第二章 文獻探討

研究者在進行實作測驗之命題研究時，應先了解測驗和評量的意義、特徵與類型，並熟悉測驗編製計畫、命題原則和預試要點，再考慮實作評量之內涵與優缺點，針對其特點進行命題工作。

第一節 測驗與評量

一、測驗與評量的意涵

測驗、測量、評鑑、評量是教育測量中常使用的術語，它們的意義有相當程度的相通之處，時常交互使用。其中評量（assessment）是指收集、綜合學生學習資料，並加以解釋、做各種教學決定的歷程；而測驗（test）可視為「透過行為樣本，以測量個人特質的有系統程序」（郭生玉，民 93），行為樣本（sample of behavior）是測驗中讓受試者回答的問題，或是操作的樣本（task），必須能包含欲測量的特質或成分，個人特質主要探究認知能力和情感特質兩個部份，有系統程序則是指測驗的標準化（standardization）過程。

二、測驗的特徵

效度（validity）和信度（reliability）是傳統標準測驗最重要的兩個特徵，效度以信度為基礎，衡量一個測驗能夠測量欲測特質的程度，而信度則檢驗測驗分數的一致性或穩定性（陳英豪、吳裕益，民 92；郭生玉，民 93）；茲將效度與信度的類型分述如下。

（一）效度

目前在教育和心理測量方面使用最廣泛的效度是 French & Michael（1966）提出的三種效度類型，這種分類法於 1974 年列入美國心理學會（American Psychological Association）發行的「教育與心理測驗之標準」（Standards for Educational and Psychological Tests）一書，包含內容效度（content validity）、效標關聯效度（criterion-related validity）以及建構效度（construct validity），其意義和考驗方法參見表 2-1。

表 2-1

效度的意義和考驗的方法

類型	意義	考驗方法
內容效度	測驗的內容能否充分代表其所欲測量的行為領域。	比較測驗的材料和所欲測量的教學目標及教材內容是否一致。
效標關聯效度	測驗成績對目前及未來某一行為表現（由其他適當的工具測量而得）預測力的高低。	求測驗分數與其他測驗成績之相關。其他測驗成績如在同時測量則為同時效度；如在往後測量則為預測效度。
建構效度	測驗的成績能以心理學的屬性來加以解釋的程度。	建立理論架構，以解釋個體在測驗上的表現；根據理論架構推演出各種假設；收集資料考驗假設是否成立。

資料來源：「測驗與評量」，陳英豪、吳裕益，民 92，頁 383。

由表 2-1 可知三種常用效度的意義和考驗方式：內容效度用來檢視編製完成的測驗，是否能測量出想要測量的特質，講求測驗內容和教材內容與目標的契合；利用效標關聯效度探討本次測驗成績與其他測驗成績的相關，其他測驗可能是目前的其他種類測驗，或是將要進行的測驗；透過建構效度來考量測驗成績所能代表的心理特質程度，尋求分數所能說明的意義。此三種效度之功能和適用的測驗類型不同，可視測驗的性質與目的選擇使用之。

了解各種效度的功能與適用類型後，需進一步探討各方面因素對效度的影響，可歸納為下列五方面（簡茂發，民 90）：

1. **測驗組成方面**

   測驗的效度取決於試題的性能，若審慎選擇測驗材料，測驗長度恰當，試題具有相當的鑑別力且難易適中，並作合理安排，則可有較高的效度。

2. **測驗實施方面**

   遵照測驗手冊規定施測，按照標準程序進行試場佈置、作答說明以及時間掌控，可避免外在因素影響測驗結果的正確性，確保測驗具有一定水準的效度。

3. **受試反應方面**

   受試者的身心狀況、動機與態度，以及其合作與努力程度，都會影響測驗結果的可靠與正確性，唯有受試者呈現真實反應，才能有效測量欲測特質，提高測驗的正確性。

4. **效標方面**

   選擇適當的效標方能真實呈現測驗的效度，一個測驗的效度係數可能因為所採用的效標不同，而有截然不同的結果；在統計上影

響效標關聯效度的因素有測驗的信度、效標的信度，以及測驗所量度者與效標所鑑定者之間的真正相關程度。

5. **樣本方面**

效度考驗所依據的樣本應能代表某測驗所要應用的全體對象，一個測驗應用於性別、教育程度或經驗背景不同的對象時，效度會隨著測驗功能不同而異；此外，樣本的異質性（sample heterogeneity）越大，效度係數也越高。

編製以及實施測驗時，應考量上述適用的效度類型，並嚴密掌控影響效度的因素，以確保測驗可有效測量受試者的欲測特質。

（二）信度

Anastasi 在 1988 年提出由測驗的一致性來探討信度，在不同時間或不同情境下，對同一受試者進行相同試題測驗，或是具有相等試題的複本測驗，由所得結果來檢驗測驗的一致性，如果測驗結果一致，表示此測驗具有穩定性、可靠性及可預測性。

若從測驗的誤差來看，信度是在估計測量的誤差有多少（郭生玉，民 93），也就是測驗分數中由特質造成的真實差異，以及由誤差造成的差異，各占有多少比率。常用的各種信度類型與其誤差來源可參見表 2-2。

表 2-2

各種信度類型與其誤差來源

信度類型	解答的主要問題	誤差來源
	一、測驗情境的影響	
重測信度	1. 相關內容樣本所得分數受到不同測驗情境的影響如何？ 2. 在不同測量時間所得分數的穩定性如何？	測驗實施
複本信度	1. 不管使用的複本測驗或實施的情境怎樣，測驗的一致性如何？ 2. 在不同測量時間所得分數的穩定性如何？	測驗實施 測驗編製
	二、不同內容取樣的影響	
複本信度（同時實施）	1. 測驗分數在相同情境下，是否受不同內容取樣的影響？ 2. 兩份仔細配合的複本測驗是否相等、平行或可交互使用？	測驗編製
折半信度	1. 測驗分數在相同情境下是否受不同內容取樣的影響？ 2. 複本形式的信度係數為多少？	測驗編製
庫李信度	1. 測驗分數在相同情境下是否受不同內容取樣的影響？ 2. 測驗的同質性如何？ 3. 每一個題目的反應一致性如何？	測驗編製
庫李信度（21 號公式）	測驗分數在相同情境下是否受不同內容取樣的影響？	測驗編製

信度類型	解答的主要問題	誤差來源
α係數	測驗分數在相同情境下是否受不同內容取樣的影響？	測驗編製
	三、不同評分者的影響	
評分者信度	1. 如果使用不同評分者，分數差異的程度如何？ 2. 測驗的客觀程度如何？ 3. 不同評分者所得的結果是否可替換？	測驗計分與解釋

資料來源：「教育測驗與評量」，郭生玉，民 93，頁 57。

在表 2-2 中依解答的主要問題來區分，將信度類型歸納出三大類，探討影響測驗的因素與誤差來源：

1. 測驗情境的影響

探討在不同時間或不同情境下施測所得結果的一致性，其誤差來源主要為測驗實施。其中重測信度使用完全相同的試題，而複本信度則使用相同試題母群體（population or universe）所抽出的試題，內容相似而不相同，以此檢驗測驗取樣分數的代表性。

2. 不同內容取樣的影響

探討在相同測驗情境下，測驗分數受內容取樣影響的程度，以測驗編製為主要的誤差來源。信度類型包含同時實施的複本信度，以及由一次施測結果即可估計信度的內部一致性方法；而內部一致性方法可分為折半信度、庫李信度以及α係數。

3. 不同評分者的影響

當測驗類型不屬於客觀測驗，亦即沒有標準答案時，測驗分數會

受到評分者判斷的影響；不同評分者對同一份試卷的評閱分數之間的相關，即為評分者信度，當評分者信度越高，則表示測驗分數受評分者間評閱差異的影響越小，是較為客觀的測驗。

進行命題研究時，應注意控制測驗編製、實施與計分解釋方面的影響因素，以期研發出可測量欲測特質且具有穩定性的測驗。

三、測驗與評量的類型

Birenbaum ＆ Dochy (1996) 依成就測驗與評量的取向不同，將測驗與評量分為傳統標準測驗（traditional standardized test）和另類評量（alternative assessment），這兩種取向的差異在於前者較重視概念學習的成果輸出，在測驗的規劃與實施方面，受試者處於被動情勢，而後者較注重評量與學習過程之結合，被評量者可主動參與評量的實施過程（Birenbaum, 1996；張惠博、黃文吟，民 89）；兩種取向的測驗與評量各包括不同的形式，分別說明如下。

（一）傳統標準測驗

由不同的測驗分數解釋方式，可將傳統標準測驗分為常模參照測驗 (norm-referenced test) 和標準參照測驗 (criterion-referenced test) 兩種形式（郭生玉，民 93）。

1. 常模參照測驗

常模參照測驗的分數解釋方式，是根據受試者的分數在團體中相對位置來加以解釋，其主要目的在於區分受試者之間的能力。

2. 標準參照測驗

標準參照測驗的分數須與教學前所訂定的標準做比較，達此標準者，稱為「精熟學習」(mastery learning)，未達此標準者則稱為

「非精熟學習」，可知道受試者會與不會的內容各為哪些；標準參照評量也稱為領域參照評量（domain-referenced assessment）、內容參照評量（content-referenced assessment），或課程本位評量（curriculum-based assessment）（Ward ＆ Murray-Ward, 1999; 郭生玉，民 93）。

（二）另類評量

常見的另類評量有檔案評量（portfolio assessment）、實作評量（performance assessment）、真實評量（authentic assessment）、直接評量（direct assessment）和建構式評量（constructive assessment）等（張惠博、黃文吟，民 89），其中以檔案評量和實作評量較為受到重視且使用頻率較高，以下分別說明這兩種評量在教學上的應用。

1. 檔案評量

檔案評量的方式主要是透過檢視學生的作品與資料，來顯示學生的努力與學習進展。能夠展現學生學習過程與進步情形的任何形式資料，都可成為檔案評量的內容物（Popham, 2005），例如書面、錄音、錄影等不同形式；檔案評量的資料收集以及評量方式皆具有彈性，可由老師、學生、家長三方面共同決定內容形式，一起進行資料收集，評量時除教師評量外，也可同時由學生自評、家長評量、同儕互評（李坤崇，民 88），評量重點在於學生對內容物的自省與統整能力。

2. 實作評量

實作評量是要求學生利用口頭、書寫、製作作品的方式，來完成特定的實作工作（performance task），並透過此成果來評量學生

的學習狀況，偏向採用由學生自己建構反應的題目，以測量學生是否能運用高層次思考能力來解答問題（郭生玉，民 93）。實作評量與 Wiggins（1989）所提出的真實評量有許多共通之處，但真實評量更為強調與相關現場發生關聯，作業的選擇不僅以學校為範圍，並將心力投注於思考如何決定教與學的評判準則，比傳統測驗有更多比重的自我評量，也常在過程中要求學生展示與介紹自己的作品，以確認學生對於相關內容的精熟程度。

另有學者認為教師不必用心區分實作評量與真實評量，而應以學生為中心，改善評量的歷程和方法，掌握實作評量的特質，方能發展出良好的教學與評量整合模式（李坤崇，民 88）。

四、測驗編製計畫

不同目的與類型的測驗編製各有其獨特之處，然而不論測驗的類型為何、欲採用何種方式檢驗試題，一個良好的測驗都需有具體可行的測驗編製計畫，此計畫中應包含下列四項（簡茂發、郭生玉，民 74）：

1. **確定測驗的範圍**

   在編製測驗時，應確定測驗的功能、目的和對象，包含受試者的生活背景與發展程度，以及測驗欲測量的特質；在這些因素所界定的範圍中進行測驗編製與施測，可確保測驗的適切性。

2. **分析測量的目標**

   每種測驗所測量的層面不同，例如在學習成就測驗中，由認知、理解、應用、分析、綜合、評鑑六個層次編擬試題，應決定測量的主要目標在於何種層次，進一步分析構成欲測特質的因素。

3. **蒐集有關資料**

確立測驗範圍與目標後，應蒐集命題相關的經驗性資料，以作為命題取材之依據；以學習成就測驗為例，命題時應蒐集課程標準、教科書、參考書、教師自編測驗等相關資料。所蒐集的資料越齊全，有助於測驗內容不致偏頗，並提高行為樣本的代表性。

4. **設計測驗的藍圖**

編製測驗有如建造房舍，需依照藍圖進行命題，命題的藍圖即為雙向細目表（two-way specification table）；以成就測驗的編製而言，應先分析教材的課程內容和行為目標，使兩者適切結合形成雙向細目表，在此以數學成就測驗為例，參見表 2-3。此外，試題編製時的雙向細目表，是判斷內容效度的最好依據（郭生玉，民 93）。

表 2-3

數學成就測驗雙向細目表

教材內容	教學目標			合計題數
	知識	理解	應用	
一、整數的加法	4	10	6	20
二、整數的減法	3	7	3	13
三、整數的乘法	5	12	3	20
四、因數與倍數	7	15	5	27
五、分數的四則運算	4	8	8	20
合計題數	23	52	25	100

資料來源：「教育測驗與評量」，郭生玉，民 93，頁 85。

五、命題原則

擬定測驗編製計畫後，即開始實行計畫與進行命題工作，在命題注意事項方面，李坤崇（民 91）參考多位學者理論，對於教師自編測驗提出八項重要命題原則：

1. **試題分佈依據雙向細目表，且題目內容具有代表性。**

   測驗內容應是教材中具代表性的內涵，而非枝微末節的內容或無關緊要的字詞。例如：教材中提及 1962 年巴特勒發現了第一種鈍氣化合物，編製試題時應著眼於「鈍氣可以與其他元素形成化合物」，而非此化合物發現的確切年份。

2. **避免使用曖昧不明和易使人混淆的言詞或語句架構。**

   試題的敘述應力求清晰易懂，避免因為用詞不當造成學生解題時的誤解。例如：照相軟片上的溴化銀膜遇光會還原成黑色的銀原子，題幹若僅寫出「溴化銀遇陽光會_____」，則可能造成學生無法選擇應填入「反應」、「進行還原反應」、「變黑」或是「變成銀原子」，命題時應完整描述「溴化銀遇陽光會進行_____反應，產生____色銀原子沉積」。

3. **敘述扼要、直接切入重點。**

   僅寫出有助於學生進行答題的訊息，其餘贅述應簡潔或刪除之。例如：欲測量學生計算電解水產生兩氣體的體積比，宜採用直述法「若電解水可得到一公升的氫氣，則同時可得到多少公升的氧氣？」，而非採用故事敘述的方式，說明進行電解水的人、事、時、地、物，模糊解題的焦點。

4. **使用字彙適合受試者。**

描述試題時所使用的字彙，應考量受測學生可理解的用字遣詞，避免學生無法理解題目而影響作答，勿使各項測驗變為字彙測驗。對於年齡層較低的學生，應以生活化用語敘述替代成語來描述試題，例如：題幹「『人非為失敗而生』這句話充滿何種精神？」對於小學生來說，是很難理解的用語，宜改為「『做事情失敗不要難過，更要努力追求成功』這句話表示什麼精神？」（李坤崇，民 91）。

5. **試題答案必須是公認的正確答案，避免爭議性。**

測驗試題常因題目描述不足、思考邏輯不同而有不同答案的爭議。例如：若題目為「何者為化學上最常用的酸」，答案會因使用者或目的不同而異，應改成「何者為化學上常用來製造鹽酸或磷酸的酸」，即可知「硫酸」為公認的正確答案。

6. **表達清楚，讓學生易於了解其任務或工作。**

試題的指導語應詳述學生如何呈現答案以及配分，例如：下列單選題對的打○，錯的打╳， 每題 2 分，共 20 分。

7. **每個試題必須獨立存在，內容不宜相互重疊。**

每個試題測量獨立的觀念，不宜和其他題目重複，或提供其他題目答題的線索。例如：「溴化銀遇陽光會形成的黑色物質為何」和上述命題原則 2 的例題觀念重複，且可由該題題幹找到本題答案，編製測驗時應避免此情形發生。

8. **不要提供正確答案的線索。**

以選擇題為例，正確答案的形式、語法，應和其他選項相近，避免凸顯答案的不同。例如：「選出和溫室效應有關的氣體」，選項

分別為一氧化碳、二氧化碳、三氧化碳、四氧化碳，可以很明顯看出答案為常見的二氧化碳，應將其他選項改為其他常見但非溫室氣體的分子式，較能測出學生對溫室氣體的認識程度。

教師或研究者在編製測驗時，宜根據上述命題原則來研發試題，並時時以此原則檢視題目的適切性，再視編製測驗的類型特質進行調整，以期發展出適宜的測驗。

六、預試

根據測驗編製計畫與命題原則完成初步試題後，應進行預試來評估試題的適切性。歐滄和（民 75、民 82）提出預試可分為兩個層次：

1. **非正式預試**

非正式預試主要用於較新穎的測驗題材，編製者可將試題影印數份，找未來預試樣本來做題目；編製者可從非正式預試中獲得的資訊主要有三點：

(1) 可以確定測驗的指導語是否夠清楚；

(2) 能夠事先得知預試時可能遇到的困難；

(3) 能事先粗略估計可做完題數及所需的時間。

進行非正式預試時，人數只需 20-30 人，利於編製者觀察或訪談預試者，藉此了解正式預試或施測時可能遇到的困難。

2. **正式預試**

當測驗材料、作答方式與背景知識充足時，可直接進行正式預試。編製者希望透過正式預試獲得實證資料，以了解每道試題的難易度或鑑別度，因此正式預試需具備兩項基本條件：

(1) 樣本具有代表性，能充分代表將來此測驗常用對象和範圍；

(2) 樣本數量要多到足以求出試題母數的穩定估計值。

進行預試時，由試題編製者擔任主試者為宜，應記錄預試者對測驗指導語的反應與疑慮、預試者對題意的疑問以及受試者答完所有試題所需時間（歐滄和，民 75），以預試蒐集的資料作為整體試題分析、選擇和施測的參考依據。

研發標準化心理測驗需經過上述程序，擬定測驗編製計畫、依命題原則進行命題，再經過預試、試題分析、建立常模等過程，才能形成正式的標準化測驗。本研究屬於實作評量之命題研究，僅呈現實作命題與心理測驗相似性較高的部份，其他較不適用於實作評量之編製程序在此不加以探討。

第二節 實作評量

傳統的評量（conventional assessment）在教學評量過程中持續佔有主導地位，直到 1990 年代由教育學者提出批評，指出傳統評量所引導的教學方式，忽視學生解決問題的獨立思考能力之培養，因而興起另類評量方式（郭生玉，民 93）。其中，實作評量是廣為使用的一種另類評量方式，研究者將實作評量的相關特質整理與說明如下。

一、實作評量的本質

如前所述，實作評量要求學生以各種方式完成一項實作任務（task），透過完成任務的過程與成果來評量學生的能力與學習狀況。Stiggins（1987）認為實作評量強調學生善用技能與知識，目的

在評量學生將知識、理解化為行動的能力；以標準參照測驗的觀點來看，則強調具體地描述「每個人所能和所不能做的是什麼」（陳英豪、吳裕益，民 92）。

實作評量著重於評量學生完成一項工作的能力，而非只知道如何進行該項工作。例如：學生即使能在傳統評量中完整描述操作實驗的步驟與注意事項，其測驗結果不一定能代表學生的實際操作實驗的能力。評量學生的實作能力時，可著重於「過程」、「作品」或兩者之組合，其偏重程度視實作活動性質而定（陳英豪、吳裕益，民 92）。

偏重實作過程、不需展出作品的實作評量有音樂演奏、體育競賽、技藝表演等活動，評分者需在過程中評量學生或表演者的實作能力，重視完成實作任務各部分動作的順序與正確性。

特別注重實作作品而非過程的實作評量有學生的文章、美術作品等項目，評分者較難觀察學生的思考歷程，或是不同過程可能完成相似品質作品時，則以實作作品為主要評分依據，評鑑時可以事先擬定的教學目標作為參考。

同時重視實作過程與作品的評量有烹飪、機械修理等活動，在學習的早期著重正確順序，在學習的後期著重成品的品質，評分者需視評量目的與學習歷程來斟酌評量重點。

IChO 的實作考試依不同試題類型區分，大多數可歸納為較注重作品或過程與作品並重的兩大類，例如：選手進行無機化合物的定性分析時，可能有不同的分析方式，同樣回答出正確答案，評量時只要批改選手的分析結果，以此檢驗選手的定性分析與問題解決能力。另在有機合成實驗中，選手操作實驗的技巧與順序固然重要，

然而 IChO 競賽屬於選手學習後期的評量，強調選手所製備出的產物品質，評量時多以產物的品質為主；此外，選手可使用不同的實驗技巧得到良好品質產物，也是偏重評量產物品質的原因之一。

二、實作評量的類型與應用

根據試題的限制程度與結構度，實作評量可分為定義清楚、定義模糊與沒有定義三類（王文中等，民 88），茲分述如下。

1. 定義清楚的問題

此類問題結構度最高、定義最清楚、解題條件充分、解題的方法明確，是解題者最熟悉的問題類型，又稱為限制反應式的實作評量。優點為問題結構性高、施測所需時間短，可增廣評量範圍，缺點為解題者反應受限，不易蒐集其整合資訊能力與原創性。

2. 定義模糊的問題

此類問題沒有清楚的定義，當結構愈模糊不清、問題情境愈新奇、條件愈少、難度愈高，愈需要解題者使用高層次思考能力來解題，內容具有彈性；定義模糊的評量強調解題者統整課程資訊與經驗的能力。

3. 沒有定義的問題

此類問題結構最低、定義最模糊、解題方法未知、解題條件不足，需要解題者統整各科能力、嘗試不同的解題方式，是解題者最不熟悉的問題類型，具有高度挑戰性。沒有定義的問題能提供更多真實而多樣性的學習成就，作為評量的參考。

三、實作評量的標準

美國匹茲堡大學學習研究發展中心與教育及經濟國家中心，共

同訂定有關科學實作評量的新標準（New Standards），包含實作標準（Performance Standards）、參照性測驗（Reference Examination）與學習歷程檔案評量（Portfolio System）三大類型；其中，由專業組織團隊所發展的實作標準可分為兩個部分，茲分述如下（邱美虹、湯偉君，民 89）。

（一）實作的描述（Performance Descriptions）

描述學生應該知道的內容，以及在不同求學階段的學生應使用的方式，以表現出在科學領域中所學到的知識及技能。科學學科的實作描述分為下列八個部份：

1. 物質科學概念（Physical Science Conception）：S1
2. 生命科學概念（Life Science Conception）：S2
3. 地球與太空科學概念（Earth and Space Science Conception）：S3
4. 科學連結與應用（Scientific Connection and Application）：S4
5. 科學思維（Scientific Thinking）：S5
6. 科學工具及技術（Scientific Tools and Technology）：S6
7. 科學的溝通（Scientific Communication）：S7
8. 科學的探究（Scientific Investigation）：S8

實作描述的前四個項目強調概念理解，後四個項目強調除了概念理解之外的其他相關素養，也需要在科學課程中特別注意。

（二）學生的作業範例及評論（Work Sample and Commentaries ）

以學生實作範例與教師評論說明實作所應該達到的程度及成果，使標準更加明確，以供教師及相關人員參考。

四、實作評量的編製

研究者綜合學者的看法與建議，將實作評量的編製過程與評分工具類型分述如後。

（一）編製過程

測驗編製主要可分為下列四個步驟：

1. 明確界定評量的目的

實作評量之目的是決定如何實施實作評量與進行評分的指標（Airasian, 1994）。編製實作評量的第一個步驟即是依教學目標與教材訂定評量目的，界定評量結果的使用方式，同時對受試者的年齡與特徵作出描述（Stiggins, 1987, 1991）。

2. 設計實作評量的任務

實作評量的任務為受試者在評量中所需完成的工作，編製者應先評估完成工作的過程中，受試者需要哪些資源與材料、需要哪些問題解決能力，試題的特定項目化，以及試題內容和難度均為設計題目時應特別注意的地方（王文中等，民 88）。

決定實作評量的任務時，也需選擇與說明實作評量進行的方式，例如：紙筆與非紙筆的實作評量、典型表現評量、長期計畫的評量、示範表演的評量、實驗的評量、檔案評量及口頭發表評量等不同方式（Nitko, 1996；郭生玉，民 93）。

3. 確定實作評量的評分標準

良好的評分標準即為成功的實作評量之核心（Airasian, 1994），編製者應根據所要評量領域的重要技能來訂定評分標準，並於編製過程中檢核此標準是否能判斷受試者的表現優劣，將試題與評

分標準妥善結合，使評量的過程與結果具有參考價值。

4. 實作評量的評分方式

依評量目的不同，評分方式可能偏重評量受試者的實作過程、作品，或是兩者的組合。Linn 和 Gronlund（1995）兩位學者建議編製者可先觀察受試者的表現，再確定適當的步驟。不論採取何種評分方式，皆需符合評分標準的規範。

測驗編製者可參考上述實作評量編製步驟，配合個別領域的需求，依序進行實作評量之編製。

（二）評分工具的類型

常用在實作評量評分與記錄的觀察工具有軼事記錄（anecdotal records）、評定量表（rating scales）以及檢核表（check lists）三種方式，另有學者在上述方法中加入主觀記錄的評分方式；上述記錄方式的優缺點列於表 2-4 （Stiggins, 1994；陳英豪、吳裕益，民 92）。表 2-4 之觀察工具名稱遵照譯者用法，名稱對照如下：量表為評定量表，事件記錄為軼事記錄。

由表 2-4 可知四種檢核表的基本定義與優缺點，測驗編製者可依測驗目的選擇適用的記錄方式。以應用於教學為例，檢核表適用於評定能細分成一系列明確而具體的動作技能，而不適合用於教師對學生人格和適應狀況的概括性評定，因為這些特質並非全有或全無，用二分法勾選有無較不適當。評定量表可用於評定學生的實作過程、作品，以及學生的「個人 - 社會」發展情形，使用時可請多人評分，避免個人偏見影響評定結果。

軼事記錄可協助教師了解學生在不同情境的行為表現，記錄學

生的知識與實際行為是否相符合，持續進行記錄與比對的軼事記錄，也可提供學生行為改變程度的參考；在理想上，軼事記錄應該像一系列的「語文攝影」（verbal snapshots），要能夠真正代表學生的實際行為，教師進行記錄時卻難以保持客觀的態度，改善方式為事先決定所要觀察的行為、將事實描述與主觀描述分開記錄，並於記錄前多加練習觀察與描述的方法。對於記錄者而言，軼事記錄可提供詳盡資料，同時也相當耗時費神，需經過詳細計畫與訓練，使軼事記錄結果更具有真實性與參考價值（陳英豪、吳裕益，民 92）。

表 2-4

觀察工具優缺點列表

	定義	優點	限制
檢核表	將優良表現的特質列出，勾選有或沒有。	快速、對大量的標準十分有用。	結果可能缺乏深度。
量表	表現在從低到高的數字量表上標出	可以將判斷和理由同時呈現。	可能需要長期密集的評分者訓練。
事件記錄	學生的表現以文字方式詳細記錄。	可提供豐富的成就描述。	閱讀、書寫和解釋都十分花時間。
主觀記錄	評量者儲存判斷並將表現的描述記下。	快速又簡單的記錄方式。	保留正確的資料有困難，尤其當時間過了以後，無從檢核紀錄的正確性。

資料來源：「Student-centered classroom assessment.」, Stiggins, 1994. / 陳玉玲譯，民 92。

五、實作評量的優點與限制

研究者整理郭生玉（民 93）提出實作評量的優缺點，以及其他學者對實作評量的看法與建議，將實作評量的優點與限制分述如下。

（一）實作評量的優點

1. 提供評量歷程與結果的方法

實作評量可評量受試者的實作過程與作品，或是兩者之結合，視測驗目的而選擇使用之；此外，觀察實作過程具有診斷的作用。

2. 可以評量實作和統整的能力

實作評量以模擬或自然情境進行施測，要求學生以實際操作或運用統整能力來解決問題（Messick, 1994）；陳文典等（民 84）學者指出：「實作評量是以很自然地方式，將學生的知識、能力和傳達能力結合起來，實作評量的情境本質就是整體的、連續性的極有意義的全方位的評量。」。在似真情境下評量學生實際操作技巧和較高層次的問題解決能力，這是實作評量的獨到之處。

3. 符合近代的學習理論

建構論主義（constructivism）強調學習者應主動建構知識，實作評量的本質即為考慮學生先備知識，使學生主動建構自己的意義，十分符合建構論主義。

4. 有助於改進教學

實作評量同時注重實作的歷程與結果，強調程序性知識（procedural knowledge）的評量，也就是指學生能做的知識，例如做化學實驗這樣的程序性操作知識（張春興，民 85），可改善學生學習偏重於陳述性知識記憶的情形。

（二）實作評量的限制

1. 費時費力的一種評量方式

實作評量在編製、實施與評分方面，均較傳統測驗耗時費力，測驗編製人員的專業能力與測驗編製技巧須在水準之上，所編製的實作評量才能有效測出受試者較高層次的操作與思考能力。

2. 無法測量到所有教學目標

實作評量應評量客觀測驗所無法測量的學習成果，測驗時所涵蓋的內容範圍較窄，可視需要搭配其他實作評量或客觀測驗，來增進評量的代表性。

3. 評量信度較低

實作評量的結果容易受評分系統的影響，而有不一致的結果，降低了評量的信度與效度；測驗編製者應審慎規劃評分系統，以期提高評量的客觀性。另有學者認為有效的運用實作評量時，高信度的必需性有待研究累積來檢驗（Moss, 1994）。

4. 評量結果的推論性低

在一個工作上的實作表現通常較難推論（generalizing）到其他的工作表現（郭生玉，民 93），例如：一個學生有優異的美術作品製作能力，卻無法由此推論他的游泳競賽表現。為改善這項限制，可採用多樣的實作工作，由不同的工作表現累積資料。

綜合上述，當學科特性與測驗目的需要使用實作評量時，相關團隊應投入大量資源與人力，擬定長期而完整的測驗編製計畫，充分發揮實作評量的優點，設法改善其侷限之處，結合多種項目的實作內容，再視需要配合客觀測驗，以期達成測驗目的。