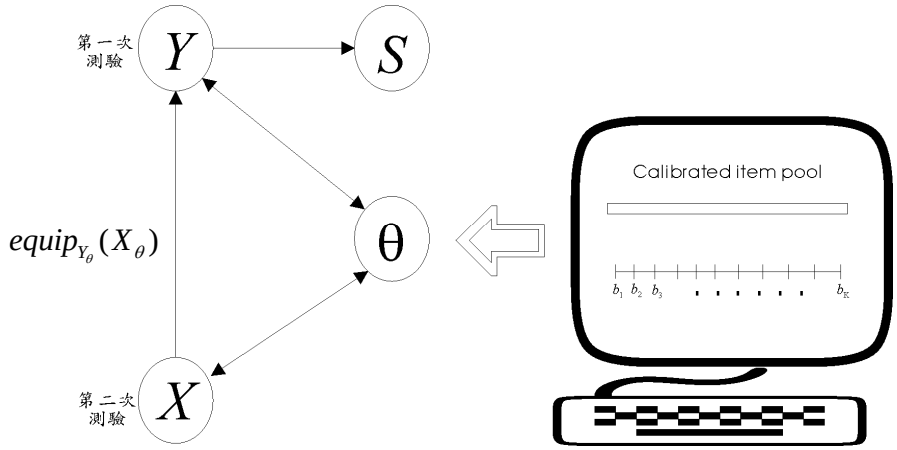
別度、難度、猜度)建立題庫，而採用單參數的Rasch模式建立題庫。基測五學科皆為四選一的選擇題試卷。Rasch模式的試題特徵曲線只反應試題難度差異，而令試題鑑別度皆相同且試題猜度為零。筆者就申請資料(90-93各年度三萬多筆)對其作適合度考驗，發現Rasch模式的單參數試題特徵曲線無法適切描述基測試題:基測試題鑑別度不同、猜度亦不為零。

心測中心校準題庫的缺失。 Rasch模式假設不滿足，**利用**公式(5)、公式(6) 建立校準題庫，題庫的試題難度指標$\hat{b}_j$的正確度亦受影響。這是因為若模式偏頗，公式(6)之A、B等式亦偏頗。此外共同試題(common items)的難度參數是估計的，非已知的(known)，這也影響題庫的試題難度指標$\hat{b}_j$的正確度。例如預試群組學生與實際考生努力作答動機（motivation）若相去甚遠，題庫的試題難度指標$\hat{b}_j$的正確度是大打折扣的。題庫資訊不精確，亦影響測驗題本達不到統計平行的基本要求。

心測中心估計$\hat{\theta}$的缺失。 心測中心宣稱預試題本隨機分配給不同地區的國三學生測試(北、中、南、東區共8校各校一班)，預試題本的每一道試題至少有240至320位學生預試過。亦即，Rasch-模式校準題庫試題難度指標$\hat{b}_j$乃依據240至320位預試學生作答估算的。心測中心將校準題庫試題難度指標$\hat{b}_j$當作已知參數，代入概率函數$\{L(u_1,u_2,.....,u_n|\theta)=\prod_{j=1}^{n}P_j^{u_j}Q_j^{1-u_j},u_j=0\text{ or }1\}$估計全國考生**的$\hat{\theta}$能力值**(估計方法或為conditional maximum likelihood estimation或為Bayesian estimation, Lord,1980)。

此種作法捨棄垂手可得的最正確的試題難度參數b_j (parameter)；因全國考生參與第一次基測，而IRT Bilog或Logist軟體皆可依據 30多萬考生之試題反映資料計算正式題本的試題難度參數b_j及考生**的$\hat{\theta}$能力值。**心測中心所公佈的統計資料甚致顯露估計91, 92, 93年度全國考生的$\hat{\theta}$能力值似乎皆以90年度題庫試題難度指標$\hat{b}_j$為基準。試問91年度基測分數是用來排序91年度考生，92年度基測分數是用來排序92年度考生，而93年度基測分數是用來排序93年度考生，為何不以該年度全國考生資料估計的正式題本試題難度參數b_j，及估計該年度考生$\hat{\theta}$能力值。

心測中心等化的缺失。基測量尺分數轉換表（conversion table）是由第一次測驗原始分數建立的，而不是由第一次測驗的θ-scale建立的，則在IRT學理架構下，正確等化程序應如圖七之示意圖而不是圖六之示意圖。



圖七 基測第二次量尺分數正確等化程序示意圖（IRT-測驗分數等化）

圖七與圖六(心測中心等化程序)不同之處如下：

1. 圖六缺少X至Y的equipercentile等化轉換（ $Equip_{Y_\theta}(X_\theta)$ ）。
2. 圖七的 $Equip_{Y_\theta}(X_\theta)$ 等化程序讓我們可以說二題本的原始分數是等同的、可交換的(equivalent and interchangeable)，**因**而對應至相同的量尺分數。
3. 圖七的等化程序反映對稱函數特質，保留二題本原始分數之間的相**等性，類似公分與英寸長度轉換，攝氏與華氏溫度轉換。**

上述題庫與等化的學理缺失，讓筆者對心測中心信誓旦旦所宣稱『考生第二次基測所得的量尺分數**絕對是合理、公平、公正的**』大打折扣。。這些學理缺失使教育部『擇優政策美意淪為擇優排擠效應』。

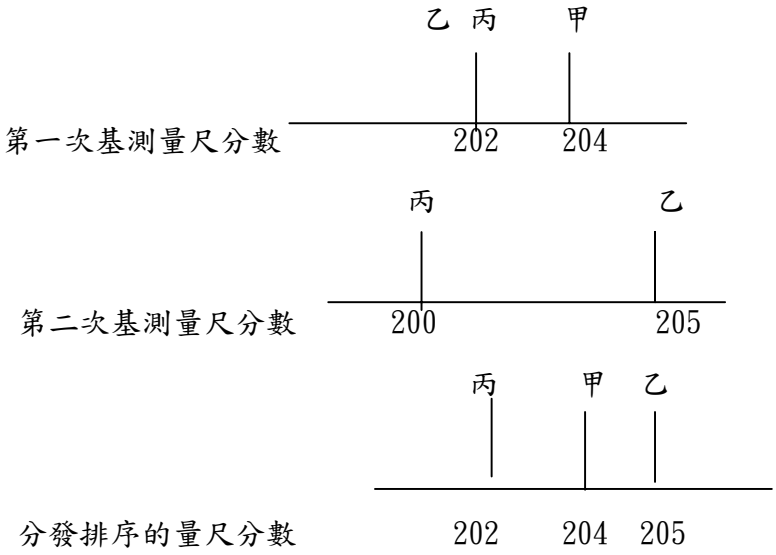
4.4 擇優名次排擠效應

教育部擇優政策允許參與第二次基測考生就其二次成績選擇較優者作為入學分發依據。教育部為避免基測淪為二次『聯考』，不鼓勵考生參加第二次基測。因此，大約40%-45%的全國考生未參加第二次基測。

擇優政策**公平基準仰賴**正確等化程序及等化條件滿足。上述學理缺失使得考生第二次基測所得的量尺分數混淆等化誤差，無法充分反映考生實質進步或退步因素。在此情況下，未參加第二次基測考生權益受害最大，遭受排擠效應機率最高，分發排序等第變化最大。而參加第二次基測考生也會遭受排擠效應。茲以下例說明。

甲、乙、丙考生第一次基測所得的量尺總分別為204分、202分、202分；其中乙、丙考生參加第二次基測。乙、丙考生所得的第二次基測量尺總分別為205分及200分。

擇優政策下，用來分發甲、乙、丙三人的量尺總分分別為204分、205分、202分；使得分發排序等第改變為乙優先甲、丙二人。



此例亦說明用來分發排序參加第二次基測考生的量尺分數是其二次分數的最大值。因而參加第二次基測考生相對未參加者，有極大機會在分發排序轉為優勢。教育部應將擇優政策隱藏的好處清清楚楚地**告訴家長及考生，避免讓**未參加第二次基測**考生或因資訊不明或因**『等化缺失』而吃悶虧。『等化缺失』的質疑可由下列資料見端倪。

二次基測量尺分數平均數之比較。 表六摘錄前後二次基測學科量尺分數平均數、標準差、總平均數；數據摘自心測中心『試題年度報告』。由數據對照，第二次基測各學科量尺分數平均值皆高於第一次基測各學科平均值。二次基測量尺分數總平均值之差：90年度為11.2，91年度為7.7，93年度為19.9。這些數值顯示參加第二次基測**考生的**量尺分數大大地高於第一次基測**考生的**量尺分數。

表六　二次基測學科量尺分數平均數、標準差、總平均數

學科		901	902	911	912	921	922	931	932
國文	平均數	30.0	33.1	30.0	32.1	30.0		30.0	33.7
	標準差	13.2	12.4	13.2	10.6	14.1		14.6	12.5
英語	平均數	29.8	31.2	29.9	31.8	32.9		29.7	33.9
	標準差	17.8	18.7	14.2	16.0	17.3		18.9	17.1
數學	平均數	30.0	32.5	29.9	32.7	29.9		29.9	34.6
	標準差	12.2	13.1	11.8	13.5	17.5		16.5	13.6
社會	平均數	29.9	31.6	30.0	31.3	29.9		29.9	34.3
	標準差	13.4	13.7	12.6	12.7	13.2		13.3	11.7
自然	平均數	29.9	32.4	29.9	29.5	29.9		30.0	33.0
	標準差	12.4	10.6	12.0	12.0	12.8		13.3	13.1
總平均數		149.7	160.8	149.7	157.4			149.5	169.5
總平均數之差			11.2		7.7				19.9

註：從試題報告得到的數據

鑑於第二次基測考生人數約為參加第一次基測考生人數之55%至62%，第二次考生平均素質可能優於第一次考生，應該以重複考生之量尺分數為對照基準。筆者從申請資料(三萬多筆)得以計算相關信息；申請資料依據母體比率抽樣而得，第二次基測考生人數亦為第一次基測考生人數之55%至62%。表七為重複考生的前後二次基測學科量尺分數的平均值、總平均值。

表七　重複考生二次基測學科量尺分數平均數、總平均數

重複考生		90年度		91年度		92年度		93年度	
(repeaters)	學科	**考生人數**	平均值	**考生人數**	平均值	**考生人數**	平均值	**考生人數**	平均值
第一次	**國文**	20048	30.9	22579	32.1	21624	33.1	17491	33.3
	英語	20048	30.8	22579	31.9	21624	33.5	17491	33.8
	數學	20048	30.7	22579	31.6	21624	33.7	17491	33.6
	社會	20048	31.0	22579	32.2	21624	32.9	17491	33.1
	自然	20048	30.8	22579	31.8	21624	32.6	17491	33.0
總平均數	未等化	20048	154.2	22579	159.6	21624	165.8	17491	166.6
第二次	**國文**	20048	33.4	22579	32.5	21624	33.6	17491	34.2
	英語	20048	31.8	22579	32.4	21624	34.5	17491	34.3
	數學	20048	32.9	22579	33.1	21624	35.7	17491	35.0
	社會	20048	32.1	22579	31.7	21624	32.2	17491	34.6
	自然	20048	32.8	22579	30.0	21624	32.1	17491	33.5
總平均數	等化	20048	163.0	22579	159.7	21624	168.1	17491	171.6
總平均數之差			8.7		0.1		2.3		5.0

從申請資料得到的數據

表七數值顯示，以重複考生的量尺分數為基準，第二次平均值仍高於第一次平均值。 90至93學年度總平均值之差依序為8.7分，0.1分，2.3分，及5.0分。換言之，除91學年外，重複考生的第二次量尺分數平均值顯著地高於第一次。第二次量尺分數經過連結非等化程序而得，這顯著的高平均值是反映整體學生真正進步現象?或是這顯著的高平均值來自等化程序的缺失? 筆者質疑後者為主因。

表八數值為重複考生的試卷原始分數的平均值及標準差，是未經過心測中心等化程序的作答能力。93學年度而言，二題本原始分數的平均值及標準差皆相近，何以第二次量尺分數總平均值卻多出5.0分； 92學年度亦然，第二次量尺分數總平均值多出2.3分。 91學年度，自然學科二題本原始分數的平均值相近，何以第二次量尺分數平均值低於第一次平均值約2.0分。90學年度，數學考科二題本原始分數的平均值相近，何以第二次量尺分數平均值高於第一次達2.2分之多，而量尺分數總平均值可增加至8.7分。

上述數據比對，處處顯示等化程序有利於參加第二次基測考生。在此情況下，擇優政策是否不利於只參加第一次基測考生? 下文分析擇優排擠效應。

表八　重複考生二次基測學科原始分數平均數、總平均數

學科		90年度			91年度			92年度			93年度		
		人數	平均值	標準差	人數	平均值	標準差	人數	平均值	標準差	人數	平均值	標準差
國文	第一次	20048	31.9	8.6	22579	34.4	9.2	21624	35.1	8.1	17491	35.0	8.7
	第二次	20048	32.9	9.0	22579	32.9	8.7	21624	34.1	9.3	17491	35.3	8.6
英語	第一次	20048	27.8	12.4	22579	25.2	11.0	21624	30.6	12.3	17491	30.9	12.3
	第二次	20048	29.6	13.4	22579	27.9	12.2	21624	29.7	11.5	17491	31.0	12.2
數學	第一次	20048	18.9	6.8	22579	17.0	6.9	21624	21.5	7.1	17491	21.4	7.8
	第二次	20048	18.9	7.4	22579	19.2	7.6	21624	21.6	7.3	17491	22.3	7.0
社會	第一次	20048	43.3	13.0	22579	40.9	11.3	21624	44.4	11.2	17491	45.1	11.4
	第二次	20048	41.3	12.7	22579	42.6	11.7	21624	43.5	11.3	17491	45.8	11.2
自然	第一次	20048	34.1	11.5	22579	35.6	11.8	21624	37.5	11.6	17491	39.1	11.3
	第二次	20048	36.0	10.8	22579	35.5	12.2	21624	38.6	11.6	17491	39.7	11.7

從申請資料得到的數據

擇優名次排擠效應分析。 令stot1及stot2 代表考生的第一次及第二次基測量尺成績總分， 而bstot代表用來分發排序之量尺成績，亦即擇優後的量尺成績總分。 stot1 stot2、bstot的意義可由下表四位考生的資料自行說明之。表中甲、丁二考生只參加第一次基測，因而其在bstot之分數不變；乙、丙 二考生參二次基測，因而其在bstot之分數為擇優分數。

	第一次 基測量尺總分	第二次 基測量尺總分	分發量尺總分 （擇優/等化）
考生	stot1	stot2(等化)	bstot
甲	200	.	200
乙	198	203	203
丙	201	197	201
丁	220	.	220

參加二次基測

依據上表意涵，重複考生(如乙、丙 二考生)有三種量尺總分(stot1，stot2，及bstot)；整體考生(如甲、乙、丙、丁四考生)有二種量尺總分(stot1及bstot)。筆者就申請資料計算:重複考生的第一次基測量尺總分(stot1)、第二次基測量尺總 (stot2)、及分發量尺總分(bstot)的平均值，平均值之差值，及進步人數及百分率(stot2-stot1及bstot -stot1)；這些數據皆摘列於表九。表九亦摘列整體考生的第一次基測量尺總分(stot1)及分發量尺總分(bstot)的平均值，及二者之差值。

表九數據顯示90至93年度，重複考生的分發量尺總分(bstot)的平均數皆大大地超過第一次基測量尺總分平均數。90至93年度二者的平均差值依序分別為11.1，5.6，7.0，及。8.5。這些高平均差值來自於大部份重複考生的第二次基測量尺總分(stot2)優於其第一次基測量尺總分(stot2>stot1)。90至93年度進步人數百分率依序分別為72.1%，49.2%，55%，及62.5%；而就stot2>=stot1的人數百分率而言，則依序分別增加為74.5%，51.9.2%，57.9%，及65.2%。

表九 重複考生、所有考生之第一次基測量尺總分及分發量尺總分的平均數及差值

							進步人數	進步人數
						平均數	及百分率	及百分率
	年度			人數	平均數	差值	#(stot2>stot1)	#(bstot>=stot1)
重複考生	90	第一次基測量尺總分	stot1	20048	154.2			
		第二次基測量尺總分	stot2	20048	163.0	8.7	14455 (72.1%)	
		分發量尺總分(擇優/等化)	bstot	20048	165.4	11.1		14936 (74.5%)
	91	第一次基測量尺總分	stot1	22579	159.6			
		第二次基測量尺總分	stot2	22579	159.7	0.1	11102 (49.2%)	
		分發量尺總分(擇優/等化)	bstot	22579	165.2	5.6		11717 (51.9%)
	92	第一次基測量尺總分	stot1	21624	165.8			
		第二次基測量尺總分	stot2	21624	168.1	2.3	11896 (55%)	
		分發量尺總分(擇優/等化)	bstot	21624	172.8	7.0		12521 (57.9%)
	93	第一次基測量尺總分	stot1	17491	166.6			
		第二次基測量尺總分	stot2	17491	171.6	5.0	10924 (62.5%)	
		分發量尺總分	bstot	17491	175.1	8.5		11408 (65.2%)
整體考生	90	第一次基測量尺總分	stot1	36233	150.6			
		分發量尺總分(擇優/等化)	bstot	36233	156.7	6.2		
	91	第一次基測量尺總分	stot1	36331	151.1			
		分發量尺總分(擇優/等化)	bstot	36331	154.6	3.5		
	92	第一次基測量尺總分	stot1	37454	150.8			
		分發量尺總分(擇優/等化)	bstot	37454	154.8	4.1		
	93	第一次基測量尺總分	stot1	30150	151.0			
		分發量尺總分(擇優/等化)	bstot	30150	156.0	4.9		

#(btot >=stot1)=20048

重複考生在擇優政策下(bstot>=stot1)促成整體考生的分發量尺總分(bstot)的平均數亦超過第一次基測量尺總分平均數(stot1)：90至93年度二者的平均差值依序分別為6.2，3.5，4.1，及。4.9。在此情況下，整體考生依據二總分的分發排序的等第是會不同的，其中只參加第一次基測考生是較吃虧的。吃虧的程度可由下文分析見端倪。

只參加第一次基測考生名次變化分析。筆者就考生的分發量尺總分(bstot)及第一次基測量尺總分(stot1)排序考生。換言之，筆者計算全體考生在此二量尺總分的名次(300分的名次為1)。只參加第一次基測考生其分發量尺總分(bstot)及第一次基測量尺總分(stot1)是不變的(因為沒有第二次基測成績)。此群考生在二量尺總分名次的變化可作為擇優排擠效應的指標。筆者進一步依照全國四個主要分發區的考生人數比率估算名次變化的範圍(range)。(四個主要分發區為台北分發區、中投分發區、高雄分發、區屏東分發區。主文只以台北分發區為例)。

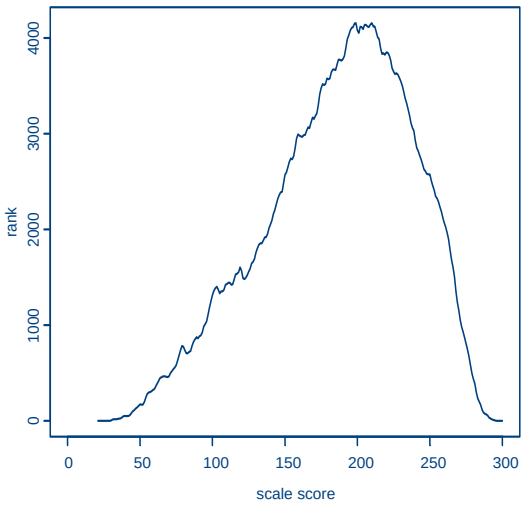
圖八為90-93年度台北分發區考生(台北縣、台北市及基隆市)名次變化分佈圖，縱座標為名次差距(rank(bstot) - rank(stot1))，而橫座標為考生的第一次基測量尺總分。圖八欲表達的信息為:在第二次基測考完後(等化/擇優後)，未參加第二次基測考

生其名次被多少人贏過，亦即其名次變化的範圍。

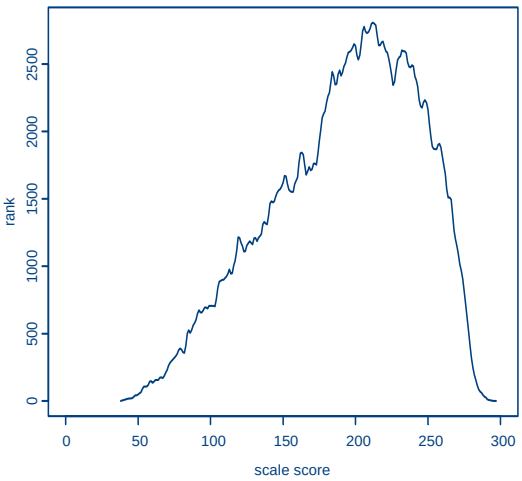
90-93年度考生名次變化的範圍依序為0-4000名，0-2500名，0-3000名、及0-3000名。這些範圍皆>=0，代表未參加第二次基測考生名次皆被往後掉。名次穩定的考生只有二群：高分群考生（stot1>=280）及低分群考生（stot1＜100）；第一次基測量尺總分界於280-100考生名次變化（被贏過）的範圍皆大於1000名至4000名（或至2500名、3000名）。換言之，這大範圍名次變化可讓考生分發進入不同學校就讀；這大範圍名次變化可造成一至三間學校差距。圖八數據處處顯示擇優名次排擠效應是存在的。擇優名次排擠效應在其他三區亦是存在的；但以台北分發區最為顯著的，最易造成較大負面影響因為差1-2分就可差一間學校。

筆者特別強調200分的量尺分數為公、私立學校分界處。而此分數名次變化的範圍，90-93年度依序約為4000名、2500名、3000名、3000名。換言之，以90年度為例，得200分的考生原以為可望進入公立學校就讀，但第二次基測考完後，因為約有4000人的新量尺分數大於200分，其名次被往後排擠約4000名，因而被分發進入私立學校就讀。

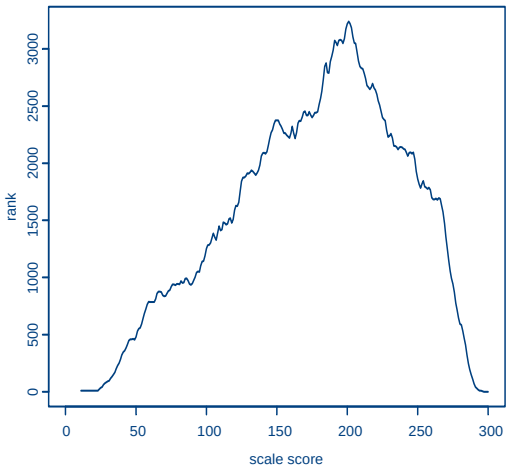
[90年度台北分發區(88206考生)]



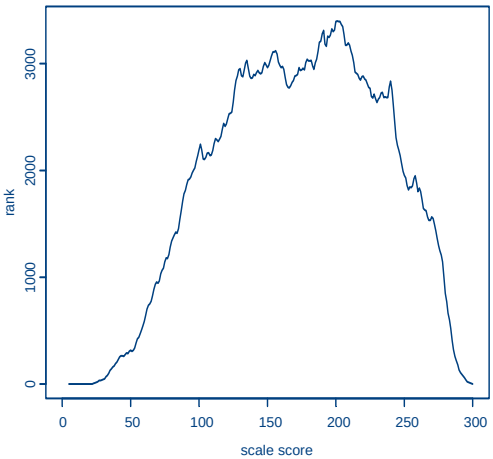
[91年度台北分發區(87525考生)]



[92年度台北分發區(91714萬考生)]



[93年度台北分發區(94223萬考生)]



圖八、 90-93年度台北分發區考生(未參加第二次基測)名次變化範圍，縱座標為名次差距(rank(bstot) - rank(stot1))，而橫座標為考生的第一次基測量尺總分。

五、結論及建議

筆者仿照基測宣導文的作法，就本文探討議題逐一作「答客問」。

1）考生基測成績通知單的量尺分數的計算是正確的嗎?

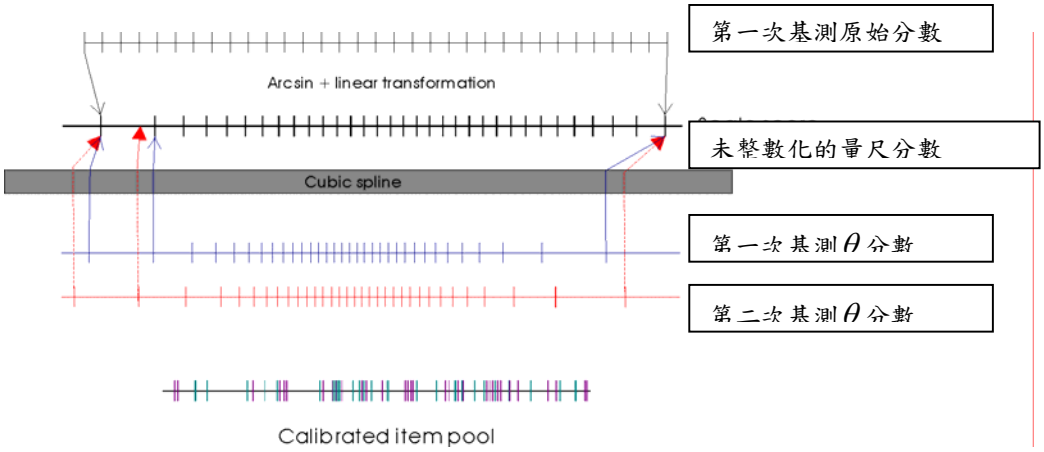
答: 錯誤的、不正確的。 考生基測成績通知單1-60分的考科量尺分數是被『調整』過的,沒有按照公告所定的計分遊戲規計算考科量尺分數,加入了莫名奇妙的『調整』: 有的考科量尺分數是以最高分64.34計算的，有的考科量尺分數是以最高分59.56計算的。如此一來，考生得到的量尺分數和正確算法(以最高分60計算)的量尺分數是不一樣的，產生加分或減分的效應。就五科量尺總分而言，加減分的範圍:[-4分至+4分]，[-11分至+14分]，[-9分至+7分]及[-4分至+3分]（90-93年度依序範圍)。

2）各考科的 1-60 分的**量尺分數代表對**分發總分影響影響力是一樣的嗎?

答:不一樣的。心測中心 1-60 分的量尺分數設計只是台面上的假像相同比重。均等條件測量標準誤量尺分數型態是藉量化函數 s^* 尋求均等量尺分數的測量標準誤(SEM)，而不是均等量尺分數標準差(SD)，無法讓基測考科對量尺總分的比重是一樣的。況且，心測中心以不正確的測量標準誤計算 1-60 分的量尺分數，考科量尺分數標準差(SD)和正確算法的標準差也不一樣，產生上述**量尺分數**加減分效應。連帶地，影響**量尺**總分的分發名次變化，變化程度可造成考生分發至不同學校就讀，變化程度可造成公、私立學校分發之落差。

3）心測中心的等化程序是真正的 IRT-等化程序嗎?

答: 不是正確的 IRT 等化程序。下圖可具體的呈現基測中心只作連結動作。第二次基測與第一次基測 IRT-θ 值皆由題庫校準試題參數估計，可合併在同一個 IRT θ-scale 上一起比較。藉 cubic spline 將第二次基測 θ 值連結至第一次基測的量尺分數轉換表。此連結程序沒有達到等化所要求的對稱、反函數的概念；而是類似於迴歸函數的預測特質，雖保留題本內原始分數、能力值、量尺分數三者之間等第(rank)的一致性，但未必保留二題本原始分數之間的等同性。因而，第二次基測的量尺分數是大約對等值，而大約的範圍是幾分之差就不得而知。考生第二次量尺分數顯著的高平均值是反映整體學生真正進步現象?或是這顯著的高平均值來自等化程序的缺失? 或是這顯著的高平均值來自人為加分動作?



4）擇優政策是否不利於只報考第一次基測的考生?

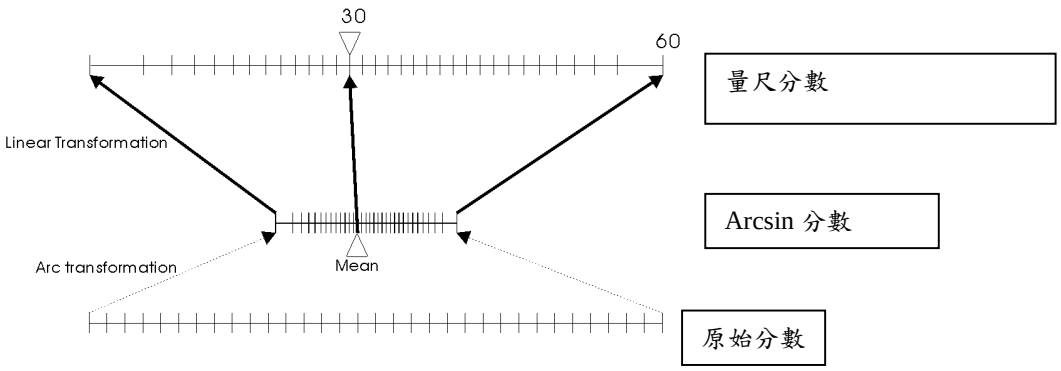
答:絕對是的。 在第二次基測考完後(等化/擇優後)，大部份未參加第二次基測考生的分發名次大大地往後掉。台北分發區考生而言，往後掉的名次足足可達 4000 名。本文發現心測中心公佈的第二次基測量尺總分，重覆考生平均值以 8.7 分，0.1 分，2.3 分，及 5.0 分高於第一次基測平均值。參加二次基測的考生有優勢的機會以較佳的第二次基測成績作為為分發成績。教育部應該將擇優好處、有利機制清楚地告訴家長及考生，不應該為了避免造成「二次聯招」之批評，而不鼓勵考生參加第二次基測。

5）心測中心真的有建構題庫嗎?

答: 題庫若有建構，充其量也是試題參數不精準的資料庫。

後者的質疑來源有二： 其一、 90-93 年度第一次基測試題是不符合 Rasch model 的試題特徵曲線；其二、就預試經驗來看，考生在預試時的作答動機不強，或故意亂答企圖影響試題參數，使題庫試題參數估計值無法正確估計正式考生前後二次基測的 θ 能力值，進而影響等化及量尺分數轉換正確度。說白一點，可以讓來源不精準的試題參數決定全國考生的入學命運嗎？

題庫是否有建構的質疑來自心測中心藉『調整』的方式來縮小滿分與錯一題的量尺分數差距。其實，只要題庫內有足夠數量的題目及精準試題參數，透過正弦反函數的不同比例轉換分數點的特性，由題庫選題組合適當難度的題本，就可縮小滿分與錯一題的量尺分數差距。茲以下圖說明要點。



例如,題數固定為31題，原始分數平均值=28、25、及22 的3個題本，經正弦反函數轉換後， 答對題數31題與30題的量尺分數相差依序分別為8分、6分、及4分。換言之，基測中心可由題庫，組合適當題本來縮小滿分與錯一題的量尺分數差距。

基測五學科的題數差異很大，建立均等條件測量標準誤的量尺分數，若題庫的參數不精確，阻礙從題庫中組合適當難度的題本時，勢必要調整量尺分數全距來縮小滿分與錯一題的量尺分數差距，然而這會造成不公平的計分，影響考生權益。

針對題庫的「答客問」，筆者建議教育部應該監督心測中心的題庫品質。試想自民國88年來，教育部每年花費至少8千萬建立題庫，應該派員或第三公正人士監督題庫品質，並督導量尺記分與等化的內部作業，不應該讓心測中心「裁判兼球員」。

筆者建議教育部誠實地面對基測成績是作為分發入學依據而不是作為門檻功用，應該以常模意義的量尺分數型態(scale with property of normative meaning)取代目前心測中心所用的量尺計分方式。筆者認為心測中心在民國90年基測第一次實施後，就應該向教育部報備基測量尺分數計分相關缺失。或至遲在民國93年，九年一貫課程實施時，就應該改變計分方式。

最後，筆者考量基測公義層面效應，呼籲基測成績計算應該回歸「簡單、公平」的計分方式。筆者下個總結:聯招讓考生痛苦，計分是公平；基測讓考生更痛苦，計分是不公平的。試問家長考生會選擇那一個，聯招或基測？

參考文獻

1 國民中學學生基本學力測驗推動工作委員會(2002)。國民中學學生基本學力測驗專輯：伍、「國民中學學生基本學力測驗」分數，355-366。

2 國民中學學生基本學力測驗推動工作委員會(2002)。國民中學學生基本學力測驗專輯：柒、九十年「國民中學學生基本學力測驗」答客問，389-404。

3 涂柏原(2002)。國中基本學力測驗量尺分數的說明（上）。飛揚第 17 期 91 年 10 月

4 涂柏原(2002)。國中基本學力測驗量尺分數的說明（中）。飛揚第18期91年11月

5 涂柏原（2003)。國中基本學力測驗量尺分數的說明（下）。飛揚第19期92年3月

6 劉長萱(1999)。一年多試分數等化及相關問題研究-子研究五：學科能力測驗級分研究。

7 涂柏原，陳柏熹，章舜雯，林世華(2000)。基本學力分數的建立。國民中學學生基本學力測驗推動工作委員會。

8 國中基本學力測驗量尺分數的計算(2006)。國民中學學生基本學力題庫發展組。飛揚第 38 期 95 年 3 月分。

9 林妙香(2005)。國中二次基測成績等化程序之探研。(NSC 93－2511－S－001－001 (2))。

10 林妙香、謝育泰(2006)。 Extended Four-Parameter Beta-Binomial Model as a Mental Testing Model ---Theoretical Development and Case Study.
http://www3.stat.sinica.edu.tw/library/c_tec_rep/2006-14.pdf。

Carlin, J. B., & Rubin, D. B. (1991). Summarizing multiple-choice tests using three informative statistics. *Psychological Bulletin*, Vol. *110*, No. *2*, 338-349.

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. (1988). Defining Score Scales in Relation to Measurement Error. *Journal of Educational Measurement*, Vol. *25*, No. *2*, 97-110.

Kolen, M. J., & Hanson, B. A. （1989）. Scaling the ACT Assessment. In R. L. Brennan （Ed.）, Methodology used in scaling the ACT Assessment and P-ACT+ （pp. 35-55）. Iowa City, IA:

ACT, Inc.

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., & Brennan, R. L.（1995）. Test Equating-Methods and Practices. New York: Springer Series in Statistics.

Holland, P. W., & Rubin, D. B. (1982). Test Equating. London/ New York:Academic Press.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, Vol. 30, No. 3,* 239-270.

Lord, F. M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Lord, F. M (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F.M. (1984). Comparison of IRT true score and equipercentile observed-score "equating". Applied Psychological Measurement, 8, 42-461.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, Norming, and Equating. Part 1: Theory and General Principles-chapter 6 in Educational Measurement (3rd edition) edited by Linn, R. L. National Council on Measurement in Education.

Maritz, J. S. & Lwin, T. (1989). *Empirical Bayes Methods* (2nd edition). London/ New York: Chapman and Hall.