國立臺灣師範大學教育心理與輔導學系
教育心理學報，2020，52 卷，2 期，415–436 頁
http://doi.org/10.6251/BEP.202012_52(2).0008

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*

謝名娟
國家教育研究院
測驗及評量研究中心

牽涉到評分的情境下，常見三種評分資料蒐集的方式，第一種為完全評分網絡設計，在此設計下所有的層面的成分（components）有完整的觀測值。第二種為不完全評分網絡設計，成分間有部份程度系統性連結，第三種為不連接評分網絡設計，各成分之間沒有任何系統性的連結，即使這種評分網絡設計具有潛在性的問題，在台灣許多重要考試在成本考量下仍使用這樣的設計。本研究以儲訓校長的口語表現評分資料為實證數據，藉由多層面 Rasch 模式（many faceted rasch model，簡稱 MFRM）的分析模式來進行參數的等化估計，探討這三種不同評分者資料蒐集設計對於各層面參數估計的影響，其研究發現評分者連結性越小，參數估計的穩定性越差，尤其在不連接評分網絡設計，雖然使用 MFRM 進行校正，其相關參數估計與受試者的能力排序存在很大的誤差，考試單位應避免使用此設計進行評分者的分數評閱。建議未來重要的考試，應至少採用不完全評分網絡設計，並以統計模型（如 MFRM）的方式進行評分者嚴厲度的校正。

關鍵詞：多層面 Rasch 模式（MFRM）、校長口語評量、等化連結設計、評分嚴厲度

*1. 本文通訊作者：謝名娟，通訊方式：hm7523@hotmail.com。
2. 本論文感謝科技部經費補助（MOST 108-2410-H-656 -002）以及審查委員、編輯團隊的辛勞，謹致謝忱。

在新課綱素養導向的風潮下，素養導向的評量著重真實的情境與真實的問題，未來台灣的評量趨勢將由傳統的選擇型試題，走向更多開放性，甚至是多元的實作評量，這樣的評量除了在班級內評量實施會越來越多，在大型的重要考試也可能會逐步推行。實作評量的成效好壞，最重要的要素之一就是評分者的評分素質，過去與評分者效應相關的研究多聚焦在評分的嚴苛程度，而此效應可能影響分數的解釋與應用，另外，還有一些公平性的議題，例如不同的評分者評分仍可以得到相同的成績。過去 Campbell（1993）與 Johnson 等人（2009）著重在評分者訓練的過程、評分資料的準備與評分者應具備的資格與能力上，其中評分者的訓練為著重在了解評分規準與建立彼此對於學生在評量中的表現共識，但即使進行了評分訓練，評分嚴厲度仍然不容易消弭（O'Neill & Lunz, 2000）。

但評分者嚴厲度的問題要解決並不困難，只要所有的評分者都評閱所有學生的表現即可，即使有嚴苛的評分者與寬鬆的評分者，若其標準一視同仁，對於受試者成績公平性的影響並不大。然而，這種作法在現實考量上較不可行，試想若是一個實作評量需要耗時 10 分鐘來進行學生操作的觀察，假使要為 100 位學生要進行評分，就需 16 個小時的評分時間，對於評審的體力耐力都是一大考驗，因此許多大型測驗，採用多個評分者同時評分。學者為了解評分者嚴厲度的問題，使用了許多統計方法，例如 Braun（1988）使用了變異數的分析方法來調整評分者嚴厲度、Longford（1993, 1994）則嘗試了貝氏的方法與架構，Lance 等人（1994）、Raymond 與 Viswesvaran（1993）、Raymond 等人（1991）與 Wilson（1988）則使用了迴歸的模式來進行調整，Myford 與 Mislevy（1995）、Lunz 與 Suanthong（2011）則使用了多層面的 Rasch 模式（many faceted Rasch model，簡稱 MFRM），這些模式都大致使用了等化的原則來調整評分者的嚴厲度。主要目的為讓受試者彼此的成績之間具有可比性，透過統計的程序將所有受試者的能力放在同一個尺度中的校正方法（Kolen & Brennan, 2014）。在試題等化程序下，受試者的能力值會依據試題在不同題本的試題難度進行調整。類似的，在評分者的等化程序下，研究者要將受試者分配給不同的評分者，評分者會評閱部分相同的，也會有部分不同的學生，學生的能力估計也會依據評分者的不同嚴厲度進行調整。國內過去對於試題等化的研究著墨較深（吳慧珉等人，2015；王暄博等人，2013），評分者等化的相關研究較為少見。

評分者等化估計的準確度受資料蒐集設計的影響。根據 Engelhard（1997）的論述，在實作評量的情境中，牽涉到評分者與實作任務之間的連結性，實務工作者常使用三種資料蒐集的設計模式，包括完全評分網絡設計（complete assessment network）、不完全評分網絡設計（imcomplete assessment network）與不連接評分網絡設計（non-linked assessment networks）。其中在完全評分網絡設計下，受試者在所有的層面均有完整的觀測值，也稱為完全的交叉設計（completely crossed designs），這種評分設計是最簡單，但是也是最花成本的一種資料設計模式。在不完全評分網絡設計中，受試者並沒有完整的觀測值，但是在各成分（component）間有系統性相互涵蓋，讓各層面之間能建立起連結的網絡，在妥善的設計下，不完全評分網絡設計也可為一種具有信效度的設計，在成本上花費較少。不完全的網絡設計下，每個測驗的元素間均具有部份程度的連結性，例如平衡不完全區塊設計（balanced incomplete block design, BIB）與部份平衡不完全區塊設計（partially balanced incomplete block design, PBIB）都是常見的模式。最後一種為不連接評分網絡設計，這種設計受試者在各層面沒有完整的觀測值，且各成分之間也沒有任何系統性的連結設計，這樣的設計對於參數估計較為困難，然而，檢視現行在台灣的大型測驗，考試單位考量到施測成本，許多的重要大型考試採單閱、或是隨機分派的方式進行閱卷，如考選部（2019）在考試通用法規中寫出在國家的考試中 ，閱卷以單閱為原則，其單閱的意涵即為每位考生的考卷，僅有一位評分者進行評分，並以此評分之成績作為考生最終的考試成績，因此即使不連結的評分網絡具有潛在性的問題，但在台灣許多考試仍廣為使用這樣的評分設計。

MFRM 為過去在評分等化中常用的統計模型，此模型已廣泛應用於歐洲共同標準架構（Common European Framework of Reference for Languages；簡稱 CEFR；Council of Europe, 2001; North & Jones, 2009），特別是各種關於語文測驗或是實作測驗上，許多研究者均使用 MFRM 來進行等化設計下評分嚴厲度的調整（Palermo 等人，2019；Breton et al., 2008; North, 2000）。然而，使用這個模式下，評分者資料蒐集的設計模式，對於參數估計，特別是受試者的能力估計與排序造成甚麼影響，目前較缺乏相關的實徵研究。本研究則以儲訓校長的口語表現評分資料為實證數據，藉由 MFRM 分析模式，來探討完全評分網絡設計、不完全評分網絡設計與不連接評分網絡設計對於各層

面參數估計與受試者能力估計的影響，並依據本文之發現，提供教育現場評量實務與未來研究之參考。具體而言，有兩個主要的研究目的（1）不同的評分設計，對於各層面（受試者能力、評分者嚴苛度、評分項目的難度）的參數估計的影響，而評估的參數包括 MFRM 的模型中的適配度、分離度、信度、卡方檢定等。（2）不同設計間，受試者能力估計的相關性為何？與其對於受試者能力排序的影響幅度，亦為本研究之重點。

一、校長儲訓班與其口語評量評分相關問題

校長儲訓課程設計包含的面向為願景形塑、策略思考、創新經營、團隊合作、溝通協調、自我覺察（林信志、謝名娟，2016），並根據此六個面向設計不同的校長專業課程。而在素養導向評量的風潮下，也從過去以傳統紙筆評量式的期末測驗，改為多元的模式，包括報告、演講、籃中演練、小組討論、校園實習、與個案研究等，希冀能透過更廣泛的面向來完整評估儲訓校長在培訓班專業成長的能力。

其中口語表達的演講課程乃訓練之重點之一。儲訓校長的演講評分之面向乃參酌 Aryadoust（2015）之研究，指出演講應該包括語言溝通（發音、語調、語詞）、非語言表達（肢體表達、臉部表情、眼神與穿著等）、與內容組織（內容與架構）能力。依據儲訓校長的實際需求，配搭焦點座談，而形成符合儲訓校長需求的口語表達評估的面向，包括內容、架構、語詞、儀態、發音、語調、時間掌控等面向。

然而在進行儲訓校長演講的評分時，遇到以下幾點困難：

（1）評分者來源不固定。評分需要經驗的累積，越有評分經驗的評分者，可以評出較為穩定的成績，然而，在儲訓班的評分或是上課講師招募過程中，遇到人員流動的問題。儲訓班的課程為每年三到五月，此時正值學校上課的期間，要能每年找到固定的評分者並不容易。

（2）評分狀況不穩定。雖然研究團隊所邀請的都是在校長領域專精的學者或資深校長，但不見得是評分專家，雖提供評分規準、多次的評分訓練，但大多數的評分者具有主觀意見，需進行不斷的溝通才能進行評分調整，且根據謝名娟（2017）的分析指出，校長儲訓班在口語評量中仍存在評分者嚴厲度的問題，有些評分者評分較為寬鬆、有些則較為嚴苛，因此需要使用統計模型進行分數校整。

（3）評分安排的時間彈性較弱。部分縣市仍以儲訓的成績當作優先分發的依據，因此成績對於儲訓校長來說是相當重要的。因此評分講究公正、公平，本研究所著重的演講訓練，每年都需要更替題庫，曾發生過去的儲訓校長會留下所謂的考古題提供後期校長練習，因此每年在初始階段，都需要進行題目的撰寫，且評量時間必須各班盡量一致，否則可能會有跨班討論題庫所造成的評量誤差。

在此限制下，如何能夠消弭評分者的嚴厲度所造成的影響，計算出公平性的成績，為儲訓課程中的挑戰之一。

二、評分者評分等化原則與設計

評分化設計牽涉到以下幾個常用的名詞，在回顧評分者等化設計之前，先以 Engelhard（1997）之定義加以說明。首先，層面（facet）為評分網絡設計下所分出的不同向度，例如在口語評量上，不同的向度包括了評分者、評分規準、受試者能力，這些在多變量分析中也稱為因子（factors），不同的層面包括了個別的元素（elements），這些元素會有不同的難度（試題）或是嚴厲度（評分者），當兩個層面交織時（crossed），則形成了成分（components），例如可為評分者和評分任務，或是評分者和受試者兩層面交織下所組成的成分，這些成分的組合則成為了評分網絡（network）。

等化的過程分成兩個步驟（Kolen & Brennan, 2014）：數據收集與統計模型的轉換。在數據蒐集的階段，可能會與如何進行抽樣、與試題難度（或是評分者的嚴厲度）有相干，而這些被蒐集的數據會用於統計模型使用。統計模型的轉換則牽涉要如何在不同的試題難度（或評分者嚴厲度）進行受試者能力值的調整，這兩步驟之間息息相關，其中數據收集的品質好壞會影響後續的統計模型調整的準確性，Mislevy（1992）的甚至指出測驗建構的模式與等化好壞不可分開，當等化可行時，大多是因為有適合的測驗建構方式。

（一）等化原則

Wolfe（2004）認為在等化中有三個原則（1）對稱性（2）團體不變性（3）相等性。這些原則要素可見表 1。在試題等化的過程中，主要的流程是試題所建構的，其三個原則的意涵包括能使用題本 A 等化到題本 B，也能用題本 B 來等化題本 A 的成績，等化的結果不會因為採用特定學生的成績進行等化而有所不同，學生也不會使用不同題本進行等化而造成成績差異。而在評分者的等化過程中，試題的題本變成了評分者，其原則為不同評分者嚴厲度的調整不會依據特定的評分者有所不同，即使採用不同的學生評分樣本，等化結果也不受影響，與學生的成績不會因為不同的評分者來評分而造成成績差異。

表 1　對於試題等化與評分者等化的需求

需求	試題等化	評分者等化
對稱性	等化的方法在不同的題本中可以相互轉換，以 A 題本來等化到 B 題本，或是由 B 題本來等化到 A 題本結果都應相同。	評分者對於學生的評分具有交換性，對於不同評分者嚴厲度的調整不會依據特定的評分者有所不同。
團體不變性	等化的結果不管來自哪些受試者，其結果都應相同。	評分等化的結果，不管使用那些受試者的成績來進行等化都是相同的結果。
相等性	受試者即使接受不同的題本，所得到的能力估計值都相同。	即使不同的評分者來進行評分，受試者得到的成績都是相同的。就是不管受試者的運氣好壞，被誰評閱，經過等化後，所拿到的成績都應該是相同的。

註：本表乃改編自 “Exploring the Effects of Rater Linking Designs and Rater Fit on Achievement Estimates Within the Context of Music Performance Assessments,” by S. A. Wind, G. J. Engelhard, and B. Wesolowski, 2016, *Educational Assessment, 21(*4), p. 279.（https://doi.org/10.1080/10627197.2016.1236676）. Copyright 2020 by the Taylor and Francis Online.

（二）評分者評分等化設計

Engelhard（1997）指出一個具有良好的評分者等化設計，必須要有一個好的分析模式與系統性的資料蒐集。若要透過統計模型來調整評分嚴厲度，則評分資料蒐集必須經過設計，才能得到較佳的等化結果。其中 Engelhard 提出了三種主要用於評分等化設計的模式，而這些設計類似測驗試題的區塊設計模式，在測驗試題的等化中，將不同的測驗分數，透過區塊設計而進行試題難度的校正，使得學生的成績不會受所施測的試題內容難度和所在班級的能力分布影響其成績（Kolen & Brennan, 2014）。與此理念相同，在評分者等化設計的模式中。則將評分者進行區塊設計，將受試者所得到的受評成績，透過評分者的區塊設計而進行評分嚴厲度的校準，使得學生的成績不會因為評分者嚴厲度與所在班級同儕的能力而受影響。在模式適配的狀況下，MFRM 可以將評分者的嚴厲度、受試者的能力、與其他所需要考慮的面向都同時放進在模型中並進行等化。在資料蒐集模式中，有三種型式的評分網絡設計：

第一種為完全評分網絡設計（complete assessment network）。在此設計中，各層面的每一個成分都是相連的，例如有兩位評分者，200 位學生，三個開放性試題，在設計中，這 200 位學生每位都做了同樣的三個試題，且都有同樣的兩位評分者進行評分。在古典測驗理論中即為單一受試的設計（single group design），這種設計可以完整基本上是測驗中最直接且簡單的設計，其結果也是最公正的，然而在大型測驗中，這種設計具有其困難性，其所需要的時間與相關經費成本也是可觀的。

第二種則為不完全的評分網絡設計（incomplete assessment networks），這種評分設計有部分的成分具有系統性的相連。例如每一位學生至少有兩個評分者以上進行評分，而評分者所評的學生有部分相同與部分不同。例如若有四位評分者，100 位考生，則可以第一位評分者 1 ～ 40、第二位評 20 ～ 60、第三位評 40 ～ 80、第四位評 81 ～ 100 與 1 ～ 20，如此一來每位考生都有兩位評分者來評分，但是卻又彼此相連，一般大型測驗以時間與經費成本的考量下，這種設計較為可行。

第三種則為不連接的評分網絡設計（non-linked assessment networks），這種設計中，網絡中所

有的成分均不相連。例如各班老師只打所負責班級學生的成績，如 A 班老師只打 A 班學生的成績，B 班老師只打 B 班學生成績，其分數呈現巢狀（nested）的設計。在傳統的古典理論下，這種設計類似相同群組的設計（equivalent group design）。Engelhard（1997）指出，這種設計的品質需檢視隨機成分的符合程度，例如學生應隨機分派給不同的評分者來評分。

建構評分者之間的連結性是重要的（Eckes, 2011; Engelhard, 1997）。Wind 等人（2016）曾指出只有在所有層面充分連結下，才能使用統計模型的方式來進行不同評分者嚴厲度的校正。然而，雖然過去有許多的研究著重在試題方面的連結性，但針對評分者評分等化的相關實證研究較少，另外不同的評分等化設計的可能造成的誤差性也值得探討。

三、多層面 Rasch 模式的應用

試題反應理論常用於測驗編制與信效度分析（趙子揚等人，2016；Kuo et al., 2015），而多層面 Rasch 模式為試題反應理論下所延伸的一種模型，此模型將估計的參數轉為羅吉斯的對數型尺度（logit scale），在評估受試者的能力時，同時考量的層面為評分項目之難度與不同評分者的嚴厲度，根據 Linacre（1989）的研究指出，影響作答反應主要的因素包括測試者能力、試題難度、評分者嚴厲度與評分者本身的偏誤，而多層面 Rasch 模式可以適度透過統計的模式來進行調節這些面向影響。

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 et al., 2008; Wolfe & Dobria, 2008），在衡量評分者的評分行為時，著重評分一致性（consistency）與分數同意度（agreement）的問題，其中一致性代表受試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，也不應因為評分者的不同而有分數的落差（Stemler & Tsai, 2008; Wolfe, 2004）。然而在現實情境中，常常會有受試者的成績排序與分數會受到不同評分者評分而有差異。這樣的差異可能來自不同的因素，包括評分者的主觀經驗、作業的不同、評分規準等（Eckes, 2011）。雖然加強評分者的訓練可能有助於避免這樣的差異，但 Elder 等人（2005）指出，評分者訓練對於評分者之間的評分一致性的提升程度不高。

MFRM 可有效用來解決與評估評分可能遇到的問題，如林小慧與曾玉村（2017），林小慧等人（2018）使用 MFRM 來進行任務難度的分析，並依據評分者嚴厲度的調整，使科學的評測不受評分者變異程度的影響。謝名娟（2013）則將 MFRM 用在標準設定上，藉以偵測標準設定成員在設定切斷分數的變異性，被給予適當的回饋機制。謝如山與謝名娟（2013）則將 MFRM 應用在數學實作評量的評分上。另外姚漢禱與姚偉哲（2007）則將 MFRM 應用於體育測量中，張新立、吳舜丞（2008）則應用在學術研討會的論文評分上，Tseng 等人（2019）則將此模型檢視英文試題效度，這些均為 MFRM 的多元應用。

方法

本研究主要探討不同的評分者設計對於校長口語評量表現評估的影響，在本研究中，採用了四位評分者，為了能更了解不同的等化設計對於評分估計模型的影響，原本使用的評分數據將進行更動來檢視評分設計的影響，而其評分等化設計參考 Engelhard（1997）所提出的設計，包括完全網絡設計、不完全網絡設計（兩種）與不連接的評分網絡設計，本研究將檢視在使用 MFRM 分析模式下，這四種網絡設計所得之參數估計。

一、研究對象

共有四位評分者參與本研究，這四位均為具有博士學位的校長學專家，其中包括大學教授與資深的在職校長，其中大學教授均曾有講授過相關校長學理論與實務課程或進行過相關研究。

參與的儲訓校長共有 85 位，其中包括 70 位國小儲訓校長與 15 位國中儲訓校長，男性儲訓校長有 52 位，女性儲訓校長則為 33 位。儲訓班別分為 A、B、C、D 四班，每班人數分別為 24、24、22 與 15 位儲訓校長，其學員隨機分配至各班。每班的儲訓校長必須接受口語演講的訓練，共有三次相關課程，前兩次為請專家學者針對演講技巧進行說明，並請學員針對輔導校長所指定的題

目進行演練，第三次使用研究團隊所設計的題目，進行抽題演練，題庫袋共有 50 個題目供學員現場抽題，學員抽過的題目則不放回題庫袋，每位學員有三分鐘來準備演講內容，並使用三分鐘進行演講，研究者在開訓前至班上宣導研究相關內容，並請每位學員簽署研究者知情同意書，同意後方進行錄音錄影，以供後續研究使用，參加儲訓的 85 位校長經溝通後全部簽署研究同意書。

二、評量工具

評分者觀看每位學員的錄影表現後，針對七個向度進行評分，包括內容、架構、語詞、儀態、發音、語調、與時間掌控。其中重點包括演說內容必須符合演說題目、演說的架構須清楚、並針對演講對象使用適切的用語。儀表態度需恰當（如在處理校園危機事件時，不能面帶微笑）。表達技巧則包括發音清楚與語調流暢有起伏，最後則為時間掌控得宜，能在指定的三分鐘內完成演說。其分數等級為 1 ～ 3，其中 1 代表未達基礎，2 代表基礎，3 則代表精熟（如表 2），此評分表乃改編自謝名娟（2017）的校長口語評分研究，然本研究再將未達基礎、基礎與精熟的表現標準水平定義加以描述，以讓評分者更了解評分規準的要求。

表 2　口語評量評分規準

審查重點	未達基礎 1	基礎 2	精熟 3
演說內容符合主旨（內容）	1. 內容與主題不符 2. 內容陳述不清或無新意 3. 離題或內容缺乏系統性	1. 內容大致恰當 2. 內容大致清楚或有新意 3. 略有離題	1. 非常符合主旨與現況 2. 內容很清楚或有新意 3. 內容具有系統性。
架構分明（架構）	1. 結構不完整 2. 開頭、論述、結尾之間無法串接 3. 開頭、論述、結尾之間比例不佳	1. 結構明確聽眾能分辨 2. 開頭、論述、結尾之間大致串接 3. 開頭、論述、結尾之間比例尚可	1. 結構明確易懂好記 2. 開頭、論述、結尾之間串接流暢 3. 開頭、論述、結尾之間比例佳
使用適當語詞、用語準確傳達意思（語詞）	1. 語詞無法明確闡述內容 2. 遣詞不甚通順 3. 不夠口語化	1. 語詞能夠明確闡述內容 2. 遣詞大致通順 3. 口語化，聽眾可理解	1. 語詞明確、豐富生動 2. 遣詞用字清晰易懂，能引聽眾專注 3. 關於細節的描述清楚深刻
儀表態度得體（儀態）	1. 穿著打扮較不符合講者身分 2. 肢體語言不適當 3. 表情無變化或不恰當	1. 穿著打扮合宜搭配演講身分與內容 2. 肢體語言大致適當 3. 表情略有變化，但不明確。	1. 穿著打扮合宜搭配講者身分內容 2. 肢體語言適當 3. 表情豐富且符合內容
發音清楚（發音）	咬字略不清楚，影響聽眾理解	咬字大致清楚	咬字清楚易懂
語調音量適切、語句流暢（語調）	1. 語調無起伏 2. 講者音量過大過小或運用麥克風不當 3. 講話語速不當	1. 語調稍有變化 2. 音量大致恰當 3. 語速大致恰當	1. 語調得宜有抑揚頓挫 2. 音量大小隨重點變化 3. 語速快慢隨內容調整
時間控制得宜、結尾不匆促（時間掌控）	1. 時間少於兩分半或超過三分半 2. 陳述不清、草草結尾	1. 時間略少於兩分半或略超過三分半 2. 大致有結尾	1. 時間控制精準（介於 2 分 50 秒到 3 分 10 秒） 2. 具有清楚的結尾

三、評分訓練

這四位評分者在進行正式評分前，均先進行評分者訓練。評分訓練目的為和評分者溝通評分規

準的定義，與各評分等級的規範，以求評分成績具有最佳的信效度。評分訓練模式為統一進行，在會議中讓評分者完成所有的訓練與練習，首先，讓評分者了解儲訓校長在抽籤時，所使用的口語評量的題目、實作時的標準作業流程，與每一個等級的定義與標準，接下來會給評分者參考過去幾期儲訓校長的表現資料作為訓練資料。

研究團隊提供各等級的定錨影片（anchor sets）讓評分者進行討論，透過這些定錨影片能讓評分者能更了解評分規準與各評分等級，每個影片中儲訓校長所表現在各向度的優缺點，會逐一討論，並請評分者判斷應該在此表現向度下，屬於何種等級。另外，訓練材料亦包括一些較特殊的受試者表現，例如在發音部分，原先的評分規準期盼儲訓校長要能發音清楚，但在現場實務中，遇到許多原住民口音的校長，雖然發音不夠清楚，但卻有個人魅力，或是夾雜了方言（如客語、閩語）的演講來貼近演說的情境，評分者會依據這些特殊的情況予以討論並決定在正式評分時，應如何進行分數的判斷。

討論完定錨影片後，則進入評分練習。每位評分者進行兩階段的評分練習，第一階段五位受試者影片，評分後由討論並提供回饋，第二階段則再評分五位，評閱後再一起討論回饋。透過兩階段的練習後評分者已大致達成共識。

四、評分設計

每位評分者均需針對所有的學員進行評分，因此原先的評量設計為完全評分網絡設計，層面包括評分者、儲訓校長的口語能力、與評分規準向度。

表 3　本研究所使用四種評分網絡設計

	設計一：完全評分網絡設計（每位學員有四位評分者）				設計二：不完全評分網絡設計（每位學員有三位評分者）			
評分者／受評班級	A	B	C	D	A	B	C	D
Expert1	X	X	X	X	X	X	X	
Expert2	X	X	X	X		X	X	X
Expert3	X	X	X	X	X		X	X
Expert4	X	X	X	X	X	X		X
	設計三：不完全評分網絡設計（每位學員有兩位評分者）				設計四：不連接評分網絡設計（每位學員有一位評分者）			
評分者／受評班級	A	B	C	D	A	B	C	D
Expert1	X	X			X			
Expert2		X	X			X		
Expert3			X	X			X	
Expert4	X			X				X

註：X 代表此班級有評分者進行評分。

每個評分者都需針對每位儲訓校長的口語能力在七個評分規準向度中進行評分。為了能夠評估研究問題，原先的數據進行調整，這些評分設計如表 3。其中設計一為最原先的設計，所有 4 位評分者都進行所有儲訓校長的評分，為完全評分網絡設計。設計二，每位學員有三位評分者的成績，例如在 A 班，有 1，3，4 號評分者來進行評閱，B 班則為 1，2，4 號評分者來評分。設計三則在四位評分者中，有其中兩位評分者的成績，如 A 班有 1，4 號評分者，B 班有 1，2 號評分者來評閱。但設計二與三評分者有跨班評分以產生連結性，這兩種設計為不完全評分網絡設計。設計四則為不連接的評分網絡設計，在此設計下，每位評分者各自評閱所被分配的班級，如 1 號評分者評閱 A 班，2 號評閱 B 班，彼此的評分成績沒有連結性。

五、數據分析

本研究以 MFRM 的分析模式來進行四種評分設計的分析模型，MFRM 為 Rasch 家族的一種統計估算模式（Linacre, 1994），在估計儲訓校長的口說能力時，將評分項目之難度與評分者的嚴厲度同時考量在模型中。

其中

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \beta_i - \gamma_j - \tau_k$$

P_{nijk} 所代表第 n 位儲訓校長，在評分向度 i，被第 j 位評分者表現評定為 k 時的可能性

$P_{nijk\text{-}1}$ 所代表第 n 位儲訓校長，在評分向度 i，被第 j 位評分者表現評定為 k-1 時的可能性

θ_n 為第 n 位儲訓校長的表現能力值（achievement）

β_i 為第 i 個評分向度的難度值（difficulty）

γ_j 為第 j 個評分者的嚴厲度（severity）

τ_k 為評定 k 與 k-1 的難度階（threshold difficulty）

當應用 MFRM 模式時，每個層面的參數均須在同一個模式下進行估算。而對於能力值來說，越高的羅吉斯參數代表受試者有較高的表現，越低則代表受試者表現較差。而對於評分向度 β 而言，較高的參數值代表該評分向度對於受試者而言是較為困難的，較低則代表該向度較為簡單。而對於評分者言，越高的參數值代表其評分越嚴苛，越低則代表較寬鬆。而並非層面，而是在評定一個等級到下一個等級之間的難度界限值。模式在估算出所有的層面的相對位置時，亦有其他的指標需要進行檢核。例如在估算每個參數的標準誤（standard error），分離係數（separation statistics）可以用來區分在常態分配的母群中，能區分出幾個具有統計顯著性差異的類群（strata）、數值越接近 0，代表評分者的嚴厲度越接近、考生的能力也越接近，而卡方則進一步針對這些差異進行顯著性的檢驗。信度（reliability）則可看出資料的穩定度，若信度值越接近 1，代表越能分辨考生的不同程度，然而，Park（2004）指出對於評分者面向，高信度值代表不同評分者有不同的評分嚴厲度，因此對評分者面向而言，低信度值反而較為適切。模式適配度（model data fit）則可以看出模式的期望值與觀測值之間的差距。大多數在 Rasch 的分析模式下都使用 Infit（the information-weighted mean-square fit statistic，簡稱 Infit）與 outfit（the outlier sensitive, mean-square fit statistics，簡稱 Outfit）的指標來探測模型和觀測值之間的差距（Smith, 2000），當許多參數的適配值小於 1 時，代表估計多仰賴在可得的資料上，若適配值多數大於 1 時，代表變異性過大，反應過於複雜（Linacre, 2005）。而 McNamara（1996）則說明當 infit 與 outfit 的均方值介於 0.7 到 1.3 之間時，代表適配度佳，使用 MFRM 的模型來分析是適切的。但是若是均方值大於 1.3，則代表估計中有許多干擾因素造成模型的不穩定估計，小於 0.7 則代表資料的獨立性不佳。Linacre（1998, 2010）、Smith（2002）和 Tennant 與 Pallant（2006）的研究指出，當模式的適配度在期望的區間內時，也可代表此模型符合單項度的假設。

Wolfe（2004）指出在 Rasch 模式進行等化時，會將學生的能力估計、評分者的嚴厲度、還有其他可能研究者有興趣的層面一起放進模型同時估算（concurrent calibration）當資料中有足夠的連結性時，相關的參數會進行校準。

本研究使用 FACETS （version 3.82.2, Linacre, 2019）來進行 MFRM 的參數估計，FACETS 為多層面 RASCH 模式最常使用的分析軟體之一，相關的資訊可參考網站 www.winsteps.com。

結果

本研究透過 MFRM 的分析模式，來呈現四種評分者等化設計下的成果：包括（1）四種模式下所產生出的參數估計與適配度指標；（2）不同設計下所產生的儲訓校長口語能力的結果相關程度與排序比較。

一、參數估計與適配度指標

四種評分等化設計的統計摘要如表 4，本研究為三個層面的 MFRM 模式，包括評分項目的難度、評分者的嚴厲度與受試者的能力表現估計，分析使用了 FACET 軟體，參考張新立與吳舜丞（2008）和謝名娟（2017）的作法，將評分項目的難度平均值定在 0，來檢視評分者的評分表現。如表 3，評分委員的參數估計 logit 估計值均為負值，代表評分委員給分寬鬆，而受試者的表現優於評審項目所設定的難度，此部分結果和過去的研究相符合（謝名娟，2017）。在此表中提供了受試者、評分者、評分項目的各項在 MFRM 分析後所呈現的統計數據包括參數估計值（logit scale measure）、適配度（infit 與 outfit）、分離度（separation）、可靠度（reliability）與卡方值等。Infit 與 outfit 的均方值在四種設計都接近 1，但在設計三與四的 infit 值在受試者能力估計時的標準差較大，代表在此評分設計中有部分受試者能力估計有模式適配的問題，資料中存在部分適配度大於 1.3，代表數據中有些干擾與不穩定性，而適配度小於 0.7 則代表資料彼此的獨立性略嫌不足。而信度越接近 1 代表資料越能區分受試者能力，而在這四個模式中，受試者的信度分布在 0.82 到 0.72 間，而模式四在評分者與評分項目的估計為 0.15 與 0.31，代表不同評分者具有類似的評分嚴謹度，而不同的評分項目的難度也類似。Facet 的分析中亦提供了分離度指標，這指標假設所有觀測值從同一個常態母體中隨機取出，而從這個常態母群中，能顯著區分出幾個具有顯出差異的類群（strata），從設計一中關於受試者的分離度有 2.16，代表 85 位學員至少能分出 2 個具有統計顯著差異性的群體。而對於評分者來說，彼此也具有顯著的差異性。卡方檢定的數值達顯著性也進一步的說明這些評分者是在評分時有顯著性的嚴厲度不同。但在設計四的不連接網絡設計，評分者與評分項目的信度值較低，且卡方檢定下已不具顯著性，和設計一的估計有明顯的落差。

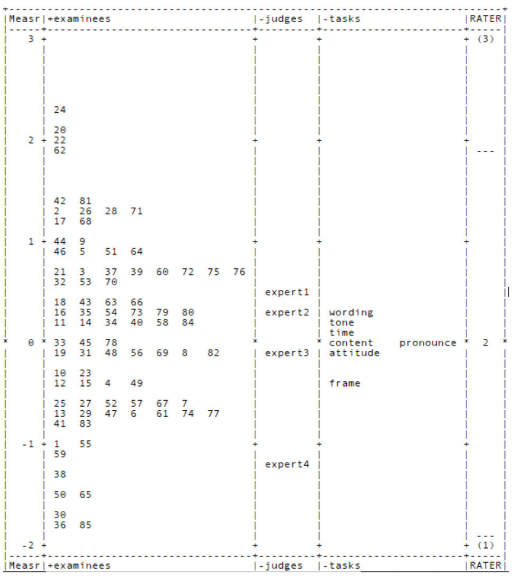
表 4　四種評分者等化設計的統計摘要

	設計一			設計二			設計三			設計四		
統計摘要	受試者	評分者	評分項目	受試者	評分者	評分項目	受試者	評分者	評分項目	受試者	評分者	評分項目
參數值												
M	0.12	–0.13	0	0.11	–0.13	0	0.13	–0.15	0	0.05	–0.05	0
SD	0.90	0.75	0.24	0.93	0.86	0.23	1.24	0.86	0.26	1.61	0.20	0.29
N	85	4	7	85	4	7	85	4	7	85	4	7
Infit												
M	1	1	1	1	1	1	1	1	1	0.98	0.99	0.97
SD	0.26	0.17	0.15	0.32	0.13	0.13	0.44	0.15	0.18	0.67	0.13	0.34
Outfit												
M	1	1	1	1	1	1	1.02	1.02	1.02	0.97	0.98	0.97
SD	0.27	0.18	0.15	0.32	0.14	0.14	0.49	0.17	0.24	0.68	0.15	0.37
分離度	2.16	9.08	1.98	1.86	8.89	1.49	1.94	6.91*	1.28	1.59	0.41	0.67
信度	0.82	0.99	0.80	0.78	0.99	0.69	0.79	0.98	0.62	0.72	0.15	0.31
χ^2	480*	251*	29.3*	371*	19.3*	19.3*	384*	157.4*	15.8*	321.*	2.90	8.70
Df	84	3	6	84	6	6	84	3	6	84	3	6

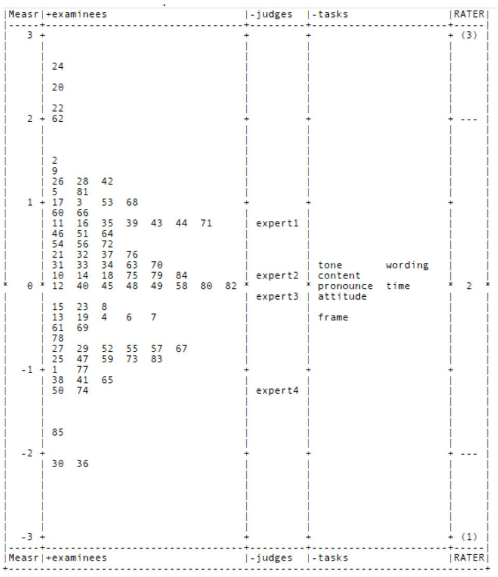
*$p < .05$.

圖 1 為針對這四種設計所繪製的變數分布圖，其分布圖可用來檢視受試者能力、評分者的嚴厲度與評分項目的相對分布。

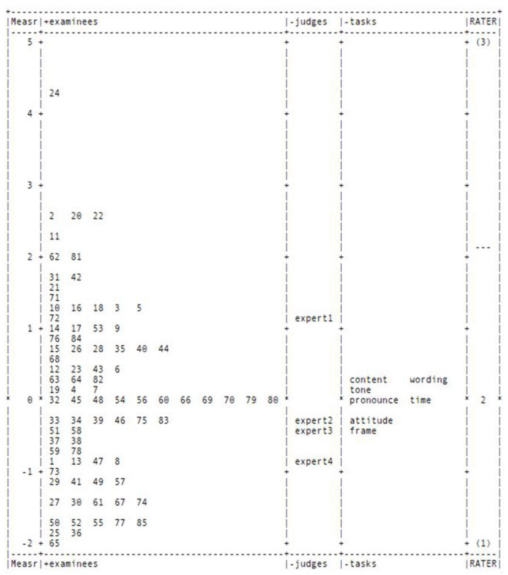
設計一： 完全評分網絡設計（每位學員有四位評分者）



設計二：不完全評分網絡設計（每位學員有三位評分者）



設計三：不完全評分網絡設計（每位學員有兩位評分者）



設計四：不連接評分網絡設計（每位學員有一位評分者）

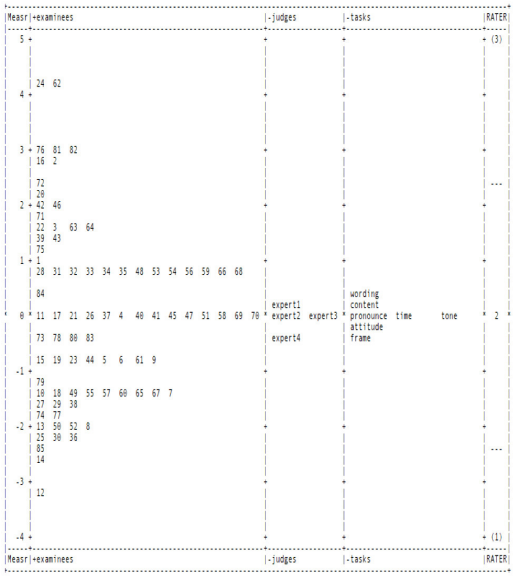


圖 1　變數分布圖

註：Measr 為參數的刻度，examinees 為學員的能力值的分布位置，judges 為評分者的嚴厲度的分布位置，tasks 則為評分向度的難度分布。

圖 1 中第一個欄位為參數的刻度，以 logit 為單位，而每個層面的刻度相對差距也可以從這個刻

度中來進行檢視。第二個欄位為 85 位學員的相對位置排序分布，代號越上面的受試者，代表其估計值越高，能力越強。由圖 2 可看出在排序上面，設計一和設計二的排名還算接近，但是和設計三和四有所落差，例如在設計四中的程度最差的第 12 號受試者，若使用其他模式進行估計，應該是屬於中間的名次。第三個欄位為評分者的嚴厲度排序，越上面代表在評分時越嚴苛，在模式一、二、三的排序上都相同，評審 1 號最嚴格、2 號、3 號次之，而最寬鬆的是 4 號，但是模式四則顯示 2 號 3 號的審查嚴厲度大致相同。第四欄位則為評分向度的難易度，越上面代表對學員而言是較困難的向度，越下面則代表越簡單，從四個模式的估計中，大致有共同的趨勢，使用適當的語詞對於儲訓校長來說是較為困難的，而具有架構性則是最簡單的一個向度，這可能是因為在課程中有上過演說的相關課程，因此儲訓校長對於架構的掌握性較足，整體而言，不同的評分設計對於受試者的能力值具有相當的影響力。

二、受試者能力估計的相關性

以下檢視不同的連結設計對於受試者表現估計的影響，設計一所估計出來的能力值，分別與設計二、三、四來進行相關分析。其結果分析如圖 2-4，整體而言，設計二、三、四的能力估計值與設計一均呈現正相關，其能力的散布圖均由左下到右上，但可看出當數據之間的連結性越弱時，相關性就越小。

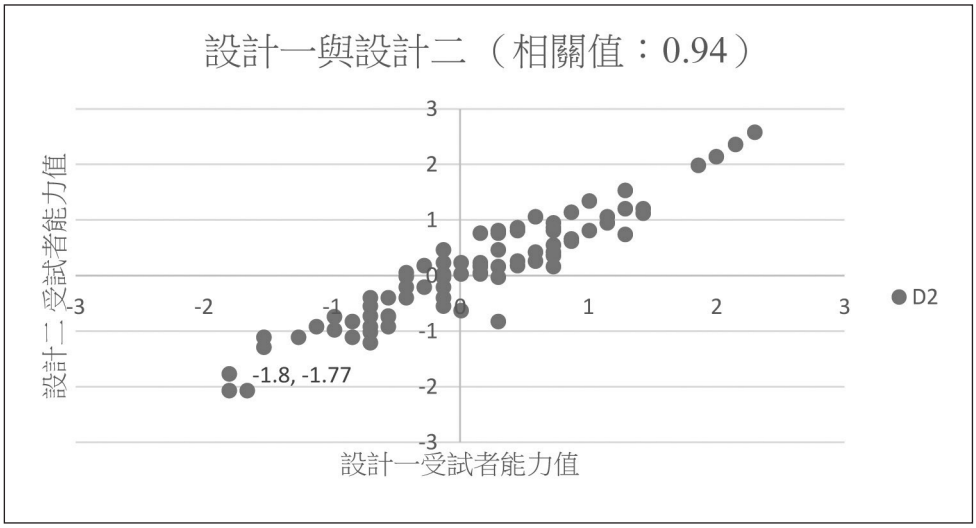


圖 2　設計一與設計二之間能力值相關分布圖

圖 5 則呈現 85 位成員在 MFRM 模型中能力值估算時，設計一與其他三種設計的差距。其橫坐標為設計一由高到低的排列，最左邊代表是在設計一中，經 MFRM 能力估算後能力最高的受試者，而最右邊則代表能力最低的受試者。這些數值越接近零值代表差距越小，正值代表設計二、三或四經 MFRM 的受試者能力估計值比設計一估計值來的高，負值則代表設計二、三或四設計比設計一估算的能力值來的低。從圖 4 中可看出設計二與三在能力值的估算上與設計一大致接近，但設計三在高分組與設計一的估計差距較大，而設計四則在高分組與低分組的有多位學員的估計有很大誤差，經檢視其最大差異正值差距達到 3.09，負值達到 -2.77。

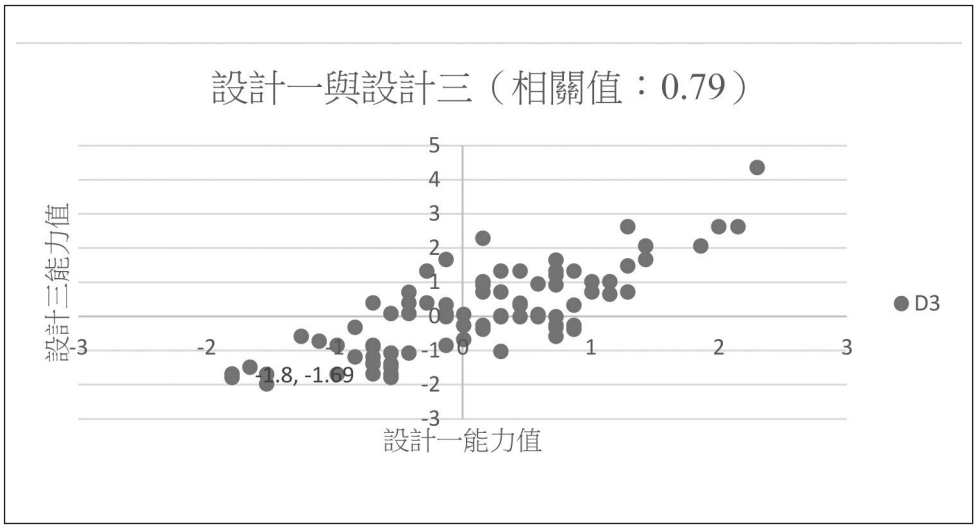


圖 3　設計一與設計三之間能力值相關分布圖

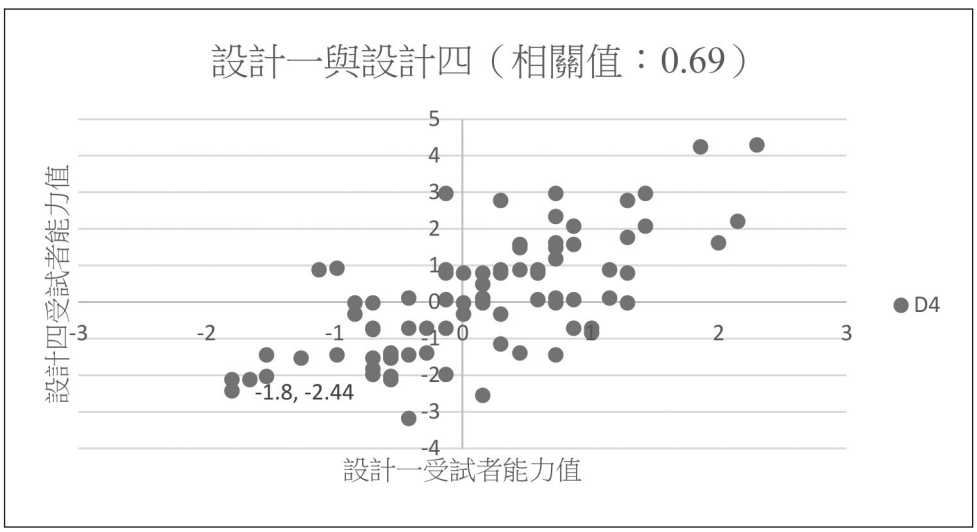


圖 4　設計一與設計四之間能力值相關分布圖

MFRM 的模式可以將評分者嚴厲度的干擾一併在模式中進行評估，進而能展現受試者的能力差異，而在儲訓校長的情境中，其成績的高低可以做為結訓成績之依據。然而，經由 MFRM 所估計之受試者能力值排序後發現，不同的評分者等化設計在排序上都有相當程度的差異。表 5 則呈現四種設計中的受試者的排序比較，其中受試者代號的排序乃根據設計一的能力值順序之前 10 名、中間 10 名與後 10 名來進行檢視。

在四種設計中，受試者 24 在四種設計中的能力估算值雖有不同，都是排名第 1，其中設計 2 與設計一的排名差距，大致小於設計三，設計三又大致小於設計四。然而卻可看到設計四即使在經過 MFRM 的估算後，其排序和設計一仍有相當大的不同，在 85 位學員中，受試者代號 59 的學員，在設計一中排序是第 79，但是在設計四居然是第 21 名，其排序差距有 58 名之多，因此在不連結的評

分者設計下，對受試者的排名估計會有較大的誤差性。

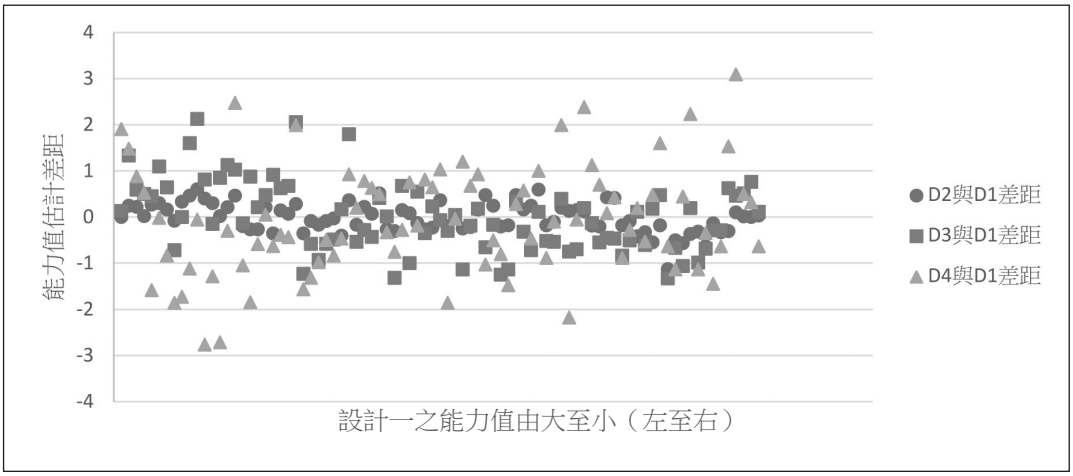


圖 5　設計一與其他三種不同設計間能力值的估計差距

表 5　MFRM 在四種設計的能力值前中後排序比較

代號	設計一		設計二			設計三			設計四		
	能力值	名次（1）	能力值	名次（2）	（2）-（1）	能力值	名次（3）	（3）-（1）	能力值	名次（4）	（4）-（1）
前十名											
24	2.30	1	2.58	1	0	4.35	1	0	4.29	1	0
20	2.15	2	2.36	2	0	2.62	2	0	2.20	9	7
22	2.00	3	2.14	3	0	2.62	2	-1	1.61	13	10
62	1.86	4	1.98	4	0	2.05	6	2	4.24	2	-2
24	2.30	1	2.58	1	0	4.35	1	0	4.29	1	0
20	2.15	2	2.36	2	0	2.62	2	0	2.20	9	7
22	2.00	3	2.14	3	0	2.62	2	-1	1.61	13	10
62	1.86	4	1.98	4	0	2.05	6	2	4.24	2	-2
42	1.43	5	1.20	7	2	1.66	8	3	2.07	10	5
81	1.43	5	1.12	11	6	2.05	6	1	2.96	3	-2
2	1.29	7	1.53	5	-2	2.62	2	-5	2.77	6	-1
26	1.29	7	1.20	7	0	0.71	24	17	-0.03	43	36
28	1.29	7	1.20	7	0	0.71	24	17	0.79	27	20
71	1.29	7	0.74	24	17	1.47	11	4	1.76	12	5
24	2.30	1	2.58	1	0	4.35	1	0	4.29	1	0

（續）

表 5 （續）

代號	設計一		設計二			設計三			設計四		
	能力值	名次（1）	能力值	名次（2）	（2）-（1）	能力值	名次（3）	（3）-（1）	能力值	名次（4）	（4）-（1）
20	2.15	2	2.36	2	0	2.62	2	0	2.20	9	7
22	2.00	3	2.14	3	0	2.62	2	-1	1.61	13	10
62	1.86	4	1.98	4	0	2.05	6	2	4.24	2	-2
42	1.43	5	1.20	7	2	1.66	8	3	2.07	10	5
81	1.43	5	1.12	11	6	2.05	6	1	2.96	3	-2
2	1.29	7	1.53	5	-2	2.62	2	-5	2.77	6	-1
26	1.29	7	1.20	7	0	0.71	24	17	-0.03	43	36
28	1.29	7	1.20	7	0	0.71	24	17	0.79	27	20
71	1.29	7	0.74	24	17	1.47	11	4	1.76	12	5
中間 10 名											
17	1.15	11	0.95	14	3	1.01	18	7	0.10	35	24
32	0.59	27	0.42	31	4	0.05	41	14	0.79	27	0
53	0.59	27	1.06	12	-15	0.94	21	-6	0.87	21	-6
70	0.59	27	0.26	35	8	-0.02	46	19	0.06	39	12
18	0.45	30	0.18	40	10	1.32	12	-18	-1.4	62	32
43	0.45	30	0.81	18	-12	0.38	34	4	1.48	17	-13
63	0.45	30	0.26	35	5	0.32	36	6	1.57	15	-15
66	0.45	30	0.86	16	-14	-0.02	46	16	0.87	21	-9
16	0.30	34	0.76	22	-12	1.32	12	-22	2.77	6	-28
35	0.30	34	0.81	18	-16	0.71	24	-10	0.79	27	-7
後十名											
83	-0.84	75	-0.83	71	-4	-0.33	56	-19	-0.34	49	-26
1	-0.98	77	-0.98	76	-1	-0.85	64	-13	0.92	20	-57
55	-0.98	77	-0.74	67	-10	-1.70	80	3	-1.45	65	-12
59	-1.12	79	-0.92	73	-6	-0.73	63	-16	0.87	21	-58
38	-1.26	80	-1.11	78	-2	-0.59	60	-20	-1.54	71	-9
50	-1.53	81	-1.29	82	1	-1.70	80	-1	-2.04	78	-3
65	-1.53	81	-1.11	78	-3	-1.98	85	4	-1.45	65	-16
30	-1.66	83	-2.07	84	1	-1.49	76	-7	-2.13	80	-3
36	-1.80	84	-2.07	84	0	-1.79	83	-1	-2.13	80	-4
85	-1.80	84	-1.77	83	-1	-1.69	78	-6	-2.44	83	-1

討論

本研究使用儲訓校長的口語評量實證數據，來說明不同的評分者連結設計對於適配性與受試者能力估計的影響。在本研究中共有三個層面與四種評分者的網絡設計，三個層面包括受試者的能力、評分項目與評分者的嚴厲度，有四種網絡設計包括從各層面連結性最強的完全網絡設計，到具有部分連接的不完全網絡設計，到無連接性的不連接網絡設計，本研究之結論與建議如下：

一、結論

（一）本研究所使用的實徵數據，大致符合模式適配度的預期

本研究所採用的四種設計模式，在 infit 與 outfit 的適配度指標都接近 1，代表使用 MFRM 的模式來進行相關數據的分析是可行的。而檢視其可靠度與分離度等指標，在設計一、二、三在可接受的範圍，但在設計四中，對於評分者與評分項目的可靠度指標與分離度指標較低，且未達顯著。

（二）評分者連結性越小，其能力值估計的穩定性越差、排序也不穩定

就研究結果而言，完全連接的網絡設計，提供各層面（受試者、評分者與評分項目）最強的連接性，而此種設計也是評分設計中最理想的情境，但此種設計最花時間與成本，因此在大型測驗中很難進行這樣的設計。退而求其次來說，第二、三種不完全的評分設計在大型測驗中是可行的，即建立評分者與評分者之間部分受試者的評閱重疊性，從分析結果可看出，連接性越低則與設計一理想的狀況會落差越大，例如受試者能力的估算上，不連結設計的能力估算相關性僅為 0.69，而在受試者的成績排序上，也可以看到與理想的完全連接網絡設計有明顯的落差。在評分者之間完全沒有任何連結性的設計四情境中，使用 MFRM 的統計模型進行校正，能力值估算與成績排序上，均會有較大的誤差現象產生。

二、建議

（一）涉及評分的測驗應採用具連結性的設計模式，並以統計模型進行評分者的嚴厲度校正

檢視現行在台灣的大型測驗，不僅很少使用統計模型進行校正，甚至許多為不具連結性的評分者設計。大型測驗只要牽涉到非選題或是作文，考試單位考量到評分成本，許多評分採單閱、或是隨機分派的方式進行閱卷，如考選部的閱卷以單閱為原則（考選部，2019），大考中心（2019）雖在非選題部分採雙閱，但閱卷乃採隨機分派，並沒有特別進行評分者的連結設計，而在評分也僅以檢視一、二閱之評分差距是否過大，大多評閱的成績則直接採用平均，並沒有進行不同評分者的分數等化與校準。然而，在這種大型的國家考試，動輒數十萬考生的情況下，動員的評分者接近百人，而過去的研究指出評分者即使受過訓練，其嚴厲度的調整仍然有限，因此，極有可能部分考生“剛好”運氣不好，遇上嚴苛的評分老師，造成分數的低落，甚至由於這一兩分的差距，造成數個志願的差距，然而，這種誤差並非在公平的考試制度應該出現的。因此建議未來在重要的考試，相關專責單位應在評分者派卷上有更完整的規劃，至少應如本文中採不完全的評分等化設計，而考生的成績在進行輸入後，應先以統計模型進行評分者的分數等化，再計算最終的成績，以避免受試者的能力估算值誤差太大。

（二）本研究採實證數據來說明不同的評分設計對於受試者能力估算的影響，未來可以考量使用模擬研究來探討其偏誤程度

在本研究中，主要採用了儲訓校長的口語評量資料，來說明不同的評分者設計對於受試者能力估計的影響，未來可考量使用模擬設計，來進行不同程度連結性比較，例如評分者連結的程度應該要到達多少的重疊性，才能達到理想的估計結果，如評分者與評分者至少應有重疊到三分之一或是四分之一的評分人數，在 MFRM 中的估計才夠準確。另外，還有不同的實驗設計也值得探討，例如受試者若與實作任務間具有巢套（nested）的關係，這種任務與受試者層面間具有巢套的關係，是否能透過試題的難度來調節能力值與評分者嚴厲度，而這些更為複雜的評分設計會對等化估計會造成甚麼影響，也值得未來進行深入的研究。

參考文獻

大考中心（2018）：107 學年度指考非選擇題評分標準說明。**選才電子報，288**。取自大考中心網站：https://www.ceec.edu.tw/xcepaper/cont?xsmsid=0J066588036013658199&sid=0J149511655048005583，2018 年 4 月 15 日。[College Entrance Examination Center. (2018). Construct response question rating standard description for 2018 school year. *CEEC E-paper*, *288.* Retrieved April, 15, 2018, from https://wwwc.moex.gov.tw/main/ExamLaws/wfrmExamLaws.aspx?kind=1&menu_id=318&laws_id=11]

王暄博、郭伯臣、呂玉如（2013）：大型測驗等化群體不變性之探究——以 2007 年臺灣學生學習成就評量資料庫國中二年級數學科為例。**測驗學刊，60**（3），489–518。[Wang, H.-P., Kuo, B.-C., & Lu, Y.-J. (2013). Exploring the population invariance of equating in the large-scale assessments: Using the Taiwan Assessment of Student Achievement as an example. *Psychological Testing, 60*(3)，489–518.]

吳慧珉、郭伯臣、許天維、陳婉寧（2015）：以可能值方法為基礎之多向度能力值垂直等化探究。**測驗學刊，62**（2），95–126。[Wu, H.-M., Kuo, B.-C., Sheu, T.-W., & Chen, W.-N. (2015). The Research in Estimating Multidimensional Traits under Vertical Equating Based on Plausible Value Method. *Psychological Testing, 62*(2), 95–126.]

林小慧、曾玉村（2017）：科學多重文本閱讀理解評量及規準之建構與信效度分析—以氣候變遷與三峽大壩之間的關係題本為例。**教育心理學報，49**（2），215–241。http://doi.org/10.6251/BEP.2017-49(2).0003 [Lin, H.-H., & Tzeng, Y.-T. (2017). Developing and validating a scientific multi-text reading comprehension assessment: Evidence from texts describing relationships between climate changes and the Three Gorges Dam. *Bulletin of Educational Psychology, 49*(2), 215–241. http://doi.org/10.6251/BEP.2017-49(2).0003]

林小慧、林世華、吳心楷（2018）：科學能力的建構反應評量之發展與信效度分析：以自然科光學為例。**教育科學研究期刊，63**（1），173–205。http://doi.org/10.6209/JORIES.2018.63(1).06 [Lin, H.-H., Lin, S.-H., & Wu, H.-K. (2018). Developing and validating a constructed-response assessment of scientific abilities: A case of the optics unit. *Journal of Research in Education Sciences, 63*(1), 173–205. http://doi.org/10.6209/JORIES.2018.63(1).06]

林信志、謝名娟（2016）：**中小學候用校長培訓混成課程模式與情境評量之發展與研究**。科技部專題研究計劃成果報告（編號：105-2410-H-656-002-MY2）。引自網站：https://rh.naer.edu.tw/cgi-bin/gs32/gsweb.cgi?o=dirproject&s=id=%22RP000000000714%22.&searchmode=basic [Lin, H.-C., & Hsieh, M.-C. (2016). *A study of competency indicators and assessment center for school candidate principals*. Ministry of Science and Technology (MOST 105-2410-H-656-002-MY2). https://rh.naer.edu.tw/cgi-bin/gs32/gsweb.cgi?o=dirproject&s=id=%22RP000000000714%22.&searchmode=basic]

考選部（2019）：**典試、監試、考試通用法規閱卷規則**。取自考選部網站：https://wwwc.moex.gov.tw/main/ExamLaws/wfrmExamLaws.aspx?kind=1&menu_id=318&laws_id=11，2019 年 4 月 15

日。[Ministry of Examination. (2019). *General rules and examination rules for the test, supervision and examination*. Retrieved April, 15, 2019, from https://wwwc.moex.gov.tw/main/ExamLaws/wfrmExamLaws.aspx?kind=1&menu_id=318&laws_id=11]

姚漢禱、姚偉哲（2007）：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.]

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.]

趙子揚、黃嘉莉、宋曜廷、郭蕙寧、許明輝（2016）：教師情境判斷測驗之編製，**教育科學研究期刊，61**（2），85-117。https://doi.org/10.6209/jories.2016.61(2).04 [Chao, T.-Y., Huang, J.-L., Sung, Y.-T., Kuo, H.-N., & Shiu, M.-H. (2016). Construction of the teacher situational judgment test. *Journal of Research in Education Sciences, 61*(2), 85–117. https://doi.org/10.6209/jories.2016.61(2).04]

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。 https://doi.org/10.6251/BEP.20160801 [Hsieh, M.-C. (2013). Who is a good Speaker? Applying multifaceted Rasch model to analyze principal three-minute impromptu speech. *Bulletin of Educational Psychology, 48*(4), 551–566. https://doi.org/10.6251/BEP.20160801]

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.]

謝如山、謝名娟（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。https://doi.org/ 10.6251/BEP.20121101.1 [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. https://doi.org/ 10.6251/BEP.20121101.1]

Aryadoust, V. (2015). Self- and peer-assessments of the oral presentations of first-year science students. *Educational Assessment, 20*(3), 199–225. https://doi.org/ 10.1080/10627197.2015.1061989

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

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1–18. https://doi.org/ 10.2307/1164948

Breton, G., Lepage, S., & North, B. (2008, June 23–25). *Cross-language benchmarking seminar to calibrate examples of spoken production in English, French, German, Italian and Spanish with regard to the six levels of the Common European Framework of Reference for Languages(CEFR)*. Paper presented at the Séminaire Interlangues Cross Language Benchmarking Seminar, CIEP, Sévres, France. https://rm.coe.int/168045a0cd

Campbell, E. H. (1993, March 31–April 3). *Fifteen raters rating: An analysis of selected conversation*

during a placement rating session. Paper presented at the 44th Annual Meeting of the Conference on College Composition and Communication, San Diego, CA, United states. https://files.eric.ed.gov/fulltext/ED358465.pdf

Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

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

Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*, 19–33.

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

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. https://doi.org/10.1207/s15434311laq0203_1

Harasym, P. H., Woloschuk, W., & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advance Health Science Education Theory Practice, 13*(5), 617–632. https://doi.org/ 10.1007/s10459-007-9068-0

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Kuo, C.-Y., Wu, H.-K., Jen, T.-H., & Hsu, Y.-S. (2015). Development and validation of a multimedia-based assessment of scientific inquiry abilities. *International Journal of Science Education, 37*(14), 2326–2357. https://doi.org/10.1080/09500693.2015.1078521

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*, 332–340. http://doi.org/10.1037/0021-9010.79.3.332

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

Linacre, J. M. (1994). *Many-facet Rasch measurement*. MESA Press.

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

Linacre, J. M. (2005). *A user's guide to Winsteps/Ministeps Rasch model programs.* MESA Press.

Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement, 11*(1), 1–10.

Linacre, J. M. (2019). *Facets Rasch measurement* [computer program]. https://www.winsteps.com/index.htm

Longford, N. T. (1993). *Reliability of essay rating and score adjustment.* Educational Testing Service

prodram statistics research (Technical Report NO.93–36). https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01563.x

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19*(3), 171–200. https://doi.org/10.3102/10769986019003171

Lunz, M., & Suanthong, S. (2011). Equating of multi-facet tests across administrations. *Journal of Applied Measurement, 12*, 124–134.

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

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Educational Testing Service policy information center. https://files.eric.ed.gov/fulltext/ED353302.pdf

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system*. Educational Testing Service center for performance assessment (Report NO.MS 94–05). https://www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf

North, B. (2000). *The development of a common framework scale of language proficiency*. Lang.

North, B., & Jones, N. (2009, Jaunary). *Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. *Language Policy Division.* https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680459fa0

O'Neill, T. R., & Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 135–146). Ablex.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *TESOL & Applied Linguistics, 4*(1), 1–21.

Palermo, C., Bunch, M. B., Ridge, K. (2019). Scoring stability in a large-scale assessment program: A longitudinal analysis of leniency/severity effects. *Journal of Educational Measurement, 56*(3), 626–652. https://doi.org/10.1111/jedm.12228

Raymond, M. R., & Viswesvaran, C. (1993). Least-squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*, 253–268. http://doi.org/10.1111/j.1745-3984.1993.tbo0426.x

Raymond, M. R., Webb, L. C., & Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions, 14*, 100–122. https://doi.org/10.1177/016327879101400107

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 12*, 199–218.

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). Sage.

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

Tseng, W.-T., Su, T.-Y., & Nix, J.-M. L. (2019). Validating translation test items via the many-facet Rasch model. *Psychological Reports, 122*(2), 748–772. https://doi.org/10.1177/0033294118768664

Wilson, H. G.(1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement, 48*, 69–81. https://doi.org/ 10.1177/001316448804800109

Wind, S. A., Engelhard, G. J., & Wesolowski, B. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment, 21*(4), 278–299. https://doi.org/10.1080/10627197.2016.1236676

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). Sage.

收稿日期：2019 年 10 月 18 日
一稿修訂日期：2020 年 04 月 02 日
二稿修訂日期：2020 年 04 月 24 日
接受刊登日期：2020 年 04 月 24 日

Bulletin of Educational Psychology, 2020, 52(2), 415–436
National Taiwan Nommal University, Taipei, Taiwan, R O.C.

Investigating the Effects of Rater Equating Designs on Parameter Estimates in the Context of Preservice Principal Oral Performance

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

A problem in performance assessments is the degree to which rater severity and leniency can affect the examinee's scores. In particular, fairness concerns related to performance systems include the exchangeability of raters. A possible resolution for addressing rater severity is for each rater to score each examinee's performance; thus, the difference in rater severity would affect each student at the same level. However, this is not always feasible in practice for fully crossed rating designs. In the context of performance assessments, equating procedures create links between raters when performing transformation with a fully crossed rating design is not feasible and could control for differences in rater severity.

An effective equating procedure involves a strong statistical model and a systematic data collection approach. The Many–Facet Rasch model (MFRM) is a commonly used approach for adjusting rater differences. Although the use of the MFRM model has gained popularity as an equating approach for rater severity, several key considerations related to data collection designs and model data fit are also crucial. In particular, it is vital to ensure a sufficient level of connectivity in the rating design; that is, the raters can be linked to other assessment components, such as other raters, examinees, or tasks.

Three types of data collection design are commonly used for equating. The first type is a complete network design, in which the data consist of complete designs with subjects of all assessment components. This is an ideal design for a rating system. The second type is an incomplete network design. Under an incomplete network design, examinees do not have scores on all assessment components, but a partial and systematic degree of connectivity exists for raters and tasks to produce a connected network of assessment components. The third type is a nonlinked network design, where no systematic linkage exists in the components of facets. Even if the unlinked scoring network has some potential problems, many important exams in Taiwan still use this rater design.

The purpose of this study was to examine the effect of differences in data collection designs that could affect parameter estimation in the performance assessment. Using empirical data, this study explicitly related the central role of consideration to data collection designs for the interpretation of results when the MFRM is applied. The study had two main research objectives: (1) To examine the impact of different data collection designs on parameter estimates for examinees' ability, raters' severity, and the difficulty of scoring criteria. The indices included infit, outfit, separation index, reliability, and the chi-square test. (2) To evaluate the correlations of ability estimates between different designs and the magnitude of their impact on the ranking of the examinees' performance level. Examinees for the top 10, middle 10, and last 10 examinees in the complete network design were selected to evaluate their ranking differences for other designs.

This study used the MFRM and oral performance score data of preservice principals to explore the effects of the three data collection designs. A total of four raters and 85 preservice principals participated. The raters scored seven criteria for each preservice principal's oral performance: content, structure, word usage, attitude, pronunciation, intonation, and time control. Each criterion was assigned a grade of 1–3, of which 1 represents the basic level, 2 represents a proficient level, and 3 represents an advanced level. The raters were trained before the actual rating was conducted. The specification of each grading level and the standards were explained; raters were also required to complete rating exercises before conducting the official rating. The anchored videos at various levels for the raters were discussed. Through these anchored videos, the raters could better understand the standards.

Four equating designs were considered in this study. Design 1 was the complete network design; four raters rated all preservice principals in this design. Designs 2 and 3 were incomplete network designs. In these two designs, some rating scores overlapped to construct the connectivity of scoring components. In design 2, each student received scores from three raters, whereas in design 3, each student received scores from two raters. Design 4 was a nonlinked network design, in which each rater only reviewed his or her assigned class; there was no connection between raters' scores. The MFRM, a statistics model of the Rasch family, was employed to perform the four equating designs. When estimating the examinee's ability level, raters' severity and scoring criteria were simultaneously considered in the model.

This study had two main findings: (1) For the incomplete and nonlinked network designs, some minor problems were related to the model fit, but overall, the infit and outfit indices were close to 1, which indicated that the use of the MFRM was feasible for analyzing the data used in this study. However, the reliability and separation indices for the nonlinked network design were low, and some chi-square tests did not reach significance—results that were quite different from the complete network design. (2) The lower the linkage between assessment components, the more biased the estimated stability of parameters. The fully connected network design provided the strongest connectivity at all levels (subjects, raters, and criteria), and this design was also the most ideal scenario for the data collection design. However, this design costs much in terms of rating time and money; thus, it is difficult to implement such a design in a large-scale test. By contrast, incomplete network designs are more feasible in large-scale tests, namely for establishing overlaps of the evaluation of some subjects of raters. The correlation between the complete network design and nonlinked network design was only 0.69, but the correlation between the complete network design and incomplete network design rose to 0.79–0.94. Moreover, a clear gap existed in participants' rankings between the ideal fully connected network design and nonlinked network design. For example, student #59 ranked 79th in the complete network design but 21st in the nonlinked network design, equaling a ranking difference gap of 58. The results revealed that even if the MFRM is used for correction, large errors will still exist in the estimation of ability values and the ranking results of examinees for a nonlinked network design.

This study provided two suggestions: (1) Examination institutions should avoid using the nonlinked network rater design. Carefully constructed network assessment designs based on effective data collection designs have the chance of obtaining objective and fair measurements within systems with multiple facets. Regarding current large-scale tests, many do not use any statistical models for rater severity correction; furthermore, they use the nonlinked rater design. It is possible that examinees can experience bad luck and encounter a severe grader, resulting in them receiving a low score. Therefore, this study recommended that important examinations in the future should adopt a more complete rating plan. (2) This study used empirical data. A simulation study can be considered to further examine the impact of different designs of component connectivity on parameter estimates. In addition, different experimental designs are worth discussing; for example, if examinees are nested within tasks, would this nested relationship affect the parameter estimates of ability and rater severity levels? The impact of more complex data collection designs are worthy of future research.

Keywords: Many-Facet Rasch model (MFRM), preservice principal oral performance, rater equating design, rater severity