教育資料與研究雙月刊
第68期　2006年 2 月　頁63-78

國中基本學力測驗使用量尺分數之檢討與省思

余民寧

摘 要

本文針對行之有年的國中基本學力測驗量尺分數的使用問題，提出學理與實務上的反省與評論。從學理上而言，現行量尺分數的使用，有其學理上的優勢與強而有力的實徵證據支持；從實務上而言，卻有不易被大眾理解與釋懷質疑之處；但從政策上而言，它卻已經部分達成當初的教改使命。因此，是否需要改變現行量尺分數的使用，各有其考量的優劣點存在，很難斷然回答是或否。不過，從本文的檢討與省思中可知，我們尚需要考量更多的研究報告、民意的觀點、與政策達成效益幾個層面，進一步審慎思考，才能做出一個完美的政策決策。

關鍵詞：國中基本學力測驗、量尺分數、試題反應理論

余民寧，國立政治大學教育學系教授
電子郵件爲：mnyu@nccu.edn.tw
來稿日期：2006年2月17日 ； 修訂日期：2006年2月19日 ； 採用日期：2006年2月21日

Reflections and Comments on the Scaled Scores Used in the Basic Competency Test for the Junior High School

Min Ning Yu

Abstract

This paper had reflected and commented on the usage of scaled scores in the Basic Competency Test for the Junior High School. Theoretically speaking, the current usage of scaled scores had many advantages and got strongly supported by the empirical evidences. In practice, the usage of scaled scores caused many doubts and was not easily understood by common layers. But from the perspective of policy making and execution, the usage of scaled scores had partially fulfilled the goals of the Education Reform. So, was it necessary for changing of the usage of scaled scores? It was really hard to answer this question: yes or no. Therefore, we needed more evidences of the empirical studies, sincerely thought about people's opinions, and considered the effectiveness of the policy execution. Then, we could make the more perfect policies again.

Keywords: basic competency test for the junior high school, scaled scores, Item Response Theory

Min Ning Yu, Professor, Department of Education, National Chengchi University

E-mail: mnyu@nccu.edn.tw

Manuscript received: February, 17, 2006 ; Modified: February, 19, 2006 ; Accepted: February, 21, 2006

國中基本學力測驗實施至今，已迄五年餘，諸多相關研究報告及問卷調查結果也紛紛出籠（余民寧、賴姿伶、劉育如，2004；余民寧，2005；余民寧、賴姿伶、劉育如，2005；余民寧、謝進昌，2006），已到了該檢討的時候了。根據上述相關文獻的評閱，國中基本學力測驗需要檢討的地方甚多，限於篇幅，本文僅著眼於「量尺分數的使用」之議題，做深入的檢討與反省，以作爲激勵未來改進升學考試之參考；至於其他議題的反省與檢討，則有待學界先進們的共襄盛舉，爲將來建置一個更公平、更客觀的升學考試方式做把關的工作。

一、量尺分數的使用現況

現行使用的國中基本學力測驗的量尺分數（scaled scores），每一考科得分的計算是從1分到60分不等的計分方式，五個考科得分加總後，爲一種5分到300分不等分佈的總得分。除了量尺分數的呈現外，考生的成績單上會同時出現根據量尺總分換算出來的百分等級分數（percentile rank），即PR值。考生即根據此份成績單去甄試、申請、或接受分發到某個高中或高職學校就讀，這種新式的考試升學方式，也已經行之五年，雖然社會大眾對此教育改革仍感疑惑不平之時，但卻也逐漸習慣於默然接受這種新式的升學遊戲規則。

根據當初的規劃（鍾叡賜，1999；林世華，2000），國中基本學力測驗量尺分數的使用，是根據試題反應理論（item response theory, IRT）爲理論基礎（李源煌、楊玉女，2000；王立行，2001），以單參數Rasch模式作爲測驗等化（test equating）的機制來建置4,600題的題庫試題，供往後一年兩試的國中基本學力測驗使用（涂柏原、章舜雯，2000；陳柏熹，2000），而於2001年開始施測後，乃依據Keats & Lord（1962）及Lord（1965, 1969）的強眞分數模式（strong true score models）理論（即假設考生的原始分數會呈現一種廣義的beta-binominal分配）爲計分公式，並輔以正弦反函數的非直線轉換（arcsine transformation）方式來穩定測量標準誤（standard errors of measurement）的計算（涂柏原，2005），將每位考生的原始得分（即答對題數）轉換到一個相同單位的量尺上，此即目前所使用的量尺分數計算的來龍去脈。

這種量尺分數的使用，有幾項優點：

（1）計分背後的學理堅強雄厚，深獲學術界的肯定與背書；

（2）各科分數單位相同，方便跨考科之間的比較；

（3）測量標準誤亦一致，可減少因爲測量誤差所造成的估計不公平現象；

（4）即使一年多試，也可以進行跨考次之間得分的比較，增進得分比較的公平性；

（5）計分程序及常模建立均標準化，即使跨年度得分之間的比較，亦是相對公平的；

（6）政府可以充分掌握有關國民教育發展素質的長期證據，作爲規劃教育優先補助區及其他教育政策決策的參考。

這些優點，使得現行量尺分數的使用在政府推行教育改革之時，爲達成當初規劃的教育改革目標之一：「高中職社區化」，而取得堅強學理的依據與支持，於是成爲教育部最後的政策

決策。

但是，此項政策的決策過程並非十分嚴謹，考慮還不夠十分周詳，不僅未經過學術界的學理辯論、未獲實徵研究數據的充分佐證、以及未獲社會大眾多數民意的支持與信任，即貿然決定實施，因而造成在民國90年剛實施初期，教育部花費不少心力與經費，到全國各地巡迴宣導、解釋、與溝通作法，至今雖然已經過了五年，但是社會大眾還是似懂非懂。但由於升學考試的辦理機構僅此一家（即官方舉辦），別無分號，社會大眾雖然還在質疑國中基本學力測驗的考試公平性問題，但是整體而言，學生與學校雙方對此政策的實施感受，雖不滿意，但還能接受（余民寧，2005；余民寧 等，2004，2005）。話雖如此，量尺分數的使用問題還是值得提出來再做檢討與省思，以期能夠做到促進未來的升學考試更加公平、更加客觀的地步，這樣子的升學政策即是一種良好政策，才能深獲全國民眾的支持與擁護。

二、為什麼需要檢討？

量尺分數的使用，真的能夠解決考試計分的公平性問題嗎？其答案是否定的。那麼，量尺分數的使用也真的一無用處嗎？其答案也是否定的。這就是筆者之所以要提出來檢討與進行省思的地方所在。

（一）從學理基礎來比較

國中基本學力測驗所使用的量尺分數，是依據試題反應理論發展而來。從測驗理論的發展角度來說，試題反應理論比古典測驗理論（classical test theory, CTT）假設更嚴謹、理論主張更雄厚堅強、推理更合乎數理邏輯、演算過程更嚴密、佐證證據更具說服力、應用廣泛且更周延；此外，試題反應理論所持的基本假設，如：大樣本、單向度、局部獨立性、參數估計方法（如：最大概似估計法（maximum likelihood estimation, MLE））等，在當前30萬名考生的大規模考試情境下，更容易被滿足而沒有違反基本假設之虞。這些都是試題反應理論的優點，也是適合使用試題反應理論的時機與場域所在，因此，在教育改革時，比較容易獲得新政策的青睞和支持，這也是理所當然的。

然而，試題反應理論比古典測驗理論所使用的數學公式深奧、難懂、複雜、且高度仰賴電腦的運算，這些都是使得試題反應理論難以被廣大社會群眾接受的原因所在。因此，在推廣上，勢必會遭遇比古典測驗理論還要大的困難與障礙。讀者可以回想一下：2001年剛實施基本學力測驗之時，國中基本學力測驗推動小組不知投注多少心力與經費，全省各地到處巡迴宣導、解釋、與溝通作法，但其成效如何呢？大家都懂了嗎？其答案想必是否定的。大家還是一樣似懂非懂的（包括教育部等政府決策機構、學校、家長、與學生），只能被動地信任測驗專家們所建議的而已，這也難怪筆者調查所得：「整體而言，學生與學校雙方對此政策的實施感受，雖不滿意，但還能接受」（余民寧，2005；余民寧 等，2004，2005），也就不足為奇了。

但是，筆者根據大數原理或中央極限定理（central limit theorem）的說法，要點出一項統計分配的基本假設（或甚至是可能的事實），那就是：在30萬名考生所形成的大規模考試情境下（即大樣本時），考生

的作答反應資料，應該都會呈現大眾所熟悉的「常態分配」（normal distribution）的形狀；也就是說，多數人的得分會集中在平均數上下一個標準差的範圍內（約佔68.34％），而在兩極端得分者會佔少數。如果30萬名考生的得分（不論是使用古典測驗理論的原始總分算法或是使用試題反應理論的能力估計值或量尺分數的算法）眞的呈現常態分配的話，那麼，選擇使用古典測驗理論或試題反應理論來計分，其得分的次數分配也就沒有明顯差別了；換句話說，30萬名考生所形成的團體，已經是構成統計學上所謂的「母群體」，而母群體所形成的次數分配，幾乎都會呈現常態分配，在此常態分配下，每位考生得分在團體中的相對位置是不會因爲使用不同測驗理論而改變的。但是，古典測驗理論淺顯易懂，試題反應理論深奧難懂，有無必要抉擇使用試題反應理論作爲計分的理論依據？實在有值得商榷之處。這一點疑問，已經是屬於價值判斷與政策決策的議題，而非測驗專業所能完全解答的了。

（二）從作答組型的實質涵義來比較

傳統聯考的計分方法，使用百分制加權法來計算考生的得分，亦即，考生答對多少題，每題各得多少分，兩者合併加權計算後，便可獲得每位考生的得分成績。這種作法很簡便、淺顯易懂，不需多大的解釋，也都容易使社會大眾明瞭上手，即使社會大眾不知道其背後的學理是根據古典測驗理論的原理而來，但僅憑每位國民具有的基礎算數概念，也都能接受這種計分方式的合理性與公平性。

然而，傳統聯考的計分方法眞的公平、合理嗎？其實也未必然如此。首先，原始得分相同的考生，其眞正的實力相同嗎？這是第一個被質疑的地方。其實，由許多測驗與評量方面的研究（余民寧，1994，2002a，2002b；余民寧、林曉芳、蔡佳燕，2001；余民寧、陳嘉成，2001；林曉芳、余民寧，2001；游森期、余民寧，2006）顯示，原始得分相同的考生，其眞正的實力未必相同，這是因爲每位考生的作答組型（response pattern）（即每位考生在一份測驗中呈現答對與答錯結果的排列方式）未必相同，並且每一試題的難度、鑑別度、和猜測度也未必一致的緣故所致，除非使用不同的評量方式（如：概念構圖法、S-P表法、路徑搜尋算則法、或知識結構診斷評量法）才能彰顯出其間的差異，否則，光是依據傳統的百分制加權法作法，是無法彰顯考生間「表面得分相同，但眞正實質能力未必相同」的實質差異問題。因此說來，古典測驗理論的傳統計分方法，未能達到實質計分精確的地步，也就不見得比較公平、合理了。當然，追求學理上所說的計分公平、實質合理與客觀，是學術研究的重點，並非是社會大眾關心的焦點所在，如果政策上決定要追求學術眞理上的實質公平，則我們還是可以透過溝通宣導的方式，讓社會大眾都能夠接受此一作法。

但是，很諷刺的是，國中基本學力測驗係以試題反應理論中的Rasch模式爲計分基礎的，根據該模式的演算法則顯示，原始答對題數相同者的能力估計值（或其轉換後的量尺分數）是一樣的，這是因爲原始得分是能力估計值的充分統計數（sufficient statistic）的緣故。這種計分方法的精神與傳統聯考的計分方法並沒有兩樣，依然都是漠視每位考生有不同的作答組型，

以及不承認每一試題有不同的鑑別度與猜測度存在的事實所致，故，原始答對題數相同者（即不論考生是在哪一（些）試題答錯），其換算後的量尺分數皆相同。或者，我們也可以反過來質疑地問：「量尺分數相同的考生，其眞正的實質能力也相同嗎？」其答案想必也是否定的，其理由也與上述者相同。

所以說來，哪一種測驗理論的計分有無比較高明之處？從上述的比較說明中可知，我們還有許多可以努力改進的空間。

（三）從測量誤差的觀點來比較

其次，古典測驗理論的傳統計分方法，被第二個質疑的地方即是假設每位考生的測量誤差是一致的。這一點假設雖然是古典測驗理論的特色，但完全遷就於背後理論的數理計算與推理方便的事實，也就成爲最被詬病的地方所在。顯然，在任何考試情境中，假設每位考生的測量誤差是一致的，是一種不合理、也不可能存在的說詞。

但，也是很諷刺的是，國中基本學力測驗的量尺分數的計算，爲了穩定測量誤差不受考生得分分佈的影響，而使用所謂的正弦反函數非直線轉換（涂柏原，2005）方式，來轉換每位考生的答對題數（即原始得分）到共同的量尺上。表面上看來，此作法的確非常嚴謹、合乎數理邏輯的演算與推理、且其背後測量理論的學理說詞亦堅強有力，也眞能說服社會大眾採信。但很顯然地，此決策的思維（維持穩定的測量誤差）亦深受古典測驗理論傳統作法的影響，無法擺脫。如此一來，試題反應理論原本的特色之一：「允許每位考生都有其不同的測量誤差存在」的事實，便成爲子虛烏有，試題反應理論的特色在此無法彰顯，讓人看不出它有比古典測驗理論更高明的地方，這豈不是很可惜！決策作法與理論期望間豈不是自相矛盾呢！

（四）從量尺分數轉換的精確性來比較

根據每年國中基本學力測驗推動小組所公布的量尺分數與原始得分（即原始答對題數）對照表顯示，每一考科的量尺分數是從1分到60 分不等，五科加總後，總量尺分數爲介於5分到300 分不等。因此，我們可以查表得知一個趨勢，那就是：每一考科若只答錯一題（不論答錯哪一題），五科加總後（共答錯五題）的量尺分數總分，會比同樣是答錯五題，但卻是集中在同一考科，而其餘四考科皆滿分的量尺分數總分還低；其後果即是，入學高中職校的志願序，便會因此而後退一、兩個志願，影響非常深遠。當然，這個現象會引發社會大眾的延伸議論：分散答錯比集中答錯的得分還低，是否意味著各科均衡發展遠比不上單科集中發展來得重要？同樣是粗心大意的作答，是否也需要訓練到集中犯錯而不是分散犯錯的地步呢？由此可知，考生於考試當時的作答壓力是十分沈重的，他（或她）們必須戰戰兢兢地作答，並且小心翼翼地檢查與核對，稍微一不留神，即有可能墜入天堂或地獄的天壤之別。這樣的升學壓力，對一位年僅十五歲左右的青少年而言，眞的是負擔不起的沈重！難怪筆者的調查所得：「國中基本學力測驗的實施，是否降低你的升學壓力或補習壓力時？絕大多數學生與學校的回答都是否定的，甚至有不少人回答更加沈重了」（余民寧，2005；余民寧 等，2004，2005），這也就不難想像得到了！

其實，從考生的原始答對題數、IRT的能力估計值、與量尺分數三項指標並列呈現，我們不難發現其間的轉換關係不僅是非直線的，並且也非完全涵蓋周延的。筆者的意思是說，每位考生的IRT能力估計值的值域，遠超過現行量尺分數的刻度之外，因此，在換算之後，有些考生的量尺分數會低於0分，而有些則會高於60分，因而基本學力測驗推動小組的政策乃決定：凡低於1分者皆以1分計算（故，社會大眾可以看見每一考科中，若答對題數低於某一數目以下時，其得分皆以1分來表示的現象），而高於60分者皆以60分計算（同理，讀者亦可見每一考科中，若答對題數高於某一數目以上時，其得分皆以60分來表示的現象）的作法。這項決議對每位考生而言，都公平、公正嗎？想當然爾，其答案一定是否定的。

根據許多實徵例子的演算結果（王立行，2001；曾建銘，2005）顯示，不論考生的得分是以原始答對題數、IRT能力估計值、或量尺分數的方式來表示，考生得分間的相對位置是不容許改變的；亦即，不論使用哪一種分數來當作考生成績，考生得分間的相對地位應該是不會改變的，也不容許改變的，不然，使用不同轉換公式豈不造成更大的計分不公平現象了嗎？因此，眞正量尺分數的使用結果，其值域應該會介於負無窮大（即－∞）到正無窮大（即＋∞）之間，但由於考量到最低得分呈現負值時的概念及作法，不容易被社會大眾所理解和接受，因而，仿照美國ACT的作法稍加改良，掐頭去尾後，故，決議使用介於1分到60 分不等的計分量尺。坦白說，這種現行作法只對中間得分考生的計分而言是公平、精確的，而對兩極端得分的考生而言，卻是不公平、也不精確的。但是我們只見到考生、家長、學校老師、及媒體新聞報導，均只注意到得高分者部分的不公平現象而已，但對於得低分部分的不公平現象，卻很少有人或機構願意出面爲他們伸張正義。

所以說來，以學術眞理的觀點來衡量，現行量尺分數的作法，對每一位考生成績的計分而言，眞的有比古典測驗理論的傳統作法更公平、更精確了嗎？其實，大家都是心知肚明。

（五）從政策達成的觀點來比較

兩年前，應考生家長及學校的集體要求，台北市政府教育局決議公布台北一考區的量尺分數組距，以三分爲一組距，公布台北一考區的考生相對成績。當初，教育部的政策堅決反對台北市的這項作法，其理由是：會引來其他縣市的競爭效尤，增加考生無謂地競爭分數，而增加升學的壓力。中央政府與地方政府爲此一政策決策之看法分歧，鬧得一時不可開交。

然而，筆者事後從台北市政府教育局的檢討會議中得知，公布組距與否，已經不太會影響考生選填第一或第二的升學志願序，充其量，僅影響少數選填第三或第四升學志願序的考生。絕大多數考生及家長的心態和作法是：「如果沒有機會考上第一或第二志願的高中職校的話，就決定就近就讀社區的高中職校（或完全中學、或綜合中學）即可」；換句話說，中央政府的教育改革目標之一：推動「高中職社區化」，在國中基本學力測驗實施多年之後，此政策俺然已經達成大部分。沒有達成的部分，也是最難以達成的部分，那就是：「降低升學壓力」。筆者認爲降低升學

壓力的問題，是無法由變更考試的計分方式來達成的，它與計分方式無關，而是與整個升學制度有關。

這幾年來，政府積極推行九年一貫課程，並且在中小學實施新式的「國民中小學學生成績評量準則」政策，採五等第九分制的評量方法，特別注重實作評量與日常生活考察的重要性，已經逐漸淡化傳統百分制分數「標籤化」的影響力。再加上，教育改革推動高中職校社區化政策，配合高中升學方式改採國中基本學力測驗的實施，以及配合推行多元入學方案政策，在在使得量尺分數的使用，已經促進達到落實「高中職社區化」的政策目標。也就是說，量尺分數的使用，已經淡化成績分數之間微分之差的白熱化競爭趨勢，而讓高中職校能夠更專注於發展有學校本位課程的特色，並且導引國民中學的教學正常化。平心而論，筆者認爲這項政策的達成，可以歸功於國中基本學力測驗使用現行的量尺分數，作爲甄試、申請、或分發入學的依據之一。模糊淡化成績分數之間的些微差距，而集中注意力於發展自身的興趣專長與特色，這就是當前量尺分數使用的最大功勞所致。雖然做得還不夠完美，但已經是瑕不掩瑜了！

綜合上述，量尺分數的使用，確實是有必須加以反省與檢討之處，它眞的能夠解決考試計分的公平性問題嗎？或者眞的一無是處嗎？答案恐怕都不是很容易回答，還需要更多的思考與反省，才能作最後的結論。

三、改變計分方法是否會更好？

量尺分數的種類有很多種，端視使用目的、方便性、與限制而定，現行的方法即是其中的一種選擇。除此之外，我們也許可以考量其他的方法與策略。

由於各考科題數不一、難易不同，甚至，考生作答時所付出的認知思考、作答時間長短、學科的重要性等因素，也都不盡相同，因此，使用標準化的分數（或稱量尺分數）來表示考生成績的好壞、優劣，是一種可行的、也是必然的合理作法。然而，使用標準化分數的表示方法也有好多種，最簡單的分法，即是分成直線轉換（linear transformation）與非直線轉換（nonlinear transformation）兩大類。關於量尺分數的使用問題，王文中、陳承德（2005）的論文有很好的討論與說明。

傳統的標準化分數，如：標準分數（即z分數）、標準九分、T分數、甚至是TOEFL或GRE分數，都是一種直線轉換的標準化分數。其轉換公式爲：

$$\text{新式量尺分數} = \alpha + \beta z$$

（公式1）

其中，α爲新式量尺分數的平均數，β爲新式量尺分數的標準差，z則爲根據原始分數所計算出的標準分數值

（即$z=\frac{x-\mu}{\sigma}$，x爲考生的原始分數，μ爲考生團體得分的平均數，σ爲考生團體得分的標準差。當30萬名考生的原始分數一起使用時，其μ和σ值幾乎即是母群體的平均數和標準差）。例如，傳統的TOEFL和GRE分數量尺，即是取平均數爲500分，標準差爲100分的一種新式量尺分數。因此，不論考生的原始分數爲幾分，一旦轉換成標準分數值（即z值）之後，即

可代入公式1，求出其所欲轉換的新式量尺分數；例如，某考生的原始答對分數爲30分，轉換成標準分數z值爲1，則經公式1的換算，即相當於TOEFL或GRE分數量尺中的600分。再代入常態分配機率的常模對照表，我們即可很快查得他的百分等級（即PR值）約爲84，即表示其原始得分高過於團體中84%的人的得分，他的成績即屬於是前16%名次內的人。

上述這種直線轉換最爲簡單、易懂，即使以試題反應理論來計分時，也可以如此使用。例如，現行量尺分數是60刻度制的分數，據國中基本學力測驗小組公告的數據，該量尺分數的平均數爲30，標準差爲7.75，故，上述公式1的轉換法即爲：

現行量尺分數 ＝ 30＋7.75 z　（公式2）

換句話說，我們只要估計出每位考生在各考科的原始能力估計值（即θ_j值），把它轉換成z值（$z=\frac{\theta_j-\bar{\theta}}{\sigma_\theta}$，$\theta_j$ 爲考生的原始能力估計值，$\bar{\theta}$ 爲考生團體原始能力估計值的平均數，σ_θ 爲考生團體原始能力估計值的標準差。）後再代入公式2計算，即可求出每位考生的現行量尺分數值。不過，從理論上來說，考生在各考科的原始能力估計值（即θ_j值）的值域爲：$-\infty \le \theta_j \le \infty$，當它被轉換成z分數後，z值的值域可以只取範圍值：$-4 \le \theta_j \le 4$ 即可，因爲如此的值域範圍幾乎會包括全部母群體在內（約佔99.99%），超出範圍者幾乎只佔全體的.001%，以30萬名考生來推估，最多只有300個人，幾乎少到可以忽略他們的存在。如果眞的遇到這群特例者時，可以專案方式特別處理，不納入電腦的自動計算中，或者，亦可將之視爲或來特別處理即可。當然，上述公式只是一種簡便的作法，現行量尺分數的算法並非使用直線轉換，而是使用非直線轉換公式（即正弦反函數轉換法）而來，那也眞的就沒有幾個人看得懂了。

如果，公式2所示的量尺刻度，我們不要限制在1到60分之間，而是採取像TOEFL或GRE分數量尺刻度的話（即取平均數爲500分，標準差爲100分），如此一來，我們便可以將考生間的些微得分差距（即根據 θ_j 值的大小即可顯示出其差異）彰顯出來。例如：甲生的 θ_j 值爲1.656，乙生的 θ_j 值爲1.650，轉換成z值後，假設分別爲 $z_{甲}=0.785$ 和 $z_{乙}=0.784$ ，則代入公式2計算，可得：

甲的量尺分數

＝ 500＋100(0.785) = 578.5≒579

乙的量尺分數

＝ 500＋100(0.784) = 578.4≒578

由此可見，考生間些微得分之差，經轉換成放大刻度的量尺分數後，我們即可輕易比較出甲乙兩位考生到底是誰的程度較高？或誰的程度較低？但問題是：我們有必要這麼做嗎？只要放大量尺分數的刻度單位，我們即可輕易地比較出任何兩位考生間的些微得分差距來，因此，30萬名考生的成績便可以切割成約4,001個類別（即五科最低分爲500分，最高分爲4,500分，其間共有4,001個差距刻度），平均每個類別內約有75名考生的得分在內。但是，如此作法可以預期得到的後果是：考生更會爲了些微分數之差距，計較得頭破血流，升學壓力的競爭更會趨向白熱化。難道，這是我們期望的教改結果嗎？目前雖然只用5分到300分刻度的量尺分數，約把30萬名考生的成績切割成約296個類別，平均每個類別內約有1,014名考生的得分在

內，分數間的些微差距較爲淡化，雖然升學壓力仍然很大，但分數些微差距之爭卻較爲緩和，也就是說，爲爭取就讀第一和第二志願序的考生會比較爲些微分數之爭而計較外，其餘較後志願序的考生則較無所謂，他們會覺得：「反正後頭還有一關大學聯考，現在可以不必那麼在乎、計較，就近入學高中職即可」。因此，前述所說教改的目標之一：「高中職社區化」，儼然已經達成政策目標了。

另外一種作法，即是筆者最建議使用的百分制量尺分數法，我們根本無須使用到任何的直線轉換或非直線轉換，只要算出每位考生的眞實分數即可。它的作法如下：我們可以先計算出每位考生的能力估計值（即 θ_j 值）（被用來估計此能力值的數學模式，可以透過測驗學界專家們的辯論與討論之後，決議使用何種計分模式，如：1PL、2PL、3PL、或其他模式皆可），再根據其作答組型，計算出他在此一考科中每一試題上的答對機率（因爲考題是來自題庫，每一試題的特徵指標（如：難度、鑑別度、或猜測度）均爲已知，因此很容易計算），再將全部試題的答對機率加總後，除以總題數，即構成該考生在此一考科上的眞實分數（true score），該分數亦稱作「領域分數」（domain score）（余民寧，1992；Hambleton & Swaminathan, 1985；Hambleton, Swaminathan, & Rogers, 1991）。其計算公式可以表示如下：

$$\pi_j = \sum_{i=1}^{n} P_{ij}(\theta)/n \qquad \text{（公式3）}$$

其中， 即爲考生j的領域分數，$P_{ij}(\theta)$ 表示考生 j 在試題 i 上的答對機率，n爲該考科的試題總題數。上述領域分數的值域爲： $0 \le \pi_j \le 1$，爲了方便說明與比較使用起見，我們可以將該領域分數乘上100，以轉換成百分制分數；換句話說，我們可以根據試題反應理論的理論優勢爲基礎，考量每位考生的作答組型，將任何考生在任何考科上的原始答對題數，轉換成具有百分制分數單位的量尺分數，如此一來，此新式的量尺分數單位與百分制分數的概念相同，無須多費唇舌解釋，社會大眾都很容易理解、接受、與認同使用，同時，它具有試題反應理論學理上的優點，並且也兼顧每位考生不同的作答組型，可說是一種最簡便不過的標準化分數的計算方法。當然，這種百分制的量尺分數刻度會放大到0分（五科全答錯或全缺考者）到500分（五科全答對者）不等，理論上來說，共計有501個類別，平均每個類別約有599名考生的得分在內。此種刻度的量尺分數比現行的放大一些，但比前述的TOEFL分數縮小很多，相對的，如果採用此量尺分數，考生會比現行的作法稍加計較考試分數的高低，區隔升學的志願序會比較容易一些，但升學壓力也會相對提高一點。

因此，改變現行量尺分數的計分方式，是否會更好？眞的需要三思而後行。

四、學術研究與政策實施的兩難

由前節所述可知，各種量尺方法都有其特殊用途之處，何種較優？實在很難回答此問題。然而，不論使用何種量尺分數，「測驗目的」還是應該擺在第一位置來思考。也就是說，我們實施教育改革的目標何在？而國中基本學力測驗的實施，是否有助於達成此目標？

根據筆者（余民寧，2005；余民寧

等，2004，2005）的調查顯示，實施五年的基本學力測驗成效，大致呈現如下的結果：

1.基本學力測驗試題之特色獲得肯定。

基本學力測驗試題大致可被認同具有所宣導的特色（即難易適中至偏簡單程度、取材均勻、具課程內容的代表性、與學生生活經驗結合、命題靈活、有變化、富創意、試題數量適中）。基本上而言，是一份良好的測驗工具。

2.基本學力測驗實施方式大致符合理想。

目前基本學力測驗的實施方式大致可行，即一年舉行兩次、兩次間隔二週至二月、考試的作答時間大致夠用。

3.基本學力測驗能否測得學生的實力，令人質疑。

基本學力測驗能否測出學生的眞正實力，是令人質疑的。因此，以其分數作爲升學的依據之一，其公正性與適當性亦遭到質疑。

4.基本學力測驗所測得學生素質較爲低落，令人擔憂。

與傳統聯考生比較起來，基本學力測驗考生均有普遍學力不足、學習表現跟不上進度、和學習落差相當參差不齊的現象。

5.基本學力測驗試題具有差異功能的問題，值得關心和研究。

仍有少數試題具有差異功能（DIF）現象，雖然數量不多，但仍宜加以重視與設法改進。

6.基本學力測驗結果有呈現雙峰分配的情形。

英語科得分有呈現雙峰分配的現象，此問題值得重視和謀求解決方法。

7.基本學力測驗無助於升學壓力之紓解。

基本學力測驗的實施無助於升學壓力之紓解，亦即無法降低升學壓力、減少補習次數、及減少填寫練習卷次數。

8.基本學力測驗的量尺分數組距應該公布。

全國受試學生、高中職教師、及國中教師，皆一致傾向認爲應該公布量尺分數的全國性和各考區組距。甚至，對於落後的學校、學區、和縣市，政府應該列爲教育資源的優先補助區，但對名列前茅者，不應該再予任何獎勵。

9.量尺分數的計分公平性問題，有待進一步研究和商榷。

量尺分數的使用可能有失公正（兩極端得分者的量尺分數比較不可信），若以此量尺分數作爲申請或分發入學的依據之一，其公平性、合理性、與客觀性，令人關心與質疑。

10.應儘早確定並公布「基本學力」爲何。

可考慮設定精熟標準，作爲畢業的門檻分數或學校評鑑的依據之一。

11.基本學力測驗的未來改進方向。

維持目前考試現況（即考五科、都是選擇題型試題、一年兩次）是多數考生和教師們期望的改變心聲。但增加可能的考試題型（如：書寫題、聽說題）以及提高試題難度，亦是考生和教師們建議的未來考試趨勢。

12.雖不滿意，但可以接受基本學力測驗。

整體而言，雖不滿意，但勉強可以接受基本學力測驗作爲一種升學考試。

由上述調查結果可知，教育改革的目標只能算是部分達成，未來還有許多值得努力改進的空間。其中，量尺分數能否促進考試計分的公平性，一直是飽受質疑

的。雖然如此，但是國中基本學力測驗實施幾年下來，量尺分數的使用已經模糊化考試成績間的區隔作用（這一點仍然飽受批評和質疑（劉約蘭，2004)），卻也逐漸促成「高中職社區化」政策的落實。功過相抵，孰重孰輕，尚有待討論的餘地，很難就此蓋棺論定。

其次，國中基本學力測驗的實施政策定位不明。原先規劃的目標是作爲探詢並確保每位國民是否都具有足夠的基本能力，以奠定具有適應二十一世紀生活能力的國民教育基礎。而以測驗與評量的觀點來看，這種規劃是屬於所謂的「效標參照測驗」（criterion-referenced test, CRT），當然，從考試政策的規劃開始、命題、計分、及成績的運用等等設計與考量，都是朝向效標參照測驗的方向前進。然而，不知這其中的政策轉變機制是如何？2001年開始實施之後，卻把國中基本學力測驗當成「常模參照測驗」（norm-referenced test, NRT）來使用（與傳統聯考的作法相同），實施結果發現考題沒有鑑別力，無法區分考生程度，50%-75%的學生都能答對每一道試題，這些現象與結果一點都不令人覺得有什麼訝異之處，因爲「效標參照測驗」的結果原本就是如此。但是我們卻把它的考試成績拿來當作升學使用，不論甄試、申請、或分發入學，彼此間的分數區隔力都明顯不足，一時之間，讓全國考生陷入一種升學恐慌之中。很顯然地，這是政策規劃與執行不一致，所造成的缺失所在。當然，模糊化分數之後，使得傳統升學的志願序重新洗牌（即明星學校的排行榜逐漸瓦解中），這一點倒是達成了目標，現在的高中職校已經愈來愈難區分出傳統升學的志願序，各校逐漸發展出其學校本位的特色，這或許就是我們所要的教改目標吧！然而，到目前爲止，我們依然不知道國民教育應該具有的「基本學力」爲何？國中畢業生應該具有多少的「基本學力」？至今，我們依然全然不知。國中基本學力測驗考不出「基本學力」，這不是很諷刺嗎！於是，近一年來，教育部又委託國立教育研究院籌備處進行有關國中小學四、六、八年級學生的「基本學力」調查，以探詢應有的「基本學力」爲何，這豈不是多此一舉、畫蛇添足嗎！學術研究成果與政策實施之間的落差何其大，大到令人難以想像得到的地步！

最後一點，有關教育部推行國中基本學力測驗的政策之一，企圖以透過改變升學考試（及其計分方式）及入學方式，就能達到降低考生的升學壓力之目標。筆者認爲這是一個永遠達不到的虛幻目標，事實上，筆者調查的結果也是顯示如此。筆者認爲只要人還活著，就一定會有壓力，更何況社會上各行各業表現傑出者，有哪一位不是面對高度壓力的？壓力與成就之間通常是成正比的，但壓力調節是其中的中介變項；換句話說，壓力可以促使人表現卓越，也可以把一個人打垮，端視個人懂不懂得調節壓力，把壓力轉變成助力。因此，我們需要教給學生知道的一件重要知識，那就是：「如何面對壓力（如升學壓力或其他壓力）？處理壓力？調節壓力？如何把壓力轉變成對自己有利的助力」，而不是一味地符應民眾、媒體、甚或是學術研究的建議與要求，而不斷地修改課程標準與簡化課程內容、降低評量或評分的要求標準、且試題愈出愈簡單化、愈生活化（目前，根據筆者調查結果顯示，已有不少學校及教師們非常憂心九年一貫

課程及國中基本學力測驗實施後的學生素質，是愈來愈不如從前），否則，到後來，應該學習的知識都被簡化掉了、都膚淺化了，學生的素質也就不如別人、不如從前，那將來拿什麼與他國的人才競爭呢？這一點潛在危機，很值得政府重視，並三思未來該有的教育改革政策。

五、未來何去何從？—代結論

其實，一個政策擬定之後，沒有經過一段時間的嘗試實施，很難斷定其結果一定是成功或失敗。政策是推動眾人朝目標方向邁進的一種決心，是屬於價值判斷多過於追求學術研究眞理的問題，雖然可以經由學術研究的佐證或背書，而取得執行時的自信心。然而，在下價值判斷之前，需要審愼地考慮、周延的配套措施規劃、與嚴格縝密的執行步驟，才是決定價值判斷後執行成敗的主因所在。可惜，近六年來，教育部長頻頻換人，這些現象都不利於長遠價值觀的塑造、制訂、與執行的，也就很難有執行成功的政策了！

著眼於本文的主題，國中基本學力測驗已經實施五年餘，現行60刻度制的量尺分數也已經逐漸被大眾所接受，並且也已經習慣它的存在了。如果現在又來一次改變，不論是改成使用何種刻度的量尺分數，其實施結果是否就會比現行的制度還好？是否就是有利於全民素質的向上提升呢？恐怕，這答案未必是肯定的。筆者認爲「考試只是一種遊戲規則」（余民寧，1998）而已，只要命題沒有出現錯誤和偏差（亦即，命題沒有出現試題差異功能（differential item functioning, DIF））現象：例如：報載大學學科考試出現火星文、考選部用閩南語來命題，皆是DIF的現象）、沒有洩題的可能、應考須知與計分方式均事先公告且沒有隱瞞或不公平對待任何考生的情事發生，則一場考試即可被視爲是公平的、公正的。至於，考試前後所衍生出來的問題，都可以經由學術研究習得改進經驗之後，而加以妥善解決。因此，考生只要遵守遊戲規則來參加一場考試即可，一場考試的得失不必看得太重，畢竟，國中基本學力測驗只是個起步而已，往後的人生還有許多挑戰需要去面對，如何習得以正確的態度和方法去面對各種挑戰，反而比斤斤計較一場考試的得失還重要。家長和考生們可以放輕鬆一點，不必過度緊張與焦慮，放鬆才能走得更遠、跳得更高、以及對未來看得更清楚！

最後，下列建議（余民寧，2005）爲筆者的一些淺見，提出來供政府有關單位及學術界的參考：

1. 國中基本學力測驗試題設計的定位與測驗目的，宜重新思考與檢討。這可以透過舉行全國性學術會議來討論及研究，並加以解決。
2. 國中基本學力測驗的公平性問題，宜持續研究與改進。這可以在每年舉行考試過後，透過委託學術界進行研究而加以改善。
3. 學校教師及家長宜合作教導學生如何面對與處理升學壓力的態度與方法，而非一味地改變考試評量方式與修改課程標準。因此，教育部的教、訓、輔政策，宜再多加宣導，加強落實。
4. 持續維護一個公平競爭、公正客觀計分的升學考試制度。目前可以考慮繼續維持現狀的考試計分方式，未來若要改變，則需要有周延的配套措施與嚴格縝

密的執行步驟的規劃才行，不能說變就變，否則，國民教育永遠是處於白老鼠的實驗階段，沒完沒了。

5.正視九年一貫課程及國中基本學力測驗實施後，學生的學業成就或能力分布呈現兩極化的現象，並及早做因應和設法補救。筆者認爲背後的經濟因素可能是問題的關鍵所在，因此，拉近貧富差距、縮小城鄉落差、並提高教育優先區的補助與弱勢族群的補救教學，可能是防止此問題繼續惡化的起點。之後，需要鼓勵教育研究學者針對此一問題進行政策性研究，並提供對策供政府參考。

6.建立長期教育成長資料庫（如：TEPS資料庫或教育研究院籌備處新建的資料庫），以持續追蹤研究學生能力成長的變化趨勢，並做爲政策規劃與制訂預防措施之參考。

7.在教育經費許可下，及早爲推行十二年國民義務教育預作規劃和準備。由於目前國內的高等教育蓬勃發展，未來幾年內可以達到百分之百的升學率，在此情況下，國中到高中階段的升學，其實已經不再需要透過國中基本學力測驗來篩選學生了，何不讓國中基本學力測驗發揮原本判定「基本學力」以謀求補救教學的測驗功能，並且鼓勵每所高中職學校都能發展辦學特色，最後，促進十二年國民義務教育正常化的發展。

參考文獻

王文中、陳承德（2005）。量尺分數之比較：以醫師高考爲例。**國家菁英**，1（1），137-156。

王立行（2001）。標準化入學考試量尺分數的心理計量問題研究。**測驗年刊**，48（1），119-140。

余民寧（1992）。試題反應理論的介紹（四）-能力與試題參數的估計。**研習資訊**，9（3），6-12。

余民寧（1994）。測驗編製與分析技術在學習診斷上的應用。載於國立政治大學教育研究所（主編）：**教育研究方法論文集**（303-327）。台北：臺灣書店。

余民寧（1998）。考試的遊戲規則。**考選情報**Go！Go！，4，30-32。

余民寧（2002a）。**教育測驗與評量—成就測驗與教學評量（第二版）**。台北：心理。

余民寧（2002b）。學科知識結構之評量研究----以「教育測驗與評量」學科知識爲例。**政大教育與心理研究**，25（中冊），341-367。

余民寧（2005）。從調查數據回顧基本學力測驗的實施。**測驗學刊**，52（1），IX-XXXVI。

余民寧、林曉芳、蔡佳燕（2001）。國小學生數學知識結構認知診斷評量之研究。**政大教育與心理研究**，24（下冊），263-302。

余民寧、陳嘉成（2001）。領域知識結構之評量研究----以「垃圾分類處理」領域知識爲例。**政大教育與心理研究**，24（下冊），393-420。

余民寧、賴姿伶、劉育如（2004）。國中基本學力測驗實施成效之初步調查：學生的觀點。**政大教育與心理研究**，27（3），457-481。

余民寧、賴姿伶、劉育如（2005）。國中基本學力測驗實施成效之初步調查：學

校的觀點。**政大教育與心理研究**，28（2），193-217。

余民寧、謝進昌（付梓中）。國中基本學力測驗之DIF的實徵分析：以91年度二次測驗爲例。**國立高雄師範大學教育學刊**。

余霖（2000）。對國民中學學生基本學力測驗之評論與期望。**教師天地**，109，30-32。

李源煌、楊玉女（2000）。建立學科評量量尺之理論基礎。**測驗年刊**，47（1），95-116。

林世華（2000）。跨世紀的測驗發展計畫--國民中學學生基本學力測驗發展研究。**教師天地**，109，4-8。

林曉芳、余民寧（2001）。國中生在數學代數概念學習之評量研究。**政大教育與心理研究**，24（下冊），303-326。

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，章舜雯（2000）。國中學生基本學力測驗的分數及相關議題。**教師天地**，109，9-17。

陳柏熹（2000）。國中基本學力測驗分數的意義與使用。**高中教育**，11，38-47。

曾建銘（2005）。國中學力測驗量尺分數之探討研究--以90年度數學科爲例。**中學教育學報**，12 ，57-89。

游森期、余民寧（付梓中）知識結構診斷評量與S-P表之關聯性研究。**政大教育與心理研究**，29（1），排版中。

雷文、邱兆偉、張拉士、謝霏霏、錢幼蘭（2001）。基本學測總體檢。**師說**，154，4-21。

劉約蘭（2004）。測驗的社會功能與影響--兼論對國民中學學生基本學力測驗的省思。**中等教育**，55（2），136-154。

鍾叡賜（1999）。「國民中學學生基本學力測驗」研究發展計畫（三年計畫）。**高中教育**，6，16-17。

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newburry Park, CA: Sage.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59-72.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30*, 239-270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika, 34*, 259-299.