彰化師大教育學報
民 106，第 31、32 輯，75-93 頁

校長情境式影片評量信效度考驗及應用情形之研究

謝名娟
國家教育研究院測驗及評量研究中心
研究員

教育者常希望使用實作評量，以越接近真實情境的試題來評斷出學生學會的知識或能力，然而，受限於人數與經費，試題要能直接在實作評量的脈絡來操作具有其困難度。本研究團隊以新的嘗試，透過校長儲訓班的期末測驗，將情境試題跳脫傳統文字敘述的模式，而是將文字中的學校情境拍攝成影片，並請學員設身處地為影片中的校長情境來著想，研擬出解決的方式與策略。在受試者接受測驗之後，以問卷方式調查受試者對於影片式情境評量的滿意度與建議，研究結果發現影片式情境評量在做法上具有創新性，但在執行上需注意細節方能達到成效。此外，影片式情境題的作答滿意度和受試者的背景有相關，雖然信效度尚在可接受的範圍，但仍有進步的空間。

關鍵字：情境評量、真實評量、影片題測驗、校長儲訓

謝名娟，國家教育研究院測驗及評量研究中心研究員

壹、緒論

一、研究背景

校務發展與學校領導者角色息息相關，領導者是學校的領航者，能進行學校改革，也主導校務發展方向，顯見校務領導者之重要性。校長是學校的靈魂人物，在學校中需扮演如校務推動、課程發展、教學改革、學校行銷、學習績效及社區溝通等多重角色，其複雜與工作繁重性不言可喻，其角色代表榮譽，也是責任，深受家長及社會大眾的肯定。不過，隨著社會變遷與發展，資訊科技普及多元化社會之來臨，校長勢必與時俱進，引導校務革新，鼓勵教職員工精進專業，以因應社會各界的期待，其任務更具挑戰性。

過去在校長的儲訓課程中，評量內容包括報告（如學校校務發展計畫報告、個案研究報告、生活札記、教育參觀心得、標準作業流程報告）、實作（三分鐘演講、學校行政實習）、紙筆學科測驗、與生活表現等。評量模式雖然很多元，然而，根據郭工賓等人（2015）研究指出，現行的校長評量方式有幾個限制：

第一、雖然作業內容繁多，但非常零散、沒有整體性。而且考試方式未能依未來校長工作所需核心職能進行評估，多侷限在傳統的報告寫作與議題申論為主。

第二、缺乏評分規準，且過多的作業與報告，讓輔導校長很難依據評分規準來進行評分，只能”大概 ”的給一個整體的分數。即使每班都有兩位輔導校長針對成績的部分進行討論，但班級間的輔導校長評分有時也有嚴厲、寬鬆的區別。為了讓每班給分趨於公平，雖建議每個班級採常態分配的給分（例如 90 以上有 10 位、80~89 有 20 位、70~79 有 10 位），然而此作法還是有所限制，某些班級學員的程度可能比另一個班級的學員程度來的好，如此一來，在平均程度較好的班級學員，某程度而言較為吃虧。

第三、學科測驗爭議大。學科測驗在學員成績中，所占的比例最高。然而，學科測驗考研習內容，為傳統的認知型試題。許多學員反應，在考主任、考校長階段其實這些學科內容都已經考過了，如果在儲訓結業階段，還要再考一次這些知識似乎有點重覆，而且會讓學員們花很多時間去背誦記憶方面的知識，同時亦造成儲訓學員的負面壓力。

這些限制中，主要問題為評量的真實性，過去的期末評量多為紙本評量，雖然在題目上盡量以情境式來敘寫，避免傳統背誦、條文式的申論題，但是離真實評量仍有相當距離。

為改良現有之評量模式，研究團隊嘗試使用實作評量的方式來更真實的評估儲訓校長的能力。

改革方案中，原欲使用評量者中心（Assessment Center）的方式來設計評量（Thornton III, Rupp, 2006），但受限於儲訓人數多達 200 人，且在有限的經費與時間下無法聘任多位評分者來進行評分，因此研究團隊採紙本評量與真實評量折衷的模式，情境試題不使用文字敘述的模式，而是將文字中的學校情境拍攝成影片，並請學

員設身處地為影片中的校長情境來著想，研擬出解決的方式與策略，然而，這樣的測驗模式，從題目設計、腳本拍攝、到實施施測之成效，可能之困難與問題是甚麼？值得深入探究。

在研究流程中，研究團隊先與命題者討論命題情境與評分規準，而後與拍攝團隊進行腳本演練，製作影片，經過專家評審委員會審議與修正之後，進行題目的正式施測。為評估影片式情境評量之效益，在受試者接受測驗之後，以問卷方式調查受試者對於影片式情境評量的滿意度與建議，學員在影片式評量的成績亦與其他學習資料進行相關性分析，來評估影片式情境評量的信效度。

貳、文獻探討

實作評量（performance Assessment） 越能接近真實的生活情境，越能評斷出學生學會的技能 （余民寧，2011）。然而，所謂真實性只是程度上的問題，例如想要評量學生日語表達的溝通技巧，最理想的狀況是真正和日本人對話，可是在課堂中，要找到一群日本人來跟學生對話並評量很困難。因此，在真正的執行層面上，只能讓師生、或是同儕間來使用日語來對話來進行評估，雖然在真實性上打了折扣，但是對學生而言，能夠提供一個學生日語口語表達的機會，會比單純使用紙筆考試，更容易讓教師更容易了解學生的口語溝通的學習情形 （王振世、何秀珠、曾文志、彭文松譯，2009）。

實作評量的類型受時間、成本等影響，都有可能限制測驗情境中的真實性，在教學情境中，實作評量可以分成以下五種類型（余民寧，2011），從真實程度較低的紙筆表現、辨認測驗、結構化表現測驗、模擬表現、到真實性最高的工作樣本。

在 1980 年代初期，實作評量被視為具有價值的教育改革方式（Linn, 1993; Resnick & Resnick, 1992; Wiggins, 1989），而其被重視的主要原因則由於現行考試著重在受試者的高層次思考和問題解決的能力，且希望所學得知識技能可應用在現實生活中。例如，美國在國家教育進展評量中（National Assessment of Educational Progress，簡稱 NAEP），將其評量重點擺在評估學生所學習的知識，是否能運用在日常生活中，而其高層次技能的評量，藉由開放式的問題，允許學生使用不同的策略來回答，甚至透過電腦模擬互動，讓學生能連結不同的知識與能力。

情境式評量衍伸自實作評量的理念，以下針對情境評量的基本理念與相關應用與未來加以說明。

一、情境式評量基本理念

情境式評量由 Latham、Saari、Pursell 與 Campion（1980）所引進，一個良好的情境式命題，著重該事件或場景形塑的真實性及代表性，命題者可預先針對自身工作經驗，以研擬該工作可能常發生的事例，作為命題的基礎；此外，亦可搭配職能分析，同時透過問卷調查或訪談校長的方式，逐步搜集各工作相關之事例，經挑選、修訂後，以成為題庫。

Trog（2009）提出情境導向評量問題發展應由課程內容或是溝通過程建構情境，讓學生使用既有的知識來處理。而後期待學生做出選擇，在每個情境中，找出一個人可以做出什麼樣的選擇。請學科專家分享這方面的經驗以及選擇之後的結果，每個選擇都必須實際但卻不明顯。避免出現顯而易見就可以刪除的選擇，這不符合真實生活情境。真實世界需要學生做出某些決定來處理某些議題。在比較複雜的情境中，並不總是需要單一正確答案和一組錯誤選項。提供一些有些對也有些錯的選項，強迫學生找出最佳選擇。選項之間的細微差別，可於稍後透過回饋訊息來加以澄清。

最後在情境評量發展中，須評估選擇的結果為何，每一個選擇都會有一個結果並且提供回饋訊息。為了支援教學之目的，建立一個程序讓學生能獲得解決問題所需的訊息。例如，做一個超連結，讓學生可以連到某個網路資源或是教學影音檔。回饋訊息的部分，可以透過直接的陳述來表達，像是「目前這個過程需要……」以便再加強某些課程內容片段。或者，也可以仿造真實生活情境，建立一組後續情境，讓學生進行更多的選擇。

二、情境式評量的應用

（一）英國政府機關人員甄選

國內外在文官培訓上的做法，也逐漸將實作評量的觀念引進。例如，在英國政府機關招募人員之各種途徑中，有一種途徑較特別、普遍受到歡迎及重視、並要求所有應徵者均須接受測驗，通過所有階段之測驗者，始得取得擔任正式永久職公務人員：快速升遷制度，由公務機關主導，透過此制度來挑選未來政府的主管。測驗內容包括線上的語言和數理能力測驗，電子公文盤測驗 （為申論題性質，受試者在電腦上閱讀一些資料，要求受試者依據情境作決定，目的為測驗受試者的分析、推理、邏輯及論證能力）、One-day Fast Stream Assessment Centre （FSAC）測驗。其中，FSAC 測驗的考科內容包括：（1）團體討論（2）政策建議練習（3）簡報發表（4）面試（劉慧娥，2013）。這些考科內容都均與文官常遇到的情境相關，例如團體討論的目的為測驗應試者是否具備建立生產力關係、產生影響力的溝通、努力達成結果等能力。團體由 5 至 6 位來自不同背景的受試者組成，團體所有成員均須參與並對計劃/方案之討論有所貢獻。在同一團體之不同受試者，會被指定代表不同的立場（觀點，扮演不同角色），受試者的任務包括為自己的立場及整個團體，獲得最佳結果。受試者不僅須堅強地代表自己的立場，亦須仔細聆聽其他人的意見，並與大家協商以獲得共識。為了讓評分盡量客觀，每個團體設有 3 位受過訓練的評分者。

（二）考選部多元化情境口試

考選部（2012a）近年來已將情境評量著墨於人力甄選上。例如考選部所訂立的口試評量原則，將考試種類區分為（1）個別口試：評分規準為依據應考人的儀態、溝通能力、人格特質、才識與應變能力來評判（2）集體口試：兩位應考人以上，使用與個別口試相同的評分規準（3）團體討論：指五位以上應考人輪流擔任會議主持人，並評量其會議主持能力、口語表達能力、組織與分析能力、親和力與感受性、決斷力、及參與討論時的影響力、分析能力、團體適應能力、壓力忍受力與積極性。其

中個別口試的時間較短為 20~60 分鐘，集體口試則為 1~2 小時，團體討論時間則為 2~4 小時。將工作情境帶入考選過程中的方式，期盼能考出學科知識內容以外的實作能力。

（三）臨床技能測驗

實作的能力在實務界的需求已相當風行，例如在 2013 年已經將臨床技能測驗直接納入醫生職照的先備考試中，現在的醫學生必須先通過這個測驗，才能參加國考（曹以會，2013）。臨床技能測驗主要藉由情境模擬實作的歷程，來評估考生應具備的能力，測驗分為 12 站，其中前 8 站是透過標準化病人演出的試題，考生依序到不同的測驗站接受測試，每個測驗站都設定一個情境，病人會有不同的身體狀況來"演出"某種疾病的症狀，考生必須在 15 分鐘中內，來進行問診、身體檢查、溝通衛教等。而後 4 站則是臨床技能的操作題，包括操作醫療器材的準確度與精確度等。

在這個測驗中，其主要的評量向度包括與病人溝通、為病人看診的態度，以及面對病人時能否表現出良好的態度與互動能力，透過這些向度，來當作評估醫學生是否合適擔任醫生的標準。雖然實施的成效還需評估，但可看出醫學界已相當重視使用實作評量，來進行評選適合的人才。

（四）警察特考之情境測驗

警察特考屬於紙筆式的情境評量，將情境訴諸文字敘述。首開先例於三等考試設置「警察情境實務」、四等設置「警察情境實務概要」考科，在警察特考情境測驗題型包括 20 題選擇題（每題二分）、3 題申論題（每題二十分），例如在選擇題的部分，三等考試「警察情境實務」第十四題（考選部，2012b）中詢問考生：

> 「你是偵查隊分隊長，率隊依法搜索竊盜通緝犯藏匿處所，搜索過程中你發現天花板有一上鎖之鐵門，疑似可通向天花板之密室，屋主表示該門久未使用，鑰匙已遺失無法開啟，此時你應如何處理？
> A. 請來鎖匠，開啟鐵門，執行搜索
> B. 請示檢察官並經同意後，請鎖匠開門，執行搜索
> C. 請示法官並經同意後，請鎖匠開門，執行搜索
> D. 繼續搜索其他房間，不搜索天花板上之密室」

這樣的選擇題試題可具體看出情境測驗有別於一般專業科目的特色（吳斯茜，2014）。

三、情境式測驗的問題與因應之道

（一）情境式評量須考量評分規準制定之嚴謹性

余民寧（2013）指出，現行國家考試，許多都使用同一份評分規準來評分，例如在口語評量中，司法官的口語評量標準，和公務員的口語評量標準居然是相同的。這樣勢必會造成評分的偏差，制定評分規準時，必須要找到職能的核心能力，工作經驗、或所需條件來擬訂評分規準較為適當。另外，口試委員除了須有專業知識之外，必須接受足夠的訓練，才能達到公平、公正、客觀的程度。

（二）情境式評量在教育上的應用可更多元

過去的教育類的口試較少透過情境式評量的方式，進行職能分析，訂立評分規

準，並依據規準來評。若要推行到教育類考試中，可以有幾種作法，例如在校長儲訓班，由於學員人數較多，無法透過實作評量的方式，能做到的還是僅限於紙筆測驗。可以的做法包括增加現場實務教學的情境題，要求考生依據某種理論，撰寫出相對應的教案，或是在教學現場出現了甚麼狀況，應使用哪種教學理論來回應。而在校長儲訓班小班教學時，由於人數較少，則可設定幾個情境式的狀況劇，例如師生衝突、校園安全等實務上的狀況題，並依據規準來進行口試，以期能了解儲訓學員是否理解學教育現場的問題。

（三）可運用科技來輔助情境式評量的進行

在情境式評量的應用上，最常面臨的是成本問題。例如情境模擬口試，需要應試者的錄音與口試委員的評分，常常需要花掉大量的成本。科技化的時代也許可以嘗試引進電腦輔助口試方式，以降低成本，並減少人為的評分誤差；例如：採行視訊口試方式、虛擬實境的操作演練方式、或非面對面錄音口試評量方式等，尤其現在moocs或是線上互動的數位課程相當風行，然而，如何運用在高風險的考試，還需未來研究來評估。

四、影片評量

影片評量是情境評量的其中一種，雖然Thorndike（1945）曾提及使用這種擬真的情境題具有其效益性，但受限於多媒體的發展，真正大規模的應用要在1990年以後。透過多媒體的傳達，受試者能夠透過演員的語言感染、面部表情與行為來讓受試者身歷情境（郭生玉，2010；Drasgow, Olson, Keenan, Moberg, & Mead, 1993），和紙本的情境評量來比較，更能夠降低不同語言、年齡、種族的次團體所造成的測驗誤差，另外，Olson-Buchanan與 Drasgow（2006）、Bauer、Truxillo、Mack與Costa（2011）指出大多數的受試者喜歡影片評量，而且具有較高工作表現預測效度。

在過去的研究指出影片評量具有其優勢，但仍面臨不少挑戰，Olson-Buchanan及Drasgow （2006）指出影片評量的試題發展比傳統紙筆式的評量模式困難，另外，播放影片的相關設備亦可能增加施測的困難度，而拍攝影片的成本也相對較高，若有情境變遷的內容，甚至需重新拍攝。計惠卿（2015）則指出即使影片評量較為擬真，但仍無法取代真實的工作場域，而受限於時間，可能必須簡化許多真實情境中的複雜情節，另外，部分的受試者可能在接受影片評量時，關注到無關的細節，而影響到評量效度。蔡璧煌、范勻蔚（2015）將影片評量運用在訓練公務人員晉升訓練中，則發現影片評量雖然比書面情境題具有較高的鑑別度，但是卻可能會影響高層次認知能力的評量，另外，在命題部分，同時需要聚焦相關情境與融入不相關的情境當作陷阱，因此在命題上面會有許多困難，評分也不容易（吳思茜，2014），需要更多的專家學者協助以提高命題品質。

參、研究方法

一、研究架構

本研究針數位課程內容，研發評量工具，包括擬真情境題的腳本撰寫與測試，並進行試題的施測與信效度評估。研究流程包括確認研究的任務和目的，而後依據儲訓校長的專業能力指標進行內容分析擬定命題架構與評分規準。施測後使用問卷的調查並訪談學員與命題者關於情境式影片評量的想法與建議。研究流程如圖1。

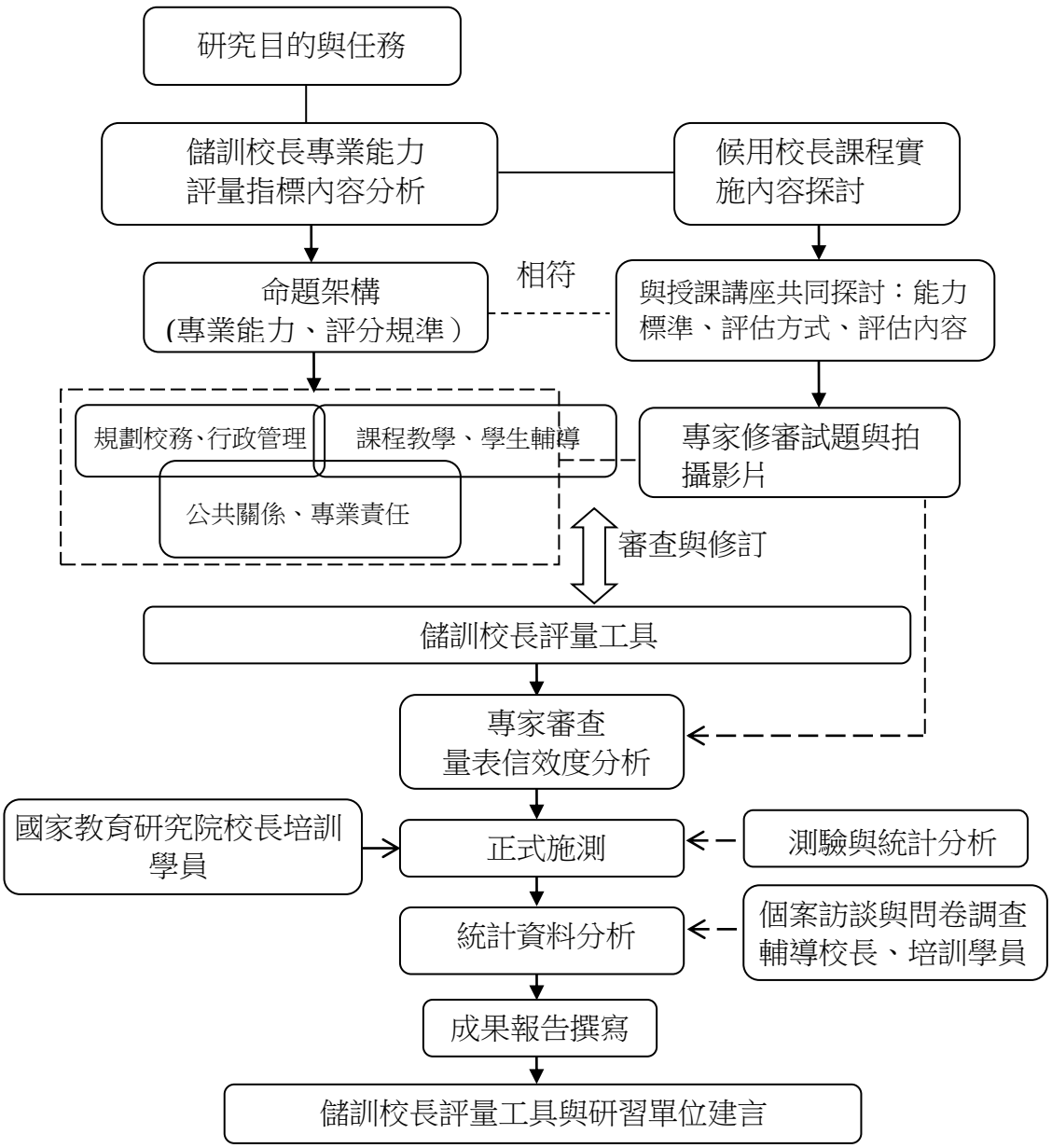

圖 1 研究流程圖

二、影片拍攝

情境式評量的擬真試題研發分為幾大工作內容，分別為：召開研究小組會議，進行試題內容發想、對話腳本撰寫、拍攝與影片製作、專家審查會議與修正等工作，以下就實際執行之內容與結果列舉說明：

（一）召開研究小組會議

本研究邀請具有實務經驗的資深校長與研究人員，擔任試題發想小組。配合儲訓校長的線上磨課師（Moocs）課程，設計校長在校園常遇到的困難與抉擇。試題的發想方向為： 甚麼樣的內容為校長常需要面對的情境？哪些是校長常遇到的衝突點？哪些是可能遇到的問題與困難？校長在社會潮流（如少子化）學生面、教師面、社區面常會需要處理的面向有哪些？

透過多次會議的凝聚共識，並與研究群討論，檢視題目設計是否恰當？經研討修訂後定案。定案後的題目大綱與拍攝導演討論劇情內容，而後由影片製作團隊撰寫腳本內容，並由研究團隊再次確認與題目大綱的符合度。

（二）製作影片

（1）研訂影片製作規格及委外規劃：研究團隊根據需求與計畫，與拍攝團隊溝通媒體製作規格。

（2）拍攝前協商：負責演出的校長與教師配合預擬拍攝日程及內容，再邀集拍攝小組、研究小組共同討論，拍攝之分場、人物與內容，並經由研究小組會議討論通過。

（3）協請學校協助錄影事宜：由拍攝小組洽請錄影學校配合支援相關事宜。

（4）準備拍攝相關道具：製作拍攝中需要使用的道具，如檔案匣、公文、海報、施工圖等資料。

（5）拍攝錄製：依據拍攝工作規畫，分別於學校校園、教室、公園、及社區進行錄製，每一情境題約 2~3 週完成。

（三）審查與修正

經過影片後製，由研究團隊先確認內容正確性，再經由專家審查會議評估影片製作的適切性。影片拍攝共完成四大主題，共計 10 題影片式情境題，由於題目將用於校長儲訓班的期末測驗，其試題具有機密性，因此在無法進行預試。本研究邀集具有校長實務研究經驗與影片拍攝的專家進行影片審查，並提供修正的建議，修正後之影片以密件方式交付首長審閱，以做為校長儲訓班之期末試題勾選參考。

三、問卷調查

調查研究（investigation method）是指經由標準化過程收集有關樣本的具信度和效度的資料，以從事統計推估或驗證假設的方法（瞿海源主編，2007）。在評估情境評量的效益部分，則發放問卷來蒐集學員對於影片式情境評量的想法與意見，問卷編製後經由 3 位專家效度之評定與修正，有效問卷為 173 份。除了詢問學員關於影片式情境評量的整體滿意度之外，並詢問受測學員關於影片的題意、音量、字幕、對話情境、長度、答題時間等滿意程度。表 1 呈現學員的相關背景資料，以男性學員為主，學歷多為碩士，在各處室服務的經驗中，以曾服務過教務處的學員最多，其次為學務

處，服務年資多為 16 年到 25 年，服務階段以小學較多。

表 1
學員背景資料

背景變項	類別	次數	百分比
性別	女	57	32.9
	男	116	67.1
最高學歷	大學	6	3.5
	四十學分班	2	1.2
	碩士	147	85
	博士	18	10.3
是否服務教務處	否	26	15.0
	是	147	85.0
是否服務學務處	否	39	22.5
	是	134	77.5
是否服務總務處	否	51	29.5
	是	122	70.5
是否服務輔導處	否	97	56.1
	是	76	43.9
服務年資	10~15 年	31	17.9
	16~20 年	68	39.3
	21~25 年	52	30.1
	26 年以上	22	12.7
服務階段	中學	41	23.7
	小學	132	76.3

四、評分架構

經由校長儲訓單位首長勾選後，結業測驗筆試部分共有 4 題，其中 2 題為影片式的情境評量，2 題為傳統申論題，每題各佔 25 分。兩片影片評量於教室投影設備公開播放，播放後各預留三十分鐘供學員填答，填答完影片題之後再填寫兩題傳統申論題，答題時間也各為 30 分鐘，總共的測驗時間為 2 小時。

針對情境評量的測驗試題，每題評分審查重點分為四個部分：論點、內容、思辨與表達。每個部份又分成卓越、精熟、基礎與未達基礎等四個等級。在這四個部分中，內容佔的比重為 10 分，其他各佔 5 分。考生共有 209 人，每題由兩位校長獨立評分，而後計算其皮爾森相關係數以評估其評分一致性。評分規準表如表 2。

表 2

情境評量評分規準表

審查重點	符合要點程度與建議配分			
	卓越	精熟	基礎	未達基礎
論點的架構具系統性，前後邏輯一致（5 分）	論點的架構具完整系統性，前後邏輯相當一致 5	論點的架構略具系統性，前後邏輯一致 4	前後尚具邏輯性 3	論述不具邏輯性 2 以下
內容完整豐富，掌握核心概念（10分）	內容切合題意，提出完整的核心概念 9~10	內容切合題意，提出部分核心概念 8	內容能切合題意，略能論述核心概念 7	內容不切題意 6 以下
思辨及創見（5分）	能用獨特的觀點貫穿整題，提出能解決問題，且具有可行性的看法 5	能用課堂中所提到的觀點貫穿整題，提出能解決問題的看法 4	能用課堂中所提到的觀點，但提出之策略未能完整解決問題 3	提出之觀點不具可行性 2 以下
文字表達能力（5分）	文字流暢、表達清晰、引用適切（著名經典、理論等） 5	文字流暢、表達清晰，但沒有引用 4	文字尚流暢 3	文字不流暢 2 以下

肆、研究結果

一、專家效度

儲訓校長的能力指標具有六大主題，包括規劃校務、行政管理、課程教學、學生輔導、公共關係、專業責任等，從中選取適合的主題與可拍攝成為試題的內容，請資深的命題校長進行題目的發想與腳本撰寫。腳本經過專業導演的評估後進行拍攝，形成評量工具。問卷則調查學員在進行情境評量的時各向度的滿意度。評量工具與問卷均經過多次專家的審查與校對後，才進行正式施測。

二、問卷部分

研究團隊雖然拍攝了四類的評量主題，但在期末測驗時，僅呈現其中兩片題目（校園建築與空間美學營造、學校特色經營與行銷）。在五點李克式問卷中，學員的平均滿意度為 3.98，標準差為 0.83，其中庫李信度為 0.96，刪除空白及無效之問卷後，有效填答份數為 173 人，為避免作答者認為問卷填答狀況會影響學員本身的期末成績，所有意見均為匿名填答。

（一）全題學員

表 3 呈現整體學員，針對這兩個影片試題在各面向的滿意程度，其中最後一題為「我不需要看影片，就能直接寫題目的答案」為反向題。可看出在學校特色經營與行銷部分，大致在各面向的平均滿意程度略高於校園建築。

表 3
學員對於影片試題的滿意度

問卷題項	校園建築與空間美學營造		學校特色經營與行銷	
	平均數	標準差	平均數	標準差
我覺得影片中的題意很清楚	4.13	0.75	4.15	0.74
我能應用上課所學來回答影片的答題內容	4.10	0.70	4.12	0.69
我覺得影片的畫質令人滿意	4.30	0.59	4.31	0.57
我覺得影片的音量適中	4.29	0.55	4.34	0.55
我覺得影片的字幕清楚	4.31	0.55	4.31	0.58
我覺得影片中的演員的對話內容貼近學校真實情境	4.21	0.68	4.28	0.67
我覺得與紙本的傳統申論題比較，影片讓我更理解題意	4.08	0.82	4.10	0.83
我覺得影片的長度適當	4.21	0.72	4.23	0.66
我覺得答題時間足夠	4.05	0.89	4.01	0.96
我不需要看影片，就能直接寫題目的答案	2.99	1.15	2.95	1.21

（二）不同背景對期末測驗採用影片情境式作答的滿意度差異

表 4 根據不同性別、最高學歷、是否服務過教務處、學務處、總務處與輔導處、服務年資與服務階段對於期末測驗採影片情境式作答的方式滿意度進行差異比較：

1. 性別、服務年資與處室與階段
   不同性別、不同服務年資、不同服務的處室與不同服務階段均無顯著差異。
2. 最高學歷
   大學學歷在影片評量的整體性滿意度顯著低於其他學歷的學員，若再深入分析其他選項，可發現「我能應用上課所學來回答影片的答題內容」顯著低於四十學分班、碩士與博士學位的學員。其他則無顯著性差異。

表4

學員對於期末測驗採影片情境式作答的方式整體滿意度之差異性統計

背景	類別	個數	平均數	標準差	*F* 值	*p*
性別	女	57	3.93	.75	0.31	0.58
	男	115	4.02	.87		
學歷	大學	6	3.33	1.21	2.84	0.04*
	四十學分	2	4.00	1.41		
	碩士	147	4.14	.72		
	博士	18	4.33	.59		
服務年資	10~15 年	31	4.13	.62	.75	0.78
	16~20 年	68	4.13	.86		
	21~25 年	52	4.12	.68		
	26 年以上	22	4.18	.73		
是否服務教務處	無	26	4.00	.94	1.62	0.21
	有	146	3.99	.81		
是否服務學務處	無	39	3.87	.94	1.61	0.20
	有	133	4.02	.81		
是否服務總務處	無	51	3.82	.77	2.31	0.13
	有	121	4.06	.85		
是否服務輔導處	無	96	3.97	.88	0.05	0.82
	有	76	4.01	.77		
服務階段	中學	41	4.10	.63	0.12	0.73
	小學	132	4.14	.78		

註： *p<.05

（三）焦點座談

焦點座談的對象為命題的資深校長與校長儲訓學員。訪談內容主要分為兩大項，其一為瞭解影片式情境評量的特色與優勢，其二為實施現況中所遇到的困難。以下分別敘述之：

1. 特色與優勢

(1) 容易理解、貼近學校情境

影片題比傳統文字敘述更容易理解，身歷其境，且影片內容和學校情境很相近，覺得這樣的考試方式相當創新。未來若能推廣到教育現場應有其意義，此外，在學校許多的危機處理事件，都是只發生一次，校長必須當機立斷的來下決策，透過影片式的試題，能更真實的反應學校現況，能讓知識、理論能轉化於情境中發展可行策略。

(2) 緩和緊張

透過影片演員對話的方式，讓題目更輕鬆有趣，也能讓受試者清楚的了解題目的脈絡，可緩和其緊張的情緒。另外，在影片中，演員互動的表情、聲音、音響的輔助，反倒能適度的提醒受試者影片中的重點與問題所在。

(3) 方式具有前瞻性與創新性

這次評量是很特別的、令人印象深刻的評量方式，未來可以學習、應用其他教育情境中（如教師專業成長、研習等），另外，這種評量方式也富即時回答的挑戰性，可讓學員更容易激發思考與想像。

2. 困難與挑戰：

(1) 可用文字書寫取代影片

每個情境影片最後都有題目，從題目的敘述中也可以直接答題。實際執行時，為了能使學員更聚焦，採先由於先發放題目，再讓學員答題，有學員反應因先公布題目，所以看影片後反倒會干擾回答架構。另外，若是用詳細的文字書寫，似乎也可以取帶影片 ，且更為省錢省力。

(2) 答題時間不夠

為了能夠清楚的表達題意，且在拍攝過程中，須考量到情節的連續性與完整，每個影片約占五分鐘，然而，每題的考試時間為 30 分鐘，看影片花了 5 分鐘，會壓縮到答題時間，學員建議未來應該要將放影片的時間單獨計算，讓學員答題的時間可以更充裕。

(3) 情境若無法清楚表達題意，或觀賞者會錯意，可能影響作答

影片題不像紙本情境題，看一遍不懂可以反覆看，若答題者恍神沒注意影片中的線索，很容易會錯意導致答非所問，因此在作答時，要更認真專注。

(4) 拍攝與命題之間的連結性具有挑戰性

在本研究中，命題的為資深校長，拍攝的為專業導演，然而，導演並不了解校長的情境脈絡與專業用詞，因此在溝通協調中，常常會出現腳本、拍攝與剪接之間的落差，由於在演員部分，要能找到適合的對象來演校長不容易，未來若是要朝影片式的情境評量做長期發展，應考量在命題部分委由外面的校長或專家來命題，但在演員與剪輯部分可成立長期的拍攝團隊來建立默契較為適當。

（四）學科成績測驗

期末學科測驗總共有四題，有兩題影片題（A 與 B），兩題傳統申論題（C 與 D），其中申論 C 屬於傳統理論性的申論題，而 D 屬於以文字敘述句有情境脈絡下的申論題。

以下針對學科測驗的成績來做分析。

1. 性別差異

透過獨立樣本 t 檢定，表 5 呈現這四題中，第一題影片題女性表現略優於男性，但其他題的表現均沒有顯著差異。

表 5
不同性別在學科測驗的表現

題項	性別	個數	平均數	標準差	*t*	*p*
影片 A	男	117	20.71	1.23	-3.68	0.00**
	女	59	21.33	1.28		
影片 B	男	117	18.55	1.48	-1.59	0.11
	女	59	18.93	1.55		
申論 C	男	117	17.36	1.93	-0.33	0.74
	女	59	17.68	2.00		
申論 D	男	117	20.88	1.57	-0.63	0.53
	女	59	20.85	1.74		

註： $^{**}p < .01$

2. 申論題與影片題相關
表 6 可看出兩個影片情境題之間相關性最高且達顯著，而兩題影片情境題均與題目 D 的紙本申論題相關性達顯著，但與題目 C 相關性較低。

表 6
傳統申論題與影片情境題的相關性

	影片 B	申論 C	申論 D
影片 A	0.39**	0.02	0.19*
影片 B		0.03	0.16*
申論 C			0.03

註： $^{*}p<.05$; $^{**}p < .01$

3. 與其他成績的相關性
最後則檢視兩個情境式影片題和各項成績的相關，從表 7 可看出不管是影片題或是傳統申論題，成績均和結業成績達顯著相關；此外，影片 B 則與研習札記成績顯著相關。影片 B 和研習札記達顯著相關的可能原因為影片 B 內容涉及學校特色的發展，而研習札記的撰寫，為學生到各校參訪，記錄其學校特色，因此和學校特色的課程內容本身就具有高度相關的內容。

表 7
影片情境題與其他成績的相關性

	影片 A	影片 B	申論 C	申論 D
校務發展計畫	0.01	-0.06	-0.08	0.07
研習札記	0.14	0.21**	-0.04	-0.02
結業成績	0.59**	0.62**	0.45*	0.42*
學校行政實習	0.07	0.03	-0.03	-0.06
口語表達實務演練	0.11	0.07	0.04	0.08

註： *p<.05; **p <.01

4. 評分者信度

經分析兩位獨立評分校長的皮爾森相關係數，如表 8 可看出，在影片 A 部分，架構部分的相關係數較低，且沒有達到顯著。在其他部分相關均達顯著，但相關性不高。

表 8
兩位評分者的皮爾森相關係數

向度	影片 A		影片 B	
	相關	*p*	相關	*p*
架構	0.08	0.25	0.36	0.00**
內容	0.293	0.00**	0.33	0.00**
思辨	0.262	0.00**	0.21	0.00**
表達	0.148	0.03**	0.30	0.00**

註： **p <.01

伍、結論與建議

本計畫的主要研究目的為檢視影片式情境評量試題研發的過程與施測、信效度的評估。在研發情境評量工具中，撰寫了擬真情境題的文字腳本、並將腳本拍攝成影片、根據專家審查會議的決定來修正影片內容。 為了解情境影片的實施成效，於校長儲訓班期末測驗時挑選其中兩部進行施測。期末並發放問卷與進行焦點座談，以評估受試者對於情境評量的滿意度。本研究的結論如下：

1. 影片式情境評量在做法上具有創新性，但在執行上需注意細節方能達到成效。

學員在影片情境評量的整體滿意度達 3.98，但標準差為 0.83，代表雖然大多數的學員對於這種新式的評量感到肯定，但是亦有學員認同度較低。其中較低的原因可能是因為影片式的情境評量在撥放題目的時間上所需較長，且限制只能看一遍，若是受試者稍沒注意，就會錯過細節，造成答題方向的誤差。而紙本的文字情境題，則可以

反覆推敲，在讀題的時間上也可以較為節省。然而，認同度高的學員則反應在學校的情境中，往往也只會發生一遍，許多訊息稍縱即逝，而影片中的情境和學校常發生的事件相符度高，能感同身受，因此對這種新型態的作答題方式感到創新與肯定。

2. 影片的情境式試題信效度均在可接受範圍

除了學校建築中的架構部分相關未達顯著之外，其他部分均達顯著相關，然而，其相關性並不高，多為.1~.3 的區間，代表兩位評分者給分之間的一致水準仍須加強。兩個情境題成績彼此相關性高，但是和申論性的紙本題相關性較低。在本次抽中的兩個影片式情境題，彼此之間的成績相關性高且達顯著，且與文字敘述申論題也算高，但與傳統性的申論題紙本題相關性相對較低，另外，和學員的其他成績做比較，在結業成績部分有達顯著相關，但與校長儲訓班其他作業相關性較低，代表就效度而言，雖達到一定的程度，但仍需持續蒐集更多影片式情境評量的更多效度證據。

本研究提出的建議如下：

1. 影片式情境試題應成立題庫與專業命題團隊。

影片題和紙本申論題雖然相近，但是影片題由於有對話、有情節、有演出與拍攝，在整個設計流程上比傳統的紙本情境體來的複雜許多。從研究結果來看，使用情境式的影片評量方式，要能夠有效的達到評量效果，受到影片本身品質的影響，而影片拍攝品質的好壞，常受細節的影響（如腳本內容、音量大小、配樂是否適切、演出是否自然）而受限。 為確保未來的題目的專業水準，應成立專業團隊，包括演員訓練、影片剪輯、配樂與字幕都能掌控品質。在命題部分，可考量經過專家命題與修審的程序後，再找一群經過訓練的演員來進行腳本演出。

2. 紙本情境題和影片情境題的相似性與相異性值得深入探討

受限於考題的範圍分布與測驗公平性的考量，在本次測驗中無法將受試者隨機分成兩組，使用同樣一個情境內容，一組使用紙本、一組使用影片來測驗。為深入比較紙本題和情境題的相似性與相異性，未來應可考量以實驗設計的方式來進行比較。

3. 可考量不同的題目範圍來探討學員背景對影片式情境題的影響

本次的研究發現，曾有服務過總務處與學務處的經驗的學員，對於影片式的情境評量認同度較高，其可能原因是首長抽到題目的內容，偏重特色經營與學校建築等議題，都和總務處和學務處所承辦的相關業務較為相關。未來可考量不同的題目範圍，例如性平、或是課程領導等範圍，來探討學員背景對於影片式情境題的影響程度。

4、加強評分者訓練

在本研究中，評分者一致性的相關性不高，其可能性由於經費的問題，要求評分者在一個下午，需評完近 200 份考卷，雖然每份卷子中只需評一題，但由於學員撰寫內容相當豐富，且寫作風格差異頗大，造成評分者在評分時相當的難以抉擇。尤其此為新式題型，過去評分者並無評過類似的申論題，因此造成評閱經驗的不足。未來建議應加強校長的評分訓練，或至少使用兩到三次的閱卷會議，來進行試卷的評分作業。

參考文獻

一、中文部分

王振世、何秀珠、曾文志、彭文松（譯）（2009）。**教育測驗與評量**（原作者：Linn, R. L., & Miller M. D.）。臺北市：雙葉。（原著出版年：2005）

考選部（2012a）。**警察情境實務（包括警察法規、實務操作標準作業程序、人權保障與正當法律程序）**，2014 年 5 月 1 日取自 http://wwwc.moex.gov.tw/ExamQuesFiles/Question/101/101080_50130（3501）.pdf。自：http://mag.udn.com/mag/edu/storypage.jsp?f_ART_ID=459982Power By udn.com

考選部（2012b）。**考選法規彙編**。臺北市：考選部。

余民寧 （2011）。**教育測驗與評量：成就測驗與教學評量**（第三版）。臺北市：心理。

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

吳斯茜（2014）。國家考試多元評量之實踐：以警察特考情境測驗為例。**文官制度季刊，6**（1），81-97。

計惠卿（2015）。促進遷移的模擬教學。**研習論壇，172**，14-23。

曹以會（2013 年 6 月 9 日）。OSCE 醫學臨床測驗今年正式舉辦 近 99%及格。**聯合報**。取自 http://news536c.blogspot.tw/2013/06/osce-99.html

郭工賓、洪啟昌、蔡進雄、黃能富、林信志、謝名娟…劉金和（2015）。**中小學校長專業發展之研究：以國家教育研究院為例**。國家教育研究院自行研究計畫（編號 NAER-104-36-C-1-02-00-1-07），未出版

郭生玉（2010）。**教育測驗與評量**（第三版）。臺中市：精華。

劉慧娥（2013）。英國政府快速升遷制度之介紹及分析：我國人才甄選可參考借鏡之處。**國家菁英，9**（2），163-209。

蔡璧煌、范勻蔚（2015）。影片評量在訓練評量之發展與運用：以薦任公務人員晉升簡任官等訓練案例書面寫作為例 。**人事月刊，358**，1-11.

二、外文部分

Bauer, T. N., Truxillo, D.M., Mack, K., Costa, A.B. (2011). Applicant reactions to technology-based selection. In N, Tippins, & S. Adler (Eds.) *Technology enhanced assessment of talent*, (pp. 1-18). San Francisco, CA: John Wiley & Sons, Inc.

Drasgow, F.,Olson, J. B., Keenan, P. A., Moberg, P. J., & Mead, A. D. (1993). Computerized assessment. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management*, Vol. 11 (pp. 163-206). Greenwich, CT: JAI press.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.

Linn, R. L.(1993). Educational Assessment: Expanded expectations and challenges.

Educational Evaluation and Policy Analysis, *15*, 1-16.

Olsom-Buchanan, J. B., & Drasgow, F. (2006). Multimedia situational judgement tests: The medium creates the message, In J. A. Weekley (Eds.), *Situational judgement tests-theory, measurement, and application,* (pp. 253-275). Hillsdale, NJ: Lawrence Erlbaum Associates.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.G. Gifford & M.C.O's Conner (Eds.). *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-55). Boston, MA: Kluwer Academic.

Thorndike, R. L. (1945). *Personnel selection*. New York, NY: Wiley.

Thornton, G. C., III, & Rupp, D. R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Trog, K. (2009). *Three steps to develop scenario-based assessment questions.* Retrieved from http://tbd-consulting.typepad.com/blog/2009/10/three-steps-t-1.html

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa*, *20*, 703-713.

An Application of Reliability and Validity of Situated Assessment on Video Mode in the Context of School Principals

Mingchuan Hsieh
National Academy for
Educational Research

Performance assessment is commonly used for education researchers to evaluate student's knowledge or ability. However, due to the limitation of resources, it is difficult to conduct performance assessment directly. Instead of using the traditional written language to describe the test items in the final exams of pre-service principals training programs. We tried to use the role play on videos to visualize the test items. The pre-service principals need to see the videos and then answer the questions. After the pre-service principals finishing the test, we surveyed their satisfaction and suggestions of this type of situated assessment. The results show that situated items displayed via video is quite creative, however there are some details need to be taken care of. In addition, the satisfaction of this type of assessment is closely related to the testee's background. Although the test result show an acceptable reliability and validity evidence, there is still room for improvement.

Keywords: Performance Assessment, Pre-Service Principal Training, Situated Assessment, Test items on Videos

Mingchuan Hsieh, Associate Research Fellow, National Academy for Educational Research