・文官制度・　第十五卷第二期
民國112年11月　頁1-20

特邀專論

評分者效應對口試評分公平性影響問題之探討

余民寧 [a]

《摘要》

口試在甄選人員的應用上已行之多年，歷年相關的研究報告亦多支持採行「結構化口試」的可行性與效用性。但在推行「結構化口試」的措施時，諸多可能干擾口試評分公平性之因素中，多數均可透過標準作業流程來加以排除或改善，唯獨其中一項「評分者效應」，很難從標準作業流程程序中加以排除或控制。因此，本文對此提出一個採行「多面向模型」（many-facet model, MFM）計分的應用建議，以客觀分析口試委員評分資料的各面向因素之估計值—含：考生的能力值、口試問題的難度值、及評審評分嚴苛/寬鬆程度值。根據過去相關文獻記載，應用 MFM 模型於此類涉及評分者效應的資料分析上，不僅可增進對考生能力值的精確估計，也可以提高整體資料分析的信度值，並得以促進口試評分更具公正性、公平性與正確性。為落實 MFM 模型能應用於爾後口試評分資料的分析，本文亦建議現成的口試評分表的評分方式必須做改變，從連續型資料屬性的等距性尺規評分（如：在 50-59 分的成績區間內評出一個分數）改變成離散型資料屬性的次序性尺規評分（如：僅將成績評定成優、佳、普通、差等四個等

[a] 國立政治大學教育學院特聘教授兼院長，e-mail: mnyu@nccu.edu.tw。

第），以期降低「評分者效應」對口試評分公平性之干擾影響。

[關鍵詞]：評分者效應、結構化口試、多面向模型、評分公平性、評分嚴苛/寬鬆程度

壹、前言

口試（oral exam）或口頭評量（oral assessment）在學校的教學評量、升學的甄試入學、就業甄試的面談，乃至國家考試的公務人員考試裡，均被廣泛使用已是不爭的事實，且有愈來愈受到重視的趨勢（余民寧，2022a；胡悅倫，2008；胡悅倫、余民寧，2009；胡悅倫等人，2008；胡悅倫等人，2009；彭錦鵬，2010；Bunting, 2007; Popham, 2008）。

針對口試或口頭評量該如何進行的問題，經過學者們多年的努力，已逐漸建構出一套口試評量方法學的架構，那就是企圖在口試的內容、實施程序與評分策略上，建立起一套共同遵守的「標準作業流程」（余民寧，2022a；余民寧等人，2011；胡悅倫等人，2010；胡悅倫等人，2009；Dixon et al., 2002）。此流程大致包括四大步驟：（一）職能分析、（二）口試問題的擬定、（三）定錨評量、及（四）口試評量訓練；通常，只要嚴格遵守此四大步驟流程所實施的口試評量，即能做到「結構化口試（structured oral exam）」的精髓（呂秋萍，2005；林文銘，2005；胡悅倫，2008；陳淑慧，2006；Arthur, 2005; Campion et al., 1997; Dipboye, 1994, 2005; Taylor & O'Driscoll, 1995），且能大幅提高口試評量的效度至 .35 ~ .62 之間（Campion et al., 1997; Huffcutt & Arthur, 1994; Marchese & Muchinsky, 1993; McDaniel et al.,1994）。這就是為什麼口試評量必須嚴格遵守「標準作業流程」的原因所在，其重要性與價值性，自是不言可喻。

就此標準作業流程來進行檢討與反思國內的國家考試作法，學者們對此提出不少的建言與反省聲音（王成基，2004；吳復新，2000，2007；余民寧，2013，2022b；彭錦鵬，2010）。相對的，考選部亦為此舉辦多次的口試方法技術研討會，進行相關考選方法與技術的精進與改良（考選部，2004，2005，2006），並制

訂個別口試與集體口試的參考手冊（考選部，2003；余民寧等人，2011；胡悅倫等人，2010）；至今，相關的考選法規已日趨完備（考選部，2012）。其中，如以「口試規則」為例來看，從 2000 年制定發布至今，歷經 2002 年、2012 年、2015 年、2016 年的 4 次修訂，及 2019 年的再次修訂，如今已成為國家考試非常完備的口試實施辦法的依據。

然而，在此考選部已實施結構化口試評量，改善並提昇口試評量的信度與效度之際，仍有許多可以繼續精益求精之處（余民寧，2013）。本文的目的，即是針對其中一項檢討建言：「盡量減少評分者效應的影響」做深入的探討，企盼透過學術研究的心得成果，繼續提供可以改善評分公平性，提高評分精確性的建議做法。

貳、正視評分者效應存在的事實

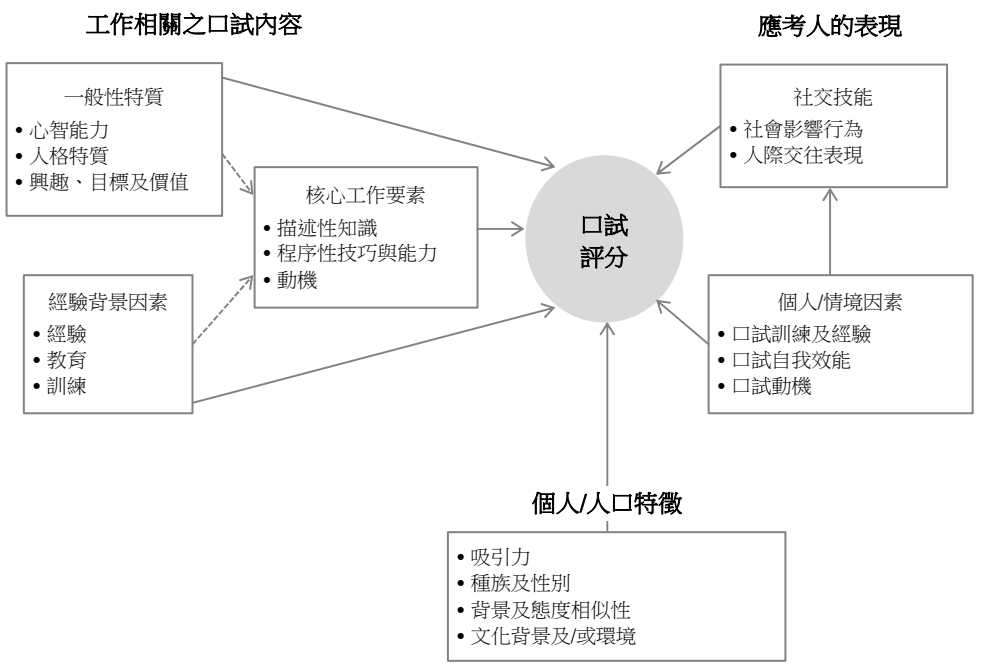
無可諱言的，即使在口試評量確實已嚴格遵守「標準作業流程」之下，要盡力做到評分公平、公正、客觀的程度，仍屬不易的事，其中的干擾因素之一，即是我們無法全面遏止「評分者效應」對口試評分的不確定影響。但是，我們可以逐漸透過學術研究的瞭解，找出到底有哪些可能產生影響的因素，然後在未來的口試評量訓練中，將這些因素納入訓練的課程範圍裡，並且在未來進行口試的評分歷程中，設法將這些干擾因素平衡掉、排除掉或控制住，以將評分者效應對評分的不確定影響，降低到最小的程度。

一般而言，口試情境中的各種干擾因素，從人員（口試委員及應考人）、評分方式（分析性評分或整體性評分）、口試情境（燈光、噪音、空調、場務布置）、突發狀況（地震、颱風、火警、各種演習）、到各種干擾因素之間的交互作用等，都可能影響口試評分的公正客觀性。例如，Huffcutt（2011, p. 63）及 Huffcutt 等人（2001）曾回顧過去的文獻，綜合提出一個影響口試評分的相關因素概念圖，如圖 1 所示。他指出在影響口試委員評分的相關因素裡，主要可以分為三個層面：與工作相關的口試內容（如：該工作專業知識）、應考人的表現（如：應考人印象管理策略）、應考人個人在人口統計學上的特徵（如：應考人的外表吸引力）。這些因素綜合起來，會讓口試委員產生主觀意識的偏好，讓口試委員在評分時，傾向於針對與自己內在或外在特質相似的應考人產生好感（即類我效應〔similar-to-me

effect〕），因而做出給予與自己特質相似（如：同性別、同畢業學校、同價值觀、同理念、同嗜好、同意識形態等）的應考人較高之得分；而給予明顯與自己特質不相似（如：不同性別、不同畢業學校、不同價值觀、不同理念、不同嗜好、不同意識形態等）的應考人較低之得分。因此，這種評分者效應確實存在於口試評分當中，它是一項不爭的事實，也是一項值得重視與研究的評分議題（Farrokhi & Esfandiari, 2011; O'Brien & Rothstein, 2011; Sears & Rowe, 2003; Thorburna & Collins, 2006; Touchie et al., 2010; Tsai et al., 2012）。

圖 1

影響口試評分的相關因素模式圖



註：影響口試評分的因素，包括與工作相關之口試內容（含應考人的一般性特質及經驗背景因素，同時也會透過核心工作要素發揮影響力）、應考人的表現（含應考人的社交技能與個人/情境因素）、與應考人個人/人口特徵等。

資料來源：修改自"An Empirical Review of the Employment Interview Construct Literature," by A. I. Huffcutt, 2011, *International Journal of Selection and Assessment*, 19(1), p.63.

其實，在測驗與評量學術領域中或教科書裡，即指出舉凡涉及到需要以「人」來當作「評審」（raters or markers），才能給予公正裁判並評分的評量情境，諸如學校的許多實作評量（performance assessment）（例如：口試、寫作、面試甄選、創造力測驗等）、藝文活動表演比賽（例如：歌唱比賽、美勞作品競賽、書法比賽、音樂表演賽等）、運動競賽等（例如：跳水、體操、各種球類比賽等），評分者效應的影響力是不容忽視的既存現象（余民寧，2020；Thorburna & Collins, 2006; Wind, 2019）。

這種仰賴評審的主觀判斷所取得的評分資料，當只有唯一的評審時，評定分數並無所謂評分寬嚴不一的存在問題，但當評審人數兩位（或超過兩位）以上時，評審之間一定會產生評分嚴苛度（rating harshness/lenience）差異的問題：有些評審是好好先生，他的評分一致性地寬鬆，他針對任何一位應考人的口試表現都一律給於評定高分（例如：所打的分數至少是80分起跳）；而有些評審是自律甚嚴的人，他的評分則一致性地嚴苛，他針對任何一位應考人的口試表現都一律給於評定低分（例如：所打的分數至多是80分為止）；但還有一些評審的評分較不穩定，他可能對某些應考人的口試表現（即類我效應較高者）給於評定高分（例如：得分最低者打高於80分以上），而對某些應考人的口試表現（即類我效應較低者）給於評定低分（例如：得分最高者只打到80分）。當考選（試）主辦單位無法事先知道每一位應考人的實際表現，也無法事先知道每一位評審的評分是否都是公平一致的情況下，任由隨機配對來組合應考人與評審的情境下，評審的評分嚴苛度問題一定會對參賽者（即應考人）的評分成績產生嚴重的判斷偏誤（judgement error）現象，因而令人質疑比賽（或口試）評分的公平性與公正性。所以，這個「評審的評分判斷公平與否」的潛在問題，一直是測驗與評量學界所關注研究的焦點之一。因此，這種來自評審評分嚴苛度不一，對評分成績造成不同影響效果的現象，便統稱為「評分者效應」（rater effects）（余民寧，2020；Wind, 2019; Wolfe, 2004）。

Myford與Wolfe（2003）、Wolfe（2004）在評閱與歸納文獻後，將評分者效應分成 12 種類型：1. 嚴苛/寬鬆效果（severity/leniency effect）、2. 月暈效果（halo effect）、3. 趨中效果（central tendency effect）、4. 侷限效果（restriction-of-range effect）、5. 錯誤評分效果（inaccuracy effect）、6. 邏輯偏誤效果（logical error effect）、7. 對比效果（contrast effect）、8. 評分者偏誤、信念、態度及人格特質的

影響效果（influences of rater biases, beliefs, attitudes, and personality characteristics effect）、9. 評分者與被評者之背景特質的影響效果（influences of rater/rate background characteristics effect）、10. 鄰近誤差效果（proximity error effect）、11. 新近/初始誤差效果（recency/primacy error effect）、及 12. 次序效果（order effect）；其中，以嚴苛/寬鬆效果、月暈效果、趨中效果、侷限效果等，是最常被提出來進行探究和測量的評分者效應（Erguvan & Dunya, 2020; Myford & Wolfe, 2003; Wolfe, 2004; Yeşilçınar & Şata, 2021）。因此，底下所述即以嚴苛/寬鬆效果為例，深入探討當我們碰到這種會影響評分公平性、公正性、客觀性的棘手問題時，又該如何處理較好。

參、評分者效應該如何處理—MFM 模型的應用

評分嚴苛度因素所反應的事實，即是「某種外在因素影響到某試題難度參數估計，或是與該試題難度產生交互作用」的一種現象，因而會干擾（可能會提升或降低）試題難度參數的估計，如果研究者不去面對處理此問題，這種測驗（或評量）結果的評分將會是不精準的（inaccurate）。也就是說，這個「面向」因素（即評分嚴苛度）不僅會干擾某些評量題目的難度估計值（即可能造成高估或低估），進而影響應考人在該評量題目上表現的評分結果。這對講求評分公平公正的競賽或評比活動而言，將會是無法被接受的「判斷偏差」（judgment bias）事件。

對此問題，Linacre（1989, 1991）提出一個測量模型—稱作「多面向模型」（many-facet model, MFM），即建立在原本的 Rasch 測量模型中，多新增一個「面向參數」（facet parameter），以用來作為衡量「評分者效應」的估計之用。這個用在二元計分（dichotomous or binary scoring）資料下的原始 MFM 測量模型公式，可以表示如下：

$$P(X_i = 1|\theta_n) = \frac{\exp(\theta_n - (b_i + \rho_m))}{1 + \exp(\theta_n - (b_i + \rho_m))} \qquad \text{（公式 1）}$$

其中，θ_n為應考人 n 的能力值參數（ability parameter），b_i為試題 i 的難度參數（difficulty parameter），而ρ_m則為評審 m 的評分嚴苛/寬鬆度參數（rating

harshness/leniency parameter）。（公式 1）所示的整體涵義即為：具有能力值為θ的應考人 n，他在一道具有難度值為 b 的試題 i 上，且在第 m 名評審評分嚴苛度為ρ的評分情境下，答對該試題 i 的機率值。

在通常的評分情境下，多半是假設每位評審對所有應考人在所有試題上的評分嚴苛度都是保持一致的。換句話說，如果某位評審屬於評分較嚴苛的人時，則他會一致性地對所有應考人在所有試題上的表現均一律地評分嚴苛，因此，他所評定的分數會提高試題的難度參數，而讓應考人在試題上的表現較傾向獲得低分或答對機率較低；而另一位評審若屬於是評分較寬鬆的人時，則他會一致性地對所有應考人在所有試題上的表現均一律地評分寬鬆，因此，他所評定的分數會降低試題的難度參數，而讓應考人在試題上的表現較容易獲得高分或答對機率較高。所以，（公式 1）是最適合在考量評分者效應影響下，用來計算口試得分的一個最基礎、最陽春的評分公式而已。因此，每當我們將（公式 1）實際應用到口試評分情境裡，我們便是在考量評分者效應（即ρ_m）及試題難度（即b_i）的條件下，估計出每位應考人應該獲得的口試能力值（即θ_n）。所以，每當口試評分結束後，我們一共可以獲得 N 位應考人的口試能力估計值、M 位評審的評分嚴苛度估計值，以及 I 道試題的難度估計值。

另一種情況，則是當我們碰到較複雜、較麻煩的口試情境，此時，評審的評分較不穩定，評審的評分會根據不同的試題難度給予不同的評分（如：遇到較容易的試題，給予較高的評分，或遇到較困難的試題，給予較低的評分）。在這種情況下，即稱作評審的評分嚴苛度與試題的難度產生交互作用，此時，我們就必須把如此的交互作用效果（interaction effect），一併考量在（公式 1）裡，而將（公式 1）擴增成為（公式 2），如下所示：

$$P(X_i = 1|\theta_n) = \frac{\exp(\theta_n-(b_i+\rho_m+\gamma_{im}))}{1+\exp(\theta_n-(b_i+\rho_m+\gamma_{im}))} \qquad \text{（公式 2）}$$

其中，（公式 2）裡的 γ_{im} 即為評審 m 的評分嚴苛度與試題 i 的難度所產生的交互作用效果參數，它的作用即是用來調整試題 i 的難度與評審 m 的平均評分嚴苛度的一個校正項；其餘符號的涵義則與（公式 1）相同。通常，當γ_{im}出現在評分結果時，即表示評分結果出現嚴重的偏差情形，不僅會干擾到評審評分嚴苛度值的估

計，也會干擾到試題難度值的估計，甚至間接影響到應考人能力值的估計。這種情形，即表示該試題出現了「差異試題功能」（differential item functioning, DIF）。當這種差異試題功能現象出現時，即表示該場考試評量對應考人能力值的估計是不準確、不公平的，若不加以人為妥適處理，該場考試結果是不可信的（Holland & Wainer, 1993; Osterlind & Everson, 2009）。在口試評分的實際情境（謝名娟，2017）下，通常評審所打的分數不是僅分成「對/錯」或「通過/不通過」這種二元計分的結果，而是評定為等第制或次序性的評分（ordinal scoring）結果，此時，評審所評定的分數通常為「優異」、「良好」、「普通」、「尚可」、「拙劣」等 k 個等第類別或次序性分數（如：1、2、3、4、5 分或 5、4、3、2、1 分不等），也就是測驗評量學界所指稱的多元計分（polytomous scoring）資料。因此，在多元計分資料的評分情境下，（公式 1）及（公式 2）通常會改採適用於多元計分資料下的另類 Rasch 測量模型來取代—無論是採用「評定量尺模型」（rating scale model, RSM）（Andrich, 1978）或「部份計分模型」（partial credit model, PCM）（Masters, 1982）均可。若此，我們則可採用對數勝算比值（log-odds ratio）的表徵概念形式，來重新表達（公式 1）的涵義內容為（公式 3）或（公式 4），如下所示：

$$log\left(\frac{P_{nimj}}{P_{nim(j-1)}}\right) = \theta_n - b_i - \rho_m - \delta_j \qquad \text{（公式 3）}$$

$$log\left(\frac{P_{nimj}}{P_{nim(j-1)}}\right) = \theta_n - b_i - \rho_m - \lambda_{ij} \qquad \text{（公式 4）}$$

（公式 3）是指口試評分在使用 RSM 公式的估計下，其涵義即為應考人 n 且在評審 m 的評分嚴苛度下，在試題 i 上獲得（被評定為）j 分（或類別）而非 $j-1$ 分（或類別）的對數勝算比值，可由等號右邊的各個參數來表示；其中，δ_j 即為被評定為第 j 類別（$j=0,1,2,\ldots,k-1$）得分的閾參數值（threshold parameter），且所有試題均共享同一組閾參數值（共有 $k-1$ 個閾參數值）。若以白話來解釋，δ_j 即是指從第 $j-1$ 評定類別跨到第 j 評定類別之間的門檻困難度值。（公式 3）中其餘符號的涵義，則與（公式 1）相同。

而（公式 4）則是指口試評分在使用 PCM 公式的估計下，其涵義即為應考人 n 且在評審 m 的評分嚴苛度下，在試題 i 上獲得（被評定為）j 分（或類別）而非 $j-1$

分（或類別）的對數勝算比值，可由等號右邊的各個參數來表示；其中，λ_{ij} 即為試題 *i* 中被評定為第 *j* 類別（*j=0,1,2,...,k-1*）得分的閾參數值，且各個試題的閾參數值各不相同（共有 *I**（*k-1*）個閾參數值）。若以白話來解釋，λ_{ij} 即是指試題 *i* 中從第 *j-1* 評定類別跨到第 *j* 評定類別的門檻困難度值。（公式 4）中其餘符號的涵義，則與（公式 1）相同。

無論是採用公式 3 或公式 4，都是適合應用在當前採用多元計分方式下各種實作評量結果的評分公式，當然也包括口試評量在內。若採用 RSM 或 PCM 為主的 MFM 模型來作為未來口試評分資料的分析公式時，一般而言，國外已有許多現成的電腦程式（如：CONQUEST、RUMM、WINSTEPS、FACETS 等）可供參考使用，並且國內心理計量學或測驗領域的專家學者人數眾多，亦可加入協助分析的行列，以及負責資料分析結果的事後解讀工作。

綜合上述，無論在二元計分資料情況下或在多元計分資料情況下，影響應考人 *n* 在試題 *i* 上被評審 *m* 評定為 *j* 分而不是 *j-1* 分的機率值（或得分），是受到應考人的能力值、試題的難度值、評審的評分嚴苛度值及閾（門檻）難度值等參數，所共同管控、決定的。過去許多國內外的相關文獻（謝名娟，2020；Gordon et al., 2021; Kermad & Bogorevich, 2022; Kogan et al., 2023; Wang & Long, 2022）均指出，使用 MFM 模型來分析口試評分情境下的資料，不僅可以增進對應考人能力值的估計精準度，也可以提高整體資料分析的信度值，更可以促進評分結果變得更具公正性、公平性與正確性。

肆、評分尺規應用的修改建議

若未來國家考試或其他類似考試承辦單位的口試評分，擬欲採用 MFM 模型來進行口試評分的話，筆者認為考選部或其他考試承辦單位仍必須採行一些配套措施來配合，才能落實具備客觀測量（objective measurement）特色的 Rasch 測量模型於口試評分中。例如，現行的「口試規則」第 5 條規定，個別口試及集體口試之評分項目及配分如下，整理如表 1 所示：

一、儀態（包括禮貌、態度、舉止、應對）二十分。

二、溝通能力（包括傾聽與表達能力）二十分。

三、人格特質（包括嚴謹性、情緒穩定性、開放性、和善性等）二十分。

四、才識（包括志趣、問題判斷、分析、專業知識、專業技術與經驗）二十分。

五、應變能力（包括理解、反應能力）二十分。

表 1

個別口試及集體口試評分表

考試名稱：　　應考人編號：

等別類科（組）：　　考試日期：　年　月　日

項　　目	配　　分	評　　分
儀態 （包括禮貌、態度、舉止、應對）	２０分	
溝通能力 （包括傾聽與表達能力）	２０分	
人格特質 （包括嚴謹性、情緒穩定性、開放性、和善性等）	２０分	
才識 （包括志趣、問題判斷、分析、專業知識、專業技術與經驗）	２０分	
應變能力 （包括理解、反應能力）	２０分	
合計	１００分	
評語：		
備註： 一、每一應考人之口試成績，以該組口試委員評分總和之平均數為實得成績。 二、口試成績未滿六十分者，應加註理由。 三、口試成績未滿六十分者，總成績雖達錄取標準，仍不予錄取。		

口試委員　　簽章

註：引自**口試規則**（https://law.moj.gov.tw/LawClass/LawAll.aspx?PCODE=R0030044）。

而為了讓口試委員在打分數時，有一個更具體的評分指引，考選部甚至也會列舉出如表 2 所示假想的評分尺規（scoring rubrics）範例，以作為更進一步規範上述評分事項內的評分指引。例如，外語口試規則第 6 條規定，分別規定外語個別口試之評分項目及配分；茲以外語導遊人員第二試口試評分表為例，擬定一份假想的評分尺規說明如下（如表 2 所示）。在此份假想的評分尺規範例中，每位口試委員在評分時，即可針對三項評定指標（即外語表達能力、語音與語調、及才識見解氣度等），各評定出應考人外語口語表達的「優」、「佳」、「普通」、「差」等程度，並在相對應得分範圍內的一個分數，評定出一個具體分數；例如，在「外語表達能力」項目下，口試委員可以打出：「59-54」、「53-48」、「47-36」、「35 以下」等分數，在「語音與語調」項目下，口試委員可以打出：「19-18」、「17-16」、「15-12」、「11 以下」等分數，而在「才識見解氣度」項目下，口試委員則可以打出：「19-18」、「17-16」、「15-12」、「11 以下」等分數。最後，再把這三項評定指標分數加總，即為應考人的口試分數。

然而，若根據表 2 所示的評分尺規來進行評分的話，其實，評審是很難打出一個具體又精確的分數的；也就是說，落在同一個等第裡的應考人表現，其實還可以再細分成某個區間段落，可以有多種分數評定的差別。例如，同樣在外語表達能力指標上表現屬於「優」等的應考人表現，他可能獲得 59、58、57、56、55、或 54 分不等的得分。但是，評審真的能區辨出給 59 分與給 58 分的應考人表現之間的差別嗎？其實很難！人類大腦在做這種精細的給分判斷時，其實是不具效能的，不僅是判斷的誤差大，且又很難下決定；充其量，我們只能粗略地判斷出別類別與類別之間的異同，而無法再進一步精細判斷出類別內的細微差別。這也就是說，口試評分的計分方式以帶有次序性尺規的離散資料（ordinal scale / discrete data）類型來編碼，遠勝過於以等距性尺規的連續資料（interval scale / continuous data）類型的編碼為宜。這不僅較符合人類大腦的判斷給分邏輯，也更能適合應用 MFM 模型來進行資料分析。

對此，筆者提出一個針對現行評分尺規應用的建議：那就是在評分時，只要求口試委員針對各項評分指標項目打出一個「優」、「佳」、「普通」、「差」等四個等第或4、3、2、1等四個類別分數（即當作次序性變項資料編碼），而不需要精細地打出一個具體的百分制分數（即當作連續性變項資料編碼）即可。如此的評分

方式，口試委員不僅較容易正確判斷出應考人表現之間的優劣、等第間的差別，資料編碼也較能符合 MFM 模型背後的客觀測量特性與演算法邏輯，正好是一舉兩得。考選單位只要收集每位口試委員對每位應考人在每一口試問題上的評分等第分數（如：4、3、2、1 分，分別代表「優」、「佳」、「普通」、「差」四個等第），再套用諸如 CONQUEST 等電腦程式，選用如公式 3 或公式 4 的估計模型，即可估計出每位應考人的能力值(θ)；同時，電腦程式也可以估計出每位口試委員的評分嚴苛度值(ρ)、口試評分項目的困難度值(b)，以及不同等第之間能否被跨越的門檻值(δ_j或λ_{ij})。接著，再將每位應考人的能力值(θ)排序或排名次後，考選單位即可根據預擬錄取的人數，決定錄取前幾個名額即可。

表 2

假想的外語導遊人員第二試口試評分尺規

評分項目、配分及量表			優	佳	普通	差
外語表達能力	60 分	聽力	能完全正確理解對方表達的語言內容及意涵	尚能正確理解對方表達的語言內容及意涵	能正確理解對方表達的語言內容，但偶而無法掌握意涵	完全無法正確理解對方表達的語言內容及意涵
		表達	簡潔、流暢、有條理	尚屬簡潔、流暢、有條理	能簡潔、流暢、有條理，但偶而出現模糊混亂情形	總是出現模糊混亂情形
		文法	完全正確無誤	尚稱正確，偶有錯誤	正確與錯誤情形相當	完全錯誤
		辭彙	用詞妥適而豐富	尚屬用詞妥適、豐富	用詞雖屬妥適但無法加以變化	完全不適當
		內容	完全切中主題	尚能切中主題	有時未能切中主題	完全不能切中主題
		建議給分	**59-54**	**53-48**	**47-36**	**35 以下**
語音與語調	20 分	發音	準確清楚	尚稱準確清楚	偶有錯誤或模糊情形	完全錯誤或模糊
		語調	音量、速度、頻率一直維持適當	音量、速度、頻率尚能維持適當	音量、速度、頻率偶而出現過或不及現象	音量、速度、頻率總是出現過或不及現象
		建議給分	**19-18**	**17-16**	**15-12**	**11 以下**

表 2（續）

評分項目、配分及量表			優	佳	普通	差
才識見解氣度	20分	分析	分析事理時，總是脈絡清楚，思考周延，快速掌握問題關鍵	分析事理時，脈絡清楚，思考周延，可掌握問題關鍵	分析事理時，脈絡尚稱清楚周延，尚可掌握問題關鍵	分析事理時，脈絡混亂偏頗，無法掌握問題關鍵
		專業才識	總是以專門知識的觀點作為立論的依據，以展現專業優點	常以專門知識的觀點作為立論的依據，以展現專業優點	有時以專門知識的觀點作為立論的依據，以展現專業優點	未以專門知識的觀點作為立論的依據，以展現專業優點
		問題判斷	始終以理性分析利弊得失後，才判斷及選擇可行的策略或方案	常以理性分析利弊得失後，才判斷及選擇可行的策略或方案	有時以理性分析利弊得失後，才判斷及選擇可行的策略或方案	未以理性分析利弊得失後，才判斷及選擇可行的策略或方案
		態度	對談態度積極，且過程中一直保持情緒穩定	對談態度積極，但過程中偶而出現焦躁情緒	對談態度尚稱積極，但過程中有時出現焦躁情緒	對談態度顯得消極，且過程中始終出現焦躁情緒
		建議給分	**19-18**	**17-16**	**15-12**	**11 以下**

資料來源：作者自行整理。

此外，若能彈性調整各項評分指標的占比，而不是以法令規範方式將其限制固定（如目前口試規則即是限制各項指標的得分均占比相同），開放各項評分指標的占比可以根據不同口試種類、不同考試的重要性、不同職能分析所強調重點評分指標的不同，再來給予考前臨時性的調整和決定的話（如表 2 假想的外語導遊人員第二試口試評分尺規所示），這不僅能使口試規則的運用更具有彈性、靈活性，也讓口試評分更能反應出口試成績在各類科考試中所占的重要性。

伍、結論

口試評量在文官制度的建置中，扮演著一個相當重要的篩選角色。因為它能夠綜合評估考生的多方面能力，符合各種不同文官職務需求，同時並提供一種公平和透明的評估機制。尤其在當今的國家考試中，我們所實施的是嚴謹的結構化口試，這是一種非常有助於選拔出最符合職位需求的優秀文官，確保公共行政推動的專業和效能的評估方法。然而，即使在嚴格遵守實施結構化口試的標準作業流程下，評

分者效應對口試委員的評分影響，仍然是一件無法避免且不容小覷的干擾事件。

因此，針對此評分者效應問題，筆者根據實作評量方法學的研究成果，建議採用一套兼具客觀測量特性與符合數理邏輯演算法的資料分析策略，以徹底解決口試評分資料的估算問題，降低評分者效應對應考人口試能力表現的干擾，並增進口試評量的信度與效度。這套資料分析策略，即是應用 Rasch 測量模型家族中的 MFM 模型於口試評分資料的分析上。

多面向模型可以採用二元計分資料的陽春 Rasch 模型，或採用多元計分資料的 RSM 模型或 PCM 模型皆可，端看口試評量所設定的給分編碼（score coding）機制而定。為了搭配這種測量模型的應用，筆者也同步建議考選部現行的評分尺規給分機制，宜改成僅打等第制或次序性資料編碼的分數，而不是打百分制的連續性資料編碼的分數。如此一來，在此給分編碼機制的配套措施下，配合 MFM 模型於資料分析的應用，便可徹底解決評分者效應對口試評分公平性的干擾問題，還原口試評量在文官制度建置中所扮演角色的重要性與價值性。

參考文獻

王成基（2004）。**加強公務人員考試口試功能之研究**。考選部。

考選部（2003）。**國家考試口試參考手冊報告**。考選部。

考選部（2004）。**考選研究系列 10—國家考試口試技術研討會會議實錄**。考選部。

考選部（2005）。**國家考試口試制度檢討報告**。考選部。

考選部（2006）。**九十五年度考選制度研討會—國家考試口試方法與技術研討會會議實錄**。考選部。

考選部（2012）。**考選法規彙編**。考選部。

呂秋萍（2005）。**國中教師甄選口試決策歷程之研究--以結構方程模式檢驗**〔未出版之碩士論文〕。國立政治大學。

吳復新（2000）。面談的問題及其改進之道：兼評高考一級口試改革方案。**空大行政學報**，（10），27-67。

吳復新（2007）。國家考試口試方法技術檢討與改進之研究。**空大行政學報**，（18），1-24。

余民寧（2013）。口試在國家考試應用之再檢討與改進。**國家菁英季刊，9**（2），87-107。

余民寧（2020）。**量表編製與發展—Rasch 測量模型的應用**。心理。

余民寧（2022a）。**教育測驗與評量：成就測驗與教學評量**（第四版）。心理。

余民寧（2022b）。有筆試就好了，為何還需要口試？**國家人力資源論壇**，（23）。https://www.exam.gov.tw/NHRF/News_EpaperContent.aspx?n=3778&s=46157&type=D080DDED9DA55705。

余民寧、謝進昌、林顯達、陳柏霖、許嘉家、湯雅芬（2011）。**國家考試集體口試參考手冊（含集體口試範例光碟）**（專題研究案結案報告）。考選部。

林文銘（2005）。**陸軍指職軍官甄選制度之研究-以情境口試建構為例**〔未出版之碩士論文〕。玄奘大學。

胡悅倫（2008）。結構化教師甄試口試之初步調查。**教育與心理研究，31**（1），65-96。

胡悅倫、余民寧（2009）。中學教師甄選口試題目圖像及其教育理念之研究。**教育與心理研究，32**（1），29-56。

胡悅倫、陳世芬、呂秋萍（2008）。教師甄選面試結構化問卷之編制。**測驗學刊，55**（1），185-212。

胡悅倫、陳世芬、莊俊儒、楊念湘、洪雅琪（2010）。**國家考試口試參考手冊**。考選部。

胡悅倫、陳皎眉、洪光宗（2009）。國家考試口試之命題與評分。**國家菁英季刊，5**（4），35-56。

陳淑慧（2006）。**在低結構化口試情境下應試者人格特質與口試結果之關係--從五大人格、自我監控、自我效能談起**〔未出版之碩士論文〕。國立政治大學。

彭錦鵬（2010）。公務人員考選制度的變革與未來展望。**國家菁英季刊，6**（1），17-40。

謝名娟（2017）。誰是好的演講者？以多層面 Rasch 來分析校長三分鐘即席演講的能力。**教育心理學報，48**（4），551-566。

謝名娟（2020）。從多層面 Rasch 模式來檢視不同的評分者等化連結設計對參數估計的影響。**教育心理學報，52**（2），415-436。

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Arthur, D. (2005). *Recruiting, interviewing, selecting & orienting new employees* (4th ed.). Amacom Books.

Bunting, S. (2007). *The interviewer's handbook: Successful interviewing techniques for the workplace*. Kogan Page.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655-702.

Dipboye, R. L. (1994). Structured and unstructured selection interviews: Beyond the job-fit model. In G. R. Ferris (Ed.), *Research in personnel and human resources management*, Vol.12 (pp. 79-123). JAI Press.

Dipboye, R. L. (2005). The selection/recruitment interview: Core processes and contexts. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 119-142). Blackwell.

Dixon, M., Wang, S., Calvin, J., Dineen, B., & Tomlinson, E. (2002). The panel interview: A review of empirical research and guidelines for practice. *Public Personnel Management*, *31*(3), 397-428.

Erguvan, I. D., & Dunya, B. A. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia*, *10*(1), 1-20.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies*, *1*(11), 1531-1540.

Gordon, R. A., Peng, F., Curby, T. W., & Zinsser, K. M. (2021). An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly*, *55*, 149-164.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.

Huffcutt, A. I. (2011). An empirical review of the employment interview construct literature. *International Journal of Selection and Assessment*, *19*(1), 62-81.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*(2), 184-190.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-

analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897-913.

Kermad, A., & Bogorevich, V. (2022). Using statistical transformation methods to explore speech perception scale lengths. *Language Teaching Research Quarterly*, *29*, 65-91.

Kogan, J. R., Dine, C. J., Conforti, L. N., & Holmboe, E. S. (2023). Can rater training improve the quality and accuracy of workplace-based assessment narrative comments and entrustment ratings? A randomized controlled trial. *Academic Medicine*, *98*(2), 237-247.

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

Linacre, J. M. (1991). *A user's guide to FACETS*. Facets.com.

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the employment interviews: A meta-analysis. *International Journal of Selection and Assessment*, *1*(1), 18-26.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599-616.

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.

O'Brien, J., & Rothstein, M. G. (2011). Leniency: Hidden threat to large-scale, interview-based selection systems. *Military Psychology*, *23*(6), 601-615.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage.

Popham, W. J. (2008). *Classroom assessment: What teachers need to know* (5th ed.). Pearson.

Sears, G., & Rowe, P. (2003). A personality-based similar-to-me effect in the employment interview: Conscientiousness, affect-versus competence-mediated interpretations, and the role of job relevance. *Canadian Journal of Behavioural Science*, *35*(1), 13-24.

Taylor, P. J., & O'Driscoll, M. P. (1995). *Structured employment interviewing*. Gower.

Thorburna, M., & Collins, D. (2006). Accuracy and authenticity of oral and written assessments in high-stakes school examinations. *The Curriculum Journal*, *17*(1), 3-25.

Touchie, C., Humphrey-Murto, S., Ainslie, M., Myers, K., & Wood, T. J. (2010). Two models of raters in a structured oral examination: Does it make a difference? *Advanced Health Science Education*, *15*(1), 97-108.

Tsai, W. C., Huang, T. C., & Yu, H. H. (2012). Investigating the unique predictability and boundary conditions of applicant physical attractiveness and non-verbal behaviours on interviewer evaluations in job interviews. *Journal of Occupational and Organizational Psychology*, *85*(1), 60-79.

Wang, J., & Long, H. (2022). Reexamining subjective creativity assessments in science tasks: An application of the rater-mediated assessment framework and many-facet Rasch model. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000470

Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159-171.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35-51.

Yeşilçınar, S., & Şata, M. (2021). Examining rater biases of peer assessors in different assessment environments. *International Journal of Psychology and Educational Studies*, *8*(4), 136-151.

A Study on the Influence of Rater Effect on the Fairness of Oral Examination Scoring

Min-Ning Yu [a]

Abstract

Oral examinations have been used in selecting personnel for many years, and relevant research reports over the years have also supported the feasibility and effectiveness of adopting "structured oral examinations." When implementing "structured oral examination" measures, most of the many factors that may interfere with the fairness of oral examination scoring can be eliminated or improved through standard operating procedures. Only one of them -- the "rater effect" -- is difficult to exclude or control from standard operating procedures. This article puts forward a suggestion for the application of "many-facet model" (MFM) scoring to objectively analyze the estimated values of various factors in the oral examination committee scoring data, including the examinee abilities, the difficulties of the oral exam questions, and the severity/lenient parameters of the raters' grading. According to relevant literature, applying the MFM model to such data analysis involving rater effects can not only improve the accurate estimation of examinee abilities, but also improve the reliabilities of the overall data analysis, and promote more accurate scoring of oral exams to achieve impartiality, fairness, and correctness. In order to ensure that the MFM model

[a] Distinguished Professor, Dean of College of Education, National Chengchi University, e-mail: mnyu@nccu.edu.tw.

can be applied to the analysis of subsequent oral examination score data, this article also suggests that the scoring method of the ready-made oral examination score sheet be changed from an interval scale score (continuous data attribute) (e.g., assign a score within the score range of 50-59 points) into an ordinal scale score of discrete data attributes (e.g., only evaluate the score into excellent, good, average, and poor grades), in order to reduce the interference of the "rater effect" on the fairness of oral examination scoring.

Keywords: rater effect, structured oral examinations, many-facet model, grading fairness, grading severity/lenient parameter