# 國中八年級自然科定期評量之性別差別試題功能（DIF）分析

蕭偉智* 傅家珍**

## 摘 要

本研究旨在探究國中八年級學生學校自然科定期評量之性別成就差異和差別試題功能（Differential Item Functioning, DIF）。研究資料係新北市某公立國中八年級 382 名學生（男性 191 名、女性 191 名）之答題反應，研究先以性別效果量、女／男標準差比值與女／男人數比值三項指標分析性別成就差異，再用 IRT Rasch Model 與 Mantel-Haenszel 法分析性別 DIF 現象。研究發現如下：1. 未配組的性別成就差異部分，全體受試並無顯著性別成就差異現象，而在高成就組（前 10%）和低成就組（後 10%）中，男性表現略高於女性；2. 將男女能力配組後，取 IRT Rasch 與 Mantel-Haenszel 結果交集，顯示自然科定期評量嚴重 DIF 的出現率為 4%且皆有利女性。筆者對 DIF 試題的內容進行進一步審視，試題特徵似乎與性別 DIF 有關聯，初步推論試題的試題附圖的提供是否可能影響 DIF 方向，但仍需更多實徵資料驗證。整體而論，本研究並未發現和測驗目標無關的因素，尚不構成試題偏誤。最後，本研究依據研究結果，對自然科測驗命題與未來研究提出建議。

**關鍵詞：**自然科成就性別差異、自然科性別 DIF、學校定期評量

---

責任編輯：楊龍立
投稿日期：2012 年 2 月 17 日，2012 年 6 月 4 日修改完畢，2012 年 10 月 15 日通過採用
*蕭偉智，國立臺灣師範大學教育心理與輔導學系碩士生，Email: toome123020@hotmail.com
**傅家珍，國立臺灣師範大學工業教育學系碩士生

# 壹、緒論

良好的測驗須能反映出受試者在欲測量特質上的差異。若一個測驗分數反映的不是原先要測量的特質，而是其它無關變項，例如：其他的認知能力、種族、性別或城鄉等因素，則運用該測驗結果做判斷時，就會產生偏誤的現象。

首屆臺北市、臺北縣及基隆市高中職聯合入學測驗（北北基聯測）於 2011 年 5 月 22 日結束後，各家媒體開始討論了哪一科試題對於男性或女性有利，研究者在此擷取一則報導為例：

> 「卓意翔說，聯測第一年考生反映國文『不易』、英語『爆難』，還認為是『史上最難』，對都會型、女性會佔優勢，因為一般語文能力女性比男性好；數學算是『小難』，與學校最後二次模擬考差不多，男女差異性不大；自然、社會命題『簡單』，這對女性有利，今年聯測考生決勝關鍵在國文、英語二科。」（”建北至少 400 分”，2011）

該篇報導充斥了性別學業成就刻板印象，也就是說，女生在語文和社會科表現較佳，而男性在數學和自然科表現較好。男女生的學業成就真的出現前述的差異嗎？研究上，若受試群體未經任何的配組程序，所觀察到表現差異稱為impact（Dorans & Holland, 1993），但是需特別注意是否可能會出現「辛普森悖論」（Simpson’s Paradox），也就是說在某個條件下的兩組資料，分別討論時都會滿足某種性質，可是一旦合併考慮，卻可能導致相反的結論。如：Simpson提出了一個治療與生存的例子（Simpson, 1951: 241），表1的資料為分別為不同性別的資料，從機率的角度檢驗「治療與否」和「生存與否」的關聯性，可發現兩變項並非相互獨立（若P(A｜B)=P(A)，則A和B事件為獨立事件）。以男性組別資料來看，先檢驗A細格，P(生存｜未接受治療)=4/7不等於P(生存)=12/20，同樣方式檢驗B細格、C細格、D細格，結果可發現在男性組別中「治療與否」和「生存與否」並非獨立事件。同理，在女性組別

資料中，兩變項亦非獨立事件。然而，將男、女資料合併，卻發現治療與否和生存與否兩變項相互獨立。

表1 辛普森悖論（Simpson's Paradox）例子

| | 男性（*n*=20） | | 女性（*n*=32） | | 全體（*n*=52） | |
|---|---|---|---|---|---|---|
| | 未接受治療 | 接受治療 | 未接受治療 | 接受治療 | 未接受治療 | 接受治療 |
| 生存 | 4(A) | 8(C) | 2 | 12 | 6 | 20 |
| 死亡 | 3(B) | 5(D) | 3 | 15 | 6 | 20 |

**資料來源：**"The interpretation of interaction in contingency tables" by E. H. Simpson, 1951, *Journal of the Royal Statistical Society*, *13*(2), 241.

再則，不同於 impact 概念的是差別試題功能（Differential Item Functioning, DIF）DIF 係指若來自不同族群，但能力相同的個人，在答對某個試題上的機率有所不同的話，則該試題便顯現出 DIF 的現象，例如：國立臺灣師範大學心理與教育測驗研究發展中心[臺師大心測中心]（2008）曾針對不同性別、社經地位（低收入戶與一般生）、身分別（原住民與一般生）、都市化程度間國中基測成績的差異提出相關說明，也就是說分析不同特性考生間是否產生不同差別試題功能，其中發現 2006 年至 2008 年各科性別 DIF 現象，國文科的 DIF 題數降至 2.08%、英語科 0.00%、數學科 0.00%、自然科 5.17%、社會科 6.35%、總測驗 3.22%，遠低於一般文獻所指出 DIF 比率標準（10%至 15%）。

2009 年，教育部公布了「擴大高中職及五專免試入學實施方案」明訂自 99 學年度起逐步擴大高中、高職及五專免試入學比率及名額，101 學年度後公立高中提供免試入學名額為 40%以上，公立高職提供免試入學名額為 60%以上，公私立五專及私立高中職提供免試入學名額皆以 70%以上，其篩選標準以學生在校的學習表現為參考依據（「擴大高中職及五專免試入學實施方案」，2009），例如：基北區免試入學採記學生前五學期學業成績全校排名百分比（「基北區 100 學年度高中職免試入學簡章彙編」，2011），所以學校定期評量變成「高風險」（high stake）

的測驗工具，然而卻鮮少研究針對學校定期評量進行差別試題功能之分析。

因此，學校的定期評量（段考）試題內容是否隱藏 DIF 現象或 DIF 試題的類型就顯得重要。鑒此，本研究包括兩部分，其一，未經任何配組程序，分析男、女性別自然科成就差異；其二，以自然科分數將男女生能力配組後，以 IRT Rasch 模式與 Mantel-Haenszel 法分析 DIF 現象。

# 貳、文獻探討

## 一、自然科成就之性別差異

吳裕益、洪碧霞、徐綺穗與葉千綺（1993）曾以國中學生5千餘人為樣本，比較男、女性在國文、數學和理化（國二和國三）三科的學業成就，分析結果顯示女性在國文科表現顯著優於男性，但在數學和理化的表現則男性優於女性。相反地，在國際性學生學習成就調查研究TIMSS 2003方面，國際間八年級生科學整體表現來說，男性平均量尺分數顯著高於女性，不過我國八年級男、女性的科學整體表現並未達顯著差異，科學分項表現來說，男性在地球科學的表現顯著高於女性，女性在化學的表現顯著高於男性，在生命科學、物理和環境科學等三項上則沒有顯著差異（邱美虹，2005）。從縱貫資料來看，Willingham、Cole、Lewis與Leung（1997）針對美國境內相關自然科學測驗分析四、八和十二年級之性別表現變化趨勢男、女之效果量和變異程度，結果發現三個年級的男性表現皆高於女性，四年級和八年級的性別效果量值約 -.10 ，在十二年級則擴大為 -.25 ，顯示性別差距隨年級增長略增加。由上述可知，自然科成就的性別差異的現象不具穩定性，可能會受到受試者的年級、樣本選取、測驗構念或能力等因素影響。此外，Willingham與Cole（1997）指出傳統運用的平均數比較男女差異，無法窺見男女差異的全貌，因此本研究將男、女學生採高、低成就分組分析，並以效果量、女／男標準差比和女／男人數比三項細部指標作自然科性別成就之差異比較。

## 二、差別試題功能（Differential Item Functioning, DIF）

DIF（Differential Item Functioning, DIF）係指對於在欲測量特質上已相配對的不同群體而言，DIF 是一種意料之外的測驗表現差異( Dorans & Holland, 1993 )。具體而言，DIF 是指根據測驗測量之構念分數將兩組受試群體加以配組（matched）後，兩組受試在試題表現上的差異。

在試題反應理論（Item Response Theory, IRT）的單向性及局部獨立性等兩項假定之下，對於兩個來自不同群體的受試者而言，如果兩者的能力相當，則他們對於同一個試題的答對機率應該相等，所以從 IRT 的角度可將 DIF 定義為：來自不同次群體、具有相同能力的受試者，對於特定試題有不相等的答對/答錯機率（王文中與陳雪珠，1999；Dorans & Holland, 1993）。另 Lord（1980）認為不同受試團體下，一個試題若具有不同的試題特徵曲線，則此試題具有 DIF。整體論之，DIF 來自於試題和受試群體之間的交互作用，使得試題對甲群體的難度高於乙群體（林奕宏與林世華，2004）。

## 三、自然科測驗之性別 DIF 研究

國內實徵研究中，對於自然科的 DIF 研究明顯少於數學科，筆者推測可能原因在於自然科的子領域（生命科學、物理、化學、地球科學、科技等）多於數學科（算術運算、代數、幾何、統計機率等），而自然科子領域彼此連貫程度低於數學科子領域，又自然科常有牽涉數學運算內容，因此相對性來說數學科之影響變因較為單純。早期針對大型評量的自然科性別 DIF 研究，僅見 Wang（1995）之研究，他利用 SIBTEST 程序調查 1991 年、1992 年與 1993 年大學聯考生物科試題（單選題及複選題）之性別 DIF 現象，結果發現在三種不同計分方式下，每年試題近半數皆呈現性別 DIF 現象，但是研究者並未進一步探討其可能原因。

近幾年來，國內多篇針對大型自然科測驗的性別DIF議題的研究都一致發現測驗有出現DIF現象但尚不構成試題偏誤，例如：盧雪梅與毛國楠（2008a）曾針對2001至2005年度「國民中學學生基本學力測驗（簡稱基測）」自然科之性別差異和差別試題功能進行研究，他們以

Mantel-Haenszel法分析發現575個試題中有52題出現DIF，DIF的出現率約為9%，健康教育、化學和生物DIF題有利女性居多；地球科學和物理DIF題則有利男性居多；臺師大心測中心（2008）以Mantel-Haenszel法和Logistic Regression法針對95至97年國中基測抽樣實徵資料進行DIF分析，自然科95年有3題有性別DIF、96年有4題有性別DIF、97年有3題有性別DIF，平均佔總題數約9%，結果顯示基測自然科試題微量的DIF比率並未影響測驗結果。

上述研究資料分別使用不同的檢定方式來探討性別 DIF 現象。研究上 DIF 檢定方法可以分為兩類：(1) IRT 取向，例如：Lord 的卡方考驗法、試題參數比較法、兩團體 IRF 或 ICC 區域面積法、概率比檢定法（likelihood ratio test, LR-IRT）等；(2) 非 IRT 取向，例如：Mantel-Haenszel 法、標準化法（standardization）、邏輯迴歸分析（logistic regression, LR）、SIBTEST（simultaneous item bias test）等（余民寧與謝進昌，2006）。本研究係以兩種不同取向的 DIF 檢定方法來進行研究：(1) Mantel-Haenszel 法、(2) 試題反應理論（item response theory, IRT）取向的試題參數差異比較法，進行自然科成就測驗的性別 DIF 檢定。

## 參、研究目的

根據前述研究動機和相關文獻探討，本研究目的如下：

一、分析學校自然科定期評量之性別成就差異。

二、分析學校自然科定期評量之性別DIF現象及出現率並探究性別DIF與試題特徵的可能關聯。

# 肆、研究方法

## 一、資料來源

本研究分析新北市某國中99學年度第2學期第二次定期評量八年級考生答題資料，該資料係向該國中教務處申請並取得該測驗命題教師同意，該校提供全體449名學生答題資料。研究者審視原始資料，將有明顯心向反應或嚴重漏答的受試資料剔除後，最後受試為382名，其中女性191名、男性191名。

## 二、研究工具

本研究之工具為新北市某國中自然科成就測驗（定期評量試題），出題範圍共涵蓋國中八年級康軒版自然課本第3冊第2章第五節「酸與鹼的反應」、第3章「氧化與還原」及第4章「反應速率與平衡」（康軒文教，2011）。資料處理以ConQuest統計程式進行IRT Rasch模式估計試題難度參數與性別DIF檢定，且使用IBM SPSS Statistics 19統計程式以Mantel-Haenszel法進行性別DIF檢定。

（一）測驗題型內容

測驗為50題四選一的選擇題（單選），答對1題得2分，總分100分。測驗作答時間45分鐘，題型內容取自康軒版自然課本第3冊課本及習作。

（二）試題分析

1. 效度分析

本研究參考國中生基本學力測驗工作推動小組所發行的「飛揚」第十三期「國中基本學力測驗自然科試題之設計理念」（國中生基本學力測驗工作推動小組，2002）一文將認知能力分成四部分：(1) 具備自然科學的基礎知識：主要包括基本的名詞、符號、科學現象、規則工具等之認識；(2) 運用資料和圖表的能力：以學生能理解數據或圖形、選用適當資料、轉換資料與圖表之能力為主；(3) 具備高層次思考的能力：主要是問題解決、分析比較關係與提出推論等；(4) 統整學科知識的能

力：主要是綜合各科相關知識、瞭解人類與環境之關係等。內容指標則依據康軒版自然第三冊第二章第五節、第三章及第四章的章節分類。根據上述的分類架構，作者與命題教師共同討論得到該測驗的雙向細目表（表 2）。

表 2 自然科成就測驗雙向細目表分析

| 內容指標 | 認知能力 | | | | |
|---|---|---|---|---|---|
| | 具備自然科學的基礎知識 | 運用資料和圖表的能力 | 具備高層次思考的能力 | 統整學科知識的能力 | 總和（題數） |
| 2-5 酸與鹼的反應 | 25, 26 | -- | 27,29,30 | 28 | 6 |
| 3-1 氧化反應 | 1, 2, 3,6,7,9 | -- | 4,8,11,13 | | 10 |
| 3-2 氧化與還原反應 | 14,15,16,18,19, 20,21,23, | -- | 10,12,17 | 5,22,24 | 14 |
| 4-1 反應速率 | 32,33,34,36,39, 40,41,42,43,44, | -- | 31,35,37, 38,45, | | 15 |
| 4-2 可逆反應與平衡 | 48,49 | -- | 50 | 46,47 | 5 |
| 總和（題數） | 28 | 0 | 16 | 6 | 50 |

**註：**本雙向細目表中之數字代表本研究成就測驗之試題題號

2. 信度分析

研究者將受試於每題試題答對計為「1」、答錯計為「0」進行分析信度分析，結果顯示該成就測驗 Cronbach's $\alpha$ 為.95，顯示測驗具備高度內部一致性。

3. 難度與鑑別度分析

試題難度 *P* 值若以「0.75 以上為容易、介於 0.50 至 0.75 之間為普通、低於 0.5 為困難」之標準來檢視，研究發現難度 *P* 值在 0.75 以上的計有 6 題、介於 0.50 至 0.75 計有 31 題、低於 0.50 計有 13 題，顯示測驗整體難度為中間普通。

鑑別度 D 值（前 27%高分組答對率－後 27%低分組答對率）若以「高於 0.30 為良好試題」為檢視標準，結果發現只有試題 29 為 0.30，其於 49 題試題皆高於 0.30，顯示該成就測驗具備良好的鑑別度。

## 三、研究程序及分析方法

（一）性別成就差異比較分析

由於直接比較全體受試之平均數差異之量數可能會犯辛普森悖論（Simpson's Paradox）之風險，因此，本研究參考Willingham與Cole（1997）以及盧雪梅與毛國楠（2008a）的分析方法，除了分析全體受試者之性別成就差異外，另將受試分為兩組：高成就組（前10%）和低成就組（後10%）（取10%的落點分數的所有人數，人數可能會略超過總人數的10%），並針對此三組進行三項差異比較量數分析：效果量（D）、女／男標準差比值和女／男人數比值。效果量計算公式（式(1)）如下：

$$D=\frac{M_{女}-M_{男}}{\sqrt{S^2{}_{pooled}}}\text{.....式(1)}$$

（二）DIF檢定- IRT Rasch模式

在社會科學的研究中，受試者的能力估計和題目的難度估計往往是彼此互相干擾，例如：受試者的程度的優劣，取決於測驗的特性，是測驗依賴（test dependent），但另一方面試題的難易程度，亦取決於受試者樣本的特性，是樣本依賴（sample-dependent），往往兩者交互作用下無法得到「等距」、「客觀」的量尺（王文中，2004）。為了解決這個問題，1960 年丹麥數學家 Georg Rasch 提出了 Rasch 測量模式（Rasch, 1960），假設受試者 $i$ 的能力值為 $\theta'_i$，試題 $j$ 的難度為 $b'_j$，作答為二元計分，Rasch 定義勝率（odds）為 $P_{ij1}/P_{ij0}$（$P_{ij1}+P_{ij0}=1$）。在 Rasch 模式中，假定 $odds_{ij}\equiv P_{ij1}/P_{ij0}=\theta'_i/b'_j$，若有兩位受試者 A 和 B，則可推導出受試者 A 和 B 於試題 $j$ 上的能力比值與題目特性無關且為比率量尺（$odds_{Aj}/odds_{Bj}=(\theta'_A/b'_j)\div(\theta'_B/b'_j)=\theta'_A/\theta'_B$）。若對 $odds_{ij}$ 取自然對數 ln，則可推導出受試者 A 和 B 於試題 $j$ 上的能力差距與題目特性無關且為等距量尺

$(\ln(odds_{Aj})-\ln(odds_{Bj})=[\ln(\theta'_A)-\ln(b'_j)]-[\ln(\theta'_B)-\ln(b'_j)]=\ln(\theta'_A)-\ln(\theta'_B))$

。最後，假設 $\ln(\theta'_i)\equiv\theta_i$、$\ln(b'_j)\equiv b_j$、$\text{logit}_{ij}=\theta_i-b_j$，可推導出受試者

答對某個試題的機率公式（式(2)）。

$$P_{ij}=\frac{e^{(\theta_i-b_j)}}{1+e^{(\theta_i-b_j)}}\text{.....式(2)}$$

在試題反應理論（Item Response Theory, IRT）的單向性及局部獨立性等兩項假定之下，對於兩個來自不同群體的受試而言，如果兩者的能力相當，則他們對於同一個試題的答對機率應該相等。換句話說，以Rasch 模式（為 IRT 理論中的一種模式）檢定來自不同群體，但具有相同能力的受試者，若發現對於某特定試題有不相等的答對機率則代表該試題有 DIF 現象。因此，本研究以 Rasch 模式檢定相同能力的女性和男性在同一試題下的 IRT 難度參數之差異，作為 DIF 試題檢定方法之一。

（三）DIF檢定-Mantel-Haenszel法

Mantel-Haenszel法通常以測驗總分為配組變項，本研究以男性為參照組（reference group）、女生為焦點組（focal group）進行Mantel-Haensze DIF分析。然而，若測驗中有DIF試題存在，測驗總分本身可能存有偏差，Holland與Thayer（1988）建議在這種情況下進行「淨化」（purification），也就是將DIF試題排除於配組變項外，再重新計算總分作為新配組變項，一直重複以上步驟，直到獲得一組完全無DIF的試題為止，接著以該總分作為研究之實際配組變項後，再進行正式DIF檢定。

Mantel-Haenszel法是一種列聯表之分析法，本研究依據總得分將群體分為六個分數層，經由前述淨化程序後，各個分數層之受試者的答題表現之人數分配可整理為一個2×2的列聯表（表3）。再者，Mantel-Haenszel法的統計虛無假設為：k個分數層的參照組和焦點組的共同勝算比（common odds-ratio）參數$\hat{\alpha}_{MH}$為1.0，$\hat{\alpha}_{MH}$的參數估計值如下（式(3)）

$$\hat{\alpha}_{MH}=\frac{\sum_k \frac{A_k D_k}{T_k}}{\sum_k \frac{B_k C_k}{T_k}}\text{.....式(3)}$$

表3 第K分層2×2列聯表

| 組別 | 得分 | | 合計 |
|---|---|---|---|
| | 1 | 0 | |
| 參照組（男） | $A_k$ | $B_k$ | $N_{Rk}$ |
| 焦點組（女） | $C_k$ | $D_k$ | $N_{Fk}$ |
| 合計 | $M_{1k}$ | $M_{0k}$ | $T_k$ |

由於統計顯著性考驗結果容易受到受試樣本人數多寡的影響，加上為了方便解釋，研究上通常會將所得 $\hat{\alpha}_{MH}$ 值取自然對數後乘上某一常數，轉換為另一種形式的DIF量數（美國ETS公司之難度量尺），稱為MH D-DIF（以下用$\Delta_{MH}$表示之），轉換公式如式(4)，其中，$\Delta_{MH}$的標準誤如式(5)。余民寧與謝進昌（2006）指出，ETS DIF分類系統較不受樣本因素影響，相對性可得到較客觀和可信的DIF指標。所以，本研究採$\Delta_{MH}$參數作為DIF判準的主要依據，當$\Delta_{MH}$值為負時，表示試題對於參照組（男性）而言較簡單，當$\Delta_{MH}$值為正時，表示試題對於焦點組（女性）而言較簡單。

$$\Delta_{MH} = -2.35\ln(\hat{\alpha}_{MH}) \text{.....式(4)}$$

$$SE(\Delta_{MH}) = 2.35\sqrt{\frac{\sum_k [A_k D_k + \alpha_{MH} B_k C_k] + [A_k + D_k + \alpha_{MH}(B_k + C_k)]}{2(\sum_k \frac{A_k D_k}{T_k})^2}} \text{.....式(5)}$$

# 伍、研究結果與討論

## 一、性別成就差異比較分析

本研究以考生的原始分數進行分析，答對1題計得2分，答錯1題計得0分，滿分100分。Willingham與Cole（1997）指出傳統運用的平均數

（效果量）比較男女差異，並無法窺見男女差異的全貌，他們另外提出兩種指標：女／男標準差比值（Standard Deviation Ratio，簡寫SDR）和女／男人數比值（Female and Male Ratio，簡寫F/M）。此外，在不同成就水準之性別差距也可能不一致（盧雪梅與毛國楠，2008a），因此，本研究參考Willingham與Cole（1997）以及盧雪梅與毛國楠（2008a）的分析方法，分別分析全體受試、高成就組（前10%）和低成就組（後10%）之三項差異比較量數：效果量（D）、女／男標準差比值和女／男人數比值。

根據Cohen（1988）的判準建議，效果量D =.20為小等級、D=.50為中等級、D = .80為大等級，全體學生性別效果量低於小等級，高分組和低分組的性別效果量達到小至中等級（$D_{高}$＝-0.417、$D_{低}$＝-0.232）且高分組和低分組D值為負值，顯示在此兩組中男性的表現較佳，其中以高分組性別效果量最為明顯（表4）。

其次，女／男標準差（SDR 值）部分，全體組和低分組男、女學生的SDR值接近1，顯示男女生分數的變異程度相似，然而高分組男女生分數的變異程度稍大（SDR＝1.261），女性分數的變異情形高於男性。

最後，女／男人數比（F/M值）部分，無論全體組、高分組或低分組的男、女比例都很接近。

表 4 自然科成就測驗性別效果量、女／男標準差比和女／男人數比

| | 女性 | | 男性 | | 效果量 | 女／男 | 女／男 |
|---|---|---|---|---|---|---|---|
| | 平均數 | 標準差 | 平均數 | 標準差 | | 標準差比 | 人數比 |
| 全體 | 58.41 | 24.88 | 59.88 | 25.31 | -0.059 | 0.981 | 1.000 |
| 高分組 | 95.10 | 2.86 | 96.19 | 2.27 | -0.417 | 1.261 | 0.952 |
| 低分組 | 22.48 | 3.80 | 23.33 | 3.52 | -0.232 | 1.078 | 1.042 |

**註：**女／男標準差比係指女生標準差對男性標準差的比值；女／男人數係指女生人數對男性人數的比值。

## 二、性別 DIF 分析

（一）試題反應理論（IRT）取向 DIF 分析

1. Rasch 模式適配性檢定

首先，Reckase（1979）建議可以使用主成分分析法來檢驗 IRT 單向度的假設，也就是考量受試者在測驗上的表現是否受到一個最主要成分或因素的影響。所以，研究者將測驗試題進行主成分分析（principal component analysis），再參照 Reckase（1979）所提出的單一向度的兩個標準：第一特徵值（$\lambda_1$）佔總變異量的 20%以上、第一特徵值（$\lambda_1$）與第二特徵值（$\lambda_2$）的比大於 4，作為判定單一向度的標準。由表 5 可以看出自然科成就測驗皆符合 Reckase（1979）建議的標準，亦即符合單向度假設。

其次，由於在使用 IRT 模式進行有關的參數估計之前，需先確認作答反應資料與模式適配的情形是否合理。本研究使用 ConQuest 軟體進行 Rasch 模式適配性分析，在適配指標上，ConQuest 軟體未加權（unweighted）及加權（weighted）兩種適配指標 MNSQ（mean squares）值。根據 Linacre 與 Wright（1994）的看法，對於評定量尺而言，MNSQ 在 0.6～1.4 之間是合理的，代表受試者在題項上的反應符合模式預期的範圍。由表 6 可知測驗試題中只有 13 題的未加權 MNSQ 不符合標準，但是此 13 題的加權 MNSQ 符合標準，依據 Chien（2006）建議當未加權 MNSQ 和加權 MNSQ 互有高低時，以加權 MNSQ（Infit MNSQ）為認定標準，因此顯示本研究所有試題皆符合 Rasch 模式假定。

表 5 自然科成就測驗之主成分分析摘要表

| | 題數 | 第一特徵值（$\lambda_1$） | 第二特徵值（$\lambda_2$） | $\lambda_1/\lambda_2$ | $\lambda_1$ 佔總變異比率 |
|---|---|---|---|---|---|
| 自然科成就測驗 | 50 | 14.697 | 2.287 | 6.426 | 29.394% |

表 6 各試題題與 Rasch 模式適配情形

| 試題 | 難度參數 | unweighted fit | | weighted fit | |
|---|---|---|---|---|---|
| | | MNSQ | *t* | MNSQ | *t* |
| 1 | 0.243(0.125) | 0.92 | -1.1 | 0.92 | -1.5 |
| 2 | -0.009(0.126) | 0.66 | -5.3 | 0.76 | -4.7 |
| 3 | -0.818(0.132) | 0.76 | -3.6 | 0.90 | -1.7 |
| 4 | -0.561(0.129) | 0.93 | -0.9 | 0.97 | -0.5 |
| 5 | 0.746(0.126) | 1.48 [a] | 5.8 | 1.36 | 5.6 |
| 6 | -1.127(0.136) | 0.58 [a] | -6.9 | 0.79 | -3.4 |
| 7 | -0.924(0.133) | 0.59 [a] | -6.6 | 0.82 | -3.1 |
| 8 | -1.186(0.138) | 0.72 | -4.2 | 0.90 | -1.6 |
| 9 | -0.662(0.130) | 0.59 [a] | -6.6 | 0.77 | -4.3 |
| 10 | -0.281(0.127) | 0.78 | -3.2 | 0.89 | -2.1 |
| 11 | 0.196(0.125) | 1.19 | 2.5 | 1.13 | -2.2 |
| 12 | -0.395(0.128) | 0.73 | -4.2 | 0.88 | -2.1 |
| 13 | 2.098(0.138) | 2.09 [a] | 11.6 | 1.18 | 2.5 |
| 14 | -0.765(0.131) | 0.88 | -1.7 | 0.97 | -0.4 |
| 15 | -0.478(0.128) | 0.66 | -5.3 | 0.82 | -3.3 |
| 16 | 0.761(0.126) | 1.02 | 0.3 | 1.00 | -0.0 |
| 17 | 1.001(0.127) | 1.56 [a] | 6.7 | 1.34 | 5.3 |
| 18 | -0.169(0.126) | 0.83 | -2.4 | 0.96 | -0.8 |
| 19 | -0.748(0.131) | 0.85 | -2.2 | 0.95 | -0.8 |
| 20 | -0.395(0.128) | 0.75 | -3.8 | 0.83 | -3.2 |
| 21 | -0.527(0.129) | 0.90 | -1.4 | 0.87 | -2.5 |
| 22 | 1.296(0.128) | 1.67 [a] | 7.8 | 1.25 | 3.9 |
| 23 | -0.428(0.128) | 0.97 | -0.4 | 1.02 | 0.3 |
| 24 | 0.384(0.125) | 0.91 | -1.2 | 0.94 | -1.1 |
| 25 | 0.921(0.126) | 0.92 | -1.2 | 0.88 | -2.1 |
| 26 | 0.777(0.126) | 1.26 | 3.4 | 1.13 | 2.2 |
| 27 | 0.841(0.126) | 1.44 [a] | 5.3 | 1.28 | 4.5 |
| 28 | 0.588(0.125) | 1.11 | 1.5 | 1.08 | 1.4 |

| | | | | | |
|---|---|---|---|---|---|
| 29 | 1.481(0.130) | 1.78 [a] | 8.8 | 1.38 | 5.6 |
| 30 | 1.001(0.127) | 1.33 | 4.1 | 1.17 | 2.8 |
| 31 | 0.509(0.125) | 1.21 | 2.7 | 1.15 | 2.6 |
| 32 | -1.407(0.143) | 0.85 | -2.2 | 0.91 | -1.2 |
| 33 | -0.714(0.131) | 0.66 | -5.4 | 0.81 | -3.5 |
| 34 | -0.395(0.128) | 1.09 | 1.3 | 1.06 | 1.1 |
| 35 | 0.148(0.125) | 0.80 | -3.0 | 0.87 | -2.4 |
| 36 | -0.185(0.127) | 1.40 | 4.9 | 1.18 | 3.1 |
| 37 | 1.834(0.134) | 2.04 [a] | 11.1 | 1.24 | 3.5 |
| 38 | -0.105(0.126) | 0.73 | -4.1 | 0.80 | -3.8 |
| 39 | -0.169(0.126) | 0.90 | -1.4 | 0.89 | -2.1 |
| 40 | -0.997(0.134) | 0.59 [a] | -6.7 | 0.81 | -3.3 |
| 41 | -0.818(0.132) | 0.78 | -3.2 | 0.86 | -2.4 |
| 42 | -0.041(0.126) | 0.73 | -4.0 | 0.85 | -2.9 |
| 43 | -1.127(0.136) | 0.54 [a] | -7.8 | 0.79 | -3.5 |
| 44 | -1.427(0.143) | 0.54 [a] | -7.8 | 0.81 | -2.7 |
| 45 | -0.009(0.126) | 0.74 | -4.0 | 0.81 | -3.6 |
| 46 | 0.070(0.126) | 0.82 | -2.6 | 0.89 | -2.0 |
| 47 | 0.857(0.126) | 1.40 | 4.9 | 1.25 | 4.0 |
| 48 | -0.281(0.127) | 1.22 | 2.9 | 1.15 | 2.5 |
| 49 | -0.105(0.126) | 0.91 | -1.2 | 0.97 | -0.5 |
| 50 | 0.384(0.125) | 1.07 | 0.9 | 1.07 | 1.3 |
| Chi-square =1913.79 *df* = 50 Sig. Level = .000 | | | | | |

**註：**上標 a 表示 MNSQ 值＜0.6 或＞1.40

2. 試題的 Rasch 模式下的難度參數估計值

本研究採用試題作答理論中的 Rasch 模式估計試題的難度參數。從圖 1 中顯示測驗整體難度參數介於 -1 和 1 之間的有 41 題（82%），難度參數高於 1 的有 4 題（8%）、難度參數低於 -1 的有 5 題（10%）。若將 IRT 難度參數高於 1 的 4 個試題與傳統測驗難度 *P* 值比較，結果顯示試題 13、試題 22、試題 29、試題 37 的 *P* 值皆低於 0.40。同理，若將

IRT 難度參數低於 -1 的 5 個試題與傳統測驗難度 *P* 值比較，結果顯示試題 6、試題 8、試題 32、試題 43、試題 44 的難度 *P* 值皆高於 0.75，與傳統測驗分析結果一致。因此，該測驗試題的整體難度為中間普通。

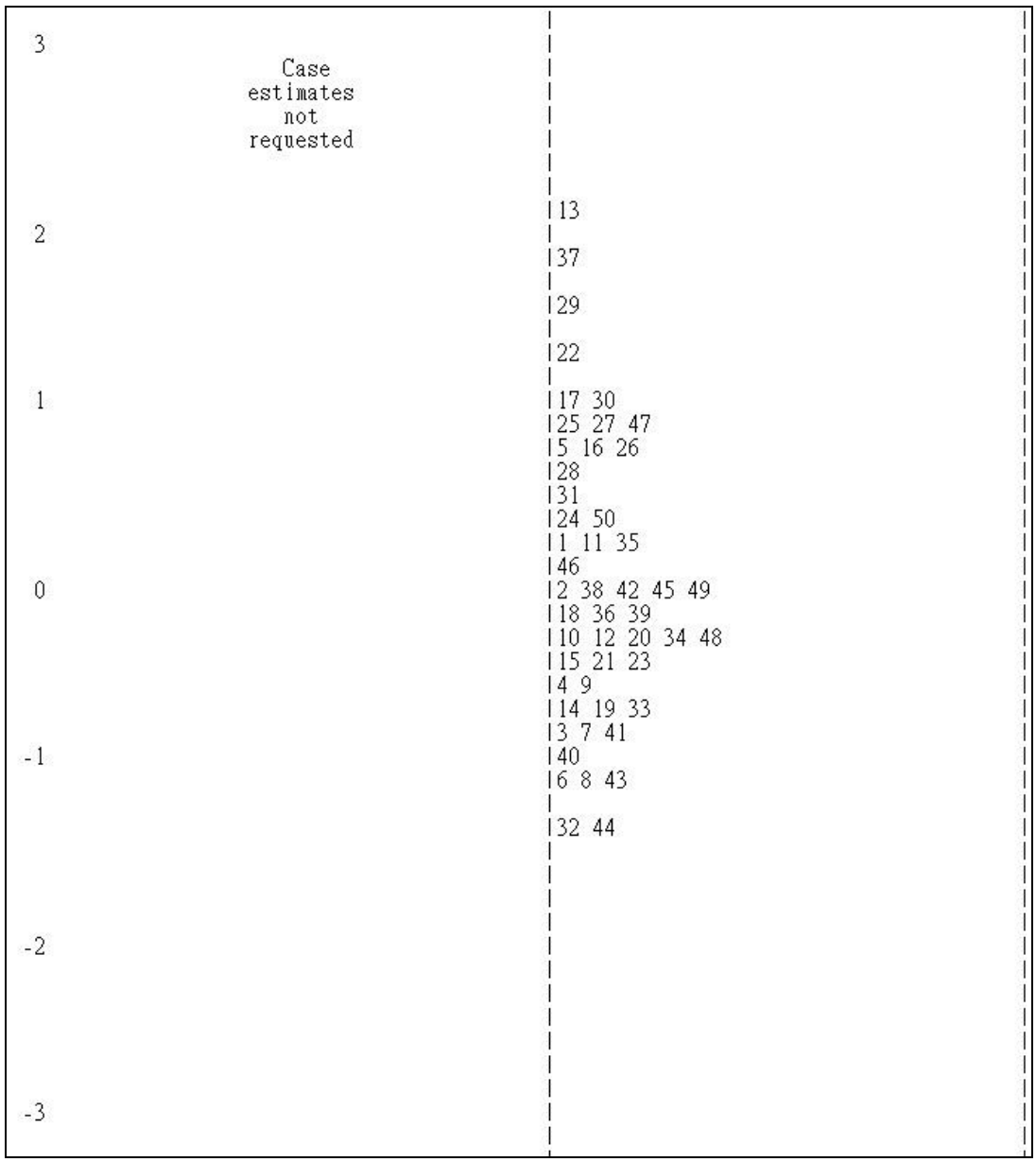


圖 1　自然科成就測驗 IRT 單參數模式之試題難度參數分配圖

3. 自然科成就測驗中試題與性別的交互作用（DIF 現象）

首先，關於自然科成就測驗男、女兩群體的能力參數估計值，由表7可知在自然科成就測驗中，女性的能力估計值為 -0.005，約略為估計標準誤0.018的 -0.278倍，參考Wu、Adams與Wilson（1998）的建議能力估計值與估計標準誤的比值絕對值大於3時，才達.05顯著水準，故本研究的女性及男性兩群體的能力差異沒有差異。

表7 女性及男性的平均能力參數估計值及估計標準誤

| 性別 | 能力參數估計值 | 估計標準誤 | 比值 |
|---|---|---|---|
| 女性 | -0.005 | 0.018 | -0.278 |
| 男性 | 0.005 | | |

第二，王文中與陳雪珠（1999）提出DIF效果可區分為A、B、C 三個等級，將其換算為logit尺度而言，則相當於若難度差異值低於0.4 logits以下的試題，歸屬於A類，即表示幾乎沒有DIF存在，若難度差異值超過0.6 logits的試題，則屬於C類，有顯著DIF效果存在，其餘則歸類為B類。由表8可知，整體各試題的性別差異值之總和為 -0.148，顯示整份測驗沒有顯著性別DIF現象。然而，進一步看各試題，A類DIF有37題、B類DIF有11題（5題微量有利女性、6題微量有利男性）、C類DIF有2題（試題31與試題50，皆有利女性）。

表 8 自然科定期評量試題性別 DIF 檢定摘要表（IRT）

| 試題 | 總難度參數 | 女性難度參數 | 男性難度參數 | 性別差異值 | 有利方向 |
|---|---|---|---|---|---|
| 1 | 0.243(0.125) | 0.015(0.125) | -0.015(0.125) | 0.030 | |
| 2 | -0.009(0.126) | -0.206(0.126) | 0.206(0.126) | -0.412 [b] | F |
| 3 | -0.818(0.132) | 0.010(0.133) | -0.010(0.133) | 0.020 | |
| 4 | -0.561(0.129) | -0.077(0.130) | 0.077(0.130) | -0.154 | |
| 5 | 0.746(0.126) | -0.020(0.126) | 0.020(0.126) | -0.040 | |
| 6 | -1.127(0.136) | -0.080(0.138) | 0.080(0.138) | -0.160 | |
| 7 | -0.924(0.133) | -0.048(0.134) | -0.048(0.134) | -0.096 | |
| 8 | -1.186(0.138) | 0.094(0.139) | -0.094(0.139) | 0.188 | |
| 9 | -0.662(0.130) | -0.146(0.131) | 0.146(0.131) | -0.292 | |
| 10 | -0.281(0.127) | -0.029(0.127) | 0.029(0.127) | -0.058 | |
| 11 | 0.196(0.125) | -0.095(0.126) | 0.095(0.126) | -0.190 | |
| 12 | -0.395(0.128) | 0.087(0.128) | -0.087(0.128) | 0.174 | |
| 13 | 2.098(0.138) | 0.078(0.140) | -0.078(0.140) | 0.156 | |
| 14 | -0.765(0.131) | -0.218(0.132) | 0.218(0.132) | -0.436 [b] | F |
| 15 | -0.478(0.128) | 0.006(0.129) | -0.006(0.129) | 0.012 | |
| 16 | 0.761(0.126) | 0.251(0.126) | -0.251(0.126) | 0.502 [b] | M |
| 17 | 1.001(0.127) | 0.239(0.128) | -0.239(0.128) | 0.478 [b] | M |
| 18 | -0.169(0.126) | 0.051(0.127) | -0.051(0.127) | 0.102 | |
| 19 | -0.748(0.131) | 0.009(0.132) | -0.009(0.132) | 0.018 | |
| 20 | -0.395(0.128) | 0.153(0.128) | -0.153(0.128) | 0.304 | |
| 21 | -0.527(0.129) | -0.044(0.129) | 0.044(0.129) | -0.088 | |
| 22 | 1.296(0.128) | -0.290(0.129) | 0.290(0.129) | -0.580 [b] | F |
| 23 | -0.428(0.128) | -0.077(0.128) | 0.077(0.128) | -0.154 | |
| 24 | 0.384(0.125) | -0.191(0.125) | 0.191(0.125) | -0.382 | |
| 25 | 0.921(0.126) | 0.124(0.127) | -0.124(0.127) | 0.248 | |
| 26 | 0.777(0.126) | 0.044(0.126) | -0.044(0.126) | 0.088 | |
| 27 | 0.841(0.126) | 0.044(0.126) | -0.044(0.126) | 0.088 | |
| 28 | 0.588(0.125) | 0.202(0.126) | -0.202(0.126) | 0.404 [b] | M |
| 29 | 1.481(0.130) | 0.167(0.131) | -0.167(0.131) | 0.334 | |
| 30 | 1.001(0.127) | -0.215(0.127) | 0.215(0.127) | -0.430 [b] | F |
| 31 | 0.509(0.125) | -0.318(0.126) | 0.318(0.126) | -0.636 [c] | F |
| 32 | -1.407(0.143) | -0.041(0.144) | 0.041(0.144) | -0.082 | |
| 33 | -0.714(0.131) | -0.234(0.132) | 0.234(0.132) | -0.468 [b] | F |
| 34 | -0.395(0.128) | -0.045(0.128) | 0.045(0.128) | -0.090 | |
| 35 | 0.148(0.125) | 0.205(0.126) | -0.205(0.126) | 0.410 [b] | M |
| 36 | -0.185(0.127) | 0.035(0.127) | -0.035(0.127) | 0.070 | |
| 37 | 1.834(0.134) | -0.040(0.135) | 0.040(0.135) | -0.080 | |
| 38 | -0.105(0.126) | 0.178(0.127) | -0.178(0.127) | 0.356 | |
| 39 | -0.169(0.126) | 0.244(0.127) | -0.244(0.127) | 0.488 [b] | M |

| | | | | | |
|---|---|---|---|---|---|
| 40 | -0.997(0.134) | -0.060(0.136) | 0.060(0.136) | -0.120 | |
| 41 | -0.818(0.132) | -0.025(0.133) | 0.025(0.133) | -0.050 | |
| 42 | -0.041(0.126) | -0.015(0.126) | 0.015(0.126) | -0.030 | |
| 43 | -1.127(0.136) | 0.111(0.138) | -0.111(0.138) | 0.222 | |
| 44 | -1.427(0.143) | 0.149(0.145) | -0.149(0.145) | 0.298 | |
| 45 | -0.009(0.126) | 0.017(0.126) | -0.017(0.126) | 0.034 | |
| 46 | 0.070(0.126) | 0.095(0.126) | -0.095(0.126) | 0.190 | |
| 47 | 0.857(0.126) | 0.092(0.127) | -0.092(0.127) | 0.184 | |
| 48 | -0.281(0.127) | 0.231(0.128) | -0.231(0.128) | 0.462 [b] | M |
| 49 | -0.105(0.126) | -0.110(0.126) | 0.110(0.126) | -0.220 | |
| 50 | 0.384(0.125) | -0.380(0.126) | 0.380(0.126) | -0.760 [c] | F |
| Chi-square =68.85 *df* = 50 Sig. Level = .040 | | | | | |

**註：**1. 上標 b 表示性別差異絕對值介於 0.4~0.6

2. 上標 c 表示性別差異絕對值高於 0.6

3.「F」為有利女性，「M」為有利男性

（二）Mantel-Haensze DIF 檢定法

由於Mantel-Haenszel法通常以測驗總分為配組變項，若測驗中有DIF試題存在，測驗總分本身可能存有偏差。因此，本研究先進行配組變項淨化，再正式進行DIF分析。言下之意，首先以自然科原始分數將男女生配組（六組）後，進行DIF分析，若出現DIF題，則將DIF題排除在外，接著重新計算無DIF試題的總分，以該總分為新的配組變項，再次進行DIF分析，重複以上步驟，直到獲得完全無DIF試題，再以最後總分為配組變項，並進行正式的DIF分析。

本研究以男性為參照組（reference group）、女生為焦點組（focal group）進行Mantel-Haensze DIF分析，再將所得$\hat{\alpha}_{MH}$值轉換為另一種形式的DIF量數，稱為MH D-DIF（以下用$\Delta_{MH}$表示之）。接著，依據ETS的標準將$\Delta_{MH}$分為三類：A類代表不顯著或輕微的DIF，B類代表中度DIF，C類則為重度DIF（若試題之$\Delta_{MH}$值未顯著異於0或$\Delta_{MH}$的絕對值小於1.0，將之歸於A類DIF；如果$\Delta_{MH}$的絕對值大於1.5且顯著大於1.0，則歸於C類DIF，其餘試題歸於B類DIF，前述統計檢定的顯著水準皆為.05）。若$\Delta_{MH}$值為負表示試題有利參照組（男性），若$\Delta_{MH}$ 為正表示試題有利焦點組（女生）。

從表 9 中可發現自然科定期評量試題，A 類 DIF 有 45 題、B 類 DIF

有 5 題（5 題皆有利女性），分別為試題 2、試題 22、試題 31、試題 33 及試題 50，沒有 C 類 DIF 試題。

表 9 自然科測驗試題性別 DIF 檢定摘要表（Mantel-Haenszel）

| 試題 | ETS 標準 | $\chi^2_{MH}$ | $\hat{\alpha}_{MH}$ | $\Delta_{MH}$ | SE（$\Delta_{MH}$） | 有利方向 |
|---|---|---|---|---|---|---|
| 1 | A | 0.020 | 0.932 | 0.165 | 0.616 | |
| 2 | B | 4.817* | 0.505* | 1.606 | 0.693 | F |
| 3 | A | 0.009 | 0.937 | 0.153 | 0.656 | |
| 4 | A | 0.391 | 0.823 | 0.458 | 0.609 | |
| 5 | A | 0.010 | 1.002 | -0.005 | 0.512 | |
| 6 | A | 0.775 | 0.723 | 0.762 | 0.738 | |
| 7 | A | 0.010 | 0.985 | 0.036 | 0.705 | |
| 8 | A | 0.080 | 1.134 | -0.296 | 0.691 | |
| 9 | A | 2.934 | 0.556 | 1.379 | 0.743 | |
| 10 | A | 0.135 | 0.878 | 0.306 | 0.616 | |
| 11 | A | 0.457 | 0.833 | 0.429 | 0.543 | |
| 12 | A | 0.051 | 1.105 | -0.235 | 0.644 | |
| 13 | A | 0.174 | 1.149 | -0.326 | 0.327 | |
| 14 | A | 2.839 | 0.621 | 1.120 | 0.618 | |
| 15 | A | 0.033 | 0.911 | 0.219 | 0.674 | |
| 16 | A | 2.373 | 1.493 | -0.942 | 0.573 | |
| 17 | A | 2.453 | 1.452 | -0.876 | 0.522 | |
| 18 | A | 0.006 | 1.057 | -0.130 | 0.618 | |
| 19 | A | 0.005 | 0.947 | 0.128 | 0.627 | |
| 20 | A | 0.551 | 1.282 | -0.584 | 0.660 | |
| 21 | A | 0.285 | 0.829 | 0.441 | 0.656 | |
| 22 | B | 3.673 | 0.617 | 1.135 | 0.557 | F |
| 23 | A | 0.435 | 0.821 | 0.463 | 0.588 | |
| 24 | A | 2.899 | 0.616 | 1.139 | 0.623 | |
| 25 | A | 0.319 | 1.206 | -0.440 | 0.630 | |
| 26 | A | 0.008 | 1.051 | -0.117 | 0.559 | |
| 27 | A | 0.071 | 1.089 | -0.200 | 0.529 | |
| 28 | A | 1.467 | 1.366 | -0.733 | 0.552 | |
| 29 | A | 1.106 | 1.297 | -0.611 | 0.611 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 30 | A | 2.177 | 0.690 | 0.872 | 0.550 | |
| 31 | B | 5.038* | 0.583* | 1.268 | 0.538 | F |
| 32 | A | 0.107 | 0.860 | 0.354 | 0.731 | |
| 33 | B | 4.902* | 0.497* | 1.643 | 0.700 | F |
| 34 | A | 0.123 | 0.891 | 0.271 | 0.571 | |
| 35 | A | 1.377 | 1.411 | -0.809 | 0.620 | |
| 36 | A | 0.001 | 1.036 | -0.083 | 0.541 | |
| 37 | A | 0.002 | 0.958 | 0.101 | 0.588 | |
| 38 | A | 0.844 | 1.358 | -0.719 | 0.679 | |
| 39 | A | 2.353 | 1.547 | -1.025 | 0.618 | |
| 40 | A | 0.631 | 0.744 | 0.695 | 0.731 | |
| 41 | A | 0.209 | 0.836 | 0.421 | 0.696 | |
| 42 | A | 0.170 | 0.857 | 0.363 | 0.656 | |
| 43 | A | 0.148 | 1.199 | -0.426 | 0.773 | |
| 44 | A | 0.659 | 1.420 | -0.824 | 0.834 | |
| 45 | A | 0.074 | 0.887 | 0.282 | 0.674 | |
| 46 | A | 0.045 | 1.095 | -0.213 | 0.620 | |
| 47 | A | 0.254 | 1.154 | -0.337 | 0.545 | |
| 48 | A | 2.252 | 1.464 | -0.896 | 0.552 | |
| 49 | A | 0.952 | 0.755 | 0.660 | 0.597 | |
| 50 | B | 7.947** | 0.500** | 1.629 | 0.559 | F |

**註：**ETS 對 DIF 效果量編碼意義：A 為可忽略，B 為中等，C 為嚴重 「F」為有利女性 「M」為有利男性

* $p < .05$ ** $p < .01$

（三）IRT Rasch Model 及 Mantel-Haensze 法之 DIF 交集討論

筆者以較嚴苛的方式於 Rasch Model 取 C 類 DIF 試題計有兩題，於 Mantel-Haensze 法取 B 類試題計有五題，並取兩者交集，也就是試題 31 與試題 50。從試題特徵曲線（Item-Characteristic curve, ICC）（圖 2 與圖 3）亦看出試題 31 和試題 50 有顯著性別 DIF 現象，該類試題佔總題數的 4%且皆有利於女性。

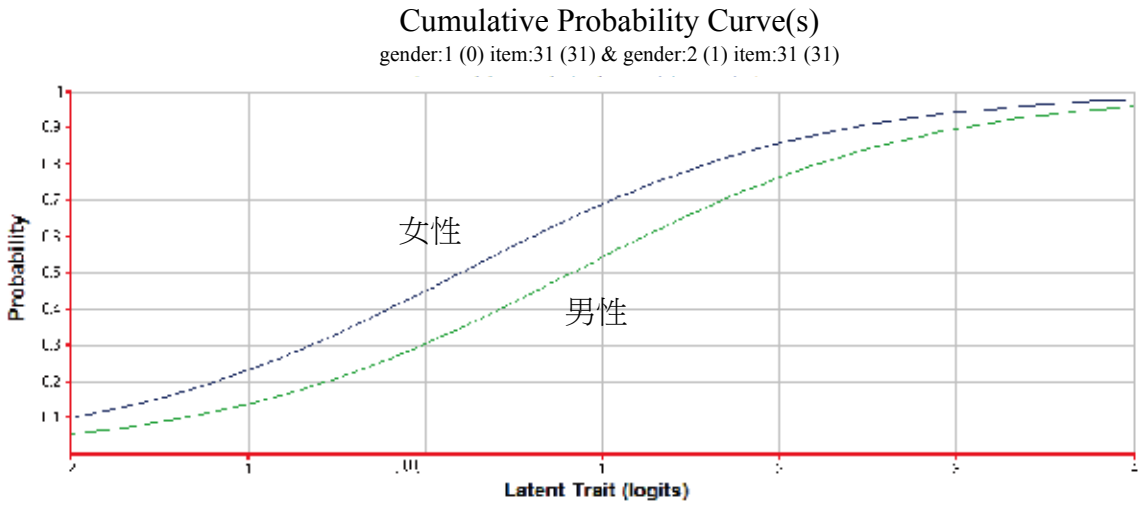


圖 2　男、女學生於試題 31 之 ICC 曲線

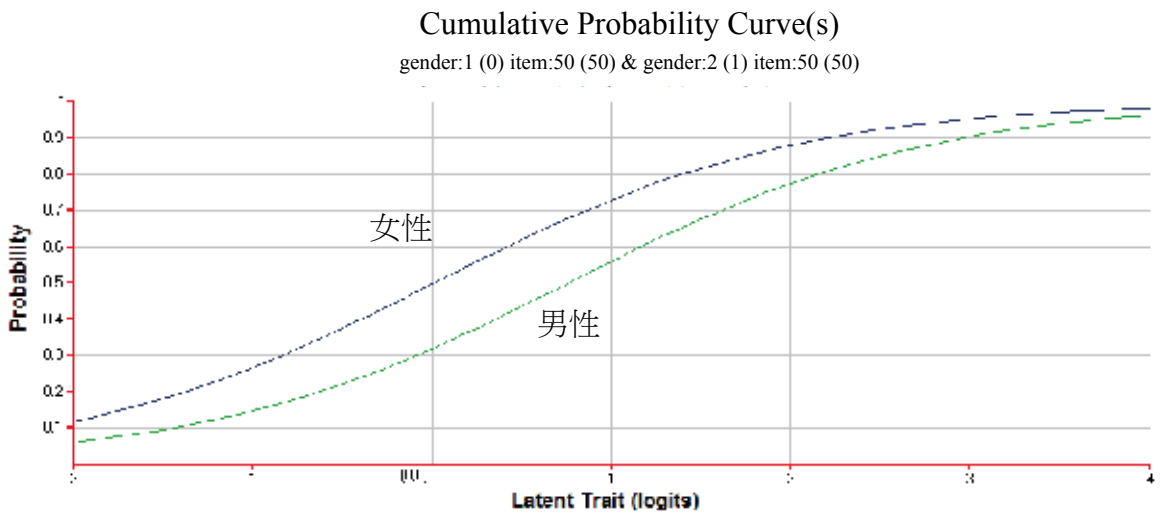


圖 3　男、女學生於試題 50 之 ICC 曲線

# 陸、結論與建議

## 一、學校自然科定期評量之內容修改建議

本研究測驗內部一致性Cronbach's $\alpha$為.95，命題題數以「具備自然科學的基礎知識」超過半數，以章節「3-2氧化與還原反應」和「4-1反應速率」各佔30%，大致符合在國中自然科章節內容的課程時數安排。

本研究測驗難度*P*值在0.75以上的計有6題、介於0.50至0.75計有31題、低於0.50計有13題，顯示測驗整體難度為中間普通，與國中基本學力測驗或北北基聯測提倡之理念一致。

本研究測驗鑑別度D值（前27%高分組答對率－後27%低分組答對率）只有試題29為0.30，其於49題試題皆高於0.30，顯示該成就測驗具備良好的鑑別度。進一步分析試題29（表10），結果發現試題29的難度偏難，中間能力與低分組考生無法從對於四個選項中判斷正確答案，對於高分組學生而言，選項C的誘答力頗大。研究者回溯國中自然課本第二章第五節酸與鹼的反應單元課文（康軒文教，2011：57-58）：

> 「**碳酸鈣**，俗稱灰石，白色固體不易容於水，為大理石和貝殼的主要成分……**碳酸氫納**，遇熱會分解成二氧化碳、碳酸納和水。」

本研究中的高分組學生由於國中課本中並未及D選項的碳酸鈣（$CaCO_3$）加熱分解產生二氧化碳（$CO_2$）（高中課程才會提及「**大理石及石灰石的主要成分均為碳酸鈣（$CaCO_3$），加熱會分解產生二氧化碳及氧化鈣**」），而選項C的碳酸鈉（$Na_2CO_3$）與碳酸氫鈉（$NaHCO_3$）又過於相近，且國中課本又提及碳酸氫鈉（$NaHCO_3$）遇熱會分解產生二氧化碳，因此造成學生產生概念性的混淆。另，低分組學生因四個選項未提供化學式，所以無法利用產生二氧化碳（$CO_2$）必須生成物就要有碳原子（C）和氧原子（O）之答題線索。綜合上述因素導致試題29之鑑別度過低。筆者建議提供學生選項的化學式，並且將超出課本範圍的原選項D「碳酸鈣」更換成「碳酸氫鈉（$NaHCO_3$）」（表10）：

表10 試題29題目分析與建議

| 選項 | A | B | C | D＊ | 難度*P* | 鑑別度*D* |
|---|---|---|---|---|---|---|
| 全體學生 | 53<br>（13.87%） | 82<br>（21.47%） | 114<br>（29.84%） | 133<br>（34.82%） | 0.35 | 0.30 |
| 高分組<br>（前27%） | 3<br>（2.83%） | 8<br>（7.55%） | 40<br>（37.74%） | 55<br>（51.87%） | | |
| 低分組<br>（後27%） | 26<br>（23.21%） | 26<br>（23.21%） | 36<br>（32.14%） | 24<br>（21.43%） | | |

試題29

何種物質遇熱會分解產生二氧化碳？

(A)氫氧化鈣　(B)氫氧化鈉　(C)碳酸鈉　＊(D)碳酸鈣

建議修正後試題29

下列何種物質遇熱會分解產生二氧化碳（$CO_2$）？

(A)氫氧化鈣（$Ca(OH)_2$）　(B)氫氧化鈉（NaOH）

(C)碳酸鈉（$Na_2CO_3$）　＊(D)碳酸氫鈉（$NaHCO_3$）

## 二、學校自然科定期評量之男女成就差異

就本研究全體學生表現而言，從效果量分析來說成就組別和性別之間沒有交互作用，高分組（前10%）的女性表現低於男性（$D_{高}$＝-0.417）、低分組（後10%）的女性表現也低於男性（$D_{低}$＝-0.232），而全體組則男、女無差異。盧雪梅與毛國楠（2008a）分析國中基本學力測驗自然科男、女性的表現，全體組和高成就組（前10%）結果與本研究一致，但是低成就組（後10%）的女性表現顯著高於男性，與本研究發現相反，探究其原因可能是因為測驗內容的不同，國中基測自然科內容包含生物、物理、化學、地球科學等範疇，而本測驗只限定化學範圍，因此本研究只能夠提供關於化學科的相關男、女成就差異實徵資料。

## 三、學校自然科定期評量之性別 DIF 分析

本研究共分析 50 題，取 IRT 分析和 Mantel-Haensze 分析，其中 2 題出現性別 DIF，出現率約 4%，兩題皆有利於女性。分析如下：

試題 31 內容包含「化學反應速率的概念」和「數學幾何空間」，而試題 31 已提供反應速率的定律，因此主要測驗學生「數學幾何空間」的能力，然而自然科有許多能力與數學能力息息相關，故此試題 31 並沒有脫離自然科定期評量的評量目標。其次，盧雪梅與毛國楠（2008b）研究發現國中基測的幾何和問題解決的 DIF 題有利男性者居多(基測的幾何題型附有圖形)且另一篇研究發現基測自然科附圖 DIF 題有利男性（盧雪梅與毛國楠，2008a）居多，但是本研究中試題 31 與幾何空間有關，但卻呈現相反結果（本研究有利於女性）。本研究試題 31 為不含圖形的純文字敘述與算則敘述（已知接觸面積增加 1 倍，反應速率增加 1 倍），且筆者反覆審視本研究的 50 題試題，僅有該試題為缺乏圖表運用之試題，所以研究者猜測「圖表」可能是 DIF 現象的關鍵特徵？若該試題提供圖例（操弄圖例給予與否），是否會改變 DIF 之現象？若會，僅是改變原本 DIF 差距？還是改變 DIF 方向？

另一個延伸議題，學生若能夠自發性畫圖圖解，是否會改變 DIF 現象？陳怡琴（2009）曾以 151 名六年級學生為對象，分數運算為主題，實驗組（圖解及多元策略教學）與控制組（傳統教學）在 A2 實作試題（未給圖表，要求受試圖示解題之試題），兩組的性別 DIF 方向相反，實驗組呈現有利於男性的 DIF 現象，控制組呈現有利於女性的 DIF 現象。亦即，在圖解教學介入後，會擴大甚至顛倒性別 DIF 之方向，然而該研究的另一 B1 實作圖解試題卻沒有明顯 DIF 現象，所以該 DIF 之改變現象仍不穩定，該研究者解釋因 B1 試題難度較低而導致沒有明顯性別 DIF 現象。

從前述研究可思考未來研究之方向：「圖形提供」或「圖解教學」是否會影響性別 DIF 現象？其中，試題的難易度是否與其有交互作用？例如：本研究的圖形為「立體幾何圖形」，改變為「平面幾何圖形」是否有影響？前述問題仍需待後續實徵研究進行驗證。

試題31

已知接觸面積增加一倍，反應速率增加一倍，某一立方體，邊長16 cm，若將其切成每邊4cm之立方體，則反應速率變為原來的幾倍？

＊(A) 4倍　(B) 8倍　(C) 16倍　(D) 64倍

試題 50 出現性別 DIF 現象，進一步審視其命題內容，未發現與測驗目標無關之因素，換言之雖出現 DIF 但仍不構成試題偏誤。

試題50

在「$Br_2+H_2O \rightleftharpoons H^+ + Br^- + HBrO$」之平衡反應中，下列何種狀況，可改變平衡使反應向左移動？

(A)加氨水溶液　＊(B)加醋酸溶液　(C)加糖水溶液　(D)加食鹽水

綜合來說，由於測驗結果常被當作「做決定」的依據，例如：教育診斷、個人升學、求職就業、資格認定、證照頒發等，所以測驗公平性（fairness of test）就變成大眾所關心的焦點，也是測驗相關領域人士重視的課題。本研究未配組的性別成就差異部分，以高、低成就組和全體組來探討自然科性別成就差異，結果顯示全體受試並無顯著性別差異出現，但是而在高分組（前 10%）和低分組（後 10%）中，都是男性表現都較佳。若進一步結合研究二的 DIF 結果，由於試題 31 和試題 50 無論任何能力的群體中都有利於女性，若排除該 2 個試題後，本研究高分組和低分組男性自然科成就表現更顯著高於女性（$D_{高}$＝-0.417 變為 $D_{new\,高}$＝-0.755、$D_{低}$＝-0.232 變為 $D_{new\,低}$＝-0.271），唯僅限於國中八年級化學科酸與鹼反應、氧化與還原反應等範疇。

再者，DIF 分析為測驗公平性分析的一種方式，目的在於篩選出有偏誤傾向試題與累積實徵性的資料作為改進依據。本研究發現試題 31 和試題 50 有顯著性別 DIF 現象，該類試題佔總題數的 4%且皆有利於女性。然而，測驗出現 DIF 題時並不代表該試題必須予以移除（林奕宏與林世華，2004），應該進一步分析該試題是否為偏誤試題、內容是否包含與測驗目標無關之因素以及 DIF 試題是否有特定關聯的特徵趨向等。因此，本研究進一步審視有 DIF 現象的試題 31 和試題 50，雖未發

現與測驗目標無關之因素，但可累積相關實徵資料後，再分析此類性別 DIF 試題可能的特徵趨勢。

# 參考文獻

王文中（2004）。Rasch 測量理論與其在教育和心理之應用。**教育與心理研究**，**27**(4)，637-694。

王文中、陳雪珠（1999）。教學觀點量表之發展與試題反應分析。**應用心理研究，2**，181-207。

余民寧、謝進昌（2006）。國中基本學力測驗之 DIF 的實徵分析：以 91 年度兩次測驗為例。**教育學刊，26**，241-276。

吳裕益、洪碧霞、徐綺穗、葉千綺（1993）。**國民中學國文、數學及理化科成就測驗編製報告**。臺灣省教育廳專案研究報告，未出版。

林奕宏、林世華（2004）。國小高年級數學科成就測驗中與性別有關的 DIF 現象。**臺東大學教育學報，15**(1)，67-96。

邱美虹（2005）。TIMSS 2003 臺灣國中二年級學生的科學成就及其相關因素之探討。載於張秋男（主編），**國際數學與科學教育成就趨勢調查 2003**（頁 7-54）。臺北：國立臺灣師範大學科學教育中心。

建北至少 400 分全國基測龍頭校最低 404 分（2011 年 5 月 23 日）。**中國時報電子報**。取自 http://life.chinatimes.com/life/100316/112011052300069.html

國中生基本學力測驗工作推動小組（2002）。國中基本學力測驗自然科試題之設計理念」。**飛揚，13**。取自 http://www.bctest.ntnu.edu.tw/

國立臺灣師範大學心理與教育測驗研究發展中心（2008）。「二〇〇八年國中基測研發成果」媒體交流茶會。**飛揚，55**，2-8。

基北區 100 學年度高中職免試入學簡章彙編（2011 年 2 月 8 日）。

康軒文教（2011）。**自然與生活科技二下**。新北：康軒出版社。

陳怡琴（2009）。**以 Rasch 模式探討性別及學習機會對數學分數實作評量「差異試題功能」（DIF）之影響**（未出版之碩士論文）。國立屏東教育大學，屏東。

盧雪梅、毛國楠（2008a）。國中基本學力測驗自然科之性別差異和差別試題功能（DIF）分析。**測驗學刊，55**(4)，725-759。

盧雪梅、毛國楠（2008b）。國中基本學力測驗數學科之性別差異與差別試題功能（DIF）分析。**教育實踐與研究，21**(2)，95-126。

擴大高中職及五專免試入學實施方案（2009 年 9 月 4 日）。

Chien（2006）。**以 Rasch 分析協助測驗同分時之測量研究**。取自 http://raschsmile.blogspot.com/2006_09_01_archive.html

Cohen, J. (1988). *Statistical power analysis for the behavioral science.* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale NJ: Lawrence Erlbaum Associates.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical of Testing Problems*. Hillsdale, NJ : Lawrence Erlbaum.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, *4*, 207-230.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *13*(2), 238-241.

Wang, C. (1995). *Differential item functioning analyses of the biology subject test of the college entrance examination of Taiwan* (Doctoral dissertation). Available from ProQuest Dissertations and Theses

Database. (UMI No. 304217336)

Willingham, W. W., & Cole, N. S. (1997). Research on gender differences. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 17-54). Hillsdale, NJ: Lawrence Erlbaum.

Willingham, W. W., Cole, N. S., Lewis, C., & Leung, S. W. (1997). Test performance. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 55-126). Hillsdale, NJ: Lawrence Erlbaum.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *ACER ConQuest user guide*. Hawthorn, Australia: ACER Press.

# Gender Differential Item Functioning in a Science Periodical Test of Eighth Graders

Wei-Chih Hsiao* Chia-Chen Fu**

## Abstract

This study investigates gender differences and differential item functioning (DIF) in a science periodical test of eighth graders. We selected 382 students (191 boys and 191 girls) from a junior high school in New Taipei City. We calculated and compared the effect size, female/male standard deviation ratio, and female/male ratio. In addition, we used the IRT Rasch model and the Mantel-Haenszel procedure for gender DIF. The results of this study are as follows: (a) in unmatched analysis, no gender differences were observed among all groups; however, boys exhibited slightly better performance to girls among both high-achieving groups (top 10%) and low-achieving groups (bottom 10%); (b) in matched analysis, the intersection of the results using the IRT Rasch model and the Mantel-Haenszel procedure showed that the average percentage of items displaying gender DIF across administrations was low, at approximately 4% (in favor of girls). The follow-up review of these DIF items indicated associations of gender DIF with item characteristics. Furthermore, charts may affect the DIF direction. Finally, this study provides suggestions for items construction and future studies on science.

**Keywords:** gender differences in science achievement, science gender DIF, science periodical test

Section editor: Long-Lih Yang
Received: February 17, 2012; Modified: June 4, 2012; Accepted: October 15, 2012
* Wei-Chih Hsiao, Master Student, Department of Educational Psychology and Counseling, National Taiwan Normal University, E-mail: toome123020@hotmail.com
** Chia-Chen Fu, Master Student, Department of Industrial Education, National Taiwan Normal University