

多面向混合極端反應風格模式在評分者中介評量的 發展與應用

宋易安¹ 黃宏宇²

摘要

為了解在評分者中介評量中，評分者是否會因為受試者潛在類別不同而有不同的反應風格，本研究旨在發展「多面向混合極端反應風格試題反應理論模式」，並檢驗其有效性。透過模擬研究的設計，在各種操弄情境下，使用貝氏估計法檢驗模式參數的回復性。模擬研究結果顯示：模式在試題參數、評分者參數、受試者能力參數、潛在類別的估計有良好的估計效果，且隨著量表點數、題項數、樣本數愈大，其估計效果愈好；但忽略潛在類別時，所有參數估計精準度皆會降低。而在實徵研究中使用泰國一所國際學校對學生的英文寫作測驗作為資料來源，以檢驗發展的模式在實務上之適配情形，結果顯示：在多面向混合極端反應風格概化部分計分模式有最佳的適配度。最後，研究者也提出相關建議供後續研究參考。

關鍵詞：混合模式、評分者中介評量、極端反應風格、試題反應理論

1. 宋易安，國立臺灣師範大學教育心理與輔導學系博士生

2. 黃宏宇，臺北市立大學心理與諮商學系特聘教授

收件日期：2022.04.29；完成修改：2022.08.18；正式接受：2022.08.18

通訊作者：黃宏宇；Email：hyhuang@go.utapei.edu.tw

地址：100234 臺北市中正區愛國西路1號 臺北市立大學心理與諮商學系

Development and Application of Many-facet Mixture Extreme Response Models for Rater-mediated Assessments

Yi-An Song¹ Hung-Yu Huang²

Abstract

In order to understand whether raters in rater-mediated assessments exhibit different response styles on rating scales due to different latent classes of ratees, this study aims to develop a "many-facet mixture extreme response style item response model" and assess the efficiency of the proposed model. This study consists of two parts: simulation study and empirical study. The simulation study results show that the model has good parameter estimation for item and person parameters and that the precision of parameter estimates declines when the misleading model was used. In the empirical study, the data that were collected from writing assessment administered to an international school in Thailand were fit to several competing models to demonstrate the application of the proposed model. The empirical study shows that the many-facet mixture extreme response style generalized partial credit model provides the best-fitting model among the competing models. Moreover, the changes in rank order between the best-fitting model and its corresponding model without considering different response styles and latent classes are not trivial. Finally, the author addresses conclusions based on the results and provides suggestions for future study.

Keywords: extreme response style, rater-mediated assessments, mixture model, item response theory

1. Yi-An Song, PhD student, Department of Educational Psychology and Counseling, National Taiwan Normal University

2. Hung-Yu Huang, Distinguished Professor, Department of Psychology and Counseling, University of Taipei

Received: 2022.04.29; Revised: 2022.08.18; Accepted: 2022.08.18

Corresponding Author: Hung-Yu Huang; Email: hyhuang@go.utapei.edu.tw

Address: No. 1, Aiguo W. Rd., Zhongzheng Dist., Taipei City 100234, Taiwan

Department of Psychology and Counseling, University of Taipei

壹、緒論

在教育評量的情境中，除了選擇反應題型外，研究者也常使用建構反應題型來了解受試者在高層次認知能力的表現情形，例如：在國中教育會考中的寫作題型、研究所入學考試的申論題型，以及藝術表演工作的實作題型等。在這些評量中，需要透過人為評分者（*rater*）給分的方式給予表現評比，而每位評分者評定的標準與趨勢不同則會影響到受試者所獲得的分數。在高風險的測驗情境中，受試者的表現分數若無法得到客觀評分，則可能影響其未來升學、就業或生涯發展。因為人為評分過程中涉及主觀經驗，即使密集的訓練也無法完全避免評分者偏誤的涉入，因此需要透過心理計量模型的發展，來適當排除與控制評分者可能產生的評分偏誤。

過去曾針對評分者偏誤進行相當多的探究，其中有兩類的偏誤常被拿來討論與分析，分別為評分者嚴格或寬容偏誤，以及評分者使用評等量尺進行評分時常見的極端或趨中偏誤（Myford & Wolfe, 2004）。在試題反應理論（*item response theory, IRT*）的架構下，評分者嚴格或寬容程度可以表徵在機率模型之中，進行調整受試者真實能力之估計；而評分者在評分過程中過度使用中間評等或極端評等，也可以透過加權參數來校正評分偏移的現象。然而，受試者的特性或答題習性，可能會影響評分者的反應風格，亦即評分者的評分趨勢可能會隨受試者群體不同而改變。Jin 與 Eckes（*in press*）的研究已發現不同評分者的嚴格度與極端反應風格會隨著受試者性別不同而有所差異，然而他們的研究侷限在可觀察的分類屬性之差異，並無法針對無法觀察的潛在群體進行分析。若能透過潛在類別的模式分析，即可以將不同受試群加以偵測，並分析評分者在不同受試群中是否存在不同的極端反應風格。但目前還沒有針對評分者極端反應風格在這方面的討論，故本研究欲透過發展新模式——「多面向混合極端反應風格模式」，來討論評分者極端反應風格與受試者之間的關係。

貳、文獻探討

一、評分者評分的偏誤

在使用評分者來進行評分的情境下，不同評分者對於測驗或評分標準的理解、評分嚴格程度、對受試者的初步印象，或被受試者的文化背景影響之程度等都會有所不同（Prieto & Nieto, 2014），因為上述原因造成這些評分者的給分偏離受試者應得分數之現象稱為評分者偏誤（rater bias）。受試者通常不希望看到與測量之構念無關的因素影響到評分結果，但評分者偏誤是使用評分者測驗裡分數產生變異的原因之一（Eckes, 2009）。Myford 與 Wolfe（2004）歸類常見的評分者偏誤有：嚴格／寬容偏誤（severity/leniency bias）、極端／趨中偏誤（extremity/centrality bias）、月暈效應（halo effect）、非一致性（inconsistency），以及似我效果（similarity）。

Myford 與 Wolfe（2004）將評分者嚴格度定義為：當考慮其他評分者對相同受試者的評分時，評分者給予的分數平均低於其他評分者之評分分數；反之則為評分者寬容度。評分者嚴格度在其給定分數時建立了一套標準，是一個非常重要的因素，因此使用評分者評定的資料必須考慮評分者嚴格度，才不會讓受試者所得的分數有不公平之現象。

在自陳式量表中，個體過度使用極端或趨中選項的情況被視為一種反應風格（response style），亦即受試者在回答問卷或量表時傾向某些反應類別的情況（Paulhus, 1991）。Baumgartner 與 Steenkamp（2001）整理了自陳量表中常見的七類反應風格，分別為：默認反應風格（acquiescence response style, ARS）、非默認反應風格（disacquiescence response style, DARS）、淨默認反應風格（net acquiescence response style, NARS）、極端反應風格（extreme response style, ERS）、反應範圍（response range, RR）、趨中反應風格（mid-point responding, MPR），以及非一致反應風格（noncontingent responding, NCR）。其中，極端反應風格最常被探討（Batchelor & Miao, 2016; Greenleaf, 1992; Huang, 2016; Ilgun Dibek, 2020; Jin & Wang, 2014, 2018）。若將此類的答題傾向應用在評分者評分的過程，便形成評分者的極端反應風格，此係指評分者在使用評定量表時傾向選擇量表中的兩端點，例如：在一至七點的李克特氏量表中，有極端反應風格的評分者在評估時較容易選擇 1 或 7 的分數

(Bolt & Newton, 2011)；反之即為趨中反應風格，亦指在回答問題時傾向選擇中間或中立的分數(Zhang & Wang, 2020)。極端反應風格表示評分者在使用評定量表時出現了變異性，而當某評分者盡量避免使用極端分數時，則代表分數大多集中在中間，亦同時降低了評定量表的效用與評定的區別性，且因為整體評定的結果變異性降低，也會造成信、效度跟著降低(Myford & Wolfe, 2003)。

二、試題反應理論在評分者資料的應用

試題反應理論是以個別試題為觀點來探討測驗分數的意義(余民寧, 2009)。在評分者中介評量的情境中，大多使用評定量尺(rating scales)來進行評分，因此在IRT模型使用上，則以多元計分的IRT模型為主。目前常見的多元計分IRT模型，包含：評定量表模式(rating scale model, RSM)(Andrich, 1978)、概化評定量表模式(generalized rating scale model, GRSM)(Wang & Liu, 2007)、部分計分模式(partial credit model, PCM)(Masters, 1982)、概化部分計分模式(generalized partial credit model, GPCM)(Muraki, 1992)。Linacre(1989)擴展多元計分的IRT模式，將影響受試者能力估計的評分者嚴格度置入機率模型之中，形成多面向IRT模式(many-facet IRT model)。本文將以多面向IRT模式為基礎，同時考慮評分者極端反應風格之測量與受試者潛在類別，發展一個新的評分者IRT模型。

在部分計分模式的框架下，Linacre(1989)將評分者嚴格度納入PCM之中，成為多面向部分計分模式(many-facet partial credit model, many-facet PCM)，其log-odds可以表徵為：

$$\log\left(\frac{P_{nirj}}{P_{nir(j-1)}}\right) = \theta_n - (\beta_i + \tau_{ij}) - \eta_r \quad (1)$$

其中， P_{nirj} 與 $P_{nir(j-1)}$ 為評分者 r 對受試者 n 在評量項目 i 選擇分數 j 及 $j-1$ 的機率； θ_n 為受試者 n 的潛在能力或特質； β_i 為評量項目 i 的難度； τ_{ij} 為評量項目 i 在分數 j 上的閾值； η_r 為評分者 r 的嚴格度。

如果IRT模式考慮試題鑑別度時，則可以擴展概化部分計分模式至多面向IRT模式，成為多面向概化部分計分模式(many-facet generalized partial credit model, many-facet GPCM)，此模式是針對二參數模式與部分計分模式的擴展，其log-odds可以表示為：

$$\log\left(\frac{P_{nirj}}{P_{nir(j-1)}}\right) = \alpha_i[\theta_n - (\beta_i + \tau_{ij}) - \eta_r] \quad (2)$$

其中， α_i 為評量項目 i 的鑑別度，其餘參數之定義與公式(1)相同。其他多元計分 IRT 模式也可以依據上述邏輯加以擴展，若每一個評量項目設相同一組閾值參數時，多面向部分計分模式可以簡化成多面向評定量表模式（many-facet rating scale model, many-facet RSM），而多面向概化部分計分模式則可簡化成多面向概化評定量表模式（many-facet generalized rating scale model, many-facet GRSM）。

三、評分者極端反應風格模式

對於自陳式量表的極端反應風格測量，已有幾個不同的模式被發展出來。Bolt 與 Johnson（2009）擴展名義反應模式（nominal response model, NRM），發展成為多向度名義反應模式（multidimensional NRM, MNRM），以隨機效果類別斜率參數的方式來表徵受試者在自陳式量表中的極端或趨中反應風格。然而，一般的李克特式量表屬於等級給分模式，並不適用於名義反應的測量模型，因此 MNRM 有其應用之侷限性（Jin & Wang, 2014）。

針對受試者在量表選項反應的不同傾向，Wang 等（2006）發展隨機閾值 IRT 模式（random-threshold IRT model），將閾值參數視為隨機效果（random effect），認為每個人對於選項的認知或觀點可能不盡相同，例如：在五點量表中，有人認為 5 與 1 之間的閾值很大，但也有人認為其閾值是很小的（Wang & Wu, 2011）。隨機閾值 IRT 模式雖然有助於了解受試者在選填量表各選項之傾向，但是此模型只有將選答的隨機性納入考慮，並無對於極端或趨中反應風格的特定反應組型進行模式化，因此仍然無法量化極端反應的趨勢。

基於上述模式對於極端反應風格控制的侷限，Jin 與 Wang（2014）發展針對極端反應風格測驗的 IRT 模式，透過隨機效果的加權參數，來描述受試者在偏好極端或中間選項之習性。以 GPCM 為例，受試者 n 在第 i 試題上第 j 個選項與第 $j-1$ 個選項機率的 log-odds 如下：

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = \alpha_i[\theta_n - (\beta_i + \omega_n \tau_{ij})] \quad (3)$$

其中， ω_n 為受試者 n 以極端反應的權重參數形式與 τ_{ij} 相乘的隨機效果，用來控制受試者在選項選擇上極端或趨中的傾向，且假設 $\omega_n \sim \log N(0, \sigma_\omega^2)$ ，服從平均數為 0、變異數為 σ_ω^2 的對數常態分配。當 ω 數值大於 1 時，閾值之間的距離愈大，表示受試者選擇極端反應的趨勢愈小，亦即趨中反應的趨勢愈大；相反的，當 ω 數值小於 1 時，閾值之間的距離愈小，表示受試者選擇極端反應的趨勢愈大；當 ω 數值等於 1 時，則簡化成為一般的多元計分 IRT 模式；其餘參數 P_{nirj} 、 $P_{nir(j-1)}$ 、 α_i 、 θ_n 、 β_i 、 τ_{ij} 定義與公式(2)相同。

雖然公式(3)的模型適合用來解釋個體極端或趨中的反應風格，但是此模式是針對自陳式量表發展而來，並不適用於評分者中介的測量。因此，為了解評分者在評定量表中的極端反應風格，Jin 與 Wang (2018) 將評分者極端反應的個別差異性考慮進去，將權重參數的概念與評分者多面向模式相結合，發展出新的多面向模式來測量評分者與極端反應風格的資料。以 PCM 為例，其 log-odds 可以表示如下：

$$\log\left(\frac{P_{nirj}}{P_{nir(j-1)}}\right) = \theta_n - (\beta_i + \omega_r \tau_{ij}) - \eta_r \quad (4)$$

其中， ω_r 表示評分者 r 在閾值上的權重參數；其餘參數 P_{nirj} 、 $P_{nir(j-1)}$ 、 α_i 、 θ_n 、 β_i 、 τ_{ij} 、 η_r 定義與前述公式相同。 ω_r 的功能則與公式(3)類似，當評分者 r 在評分者過程中偏好趨中評分，則其 ω_r 會大於 1；當評分者 r 在評分者過程中偏好極端評分，則其 ω_r 會小於 1；若評分者 r 在評分者過程中無極端或趨中傾向時，其 ω_r 會等於 1，即為一般多面向評分者 IRT 模型。

四、多面向混合極端反應風格試題反應理論模式

在評分者中介的評量中，試題分析需要考慮三個面向，模式參數包括了受試者能力參數、試題參數、評分者參數，在大部分情況下會假設這三個參數是互相獨立的 (Jin & Wang, 2017)，但是當評分者參數因為不同受試者而有不同的評定時，則會出現差異評分者功能 (differential rater functioning, DRF) (Du et al., 1996)。目前對於 DRF 的討論僅限於評分者嚴格度，但極端反應風格也是在使用評分者資料中常被討論的偏誤，故本研究假設評分者的極端反應風格也有可能會有 DRF 的現象，意即評分者之反應風格可能會因為受試者所屬的族群而有所差異。若這些受試者群體無法直接透過外在觀察，而是發生在未知的潛在類別中時，通常會使用混合試題反應理論模式來

處理這些潛在群體（Du et al., 1996）。因此，本研究假設評分者的極端反應風格會因為受試者所屬的潛在類別而有不同之變化，並發展多面向混合極端反應風格試題反應理論模式來識別出影響評分者反應風格的受試者群體。

混合試題反應理論模式（mixture IRT model）是一種將試題反應理論與潛在類別模型（latent class model, LCM）結合的模式，主要用於解釋測驗中重要的影響因子，例如：不同的反應策略與差別試題功能的檢測（Cohen & Bolt, 2005），而透過潛在類別模型的分析即可以將有相似反應模式的人識別出來（Bertrand & Hafner, 2014）。針對多元計分的極端反應風格在混合試題反應理論的擴展上，Huang（2016）以 GPCM 進行延伸發展了混合極端反應風格模式（mixture ERS-IRT model），以區別受試者極端反應風格及欲測量的能力在試題功能之影響，此擴展的潛在類別考慮了三種不同的受試者極端反應風格（正常、極端、趨中），並且假設試題功能可能會因為潛在類別不同而有不同的反應類型，其 log-odds 為：

$$\log\left(\frac{P_{ngij}}{P_{ngi(j-1)}}\right) = \alpha_i \theta_{ng} - (\beta_{ig} + \omega_{ng} \tau_{ij}) \quad (5)$$

其中， θ_{ng} 表示潛在類別 g 中的受試者 n 之能力，假設服從平均數為 μ_{θ_g} 、變異數為 $\sigma_{\theta_g}^2$ 的常態分配； ω_{ng} 表示潛在類別 g 中的受試者 n 之極端反應權重參數，假設服從平均數為 μ_{ω_g} 、變異數為 $\sigma_{\omega_g}^2$ 的對數常態分配； P_{ngij} 與 $P_{ngi(j-1)}$ 表示在潛在類別 g 中的受試者 n 在試題 i 裡選擇的 j 選項與 $j-1$ 選項之機率； α_i 為試題 i 的鑑別度，並假設鑑別度參數不會因潛在類別的不同而改變，所以沒有下標 g ； τ_{ij} 為試題 i 裡 j 選項的閾值。

基於上述模式的發展，本研究結合了 Jin 與 Wang（2018）測量評分者極端反應風格的新多面向模式以及 Huang（2016）的混合極端反應風格模式，發展出測量評分者極端反應風格的「多面向混合極端反應風格試題反應理論模式」（many-facet mixture extreme response style item response model, MFMixERS-IRTM）。在 MFMixERS-IRTM 的架構之中，承襲 Huang 以 ERS-GPCM 延伸的混合極端反應風格模式，認為每位評分者會因為受試者來自不同潛在類別，其反應風格的閾值權重參數也會不同，但考量到參與測驗的評分者通常都受過專業訓練，其反應模式不太會在極端反應與趨中反應間移動，故本研究將反應風格分為「無反應風格組」、「有反應風格組」兩大類，有反應風格組可能為極端反應或趨中反應，例如：第一位評分者在面對

第一類受試者時，並沒有反應風格的趨勢，但在面對第二類受試者時則產生極端或趨中的反應風格。為了因應這樣的情況，本研究在閾值權重參數下標 ω 改為潛在類別 r_g ，用來表徵受試者在評分者 r 評分之下可能分屬不同的潛在類別，當 MFMixERS-IRTM 應用在 GPCM，稱之為「多面向混合極端反應風格概化部份給分模式」（MFMixERS-GPCM），其 log-odds 如下：

$$\log\left(\frac{P_{nir_{sj}}}{P_{nir_{s(j-1)}}}\right) = \alpha_i[\theta_n - (\beta_i + \omega_{r_g}\tau_{ij}) - \eta_r] \quad (6)$$

$$P(x) = \sum_{g=1}^2 \pi_g \times P(x|g) \quad (7)$$

其中， $P_{nir_{sj}}$ 與 $P_{nir_{s(j-1)}}$ 為受試者 n 被評分者 r 評分時，因其隸屬第 g 個潛在類別下（ $g = 1$ 為有反應風格， $g = 2$ 為無反應風格），評分者於試題 i 選擇選項 j 及 $j-1$ 的機率； ω_{r_g} 表示評分者 r 評分下因受試者 n 之潛在類別 g 影響的閾值權重參數； θ_n 、 β_i 、 α_i 、 τ_{ij} 、 η_r 定義與前述公式相同； $P(x)$ 為某一反應組型下的聯合機率； $P(x|g)$ 為在潛在類別 g 之下的條件機率； π_g 表示受試者在潛在類別 g 的機率，且 $\sum \pi_g = 1$ 。

若採用的多元計分模型為 PCM、GRSM、RSM，則其對應的混合極端反應模型可以標示為 MFMixERS-PCM、MFMixERS-GRSM、MFMixERS-RSM。在 IRT 模型參數估計過程中，會面臨量尺不定性的問題，因此需要透過特定的定錨（anchoring）來達到模式的正定（identification）（Embretson & Reise, 2000）。在本研究發展的 MFMixERS-GPCM 之中，研究者設定 θ_n 服從標準常態分布（平均數為 0，變異數為 1），評分者嚴格度 η_r 的平均數為 0，每一試題內的閾值平均數為 0，潛在類別為無反應風格評分群的權重參數為 1。此外，在 Huang（2016）的混合極端反應風格模式中（公式 5），他將鑑別度參數與能力參數相乘，亦即 $\alpha \times \theta$ 與試題難度參數（ β ）、閾值加權參數（ $\omega \times \tau$ ）會置於同一量尺上。然而，在研究者發展的模式中，鑑別度參數是乘上能力參數與試題參數結合後的數值；換句話說， θ 、 β 、 $\omega \times \tau$ 與 η 是置於同一量尺，在不失一般性的統計特質前提之下，透過再參數化的轉化，兩種公式的表徵可以獲得一致性的估計結果。

最後，研究者也提出這個新模型與過去模型之差異。MFMixERS-IRTM 和 Jin 與 Wang（2018）的主要差異在於，本研究假設評分者在閾值上的權重參數會隨著不同潛在群體而改變；和 Huang（2016）的差異在於，本研究提

出的模型採用固定效果之潛在類別分析，亦即當受試者被歸類在某一潛在群體時，他們在權重參數上不會具有隨機的變異量，以模型上來說較為簡易，可以加速模式參數估計的效能。為檢驗 MFMixERS-IRTM 參數估計的有效性與實際應用性，本研究包含模擬研究與實徵研究兩個部分。首先，使用模擬研究了解本研究發展之新模式在試題參數、受試者參數、評分者參數、潛在類別上的估計效果；其次，再透過實徵研究了解新模式在實際資料上的應用情形。

參、研究方法

本研究旨在發展多面向混合極端反應模式，來討論外部評分者在評分過程中可能的異質反應型態。因部分計分模式相較於評定量表模式，其每道試題的閾值不盡相同，較為符合真實情境，且試題鑑別度也是試題參數裡一個重要的指標，故選擇以 MFMixERS-GPCM 來進行主要分析。研究包含兩個部分：模擬研究與實徵研究。研究者於模擬研究使用 MFMixERS-GPCM 操弄變項並產生模擬資料，再將產生之資料使用 MFMixERS-GPCM 與 Many-facet GPCM 進行參數回復性分析，操弄之變項包含：受試者人數、量表點數、題目數，且為了更貼近真實情境資料，也操弄有遺漏值的情境，檢驗新模式在有遺漏值的資料下各參數的估計效果為何，而遺漏值情境只使用 MFMixERS-GPCM 進行分析。實徵資料則使用泰國一所國際學校對於學生寫作能力的評估，檢驗本研究發展的模式在實際資料上之適配情形。

一、模擬研究

本研究操弄的變項與參數設定參考 Jin 與 Wang (2018) 及 Huang (2016) 的模擬研究，受試者人數：500 人、1000 人；量表點數：四點量表、六點量表；試題長度（或為評分標準）：三題、六題；評分者人數固定為五人，各情境重複 30 次。使用 MATLAB (R2015b) 軟體產生資料，在參數設定方面，試題難度為介於 -1.5~1.5 之間且平均數為 0 的均勻分配；試題鑑別度為介於 0.5~1.5 之均勻分配；試題閾值參數在四點量表中設為 -0.6、0、0.6，在六點量表中設為 -1.0、-0.6、0、0.6、1.0；五位不同等級之評分者嚴格度設為 -1、-0.5、0、0.5、1。兩種潛在類別參考 Huang 的研究分為無反應風格、有反應風格，其發生機率分別為 0.6、0.4，並設定第一位、第二位

評分者於具有反應風格類別且為極端反應風格，第三位評分者的兩類潛在類別皆無反應風格，第四位、第五位評分者具有反應風格類別且為趨中反應風格，在閾值權重參數設定上無反應風格為 1、極端反應風格為 $exp(-0.5)$ 、趨中反應風格為 $exp(0.5)$ 。在遺漏值操弄設定方面，受試者 500 人之操弄為：第 1~100 位受試者只被第一、二位評分者評分，第三、四、五評分者為遺漏值；第 101~200 位受試者只被第二、三位評分者評分，第一、四、五評分者為遺漏值，依此類推。受試者 1000 人之操弄為：第 1~200 位受試者只被第一、二位評分者評分，其餘為遺漏值；第 201~400 位受試者只被第二、三位評分者評分，其餘為遺漏值，依此類推。

二、實徵研究

研究資料來源為泰國一所國際學校在 2012~2018 年間針對參加英語課程的 3~12 年級學生實施的英語寫作測驗 (Conrad II, 2020)。學生寫作分數由兩位國家評分服務機構的評分員評定，寫作題目根據奧勒岡寫作六要素 (Oregon 6 Traits Rubric) 的六項評分標準 (亦即六題)，包含：想法 (ideas)、組織 (organization)、語音 (voice)、字詞選擇 (word choice)、語句流暢度 (sentence fluency)，以及寫作規約 (conventions)，每項評分標準為六點量表。因每年之寫作指引不同，故本研究只採用 2012 年之學生數據，受試者人數共 744 位，評分者共兩位，並在 Many-facet RSM、Many-facet GRSM、Many-facet PCM、Many-facet GPCM、MFMixERS-RSM、MFMixERS-GRSM、MFMixERS-PCM、MFMixERS-GPCM 進行八種模式的模式比較，分析模式適配度，再估計各參數。本研究為混合試題反應理論模型，在實徵研究的模式比較上使用 BIC (Bayesian information coefficient) (Schwarz, 1978) 作為模式比較的適配度指標，且因每次疊代產生出的各參數估計值可能都不同，所以需要在每次疊代時監控概率 (Cho & Cohen, 2010)。其公式如下：

$$BIC = -2\log(L)' + m\log(N), \quad (8)$$

其中， $\log(L)'$ 表示在第 L 次疊代中概率的 log 值； m 表示參數的數量； N 表示樣本數。當 BIC 值愈小，代表其模式適配度愈好。

三、資料分析

本研究使用 JAGS (Plummer, 2015) 軟體來進行模擬研究與實徵研究的後續參數估計，其算則透過貝氏估計 (Bayesian estimation) 建立馬可夫鏈蒙地卡羅 (Markov chain Monte Carlo techniques, MCMC) 來產生參數的後驗分布，再從其中進行抽樣，以其期望值作為參數估計值。JAGS 的使用方式相當彈性，可以處理複雜困難的模式，能提供許多參數估計的訊息，也被使用在極端反應風格試題反應理論模型之中 (Jin & Wang, 2018)。在參數先驗分布設定方面，本研究參考 Jin 與 Wang (2018) 及 Huang (2016) 於模擬研究使用的先驗分布，將受試者能力或特質、試題鑑別度、試題難度、閾值參數、評分者嚴格度、閾值權重參數各分別設為： $\theta_n \sim N(0, 1)$ 、 $\alpha_i \sim \text{lognormal}(0, 1)$ 、 $\beta_i \sim N(0, 4)$ 、 $\tau_{ij} \sim N(0, 4)$ 、 $\eta_r \sim N(0, 4)$ 、 $\omega_r \sim \text{lognormal}(0, 1)$ 、 $\pi_g \sim \text{Dirichlet}(0.6, 0.4)$ 。值得注意的是，能力參數的常態分布是屬於群體特質的先驗分布假設，如同前面所述，為了達到模式正定的要求，而設定為標準常態分布。然而，在貝氏估計過程中，每一個估計參數都需要賦予一個合適的先驗分布；一般而言，在不確定的情境下，分布的變異數宜設較大的數值，以降低主觀先驗的影響，因此研究者採用平均數為 0、變異數為 4 的常態分布，做為試題難度參數、閾值參數、評分嚴格度參數的先驗分布。疊代部分將進行 15000 次疊代，且因為前 5000 次疊代通常較不穩定會將其捨棄，故只取後 10000 次疊代作為本研究參數估計的結果。經由圖形檢視，顯示 10000 次疊代的結果已趨近收斂，適合作為貝氏統計的參數估計值。

參數回復性的檢驗使用平均偏誤 (bias) 與均方根誤差 (root mean square error, RMSE) 作為評估指標，若其估計誤差值愈小，代表其估計效果愈好。試題參數與評分者參數之評估指標公式分別如下：

$$\text{Bias}(EAP(\xi)) = \frac{\sum_{r=1}^R (EAP(\xi_r) - \xi)}{R} \quad (9)$$

$$\text{RMSE}(EAP(\xi)) = \sqrt{\frac{\sum_{r=1}^R (EAP(\xi_r) - \xi)^2}{R}} \quad (10)$$

其中， R 為每個模擬情境重複的次數，本研究設定為 30 次； ξ 為模擬研究產生之參數真值； $EAP(\xi_r)$ 為期望後驗分布的估計值，亦為參數在第 r 次重複實

驗的估計值。受試者能力參數之評估指標公式如下：

$$RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{n=1}^N (\hat{\theta}_n - \theta_n)^2}{N}} \quad (11)$$

其中， N 為受試者人數； θ_n 為受試者 n 的能力真值； $\hat{\theta}_n$ 為受試者 n 的能力估計值。在能力參數的回復性檢驗方面，研究者只計算 RMSE，只要原因在於 bias，即容易產生正負值抵銷的情形，不適合用來檢驗能力參數的估計效果，而試題參數則可利用 bias 的計算，進而觀察是否存在系統性的正向或負向偏差。

肆、研究結果

一、模擬研究

本研究發展的 MFMixERS-GPCM 在無遺漏值情境中，試題參數、評分者參數、受試者能力參數之 bias 平均值與 RMSE 平均值與潛在類別一致度，如表 1 所示。整體來說，試題參數與評分者參數估計值的 bias 與 RMSE 都很趨近於 0，表示估計效果良好。試題參數 RMSE 在各情境下大樣本（ $N = 1000$ ）比小樣本（ $N = 500$ ）小，且試題鑑別度估計值之 RMSE 在大樣本情境下，不太受量表點數及題項數影響；對試題難度來說，在四點量表時的六題情境之估計值 RMSE 較小，相反地，在六點量表時的三題情境之估計值 RMSE 較小，本研究推測此現象應和量表點數與試題數量相互影響有關，此現象與 Jin 與 Wang（2014）有類似結果；閾值參數方面顯示，在量表點數多的情況下，題目愈多，其 RMSE 會較小，意即估計效果會比較好。閾值權重參數、評分者嚴格度的 RMSE 值在大樣本情境下比小樣本情境小，六題情境比三題情境小，六點量表情境比四點量表情境小。受試者能力參數的 RMSE 值在大樣本情境下比小樣本情境小，六題情境比三題情境小，六點量表情境比四點量表情境小。潛在類別的一致度介於 0.611~0.757（61.1~75.7%），情境愈複雜，其受試者的潛在類別估計出來與真值的一致度愈高。

表 1 MFMixERS-GPCM 在無遺漏值情境中各參數之 Bias 與 RMSE 平均值摘要

		四點量表				六點量表			
		三題		六題		三題		六題	
		$N = 500$	$N = 1000$	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
α_i	bias	-0.011	-0.010	-0.011	-0.002	-0.017	-0.012	-0.031	-0.012
	RMSE	0.051	0.035	0.050	0.036	0.044	0.034	0.056	0.034
β_i	bias	-0.016	-0.014	-0.005	-0.007	-0.015	-0.004	-0.023	-0.008
	RMSE	0.067	0.048	0.065	0.045	0.054	0.033	0.073	0.043
τ_{ij}	bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RMSE	0.079	0.060	0.069	0.047	0.104	0.080	0.069	0.043
η_r	bias	-0.004	0.002	-0.007	-0.002	-0.005	0.000	-0.015	-0.006
	RMSE	0.041	0.030	0.031	0.022	0.033	0.026	0.032	0.019
ω_{r_s}	bias	0.105	0.090	0.056	0.035	0.030	0.017	0.010	0.005
	RMSE	0.294	0.249	0.190	0.134	0.145	0.094	0.078	0.057
θ_n	RMSE	0.386	0.379	0.258	0.254	0.304	0.300	0.209	0.204
g	一致度	0.611	0.609	0.643	0.649	0.671	0.683	0.757	0.757

註： α_i 為試題鑑別度； β_i 為試題難度； τ_{ij} 為閾值參數； η_r 為評分者嚴格度； ω_{r_s} 為閾值權重參數； θ_n 為受試者能力參數； g 為潛在類別估計值與真值一致度。

FMixERS-GPCM 在有遺漏值情境中，試題參數、評分者參數、受試者能力參數、潛在類別一致度的 bias 與 RMSE 平均值，如表 2 所示。整體來說，各參數估計值之 bias 都很趨近於 0。在各情境下每個試題參數在大樣本（ $N = 1000$ ）之 RMSE 值比小樣本（ $N = 500$ ）之情境小。試題鑑別度估計值的 RMSE 在小樣本情境下，不太受量表點數及題項數影響，在大樣本情境下較有影響，但差異不大；對試題難度來說，在四點量表時的六題情境之估計值 RMSE 較小，但在六點量表時的三題情境之估計值 RMSE 較小，此部分與無遺漏值情境中相同；閾值參數方面顯示，在量表點數多的情況下，題目愈多，其 RMSE 會較小，意即估計效果會比較好。閾值權重參數、評分者嚴格度的 RMSE 在大樣本情境下比小樣本情境小，六題項情境比三題項情境小，六點量表情境比四點量表情境小。在受試者能力參數估計值的 RMSE 方面，大樣本情境下比小樣本情境小，六題項情境比三題項情境小，六點量表情境比四點量表情境小。潛在類別的一致度介於 0.604~0.670（60.4~67%），情境愈複雜，其受試者的潛在類別估計出來之一致度愈高。就無遺漏值與有遺漏值的兩個情境來說，所有參數與潛在類別一致度在無遺漏值情境中的估計效果比較好。

表 2 MFMixERS-GPCM 在遺漏值情境中各參數之 Bias 與 RMSE 平均值摘要

		四點量表				六點量表			
		三題		六題		三題		六題	
		$N = 500$	$N = 1000$	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
α_i	bias	-0.018	-0.020	-0.014	-0.008	-0.030	-0.010	-0.035	-0.014
	RMSE	0.075	0.064	0.075	0.056	0.086	0.056	0.075	0.051
β_i	bias	-0.019	-0.010	-0.003	-0.017	-0.019	-0.002	-0.021	-0.013
	RMSE	0.085	0.064	0.079	0.058	0.078	0.044	0.081	0.052
τ_{ij}	bias	0.000	0.000	0.000	0.000	-0.011	0.000	0.000	0.000
	RMSE	0.118	0.099	0.112	0.077	0.183	0.120	0.108	0.072
η_r	bias	-0.003	-0.001	-0.009	-0.007	-0.009	0.003	-0.012	-0.006
	RMSE	0.078	0.046	0.052	0.036	0.074	0.040	0.043	0.034
ω_{r_s}	bias	0.172	0.064	0.084	0.060	0.063	0.067	0.031	0.013
	RMSE	0.454	0.338	0.256	0.195	0.269	0.195	0.140	0.100
θ_n	RMSE	0.550	0.548	0.388	0.385	0.475	0.449	0.321	0.311
g	一致度	0.607	0.604	0.607	0.617	0.622	0.623	0.658	0.670

註： α_i 為試題鑑別度； β_i 為試題難度； τ_{ij} 為閾值參數； η_r 為評分者嚴格度； ω_{r_s} 為閾值權重參數； θ_n 為受試者能力參數； g 為潛在類別估計值與真值一致度。

接下來，研究者以 Many-facet GPCM 來分析 MFMixERS-GPCM 產生的資料，透過此方法來探討當資料反應存在潛在類別與極端反應而被忽略的後果。在 many-facet GPCM 且無遺漏值情境中的試題參數、評分者參數、受試者能力參數、潛在類別一致度之 bias 與 RMSE 平均值，如表 3 所示。FMixERS-GPCM 與 Many-facet GPCM 兩模式的比較，發現大多參數在 MFMixERS-GPCM 之參數回復性較佳，其中試題鑑別度的參數回復性在 Many-facet GPCM 下較容易有不穩定之情況，因其在四點、六題、大樣本情境下的參數回復性為最好，但通常會預期參數回復性在較複雜、資訊較多的情境下會比較好。在評分者嚴格度方面，FMixERS-GPCM 的評分者嚴格度參數回復性隨著情境複雜度而趨佳，但 Many-facet GPCM 的評分者嚴格度參數回復性在六點量表情境下不如四點量表，故推論在未考量評分者閾值權重參數的 Many-facet GPCM 中，其評分者嚴格度較容易受各情境的不同而影響其回復性。在受試者能力方面，兩模式之參數回復性與其趨勢非常相近，但 MFMixERS-GPCM 估計的參數回復性還是比 Many-facet GPCM 略好一點。綜上所述，本研究推論模式之間的差異在試題參數（含試題難度、試題鑑別度、試題閾值參數）與評分者參數上較容易顯現出來，且相較於

MFMixERS-GPCM，模式較為簡易的 Many-facet GPCM 在試題參數與評分者參數估計上會產生偏誤的估計情況；然而，兩模式在受試者能力參數上的差異較小，可能來自於研究者操弄情境之影響，亦即模擬研究中設計各兩位評分者為極端反應風格或趨中反應風格，且加權參數影響設定相同（影響方向不同），因此可能導致受試者能力參數在最終的估計過程中，平衡了極端或趨中反應的影響，導致兩個模式在能力估計上差異較小的原因。

表 3 Many-facet GPCM 在無遺漏值情境中各參數之 Bias 與 RMSE 平均值摘要

		四點量表				六點量表			
		三題		六題		三題		六題	
		$N = 500$	$N = 1000$	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
α_i	bias	-0.016	-0.014	-0.012	-0.004	-0.036	-0.028	-0.041	-0.024
	RMSE	0.053	0.039	0.051	0.037	0.060	0.051	0.063	0.045
β_i	bias	-0.025	-0.027	-0.015	-0.019	-0.043	-0.037	-0.054	-0.040
	RMSE	0.068	0.052	0.068	0.050	0.067	0.049	0.089	0.059
τ_{ij}	bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RMSE	0.086	0.071	0.086	0.067	0.125	0.102	0.096	0.066
η_r	bias	0.009	0.018	-0.009	-0.006	0.022	0.029	-0.023	-0.013
	RMSE	0.040	0.036	0.033	0.027	0.040	0.040	0.045	0.033
θ_n	RMSE	0.386	0.379	0.259	0.255	0.305	0.301	0.212	0.206

註： α_i 為試題鑑別度； β_i 為試題難度； τ_{ij} 為閾值參數； η_r 為評分者嚴格度； θ_n 為受試者能力參數。

二、實徵研究

實徵研究主要使用泰國一所國際學校在 2012~2018 年參加英語課程學生的英文寫作測驗（Conrad II, 2020）。實證資料分析之目的主要在於探討評分者中介評量中，評分者是否存在於不同的潛在類別受試群中，表現不同的 ERS 傾向，因此除了以新的模式來分析資料外，研究者另外也以傳統的評分者模式進行資料分析，藉以比較兩種不同取向模式觀點的差異。基於此，共使用八種分析模式來進行資料分析。有關 FMixERS-RSM、FMixERS-GRSM、FMixERS-PCM、FMixERS-GPCM、Many-facet-RSM、Many-facet-GRSM、Many-facet-PCM、Many-facet-GPCM 共八個模式的適配情形，如表 4 所示。BIC 值最小的模式為 FMixERS-GPCM（BIC = 11,640），代表 FMixERS-GPCM 在八種模式中的適配度檢驗下有最好之模式效果。在

MFMixERS-GPCM 中，各參數估計值的分布情形如表 5 所示，其中潛在類別機率在無反應風格組與有反應風格組的比率約為 0.36、0.64，也就是被歸類於無反應風格的受試者約占全部受試者 36%，而有反應風格組的受試者約占 64%。而各參數估計值皆在合理範圍，且當受試者的潛在類別為有反應風格組時，評分者會有偏向趨中的反應風格。

表 4 實徵資料在各模式下之 BIC 適配度指標

模式	BIC
多面向混合極端反應模式	
MFMixERS-RSM	12,550
MFMixERS-GRSM	12,170
MFMixERS-PCM	12,260
MFMixERS-GPCM	11,640
多面向模式	
Many-facet-RSM	13,420
Many-facet-GRSM	12,780
Many-facet-PCM	13,390
Many-facet-GPCM	12,770

註：粗體數值表示最小 BIC 值。

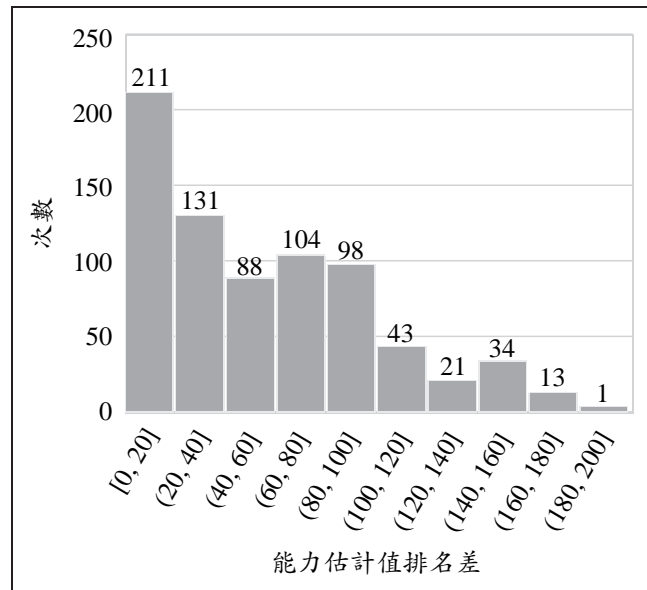
表 5 實徵資料在 FMixERS-GPCM 之各參數分布情形

	α_i	β_i	τ_{ij}	η_r	ω_{r_s}	θ_n
平均值	3.582	-1.403	0.000	0.000	1.911	-0.001
標準差	1.390	0.803	2.899	0.024	0.030	0.949
最大值	5.430	-0.538	4.645	0.024	1.940	2.715
最小值	1.280	-2.725	-10.050	-0.024	1.881	-3.105

註： α_i 為試題鑑別度； β_i 為試題難度； τ_{ij} 為閾值參數； η_r 為評分者嚴格度； ω_{r_s} 為閾值權重參數； θ_n 為受試者能力參數。

除了以 BIC 值的方式檢驗模式適配度，本研究也使用 FMixERS-GPCM 與 Many-facet-GPCM 模式之間的能力估計值排序差，來檢驗 FMixERS-GPCM 確實有較好的估計效果（如圖 1 所示）。經能力排序差檢驗發現，兩模式之間的能力估計值排序確實有明顯差異，且排序差異最大可以到 181，意指若使用錯誤的模式進行估計，受試者的能力有可能會被錯估很多，在實

圖 1 實徵資料在 MFMixERS-GPCM 與 Many-facet-GPCM 之受試者能力估計值排序差



務上可能會嚴重影響到受試者的能力評估及名次，故可以顯示出本研究發展的多面向混合極端反應風格試題反應理論模式在資料分析上之必要性。

對於混合試題反應理論的應用上，已有研究指出若實徵資料內包含受試者的背景變項，可以透過分析背景變項與潛在類別之間的關聯來了解潛在類別發生之原因（Cho & Cohen, 2010）。因此，本研究使用實徵資料內包含的背景變項：性別、年級，與MFMixERS-GPCM分析出的潛在類別進行統計考驗，來了解背景變項與潛在類別之間的關聯，如表 6 所示。受試者性別在潛在類別之間的差異未達顯著性（ $\chi^2 = 1.092$ ， $p = 0.296$ ），表示受試者影響評分者有不同反應風格的原因應與受試者性別無關。但在受試者年級方面的潛在類別達顯著差異（ $\chi^2 = 4.678$ ， $p < .05$ ），表示受試者影響評分者有不同反應風格的原因可能與受試者的年級有關，大致上各年級之學生被歸類於有反應風格者較多，其中 5 年級之受試者在兩種潛在類別的差異最大，本研究根據此結果推論評分者有不同反應風格的原因可能與不同年級學生在此測驗的寫作策略或其他相關背景有關，而因此產生在潛在類別上的差異。

表 6 受試者性別與年級在潛在類別之差異情形

		潛在類別		總計	χ^2
		無反應風格組	有反應風格組		
性別	男	127 (34.6%)	240 (65.4%)	367	1.093
	女	144 (38.3%)	232 (61.7%)	376	
	總計	271	472	743	
年級	3	14 (36.8%)	24 (63.2%)	38	24.678*
	4	15 (22.1%)	53 (77.8%)	68	
	5	9 (16.1%)	47 (83.9%)	56	
	6	28 (31.8%)	60 (68.2%)	88	
	7	29 (39.2%)	45 (60.8%)	74	
	8	30 (41.1%)	43 (58.9%)	73	
	9	38 (42.2%)	52 (57.8%)	90	
	10	32 (36.4%)	56 (63.6%)	88	
	11	39 (43.3%)	51 (56.7%)	90	
	12	37 (46.8%)	42 (53.2%)	79	
	總計	271	472	744	

* $p < .05$; ** $p < .01$; *** $p < .001$

伍、結論與建議

在社會科學研究領域中，使用外部評分者作為觀察分數或表現評量的依據，是非常普遍的研究設計。然而，評分者的主觀意識或評分傾向，經常是造成評分者偏誤的變異來源。過去在 IRT 的架構下，評分者嚴格度的測量與極端反應風格的控制皆已發展出適合的心理計量模型。本研究進一步擴展先前的 IRT 模式，納入混合模式的概念，檢驗受試者群的背景或答題策略可能造成的潛在類別，而影響評分者在評量項目選項上的使用傾向，造成極端或趨中反應風格。

在本研究發展的 MFMixERS-GPCM 中，模擬研究顯示參數回復性估計值皆良好，各參數整體來說在無遺漏值情境比有遺漏值情境的估計效果較好，在大樣本的估計效果比小樣本情境較好。試題參數在題項數、量表點數、樣本數容易互相影響而有不同變化，推論當量表點數多的時候，題項數與樣本數也要夠多時，才有足夠資訊可以進行有效估計。其中，試題難度在

各情境的趨勢和 Jin 與 Wang (2014) 之研究有相似結果，而其他參數在題項多及量表點數多的情境下，其估計效果比較好。在潛在類別一致度方面也會隨著情境的複雜度而有系統性的變化，提供資訊愈多的情境，其一致度會愈高。在 Many-facet GPCM 與 MFMixERS-GPCM 的模式比較上，發現試題參數與評分者參數在 MFMixERS-GPCM 的估計效果都比 Many-facet GPCM 良好，且在 Many-facet GPCM 中，試題參數更容易受到量表點數、題項數、樣本數之間的影響，這和 Jin 與 Wang (2018) 的研究討論閾值權重參數時有類似之結果。而受試者能力參數雖然 MFMixERS-GPCM 略好一點，但與 Many-facet GPCM 的差異並不大，本研究推論因在進行模擬資料產生時設定正常組與異常組的比率為 0.6、0.4，表示被歸類為正常組的受試者比異常組的受試者多，且正常組之權重參數為 1，換句話說當受試者被歸類為正常組時，使用 MFMixERS-GPCM 估計與使用 Many-facet GPCM 估計本來的差異就不大。除此之外，在產生模擬資料時，本研究對於評分者嚴格度與閾值權重參數的設定沒有進行系統性配對，可能因此低估了評分者嚴格度與反應風格的交互影響，因而影響到受試者能力的估計。

實徵研究的適配度檢驗結果顯示，在 MFMixERS-GPCM 的 BIC 值最小，亦指資料在此模式下有較好的適配度，且評分者確實會因為受試者所屬的潛在類別而有不同之反應風格，且對於被歸類為異常組的學生來說，評分者在評分時有趨中反應風格的傾向。除此之外，FMixERS-GPCM 與 Many-facet GPCM 的受試者能力估計值排序差異比較結果發現，使用多面向概化部分計分模式估計受試者能力確實有比較大的差異，故 MFMixERS-GPCM 不僅可以識別出受試者的潛在類別，更發現在評分者測驗中，評分者的確會因為受試者所屬潛在類別而有不同之反應風格，此模式在實務上也有良好的適配情形。最後，針對受試者背景變項與潛在類別的關聯分析上發現，不同反應風格的潛在類別可能與受試者性別無關但與年級有關，不同年級的受試者在寫作策略上之不同，可能影響其被歸類至不同潛在類別中。

在混合模型的研究中，受試者人數影響甚大，因此本研究發展的模式並不適合小樣本的情境，此乃本研究侷限之一。此外，本研究因研究設備與時間考量，未能將 MFMixERS-RSM、FMixERS-GRSM、FMixERS-PCM 納入此次的模擬研究並檢驗各模式之參數回復性，故建議未來若能克服設備及時間的問題後，將四種標準模式納入模擬研究，就更能了解四種模式之間參數的估計效果與其差異，讓研究者與讀者對於發展的模式有更全面的了

解。且因研究時間的考量，未能將評分者人數納入情境控制變項，建議能將評分者人數加入情境變項的討論，以了解面對不同的評分者人數時，其模式的檢驗效果有何不同，讓評分者測驗的研究有更全面的了解。除了單向度的討論，未來也可以將研究延伸至多向度試題反應模式，了解模式在不同向度之間估計效果的差異。

謝誌

本研究係為第一作者碩士論文改寫，感謝論文口試委員：國立臺灣師範大學教育心理與輔導學系陳柏熹教授、國家教育研究院測驗及評量研究中心謝名娟研究員提出的寶貴建議；另感謝匿名審查委員的耐心審閱與修改意見。該論文亦獲得2021年中國測驗學會碩士論文獎之獎勵，特別感謝中國測驗學會對於新進學者之鼓勵。

本研究部分接受行政院國家科學委員會研究計畫補助，計畫編號：109-2410-H-845-015-MY3，謹此致謝。

參考文獻

中文部分

余民寧（2009）。試題反應理論（IRT）及其應用。心理。

英文部分

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. <https://doi.org/10.1007/BF02293814>
- Batchelor, J. H., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology*, 16(2), 51-62. <https://articlegateway.com/index.php/JOP/article/view/1790>
- Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143-156. <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Bertrand, A. M., & Hafner, C. M. (2014). On heterogeneous latent class models with applications to the analysis of rating scores. *Computational Statistics*, 29, 307-330. <https://doi.org/10.1007/s00180-013-0450-5>
- Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, 33(5), 335-352. <https://doi.org/10.1177/0146621608329891>
- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, 71(5), 814-833. <https://doi.org/10.1177/0013164410388411>
- Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, 35(3), 336-370. <https://doi.org/10.3102/1076998609353111>
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2), 133-148. <https://doi.org/10.1111/j.1745-3984.2005.00007>
- Conrad II, C. J. H. (2020). Direct writing prediction models identify at-risk writers.

- THAITESOL Journal*, 33(1), 57-71.
- Du, Y., Wright, B. D., & Brown, W. L. (1996). *Differential facet functioning detection in direct writing assessment*. American educational Research Association Press.
- Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages; Learning, teaching, assessment (Section H)* (pp. 1-52). Council of Europe/Language Policy Division.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, 56(3), 328-351. <https://doi.org/10.1086/269326>
- Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, 7, 1706. <https://doi.org/10.3389/fpsyg.2016.01706>
- Ilgun Dibek, M. (2020). Effect of extreme and acquiescence response style in TIMSS 2015. *Eurasian Journal of Educational Research*, 87, 199-219. <https://doi.org/10.14689/ejer.2020.87.10>
- Jin, K. Y., & Eckes, T. (in press). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement*. <https://doi.org/10.1177/00131644211043207>
- Jin, K. Y., & Wang, W. C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, 74(1), 116-138. <https://doi.org/10.1177/0013164413498876>
- 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. <https://doi.org/10.1080/00273171.2017.1299615>
- Jin, K. Y., & Wang, W. C. (2018). A new facets model for rater's centrality/extremity response style. *Journal of Educational Measurement*, 55(4), 543-563. <https://doi.org/10.1111/jedm.12191>
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. MESA Press.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. <https://doi.org/10.1007/BF02296272>.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, 16, 159-176. <https://doi.org/10.1002/j.2333-8504.1992.tb01436.x>

- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.
- Paulhus, D. L. (1991). Measurement and control of response bias. In *Measures of personality and social psychological attitudes* (pp. 17-59). Academic Press. <https://doi.org/10.1016/B978-0-12-590241-0.50006-X>
- Plummer, M. (2015). *JAGS Version 4.0.0 user manual*. https://www.uvm.edu/~bbeckage/Teaching/PBIO_294/Manuals/manual.jags.pdf
- Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, 35(2), 385-397.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464. Institute of Mathematical Statistics Press.
- Wang, W. C., & Liu, C. Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, 67(4), 583-605. <https://doi.org/10.1177/0013164406296974>
- Wang, W. C., & Wu, S. L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement*, 48(4), 441-456. <https://doi.org/10.1111/j.1745-3984.2011.00154.x>
- Wang, W. C., Wilson, M., & Shih, C. L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement*, 43(4), 335-353. <https://doi.org/10.1111/j.1745-3984.2006.00020.x>
- Zhang, Y., & Wang, Y. (2020). Validity of three IRT models for measuring and controlling extreme and midpoint response styles. *Frontiers in Psychology*, 11, 271. <https://doi.org/10.3389/fpsyg.2020.00271>