國立臺灣師範大學教育心理與輔導學系
教育心理學報，2022，54 卷，2 期，333–356 頁
https://doi.org/10.6251/BEP.202212_54(2).0004

校務檔案怎麼評？校務經營檔案之信效度分析*

謝名娟

國家教育研究院
測驗及評量研究中心

校務經營能力為候用校長的必備職能之一。過去儲訓校長班繳交許多作業與文檔，內容聚焦為校長校務經營的各種面向的培養，但卻缺乏系統性的蒐集。本研究設計候用校長設計校務經營的檔案評量，並探討其信效度。評分者的評分嚴苛度、評分規準、評分者是否為師傅校長或是外聘專家對於受試者的影響亦加以探討。其結果顯示候用校長在校務檔案的製作中，對於學校背景資料掌握較好，對於創新經營的面向較差，外聘專家比師傅校長評分較為嚴格，但外聘與師傅校長評分均受到誤差的影響。總體而言，校務經營檔案評量具有大致良好的幅合效度證據，但區辨效度較不理想。

關鍵詞：校務經營、檔案評量、信度、效度

* 1. 通訊作者：謝名娟，mhsieh@mail.naer.edu.tw。
2. 本研究感謝科技部補助（計畫編號：MOST 108-2410-H-656-002）。

檔案評量在 1980 年代後期才逐漸為學者所重視，所具備的訊息和傳統的測驗題型有所不同，得到的資訊較為整合，尤其對學生的學習歷程可提供較為全面的回饋訊息（張郁雯，2008）。近幾年由於大學考試制度的改變，學生的學習歷程檔案也逐漸受到重視，電腦應用的普及有助於資料保存與交換，許多學術單位都將數位化的學習檔案，融入課堂教學與評量中（張基成、彭星瑞，2008；Singh & Ritzhaupt, 2006）。以檔案為基礎的學習方式，可透過系統性的作業收集，讓學習者去自我監控進度與學習目標達成的狀況。在檔案的學習歷程中，教師的角色為指導學生去建構一個檔案，設定學習的目標、寫出反思的歷程、提供自我評量與同儕回饋等（Chang et al., 2018; Higher Education Funding Council for England [HEFCE], 2008）。檔案不是將文檔或作業散亂的集結在一起，而是透過這些檔案讓教與學更有效率（HEFCE, 2008）。

檔案的學習歷程包括了計畫、監控、反思，而這些都能支援學生的學習歷程與表現、並提供自我成長的機會（Chau & Cheng, 2010）。其中包括了目標設定、資料蒐集、反思、修正、整理、作品發表、自我評估、同儕評估、同儕觀察、同儕回饋（Chang et al., 2018; Coombe & Barlow, 2004; Hughes, 2008; Joyes et al., 2010; Sharma, 2007），而這些和知識建構與產出都有密切相關（Chang et al., 2013; Chau & Cheng, 2010）。過去臺灣有許多學者專家在大學與中小學階段，建構了學習檔案的豐富理論基礎與實務研究（林素卿、葉順宜，2014；張美玉，2000；張郁雯，2008；張基成、吳炳宏，2012；張基成、林俊宇，2015；張基成、廖悅媚，2013；張麗麗，2002）。本研究著重在統整性、職能性的成人檔案評量，探討檔案評量的信效度，另外聘分者的評分嚴苛度、評分規準、評分者是否為內聘或是外聘專家對於受試者的影響亦加以探討。

在現在的中小學候用校長班，每期課程學員均繳交許多作業與文檔，雖內容聚焦為校長校務經營的各種面向的培養，然而卻缺乏針對職能指標來系統性的蒐集。尤其在實務執行時，有不少學員表示作業過多、且彼此有重複性的問題。根據謝名娟與林信志（2014）的研究指出，候用校長的整體性的作業很多，但是卻相當的分散，缺乏整體性，和職能指標的連結性也不夠強，這些散亂且缺乏系統性的作業，對於業務單位、候用校長與評分者都是負擔，甚至作業還有疊床架屋的情形。另外，目前的主要評分人員為各班的師傅校長，然而，由於師傅校長評分有嚴厲和寬鬆的差別，這樣的問題曾呈現於謝名娟（2017）的研究中，因此，如果單靠各班的師傅校長的評分，而缺乏多方的證據的相互驗證，可能會因為月暈效應的影響，造成評分的不公與誤差，因此需要外聘專家的評分，並將每班的成績進行等化與串連（謝名娟，2020）。然而，過去對於成人的歷程檔案的相關研究較為缺乏，而本研究聚焦在儲訓校長的校務經營檔案，探討內外聘專家在評分上是否具有一致性？另外，檔案評量耗時耗力，但其效度證據是否足夠？這些乃是本研究主要探討的問題。

文獻探討

（一）候用校長的專業標準與職能指標

職能一詞由哈佛心理學家 McClelland 在 1973 年所提出之名詞，他透過一系列在職場上的研究，發現決定職員工作績效的原因相當多元，例如態度、知識與人格特質等，而非只有智力（Spencer & Spencer, 1993）。在工作的場所中，個人所具有的基本能力與特質，更可以用來預期工作者在職的實際表現與績效。由於現在的工作瞬息萬變，大多數工作的核心職能不會由單一的能力或技術所組成，而是有一組複雜的關鍵能力。校長則是學校最主要的領導人物，在學校負責各種校務的執行、政策的發展、教學引導，學校的行銷與績效管理，與校院內外的溝通協調等，其工作的複雜度須與時俱進，日日更新。尤其隨社會的發展與變遷、十二年國教的推行，家長、學生、教師、甚至是社區鄰里的期待，校長的職能培養更具有其重要性（石文傑等人，2014；陳宏彰，2017；劉祥熹等人，2016）。林信志與謝名娟（2018）所提出之未來學校領導方案中，再進一步將過去研究將以彙整與融合，明確訂立中小學候用校長職能指標為 6 大面向，包括願景型塑、策略思考、團隊合作、溝通協調、創新經營與自我覺察等（如表 1）。

表 1
中小學候用校長專業標準與職能指標

領導職能	專業標準	職能指標
A. 願景形塑	致力於學生學習和教師專業發展，並能結合各項資源，明確勾勒學校的未來發展。	A1. 能理解學校文化與使命，發展學校願景 A2. 能明確勾勒學校未來發展的定位與方向 A3. 能掌握校內外各項資源，並妥善分配、應用與創造其價值 A4. 能以學生學習和教師專業發展為一切決策與行動的基礎
B. 策略思考	善用系統思考，並能明辨校務各項問題的成因，且能洞察時勢做出校務發展的最佳決策。	B1. 善用系統思考的工具或方法，分析歸納尋繹具優勢的校務策略 B2. 能找出校務資料間之關聯性，推演預測其趨勢 B3. 能由學校各種分歧的事件中辨識問題產生的特徵與因果關係 B4. 能洞察時勢調整校務發展，快速回應環境之挑戰與需求
C. 團隊合作	知人善任，並能與那些有能力改善校務的夥伴一起協作，創造正向有效能的學校組織文化。	C1. 能知人善任適才適所，給予學校各位同仁最佳位置 C2. 能分析自己與學校之狀況，找出需要建立或加強的夥伴關係 C3. 能組織校內外的結盟，建立不同協力關係 C4. 能與合作夥伴釐清雙方期望及合作範圍，確保符合彼此需求
D. 溝通協調	善於溝通及尊嚴傳達，並能激勵團隊，採取創造性協調方式達成最佳共識。	D1. 能激勵引領同仁相互支持，鼓舞其士氣，形成團隊共識 D2. 能尊重學生、同仁和其他利害關係人的權利、看法和信念 D3. 能針對不同對象選擇適合的口語或非口語表達方式，切要且條理分明地告知訊息 D4. 於嘗試溝通後，若還是無法與對方達成共識，能運用多元管道協商共識
E. 創新經營	能跳脫既有框架模式，在堅持傳統與開放改革間取得平衡，建置鼓勵創新經營的機制，引領學校革新。	E1. 能跳脫既有框架模式，找出校務最佳運作模式 E2. 能接納不同、具創意的工作方式，並建置鼓勵創新經營的機制 E3. 能鼓勵同仁突破工作現狀，提出新模式，落實校務創新經營 E4. 能在堅持傳統與開放改革間取得平衡，引領學校創新與變革
F. 自我覺察	能不斷反思自身和他人的實踐，具備社會意識並能調適自我情緒與壓力，困境中依然能正向發展出健康因應策略。	F1. 能不斷反思自身和他人的實踐，及情境脈絡中自身的定位 F2. 能理解學校利害關係人的期待，並能接納少數族群與特殊利益團體的觀點 F3. 遇到困境時能正向思考，反思辦學契機 F4. 能調適自我情緒與壓力，於危機或壓力情境中發展出健康的因應策略

註：引自《中小學候用校長職能指標系統與評鑑中心法之發展與研究》（頁 5—7），林信志、謝名娟，2018，科技部補助專題研究計畫。

（二）檔案評量的評分

若將數個作業的內容系統性的集結並進行歸類，可變成檔案評量。根據張美玉（2000）的文章指出，其檔案的內容，須依據評量目的而決定，有了具體的評量目的之後，才決定評量的內容。而檔案中，則須包括學生自己撰寫並選擇的內容、自我評估的證據等 （Arter & Spandle, 1992）。張基成與吳炳宏（2012）指出，檔案的內容項目應該符合教學者的教學目的、需要與可行性，且學習者的檔案必須具備共同與最低要求的內容項目，讓教師容易進行評量。

依據檔案評量的目的，會有不同的評量指標與評分要項。例如在 LeMahieu 等人（1995）的寫作檔案中，其檔案的規準分為三個面向，包括寫作表現、寫作過程與策略應用、寫作者成長與發展能力的評估。Skawinski 與 Thibodeau （2003） 指出檔案評量的規準可分為一般性、特殊性與綜合性的規準。一般性的可提供反思與進步的證據，特定性可要求特別的學習績效重點，而綜合性則包

括整體品質。張基成與吳炳宏（2012）建議可以將評量規準分成整體與分項，並使用五等量表的方式呈現不同等級，在 1 分、3 分、5 分的部分分別寫下表現標準描述，並提供範例、證據或具體的作品。這樣的評分規準可以作為教學者評分的依據與學習者準備資料的參考。而在表現的標準描述部分則應力求具體，提醒評分者避免偏誤的現象。

（三）檔案評量的信效度

信度是指評量結果的穩定性（stability）及一致性（余民寧，2011）。評量結果的穩定性可由再測信度來評估，然而，由於實施檔案評量會耗費較大的人力和時間，大多都只能施測一次。因此，大多使用評分者的一致性來檢視檔案評量的信度。張麗麗（2002）指出檔案評量的信度可以透過不同的檔案、不同的評分者、不同的時間與施測情境來檢視評量結果的一致性。但由於檔案多為主觀型的計分，檔案內容的項目數量也不能太多，加上內容複雜、資料蒐集時間較長的相關因素影響，分數的誤差值容易提高，進而影響到信度。

受限於大多數的課室內評量，評分者僅限為教師自己，這種只有單一評分者的狀況無法計算評分者一致性。然而，在有限的經費人力下，也可以考量同儕互為評分者，或是和搭擋的教師一起評分，即可以評分者信度來進行檔案評量的信度。評分者在進行評分時，常常會有盲點，例如對某些學生平常的印象很好，即使在此測驗的表現不好，也會因為印象分數而給高分。這些干擾的因素，都可能會影響到測驗的結果，要如何提高檔案評量的信度，可以透過較為清楚、明確的評分準則與適當的評分者訓練，也有研究者透過系統的掌控，來監控是否有評量偏誤的現象存在，若評分者之間的一致性過低，則不讓評分者繼續評分。張郁雯（2008）以檔案評量探討國小學生的資訊素養能力，其科目包括語文、自然與生活科技、社會等三個學習領域，每個領域包括 2—3 個作品，信度指標使用類推性係數與可靠性係數來推估，其結果顯示即使只有一位評分者，其類推性係數仍能達到 0.70—0.91 的水準，另外學生在解決資訊問題的各歷程能力程度不同，透過適當的作業設計可得知學生在各向度的相對強弱能力。在評分者一致性，則發現教師間高於教師與助理間的一致性，其可能原因是助理教學經驗較為不足，且對於學童的程度較無法掌握。另外，作業的難度與評分者背景的落差會影響到檔案的信度，若要確保信度則應發展良好清楚的評分規準。

效度是用來評鑑評量結果的解釋與使用的合適性。效度分為很多種，包括內容效度、效標關聯效度、建構效度等（余民寧，2011）。每種效度的證據若能都蒐集是最好的，但要找到這些證據較為困難，尤其就張麗麗（2002）的研究指出，探討檔案評量的效度方面的研究仍須累積，過去大多以學生或教師自陳的方式來探討學生的學習成效。但應該更為系統性的蒐集效度的成效，例如在內容效度的部分，可以透過清楚界定的目標、強化轉化檔案內容與目標之間的連結性。並將計分準則清楚的訂定，而在效標關聯效度部分，則可依據學生在班上的其他表現，如紙筆的學習成績等作為效標，例如 Diperna 與 Derham（2007）、Gadbury-Amyot（2003）等均使用學生測驗分數當作外在效標，並以效標關聯效度來檢視學業表現與檔案評量間的分數。

張基成與吳炳宏 （2012）探討網路評量中的信效度，並指用高職生作為研究對象，其系統包括個人檔案的製作、觀摩、自我評鑑與同儕互相評鑑，其結果顯示其網路化的評量具有足夠的信效度，檔案分數與學生的測驗成績有高相關。而其結果亦顯示檔案評量技能面（如檔案製作）會有較高的效度，而在情意面（如個人反思）則一致性較低。而 Diperna 與 Derham（2007）針對師培教生檢驗數位檔案評量的信度，其結果發現檔案分數與學生的學業表現相關係數未達顯著，而兩位評分者之間的評分一致性也不佳。檔案的信效度高低，可能受研究對象的教育程度、學科屬性、評量規準、數位檔案的內容、評分訓練等影響造成。

張郁雯（2010）則著重以 Messick（1989） 所提出建構效度中的幅合與區辨效度、後果效度以及實質面來探討國小學生運用資訊能力的檔案評量，其結果顯示檔案有可接受的幅合效度，但區辨效度則須提升。

方法

（一）研究對象

本研究的參與受試者為在研究機構受訓的候用校長。其受訓為每年三月開始，為期八周的課程。本研究蒐集的樣本共有 61 位候用校長，男性為 45 位，女性為 16 位，平均年齡為 44 歲。每班配有資深的師傅校長兩名，其任務類似帶班導師，除了解決候用校長在受訓期間關於課程的相關疑問之外，亦需擔任評分員的角色。儲訓班每一期均有系列發展主軸課程，而這些主軸課程延伸下，學員也需要繳交相關的作業，其作業會由師傅校長與外聘專家一起評分。外聘專家的身分大多為資深校長、或是校長學領域的學者專家。

（二）檔案評量作業規劃流程

候用校長的課程主要包括校務發展、行政管理、專業責任、公共關係、課程與教學領導、教育參訪、師傅學習、博雅通識與綜合活動等。其中在校務發展與師傅學習為候用校長的主軸項目，校務發展的重點包括如何做校園的願景規劃、學校領導、校園建築、校務研究與自我改善、學校特色發展，而在師傅學習部分則透過案例分析、透過師傅與學員的對話、分享來協助校務發展計畫的擬定與執行。這部分的課程佔的時數為 78 小時，占總課程約 1 ／ 3 時數，除了課程外，師傅校長亦需要針對校務發展計畫中，所需要之內涵與實踐進行六次指導與交流。除了每部份的書面作業繳交之外，亦須進行個別分享，師傅校長則針對分享的內容進行學員的回饋。

每位學員必須產出校務經營的一系列相關作業，內容包括五大項：（1）學校背景資料的分析；（2）整體目標與行動方案；（3）學校課程發展與設計；（4）學校革新與特色經營；（5）自我檢討與反思。其中這五大部分對焦了校務發展的重點。例如在學校背景資料的分析部分則包括了校園的願景規劃、校園建築的盤點，整體目標與行動方案則顯示了校長對於學校領導理念與校務研究的能力。學校革新與特色經營則可以了解學校特色發展，最後的自我檢討與反思則可以看出學員如何透過對話、分享與回饋中進行自我反思與改進。

在開訓前即進行師傅校長的培訓，由研究者先介紹候用校長的職能，應包括那些內涵，而後則介紹校務檔案的五大部分。開訓前，則由各班的兩位師傅校長於班會時間來介紹職能與校務檔案的內涵，並說明每個部分應撰寫的重點內容，並配合課程的進行要求學員逐漸上傳檔案內容。例如在「資料驅動的校務研究分析」講座課程後，師傅校長則會引導學員進行校務資料的盤點，然後透過問題的引導與表格的填寫，引導學員來完成。在每個步驟下，師傅校長均會進行同儕分享、討論、而後給予建議，然後再進行下一部份的撰寫。直到學員完成完成整份檔案後，才會進行整體的評分。在附錄一中呈現校務檔案中的「背景說明」中應包含的內容與表格，每一部分完成的內容會請學員上傳到自己個人的檔案系統平台中，在結訓前則會請學員報告與分享。所有的學員在結訓前均完成檔案中的所有部分，共計 61 份。

（三）評分流程

檔案主要由每一班的 2 位師傅校長、與 2 位外聘專家進行評分，有 3 個班級，因此共有 12 位專家學者會參與評分的工作。在本研究擔任評分者均會經過評分訓練，在研究開始與結束接受資料蒐集的訪談，由研究者解說檔案製作的設計原則，接著會由研究者協助規劃作業，並一起討論評分規準的適切性，在儲訓期間將進行檔案實施的分享、檢討與修正。

進行評分時，先與評分者說明評分標準，並請它們進行試評，針對一致性進行檢視，有評分不一致的情形則深入討論其原因，若是因為對評分規準理解有所差異，則重新進行定義釐清或修正，修正後的評分規準如表 2。

表 2
候用校長校務檔案評量評分表

評分項目	未達基礎 1（略嫌簡陋）	基礎 2-3（大致正確、仍須調整）	精熟 4-5（清楚完整、彼此呼應）
背景說明	1. 學校資源的盤點簡陋。 2. 學校資源與學校優弱勢分析與戰略思維的關聯性不足。	1. 能大致盤點現有學校資源（人力、軟硬體、家長、與其他外部力量）。 2. 能依據學校背景與分析資料，提出相對應戰略思維。	1. 能完整盤點現有學校資源（人力、軟硬體、家長、與其他外部力量）。 2. 能依據學校背景與分析資料，準確說明學校的對應的戰略思維。
整體目標與行動方案	1. 願景與目標的完整性不足。 2. 方案執行可行性偏低。 3. 目標、方案與願景的關聯性不足。	1. 願景與目標大致完整性。 2. 方案大致有可行性。 3. 目標、方案與願景有大致的關聯性。	1. 願景具有完整性，目標則能看出跨面向的豐富性。 2. 方案具體且有可行性。 3. 目標、方案與願景間具有明確的關聯性。
學校課程發展與設計	1. 課程規畫主軸未能與學校之整體架構或圖像相呼應。 2. 彈性學習課程計畫與課程檢核表撰寫需進行大幅調整。	1. 課程規畫大致能與學校之整體架構與圖像呼應。 2. 彈性學習課程計畫與課程檢核表大致撰寫妥適但有調整之疑慮。	1. 課程規畫能與學校之整體架構與圖像呼應。 2. 彈性學習課程計畫與課程檢核表撰寫適切。
學校革新與特色經營	1. 未能針對學校背景，提出適合的特色經營模式。 2. 系統思考圖撰寫不完整或有錯誤。	1. 能針對學校背景，大致提出適合的學校革新與特色經營模式。 2. 系統思考圖大致撰寫正確，但仍有調整之疑慮。	1. 能針對學校背景，提出適合的學校革新與特色經營模式。 2. 系統思考圖撰寫正確且能彼此呼應。
個人省思	1. 未能針對原計畫提出具體修改之項目。 2. 提出之未來之領導作為與本計畫無法呼應。	1. 能大致針對原計畫提出具體修改之項目。 2. 提出之未來領導作為部分與本計畫呼應。	1. 能考量學校原計畫所有需求下，提出具體修改之項目。 2. 提出之未來領導作為能與本計畫呼應。

評分時，校務計畫中各部分的作業內容會逐一評分，而這些規準的內容則配搭校長的課程與職能進行設計。

（四）信效度考驗

在信度的估計部分，探討不同評分者的評量結果是否一致？尤其檔案的評分有分為師傅校長與外聘專家，雖然都照評分規準來訓練評分，但其評分者本身的嚴苛度不同、評分規準的誤解，是否仍會影響評分的準確性？不同作業間的信度是否相同？

這些關於信度的問題，研究者 用多層面 Rasch 模式（Multifacet Rasch Model）來回答評分行為的研究問題。並搭配 FACETS 軟體（Linacre, 2014）進行分析。FACET 可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苛度。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， 評分者身分為 m, 評分者的嚴苛度為 j，在表現層級為 j 時，針對這個檔案成績，學員被評分者評定為 k 等級之對數勝算比可表示為：

$$\ln\left(\frac{P_{nlmjk}}{P_{nlmjk-1}}\right) = \theta_n - \delta_l - \alpha_j - \beta_m - \tau_k$$

其中 P_{nlmjk} 為第 n 位成員者，評分項目是 l，評分者身分為 m，評分者的嚴苛度為 j，在表現層級為 j，得分是 k 的可能性。

其中 $P_{nlmjk\text{-}1}$ 為第 n 位成員者，評分項目是 l，評分者身分為 m，評分者的嚴苛度為 j，在表現層級為 j，得分是 $k\text{-}1$ 的可能性。

θ_n 為受試者 n 的能力；

δ_l 為評分項目 l 的難度；

α_j 為評分者 j 的嚴苛度；

β_m 為評分者身分為 m 的嚴苛度 ,

τ_k 指評定 k 或是 $k\text{-}1$ 之間的難度界線，也稱為難度階（threshold difficulty）。

由此模型可見，受試者本身的能力、評分項目、評分者嚴苛度為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 均方值做為指標，當這兩個只介於 0.7 到 1.3 之間（McNamara, 1996），代表資料適合使用此模式進行分析，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，則代表資料可能為相依樣本，資料的獨立性不足。

模式所估出之可靠度（reliability）也可用以檢視評分資料的穩定性（Wright & Masters, 1982），越接近於 1，代表評分資料越穩定。另外，MFRM 既為 Rasch 家族中的模型之一，亦要符合單向度的假設，其中一種有效檢測單向度的方法為檢視 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向度的證據 （林小慧等人，2018；Linacre, 1998; Smith, 2002; Tennant & Pallant, 2006）。

效度部份，對於校務經營的檔案研究最好的效標是候用校長實際到校後，觀察校長在校的經營狀況與理念，或是訪談該學校的教師、學生、在地的縣市主管，來做為校務經營檔案的效標。然而，這樣的方式在實際操作有很大的困難，大多候用校長離開儲訓單位後，其相關資料基本上無法強制蒐集。因此本研究參考 Messick（1989）與張郁雯（2010）所使用的效度面向，檢驗幅合與區辨效度。其中外在面的幅合與區辨效度，將蒐集候用校長的檔案成績、考試成績與其他作業成績、師傅校長對於其生活表現的綜合評比等，以校務經營檔案評量分數與這些成績來做相關，以檢視其效度。

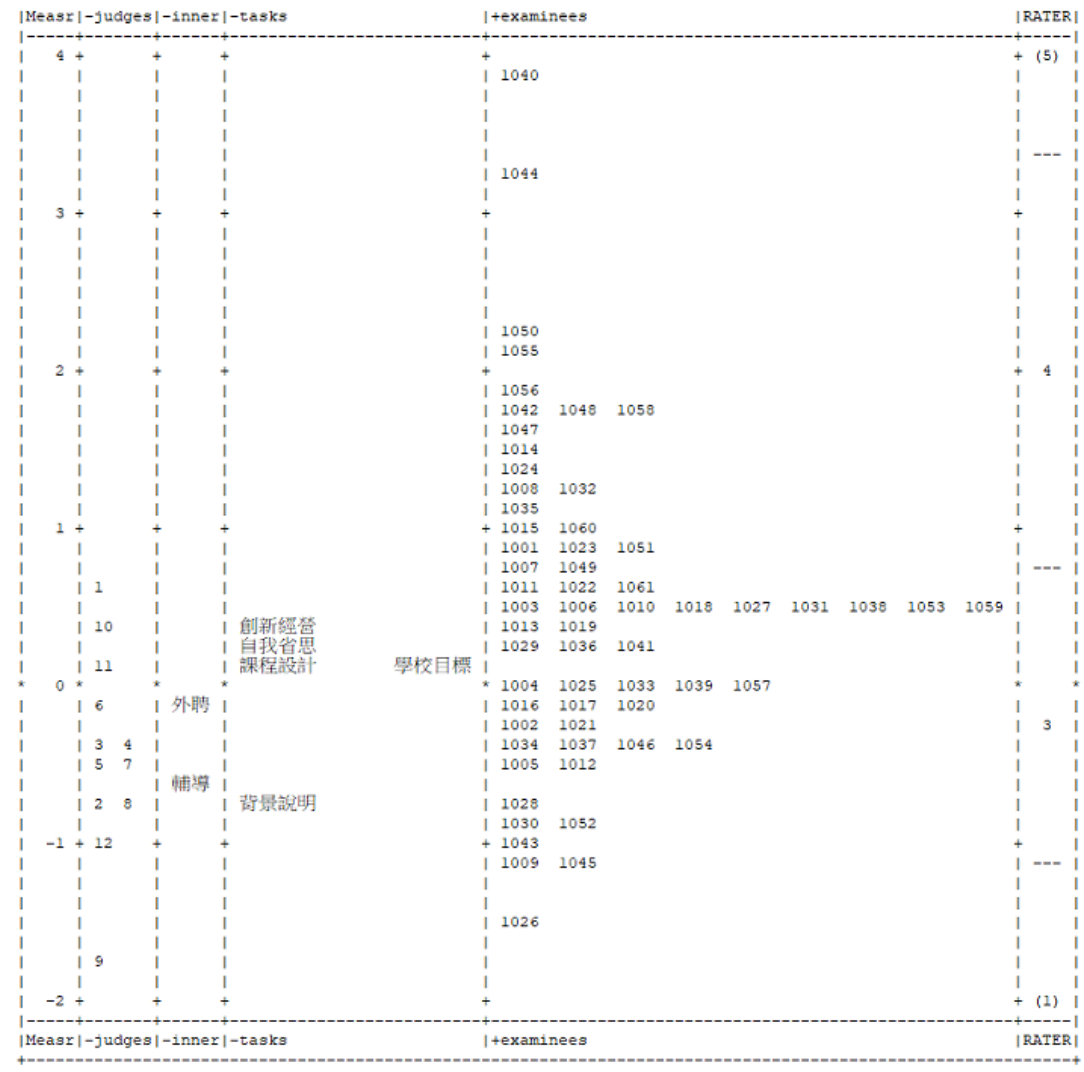
結果

根據在研究機構受訓的候用校長，所撰寫校務經營檔案報告，資料共有 61 位中小學的候用校長的資料進行評估。主要呈現兩項成果（1）透過四層面的多面向 Rasch 模型來估計參數與適配度指標。（2）校務檔案中的區辨效度與幅合效度。

（一）參數估計與適配度指標

利用 Rasch 模型來進行分析，所估計的參數包括評分項目的難度、評分者的嚴苛度、評分者身分（是否為外聘），與候用校長的表現等都融入在模型中。其中在分析上採用 FACETS 軟體，並參考謝名娟（2017）、張新立與吳舜丞（2008）的研究，在 MFRM 的模式中，將評分項目難度設為 0，以評估評分者的嚴苛度、內聘與外聘委員的嚴苛度與學員的程度表現的相對程度。

圖 1
變數分布圖（variable map）



註：從這個分布圖中，可以看出本研究所考慮的各層面中變數分布的概況。其中 Measr 為參數的刻度，judges 為評分者嚴苛度的分布位置，inner 為評分者身分的嚴苛度分布位置，tasks 為評分項目的難度分布。

圖 1 為校務檔案中的參數分布圖，這個圖主要用以呈現每一個變數（包括不同的評分者的嚴苛度、評分者身分是否為輔導校長或是外聘專家的嚴苛度、作業任務的難易度、學員能力高低分布）等概況。圖 1 中的第一個欄位是顯示參數刻度，以 logit 為單位。在每一個層面中所展示的相對位置也可以從這個刻度中進行檢視。第二個欄位是所有 12 位評分者的嚴苛度排序，越上面代表越嚴格，從中可看到編號 1 評分者是最嚴格，而編號 9 的評分者最為寬鬆。第三個欄位則顯示評分者身分的影響，從中可以看到外聘的校長較為嚴格，而具有輔導員身分的師傅校長在評分上較為寬鬆。第四個欄位則為評分向度的難易度，越上面代表對於學員越困難，越下面代表越簡單，從中可以看出創新經營的部分對於學員來說較為困難，而最為簡單的是背景說明。第五個欄位則顯示 61 位學員的相對位置排序分布，在圖形中越上面的學員表現越好。

由前所述，本研究參考張新立與吳舜丞（2008）的研究，將評分項度的難度參數設為 0，以進行 MFRM 中各種參數高低之比較。從表 3 顯示候用校長的表現參數平均值為 0.49，其參數平均值高於評分項目 0，評分者嚴苛度 -0.43，與內外聘評分者的嚴苛度 -0.39，這顯示候用校長的表現，高於評分者的平均嚴苛度，也代表評分者的評分對於候用校長來說較為寬鬆。各層面的 infit 與 outfit 均方值接近於 1，代表適配度良好。但從表 3 中可看出候用校長表現程度的信度值較低，其中在 Rasch 模式中，信度代表測量值是否具有可複製性（reproducibility），若高信度代表能力值的受試者，在真實能力上也有較高的可能性表現較好，而候用校長的能力估計上信度值較低可能是因為候用校長彼此之間同質性高、能力相近的緣故。

表 3
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	信度 Reliability	p-value
評分項目難度	.00	.10	1.01	0.98	.93	.00
評分者嚴苛度	-.43	.15	0.98	0.97	.94	.00
內外聘嚴苛度	-.39	.06	1.00	0.98	.92	.00
候用校長能力	.49	.34	1.01	0.98	.88	.00

如表 4 在各評審向度中，可看出背景說明的參數值最低，為 -0.71，代表此項目對於候用校長來說最為簡單，最困難的是創新經營的向度，其參數值最高，為 0.33，代表對候用校長來說，要能在校務報告中展現出創新經營的理念較為困難。其可能原因是候用校長自認為已經提出創新的想法，屬於較為高層次的能力，對於候用校長來說較為不容易掌握，尤其評分者多為資深校長或是專家學者，他們已見多識廣，並不覺得候用校長所提出的創新叫做創新，或是認為其提出的創新並不符合學校的背景所需。因此在此項目評分都不高。但是在背景說明部分，由於多以數據方式呈現，屬於事實性的資料陳述，將校務現況與相關資源盤點清楚即可得分，因此對於候用校長來說算是簡單。另外，鑑別度為 Rasch 模式對於鑑別度的適配指標之一（Linacre, 2014），理想的適配鑑別度為 1，代表這項目的鑑別度符合 Rasch 模型的預期，若是大於 1 代表對於高分者與低分者的鑑別度超過了 Rasch 模型的預期，低於 1 則代表此鑑別度低過 Rasch 模型的預期。鑑別度合理範圍界定為 0.5 到 1.5 之間，從表 4 中可看出各評審項目在鑑別度上的適配度都在可以接受的範圍之內。

表 4
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
背景說明	-0.71	0.15	0.88	0.87	1.15
學校目標	0.11	0.14	1.20	1.14	0.83
課程設計	0.08	0.15	1.10	1.07	0.91
創新經營	0.33	0.17	0.91	0.90	1.10
自我省思	0.20	0.19	0.94	0.95	0.56

從表 5 看出其評分者嚴苛度參數值大多為負值，不管為內聘、外聘人員均為寬鬆取向，較為嚴格的評分者為 1、10、11 三位。Infit 與 outfit 的期望值是 1，其範圍為 0 到∞，若 infit 與 outfit 值大於 1，代表評分的變異性超出預期，若小於 1 則代表變化性小於預期。

根據 Linacre（2014）建議，infit 與 outfit 合理的範圍值為 0.7 到 1.3 之間。評分者的 Infit 與 Outfit 的均方值大多落在合理的範圍內，但評分者 3、4 的均方 值過低，代表評分的分數變化不大，其評分過於容易預測（too predictable），經檢視其大多數的評分不論為甚麼向度，所有候用校長都是給 3 ～ 4 分，其評分具有趨中現象，而這兩位評分者為內聘的師傅校長。

然而對於 6、11 號評分者其 outfit 值已經超過了 1.3，代表這兩位評分者給的分數變異性較大，已超過模型的預期。

表 5
評分者嚴苛度分布

評審代號	參數值	標準差	Infit 均方	Outfit 均方	身分
1	0.66	0.15	1.33	1.33	外
2	-0.79	0.16	1.00	1.00	外
3	-0.40	0.16	0.41	0.41	內
4	-0.35	0.16	0.45	0.44	內
5	-0.52	0.14	1.01	1.01	外
6	-0.18	0.14	1.74	1.67	外
7	-0.47	0.15	1.02	0.99	外
8	-0.81	0.15	0.90	0.92	內
9	-1.72	0.16	0.85	0.87	內
10	0.31	0.14	0.79	0.81	外
11	0.18	0.15	1.52	1.52	內
12	-1.06	0.16	0.72	0.68	內

註：粗體代表超過合理範圍。

若將候用校長的表現由高分排列到低分（圖 2），並將得到的原始觀測平均與模型預估的期望平均值進行比較，可看出對於高分組的候用校長來說，觀測平均略低於期望平均值，且中間與後段的候用校長表現實際表現與模型預估的分數落差較大。

圖 2
觀測平均跟期望平均值的差異

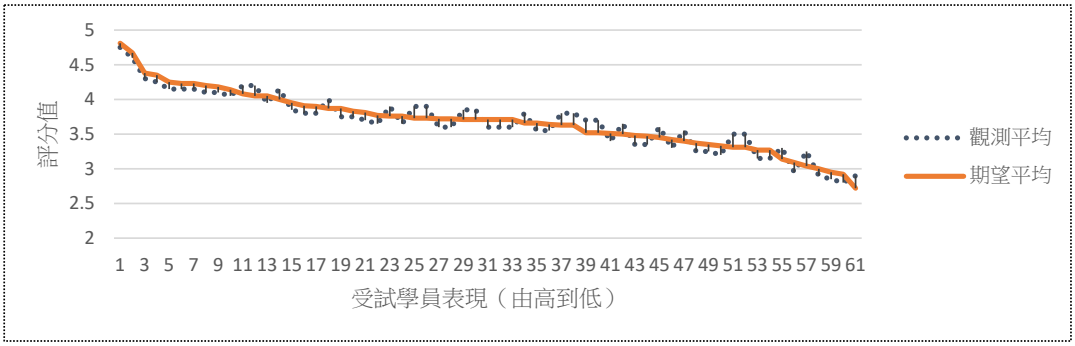


表 6
外聘專家與師傅校長的評分差異

評審代號	參數值	標準差	Infit 均方	Outfit 均方	鑑別度
外聘專家	-0.17	0.06	1.15	1.13	0.84
師傅校長	-0.61	0.06	0.84	0.84	1.17

從表 6 來看，外聘專家與師傅校長評分的參數值均為負數值，但外聘的參數值 -0.17 仍高於師傅校長的 -0.61，代表外聘專家較師傅校長略為嚴格。但兩者在評分上的鑑別度上都在可接受之範圍，代表應能有效鑑別高分組與低分組表現的候用校長。

表 7 呈現各向度中標準化殘差的絕對值大於 2 的次數統計，這殘差值過大也可以顯示評分者的

評分與模式的預期值差異較大，其中對於外聘委員來說，次數最多的在學校目標與方案設計的向度中，其次為課程設計，其可能原因師傅校長在八周指導候用校長，經過六次課程的討論，對於他們所提出的學校目標、課程設計方式已有許多回饋、候用校長也依據師傅校長的意見進行修正，因此在最後作品與師傅校長的期望多為符合，在評分部分也較為一致。然而，師傅校長所指導的方向，卻不見得與外聘專家相符，例如在評分規準中，要求候用校長所提方案應為具體且有可行性，但是對於專家學者來說，其方案的可行性為見仁見智。另外在課程設計的面向中，要求候用校長在校務報告中的課程規畫須與學校之整體架構與圖像呼應，並需要將彈性時數與課程檢核都能夠在校務報告中撰寫清楚，但由於外聘委員對於候用校長學校的脈絡可能掌握並不夠清楚，因此對於課程規劃適切性也僅能憑過去自身的經驗來評斷。

表 7
標準化殘差絕對值 ≥ 2 以上之次數分布

評分向度	外聘委員	師傅校長	總計
背景分析	4	4	8
學校目標	18	0	18
課程設計	11	2	13
創新經營	6	2	8
自我省思	4	7	11
總計	43	15	58

（二）區辨效度與幅合效度

候用校長班除了需要撰寫校務經營的檔案內容之外，亦需要繳交其他作業與參加期末考試。作業類包括團體的案例分析，針對校務所遇到的特殊狀況，以小組的方式提出解決策略。另外還有共需繳交兩次的研習省思，其內容為撰寫在研習八周過程中關於學習的個人心得感想與省思，評分者主要為師傅校長。另外在期末評量部分，則包括進行模擬校長遴選場景的 6 分鐘簡報，校長領域專業知識紙筆測驗，與進行溝通表達的實作評量（其形式為模擬角色扮演，以校長的身分來進行親師之間的溝通或是媒體發言等）。其中在案例分析、校長遴選簡報評分者為師傅校長與外聘專家，研習省思與師傅校長行為觀察則為師傅校長評分，紙筆測驗與溝通表達實作評量則全由外聘專家評分。

若以本研究所著重的校務經營檔案來看，檔案評量的分數與校長遴選的簡報相關度最高（$r = 0.49, p < .01$），其可能原因為校長遴選簡報為儲訓校長要模擬校長遴選情境，做一個六分鐘的簡報，其中部分會針對校務進行說明與發表，並提出未來要進行的改革與方向，而其報告的部分內容和校務檔案中的內容相似。只是檔案主要為文字書寫的內容，而校長遴選簡報則主要以口頭發表。

另外，研習省思的相關性也較高，其內容也和校務檔案中自我省思的向度類似，雖然省思的內容不同，但是能力略有類似。然而，案例分析的內涵也是圍繞在校務處理方面，但是由於這是小組團體討論下，共同完成作業，因此和校務檔案的幅合效度並不高（如表 8）。

在區辨效度部分，校務檔案和溝通表達類的實作評量、紙筆測驗的相關都不會太高，代表本研究也具有可接受的區辨效度。但值得注意的是師傅校長的行為觀察和校務檔案具有顯著相關，然而，行為觀察主要是由師傅校長針對候用校長平日的上課態度、參與班務的狀況進行評分，理應跟校務經營檔案的成績相關不高，但在此卻發現校務檔案與師傅校長的行為觀察成績具有顯著相關，也隱喻校務檔案的評分可能多少受到月暈效應的影響。

表 8
校務檔案成績與其他成績的相關

	研習省思	案例分析	師傅校長行為觀察	校長遴選簡報發表	紙筆測驗	溝通表達實作評量
校務檔案	.37**	.13	.28*	.49**	.07	.24

* $p < .05$. ** $p < .01$.

結論與建議

校長的職能多為抽象的能力，需要設計出具體可評量的指標，並配搭可操作的作業內容與容易執行的檔案評量模式，這種模式對於儲訓校長的作業能夠建構系統性地蒐集、避免雜亂作業繳交模式，對於候用校長的檔案儲存、搜尋，具有很大的助益。在本研究中，蒐集了 61 位研究機構的候用校長的檔案資料進行分析。其結論與建議如下：

（一）結論

1. MFRM 統計模型適合用於評量候用校長的校務檔案資料，較能與本研究之研究結果相契合

本研究主要使用 MFRM 來進行檔案評量評分，並分析了在檔案評量中每個項目的難度、評分者的嚴苛度與候用校長的表現水準，雖然樣本數並不大，但是在模型上的適配度大多可接受，代表在本研究的資料適合使用 MFRM 的模型進行分析。

2. 候用校長在創新經營的面向感到較為困難，但對背景分析的面向掌握較佳

在候用校長的校務檔案中，發現候用校長在撰寫創新經營內容表現較差，但對於學校背景分析表現較優。其可能原因是候用校長過去的經歷大多是主任，為中層領導者，但決斷校務方向的主導者是校長。因此對於僅具有主任背景的候用校長來說，學校創新經營的面向較無把握。但對陳述學校背景資料，進行分析則較為容易。

3. 外聘比師傅校長評分嚴苛，而兩者的誤差不同

從分析來看，外聘專家評分較師傅校長來的嚴苛，其可能原因是師傅校長即為候用校長八周的導師，朝夕相處下很難給作業打下低分，在「沒有功勞也有苦勞」的想法下，評分有偏高的現象。另外，infit 與 outfit 值較小，也顯示給分較為集中，分數差異不大。而外聘專家給分較為嚴苛，但在學校目標、課程設計兩個面向的標準化殘差值也有誤差較大的現象，可能的原因是不夠了解檔案內容的前後脈絡，因此造成向度評分的誤差現象。

4. 檔案評量具有大致良好的幅合效度證據，但也可能受到月暈效應的影響而產生評分誤差

研究中發現校務檔案評量與校長個人的口頭遴選簡報發表具有高度的相關，且與研習省思也達顯著相關，代表具有不錯的幅合效度，但在區辨效度卻發現校務檔案與校長的行為表現也有顯著相關，其可能原因是校務檔案與行為表現的評分者均為內聘的師傅校長，因此可能受到月暈效應的影響，而造成區辨效度較不理想。

（二）建議

1. 評分者的評分偏誤行為不易預測，未來應針對評分者訓練模式進行探討

評分者雖為領域專家，但在評分上仍需要加強訓練。在評分過程中，即使有分數的評分規準設計，但仍存在評分偏誤，未來應以嚴謹的方式來設計評分者訓練，例如在評分過程中，可以安插部

分的校準案例，來檢視評分者的偏誤情形，並即時進行調整。但要如何能設計更為方便、簡單操作的評分訓練模式，增加不同評分者之間的共識，以減少主觀性的評分偏誤值得探討。另外，本研究中發現在校務檔案中雖然有不錯的幅合校度，但是區辨效度仍不高，要如何提高亦值得後續之研究。

2. 內評或外聘專家存在不同誤差，因此檔案評量的評分者應考量背景的多元性

不管是內評或外聘專家，在評分上均有其誤差性，本研究發現，外聘專家在評脈絡性的資料偏誤較大。因此未來可考量外聘專家可以針對事實性的資料進行評分，而內評專家（如師傅校長）可針對需要脈絡性，如學生進步幅度、報告設計的前後一致性等資料進行評分。如何減少師傅校長在評分上的月暈效應，也值得後續探討，也許可考量將學員報告以匿名的方式進行跨班的互評，或是融入部分同儕評量的評語或對同儕學員彼此的回饋建議，作為評分的參考依據。 然而，若須將檔案評量推展到高風險的考試（如升學考試依據），仍需謹慎的考量其信效度。

3. 可提出更為創新方案的相關設計來提升候用校長的創新能力

本研究發現候用校長在創新經營方面的面相較為薄弱，為加強此能力，可以與企業界有創新實務能力的講師進行跨域合作，以提升候用校長的創新思維。或是透過與學校現場作結合，以小組團隊合作的模式來腦力激盪，共同解決現有學校議題，除了可檢視候用校長將創新的能力應用在校務發展上，也可以將所學完整的應用出來。

參考文獻

石文傑、馮啟峰、劉偉欽、羅聰欽（2014）：〈高職校長科技領導能力指標之探討〉。《科技與工程教育學刊》，*47*（2），1–14。[Shyr, W.-J., Feng, C.-F., Liu, W.-C., & Lo, T.-C. (2014). The exploration of principal technology leadership competency indicators epistemologies for vocational high school. *Journal of Technology and Engineering Education, 47*(2), 1–14.] https://doi.org/10.6232/JTEE.201412_47(2).0001

余民寧（2011）：《教育測驗與評量—成就測驗與教學評量（第三版）》。心理出版社。[Yu, M.-N. (2011). *Educational testing and assessment* (3th ed.). Psychological Publishing.]

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。[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.] https://doi.org/10.62091/fJORIES.2018.63(1).06

林信志、謝名娟（計畫主持人）（2018）：《中小學候用校長職能指標系統與評鑑中心法之發展與研究》（計畫編號：MOST 107-2410-H-656-010- ）。科技部補助專題研究計畫成果報告，科技部。https://rh.naer.edu.tw/handle/vnbw8 [Lin, H.-C., & Hsieh, M.-C. (Principal Investigator). (2018). *A study of competency indicators and assessment center for school candidate principals* (Report No. MOST 107-2410-H-656-010-) (Grant). Ministry of Science and Technology. https://rh.naer.edu.tw/handle/vnbw8]

林素卿、葉順宜（2014）：〈檔案評量於國中英語教學應用之個案研究〉。《教育科學研究期刊》，*59*（2），111–139。[Lin, S.-C., & Yeh, S.-I. (2014). Applying a portfolio assessment of English

teaching in a junior high school: A case study. *Journal of Research in Education Sciences, 59*(2), 111–139.] https://doi.org/10.6209/JORIES.2014.59(2).05

陳宏彰（2017）：〈跨域社會化經驗：候用校長教育局處行政實習實踐之分析與反省〉。《當代教育研究季刊》，*25*（3），1–40。[Chen, H.-C. (2017). The cross-field socialization experience: Exploring aspiring principals' administration internship practice in three local education bureaus. *Contemporary Educational Research Quarterly, 25*(3), 1–40.] https://doi.org/10.6151/CERQ.2017.2503.01

張美玉（2000）：〈歷程檔案評量的理念與實施〉。《科學教育》，*231*，58–63。[Chang, M.-Y. (2000). Theoretical and practical aspects of portfolio assessment. *Science Education Monthly, 231*, 58–63]. https://doi.org/10.6216/SEM.200006_(231).0015

張郁雯（2008）：〈國小學童資訊素養檔案評量之信度研究〉。《教育心理學報》，*39*，43–60。[Chang, Y.-W. (2008). Developing portfolio to assess the information literacy of elementary students. *Bulletin of Educational Psychology, 39*, 43–60.] https://doi.org/10.6251/BEP.20090108

張郁雯（2010）：〈國小學童資訊素養檔案評量之發展研究〉。《教育心理學報》，*41*，521–550。[Chang, Y.-W. (2010). Developing portfolio to assess the information literacy of elementary students. *Bulletin of Educational Psychology, 41*, 521–550.] https://doi.org/10.6251/BEP.20090108

張基成、吳炳宏（2012）：〈網路化檔案評量環境下教學者評量之信度與效度〉。《科學教育學刊》，*20*，393–412。[Chang, C.-C., & Wu, B.-H. (2012). The reliability and validity for the teachers' assessment under the online portfolio assessment environment. *Chinese Journal of Science Education, 20*, 393–412.] https://doi.org/10.6173/CJSE.2012.2005.01

張基成、林俊宇（2015）：〈網路化檔案評量系統內反思機制之設計及其對自我調整學習影響之評估〉。《教育與心理研究》，*38*（1），31–64。[Chang, C.-C., & Lin, J.-Y. (2015). Design of reflective mechanisms in web-based portfolio assessment system and evaluation of their influence in self-regulated learning. *Journal of Education & Psychology, 38*(1), 31–64.] https://doi.org/10.3966/102498852015033801002

張基成、彭星瑞（2008）：〈網路化檔案評量於國中電腦課程之使用及成效〉。《師大學報：科學教育類》，*53*（2），31–57。[Chang, C.-C., & Peng, S.-R. (2008). Use and effects of web-based portfolio assessment on computer course of junior high schools. *Journal of Taiwan Normal University: Science Education, 53*(2), 31–57.] https://doi.org/10.6300/JNTNU.2008.53(2).02

張基成、廖悅媚（2013）：〈數位化學習歷程檔案之學習目標設定對自我調整學習之影響〉。《科學教育學刊》，*21*，431–454。[Chang, C.-C., & Liao, Y.-M. (2013). Influences of e-portfolio in self-regulated learning-effect of learning goal setting. *Chinese Journal of Science Education, 21*, 431–454.] https://doi.org/10.6173/CJSE.2013.2104.03

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*，105–128。[Chang, H.-L., & Wu, S.-C. (2008). A multi-facet Rasch analysis on rating the academic scientific papers. *Psychological Testing, 55*, 105–128.] https://doi.org/10.7108/PT.200804.0105

張麗麗（2002）：〈檔案評量信度與效度的分析—以國小寫作檔案為例〉。《教育與心理研究》，

25（1），1–34。[Zhang, L.-L. (2002). Reliability and validity of writing portfolio assessment. *Journal of Education & Psychology, 25*(1), 1–34.]

劉祥熹、陳玉娟、鄭筱慧（2016）：〈學校創新經營對家長選校意願影響之研究－以服務品質與學校形象為中介變項〉。《教育科學研究期刊》，*61*（4），59–88。[Liu, H.-H., Chen, Y.-C., & Cheng, H.-H. (2016). Impacts of school innovation management on school selecting intentions: Service quality and school image as mediators. *Journal of Research in Education Sciences, 61*(4), 59–88.] https://doi.org/10.6209/JORIES.2016.61(4).03

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*，551–566。[Hsieh, M.-C. (2017). Who is a good speaker? applying multifaceted Rasch model to analyze principal three-minute impromptu speech. *Bulletin of Educational Psychology, 48*, 551–566.] https://doi.org/10.6251/BEP.20160801

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*，415–436。[Hsieh, M.-C. (2020). Investigating the effects of rater equating designs on parameter estimates in the context of preservice principal oral performance. *Bulletin of Educational Psychology, 52*, 415–436.] https://doi.org/10.6251/BEP.202012_52(2).0008

謝名娟、林信志（計畫主持人）（2014）：《中小學校長培訓與專業發展評鑑模式之研究》（計畫編號：NAER-101-36-C-1-02-05-1-07）。國家教育研究院計畫成果報告，國家教育研究院。https://www.naer.edu.tw/PageManpower/projectDetail/RP000000000184 [Hsieh, M.-C., & Lin, H.-C. (Principal Investigator). (2014). *The study of training and professional evaluation models of junior and elementary school principals* (Report No. NAER-101-36-C-1-02-05-1-07) (Grant). National Academy for Educational Research. https://www.naer.edu.tw/PageManpower/projectDetail/RP000000000184]

Arter, J. A., & Spandle, V. (1992). NCME instructional module: Using portfolios of student work in instruction and assessment. *Educational Measurement, Issues and Practice, 11*(1), 36–43. https://doi.org/10.1111/j.1745-3992.1992.tb00230.x

Chang, C.-C., Liang, C., Chou, P.-N., & Liao, Y.-M. (2018). Using e-portfolio for learning goal setting to facilitate self-regulated learning of high school students. *Behaviour & Information Technology, 37*(12), 1237–1251. https://doi.org/10.1080/0144929X.2018.1496275

Chang, C.-C., Liang, C., Tseng, K.-H., Tseng, J.-S., & Chen, T.-Y. (2013). Were knowledge management abilities of university students enhanced after creating personal blog-based portfolios? *Australasian Journal of Educational Technology, 29*(6), 916–931. https://doi.org/10.14742/ajet.462

Chau, J., & Cheng. G. (2010). Towards the understanding the potential of e-portfolios for independent learning: A qualitative research. *Australasian Journal of Educational Technology, 26*(7), 932–950. https://doi.org/10.14742/ajet.1026

Coombe, C., & Barlow, L. (2004). The reflective portfolio: Two case studies from the United Arab Emirates. *English Teaching Forum, 42*(1), 18–23.

Diperna, J., & Derham, C. (2007). Digital professional portfolios of preservice teaching: An initial study of score reliability and validity. *Journal of Technology and Teacher Education, 15*(3), 363–381.

Gadbury-Amyot, C. C., Kim, J., Palm, R. L., Mills, G. E., Noble, E., & Overman, P. R. (2003). Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program. *Journal of Dental Education, 67*, 991–1002. https://doi.org/10.1002/j.0022-0337.2003.67.9.tb03697.x

Higher Education Funding Council for England. (2008). *Effective practice with e-portfolios: Supporting 21st century learning.* JISC. https://research.qut.edu.au/eportfolio/wp-content/uploads/sites/186/2018/04/JISC_effective_practice_e-portfolios.pdf

Hughes, J. (2008). E-portfolio-based learning: A practitioner perspective. *Enhancing Learning in the Social Sciences, 1*(2), 1–12. https://doi.org/10.11120/elss.2008.01020005

Joyes, G., Gray, L., & Hartnell-Young, E. (2010). Effective practice with e-portfolios: How can the UK experience inform implementation? *Australasian Journal of Educational Technology, 26*(1), 15–27. https://doi.org/10.14742/ajet.1099

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1995). Portfolios in large-scale assessment Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11–28. https://doi.org/10.1111/j.1745-3992.1995.tb00863.x

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions, 12*(2), Article 636.

Linacre, J. M. (2010). *A user's guide to WINSTEPS: Rasch-model computer programs* [Computer software manual]. Winsteps.com.

Linacre, J. M. (2014). *Facets Rasch measurement* [computer software]. Winsteps.com.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3th ed., pp. 13–103). Macmillan.

Sharma, S. (2007). From chaos to clarity: Using the research portfolio to teach and assess information literacy skills. *The Journal of Academic Librarianship, 33*(1), 127–135. https://doi.org/10.1016/j.acalib.2006.08.014

Singh, O., & Ritzhaupt, A. D. (2006). Student perceptive of organizational uses of eportfolios in higher education. In E. Pearson & P. Bohman (Eds.), *Proceedings of edmedia* (pp. 1717–1722). Association for the Advancement of Computing in Education (AACE).

Skawinski, S. F., & Thibodeau, S. J. (2003). A journey into portfolio assessment. *The Education Forum, 67*(1), 81–88. https://doi.org/10.1080/00131720208984537

Smith Jr., 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*(2), 205–231.

Spencer, L., & Spencer, S. (1993). *Competency at work: Models for superior performance.* John Wiley & Sons.

Tennant, A., & Pallant, J. (2006). Unidimensionality matters (a tale of two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048–1051.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* MESA Press.

收稿日期：2021 年 11 月 03 日
一稿修訂日期：2021 年 11 月 08 日
二稿修訂日期：2022 年 01 月 06 日
三稿修訂日期：2022 年 02 月 03 日
四稿修訂日期：2022 年 02 月 09 日
接受刊登日期：2022 年 02 月 14 日

Bulletin of Educational Psychology, 2022, 54(2), 333–356
National Taiwan Normal University, Taipei, Taiwan, R. O. C.

Portfolio Assessment: Reliability and Validity in School Management

Ming-Chuan Hsieh

National Academy for Educational Research

Research Center for Testing and Assessment

Portfolio assessment provides an opportunity to understand characteristics about a student that cannot be gleaned from responses to traditional test questions; the characteristics obtained through portfolio assessment are more integrated and reflect a more comprehensive picture of a student's learning experiences. Portfolio assessment has become more popular in recent years as a tool for measuring a student's suitability for admission into university. Portfolio assessment allows a learner to self-monitor their progress and achievement of learning goals through a systematic collection of assignments. The role of an instructor is to guide a student in constructing a portfolio, to set learning goals, and to provide opportunities for self-assessment and peer feedback.

Many scholars and experts in Taiwan have built a rich theoretical foundation and performed practical research on portfolio assessment at the university, elementary, and secondary levels; few studies have investigated the use of portfolio assessment in other contexts. This study investigates the reliability and validity of portfolio assessment in the context of school management and, in particular, their use by preservice principals. The influence of scorer severity, scoring rubrics, and the influence of internal or external experts was investigated. In current preservice principal training courses, preservice principals are given many assignments. Although the assignment content focuses on the various aspects of school management, a systematic collection of assignments is lacking. In particular, challenges with practical implementation include the availability of too many assignments and the assignments having overlapping purposes. These disorganized and unsystematic assignments are a burden to both preservice principals and scorers. Based on the six professional competency indicators of preservice principals (visioning, strategic thinking, teamwork, communication and coordination, innovative management, and self-reflection), this study systematically constructs a course that uses portfolio assessment. However, evidence of the validity of the use of portfolio assessment in the context of school management must be further investigated.

Preservice principal training courses cover a variety of topics, including school development, administration, professional responsibility, public relations, curriculum and instructional leadership, educational visitation, teacher learning, liberal arts, and integrated activities; such courses also often incorporate aspects of mentorship. Mentorship learning is designed to help preservice principals finish their assessments. Typically, six mentoring sessions from the senior school principals are provided on the content and practices required in the school development portfolio. In addition to written assignments, individual sharing is required, and mentors provide feedback.

In total, 61 preservice principals who enrolled in the 8-week training course participated in this study. Each student was required to produce a portfolio of assignments related to school management that covered five major topics: (1) school background information, (2) goals and action plans, (3) school curriculum design, (4) school innovation plan, and (5) self-reflection. These five components are a primary focus of school development. Each class had two coaches and two external experts who scored the portfolios. A total of 12 experts were involved in the scoring process. At the start of the course, the coaches of each class introduced the function of the portfolios and explained the key content that must be included in each

section. Students were asked to upload the content of their portfolios two to three times to the system as the course progressed. For example, after the "Data-Driven School Research" seminar, students were asked to complete an inventory of school records, followed by questions and forms to guide them through the process . At each step, the coaches shared, discussed, and advised students before moving on to the next section. An overall grade was given after completion of the entire portfolio.

Reliability was assessed using the multifacet Rasch model (MFRM). Item difficulty and scorer severity were both considered when assessing ability. The validity dimension examined convergent and discriminant validity. Validity was examined by correlating portfolio scores with test scores and other assignment scores, and scores were based on daily observations by the coaches .

The mean square of infit and outfit for each level are close to 1, which represents goodness of fit. In the Rasch model, reliability represents the reproducibility of the measure; ability as measured in a test with high reliability is more likely to be consistent with ability in the real world. Most reliability indices are relatively high; the lowest values occur at the estimation of the ability of preservice principals, which may be due to the high homogeneity and similarity of the preservice principals.

Although the sample size was not large, the fit of the model was acceptable, which means that the data in this study are suitable for analysis using the MFRM. Students performed poorly in creating innovative plans but performed well in school background information, which may be because most of these students tend to come from backgrounds without much opportunity to make decisions on the direction of school affairs. For example, having a background as a director predisposes an individual to being uncertain about the direction of school innovation, whereas it is trivial for a student of any background to present information related to a school's background. The external experts were more severe than the coaches, probably because as mentors, the coaches developed close relationships with the students, and this made it difficult for the coaches to give the students low scores. The final study found a strong correlation between portfolio scores and oral scores. Oral tests were intended to provide an opportunity for students to present their school management philosophies. A significant correlation between portfolio scores and weekly reflection journal scores was also observed and is an indication of validity. A significant correlation between portfolio and behavioral performance scores was also observed, possibly because scoring of both of these items was performed by coaches and may therefore have been affected by the halo effect, resulting in lower discriminant validity.

Although the scorers are experts in the field, they still need to be trained in scoring. In the future, we should design a rigorous scorer training program. For example, we can place some calibration cases in the scoring process to examine rater bias and make adjustments immediately. Exploring how to design a more convenient and easier-to-use model of scorer training to increase consensus among scorers and reduce subjective bias is noteworthy.

In this study, both internal and external experts had different sources of errors in scoring; external experts were more biased when evaluating contextual data. Therefore, in future courses, external experts should be tasked with scoring factual information and internal experts (e.g., mentor principals) should be tasked with scoring contextual information, such as the rationale for or consistency of a portfolio. Exploring how the halo effect might be avoided is another potential area of improvement. Incorporating anonymous cross-class grading or peer-graded comments and suggestions might help avoid the halo effect. Cross-disciplinary collaboration should be considered as a way of enhancing innovative thinking. Team activities, such as brainstorming as a team, may be useful in helping students generate innovative ideas. In conclusion, this study evidenced the reliability and validity of portfolio assessment in the context of school management. However, careful consideration is still required to determine if it is suitable as an examination tool.

Keywords: school management, portfolio assessment, reliability, validity

附錄 檔案作業範例：校務檔案計畫背景說明

本作業請依據候用校長原校之校務資料為基礎脈絡（若無計畫則參考其他校務資料）為基礎脈絡，搭配講座課程來完成。請撰寫學校基本資料、班級數、學生及教師人數、弱勢學生人數表與趨勢折線圖、學校近年與預估未來之班級數、教職員編制、教師學歷統計、師資結構現況與人力進用規劃、校內課程發展相關組織、校內外學習空間盤點、依據前述相關背景資料，描述學校願景（如空間營造、課程教學、教師專業成長、親師溝通與公共關係、學習品保等面向）。其參考格式如下

（一）學校 109 學年度基本資料

學校名稱		學校類型（請勾選）	□一般 □非山非市 □偏遠 □特偏 □極偏		
地址		電話		傳真	
編制內教師數					
班級數及學生人數概況	班級類別	班級數	學生數		
	總計				

（二）班級數、學生及教師人數

1. 近 3 學年（107—109 學年）及未來 4 學年（110—113 學年）之班級數與學生數一覽表（可以折線圖呈現）

學年度	普通班			體育班			特教班			藝才班			合計		
	班級	學生	教師	班級	學生	教師	班級	學生	教師	班級	學生	教師	班級	學生	教師
107															
108															

109															
110															
111															
112															
113															

2. 近 3 學年（107—109 年）弱勢學生人數表、趨勢折線圖與說明

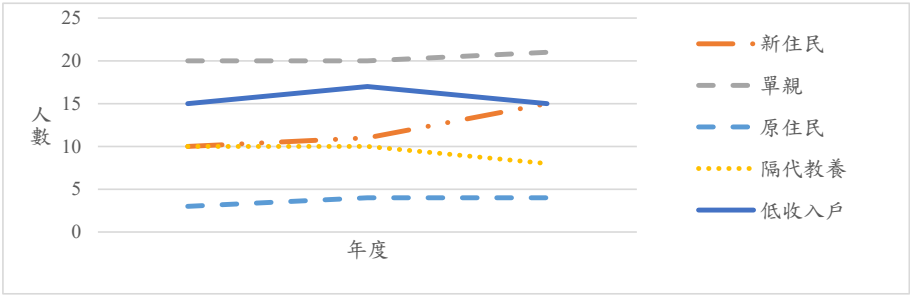
依據學校實際可得資料，分別探討原住民、新住民、單親、隔代教養、低收入戶等弱勢學生人數趨勢變化折線圖（**若學生有雙重弱勢身分，請重複計算在不同類別**）

範例

（1） 近 3 年（107—109 年）弱勢學生人數表

年度 / 弱勢身份	107	108	109
原住民	3	4	4
新住民	10	11	15
單親	20	20	21
隔代教養	10	10	8
低收入戶	15	17	15

（2）近 3 年（107—109 年）弱勢學生趨勢折線圖



（3）說明

3. 教職員編制表

109 學年度教職員編制表								
每班教師員額編制○人								
現行教師編制有教師○○人，含代理代課教師○○人								
人數	校長	導師	科任	兼任主任	兼任組長	資源班教師	行政職員	合計
	1							
備註： （1）學校如有法定編制不足，請說明原因： （2）學校是否屬於學校組織創新： □是 □否								

4. 師資結構現況與人力進用規劃

109 學年度師資結構現況與人力進用規劃				
專長領域	現有教師數（人）	課程需求數（人）	人力現況	說明
國語	8	6	充足	人力充足
特教	2	3	-1	增聘代理教師 1 人

5. 學生學力品質表現（會考成績、學力檢測成績、段考成績等）

6. 校內外學習空間盤點

	編號	場地／品項名稱	數量	狀態	領域／彈性課程融入規劃
現況 近 3 年（至多五項）	1	情境教室	1	堪用	英語：行動學習 彈性：統整性主題／專題／議題（國際教育視訊設備）
	2	生活科技教室	2	建置中	生活科技／社團活動或技藝課程（3D 創課社）

未來 4 年亟需規劃之空間（至多三項）	編號	場地／品項名稱	數量	亟需原因	課程需求
	1	風雨操場	1	學校位處多雨山區	雨天戶外健體課程

7. 請說明學校願景

（1） 學校願景（文字或圖表示）

a. 願景及意涵

b. 願景形塑過程