國立臺灣師範大學教育心理與輔導學系
教育心理學報，2017，48 卷，4 期，551-566 頁
DOI：10.6251/BEP.20160801

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*

謝名娟
國家教育研究院
測驗及評量研究中心

在儲訓校長的培訓課程中，即席演講爲重要的培育能力之一。過去在進行演講評分，大多使用所有評審分數的加總或是平均當作最終的分數，然而，當比賽的人數眾多必須分組時，每一組評審的嚴苛度可能會對成績造成影響。本研究將評分項目、評分委員的嚴苛度放進多層面 Rasch 模型中估計，進而以較爲客觀的方式來評估儲訓校長的口語表達能力。研究結果發現，即使有受過訓練的評分者，在評分過程中，還是有嚴苛度的不同，而儲訓校長在演說上，對於內容、架構，語詞使用、與時間掌控等部份較感困難，而對於發音標準、演說中有合適的語調，具有良好的儀表態度等向度感到較爲容易。此外，若忽視納入可能影響的層面而直接使用原始分數總分或是平均做爲受試者的成績，可能會造成分數上的誤差與排名的不公。

關鍵詞：多層面 Rasch、即席演講、校長評量

* 本篇論文通訊作者：謝名娟，通訊方式：hm7523@hotmail.com。

壹、前言

校長在國家推動教育改革和追求學校進步的過程中，爲落實教育活動的靈魂人物之一，牽動著學校的發展方向（黃姿霓、吳清山，2010），其領導的良窳與學校效能高下有密不可分之關聯（秦夢群，2007）。以先進各國校長培訓與專業發展課程的教學與學習觀之，可知爲達到目標，提升教學及學習之成效，先進各國課程皆大量應用研討會、團體會議、線上學習、師傅教導、自我學習、實地參訪、學校實習等多元教學型式來引導學員校長學習，課程講師也多由具有實務經驗之校長或退休校長來擔綱（秦夢群，2007；NPQH，2016）。基本上，校長培訓課程之教學法必須多元化，傳統講述法或上對下之單向知識傳遞已不符所需，唯有學員與實際情境進行學習互動，方能提升其專業發展之效果。

過去對於校長培訓的課程模式有許多相關的研究，但是針對校長培訓的評量模式較少著墨。然而，研究團隊於前導研究針對評量模式進行多次的學員校長與輔導校長的焦點訪談以及問卷調查，發現了現有的校長儲訓評量方式具有諸多限制，例如多數學員校長並不認同期末紙筆測驗，期望減少作業與報告，以及學習與測驗內容更能貼近教育現場。職是之故，相關結果引發研究團隊研發實作的情境式評量之興趣。並配合校長儲訓課程，發展評量指標，並選擇部分內容，進行擬真情境題的研發與試驗。

在諸多校長所應具備的多元能力中，即席演講爲校長必備的能力之一。在台灣的社會，校長的社會地位相對崇高。在大大小小的場合，像是校園的運動會、畢業典禮、開幕式，甚至是結婚典禮、募款餐會、參加地方活動等，只要有校長出現的場合，總是需要校長上台致詞，有時會會讓校長事先準備，但常常是必須要能隨時上台，因此即席演講的能力，在校長儲訓的課程中，也是重點之一。然而，即席演講的能力，要如何評估? 甚麼特質演講者，是好的演講者? 在評估的過程中，遇到了甚麼樣的問題? 評分者的評分一致性、嚴厲的評分者與寬鬆的評分者，所造成的受試者可能的影響爲何? 這都是本研究欲探究的問題。

在過去國內的研究中（張新立、吳舜丞，2008；姚漢禱、姚偉哲，2008），曾應用了多層面的 Rasch 模型（Multi-Facets Rasch Model, MFRM）來衡量試題難度與評審嚴厲度，以客觀的評估受試者能力，由於多層面 Rasch 模型能將原始的順序或是等級尺度的分數轉化成對數（logit）尺度的分數，因此在各個層面所估出的參數值都可以進行統計分析與比較。在本研究中，將透過校長的三分鐘即席演講的資料，使用 MFRM 的分析，探討使用原始分數的誤差。而這一套口語評量的評估方法也可做爲後續學者進行相關實作評量的參考。本研究所蒐集的影音資料，亦用以整理歸納校長之演講特質。

貳、文獻探討

一、校長培訓

學校校長常須面對各種不同的責任、挑戰。然而，應該要如何培訓校長，讓校長具有應有的能力，爲各校長培訓機構之重點。Hess 與 Kelly（2007）調查了 56 個校長培訓機構，發現大多數的培訓機構都相當類似，其中有 42%的課程內容聚焦在學校律法、學校財務管理等專業性課程，有 11%的課程則爲統計數據管理等實務研究課程，11%爲教學管理議題，如課程發展、教室管理、學習理論等。其他類別則有人力資源管理、教師任用、資遣、選擇、公共關係經營等。此外，有 12%的內容價值判斷的課程，內容包括社會公平、學校改革、性平、同性戀議題等。然而，這類的課程，容易出現意識形態的誤差。就 Hess 與 Kelly 的研究報告指出，校長學課程需要更與時俱進，貼近學校的需求，尤其以證據爲導向的社會，應在課程中融入更多資料、科技與實徵研究的相關課程。

反觀臺灣國中小校長儲訓制度，主要由培訓單位針對候用校長，設計一系列專業的課程、實習及評量，根據陳木金、李俊湖（2006）「國民小學校長培訓模式之研究」研究報告與陳木金（2009）所提出的文獻指出，可將國家教育研究院之培訓，大致區分爲三時期-思想教育培訓、角色任務培訓、專業發展培訓。

第一個部分爲爲思想教育培訓期（1965 年-1988 年），在此時期，培訓包含「板橋期良師興國階段」、「板橋期中興思想階段」、「板橋期愛國精神階段」。這時期的課程特色反映了政局影響，尤其在課程設計及制度規劃上，強調反共精神、時局政策爲導向，受中興思想、良師興國、愛國精神影響，因此在培訓設計理念對於校長個人修養、生活禮節、品德修養、教育專業能力養成均相當重視。

第二個部分爲校長角色任務培訓時期（1989 年-1998 年），內容包含「板橋期民主開端階段」、「板橋期博雅教育階段」、「板橋期教育改革階段」。本時期受到政局影響在課程設計及制度規劃上，培訓設計理念強調民主思潮爲導向，受教育改革、博雅教育、民主開端影響，因此在培育課程上設計基本修養、通識教育、博雅教育、教育專業知能、教育政策、領導才能、校務行政理論、休閒活動等方面。

第三部分爲校長專業發展培訓時期，從 1999 年迄今，包含「三峽期行政管理階段」、「三峽期專業成長階段」、「三峽期師傅校長階段」、「國教院特色培訓課程發展階段」。受到社會、經濟、文化、政治之影響，在課程設計及制度規劃上，培訓設計理念強調校長專業能力發展，延續前階段民主思潮影響，課程上較爲多元、開放，以師傅教導、實務實習爲核心理念，因此在培育課程設計出通識教育、教育政策、教育專業知能、休閒活動、教務行政理念與實務、行政理論、綜合活動等方面，以養成校務發展、行政管理、教學領導、公共關係、專業責任之能力。

二、校長專業能力

諸多學者對於校長的「系統性知識體系」的專業能有多有詮釋，高慧容（2007）指出校長須具備三種能力：專業價值、管理功能及專業能力。以「why」來發展專業價值，包括價值、學習、知識；以「what」來發展管理功能，包括管理學習和教學、管理人員、管理政策和計畫、管理資源和財政；以「how」來發展專業能力，包括人際關係和智能。張榮輝（2009）則認爲國中小校長的主要工作內容包括政策執行與學校經營。政策執行爲依據國家上及法令政策推動校務，擬定校務計畫、舉辦學校活動並確實執行；學校經營則分爲教學領導與學校行政領導兩方面。首先教學領導工作包括教學視導與輔導、課程研究發展、教學指導與革新，學生事務與服務等四大層面。其次學校行政領導工作則包括：學校計劃與組織、行政決策、管理與執行、監督與評鑑、設備與經費，公共關係等。

新北市卓越學校則明列校長領導的向度有 12 條指標（新北市教育局，2010）。其指標如下（一）前瞻的辦學理念，校長教育理念能反映社會發展趨勢，辦學理念能結合教育改革潮流、展現學校教育願景。（二）高度的專業素養：校長熟悉主要教育理論與原理，並具備豐富的行政管理、課程與教學領導知能、能參與專業組織，並進行專業成長。（三）創新的領導作爲：校長能洞察時勢，了解校務發展需求，善用領導策略，引領學校變革與創新，並能激勵成員，積極解決校務問題。（四）優質的領導效能，校長辦學具有績效贏得認同，能提升團隊能量，落實學校教育目標，能整合各界資源，支援學校教育運作等。

國內外學者對於校長工作職責與內容的界定並不太一致。校長的職司，可以說相當複雜。茲綜合國內外學者專家的意見將校長的職責歸納依屬性分爲以下五方面：

（一）人事管理：學校人員聘任、考核、監督等。

（二）教學輔導：課程發展、課程教學、學生生活指導。課程目標與設定、

（三）課程組織與編製、課程執行、教材選擇、教學時間編排、以及教師分配工作等。

（四）公共關係：校內外師生、家長、社區人士、媒體與上及各機關之溝通聯繫。

（五）學校建設：除了校舍規劃與建設、軟硬體設備的維護與使用等物質建設之外，更要加強文化與精神建設，建立生命共同體意識，塑造優質的學校氣氛。

三、演講的要素

在口語演講中，有幾個重要的要素需要評估，Aryadoust（2015）指出，除了聲音、音調、非語言的表達能力、手勢、臉部表情，都會影響演講者所傳達的訊息（Sundrarajun & Kiely, 2010），根據 Aryadoust 所歸類，在演講中決定好壞的要素有三項，包括語言溝通能力、非語言溝通能力與內容組織能力。

（1）語言溝通能力

決定演講者演說能力最重要的就是口語表達能力，爲了讓聽眾能充分了解演講的內容，演講者需要有清楚的發音、講話的速度要適中。演講中可能會有一些短暫的停頓或錯誤，但是不能夠影響到聽眾的專注與理解的連貫性。在語音表達中，也應該有抑揚頓挫，來強調演講的重點與方向（Kormos & Denes, 2004; Pickering, 2004）。文法與修辭也影響語言溝通的能力，使用豐富多變的字彙與不同的修飾語句來表達有助於吸引聽眾的專注力，讓聽眾更能了解演講者所要傳達的意思（Luoma, 2004; Jensen & Harris, 1999）。

（2）非語言表達能力

非語言表達能力包括肢體語言、穿著、眼神、臉部表情和手勢（Aryadoust, 2015），語言表達可直接傳達訊息，但是適度的非語言的表達有助於聽眾對於訊息的理解。而眼神、臉部表情等都能有助於聽眾對於訊息的理解，

（3）內容和組織能力

內容和組織能力對於演講的影響並不直接，但是對於學術性的演說，有組織性的演講內容、每個段落適度的標題和重點、有助於聽眾更加理解內容（Aryadoust, 2016），透過有組織的內容，可以讓聽眾更聚焦演講者的內容，並正確的理解訊息。

四、多層面的 Rasch 模型

近年來 MFRM 的使用相當頻繁，尤其在語言測驗、教育與心理測量、醫療科技領域等（Bond & Fox, 2015; Engelhard, 2012; Harasym, Woloschuk, & Cunning, 2008; Wolfe & Dobria, 2008），例如在歐洲語言委員會所建構的歐洲共同語言參考標準（Common European Framework of Reference for Languages，簡稱 CEFR）中，MFRM 爲其中一個使用的統計方法來建構量表與口語評量的尺度，North（2000）更指出 MFRM 這個統計方法，對於 CEFR 的發展具有獨特的相關性。

在口語評量的應用上，常需面臨的是評分一致性（consistency）與分數同意度（agreement）的問題，一致性代表一群學習者的成績排序不應受到評分者的影響，同意度則是成績的高低，也不應因爲評分者的不同而有分數的落差 （Stemler & Tsai, 2008; Wolfe, 2004）。然而，評分者的評分會受到許多因素的影響，例如在演講的評分中，評分者的主觀經驗、選擇的講題難度、評分的規準、或是其他隱藏的層面都可能對於受試者的成績造成影響（Eckes, 2011）。雖然在評量的過程中，可透過加強評分者的訓練，來強化評分者的一致性，但 Elder, Knoch, Barkhuizen 與 von Randow（2005）指出，即使有進行評分者訓練，但對於評分者之間的評分一致性的提升仍相當有限。

例如在演講評分過程中，某些評分委員認爲演講者發音標準相當重要，因此對於發音是否標準，會給予較多的權重，然而，其他委員可能覺得內容比發音更重要，只要聽眾聽懂就好，發音是否正確，相對就不是那麼重要了。而在同意度部分，有些寬鬆評分者喜歡給高分，有些嚴苛的評分者偏好給低分，也有些中庸的評分者總是給趨中的分數，這些給分的偏好，也容易造成了較低的評分同意度（Eckes, 2011）。

MFRM 可用來解決這類的評分者問題，Aryadoust（2016）使用 MFRM 來評估學生口頭發表時，同儕評分的評分者信度。在其研究中，共有 66 名上課的學生，根據同學的口頭報告進行互評，每位同學互評 10~14 人。其評分同時考量評分者與發表者的性別與主修，研究發現，若不進行 MFRM 的校正，當發表者表現相同時，評分者與發表者性別相同時，給分較低，性別不同時，給分較高。而評分者與發表者的主修相同時，會給較低分，但對與自己主修不同的發表者會給較高分。因此在其研究中，特別將評分者與發表者的性別與主修加入層面中進行分析，並用以校正分數。

Busturk（2008）則將 MFRM 應用在學生的口頭簡報評分上，其模式包括三個層面（整組的發表能力、評審者的嚴苛度、與被指派議題的困難度），研究者並檢視模式中的適配度指標（包括 Infit 和 Outfit），提出在評分過程中，根據適配度較差的項目加以討論，並透過這種模式進行分數的校正。

國內也有相關學者進行 MFRM 在各種評分的應用，如張新立、吳舜丞（2008），應用在學術研討論的論文評分，姚漢禱與姚偉哲（2008）應用在雙不定向飛靶優秀選手的射擊技術，謝名娟（2013）用於評估標準設定成員之間，在設定標準分數的變異性，謝名娟與謝如山（2013）則應用在評估數感及生活應用能力中，實作評量的分析。這些研究都提出 MFRM 合理有效的應用。

參、研究方法

一、研究對象

本研究的對象爲通過校長考試，在研究機構接受校長儲訓的學員，在儲訓課程中，有三次的三分鐘演講課程，前兩次的課程爲訓練，輔導校長會針對每位學生上台的講評，而最終一次的表現則當作評量，其成績占儲訓期間學要表現的 5%。雖然所占比重並不高，但由於儲訓成績在某些縣市爲學校分發之依據，因此大多的學員都極力表現，以求高分。

學員在上台演講前，先從題目袋中，隨機抽取一個題目，而後就這個抽到的題目，準備三分鐘，上台演說。針對最後一次的訓練課程，研究團隊進場錄音錄影，錄製好的影片經由團隊匿名處理分成四組，每一組的學員交由兩位資深校長進行評分，總共有 128 學員、八位評分者進行評分。

演講的題目爲資深校長依據校長常需要即席演講的內容彙整而成。如各類活動致詞，包括休業式、校長對學生談假期生活安排、畢業典禮校長致詞、學校校慶開幕致詞、新任校長交接典禮致詞等，題目從題庫中選出，多爲資深校長根據校務經營經驗，常遇到的演講題目所設計而成。

二、評分規準

根據兩次與資深校長的焦點座談，擬出之評分的標準如表 1，共分成四大項：內容、儀態、表達技巧及時間掌控。其中內容包括符合主題、架構分明，並針對不同對象使用適當的語詞。例如對於學生演講，言詞應該要盡量深入淺出，多使用一些範例、故事來吸引聽眾的注意，而對於教師演講用語可以較爲學術性。在儀態部分，則爲儀表態度得宜，原先儀表包括服裝儀容，但是在儲訓學員中，服裝都相當端莊得宜，在此項目中不具區辨性，因此向度改成根據演講時的態度是否符合情境所需。在表達技巧部分，分成發音和語調。發音標準，語調與語速要合宜。 最後在時間掌控部分，也是評分的要點之一。因爲很多演講者一上台就忘了時間，尤其校長常常犯了多話的毛病，因此時間掌控在校長演講的評分中，也占了相當的比重。

原本的表格設計要求評審在每個細項都給具體分數，使各項加總成績爲 100 分。然而，在執行的過程中，發現能考上的校長程度都相當高，要能進行些微分數間的區辨，有相當的困難度。

因此，在實際執行演講評分表時主要是採效標參照的精神，請資深校長判斷儲訓校長針對在各個面向中所達成的程度是屬於完全符合、部分符合或少部分符合。

表 1　校長儲訓班即席演講評分表

審查重點	符合要點程度		
	少部份符合	部分符合	非常符合
內容			
1、演說內容符合主旨（內容）			
2、架構分明、井然有序（架構）			
3、針對對象使用適當語詞（語詞）			
儀態			
4、儀表態度得體（儀表）			
表達技巧			
5、發音標準（發音）			
6、語調音量適切、語句流暢（語調）			
時間掌控			
7、時間控制得宜、結尾不匆促（時間）			

三、評分者訓練

在資深校長評分前，研究者會先行說明研究的目的，請評分者務必依照評分規準來進行評分，而後，則先取出其中五位學員的影音檔，做爲評分訓練使用。資深校長各自依據評分規準，獨立評分五位學員的分數，而後將差異較大的評分項目進行討論。例如在發音部分，是否標準每個評分者的落差很大，有些校長對於標準的定義是要像新聞主播一樣的發音才叫標準，但是對於其他校長而言，有一些台灣國語的腔調反而能讓演講更平易近人。因此在發音部分，大家的共識爲只要說話能讓人聽得懂，就可稱爲發音標準。然而，有部分的學員用本土語進行演講，評分者認爲這類的學員演講評分應該和使用國語演講的學員評分予以區隔，因此若使用閩語演講的學員，則不納入本研究中。

另外，爲了更了解校長的演講特質，研究團隊亦要求每位評分員，將每位演講者的優缺點記錄下來。其中特別針對前 1/3 的高分者與後 1/3 的低分者特質則進行歸納。同時，爲求正確性，研究團隊亦反覆檢視影帶中演講者的特質以求正確。主要整理的內容可做供研究團隊未來設計相關演講訓練課程的參考。

四、多層面 Rasch 模型

在本研究共有 8 位評分者，每位評分者必須就其所分到的學員三分鐘影像檔，進行評分，經刪除不適切的樣本後（如使用本土語演講、影音檔不清楚等）總共有 128 位學員納入，每位學員均有兩位評審的評分。在分析上，使用軟體 FACETS（Linacre，2014），此 FACETS 程式常用來分析多層面 Rasch 模式，可將估計的參數轉換成對數型尺度（logit scale）。爲了讓數據更符合 FACET 軟體的需求，在每個評分項目中，非常符合編碼成 3，部分符合編碼成 2，少部分符合編碼成 1。

本研究在評估受試者的口說能力時，同時考量的層面爲評分項目之難度與評分者的嚴厲度。對於第 n 位受試者而言，在接受評分是 K 等級，其評分項目是 l，評審的嚴苛度爲 j，在表現層級爲 j 時，針對這個題目，被此成員者評定爲 k 等級之對數勝算比可表示爲：

$$\ln(\frac{P_{nljk}}{P_{nljk-1}}) = \theta_n - \delta_l - \alpha_j - \tau_k$$

其中 P_{nljk} 爲第 n 位成員者，評分項目是 l，評審的嚴苛度爲 j，在表現層級爲 j，得分是 K 時的可能性。

其中 P_{nljk-1} 爲第 n 位成員者，評分項目是 l，評審的嚴苛度爲 j，在表現層級爲 j，得分是 K-1 時的可能性。

θ_n 爲受試者 n 的能力；

δ_l 爲評分項目 l 的難度；

α_j 爲評審 j 的嚴苛度；

τ_k 指評定 K 或是 K-1 之間的難度界線，也稱爲難度階（threshold difficulty）。

由此模型可見，受試者本身的口語能力、評分項目、評審嚴苛度爲本研究要考量的層面。在評估模型適配度方面，可使用 infit 與 outfit 均方值做爲指標，當這兩個只介於 .7 到 1.3 之間（McNamara, 1996），代表資料適合使用 MFRM 進行分析，若均方值大於 1.3，代表數據存在許多干擾與不穩定性，若小於 0.7，則代表資料可能爲相依樣本，資料的獨立性不足。模式所估出之可靠度（reliability）也可看出資料之穩定性（Wright & Masters, 1982），越接近於 1，代表資料越穩定。另外，MFRM 既爲 Rasch 家族中的模型之一，亦要符合單項度的假設，其中一種有效檢測單向度的方法爲檢視 infit 與 outfit 等適配度指標，若在於適合的區間內，也可代表模型具有單項度的證據 （Linacre, 1998, 2010; Smith, 2002; Tennant & Pallant, 2006）。

肆、研究結果

本研究爲根據研究機構參加校長儲訓的學員，參與三分鐘演講的課程，所蒐集到的 128 位儲訓校長演講資料進行評估。採三層面的 MFRM 模型，包括評審項目的難度、評審委員的嚴厲度與受試者的表現，分析上使用 FACETS 軟體，並參考張新立與吳舜丞（2008）所使用的方式，將評審項目難度定在 0 logit 處，進行評審委員嚴厲度和受試者表現程度的估計。

如表 2 所示，評審委員的嚴厲度遠低於平均之評審項目難度，代表評審委員給分相對屬於寬鬆，而受試者的表現高於評審項目難度，代表大多數的受試者表現都是相當優秀的。能考上校長，在儲訓班受訓的校長，大多已具有一定的水準，程度分布屬左偏，因此校估的結果符合預期。另外，在適配度方面，Infit 和 Outfit 的均方值介於 .7 到 1.3 之間，代表使用 MFRM 來進行分析應屬適合。然而，在受試者表現的可靠度數值略低。

表 2　各層面模型估計與適配狀況

	參數平均值	參數標準差	Infit 均方	Outfit 均方	可靠度 Reliability	*p*-value
評審項目難度	.00	.16	.97	1.28	.96	.00
評審委員嚴厲度	-2.74	.20	.99	1.26	.97	.00
受試表現程度	.41	.82	.98	1.18	.77	.00

表 3 則呈現各評審項目之難度參數估計值，評分規準共有七個向度，在這七個向度中，參數值越高，代表這個項目對於受試者而言越難達成。架構分明、井然有序的參數值最高爲 1.06，代

表儲訓校長對於要能在短短三分鐘，具有系統的架構演說內容，並能清楚的呈現感到最為困難。在內容符合主旨、針對對象使用適當語詞、時間掌控部份，儲訓校長也不容易掌握。

相對而言，發音標準、演說中有合適的語調，具有良好的儀表態度對於儲訓校長而言，是相對容易達到的，其參數值較低。在 FACET 的分析中，亦提供分離度（Separation）的指標，這個指標假設為所有的觀測值是從一個常態分布的母群所隨機抽取出來的，而此母群中的統計特徵和觀測值完全相同，而從這種特性之常態分布母群中，能夠區分出幾個具有統計顯著差異性的類群（Strata）。如表 3 所示，評審項目的分離度為 5.67，代表評審項目中所設定的三個層級，有明顯的不同。此外，在時間部份的 Outfit 均方值為 3.27，此值大於 1.3，代表大多數的校長在時間掌控部分都很一致，變化性較少，但也可看出模式的整體適切度（model fit）稍有問題。

經檢視表 4 類別結構估計，可看出平均的測量值在跨個項目中逐漸擴大，且評審項目的類別結構測量值（Rasch-Andrich Thresholds），均逐漸增加且相鄰的類別結構測量值變大幅度小於 5 logits（Linacre, 2002），代表使用 3 個類別來區辨評審項目應屬適當。

表 3　各評審項目之難度參數估計值

評審項目	參數值	標準差	Infit 均方	Outfit 均方	鑑別度
1. 內容	.73	.15	.89	.73	1.16
2. 架構	1.06	.14	.85	1.02	1.15
3. 語詞	.66	.15	.92	.87	1.09
4. 儀表	-0.71	.17	1.06	1.02	.93
5. 發音	-1.28	.19	.98	1.16	.95
6. 語調	-0.86	.17	.9	.89	1.09
7. 時間	.40	.15	1.22	3.27	.64

註：Model, Populn: RMSE .16 Adj (True) S. D. .84 Separation 5.24 Strata 7.31 Reliability .96
Model, Sample: RMSE .16 Adj (True) S. D. .91 Separation 5.67 Strata 7.89 Reliability .97
Model, Fixed (all same) chi-square: 187.7 d. f.: 6 significance (probability): .00
Model, Random (normal) chi-square: 5.8 d. f.: 5 significance (probability): .32

表 4　類別結構估計測量

類別	平均測量值	Outfit 均方	類別結構		類別中點
			測量值	標準誤	
少部分符合（1）	-0.74	1	無	--	--
部分符合（2）	1.21	1.5	-1.76	.14	-1.79
非常符合（3）	3.84	1	1.76	.07	1.78

從表 5 可看出，大多數的評審評分都是寬鬆的，但是還是有相對性的較為寬鬆或較為嚴格的評審，其中 2、7 號評審較為寬鬆，而 1、5 號評審較為嚴格。鑑別度亦為模式適切度之一（Linacre, 2010），其合理範圍應介於 .5 至 1.5 中間，各評審項目中之鑑別度均在可接受的範圍內。然而，2 號評審的 Outfit 均方為 2.78，落出可接受的範圍外（.7~1.3），代表這位評分者所給的分數太過一致，大多都給高分，變化性較少，從此也可看出模式的整體適切度（model fit）稍有問題。表 5 所可看出評審之間的分離度為 6.33，代表若是有一個與本研究的組成評審相似的常態分布母群，其評判的嚴厲度差異至少可以被分成 6 個層級，這也隱含即使受過訓練與討論的評審，在進行評斷時，還是很難脫離原先自己的主觀性。此外，經計算評審之 Rach-Cohen's Kappa 值為-0.08（其中 Obs% = 43%，Exp% = 44.8），由於此數值相當接近於 0，代表評審應具有獨立評分（independent raters）的特性（Eckes, 2011）。

表 5　各評審者之嚴苛度參數估計值

評審代號	參數值	標準差	Infit 均方	Outfit 均方	鑑別度
1	-1.18	.15	.81	1.04	1.16
2	-5.75	.46	1.13	2.78	.82
3	-3.3	.18	.8	.69	1.21
4	-2.05	.15	.79	.9	1.25
5	-1.92	.14	1.23	1.25	.76
6	-2.15	.14	1.07	1.27	.85
7	-3.56	.2	1.27	1.16	.78
8	-2.02	.16	.84	1.02	1.14

註：Model, Populn: RMSE .22 Adj (True) S. D. 1.33 Separation 5.91 Strata 8.22 Reliability .97
Model, Sample: RMSE .22 Adj (True) S. D. 1.42 Separation 6.33 Strata 8.78 Reliability .98
Model, Fixed (all same) chi-square: 193.6 d. f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.6 d .f.: 6 significance (probability): .36

MFRM 在進行估計時，將評審的嚴厲度與各評審向度的難度都考量進去，進而呈現儲訓校長本身表現的優劣，彼此的表現可以參考比較。然而，將這兩個層面考量進去，和使用傳統的總分來比高下，到底差異有多大?

圖 1 呈現 MFRM 的估算分數與每位校長在兩位評審者平均分數之落差，就此圖可看出分數的分布爲左偏，大多的校長都落在 2~3 的高分群，只有少數幾位低於 2 分。

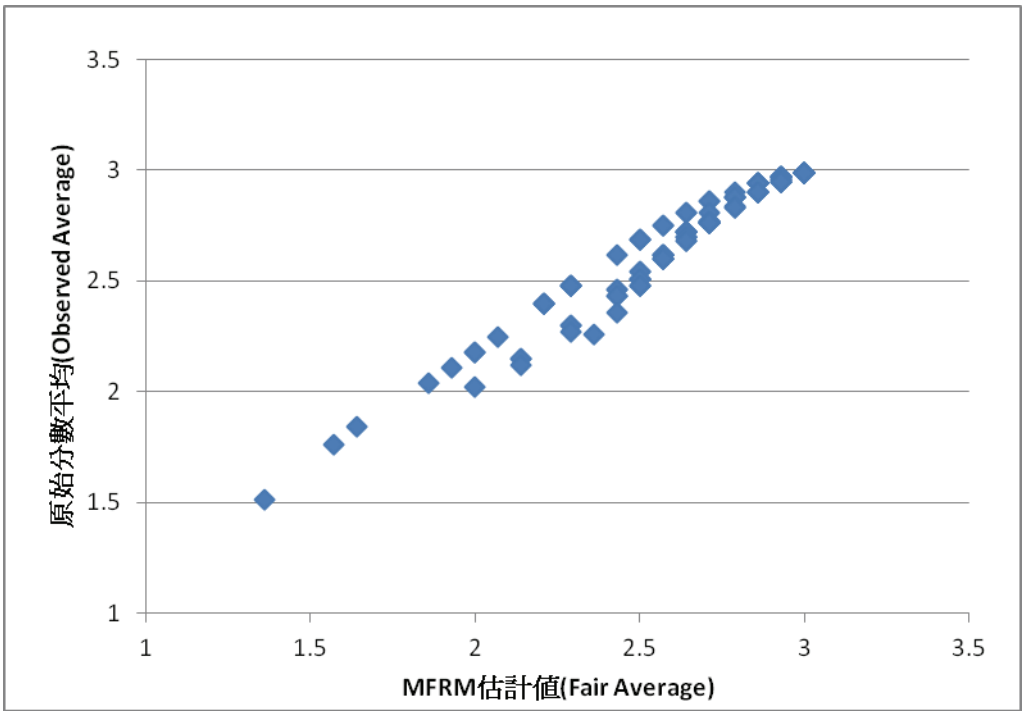


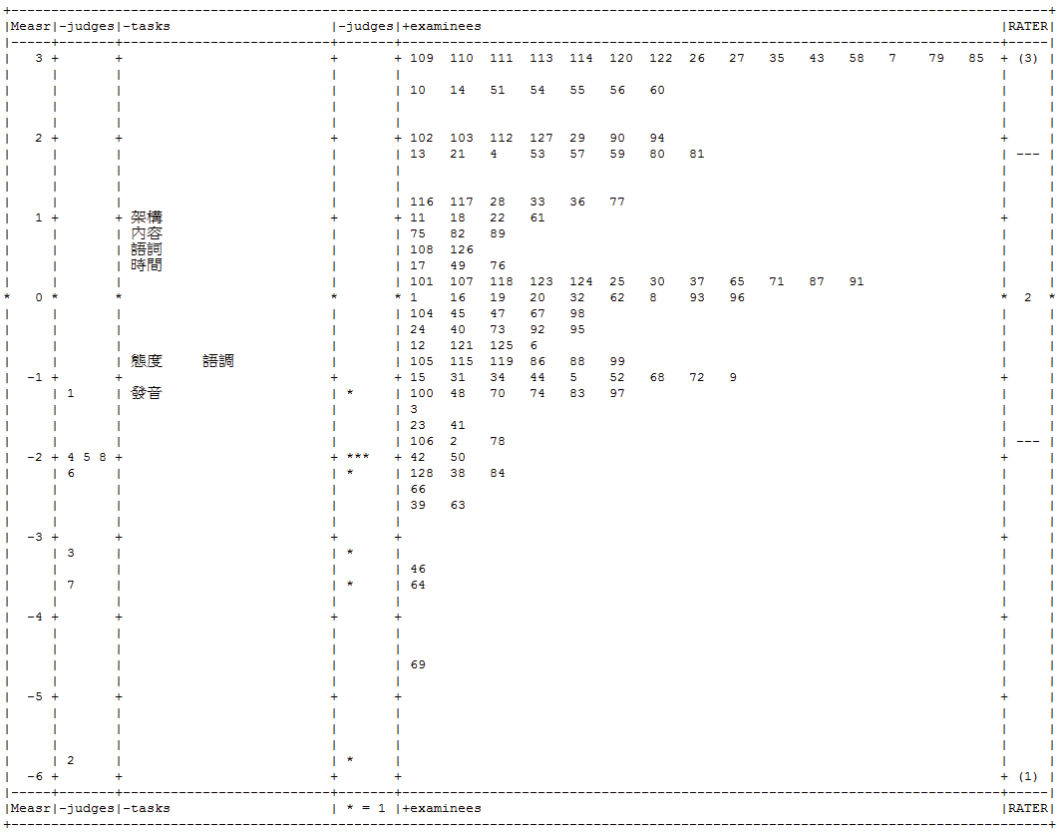
圖 1　MFRM 的估算分數分布圖

表 6 更近一步比較 MFRM 的量測分數和傳統原始的計分下所產生的排名差異。考量 MFRM 之後，在排名的分數上有相當大的不同，例如原先的 27 和 35 號，其原始分數是第 1 名，但是 MFRM 的排名是第 5 名。而在 7 與 58 號，MFRM 的分數是排序第一名，但是在傳統原始計分下卻是第

13 和第 5 名。落差較大的還有第 26 號與 114 號，MFRM 分數是第 5 名與第九名，可是在原始的計分卻是第 23 名與 45 名。

表 6　前十五名 MFRM 與原始成績排序之比較

受試者代號	MFRM 量測分數		傳統原始計分方式	
	分數	排名	原始分數	原始排名
7	4	1	170	13
58	4	1	177	5
43	3.85	3	171	12
79	3.85	3	178	3
26	3.38	5	166	23
27	3.38	5	179	1
35	3.38	5	179	1
85	3.38	5	173	8
114	3.35	9	159	45
122	3.35	9	172	10



註：從這個分布圖中可以看出本研究中所考慮各個層面中變數分布的概況。

圖 2　為變數分布圖（variable map）

除了量化的分析之外，本研究要求審查委員依據各演講者的演講特質，寫出質性的描述。本研究根據 MFRM 所分析出較佳與較差的演講者，對照審查委員在各評審項目下所寫的質性描述，

所呈現的演講特質如表 7。由這些特質可看出，在資深校長眼中，一個好的演講者，應該要能與觀眾互動、具有幽默感，會用關懷、親切的口吻、能適當但不用過度的使用肢體語言等，來進行演講，和傳統的高高在上、嚴肅的校長形象有所不同。另外，時間控制上對於校長來說也是相當看重的，要如何在一定的時間內，把該講的內容講完，不匆匆結束，也是一個好的演講者應該具備的重點。

表 7　高分組與低分組的演講特質

項目	高分組特質	低分組特質
內容	能利用小故事帶動氣氛，吸引目光，與聽眾互動。 有幽默感 分享自己的作法與其過程	太緊張，講話會一直停頓、中斷 口語贅詞、贅字太多 低頭看演講稿或題目的頻率太高
架構	內容架構、流程清楚 演講條理分明，讓人有種舒適感 最後做整個總整理來重新提醒聽眾該注意的事項	停頓不明顯，容易讓聽眾難以消化所有內容
合適詞句	演講的對象準確，抓到重點 透過詢問來與觀眾互動 感謝的態度與親切的口吻	有些場合客氣話不適合 互動到有點過頭，導致聽眾搞不清楚演講者想表達概念 少互動
儀表	肢體語言，手勢使用適當 表情和語調變化都很豐富 不斷的感謝與溫和的態度，大大強化了聽眾對演講者的認同感	幾乎沒有手勢，或是手只是在亂揮，無法與演講配合 表情僵硬，或面對學生，表情卻太嚴肅
發音	聲音宏亮，台風穩健	聲音過大，導致觀感不佳 講話停頓怪異，可能是緊張導致斷斷續續的停頓 聲音過小緊張而口吃，有時候會重複同樣字詞
語調	聲音抑揚頓挫有致，來強調理念	講話太快，讓聽眾有緊張感 聲音平淡，易使人難以集中注意力 有習慣性尾音用語 爲了強調而過分大聲 停頓語氣怪異
時間掌握	控制時間適當，不會有突兀感	時間掌控不佳，例如說舉 5 點報告，最後卻只講了 3 點 被鈴聲影響，導致整體演講中斷而草草結束

伍、結論與建議

本研究使用了多層面的 Rasch 模型（Multi-Facets Rasch Model, MFRM）來估算受試者能力，研究共使用 128 位儲訓校長的三分鐘即席演講影音資料與 8 位評分校長的評分數據來做分析，根據本研究所獲得的結論與建議如下。

（一）結論

1、MFRM 可協助口語評量更客觀的分析

本研究中使用 MFRM 來探究校長三分鐘演講的評分，透過考量評分項目、評分委員的嚴苛度，進而以較爲客觀的方式來評估儲訓校長的口語表達能力，未來口語測驗，也可考量使用 MFRM 來進行更爲客觀的分析。

2、本研究所用的數據符合適配度與可靠度的預期

根據本研究的結果，評分項目難度、評審委員的嚴苛度與口語表現的能力等層面，其適配度與可靠度都達到預期，代表本研究所使用的資料具有穩定性、符合 Rasch 模型單項度的假設，並且適合 MFRM 的分析。

3、校長在架構、內容、適當語詞與時間掌控上較感困難

在校長口語評分的各種評分項目中，校長在架構分明，內容符合主旨、針對對象使用適當語詞、時間掌控等部份較感困難，而對於。發音標準、演說中有合適的語調，具有良好的儀表態度等向度感到較爲容易達成。

4、使用原始分數的平均或總分會造成公平性的誤差

從評審嚴苛度分析可看出，大多數的資深校長評分較屬寬鬆，但每個評審的嚴苛度仍確有不同，若使用原始的評分排名，和使用 MFRM 考量各層面下的排名，卻有所不同，使用原始分數的加總或平均方式，可能造成公平性的誤差。

5、好的校長演講應具親民的特質

從演講者特質中，可看出資深校長對於一個好校長演講，應該是能使用親切有趣的方式、能跟聽眾互動，使用適當的手勢、語氣、語調與語詞來輔助演講的進行，而不是以嚴肅、生硬的八股內容、高高在上的校長形象來進行演說。

（二）建議

1、增加更多元身分的評分者

在本次研究中，評分者的身分僅限於資深校長，並依據資深校長多年的治校經驗來進行評斷，然而不同的評分者，對於演講內容的喜好與接受程度可能會有所不同。例如資深校長在看一個好的畢業致詞演講，可能看他有沒有講出對於學生的期望與期許、未來的升學建議等。但是對於學生來說，這些內容對他們可能很枯燥無味，他們反倒喜歡聽一些小故事、甚至是校長自己的一些經驗等等。未來的研究中，可以考量這些不同評分者的多元身分對於評分的影響與探索身分與評分的交互作用。

2、自動口語評分系統的推展與研發

執行口語評分時，最難克服的爲評分人力的問題。即使透過評分者訓練，但客觀性還是很難維持。除此之外，面對大量的評分作業，要能保持同樣的標準、體力、專注度都相當困難。本研究的研究對象僅有 128 人，但是爲了能讓評分更具公平性，將每位學員的演講資料錄音錄影，並動用了 8 位校長，在反覆看錄影帶的過程中來評分，每位評分者花費的時間至少超過 3 個小時，所需花費的成本甚巨!然而，在推動語言教育上，客觀且迅速的評估學生的口語的表達能力是測驗界上所須面對的問題之一，在科技發達的時代，應可考量機器自動的評分的輔助，然而，這樣的方向需要更多的經費和努力才能實現。

參考文獻

姚漢禱、姚偉哲（2008）：應用多層面 Rasch 模式分析雙不定向飛靶優秀選手的射擊技術。**測驗學刊，55**（1），89-104。[Yau, H. D., & Yao, W. C. (2007). Application of many-facet rasch model to analyze the skills of elite athletes in double trap. *Psychological Testing, 55*(1), 89-104.]

陳木金（2009）：我國國民小學校長儲訓模式的回顧與展望。**學校行政雙月刊，60**，98-120。[Chen, M. J. (2009). Reviews and prospects for principal preparation training model for elementary school in Taiwan. *Journal of School Administration, 60*, 98-120]

張榮輝（2009）：臺北縣卓越學校指標的發展歷程, 載於臺北縣政府教育局（主編），**2010 臺北縣邁向卓越學校：指標系統與行動方案**（9-13）。臺北：臺北縣政府教育局。[Chang, J. H. (2009). Development Procedure for Excellent School Indices in New Taipei city. In 2010 *Excellent Schools in New Taipei City: Indices and Action Plan* (pp. 9-13). Taipei, Taiwan: New Taipei City Education Bureau.]

張新立、吳舜丞（2008）：多層面 Rasch 模式於學術研討會論文評分之應用。**測驗學刊，55**（1），105-128。[Chang, H. L., & Wu, S. C. (2008). A multi-facet rasch analysis on rating the academic scientific papers. *Psychological Testing, 55*(1), 105-128.]

高慧容（2007）：蘇格蘭校長培訓制度對我國國小校長培育課程之啓示。**學校行政雙月刊，47**，27-42。[Kao, H. J. (2007). The implication of scottish qualification for heads/up in curriculum of principal preparation for elementary schools of our country. *Journal of School Administration, 47*, 27-24.]

黃姿霓、吳清山（2010）：美國證聯會 2008 年校長領導國家層級新標準及其對我國國民中小學校長培育制度之啓示。**教育研究與發展期刊，6**（1），199-228。[Huang, Z. N., Wu, C. S. (2010). A study on the national standards for educational leadership of ISLLC 2008 in the United States and its implication for preparation programs of the elementary and secondary school principals in Taiwan. *Journal of Educational Research and Development, 6*(1), 199-228.]

秦夢群（2007）：校長培育制度之趨勢分析：以英、美及新加坡爲例。**學校行政雙月刊，51**，1-18。[Chin, M. C. (2007). Administrative preparation of principals: Issues and perspectives in USA, Great Britain, and Singapore. *Journal of School Administration, 51*, 1-18.]

陳木金、李俊湖（2006）：**國民小學校長主任培訓模式之研究**（國家教育研究院籌備處專題研究案）。臺北：國家教育研究院籌備處。[Chen, M. J., & Lee, J. H. (2006). *Study of elementary school principals training models*. Taipei, National Academy for Educational Research]

新北市政府教育局（2010）：**新北市卓越學校指標**。新北市：新北市政府教育局。[Education Department in New Taipei City Government, *Excellent School Indices in New Taipei City.* Taipei, Taiwan: New Taipei City Education Bureau.]

謝名娟（2013）：以多層面 Rasch 分析的角度來評估標準設定之變異性。**教育心理學報，44**（4），793-811。[Hsieh, M. C. (2013). Evaluating the variability in standard setting using many faceted rasch model. *Bulletin of Educational Psychology, 44*(4), 793-811.]

謝名娟、謝如山（2013）：多層面 Rasch 模式在數學實作評量的應用。**教育心理學報，45**（1），1-18。[Hsieh, M. C., Hsieh, J. S. (2013). An application of many-facet rasch model to evaluate mathematics performance assessment. *Bulletin of Educational Psychology, 45*(1), 1-18.]

Aryadoust, V. (2015). Self- and peer-assessments of the oral presentations of first-year science students. *Educational Assessment, 20*(3), 199-225.

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly, 13*(1), 1-24.

Basturk R. (2008). Applying the many-facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment & Evaluation in Higher Education, 33*(4), 431-444.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). London, UK: Erlbaum.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training. Does it work? *Language Assessment Quarterly, 2*(3), 175-196.

Engelhard, G., Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Harasym, P. H., Woloschuk, W,. & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education, 13*, 617-632.

Hess, F. M., & Kelly, A. P. (2007). Learning to lead? What gets taught in principal preparation programs. *Teachers College Record, 109*, 244-274.

Jensen, K., & Harris, V. (1999). The public speaking portfolio. *Communication Education, 48*(3), 211-227.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145-164.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions, 12*(2), 636. Retrieved from http://www.rasch.org/rmt/rmt122m.htm

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2010). A users guide to Winsteps Rasch model computer program: Program manual 3.70. Chicago, IL: Winsteps.

Linacre, J. M. (2014). *Facets Rasch measurement* [computer program]. Chicago, IL: Winsteps.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

North, B. (2000). The Development of a common framework scale of language proficiency. New York, NY: Peter Lang.

NPQH (2016). *National professional qualification for headship* (NPQH). Retrieved from https://www.gov.uk/guidance/national-professional-qualification-for-headship-npqh

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes, 23*(1), 19-43.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Los Angeles, CA: Sage.

Sundrarajun, C., & Kiely, R. (2010). The oral presentation as a context for learning and assessment. *Innovation in Language Learning and Teaching, 4*(2), 101-117.

Tennant,A., & Pallant, J. (2006). Unidimensionality matters (a tale of two Smiths?). *Rasch Measurement Transactions, 20*, 1048-1051.

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

Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In j. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71-85). Los Angeles: Sage.

Wright B. D., & Masters G. N. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

收 稿 日 期：2016 年 06 月 02 日
一稿修訂日期：2016 年 07 月 21 日
二稿修訂日期：2016 年 07 月 26 日
接受刊登日期：2016 年 08 月 01 日

Bulletin of Educational Psychology, 2017, 48(4), 551-566
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Who is a Good Speaker? Applying Multifaceted Rasch model to analyze Principal Three-minute Impromptu Speech

Ming-Chuan Hsieh
Research Center for Testing and Assessment
National Academy for Educational Research

In the pre-service training program, impromptu speech is one of the most important ability to develop for school principals. To assess the speech ability, most researchers use the total or average of the raw scores as final scores. However, when there are many students attending the test, raters need to be divided into several groups to provide grading, and the severity of the raters for different groups may impact students' score. In this study, the scoring criteria, rater's severity and person's ability are all considered into the multifaceted Rasch model. The results show that, even for the trained raters, there still exist subjectivity and different level of rater severity. On the other hand, school principals feel most difficult in content, framework, proper usage of words and time control and feel relative easy in pronunciation, prosody, and appropriate manner in the context of impromptu speech. The ignorance of facets and use the raw total or average score as final score may cause bias or unfairness of the score ranking.

KEY WORDS: multifaceted Rasch model, school principals evaluation, speech