教育實踐與研究
第 21 卷第 2 期，95-126，2008 年 9 月

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

盧雪梅、毛國楠*

本研究分析90至94年度國民中學學生基本學力測驗數學科之性別差異和差別試題功能（Differential Item Functioning，簡稱 DIF），各次測驗資料係隨機從該次全體考生中抽取之 5000 名受試的答題反應。在性別差異分析部分，本研究分析全體受試、低成就組（後 10%）和高成就組（前 10%）之數學總分的性別效果量、女／男標準差比和女／男人數比三項指標，此外，也分析數與量、代數、幾何、統計與機率四項內容，及程序知識與執行、概念理解和問題解決三類認知歷程的性別效果量。整體數學表現的性別差異分析結果顯示：全體受試和低成就組無顯著的性別差距出現，高成就組的男生表現略高於女生；此外，男生分數的變異程度較女生大，低成就組和高成就組的男生人數比例都高於女生，高成就組又比低成就組更高些。分項表現的性別差異分析結果顯示：全體受試的男生在問題解決的表現略高於女生，其餘分項皆無顯著的性別差距出現；高成就組男生在幾何的表現略高於女生，其餘分項皆無顯著的性別差距出現；低成就組在所有的分項皆無顯著的性別差距出現。在性別 DIF 分析部分，本研究採 Mantel-Haenszel 法評估 DIF，分析結果顯示：基測數學科性別 DIF 出現率為 2.5%，DIF 題數量雖不多，但性別 DIF 與試題特徵似有關聯，具體而言，代數 DIF 題傾向對女生有利，幾何和問題解決 DIF 題傾向對男生有利。

關鍵詞：國中基本學力測驗、數學性別差異、數學性別 DIF

* 盧雪梅：臺灣師範大學教育心理與輔導學系副教授
smlu@ntnu.edu.tw
毛國楠：臺灣師範大學教育心理與輔導學系教授

Journal of Educational Practice and Research
Vol. 21, No. 2, 95-126, Sep. 2008

Gender Differences and Differential Item Functioning in Mathematics Basic Competence Test for Junior High School Students

Sheue-Mei Lu & Kuo-Nan Mao*

This research investigated gender differences and differential item functioning (DIF) on Mathematics parts of the Basic Competence Test for Junior High School Students (BCTEST) from its 2001 to 2005 administrations. Data for each administration were a sample of 5000 examinees randomly drawn from its examinee population. In the study of gender differences, we calculated and compared the effect size, female/male standard deviation ratio and female/male ratio for total-group, low- achieving (bottom 10%) and high-achieving groups (top 10%) across test administrations. Results based on overall performances indicated that there were no visible gender differences among total groups and low-achieving groups. However, males performed slightly better than females amongst high-achieving groups. Results based on performance of math contents and processes indicated that there were no visible gender effect sizes among total, low-achieving and high-achieving groups with only a few exceptions. The exceptions were that males performed slightly better on problem solving among total groups and on geometry among high-achieving groups. In the study of gender DIF, Mantel-Haenszel procedure was used. And results showed that the average percentage of items displaying gender DIF across administrations were low, at about 2.5%. There appeared associations of gender DIF with item characteristics, DIF in favor of females on algebraic items, and DIF in favor of males on geometry and problem solving items.

Keywords: ***Basic Competence Test for Junior High School Students, gender differences in mathematics, gender DIF in mathematics***

* Sheue-Mei Lu: Associate Professor, Department of Educational Psychology and Counseling, National Taiwan Normal University
smlu@ntnu.edu.tw
Kuo-Nan Mao: Professor, Department of Educational Psychology and Counseling, National Taiwan Normal University

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

盧雪梅、毛國楠

壹、緒論

國民中學學生基本學力測驗（以下簡稱基測）於民國 90 年開辦，每年舉辦兩次。自開辦以來，每年的第一次測驗都有三十餘萬考生參加，接近應屆國中畢業生的母群體數，第二次也都有五、六成的考生再參加，此外，基測是經過嚴謹發展程序的標準化測驗，故基測是了解我國國中畢業生學習成就最豐富、也最重要的資料庫。

一、研究緣起

94 年度是九年一貫課程實施後的首次基測，學生表現如何，備受矚目。在高中、高職、五專聯合登記分發放榜後，發現若干傳統明星學校的最低錄取分數和 93 年相比，男校維持或出現小幅下降，女校卻出現較大的降幅，媒體對於這個現象有不少報導，研究者在此擷取兩則報導爲例，「今年國中基測雖沒有因為九年一貫課程出現轉型落差，卻出現『男女性別歧視』？」（自由時報記者，2005），「台北縣多所國中數學老師說，今年基測考題，自然科、數學科都比較難，更別提第二次的數學又更難，對女生較不利；而女生拿手的英語、國文、社會科等，又較簡單，男生也好拿分。」（孫蓉華，2005）。前篇報導有過於聳動之嫌，後篇報導似乎充斥了性別學業成就刻板印象，也就是說，男生在數學和自然科表現較好，女生在語文和社會科表現較佳。男女生的學業成就果真出現前述的差異嗎？基測試題對於不同性別考生有不公平的現象出現嗎？

「性別平等」課題近來備受重視，男女生學習成就差距爲國際間教育界所關注，一些國際性的學生學習成就調查研究，如 TIMSS（Trends in Mathematics and Science Study 的簡稱）和 PISA（Programme for International Student Assessment 的簡稱），都

專章或專節報告男女生差異分析結果，反觀國內性別學習差異研究則較受到忽略。此外，測驗結果通常攸關考生的權益，因此測驗對不同背景應試群體的公平性是測驗發展機構關切的焦點之一，這也是差別試題功能（Differential Item Functioning，簡稱 DIF）研究受重視的原因。

由於數學成就對於學生後續學習、生涯選擇和專業成就有重要影響，所以在眾多學科中，男女生之數學成就差異研究是最受關注的（Fan, Chen, & Matsumoto, 1997）。之前提及，基測資料是了解我國國中畢業生學習成就的重要資料庫，職此，本研究擬分析 90 至 94 年度基測數學科成就性別差異和性別差別試題功能，了解國中畢業生數學科學習之性別差異情形，累積國內性別學習成就之實徵資料，並根據分析結果提出建議，供有關人員參酌。

二、差別試題功能（DIF）的意義

Dorans 和 Holland（1993）將 DIF 界定爲兩組能力或表現相配比的群體，在答題表現上出現顯著的差異。學者今多以 DIF 這個詞來取代先前研究所謂的「試題偏誤」（item bias），並且對 DIF 和試題偏誤做進一步區分（Camilli & Shepard, 1994），不再視二者爲同義詞，DIF 試題不等同於偏誤試題（biased item），DIF 試題經進一步審視和判斷後，如果發現確有與測驗擬測構念（construct）無關的因素造成試題難度對不同背景的應試群體不相等的情形，方能宣稱該試題爲「偏誤試題」。舉例來說，一道以球賽爲背景的數學推理問題，除了測量學生數學知識外，也許還測量到棒球方面的知識，如果棒球知識不是這道試題擬測量的能力，但因不同群體（如不同性別）考生對棒球知識不一，造成了試題對兩組考生產生不同的功能，如果原因確立，方稱該試題爲偏誤試題。總之，DIF 僅是統計分析或數量分析的結果，DIF 是試題偏誤的必要但非充分條件，偏誤試題之斷定須兼具量與質方面的證據。

此外，DIF 和因受試群體能力差距所造成的表現差異要加以區別，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，若受試群體未經任何的配組程序，所觀察到表現差異英文稱爲 impact。Dorans 和 Holland（1993）以「辛普森的反論」（Simpson's Paradox）來說明 DIF 和 impact 的區別，表 1 是 A 和 B 兩組學生在某一試題的表現，A 組 2400 受試當中，有 1440 人答對此題，B 組 2400 人當中，有 1200 人答對此題，換言之，A 組的通過率爲 60%，B 組爲 50%，兩組答題表現的差異爲 0.1，A 組表現較好。但仔細觀察表 1 裡 A 和 B

組之三種不同能力水準之受試的表現，可以發現在各能力層，B 組受試答對率均高於 A 組 0.1，這個試題實際上是對 B 組有利，對 A 組不利，這個例子再次強調將兩組受試者能力加以匹配在 DIF 檢定上的重要性。

表 1 兩組學生在某試題的答題表現

	A 組			B 組		
	人數	答對人數	通過率	人數	答對人數	通過率
低能力組	400	40	0.1	1000	200	0.2
中能力組	1000	500	0.5	1000	600	0.6
高能力組	1000	900	0.9	400	400	1.0
全　體	2400	1440	0.6	2400	1200	0.5

資料來源：”DIF detection and description: Mantel-Haenszel and standardization” (p.37), by N. J. Dorans & P. W. Holland, 1993, in P.W. Holland & H. Wainer (Eds.), *Differential item functioning*, Hillsdale, NJ: Lawrence Erlbaum Associates.

本研究包括兩部分，其一，未經任何配組程序，分析男女考生數學科成就差異，即 Dorans 和 Holland（1993）所稱之 impact 分析；其二，以數學科分數將男女生配組後，進行 DIF 分析。

三、DIF 檢定方法

DIF 檢定方法可以分爲觀察分數（observed score）和潛在變項（latent variable）兩種取向。潛在變項取向的 DIF 程序主要是應用試題反應理論（item response theory，簡稱 IRT），雖具較堅實的理論依據，但實施上比較費時費力，同時需要大量的樣本，才能得到穩定的試題參數估計值，晚近發展出來的 DIF 程序傾向於觀察分數取向（Clauser & Mazor, 1998）。目前較普遍的 IRT 取向的 DIF 程序有 Lord χ^2 考驗法、兩組受試群體之試題反應函數（item response function）間之區域量數（area measure），及概似率考驗法（likelihood ratio test）。觀察分數取向的 DIF 程序在計算上較簡易、不需要大量樣本、容易實施，在實際上應用較受歡迎，目前較普遍使用的觀察分數 DIF 程序有 Mantel-Haenszel 法、標準化法（Standardization）、羅吉式迴歸分析法（logistic regression）和 SIBTEST（全名爲 simultaneous item bias test）程序。上述 DIF 程序的原理和特性可進一步參考 Camilli 和 Shepard（1994）、Clauser 和 Mazor（1998）、Dorans

和 Holland（1993）的專書或專篇。

本研究採 Mantel-Haenszel 法（簡稱 MH，Dorans & Holland, 1993；Holland & Thayer, 1988；Mantel & Haenszel, 1959），MH 法是美國教育測驗服務社（Educational Testing Service，簡稱 ETS）研發和使用的方法，屬觀察分數 DIF 程序，是美國測驗業界使用最爲普遍 DIF 檢定方法（Roussos & Stout, 1996），受歡迎的理由包括：(1)計算簡便，(2)不需要大量樣本，(3)有顯著性考驗統計。此外，MH 也提供一項 DIF 量數指標，稱爲 MH D-DIF，來描述 DIF 幅度的大小，並有一套 DIF 嚴重度分類系統，這都是 MH 法受到實務工作者青睞的原因。

四、相關研究

根據本研究旨趣，擷取若干大型數學學習成就評量或統合分析（meta-analysis）研究進行探討。

（一）數學學習成就之性別差異

簡茂發曾計畫主持「國民教育階段學生基本學習成就評量研究」並發表系列研究報告，其中關於數學性別差異分析結果摘述如後。簡茂發等人（1995）以國小 5 年級學童男女共約 3600 餘人爲樣本，研究結果顯示：在數學內容方面，男生機率的平均數顯著高於女生，但在算術運算、數的關係、幾何、度量、統計、類型與關係、代數等項目則無顯著性別差異；在數學歷程方面，女生的數學解題平均數顯著高於男生，但在數學溝通、數學推理和數學聯結等項目則無顯著性別差異。簡茂發等人（1996）以國中 2 年級男女生近 4200 餘人爲樣本，研究結果顯示：在數學內容方面，男生在幾何、度量和代數的平均數顯著高於女生，但在算術運算、數的關係、統計、機率、類型與關係等項目上則無顯著性別差異；在數學歷程方面，男生的數學解題平均數顯著高於女生，不過在數學溝通、數學推理和數學聯結等項目則無顯著性別差異。簡茂發等人（1999）年以國小 3 年級學童男女共約 3600 餘人爲樣本，研究結果顯示：在所有的數學內容項目：估算、算術運算、數的關係、幾何與度量、統計、機率、類型與關係，女生平均數都顯著高於男生；在所有數學歷程項目：數學溝通、數學推理和數學聯結，女生平均數也都顯著高於男生。綜合上述分析結果可看出：國小 3 年級女生在所有的分析項目的表現都顯著高於男生；國小 5 年級男生機率表現優於女生，女生數學解題表現優於男生，其餘項目無顯著性別差異出現；國中 2 年級男生在幾何、

度量、代數和數學解題表現優於女生，其餘項目無顯著性別差異出現。也就是說，隨著年級增長，出現女生數學優勢漸減，而男生漸長的現象。

在國際學生學習成就表現調查研究方面，舉 TIMSS 2003 來說，我國 4 年級男女生的數學整體表現並未有顯著差異；就 25 個參與國家或地區估得的國際平均數來說，男女生的數學整體表現也沒有顯著差異（林碧珍、蔡文煥，2005；Mullis, Martin, Gonzalez, & Chrostowski, 2004）。在分項主題表現方面，我國 4 年級女生在資料呈現與分析的表現顯著高於男生，在數、數型與關係、幾何和測量等主題上，男女生的平均數並沒有顯著差異。就國際分項平均數來說，男生在測量的表現顯著高於女生，女生在資料呈現與分析和幾何的表現顯著高於男生，在數和數型與關係上，男女生的平均數則沒有顯著差異（林碧珍、蔡文煥，2005；Mullis et al., 2004）。就我國 8 年級學生在 TIMSS 2003 的整體表現來說，男女的平均數並未有顯著差異；就 46 個參與國家或地區估得的國際平均數來說，男女的平均數也沒有顯著差異（曹博盛，2005；Mullis et al., 2004）。在分項主題表現上，我國 8 年級女生在代數和幾何的表現顯著高於男生，在數、統計和測量等主題上，男女生的平均數並沒有顯著差異。就國際分項平均數來說，男生在測量的平均數顯著高於女生，女生在代數的平均數顯著高於男生，在數、幾何和統計上，男女生的平均數則沒有顯著差異（曹博盛，2005；Mullis et al., 2004）。綜合上述，無論是我國或是國際平均數皆顯示 4 和 8 年級學生在 TIMSS 2003 數學表現無性別差異，不過在若干分項表現上出現性別差異，就我國來說，4 年級女生在資料呈現與分析的表現顯著高於男生，8 年級女生在代數和幾何的表現顯著高於男生。就國際分項表現來說，4 和 8 年級男生的測量表現皆高於女生，4 年級女生在資料呈現與分析和幾何的表現顯著高於男生，8 年級女生代數表現高於男生。

接下來，就國外若干統合分析進行探討。Maccoby 和 Jacklin（1974）對 1500 餘篇項兩性差異實徵研究進行統合分析，出版《性別差異心理學》（*The Psychology of Sex Differences*）一書，指出男女差異主要在語言能力、數學能力、視覺空間能力和攻擊性四範疇，除語言能力外，其餘三個範疇皆是男性高於女性。Friedman（1989）對 98 篇 1974 年之後出版的兩性數學表現實徵研究進行統合分析，由於分析對象跨的年級層範圍頗大，不易歸納出明顯的性別差異組型，Friedman 特就其中 77 篇以年齡較輕者爲對象的研究進行分析，因爲這群樣本較不受選擇性偏誤（selection bias）影響，得到的效果量平均值爲-.02，負值表示男生平均數較高些，Friedman 結論說：整體來說，男女生的數學表現差距非常小，此外，Friedman 也指出近年來數學性別差距有縮

小的趨勢。

Hyde、Fennema 和 Lamon（1990）對 100 篇數學性別研究進行統合分析，Hyde 等人指出受試者的年齡、測驗的認知層次和樣本的選擇性（selectivity of sample）皆會影響到性別效果量。就年齡來說，在小學和中學階段，女生表現略優於男生，在高中和大學階段，男生表現優於女生。就認知層次來說，女生在計算表現較佳，男生在問題解決的表現較佳，男女生在數學概念理解則無明顯差距。Hyde 等人認爲男生在高中和大學階段表現優於女生的原因，係因爲高中和大學階段數學課程較強調問題解決所致。Hyde 等人將樣本選擇性分爲一般、中度和高度選擇，一般選擇樣本指受試取自接受一般共同課程者，如中小學的學生，或隨機抽樣而來的樣本，如學習成就調查或測驗常模樣本。選擇性樣本指來自選修或主修相關課程的受試者，視其修習相關課程多寡分中度和高度，或指自行報考的樣本，如參加 SAT 和 GRE 等入學考試。Hyde 等人（1990）統計出一般、中度和高度選擇樣本之效果量平均值分別爲.05、-.33 和-.54，也就是說，樣本選擇性越高，男女的數學差距越大。Hyde 等人特別指出：樣本選擇性通常和年齡與認知層次是有關聯的，舉例來說，高中和大學選修數學課程的以男性居大多數；此外，Hyde 等人也指出數學性別差距近年來有縮小的趨勢，具體言之，在 1974 年前出版的研究論文之效果量平均值爲 -.31，然而 1974 年到 1990 年間出版的論文的平均值降爲 -.14。

Willingham、Cole、Lewis 和 Leung（1997）收集了美國境內若干標準化測驗、大型評量方案和資料庫，計 14 種資料來源、74 種測驗，他們將這些測驗分成 15 大類，對 12 年級男女表現進行統合分析，此外，另就其中 10 類測驗進行 4、8 和 12 年級的性別表現趨勢變化分析，以下摘述當中有關數學表現分析的發現。12 年級的統合分析顯示女生在計算表現較佳，而男生在數學概念（含推理）的表現較佳。在年級變化趨勢分析上，Willingham 等人指出：無論是 4、8 或 12 年級，女生計算的表現一直都較男生好，效果量平均值介於.15 到.20 之間；在數學概念方面，4、8 和 12 年級的效果量平均值分別爲.02、-.00、-.07，換言之，隨著年級增長，出現女生優勢漸退，而男生漸長的現象。此外，Willingham 等人（1997）也指出男生數學的變異程度比女生大，Cleary（1991）、Fan 等人（1997）、Han 和 Hoover（1994）、Zhang 和 Manon（2000）也有相同的發現。

若干研究指出性別差距和能力水準有交互作用的現象（如 Cleary, 1991；Han & Hoover, 1994；Zhang & Manon, 2000）。Cleary（1991）分析百分等級 90、50 和 10 者

的性別效果量，發現摘述如下：就 9-11 歲和 12-14 歲組（非選擇性樣本）來說，PR10 和 PR50 組的女生表現優於男生，PR90 組則無性別差異出現。就 15-18 歲（非選擇性樣本），PR10 組無性別差異出現，PR50 和 PR90 組則男生表現高於女生，其中 PR90 組的性別差距又更大些，至於 15-18 歲（選擇性樣本），無論是 PR10、PR50 或 PR90 組，男生表現皆優於女生，其中 PR10 的差距較小些，PR50 和 PR90 組的性別差距幅度非常接近。Han 和 Hoover（1994）、Zhang 和 Manon（2000）都發現在低分的一端，女生成就高於男生，但在高分的一端，男生成就高於女生。Fan 等人（1997）也發現高成就組的男生人數比例高於女生，而且比例隨年級增長增加。

綜合上述，數學成就是否有性別差異存在？難以獲得明確和一致的結論，如 Willingham 和 Cole（1997）所言：受試者的年級（或年齡）、樣本選取方式（如隨機抽取代表性樣本、報名考生、方便取樣）、測驗構念和形式（前者如知識內容、認知歷程；後者如選擇題、開放式問題、實作評量等）等因素都可能影響到結果，外此，不同成就水準組的性別差距也可能不一致。之前提及，基測是了解我國國中畢業生學習成就的重要資料庫，然而迄今基測表現性別差異的實徵研究幾乎闕如，職此，本研究將對基測數學科成就性別差異進行比較有系統的分析，提供研究結果供相關人員參考，並累積國內數學性別學習成就之實徵資料。

（二）數學測驗之性別 DIF 研究

根據本研究的旨趣，針對國內外探究數學性別 DIF 與試題特徵關聯之實徵研究進行探討。

黃財尉和李信宏（1999）以中部地區國中一年級學生 1440 人爲對象，以 SIBTEST 程序檢定其自編成就測驗的性別 DIF，其研究目的之一在探討性別 DIF 和試題內容是否有關聯。他們根據內容將試題分群，進行差別題群功能（differential bundle functioning，簡稱 DBF）分析，DBF 是 DIF 的延伸，不是以試題爲分析單位，而是以題群爲分析單位。黃財尉和李信宏發現：算術內容之分數運算題群對女生有利，算術內容之整數加減題群對男生有利，但代數內容之分數運算、幾何內容之距離觀念和應用消費問題三類題群並無顯著性別 DBF 出現。

曾建銘（2004）以兩種 IRT 取向 DIF 程序--有符號區域量數（signed area measure）和概似率考驗法（likelihood ratio test），檢定 90 年度第 1 次基測數學科的性別 DIF，兩種 DIF 程序的檢定結果頗爲一致，在 32 題中，兩種方法同時檢定爲 DIF 者有 2 題，

出現率約 6%，其中一題對女生有利，內容屬於代數（第 13 題）；另一題對男生有利，內容屬於幾何（第 26 題）。

林奕宏和林世華（2004）以 MH 法和一種 IRT 取向 DIF 程序研究自編數學成就測驗的性別 DIF，以某國小六年級全體學生 677 名爲樣本，其測驗編有 2 份題本，每份有 28 題，共計 56 題。他們以兩種 DIF 程序同時檢定呈現 DIF 爲判斷標準，共判定 2 題出現 DIF，都對男生有利，一題內容爲數列規律，另一題爲位值關係。

Harris 和 Carlton（1993）研究 SAT 數學測驗之性別 DIF 組型，利用 MH 法分析 6 份題本的考生答題資料，各份題本由 40 題問題解決（problem solving）題和 20 題數量比較（quantitative comparison）題組成，共計 360 題，他們根據將試題依數學內容（如幾何、代數……等）、認知複雜度（如事實知識、計算、概念、例行問題……等）、試題內容特徵（如文字題、抽象情境、與教科書習題相似度、變數出現的位置..等）和形式特徵（如題幹字數、閱讀難度、題目是否出現人物以及其性別……等）共 37 項特徵將試題分類。利用單因子變異數分析探究各特徵對 MH DIF 量數的影響如何，分析發現，將男女依數學總分配組後，幾何、幾何／算術題、認知複雜度較高、真實生活情境（應用文字題）、題幹字數較多和非教科書習題式的試題較傾向於有利男生，另一方面，代數／算術和雜項（如數系、集合等）、認知複雜度較低、抽象的情境（符號運用）、題幹字數較少，與教科書習題雷同的試題則傾向於有利女生。Doolittle 和 Cleary（1987）以標準化法（standardization）分析 ACT 數學測驗的性別 DIF，也發現對數學表現相等的男女生來說，幾何題和文字式的問題解決題傾向對男生較有利，代數題則傾向對女生較有利。O'Neill、Wild 和 McPeek（1989）以 MH 法分析 GRE 數學測驗，發現文字式的問題解決題對男生有利，計算題則對女生有利。

Ryan 和 Fan（1996）以參與 1985 年第二次國際數學調查研究（Second International Mathematics Study）8 年級男女生共 6 千餘人爲研究對象，以 SIBTEST 程序分析 4 份題本的性別 DBF，發現就 SIBTEST 檢定值（β）方向顯示，應用問題對男生有利，4 份題本都達顯著水準，幾何也對男生有利，當中有 3 份題本達顯著水準；代數題、計算題對女生有利，有 3 份題本達顯著水準，算術（數與量）也對女生有利，但只有 2 份題本達顯著水準。Ryan 和 Chiu（2001）以 SIBTEST 程序檢定 Midwestern Mathematics Placement Exam（簡稱 MMPE）的性別 DBF，MMPE 是大一學生數學安置評量，Ryan 和 Chiu 發現文字題（word problems）、幾何題，以及需要高層思考的代數問題對男生比較有利。Mendes-Barnett 和 Ercikan（2006）以加拿大約 9000 名 12 年級學生爲對象，

以 SIBTEST 分析他們在 British Columbia Provincial of Mathematics Examination 的性別 DBF，研究發現男生在問題解決和高層認知試題表現較佳，女生在無方程式的計算題表現較佳，不過幾何試題並無性別 DBF 出現。

上述研究旨趣大都在探討性別 DIF 與試題特徵的關聯，研究發現雖然沒有完全一致，但有一些頗爲類似的傾向出現，特別是國外的研究，具體言之，問題解決、幾何和高層次思考試題傾向對男生有利，代數和純計算試題傾向對女生有利（如 Harris & Carlton, 1993；Doolittle & Cleary, 1987；O'Neill et al., 1989；Ryan & Fan, 1996；Ryan & Chiu, 2001；Mendes-Barnett & Ercikan, 2006）。本研究將對 90-94 年度 10 次基測數學題進行性別 DIF 分析，一方面調查 DIF 出現率爲何，另一方面，10 次測驗共累積 317 道試題，擬觀察是否有與性別 DIF 較有關聯的試題特徵出現，期望研究結果供各種考試試題研發人員參考暨相關教育人員參酌。

五、研究目的

根據前述研究動機和相關文獻探討，本研究目的臚列如下：

（1）分析 90 至 94 年國中基測數學科之男女成就差異，包括效果量、女／男標準差比和女／男人數比等三項指標，並比較高成就組和低成就組之性別差異組型。

（2）分析 90 至 94 年國中基測數學科之性別 DIF，調查性別 DIF 出現率並探究性別 DIF 與試題特徵的關聯。

貳、研究程序

一、資料來源

本研究分析之 90 到 94 年度考生答題資料係向「國民中學學生基本學力測驗工作推動委員會」申請，根據該委員會的資料釋出規定，各科各次測驗提供隨機抽取之 5000 名考生答題反應資料。研究者就各次測驗 5000 名考生和基測委員會提供之全體考生的原始分數平均數與標準差進行比較，結果顯示 5000 名考生資料已經可以獲得相當精確的估計值。

二、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，基測答對 1 題計得 1 分，答對題數即爲原始分數。Willingham 和 Cole（1997）指出：傳統運用的平均數（即效果量）比較男女差異，並無法窺見男女差異的全貌，他們另提出女／男標準差比（Standard Deviation Ratio，簡寫 SDR）和女／男人數比（Female and Male Ratio，簡寫 F/M）兩種量數。此外，亦有研究指出男女生表現分配變異程度不一，不同成就水準之性別差距也不一致，因此，Willingham 和 Cole 也對成就在分配兩端者（前 10%和後 10%）的表現進行分析。本研究循 Willingham 和 Cole（1997）的分析方法，分別分析全體受試、低成就組和高成就組之三項差異比較量數：效果量（D）、女／男標準差比和女／男人數比，爲了使效果量與 DIF 量數方向一致，本研究之效果量計算公式如下：

$$D = \frac{M_{女} - M_{男}}{\sqrt{S^2_{pooled}}}$$

S^2_{pooled} 爲兩組之合併變異數，D 爲正值表示女生的表現較佳，爲負值表示男生的表現較佳，根據 Cohen（1988）的評鑑規準，效果量 D =.20 爲小等級、D=.50 爲中等級、D = .80 爲大等級。此外，女／男標準差比值若大於 1，表示女生分數變異程度較大，若小於 1，則表示男生分數變異程度較大。又，女／男人數比若大於 1，表示女生人數比例較高，若小於 1，則表示男生人數比例較高。

本研究高、低成就組的選取方式說明如後，低成就組由全體受試中從最低分開始往上取 10%的受試者，高成就組則是從最高分開始往下取 10%的受試者，由於同分的關係，高、低成就組人數未必正好是 10%，此時取最接近 10%的人數，有時可能略不足 10%，有時可能稍超過 10%。

本研究計算全體受試、低成就組和高成就組在基測數學整體表現之三項比較量數：效果量、女／男標準差比和女／男人數比，此外，另分析全體受試、低成就組和高成就組在各數學內容和認知歷程類別的性別效果量，內容向度分數與量、代數、幾何和統計與機率等四類，認知歷程分程序知識與執行、概念理解和問題解決等三類。本試題分類係採用鄭蕙如（2006）的架構，鄭蕙如參酌若干國內外大型數學評量的架構和九年一貫課程綱要，從內容和認知兩向度將基測數學試題進行分類，並進行過試題分類一致性研究。

三、性別 DIF 分析

（一）配組變項的淨化

Mantel-Haenszel 法通常以測驗總分爲配組變項，若測驗中有 DIF 試題存在，測驗總分本身可能存有偏差，將無法有效將兩組受試者的能力匹比起來，在這種情況下可藉由「淨化」過程（purification）將 DIF 試題排除於配組變項外，以獲致一組完全無 DIF 的試題爲配組變項（Holland & Thayer, 1988）。因此，本研究先進行配組變項淨化，再正式進行 DIF 分析。具體言之，首先以數學科原始分數將男女生配組後，進行 DIF 分析，若出現 DIF 題，則將 DIF 題排除在外，重新計算無 DIF 試題的總分，以之爲新的配組變項，再次進行 DIF 分析，重複以上步驟，亦即應用迭代（iteration）程序，直到獲致一組完全無 DIF 試題，以其總分爲配組變項，方進行正式的 DIF 分析。

（二）Mantel-Haensze DIF 檢定法

本研究以 MH 法進行 DIF 分析，並以男生爲參照組（reference group）、女生爲焦點組（focal group），分析時以無 DIF 試題之總分爲配組變項，MH 法是一種列聯表（contingency tables）分析法，具體言之，在 K 個分數層中，各個分數層之受試者答題表現資料可整理成一個 2x2 的列聯表，如表 2，表中的 T_k 代表得分爲 k 的人數,n_{Rk} 和 n_{Fk} 分別是參照組和焦點組的人數，m_{1k} 是答對此題的人數，m_{0k} 是答錯的人數，每道試題共計可以得到 2×2×K 個列聯表，其中 K 爲無 DIF 試題之總分。

表 2　得分爲 k 之 2 × 2 列聯表

	得　分		
組　別	1	0	合計
參照組	A_k	B_k	n_{Rk}
焦點組	C_k	D_k	n_{Fk}
合　計	m_{1k}	m_{0k}	T_k

MH 法檢定的虛無假設爲：這 K 個分數層的參照組和焦點組的共同勝算比(common odds-ratio)參數 α_{MH} 的值等於 1.0。α_{MH} 的估計值如下：

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k}$$

Mantel 和 Haenszel（1959）提出一卡方統計數來考驗 α_{MH} 等於 1.0 的假設，公式如下：

$$\chi^2_{MH} = \frac{\left(\left|\sum_k A_k - \sum_k E(A_k)\right| - 0.5\right)^2}{\sum_k Var(A_k)}$$

上式中 $E(A_k)$和 $Var(A_k)$的定義分別爲：$E(A_k)=\frac{n_{Rk}\, m_{1k}}{T_k}$，$Var(A_k)=\frac{n_{Rk}\, n_{Fk}\, m_{1k}\, m_{0k}}{T^2{}_k\,(T_k-1)}$。

在虛無假設下，χ^2_{MH} 從自由度爲 1 的卡方分配，拒絕虛無假設代表試題呈現 DIF。

實際應用上，爲了便於解釋 DIF 分析的結果，通常將 $\hat{\alpha}_{MH}$ 值轉換爲另一種形式的 DIF 量數，稱爲 MH D-DIF（以下用$\triangle_{MH}$表示之），轉換公式如下：

$$\triangle_{MH} = -2.35 \ln(\hat{\alpha}_{MH})$$

$\triangle_{MH}$是以 ETS 的難度量尺（Delta，以△表之）來解釋焦點組和參照組的難度差異，$\triangle_{MH}$爲一等距變數。$\triangle_{MH}$值爲負表示試題有利參照組（男生），$\triangle_{MH}$爲正表示試題有利焦點組（女生）。$\triangle_{MH}$的標準誤的估計公式如下：

$$SE(\triangle_{MH}) = 2.35 \sqrt{Var(\ln(\hat{\alpha}_{MH}))}$$

上式之 $Var(\ln(\hat{\alpha}_{MH})) = \frac{\sum_k \frac{U_k V_k}{T^2{}_k}}{2\left(\sum_k \frac{A_k D_k}{T_k}\right)^2}$，其中 $U_k = (A_k D_k) + \hat{\alpha}_{MH}(B_k C_k)$，

$V_k = (A_k + D_k) + \hat{\alpha}_{MH}(B_k + C_k)$。

（三）ETS DIF 嚴重度的分類系統

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，當樣本人數很大時，即使是微小的差異，也可能達到統計上的顯著性（statistical significance），但不見得具有實際上的顯著性（practical significance），因此，ETS 兼顧統計顯著性考驗結果和$\triangle_{MH}$的幅度發展了一套 DIF 分類系統（Dorans & Holland, 1993）。如果試題之$\triangle_{MH}$值在統計上未顯著異於 0 或$\triangle_{MH}$的絕對值小於 1.0，將之歸於 A 類 DIF；如果$\triangle_{MH}$的絕對值大於 1.5 且統計上顯著大於 1.0，則歸於 C 類 DIF，其餘的試題歸於 B 類 DIF，上述統計檢定的顯著水準皆爲.05。A 類代表不顯著或輕微的 DIF，B 類代表中度 DIF，C 類則爲重度 DIF。

本研究對 DIF 題的判定之不單以χ^2_{MH}檢定結果爲據，因爲男女兩組樣本人數在 2500 人上下，僅依據統計檢定結果容易得到大量的 DIF 題，因此採用 ETS DIF 嚴重度分類的結果，以 B 和 C 類 DIF 題爲 DIF 題。此外，根據盧雪梅（2000）的模擬研究結果，即使在焦點組和對照組能力差距達 1 個標準差時，ETS DIF 分類結果在大多數情況下皆能有效控制第一類型錯誤率。余民寧和謝進昌（2006）也指出 ETS DIF 分類系統較不受樣本因素影響，可得到客觀和可信的 DIF 指標。爲更清楚說明 DIF 的定義和分析的基本原理，在此輔以兩個實例說明，圖 1 爲某一 B 類 DIF 題的試題特徵曲線（M901-26，數學科 90 年 1 次基測第 26 題），由圖可看出：分數相同之男女生的答題通過率出現明顯差異，此題明顯有利於男生，因爲男生通過率明顯高於同分的女生。圖 2 爲某一 A 類 DIF 題的實徵試題特徵曲線（M901-27），圖中顯示：分數相同之男女生的答題通過率無明顯差異。

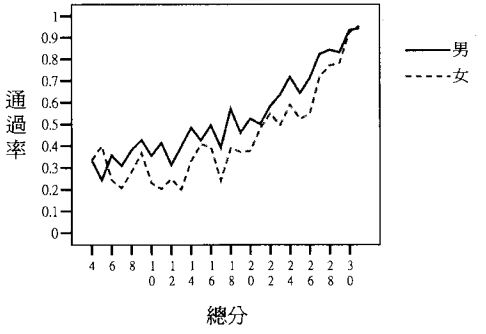

圖 1　某一 B 類 DIF 題的試題特徵曲線

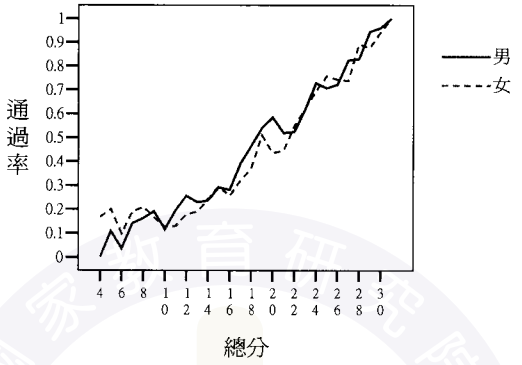

圖 2　某一 A 類 DIF 題的試題特徵曲線

（四）DIF 量數和試題特徵關聯之分析

本研究參考 Harris 和 Carlton（1993）的分析方式，利用單因子變異數分析探討試題特徵對$\triangle_{MH}$的影響和變異解釋量。本研究探討之試題特徵包括數學內容和認知歷程，內容分爲數與量、代數、幾何和統計與機率等四類，認知歷程分爲程序知識與執行、概念理解和問題解決三類。此外，也分析各類試題在 ETS DIF 嚴重度分類系統的分配，並對 B 和 C 類 DIF 試題的內容做進一步檢視，嘗試歸納出與 DIF 較有關聯的特徵。

參、結果與討論

一、基測數學科成就之性別差異分析結果

表 3 呈現各次測驗全體受試、低成就組和高成就組受試之性別效果量、女／男標準差比（SDR）和女／男人數比（F/M），表中之測驗年次的前 2 碼代表年度，第 3 碼代表次別。首先，看全體受試的分析結果，效果量介於-.12 至.09 間，平均值爲-.00，標準差約爲.06，以單一樣本 t 考驗檢定 10 次效果量之平均值是否顯著異於 0，結果未

達顯著水準（$t_{(9)}$ = -0.05，p >.05），整體來說，基測數學科成就並無顯著的性別差異出現。再看女／男生標準差比（SDR），SDR 值介於.90 至.94 間，皆小於 1，顯示男生分數變異程度略大於女生。另外，從女／男人數比可以看出考生中男生人數略多些。

其次，看低成就組的分析結果，效果量介於-.15 至.18 間，平均值約爲.04，標準差約爲.11，效果量平均值之 t 檢定結果未達顯著水準（$t_{(9)}$ = 1.22，p >.05），整體來說，低成就組基測數學科成就亦無顯著的性別差異出現。再看 SDR 值，SDR 值介於.86 到 1.08 間，亦即男女生分數的變異程度大小互見。接著看女／男人數比（F/M），F/M 值介於.66 到.95 間，皆低於該次測驗全體樣本觀察到的比值，由於全體受試組成以男生略多些，以χ^2 適合度檢定低成就組男女比例分配是否顯著不同於全體樣本，10 次測驗中有 7 次的χ^2 值達顯著水準，再佐以百分比差距值之區間估計結果，亦即以低成就組男女百分比差距建構 95%的信賴區間，看是否包含全體受試男女百分比差距值，若不含此點，則可推論男生百分比顯著高於女生。結果發現除 911、922 和 942 次測驗外，其餘次別男生人數百分比皆顯著高於女生。低成就組各次測驗之男女生人數百分比和χ^2 檢定結果詳見表 4。綜合言之，雖然組間差距整體未達顯著水準，但在多數時候，低成就組男生人數百分比顯著高於女生。

最後，看就高成就組的分析結果，效果量介於-.24 至.24 間，平均值爲 -.11，標準差約爲.12，效果量平均值之 t 檢定結果達顯著水準（$t_{(9)}$ = -2.46，p <.05），整體來說，高成就組基測數學科成就性別差異達顯著水準，男生表現高於女生。再看 SDR 值，10 次測驗的 SDR 值介於.89 到 1.08 間，亦即男女生分數變異程度大小互見。接著看女／男生人數比(F/M)，F/M 值介於.51 到.75 間，皆低於該次測驗全體樣本觀察到的比值，以χ^2 適合度檢定高成就組男女比例分配是否顯著不同於全體樣本，10 次測驗之χ^2 值達顯著水準，再佐以百分比差距值之區間估計結果，結果發現 10 次測驗男生人數百分比皆顯著高於女生。高成就組各次測驗男女生人數百分比和χ^2 檢定結果詳見表 4。綜合言之，高成就組男生整體表現顯著高於女生，惟平均效果量並不大，仍未及 Cohen（1988）的「小等級」標準。此外，各次測驗高成就組男生人數百分比皆顯著多於女生。

表 3　基測數學科的性別效果量、女／男標準差比和女／男人數比

測驗年次	題數	組別	效果量	標準差比	人數比
901	32	全體	.02	.93	.97
		低成就	.17	.88	.81
		高成就	-.24**	.93	.75
902	31	全體	-.02	.93	.93
		低成就	.11	.98	.77
		高成就	-.15	1.08	.65
911	31	全體	-.05	.93	.97
		低成就	.03	.95	.95
		高成就	-.16*	.89	.70
912	31	全體	.00	.94	.88
		低成就	-.07	.93	.72
		高成就	-.11	.99	.61
921	31	全體	.09**	.92	.95
		低成就	.04	.88	.68
		高成就	.24**	1.03	.74
922	31	全體	-.01	.94	.93
		低成就	.00	1.06	.80
		高成就	.00	1.00	.72
931	32	全體	.05	.94	.93
		低成就	-.15	1.08	.72
		高成就	-.22*	.97	.66
932	32	全體	-.03	.90	.92
		低成就	.18*	1.02	.72
		高成就	-.14	.97	.62
941	33	全體	.06*	.93	.88
		低成就	.02	1.02	.66
		高成就	-.14	1.00	.66
942	33	全體	-.12***	.93	.90
		低成就	.06	.86	.87
		高成就	-.15	.99	.51

*p < .05. **p < .01. ***p <.001.

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

表 4　基測數學科低成就組和高成就組男女生人數百分比

	低成就組			高成就組		
測驗年次	女	男	χ^2	女	男	χ^2
901	44.6	55.4	3.91*	42.7	57.3	7.85**
902	43.4	56.6	4.50*	39.4	60.6	16.89***
911	48.7	51.3	0.43	41.3	58.8	11.92**
912	41.8	58.2	5.93*	37.9	62.1	17.53***
921	40.4	59.6	16.51***	42.5	57.5	8.43**
922	44.4	55.6	3.08	41.9	58.1	11.70**
931	41.9	58.1	7.58**	39.7	60.3	13.93***
932	41.8	58.2	8.32**	38.3	61.7	18.36***
941	39.6	60.4	11.81**	39.8	60.2	10.78**
942	46.6	53.4	0.11	33.9	66.1	36.75***

*p < .05. **p < .01. ***p <.001.

註：χ^2 檢定結果顯著表示組內男女生之人數比例與全體樣本有顯著差異，此外，佐以百分比差距值之區間估計結果，發現χ^2達顯著者其男生人數百分比亦皆顯著高於女生。

表 5 呈現全體受試、低成就和高成就組各次測驗各類試題之性別效果量，從效果量平均值方向可以看出高成就組男生在各類試題表現皆較女生佳，低成就組和全體受試則顯出男女生在各類試題表現強弱不一，進一步以單一樣本 t 考驗檢定各類試題效果量平均值是否顯著異於 0。全體受試者的分析結果顯示問題解決試題之檢定結果達顯著水準（$\overline{D}$ = -.06，$t_{(9)}$ =-3.10，p < .05），男生整體表現優於女生，其他類試題之性別效果量平均值皆未顯著異於 0；低成就組的分析結果顯示各類試題的性別效果量平均值皆未顯著異於 0；高成就組的分析結果顯示幾何試題之檢定結果達顯著水準（$\overline{D}$ = -.11，$t_{(9)}$ = -2.98，p < .05），男生整體表現優於女生，其他類試題之性別效果量平均值皆未顯著異於 0。

表 5　基測數學科內容和認知分項的性別效果量

測驗年次	數與量	代數	幾何	機率／統計	程序知識	概念理解	問題解決
			全體受試				
901	.06*	.05	-.03	--	.02	.03	-.02
902	.05	.00	-.08**	--	.01	-.05	-.03
911	-.03	-.04	-.06*	-.04	-.01	-.02	-.12***
912	.04	.01	-.03	-.09**	.01	.00	-.04
921	.09**	.12***	.05	.04	.11***	.10***	-.01
922	-.05	.06*	-.03	-.04	.03	-.02	-.09**
931	.09**	.01	.04	.05	.08**	.06*	-.02
932	-.01	-.02	-.06*	-.06*	-.03	.05	-.17***
941	.10***	.09**	.02	.03	.09**	.05	.02
942	-.18***	-.09**	-.08**	-.03	-.14***	-.03	-.10**
平均數	.01	.02	-.03	-.02	.02	.02	-.06*
			低成就組				
901	.29**	.07	-.12	--	.11	.00	.08
902	.36***	-.07	-.16	--	.22*	-.23*	.18*
911	.03	.07	-.03	-.10	.01	-.05	.08
912	-.03	.00	.02	-.21*	.04	-.18*	.02
921	.06	.04	.02	-.09	.08	.04	-.08
922	-.20*	.22*	-.06	.01	.04	.00	-.08
931	.21*	-.23*	-.10	-.11	-.03	-.08	-.05
932	.15	.11	.01	.03	.05	.29**	-.16
941	.01	.10	-.08	.02	.17*	.12	-.29**
942	-.03	.03	.11	-.08	-.29**	.27**	.21*
平均數	.09	.03	-.04	-.07	.04	.02	-.01
			高成就組				
901	-.14	-.05	-.15	--	-.03	-.15	-.17*
902	.05	-.10	-.15	--	-.21*	.01	-.02
911	.01	-.18*	-.07	-.20*	-.02	-.09	-.14
912	-.06	.03	-.11	-.16	-.14	.02	-.01
921	.18*	.05	.00	.19*	.10	.19*	.01
922	-.09	.09	.08	-.09	-.02	-.05	.06
931	.05	-.11	-.21*	-.12	.06	-.20*	-.13
932	.16	-.09	-.17*	.02	.05	-.08	-.15
941	.03	.17*	-.29**	-.01	-.12	-.14	.04
942	-.20*	-.05	.00	.00	-.17*	-.02	.00
平均數	-.00	-.02	-.11*	-.05	-.05	-.05	-.05

*p < .05. **p < .01. ***p <.001.

綜合上述，就基測數學整體表現來說，效果量分析結果顯示全體受試無顯著的性別差距出現（$\overline{D}$ = -.00，p>.05），低成就組也無顯著的性別差距出現（$\overline{D}$ = .04，p > .05），高成就組則男生表現略高於女生（$\overline{D}$ = -.11，p < .05），惟根據 Cohen 的標準，效果量仍屬微小。此外，男生分數的變異程度較大，成就居分配兩端者以男生居多，10 次測驗高成就組男生人數百分比皆顯著高於女生，低成就組也有 7 次測驗男生人數百分比顯著高於女生。在分項表現上，全體受試者效果量分析結果顯示男生僅在問題解決的表現略高於女生（$\overline{D}$ = -.06，p < .05），其餘分項則無顯著之性別差距出現，低成就組在所有分項皆無顯著之性別差距出現，高成就組男生在幾何的表現高於女生（$\overline{D}$ = -.11，p < .05），其餘分項也無顯著之性別差距出現。

大致來說，男女生在基測數學的表現差距不顯著或者偏小，不過男生整體表現的變異程度較大且成就居分配兩端者以男生居多的現象卻相當穩定，特別是各次測驗高成就組男生人數比百分比皆高於女生的現象，此現象和 Han 和 Hoover（1994）、Zhang 和 Manon（2000）、Fan 等人（1997）、Willingham 等人（1997）的發現相似。

二、基測數學科之性別 DIF 分析結果

（一）試題特徵對△$_{MH}$ 量數的影響和變異解釋量

研究者利用單因子變異數分析探討試題特徵對△$_{MH}$ 量數的影響，表 6 呈現各類試題△$_{MH}$ 量數的平均數、標準差、平均數的 95%信賴區間和變異數分析結果。數學內容之變異數分析結果達顯著水準（$F_{(3, 313)}$ = 3.53，η^2 = .03，p < .05），數學內容可以解釋△$_{MH}$ 總變異量約 3%。雖然△$_{MH}$ 平均數方向顯示數學能力相等的男女生在不同內容表現強弱略有不一致的現象，不過△$_{MH}$ 平均數的 95%信賴區間顯示各類內容之區間皆包含 0 於內，也就是說，數學表現相等的男女生在數與量、代數、幾何和統計與機率類試題之整體答題表現，就如理論所預期一般，沒有顯著差異。

認知歷程之變異數分析結果亦達顯著水準（$F_{(2, 314)}$ = 6.09，η^2 = .04，p < .01），認知歷程可以解釋△$_{MH}$ 總變異量約 4%。△$_{MH}$ 平均數之 95%信賴區間顯示只有問題解決類的區間沒有包含 0 於內，亦即對數學科表現相等的男女生而言，問題解決試題對男生相對較簡單些，男女平均難度相差約.17 個△單位，不過在程序知識和執行和概念理解類試題上，表現相同的男女生之整體作答表現並無顯著差異。本分析結果顯示

男生在問題解決答題表現較同能力的女生爲佳，此項結果和 Harris 和 Carlton（1993）、Doolittle 和 Cleary（1987）、O'Neill 等人（1989）、Mendes-Barnett 和 Ercikan（2006）的發現相似。

表 6　基測數學科$\triangle_{MH}$量數的平均數、標準差和變異數分析結果

類　別	題數	平均數	標準差	95%信賴區間	*F*	η^2
數與量	81	-.00	.49	-.11 ~ .10	3.53*	.03
代數	105	.08	.43	-.00 ~ .17		
幾何	118	-.07	.41	-.15 ~ .00		
機率與統計	13	-.24	.44	-.51 ~ .03		
程序知識與執行	135	.05	.42	-.02 ~ .13	6.09**	.04
概念理解	117	.01	.46	-.08 ~ .09		
問題解決	65	-.17	.42	-.28 ~ -.07		
全　體	317	-.01	.44			

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

（二）DIF 嚴重度分類結果

表 7 呈現基測數學科各類試題 ETS DIF 試題嚴重度分類分配和 DIF 出現率，本研究對 DIF 的判定以 B 和 C 類 DIF 爲據。表中 A 類代表不顯著或輕微的 DIF，B 類代表中度 DIF，「＋」號表示對女生有利，「－」號表示對男生有利，本分析未發現 C 類的重度 DIF 題。

首先看整體的分析結果，317 道題中共 8 題出現 DIF，皆爲 B 類 DIF，出現率約爲 2.5%，當中有利女生者共 3 題，計 0.9%，有利男生者共 5 題，計 1.6%。綜合言之，基測數學科性別 DIF 出現率相當低，此外，DIF 題當中有利男生者比例略高些。

再就數學內容的分析結果來看，數與量類共有 4 題出現 DIF，出現率爲 4.9%，其中有利男生者 3 題（3.7%），有利女生者 1 題（1.2%）。次高的是代數類，共有 2 題出現 DIF，出現率爲 1.9%，2 題均有利女生。再其次是幾何類，共有 2 題出現 DIF，出現率爲 1.7%，2 題皆對男生有利。機率與統計類試題並無 DIF 試題出現。

最後，就認知歷程的分析結果來看，概念理解類共有 4 題出現 DIF，出現率爲 3.4%，其中有利男女生者各 2 題；問題解決類有 2 題出現 DIF，出現率爲 3.1%，均對

男生有利；程序知識和執行類共有 2 題出現 DIF，出現率爲 1.5%，有利男女生者各 1 題。

截至目前分析爲止，基測數學科的性別 DIF 題雖不多，但 DIF 題的特徵和若干國外研究發現相似（如 Doolittle & Cleary, 1987；Harris & Carlton, 1993；O'Neill et al., 1989），具體言之，代數的 DIF 題有利女生居多，幾何和問題解決的 DIF 題有利男生者居多，不過由於目前觀察到的 DIF 題相當仍有限，不宜過度推論，有待更多實徵分析結果來佐證這些傾向。

表 7　基測數學科試題 ETS DIF 分類結果

	A		DIF		B+		B−	
類　別	題數	%	題數	%	題數	%	題數	%
數與量	77	95.1	4	4.9	1	1.2	3	3.7
代　數	103	98.1	2	1.9	2	1.9	0.0	0.0
幾　何	116	98.3	2	1.7	0	0.0	2	1.7
機率與統計	13	100.0	0	0.0	0	0.0	0.0	0.0
程序知識與執行	133	98.5	2	1.5	1	0.7	1	0.7
概念理解	113	96.6	4	3.4	2	1.7	2	1.7
問題解決	63	96.9	2	3.1	0	0.0	2	3.1
全　體	309	97.5	8	2.5	3	0.9	5	1.6

註：A 代表不顯著或輕微的 DIF 題，B+代表對女生有利之中度 DIF 題，B－代表對男生有利之中度 DIF 題。

（三）各次測驗之 DIF 出現率和 DIF 試題特徵

表 8 呈現本分析之 DIF 題的資訊，按測驗年次和題號排列，提供資訊包括試題所屬內容領域和認知歷程、χ^2_{MH}、$\triangle_{MH}$、SE（$\triangle_{MH}$）、DIF 類別和試題內容主題。10 次測驗中計有 4 次測驗未出現 DIF 題，分別是 911、912、932 和 942 次；有 4 次出現 1 題 DIF 題，分別是 901、902、931 和 941 次，DIF 出現率約爲 3.0%左右；有 2 次出現 2 題 DIF，分別是 921 和 922 次，DIF 出現率約爲 6.5%。綜合言之，10 次測驗的 DIF 出現率在 0%到 6.5%之間，平均出現率爲 2.5%，整體來說，基測數學科性別 DIF 出現率相當低。

研究者和同僚對 DIF 試題的內容做進一步審視，初步並未發現和測驗目標無關的因素，也就是說，雖出現 DIF 現象但尚不構成試題偏誤，無危害試題公平性之虞，不過有一些值得一提的發現如下所述。

表 8 顯示有利女生的 3 題 DIF 中，其中 1 題屬於「數與量」的程序知識執行，考分數四則運算，是一道純計算題（M931-2）。另外 2 題是「代數」的概念理解題（M921-1、M941-1），這兩題有一個共同特徵，就是將題幹文字敘述轉成未知數列式，「小玲的錢包內有佰元鈔票 x 張，拾元硬幣 y 個，請問錢包內有多少元？」（M921-1），「某人帶了 400 原到市場買水果，如果他買 3 個蘋果、5 個水梨，則剩下 30 元；....。設蘋果每個 x 元，水梨每個為 y 元，則依題意可列出下列哪一組聯立方程式？」（M941-1）。

有利男生的 5 題中，有 1 題屬於「數與量」的知識，考平方根數的大小比較。有 2 題屬於「幾何」的概念理解題（M901-26、M902-18），一題考相似形的概念，另一題考圓切線的性質應用，其中 M901-26 在曾建銘（2004）也被檢定為 DIF 題。其餘 2 題是「數與量」的問題解決題（M921-19、M922-29），很巧的是，這兩題都是比例式的應用，「某校一年級與二年級的學生人數比為 3：2，已知一年級的學生中，有 40%視力良好，二年級的學生中，有 30%視力良好。請問一、二年級所有學生中有多少比例的學生視力良好？」（M921-19），「兩個罐子裝有相同重量的酒精溶液，其中水與酒精的重量比分別為 3：1 和 1：1，若將這兩罐溶液全倒入一個較大的容器中且沒有溢出，則後來所得的混合液中，水與酒精的重量比為何？」（M922-29）。

表 8　基測數學科 DIF 題之資訊

測驗年次	題號	內容	認知	χ^2_{MH}	$\triangle_{MH}$	$SE_{(\triangle MH)}$	DIF	內容主題
901	26	幾何	概念理解	38.42	-1.27	.20	B-	相似形
902	18	幾何	概念理解	17.51	-1.01	.24	B-	圓切線性質
921	1	代數	概念理解	3.95	1.32	.24	B+	未知數列式
921	19	數與量	問題解決	59.44	-1.31	.17	B-	比例式應用
922	9	數與量	程序知識	29.39	-1.05	.19	B-	平方根值比較
922	29	數與量	問題解決	39.03	-1.04	.17	B-	比例式應用
931	2	數與量	程序知識	26.32	1.06	.21	B+	分數四則運算
941	1	代數	概念理解	23.81	1.22	.25	B+	二元一次方程式列式

綜合上述，有利女生的 3 題 DIF 中，其中 2 題考的是將題幹文字敘述轉成未知數列式，有利男生的 2 題問題解決 DIF 題中，主題內容都是比例式的應用，是巧合或者有其特殊意涵，由於 DIF 題數量仍然有限，仍有待更多的實徵資料累積，若雷同例子一再出現，未來可設計「驗證性」取向的研究，操弄試題特徵，觀察其是否引發 DIF 出現。

肆、結論與建議

Willingham 和 Cole（1997）在《性別與公平的評量》（*Gender and Fair Assessment*）一書序言中曾提到：知道更多男女生在測驗表現異同的資訊，其目的之一在設計更公平的評量；此外，了解和縮小族群間的學習成就差距（包括性別差距）是許多國家教育政策關注所在，尤其數學科常常被視爲男性的科目。國中基測是了解國內國中畢業生學習重要資料庫，本研究分析 90 到 94 年度 10 次基測之數學科性別成就差異和性別 DIF，呈現國內國中畢業生數學科學習性別差異實況，以下摘述主要研究發現並提出建議，供有關人員參酌。

一、研究發現

（一）基測數學科成就之性別差異分析

就基測數學科整體表現來看，效果量分析結果顯示全體受試無顯著的性別差距出現（$\overline{D} = -.00$，$p > .05$），低成就組也無顯著的性別差距出現（$\overline{D} = .04$，$p > .05$），不過高成就組男生表現略高於女生（$\overline{D} = -.11$，$p < .05$），惟效果量仍屬微小。此外，成就居分配兩端者以男生居多，高成就組又比低成就組更明顯。

就基測數學科分項表現來說，全體受試者的效果量分析結果顯示男生在問題解決的現較高於女生（$\overline{D} = -.06$，$p < .05$），惟效果量仍屬微小，在其他分項如數與量、代數、幾何、統計與機率、程序知識與執行和概念理解上，皆無顯著之性別差距出現。低成就組在所有分項皆無顯著之性別差距出現。高成就組男生在幾何的表現高於女生（$\overline{D} = -.11$，$p < .05$），其餘分項也無顯著之性別差距出現。

數學科往往被視爲男生較擅長的科目，不過本分析結果顯示此印象屬刻板印象成份居多，大抵而言，男女生在基測數學科表現並無明顯的性別差距出現。

（二）基測數學科之性別 DIF 分析

本研究分析 90 到 94 年度 10 次基測數學科計 317 題的性別 DIF，共 8 題出現 DIF，皆爲 B 類 DIF，平均 DIF 出現率爲 2.5%，整體來言，基測數學科之性別 DIF 出現率相當低。

變異數分析結果顯示數學內容（$F_{(3,313)} = 3.53$，$p < .05$）和認知歷程（$F_{(2,314)} = 6.09$，$p < .01$）皆可顯著影響$\triangle_{MH}$的表現，分別可解釋$\triangle_{MH}$變異量 3%、4%。各類試題之$\triangle_{MH}$平均數的 95%信賴區間顯示，就數學表現相同的男女生來說，男生除在問題解決題表現略高於同分女生外，在其他類別試題上，如預期的，男女整體答題表現皆無顯著差異。

DIF 分析是篩選有偏誤傾向試題的統計機制，研究者對 DIF 試題的內容進行進一步審視，初步並未發現和測驗目標無關的因素，也就是說，雖出現 DIF 現象但尚不構成試題偏誤，無危害試題公平性之虞。不過$\triangle_{MH}$平均數的方向和 DIF 題的方向大致顯示代數 DIF 題傾向對女生有利，幾何和問題解決 DIF 題傾向對男生有利，這和國外文獻的性別 DIF 傾向頗爲相符，由於本研究觀察到的 DIF 題數量仍然相當有限，不宜過度推論，上述之試題特徵與性別 DIF 的關聯傾向仍有待更多實徵資料支持佐證。

二、建議

（一）數學性別學習差距變化趨勢的研究

數學科一般被視爲是男性科目，然而根據 90 到 94 年度基測數學科性別表現分析結果，國中畢業階段男女生數學表現並無明顯差距出現。不過簡茂發等人（1995，1996，1999）、Hyde 等人（1990）、Willingham 等人（1997）皆發現隨著年級增長，出現女生數學優勢漸減而男生漸增的現象，此外，在實際情境裡，無論是高中組別、大學科系和生涯專業，選擇數理相關領域者仍以男性居多數，因此，站在鼓勵適性發展的立場，特別是女性，不同階段學生之數學性別差距的變化趨勢，值得系統探究。未來研究可就已有之大型考試和資料庫進行系統分析，例如歷年的大學入學之學科基本能力測驗和指定科目考試，以及「臺灣學生學習成就資料庫」，其實施對象包括國小四和六年級，國中二年級和高中（職）二年級。除此之外，也可進行跨不同研究的

整合分析，相信對國內各學習階段學生數學科學習性別差距變化趨勢的了解和實徵研究資料累積，以及不同學習階段的課程和教學規畫和實施，將有莫大的益處。

（二）數學科之性別 DIF 和關聯因素研究

測驗結果常被用來做爲個人升學、就業、資格認定、證照頒發等決定的依據，因此，測驗公平（fairness of test）是眾所關心的焦點，也是測驗專業社群重視的課題，如由美國數個專業團體聯合頒布的《教育測驗實務公平性準則》（*Code of Fair Testing Practices in Education*）（Joint Committee on Testing Practices, 1988）和《教育與心理測驗標準》（*Standards for Educational and Psychological Testing*）（American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999）皆對測驗公平課題提出規範或準則。測驗公平是相當複雜的概念，DIF 分析是測驗發展者展現對追求測驗公平的作爲之一，除以之作爲客觀篩選出有偏誤傾向試題外，測驗發展機構也藉由 DIF 分析結果資料的累積，探究 DIF 可能的原因，做爲改進命題品質的參考。舉例來說，美國 ETS 從累積 DIF 分析結果中發現涉及軍事和運動的內容或主題，特別容易出現性別 DIF 的現象，對男性有利，因此，命題指南建議命題者少以軍事和運動爲主題（Educational Testing Service, 2003）。判斷 DIF 題是否爲偏誤試題時，測驗目標的確認是重要依據之一，再進一步舉例說，若戰爭爲歷史科的測驗目標，以之爲命題材料並無不妥，然而在其他科目，如國文和英語的閱讀短文，應避免以戰爭爲試題材料，以避免試題對不同性別考生有失公平之虞。

本研究分析了 90 到 94 年度基測數學科共 317 道試題，其中僅 8 題出現性別 DIF，出現率相當低，不過這些 DIF 題特徵和國外若干研究的發現頗爲相似，具體言之，代數 DIF 題有利女生居多，幾何和問題解決 DIF 題有利男生居多。此外，研究者進一步觀察試題內容，發現一些有趣值得注意的特徵，有利女生的 3 題 DIF 題中有 2 題是文字敘述轉換成未知數列式的題目，有利男生的 2 題問題解決題都是比例式的應用題，是巧合或者有其特殊意涵，仍無從判斷起。由於本研究屬於探索性研究且偵測到 DIF 題數量相當有限，未來研究可繼續累積數學科測驗之性別 DIF 與其關聯因素的實徵資料，若與前述雷同之例子一再出現，可進一步設計「驗證性」取向的研究，操弄試題特徵，觀察其是否引發 DIF 出現。此外，本研究只以數學內容和認知歷程將試題分類，此爲最典型的分類方式，未來的研究可仿效 Harris 和 Carlton（1993）用更細緻的特徵

（如題幹字數、題目是否出現人物及其性別、生活情境、抽象情境、與教科書習題相似度……等）將試題分類，更周全探究與性別 DIF 可能有關的因素。

誌　謝

本研究分析之資料由「國民中學學生基本學力測驗工作推動委員會」提供，在此誌謝，本文內容完全由作者負責。

參考文獻

自由時報記者（2005，8 月 6 日）。考倒女生？杜正勝指示研究。**自由時報**，A8 版。

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

林碧珍、蔡文煥（2005）。TIMSS 2003 臺灣國小四年級學生的數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 125-164）。臺北市：國立臺灣師範大學科學教育中心。

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。**台東大學教育學報**，**15**（1），67-96。

孫蓉華（2005，8 月 6 日）。竹女降 7 分 彰女降 6 分 屏女降 13 分－老師說：數理難不利女生。**聯合報**，A3 版。

曹博盛（2005）。TIMSS 2003 臺灣國中二年級學生的數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 55-94）。臺北市：國立臺灣師範大學科學教育中心。

黃財尉、李信宏（1999）。國中數學成就測驗性別 DIF 之探討: Poly-SIBTEST 的應用與分析。**中國測驗學會測驗年刊**，**46**（2），45-60。

曾建銘（2004）。Differential item functioning on basic mathematics achievement test for middle school in Taiwan。**中華教育學報**，**11**，331-354。

盧雪梅（2000）。Mantel-Haenszel DIF 程序之第一類錯誤率和 DIF 嚴重度分類結果研究。**中國測驗學會測驗年刊**，**47**（1），57-71。

鄭蕙如（2006）。**國中生數學內容知識與數學認知能力之混合 Rasch 模式分析研究**。國立臺灣師範大學教育心理與輔導研究所博士論文，未出版，臺北市。

簡茂發、李虎雄、陳昭地、林保平、曹博盛、王淑真、鄭再添、張敏雪、陳文典、陳義勳、蕭志明、莊玉梅、黃長司、黃萬居、朱玲玲、鄭美雪、曾文雄、吳美麗、李秀玉、卓娟秀、張武昌（1995）。**教育部八十四年度國民教育階段學生基本成就學習評量研究研究報告**。台北：國立臺灣師範大學中等學校研習中心。

簡茂發、李虎雄、陳昭地、林保平、曹博盛、楊瑞智、王淑真、鄭再添、張敏雪、唐書志、陳文典、陳義勳、莊玉梅、蕭志明、黃長司、黃萬居、朱玲玲、鄭美雪、曾文雄、吳美麗、李秀玉、卓娟秀、張武昌（1996）。**教育部八十五年度國民教育階段學生基本成就學習評量研究研究報告**。台北：國立臺灣師範大學中等學校研習中心。

簡茂發、李虎雄、江永明、朱玲玲、李秀玉、吳美麗、卓娟秀、林靜雯、唐書志、莊玉梅、曹博盛、曾文雄、陳文典、陳昭地、陳義勳、陳麗巧、黃長司、黃萬居、張武昌、張敏雪、蕭志明、鍾靜（1999）。**教育部八十六、八十七年度國民教育階段學生基本成就學習評量研究研究報告**。台北：國立臺灣師範大學科學教育中心。

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Cleary, T. A. (1991). *Gender differences in aptitude and achievement test scores*. Paper presented at the 1991 ETS Invitational Conference on Sex Equity in Educational Opportunity, Achievement, and Testing, Princeton, NJ.

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

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, *24*(2), 157-166.

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 Associates.

Educational Testing Service (2003). *Fairness review guidelines*. Retrieved August 22, 2006, from http://www.ets.org/Media/About_ETS/pdf/overview.pdf.

Fan X., Chen, M., & Matsumoto, A. (1997). Gender differences in mathematics achievement: Findings from the National Education Longitudinal Study of 1988. *The Journal of Experimental Education*, *65*(2), 229-242.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, *59*, 185-213.

Han, L., & Hoover, H.D. (1994). *Gender differences in achievement test scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 369 816)

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the SAT. *Applied Measurement in Education, 6*, 137-151.

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.

Hyde, J., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139-155.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education.* Washington, DC: Author.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from respective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289-304.

Mullis, I. V. S., Martin, K. D., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eight grades*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

O'Neill, K. A., Wild, C. L., & McPeek, W. M. (1989). *Gender-related differential item performance on graduate admissions tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Roussos, L., & Stout, W. (1996). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*, 73-90.

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice*, *15*(4), 15-20, 38.

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 Associates.

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 Associates.

Zhang, L., & Manon, J. (2000). *Gender and achievement-understanding gender differences and similarities in mathematics assessment*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

投稿收件日：2008 年 4 月 15 日
接受日：2008 年 8 月 18 日