活化測驗方式的另一個選擇—實作評量

➢ 謝名娟

活化測驗方式的另一個選擇—實作評量

謝名娟
國家教育研究院副研究員

壹、前言

十二年國教即將上路，教育部長蔣偉寧曾指出-十二年國教的重點不在考試，而在學生是否能主動學習。因此，老師除了發展自己的特色課程、特色教學之外，也應思考如何使用有別於傳統的紙筆考試，以刺激學生主動學習的能力。尤其是重視操作能力的世代，有些能力是紙筆測驗較不容易進行施測的。例如，報導（曾薏蘋，2012）指出，透過高考三級進用的電機工程人員，去鄉公所任職時，完全不知道應該如何修檢發電機；也有醫事技術類別的公務員被派去醫院進行業務督導，卻因為沒有證照，被當地醫院認為是外行領導內行。因此，目前的公務人員考試制度偏重於紙筆考試，而非實作的能力，即使有些考生很會寫題目，但不見得會動手操作。在現今的教育現場，考甚麼人們就重視甚麼，由於我們的考試並不重視實作的能力，只重視傳統紙筆測驗的能力，因此電機工程人員不會修電機，也是意料之中的事。

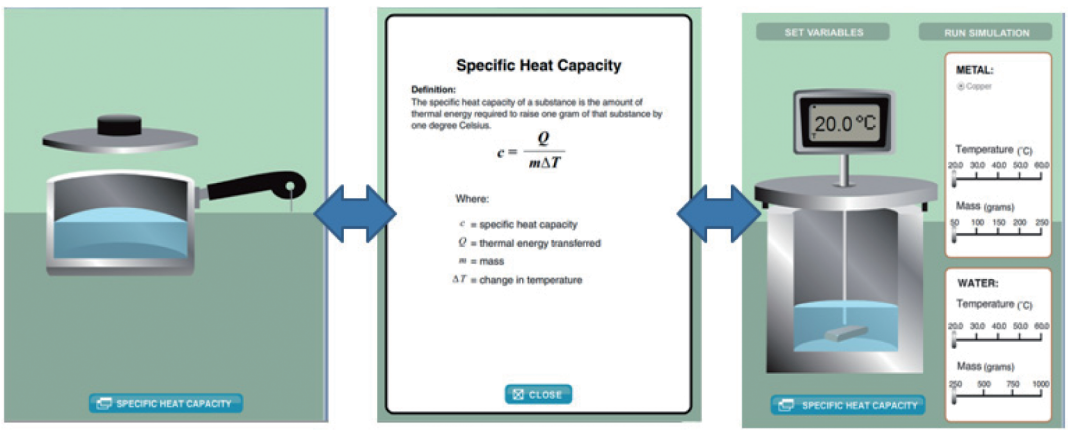
在1980年代初期，實作評量被視為具有價值的教育改革方式 (Linn, 1993; Resnick & Resnick, 1992; Wiggins, 1989)，而其被重視的主要原因則由於現行考試著重在受試者的高層次思考和問題解決的能力，且希望所學得知識技能可應用在現實生活中。例如，美國在國家教育進展評量中 (National Assessment of Educational Progress，簡稱NAEP)，將其評量重點擺在評估學生所學習的知識，是否能運用在日常生活中，而其高層次技能的評量，藉由開放式的問題，允許學生使用不同的策略來回答，甚至透過電腦模擬互動，讓學生能連結不同的知識與能力。例如在科學的測驗中，相當著重學生動手做實驗的能力，受限於空間經費，很難有真實的器材供給每個學生進行實驗。因此會用模擬的情境來測試學生操做實驗的能力，如圖1的範例所示，要請學生回答溫度、水分子與銅原子變化的一些問題。在操作過程中，須設定銅塊的質量（圖中銅塊的大小隨著質量的改變而增大或縮小）、銅塊的溫度、水的質量 （水的質量改變，圖中的水位隨之上升或下降）、設定水的溫度 （水的溫度改變，圖中溫度計顯示數值同時配合改變）、模擬實驗儀器的數據輸出說明，來執行並找出實驗數據 。而後，透過相關問題，如：

根據你模擬過程得到的數據，銅塊和水的比熱何者較大？

A. 銅塊

*B. 水**

解釋你選擇的選項，並針對你模擬得到的數據作為解釋的依據。



這個任務中，你將研究製作平底鍋時，鍋底部分最適合的金屬材料爲何。爲了充分加熱食物，烹調過程中鍋底部分將加熱至高溫。最適合的金屬材料是外界輸入熱能時，金屬可到達的溫度較高。

比熱(Specific heat capacity)爲挑選鍋底金屬材質時考慮的特性之一。這個任務中，你將研究二種可能用於製造鍋底的金屬材料的比熱。

任何時間點選 SPECIFIC HEAT CAPACITY 可顯示比熱的定義及方程式

圖爲熱量計模擬實驗。二物質接觸時的熱能轉換可用熱量計量測。熱量計經過特殊設計以避免系統內部與外在環境間的熱能轉換。

將不同溫度的放入熱量計內部時，可以嘗試操作模擬的熱量計，研究熱量計中水的溫度變化量。

圖 1　NAEP 科學測驗範例試題

實作評量將學生從被動的、被給予的角色，轉換成主動的、積極的學習者。傳統的選擇題試題具有標準答案，選到正確的答案才能得分，選到不正確的答案則失分，而在實作評量中，許多答案都可以是正確的，完全是看學習者要如何建構自己的答案。實作評量鼓勵學生積極的展現成就，進行高層次思考與問題解決技巧；並運用他們的問題解決能力。學生在進行實作評量所需要的時間會依據問題情境有所不同，可能從幾分鐘到幾天甚至更久的時間 (Aschbacher, 1991; Baron, 1991; Herman, Aschbacher, & Winters, 1992; Madaus & O'Dwyer, 1999; Stiggins, 1987)，而進行的方式，除了個人完成之外，也可以透過小組合作。

實作評量在台灣教育領域中已逐漸受到重視，然而，真正應用在教學中的現場老師還是相當有限，其主要原因並非設計上的困難，而是對於要如何進行實作評量

的步驟與準則不夠熟悉，本文的目的，則希望能提供一份關於實作評量的簡單食譜，讀者只要能依樣畫葫蘆，照著食譜的步驟依序進行，即使菜餚不甚美味，但至少能呈盤上桌，而未來只要透過不斷的練習與修正，則可以設計一份不錯的實作評量教案。

本文包括幾個部份，首先簡述實作評量的內涵與定義、實施步驟、評量規準等，而後則依據測驗理論，提出應該如何評估實作評量的信度與效度，當然，對於現場老師而言，信效度的評估可能過於困難，然而以測驗評量領域而言，信效度的評估就像是做菜要試味道，不是只把菜做完就結束了，最重要的透過顧客的品嘗，給與菜餚的回饋，以做為改進的依據。信度的評估可看出評分者的一致性，並顯示評分規準的潛在問題，效度評估則是評估測驗內容和評量目的之一致性，本文提出簡單的信效度評估方法，供讀者參考。最後，提出一個實際教案來展示應如何進行實作評量的完整過程。

貳、文獻探討

教育及心理測驗的標準（全美教育研究協會、美國心理協會及教育研究協會[AERA, APA, & NCME], 1999）宣稱實作評量的測驗情境和現實生活中是很類似的，透過學生在這些從特定領域表現中取出的樣本，用以解釋了這領域典型或預期表現。 實作評量的主要特徵，就是表現能直接被觀察到，而其表現要能將目標及分數解釋進行連結。例如，在藝術理論課程中，實作評量可以是要求學生寫一篇關於繪畫理論的論文，然而，若是在一個繪畫教室，要來評估學生的繪畫能力，則不可能要求寫一篇論文，而是應該要學生真正的動筆，來畫出一幅畫來。

Larkin, McDermott, Simon, & Simon (1980) 指出實作評量應與測驗目的緊密連繫。測驗目的可以透過反思教學目標，與學生最後應該表現出來的能力做緊密連結。舉例來說，寫作過程在寫作中是一個相當重要的部分；因此，寫作實作評量最好應包含寫作過程的各方面因素，如何構思、如何校對與編輯修改的草稿等，另一個例子，若是數學的實作評量，則內容可包含問題解決、理由及證明，在解答中，也應指出需要落實和調整各種不同的策略以解決問題、發展和評估數學證據和論點，組織、摘要和傳達他們數學思考過程。所有應具備的表現能力應該都定義清楚，並將具體行為羅列出來。

一、實作評量的任務發展

在1990年代初期，研究者提出對於設計實作評量任務的一般準則至今仍適用(Baron, 1991; Herman et al 1992; Linn et al., 1991)。Baron (1991) 主張表現任務設計的內容必須很豐富，學生進行評量的過程中，老師必須引導學生參與教學活動，且使他們深入的了解，並在學年度中就讓學生清楚知道評量的準則，使他們能有機會對

自己的學習成果進行自我檢視。在討論有效的任務需具備哪些特性中，Baron認為表現任務包含以下幾個面向：需要學生替自己的問題想出對策、允許多樣化的策略和解決方法、使學生能夠使用他們先備知識及互助合作。再者一些任務可能需要持續工作好幾個星期甚至好幾個月，讓學生可控制如何解決問題及調查，也需要學生設計和與他們的調查連結，且需要自評及自我檢視（Baron, 1991）。最後，此任務必須對學生是有意義的，有挑戰性的，並且在真實社會背景下是切合的，使學生能將他們的理解和技巧轉換到相關的任務上。

Marzano et al.（1993）認為，實作評量是讓學生透過在不同的情境下，來完成某些工作任務，以展現學生對知識的理解、技巧的運用及思維的習慣。實作評量一般包括動手量度、圖形製作等。實作評量具有一個特定的模式。這個模式的第一步是確認內容的標準，標準可能是建構在教學目標或是評量架構中，希望學生達成具體、且可測量的評量目的。舉例來說在第一步驟中，想讓學生討論一個關於西安事變的觀點：「張學良是否應該脅持蔣介石以達到與日本宣戰的目的」以這個內容而言，許多議題可被拿來討論，學生可以討論並決定，「到底是不是應該與日本宣戰?或是有沒有其他不用脅持蔣介石，就可以讓蔣介石對日本宣戰的替代方案」。

第二步是建構相關的知識，以完成任務，在第二個步驟，內容變得更明確，任務是找尋關於此議題的相關內容，包括去圖書館找文獻、上網找資料，或是訪談耆老等。

第三步則建構前兩步驟的資訊內容。這個過程可能會發展出以下任務：你是張學良的幕僚，九一八事變後，日軍慘無人道屠殺東北人民。然而，由於蔣介石為當今時局最高領導人，面對共產黨日益壯大，蔣介石決定要與日本保持和平，提出先安內、後抗日的政策，因此不願出兵與日本對抗。張學良已多次和蔣介石建議應與日宣戰，然而均無法達成目的。你是張學良的幕僚，你必須權衡是否應該要發動西安事變，迫使蔣介石與日軍宣戰。你必須解釋你的決定，並解釋是否有其他方式，無須脅持蔣介石，就可以讓蔣介石對日本宣戰的替代方案。

第四步驟為呈現成果並使用標準進行評鑑，對於學生可以使用不同的方式來呈現成果，舉例來說，學生以兩種以上的不同方式發表他們的發現，例如根據當情境，進行歷史文獻蒐集，並撰寫一篇報告等。作者也提醒為了使評量能夠被我們掌握，在一個特定的任務中，不要使用超過三到四個評量指標（亦即你會從那些角度去看學生的作品），各評量指標應寫下如何評分，例如用三點量尺評定，「3-優異」、「2-普通」和「1-要加油」，以描述作品的特徵或展現的品質。

除了使用單一實作評量，來評鑑學生的學科能力之外，也可以將數個實作評量串起來評鑑，Shavelson & Ruiz-Primo（1998）描述一個對於科學實作評量的框架，在這計劃中，他們以各種不同的評分標準跨越了各種任務。任務的類別有：調查較

有用的方法、識別主要構成要素、分類調查及觀測調查。分析評分計劃由以下四方面組成：基於過程、基於證據、基於分類合理性，以及基於數據準確性。表格1為上述四個要件，提供了關於實作類型、任務、回答格式及評分要素各種不同類型的例子。

表 1　科學實作評量類型

實作類型	任務	回答格式	評分要素
比較性的探索： 溶解實驗	給與學生三種粉末，學生須判定哪一種粉末最容易溶解於水中。	學生必須寫下如何進行實驗，與發現的結果。	歷程導向: 學生進行實驗的過程是否合理，最後解答的正確性。
成份辨識： 神秘粉末	一個紙袋裝了幾種粉末，學生須判定紙袋中有哪些粉末。	學生須寫下如何進行實驗來決定袋子中所具有的粉末，與最後的辨識結果。	證據導向:如何判斷粉末存在的證據，與最後結果的正確性。
分類： 石頭圖鑑	提供數種石頭，學生須將這些石頭進行分類，並指出各種石頭的特性。	學生須展示如何進行各種石頭的分類，並指出各種石頭的不同特性。	分類合理性:分類特質的正確性與精準性。
觀察： 地質調查	依據石頭的出產地進行實地訪查，並描述當地地質、地形、氣候等。	學生需提供觀察的歷程與結果。	數據準確性:如何蒐集證據與描述之準確性。

思考及推論解決的任務過程也被用來設計評量，舉例來說，分析認知型的任務可使用放聲思考（Ericsson & Smith, 1991），目前已被運用在醫學領域上（Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999）。透過受試者思考、知識、過程及提問的特徵，都可用來評估此領域技術的專業程度（Glaser, Lesgold, & Lajoie, 1987），這些特徵更可以用來嵌入不同的評分規準。

二、評分規準與原則

設計評分項目是一個不斷反覆的過程，教師設計完之後，可先進行測試，看看評分項目是否學生所有的行為表現，而後透過幾次的修改，才能讓評分的項目較為完整。設計評分項目需要明確的標準，不管是對於評論表現品質或選擇一個評分步驟（例如：分析型的或整體型的）。基本上來說，評分標準是由一群專家學者，藉由他們在不同領域的知識，以及身為教育家的經驗過程發展而來，這些專家也同樣參與設計表現任務，且擁有關於學生在不同階段所表現的不同精熟程度的知識。Clauser (2000) 指出有幾種可能的方法用來確認標準是否合宜，例如請專家進行放聲思考，或解析對任務應有的可能回應。

評分標準說明在各個分數階段和被測量的架構是有關的，事實取決於包括是否

為一個成果或過程的測驗、測驗的任務需求、參與測驗人數及測驗目的和分數解釋。得分水平的數量取決區分多少個不同層次的表現。然而，得分的水平設定不宜太多，三到四個層次即可。

評分規準（rubric）有三種主要類型：總結性、分析性及檢核表（Huot,1990; Miller & Crocker, 1990; Mullis, 1984）。

選擇一個特定測驗的評分程序取決於測驗的目的與分數的詮釋，使用總結性評分，測驗者會根據寫作的品質做出單一且全面性的評論並打一個分數，例如表2所陳述的為國中基本學力測驗中的寫作測驗的評分規準摘錄，其採用規準為總結性評分，受試者在收到成績時，只會拿到一個來代表自己的成績。這種評分對於受試者人數多，如國中基測的作文考試，動輒數萬考生，較為適合。而分析式的評分則如表3，評分者評量寫作是根據數個面向，例如：立意取材、結構組織、遣詞造句、錯別字、格式與標點符號。而評分要呈現出每個項目的品質。根據 Mullis （1984）的摘要指出「總結性的評分是設計來描述全盤性的；或是各部份的總結，而分析式的評分則設計為描述單一特徵或部份主題，並以加總方式達成全面性的評分。」分析性各部份的評分高低，取決於內容的相對重要性，若是這項指標，是評量的重點，則給予這個分向度的配分可以較高。然而，分析性評分對於教師而言較為費時費力，但可提供受試者表現強項與弱項資訊的回饋訊息。

表 2　總結性評分範例：國民中學學生基本學力測驗寫作測驗評分規準摘要表

級　分	國民中學學生基本學力測驗寫作測驗評分規準
六級分	六級分的文章是優秀的，這種文章明顯具有下列特徵： ※立意取材：能依據題目及主旨選取適切材料，並能進一步闡述說明，以凸顯文章的主旨。 ※結構組織：文章結構完整，脈絡分明，內容前後連貫。 ※遣詞造句：能精確使用語詞，並有效運用各種句型使文句流暢。 ※錯別字、格式與標點符號：幾乎沒有錯別字，及格式、標點符號運用上的錯誤。

註：內容摘取自國中基測網站http://www.bctest.ntnu.edu.tw/writing.htm

表 3　分析性評分範例: 改編自國民中學學生基本學力測驗寫作測驗評分規準

項　　目	分項得分	教師評分
立意取材	25%	
選取適切材料	15	
能闡述說明，以凸顯文章的主旨	10	
小計	25	
結構組織	25%	
文章結構完整	10	
內容前後連貫	15	

小計	25
遣詞造句	25%
能精確使用語詞	15
有效運用各種句型使文句流暢	10
小計	25
錯別字、格式與標點符號	25%
沒有錯別字	20
標點符號運用正確	5
小計	25

檢核表則列出表現或成果的測驗向度、在適當空格中做記號 ，教師可將所有需評鑑的要點寫下來，而後再來檢核是否學生的表現符合標準（如表4）。檢核表適合用在只想檢視學生達成某項任務與否，而不想用分數來區分學生成績的高低情況。

表 4　檢核表範例: 改編自國民中學學生基本學力測驗寫作測驗評分規準

項　　目	是否達成
選取適切材料	
能闡述說明，以凸顯文章的主旨	
文章結構完整	
內容前後連貫	
能精確使用語詞	
有效運用各種句型使文句流暢	
沒有錯別字	
標點符號運用正確	

總體而言，製作評分規準的步驟如下:

1. 參考其他Rubric範例，選擇適合的類型。
2. 依據需求來選擇總結性、分析性或檢核表的評分規準。
3. 配合教學目標，來思索受試者應該要表現出的行為能力。
4. 根據行為能力來定義評分規準。
5. 羅列學生在每個規準上的不同表現方式或是程度。
6. 訂出不同的程度等級、或界定每個等級的分數範圍。
7. 制定適合的表格。
8. 依據表格評分，檢討是否有不足之處並進行相關修正。

制定評分方式的嚴謹度需依考試的風險度而定。若是班級課堂評量使用，則可由任課老師，依據現有的評分表格進行修改。若是風險較高，重要的國家考試，則須由一群專家學者，共同制定出評量要點，透過不斷的修正與改進，才能制定出一個較為完善的評分方式。

國外有一些現成的網站可以使用，新手老師可以參考這些網站的內容，來進行評分規準的設計。以下大致介紹這些網站，有興趣的讀者可以深入研究網站的內容。

1. Rubistar (網站為http://rubistar.4teachers.org)

這個網站可以下載別人已經制訂好的歸準，也可以用裡面的程式來寫制定自己要用的規準。其中可以直接下載的評分規準內容涵蓋口頭報告、作品、多媒體、科學作業、寫作、工作技能、數學、藝術、音樂、閱讀等領域。然而，別人寫好的評分規準不見得可以直接拿來用，所以網站將各類別的評分規準予以統整，並將所有主要的子項目進行歸類，使用者在使用其系統時，不用重新去想評分項目，而是可以點選系統內已經設好的向度進行修改即可，對於老師而言，應該可以節省不少時間。

2. Authentic Assessment Toolbox (網站為http://jfmueller.faculty.noctrl.edu/toolbox/)

這個網站為美國North Central College in Naperville大學教授Dr. Jon Mueller所架設，除了有豐富的理論介紹外，例子多都搭配作業的描述，所以讀者可以清楚看到作者原先所設計的實作評量，與其搭配的規準為何，除此之外，還提供檔案評量的作業範例。其範例涵蓋國小、國中、高中與大學階段各式各樣的實作評量，不僅評量者可以使用其評量規準，教學者亦可以參考其實作評量的範例融入課程使用。

3. Rubric Library (網站為http://www.fresnostate.edu/academics/oie/assessment/rubric.html)

此網站提供許多評分規準的範例，尤其是非學科的部分，包括政策性的寫作、領導才能、口試、計畫案、批判性思考、戲劇寫作等。

三、信效度

教室評量中，很少能對實作評量的信效度進行評估，然而，評量設計者可借由信效度的評估來檢視評量的成效。信效度亦可讓教師對於自己所設計的評量品質更加了解。

信度是指評量結果的穩定性（stability）及一致性（equivalence; consistency）。評量結果的穩定性可由再測信度來評估，然而，由於實施實作評量費時費力，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實作評量的信度。

大多數的教室評量，評分者侷限為教師一人，這種情況無法計算評分者一致性。然而，若是能選出一些學生當作評分者，或是在進行合作教學時，和搭擋的教師一起評分，此時即可以使用評分者信度來進行實作評量的信度評估。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，透過其他評分者的評分，可進一步檢視評分的客觀性。

肯德爾和諧係數常用評估評分者的信度，此係數用來評估 K 位評分者，針對 N 位受試者表現評比時的評分一致性，也可以視同一個評分者先後 K 次評 N 個對象。其計算公式如下，其中 K 代表評審者的數目，N 代表受試者的數目，∑Ri 為每個被評對象所評等級之和。$\sum Ri^2$為每個被評對象所評等級之平方和。

$$W = \frac{\sum Ri^2 - \frac{(\sum Ri)^2}{N}}{\frac{1}{12}K^2(N^3 - N)}$$

計算時，除了可以使用現成的統計軟體之外，也可以用excel甚至計算機來進行計算，其計算步驟如下。

步驟1：陳列每位評分者對每位選手的評分總分。假設有五位評分者ABCDE，他們針對三位選手歌唱表現進行評分，而下面則是每位評分者所給的總分，第一位評分者A，給選手甲75分，給乙選手73分，給丙選手63分…以此類推。

	甲選手	乙選手	丙選手
評分A	75	73	63
評分B	67	73	65
評分C	86	82	69
評分D	54	70	66
評分E	83	77	87

步驟2：把分數排序。每位評分者的成績進行排序，例如，對於評分者A來說，給甲選手的成績最高，所以排序為一，乙選手的成績次之，所以排序為二，丙選手為第三，排序為三。而 ∑Ri 則是將排序的成績加總起來，而 $\sum Ri^2$ 則是將Ri先平方再進行加總。

	甲選手	乙選手	丙選手
評分A	1	2	3
評分B	2	1	3
評分C	1	2	3
評分D	3	1	2
評分E	2	3	1
Ri	9	9	12
Ri^2	81	81	144

步驟3：帶入公式

K = 幾位評審 = 5
N = 幾位受試者 = 3

$$W = \frac{\sum Ri^2 - \frac{(\sum Ri)^2}{N}}{\frac{1}{12}K^2(N^3 - N)}$$

$$= \frac{(81+81+144) - \frac{(9+9+12)^2}{3}}{\frac{1}{12}5^2(27-3)}$$

$$= 0.12$$

評分一致性為0.12，和諧係數w越大則一致性越高。一般來說，W值介於0.9~1.0代表評分者之間的評分非常高相關， 0.7~0.9代表高相關，0.5~0.7代表中等相關，0.3~0.5代表低相關，0.0~0.3代表微相關。因此在本例中，評分者的信度較低，代表評分者評分分數之間的相關性很弱。

效度是用來評鑑評量結果的解釋與使用的合適性， 在許多考試中，效度的評估是極為重要的。若是可以的話，效度的證據可多方蒐集，用以評估評量的成效。效度有分為很多種，包括內容效度、效標關聯效度、構念效度等。內容效度指的是指測驗的內容是否符合測驗的目的，若是題目的內容符合教學目標、所選的教材也有代表性, 則我們稱測驗的內容效度是良好的。教師在設計實作評量時，可以將設計好的教案，請其他老師檢視看看，甚至可以請相同領域的學者專家來看看是否所設計的實作評量任務，能夠與教材內容所涵蓋的範圍與教學目標相符。

效標關聯效度指測驗的分數與其他相關測驗或指標的相關性。如果設計的為數學的實作評量測驗，那麼可選用的效標為學校老師給定的成績、學生課堂時的表現、其他相關的數學測驗（像是學生的期中考數學成績）、或是學生是否在其他數學競賽中的得到優異的成績。如果測驗分數和外在效標之間的相關越高，則代表效標關聯效度越高，也就是越能用測驗分數來有效解釋及預測外在的效標。

建構效度則是建立在構念之上， 而構念是在心理學或社會學上所存在的一個理論上的構想特質，不容易觀察，也很難被測量，但是我們卻可以假想這是存在的。建構效度的建立，必須由研究者先提出假說，並蒐集資料去驗證 並反覆檢討、修正整個建構的過程， 直到建構效度可以成立為止。而內容效度與效標關聯效度的建構方法與結果，都可用來當作建構效度的證據。建構效度的驗證方法種類繁多, 有內部一致性的分析法、外在效標分析法、因素分析法、結構方程式模式、多特質-多方法分析法等。

每種效度的證據若能都蒐集是最好的，可一般教師在課堂上要能找到這些證據較為困難，比較可行的應該是蒐集內容效度與效標關聯效度的相關證據，內容效度的獲得可透過其他教師的相互討論，而效標關聯效度則是依據學生在班上的其他表現，如學習成績等。而建構效度，需要一些較為複雜的統計方法與統計軟體來執行，有興趣的讀者可參閱余民寧（2011）專著。

此外，針對教室評量其相關更深入的信、效度議題可參考Brookhart（2003）。

四、實作評量應用

(一)大型測驗

實作評量在操作上較為費時費力，但還是可能以大規模的方式進行施測。若將數個實作評量集結起來，並進行系統性的歸類，則可變成檔案評量。LeMahieu et.al（2005）所進行檔案評量，即是將數個實作評量的內容集結起來。實驗中將美國匹茲堡的學區內六年級到十二年級的學生要求進行寫作的檔案製作，並對其中部份的檔案進行隨機抽樣，取出了1250份的檔案，針對寫作的三個面向，來進行學生寫作能力的評估。每份檔案須包含以下幾份作品

(1) 依據自己的標準，選擇一份最重要的作業
(2) 一份自己最滿意的作業
(3) 一份自己最不滿意的作業
(4) 一份自選作業，但須寫出選這份作業的理由
(5) 若是班級教師覺得學生所選擇的作業不夠具有代表性，則可以再為學生選一份作業。

除了這五份作品之外，檔案中還需提供一份目次表，來描述檔案中作品的內容與製作日期，一份寫作的問卷來描寫個人成為一個寫作者的經驗，及一個反省回顧的描述，來記錄學生過去一年來的寫作能力的變化。從這份檔案的內容的描述可以看出，學生對於自己檔案的內容有充分的決定權，某學生可以選擇寫新詩，而另一位學生可以選擇寫短文，在所有作品選擇中，並沒有硬性的規定學生應選取何種作業放進檔案中。為了避免學生會對所選取的文章進行過多的修飾，抽到的學生在一個禮拜前才會進行通知。

此檔案的評分從三個面向來進行探討（見表5），而每一個面向都有六分，得0分代表學生的表現不足（inadequate）而6分代表學生的表現卓越，若是評分者覺得檔案內容不足，無法對某個面向進行評分，也可以評證據不足（no evidence）。第一個面向為寫作的成就。內容包括學生的寫作品質，對於寫作能力、技巧、架構、文章標題的了解程度與語言的表達能力都與以評價，這個面向與一般對寫作的要求相同。第二個面向為評估學生對於寫作過程及策略的運用能力，內容包括有效的使用

預寫的策略，使用草稿來形成自己的想法，並利用外界的資源（例如同儕討論、讀者或是其他成人的回饋）來進行文章的修正。第三個面向為寫作者的成長與發展的能力評估。學生必須展現對於寫作的熱誠與態度、如何看出個人寫作的優缺點並給予評鑑，並能夠對不同的目的、題材與對象來寫作。

表5　匹茲堡寫作檔案評分向度與內容

面向一:寫作成就
(1) 達到具有價值性的挑戰
(2) 建構及維持目的
(3) 使用技巧及選擇題材
(4) 控制慣用語、字彙與語句結構
(5) 瞭解讀者的需求（組織、發展與使用細節）
(6) 使用語言、聲音、圖片與語態
(7) 幽默感、比喻、有趣性
面向二:寫作過程及策略運用
(1) 有效的使用預寫的策略（prewriting）
(2) 使用草稿來發現並修正想法
(3) 使用討論的機會來修正寫作（同儕、成人、或讀者）
(4) 有效的進行修改（改造、重新聚焦與修改）
面向三:身為一個作者的成長、發展與專注
(1) 對寫作任務投入性的證據
(2) 增進對寫作的投入性
(3) 發展身為作家的感覺
(4) 個人寫作標準的演進
(5) 能看出某人寫作的優點及需要
(6) 能在進行寫作作業上展現冒險與創新
(7) 可以使用不同目的、題材或是對象來寫作
(8) 最早的作業和最近的作業之間的進步、成長與發展

註: 翻譯自LeMahieu et al.（1995）

此研究共有25位評分者參與檔案的評鑑。其中，12位進行國中的檔案評鑑，而13位進行高中的檔案評鑑。評分者的組成份子為教師與閱讀或是寫作領域的專家，國中組的專家每人須評鑑99份的檔案，而高中的專家則須評鑑78份的檔案。進行正式評分前，每位評分者都須經過訓練並使用幾個範例檔案進行評分練習。透過不斷對於各面向的討論與實例探討，評分者對於規準的了解程度與實例檔案的各面向評分達到一致性之後，才正式對學生的檔案進行評分。

每份檔案須有兩位評分者對三個面向進行評分，若是兩位評分者所評的分數差異性在一分之下，則以兩位評分者所給分的加總作為這份檔案的分數。如果差異性在一分以上，則有第三位評分者進行仲裁評分。

評分過程經歷整整一周的時間，得到相當高的信度，三個向度的信度約為0.74到0.87之間，尤其以寫作成就的信度最高。而對於評分者之間的信度（inter-rater reliability），也達到0.80到0.84之間。從這個研究可以看出，只要能夠對實作評量進行充分的規劃，對於評量的內容詳細加以說明，並設定嚴謹的評分程序與對評分者的訓練，即使讓受試者自由選擇檔案中的內容，亦可以得到良好的信度。

（二）醫學臨床技能測驗

實作的能力在醫學界的需求已相當風行，且考選部在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)。

這項測驗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。 在所有的12站中，受試者必須通過七站才算合格。

這項測驗動員了龐大的人力，包括768位主治醫師擔任考官， 512位標準化病人配合測驗，還有眾多的試務工作人員。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。

參、實作評量範例

在生活課程中，著重讓學生體驗各種姿態、表情動作的美感，並表達出自己的感受。因此，在這位老師的實作評量設計中，讓學生親自體驗當模特兒，透過走台步和產品代言拍攝來展現自己。

一、設計理念

這是一份讓同學們親自體驗當模特兒的實作評量，評量內容包含走台步和產品代言拍攝兩大部分，這兩項測驗都是以測驗學生是否具有模特兒必備的專業能力為目的，畢竟模特兒的專業不能只靠書本知識的吸收，能否將吸收的知識展現在這兩個測驗項目，才是能否成為專業模特兒的關鍵，透過這樣的實作評量，同學們才能知道自己不足之處，進而能透過本次的評量自行調整改善。

二、指導語和作業說明

(一)指導語

通常在實作評量的開始進行前，或有簡單的指導語，來告知同學接下來要做甚麼樣的活動。說明活動的目的與正確的回答方法，讓受訪者能認真的據實回答，有助於增加評量的效度。若活動中有特殊需求及回答方法，也應事先說明。

指導語須包含以下幾個要點：

1.敘述實作評量的題目與認真作答的重要性
2.告訴受試者評分的重點
3.評量活動進行的方式與可能需要花費時間
4.如果有疑問，應如何尋求幫助

在此處所提供的範例，所提供的指導語如下:

給未來的模特兒們：

在過去專業訓練下，準備好要show出你們努力的成果了嗎？記住，唯有跨過重重難關的模特兒才能站在群眾的面前發光發亮！準備好要接受考驗了嗎?這次的成果驗收將從由兩個不同層面分開檢視評分，分別是走台步和產品拍攝代言的部分。走台步方面，透過不斷的練習，希望學生能在走秀時態度自信，在肢體動作上展現模特兒的水準。另外，在產品代言攝影的部份，也期待學生能夠發揮所學，構想出符合產品功能及風格，且與肢體能相搭配的和諧畫面，以下將有更詳細的說明。

台步的展示將會是第一個施測的項目，同學需沿著地上的直線行走，且在標記處擺出一兩個姿勢後轉身回到出發時的標記處再次擺出姿勢後離開。在這樣的過程當中，眼神除了要透露出自信之外，也要盡量保持平視，避免東張西望或是看地板，在最後的姿勢方面則要記得自己身體的優勢，擺出能夠展現優勢、掩蓋缺點的姿勢。轉身的時候也要記得保持平衡，抬頭挺胸，這些都會列入評分的項目之中。至於攝影的部分，受試者需要在限制的四樣產品當中（分別是雨傘、包包、水壺和手機），選擇自己喜愛且能發揮的產品進行拍攝，一個產品將會拍攝三張照片，在拍攝過程中，將由同一位攝影師進行拍攝，但模特兒可有自己的想法與攝影師溝通拍攝與取景角度。在拍完兩個商品共六張照片之後，評審們將會根據攝影的過程和拍攝出來的照片進行評分。除了能夠明確表達出該產品的核心概念外（例如：雨傘的功能是遮雨），肢體和表情是否和產品有互相搭配也是這個測驗項目的評分重點之一。另外，照片當中是否有使用不同的姿勢和表情來傳達商品也會在評分考量內。這是我們蒐集的資料，僅當做研究使用，且會對你的個人絕對保密。如果你有甚麼問題，可以問身邊協助的同學喔!

(二)評分規準

採用評定量表作為判斷模特兒表現的評量工具，量尺類型屬於數字型評定量表。評分者依據情境角色所表現出的程度圈選適當表現該特質程度的數字；「1」代表不佳、「2」代表有待加強、「3」代表尚可、「4」代表優良、「5」代表非常優秀。

級分	評分規準
5	模特兒表現優異，符合以下特徵： (1) 肢體動作能自由駕馭，態度自信大方，不扭捏。 (2) 代言充分呈現產品特色與核心概念，模特兒充滿個人魅力與特質，創意度極高。 (3) 照片具有吸引觀眾購買之潛力。
4	模特兒表現已在水準之上，符合以下特徵： (1) 肢體動作表現得宜，偶有不流暢之處，但不致影響評分。 (2) 代言已能掌控產品特質，經提示後，能發揮創意表演。 (3) 照片可激發大部分觀眾購買意願。
3	模特兒表現已達一般水準，符合以下特徵： (1) 肢體動作表現尚稱完整，偶有銜接不順之處。 (2) 代言尚可，偶需旁人提點產品核心概念，創意度尚可。 (3) 照片效果尚可，觀眾購買意願持平。
2	模特兒表現未達基本水準，符合以下特徵： (1) 肢體動作斷斷續續，銜接不順暢。 (2) 代言表現度待加強，大部分需旁人提點產品核心概念，無法自行發揮創意表現產品特質。 (3) 照片效果待加強，觀眾購買意願低。
1	模特兒表現不佳，符合以下特徵： (1) 肢體動作不協調，神情渙散沒有自信。 (2) 代言表現度不佳，經提示及示範後，仍無法呈現產品特質。 (3) 照片效果不好，觀眾無購買意願，非模特兒人選。

共5位評分者 (k = 5) 及10位被評者 (N = 10)，經過總分計算，每位評分者對 10 位受試者的評分排序如下：

評分者（k=5）	學生作品（N=10）									
	甲	乙	丙	丁	戊	己	庚	辛	壬	癸
A	5	6	4	1	2	3	5	9	8	10
B	6	7	3	2	8	1	9	4	5	10
C	7	9	2	1	5	3	4	8	6	10
D	2	9	6	1	4	7	5	8	3	10
E	7	8	2	1	6	3	9	5	4	10
Ri	27	39	17	6	25	17	32	34	26	50
Ri2	729	1521	289	36	625	289	1024	1156	676	2500
$\sum Ri=27+39+17+6+25+17+32+34+26+50=273$										
$\sum Ri2=272+392+172+62+252+172+322+342+262+502=8845$										

代入和諧係數公式可得和諧係數為為=0.674

以和諧係數來計算評分者間的信度係數，由這五位評分者對十位參賽者的評分結果以等第分數評定後，計算其評分者間信度係數為0.674。

肆、結語

在本文中提出的內容，希冀能提供相關單位參考，最後，提出幾點建議。

一、評量帶動教學的改變

考試領導教學是大多數學者所反對的一個方向，然而，不諱言的在台灣升學主義下，往往是升學考試要考甚麼，老師就教甚麼，在傳統選擇題的紙筆考試下，著重的往往是某部份、片面性的知識，而且要能夠廣泛的考到所有的內容非常困難，而這些部份性的、片面性的知識卻要用以代表學生所有的學習成效，因此老師在教學時，只教要考的重點，對於考試不考的內容就跳過，然而，這些片段式的知識往往是見樹不見林，學生記幾個重點公式，會看關鍵字套公式，卻不了解整個內容來龍去脈，因此可能考完就忘記，沒辦法學習到整體性的知識。

在實作評量是鼓勵考試來帶動教學的改變，先把評量規準告訴學生，告訴他們這學期要評鑑的重點就是依照規準來評，因此老師按照規準的重點來教，學生也要依重點來學，如寫作測驗中，評量歸準可能包含大綱的訂定、文獻搜集的深廣度、段落的清晰、文句排列的邏輯性等，老師在教學時，依據這些重點來教，而學生也照這些重點來學，一整個學期下來，學生大概就會知道寫作的重點有甚麼了，也能夠寫出一篇像樣的文章。因此實作評量上而言，學生需要如何去操作一個實際的任務，而且事先可以先看到評分規準，知道甚麼是好的表現並依據這樣的要求去完成，並在整個學習過程中，可以培養完整的知識與技巧。

二、實作評量的任務設計

實作評量搭上多元評量的列車，許多學校都在推行，然而，在任務執行中，須隨時檢視評量最終的目的，是為了解學生學習的成效，應以學生的學習為中心，且評量應與教學目的、課程內容緊密結合。因此在設計實作評量任務時，若能與相同領域的老師共同討論，甚至能跨不同學科共同設計，方能達其效果。而目前對於實作評量的理解與推行上仍有努力的空間，希冀相關單位能藉由辦理研習課程、來推廣這方面的知能。

三、標準本位評量的趨勢

過去的教室評量結果重試排名，只知道學生的名次，卻不知學生在學習上是哪

裡不懂，也不知理解程度為何，十二年國教推行後，標準本位評量將為推行重點所在，重視的是學生了解學科內容的程度，依據學科習得知識的了解與應用程度，將學生的表現分成幾個等級，而每個等級對應出甚麼樣的表現水準，都有明確的指出方向。例如在每個學習階段、學科領域都有各自的課程綱要與對應出的基本能力指標，而透過這些能力指標將其轉換成學習內容方向、並設計相關的實作評量任務，發展評分規準，而據此評量與了解學生的表現。這樣的標準本位評量模式，將是未來評量的新趨勢。

根據美國視導與課程發展學會（The Association for Supervision and Curriculum Development, ASCD）研究指出，在各種考試形式中，以學生檔案評量效果最佳，其次為實作評量、上台報告、期末考試與州立大型標準化測驗（ASCD Smartbrief, 2012），而檔案評量和實作評量相當類似，都是透過真實性評量來評估學生的學習結果，然而，這樣的考試進行方式，需要相關教育單位、教師、家長與學生共同的努力與配合，方能推廣與執行。

伍、參考文獻

余民寧 (2011)。**教育測驗與評量：成就測驗與教學評量**（第三版）。台北：心理出版社。

曾蕙蘋 (2012) 。奶嘴公務員鬧笑話公職考試尋專才。**中國時報**。取自：http://blog.udn.com/baogon/6339187

曹以會 (2013) 。OSCE醫學臨床測驗今年正式舉辦 近99%及格。**聯合報**。取自:http://mag.udn.com/mag/edu/storypage.jsp?f_ART_ID=459982Power By udn.com

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

ASCD Smartbrief. (2012). Which do you think provides the most accurate summative assessment of student learning? Retrieved February, 4, 2013 from http://www.smartbrief.com/news/ascd/poll_result.jsp?pollName=89A63917-5867-47C5-8F56-1B50832D90D3&issueid=1318C04D-9D20-4609-9608-FA30F3D9363A

Aschbacher, P. R. (1991). Performance assessment: State activity, interest and concerns. *Applied Measurement in Education,* 4 (4) , 275-288.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education,* 4 (4) , 305-318.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment

purposes and uses. *Educational Measurement: Issues and Practice, 22* (4) , 5-12.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement,* 24 (4) , 310-324.

Ericsson, L. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An instruction. In LA. Ericsson & J. Smith (Eds.) , *Toward a general theory of expertise: Prospects and limits,* 1-38, Cambridge, MA: MIT Press.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning J. Glover, J.C. Conoley, & J. Witt (Eds.) , *The influence of cognitive psychology on testing and measurement.* The Buros-Nebraska Symposium on measurement and testing, 3, 41-875, Hillsdale, NJ: Lawrence Erlbaum.

Herman, K. L., Aschbacher, P. R., & Winters, L (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Huot, B. (1990). The literature of direct writing assessment major concerns and prevailing trends. Review of Educational Research, 40 (2) , 237-263.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science,* 208, 1335-1342.

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.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher,* 20 (8) , 15-21.

Linn, R. L. (1993). Educational Assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis,* 15, 1-16.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessments. *Phi Delta Kappan,* 80 (9) , 688-696.

Marzano, J., Schmitt, A., Bleistein, C. (1993). *Sex-related performance differences on constructed-response and multiple choice sections of the Advanced Placement Examinations* (RR-93-5). Princeton, NJ: Educational Testing Service.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education,* 3 (3) , 285-296.

Misley, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, LA. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior,* 15-335-374.

Mullis, I.V.S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement :Issues and Practice,* 3 (1) , 16-18.

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,* 37-55, Boston: Kluwer Academic.

Shavelson, R. J., & Ruiz-Primo, M. A. (1998).On the assessment of science achievement conceptual underpinnings for the design of performance assessments: Report of year 2 activities (CSE Technical Report 481). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices,* 6 (1) , 33-42.

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