

HIRT 電腦化測驗-以六年級數學幾何單元為例

孫長蓀

臺中教育大學教育測驗統計研究所研究生
charng.suen@yahoo.com.tw

張素珍

臺中教育大學教育測驗統計研究所研究生
jk86jk88@yahoo.com.tw

郭伯臣

臺中教育大學教育測驗統計研究所教授
kbc@mail.ntcu.edu.tw

林佳樺

臺中教育大學教育測驗統計研究所博士生
hua_2812@hotmail.com

摘要

近年來為測得學生較高階的能力值，測驗編製者逐漸朝向設計評量架構為階層式的測驗，然而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，亦可獲得較好的估計結果。本研究根據 NAEP 的數學評量架構，編製一份六年級數學「幾何」測驗，並製作成電腦化試題，利用實證資料比較 HIRT (hierarchical item response theory)、MIRT (multidimensional item response theory) 和 UIRT (unidimensional item response theory) 三種模式對數學能力值的估計是否有差異，影響如何？以作為數學評量模式的參考。

此份測驗信度是 0.82，測驗結果以 HIRT、MIRT 和 UIRT 三種模式進行分析和比較。其中模式適配度指標(model fit indexes)AIC、BIC、DIC 三種指標值都顯示 HIRT 的模式較適配多階層評量架構；分析結果 HIRT 估計出來的能力與答對題數有高相關。因此透過實證方式證實 HIRT 模式可以提供較多訊息，估計效果較佳。

關鍵詞：階層式試題反應理論、電腦化測驗
幾何能力、NAEP

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 Geometry is designed and computerized for six-grade students based on the

mathematical assessment framework of NAEP.

The reliability of the assessment is 0.82. 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 and has high correlation between students' abilities and the number of questions correctly answered. Therefore, the result of the empirical study confirms HIRT model can provide more information and has better estimation.

Keywords: hierarchical item response theory, computer-based test, geometry-abilities, NAEP

第一章 研究動機與目的

學生的數學能力一直是我們國人非常注重的問題，因此學生參加各種大型數學競賽或數學測驗日益頻繁。如：國際學生評量 (The Programme for International Student Assessment, 簡稱 PISA)、國家教育進展評量 (The National Assessment of Educational Progress, 簡稱 NAEP) 和國際數學趨勢研究 (The Trends in International Mathematics and Science Study, 簡稱 TIMSS) [6]。這些大型測驗通常都會有總測驗的分數 (主要量尺) 及分測驗的分數 (次級量尺)，皆屬於高階層測量模式，但這些大型測驗卻是使用單向度試題反應理論或是多向度試題反應理論等測驗模式進行分析，可能會因違背其假設而使高層次量尺分數估計不準確，或是當低層次量尺分數所對應的題數較少時，會造成估計效果不可靠 [8]。

為因應較複雜之評量架構，林佳樺探討適用於高階層評量架構之測量模式，以 PISA 之評量架構 (如圖 1) 作為基礎，設計高階層試題反應理論模式，提出可以同時估計主要量尺 (total score) 及次級量尺 (domain score) 的

完整模式，且估計誤差皆接近或優於 PISA 之估計方式[6]。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。

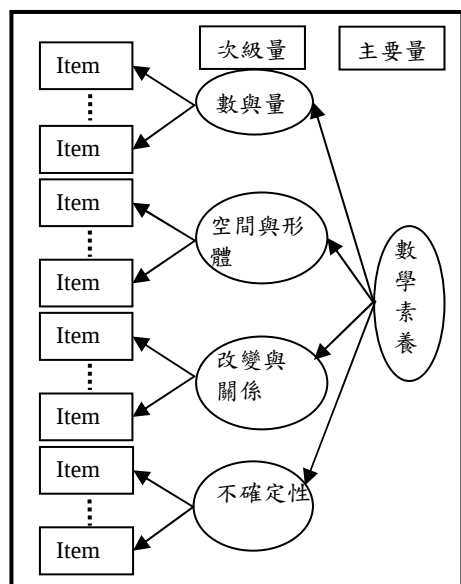


圖 1 PISA 數學科評量架構[6]

基於上述動機，並依據林佳樺的研究[6]，所發展出的 HIRT 模式可以同時估計主要量尺及次級量尺，適用於高階層評量架構。而數學幾何單元，是六年級學童必備非常重要的數學知識。所以本研究的目的如下：

- 1 依據 NAEP 數學評量架構，編製一份數學幾何電腦化測驗。
- 2 比較 HIRT、MIRT 及 UIRT 模式的分析結果，驗證 HIRT 模式是否可以應用於數學幾何的測驗，並作為模式選用之依據。
- 3 瞭解學童在數學幾何測驗的表現情形。

第二章 文獻探討

本研究主要目的是以 NAEP 架構編製一份國小六年級學童數學幾何能力之測驗，透過 HIRT 模式了解學童的幾何概念能力、幾何程序能力及幾何問題解決能力間的關係；比較 HIRT、MIRT 及 UIRT 模式的分析結果，驗證 HIRT 模式是否可以應用於數學幾何能力測驗，及瞭解學童在數學幾何測驗上的表現情形。因此本章將分成四節來加以闡述：第一節為 NAEP 的數學架構；第二節為數學幾何的意

義及國小幾何教材；第三節為電腦化測驗；第四節為試題反應理論模式。

第一節 NAEP數學評量架構

NAEP是指美國國家教育進展評量委員，NAEP數學領域之評量架構主要包含三大向度：內容成分（content strands）、數學能力（mathematical abilities）、數學力（mathematical power），每一向度各有其所屬之子向度（如圖2）[22]。

內容成分包含：數學的概念、性質與運算（number sense, properties, and operations）、測量（measurement）、幾何與空間觀念（geometry and spatial sense）、資料分析、統計與機率（data analysis, statistics, and probability）、代數與函數（algebra and functions）

數學能力包含：概念的了解（conceptual understanding）、程序性知識（procedural knowledge）、及問題解決（problem solving）。

數學力包含：推理（reasoning）、連結（connections）、及溝通（communication）。

本研究將以數學內容—幾何，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。瞭解學童在幾何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—數學內容—幾何的影響情形，如圖 2。如此可更完整的知道學童在整份數學幾何測驗中的表現。

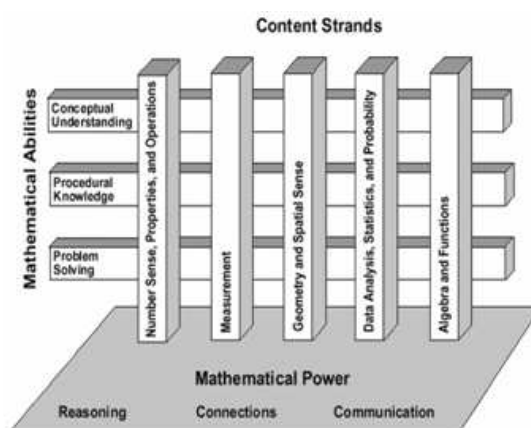


圖 2 NAEP 數學科評量架構[22]

第二節 幾何的意義及國小幾何教材

壹、何謂幾何

Geometry 一詞的拉丁文是 geometria, geo 是土地的意思, metria 是測量的意思, 所以它的原意就是土地測量, 研究物體形狀、大小、位置以及它們相互關係的學科。包括點、直線、圓、曲線、平面與立體等, 以前也稱之為「形學」。

貳、幾何能力

本研究主要在評量國小六年級的幾何內容, 根據九年一貫課程能力指標, 國小六年級屬於數學領域的第二階段, 其「幾何」課程需達到的能力指標如下:

1. 能運用簡單幾何形體的組成要素, 作不同形體的分類。
2. 能理解垂直與平行的意義。
3. 能透過操作, 認識簡單平面圖形的性質。
4. 能認識平面圖形全等的意義。
5. 能理解旋轉角的意義。
6. 能理解平面圖形的線對稱關係。
7. 能理解長方形面積、周長與長方體體積的公式。
8. 能運用切割重組, 理解三角形、平行四邊形與梯形的面積公式。

參、國小幾何教材分析

根據92課綱, 數學領域將九年國民教育區分為四個階段, 另將數學內容分為「數與量」、「幾何」、「代數」、「統計與機率」、「連結」等五大主題。國小五、六年級屬於數學領域的第二階段, 幾何課程的主要內容單元分別是「形體的性質」、「柱體與錐體」、「圓形圖」。

第三節 電腦化測驗

電腦化測驗的優點是學生藉由電腦試題作答可以得到及時回饋, 且計分快速。因電腦能儲存大量訊息之優勢, 在記錄學生相關作答歷程的同時, 可以獲得更多的學生作答訊息, 並更精確的估計學生之相關能力。NetPASS是一套由Cisco Networking Academy Program (CNAP) 研發的系統, 用於電腦網路設計、結構和偵測錯誤的評量及訓練。為了評量學生是否可以讓網路運作, 所以評量的設計必須在深度及廣度上都必須超越其他的線上測驗。此系統可供學生練習電腦網路技巧並從學生作答反應得到回饋, 亦可由學生作答反應得到之

證據, 並可估計及儲存學生多次作答反應之技能及能力。NetPASS所建立的模擬環境是以心理學為基礎用以支援學習, 其中包括知識的建構、多元評量的使用、真實複雜測驗的表現、某些抽象概念及反省的支持。當學生第一次登入時, 可於頁面上選擇範圍(設計、結構或偵測錯誤)及不同的難度等級(簡單、中級、困難); 在學生作答完成後, NetPASS針對學生的錯誤予以回饋。[5]

本研究的電腦化測驗是以貝氏網路為基礎之適性系統 (Bayesian Network based Computerized Adaptive Test, 簡稱BNAT), 此系統是由郭伯臣、曾彥鈞 (2007) 所架設之線上測驗平台, 再以貝氏網路(Bayesian Network)機率更新的技術提供錯誤類型診斷的工具。此線上測驗, 能診斷出學生所犯的錯誤類型和迷失概念, 使得評量、診斷與補救教學皆能達到個別化與適性化的目的[11]。

第四節 試題反應理論模式

壹、單向度 IRT 之 Rasch 模式

IRT 建立在兩個基本概念上: (一)受試者 θ_i 在測驗試題上的答題表現, 可由潛在特質 (latent traits) 或能力 (abilities) 來加以預測或解釋; (二)受試者的答題表現與其能力間的關係, 可透過試題特徵曲線 (item characteristic curve, ICC) 來加以解釋。任何一條試題特徵曲線是由受試者能力和試題的特性共同決定, 代表受試者答對某一試題的機率[4]。

本研究只使用單參數對數模式, 僅針對單參數對數模式進行介紹, 亦即 Rasch model。

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

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

貳、多向度 IRT 之 MRCML 模式

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[1]。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；在測驗裡的每一試題不只測量單一能力，也就是試題內包含多個向度，稱此測驗為題內多向度測驗 [6]。

其模式定義如下：

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

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）

參、HIRT 模式

階層式試題反應理論模式（hierarchical item response theory, HIRT），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[8]。

測驗主要可觀察多個單向度的子測驗，即低層次量尺分數 $\theta_i^{(d)}$ ， $\theta_i^{(d)}$ 表示第 i 位受試者在低層次量尺 d 的表現，其中， $d=1,2,3,\dots,D$ 。當不同低層次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，而不同低層次量尺間有關聯，則會藉由一高階層能力量尺分數 θ_i 來連接這些低層次量尺分數，其中 θ_i 為第 i 位受試者在高層次量尺分數，其中並假設低層次量尺分數為高層次能力量尺分數的一線性函數[21]。

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

其中， $\lambda^{(d)}$ 為回歸參數， 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，更假設在已知高層次量尺分數，則低層次量尺間會互相獨立；其中 $\lambda^{(d)}$ 更可表示高層次量尺分數與低層次量尺分數間的相關，而低層次量尺 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高層次量尺及低層次量尺間的相關皆為正的[21]。

圖 4 為該模式的模式圖，第一層表示第 i 位受試者在低層次量尺 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 $\beta_j^{(d)}$ 連結到低層次量尺分數，第三層表示受試者的低層次量尺分數透過回歸參數 $\lambda^{(d)}$ 連結到相對應之高層次量尺分數 θ_i 。

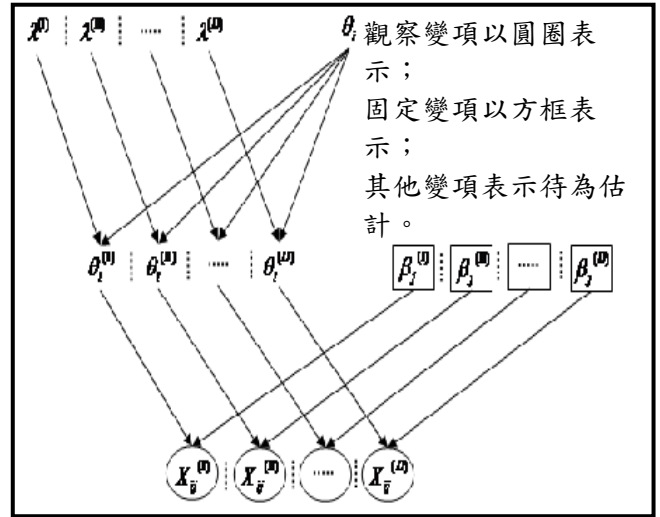


圖 3 應用於一個 D 維度的測驗[21]

第三章 研究方法

本章共分為四節作說明，分別介紹研究設計、研究流程、研究範圍與對象、研究工具。

第一節 研究設計

本研究使用林佳樺[6]發展之 HIRT 模式，透過實證資料方式探討完整估計與分開估計之估計效果。使用 WinBUGS 軟體進行單一階層（各階層分開估計）之估計及兩階層（完整估計）之估計；再比較 UIRT 和 MIRT 的分開估計與 HIRT 完整估計的成效。

壹、計分方式與測驗型態

測驗計分型態一般可分為二元計分與多點記分，多元評量大都屬於多向度測驗，多向度測驗又分成題間多向度測驗及題內多向度

測驗兩種[1]。本研究自編之國小六年級學童數學幾何測驗為二元計分，測驗試題設計為題間多向度測驗。以 BNAT 適性診斷測驗系統進行施測如圖 4。



圖 4 電腦化測驗介面

貳、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幾何內容 (H)，次要量尺為數學能力-概念 (L_1)、程序 (L_2) 及問題解決 (L_3)。(如圖 5)

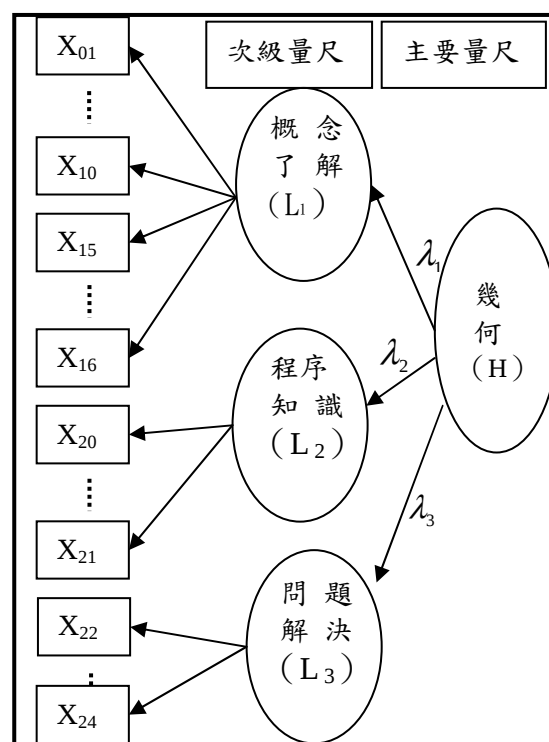


圖 5 HIRT 多向度模式

第二節 研究流程

本研究流程，首先研讀相關文獻以擬定研究主題，而後就研究主題蒐集相關資料做為理論基礎，將收集到的電腦化試題測驗結果進行分析，並做出結論，最後撰寫研究報告。

第三節 研究範圍與對象

電腦化測驗採立意取樣，對象為九十八學年度六年級學生，包括中部四縣市共 16 個班級，有效樣本共計 434 人。人數來源如表 1：

表 1 樣本人數來源表

學校	班級數	男生	女生	小計
A 國小	6	71	69	140
B 國小	2	38	31	69
C 國小	1	14	11	25
D 國小	1	16	18	34
E 國小	6	79	87	166
合計	16	218	216	434

第四節 研究工具

本研究使用的工具有自編之數學幾何能力測驗、Bilog-mg 軟體以分析試題參數、MATLAB 軟體計算模式適配度指標、WinBUGS 軟體分析試題參數及受試者能力值、SPSS 軟體估計測驗信度。

第四章 研究結果

本章主要是呈現資料分析之結果與討論，共為四節：第一節為數學幾何電腦化測驗分析；第二節比較 HIRT、MIRT 及 UIRT 三種模式之分析結果；第三節為學童在數學幾何測驗的表現情形；第四節為數學幾何與概念、程序、問題解決間的關係

第一節 數學幾何電腦化測驗分析

試卷經施測後分析作答情形，分析測驗內部一致性的數值，Cronbach α 係數值為0.817，顯示有良好的測驗信度。

本研究之測驗以通過率表示難易度，測驗之試題之平均難度指數 0.44。本測驗試題鑑別度值介於 0.481~1.495。

第二節 比較 HIRT、MIRT 及 UIRT 模式的分析結果

一、適配度的指標分析

HIRT、MIRT 及 UIRT 三種模式之適配度指標 AIC、BIC、DIC 的分析數據如表 2，數值較低者為佳。AIC、BIC 與 DIC 所獲得的結果表示模式選取效果，其結果顯示 HIRT 模式最適合用來分析階層式評量架構之測驗。

表 2 各模式指標的分析結果

指標	HIRT	UIRT	MIRT
AIC	13519	14111	16865
BIC	13277	14229	16983
DIC	13403	16442	19865

二、各模式在試題難度參數的估計比較

將 HIRT、MIRT 及 UIRT 與試題通過率來比較，可發現各模式在試題難度參數的估計上有一致性。各模式皆顯示試題 1、5、12 是最容易的試題，試題 22 難度最高，如表 3。

表 3 各模式試題（難度）參數估計值

題號	HIRT	UIRT	MIRT
1	-1.548	-1.484	-1.508
2	-0.533	-0.492	-0.489
3	0.341	0.374	0.372
4	-0.417	-0.385	-0.374
5	-1.250	-1.215	-1.218
6	-0.627	-0.585	-0.590
7	0.062	0.094	0.098
8	-0.639	-0.608	-0.613
9	0.035	0.067	0.054
10	-0.460	-0.442	-0.449
11	-0.369	-0.343	-0.339
12	-1.372	-1.330	-1.351
13	0.184	0.216	0.203
14	-0.191	-0.162	-0.153
15	-0.485	-0.454	-0.451
16	-0.783	-0.738	-0.737
17	-0.270	-0.239	-0.235
18	0.257	0.283	0.279
19	-0.473	-0.444	-0.438
20	0.147	0.186	0.180
21	0.013	0.043	0.047
22	1.459	1.478	1.458
23	0.518	0.544	0.549
24	0.210	0.235	0.233

二、學生在幾何測驗中主要量尺的能力表現

比較受試者在 HIRT 和 UIRT 主要量尺（幾何內容）的估計結果，發現 HIRT 和 UIRT 的相關值為 0.89，所以 HIRT 和 UIRT 估計有一致性。

受試者在 HIRT 和 UIRT 主要量尺（幾何內容）平均能力值分別為-0.016 和 0.017。可見國小六年級學童的幾何能力偏低。

三、學生在幾何測驗中概念、程序、問題解決的能力表現

在本研究中，次級量尺分別代表 L1 是幾何的概念、L2 是幾何的程序性知識、L3 是幾何的問題解決。分析 HIRT 和 MIRT 在估計學生的次級量尺能力值的相關，顯示 HIRT 和 MIRT 對 L1 的相關最低，L2 和 L3 的相關都很高如表 5。由模式適配度指標分析中得知 MIRT

不適合階層式的評量模式，因此可能造成 HIRT 和 MIRT 在 L1 的估計上不一致。

表 5 HIRT 與 MIRT 對次級量尺之相關

次級量尺	L1	L2	L3
評量模式	HIRT	HIRT	HIRT
MIRT	0.549	0.883	0.867

所以從 HIRT 模式中發現，在估計學生的次級量尺能力值時，學生在 L2 程序性知識和 L3 問題解決的能力是較優於 L1 概念的了解，如表 6。

表 6 HIRT 與 MIRT 對次級量尺之估計結果

能力值	L1	L2	L3
HIRT	-0.0175	-0.0158	-0.0157
MIRT	0.0031	0.0123	0.0040

第三節 數學幾何與概念、程序、問題解決的關係

HIRT 模式中迴歸參數 (λ) 表示主要量尺與次級量尺間的相關， λ 的範圍在 0~1 之間， λ 值越大表示次級量尺對主要量尺之貢獻量越多，相關越高。由表 7 可發現 HIRT 三個迴歸參數值都在 0.9 以上，顯示本測驗的主要量尺—數學幾何內容與三個次級量尺—概念、程序、問題解決的相關程度高，在估計精準度有一定的精確度。而且在迴歸參數 λ_1 的值最高，顯示第一個次級量尺 (λ_1)—概念與數學幾何的相關比其他二個次級量尺—程序及問題解決要來的高，亦即在數學幾何中，概念對學童數學幾何的表現影響較大。

表 7 HIRT 參數迴歸參數 (λ) 表

HIRT	λ_1	λ_2	λ_3
題間多向度	0.99925	0.93958	0.93800

第五章 討論與建議

本章分二節，第一節為本研究之結論，第二節就本研究未盡完備之處，提供相關研究建議，茲分述如下：

第一節 結論

壹、自編數學幾何測驗信、效度佳

一、信度

本研究所編製之數學幾何能力測驗，其內部一致性的數值，Cronbach α 係數值為 0.81，顯示有很好的測驗信度。

二、效度

本測驗之效度採用的是內容效度及專家效度分析。在編製測驗過程中，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，共同審核與修訂。試題完稿後再經檢核試題編製合宜，並提供修改測驗工具的建議，作為內容效度及專家效度的證據。

貳、HIRT 模式顯示數學幾何測驗中，數學幾何內容與幾何概念的相關較高

HIRT 的三個迴歸參數值都在 0.9 以上，顯示本測驗的主要量尺—數學幾何與三個次級量尺—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)—概念與幾何能力的相關比其他二個次級量尺要來的高，亦即在數學幾何能力中，概念的理解對幾何能力的影響是較大。

參、學生的幾何能力表現

本研究的實證資料研究結果發現，在幾何的測驗中主要量尺值為-0.016，學生的幾何能力偏低。學生的程序性知識和問題解決的能力表現較好，而對於幾何的概念了解表現較差。

肆、HIRT 模式最適合分析階層式評量架構之測驗

比較 HIRT、MIRT 及 UIRT 三種模式的 AIC、BIC 和 DIC 指標，顯示 HIRT 模式最適合用來分析階層式評量架構之測驗。

第二節 建議

本節就本研究未盡完備之處，提出一些研究建議，供後續研究者參考。

壹、HIRT 模式

本研究僅用 HIRT 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。

貳、計分方式

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

參、試題編製

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。

參考文獻

- [1] 王文中，”Rasch 測量理論與其在教育上的應用”，教育與心理研究，pp.637-694，2004。
- [2] 余民寧，”試題反應理論的介紹（一）-測驗理論的發展趨勢”，**研習資訊**8，pp.13-18，1992a。
- [3] 余民寧，”試題反應理論的介紹（二）-測驗理論的發展趨勢”，**研習資訊**9，pp.5-9，1992b。
- [4] 余民寧，”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，pp.6-10，（1992c）。
- [5] 吳任婕，”「分數的乘法」數學單元之建構反應題電腦化測驗”，**全球華人計算機教育應用大會 GCCCE**，2009。
- [6] 林佳樺，”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**，2009。
- [7] 莊月嬌，”九年一貫課程小學幾何教材內容分析研究”**國立臺北師範學院數學教育研究所碩士論文**，2004。
- [8] 張勝凱，”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**，2010。
- [9] 陳于倩、姚如芬”van Hiele 幾何思考層次與 SOLO 分類法之比較”，**屏師科學教育**，2002。
- [10] 曾明義，”NEAP2007評量架構在台灣數學成就評量發展模式支應用”，**國立臺北教育大學數學教育研究所碩士論文**，2008。
- [11] 蔡幸如、張永鑫、郭伯臣、蘇少祖”資訊科技融入國小五年級數學「數量關係」數位系統教材研發”，**資訊管理際商務科技研討會學術論文**，2010。
- [12] 盧雪梅，”美國的「全國教育進展評量」（NAEP）”，**人文及社會學科教育通訊**11，pp.143-153，2000。
- [13] 謝典佑、林佳樺、郭伯臣、施淑娟，”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**，國立臺中教育大學，2009。
- [14] 譚寧君，”兒童的幾何觀—從Van Hiele 幾何思考的發展模式談起”，**國民教育**，1993。
- [15] Adams, R. J., Wilson, M., & Wang, W. C., “The multidimensional random coefficients multinomial logit model”***Applied Psychological Measurement***, Vol.21, pp.1-23,1997.
- [16] Clements,D.H.,Battista,M.T, “ Geometry and spatial reasoning “, **In D.A.Grouws (Ed.), Handbook**,1992.
- [17] Cowles, M. K. (2004). “ Review of WinBUGS 1.4. “,**The American Statistician**, Vol.58, pp.330-336,2004.
- [18] Dane, F. C“.Research Methods Pacific Grove “,**California: Brooks/Cole**,1990.
- [19] Hargreaves, “M., Shorrocks-Taylor, D., & Threllfall, J. Children's strategies with number patterns“, **Educational Studies**, Vol24(3), pp.315-331,1998.
- [20] Hoskens, M., & De Boeck, P.“A parameteric model for local dependence among test items“,**Psychological methods**,Vol.2,pp. 261-277,1997.
- [21] Song,H,“A higher-order item response model: development and application“. Unpublished doctoral dissertation, **The State University of New Jersey**,2007.
- [22] “mathematics framework for the 2003: nationa assessment of educational progress“.,2003.