

應用 HIRT 於實徵資料分析 －以國小六年級數學小數的除法單元為例

張素珍 李佩瑾 郭伯臣 林佳樺

國立臺中教育大學教育測驗統計研究所

摘要

近年來測驗的編製逐漸朝向階層式的評量架構，期能測得學生較高階的能力值，加上使用適當的模式進行分析，不僅可同時獲得欲求得的能力估計值，還能獲得較多的訊息量。本研究根據 NAEP 的數學評量架構，編製一份六年級數學小數的除法測驗，利用實徵資料比較 HIRT、MIRT 和 UIRT 三種估計模式對學生小數除法的能力值估計是否有差異，影響如何？以作為階層式評量架構之測量模式選用依據。

本研究測驗的結果顯示此份測驗具備良好的信度（Cronbach α 係數值為 0.79）、效度和鑑別度；透過實徵方式在模式適配度指標（model fit indexes）AIC、BIC、DIC 三種指標值都顯示 HIRT 之完整模式較適合用於估計階層式評量架構；HIRT 模式中小數的除法能力及概念、程序、問題解決間的迴歸係數在 0.8 以上，尤其概念在解釋小數的除法能力時影響力最大；HIRT 之完整模式也可以提供較多訊息，且估計效果較佳。

關鍵字：階層式試題反應理論、小數的除法、NAEP

Applied HIRT to Empirical Data for Sixth Grade Students- Using Decimal Division as An Example

Su-Chen Chang Pei- Jin Li Bor-Chen Kuo Chia-Hua Lin

Graduate Institute of Educational Measurement and Statistics,
National Taichung University

Abstract

In recent years, for estimating students' higher abilities, the framework of assessment gradually turns into large-scale standardized assessment framework. Suitable model not only tells us the ability estimates wanted, and gets the better estimation result. By means of empirical study, the main purpose of the study is to compare if there is difference in mathematical ability estimation by HIRT(hierarchical item response theory), MIRT (multidimensional item response theory) and UIRT (unidimensional item response theory) and what their influences are as the reference of mathematical assessment model. The assessment on Decimal division is designed for six-grade students based on the mathematical assessment framework of NAEP.

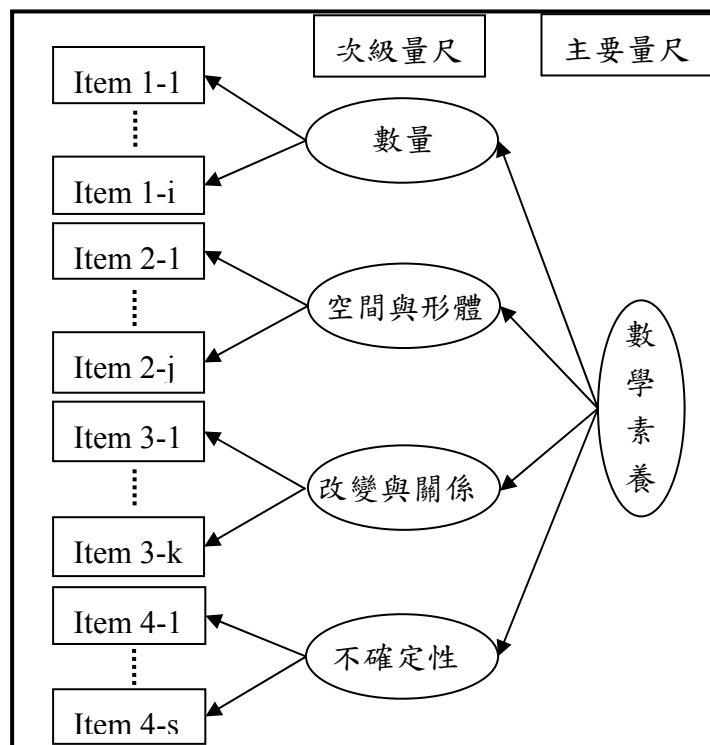
The reliability on the assessment is 0.79. The result is analyzed and compared by HIRT, MIRT and UIRT models. According to the model fit indexes (AIC, BIC and DIC), it shows that HIRT model is suitable to large-scale standardized assessment framework. In HIRT pattern, the coefficients of Decimal division, and conceptual understanding, procedural knowledge, problem solving inference regression are higher than 0.7, especially conceptual understanding influence the Decimal division. Therefore, the result of the empirical study confirms HIRT model can provide more information and has better estimation.

Keywords: hierarchical item response theory, decimal division , NAEP

壹、緒論

許多先進國家的教育系統深切的關注學生基本能力的表現，為了提升國民的素質及國際競爭力，都提供了許多良好的評量架構範例，如：國家教育進展評量（The National Assessment of Educational Progress，簡稱 NAEP）、國際學生評量（The Programme for International Student Assessment，簡稱 PISA）和國際數學趨勢研究（The Trends in International Mathematics and Science Study，簡稱 TIMSS）（林佳樺，2009）。這些大型測驗通常都會有總測驗的分數（主要量尺，overall ability）及分測驗的分數（次級量尺，domain ability），皆屬於階層式測量模式，但這些大型標準化測驗，卻是使用單向度試題反應理論（unidimensional item response theory，UIRT）或是多向度試題反應理論（multidimensional item response theory，MIRT）等測驗模型，可能會因違背其假設而使主要量尺估計不準確，或是當次級量尺所對應的題數較少時，會造成估計效果不可靠，因此 Song（2007）提出單因子階層試題反應理論（one-factor higher-order item response theory，HO-IRT）模式，此模式同時包含主要量尺分數與次級量尺分數，可視為階層式的一般化模式，故亦稱為階層式試題反應理論（hierarchical item response theory，HIRT）模式。

為因應較複雜之評量架構，林佳樺（2009）探討適用於階層式評量架構之測量模式，以 PISA 之評量架構作為基礎，設計階層式試題反應理論模式，其模式可以同時估計主要量尺及次級量尺，稱為完整估計，且估計誤差皆接近或優於 PISA 之估計方式，如圖一。但其研究使用模擬資料進行研究，故本研究將以實證資料來探究其提出之完整模式應用於實際測驗上是否也有相同的成效。



圖一 PISA 數學科評量架構（引自林佳樺，2009）

此外，在國小的數學課程中，小數是數學內容裡「數與量」的一環，是整數十進數結構的延伸和分數的另一種表示方式，其學習十分抽象與複雜的，在學習上學生也常遭遇困難。最近十多年的一些研究結果和評量報告顯示，學生在學習小數方面表現的並不理想（吳昭容，1996；陳永峰，1998；劉曼麗，2004）。

基於上述動機，本研究之研究目的如下：

- 一、依據 NAEP 數學評量架構，編製一份小數的除法測驗，並驗證其成效。
- 二、HIRT 模式探討國小六年級學童小數除法之能力和次級量尺間的關係。
- 三、驗證 HIRT 模式是否可以應用於小數的除法測驗。
- 四、檢驗 HIRT、MIRT 及 UIRT 三種模式之適配度，作為模式選用之依據。

貳、文獻探討

一、NAEP數學評量架構

NAEP是美國之「國家教育進展評量委員會」，其成立目的在瞭解學生學習進展情況，藉以促進教育改革與課程教學革新，並提供了解影響教育表現之因素。NAEP為建置一個有意義的數學評量，將內容分為五大領域，分別為「數的概念、性質與運算」、「測量」、「幾何與空間」、「資料分析、統計與機率」、「代數與函數」（NAGB, 2002）。

NAEP（2003）將數學能力劃分成三個向度，分別為內容向度（content strands）、數學能力（mathematical abilities）與數學力（mathematical power）。其中，數學能力分為三種類型：

（一）概念的了解（conceptual understanding）

1. 能辨認、歸類、產生概念的例子及非例子。
2. 能使用相關的模式、圖表、操作方法，及改變概念的表現方式。
3. 辨認和應用原理原則。
4. 能知道及運用事實及定義。
5. 能比較、對照及整合相關的概念及原理原則，以擴展原有概念及原理原則。
6. 能辨認、解釋及應用來表示概念的符號及術語。
7. 能詮釋在數學情境下相關概念的假設和關係。

（二）程序性知識（procedural knowledge）

1. 正確的選擇和應用程序。
2. 使用具體的模式或象徵性的方法證明程序的正確性。
3. 擴展或修正程序以處理問題情境中原有的因素。

（三）問題解決（problem solving）

1. 能以確認及規劃解決問題。
2. 決定資料的充分性及一致性。
3. 能使用策略、資料、模式及相關的數學。
4. 產生、擴展或修正程序。
5. 在新的情境中能推理。
6. 判斷結果的合理性及正確性。

另外，數學力分為三種類型：

（一）推理（reasoning）

（二）溝通（communication）

（三）連結（connections）

本研究將以數學內容—小數除法，作為主要量尺來編製一份測驗，將數學能力分為概念、程序、問題解決三個次級量尺，以 HIRT、MIRT 及 UIRT 模式探討其間的關係。瞭解學童在小數除法測驗中數學能力的表現情形及測驗中次級量尺—概念、程序、問題解決對主要量尺—小數除法的影響情形。如此可更完整的知道學童在整份小數的除法測驗中的表現。

二、小數的除法概念和教材分析

小數概念及四則運算的學習是國民小學數學的核心課程之一，學生學習小數概念及運算的發展順序是，首先學習整數的概念及四則運算，接著擴展至分數概念及相關運算，然後才發展小數意義及四則運算。雖然小數概念發展在分數之後，但由於小數系統與整數的表徵與十進制結構類似，且社會生活中，使用小數的機會遠大於分數，也就是說，學生在未來生活中使用小數的機會遠超過分數，且小數概念也是建立數學知識的重要基礎，因而學生發展小數概念與運算越顯重要。在生活情境中，因使用整數為多，小數的生活情境較少，也會影響學生對小數的了解，尤其是小數四則運算的學習，學生偏向算則，無法賦予小數四則運算有意義的學習。例如小數的加減是要對齊小數點，而非末位；小數的乘法卻是對齊末位，而非小數點，以及小數的乘法是將它視為整數乘法後再計數被乘數及乘數小數位數的和作為積的小數位數。因此，小數的學習是小學數學教育中，極有挑戰性的教學主題。

九年一貫課程（教育部，2003）中，國民小學階段數學領域的目標（三），規定在小學畢業前，學生能熟練小數與分數的四則計算；能利用常用數量關係，解決日常生活的問題。92課綱中，數學領域的分年細目能力指標，也指出學生要能用直式處理除數為小數的計算，並解決生活中的問題。所以，小數的除法在國小的數學領域中，也算是重要的一環。

國小六年級學童的小數知識，其內容包括：小數的概念、小數的計算和小數的應用等三大部分。

- （一）小數的概念：包括小數意義、化聚、位值、比較大小、稠密性、度量衡單位換算和分數的關係。
- （二）小數的計算：包括小數加法、減法、乘法和除法。
- （三）小數的應用：包括小數加法、減法、乘法和除法文字題。

在六年級的階段中，小數內容主要著重在計算部分，在量的情境下陸續進行乘數、除數是整數的小數乘除計算及乘數、除數是小數的小數乘除計算。吳昭容（2003）認為，運算背後的原理其實攜帶著豐富的內涵，所以解題流程的每個步驟都是有意義的，不能單靠硬背。因此，概念性知識與程序性知識之間必須有所聯繫，了解意義後再熟練計算，如此才能真正地減輕計算時的認知負荷。劉曼麗（2002）在針對台灣地區國小學童所做的小數概念研究中指出：在教導學生程序性知識之前，應讓學生先建構相關的概念性知識，讓小數的概念更穩固、更紮實，如此一來，學生才能正確的運用小數的運算符號。然而，許多研究指出學童普遍缺乏小數概念，例如：學童對於小數的位值缺乏理解、欠缺小數符號之語意意義。

陳永峰（1998）研究發現學童在小數的表現上並不理想，他們所獲得的小數知識似乎都偏向程序性的了解或以記憶性的居多。學童常犯的錯誤在概念部分有：「乘會變大，除會變小」的迷思概念，以及用「大的數」÷「小的數」來解題；在計算部分有：求餘數問題中常以四捨五入法求商，在有餘數的除法中，會忽略餘數的小數點，或是將餘數的小數點對齊移位後的被除數小數點。劉曼麗（2008）

研究也發現學生不易掌握除數為純小數的除法意義，且容易受到「大的數÷小的數」或「乘變大、除變小」迷思概念的影響，因而在除法文字題的解題列式上產生困難。此外，學生在處理小數除法計算問題時，容易在商和餘數小數點的處理上產生錯誤。賴文溥（2009）的研究中發現學童在進行小數除以小數之計算時，小數單位轉換策略中，對商和餘數分別代表的意義不清楚，也就是對餘數及商之單位解讀產生錯誤；在處理沒有餘數的題目時，表現比有餘數的題目還要來得好，顯示許多學童對於小數點位置之意義沒有完全瞭解；在處理小數除法時，對於「位值」及「不夠除時，商要先補零」的概念不清楚。本研究將以國小六年級小數的除法單元為主，針對學童在概念的了解、程序性知識及問題解決上，常會出現的迷思，設計一份階層式之評量架構的測驗。

三、試題反應理論模式

（一）UIRT模式

試題反應理論中的各種模式有個最常用的共同假設，那就是測驗中的各個試題都在測量同一種共同的能力或潛在特質；這種單一能力或潛在特質（因素）必須包含在測驗試題裡的假設，便是單向度的假設。適用於含有單一主要因素測驗資料的試題反應模式，便稱作UIRT模式；適用於含有多種主要因素的試題反應模式，就叫作MIRT模式，透過試題反應理論的理論基礎可發展各領域之題庫、不同測驗間的等化、適性化測驗的建置及發展特定目的之測驗。有關IRT的研究，國內外都有許多相關之應用，如美國的TOEFL、TOEIC、GRE測驗和我國的國中基測。此外，ETS、PISA、NAEP、TIMSS與PIRLS等知名的測驗，亦使用IRT之相關理論作為其學術研究的主要工具之一。本研究只使用單參數對數模式，亦即Rasch模式（Rasch, 1960）。其模式定義如下：

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}} \quad i=1,2,3,\dots,n \quad (1)$$

其中， $P_i(\theta)$ ：能力為 θ 之受試者，答對第 i 題的機率

b_i ：第 i 題的試題難度參數

n ：測驗長度

（二）MIRT模式

MIRT模式可測量多個不同能力量尺的模式，對不同能力之次級量尺進行估計。多向度測驗可以分為題間多向度測驗（between-item multidimensional test）與題內多向度測驗（within-item multidimensional test）兩種（Adams, Wilson & Wang, 1997）。在測驗中的每一試題只測量一種能力，即為單向度的試題，若整份測驗包含多個測量不同能力的單向度試題，則稱此測驗為題間多向度測驗，如國中基本學力測驗中的自然科，測量了包含物理、化學、生物、地球科學等學科能力；在測驗裡的每一試題不只測量單一種能力，也就是試題內包含多個向度，稱此測驗為題內多向度測驗，如數學成就測驗裡的情境題，一個題目不僅測量了問題表徵能力還有計算能力。王敏嫻（2010）利用TASA 2006數學科實證資料，以MIRT探討不同相關程度背景變項以及不同模式下，對於群體參數估計之影響。邱美珍（2008）以多向度試題反應理論模式，探討兒童獨自步行上下學之能力。

MRCMLM（multidimensional random coefficients multinomial logit model）是延伸Rasch模式而成之MIRT模式（Hoskens & De Boeck, 1997; Wang, Wilson, &

Cheng, 2000; Wilson & Adams, 1995)，如PISA數學能力之測量模式即是使用MRCML模式。其模式定義如下：

$$P(X_{ik} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}'_{ik} \boldsymbol{\theta} + \mathbf{a}'_{ik} \boldsymbol{\xi})}{\sum_{k=1}^{K_i} \exp(\mathbf{b}'_{ik} \boldsymbol{\theta} + \mathbf{a}'_{ik} \boldsymbol{\xi})} \quad (2)$$

其中， X_{ik} ：受試者之作答反應組型

K_i ：第 i 試題的計分類別數

$\boldsymbol{\theta}$ ：受試者的能力參數矩陣（多向度能力）

$\boldsymbol{\xi}$ ：試題參數向量

$\mathbf{a}_{ik}$ ：第 i 題中第 k 個反應類別的設計向量（design vector）

$\mathbf{b}_{ik}$ ：第 i 題在第 k 個反應類別上的計分向量（scoring vector）

$\mathbf{A}$ ：整份測驗的設計矩陣（design matrix）

$\mathbf{B}$ ：整份測驗的計分矩陣（scoring matrix）

（三）HIRT 模式

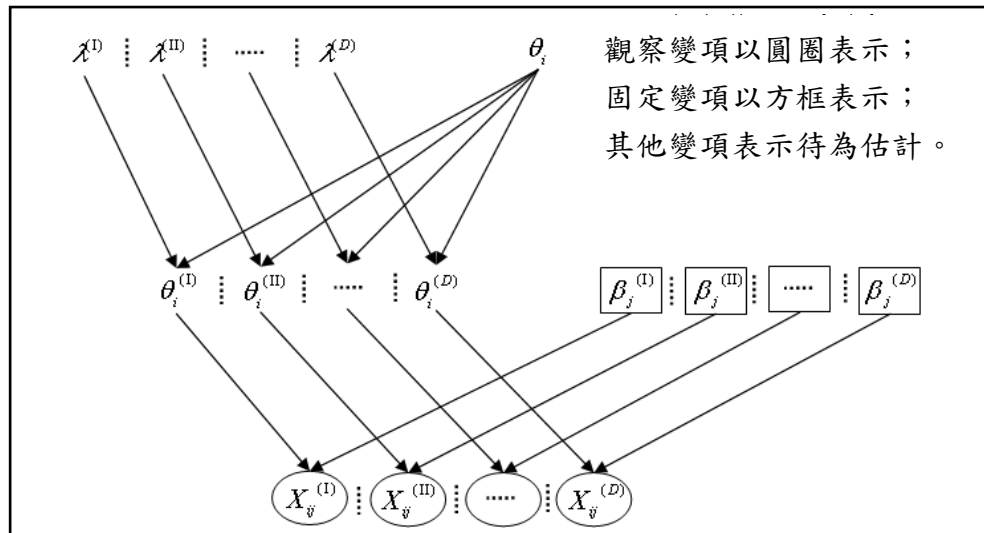
HIRT 模式，包含兩階層的能力量尺，第一層的能力量尺是測量學生在不同指標下的能力表現，稱為次級量尺；第二層的能力量尺是整合次級量尺預測量之高階的學科能力，稱為主要量尺。

測驗主要可觀察多個單向度的子測驗（引自林佳樺，2009），即次級量尺 $\theta_i^{(d)}$ ， $\theta_i^{(d)}$ 表示第 i 位受試者在次級量尺 d 的表現，其中， $d = 1, 2, 3, \dots, D$ 。當不同次級量尺測量相同的能力時，則整份測驗被認為是單向度的測驗，而不同次級量尺間有關聯，則會藉由一主要量尺 θ_i 來連接這些次級量尺，其中 θ_i 為第 i 位受試者在主要量尺的表現，其中並假設次級量尺為主要量尺的一線性函數：

$$\theta_i^{(d)} = \lambda^{(d)} \theta_i + \varepsilon_{id} \quad (3)$$

其中， $\lambda^{(d)}$ 為迴歸參數， ε_{id} 為誤差項， ε_{id} 假設服從平均數為 0 且變異數為 $1 - \lambda^{(d)2}$ 的常態分配，其中 $|\lambda^{(d)}| \leq 1$ 。根據這些假設可得知 $\theta_i^{(d)}$ 的分配與 θ_i 相同屬於標準常態分配 $N(0, 1)$ 。此外，假設已知主要量尺，則次級量尺間會互相獨立；其中 $\lambda^{(d)}$ 更可表示主要量尺與次級量尺間的相關，而次級量尺 d 與 d' 間的相關則為 $\lambda^{(d)} \times \lambda^{(d')}$ 。雖然 $\lambda^{(d)}$ 可為負數，但在教育測驗的應用上主要量尺及次級量尺間的相關皆為正的。林佳樺（2009）以模擬實驗方式，探討 HIRT 之完整模式估計方法與 PISA 中所使用之分開估計方法的成效差異。張勝凱（2010）使用 HIRT 模式建立國小六年級數學推理能力測驗，並驗證 HIRT 模式可應用於數學推理能力測驗。

圖二為該模式的模式圖，第一層表示第 i 位受試者在次級量尺 d 中的第 j 題試題之反應情形 $X_{ij}^{(d)}$ ，第二層表示受試者的反應透過 IRT 模式中的試題參數 $\beta_j^{(d)}$ 連結到次級量尺，其中試題參數 $\beta_j^{(d)}$ 假設為已知（Song, 2007），第三層表示受試者的次級量尺透過迴歸參數 $\lambda^{(d)}$ 連結到相對應之主要量尺 θ_i 。



圖二 應用於一個 D 維度的測驗 (Song, 2007)

參、研究方法

一、研究設計

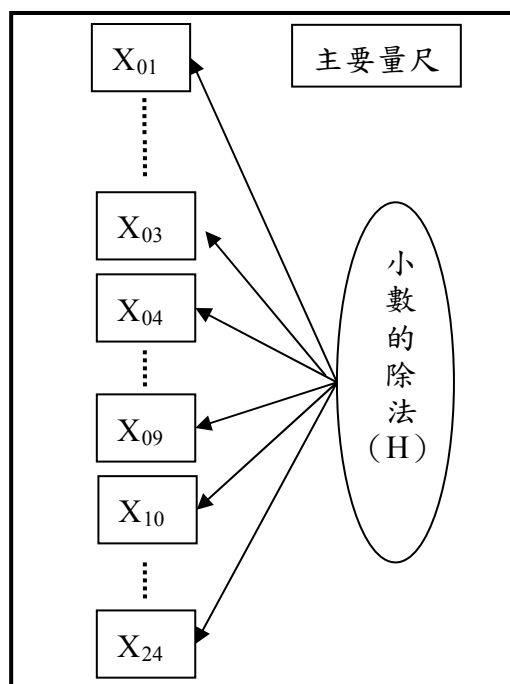
本研究自編一份國小六年級學童小數的除法測驗，並使用林佳樺 (2009) 之 HIRT 模式，透過實徵資料方式探討完整估計與分開估計之估計效果，並以 Akaike's information coefficient (AIC)、Bayesian information coefficient (BIC) 與 deviance information coefficient (DIC) 來檢驗 HIRT、MIRT 和 UIRT 三種模式之適配度，作為模式選用之依據。用統計軟體進行各階層之分開估計及兩階層之完整估計，比較 UIRT 和 MIRT 的分開估計與 HIRT 完整估計的成效。

(一) 自編小數的除法測驗

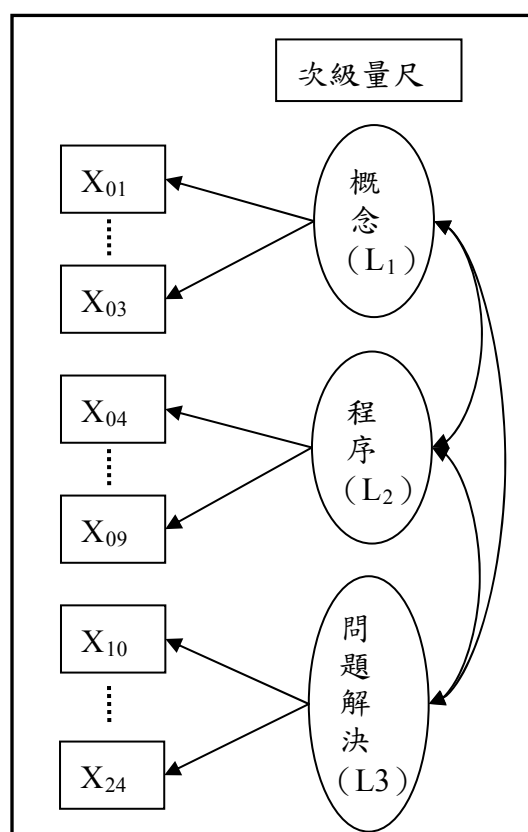
本研究自編之國小六年級學童小數的除法測驗為二元計分，共有 24 題，測驗試題設計為題間多向度測驗。

(二) 模式估計

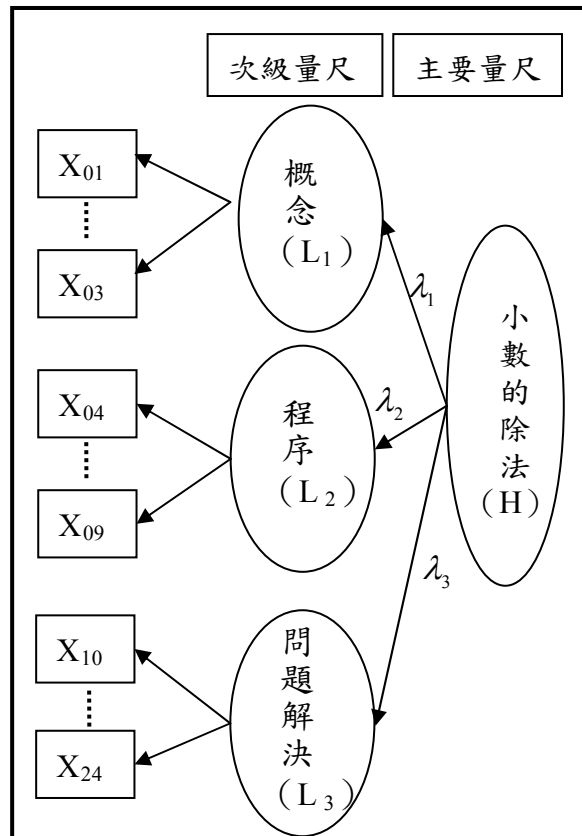
本研究應用來做測驗參數估計的模式有 UIRT、MIRT 及 HIRT 三種。本研究所指之主要量尺為數學小數的除法能力 (H)，次級量尺為數學能力--概念 (L_1)、程序 (L_2) 及問題解決 (L_3)，以 UIRT 來估計主要量尺能力，如圖三；MIRT 來估計次級量尺能力， L_1 包含 X_{01} 到 X_{03} ； L_2 包含 X_{04} 到 X_{09} ； L_3 包含 X_{10} 到 X_{24} ，如圖四；HIRT 同時估計主要量尺和次級量尺之能力， X_{01} 到 X_{03} 是估計 L_1 ； X_{04} 到 X_{09} 是估計 L_2 ； X_{10} 到 X_{24} 是估計 L_3 ， λ_1 、 λ_2 和 λ_3 分別代表 L_1 、 L_2 和 L_3 與 H 間的相關係數，如圖五。最後將 UIRT 和 HIRT 估計出的參數值及 MIRT 和 HIRT 估計出的參數值做相互比較，探討不同模式在估計上是否有一致性，HIRT 之完整估計成效是否較佳。



圖三 UIRT 模式



圖四 MIRT 之題間多向度模式



圖五 HIRT 之題間多向度模式

(三) 模式檢定

應用試題反應理論模式進行測驗資料分析，若想要獲得該模式所提供的優勢，則必須先確定資料與模式間匹配的程度是合適的。謝典佑等（2009）利用貝氏架構（Bayes framework）下的 Akaike's information coefficient（AIC; Congdon, 2003）、Bayesian information coefficient（BIC; Congdon, 2003）與 deviance information coefficient（DIC; Spiegelhalter, Best & Carlin, 1998）等指標驗證高層次試題反應理論架構下的測驗資料，以了解 AIC、BIC 及 DIC 在高層次試題反應理論之試題反應函數（item response function, IRF）選擇的效果。本研究採用基於貝氏架構的 AIC、BIC 與 DIC，探究 HIRT、MIRT 和 UIRT 模式下何者效果佳。

二、研究對象

本研究測驗採立意取樣，對象為九十八學年度六年級學生，包括中部四縣市共 18 個班級，有效樣本共計 538 人。人數來源如表 1：

表 1 樣本人數來源表

學校代碼	樣本人數			
	班級數	男生	女生	小計
台中市 A 國小	9	154	120	274
南投縣 B 國小	4	63	47	110
台中縣 C 國小	3	47	48	95
台中縣 D 國小	2	35	24	59
合計	18	299	239	538

三、研究工具

本研究使用的工具有 MATLAB 軟體、WinBUGS 軟體、SPSS 軟體與自編之六年級數學領域小數的除法測驗。

(一) MATLAB 7

本研究使用 Matlab 撰寫程式。因為 Matlab 的程式語法簡單，並具有強大的函數庫功能，可評估測驗模式之成效。

(二) WinBUGS

WinBUGS 統計軟體使用方法很彈性。WinBUGS 軟體可應用的模式相當的廣泛，包括一般線性和非線性模式、處理連續和非連續性資料及多變量模式 (Cowles, 2004; Qiu, Song, & Tan, 2002; Sturtz, Ligges, & Gelman, 2005)。本研究使用 WinBUGS 軟體進行單一階層能力估計與完整估計。

(三) SPSS

本研究以電腦統計套裝程式 SPSS 軟體進行本測驗之信度分析，分析測驗內部一致性的數值。

(四) 自編測驗

本研究採用自編之六年級數學領域小數的除法測驗。在編寫測驗之前，以 NAEP 評量架構中的數學能力向度為依據，並確認單元教學目標內所包含的節點進行命題。試題編製完成後，敦聘數學教育專家及數位國小教師進行審題，接著進行組卷、施測，最後做資料、模式比較與分析。本測驗 Cronbach α 係數值為 0.79，顯示有良好的測驗信度，效度採用的是內容效度及專家效度。

肆、研究結果

一、小數的除法測驗分析

本研究以現行九年一貫課程數學領域中，六年級的小數除法為主測驗內容，試卷經施測後分析作答情形，Cronbach α 係數值為 0.79，顯示有良好的測驗信度。

本研究之測驗通過率介於 0.182~0.905，平均為 0.549。其中試題 11、13 通過率較低；試題 4、6 通過率較高。本測驗古典測驗理論鑑別度值介於 0.037~0.459。

二、主要量尺與次級量尺間的關係

HIRT 模式中迴歸參數 (λ) 表示主要量尺與次級量尺間的相關，在教育測驗的應用上，主要量尺及次級量尺間皆為正相關，範圍在 0~1 之間。由表 2 可發現 HIRT 的三個迴歸參數值都在 0.8 以上，顯示本測驗的主要量尺--小數的除法能力與三個次級量尺--概念的了解、程序性知識、問題解決間的相關程度高，在估計精準度有一定的精確度。而且在迴歸參數 λ_1 的值較高，顯示第一個次級量尺 (λ_1) --概念的了解與小數的除法的相關比其他二個次級量尺--程序性知識、問題解決要高，亦即在小數的除法中，概念的了解對小數的除法影響較大。

表 2 HIRT 迴歸參數 (λ) 比較表

	λ_1	λ_2	λ_3
HIRT	0.9634	0.8199	0.9052

三、HIRT 完整估計與 MIRT 及 UIRT 分開估計之模式分析結果

本研究使用 UIRT、MIRT 及 HIRT 三種模式來估計學童在小數的除法測驗的主要量尺、次級量尺及試題難度參數，並做相互比較。使用 UIRT 進行主要量尺--小數的除法及試題難度參數之估計，使用 MIRT 進行三個次級量尺--概念的了解、程序性知識、問題解決及試題難度參數之估計，使用 HIRT 進行完整估計主要量尺--小數的除法和三個次級量尺--概念的了解、程序性知識、問題解決及試題難度參數。綜合比較結果如下：

(一) 各模式在試題難度參數的估計比較

將 UIRT、MIRT 及 HIRT 與試題通過率來比較，發現各模式與試題通過率的相關係數皆為-1.0，顯示在試題難度參數的估計上有很高的一致性，如表3。各模式皆顯示試題4、6是較容易的試題，試題11難度較高。再把學童測驗的試卷拿來比對，學童在試題4、6的答對率高達80%以上；在試題11的答對率僅有18%。

表 3 各模式通過率比較表

試題編號	通過率	UIRT	MIRT	HIRT
4	0.905	-2.5242	-2.6007	-2.6052
6	0.812	-1.6852	-1.7313	-1.7245
5	0.766	-1.3591	-1.4070	-1.4046
1	0.764	-1.3544	-1.3959	-1.3953
2	0.755	-1.2970	-1.3451	-1.3383
7	0.749	-1.2573	-1.2985	-1.3046
10	0.732	-1.1692	-1.1995	-1.1883
3	0.721	-1.0866	-1.1320	-1.1320
19	0.643	-0.6820	-0.7033	-0.7121
15	0.563	-0.2894	-0.3227	-0.3244
14	0.539	-0.1727	-0.1942	-0.1937
18	0.517	-0.0630	-0.0936	-0.0984
9	0.496	0.0294	0.0332	0.0361
17	0.474	0.1361	0.1119	0.1105
22	0.457	0.2130	0.1911	0.1868
24	0.437	0.3219	0.2980	0.2886
20	0.429	0.3502	0.3227	0.3246
23	0.413	0.4356	0.4044	0.4093
8	0.398	0.5076	0.5121	0.5167
16	0.392	0.5295	0.5136	0.5014
12	0.387	0.5586	0.5339	0.5319
21	0.377	0.6003	0.5957	0.5806
13	0.279	1.1384	1.1188	1.1125
11	0.182	1.7603	1.7567	1.7379
與通過率的相關係數		-1.00	-1.00	-1.00

(二) 各模式在主要量尺及次級量尺的估計比較

1. UIRT 與 HIRT 在主要量尺的估計參數比較

發現學生在主要量尺--小數的除法能力估計值上，UIRT 和 HIRT 差異不大。受試者得分相同時，UIRT 的估計值都一樣。依據學生測驗得分排序與 UIRT 及 HIRT 對學生小數的除法能力估計值相比較，發現得分低的學生在 UIRT 及 HIRT 估計下，其小數的除法能力估計值也低；得分高的學生其小數的除法能力估計值

也高，且 UIRT 與 HIRT 估計的主要量尺相關係數為 0.99，顯示 UIRT 與 HIRT 在主要量尺--小數的除法能力的估計上有一致性。

2. MIRT 與 HIRT 在三個次級量尺的估計參數比較

MIRT 與 HIRT 在學童的三個次級量尺--概念的了解、程序性知識、問題解決能力的估計值上差異不大。依據學童在測驗部份得分排序與 MIRT 及 HIRT 對學童概念、程序、解題能力估計值相比較，發現得分低的學童在 MIRT 及 HIRT 估計下，其次級量尺--概念、程序、解題能力估計值也低，得分高的學童其概念、程序、解題能力估計值也高，在 MIRT 與 HIRT 的 L1、L2、L3 的相關值分別為 0.70、0.88、0.97，雖然 L1 的相關值只有 0.70，可能是題數只有 3 題而受影響，但仍屬於高相關，顯示 MIRT 與 HIRT 在次級量尺的估計上也有一致性。

依據本研究的實證資料分析結果發現，UIRT、MIRT 及 HIRT 在主要量尺--小數的除法能力及三個次級量尺--概念的了解、程序性知識、問題解決能力及試題通過率的估計有一致性。

3. 從個別受試者作答情形來比較

- (1) 受試者 6338 號和 6188 號在本測驗的總分相同，如表 4。兩者在 UIRT 小數的除法能力的估計值是一樣的，但在 HIRT 是受試者 6188 號較高。從兩者的作答反應情形來探究，兩位同學有 4 題作答反應情形不相同，受試者 6338 號答對 12、15，受試者 6188 號答對 7、10，從試卷試題題目來分析，試題 10、12、15 是問題解決，試題 7 是程序性知識，受試者 6188 號在整份小數的除法測驗中具備較多的能力，因為他答對的小數的除法類型題目較多，相對的比受試者 6338 號具備較多小數的除法能力，由受試者 6338 號和 6188 號可知 HIRT 在小數的除法能力的估計值較 UIRT 佳。

表 4 UIRT 與 HIRT 在 (H) 的估計參數表 (受試者 6338 號和 6188 號)

受試者		UIRT	HIRT
編號	得分 (%)	theta H	
6338	63	0.3159	0.2849
6188	63	0.3159	0.4251

- (2) 受試者 6207、6224 和 6225 號在本測驗的能力值 0.6175 皆相同，如表 5。從三者的作答反應情形來探究，三位同學在試題 1、2、3 屬於概念的了解上作答全對，MIRT 的估計值是 -0.1855 至 -0.1210；HIRT 的估計值是 0.3951 至 0.5028，由此可知，HIRT 的估計模式比 MIRT 更能適切反映原始的作答反應。另受試者 6002 號在本測驗的能力值是 1.4995，MIRT 在其概念的了解估計值是 -0.2020；HIRT 的估計值是 1.3196，從 6002 號的作答反應情形來探究，屬於概念的了解試題 1、2、3 上，答錯試題 3，而試題 3 的通過率是 0.755，由此可知，HIRT 的估計模式比 MIRT 更能真實反映，受試者 6002 號在測驗上的原始作答反應。

表 5 MIRT 與 HIRT 在 (L1、L2、L3) 的估計參數表 (部分)

受試者			MIRT	HIRT	MIRT	HIRT	MIRT	HIRT
編號	CTT θ	得分	theta L1		theta L2		theta L3	
6207	0.6175	66.	-0.1855	0.3951	0.9862	0.6495	0.2767	0.4111
6224	0.6175	66.	-0.1210	0.4847	1.1007	0.7687	0.2678	0.4255
6225	0.6175	66.	-0.1595	0.5028	0.9912	0.8078	0.3171	0.4544
6002	1.4995	87.	-0.2020	1.3196	0.9222	1.4083	1.5589	1.5763

四、模式適配度指標的分析比較

HIRT、MIRT 及 UIRT 三種模式之適配度指標 AIC、BIC、DIC 的分析數據，如表 6，數值較低者為佳。AIC、BIC 指標皆顯示 HIRT 較適合用來分析階層式評量架構之測驗，而 DIC 是 UIRT 比較低，但與 HIRT 差異不大，以整體而言，HIRT 模式較適合用來分析階層式評量架構之測驗。

表 6 各模式指標的分析結果

指標	UIRT	MIRT	HIRT
AIC	17160	19706	16021
BIC	17284	19830	16145
DIC	17395	23042	17636

伍、結論與建議

一、研究結論

(一) 自編小數的除法能力測驗信度和效度佳

本研究使用 HIRT 模式建立之小數的除法能力測驗，其內部一致性的數值，Cronbach α 係數值為 0.79，顯示有很好的測驗信度。效度採用的是內容效度及專家效度分析。在測驗編製過程中，均與有測驗編製經驗的專家學者及現任國小教師一同開會討論，進行試題內容與測驗編排的審核與修訂，試題完稿後再經討論檢核試題編製。

(二) HIRT 模式顯示小數的除法能力與概念的了解相關較高

HIRT 的三個迴歸參數值都在 0.8 以上，顯示本測驗的主要量尺--小數的除法與三個次級量尺--概念、程序、問題解決間的相關程度高。而且在迴歸參數 λ_1 的值較高，顯示第一個次級量尺 (λ_1) --概念與小數的除法相關比其他二個次級量尺要來的高，亦即在小數的除法能力中，概念的理解對小數的除法影響是較大。

(三) HIRT 模式的分析結果佳

本研究的實證資料研究結果發現，HIRT、MIRT 及 UIRT 在主要量尺--小數的除法及三個次級量尺--概念、程序、問題解決及試題難度參數的估計有一致性；在主要量尺的估計上 HIRT 的效果較 UIRT 佳，在次級量尺的估計上 HIRT 的效果也比 MIRT 佳，顯見使用 HIRT 模式來估計是較佳的選擇。

(四) HIRT 模式較適合分析階層式評量架構之測驗

比較 HIRT、MIRT 及 UIRT 三種模式適配度的 AIC、BIC 和 DIC 指標，AIC、BIC 指標皆顯示 HIRT 較適合用來分析階層式評量架構之測驗，而 DIC 是 UIRT 比較低，但與 HIRT 差異不大。所以，整體而言，HIRT 模式較適合用來分析階

層式評量架構之測驗。

二、建議

(一) 試題編製

本研究在測驗的編製內容上，只針對數學內容中數與量下的有理數部份進行出題，建議未來研究者可擴大出題範圍，在數與量這方面進行探究。另外，本測驗的電腦化試題都是選擇題，後續研究者可以朝向編制電腦化的建構反應試題，來探討模式的適配度是否有差異。

(二) 測驗使用的 HIRT 和 MIRT 估計模式

本研究測驗的估計模式，僅使用 HIRT 和 MIRT 模式中的題間多向度作分析，未來可再加入題內多向度的比較。另外，本研究只以 NEAP 數學能力中的概念、程序、解題作為次級量尺，所以，後續研究者可再加入推理、溝通、連結，完整的估計學生的數學能力。

(三) 計分方式

本研究之計分型態屬於二元計分，所以後續研究者可延伸研究至多點計分或二元計分與多點計分混合，來探討模式參數估計的精準度是否有差異。

參考文獻

中文部分

- 王敏嫻 (2010)。多向度試題反應理論之可能值方法對大型測驗中群體平均數估計之影響—以TASA2006數學科為例。**測驗統計年刊**，18(1)，47-68。
- 余民寧 (1992)。試題反應理論的介紹(三)-試題反應模式及其特性。**研習資訊** 9(2)，6-10。
- 吳昭容 (1996)。先前知識對國小學童小數概念學習之影響。國立台灣大學心理學研究所博士論文。
- 吳昭容 (2003)。「理解」與「計算」，有何兩難？國立台北師範學院教育心理與輔導學系、數學教育學系聯合主辦「你建構我運算，孩子會了什麼？」數學教育之趨勢研討會，台北。
- 林佳樺 (2009)。高階層試題反應理論及其成效探討。國立臺中教育大學教育測驗統計研究所碩士論文。
- 邱美珍 (2008)。以多向度試題反應理論量測兒童步行上放學之能力。國立交通大學運輸科技與管理學系碩士論文。
- 教育部 (2003)。國民中小學九年一貫課程綱要數學學習領域。台北市：教育部。台國字第0920167129號。
- 陳永峰 (1998)。國小六年級學童小數知識之研究。**國民教育研究**，3，337-373。
- 張勝凱 (2010)。使用HIRT模式建立國小六年級學童數學推理能力測驗。國立臺中教育大學教育測驗統計研究所碩士論文。
- 劉曼麗 (2002)。台灣地區國小學童小數概念研究(Ⅱ)國小學童「小數與小數運算」概念之調查研究。國科會補助之專題計畫成果報告。編號：(NSC 90-2521-S-153-003)。
- 劉曼麗 (2004)。九年一貫數學領域分數與小數能力指標的詮釋：子計畫三——小數。行政院國家科學委員會專題研究計畫成果報告。編號：(NSC

92-2521-S-153-008)。

- 劉曼麗 (2008)。小數除法的學與教。科學教育月刊，314，27-38。
- 賴文溥 (2009)。國小六年級學童小數除法概念結構分析之研究。國立臺中教育大學數學教育研究所碩士論文。
- 謝典佑、林佳樺、郭伯臣、施淑娟 (2009)。單因子高層次IRT模式適合度檢定之研究－以TASA數學科為例。「大型教育資料庫建置及相關議題」學術研討會，2009年9月，國立臺中教育大學。

英文部分

- Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Congdon, P. (2003). *Applied Bayesian modelling*. New York: John Wiley.
- Cowles, M. K. (2004). Review of WinBUGS 1.4. *The American Statistician*, 58, 330-336.
- Hoskens, M., & De Boeck, P. (1997). A parameteric model for local dependence among test items. *Psychological methods*, 2, 261-277.
- National Assessment Governing Board(2002). Mathematics framework for the 2003 national assessment of educational progress. National Assessment Governing Board U.S. Department of Education.
- Qiu, Z., Song, P. X.-K., & Tan, M. (2002). Bayesian hierarchical models for multi-level repeated ordinal data using WinBUGS. *Journal of Biopharmaceutical Statistics*, 12, 121-135.
- Rasch, G. (1960). *Probability models for some intelligence and attainment tests*. Copenhagen Denmark: Danmark's Paedagogiske Institute for Educational Research.
- Song, H. (2007). *A higher-order item response model: development and application*. Unpublished doctoral dissertation, The State University of New Jersey.
- Spiegelhalter, D., Best, N., & Carlin, B. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Technical report, Division of Biostatistics, University of Minnesota. Research Report 98-009.
- Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12, 1-16.
- Wang, W., Wilson, M., & Cheng, Y. (2000). Local Dependence between Latent Traits when Common Stimuli are Used. *Paper presented at the International Objective Measurement Workshop*, New Orleans, LA
- Weiss, D. & Yoes, M. (1991). Item Response Theory. In R. K. Hambleton & J. Zall (Eds.), *Advances in educational and psychological testing*. Boston: Kluwer-Nijhoff.
- Wilson, M. & Adams R. J.,(1995). Rasch models for item bundles. *Psychometrika*, 60, 181-198.