從TIMSS 2007臺灣八年級學生數學科作答反應檢視古典測驗理論和試題反應理論特性和測驗分析結果

蘇旭琳* 陳柏熹**

摘 要

本研究之主要目的在於透過TIMSS 2007臺灣八年級學生數學科的作答反應瞭解古典測驗理論和試題反應理論的特性，並且綜合兩種測驗理論取向檢視TIMSS 2007之數學科試題，希望透過試題分析結果提供試題編制和教育相關領域教學啟發之參考。資料庫人數共287人，以題本組合一為研究工具，利用軟體R 2.13.1、Excel 2003、SPSS 12.0與ConQuest 2.0等，執行各項數值運算和圖表繪製。結果發現，兩種測驗理論均顯示測驗試題呈現難度偏易情形、兩題誘答選項呈現中間組作答比率高於低分組的情形，試題特徵曲線顯示有四題多重計分題呈現階難度失序情形。整體試題品質相當良好，少數題目呈現無法區辨中間能力者樣貌，且試題對於學生較為容易。在理論特性方面，古典測驗理論可以從信度、構念效度、試題參數、項目分析和測驗參與者總分，對應試題反應理論的測驗訊息量、構念效度和模式適合度、試題參數、試題和類別特徵曲線和測驗參與者能力等概念，試題反應理論的優勢展現在測驗訊息量、潛在特質假設、參數不變性和試題等化上。在提供試題編制和教育相關領域啟發方面，研究者對兩題試題提出調整方式建議，並認為在某特定試題上，讓學生感到作答困難的原因，或許可以和數感教育有所連結。

關鍵詞：古典測驗理論、試題反應理論、PCM、TIMSS

DOI: 10.3966/199679772014123102003
責任編輯：楊叔卿
投稿日期：2013年12月31日，2014年3月25日修改完畢，2014年4月10日通過採用
*蘇旭琳，國立臺灣師範大學教育與心理輔導學系博士生，E-mail: aguri.su@gmail.com
**陳柏熹，國立臺灣師範大學教育與心理輔導學系副教授

壹、緒論

一、研究動機

測驗理論是一種解釋測驗資料間實證關係的有系統的理論學說，它的發展，迄今已邁入不同的新紀元，測驗理論學者通常把它劃分成二大學派：一為古典測驗理論（classical test theory, CTT）（Gullikson, 1987; Lord & Novick, 1968）——主要是以真實分數模式（true score model）為骨幹；另一為當代測驗理論（modern test theory）（Lord, 1980）——主要是以試題反應理論為架構（item response theory, IRT）（余民寧，1991）。為了克服古典測驗理論的限制，試題反應理論隨之誕生。前者之限制如：古典測驗理論計算出的難度、鑑別度、信度具有樣本依賴的特性；每位測驗參與者均具有相同的測量標準誤；得分相同受試者具有相同的能力分數等（Embretson & Reise, 2000）。希望深入瞭解古典測驗理論和試題反應理論的特性，即為研究者的動機源起。

二、研究目的

國際數學與科學教育成就趨勢調查（Trends in International Mathematics and Science Study, TIMSS）是由國際教育學習成就調查委員會（International Association for the Evaluation of Educational Achievement, IEA）所帶領進行的國際間教育成就比較性評估之研究，致力於提升世界各國數學和科學的教與學，實施時間為每四年一循環，參與對象為四年級和八年級之學生。為瞭解學生學科表現，研究報告內容提供學生在該科目之內容層面和認知層面的表現；報告中也囊括了學生的學習脈絡調查，例如：課程、學校、教師及準備、課室活動及其特徵、學生等相關變項的資料，俾便充分的了解教學和學習過程。在TIMSS 2007的資料庫中，除了包含各個參與國家的數學和科學的成就測驗表現，尚包含了學生問卷、教師問卷、學校問卷、課程問卷四項問卷調查資料。研究者認為透過實證資料是深入了解兩種測驗理論的途徑，因此藉由TIMSS資料庫之實際作答反應探討古典測驗理論和試題反應理論的特性，並進行測驗和試題分析。

綜合前文所述，研究者希望運用TIMSS 2007臺灣八年級學生數學科的作答反應瞭解古典測驗理論和試題反應理論的特性，並且綜合兩種測驗理論取向檢視TIMSS 2007之數學科試題，希望透過試題分析結果提供試題編制和教育相關領域教學啟發之參考。研究目的有三：

（一）以古典測驗理論和試題反應理論為基礎，對於TIMSS 2007數學科試題進行測驗和試題分析。

（二）藉由測驗和試題分析過程和結果，瞭解古典測驗理論和試題反應理論之特性。

（三）透過試題分析結果，提供試題編制和教育相關領域教學啟發之參考。

三、對應前述內容之研究問題為：

（一）以古典測驗理論和試題反應理論為基礎，TIMSS 2007數學科試題之測驗和試題分析結果為何？

1. 以古典測驗理論為基礎，其在內部一致性、構念效度、試題參數、項目分析、測驗參與者總分之各項分析結果為何？
2. 以試題反應理論為基礎，其在測驗訊息量、構念效度和模式適合度、試題參數、試題和類別特徵曲線、測驗參與者能力之各項分析結果為何？

（二）藉由測驗和試題分析過程和結果，古典測驗理論和試題反應理論展現並印證理論之特性為何？

（三）本研究之試題分析結果，提供試題編制和教育相關領域教學啟發之參考為何？

貳、文獻探討

此部分將依序描述「古典測驗理論和試題反應理論之特性」、「以古典測驗理論和試題反應理論進行測驗和試題分析之指標」，並歸納前兩部分提出「小結」。

一、古典測驗理論和試題反應理論之特性

在古典測驗理論中，假定個人在測驗上的實得分數是由兩部分所組成：一部分是真實分數，另一部分是誤差分數（郭生玉，1999），透過總變異量、共同因素的變異量、獨特變異量和誤差變異量等概念解釋測驗效能，賦予測驗實質意義；試題反應理論又被稱為潛在特質理論，是以模式為基礎的測量理論，特質的水準估計來自於測驗參與者的答題反應和試題特徵兩方面（Embretson & Reise, 2000）。試題反應理論主要是用來描述試題特性（難度、鑑別度、猜測度）與受測者的能力（潛在特質）如何影響其答題反應的一種數學模式（陳柏熹，2006）。Embretson和Reise（2000）提到古典測驗理論和試題反應理論的差異有十點，但也強調這十點是「極端狀況」下的差異，將其整理於表1：

表1 古典測驗理論和試題反應理論之差異一覽表

差異來源	古典測驗理論	試題反應理論
測量標準誤	針對特定群體下、所有分數共用相同的測量標準誤	可類推不同群體、不同分數具有不同的測量標準誤
測驗的長度和信度	較長測驗信度高於較短測驗	較短測驗信度可以高於較長測驗
可交換的測驗形式	當測驗間為平行題本時，比較測驗間參與者的測驗分數是最理想的	當題目難度是分散的，比較測驗間的參與者的分數是最理想的
試題屬性的不偏評量	試題屬性的不偏估計仰賴代表性的群體	試題屬性的不偏估計不需仰賴代表性的群體獲得
建立有意義的量尺分數	試題分數需要透過常模團體加以比較獲得意義	試題分數即使採用不同題目也可以相互比較且具有意義
建立量尺屬性	常態分數分配下可以達到等距量尺屬性	應用適當測量模式可以達到等距量尺屬性
混合試題形式	混合的試題形式對於測驗整體分數會產生不平衡的影響	混合的試題形式可以產生更能反映能力的測驗分數
改變分數的意義	起始分數水準不同，改變的分數不能有意義的比較	起始分數水準不同，改變的分數可以有意義的比較
二元計分試題的因素分析	二元計分的因素分析可能跑出非實際因素的成分出現	原始資料的因素分析可以產生充分的訊息
試題刺激特性的重要性	試題刺激相較心理測量屬性之下是不重要的	試題刺激可以和心理測量屬性直接對應

資料來源：整理自*Item response theory for psychologists,* by S. E. Embretson & S. Reise, Mahwah, NJ: Lawrence Erlbaum Associates.

表1中的「建立量尺屬性」、「改變分數的意義」、「試題刺激特性的重要性」，在試題反應理論中必須仰賴特定模式才能夠達到等距量尺和有意義比較改變分數（如：Rasch model）、試題刺激和心理屬性相互對應（如：Linear logistic test model, LLTM）的目標，而「試題屬性的不偏評量」和「混合試題形式更能反映能力」則是多數模式均具有的特性。除了表1臚列項目之外，試題反應理論是針對單一試題的作答反應所提出的數學模式，測量精確度的評估是以題目為單位來計算再加總起來，因此受試者的測量精確度（訊息量）是隨著受試者的能力以及所接受的題目特性而有所不同，且可以應用在：編製測驗（量表）、分數等化、編製題庫、電腦化適性測驗、組合測驗等層面；古典測驗理論則是針對測驗總分所提出的數學模式，模式簡單易理解，能力與試題參數容易計算，但是在應用層面主要以編製測驗（量表）為主，較為有限（陳柏熹，2006）。

綜上所述，試題反應理論相較於古典測驗理論可能具有的特性包含：能夠依據能力提供合理的測量標準誤、測驗的長度對於信度影響較小、參與不同測驗者能力可以相互比較、降低樣本依賴特性和減少依據常模解釋試題意義、具有等距量尺特性、試題形式對於能力估計影響較小、較能有意義的比較分數的改變、從原始資料中可以獲得較充分的試題訊息、能夠連結試題和心理測量屬性於同一量尺等特性，同時，透過試題和測驗訊息量的計算能夠經由數值估計測量精準度，在應用層面上也更為寬廣。由此可知，試題反應理論能夠清楚的呈現試題和測驗參與者之間的對應關係，使用時必須基於適合模式基礎和適當數量的測驗參與者；而古典測驗理論相對而言較不設限於模式假定，且具備簡單容易理解的特性。儘管古典測驗的應用層面相對受限，但學者們仍肯定古典測驗理論的有效性（例如：Embretson & Reise, 2000）。研究者可視自身研究目的，並在考量經費、人力、資源等狀況後，選擇合適的理論依據進行研究。

二、以古典測驗理論和試題反應理論進行測驗和試題分析之指標

若以古典測驗理論為基礎進行測驗和試題分析，通常會提到測驗的信度、效度和試題分析等指標。其中信度指標包含重測信度、複本信

度、內部一致性信度和評分者信度等；效度方面，American Educational Research Association、American Psychological Association與National Council on Measurement in Education [AERA, APA, NCME]（1999）認為效度是指證據或理論支持測驗分數的程度，可以從測驗內容、反應歷程、內在結構、與其他變項關係、測驗結果影響等層面收集證據作為效度資訊；在試題分析部分，郭生玉（1999）提到可以從難度、鑑別度、選項有效性分析來檢視試題。

試題反應理論在Lord於1980年代發表試題反應理論一書後正式誕生。儘管其理論取向和古典測驗理論有所區別，但仍有可相互指涉的概念，例如：試題反應理論中的試題訊息量的概念接近信度概念，模式適合度接近效度概念且傳統測驗的內容效度亦為受該理論重視，試題難度和鑑別度亦有其估計方式，運用能力估計值代表受測者能力，在項目分析方面可透過類別特徵曲線呈現各選項的能力和作答機率關係。前述概念分別敘述如下：

（一）測驗訊息量

表示試題在不同能力點上的測量精準度。訊息量愈高表示試題對該能力點的測量精準度愈高，從另一方面來解釋，訊息量也反映出試題在不同能力點的測量誤差（standard error, SE），訊息量愈高表示測量誤差愈小，此外，根據IRT的局部獨立性的假設，各題目所提供的訊息量彼此是沒有關聯的。因此可以將測驗中所有題目的訊息量加總得到測驗訊息量（陳柏熹，2006）。透過測驗訊息函數的計算和圖表繪製，可以做為描述試題或測驗、挑選測驗試題、以及比較測驗的相對效能的實用方法，作為建立、分析、與診斷測驗的主要參考依據（余民寧，1992b）。

（二）模式和模式適合度

在試題反應模式部分，若依其試題參數的數目多寡來命名，可分為單參數、二參數和三參數模式；若試題計分方式採非二元計分，亦可依據資料特性選擇適當模式，如：名義反應模式、等級反應模式、評定量表模式、部份給分模式等。

在模式適合度檢定部分，可以從試題反應理論的主要假定予以檢視，如單向度、局部獨立等（Embretson & Reise, 2000）；另外，適配性統計量可以描述資料符合模式期待的程度。Bond和Fox（2007）提到以Rasch模式分析為基礎的軟體通常會報告兩種適配性統計量：infit和outfit均方統計量，透過適配性統計量可以檢測模式和資料的適配程度。其中，outfit均方統計量是指未加權的標準化殘差的平方之平均值，當作答反應和預期反應相距越遠，該指標受到的影響越大；infit均方統計量則是透過變異量對於殘差進行加權，對於試題難度與測驗者能力相近的非預期作答反應較為敏感。前述兩種指標期望值皆為1，當數值偏離1時，代表觀察資料和模式所預期之作答反應差距越大，在一般二元計分的狀況下，可被接受的範圍在1±0.3之間。

（三）試題參數和能力參數

難度、鑑別度和猜測度均為試題參數，每位測驗參與者均有估計之能力值，在試題反應理論當中試題參數和能力參數的單位均為logit，由於題目參數的估計通常必須與受試者的能力一起估計，因此基於不同假設便須使用不同的參數估計方法，常見的方法有：聯合最大概似法（joint maximum likelihood, JML）、邊際最大概似法（marginal maximum likelihood, MML）、條件最大概似法（conditional maximum likelihood, CML）。關於這些估計方法的假定與介紹可以參閱相關書籍，如Embretson和Reise（2000）。

（四）試題特徵曲線和類別特徵曲線（category characteristic curves）

依據試題反應理論，可以將考生的表現情形與潛在特質間的關係，透過一條連續性遞增的函數來加以詮釋，這個函數便叫作試題特徵曲線。潛在特質的程度愈高（或愈強），其在某一試題上的正確反應機率便愈大。隨著選用的試題反應模式所具有的參數個數及其數值的不同，所畫出的試題特徵曲線形狀便不同（余民寧，1992a）。而類別特徵曲線則可描繪多重選擇題各選項的能力和機率曲線。

三、小結

綜合上述內容，研究者須考量TIMSS官方網站資料庫實際能夠蒐集到的資訊進行兩種理論基礎的特色說明並進行測驗和試題分析。因此，在古典測驗理論部分收集並呈現：內部一致性的信度資料、以內容效度為主的構念效度資料、試題參數（難度、鑑別度）、項目分析和測驗參與者總分等結果；在試題反應理論部分，收集並呈現：測驗訊息量、以內容效度為主的構念效度資料和模式適合度、試題參數（難度、鑑別度）、試題和類別特徵曲線以及測驗參與者能力等結果。

參、研究方法

此部分依序介紹研究架構、資料庫之研究參與者、研究工具，並說明研究程序和資料分析方式。

一、研究架構

本研究目的在於以古典測驗理論和試題反應理論為基礎，對於TIMSS 2007數學科試題進行測驗和試題分析，透過測驗和試題分析過程和結果，瞭解兩種理論之特性，並希望透過試題分析結果提供試題編制和教育相關領域教學啟發之參考。研究架構如圖1所示：

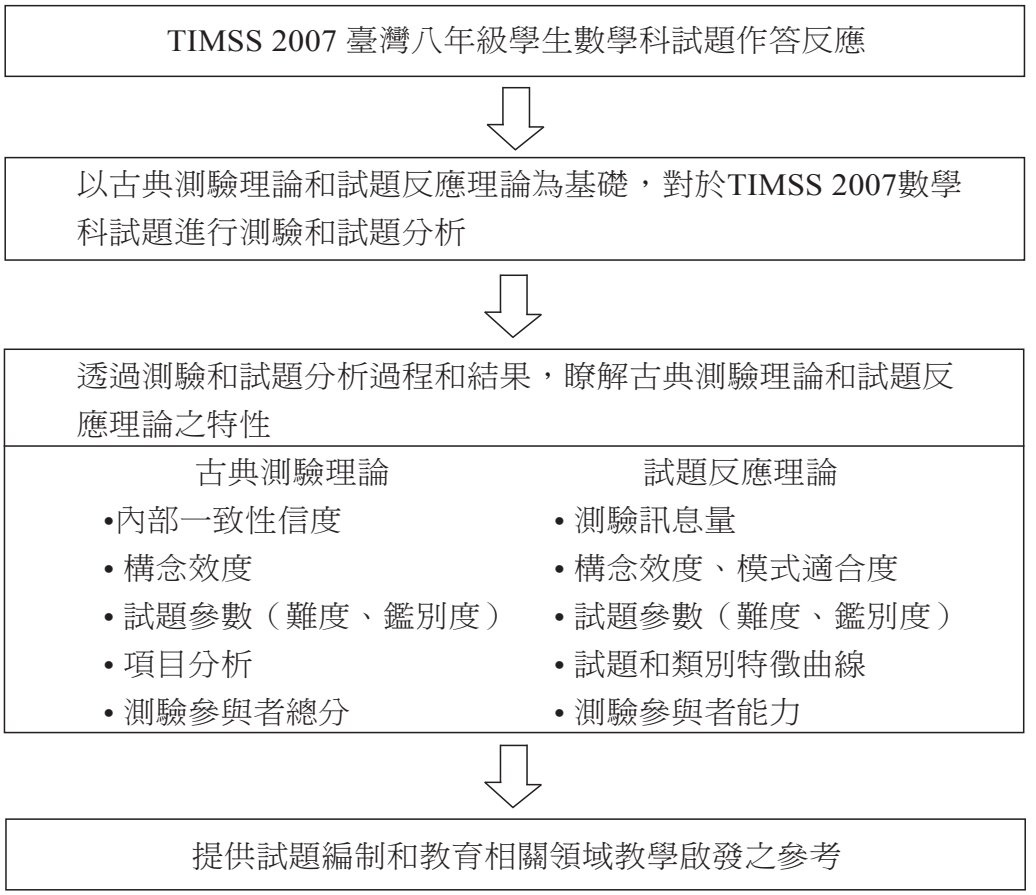

圖1 研究架構圖

基於兩種不同理論基礎所收集的資料如圖1。總之，研究者先對TIMSS 2007數學科試題進行測驗和試題分析，再藉由測驗和試題分析過程和結果，瞭解古典測驗理論和試題反應理論之特性，最後，透過試題分析結果，提供試題編制和教育相關領域教學啟發之參考。

二、資料庫之研究參與者

參與TIMSS 2007臺灣八年級數學科評量的學生總計4,046人。參與評量的學生為八年級學生，平均年齡年滿13.5歲。抽樣方式為二階段的階層叢集設計，第一階段先抽樣學校，第二階段再抽樣班級。每個國家均取樣約150所學校，每所學校再抽取1或2個班級，施測時不同學生可能作答不同題本。本研究由於必須同時比較古典測驗理論和試題反應理

論分析結果，因此選擇題本組合一（booklet 1）的數學科試題作為研究工具，實際分析之作答反應共有287人。

三、研究工具

研究工具為題本組合一，包含M01和M02兩種數學題本。M01為TIMSS 2003試題，有13題；M02為TIMSS 2007新編試題，有16題，合計29題。施測時間為45分鐘。題目區分為內容維度和認知維度，題本組合一的題目內容屬性整理於表2：

表2 TIMSS 2007題本組合一數學題目之內容維度、認知維度數量分析一覽表

內容領域	數量		代數		幾何		資料和機率		列小計	
認知領域	M01	M02	M01	M02	M01	M02	M01	M02	M01	M02
知識	2	2	1	2	0	1	0	1	3	6
應用	4	3	0	0	4	1	1	2	9	6
推論	0	0	0	2	1	1	0	1	1	4
欄小計	6	5	1	4	5	3	1	4	總計29	

數學科題目分為兩種類型：多重選擇題型和建構反應題型。其中，多重選擇題型均為二元計分（正確1分，錯誤0分），小計16題；建構反應題型計分方式分為二元計分（正確1分，錯誤0分）和多重計分（完全正確2分，部分正確1分，錯誤0分），建構反應之二元計分題型有9題，建構反應之多重計分題型有4題，小計13題。試題總計29題。

四、研究程序

本研究之研究程序包含下列步驟：

（一）取得TIMSS 2007臺灣八年級學生數學科之作答反應。登入TIMSS的官方網頁（http://timss.bc.edu/index.html），點選所需的資料庫項目進行下載。

（二）將前述資料，以古典測驗理論和試題反應理論為基礎，透過相關軟體進行各項數值運算，進行測驗和試題分析。

（三）根據前述過程和結果並參考相關文獻進行兩種理論之特性探討，之後撰寫研究報告。

五、資料分析

TIMSS 2007資料庫作答反應編碼顯示，部分題目參與者有未作答的情形，此類作答反應計為0分。以下分別描述古典測驗理論和試題反應理論資料分析方式：

（一）以古典測驗理論為基礎進行測驗和題目分析。

在古典測驗理論部分，收集內部一致性的信度資料、以內容效度為主的構念效度資料、難度（通過率）、鑑別度、項目分析和測驗參與者總分等結果，前述資料使用軟體R的CTT套件和Excel 2003進行數值計算；在試題反應理論部分，收集測驗訊息量、以內容效度為主的構念效度資料和模式適合度、試題參數、試題和類別特徵曲線以及測驗參與者能力等結果。前述資料使用軟體R的eRm套件、SPSS 12.0、Excel 2003、ConQuest 2.0 （Wu, Adams, & Wilson, 2007）進行數值計算和曲線繪製。

試題反應理論分析須考量模式選擇，由於數學測驗題目同時包含二元計分和多重計分題型，且研究對象人數為287人，因此選擇單參數的部分給分模式（partial credit model, PCM）（Masters, 1982）作為分析模式。在部份給分模式當中，假設θ代表能力，第i題分數為x（$x = 0,\cdots,m_i$），因此第i題有m_i+1個反應類別，則當$x = j$時，其公式可以表示如下

$$P_{ix}(\theta)= \frac{\exp[\sum_{j=0}^{x}(\theta - \delta_{ij})]}{[\sum_{r=0}^{mi}\exp[\sum_{j=0}^{r}(\theta- \delta_{ij})]} \qquad (1)$$

其中$\sum_{j=0}^{0}(\theta - \delta_{ij})\equiv 0$

δ_{ij}（$j = 1,\cdots, m_i$） 是指類別分數為j時的階難度（step difficulty），當階難度值越高代表難度越高（Embretson & Reise, 2000）。換言之，當能力值為θ時，其在第i題的第j個得分類別之機率，可以用該公式加以

運算並描繪為試題反應曲線，δ_{ij}在曲線上是指各個類別分數機率交點所對應的難度值，若計分有0、1、2三種分數，則會有兩個階難度值。

完成模式選擇之後，再收集模式適合度檢定、試題參數和能力參數估計結果、試題和類別特徵曲線、測驗訊息量等資料。在模式適合度檢定方面，利用SPSS 12.0視窗中文版進行因素分析，檢視單向度假定，並使用軟體R的eRm套件計算infit和outfit等適配性統計量指標；在試題參數和能力參數估計結果方面，採用eRm套件完成數據運算，在估計參數的方法選擇上，根據陳柏熹（2006）提到，條件最大概似法主要應用於Rasch模式，因為只有在Rasch模式中，答對題數才是受試者能力值的充份統計量，也就是答對題數相同者的能力值會相同。由於本研究所欲探討之範疇亦以單參數為主，故而採用條件最大近似值估計法進行試題和能力之參數估計；在試題特徵曲線和類別特徵曲線，採用eRm套件和ConQuest 2.0完成曲線繪製；在試題訊息量方面，採用eRm套件完成數據運算和曲線繪製。其中，試題訊息量的計算公式為：

$$I_i\theta = \frac{[p'_i(\theta)]^2}{P_i(\theta)\,Q_i(\theta)} \quad i=1,\cdots,n \qquad (2)$$

公式（2）中，$I_i(\theta)$代表試題i在能力為θ上所提供的訊息，$p'_i(\theta)$為在θ點上的$P_i(\theta)$值的導數，而$P_i(\theta)$為能力θ在試題i上的試題反應函數，$Q_i(\theta)=1-P_i(\theta)$（余民寧，1992b）。從另一方面來解釋，訊息量也反映出試題在不同能力點的測量誤差（standard error, SE），訊息量愈高表示測量誤差愈小（陳柏熹，2006）。一份測驗的訊息函數是指它在某一個θ值上所提供的訊息量，該訊息量剛好是在θ值上的試題訊息函數之總和（余民寧，1992b）。綜合前述內容，本研究採用部分給分模式、探索性因素分析、條件最大概似法和前述試題訊息量公式，進行試題反應理論之各項數值估計運算。

肆、結果與討論

在結果與討論部分，從第一至第五部分將測驗和試題分析結果嵌入古典測驗理論和試題反應理論可供對照之特性（可參見圖1）加以描

述，並輔以相關文獻進行討論，第六部分以小結整合前述內容，並描述提供試題編制和教育相關領域教學啟發參考之處。

一、內部一致性信度 / 測驗訊息量

古典測驗理論當中，測驗結果的一致性使用信度的概念加以描述；試題反應理論當中，訊息量說明了測驗的精準程度，前者透過信度數值、後者透過測驗訊息量數值，可以計算測量標準誤。在古典測驗理論部分，透過軟體計算之Cronbach's alpha係數為0.90，顯示測驗整體具有良好的內部一致性，刪題後Cronbach's alpha係數均能夠維持在0.90以上，顯示各試題均具有良好的題目間同質性的表現。

在試題反應理論部分，隨著測驗參與者的能力值不同，測驗訊息量亦有所改變，當題目的難度越符合參與者的能力值時，就能提供較高的測量精準度。圖2為測驗訊息量示意圖，圖中橫軸為數學能力，縱軸為測驗訊息量。由圖可知，測驗整體對於中間能力者所提供的訊息量較高，其餘則遞減。不同能力可以計算出不同的標準誤（參見表10）。

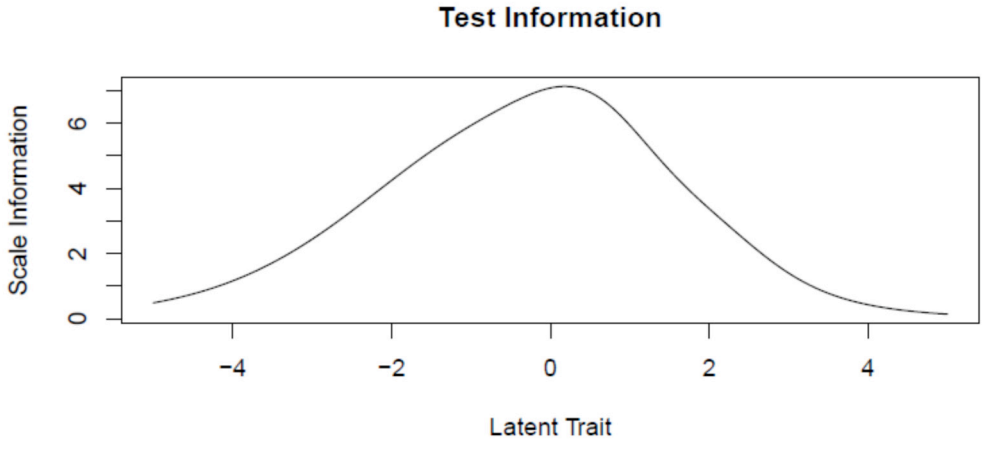

圖2 測驗訊息量示意圖

根據前述結果可知，古典測驗理論從信度描述測驗整體品質，試題反應理論藉由訊息量了解測驗精準程度，兩種測驗理論對於測驗品質的描述觀點有所不同。古典測驗理論認為信度就是指相同的個人在不同的時間，以相同的測驗測量，或以複本測驗測量，或在不同的情境下測

量，所得結果的一致性（郭生玉，1999）；在試題反應理論中，不論是單一題目或整份測驗，對不同能力的受試者會提供不同的測量精準度（陳柏熹，2006）。如同本研究所得結果，Cronbach's alpha係數反映了測驗的分數當中真實量數或是沒有誤差的程度，而測驗訊息量則反映了測驗對於不同能力測驗參與者具有不同的精準程度。

若從測量誤差的角度審視，Embretson與Reise（2000）提到，古典測驗理論對於所有的分數共用相同的測量標準誤；試題反應理論給予不同能力的測驗參與者不同的測量標準誤。在本研究中，以古典測驗理論為基礎透過公式計算出的測量標準誤只有一個，而試題反應理論則可以針對不同能力者提供不同的測量標準誤。余民寧（1991）提到古典測驗理論以一個相同的測量標準誤，作為每位受試者的測量誤差指標，這種作法並沒有考慮受試者能力的個別差異，對高、低能力兩極端組的受試者而言，這種指標極為不合理且不準確，致使理論假設的適當性受到懷疑，試題反應理論能夠針對每位受試者，提供個別差異的測量誤差指標，而非單一相同的測量標準誤，因此能夠精確推估受試者的能力估計值。

二、構念效度 / 構念效度和模式適合度

在效度部分，兩種理論均需考量內容效度。TIMSS 2007的技術報告手冊（Olson, Martin, & Mullis, 2008）提到其試題評估，歷時兩年以上時間，參與試題發展的有來自各國的教育專家、測驗發展專家，參與機構有TIMSS & PIRLS International Study Center和Science and Mathematics Item Review Committee [SMIRC]，彼此相互合作達到回顧與修訂試題、檢視正確性、確認試題是否符合雙向細目表的架構等三項目的。而且，試題發展具有詳細的雙向細目表，分為內容維度和認知維度。在內容維度部分，分為數量、代數、幾何、資料和機率四部分；在認知維度部分，分為知識、應用、推論三個部分，並明確設定各維度所佔題目之比例。茲將TIMSS 2007的試題編制過程對照Crocker和Algina（1986）所提到的測驗理想編制步驟，說明如表3：

表3 TIMSS 2007和Crocker & Algina（1986）測驗編制步驟對照一覽表

Crocker和Algina（1986） 測驗編製步驟	TIMSS 2007 測驗編製步驟對照
1. 確認測驗分數使用的主要目的。	測驗分數的用途在於國際間教育成就之研究，目的在於提升世界各國數學和科學的教學和學習。
2. 確認行為代表的構念或定義的範圍。	透過第一次National Research Coordinators [NRC]的會議，參與人員修訂評量架構，
3. 準備一組測驗的詳細說明，描繪出項目在每種行為上應有的比例。	以八年級數學科為例，測驗維度分為內容維度和認知維度，題型、各領域得分比例也都事先明訂。
4. 建構初步的項目。	在第二次NRC會議中，相關人員參與試題寫作工作坊，接受試題寫作之指導，依據評量架構編製試題。初步完成試題編制後，成員重新檢視評量架構修正或補足不足項目試題。
5. 項目複查。	試題初步完成後，由SMIRC檢視試題的內容正確性、年級適當性、評量架構適配性。在田野測驗前，事先回顧試題並予以分組，試題會在第三次NRC會議中加以檢視、修正。
6. 進行項目初試。 7. 在應試者中選擇樣本來測試。	八年級部分共45個國家、每國至少25校的隨機樣本，參與田野測驗。
8. 決定項目分數的統計特性，淘汰不符合先前標準的項目。	根據田野測驗結果，進行試題選擇。選擇鑑別度、難易適合度兼備、涵蓋各項認知內容維度等試題。
9. 為最後的測驗形式計畫信度和效度。	第五次NRC會議，再度針對試題進行討論，決定八年級試題共429題。
10. 發展測驗分數的管理、計分和解釋的指導方針。	1. TIMSS 2007的計分，錯誤類型也可從紀錄的編碼略見端倪。 2. 在進行田野測驗之前，為使計分手冊更為完備，先以英語語系國家進行前測，以建構反應題型為主，收集學生反應以利計分者參考，藉由學生真實反應修訂試題。 3. 完成前測之後，在第四次NRC會議中進行計分訓練，訓練包含計分的教導討論和實務演練。 4. 最後定版之試題和計分指導手冊，由NRC和SMIRC再次確認。

整體來說，TIMSS在試題編制的程序部分，符合Crocker和Algina（1986）所述的試題編制內容，透過各領域專家參與、不斷地修訂和討論、和實際進行施測，確認試題是否符合需求，亦有問卷調查試題內容是否符合各國教學課程；在試題編制的內容部分，具有明確的評量架構，擬定結合內容維度和認知維度的雙向細目表，透過各領域專家在試題編制過程的數次討論、修訂和實際實施，確認題目測量之能力意涵是各國所認同的，用以確保題目能夠測量所欲測量的能力。透過層層檢驗可以降低「構念的代表性低落」和避免測驗測量到「與構念無關的變異量」。

此外，古典測驗理論和試題反應理論皆可透過因素分析檢視測驗結構是否符合預期。本研究使用探索性因素分析，採用主軸法抽取因素，結果顯示，Bartlett 球形檢定近似卡方分配值為2,918.64（p = .000），因素間的獨立互斥性達到顯著水準，若選擇特徵值＞1的因素，共有7個因素，第一因素的特徵值為9.14，佔總變異量31.51%，第二因素的特徵值為1.34，佔總變異量4.64%。因素分析結果之陡坡圖如圖3所示：

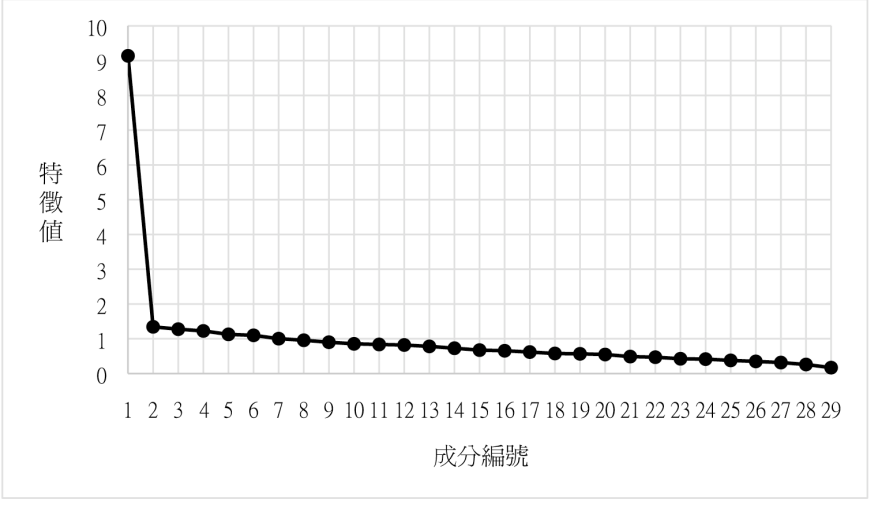

圖3 因素分析結果陡坡圖

依據Reckase（1979）建議，可利用第一因素的特徵值與第二因素特徵值的比值是否大於3作為檢測標準，本研究中第一因素和第二因素的特徵值比值為6.82（＞3），符合Reckase（1979）所建議測驗單向度的評鑑標準。除了運用因素分析檢測單向度假定之外，Hattie（1985）建議利用測驗內部一致性等資料判斷是否測驗題目均測量到相同特質。

從因素分析結果對照Cronbach's alpha係數結果0.90，可知測驗題目大體上均測量到相同特質並符合單向度假定。

除了因素分析之外，試題反應理論的效度概念亦包含資料和模式適配度，有關資料和模式適配度指標的統計數值，呈現如表4：

表4 資料和模式適配度之統計數值一覽表

題號	Outfit MSQ	Infit MSQ	題號	Outfit MSQ	Infit MSQ
M022043	1.69*	1.30	M042018	0.64*	0.89
M022046	0.87	0.99	M042055	0.97	1.01
M022049	1.44*	1.29	M042039	1.31*	1.31*
M022050	0.85	0.93	M042199	0.42*	0.81
M022055	0.81	0.81	M042301A	0.93	1.05
M022057	1.47*	1.35*	M042301B	0.64*	0.79
M022257	0.61*	0.82	M042301C	0.53*	0.62*
M022062	0.97	0.91	M042263	0.84	0.87
M022066	0.33*	0.70	M042265	1.13	1.11
M022232	1.17	0.95	M042137	0.90	1.02
M022234A	1.06	1.01	M042148	0.70	0.93
M022234B	0.91	1.01	M042254	0.76	0.98
M022243	0.74	0.80	M042250	0.38*	0.85
M042003	1.70*	1.16	M042220	1.47*	1.20
M042079	0.78	0.82			

註：*表示超出1±0.3範圍的數值。

表4所提供的資料和模式適配度的判斷指標為以殘差為基礎的適配性統計量：Outfit MSQ和Infit MSQ。期望值為1，數值與1相距越遠代表適配程度較差，若數值落於0.7~1.3之外，則代表適配程度較差（Bond & Fox, 2007）。表中顯示部分數值超出建議範圍，研究者僅繪製兩種指標數值均超出範圍者之試題特徵曲線，題號分別為：M022057、M042039、M042301C。請參考圖4至圖6。

由圖4可知，在題目M022057中，部分中低能力者的實際作答資料和理想特徵曲線的差異，似乎是造成適配性統計量數值超出0.7~1.3範圍的主要因素，類似的情形也發生在圖6的題目M042039當中；而圖5的題目M042301C，參考表5可知其兩項指標均低於0.7，當數值低於0.7時代

表資料和模式的變異性過小，呈現過度適配情形，因此試題能夠提供的訊息量較為不足。就測驗整體而言，資料和模式適配度大致良好。且綜合前述因素分析和適配度指標的相關數據，可知資料和部分給分模式尚稱適配，可以支持研究者選擇適當模式進行資料分析的合理性。

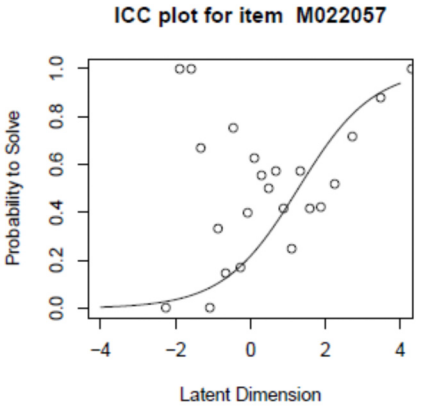

圖4 M022057之原始資料和理想試題特徵曲線樣貌圖

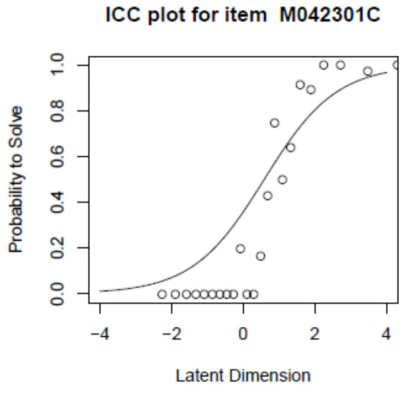

圖5 M042301C之原始資料和理想試題特徵曲線樣貌圖

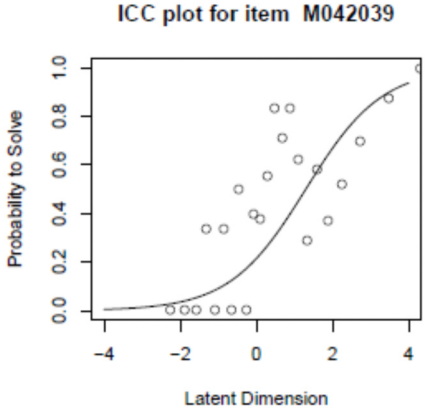

圖6 M042039之原始資料和理想試題特徵曲線樣貌圖

註：圖4至圖6中黑線代表理想特徵曲線，空心圓點代表實際作答反應。

由上述結果可知，古典測驗理論和試題反應理論均需檢視測驗之構念效度證據，古典測驗理論可適時運用因素分析確認結果是否符合測驗架構，試題反應理論可運用因素分析確認測驗是否符合單向度假定。古典測驗理論不需考慮資料和模式之間的適配度；而試題反應理論必須立基於模式的適合度上分析測驗參與者和試題特性的相互關係。本研究

採用試題反應理論進行分析時，必須考慮作答反應計分方式、試題參數需求、樣本人數等條件，選擇適當的模型和參數估計方法。恰如余民寧（1991）所提，古典測驗理論的內涵，主要是以真實分數模式為理論架構，依據弱勢假設（weak assumption）而來，當代測驗理論的內涵，主要是以試題反應理論為理論架構，依據強勢假設（strong assumptions）而來。換言之，古典測驗理論主張各種不同計分方式與作答方式的測驗都使用同一套模式，而現有各種不同的IRT模式，適用於不同計分方式與作答方式的測驗中（陳柏熹，2006）。

三、試題參數（難度、鑑別度）

以下依序分述兩種理論的試題參數計算結果。在古典測驗理論部分，說明難度、鑑別度結果；試題反應理論部分，由於部分給分模式假定鑑別度為1，故而僅說明難度結果。

關於古典測驗理論的難度計算，多重選擇題型和建構反應之二元計分題目，則計算「反應正確的答對人數比率」，建構反應之多重計分題目，則採用「得分平均數／每題最高分數」的公式計算。結果顯示：在29題中有14題通過率超過0.80、14題通過率在0.60~0.80之間，只有一題（題號M022232）的難度為0.30。由此可知，臺灣學生在大多數題目中表現均相當良好。且參考多重計分題的選答人數比例，可知學生在建構反應的多重計分的題目上，大多呈現完全正確或是錯誤的情形，部分正確的比例較低。就古典測驗理論觀點而言，測驗可能缺乏較難題目，以區別能力較高的學生，郭生玉（1999）提到難度集中在0.50左右的測驗，信度最高，在這種情況之下，測驗應增加困難的題目，以使測驗分數的分配能接近常態，測驗才能達到區分各種能力水準的最大作用。在鑑別度方面，計算「該題得分和總分之相關係數」和「總分排除該題與該題得分之相關係數」兩項。結果顯示：在「該題得分和總分之相關係數」當中，29題題目當中有28題的相關係數高於0.3，只有1題（題號M022049）略低於0.3；在「總分排除該題與該題得分之相關係數」當中，無論刪除任何一題，均會降低相關係數。由此可知，幾乎各個題目均具有良好的鑑別度，且沒有試題應予以刪除。

關於試題反應理論的試題難度參數估計結果，由易到難排列呈現如表5：

表5 TIMSS 2007八年級臺灣學生數學科試題難度參數估計一覽表

試題	平均難度	階難度1	階難度2	試題	平均難度	階難度1	階難度2
M042250	-2.89			M042137	-0.49		
M042254	-2.04			M042301B	-0.44		
M022066	-1.62			M042301C	-0.17		
M042199	-1.58			M022243	-0.12		
M022046	-1.53			M042265	-0.03		
M042079	-1.44			M022050	0.16		
M042148	-1.44			M022055	0.36		
M042003	-1.20			M042263	0.39		
M022049	-1.02			M022057	0.49		
M042018	-0.92			M042039	0.51		
M042301A	-0.92			M022232*	2.14	4.32	-0.05
M022257	-0.82			M022234A*	0.83	1.88	-0.22
M042055	-0.79			M022234B*	0.69	2.45	-1.08
M022043	-0.67			M042220*	0.22	2.37	-1.92
M022062	-0.52						

註：*表示試題為多重計分題，計分方式為0、1或2分，故而有兩個階難度估計值和一個平均難度估計值。其餘題目皆為0、1計分，難度值僅有一個估計值。

由表5可知，二元計分和多重計分題型相較之下，二元計分題型難度相對較低。在二元計分題目當中，題目的難度分布於-3 ~ 0.5 logit左右，顯示題目難度為中等偏易的趨勢；多重計分題目平均難度約分布於0 ~ 2 logit之間，各題的階難度均呈現跨越第二個階難度比跨越第一個階難度較為容易的情形，顯示作答1分人數較少，意味著試題可能無法區分1分（中間能力）的群體，導致階難度的跨越順序呈現相反的情形，請參見圖7至圖10。Wu和Adams（2007）提到難度的失序只能意味選擇1分的人數較少、後來的階難度相較於先前的階難度較為容易跨越，並不代表試題必定產生瑕疵。但回頭思考TIMSS 2007編制多重計分題選項的目的，原本即在於區分不同數學能力的學生，因此多重計分題似乎未能滿足預期目的。在鑑別度方面，本研究選用部分給分模式，故而鑑別度

假定為1。由於先前數據顯示資料和模式適配度尚可，因此研究者並未改採用其他模式，若要選用具有鑑別度估計數值的多重計分模式，可採用二參數的通用部分給分模式（Generalized partial credit model, GPCM）（Muraki, 1992）。研究者將兩種理論算出的難度值求取相關，結果為-0.96，顯示古典測驗理論和部分給分模式計算出的難度值具有高度相關。

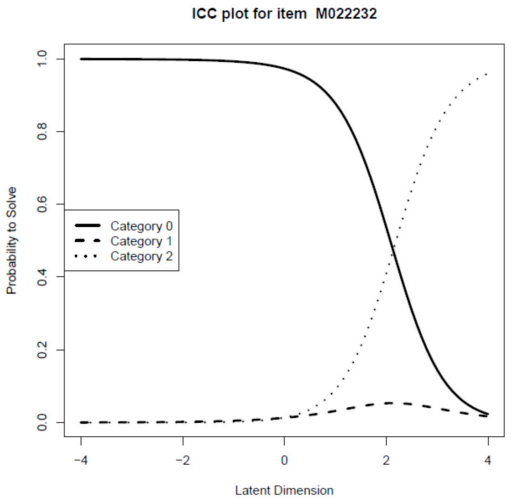

圖7 M022232試題特徵曲線

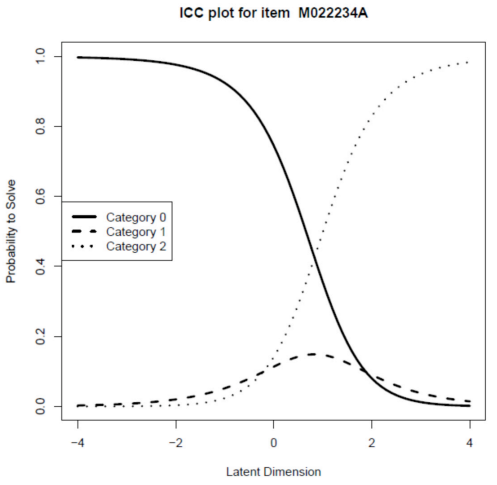

圖8 M022234A試題特徵曲線

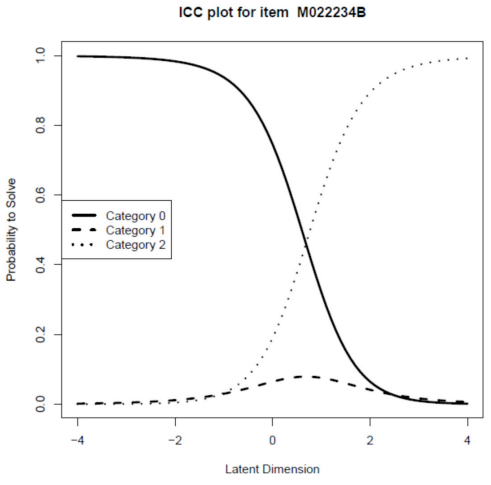

圖9 M022234B試題特徵曲線

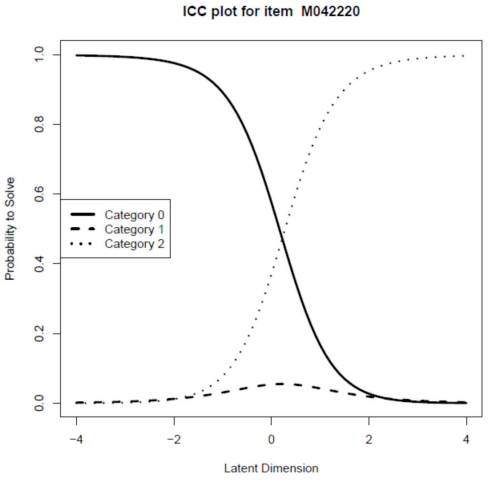

圖10 M042220試題特徵曲線

由前述結果可知，古典測驗理論透過計算難度、鑑別度，作為了解試題特性的方式，試題反應理論透過適當模式的選擇，估計試題難度、鑑別度，並可描繪試題特徵曲線，呈現試題和不同能力的相互關係。陳柏熹（2006）提到，在古典測驗理論中，試題是難還是簡單，完全取決於抽樣時所選到的受試群體能力高低，因此樣本的代表性對試題參數的估計有很重要的影響力。同樣地，試題鑑別度也會明顯地受到受試群體的能力分散程度所影響。在IRT中，題目參數的估計不會受到受試者能力所影響。這主要是因為在IRT中已經將試題參數與受試者能力同時納進其模式裡，因此在估計試題參數時已經考量了受試者能力的影響，因此所估計出來的試題參數不會受到受試者能力所影響。

四、項目分析／試題和類別特徵曲線

從古典測驗理論進行項目分析，在各題項之選擇比例方面，多重選擇題型能夠提供各題項作答比例之分析資料，題本組合一當中的多重選擇題型共有十六題。所有題目均能符合後續條件：（一）總分較高者，選擇正確項目比例較多；（二）總分較低者，選擇錯誤項目比例較多；（三）各個錯誤項目，均有總分較低者選擇。唯其中兩題，研究者發現誘答選項呈現中能力者選擇比例高於低能力者的情形。此兩題題目如圖11、12，項目分析結果列於表6：

M022057
Content domain：number
Cognitive domain：applying
Maximum points：1
key：C

有一年某一家公司賣出了1426公噸的肥料，第二年這家公司少賣了15%。下列哪一個數字最接近這家公司第二年所賣出肥料的公噸數？

(A) 200
(B) 300
(C) 1200
(D) 1600
(E) 1700

圖11 試題M022057

註：中文題目來源為TIMSS 2007國際數學與科學教育成就趨勢調查資料庫，網址：http://www.dorise.info/DER/01_timss_2007_html/index.html

M042039
Content domain：number
Cognitive domain：applying
Maximum points：1
key：A

一件外套平常賣60元。在減價30%的時候，阿倫買了這件外套。
請問阿倫省了多少錢？

(A) 18元
(B) 24元
(C) 30元
(D) 42元

圖12 試題M042039

註：中文題目來源為TIMSS 2007國際數學與科學教育成就趨勢調查資料庫，網址：http://www.dorise.info/DER/01_timss_2007_html/index.html

表6 TIMSS 2007 八年級數學科題本組合一之多重選擇題「可能需要適度調整題目」項目分析資料一覽表

題目一	答案	低分組	中間組	高分組
M022057	A	0.17	0.25	0.00
	B	0.15	0.03	0.00
	C*	0.40	0.67	1.00
	D	0.22	0.02	0.00
	X	0.06	0.02	0.00
題目二	答案	低分組	中間組	高分組
M042039	A*	0.45	0.62	1.00
	B	0.06	0.01	0.00
	C	0.27	0.02	0.00
	D	0.22	0.35	0.00
	X	0.00	0.00	0.00

註：1. X代表測驗參與者未作答。

2. *表示該選項為標準答案。

儘管試題M022057和M042039亦符合前述三項標準，但試題M022057的A選項和M042039中的D選項均顯示中間組的選答人數比例稍微高於低分組，顯示該選項無法明確區辨中間組和低分組。

試題反應理論亦可進行項目分析，研究者將多重選擇題型繪製成曲線，以了解測驗參與者在各個選項上的作答情形。依照常理推斷，隨著能力增加，選擇誘答選項機率應該逐漸降低。但同樣發現試題M022057的A選項和M042039中的D選項呈現作答機率隨著能力增加有略微升高、再下降的情形，兩題的試題特徵曲線和類別特徵曲線如圖13和圖14所示：

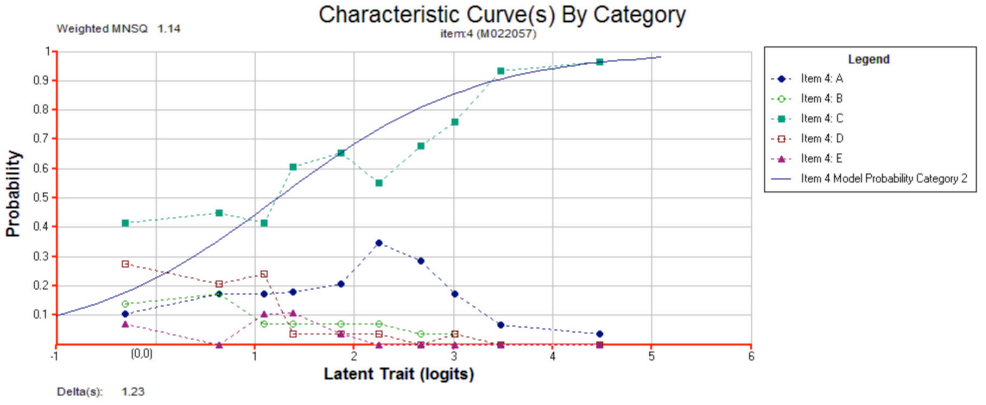

圖13 M22057試題特徵曲線暨類別特徵曲線

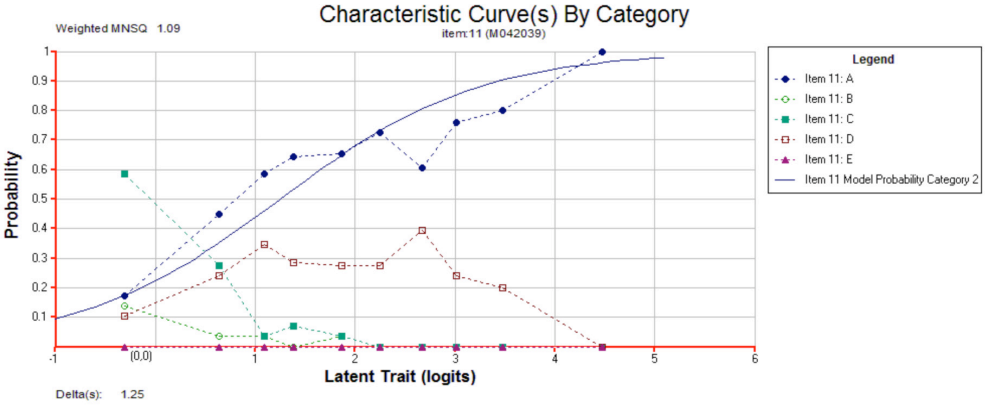

圖14 M042039試題特徵曲線暨類別特徵曲線

註：ConQuest 2.0預設十組能力點繪製各選項之類別特徵曲線

總之，前述試題分析和類別特徵曲線具有一致性的結果：分數較高或是能力較高選答此兩題誘答選項的機率反而較高。進一步檢視試題內容，試題參見圖11和圖12。就試題M022057而言，推測中間組學生可能沒有注意到題目問的是「哪一個數字最接近這家公司第二年所賣出肥料的公噸數？」，而把注意力集中於「第二年這家公司少賣了15%」，因此造成中間能力學生可能因此誤選A選項。而試題M042039，推測中高能力學生可能是沒有注意到題目問的是「請問阿倫省了多少錢？」而直接認為是「請問阿倫花了多少錢？」，所以產生誤選D選項的情形。事實上，這兩題的狀況有些類似，研究者認為題目或許可以在關鍵字「賣出」和「省了」加上底線標示，避免可能有數學能力之外的其他因素影響作答反應。

五、測驗參與者總分／測驗參與者能力

古典測驗理論和試題反應理論分別運用原始總分和能力估計值描述學生能力。測驗參與者之原始分數和能力估計結果，列出如表7：

表7 TIMSS 2007八年級臺灣學生數學科原始分數和能力參數一覽表

原始分數	能力估計值	標準誤	原始分數	能力估計值	標準誤
3	-3.01	0.64	19	0.13	0.37
5	-2.34	0.53	20	0.27	0.37
6	-2.08	0.49	21	0.41	0.38
7	-1.85	0.47	22	0.56	0.38
8	-1.64	0.45	23	0.71	0.39
9	-1.44	0.44	24	0.86	0.40
10	-1.26	0.42	25	1.02	0.41
11	-1.08	0.41	26	1.20	0.43
12	-0.91	0.41	27	1.40	0.45
13	-0.75	0.40	28	1.62	0.48
14	-0.59	0.39	29	1.87	0.52
15	-0.44	0.39	30	2.17	0.58
16	-0.29	0.38	31	2.55	0.67
17	-0.15	0.38	32	3.16	0.93
18	-0.01	0.38	33	3.82	NA

註：NA表示沒有標準誤的計算值。

在古典測驗理論當中，以測驗總分來了解測驗參與者的數學表現，試題反應理論則計算能力估計值來描述其數學表現，隨著不同能力估計結果，標準誤亦隨之改變。由表7可知，某些低分原始分數並無得分者，中間原始分數的標準誤較小，原始分數越高或是越低則標準誤愈大，顯示對於中間能力測驗參與者的能力估計值較為準確。此外，由於試題參數和能力估計參數整合的單位相同，可以透過軟體繪製試題參數估計和能力參數估計對照圖，呈現兩者之間的關係。

六、小結

茲將前述各項結果，整理於表8。

表8 古典測驗理論和試題反應理論之分析結果對照表

測驗理論 資料項目		古典測驗理論	試題反應理論
測驗整體	信度	● 內部一致性係數 從Cronbach's alpha係數和刪題後Cronbach's alpha係數，得知各題目間同質性相當一致。	● 測驗訊息量 由測驗訊息量可以得知，試題對於中間能力的參與者具有較佳的測驗精準度。
	效度	● 構念效度 具有完整的雙向細目表和完善的試題編製過程。 由因素分析結果和內部一致性係數數據可知，測驗題目大致測量到相同特質。	● 構念效度 具有完整的雙向細目表和完善的試題編製過程。 ● 模式適合度 由因素分析結果可知測驗題目大體上符合單向度假定。 ● 資料與模式適配度 資料與模式適配度大致良好，試題M022057、M042039、M042301C數值落於理想範圍之外。
測驗試題分析	難度	題目的難度均在0.6以上，顯示測驗可能缺乏較難題目	試題難度對於參與者而言呈現中等偏易的樣貌。
	鑑別度	計算點二系列相關，題目均具有良好的鑑別度。	部份給分模式假定鑑別度為1，由於資料和模式適配度尚可，故而代表各題鑑別度尚稱一致。
	項目分析	透過百分比進行項目分析，發現試題M022057、M042039的誘答選項呈現中間能力作答比例稍微高於低分組情形。	● 建構反應題型（多重計分 透過試題特徵曲線發現，試題M022232、M022234A、M022234B、M042220有階難度相反情形。 ● 多重選擇題型 透過類別特徵曲線可知，試題M022057、M042039的誘答選項呈現中間能力作答比例稍微高於低分組情形。

回應研究問題一，就古典測驗理論而言，在測驗信度部份，題本組合一的Cronbach's alpha係數為0.90，而「刪題後Cronbach's alpha係數」顯示無論刪除何題，該係數均能夠維持在0.90以上，顯示各題目均具有良好的題目間同質性的表現；效度部份，TIMSS 2007具有審慎的測驗編製過程，並依據內容維度和認知維度訂定明確的雙向細目表，因素分析結果顯示測驗題目大致測量到相同特質；難度資料結果顯示，幾乎所有題目的難度（通過率）均在0.6以上，顯示測驗可能缺乏較難題目，以區別能力較高的學生，也許測驗應適當增加部分較為困難題目，才能達到區分各種能力水準的最大作用；鑑別度資料結果顯示，幾乎各個題目均具有良好的鑑別度，沒有試題應予以刪除；在項目分析資料結果顯示，幾乎所有題目都能符合：總分較高者，選擇正確項目比例較多、總分較低者選擇錯誤項目比例較多、各個錯誤項目均有總分較低者選擇等條件，唯試題M022057、M042039的誘答選項呈現中間能力作答比例稍微高於低分組情形，似乎無法明確區辨中間組和低分組考生。大體而言，若以古典測驗理論作為分析基礎，雖然有難度較低的情形，但整體而言測驗的品質相當良好。

就試題反應理論而言，在測驗訊息量方面，整體測驗對於中間能力的參與者能夠提供較佳的測驗精準度；在效度方面，內容效度之檢測為兩種測驗理論兼備，至於模式適合度檢測，因素分析結果支持該測驗為單向度，資料和模式適配性統計量結果大致上適配的情形，試題M022057、M042039、M042301C數值落於理想範圍之外，除試題M042301C呈現過度適配，另外兩題題目呈現可能不適配情形，選擇部分給分模式進行資料分析尚稱恰當；在試題難度參數方面，試題對於參與者而言呈現難度偏易（分布於-3 ~ 0.5 logit左右，顯示題目難度為中等偏易）的情形；鑑別度方面，研究者選擇之模式假定鑑別度均為1，尚有其他具有鑑別度之試題反應模式可以選用；在能力參數估計方面，相對試題難度偏易，學生以中高能力者人數較多；在項目分析方面，繪製試題特徵曲線發現多重計分題M022232、M022234A、M022234B、M042220有階難度相反情形，亦即作答記為1分者人數較少；類別特徵曲線亦發現試題M022057、M042039有誘答選項機率與能力趨勢不甚符合常理的情形。大體而言，若以試題反應理論作為分析基礎，亦呈現試題對於測驗參與者較為容易的情形，並且顯示多重計分題可能無法區辨

中間能力者。

綜合前述結果，可以反映試題品質大致良好，僅有少數題目呈現試題功能未能十全十美，無法區辨中間能力者樣貌，且試題對於學生較為容易。

回應研究問題二，兩種理論的信度資料顯示不同特性。傳統測驗理論其測量精確度的評估是以測驗為單位所計算出來的，也就是測量標準誤，由於接受相同測驗的受測者其信度都相同，因此測量標準誤也被視為相同（陳柏熹，2011）。而在試題反應理論利用測驗精準度概念取代信度主張，試題訊息量和測驗訊息量隨著能力不同而有所改變，因此，測驗能夠對不同能力者提供不同測量精準度，此點相較於古典測驗理論更具合理性和實質意義。在效度概念方面，兩種測驗理論均需考量測驗能夠被理論或是證據所支持的程度，也就是構念效度，但在試題反應理論方面，研究者尚須考量作答反應類型、樣本人數等特性選擇適合模式，並進行資料和模式的適配度分析，而古典測驗理論則基於真實分數假設，相對限制較少，呼應了試題反應理論是以模式為基礎的測量理論（Embretson & Reise, 2000）。

至於在試題參數方面，本研究為同時比較兩種測驗理論的分析結果，僅能選擇題本組合一作為研究工具。此點顯示了古典測驗理論樣本依賴之特性，故而在大型測驗題庫建置和施測時，若由不同測驗參與者作答不同題本，便無法比較試題參數和測驗參與者能力，而試題反應理論則基於試題和能力參數不相互影響的特性，經由等化技術，即使是不同測驗參與者作答不同試題，題目難度和測驗參與者能力亦可以置於同一量尺上相互比較。在項目分析方面，古典測驗理論可以分低、中、高能力組分析各項目作答比例；試題反應理論採用類別特徵曲線呈現隨能力改變、在各選項作答機率的變化情形，針對此點，研究者認為兩種理論並無孰與優劣之分。最後，在試題反應理論中影響個體作答反應的是潛在特質，也因此本研究中獲知的是個體的數學能力，而以古典測驗理論計算的是個體的數學總分。總之，試題反應理論假定作答反應受到潛在特質影響，古典測驗理論強調觀察分數為真實分數和誤差之和，試題反應理論的潛在特質假定將分數和影響分數背後的潛在特質加以區隔，就理論價值上更具意義。

綜合前述內容，試題反應理論的優勢展現在測驗訊息量、潛在特

質假設、參數不變性和試題等化上，然其必須考量模式選擇、模式適配度等問題，且涉及較為複雜的數學理論基礎，另外，在考生的能力解釋上，亦不同於古典測驗理論之假定。

回應研究問題三，在提供試題編制參考方面，由於試題對於學生而言較為容易，或許可以考慮增加較難試題，以利區分中、高能力者。對於少數題目呈現試題功能未能十全十美的情形，研究者認為試題M022057、M042039或許可以考慮在關鍵字加註底線。在提供教育相關啟發參考之處，儘管整體題目對於考生相較容易，但唯獨在試題M022232呈現題目偏難的情形（通過率0.3，IRT難度值2.14；作答0分者有69%，1分者有3%，2分者有28%），與同屬「數量向度」的其他題目表現明顯具有差異（其餘題目通過率分布為0.64~0.9，IRT難度值分布為-1.62~0.51）。該試題在分類架構屬於「數量」和「應用」向度（分類架構可參見表2），以建構反應作答方式進行，題目如圖15：

M022232
Content domain：number
Cognitive domain：applying
Maximum points：2

君君做了一個表，用來記錄燒杯裡的水從95℃冷卻到70℃所花費的時間。他以5℃為間隔來測量水冷卻所需的時間。

下降溫度區間讀數	冷卻所需的時間
95℃ — 90℃	2分10秒
90℃ — 85℃	3分19秒
85℃ — 80℃	4分48秒
80℃ — 75℃	6分55秒
75℃ — 70℃	9分43秒

請估計一下：全部的水從95℃冷卻到70℃所需花費的時間最接近多少分鐘？並請解釋你是如何得到這個估計值的。

估計：_____

解釋：

圖15 試題M022232

註：中文題目來源為TIMSS 2007國際數學與科學教育成就趨勢調查資料庫，網址：http://www.dorise.info/DER/01_timss_2007_html/index.html

由於研究者無法獲知學生的實際作答反應，若從自身教學經驗推測，未能瞭解題目所具意義、曲解表格閱讀方式、估計數字計算錯誤或是無法明確解釋自己使用方法等，均有可能為無法正確作答之原因。儘管造成學生未能正確作答因素可能性很多，但該題所欲測量的主要意涵可以和數感教育有所連結。數感（number sense）一詞，據美國數學教師協會（National Council of teachers of Mathematics, NCTM）在其1989年出版之數學課程與評量標準中的解釋是「根據數所代表的不同意義，產生一種對於數的一份直覺」（支毅君，1997）。擁有數字常識能力的人，對數字有較強的直覺感覺，能以各種不同的方式使用和解釋數字，並能創新數字的形式來解決問題，在不須使用紙筆計算的情境下，能以各種不同形式的數字瞭解方式做數學上的判斷，並能運用有效的策略來處理所面對的數字情境（曹雅玲，2006）。對於教師而言，可以參考Gurganus（2004）建議之二十項增進數感之方式進行教學活動，以利學生發展較佳數感能力。在Gurganus建議項目當中可能有利於此題題目之教學活動包含：計畫有力的估計經驗、測量並進行估計、收集資料並用圖表呈現。

伍、結論與建議

此部分針對本研究目的和結果提出結論和建議，再說明研究限制和未來研究方向。

一、結論與建議

針對研究問題一，本研究以古典測驗理論為基礎，透過內部一致性信度、構念效度、難度、鑑別度、項目分析和測驗參與者總分等資料，發現整體測驗和試題呈現難度較低的狀況，其中兩題項目分析發現誘答選項有中間組作答比率高於低分組的情形；若以試題反應理論為基礎，透過測驗訊息量、構念效度和模式適合度、試題參數、試題和類別特徵曲線和測驗參與者能力等資料，發現整體測驗和試題呈現難度偏易的狀況，有四題多重計分題呈現階難度失序情形，在試題和類別特徵曲線部分，與古典測驗理論發現同樣兩題有誘答選項有中間組作答比率高於低

分組的情形。兩種理論同樣顯示，試題雖有較為容易的現象，但整體試題品質相當良好。

針對研究問題二，本研究經比較兩種測驗理論為基礎之試題、測驗結果，歸納古典測驗理論提供的信度結果是整份測驗的一致性資料，而試題反應理論則是透過訊息量的觀點進行測驗和試題分析；在效度方面，兩種理論基礎均需透過合理的測驗編製過程和事先擬定的雙向細目表加以確保，而試題反應理論尚須考量模式選擇、模式和資料適配度等議題；在試題參數部分，兩種理論對於難度、鑑別度的主張有所不同，古典測驗理論具有樣本依賴情形，而試題反應理論則具備試題參數不變性的特色；在項目分析方面，兩種理論可進行百分比分析或是類別特徵曲線分析選項的合理性；在描述測驗參與者能力方面，古典測驗理論透過總分加以描述，而試題反應理論透過潛在特質能力加以描述。總之，試題反應理論的優勢展現在測驗訊息量、潛在特質假設、參數不變性和試題等化上，然使用時須考量模式選擇、模式適配度等問題，並充分理解其數學理論基礎，且在解釋考生能力意義時不同於古典測驗理論。

針對研究問題三，在試題編制方面，研究者認為或許可以考慮增加較難試題，以區辨中、高能力者，並在試題M022057、M042039中考慮在關鍵字加註底線，避免無關能力影響作答反應。在教學啟發之參考方面，研究者探討造成學生在試題M022232呈現題目偏難現象的可能原因，除了未能瞭解題目所具意義、曲解表格閱讀方式、估計數字計算錯誤或是無法明確解釋自己使用方法等因素，該題內容或許可以提供國內數感教育相關啟發，透過計畫有力的估計經驗、測量並進行估計、收集資料並用圖表呈現，或許是可以提升作答反應正確性的具體教學方法。

二、研究限制和未來研究方向

研究者認為研究限制和未來研究方向可以包含下列項目：

（一）模式的考量

試題反應理論係依據強勢假設，因而選擇適合的模式更為重要。本研究在試題反應理論部分，以部分給分模式進行試題參數估計。TIMSS的作答反應分析方式，使用試題反應理論的二、三參數、部份給分三種

模式：在多重選擇題（二元計分）的部分使用三參數模式、在建構反應題型二元計分題使用二參數模式、在建構反應題型多重計分題使用GPCM（General Partial Credit Model）（Muraki, 1992）（Olson et al., 2008）。事實上，在題本組合一當中具有兩組題組題型（共五小題），而題組題型可能違反局部獨立性假定，或許加入題組反應模式（Testlet Response Theory, TRT），可以使測驗分析更符合實際的試題特徵，讓研究結果更為嚴謹。

（二）試題的修改與調整

本研究找出少數試題或許可以再行調整，以達到明確區辨中間能力者的目標。然受限於研究者之人力和資源，只能夠完成初步測驗分析，並針對部分題目提出建議。未來或可繼續針對該類試題進行討論和施測，透過與測驗專家、教育專家、授課教師和受測學生等進行討論，輔以實際施測，以提升測量精準度。此外，或許可以透過比較資料庫中其他國家試題分析結果，了解試題是否顯現同樣的現象，作為調整依據將更具說服力。

（三）對於教育領域教學的啟示

本研究認為數感教育或許是可以提升學生在特定試題作答反應正確性的可能方法之一，此點或可提供數學教育參考，但造成題目作答表現不如其他題目的原因，應當再行核對原始作答反應結果，或是再收集實證資料，以利後續針對題目特徵和教育啟示的深入討論。未來研究方向或許可以考量進行實驗研究，或從個別差異的角度檢視個體在數學認知結構上所展現的特性，屆時方能提供教育相關領域較多助益。

誌　謝

本文感謝教育部「邁向頂尖大學計畫」與科技部「跨國頂尖研究中心計畫」（NSC103-2911-I-003-301）支持。

參考文獻

支毅君（1997）。我國國小三年級數感教學研究。**臺東師院學報，8**，83-116。

余民寧（1991）。測驗反應理論的介紹（一）——測驗理論的發展趨勢。**研習資訊，8**(6)，13-18。

余民寧（1992a）。測驗反應理論的介紹（二）——基本概念和假設。**研習資訊，9**(1)，5-9。

余民寧（1992b）。測驗反應理論的介紹（七）——訊息函數。**研習資訊，9**(6)，5-9。

曹雅玲（2006）。數感教學重要性之探討。**國民教育，46**(3)，101-110。

郭生玉（1999）。**心理與教育測驗**。臺北市：精華。

陳柏熹（2006）。**IRT在量表編製上的應用（上）**。取自http://www.rcpet.ntnu.edu.tw/IRT%E5%9C%A8%E9%87%8F%E8%A1%A8%E7%B7%A8%E8%A3%BD%E4%B8%8A%E7%9A%84%E6%87%89%E7%94%A8(%E4%B8%8B)95.1.2.doc

陳柏熹（2011）。**心理與教育測驗：測驗編製理論與實務**。新北市：精策教育。

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME] (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Holt, Rinehart and Winston.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gullikson, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gurganus, S. (2004). Promote number sense. *Intervention in School and Clinic, 40*(1), 55-58.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 Technical Report.* International Study Center, Boston College, Chestnut Hill, MA: TIMSS & PIRLS.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne, Australia: Educational Measurement Solutions.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest: Generalized item response modeling software* (2nd ed.). Hawthorn, Australia: Australia Council for Educational Research.

Inspecting the Characteristics of the Classical Test Theory and Item Response Theory by Using Test Analysis Results and the Responses of Taiwanese Eighth-Grade Students in the TIMSS 2007 Database

Hsu- Lin Su* Po-Hsi Chen**

Abstract

The purpose of this study was to investigate the characteristics of classical test theory (CTT) and item response theory (IRT) by using the responses given by eighth-grade Taiwanese students in the TIMSS 2007 database to conduct test and item analysis from 2 distinct test perspectives, to provide benefit for test design and education settings. A total of 287 students were included in the research, and Booklet 1 was selected as the research tool. Software such as R 2.13.1, Excel 2003, SPSS 12.0, and ConQuest 2.0 were used during data analysis and curve drawing. The results showed that the test difficulty ranged from medium to easy, and the 2 distractors embedded respectively in the 2 items showed that the proportion of average students was more than that of below-average students, according to the aforementioned theories. Moreover, item characteristic curves of 4 multiple-choice items were not ordered. In general, the test quality was high, despite slight flaws meaning that some items could not differentiate average students. However, items were easy for students. The reliability, construct validity, item parameters, category analysis, and overall scores of the test takers in CTT corresponded to concepts of test information function, construct validity, and model fit assessing, item parameters, item and category characteristic curves, abilities of participants in the IRT. The relative strengths of the IRT lie in test information function, latent trait assumption, and parameter invariance and test equating. Regarding the test design and educational implications, we suggest modifying presentations of 2 distractors and making connection to number sense education owing to a specific difficult item.

Keywords: classical test theory (CTT), item response theory (IRT), PCM, TIMSS

DOI: 10.3966/199679772014123102003

Section editor: Shwu-Ching Young

Received: December 31, 2013; Modified: March 25, 2014; Accepted: April 10, 2014

* Hsu-Lin Su, Doctoral student, Department of Educational Psychology and Counseling, National Taiwan Normal University, E-mail: aguri.su@gmail.com

**Po-Hsi Chen, Associate Professor, Department of Educational Psychology and Counseling, National Taiwan Normal University