國立臺灣師範大學教育心理與輔導學系
教育心理學報，2018，50 卷，2 期，167-188 頁
DOI：10.6251/BEP.201812_50(2).0001

DIF 成因之初探：試題特徵與差異試題功能之關聯 *

孫國瑋
高雄醫學大學
校務研究辦公室

陳承德
國立清華大學
教育心理與諮商學系

施慶麟
國立中山大學
師資培育中心
評估研究中心

近年來，研究者對於差異試題功能（differential item functioning, DIF）議題的探討，已由「檢測」DIF 轉變為「解釋」DIF。以往對於 DIF 試題的解釋，多有賴於專家質性審查的方式。然而，如果能有量化分析的證據輔助專家審查，可對 DIF 成因的判斷有所幫助。本研究透過分析 DIF 試題之特徵，找出試題特徵與 DIF 之關聯，作為後續專家審查時判斷 DIF 成因的參考。為此，本研究採用線性邏輯斯測驗模式（linear logistic test model, LLTM）及隨機效果線性邏輯斯測驗模式（random effects linear logistic test model, LLTM-R）針對測驗中各試題特徵進行所謂的差異層面功能（differential facet functioning, DFF）之檢測，藉以說明試題特徵與 DIF 之關聯。模擬研究結果顯示試題的 DIF 程度受到該試題特徵的 DFF 效果之影響。此外，測驗的 Q 矩陣密度較高時（例如 60%），可能因型一誤差之膨脹而檢測出高比例的 DIF 試題；本研究另以實徵資料說明如何針對試題進行 DFF 分析，藉以找出與 DIF 有關的試題特徵，並作為後續試題修正之方向。根據結果，本研究建議採用 LLTM-R 進行 DFF 檢測，可有助於釐清試題特徵與 DIF 之關聯。

關鍵詞：DIF 成因、差異試題功能、差異層面功能、線性邏輯斯測驗模式、隨機效果線性邏輯斯測驗模式

* 1. 本篇論文通訊作者：陳承德，通訊方式：chengte@mx.ntnu.edu.tw。
2. 本篇論文獲科技部計畫經費補助（計畫編號：MOST 106-2410-H-110-024）僅此誌謝。

自 20 世紀初起，測驗公平性的議題便日益引起關注（Angoff, 1993），因此研究者逐漸重視差異試題功能（differential item functioning, DIF）的檢測。一道試題如果有 DIF，則其在顯現特定群體（例如：特定性別或特定國家）的能力上將有偏誤，若一道試題或是整份測驗的測量對特定群體有偏誤，則測驗的結果便無法有效反應受試者的真實能力，也將進而影響研究者在群體比較上之結果。

過去數十年間，DIF 領域研究者主要鑽研於 DIF 的檢測方法，目標在於有效控制型一誤差（Type I error）的前提下，盡可能正確找到真正有 DIF 的試題。至於檢測出 DIF 試題之後進一步探究 DIF 原因的研究，則大致可回溯到 Angoff（1993）提到「測驗發展者常常無法理解看似完美的試題，為何會存在頗大的 DIF 量」（p. 19）。故而，在 DIF 檢測程序日趨發展成熟之下，Zumbo（2007）主張 DIF 研究的角度應由 DIF 試題之「檢測」轉移至對其原因之「解釋」。因此近年來陸續有學者針對 DIF 成因進行探究（Ercikan, 2002; Gierl & Bolt, 2001; Gierl & Khaliq, 2001; Mendes-Barnett & Ercikan, 2006; Oliveri & Ercikan, 2011），可作為對於探討 DIF 成因有興趣的研究者在不同取向方法上之參考。

在 DIF 的研究中，有相當一部分是探討試題難度（difficulty）上的 DIF，也就是同一道試題對於兩群體有不同的難度，亦即所謂的一致性 DIF（uniform DIF）。如果可以有效的將試題難度分解為數個試題特徵（item property）難度的和，研究者便可透過 DIF 與試題特徵間的關聯，進一步掌握後續探討 DIF 成因的基礎。在試題反應理論的架構中，線性邏輯斯測驗模式（linear logistic test model, 以下簡稱 LLTM）及隨機效果線性邏輯斯測驗模式（random effects linear logistic test model, 以下簡稱 LLTM-R）均可用以分解試題難度，將試題難度拆解成為試題特徵難度的線性組合。故而本研究擬透過模擬研究，分別以 LLTM 及 LLTM-R 對 DIF 試題進行差異層面功能（differential facet functioning, DFF；細節詳述於後）分析，藉以比較兩個模式在探究 DIF 試題與試題特徵的關係上之效能，此為本研究目的之一。

回顧近 20 年來國內在實徵資料上進行 DIF 檢測的相關研究時，對於 DIF 試題形成原因的探討多採用質性分析的方式，量化技術的探討則較少。然而，Ercikan（2002）主張如果能有量化分析的證據輔助專家審查，可對 DIF 成因的判斷有所幫助。故而本研究擬以一筆實徵資料，說明如何針對試題特徵層面進行 DFF 檢測，希冀藉由量化角度提供資訊，作為後續解釋 DIF 成因的基礎，此為本研究目的之二。

DIF 成因的研究

DIF 研究依據不同階段著重的議題不同，大致可區分為三個時期（Zumbo, 2007），為協助讀者對於 DIF 研究的發展過程有較為完整的認識，以下簡要說明之。

在第一個時期，研究者開始對於測驗偏誤（test bias）議題形成研究動機，此時期所謂的「試題偏誤」與現階段的「DIF」之意涵並不相同。「試題偏誤」是指試題在某些情況下對於一個群體不公平，DIF 則是一個統計程序上所顯現的群體表現差異指標（Zumbo, 1999）。第二個時期開始，「DIF」一詞被普遍使用，也針對兩群體受試者平均能力的差異量（以下簡稱為 impact）與 DIF 進行區分，同時許多研究者嘗試以模擬研究方式探究 DIF 檢測方法在各種情境下的效能，特別是關於型一誤差的控制以及檢測力（power）的提升。由於各方法在檢測 DIF 上的表現越來越接近、改善幅度也相對有限，因此部分研究者開始投入尋找產生 DIF 原因，進而進入第三個時期。誠如 Zumbo（2007）所述，DIF 的第三個時期最重要的特徵係對於 DIF 研究思維的改變，例如 Oliveri 與 Ercikan（2011）發現不同語言版本的 PISA（Programme for International Student Assessment）測驗使用的字詞難度、語句結構、語句長度之差異等變項，足以解釋試題的 DIF。除了 DIF 檢測方法的精進外，研究者也逐漸投注心力在對於 DIF 現象的解釋上。

針對 DIF 成因的探討，研究者陸續提出不同的策略，可分為質性與量化兩類分析取向，謹就兩種取向分別介紹如下：

一、質性取向

文獻上採用的質性分析策略主要包含專家審查（expert reviews）以及放聲思考法（think-aloud protocols）兩種，並以前者最廣為使用（Ercikan, 2002）。其作法是透過領域專家（如課程、語言以及文化等領域）針對 DIF 試題進行審查，探究 DIF 是否源自於與測驗無關的因素，以判斷試題應該被保留、修正或是刪除（Oliveri & Ercikan, 2011）。在 Drabinová 與 Martinková（2016）的研究中，有一道關於兒童疾病的試題顯示對捷克的女性較為有利，經專家審查後認為原因在於捷克的女性和孩童有較多的時間相處，因而對於兒童疾病有較多經驗。蘇旭琳與陳柏熹（2008）曾對於 DIF 試題進行質性分析，推測圖表的複雜度可能影響視障生作答，然普通生在某些試題上可能會依靠視覺觀察而選擇錯誤答案。此外，部分學者除了對內容進行質性檢視外，另將 DIF 的試題分別彙整至內容領域、認知成分、問題類型以及文章主題等方面，嘗試推論可能造成的原因（蕭偉智、傅家珍，2012；廖彥棻，2015）。然而，研究者也發現專家審查的成效經常受到特定因素的影響，例如：審查過程是否標準、DIF 試題在題庫中的數量以及審查者是否意識到哪些是真正具有 DIF 的試題等（Ercikan, 2002），且研究顯示專家審查能確定 DIF 成因的比例約在 44%至 80%之間（Gierl & Khaliq, 2001）。

放聲思考則是在教育研究中經常用於當學生進行解決問題、解釋圖表以及閱讀短文或是完成一項活動時，要求他們用語言表達自己的想法及理解過程。使用放聲思考做為判別 DIF 成因時，可藉由受試者對於題目的理解以及回應試圖找出 DIF 可能的原因，是屬於較新穎的方法，Ercikan 等人（2010）於研究結果中指出，在題數不超過 20 題之情況下，放聲思考可以用以確認語言差異是 DIF 的成因，此結果與專家審查模式相同。根據前述說明，放聲思考的適用情境有限，無法用於題數過長時；而能夠用專家審查來檢核 DIF 成因的比例亦可能有審查者主觀意見造成之差異，故而質性取向若能藉由量化方法之輔助，預期可在試題修正上更具相輔相成之效。

二、量化取向

近年來有諸多研究者嘗試以量化方法解釋 DIF 成因，Zumbo 等人（2015）將方法分為以下四種取向：列聯表取向的試題反應模型以及迴歸模型（item response modeling via contingency tables and/or regression models）、試題反應理論（item response theory, 以下簡稱 IRT）、多向度模式（multidimensional models）以及潛在類別方法（latent class methods），然由於本研究所要使用的方法主要是基於試題反應理論進行，因此以下僅針對其中的「試題反應理論」及「多向度模式」兩者分別介紹之。

（一）試題反應理論

在此取向中，本文將介紹在文獻上採用頻率較高之差異題群功能（differential bundle functioning）、差異誘答項功能（differential distractor functioning）以及差異層面功能（differential facet functioning, 以下簡稱 DFF）。

差異題群功能可視為 DIF 的延伸，針對試題進行 DIF 檢測時，為避免個別試題對於群體的些微差異未能被有效偵測，可將試題結合為題群，此時有利於同一群體的許多試題之些微差異便會累積成可被偵測的差異（Gierl & Bolt, 2001）。透過 DIF 的累積以及與質性審查的結合，差異題群功能分析可用以確認一群試題的 DIF 成因與潛在來源（Douglas, Roussos, & Stout, 1996; Shealy &

Stout, 1993）；Bolt（2002）也指出差異題群功能的優點在於能增加統計檢測力且有效控制型一誤差。Mendes-Barnett 與 Ercikan（2006）針對 12 年級的數學測驗，以差異題群功能探究造成性別 DIF 的成因，結果發現 DIF 與試題內容的特徵有關，例如解題所需的認知技能層次、試題是否包含公式或特定內容、是否以故事型態呈現等，同時也發現以差異題群功能檢測會較 DIF 檢測找出更多具有性別差異的試題。

差異誘答項功能主要是分析選擇題中誘答選項之作答情形，其目的在於比較不同群體在各誘答選項上的選答機率是否具有差異。若各誘答選項的選答機率均相近，代表 DIF 效果可能因正確選項而造成；然而，然若其中一個誘答選項呈現選答機率的差異，則 DIF 效果可能與此誘答選項有關，可進一步審查其內容，以確定其與 DIF 成因之關聯，進而協助修改試題（Gierl & Bolt, 2001）。

在 IRT 的架構中，受試者在試題上的得分會受到不同元素的影響，這些元素被稱為層面（facet）（Linacre, 1989），最常被提到的層面是試題和受試者，若層面間具有交互作用則通稱為 DFF（Engelhard, 1992）。由於 DIF 分析檢測的是受試者與試題間的交互作用，因此 DIF 可視為 DFF 之特例（Jin & Wang, 2017）。Xie 與 Wilson（2008）針對 PISA2003 的數學試題進行 DIF 分析，同時也以數學構念中的三個領域進行 DFF 檢測，研究結果發現 DFF 檢測的結果能協助解釋 DIF。在測驗編製過程中，由於試題特徵是組成試題的基礎，透過這些試題特徵做為開發試題之依據，再針對這些特徵進行 DFF 檢測，可以使得試題特徵之訊息得在試題分析過程中被使用，當然也包含 DIF 成因的探究。

（二）多向度模式

多向度（multidimensional）模式主要用來分析同時測量多個能力的測驗資料，由於 DIF 的發生也可被視為測驗中測量到了非預定的向度，若這些向度與測驗欲測量的構念無關，便會被視為是一種干擾（nuisance）（Roussos & Stout, 1996; Shealy & Stout, 1993）。例如語言能力影響考生對於數學能力測驗中試題的題意理解，然由於語言能力並非數學能力測驗想要測量的向度，此時語言能力便可視為一種干擾。因此，如果相同數學能力的考生，由於語言能力上的差異導致在應用題上的作答機率不同，透過多向度模式的分析，如能找出 DIF 試題上的干擾向度，便可能是 DIF 發生的原因。Gierl、Bisanz、Bisanz 與 Boughton（2003）曾應用多向度模式與差異題群分析探討測驗中是否包含性別差異以及試題間是否存在內容與認知上的差異，研究發現當相較於以試題為分析單位，採用題群進行分析時更有利於檢核出多向度以及解釋群體差異。

國內的研究在納入 DIF 檢測做為量表與測驗品質的檢驗指標之一時，在題量充足的情況下，研究者較常採用的方式是刪除 DIF 試題、或者考量測驗的整體性而保留 DIF 效果量較小的試題（王佳琪、何曉琪、鄭英耀，2014；侯雅齡，2013）。然以 DIF 檢測探討應試群體的公平性時，研究者會嘗試更進一步發掘 DIF 成因（蕭偉智、傅家珍，2012；廖彥棻，2015），其中多數研究者採行質性審查方式，亦有部分研究者輔以其他量化的方法，如賴姿伶與余民寧（2015）除進行 DIF 檢測之外，並使用差異題群功能以及差異測驗功能（differential test functioning）對於人格測驗中的試題組合以及量表進行多層次之檢測，試圖多面向的釐清造成群體（應徵者與在職者）差異的原因；曾明基與邱皓政（2015）在研究中使用 DIF 分析、潛在類別 DIF 分析以及多群體潛在類別 DIF 分析，探討研究生與大學生的教師教學評鑑結果，發現除了外顯群體外，外顯群體背後的潛在異質差異也同時會對教師教學評鑑結果具有影響力。

DFF 檢測

以 Rasch 模式（Rasch, 1960）分析測驗資料時，僅以一個參數（即難度）來描述一道試題，故而各試題間的差異也就主要顯示在試題難度上。為了進一步建立難度與試題特徵之間的連結，拆解出各特徵對難度造成之效果，Fischer 於 1973 年提出線性邏輯斯測驗模式（Linear logistic test model, 以下簡稱 LLTM），將試題難度拆解成為試題特徵的線性組合，也曾被研究者用於進行 DFF 檢測。基於本研究聚焦於如何以 IRT 取向中的 DFF 檢測掌握更多 DIF 成因資訊，以提升命題者修題之效能，而 DFF 檢測的是群體與層面間的交互作用，研究上曾被探討會對測驗分數有影響的層

面，如試題所共同擁有的特徵（如試題類型、內容領域、詞彙知識等）、評分者（rater）等（Gierl & Khaliq, 2001; Jin & Wang, 2017; Oliveri & Ercikan, 2011），均可納入作為 LLTM 模式中的試題特徵，進而進行 DFF 分析。以下將分別說明 LLTM 與加入隨機效果之 LLTM，以及如何使用模式進行 DFF 檢測。

一、線性邏輯斯測驗模式（LLTM）

試題反應理論以機率方式描述受試者面對問題的答對機率，以 Rasch 模式為例，其可表徵如下：

$$P_{ij} = \frac{1}{1+\exp[-(\theta_j - \beta_i)]} \quad (1)$$

P_{ij} 是受試者 j 在試題 i 上答對的機率，θ_j 為受試者 j 的能力參數，β_i 為試題 i 的難度參數。將第（1）式轉換成 logit 形式時，可表徵如下：

$$\text{logit} = \theta_j - \beta_i \quad (2)$$

LLTM假設Rasch模式中的試題難度參數可表徵為試題特徵的線性組合如下：

$$\beta_i = \sum_{k=0}^{K} \eta_k Q_{ik} \quad (3)$$

將（3）帶入（2）可得到 LLTM 的 logit 如下：

$$\text{logit} = \theta_j - \sum_{k=0}^{K} \eta_k Q_{ik} \quad (4)$$

其中 η_k 為試題特徵 k 的難度參數，η_o 為截距項，K 為試題特徵的總數，Q_{ik} 則是第 i 道試題於試題特徵 k 上的權重，若試題 i 上具有試題特徵 k 的話，可設定 Q_{ik} 為 1，反之則為 0，所有的 Q_{ik} 便形成所謂的 Q 矩陣（Q-matrix）。在 LLTM 中，所有試題與層面（即試題特徵）之間的關聯，便是以 Q 矩陣描述；Q 矩陣中，有些試題僅具有一個試題特徵，也會有試題具備 2 個以上的試題特徵（可參考表 1）。

以 LLTM 進行 DFF 檢測時，模式中除了試題特徵的效果之外，須再加入群體的主要效果以及試題特徵與群體的交互作用，可表徵如下：

$$\text{logit} = \theta_j + Z_j\gamma_o - \sum_{k=0}^{K} \eta_k Q_{ik} + Z_j \sum_{k=1}^{K} \gamma_k Q_{ik} \quad (5)$$

其中 γ_o 為 impact，Z_j 為受試者 j 的群體編碼，受試者屬於焦點群體時，Z_j 設定為 1，參照團體則設定為 0。γ_k 參數則是用以判別 DFF 效果量，當 γ_k 值達顯著且大於 0，表示該試題特徵 k 有 DFF 且對焦點群體較有利；若 γ_k 值達顯著且小於 0，則試題特徵 k 有 DFF 且對參照群體較有利。

在 LLTM 中，由於試題特徵對於試題難度具有相當的解釋作用，透過分析試題特徵的 DFF 效果，也許可以適度解釋試題的 DIF 效果，進而作為探討 DIF 成因的基礎。如同 DIF 效果，DFF 效果也包含方向及效果量。如以試題為單位，相同方向之 DFF 的效果量可相加，會凸顯在試題難度的差異上，因而凸顯該題的 DIF 效果；而相反方向的效果量則會抵銷，從而削弱該題的 DIF 效果。

二、隨機效果線性邏輯斯測驗模式（LLTM-R）

由於公式 3 中並未包含誤差項，表示 LLTM 假設試題難度可以完全的被試題特徵解釋，然而 Janssen、Schepers 與 Peres（2004）認為試題難度無法完全被現有的資訊完整解釋，因此於 LLTM 中將無法被解釋的部份視為隨機誤差，進而提出隨機效果線性邏輯斯測驗模式（random effects linear logistic test model, 以下簡稱 LLTM-R），並將試題難度 β_i 的定義如下：

$$\beta_i = \sum_{k=0}^{K} \eta_k Q_{ik} + \varepsilon_i \qquad (6)$$

其中 ε_i 為試題 i 上的隨機誤差項，且 $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ 。將（6）代入（2），可得到 LLTM-R 中的 logit 可表示如下：

$$\text{logit} = \theta_j - (\sum_{k=0}^{K} \eta_k Q_{ik} + \varepsilon_i) \qquad (7)$$

與LLTM相比，LLTM-R除了考量受試者與試題特性之外，也納入了試題總群體（item population）的概念，而施測的試題可以被視為巢套於試題總群體內（Janssen, 2010）。隨著研究者證明隨機試題效果模式有助於瞭解DIF效果的來源（Meulders & Xie, 2004; Van den Noortgate & De Boeck, 2005），研究者也嘗試在LLTM-R中加入impact及試題特徵與群體的交互作用效果，以便進行DFF檢測（Beretvas, Cawthon, Lockhart, & Kaye, 2012; Janssen, 2010;Van den Noortgate & De Boeck, 2005），故而以LLTM-R模式進行DFF檢測可表示如公式（8）：

$$\text{logit} = \theta_j + Z_j\gamma_o - \sum_{k=0}^{K} \eta_k Q_{ik} + Z_j \sum_{k=1}^{K} \gamma_k Q_{ik} + \varepsilon_i \qquad (8)$$

在公式（8）中，LLTM-R模式也是透過判斷 γ_k 參數來判斷DFF的效果量，對於DFF的判定準則會與前述LLTM相同。

綜上所述，LLTM 與 LLTM-R 均可使用於 DFF 檢測，前者將試題難度拆解為試題特徵之線性組合，試圖以試題特徵詮釋難度；後者則著重於對資料的解釋及適配情形。為了深入探討前述兩模式進行 DFF 檢測來找出試題特徵與 DIF 關聯性上之效能，本研究將先透過模擬研究評估兩模式在不同情境下之表現，再以實徵資料示範說明如何應用 DFF 檢測找出試題特徵與 DIF 關聯，進而協助解釋 DIF 的可能成因。基於本研究中同時安排模擬研究與實徵資料分析，加上研究主要目的在於展現應用 LLTM 模式於解釋試題特徵與 DIF 試題之關聯，為了使目的聚焦以及避免篇幅過大，因此以下模擬研究的情境較為精簡，LLTM 及 LLTM-R 模式在此議題下更為細緻的比較，就留待後續研究進行探討。

研究一　以模擬資料探討 DIF 與 DFF

一、研究設計

本模擬研究希望透過模擬不同情境，以便在仿真的情境下讓 LLTM 及 LLTM-R 透過 DFF 檢測探討試題特徵與 DIF 之關聯。故此，本研究操弄四個獨立變項，分別為資料型態、DFF 檢測模式、矩陣密度以及 DFF 效果量，分述如下：

（一）資料型態

使用 LLTM 以及 LLTM-R 模式產生兩種型態之模擬資料，在產生 LLTM-R 模式的模擬資料時，於試題參數中加入平均數為 0、標準差為 0.4 的標準常態分配之隨機效果。

（二）DFF 檢測模式

使用 LLTM 與 LLTM-R 模式分別對前述兩種資料進行檢測，預期兩者在分析 LLTM 資料時，兩者效果會相似；然由於 LLTM-R 可處理試題隨機效果，預期以 LLTM-R 進行 LLTM-R 資料的 DFF 檢測時，檢測效果將較 LLTM 來得佳。

（三）矩陣密度

Baker（1993）指出密度越高的 Q 矩陣，在分數的配對準則（matching criterion）上，受到試題 DIF 效果的影響較大。為探討矩陣密度對於 DFF 檢測之影響，本研究於表 1 操弄二組不同的 Q 矩陣密度，表 1 左側的矩陣參照 Fischer（1973）的研究，在 30 道試題、每一題至少需有一個特徵的情境下，將三個試題特徵均設定存在於 12 道試題中，亦即矩陣密度為 40%；右側的矩陣則參照 Green 與 Smith（1987）的研究，將三個試題特徵均設定存在於 18 道試題中，亦即矩陣密度為 60%。

（四）DFF 效果量

在 LLTM 的模式下，若某試題中僅有單一試題特徵具有 DFF 效果時，則 DFF 效果量即等同於此題的 DIF 效果量。為比較不同 DFF 效果量之影響，本研究產生焦點群體之模擬資料時，僅於單一試題特徵之參數上分別增加三種程度的 DFF 效果量：0.3、0.6 以及 0.9，即產生對焦點群體程度不一的不利情形。

關於其他研究設計部分，本研究參考之前的研究（Green & Smith, 1987），僅操弄 3 個試題特徵。另設定 3 個試題特徵間之相關均為-0.389，排除試題特徵間相關不同之影響以利聚焦討論。本研究參考 Fischer（1973），將 3 個試題特徵參數值設定為-0.061、0.388 及 1.75。由於本研究將以 Mantel-Haenszel（Holland & Thayer, 1988；以下簡稱 MH）法進行 DIF 檢測，復以文獻建議 MH 法以及 LLTM 模式之樣本數皆應大於 200 人以上（Green & Smith, 1987; Mazor, Clauser, & Hambleton, 1992），因而本研究的樣本數設定為 R500/F250，R 與 F 分別代表參照群體及焦點群體，試題數則設定為 30 題。參照群體的能力分配設定來自平均數為 0、標準差為 1 的標準常態分配，焦點群體則來自平均數為-1、標準差為 1 的常態分配。為避免抽樣偏誤，所有情境的資料皆重複模擬 100 次。

研究一的依變項共有四個，分別為 DIF 檢測的型一誤差與檢測力以及 DFF 檢測的型一誤差與檢測力。DIF 檢測的型一誤差是將沒有 DIF 的試題誤判為 DIF 試題的比率，檢測力則是能正確檢測出 DIF 試題的比率；DFF 檢測的型一誤差代表將沒有 DFF 的試題特徵誤判為具有 DFF 之比率，檢測力則是能正確檢測出具有 DFF 的試題特徵之比率。表 2 為依變項之計算範例，於研究結果中所呈現之型一誤差與檢測力皆為 100 次模擬下之平均數值。

表 1　研究一之 Q 矩陣設計

	Q 矩陣密度＝40%			Q 矩陣密度＝60%		
	試題特徵			試題特徵		
試題	1	2	3	1	2	3
1	1	0	0	1	0	0
2	1	0	0	1	0	0
3	1	0	0	0	1	0
4	1	0	0	0	1	0
5	1	0	0	0	0	1
6	1	0	0	0	0	1
7	1	0	0	1	1	0
8	1	0	0	1	0	1
9	1	1	0	0	1	1
10	1	1	0	1	0	1
11	0	1	0	1	1	0
12	0	1	0	0	1	1
13	0	1	0	1	0	1
14	0	1	0	0	1	1
15	0	1	0	1	1	0

16	0	1	0	0	1	1
17	0	1	0	1	1	0
18	0	1	0	1	0	1
19	0	1	1	0	1	1
20	0	1	1	1	1	0
21	0	0	1	1	0	1
22	0	0	1	0	1	1
23	0	0	1	1	1	0
24	0	0	1	1	0	1
25	0	0	1	0	1	1
26	0	0	1	1	1	0
27	0	0	1	1	0	1
28	0	0	1	0	1	1
29	1	0	1	1	1	0
30	1	0	1	1	0	1

二、研究方法

本研究希望透過 DFF 探討試題特徵與 DIF 的關係，因此需先進行 DIF 檢測。由於 MH 法採用測驗總分作為配對變項，在使用上較為便利，因而仍是目前最為普遍使用的 DIF 檢測方法之一，因此本研究將以 MH 法進行 DIF 檢測，MH 法的檢測作法將詳述於後。因為本研究是以 LLTM 及 LLTM-R 產生資料，在此二個模式下，試題難度為試題特徵難度之線性組合，故而試題的 DIF 效果會受到試題特徵 DFF 效果之影響，因此預期隨著操弄之 DFF 效果量增加，將使 DIF 檢測力隨之提高。使用 LLTM 以及 LLTM-R 進行 DFF 檢測時，可透過型一誤差與檢測力，瞭解兩個模式在不同情境下檢測 DFF 之效能。

表 2　DIF/DFF 檢測平均型一誤差與檢測力計算範例

	試題1	試題2	試題3	試題4	型一誤差	檢測力
	無DIF	無DIF	無DIF	DIF		
模擬次數						
第1次	1	0	0	1	1/3=0.333	1
第2次	0	0	0	1	0	1
第3次	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
第99次	1	1	0	1	2/3=0.666	1
第100次	0	0	0	1	0	1
平均					0.009	0.99

註：1為檢測出具有DIF、0為無檢測出具有DIF。

MH 法檢測 DIF 的原理為將具有相同分數或相近分數的兩群體（例如男生、女生）受試者視為具有相同能力者，對相同總分的受試者而言，若男女生在該題的答對／答錯比例均相當，則該試題對於男女生而言是公平的，亦即該試題無 DIF，若反之則存在 DIF 現象。使用 MH 法檢測 DIF 時，先以總分將所有受試者分為 k 個組別，每個組別中再分別計算參照群體及焦點群體在待檢測試題上的答對或答錯情形，製作該得分組別的 2×2 列聯表如下表 3。

表 3　第 *m* 個得分組別之 2×2 列聯表

	第 *i* 題上之得分情形		
群體	1	0	合計
參照群體（*r*）	N_{r1m}	N_{r0m}	N_{r+m}
焦點群體（*f*）	N_{f1m}	N_{f0m}	N_{f+m}
合計（*t*）	N_{+1m}	N_{+0m}	N_{++m}

若對不同的得分組別而言，男生答對該題的比例與女生答對相同或非常相近時，該試題將被視為沒有 DIF。對第 *i* 題而言，可從 *k* 個 2×2 列聯表中計算並校正得到 MH 統計量如下：

$$\chi^2_{MH}=\frac{\left[\left|\sum_m N_{\mathrm{r1m}}-\sum_m E(N_{\mathrm{r1m}})\right|-0.5\right]^2}{\sum_m Var(N_{\mathrm{r1m}})} \qquad (9)$$

其中，$E(N_{\mathrm{r1m}})=\frac{N_{\mathrm{r+m}}N_{\mathrm{+1m}}}{N_{\mathrm{++m}}}$，$Var(N_{r1m})=\frac{N_{\mathrm{r+m}}N_{\mathrm{f+m}}N_{\mathrm{+1m}}N_{\mathrm{+0m}}}{N^2_{\mathrm{++m}}(N_{\mathrm{++m}}-1)}$

上述 MH 統計量 χ^2_{MH} 服從自由度為 1 的卡方分配，若 $\chi^2_{MH}>\chi^2_{(1)}=3.84$，則該試題將被判定為 DIF，而 DIF 的效果量 α_{MH} 則可透過下式進行估計之：

$$\alpha_{MH}=\left[\frac{\sum_m N_{\mathrm{r1m}}N_{\mathrm{f0m}}}{N_{\mathrm{++m}}}\right]\Big/\left[\frac{\sum_m N_{\mathrm{f1m}}N_{\mathrm{r0m}}}{N_{\mathrm{++m}}}\right] \qquad (10)$$

本研究以免費軟體 R（R Core Team, 2015）之套件"difR"（Magis, Beland, Tuerlinckx, & De Boeck, 2010）中的"difMH"函數對各試題進行 DIF 檢測，由於"difMH"在 DIF 檢測中納入量尺淨化（scale purification）程序，因此本研究均量尺淨化後之結果計算相關數據。

在 DFF 檢測部分，本研究根據前述公式（8）之公式，對測驗進行 DFF 檢測，在本模擬研究中，假設試題特徵及 Q 矩陣均被正確的界定。接著分別以 LLTM 及 LLTM-R 兩個模式進行 DFF 檢測，採用之工具為軟體 R 的套件"lme4"（Bates, Maechler, Bolker, & Walker, 2015），由於"lme4"中可使用廣義線性混合模式（generalized linear mixed models）估計多層次以及 Rasch 模式的二元資料，因此得以作為估計 LLTM 以及 LLTM-R 模式的工具。

三、研究結果

研究一的結果整理如表 4，表 4 左側與右側分別呈現 DIF 檢測及 DFF 檢測之型一誤差與檢測力。

（一）DIF 檢測結果

在本研究操弄的情境下，DIF 檢測之型一誤差皆過度膨脹，使得檢測力變得無意義。隨著 Q 矩陣的密度或 DFF 效果量的增加，DIF 檢測的型一誤差便升高，特別是在 Q 矩陣密度為 60%、DFF 效果量為 0.9 時，不論資料來自何種模式，型一誤差均膨脹至 0.9 以上。此結果可能是因為在 Q 矩陣密度高可類比於測驗中 DIF 試題的百分比高，使得整個量尺受到的污染較為嚴重，當 DFF 效果量大時，會使情況更為嚴重，加上兩群體之平均能力相差一個標準差，也會對於能力配對有所干擾，因此型一誤差容易失控。

（二）DFF 檢測結果

由於資料產生來自兩種模式，因此以下將結果分為資料型態為 LLTM 與 LLTM-R 兩種情境進行說明：

1. 當資料型態為 LLTM 時，使用 LLTM 或 LLTM-R 進行 DFF 檢測之型一誤差均可控制在 0.09 以內，在 DFF 效果量在 0.6 以上時，兩者均能正確地判別具有 DFF 的試題特徵，即便在 DFF 效果量較小時（0.3），DFF 檢測力也能達到 0.86 以上，顯示檢測結果相當穩定且可信。Q 矩陣的密度在 DFF 效果量大時並無影響，在 DFF＝0.3 時，對於檢測力僅有微幅的影響。

2. 當資料型態為 LLTM-R 時，以 LLTM 進行 DFF 檢測僅在 Q 矩陣密度較高且 DFF 效果量較低時，其型一誤差可以收到良好控制，其他情境下均會呈現較高的型一誤差。反之，以 LLTM-R 進行 DFF 檢測時，型一誤差均可受到良好控制，且檢測力均在 0.76 以上，顯示檢測結果亦屬穩定可信，惟在 DFF 效果量小（0.3）、Q 矩陣密度低（40%）時，檢測力較為不足。

為深入釐清型一誤差發生的情況，輔以計算 DFF 參數估計上之均方根誤差（root mean square error, 以下簡稱 RMSE）與估計偏誤（bias），藉以瞭解模式在 DFF 檢測上的準確度。本研究之 bias 及 RMSE 的計算公式如下：

$$\text{Bias} = \frac{1}{r}\sum_{n=1}^{r} DIF_n - TrueDIF \qquad (11)$$

$$\text{RMSE} = \left[\frac{1}{r}\sum_{N=1}^{r}(DIF_n - TruDIF)^2\right]^{\frac{1}{2}} \qquad (12)$$

公式（11）與（12）中 DIF_n 為在樣本 n 下估計的 DIF 效果量，「*True DIF*」則為研究設計中預設的 DIF 效果量，r 為重複模擬次數，此公式參照自 Sinharay、Dorans、Grant 與 Blew（2009）。由於在研究一當中，每道 DIF 試題中僅設計試題特徵二具有 DFF 效果量，因此公式中的 DIF 效果量即可視為本研究中操弄之 DFF 效果量。

DFF 檢測之參數估計的 RMSE 與 bias 列於表 5，以 LLTM 產生資料時，以 LLTM 或 LLTM-R 進行 DFF 分析的參數估計，其 bias 及 RMSE 相差不多，不過對於無 DFF 的試題特徵一及試題特徵三，兩者均是略微高估，反之試題特徵二則是呈現低估；再者，在 DFF 效果量為 0.6 以上時，試題特徵二的 RMSE 有略微呈現增加的狀況，但差異不大。比較特別的是，以 LLRM-R 產生資料時，用 LLTM 進行 DFF 分析的話，當 DFF 效果量為 0.6 以上時，bias 及 RMSE 均呈現明顯增加的情形，也可能因此造成型一誤差的膨脹。

表 4　模擬資料 DIF 檢測與 DFF 檢測結果

	DIF 檢測		DFF 檢測			
	Mantel-Haenszel 法		LLTM		LLTM-R	
	型一誤差	檢測力	型一誤差	檢測力	型一誤差	檢測力
資料模式：LLTM						
Q 矩陣密度＝40%						
DFF 效果量＝0.3	0.108	0.121	0.070	0.920	0.055	0.860
DFF 效果量＝0.6	0.229	0.353	0.050	1.000	0.050	1.000
DFF 效果量＝0.9	0.291	0.513	0.080	1.000	0.070	1.000
Q 矩陣密度＝60%						
DFF 效果量＝0.3	0.135	0.061	0.070	0.880	0.050	0.870
DFF 效果量＝0.6	0.667	0.083	0.045	1.000	0.090	1.000
DFF 效果量＝0.9	0.977	0.042	0.050	1.000	0.060	1.000
資料模式：LLTM-R						
Q 矩陣密度＝40%						
DFF 效果量＝0.3	0.102	0.112	0.115	0.960	0.070	0.760
DFF 效果量＝0.6	0.246	0.352	0.545	1.000	0.050	1.000
DFF 效果量＝0.9	0.290	0.498	0.975	1.000	0.095	1.000
Q 矩陣密度＝60%						

DFF 效果量＝0.3	0.145	0.078	0.030	0.890	0.055	0.840
DFF 效果量＝0.6	0.617	0.104	0.355	1.000	0.040	1.000
DFF 效果量＝0.9	0.983	0.049	0.405	1.000	0.040	1.000

綜合上述結果，如果某個試題特徵發生 DFF 時，如果 DFF 效果量大、且 Q 矩陣密度高的情況下，將會影響 DIF 檢測的型一誤差，如果能輔以 DFF 檢測，瞭解是否有試題特徵具有 DFF，可以提供研究者額外訊息，除了協助評估 DIF 檢測的正確性，也可以透過質性審查判斷試題特徵與 DIF 試題之間的關聯，甚至判斷 DIF 的成因。在進行 DFF 檢測時，建議可採用 LLTM-R，以隨機效果的角度來處理資料，可使 DFF 檢測的型一誤差獲得控制，並使檢測力維持在一定的水準。

表 5 DFF 檢測之參數估計的 RMSE 與 bias

	LLTM						LLTM-R					
	γ_1（Non-DFF）		γ_2（DFF）		γ_3（Non-DFF）		γ_1（Non-DFF）		γ_2（DFF）		γ_3（Non-DFF）	
	RMSE	bias	RMSE	Bias	RMSE	Bias	RMSE	bias	RMSE	bias	RMSE	bias
資料模式：LLTM												
Q 矩陣密度＝40%												
DFF 效果量＝0.3	.081	.066	.079	-.035	.110	.091	.086	.073	.083	-.025	.111	.090
DFF 效果量＝0.6	.083	.068	.097	-.052	.094	.078	.092	.076	.109	-.057	.099	.078
DFF 效果量＝0.9	.085	.068	.120	-.079	.107	.088	.090	.072	.126	-.082	.114	.093
Q 矩陣密度＝60%												
DFF 效果量＝0.3	.072	.058	.082	-.040	.101	.081	.076	.060	.085	-.042	.109	.088
DFF 效果量＝0.6	.069	.053	.106	-.056	.092	.075	.082	.065	.108	-.062	.107	.085
DFF 效果量＝0.9	.071	.056	.113	-.077	.102	.084	.086	.068	.115	-.071	.106	.085
資料模式：LLTM-R												
Q 矩陣密度＝40%												
DFF 效果量＝0.3	.109	.086	.091	.021	.083	.067	.100	.082	.105	-.044	.101	.079
DFF 效果量＝0.6	.176	.159	.254	-.242	.205	.189	.088	.070	.117	-.072	.104	.086
DFF 效果量＝0.9	.340	.329	.462	-.456	.336	.324	.097	.074	.131	-.090	.114	.092
Q 矩陣密度＝60%												
DFF 效果量＝0.3	.079	.067	.082	-.040	.087	.071	.089	.070	.090	-.039	.108	.087
DFF 效果量＝0.6	.133	.118	.221	-.207	.172	.154	.079	.063	.110	-.066	.099	.077
DFF 效果量＝0.9	.143	.124	.246	-.236	.188	.171	.091	.073	.113	-.079	.097	.078

研究二　以實徵資料探討 DIF 成因

一、資料說明

本研究使用的實徵資料來自 De Boeck 與 Wilson（2004）《Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach》一書中所提供之言語攻擊（verbal aggression）問卷資料，讀者可於下列網站（http://bearcenter.berkeley.edu/page/materials-explanatory-item-response-models）下載，同時亦為免費軟體 R 中的"difR"以及"lme4"模組中的附件，屬於公開資料，且已在其他研究中使用（Choi & Wilson, 2015; De Boeck & Wilson, 2004）。

問卷資料內容主要敘述在挫敗的情境下可能伴隨的言語攻擊反應，共有 24 道二元計分試題，將作答「否」者記為 0，回答「或許」及「是」者記為 1。資料中有 316 名受試者，包含 243 位女性以及 73 位男性。量表中包含的試題特徵及其定義如下：

（一）行為模式（Behavior Mode）

分為兩個不同的層次：想要進行（Want）以及實際上進行（Do）。

（二）情境類型（Situation Type）

分為兩個層次：歸咎他人（Other-to-blame）以及歸咎自己（Self-to-blame）。

（三）行為類型－指責（Behavior Type-blame）

本項特徵與「行為類型-表達」均由詛咒（Curse）、罟罵（Scold）以及咆哮（Shout）三個層次的行為類型分類而來，詛咒與罟罵屬於此類型。

（四）行為類型-表達（Behavior Type-express）

包含前述的罟罵與咆哮。

本問卷的試題特徵編碼方式詳如表 6，且參考 De Boeck 與 Wilson（2004）提供各題包含之試題特徵，編碼後顯示如表 7 之 Q 矩陣。

表 6　言語攻擊問卷的試題特徵編碼方式

試題特徵	編碼方式	
行為模式	實際上進行= 1	想要進行= 0
情境類型	歸咎他人= 1	歸咎自己= 0
行為類型－指責	詛咒、罟罵= 0.5	咆哮= -1
行為類型－表達	罟罵、咆哮= 0.5	詛咒= -1

資料來源：Wilson, M., & De Boeck, P. (2004). *Descriptive and explanatory item response models. In Explanatory item response models* (p. 63). Springer New York.

二、研究方法

研究二的主要目的在於透過 DFF 檢測的結果，得以探討 DIF 是否由某些試題特徵所造成，以便作為後續試題質性修正的方向。本研究樣本中男性人數較少，因而分別將男性與女性設定為焦點群體及參照群體，並將男性編碼設定為 1，若 DFF 效果量顯著大於 0，則代表男性在此試題特徵上較為有利。進行 DFF 分析時，本研究依循以下步驟進行之：

（一）建立測驗的 Q 矩陣

應先確認測驗當中之試題特徵，此時應注意的是，Q 矩陣之試題特徵間應該盡量有清楚的區隔，避免發生所謂的共線性（collinearity）的問題，如此一來除了可以避免試題特徵間因高度相關而難以解釋之外，也可避免具有 DFF 的試題特徵影響其他高度相關的試題特徵之參數估計，進而造成 DFF 檢測的型一誤差膨脹。

（二）進行 DFF 檢測

使用 LLTM 或 LLTM-R 模式，分別依據公式（5）及公式（8）進行 DFF 檢測。

（三）判讀 DFF 檢測結果，並建立試題特徵與 DIF 之間的連結

1. 未檢測出具 DFF 之試題特徵：這些試題特徵與 DIF 試題之間的關聯並不明顯。

2. 多數試題特徵皆顯示具有 DFF：此時可能發生 DFF 檢測的型一誤差膨脹的狀況，研究者可回到步驟 1 檢查試題特徵之間是否有共線性問題；再者，如果是使用 LLTM 模式進行 DFF 檢測，則可考慮使用 LLTM-R 模式。

3. 部分試題特徵具有 DFF：可進一步分析 DIF 試題與 DFF 試題特徵間之對應，是否 DIF 試題具有特殊的試題特徵組合，並將相關發現與 DIF 分析結果提供質性審查及修題時參考。

研究二之 DIF 及 DFF 檢測方法均與研究一相同，關於分析過程中對於各模式的評估，本研究採用 Akaike's information coefficient（AIC; Sakamoto, Ishiguro, & Kitagawa, 1986）及 Schwarz's Bayesian information coefficient（BIC; Schwarz, 1978）等適配度指標，兩項指標的計算公式分列如下：

$$AIC = -2\log L(\theta) + 2k \qquad (13)$$

$$BIC = -2\log L(\theta) + k\log n \qquad (14)$$

其中 $L(\theta)$ 為likelihood，k為模式中估計的參數個數，n代表樣本數。兩種指標均可用於檢驗LLTM與LLTM-R兩種模式與資料的適配度，數值越接近0者愈佳（Kaplan, 2009），其中BIC可修正AIC指標過度配適的問題，在樣本數較大時，較能正確地選取模式。

三、分析結果

研究二中各試題之 DIF 分析結果呈現於表 7 中，結果顯示第 6、8、14、16、17、19、20、22 與 23 題具有 DIF，DIF 試題占總試題數的 37.5%。在探討 DIF 效果可能來自於哪一個試題特徵時，可先觀察量表的 Q 矩陣，若 DIF 試題都僅具有特定的試題特徵，即可推論 DIF 效果可能是來自於該些試題特徵。然觀察本研究之 Q 矩陣可發現，部分 DIF 試題同時具有四個試題特徵，因而無法直接由 Q 矩陣推論 DIF 效果與哪一個試題特徵的關聯性最高。故此，本研究將同時採用 LLTM 以及 LLTM-R 進行 DFF 檢測，以掌握更多與 DIF 相關的資訊。

表 7　言語攻擊量表的 Q 矩陣與 DIF 檢測結果

試題特徵	Q 矩陣				統計量	*p* 值
	行為模式	情境類型	行為類型－指責	行為類型－表達		
試題 1	0	1	0.5	0.5	0.007	.934
試題 2	0	1	0.5	-1.0	0.038	.846
試題 3	0	1	-1.0	0.5	0.009	.926
試題 4	0	1	0.5	0.5	0.888	.346
試題 5	0	1	0.5	-1.0	0.111	.739
試題 6	**0**	**1**	**-1.0**	**0.5**	**4.268**	**.039***
試題 7	0	0	0.5	0.5	0.099	.753
試題 8	**0**	**0**	**0.5**	**-1.0**	**4.372**	**.037***
試題 9	0	0	-1.0	0.5	0.345	.557
試題 10	0	0	0.5	0.5	0.141	.708
試題 11	0	0	0.5	-1.0	1.685	.194
試題 12	0	0	-1.0	0.5	1.077	.300
試題 13	1	1	0.5	0.5	2.096	.148
試題 14	**1**	**1**	**0.5**	**-1.0**	**6.274**	**.012***
試題 15	1	1	-1.0	0.5	0.017	.896
試題 16	**1**	**1**	**0.5**	**0.5**	**9.667**	**.002***
試題 17	**1**	**1**	**0.5**	**-1.0**	**11.944**	**.001***
試題 18	1	1	-1.0	0.5	0.700	.403
試題 19	**1**	**0**	**0.5**	**0.5**	**9.464**	**.002***
試題 20	**1**	**0**	**0.5**	**-1.0**	**6.436**	**.011***
試題 21	1	0	-1.0	0.5	1.419	.234
試題 22	**1**	**0**	**0.5**	**0.5**	**3.932**	**.047***
試題 23	**1**	**0**	**0.5**	**-1.0**	**5.799**	**.016***
試題 24	1	0	-1.0	0.5	0.322	.570

*$p < .05$.

表 8 列出 LLTM 以及 LLTM-R 的 DFF 檢測結果，根據 AIC 與 BIC 指標，均顯示以 LLTM-R 模式進行 DFF 檢測與此份資料的適配度較佳，因此以下將以 LLTM-R 的結果進行說明。LLTM-R 的分析結果顯示試題特徵「行為模式」與性別之交互作用，亦即「行為模式」之 DFF 效果量為 0.811；試題特徵「行為類型-指責」之 DFF 效果量為 0.317，兩者均達顯著，意謂對男性而言，在兩個試題特徵上傾向比女性得到高分。得到前述結果後，可對照表 7 的 Q 矩陣，發現被檢測為 DIF 的試題中，除了試題 6 及試題 8 之外，其餘 DIF 試題在「行為模式」上的編碼均為 1（代表實際上進行），此外，除了第 6 題之外，其餘 DIF 試題的「行為模式—指責」的編碼均為 0.5（代表詛咒、罟罵）；針對所有「實際上進行」且「指責方式為詛咒、罟罵」的試題，僅有第 13 題並未出現 DIF，根據 DIF 統計量，發現這些試題均對男生比較有利，亦即男生比較容易得高分，這是相當關鍵且值得注意的資訊。故而，透過 DFF 檢測，可以更精準的掌握試題特徵與 DIF 試題間的關聯，進而作為後續試題修改上的建議，並提升修題的效率。尤有甚者，如能進一步探討得知這些特徵與 DIF 成因之關聯，更可作為教師進行教學與命題時的參考，具有相當重要的意義。

表 8　言語攻擊量表的 DFF 檢測結果

參數項	參數估計值	
	LLTM	LLTM-R
固定效果參數		
截距項	0.312*	0.330*
性別	0.039	0.030*
行為模式	0.859*	0.898*
情境類型	-1.067*	-1.096*
行為類型－指責	-1.294*	-1.336*
行為類型－表達	-0.739*	-0.742*
行為模式*性別	0.790*	0.811*
情境類型*性別	-0.127	-0.139
行為類型－指責*性別	0.322*	0.317*
行為類型－表達*性別	-0.122	-0.123
隨機效果參數		
受試者	1.810	1.901
試題		0.120
AIC	8202.7	8115.3
BIC	8279.0	8198.5

AIC = Akaike's information coefficient；BIC=Schwarz's Bayesian information coefficient
*p < .05.

結論與建議

回顧國內對於 DIF 分析的研究，發現藉由 DIF 檢測探討應試群體上的公平性或差異者為數眾多，顯示國內對此議題的關注程度。其中，部分研究者會進一步探討與 DIF 成因有關之議題，其中主要採用的多是質性方法，惟使用單一方法進行 DIF 成因的探討尚有不足，如果能有量化分析的證據輔助專家審查，可對 DIF 成因的判斷有所幫助（Ercikan, 2002）。因此，本研究希望透過進行 DFF 檢測，掌握試題特徵與 DIF 之間的關聯，藉以提供更多與 DIF 成因有關的資訊，並且對於未來命題時如何避免 DIF 試題的出現也有所幫助。

以往探討 DIF 成因時，多以執行審查與 DIF 成因有關的量化方法包含差異題群分析以及差異誘答項功能。透過試題層次之 DIF 效果的累積以及結合質性審查，差異題群功能分析可用以確認一群試題的 DIF 成因與潛在來源。相較於 DIF 檢測，差異題群功能檢測可以找出較多具有性別差異的試題（Mendes-Barnett & Ercikan, 2006）。差異誘答項功能檢測主要是比較不同群體在各誘答選項上的選答機率是否具有差異，如果兩群體在其中某個誘答選項呈現選答機率的差異，DIF 效果可能與此誘答選項的概念有關。

除了上述兩個方法之外，也可透過本研究所介紹的 DFF 來建立探討 DIF 成因的基礎。相較於前述兩個方法，利用試題特徵層面進行 DFF 分析具有以下幾項優點：首先，相較於試題層面，試題特徵層面與測驗層面更為適配，在檢測上可提供更可靠之分析結果（Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001）；其次，使用 LLTM 或 LLTM-R 進行 DFF 分析，可以找出具有 DFF 的試題特徵，如能進一步結合 DIF 檢測的結果，便可建立試題特徵與 DIF 試題之關聯，甚至找出 DIF 成因，俾便作為命題者進行修題時的參考；第三、透過檢測出具有 DFF 之試題特徵，可協助命題者瞭解哪些試題特徵較易產生 DFF，可於爾後測驗編製時特別留意而減少試題之 DIF 現象。

本研究採用模擬資料及實徵資料分別進行 DFF 檢測，模擬研究結果顯示 DIF 檢測會受到試題特徵之 DFF 效果量的影響，DFF 效果量越大，DIF 檢測的型一誤差便越膨脹，特別是矩陣密度較高（例如 60%）的時候。此外，如果資料中存在著隨機變異，但研究者以 LLTM 模式進行 DFF 分析，而未考量到這些隨機變異時，在 DFF 效果量為 0.6 以上的情況下，會使得參數估計的 bias 及 RMSE 增加，進而也會造成 DIF 檢測之型一誤差的膨脹。另一方面，如果改以 LLTM-R 進行 DFF 檢測，由於模式中有考量到隨機變異，因此 LLTM-R 可以相當正確地判斷具有 DFF 的試題特徵，從模擬研究結果也可發現以 LLTM-R 進行 DFF 檢測時，無論資料服從 LLTM 或是 LLTM-R，LLTM-R 均可得到較為精準的分析結果。從而可見，在研究者可以界定出測驗中的試題特徵、進而得到 Q 矩陣的情況下，同時針對測驗進行 DFF 與 DIF 檢測，可以對於測驗品質以及測驗公平性的控制發揮互補效果，也有機會可進一步達到從命題端控制具有 DFF 之試題特徵的出現頻率，進而使測驗更為公平。

實徵資料部分，本研究採用言語攻擊問卷資料檔為範例，以經常被使用的 MH 法進行性別的 DIF 檢測，結果發現具有 DIF 現象的試題占總試題數之 37.5%，屬於高比例 DIF 的測驗情境，為了使測驗所測量的潛在特質對於不同群體具有相同的意涵，建議應該針對這些試題進行修改；為了對修題者提供更多有效的訊息以協助修題，本研究接著分別以 LLTM 與 LLTM-R 進行 DFF 檢測，透過模式—資料適配度指標 AIC 及 BIC 所提供的訊息，這筆測驗資料與 LLTM-R 較為適配，從而根據 LLTM-R 的估計結果，判斷「行為模式」以及「行為類型─指責」兩個試題特徵具有 DFF。進一步將 DIF 及 DFF 檢測結果對應到測驗的 Q 矩陣，可找出試題特徵與 DIF 試題間的可能關聯。舉例而言，在研究二中可發現，針對所有「實際上進行」且「指責方式為詛咒、罟罵」的試題（亦即行為模式上編碼為 1 且行為類型—指責的編碼為 0.5）中，發現僅第 13 題並未出現 DIF，其餘 8 題均有 DIF，且這些試題都是男生比較傾向得到高分。由此推論，具有「實際上進行」且「指責方式為詛咒、罟罵」特徵的試題，應該與 DIF 之間存在著相當程度的關聯，是否是試題的敘述或是情境對於男生較為熟悉，以致男生傾向得到高分，值得作為研究者後續探討此一領域之 DIF 成因上之基礎。

使用 LLTM 或 LLTM-R 時，研究者對於試題特徵的掌握實為最關鍵的部份。回顧以往的研究，若於編製測驗時便已經考量其試題特徵，則研究者經常採用已掌握的試題特徵做為 Q 矩陣（林月仙，2013；黃宏宇、洪素蘋，2009）；如若研究者無法確認試題特徵，則可透過其他統計方式（如多元迴歸）確認試題特徵對於試題難度的解釋力後，再進行 LLTM 的分析（張銘秋、謝秀月、徐秋月，2010）。此外，由於 LLTM-R 中將隨機效果加以模式化，通常會與真實資料更為適配，因此建議研究者亦可同時使用 LLTM 與 LLTM-R 對資料進行 DFF 檢測。

進行 LLTM 的參數估計時，現時有許多軟體可供選擇，IRT 領域的軟體如 ConQuest（Wu, Adams, & Wilson, 1998）以及 Winsteps（Linacre, 2017）均可進行 LLTM 的分析。也有研究者採用 SAS 的 NLMIXED 程序（Xie & Wilson, 2008；黃宏宇、洪素蘋，2009 等）以及 WinBUGS（Spiegelhalter, Thomas, Best, & Lunn, 2003; De Boeck, 2008）。另一方面，在估計 LLTM-R 的部分，由於模式較為複雜，需要估計隨機變異參數，在進行模式分析與 DFF 檢核上可採用 SAS 中的 GLIMMIX 程序進行。然而，如果要同時能夠估計 LLTM 及 LLTM-R，除了多層次模式分析軟體 MLwiN（Rasbash, Charlton, Browne, Healy, & Cameron, 2009）（Van den Noortgate & De Boeck, 2005; 林月仙，2013）之外，自由軟體 R 中的套件“lme4”也可同時分析 LLTM 與 LLTM-R，這些軟體均可協助研究者以 LLTM 或 LLTM-R 進行 DFF 檢核。

本研究嘗試以 LLTM 及 LLTM-R 來探討試題特徵與 DIF 試題之間的關聯，作為探究 DIF 成因的基礎，此部分以往探討的研究者並不多，故而在本研究的模擬研究中僅初步操弄 3 個試題特徵與 30 道試題，產生兩種密度之設計矩陣，同時所有設定皆在兩群體具有平均能力差異之情境，因此本研究結果之推論性將相對較為受限。在後續研究上可嘗試建構其他真實情境，進一步操弄如矩陣密度型式、題數、試題特徵的解釋力、試題特徵 DFF 效果量之形式以及試題特徵間相關等變項，以進一步釐清其他試題特徵與 DIF 試題之間的關聯。再者，雖本研究中以試題特徵層面來進行 DFF 檢測，但 Zumbo 等人（2015）亦曾提出 DIF 研究可以依照生態歸納為五個層次，建議未來研究者可透過上述對測驗分數有系統性影響之層面進行 DFF 檢測，以獲得更多 DIF 成因資訊。

參考文獻

王佳琪、何曉琪、鄭英耀（2014）：「科學創造性問題解決測驗」之發展。**測驗學刊，61**（3），337-360。[Wang, C. C., Ho, H. C., & Cheng, Y. Y. (2014). Development of the children scientific creative problem solving test. *Psychological Testing, 61*(3), 337-360.]

林月仙（2013）：中文色塊測驗之認知成分分析：LLTM 與 SEM 取向。**教育與心理研究，36**（2），113-144。[Lin, Y. H. (2013). Validation of cognitive structures for the mandarin token test: The linear logistic test model and structural equation modeling. *Journal of Education & Psychology, 36*(2), 113-144.]

侯雅齡（2013）：高級中學自然科學術性向測驗編製。**科學教育學刊，21**（2），189-213。[Hou, Y. L. (2013). The development of natural science academic aptitude tests for high school students. *Chinese Journal of Science Education, 21*(2), 189-213.]

張銘秋、謝秀月、徐秋月（2010）：PISA 科學素養之試題認知成份分析。**課程與教學，13**（1），1-20。[Chang, M. C., Hsieh, H. Y., & Shyu, C. Y. (2010). A cognitive component analysis for PISA science literacy. *Curriculum & Instruction Quarterly, 13*(1), 1-20.]

曾明基、邱皓政（2015）：研究生評鑑教師教學的結果真的可以與大學生一起比較嗎？多群組混合 MIMIC-DIF 分析。**測驗學刊，62**（1），1-23。 [Tseng, M. C., & Chiou, H. J. (2015). Graduate student in SRI can really compare with the university student? Multi-group mixture MIMIC-DIF analysis. *Psychological Testing, 62*(1), 1-23.]

黃宏宇、洪素蘋（2009）：建構效度檢驗之線性與非線性取向：以學生創意自我效能量表為例。**屏東教育大學學報－教育類，33**，489-513。[Huang, H. Y., & Hung, S. P. (2009). A study of examining construct validity on the scale of creative self-efficacy for students through both linear and nonlinear approaches. *Journal of Pingtung University of Education: Education, 33*, 489-513.]

廖彥棻（2015）：英文學科能力測驗選擇題之性別差異與差異試題功能分析。**東吳外語學報，41**，21-59。[Liao, Y. F. (2015). Gender differences and differential item functioning on the English GSAT multiple-choice questions. *Soochow Journal of Foreign Languages and Literatures, 41*, 21-59.]

賴姿伶、余民寧（2015）：應徵者與在職者在多分題人格測驗的作答差異之研究：試題層次與試題組合層次的分析。**人力資源管理學報，15**（4），91-120。[Lia, T. L., & Yu, M. N. (2015). Response variation on polytomous personality measures between applicants and incumbents: Analyses on item-level and item-composite level. *Journal of Human Resource Management, 15*(4), 91-120.]

蕭偉智、傅家珍（2012）：國中八年級自然科定期評量之性別差別試題功能（DIF）分析。**新竹教育大學教育學報，29**（2），35-64。[Hsiao, W. C., & Fu, C. C. (2012). Gender differential item functioning in a science periodical test of eighth graders. *Educational Journal of NHCUE, 29*(2), 35-64.]

蘇旭琳、陳柏熹（2008）：DIF 分析在小樣本情境中的偵測效果－以視障生和普通生在國中基測數學科之 DIF 爲例。**測驗學刊，55**（4），761-791。[Su, H. L., & Chen, P. H. (2008). Detecting differential item functioning in small sample size conditions: An empirical study of the DIF detection in basic competence test for junior high school students. *Psychological Testing, 55*(4), 761-791.]

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillside, NJ: Lawrence Erlbaum.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17*, 201-210.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version, 1*(7), 23. Retrieved March 18, 2015, from http://CRAN. R-project. org/package= lme4.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated item scores a comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement, 72*(5), 754-773.

Bolt, D. (2002, Aril). *Studying the potential of nuisance dimensions using bundle DIF and multidimensional IRT analyses.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans: LA.

Choi, I. H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and Psychological Measurement, 75*, 78-101.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-Bundle DIF Hypothesis Testing: Identifying Suspect Bundles and Assessing Their Differential Functioning. *Journal of Educational Measurement, 33*(4), 465-484.

Drabinová, A., & Martinková, P. (2016). *Detection of differential item functioning with non-linear regression: Non-IRT approach accounting for guessing*. Retrieved May 11, 2017, from http://hdl.handle.net/11104/0259498.

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

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3-4), 199-215.

Ercikan, K., Arim, R. G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice, 29*(2), 24-35.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40*(4), 281-306.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26-36.

Gierl, M. J., & Bolt, D. M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1*(3-4), 249-270.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164-187.

Green, K. E., & Smith, R. S. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369-381.

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

Janssen, R. (2010) Modeling the effect of item designs within the Rasch model. In. S.E. Embretson (Ed.), *Measuring psychological constructs: Advances in modelbased approaches* (pp. 227-245). Washington, DC: American Psychological Association.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.). *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189-212). New York, NY: Springer-Verlag.

Jin, K. Y., & Wang, W. C. (2017). Assessment of Differential Rater Functioning in Latent Classes with New Mixture Facets Models. *Multivariate Behavioral Research, 52*(3),391-402.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions (2nd ed.)*. Los Angeles, CA: Sage.

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

Linacre, J. M. (2017). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps. com.

Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.

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.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 213-240). Springer New York.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education, 24*, 1-18.

R Core Team. (2015). *R: A language and environment for statistical computing* [Interne]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009) *MLwiN Version 2.10*. Centre for Multilevel Modelling, University of Bristol.

Rasch, G. (1960). *Probabalistic models for some intelligence and attainment tests.* Copenhagen, Denmark: The Danish Institute for Educational Research.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics.* Dordrecht, The Netherlands: D. Reidel.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009). Using past data to enhance small sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34*, 74-96.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D. (2003). *WinBUGS version 1.4 users manual.* MRC Biostatistics Unit, Cambridge. Retrieved 6 June, 2005, from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*, 443-464.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189-212). New York, NY: Springer-Verlag.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *ACER ConQuest: Generalized item response modelling software manual*. Melbourne, Victoria: The Australian Council for Educational Research Ltd.

Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: an international testing context. *Psychology Science, 50*(3), 403.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, Ontario, Canada: Directorate of human resources research and evaluation, department of National defense.

Zumbo, B. D. (2007). Three generation of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly: An International Journal, 4*, 223-233.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*(1), 136-1.

收 稿 日 期：2017 年 03 月 31 日
一稿修訂日期：2017 年 04 月 07 日
二稿修訂日期：2017 年 11 月 03 日
三稿修訂日期：2017 年 12 月 06 日
接受刊登日期：2017 年 12 月 06 日

Bulletin of Educational Psychology, 2018, 50(2), 167-188
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Investigating Sources of Differential Item Functioning: the Relationship Between Item Property and Differential Item Functioning

Guo-Wei Sun
Office of Institntional Research
Kaohsiung Medical University

Cheng-Te Chen
Department of Educational Psychology and Counseling
National Tsing Hua University

Ching-Lin Shih
Center for Teacher Education
Assessment Research Center
National Sun Yat-sen University

Because assessment methods for differential item functioning (DIF) have been developed and thoroughly investigated, the focus in DIF research has shifted to explaining DIF phenomena. Experts in this field are recruited to tap possible sources of DIF. Quantitative analysis results help experts reviewing DIF to locate sources for DIF items. This study aimed to demonstrate the use of the differential facet functioning (DFF) procedure implemented using the linear logistic test model (LLTM) and random effects linear logistic test model (LLTM-R) to explain possible DIF sources. The efficiency of LLTM and LLTM-R in detecting DFF under various conditions was also evaluated. The simulation results indicated that the DIF effect was significantly influenced by the DFF effect of item properties. Moreover, as the design matrices had a high density (e.g., 60%), Type-I error rates of DIF assessment were seriously inflated. We also demonstrated the procedure of DFF analysis with an empirical data. The result showed that most DIF items were related to two item properties, which would be provided as possible DIF sources in the item-review meeting. Researchers should implement DFF assessment using LLTM-R to help explain DIF sources.

KEY WORDS: DIF source, Differential item functioning, Differential facet functioning, Linear logistic test model, Random effects linear logistic test model