教育部臺灣省中等學校教師研習會九十三年度研究報告

國中學力測驗兩次測驗量尺分數與等化之探討研究—以90年度數學科為例

曾 建 銘 著

教育部臺灣省中等學校教師研習會　出版
中華民國九十三年十二月

F0039407

國中學力測驗兩次測驗量尺分數與等化之探討研究—以 90 年度數學科為例

摘要

本研究主要探討民國九十年第一次學測與第二次學測，兩次學測間數學科量尺分數與等化，在 Rasch、1PL 和 3PL 等不同模式下，所估計的差異。具體的研究項目包括下列六項：一、探討民國 90 年第一次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。二、探討民國 90 年第二次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。三、探討民國 90 年第一次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。四、探討民國 90 年第二次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。五、探討民國 90 年第一次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現的真分數分佈情形。六、探討民國 90 年第二次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現的真分數分佈情形。

針對此等研究項目，本研究以國民中學學生基本學力測驗推動工作委員會

所提供之資料，以 1PL 和 3PL 等不同模式進行分析，研究工具包括：一、民國九十年國中基本學力測驗第一次和第二次數學科試題，二、電腦軟體：BILOG-MG、MATLAB 和 SPSS。經統計分析後，獲得下列發現：

1. 第一次學測樣本資料經過 1PL 模式校準後，結果發現答對題數在 19 題以下，除了答對 13 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，答對題數高於 20 題的情形者剛好相反，因此量尺分數若是以此種情形呈現，將有利於前半段能力較好的考生，若以變化率而言，此種估計與學測只有 36.36%(12/33)相同，真分數範圍為 0.98~31.46。

2. 第一次學測以 3PL 模式校準估計後，結果發現在答對 17 題之前，大都低於或等於學測之量尺分數，只有答對 2 或 3 題例外，在答對 17 題之後則相反，若以變化率而言，此種估計與學測只有 30.3%(10/33)相同，真分數範圍為 8.19~30.87。

3. 第二次學測樣本資料經過 1PL 模式校準後，結果發現除了答對 6 與 30 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，因此量尺分數若是以此種情形呈現，將使得大部份的學生的量尺分數低於學測的量尺分數，若以變化率而言，此種估計與學測只有 28.13%(9/32)相同，真分數範圍為.97-29.47。

4. 第二次學測以 3PL 模式校準估計後，結果發現除了答對 1、3、19、20 及 21 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，因此量尺

分數若是以此種情形呈現，將使得大部份的學生的量尺分數低於學測的量尺分數，若以變化率而言，此種估計與學測只有 15.63%(5/32)相同，真分數範圍為 8.22~30.3。

5. 根據學測所提供之數學科統計資料，發現第二次考試平均難度高於第一次，但第二次學測的整體原始與量尺平均分數都高於第一次學測成績。

國中學力測驗兩次測驗量尺分數與等化之探討研究—以90年度數學科為例

目 次

摘要 I

目 次 IV

表 次 VI

第一章 緒 論 1

第一節 研究動機與目的 1

第二節 研究問題與假設 3

第三節 名詞解釋 5

第二章 文獻探討 9

第一節 常用之試題反應理論模式簡介 9

第二節 測驗等化 10

第三節 國中基本學力測驗與聯考實務作法的差異 23

第三章 研究方法 29

第一節 研究對象 29

第二節 研究工具 30

第三節　實施程序及資料處理 32
第四章　研究結果與討論 34
第一節　民國九十年第一次國中基本學力測驗數學科校準後所呈現量尺分數之研究結果 34
第二節　民國九十年第二次國中基本學力測驗數學科校準後所呈現量尺分數之研究結果 39
第三節　綜合討論 44
第五章　結論與建議 47
第一節　結論 47
第二節　建議 49

表　次

表一　第一次國中學測母體與樣本資料統計表 34

表二　第一次國中學測 Rasch、1PL 及 3PL 模式之試題參數估計值 35

表三　第一次學測答對題數與量尺分數、1PL 與 3PL 模式之能力、真分數對照表 37

表四　第一次國中學測母體與樣本量尺分數統計表 39

表五　第二次國中學測母體與樣本資料統計表 40

表六　第二次國中學測 Rasch、1PL 及 3PL 模式之試題參數估計值 40

表七　第二次學測答對題數與量尺分數、1PL 與 3PL 模式之能力、真分數對照表 43

表八　第二次國中學測母體與樣本量尺分數統計表 44

第一章　緒 論

第一節　研究動機與目的

回顧我國數十年的教育評量實務，例如高中職、五專及大學聯考，乃至民國九十年開始舉辦的國中學力測驗，這些考試決定大部份國中生進入高中職，及高中生進入某所大學的標準。學生努力了三年，就是為了畢其功於此一役—聯考，此對於考生造成很大的壓力。而其他不可抗拒的因素，如考試當時學生的身心狀況皆可能影響學生作答的成績。此種一年一次的大會考相當不符合人性。假如此種決定性的考試能改為一年多試，不僅能減少考生的壓力，且讓考生有更多的機會來測試自己的能力，此種考試才能更公平、更真實地測出考生的能力。

自從民國九十年起所實施的國民中學學生基本學力測驗，其目的是要評量學生在經過國中三年後所應具備的基礎的、核心的、重要的知識與能力，學生測驗所得的分數將取代過去的高中職及五專聯招的分數作為申請進入高中、高職或五專的依據。國民中學學生基本學力測驗之取材，是以學生學習及生活經驗為主(國民中學學生基本學力測驗推動工作委員會，民 90 年)，其進行方式不同於以往的考試形式，是採取事先命題，經過學科、測驗專家多次修題、審題，再進行預試後，經試題分析取得試題的參數及相關訊息，然後再將合乎標準的優良試題納入題庫之中。在入闈時根據事前公布的測驗目標，從題庫中選取試

題，組成正式題本進行測驗。其目的在於提昇測驗品質、降低測量的誤差，以增加測驗的信度、效度。國民中學學生基本學力測驗與過去的聯考有一很大的不同，是一年舉辦兩次，測驗分數是利用現代測驗理論 IRT 所建立的量尺分數，其測驗結果是客觀、公平的，且兩次測驗的結果是利用測驗等化技術得到的，彼此是可以互相比較的。因此，學生可以選擇報考兩次，或只選擇其中一次報考，然後選擇較優的成績作用申請入學用，以避免一年一試來減輕考生的壓力。面對如何將相同或不同的考生，於第一次學測或第二次學測的測驗結果，校準於同一量尺上作直接比較的問題，正是測驗等化所要解決的問題，也是一般參加國中基本學力測驗學生、家長及老師，所一致關心和重視的問題。

例如，美國 ETS 在全世界所辦的 TOFEL、GRE 和 GMAT 等考試有多試的複本測驗，可以在一年內實施多次測驗，不同受試者可接受不同複本測驗，這些複本測驗經過等化後，測驗的成績即可互相比較。

已有很多關於測驗等化方法的文章及書廣為發表及討論，其詳細內容擬於文獻探討中討論之。因利用試題反應測驗理論(IRT)所建立之等化測驗，已為現代測驗界所廣泛應用，而其中因方法不同所做的等化結果亦不盡相同，本研究擬利用 IRT 之 1-PL、3-PL 的模型，來探討國中學力測驗—數學科兩次測驗間的量尺與等化結果，再與目前的學力測驗等化結果做比較及討論。另外擬應用 IRT 真分數等化法，來探討其與學力測驗原始分數的差異。因此，本研究之主要目的有下列六項：

1. 探討民國 90 年第一次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。

2. 探討民國 90 年第二次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。

3. 探討民國 90 年第一次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。

4. 探討民國 90 年第二次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。

5. 探討民國 90 年第一次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現的真分數分佈情形。

6. 探討民國 90 年第二次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現的真分數分佈情形。

第二節　研究問題與假設

一、研究問題

根據上述之研究動機與目的，本研究擬探討的問題如下：

1. 民國 90 年第一次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變率為多少？

2. 民國 90 年第二次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變率為多少？

3. 民國 90 年第一次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變率為多少？

4. 民國 90 年第二次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變率為多少？

5. 民國 90 年第一次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現的真分數分佈為何？

6. 民國 90 年第二次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現的真分數分佈為何？

二、研究假設

為回答上述之研究問題，本研究之研究假設如下：

1. 民國 90 年第一次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數具有改變。

2. 民國 90 年第二次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數具有改變。

3. 民國 90 年第一次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數具有改變。

4. 民國 90 年第二次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數具有改變。

5. 民國 90 年第一次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值的校準後，所呈現的真分數與學力測驗的原始分數不同。

6. 民國 90 年第二次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值的校準後，所呈現的真分數與學力測驗的原始分數不同。

第三節　名詞解釋

茲將本研究所涉及之重要名詞解釋如下：

一、測驗等化

測驗等化是利用統計過程，來調整不同測驗間的分數，使得不同測驗間的分數可以彼此交換，但這些不同測驗的內容與難度是類似的(Kolen & Brennan, 1995)。

等化可分為水平等化與垂直等化，水平測驗是指兩個或兩個以上測量某一特質的測驗，其考生的能力分配與試題難度相似，其不同測驗間原始分數互相轉換的過程。如國中基本學力測驗，可在一年內實施兩次，這兩次測驗經過等化的程序後，其測驗成績是可以互相比較。垂直測驗是指兩個或兩個以上測量某一特質的測驗，其考生的能力分配與試題難度皆不同，其原始分數互相轉換

的過程，即測驗內容程度高低的等化。如美國加州成就測驗。

二、國中基本學力測驗

國中基本學力測驗發展概述

國中基本學力測驗之緣起係伴隨著高中多元入學理念而來，自民國八十三年第七次全國教育會議後，教育部於八十四年提出「中華民國教育報告書」，即揭示高中教育遠景之一為建立多元入學管道，導引國中教學正常發展。至八十五年十二月行政院教育改革審議委員會提出教育改革總諮議報告書中，亦建議推動多元入學制度。教育部遂於八十七年七月正式公佈高級中學多元入學方案，同年九月公布高級職業學校多元入學方案，同時委託臺灣師大心測中心，以三年時間研究與發展基本學力測驗，並自九十學年度起正式實施(教育部，2003；鄭同僚、張原禎，2001)。

國民中學學生基本學力測驗的緣起及施行理念

民國八十五年教改總諮議報告書中，對入學政策的主張是「推動多元入學制度」，對評量的積極主張則是「基礎學科能力的評量」與「其他項目表現的評量」。國中基本學力測驗就是在前述的前提下所形成的，其目的在評量國民中學學生能力表現及其發展潛能。其主要目標則是在維持制度公平的前提下，消除入學考試對於國民中學教育的不利影響，進而充分發展學生的潛能。因此基本學力測驗的命題方向，偏重在對學生未來的學習與生活有所幫助之基礎的、核

心的、重要的知識與能力；命題概念是完整的、周延的，而非偏狹的、殘缺的。至於測驗難度則是以國中學生平均能力為主，因此過去以「過度學習」和「機械式練習」來提升考試分數的學習方法，將無法在本測驗中獲得預期的效果。希望在這樣的制度與測驗方式之下，能夠還給學生一個理想、正常的學習空間。

國中基本學力測驗的特色(國民中學學生基本學力測驗推動工作委員會，2003)

1. 標準化的

國中基本學力測驗是結合測驗專業與各學科領域所發展出來的標準化測驗。所謂標準化測驗可以從兩個方向來看：第一個部分是測驗發展程序的標準化，這是指測驗發展的過程完全依據標準的測驗編製程序來建立，從各學科領域雙向細目表的擬定、試題的編製與取樣、預試、分析試題以及信、效度的建立，完全依照標準化的測驗編製程序來進行，其主要目的是使測驗題目能有較好的品質。第二部分是施測與計分的標準化，包括固定而且明確的施測程序與指導語、計分標準、分數的計算方式等，其目的是降低測量的誤差，增加測驗分數的客觀性。

2. 可比較的

基本學力測驗使用測驗等化技術所提供的測驗分數，對所有的考生來說，即使不同考生參加不同次別的基本學力測驗，其結果仍然是客觀、公平而且可以互相比較的。

3. 一年多試、一試多用的

基本學力測驗一年舉辦兩次，學生可依自己的意願選擇報考兩次或報考其中一次，測驗分數亦可擇優使用，避免一試定終身之憾。測驗分數可供高中、高職、五專多元入學參採之用。

4. 能力導向的

基本學力測驗以能力為導向，旨在評量學生的基本學力，評量的重點是學生能夠帶著走的能力。

5. 對教學有良性影響的

基本學力測驗的命題理念與實施方式，能夠對國中教育發揮正面的影響，讓老師正常教學、學生快樂學習。

第二章　文獻探討

第一節　常用之試題反應理論模式簡介

在社會科學領域裡，我們常藉由考生作答的反應來估計其潛在的能力，而考生的能力與其在題目反應的關係可藉由試題反應理論來建立(Lord, 1980)，此理論的特色有(1)試題的參數估計值不變性(invariance)(2)能力的參數估計值不變，和(3)能力估計值的測量誤差大小，隨能力不同而異。

現代測驗理論發展至今，已有許多試題反應理論(IRT)模式被發展出來，因為國中學力測驗的內容皆是四選一的單選選擇題，在此，我們僅介紹大多適用於在大型學業成就評量的模式。

(一)1-PL 模式：

1-PL 模式相通於 Rasch 模式，因此又可稱為 Rasch 模式，Rasch 認為受試者的潛在能力 θ 與受試者對試題 i 的反應可以用以下之試題特徵函數表示：

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}$$

θ 表示考生的能力，b 代表題目的難度，若在 Rasch 模式中加入常數 D，通常 D 值為 1.7，則原來的 Rasch 模式則稱為 1-PL 模式，其試題特徵函數表示為：

$$P_i(\theta) = \frac{\exp D(\theta - b_i)}{1 + \exp D(\theta - b_i)}$$

(二) 2-PL 模式：

2-PL 模式比 1-PL 模式多考慮一個試題參數 a，即鑑別度參數，其試題特徵函數表示為：

$$P_i(\theta) = \frac{\exp Da_i(\theta - b_i)}{1 + \exp Da_i(\theta - b_i)}$$

（三）3-PL 模式：

此模式又比 2-PL 模式多使用了一個猜測參數 c，來描述試題，其試題特徵函數表示為：

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp Da_i(\theta - b_i)}{1 + \exp Da_i(\theta - b_i)}$$

c 參數為能力極低時仍有答對該題的機率。

第二節　測驗等化

由於測驗等化牽涉到不同測驗或不同受試者之間的關係，為了使得不同的估計數值可轉變成相同，或容易解釋與應用，測驗學家發展各種測驗等化的理論與技術。

一、建立評量量尺的重要性

若要申請美國研究所(除了商學院外)，大部份的學校都會要求 GRE 的成績，作為他(她)們申請入學之必要條件。假如甲生在今年 1 月份在計量部份考 700 分，語文考 400 分。乙生在今年 3 月份在計量部份考 600 分，語文考 500 分。雖然兩者的考試時間與試卷皆不同，但對於以上兩位考生的成績，我們將會有

一致的看法：甲生的計量部份比乙生好；乙生的語文部份比甲生好。此外，從甲生的計量得分，我們可以推論甲生的計量成績是屬於上，但語文成績部份則為中下，而甲生的數學能力可能比語文好。從乙生的得分，我們可以推論其數學能力是中上，語文能力是中等，同樣地，乙生的數學能力可能比語文好。且間隔一段期間後(如二個月)，甲生與乙生再重考一次，他們的得分應不會有太大的改變，除非在間隔的這一段時間，他們非常用功地準備與復習，或在兩次考試時之身心狀況有極大的差異。

二、等化的設計(王寶墉，民 84；余民寧，民 82；Hambleton & Swaminathan, 1985; Kolen & Brennan, 1995)

1. 單團體設計(single group design)：讓同一組受試者，接受兩個或兩個以上的測驗，是最簡單的一種設計。但易受重覆練習和疲倦因素的影響。
2. 等團體設計(equivalent group design)：以隨機選取或配合抽取能力分配相等的不同受試團體，接受兩個不同的測驗而進行等化。
3. 等測驗設計(equivalent test design)：隨機分派題庫或很多要校準的題目成不同的測驗組合，但有如傳統測驗的平行複本試題，將此平行測驗施測於受試能力分配不必相同的團體，但嚴格的平行測驗實際上不易達成。
4. 校準團體設計(anchor group design)：此設計是以不同測驗施測於不同團體，但校準團體接受每個測驗的施測，而其他團體則只施測一種測驗。此設計的關鍵在校準團體的樣本數決定估計數的穩定性。

5. 校準測驗設計(anchor test design)：此設計是將校準試題放入不同的測驗中，而對不同的團體施測。但將校準試題放入不同的測驗，有下列幾種方式：

甲、內在校準測驗(internal anchor test)：將校準試題放入不同的測驗內。

乙、外在校準測驗(external anchor test)：除了已有的測驗外，另外加測校準試題。

丙、分段校準測驗(cascading anchor test)：不同校準試題放入不同的測驗內。

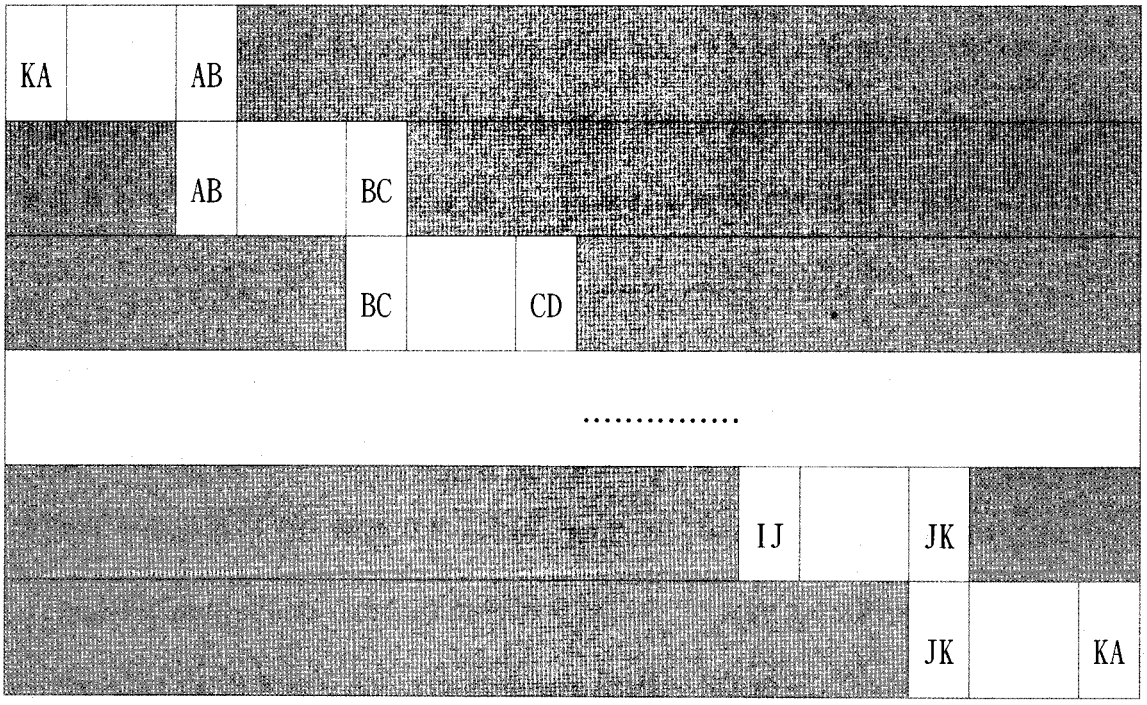
丁、一致校準測驗(uniform anchor test)：相同校準試題放入不同的測驗內。

6. 校準試題矩陣抽樣設計(matrix sampling design with overlapping items)：此設計主要有兩種型態，一是區塊共同試題多重矩陣抽樣設計，另一種是分散型共同試題多重矩陣抽樣設計。兩種將校準試題、非共同試題及不同的受試團體，靈活交互應用，是建立題庫時有效的設計。

7. 同時校準法:藉由測驗資料的收集設計與 IRT 電腦軟體之應用結合，將所收集之數筆測驗資料同時執行試題校準，經過校準之後，所有題目的參數估計值，皆有相同的量尺單位，此乃同時校準法之涵意。

將所收集到的試題，經專家學審題後，組成數份不同之試卷，且試卷裡之

部份題目(定錨試題)是相同的(比率約 15%)(Vale, 1986)，將其施測於預先選好其能力分佈與學測相似的不同學校中，在蒐集並整理測驗的資料中，將幾組樣本對幾份試卷反應資料做調整，並整理成一筆測驗資料矩陣，如下圖。然後以 IRT 電腦軟體，如 BILOG-MG(Zimowski, Muraki, Mislevy & Bock, 1995)，對此一資料矩陣進行試題校準工作。假設此資料矩陣含有 300 道題目，則此 300 題之參數估計值將對應於相同量尺上。經過校準後的試題，必須能夠通過適合度的考驗者，方可被保留在題庫裡，因為它們可以被適當的反應模式所解釋。



灰色部份通常假設其為考生未完成部份。

一般測驗機構所採用的量尺分數根據其產生的方式，大概可以分為兩種：(1) 將原始分數常態轉換後所得 (normalizing raw scores) 之量尺分數。例

如美國教育測驗服務社（ETS）的GRE或TOEFL測驗分數，就是將原始分數常態化轉換後的量尺分數。智力測驗中的比西量表分數（Stanford-Binet Intelligence Scale）及魏氏兒童智力量表分數（Wechsler Intelligence Scale for Children-Revised Form; WISC-R）等也是一種將原始分數常態化轉換後的量尺分數。（2）均等測量標準誤（equalizing measurement error variability）之量尺分數。這是E. L. Lindquist在發展「愛荷華教育發展測驗」（Iowa Tests of Educational Development, ITED）時所提出來建立量尺分數的方法，也是美國知名測驗機構ACT公司的ACT Assessment Test（ACT, 1997）所採用的量尺分數型態。基本上，這是在原始分數轉換成量尺分數的同時，利用數學的方法將每一個量尺分數點上的測量標準誤（或稱測量誤差）調整成相等或是非常接近（Kolen & Hanson, 1989; Kolen, Hanson, & Brennan, 1992）。

測驗等化的方法可分為古典測驗理論與現代測驗 IRT 的方法。古典測驗理論等化法，是利用原始總分進行等化，又可分為線性等化(linear equating)和等百分位數等化(equipercentile equating)。現代測驗IRT的等化法常用的有迴歸法(regression method)、平均數與標準差法(mean and sigma method)、強韌平均數與標準差法(mean and sigma method)、真分數等化法(true score equating)及特徵曲線法(characteristic curve method)(王寶墉，民84；Angoff, 1971 ; Kolen, 1988；Hambleton & Swaminathan, 1985)。

一、古典測驗理論等化的方法

1. 線性等化法：

兩個測驗的分數分配相同，於各測驗中找出對應相同標準分數之觀察分數，即兩測驗之觀察分數可置於相同量尺比較，亦即除了平均數和標準差不同外，X 和 Y 分數的分佈是相同的。如果這個假設能夠成立，那就能從 X 和 Y 找出具有相同的 z 分數配對，這些分數即為等化分數。亦即如果

$$(X-\mu_X)/\sigma_X=(Y-\mu_Y)/\sigma_Y$$

則 X 與 Y 等值。其中 X 與 Y 分別代表測驗 X 和 Y 的原始分數，μ_X、μ_Y 分別代表兩測驗的平均數，σ_X 與 σ_Y 分別代表兩測驗的標準差。其直線轉換為：

$$Y=aX+b$$

$$\mu_Y=a\mu_X+b$$

$$\sigma_Y=a\sigma_X$$

$$Y=(\sigma_Y/\sigma_X)X+[\mu_Y-(\sigma_Y/\sigma_X)\mu_X]$$

2. 等百分位數等化法

等百分位數等化法是在決定兩個測驗的哪些分數具有相同的百分等級。第一步是決定兩個測驗分數分佈上的百分等級。然後將兩個測驗的百分等級對原始分數的散佈圖畫出來。

在連結各資料點時必須利用直線內插法。如假設在 A 測驗卷的原始分數是 a，其對應的百分等級為 P，而在 B 測驗卷，從百分等級 P 所對應的原始分數是

b。則我們可以說一個人在A測驗的得分a與在B測驗的得分b是具有相同能力的。

Lord(1980)認為測驗分數要能公平的等化，需滿足下列需求：

A. 測量不同特質或能力的測驗不能等化。

B. 信度不相等的測驗之原始分數不能等化。

C. 難度不相等的測驗之原始分數不能等化。

D. 除非兩測驗是完全複本測驗，否則兩測驗不能等化。

E. 具有完全信度的測驗分數才可以等化。

此外，對稱性(symmetry)與不變性(invariance)等二條件亦是進行測驗分數等化所必備的。對稱性是指無論是X轉化為Y或Y轉化為X，等化結果必須相同。不變性則指等化的程序是獨立的，不受所選用樣本的影響。

由於古典測驗理論是依據弱勢假設而來，理論淺顯易懂、便於計算，但卻有下列的缺失與限制：

A. 難度、鑑別度與樣本相依，即難度、鑑別度隨樣本的不同而變化。

B. 不同測驗不易比較不同受試者的能力。

C. 理論中假設所有受試者等測量標準誤與實際不符。

亦即以原始分數來進行等化有其限制存在，如公平性、對稱性與不變性等要求常無法獲得滿足。

二、試題反應理論的等化方法

常用的試題反應理論的等化方法有下列幾種：迴歸法(regression method)、平均數與標準差法(mean and sigma method)、強韌平均數與標準差法(robust mean and sigma method)、真分數等化法(true score equating)及特徵曲線法(characteristic curve method)（王寶墉，民84；Angoff, 1971；Kolen, 1988；Hambleton & Swaminathan, 1985)。

1. 迴歸法(regression method)

兩測驗分數的直線關係，可由下列迴歸式來表示：

$y = \alpha x + \beta + e$ ，$\alpha = r_{xy}\frac{s_y}{s_x}$，$\beta = \bar{y} - \alpha\bar{x}$

y，x 可以是能力，即 $y = \theta_y, x = \theta_x$，也可以是難度，如 $y = b_y, x = b_x$。誤差 e 則視為獨立而相等的隨機變項，迴歸係數α與β是斜率與截距。r_{xy}是兩測驗相關係數，s_x, s_y是標準差，$\bar{x}$與$\bar{y}$是平均數。迴歸法的缺點是不對稱，只有 Rasch 模式例外。

2. 平均數與標準差法(mean and sigma method)

平均數與標準差法是假如 $y = \alpha x + \beta$ ，則 $\bar{y} = \alpha\bar{x} + \beta$ 且 $s_y = \alpha s_x$ ，

所以得到 $\alpha = s_y / s_x$，$\beta = \bar{y} - \alpha\bar{x}$ ，此關係是對稱的。

3. 強韌平均數與標準差法(mean and sigma method)

如果連結是包含兩個以上的參數，使用平均數與標準差法的估計誤差變化很

大，而且離異值(outlier)會影響係數的計算。因此，Linn, Levine, Hastings, & Wardrop(1981)提出強韌平均數與標準差法。

強韌平均數與標準差法認為應給每一對(x, y)的能力或參數加權，此加權得自兩個難度參數估計值的變異數中較大一個的倒數，其步驟為：

a. 決定每個配對(x_i, y_i)的加權植w_i，

$w_i = \max\{Var(x_i), Var(y_i)\}$, j=1, 2, ⋯, k,

b. 加權值量尺化，新加權值： $w_i' = w_i / (\sum_{i=1}^{k} w_i)$

c. 計算新的 x 與 y 值

$x'_i = w'_i x_i$ ， $y'_i = w'_i x_i$ j=1, 2, ⋯, k,

d. 求出加權後的平均數與標準差：

$\bar{x}', \bar{y}', s'_x, s'_y$

e. 以加權後的平均數與標準差求出α，β：

$\alpha = s'_y / s'_x$ $\beta = \bar{y}' - \alpha\bar{x}'$

4. 特徵曲線法(characteristic curve method)

上述所提的各種等化方法，都只考慮能力與難度指數，其餘之鑑別度及猜測度都未列入考慮，因此，Haebara(1980)及 Stocking & Lord(1983)提出特徵曲線法，方法如下：

具有θ_a能力的受試者在測驗 X 的真分數為ξ_{xa}

$$\xi_{xa} = \sum_{i=1}^{n} P(\theta_a, a_{xi}, b_{xi}, c_{xi})$$

具有θ_a能力的受試者在測驗 Y 的真分數為ξ_{ya}

$$\xi_{ya} = \sum_{i=1}^{n} P(\theta_a, a_{yi}, b_{yi}, c_{yi})$$

試題參數是：

$$b_{yi} = \alpha b_{xi} + \beta,$$
$$a_{yi} = a_{xi} / \alpha,$$
$$c_{yi} = c_{xi}$$

為求兩真分數間差異的最小值，其函數為

$$F = \frac{1}{N} \sum_{a=1}^{N} (\xi_{xa} - \xi_{ya})$$

用 Newton-Raphson 解出以下偏導數方程式，即可得兩真分數的最小值：

$$\frac{\partial F}{\partial \alpha} = \frac{\partial F}{\partial \beta} = 0$$

因特徵曲線法涉及三個試題參數，等化的效果最佳。

5. 真分數等化法(true score equating)

IRT 能力分數之評量單位不易為一般人所接受，而且長久以來大眾皆習慣採用原始分數來計分。有鑑於此，IRT 原始分數(IRT observer score equating)與真實分數等化法(IRT true score equating)乃被發展運用來克服此問題。IRT 真實分數為考生在 n 題題目答對機率的總和，公式如下(Lord, 1980)：

$$T \mid \theta = \sum_{i=1}^{n} P_i(\theta)$$

真實分數等化方法包括兩個主要步驟(Kolen & Brennan, 1995)：

A. 將兩份試卷的試題參數估計值對應在相同之量尺上。

B. 使用 IRT 分數為橋樑，製做兩份試卷真實分數的對應表。

第二步驟之原理為：假定兩組(分為基礎組與等化組)樣本之能力參數之單位相同，且兩份試卷的試題參數值之單位亦同，將基礎組每一位受試者的 IRT 能力分數轉化為真實分數；同樣地，也等化組每一位受試者的 IRT 能力分數轉化為真實分數。然後以 IRT 分數為橋樑，製造出兩份試卷真實分數的對應表。IRT 真實分數等化方法之最大優點為：當我們從題庫裡選定試題樣本後，即可開始製造試卷間之分數轉化表，而不需要等到考完試取得考生作答資料再來做。但先決條件為題庫裡試題的參數估計值要準確。

6. 觀察分數等化法

真實分數ξ和觀察原始分數 r 有相同的量尺，而

$$r = \sum_{i=1}^{n} U_i$$

若項目反應理論是有效的，則

$$E(r) = \xi$$

一個測驗理論上的觀察分數分佈 $f(r \mid \theta)$ 可以由下列等式中獲得

$$\sum_{r=0}^{n} f(r \mid \theta)\, t^n = \prod_{i=1}^{n} [Q_i(\theta) + tP_i(\theta)]$$

若θ的值已知，如N個考生的能力為$\theta_1,\theta_2,...,\theta_a,...,\theta_N$，則$f(r)$的邊際分配為

$$f(r)=\sum_{a=1}^{N} f(r \mid \theta_a)$$

我們可以利用下列例子來做說明：

假設有一個2參數的logistic模式，有三個題目，其

難度 $b=[b_1,b_2,b_3]=[1.0,0.0,-1.0]$

鑑別度 $a=[a_1,a_2,a_3]=[1.5,1.0,0.5]$

給定下列5個能力 $\theta=[\theta_1,\theta_2,\theta_3,\theta_4,\theta_5]=[-2,-1,0,1,2]$

根據上面的條件，我們可以獲得在不同能力的理論上觀察分數分佈如下：

	條件相對次數$f(r \mid \theta)$				
原始分數	$\theta=-2$	$\theta=-1$	$\theta=0$	$\theta=1$	$\theta=2$
0	.678	.420	.139	.012	.000
1	.313	.500	.475	.143	.009
2	.010	.080	.361	.488	.158
3	.000	.000	.025	.357	.833

假設總共的考生有100人，而且具有能力θ=-2, -1, 0, 1, 2的人分別有5, 15, 30, 40, 及10人，則原始分數的理論邊際次數分配可計算得如下表：

	條件相對次數 $f(r \mid \theta)$					邊際次數分配 $f(r)$
原始分數	$\theta=-2$	$\theta=-1$	$\theta=0$	$\theta=1$	$\theta=2$	
0	3	6	4	0	0	13
1	2	8	14	6	0	30
2	0	1	11	20	2	34
3	0	0	1	14	8	23
考生人數	5	15	30	40	10	100

根據以上，觀察分數等化法的步驟如下：

a. 根據 $\sum_{r=0}^{n} f(r \mid \theta)\, t^n = \prod_{i=1}^{n} [Q_i(\theta) + tP_i(\theta)]$ 及考生的能力及測驗 X 的項目參數，可以得到條件次數分配

b. 再從 $f(r) = \sum_{a=1}^{N} f(r \mid \theta_a)$ 得到邊際次數分配

c. 對於測驗 Y 重覆 a, b 的步驟

d. 用等百分位數來等化此兩個測驗的原始分數

與觀察分數等化法比較，IRT 真實分數等化方法之計算過程較為簡易，此方法無須依賴考生能力之分佈狀態，而且經過 Lord & Wingersky (1983)的研究發現，此兩種方法的等化結果相似，況且 ETS 測驗公司亦採用 IRT 真實分數等化方法。因此，本研究擬採用 IRT 真實分數法來對兩次學測的數學科做等化，然後與國中基本學力測驗中心的結果比較做探討。

第三節　國中基本學力測驗與聯考實務作法的差異

基本學力測驗(以下簡稱基本學測)與傳統聯招考試在作法上是有實質上的不同。這些差異主要是圍繞在四個相關聯的議題：(一)、基本學測的內容（基本學測到底是什麼？）；(二)、入闈組題；(三)、量尺分數；(四)、一年多次（目前是一年兩次）測驗等化的問題(林世華，民91)。

基本學測與傳統聯考最大的差異是：測驗內容與測驗工具發展方法的差異。基本學測主要是設計來評量國中學生在修完國中三年後，所學習到的基本知識與所培養的基本能力。所謂「基本」意指核心的、重要的、統整的。國中三年老師在學校所教的，學生所學習的，包羅萬象，其中主要的當然是基本的範疇，當然也不乏複雜專業方面的涉略。基本學測是排除複雜專業的部分。基本學測的發展方法是以標準化心理測驗發展的科學方法為依據。科學方法步驟中，包括：清楚界定基本學力；測驗試題編寫，審查與試驗研究；測驗分數如何使用等。基本學測研究團隊從八十七年起便著手建造題庫，其中每一試題均需經歷嚴密審查並預試，以便清楚掌握試題的品質。題庫中的試題是不停地在增加與淘汰。這與聯考在考前入闈後才開始命題形成測驗是截然不同的。

其次是入闈組題的議題。基本學測是以持續發展的題庫入闈，並在闈場以自行開發的組題電腦系統來選題，組成當次的基本學測。理論上，系統會組成

哪一種測驗試題，事先大家並不清楚。所能清楚的是組題的原則。由於基本學測的界定相當清楚，所以組題原則也隨之具體。幾個主要原則如：試題評量符合基本學測目的；試題取材均勻分佈（到國三下第二次段考前的範圍），生活化、統整化素材優先；試題難度以50-75%通過率為主。上述組題原則的目的是為了保證所組出來的測驗能吻合基本學測的原有目的。這也等同於是說，基本學測有一定的設計規格，基本上來說，基本學測是朝穩定性的方向來發展。傳統聯考中易出現「去年簡單今年會難」、「去年難今年會簡單」的現象，在基本學測中這樣的情況會比較少出現，但預測難度及真實難度若有差別的話則是在所難免。

至於量尺分數則是基本學測分數使用上的考量。測驗分數的使用有兩大重要考量；第一是，測驗相對機制的建立（專業術語稱建立常模），它是希望能建立一個類似於「IQ100是中等」，「IQ100以上是中等資質以上」的具體意義。這套作法，在理論上還可以將量尺分數作跨年的比較，然而由於教育部所公佈的命題範圍會因不同學年而異等因素，以測驗的角度來看量尺分數作跨年的比較，暫不可完全實行，但有其參考的價值。第二是測驗分數的採計與運算。基本學測是以答對題數作為原始分數。原始分數在測量性質上是屬於名次性質，它會有在不同區段中名次間的差距看似等距離，其實不是等距離（如第1與第2名；第15與16名，間距均為1，但實質並不是相等的1），因此將其轉換成一

個可以採計與運算的量尺分數（基本學測量尺分數是1～60）是一個測驗標準化方法中的重要步驟。

最後基本學測是一年多次的（目前是一年兩次），國人也不易接受。來自各方的質疑有兩方面，第一、「學生會因考到簡單題而佔便宜，也會因考到難題而吃虧」。其實基本學測採用量尺分數的目的之一，也是因應此一問題。不過處理這問題的基本作法有二：第一、不同次的基本學測內容不會一樣，但是組題分佈、難易程度均要控制成一樣（再一次保證了基本學測穩定性的必要）；第二、如果在控制上出現小瑕疵，則另一個專業機制：等化，便可以派上用場。簡單的說測驗等化設計可以讓學生在考到稍難試題時，只要答對較少題時，就可以得到考簡單題時要答對較多題時的量尺分數，舉例來說：第一次數學科測驗較簡單，第二次數學科測驗較難，那是有可能透過等化，第一次對10題與第二次對9題的量尺分數是一樣的。另外一個質疑則是兩次基本學測考生不同（有第二次考生資質較優，亦有較低之說），所建立量尺分數在相對機制上有公平性的問題。確實，基本學測如果每次測驗均獨立建造常模，那會是問題。但基本學測其實不是這樣做。以90年為例：只有第一次基本學力測驗真正建立量尺分數對照表，第二次其實是以第二次答對題數等化到第一次答對題數，再由第一次答對題數對照量尺分數。

基本學力測驗的分數等化過程(涂柏原，民92)：

（1）由於幾乎所有的國三學生都參加了第一次國中基本學力測驗，第一次測驗的結果非常適合用來建立量尺分數。因此，根據前面所提到的原始分數轉換成量尺分數的方法，我們先將原始分數與量尺分數對照表算出來。如此，我們可以得到每一位考生的原始分數與量尺分數等資料。因為要用 IRT 的方法來進行等化，所以我們得進一步用 IRT 的方法來估算每一位考生的能力參數，根據全部考生的資料，經過綜合整理之後，我們可以得到下表的第二、三及四欄的資料，也就是答對題數、能力值與量尺分數的對照表。

（2）根據參加第二次學力測驗的考生在第二次測驗上面的作答資料，用 IRT 的方法來估計考生的能力值。如此，每一位考生的答對題數原始分數與其能力估計值都可以被收集到，而將這些資料與從第一個步驟所得到的資料合併，我們可以得到如下的對照表範例。

（3）第二次國中基本學力測驗不再建立新的量尺分數，而是使用我們在上一個步驟所得到的表，來將考生在第二次學力測驗所得到的原始分數，對應到在第一次學力測驗時所建立的量尺分數。舉例而言，如果某一考生在第二次基本學力測驗某學科的原始分數為 58 分，從下表我們可以得知他所對應的量尺分數為 51，相當於在第一次基本學力測驗該科的原始分數為 62。換言之，對該科來說，在第一次基本學力測驗原始分數為 62 分的考生，與在第二次基本學力測驗原始分數得到 58 分的考生，有相同的量尺分數 51 分。

某科量尺分數對照表範例：

第二次答對題數	能力值	第一次答對題數	量尺分數
63⟶	4.93	66	60
	4.4		59
62⟶	4.1		58
	3.71	65	57
61⟶	3.35		56
	2.99	64	55
60⟶	2.77		54
	2.56	63	53
59⟶	2.4		52
58⟶	2.25	62	51
:	:	:	:
:	:	:	:

IRT 最為有力的地方，是同一個考生若在同一個時段考兩次試，即使作答不同的試題，所估計出來的兩個能力參數的值，理論上是一樣大小的。如果考完第一次之後隔了一些時日再考第二次，以至於兩次考試之間有成長（或進步）的現象存在，那麼根據第二次測驗的結果所估計得到的能力參數值，將會大於

第一次測驗後所估計得到的。因為 IRT 具備這樣的特性，因此透過用 IRT 方法所進行的等化之後，我們有信心考生在第二次測驗所得到的量尺分數絕對是合理、公平、公正的；而且從兩次測驗所得到的量尺分數之間也是可以相互比較的。

影響整個基本學力測驗等化程序的關鍵點，其實是在於各個題目的試題參數（item parameter）是否被精確地估計出來。因為這裡所說明的等化方法在某個程度之下，其原理與目前常見的電腦化適性測驗（computerized adaptive testing, CAT，如 TOEFL、GRE 以及美國的護士執照考試等）原理是一樣的，是在試題參數已知的情形下估計考生的能力參數（ability parameter）（註：在 IRT 中，每一個試題會有幾個試題參數來描述該題的特性，而也有一個能力參數來描述考生的能力）。目前國中基本學力測驗每一道試題皆經過至少 240 至 320 位不同地區的國三學生「預試」過，截至目前為止，參與過預試的國三學生已超過數萬人，涵蓋全國各縣市的國民中學。以此大規模的預試工作所得的答題反應資料，可以用 IRT 來估計出每一個題目的試題參數（難度、鑑別度）等。經過實際驗證，九十年預試所得的題目難度與該年度第一次測驗三十萬人的資料，算出來的題目難度相當接近。即使如此，學力小組仍然依據從三十萬人的資料所得到的試題參數，來將題庫中每一個題目的試題參數加以校正。

第三章　研究方法

第一節　研究對象

本研究為探討民國九十年國中基本學力測驗數學科兩次測驗間等化問題，研究之對象包含當年度第一次有效應考人數 299368 人，第二次有效應考人數 167440 人，經向國民中學學生基本學力測驗工作推展委員會申請資料，得到第一次考生的答對狀況資料計 5000 人，第二次的資料 5000 人，這些資料是國中基本學力測驗委員會利用隨機抽樣自所有應答的考生答題狀況中抽出，資料中包含學生的性別、區域及答對題目的狀況(1 表示答對該題，0 表示答錯)，另外國中基本學力測驗委員會也提供了兩次考試的有關統計資料，如第一次測驗 alpha 信度為 0.894，原始分數的總平均 18.45，原始分數的標準差 7.26，量尺分數的總平均 30，量尺分數的標準差 12.3，以及利用 Rasch 模式所估計出的測驗難度值；第二次測驗原始分數的總平均 18.67，原始分數的標準差 7.40，量尺分數的總平均 32.54，量尺分數的標準差 13.12，以及利用 Rasch 模式所估計出的測驗難度值。

第二節　研究工具

本研究所使用的工具依其性質可分為：一、民國九十年國中基本學力測驗第一次和第二次數學科試題，二、電腦軟體，茲分述如下：

一、民國九十年國中基本學力測驗第一次和第二次數學科試題

「基本學力測驗」數學科的題目有下列幾個特色：(國民中學學生基本學力測驗推動工作委員會，民89)

1、 是基本的、核心的、重要的：基本並不代表簡單，而是學生在國中階段的數學課程中應該學會的重要觀念，命題小組把國中教材中的重要概念一一收羅在題目之中，期望學生在國中的數學課程裡，了解基本的、核心的數學理念，進而培養對數學的喜好與懂得如何去思考。

2、 跳脫過去參考書題目的窠臼：有許多參考書的題目艱深難懂，且須應用一些高中階段的觀念來解題，否則就會很難求得正確答案。且不少國中學生為了怕考不好，做了太多同樣又重複類型的題目，形成過度學習而不自知，甚至有些題目只要一看敘述就知道答案，根本不必思考或計算。這些情況造成學習的盲點，基本學力測驗的命題將改變這些弊端，跳脫傳統參考書題目的窠臼，幫助學生正常學習數學。

3、 題目語句敘述完整：基本學力測驗每一題題目的敘述都很完整，明確地讓學生知道問題是什麼。命題小組在設計題目時，會把各種可能的情形都考慮進去，設計的題目避免學生因城鄉差距而造成作答困難，同時使學生不會因生活型態或生活環境背景的不同而誤解題目的意義。

4、 作答不強調特殊解法、不需死背公式：許多學生認為學數學只要背公式，然後把數字代進公式就可以解題了；有的人碰到難題時，不運用最基本的觀念去思考，而只希望能學會特殊解法，利用很快速的步驟來解題，至於「為什麼這麼做」，根本不在他們的思考範圍內。命題小組完全避除這一類型的題目，基本學力測驗的題目在作答時，不強調特殊解法，也不需死背公式，只要具有基本的、正確的數學能力自然能夠解題。不過，並不代表基本學力測驗沒有難題，只是這些難題都是由命題小組精心思考、設計的，傳統的參考書或測驗卷曾出現過。

5、 不超出課程範圍：在一些參考書或講義的題目中，偶而會利用一些課外公式解題，如：利用行列式求面積等。這些公式用起來也許很方便，但若要深究運用的道理與原因，學生卻完全不知道，若數學課程只是在學習這些課外公式，那根本不用數學老師的教導，只要有一本參考書或一大堆公式即可。因此為避免加重學生學習的負擔，基本學力測驗命題絕不超出課程範圍，學生只要具備國中生應有的基本能力，參加基本學力測驗時，就能從容應考。

6、 題庫的每一道題目皆經過測試：為了要達到兩次「基本學力測驗」的客觀性與公平性，每一道題目都經過測試，再由命題小組依據難易度與章節層次分別納入題庫系統，在舉行基本學力測驗時，再以電腦依各項比例選題，配置成一張試題。

第一次測驗共有 32 題，第二次測驗共有 31 題，皆為四選一之選擇題，第一次測驗日期為民國九十年三月三十一日及四月一日，第二次測驗日期為民國

九十年六月九日及十日，考試時間共 60 分鐘，兩次考試間隔約 70 天。

二、電腦軟體

本研究採用的電腦軟體有三種；BILOG-MG, MATLAB, 和 SPSS，茲分述如下：

BILOG-MG 是適用於二元計分(對與錯)試題 logistic 模式之試題參數及考生能力之估計的套裝軟體。由美國 Scientific Software, Inc 發行，能處理單參數、二參數及三參數模式的資料。BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). 使用的統計法有最大相似(Maximum Likelihood)法、後面期望的貝氏法(Expected A Posteriori)及後面最大的貝氏法(Maximum A Posteriori)可供選擇，在此研究我選用後面期望的貝氏法，因為此法可得到精確及穩定的估計(Embretson, & Reise, 2000)。

MATLAB 是由 Math Works 公司於 1984 年推出的數學軟體，此研究利用 MATLAB 來截取 BILOG-MG 所輸出的試題參數及考生能力之估計。

SPSS 是社會統計軟體程式(Statistical Package for the Social Sciences, SPSS)，是 Nie, Hull, Jenkins, Steinbrener 和 Bent 為資料處理而發展，由美國 SPSS, Inc 發行，此研究利用 SPSS 來處理真分數的計算。

第三節　實施程序及資料處理

本節茲分為資料蒐集及資料處理等二部份，說明如下：

一、蒐集資料

本研究為探討民國九十年兩次國中學力測驗數學科分數以量尺分數呈現後，與國中學力測驗中心所公布的量尺分數間差異的探討。國中學力測驗於民國九十年首次舉行，一年辦兩次，兩次間隔時間為 70 天，第一次考生 299368 人，第二次 176416 人，此研究所需之資料形態為學生的作答資料，即答對題目的情況顯示，所有作答資料以 0、1 二元計分表之，1 代表答對該題，相反地若是該題答錯則以 0 表示，因國民中學學生學力測驗委員會有提供學生作答資料申請，於是向委員會申請兩次測驗數學科作答資料，各得 5000 份資料，經整理後，第一次實際有效用來分析的資料有 4999 人，第二次資料有 4737 人，資料尚含學生的性別及應考的區域，在此研究中因不具意義，所以予以刪除不用。

二、資料處理

取得資料後，逐步進行資料處理工作，首先完成 BILOG-MG 的統計分析，估計得題目參數及學生能力估計，然後利用 MATLAB 將題目參數及學生能力從 BILOG-MG 結果中截取並存成文字檔，最後利用 SPSS 算出能力估計值從-2.5 至 2.5 之間每隔 0.01 的所有真分數。

第四章　研究結果與討論

本章分為三節，第一節民國九十年第一次國中基本學力測驗數學科校準後所呈現量尺分數之研究結果，第二節民國九十年第二次國中基本學力測驗數學科校準後所呈現量尺分數之研究結果，第三節綜合討論。

第一節　民國九十年第一次國中基本學力測驗數學科校準後所呈現量尺分數之研究結果

為了解所有考生與本研究所獲得資料間的差異，茲將敘述性的統計列於表一，民國九十年第一次國中基本學力測驗數學科所參加的總人數為 299368 人，若以每題答對得一分來計算，平均分數為 18.45，分數的標準差為 7.26，此次考試的信度 alpha 為 0.89，而此研究所獲得的樣本數為 4999 人，平均分數為 18.62，分數的標準差為 6.74，信度 alpha 為 0.88。因此樣本平均數略高於母體 0.17，而標準差的差異則較母體少了 0.52，因此樣本分數的分布將較母體的分數集中。

表一　第一次國中學測母體與樣本資料統計表

	母體	樣本
總人數	299368	4999
原始平均分數	18.45	18.62
原始分數的 SD	7.26	6.74
信度 alpha	0.89	0.88

第一次國中學測所取得的樣本資料經過 BILOG-MG 以 1PL 及 3PL 模式估計後，和學測所題供之 b 值呈現於表二中，本研究經 1PL 模式校準後估計所得 b 值與學測所提供之 Rasch 模式 b 值頗為相近，平均難度相差 0.01，其中以第 26 題差異最大為 0.14，第 8、10、25 和 29 幾乎相等，標準差相差為 0.02。

表二　第一次國中學測 Rasch、1PL 及 3PL 模式之試題參數估計值

	學測估計	樣本估計			
題號	Rasch_b	1PL_b	3PL_a	3PL_b	3PL_c
1	-1.52	-1.45	0.85	-0.55	0.39
2	-2.04	-1.99	1.46	-1.15	0.20
3	-1.51	-1.50	1.09	-0.82	0.23
4	-2.06	-2.02	0.83	-1.49	0.20
5	-1.62	-1.57	0.81	-0.71	0.38
6	-1.70	-1.75	0.86	-1.26	0.17
7	-1.00	-0.97	1.63	-0.18	0.30
8	-1.55	-1.55	1.85	-0.81	0.17
9	-1.46	-1.38	1.00	-0.79	0.21
10	-0.35	-0.35	1.18	0.32	0.28
11	-0.49	-0.45	1.09	0.04	0.20
12	-0.39	-0.43	1.73	-0.10	0.11
13	-0.80	-0.85	1.45	-0.23	0.23
14	-0.08	-0.07	1.06	0.63	0.27
15	-1.11	-1.14	0.83	-0.59	0.23
16	0.03	0.06	0.77	1.23	0.33
17	-0.51	-0.55	1.59	0.09	0.26
18	0.14	0.09	1.56	0.29	0.11
19	-0.43	-0.42	0.49	-0.10	0.14
20	0.32	0.29	1.22	0.65	0.18
21	0.19	0.14	1.12	0.69	0.23
22	0.13	0.09	1.23	0.51	0.19
23	-0.19	-0.12	1.26	0.46	0.25

24	0.43	0.40	0.85	1.47	0.29
25	-0.41	-0.41	1.85	0.10	0.21
26	0.06	0.20	1.18	0.96	0.28
27	0.29	0.28	0.90	0.56	0.13
28	0.17	0.25	1.17	1.14	0.31
29	0.70	0.70	0.94	1.07	0.15
30	0.74	0.70	1.01	1.20	0.19
31	1.12	1.09	1.50	1.73	0.22
32	0.62	0.57	1.06	1.52	0.28
平均	-0.45	-0.44	1.17	0.18	0.23
SD	0.88	0.86	0.34	0.87	0.07

樣本資料經過 1PL 模式校準後，所得答對題數之能力估計如表三，1PL 所估得之能力範圍為-3.05 至 2.41 之間，其所採用的量尺平均為 0、標準差為 1，轉換為平均 30、滿分為 60 後的量尺分數呈現於表三的第 4 欄，結果發現，答對題數在 19 題以下，除了答對 13 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，答對題數高於 20 題的情形者剛好相反，因此量尺分數若是以此種情形呈現，將有利於前半段能力較好的考生，若以變化率而言，此種估計與學測只有 36.36%(12/33)相同。

若以 1PL 模式校準後，再將估得之能力配合題目難度估出真分數，所呈現之真分數分佈如表三之第 5 欄，真分數範圍為 0.98~31.46。此與原始答對題數範圍(0~32)不同，這是因為除非考生的能力值為-∞或∞才有可能真分數達到 0 或 32。

因為學測所有的題目皆是四選一的選擇題，以試題反應理論而言，若鑑別度 a 無法控制在一小範圍內(0.8~1.25)(王寶塘，民 85)，且猜測度 c 又無法避

免，則最好的模式應是 3PL 模式(Kolen, & Brennan, 1995)。因此，本研究亦採用 3PL 模式來校準估計題目及學生能力參數，所得之結果如表二第 4, 5, 6 欄與表三第 7 欄，由表二之估計結果可以看出，事實上鑑別度及猜測度是存在的。鑑別度從 0.49~1.85，難度範圍為-1.49~1.73，猜測度為 0.11~0.39，其平均分別為 1.17, 0.18, 0.23。能力估計值的範圍則是-2.28~2.38。以 3PL 模式校準估計，對於答對題數相同，但其答對題目不同者，會產生能力估計不同的情形發生，甚至於有可能會有答對題數較多反而能力估計值較低的情況出現，如本研究答對 3 題之能力-2.03 大於答對 4 題之能力-2.04 的情況。為了以答對題數的方式來呈現能力估計值，因此，研究者將答對相同題數的能力作加權平均後呈現出。最後再藉由此能力分布轉換成量尺分數如表三第 8 欄，在答對 17 題之前，大都低於或等於學測之量尺分數，只有答對 2 或 3 題例外，在答對 17 題之後則相反，若以變化率而言，此種估計與學測只有 30.3%(10/33)相同。若再配合估得之題目參數，可計算得配合答對題數的 3PL 真分數分佈，如表三第 9 欄。我們發現其範圍為 8.19~30.87，與 0~32 有很大的差異，這是因為我們採用 3PL 模式，猜測度存在的原因，即使學生能力再低，他都有猜對該題的機率。

表三　第一次學測答對題數與量尺分數、1PL 與 3PL 模式之能力、真分數對照表

答對題數	學測	1PL 能力估計	1PL 量尺	1PL 真分數	3PL 能力估計	3PL 量尺	3PL 真分數
0	1	-3.05	1	0.98	-2.28	1	8.19
1	1	-2.75	1	1.37	-2.28	1	8.19
2	1	-2.49	1	2.15	-2.07	4	8.47
3	3	-2.30	1	2.78	-2.03	4	8.53

4	6	-2.14	3	3.42	-2.04	4	8.51
5	8	-1.94	6	4.36	-1.91	6	8.73
6	10	-1.71	9	5.63	-1.76	8	9.03
7	11	-1.52	11	6.83	-1.63	9	9.33
8	13	-1.41	12	7.58	-1.54	11	9.57
9	15	-1.33	13	8.15	-1.42	12	9.94
10	17	-1.23	15	8.89	-1.24	14	10.57
11	18	-1.08	17	10.05	-1.07	17	11.28
12	20	-0.87	19	11.79	-0.87	19	12.25
13	21	-0.67	22	13.54	-0.69	21	13.25
14	23	-0.54	23	14.73	-0.57	23	13.99
15	24	-0.47	24	15.37	-0.49	24	14.51
16	26	-0.42	25	15.84	-0.40	25	15.13
17	27	-0.34	26	16.59	-0.30	26	15.85
18	29	-0.21	27	17.82	-0.09	29	17.47
19	30	0.00	30	19.8	0.12	32	19.19
20	32	0.20	33	21.64	0.30	34	20.66
21	33	0.34	34	22.87	0.40	35	21.47
22	35	0.43	35	23.63	0.44	36	21.78
23	36	0.50	36	24.19	0.48	36	22.09
24	38	0.61	38	25.04	0.57	37	22.78
25	40	0.79	40	26.31	0.75	39	24.07
26	42	1.02	43	27.7	1.02	43	25.8
27	44	1.22	45	28.69	1.15	45	26.55
28	46	1.38	47	29.34	1.33	47	27.49
29	48	1.55	49	29.92	1.44	48	28.01
30	51	1.79	52	30.55	1.67	51	28.96
31	54	2.09	56	31.09	2.02	55	30.09
32	60	2.41	60	31.46	2.38	60	30.87

根據學測提供之資料，若作答資料以 Rasch 模式估計，以 0~60 量尺分數呈

現，則平均量尺分數為 30，量尺分數的標準差為 12.3；以 1PL 模式估計，則平均量尺分數為 30，量尺分數的標準差為 11.81; 以 3PL 模式估計，則平均量尺分數為 30，量尺分數的標準差為 11.91。，統計結果如表四。

表四　第一次國中學測母體與樣本量尺分數統計表

	量尺平均分數	量尺分數的 SD
母體	30	12.3
樣本_1PL	30	11.81
樣本_3PL	30	11.91

第二節　民國九十年第二次國中基本學力測驗數學科校準後所呈現量尺分數之研究結果

為了解所有考生與本研究所獲得資料間的差異，茲將敘述性的統計列於表五，民國九十年第二次國中基本學力測驗數學科所參加的總人數為 167440 人，若以每題答對得一分來計算，平均分數為 18.67，分數的標準差為 7.40，此次考試的信度學測中心並無提供。而此研究所獲得第二次的樣本數為 4737 人，平均分數為 18.76，分數的標準差為 7.35，與母體的差異均很小，分別只有 0.09 與 0.05，信度 alpha 為 0.90。

若作答資料改以 Rasch 模式估計，以量尺分數呈現，則平均量尺分數為 30，

量尺分數的標準差為 12.3，

表五　第二次國中學測母體與樣本資料統計表

	母體	樣本
總人數	167440	4737
原始平均分數	18.67	18.76
原始分數的 SD	7.4	7.35
信度 alpha	無	0.9

第二次國中學測所取得的樣本資料經過 BILOG-MG 以 1PL 及 3PL 模式估計後，和學測所題供之 Rasch 模式 b 值呈現於表六中，本研究經 1PL 模式校準後估計所得 b 值與學測所提供之 b 值差異大於第一次學測，平均難度相差達 0.16，其中以第 16 題差異最大為 1.00，第 23 題最小差異為 0.01，標準差相差也較第一次學測大，相差達 0.37。

表六　第二次國中學測 Rasch、1PL 及 3PL 模式之試題參數估計值

	學測估計	樣本估計			
題號	Rasch_b	1PL_b	3PL_a	3PL_b	3PL_c
1	-1.61	-1.24	1.62	-0.67	0.20
2	-0.35	-0.50	1.18	0.23	0.31
3	0.13	-0.29	0.91	0.29	0.24
4	-1.80	-1.03	1.40	-0.51	0.20
5	-1.01	-0.92	1.39	-0.47	0.17
6	-1.27	-0.98	2.05	-0.33	0.26
7	-0.15	-0.55	1.48	0.12	0.29
8	0.04	-0.24	0.91	0.02	0.10

9	-2.15	**-1.61**	1.25	-0.73	0.42
10	-0.65	-0.54	0.92	0.15	0.29
11	-0.76	**-1.16**	1.87	-0.13	0.45
12	-0.87	**-0.82**	0.84	-0.46	0.18
13	0.66	**0.15**	1.60	0.52	0.18
14	0.07	**-0.50**	1.71	-0.14	0.13
15	-1.57	**-1.33**	1.45	-0.46	0.39
16	-2.82	-1.82	1.17	-1.22	0.26
17	-0.34	-0.53	1.15	-0.10	0.18
18	-0.80	-0.70	1.44	-0.23	0.19
19	0.14	-0.17	1.56	0.06	0.08
20	-1.24	-1.47	1.10	-0.85	0.29
21	0.03	-0.30	1.59	0.29	0.26
22	0.71	0.12	2.15	0.68	0.25
23	0.35	0.36	2.10	0.60	0.15
24	-0.39	-0.44	1.58	0.05	0.21
25	0.41	-0.14	1.89	0.86	0.38
26	0.42	0.04	1.43	0.63	0.25
27	1.02	0.38	2.06	1.02	0.27
28	0.76	0.19	1.24	1.46	0.37
29	1.04	0.37	1.64	0.68	0.16
30	0.67	0.16	1.18	1.20	0.33
31	0.64	0.10	0.94	0.78	0.24
平均	-0.34	-0.50	1.44	0.11	0.25
SD	0.99	0.62	0.38	0.64	0.09

樣本資料經過 1PL 模式校準後，所得答對題數之能力估計如表七，1PL 所估得之能力範圍為-2.82 至 2.10 之間，為了配合第一次學測的能力範圍，因此將其拓寬使其相符後，再轉為滿分為 60 的量尺分數，結果呈現於表七的第 4 欄，結果發現除了答對 6 與 30 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，因此量尺分數若是以此種情形呈現，將使得大部份的學生的量尺

分數低於學測的量尺分數，若以變化率而言，此種估計與學測只有 28.13%(9/32) 相同。答對 2 題空白，乃所得資料並無恰好答對 2 題者。另以 3PL 模式估計，結果答對 1 者，其能力竟高於 3 題與 4 題者，此乃樣本資料答對 1 題的只有 1 人，而答對 3 與 4 題者其平均反而小的緣故。

若以 1PL 模式校準後，再將估得之能力配合題目難度估出真分數，所呈現之真分數分布如表七之第 5 欄，真分數範圍為 0.97~29.47。

同樣地，本研究亦採用 3PL 模式來校準估計題目及學生能力參數，所得之結果如表六第 4,5,6 欄與表七第 6 欄，由表六之估計結果可以看出，事實上鑑別度及猜測度亦是存在的。鑑別度從 0.84~2.15，難度範圍為-1.22~1.46，猜測度為 0.08~0.45，其平均分別為 1.44,0.11,0.25。能力估計值的範圍則是 -2.01~2.00。若先轉換成與第一次學測所估的能力範圍，再轉成 0~60 的量尺，結果呈現於表七的第 7 欄，結果發現除了答對 1、3、19、20 及 21 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，因此量尺分數若是以此種情形呈現，將使得大部份的學生的量尺分數低於學測的量尺分數，若以變化率而言，此種估計與學測只有 15.63%(5/32)相同。若再配合估得之題目參數，可計算得配合答對題數的 3PL 真分數分布，如表七第 8 欄，真分數範圍為 8.22~30.3。

表七　第二次學測答對題數與量尺分數、1PL 與 3PL 模式之能力、真分數對照表

答對題數	學測	1PL 能力估計	1PL 量尺	1PL 真分數	3PL 能力估計	3PL 量尺	3PL 真分數
0	1	-2.82	1	0.97	-2.01	1	8.22
1	1	-2.50	1	1.58	-1.74	5	8.55
2	1						
3	4	-2.13	2	2.7	-1.85	5	8.4
4	6	-1.92	4	3.6	-1.76	5	8.52
5	8	-1.66	8	5	-1.67	6	8.67
6	10	-1.47	11	6.25	-1.60	7	8.8
7	12	-1.38	11	6.91	-1.52	8	8.96
8	14	-1.34	12	7.21	-1.46	9	9.1
9	16	-1.29	13	7.6	-1.32	11	9.47
10	18	-1.19	15	8.42	-1.16	14	10.01
11	20	-0.99	17	10.19	-0.98	16	10.77
12	21	-0.74	21	12.61	-0.85	18	11.45
13	23	-0.57	23	14.33	-0.62	22	12.91
14	25	-0.49	24	15.15	-0.52	23	13.65
15	26	-0.46	25	15.46	-0.47	24	14.05
16	28	-0.44	25	15.67	-0.43	25	14.37
17	29	-0.41	25	15.97	-0.30	26	15.48
18	31	-0.33	26	16.8	-0.13	29	17.03
19	32	-0.15	29	18.61	0.12	33	19.39
20	34	0.10	32	21	0.31	35	21.19
21	36	0.30	35	22.72	0.40	37	22.03
22	37	0.40	37	23.5	0.44	37	22.4
23	39	0.44	37	23.8	0.45	37	22.49
24	41	0.49	38	24.16	0.46	38	22.58
25	43	0.59	39	24.84	0.52	38	23.12

26	44	0.80	42	26.08	0.72	41	24.86
27	47	1.06	46	27.28	1.06	46	27.31
28	49	1.27	49	28.01	1.28	49	28.46
29	51	1.46	51	28.51	1.37	51	28.83
30	54	1.72	55	29.02	1.54	53	29.39
31	60	2.10	60	29.47	2.00	60	30.3

表八　第二次國中學測母體與樣本量尺分數統計表

	量尺平均分數	量尺分數的 SD
母體	32.54	13.12
樣本_1PL	31.03	13.42
樣本_3PL	30.65	14.01

第三節　綜合討論

本研究的量尺分數計算，是基於所有題目皆來自題庫，且題庫中的題目參數都具有相同的量尺的假設，即所有試題的參數皆已知並具相同的量尺，在此假設下，其實兩個測驗間即不需做等化，因將此兩測驗施測於學生，獲得作答資料後，即可利用軟體求得學生的能力估計值，一般的測驗軟體，都會將考生的能力估計設定於平均數為 0、標準差為 1 的分佈。所以，所估得之能力量尺再

經過平均數 30、全部答對題數設為 60 的直線轉換，即可轉換為學測的量尺分數。但是在本研究因母群體之資料無法獲得，且研究者認為 3-PL 模式較符合四選一之選擇題，因此，擬用所獲得之樣本重新校準題目之參數，卻發生因為兩次考試題數不同，透過 BILOG-MG 估計後所得到之能力值範圍亦不同的情形，用 1PL 模式的估計，第一次學測的範圍為-3.05~2.41，第二學測的範圍則縮小為-2.82~2.10 之間；利用 3PL 模式的估計，第一次學測的範圍為-2.28~2.38，第二學測的範圍則縮小為-2.01~2.00 之間。為了等化量尺，因此將第二次學測的能力範圍皆調成與第一次的相等後，並將滿分調為 60 後呈現出。

原擬用的真分數等化法，亦因兩次考試不同，而無法做等化，由表三與表七的結果顯示，無論 1PL 模式或 3PL 模式呈現之量尺分數都與學測中心有不小的差異，這是因為項目與學生能力參數校準估計不同所使然，歸納其原因如下：

1. 項目與學生能力參數的校準估計與資料的內容有很大相關，以第一次學測的樣本而言，其樣本數不到母體的 2%，第二次學測的樣本數也不及 3 %，樣本在這麼小的比例下，產生誤差的機率自然大增。

2. 項目與學生能力參數的校準估計和所使用的模式有關，本研究所採用之 1PL 模式或 3PL 模式皆與學測中心所採用的 Rasch 模式不同，因此，估計出來的項目參數與能力值也不相同，量尺分數自然有差異。

3. 項目與學生能力參數的校準估計和所使用的測驗軟體有關，本研究所採用的測驗軟體為 BILOG-MG，統計估計是用 EAP 法，若不同的軟體或統計估計法所

求出的參數，亦會不同。

由以上的經驗，我們得知作答資料的大小會影響題目參數的估計，進而影響測驗的等化，雖然學測中心報告中顯示，題庫中每道題都經過 240~300 位學生預試估計而得，但比起真正考試的母群體約 30 萬人，還是微不足道。因此，真正所得資料校準後估計的題目參數與原先題庫的參數值間的差異值得研究，若兩次考試中有任何一次產生顯著差異，則等化的問題即產生，在此狀況下，若兩次考試間又無定錨問題，或兩次都參加考試的人又無法篩選出，而另外蒐集做估算，則等化就值得商確。為了避免此問題發生，而此考試牽涉到事後試題的公開，因此定錨問題似乎不可行。考試的學生有一現象，即幾乎所有參加第二次學測的學生都有參加第一次學測，若能請學生於報考第二次時加註第一次學測的編號，以利追蹤比對兩次的結果。但若不再重新校準估計題目參數，直接估計能力值也是一種方法。

另外由學測所給的統計資料中顯示第一次測驗難度平均為-0.45，而第二次測驗難度平均為-0.34，因此第二次的題目較第一次難些，但第二次雖只考 31 題，可是原始平均分數 18.67 確比第一次 18.45 高，量尺平均分數也呈現同樣情況，第一次量尺平均為 30，第二次為 32.54，其中原因是學生因間隔 70 天能力增加了，還是成績較不好的大都放棄第二次考試的機會，或是尚有其他因素，值得探討。

第五章　結論與建議

本章將本研究作一整體性的總結，並提出建議。本章共分為兩節，第一節結論，第二節建議。

第一節　結論

本研究主要探討民國九十年第一次學測與第二次學測，兩次學測間數學科量尺分數與等化，在 Rasch、1PL 和 3PL 等不同模式下，所估計的差異。具體的研究項目包括下列六項：一、探討民國 90 年第一次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。二、探討民國 90 年第二次學測，數學科利用 1-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。三、探討民國 90 年第一次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。四、探討民國 90 年第二次學測，數學科利用 3-PL 模式的校準後，所呈現的量尺分數與學力測驗的量尺分數改變情形。五、探討民國 90 年第一次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現真分數分佈情形。六、探討民國 90 年第二次學測，數學科利用 1-PL 與 3PL 模式校準後之題目難度估計值，應用 IRT 真分數法所呈現真分數分佈情形。

針對此等研究項目，本研究以國民中學學生基本學力測驗推動工作委員會所提供之資料，以 1PL 和 3PL 等不同模式進行分析，研究工具包括：一、民國

九十年國中基本學力測驗第一次和第二次數學科試題，二、電腦軟體：BILOG-MG、MATLAB 和 SPSS。經統計分析後，獲得下列發現：

1. 第一次學測樣本資料經過 1PL 模式校準後，結果發現答對題數在 19 題以下，除了答對 13 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，答對題數高於 20 題的情形者剛好相反，因此量尺分數若是以此種情形呈現，將有利於前半段能力較好的考生，若以變化率而言，此種估計與學測只有 36.36%(12/33)相同，真分數範圍為 0.98~31.46。

2. 第一次學測以 3PL 模式校準估計後，結果發現在答對 17 題之前，大都低於或等於學測之量尺分數，只有答對 2 或 3 題例外，在答對 17 題之後則相反，若以變化率而言，此種估計與學測只有 30.3%(10/33)相同，真分數範圍為.97~29.47。

3. 第二次學測樣本資料經過 1PL 模式校準後，結果發現除了答對 6 與 30 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，因此量尺分數若是以此種情形呈現，將使得大部份的學生的量尺分數低於學測的量尺分數，若以變化率而言，此種估計與學測只有 28.13%(9/32)相同，真分數範圍為.97~29.47。

4. 第二次學測以 3PL 模式校準估計後，結果發現除了答對 1、3、19、20 及 21 題高於學測所公布的量尺外，餘皆等於或小於學測所公布的量尺，因此量尺分數若是以此種情形呈現，將使得大部份的學生的量尺分數低於學測的量尺

分數，若以變化率而言，此種估計與學測只有 15.63%(5/32)相同，真分數範圍為.97~29.47。

5. 根據學測所提供之數學科統計資料，發現第二次考試平均難度高於第一次，但第二次學測的整體原始與量尺平均分數都高於第一次學測成績。

第二節　建議

根據以上的結論，提出下列建議：

一、測驗應用上的建議

1. 因為是從同一量尺的題庫中選取題目，建議除了符合雙向細目表的內容外，亦應事先考慮兩次考試測驗平均難度及總字數，希望能相近。

2. 兩次考試的時間既然相同，試題的題數也能相同，可避免兩次估出能力值範圍差距太大的問題，而且在國際上大型的考試，如托福、GRE 與 GMAT 的考試，相同類型的考試題數似乎也都一樣。

3. 試題的型式皆是選擇題，因此猜測因素是無法避免，而且所有試題要做到鑑別度相同，似乎也不容易，將來題庫的建立與資料的校準可以朝 3PL 模式的方向建立，較符合試題的型態，但相對的採用 3PL 模式所面對的問題亦有待克服。

4. 為了等化的精確性，建議可以在考生第二次報告時，請考生加註第一次考試

編號，以利追蹤比較兩次成績或作為等化之用。

二、未來研究的建議

1. 在可以取得有效且足夠大的樣本後，以 Rasch 與 3PL 模式校準估計後所產生量尺，對考生影響程度的比較。

2. 可以擴大到其它科目進行研究分析。

3. 兩次間隔時間長短對學生考試成績的影響。

參考資料

王寶墉(民 84)。*現代測驗理論*。台北：心理出版社。

余民寧(民 82)。試題反應理論的介紹(九)—測驗分數的等化(上)。*研習資訊，10(2)*，6-11。

林世華(民 91)。國中基本學力測驗與聯考實務作法的差異。*飛揚，14*。

涂柏原(民 92)。國中基本學力測驗量尺分數的說明（下）。*飛揚，19*。

教育部(2003)。*多元入學方案*。2003 年 11 月 23 日，取自 http://140.111.1.192/high-school/bbs/versatile-1.htm

國民中學學生基本學力測驗推動工作委員會(2003)。*九十三年國民中學學生基本學力測驗問與答*。2003 年 11 月 23 日，取自 http://www.bctest.ntnu.edu.tw/93bctest_q&a.htm

鄭同僚、張原禎(2001)。國民中學學生基本學力測驗—專訪臺灣師大心測中心林世華主任。*教育研究月刊，83*，4-16。

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) *Educational measurement* (2nd ed.) (pp.508-600). Washington, DC: American Council on Education.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36.

Brennan, R. L., & Kolen, M. J. (1989). Scaling the ACT assessment and P-ACT+: Rationale and goals. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 1-17). Iowa City, IA: ACT, Inc.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates.

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

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Wingersky, M. S. (1983). Comparison of IRT observed-score and

true-score "equating." *Research Bulletin 83-26*. Princeton, NJ: Educational Testing Service.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333-344.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintence for binary for binary items.* Mooresvilk IL: Scientic Software.

國家圖書館出版品預行編目資料

國中學力測驗兩次測驗量尺分數與等化之探討研究—以90年度數學科為例 / 曾建銘著.--
臺中縣豐原市：教育部中教研習會，民93
面；　公分
教育部臺灣省中等學校教師研習會九十三年度研究報告
參考書目：面
ISBN 957-01-9209-7(平裝)

1.數學 - 教學法 2. 中等教育 - 入學考試

524.32　　93022239

書　名：國中學力測驗兩次測驗量尺分數與等化之探討研究
　　—以90年度數學科為例
發行人：黃新發
編著者：曾建銘著
出版者：教育部臺灣省中等學校教師研習會
電　話：04-25227929
傳　真：04-25255440
地　址：台中縣豐原市師範街67號
網　址：本書同時登載於本會網站 http://www.isst.edu.tw
印刷所：穎弘文具印刷有限公司
地　址：台中市東區建成路670號
電　話：04-22818934
傳　真：04-22861605
出版年月：中華民國93年12月
定價：NT$ 180

GPN:1009304176

ISBN:957-01-9209-7