國立臺灣師範大學教育心理與輔導學系
教育心理學報，2013，45 卷，1 期，1-18 頁
DOI 10.6251/BEP.20121101.1

多層面 Rasch 模式在數學實作評量的應用*

謝如山
國立台灣藝術大學
師資培育中心

謝名娟
國家教育研究院
測驗評量研究中心

本研究採用多層面的 Rasch 分析，來評估學生在實作評量所具備的潛在能力。本研究所使用的測驗內容在測量學生對數學的數感及生活應用的數學能力，總共有四個考題，三百一十四位四到六年級的學童，與三位評分者參與本研究。結果顯示即使評分者接受訓練並依據規準來進行評分，不同評分者之間的嚴厲度也確實存在差異，透過多層面的 Rasch 模式分析，可以將評分嚴厲度加以考量，並能協助研究者，偵測評分者不合常規的評分現象，而數據中所呈現之標準化殘差值大致爲隨機分布，顯示評分並無明顯的系統性誤差存在。由本研究結果可看出多層面 Rasch 模式在分析實作評量上相當有用，值得其他教育工作者參考。

關鍵詞：多層面 Rasch 模式、評分者信度、實作評量

選擇題是應用最廣的一種測驗題型，可以有效的測量各種不同知識與學習成果。選擇題之所以應用廣泛的原因眾多，其中包括評分容易，可以很快的辨認學生作答反應的型態，評分也不會因爲其他因素，如字跡潦草、寫錯字、或是其他因素造成評分的不公平。選擇題較容易有高信度，因爲就測驗題數而言，可以比簡答題題數來的多，因而容易提高信度。且選擇題的題目可以互相獨立，所以測量到的構念，可以廣泛的覆蓋多元的教學內容。再者，選擇題比其他種題型（例如簡答題、是非題、配合題）較易建構高品質的題目，分析角度而言也比較容易，有許多可使用的軟體快速的爲選擇題進行分析。

雖然選擇題具有許多優勢，但有其限制，例如，選擇題並不適合測量所有的知識與能力，因爲這種題型只能要求學生選出正確的答案，但是對於一些高階的數學或科學方面的問題解決與組織思考的能力，或是對於音樂、藝術等表演的能力，是無法測量的。此外，設計選擇題時，較不易找到看似有理卻不正確的誘答選項，一些較機靈的學生，可能會使用答題的技巧，來找出問題的答案，有可能學生並不知道要如何計算正確的答案，卻可以用巧妙的猜測來得分。反觀來說，

* 本篇論文通訊作者：謝名娟，通訊方式：hm7523@hotmail.com。

開放性問題不易猜測，而從答題中可以看出，學生對於知識理解的困惑之處，並依據所填答的內容，而給予適度的補救教學。再者，選擇題型式的考題，不見得適用所有的學生，尤其對於圖像式思考或是需要藉由引導步驟來循序漸進式找出答案的學生，可能並不合適。

相對應於選擇題的考試型式有很多種，而實作評量爲一種新興的考試型態。余民寧（2002）指出，在教學情境下常用的實作評量可分爲五種類型（1）紙筆表現（2）辨認測驗（3）結構化表現測驗（4）模擬表現（5）工作範本。實作評量有很多種形式存在，例如藝術與表演領域，可以要求學生演唱一首民謠或做一件手工藝品，並依據學生的演唱技巧、或是手工藝品的精細程度來進行評量。對於職業與工程教育課程，可以要求學生進行汽車維修、做一件木工來進行評量。當然，一般的學科測驗，例如數學、科學與語文等，也可以藉由實作評量，來評斷學習成果。就數學而言，可以測試學生是否能夠運用數學，來解決現實生活中所遇到的問題，也可能是在紙筆表現中，評估受試者是否能在模擬情境中應用知識與技能。而科學方面，可以藉由觀察、形成假設，蒐集資料以獲得科學的素養與能力。至於語言，雖說一般的辭彙與用法可以藉由筆紙測驗來測量，但是對於口語表達或是演說的能力，則是需由實作評量來完成的。

多元智慧理論（Theory of Multiple Intelligence）中，Gardner（1993）提出了人類的智慧有很多種，包括音樂、語言、數學、空間、肢體運動、知己、存在主義以及知人的八種智慧，對於學校的教學，應著重各個面向，而非只著重在智能上的評估。而對於評量方面，也應採多元的評量方式，方能測出學生真實的能力。雖然選擇題的題型有其優勢存在，但對於某些面向的智慧評估來說，是很困難的。

近年來，IEA（International Association for the Educational Achievement）所主辦的國際數學與科學趨勢研究（Trends in International Mathematics and Science Study，TIMSS）頗受國際注目，TIMSS 主要是將學齡的學童數學與科學的成就進行國際間的排名比較，在 TIMSS 的測驗中，除了進行一般傳統的選擇題式的筆紙測驗之外，亦採用實作評量的題目，要求學生藉由操作教具並將結果寫在評量單，評分者會依據學生所填答的內容進行評分，以了解學生的數理能力。由 TIMSS 的做法可得知，實作評量可以運用在大型考試中的，只要能夠建立評分的公平性、進行準確的時間控制、選取適當的評分人員，亦可以得到公平的結果，對於學生的成就表現，也可以藉由實作評量的評估，有更深入的瞭解。

實作評量的題目大多採多分題的形式，例如 4 分，學生答對部分概念得到 1 至 2 分，略有錯誤但大致正確得 3 分，完全正確得 4 分，而對於這種試題的分析，則多侷限在古典測驗理論的框架中，鮮少有文獻採用試題反應理論，來探索如何有效測量受試者的潛在能力。古典測驗理論下的分析所採用的指標，像是難度、鑑別度、信度，都會受樣本的影響（sample dependent），也就是說，所得到的這些指標，會依據受測樣本的不同而有所差異，同一份試題施測於兩個學校，在 A 校所得到的試題參數，和在 B 校所得到參數有可能會截然不同。另一方面，古典測驗理論下來估計受試者的能力值，容易受到試題的影響（item dependent），同一位學生在一份簡單的考試卷可得高分，而在一份難的考卷卻會得低分，只要測驗的題目不同，受試者所得到的能力估計值就會有很大的差異，在不同試卷的題目，即使內容相近，不同受試者之間的成績也無法做直接的比較與對照。然而，試題反應理論能補足古典測驗理論分析的缺點，進行較嚴謹的分析。

多層面的 Rasch 模式，是延續 Rasch 模式發展而來（Linacre, 1989），由於可以同時考量試題難度和評分者嚴厲度之間的關係，被廣泛的使用在許多不同的領域。例如，姚漢禱與姚偉哲（2007）將多層面的 Rasch 模式應用分析在雙不定向飛靶優秀選手的射擊技術上，結果發現多層面的 Rasch 模式可以提供教練在選手訓練建議，並找出選手的弱點，進而提升選手的成績表現。張新立與吳舜丞（2008）則將多層面 Rasch 模式應用在學術研討會論文評分，總共有 131 位評審，每篇論文之評分工作則由三名相關領域專家進行同儕評審。他們的研究中發現不同學術分組的評審委員之間，存在嚴厲度的差別，也就是某些組的評審委員，會評分偏高，有些則一致性的給分較低，若是使用原始總分來評定論文表現優劣，會造成不公平的現象。藍珮君（2012）則將多層面 Rasch 模式應用在華語文口語能力測驗中，其研究結果顯示，評分者經過多次訓練後，評分者之間嚴格度的差距有縮小，但還是有顯著性的不同，但是評分者本身給分一致，適配度在合理且可接受的範圍之內。

測驗領域也廣泛使用此理論進行分析，像是分析寫作考試（Engelhard, 1992; Twing & Williams, 1992）與臨床醫學上的證照考試等（Lunz & Stahl, 1990; Lunz, Wright, & Linacre, 1990; Lunz, Wright, Stahl, & Linacre, 1989）。多層面 Rasch 模式在實作評量領域也有許多應用（Smith & Kulikowich, 2004; Basturk, 2008），例如，Basturk（2008）用此模式來評斷學生在做報告時的表現，學生報告分成六組，每組會依據六個向度: 緒論、內容、描述、整體架構、圖表呈獻與表達力來做評分，在其研究中探討三個層面之間的關係，包括每一組的表現能力，六個向度的難度、與評審的嚴厲度，研究結果指出多層面 Rasch 特別有利於分析多分題的數據，尤其是牽涉到需要評分者來進行評分，能有效的把評分者個人的嚴厲進行考量，而對受試者的能力進行客觀的比較。此外，透過此分析亦能將原本順序尺度的數據，轉換成具有可進行加減的等距量尺（interval scale），進行轉換後的 logit 尺度，可以展現能力在各個向度的強弱，與進行各組別之間的比較（例如性別差異）。分析後所得到的參數更可進一步的分析，雖然 Rasch 模式廣泛的爲教育及心理計量領域所使用（王文中，2004），但是國內對於多層面 Rasch 的模式分析，尤其是在實作評量上的應用較少，因此，本研究將利用多層面的 Rasch 來分析實作評量的數據資料。

受試者在實作評量上的分數大多由評分者來進行給分。評分者之間的異質性可以透過嚴密的訓練，與完整的評分歸準，使其盡量達成一致，然而，人和人之間還是存在明顯的差異性，即使用客觀的框架將評斷的標準限制住，主觀的意識與個人好惡還是會在評分的過程中造成影響。尤其是在大型測驗上，幾千人的考題，需要上百位評分者進行評分，某些嚴格的評分者，給的分數偏低，而某些寬鬆的評分者，則給的分數偏高，這些評分者嚴厲度的差別，會嚴重影響到學生成績的公正性。然而，這種評分者嚴厲度差異的問題，能夠藉由多層面的 Rasch 分析來解決，使學生的成績，不再受到評分者個人嚴厲度的影響，而能夠公平的進行比較。本研究將使用此模式，並針對四個數學實作評量的題目進行分析，施測對象爲國小四年級、五年級、及六年級共三個年級 314 位學生，其中，三位評審依據評分歸準與學生的作答，進行 0~3 分的評分，希望藉由本研究的結果，提供未來分析實作評量數據的參考。

文獻探討

一、實作評量的簡介與基本原則

以往教學爲了顧及大規模施測的方便性與評分的客觀性，大多是使用選擇題的方式來評鑑學生成效，但近年來，社會變遷與提倡多元智慧的教學目標，使原本著重於基礎記憶知識的教學，逐漸轉變爲強調學生高層次的知識與問題解決能力的培養，因此，傳統的選擇題式的題型，無法完全的因應所有評量的需求。張麗麗（2002）指出，選擇題型在評量高層次與統整能力上有其困難度。而很多的選擇題目，指將知識進行分割與去情境化，使測驗內容與生活脫節，無法將所學應用在生活中。此外，選擇題的考試著重結果，不重視歷程，使得學生無法在學習歷程中自我建構出有意義的學習，並逐漸成爲被動的學習者。爲了克服這些問題，實作評量逐漸受到測驗界的重視。

實作評量與另類評量（alternative assessment）及真實評量（authentic assessment）十分類似，但另類評量爲廣泛的指有別於傳統的紙筆測驗，真實評量則著重於評量內容於現實生活的結合，實作評量則是重視學生參與建構、並進行實作的過程與結果。實作評量的種類有很多，例如，可以使用具有結構限制的紙筆形式試題，要求學生完成個人或小組的計畫，進行日誌與省思，或由教師對於學生表現進行正式與非正式的觀察記錄，或是藉由科學實驗、展示或檔案評量的方式進行。

Shavelson、Baxter 與 Pine（1992）提出了一些關於實作評量實施的基本原則：

（一）實作評量的內容需要超脫記憶與零碎的片段性知識，或是從多個變項中，選出一個單一正確答案的模式。評量要能夠捕捉學生對於學科知識的理解、認知與展顯問題解決的能力，並能夠鼓勵學生提供新穎、具有創造力的答案。

（二）學生必須依據操弄性與實驗性的教具來回答評量的問題，一些需要動手實作的評量內容必須要客觀且標準化。某些需要長期進行的計畫案或許無法以一個測驗時段來完成。

（三）長時間進行的計畫案與動手做的實作評量是較爲昂貴且費時的，因此若能加上電腦化科技的輔助是較佳的。

（四）實作評量要能夠反映認知方面的發展，特別是要著重於心智方面，以用於評鑑學生的知識架構與了解學生的迷思之處。

（五）實作評量須與課程改革緊緊相扣。因爲測驗分數主要是依據考試內容來做解釋，因此測驗分數應要能在學科的指標中做有意義的解釋。

二、實作評量的步驟

實作評量講求學生透過學生的動手操作，讓學生展現所學來解決問題。Stiggins（1994）建議在建構實作評量的試卷時，應考量下面的步驟：

（一）確定評量的原因與目的

必須先確立爲何要實施實作評量，例如，評量的結果是要進行學生的分組以鑑定個別能力、還是僅供教學的參考。

（二）設定所欲評量的實作表現

依據評量教學中重要的概念與技能，設計出能夠測量學生真實能力的內容，並選擇所要評量的表現，是針對結果、過程還是對過程和結果兼顧來進行評鑑。依據評量領域中所需表現的知識或技能，設定評量表現的標準。

（三）設計實作評量的內容

需決定作業的形式，實施評量的情境，與確定所要編制的試題數量。所選擇的內容應具有代表性。

（四）確定實作評量成績計算的標準

須決定評分的方式，採用分析式評分或是整體式評分，須對評分者進行專業的訓練，並使每個人對於評分歸準有所共識，最後，要決定記錄評分的方法，方法可採用檢核表、評定量表或是軼事記錄表等。

三、國內外相關研究

國內外有許多論文都對數學科的實作評量進行相關研究，且相當著重於量化的實驗研究。詹元智（2002）以小六學生爲對象，以兩個月的時間，採用實驗組與控制組的方式進行研究。對於控制組的學生，進行傳統教學、學生解題與練習。而實驗組則是進行數學實做評量教學以及八次的實作作業練習。教師以生活情境佈題、小組討論的方式，針對學生的問題給予實作評量的練習，並提供課後輔導與回饋。兩組學生在實驗前後須接受筆紙測驗與數學實作測驗。研究發現兩組學生在傳統筆紙測驗上的表現無顯著差異，但是在數學實作評量表現上，接受實作評量教學與練習的實驗組學生表現較佳。

曾安如（2004）則是探討數學日記寫作活動對於數學成就與態度的影響。研究採準實驗設計，實驗組的教學內容包括回憶上課摘要、解決生活情境的應用、對每個數學單元的感想、找出題目

中錯誤的概念並進行澄清，將數學概念融入故事進行寫作等。結果發現實驗組學生在數學成就與數學態度的表現上均優於控制組學生。

Jurdak 與 Zein（1998）以中學生爲對象，探討數學日誌對學生數學成就及學習態度的影響。其準實驗設計中，實驗組在每周數學課結束後進行數學日誌寫作，學生必須完成題目的寫作、閱讀與數學學習有關的內容，並寫出答案。控制組的學生則是進行課內測驗的計算與練習。研究發現實驗組學生在數學理解、程序上的知識與數學溝通上的表現均優於控制組，但是對問題解決、數學態度及在校數學成績上而言沒有顯著差異。

蔡正濱（2006）則探討實作評量中的評分者一致性。採用不同類型的計分歸準、不同複雜度的作業及受試者是否熟悉實作評量計分歸準的經驗三個面向來探討對評分者一致性的影響。研究發現對影響實作評量一致性的主要變異來源爲受試者與作業項目之間的交互作用（*pxt*）變異，其是爲受試者本身的變異（*p*），再其次爲受試者、作業與評分者三者之交互作用（*pxtxr*）。此外，不同的計分歸準對評分的一致的影響不大，高複雜度的作業比低複雜度的作業不易達到一致性，且分析式的計分影響小於整體式的計分歸準。

根據過去的文獻探討，大多數的研究在分析實作評量的數據時，採用古典測驗理論下的模式進行分析，然而，在古典測驗理論下分析所得的試題難度，會因受試者能力分配的不同，而有不一樣的難度參數，同樣的問題也會發生在考生能力值的估計上，能力值的估算會受到所使用的測驗題不同而有所差異。然而，這種問題可以透過試題反應理論的統計方式獲得解決。

四、試題反應理論與多層面 Rasch 模式

試題反應理論（Item Response Theory，IRT）爲近年進行試題分析的主要理論之一，此理論是以個別試題的觀點，來對學生所得到的測驗分數進行解釋，這個理論認爲，學生在試題上答對與否，與其自身所具備的能力或特質息息相關，要使用 IRT 來進行試題分析，則需要先檢驗兩大基本假設。第一個假設爲單向度（unidimensionality），在此假設中，測驗試題必須爲單向度，即爲學生在某一測驗試題上的做答表現，都是單一的能力所主導與決定。也就是說，一份考題，只測量單一的構念。數學能力測驗，所測得的就是數學能力。若數學能力中的應用題文字敘述很多，則語文能力可能會成爲另一個向度，這樣夾雜語文能力的測驗，需使用多向度的試題反應理論（multidimensional item response theory）來進行分析。第二個假設爲局部試題獨立性（local item independence），當影響測驗能力的因素不改變時，同一個考生在測驗題目任兩題的作答反應應該是獨立的。這個假設常常是在回答題組題的時候被破壞，像是閱讀一篇文言文測驗，然後依據這個文章內容來找答案，則往往出現上一題的答案，會影響到下一題的作答。尤其是有些「承接上題」的題目，則是極有可能違反局部試題獨立性的假設。

單向度的試題反應模式有很多種。可先分爲兩大類，第一大類爲二元化記分題（dichotomous scoring），第二大類爲多元化記分（polytomous scoring）。在二元化記分中，學生的作答評分只可能有兩種。第一種爲答對，得分爲 1 分，第二種爲答錯，得分爲 0 分。這種最常出現在選擇題或是非題的給分。而多元化記分，則給分的範圍就不限制在 0 或 1，而是可以有多層次的給分。譬如在實作評量中，我們可以依據學生歌唱的表現給分，總分爲 10 分，給分的層次，可以有 1 分，代表表現得不好，一直給到 10 分，代表表現得非常好。更常見的應用在態度或情意量表上，使用李克式量表的五點記分。分別代表 「非常滿意」、「滿意」、「無意見」、「不滿意」、「非常不滿意」等五種態度的層面。一般研究者常把「非常滿意」當成是 5 分，「滿意」爲 4 分，「無意見」給三分，「不滿意」給 2 分，「非常不滿意」給 1 分等依順序來給分。多元計分的試題反應模組的計算十分複雜, 也牽涉到需多數學的模式。

在二元化計分中，常用的統計模式有潛在線性模式、完美量尺模式、潛在距離模式、常態肩型模式（包括一、二、三參數）、常態肩型模式（包括一、二、三參數）、參數對數型模式（包括一、二、三參數）、四參數對數型模式。而常見的多元化計分模式包括等級反應模式、名義反應模式、評定量表模式與部份計分模式。這些模式的詳細介紹，請詳見余民寧（2009）的專著。

多層面 Rasch 模式（Many faceted Rasch model）從試題反應理論發展而來，爲部份計分模式的延伸。多層面 Rasch 模式可以解決一些傳統古典測驗理論假設下的問題。古典測驗理論下所計算出的試題參數與統計指標（如難度、鑑別度、信度），均有樣本依賴的特性，會因爲測驗的受試者樣本而不同。此外，古典測驗對於測驗功能相似的測驗分數之間，無法提供合理的比較，有意義的比較受限在相同測驗的前後測分數或是平行複本的分數之間（余民寧，2009）。根據 Linacre（1989）指出，多層面 Rasch 模式有多項優點。第一，此模式將所有需考量的層面放在同一個尺度上，例如受試者各組的表現、試題難度、評審者的嚴厲度等。第二，因爲多層面 Rasch 模式建構於試題反應理論下，因此具有試題獨立、與受試者獨立的優點。即使是使用不同的題目、來進行施測，受試者的能力也不會受試題難度而有影響，此外，若要分析試題參數，不同的受試者來進行估計，所得到的試題參數也會維持一致。這對於實作評量的分析特別有利，尤其是在大型測驗上，例如基測的作文評分，不大可能同一個評分者可以評比所有的項目，而不同的評分者，即使受到良好的訓練，也會有個別差異，某些評分者評的較嚴、而某些則評的較鬆，而這些較嚴格或是較寬鬆的評分，會直接影響到受試者的成績，而透過 Rasch 的分析，則能將這種鬆嚴不一的情況考慮進去，並將受試者的能力做合理的推論。第三，Rasch 模組中所提供「適配度」的指標可以用來找出哪些評分者的評分或是那些受試者的作答反應有不一致的情形。例如，若是某一位受試者，所有難的題目都答對，卻有幾題很簡單的題目沒有答對，這種情形下，Rasch 模組可以偵測這些受試者的作答反應，研究者可以找出這些作答者，並以深入了解原因（Engelhard, 1992; Linacre, 1999）。

多層面 Rasch 模式透過統計的模組，來探討試題的難度和受試者能力之間的關係。此模組將試題難度與受試者的能力放在同一個尺度上，並將分數轉化成 logit 分數，logit 分數爲一等距尺度的分數，可以進行統計運算，並能展現受試者能力的強項和弱項，或是對不同的族群進行比較。本研究使用多層面 Rasch 模式來進行分析，有兩大主要原因。其一是由於實作評量牽涉到評分者來進行評分，雖然在本研究所使用的數據中，三位評分者對全部的受試者都進行評分，但是這種情形在一般實務界並不常見。較爲常見的情形是一份試卷，經由兩位評分者來評，如果分數沒有出入，則採用兩位評分者評分的平均來當作受試者的最終分數，相反的，如果差距大於某個指標，則會再找第三位評分者來評分。但是這些評分制度雖然客觀，但還是無法避免某些評分者評分嚴厲，某些則較爲寬鬆的情形。而本研究所使用的多層面 Rasch 模式的方法，則可以將評分者的嚴厲度進行考量，使受試者所得的分數更爲客觀。其二爲在國內用來評估實作評量主要分析方法多局限在古典測驗理論下，較少文獻使用多層面 Rasch 模式來進行實作評量的數據分析，本文主要是探討如何以多層面 Rasch 分析來分析數學實作評量。除了示範分析的結果之外，並詳加解釋相關參數。其結果可供未來實務研究者的參考。

研究方法

以下依研究對象、資料來源、課程設計與統計方法分別論述。

一、研究對象

本研究從桃園縣取樣四、五與六年級，四年級有四個班級，其他年級各三個班級，其中四年級的有 154 位，五年級的有 64 位，六年級的有 96 位，共 314 位學生。本研究取樣的學校爲中壢市市中心的小學，學區屬性較爲中上，四至六年級每學年有十一班，均爲常態分班。

二、研究工具

本研究設計實作評量試題，試題性質爲應用、分析、評鑑與創造等向度。每題均有三到四子題，每一題組以 3 分制進行評分，最高爲 3 分，最低爲 0 分，以進行設計。得 3 分者完全正確，2 分者爲少部份錯誤，1 分等級者爲部份正確，0 分者爲完全錯誤，評分標準如表 1。研究者針對評分標準對三位評分者進行訓練，並使用五份學生試卷作爲練習，以確保評分者完全瞭解評分標準的內容。

表 1　評分標準

分數	評分指標
3	完全正確、清楚 能使用適當的策略與步驟來完成解題 能展現出學生有清楚的概念與解題的邏輯 學生的解釋很清楚 學生能使用圖像或其它方式來表達解題的過程 能正確與合理的回答三個子題問題
2	部份正確、清楚 某些問題能清楚的回答，但並不完整 學生在某些概念是很清楚的，但並不是完全清楚 有些學生的解釋很清楚，但需再多些解釋才會更加完整 學生的圖像或是其它的表現方式，對解題並未有明顯的幫助 有些答案是正確的，但有些答案是錯誤的
1	少數正確、清楚 有些解題的過程省略或是不正確，僅有少部份的答案是正確的 少量的證據顯示學生了解這些概念和步驟 學生的解釋和表現很難理解，有些解題的過程需再多做說明 學生的圖像或是其它的表現方式，對解題可能無關 大多的答案都是不正確的
0	不正確，不完整、不清楚 所有的問題都是錯誤的 學生並未了解關鍵的數學概念 學生的解題策略與方法令人無法理解 學生的圖像或是其它的表現方式，對解題無關 所有的答案都是不正確的

本研究設計四題組，其設計面向有數字的運算、圖表的判讀與製作、時間的量感與數字的關係等不同目標。第一個題組意爲檢視學生的數與計算的概念，要求學生能正確的進行加減法計算，從題目的意義中發現數字運算的基本概念。第二個題組的目的在於讓學生判讀圖表，製作圖表，進而能發現圖表中的問題，如學生可用不同的想法來贊同或反對此一圖表的應用。第三個題組旨在了解學生在時間的量感，要求學生計算時間、畫出時鐘、設計行程表等，了解學生在時鐘的觀念與主動安排行程的規劃能力。第四個題組爲評量學生對乘除法關係的了解，如因倍數的概念與發現乘法與除法的互逆關係等。

各題目間之能力指標對應如下。

題組一：N2-2 延伸加、減、乘、除 與情境的意義，使能適用來解決更多的生活情境問題，並能用計算器械處理大數的計算。

題組二：D2-2 能將分類資料整理成長條圖，並抽取長條圖中有意義的資訊加以解讀。

題組三：N2-8 能報讀（鐘面上的）時刻以及點算兩時刻間的時間；能理解 24 時制並應用在生活中。

題組四：N2-14 能在情境中，理解乘法交換律、等號的對稱性、「<、=、>」的遞移性、加法和乘法的結合律與分配律，以及乘法和除法的相互關係。

本題組設計方式，參考國際數學測驗 PISA（programme for international student assessment），設計數與量、統計圖表、時間與乘除法關係的試題。每一題組設計三到四小題的子題，以開放性問題爲主，對學生的數學認知程度進行了解。

三、多層面 Rasch 分析

本研究所使用的數據包括了三位評分者，針對四年級的 154 位學生、五年級的 64 位學生與六年級的 96 學生進行多層面的 Rasch 分析。主要使用的軟體爲 FACETS（Linacre, 2006），FACET 使用「聯合的最大概似估計法」（JMLE）對模式之各項參數進行估算，在此估計中，考生的能力參數，與其他參數一起進行同步的估算。

多層面 Rasch 模式中，學生在測驗中的答題表現，除了受試題難度和本身能力的影響之外，同時也考量到其他層面的影響性。例如評分者的嚴厲度，因爲在測驗領域中，有可能同一個考生答案，在不同評分者的評分下，還是會得到不一樣的分數。而此模式可以有效的處理試題難度及評分者嚴厲度，以達到公平評估受測者能力的功能。

本研究考量的層面爲受試者能力、試題難度、評分者的嚴厲度與受試者的年級。則第 n 位受試者爲 m 年級的學生，他在第 i 題，被第 j 個評分者評定爲 x 分數之對數勝算比可表示爲：

$$\ln\left(\frac{P_{nijmx}}{P_{nijm(xij)}}\right)=\beta_n-\delta_i-\omega_j-\alpha_m-\tau_x$$

其中 P_{nijmx} 代表受試者得到分數爲 x 的機率；

$P_{nijm(xij)}$ 代表受試者得到分數爲（x-1）的機率；

β_n 爲受試者能力值，；

δ_i 爲第 i 題的試題難度；

ω_j 爲第 j 個評分者的嚴厲度，

α_m 爲受試者所就讀的年級爲 m，

τ_x 爲從分數（x-1）到 x 的所需增加的困難度。

由此模式可見，受試者、試題、年級、評分者都是要考量的層面，而各個層面之間的關係緊緊相扣，受試者的分數是否能夠順利的進級（例如從 2 分進步到 3 分）和本身的能力，試題的難度、所在的年級、與評分者的嚴厲度都有關聯性。此模組將原本單純的計分方式，進而分離各種可能影響得分的因素，使實作評量的評分方式能夠更爲精準。

在每一個層面，FACET 都會估算參數平均值，參數標準差、不同的 fit 參數（如 infit 均方值、outfit 均方值等）。所有經過估算的參數值，均轉換成對數型尺度（logit scale），而其學理上的範圍爲正無限大到負無限大之間。而這些參數的大小估計，不受樣本影響（sample free），也就是這些參數的獲得，不會因爲所選出接受測驗的受試者樣本不同而有所差異。由於多層面 Rasch 模式的估計採用累積機率比，變數均轉換成以 logit 爲單位之等距尺度，所以在變數分布圖上，參數間的大小關係，可以互相進行意義性的比較。一般而言，越高的參數，代表受試者的能力越高，題目越難，評分者越嚴厲（Linacre, 2006）。

不管是多完美的統計模式，實徵數據的資料，也不可能完全符合模式的設計。而在 Rasch 模式中假設所有的得分狀況，主要受到模式中構面的影響，高能力者，在題目中得分應該較高，而低能力者，得分較低。若樣本的作答反應隨機性過高，或是評分者的給分不穩定，則數據結構則

會偏離模式的假設，代表 Rasch 分析不適合分析此數據。在進行 Rasch 模式估計時，要檢視是否資料本身之適配度合乎要求，才能對後續的參數估計有意義的解讀。

因此，檢視數據與模式之間的適配度，是非常重要的一環，適配度的指標，爲受試者在試題上的實際表現與 IRT 模式所估計出的預期表現之間的比較，若觀測值與模式預期值之間的差異性越小，則適配度會越佳。多層面分析中常用來檢視適配度的統計指標爲 Infit 均方值與 outfit 均方值，這兩種指標均用來分析資料上各層面是否偏離預定值之檢測依據。de Ayala（2009）建議這兩種均方值的理想範圍爲 0.5 到 1.5 之間，越接近 1 代表適配度越好。若均方值高過 2，則代表適配度有問題，需要進一步審核試題與受試者能力之間的關聯性。

然而，對於單向度的檢測，Tennant 及 Pallant（2006）認爲 Outfit 或 Infit 的均方誤仍有不足，應使用 Rasch 殘差主成份分析來佐證單向度的證據，依 WINSTEPS 的操作手冊指出，只要解釋變異量大於 60%，第一殘差特徵值小於 3.0 或第一因素殘值變量佔殘差總量 5%內，符合任何一項條件時，表示資料符合 Rasch 模式單向度的假設。

除此之外，進行多層面 Rasch 分析時，亦需考量 Rasch 參數估計的分離性信度係數（The reliabilities of separation coefficients）。此係數的用途和 KR20 與庫李信度用途相似，均爲評估數據的內部一致性，但在解釋上略有不同。這個顯示觀測值之間存在的差異性有多大，當分離性信度越大，代表在同個變數內的觀測值越不同。Linacre（2006）指出，一般而言，對於受試者的能力參數與試題的參數估計，會希望分離性信度係數大（接近 1），因爲代表這個測驗的題目很多元，可以測到各種能力分布的受試者。然而，對於評分者的分離性信度則希望越低越好，因爲接近 0 反倒是代表評分者的評分具有相當的一致性。在多層面分析中，分離度指標（separation indices）更進一步的顯示，觀測值可以被分幾個層次，當分離度指標越大，代表觀測值內部存在的差異性越大，可以分成越多層次。而透過卡方檢定，可以進一步檢視觀測值的差異性有沒有達到顯著。例如，研究者想要知道試題的難度是否一致，則可查看卡方檢定的結果，若卡方檢定不顯著，代表試題間的難度是一致的，而顯著的卡方值則隱含試題間的難度，是存著差異性的。

結果

一、描述性統計值

表 2 所呈現的是三位評分者，在每一個題目上所給的平均成績，例如，對於四年級的學生，第一位評分者在第一題數與量所給的分數之平均爲 2.27 分，而在第二題統計圖表的給分的平均分數則爲 1.27。由表 2 可知，第四題乘除法的難度最高，三位評分者所給的分數在此題都是最低，而第一題是最簡單。而就評分者而言，每位評分者的嚴厲度略有不同，評分的差異性約爲 0.2 分左右，最大的差異爲四年級的第二題，第一位評分者所給分數之平均爲 1.27，而第三位給的平均分數則爲 1.53。而就各年級的表現來說，則符合預期，大體而言六年級在各題的表現優於五年級，而五年級學生表現優於四年級，唯一的例外是五年級在第一題的得分略低於四年級的學生，此外，對於簡單的題目，年級間表現差異較小，而難的題目，差異較大。

表 2　各年級與試題的平均得分表

四年級	題 1	題 2	題 3	題 4
評分者 1	2.27	1.27	0.98	0.45
評分者 2	2.21	1.52	0.97	0.42
評分者 3	2.28	1.53	1.04	0.60
四年級分數平均	2.26	1.44	1.00	0.49
五年級				
評分者 1	2.22	1.52	1.36	0.66
評分者 2	2.17	1.55	1.38	0.69
評分者 3	2.22	1.70	1.55	0.84
五年級分數平均	2.20	1.59	1.43	0.73
六年級				
評分者 1	2.26	1.63	1.37	0.81
評分者 2	2.21	1.77	1.47	0.91
評分者 3	2.36	1.87	1.58	1.04
六年級分數平均	2.28	1.76	1.47	0.92
總平均	2.25	1.57	1.23	0.67

二、多層面 Rasch 分析

本研究爲使用多層面 Rasch 模式，並使用 FACETS 軟體進行分析，由於 FACETS 軟體建基於單向度之假設，所以應先行進行單向度的假設檢驗。首先將數據放置 Winsteps 軟體中，並藉由主成分分析，來檢視是否呈現單向度。依 WINSTEPS 的殘差主成份分析報表顯示，解釋變異量爲 76.9%，第一及第二殘差特徵值(eigenvalue)爲 1.4 與 1.3，第一及第二因素分別只解釋 8.3%及 7.6%的殘值變異量，由於解釋變異量大於 60%、第一殘差特徵值小於 3.0，即表示資料符合 Rasch 模式單向度的測量。

圖 1 爲變數分布圖（variable map），由這個分布圖中可以看出本研究中所考慮各個層面中的相對難易度散布狀況。

圖 1 最左邊的欄位是刻度，單位爲 logit，logit 值越高，代表學生表現越好、評分者越嚴厲、或是題目越難，logit 的單位爲等距尺度，爲一具有連續性、單位又相等的數值。logit 的大小不但能顯示大小的順序，而且數值之間具有相等的距離。第二個欄位是評分者，而第三位評分者的 logit 值最低，代表評分最鬆。而第三個欄位則爲每個試題之間的難度差距估計，第四題關於乘除法的題目最難，所估計的難度值爲 1.8，而第一題數與量最簡單，難度約爲-1.6，此外，第二題統計圖表和第三題關於時間概念題的難度介於第一題和第四題之間，第五個欄位是學生能力的估計分布，學生能力分布約爲常態分布，大多集中於-1 至 1 之間。透過多層面 Rasch 分析，所有的變數都轉換成同樣的尺度與單位，彼此相互的對應關係可以直接從圖中進行檢視，由此圖中可看出很少人得到滿分，因爲第四題的難度接近 2，已經超過大多數學生的能力範圍，代表大多數受試學生對於乘除法的相關概念感到困難。

表 3 爲各層面分析參數整理，總共探討三個層面，並將評分者與試題估計的難度固定到 0，並觀察受試者能力的估計值。由此可知，此份試題對於受試者而言，是稍感困難的，受試者的平均能力的估計參數爲 0.12，而就適配度 infit 與 outfit 來看，每個層面的適配度都在 0.5~1.5 的範圍內，代表使用的 Rasch 模式來進行估計應該是適合的。此外，每個層面的信度值都高達 0.8~1 之間，代表此數據在模組估計上相當穩定。然而，評分者的分離性信度高達 0.94 也代表評分者的評分具有相當的大的異質性。

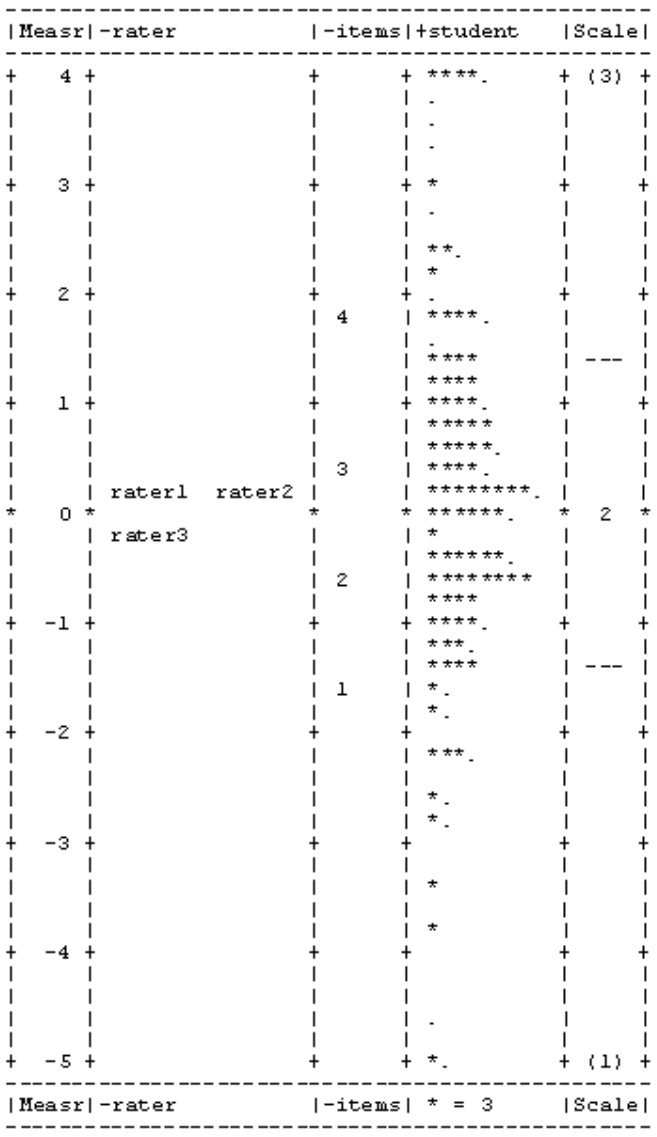

圖 1　變數分布圖

分離度爲一種依據假設所算出統計指標，而此假設爲所有的觀測值是從一個常態分布的母群所隨機抽取出來的，而此母群中的統計特徵和觀測值完全相同。分離度指在這種特性之常態分布母群中，可以分辨出幾種具有統計顯著差異性的類群（strata）。分離度（separation）越高，則代表越能將層面的類別區隔出來，如表 3 所示，試題的分離度爲 20.7，若是有一個與本研究所探究的試題參數相似的常態分布母群，其試題的難度至少可以被分成 20 個層級，也就代表難度差異性很大，有些題目很難、有些題目很容易。 然而，相較之下，受試者的相似度較高，但也至少可以區分爲 2 類程度的受試者。 而值得注意的是評分者也出現差異性，分離度爲 3.95，而卡方檢定亦達顯著，代表不同的評分者之間，評分的嚴厲度具有顯著性的不同。

表 3 各層面估計概況整理

	評分者	試題	受試者
Rasch 參數			
平均	0	0	0.12
標準差	0.25	1.52	1.89
N	3	4	313
Infit			
平均	1.08	1.10	1.07
標準差	0.08	0.11	0.64
Outfit			
平均	1.11	1.13	1.09
標準差	0.06	0.13	0.82
信度	0.94	1	0.81
分離度	3.95	20.7	2.04
卡方值	33.2*	1186*	1859.4*

＊代表 $p < 0.05$

表 4 進一步將所有層面的 Rasch 估計值陳列，其中參數平均值以 logit 爲單位，爲等距尺度的單位，可進行參數之間距離的比較，其中第一位評分者的嚴苛度最高，其估計值爲 0.18，高於第二位和第三位的 0.11 和-0.28，就試題難度而言，與圖 1 結果相呼應，第四題的難度最高，難度值遠高於其他三題，相對而言，第一題和第二題的難度較低，是屬於比較簡單的題目。 每一個層面的 Infit 與 outfit 均方值都座落在 0.5 與 1.5 之間，代表使用 Rasch 的模式來進行分析應該是適當的。

表 4 各層面模式估計結果與配適狀況

	參數平均值	參數平方差	Infit 均方	Outfit 均方
評分者				
評分者 1	0.18	0.06	0.96	1.03
評分者 2	0.11	0.06	1.13	1.12
評分者 3	-0.28	0.06	1.14	1.19
試題				
題 1	-1.67	0.09	1.25	1.32
題 2	-0.63	0.07	0.96	0.99
題 3	0.41	0.06	1.06	1.03
題 4	1.89	0.07	1.13	1.18

表 5 所呈現的是三位評分者，在各題評分的標準化殘差值分布狀況。標準化殘差值所代表的爲評分者所給定的分數，和多層面 Rasch 模式之下所估計出的期望評定分數之間的差異。兩者的值差異越大，代表評分者所給定的評分和依據 Rasch 模式所估算出的評分越不一致。當標準化殘差值大於 0，代表這位評分者，給這位受試者的分數高於模式所預期的分數，也就是給分過高，若是低於 0，則代表這位評分者給分低於模式所預期分數，也就是有給分過低的現象。當標準化殘差值差距大於 2 或小於-2，稱此評分爲非預期的觀測值（unexpected observation）。由表 5 可知，評分者在第四題中，有較多給分過高的情形出現，而在第一題則傾向給分過低，其中尤其是第三位評分者，在第四題的評分其殘差值大於 2 的比率高達 0.1，代表這位評分者在這一題給分過高的出現情形較多，須進一步檢視是否因爲此評分者對於評分規準不清楚或是粗心評分所造成的，但就表 5

所呈現，標準化殘差值的分布情形大致爲隨機分布，並無明顯的系統性誤差存在，此外，殘差值大於 2 或是小於-2 的比率大致分布在 0.05 左右，應屬合理範圍之內。

表 5　評分者之標準化殘差值分布情形

殘差值	評分者 1				評分者 2				評分者 3			
	題 1	題 2	題 3	題 4	題 1	題 2	題 3	題 4	題 1	題 2	題 3	題 4
5	0	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	2	0	0	0	1
3	0	0	3	1	0	0	1	2	0	1	0	4
2	2	5	7	11	1	6	7	13	0	4	8	25
|殘差| < 2	154	274	277	295	144	260	269	290	149	241	270	277
-2	2	2	0	0	6	1	0	0	4	7	1	0
-3	2	1	0	0	3	0	0	0	2	0	0	0
-4	1	0	0	0	1	0	0	0	1	0	0	0
|殘差值| > 2 之比率	0.04	0.03	0.03	0.05	0.07	0.03	0.03	0.06	0.04	0.05	0.03	0.10

而透過標準化殘差分布圖可以進一步檢視評分者對每一位學生在實際給分與模式預估的差距性。如前所述，給分的殘差值在-2 至 2 之間代表觀測到的值屬於正常可接受的範圍。但大於 2 或小於-2，則代表此給分超過預期，需要進一步找出造成此現象的原因。例如，圖 2 爲評分者 3 在第一題的殘差分布圖，若單單檢視五年級 64 位學生的給分情形，則可以看評分者 3 給分時，對於第 3、4、5、7、8、37、50、55、56 這幾位學生的評分給分的殘差值低於-2，研究者可以針對這幾個學生的評分狀況進行深入檢視。就第四位學生來說，其殘差值爲-3.79，評分者給的分數是 2 分，而模式所預估的分數爲 3.75 分，代表此評分者在這一題的評分上，對這位學生的評分可能有過低的現象。而深入進行探究，發現這一位學生在第二題的得分爲 3 分，第三題的得分爲 4 分，第四題的得分爲 3 分，而試題分析顯示，第四題的題目最難，其次爲第三題、第二題，第一題的題目是最簡單的，然而，這位學生在第一題的得分卻低於其他的題目。這種難題得高分、簡單題卻得低分的不合理現象，是主要造成殘差值過大的原因。研究者可以進一步深入分析，是否這樣的現象，是由於評分者誤判，還是純粹由於受試者粗心作答造成的。殘差分布圖可以清楚的檢視評分者評分上的誤差情形，也能供給評分者，作爲修正評分的參考。

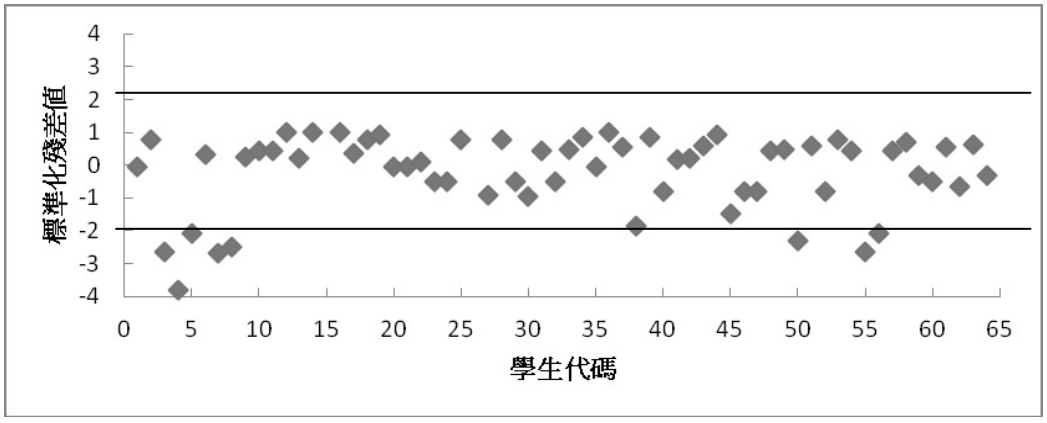

圖 2　評分者 3 在題一之殘差分布圖（五年級）

結論與建議

一、結論

本研究使用試題反應理論中發展出來的多層面 Rasch 分析，來剖析在實作評量中所獲得的實徵數據。綜合本研究的實徵研究所獲得之結果，有以下幾點結論：

（一）就本研究所測的四個數學概念來看，第四題關於乘除法的題目最難，而第一題數與量最簡單，此外，第二題統計圖表和第三題關於時間概念題的難度介於第一題和第四題之間。

（二）多層面 Rasch 分析，除了具有 IRT 分析方式的優點之外，更能進一步考量影響測驗結果的相關因素，並將這些研究者所需考量的因素放進模組中。例如評分者嚴厲度等，透過分離這些因素的影響，客觀的估算受試者之能力表現。

（三）根據實徵研究的結果，試題難度、評分者的嚴厲度、受試者能力表現等構面之適配度均佳，顯示此實徵數據具有穩定性，使用多層面 Rasch 分析應屬恰當。此外，透過 Rasch 模式所估計出的參數值，也具備進一步比較分析之統計推論的特性。

（四）在評分者之評分嚴厲度分析中發現，評審委員之評分嚴厲度的適配性符合 Rasch 模式之假設，代表整體評分之專業性應值得肯定。然而，評分者的高分離度指標顯示不同評分者的嚴厲度有差距，其可能原因爲在本研究中，規準是單一的整體性標準，並沒有依據各題訂定細部的評判標準；而在各題組織間，特質差異大，且各種答案的可能組合不少，增加評判的難度，因而產生主觀性的評分不一致的現象，應訂定題組個別的評分標準，或許能改善評分者信度。

（五）在大型測驗中，評分者若要評閱多份試卷，常有可能發生失誤的現象，透過殘差分布圖，研究者可以清楚的檢視評分者評分上的誤差情形，此圖也能供給評分者，作爲修正評分的參考。

二、建議

由理論和實徵數據顯示，多層面 Rasch 模式在實作評量應用上爲一種有效的方式，且提供研究者不同面向的訊息。對於未來的研究方向，有兩點建議：

（一）本研究將多層面 Rasch 分析應用在實作評量上，然而，此理論除了可以加以應用在具有評分機制的筆紙測驗之外，在其他學術領域上的應用值得更加推廣與發展。其他像是體操選手的評分、或是音樂比賽的評分也可加以應用。

（二）透過多層面 Rasch 分析中發現評分者在評分中出現差異頗大的情形，代表評分者在評分時具有相當的主觀性。然而，未來研究者可以進一步推敲其成因爲何？是評分歸準設定的不夠清楚？還是單純是因爲受試者粗心或是評分者的誤評，所造成之隨機誤差？研究者可以透過質性訪談，了解原因，並用以精進評分人員的訓練課程與改善現有評分歸準之依據。

參考文獻

王文中（2004）：Rasch 測量模式與其在教育與心理之應用。**教育與心理研究，27**（4），637-694。[Wang, W. C. (2004). Rasch Measurement Theory and Application in Education and Psychology. *Journal of Education & Psychology, 27*(4), 637-694.]

余民寧（2009）：**試題反應理論及其應用**。台北：心理。[Yu, M. N. (2009). *Item response theory*. Taipei, Taiwan: Psychological Publishing.]

姚漢禱、姚偉哲（2007）：應用多層面 Rasch 模式分析雙不定向飛靶優秀選手的射擊技術。**測驗學刊，55**（1），89-104。[Yau, H. D., & Yao, W. C. (2007). Application of Many-Facet Rasch Model to Analyze the Skills of Elite Athletes in Double Trap. *Psychological Testing, 55*(1), 89-104.]

張麗麗（2002）：從分數的意義談實作評量效度的建立。**教育研究月刊，98**，37-50。[Chang, L. L. (2002). Establishing the Validity of Performance Assessment from the Meaning of the Testing Scores. *Journal of Education Research, 98*, 37-50.]

張新立、吳舜丞（2008）：多層面 Rasch 模式於學術研討會論文評分之應用。**測驗學刊，55**（1），105-128。[Chang, H. L., & Wu, S. C. (2008). A Multi-Facet Rasch Analysis on Rating the Academic Scientific Papers. *Psychological Testing, 55*(1), 105-128.]

曾安如（2004）：**國小二年級學童數學寫作活動、數學成就與數學態度之相關研究**。國立台中師範學院教育測驗統計研究所碩士論文。[Tseng, A. R.(2004). A Study of the Relationship between mathematical writing activity, mathematics achievement and mathematics attitude for second Grade Students (Master's thesis). National Taichung University of Education.]

詹元智（2002）：**國小數學科實作評量之效度探討**。屏東師範學院教育心理與輔導研究所碩士論文。[Zhan,Y. C.(2002). *Validity of Mathematics Performance Assessment* (Master's thesis). National Pingtung University of Education.]

蔡正濱（2006）：**國小數學科實作評量評分者一致性相關因素探討**。國立屏東教育大學教育心理與輔導學系碩士論文。[Tsai, C. P. (2006). *Factors Influencing Rater Consistency on a Mathematics Performance Assessment* (Master's thesis). National Pingtung University of Education.]

藍珮君（2012）：以多面向 Rasch 測量模式分析 TOCFL 口語測驗評分者訓練效果。**永續教育發展-創新與實踐論文集：2010 年國際學術研討會-測驗及評量論文專輯**，125-139。新北市：國家教育研究院。[Lan, P. J. (2012). Using many-facet Rasch measurement to examining rater training effects of TOCFL Speaking. In 2010 NAER Conference: Education for Sustainable Development-Innovation and Implementation (125-139). New Taipei City, Taiwan: National Academy for Educational Research]

Basturk, R. (2008). Applying the many-facet Rasch model to evaluate powerpoint presentation performance in higher education. *Assessment & Evaluation in Higher Education, 33*(4), 431-444.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Engelhard, G. J. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences.* New York, NY: Basic Books.

Jurdak, M., & Zein, R. A. (1998). The effect of journal writing on achievement in and attitudes toward mathematics. *School Science & Mathematics, 98*(8), 412-419.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, J. M. (2006). *Winsteps: Rasch model statistical software.* Chicago, IL: MESA.

Linacre, J. M. (2006). *FACETS: Many-facet Rasch measurement computer program.* Chicago, IL: MESA.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*, 435-444.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge seeerity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Lunz, M. E., Wright, B. D., Stahl, J. A., & Linacre, J. M. (1989). *Equating practical examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessment: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many facet Rasch measurement using a complex problem solving skills assessment. *Educational and Psychological Measurement, 64*(4), 617-639.

Stiggins, R. J. (1994). *Student-centered classroom assessment.* New York, NY: Macmillan.

Tennant, A., Pallant, J. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions 2006, 20*(1), 1048-1051.

Twing, J., & Williams, K. T. (1992). *An investigation of writing assessment using a many-faceted Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

收 稿 日 期：2011 年 11 月 04 日
一稿修訂日期：2012 年 04 月 10 日
二稿修訂日期：2012 年 10 月 25 日
三稿修訂日期：2012 年 11 月 01 日
接受刊登日期：2012 年 11 月 01 日

Bulletin of Educational Psychology, 2013, 45(1), 1-18
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

An Application of Many-Facet Rasch Model to Evaluate Mathematics Performance Assessment

Ju-Shan Hsieh
National Taiwan University of Arts
Teacher Education Center

Ming-Chuan Hsieh
National Academy for Educational Research
Research Center for Testing and Assessment

The purpose of this study is to evaluate student's potential ability on mathematics ability using Many-facet Rasch Model. The test students took consisted of four ranking levels. Three hundred and fourteen elementary students, and three raters were participated in this study. The results show that even with training and delineating a standard for grading, there remained differences in grader severity among the raters. Many-facet Rasch analysis enabled calibration of grader severity for researchers to detect grader irregularities. Data showed that the standard residuals were randomly distributed, indicating that there were no obvious systematic errors. Overall, the study show Many-facet Rasch model can be quite a useful analytic tool in performance assessment.

KEY WORDS: many- facet Rasch, performance assessment, rater reliability