

科學學習成就評量

——I. 命題與閱卷——

國立臺灣師範大學生物系 鄭湧涇

壹、緒 言：

科學學習成就的評量 (Assessment) 一向是我國各級學校科學教師深感困擾卻又畏於深入探討的問題。國科會科學教育組有鑑於此，乃藉舉辦「科學教育研討會」的機會，多次箋邀國內科學教育界諸先進們，就科學教學評鑑的問題，做了幾次深入的研討。如今，他們的遠見及努力已經有了初步的迴響（潘音利，民70年），這是一個很好的開端，相信在最近的將來，這些問題將會因科學教育界先進及同仁們，無所避諱的深入研討，而能獲致若干改進，使我國各級學校的科學學習成就評量，能趨向更合理、更公平、更能發揮應有功能而向更人性化邁進。

筆者於本刊第36期中，曾就科學教學評鑑的一般理論與實際運用原則做了簡單的介紹（鄭湧涇，民69年），因此，本文將僅就科學學習成就評量的實用技巧方面，做較深入的探討。「野人之愚」之所以不揣粗陋，再度「獻曝」，意在「引玉」，假若能獲致些許共鳴或批評，則不僅是鄙人之得，也應是探啓學習成就評量「寶盒」的嘗試。

貳、科學學習成就評量的意義與試題製備：

一、對學習成就評量應有的認識：

學習成就評量的成敗，繫於工具之良窳甚鉅，所謂「工欲善其事，必先利其器」，因此，於實施評量時，工具之選擇及製作應屬第一要務。由於我國的教育環境至為特殊，評量學生學習成就的目的，幾乎絕大部分都是為了「評定等第」。雖然「評定等第」是評量學生學習成就的目的之一，但是無可諱言的，它至少應該還具有其更積極的目的，亦即診斷教學缺失和學習困難以引導學習的功能才對，學習成就評量之能否達成本項功能，與評量工具之選擇和製作是否得當有極密切關係，因此，在決定評量學生的學習成就以前，應該先對各種評量工具的特徵及製作，有深入的瞭解，方能「利其器而善其事」。

二、實施評量前的準備：

在實施評量前應先決定：

1. 評量的目的：評量的結果作何用途，是為診斷教學缺失和學習困難或是為評定等第。
2. 評量的方式：針對評量的目的，選擇適當的評量方式，例如：假若評量的目的是為評定等

第，便應先決定究應採用「常態評量」或是「偏態評量」。

3. 評量的依據：評量的基準是擬將學生和一般「常模」(Norm group)做比較，來評定等級，即所謂「常模參考評量」(Norm-referenced Assessment)抑或是擬將學生的實際學習成就與其應達到之學習標準(Criterion)比較，即所謂「標準參考評量」(Criterion-referenced Assessment)。

4. 試題的類型：在每次評量時，擬使用「主觀型試題」(Subjective item)或是「客觀型試題」(Objective item)，應先決定。主觀型試題問答題，申論題和解釋名詞等，而客觀型試題則包括是非題、選擇題、配合題等。我們應該因應評量的目標和擬評量的內容來選擇，以發揮各類型試題之長處，而避其短處。

5. 何時進行評量？需用多少時間？：就學習的心理來說，實施學習成就評量前，應事先通知學生實施評量的日期以及所需的時間，讓學生有充分的心理準備，而能進行有效的學習(Effective learning)。

6. 擬評量的範圍和各部分內容所佔的比例：在製作試題以前，假若不事先計劃妥當，往往會於事後才發現，試題的比重殊不平均，無意的特別強調了某一範圍，而忽略了其他部分，因此，在製作試題以前，先製備一份「命題綱要」(Specification)或周詳的「評量計畫」(Test Plan)將大有助益(鄭湧涇，民69年)。

三、選擇試題類型應有的認識：

在學習成就的評量上，如何適當製作和善用每一種類型試題，比究應選擇何種類型試題來評量，更為重要，因為，實際上，兩種類型試題均能有效的評估學習成就。亦即如何提高試題的品質，以提高評量工具的信度，比究應選用申論題或選擇題來評量，更為重要。

由於主觀和客觀型試題各有其特性，因此，在討論命題的原則和注意事項之前，筆者擬先就兩類試題，可以評量的成就，稍加討論。

1. 在下列情況下，以使用主觀型試題來評量為宜。

① 擬評量學生的創意、組織或綜合科學知識或語文表達等比較複雜的成就，以客觀型試題通常比較不容易評量者。

② 擬評估學生對某些科學知識的意見或態度。

③ 擬評估學生分析、應用或評鑑整體科學概念的能力。

④ 受試人數不多而且試題擬只使用一次或有充足時間閱卷而無足夠時間命題。

⑤ 教師個人於評閱答案時，能明確訂定恒一的給分標準。

2. 在下列情況下，則以使用客觀型試題來評量為宜。

① 試題擬重複使用，而且受試者人數眾多。

② 要求測驗結果具有高信度。

③ 評量結果必須儘可能公平、公正。

④ 有充足時間命題，而缺乏足夠時間閱卷。

- ⑤ 教師個人於評閱答案時，無法把握明確而恒一的標準。
3. 下列情況下，假若命題和閱卷適當，則兩類試題均可達成預期目標。

- ① 評量學生的各方面學習成就。
- ② 考驗學生是否了解並能運用科學定律。
- ③ 考驗學生應用科學知識以解決問題的能力。
- ④ 考驗學生的批判思考能力。
- ⑤ 考驗學生記憶和運用科學知識的能力。
- ⑥ 考驗學生分析和綜合科學知識的能力。

參、實施評量時應注意的命題原則：

各類型試題之製作，均有一些基本原則可供遵循，雖然這些原則並非“定律”不過假若能於命題後，根據這些原則稍加檢討，定將有助於提高試題之品質。

一、主觀型試題之製作原則：

1. 儘量使問題的「預期答案」(Expected response)，在內容和型式上都能明確界定。千萬不要製作“讓學生儘量發揮其個人意見，而缺乏明確內容”的試題。

2. 在問題中，明確指出學生應回答的範圍。請比較下列兩個題目，何者較佳？

- ① 水分在植物體內如何上升？
- ② 試列舉三種解釋水分在植物體內如何上升的學說。

顯然的，第一個問題太籠統，學生可能誤解，答案範圍也未明確界定，將來閱卷將會很難把握一恒一的標準。而第二個問題則明確指出命題者的意圖，學生的反應有一定的方向可資遵循。

3. 不要製作「預期答案」太長的試題。因為預期答案太長，往往造成題意模糊，學生無從下筆或不知何處結束，閱卷時，也不易秉持某一標準，因而造成不公平，以致降低試題之信度。

4. 於每一試題之後，註明回答該題所需的時間和該題的分數，讓學生能夠根據命題者建議的時間和題分，斟酌其答案的繁簡和內容深度。例如：

「說明影響光合作用速率的四種環境因素。」(8分鐘；8分)

5. 命題完畢後，先自己試做標準答案，以免題意不清或答案太繁。
6. 最好不要有選題，例如：五題選答三題；因為，選題意指評量基礎已經不公平。

二、選擇題之製作原則：

選擇題是最常用的客觀型試題，假若試題的製作良好而且講求技巧的話，它幾乎可以評量任何科學知識和學習成就。它的最大特性是閱卷快速且客觀，因此在學習成就的評量上，是最具效率的試題之一。但是，假若試題製作不當，則效果將大打折扣，因此，其製作技巧需特別講求。

大體說來，選擇題因其答案性質的不同可分兩類，即「正確答案型」(Correct answer)和「最佳答案型」(Best answer)，所謂「最佳答案」是在兩個或兩個以上都是正確的答案中，選出一

個最好的答案。選擇題的結構分「題幹」(Stem)與「可能答案」(Alternatives)兩部分,「題幹」可用「完整問題式」(Direct question)或「未完成敘述式」(Incomplete statement),而「可能答案」則包括「標準答案」(Answer)和「擾亂答案」(Distracters)。其製作原則如下:

1. 題幹本身應為一具有完整意義的句子。請比較下列兩個題目。

① 葡萄糖

- A. 為植物體內養分之運輸態。
- B. 與本氏液反應,產生紅色沉澱。
- C. 分子量為 180,故不能透過細胞膜。
- D. 完全氧化了二氧化碳可產生 35 個 ATP。

② 下列何者為高等動物體內醣類之運輸態?

- A. 果糖。
- B. 蔗糖。
- C. 麥芽糖。
- D. 葡萄糖。

顯然的,就本項原則來說,後者比前者為佳。

2. 題幹句子應長些,而可能答案的句子則儘量短些。如:前例之後者即比前者為佳。

3. 除非擬評量非常重要的科學概念,否則題幹以不使用「否定」為宜。必須使用否定時,亦應於否定詞下劃線,以提醒學生注意,否則假若學生因疏忽否定字眼而答錯該題,實非評量成就之本意。

例題:下列何者不屬於維管束植物?

4. 每一題幹之「可能答案」以四或五個為宜,其中應有而且最好“只有”一個「正確的」或「最好的」答案。

5. 每一「可能答案」在文法、語氣和句子長度等方面,應儘可能保持一致。有些命題者,往往為了使正確答案達到完全正確,無懈可擊,因此往往將句子堆砌得特別長,其餘的「擾亂答案」,反正是不對的,就草草湊數,句子就短了些,於是“有技巧”的受試者,只要抓住命題者的習慣,便可“猜對”大部分題目了。

6. 所有的「可能答案」都應合理,不可不合邏輯,否則便失去擾亂效果,學生也就會很容易猜出何者才是正確答案了。

7. 正確答案應以任意次序出現,平均分配,不可為了閱卷方便以一定次序排列,如:1、2、3、4、1、2、……或1、1、2、2、3……等。

8. 可適當的使用「以上皆是」或「以上皆非」為「擾亂答案」。

三、是非題之製作原則:

是非題旨在評量學生鑑別正確敘述、事實、概念或名詞意義等的能力,因此,在命題時,必須特別注意其明確性,不可稍有含糊,否則將會造成是非不明,也因此,在使用其他類型試題來評量某一成就可能比較適當時,最好不要使用是非題。

在製作是非題時，應注意的地方是：

1. 避免使用一般性敘述來命題，而應儘量詳細說明題意，以免題意不明，學生無法明確判別。請比較下列兩個試題：

① () 細胞壁的成分是纖維素。

② () 維管束植物之細胞壁，主要由纖維素所構成。

顯然的，後者就題意來看，要比前者明確，也比較不會引起爭議，因為有些真菌之細胞壁並非由纖維素構成。

2. 避免零碎、不重要或純記憶性的命題，例如：

() 華特森 (Watson) 和葛立克 (Crick) 在 1966 年發表 DNA 雙螺旋的結構。

3. 避免使用否定，尤其是雙重否定的命題，例如：

() 蝴蝶“不是”一級消費者。

4. 避免以冗長而複雜的敘述來命題。

5. 避免在一命題中，包括兩個或兩個以上的科學概念，除非其有因果關係存在。請比較下列兩個命題。

① () 蚯蚓缺乏四肢，體表佈有剛毛，不能感受光源。

② () 蚯蚓藉剛毛運動。

後者顯然比較清晰的表達了命題者的意圖，因此學生不易混淆，答案也明確。

6. 答案為「是」與「非」的命題應約各佔一半，否則相差太多時，有考試技巧的學生便可藉是與非出現的頻率來猜題，使評量結果信度偏低。

四、配合題之製作原則：

使用配合題主要是為了評估學生鑑別科學知識之間的相互關係的能力，因此，在內容上，以涵蓋與某一科學概念有關的零碎小概念，最為恰當。一組配合題，通常可分為「前提」(Premises)和「選目」(Responses)兩個部分，命題時，要求學生由「選目」中，選出與某一「前提」有關的答案，因此，其特徵在要求學生尋求兩件事實或敘述之間的關係。於製作時，應注意的原則如下：

1. 在一配合題中，「前提」與「選目」的內容均應是同質的(Homogeneous)，不可將不相關的內容勉強湊合。

2. 「前提」與「選目」的數目應不同，同時應說明清楚「選目」可以使用的次數，如：每一個「選目」只能使用一次或兩次或兩次以上。

3. 「前提」的句子通常應比「選目」長，而且將「選目」置於右方。

4. 「選目」應依合理或合乎邏輯的方式排列，避免與「前提」之間有任何順序上的相關存在。

5. 每一配合題應附一作答說明，使學生對作答方式能一目瞭然。

6. 一配合題的所有「前提」和「選目」應置於同一頁內，不要分開在兩頁上，以免增加學生答題時翻頁的麻煩，甚至因疏忽而造成錯誤。

五、簡答題之製作原則：

簡答題是指要求學生填寫幾個很短的字或詞句的命題，「填充題」即屬此類。這類命題主要用來評估學生對於非常重要的名詞、原理或概念是否清楚，因為這些知識很重要，所以往往需要理解和記憶，簡答題即在評量其理解或記憶的能力。大體上，它有兩種方式呈現，一為「問題型」，另一為「填充型」，請比較下列兩題：

① 酵母菌以何種方法進行無性生殖？（出芽生殖）

② 馬鈴薯利用（塊莖）行無性生殖。

在命題時，應注意下列諸原則：

1. 命題宜愈明晰愈好，以免產生誤會，請比較下列兩個試題：

① 掠食其他動物的動物稱為（ ）。

② 掠食其他動物的動物，在生態營養階層上，稱為（ ）動物。

後者比前者，在答案的要求上，更為明確。

2. 不可由教科書上摘錄一段句子來命題，這是大多數教師，為了節省時間，常犯的毛病。正確的作法是，應該以自己的語辭，從新造句。

3. 一般來說，「問題型」的命題，對答案的要求，較「填充型」為明確，故宜多用「問題型」命題。

4. 假若期望的答案，帶有單位時，應將單位標示出來。

例如：人類唾液中的澱粉酶，作用的最適溫度是幾度？ 答：（ ） $^{\circ}\text{C}$ 。

5. 將填寫答案的空格，放在最右邊，線條或括弧應等長。例如：

① 植物的光合作用在那一個細胞器內進行？ 答：_____

② 洋菇屬於真菌門的那一綱？ 答：_____

③ 答：_____

6. 假若題幹使用「填充型」，原則上以一題一空格為原則儘量不要在一個題目內，要求學生填兩個或兩個以上的空格。因為空格愈多，通常題意模糊。例如：

① 維生素_____與_____的代謝有關，缺乏的話，會產生_____病。

② 維生素D與人體內，_____的代謝有關。

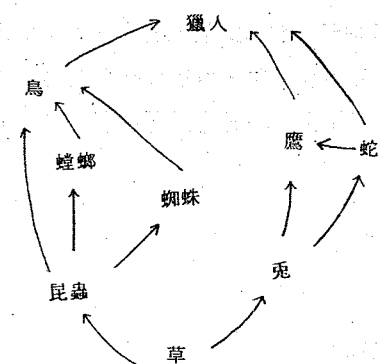
後者答案比前者明確得多，而前者則可能有許多答案存在。

六、「解釋題」或「實驗題」的製作要領：

此類試題在評量學生應用科學定律、綜合資料和解釋實驗數據等的能力，主要由一段科學事實、數據或實驗結果以及以這段資料為基礎的客觀型試題所構成。基本上，它還是屬於客觀型試題，不過在內容上，必須以某些已知資料為基礎來作答而已。

於製作這類型試題時，先設計或轉錄一段科學概念，科學實驗結果或圖表等，再根據其內容，設計數題客觀型試題。例如：

※以下是一個生態系內的食物網，請根據此食物網回答第1—4題。



1. 在本食物網中，那一種生物之生物量最大？

答：_____

2. 在本食物網中，蛇是屬於那一種營養階層？

- A. 生產者。
- B. 一級消費者。
- C. 二級消費者。
- D. 三級消費者。

答：_____

3.

答：_____

肆、試卷的評閱：

命題與閱卷是學習成就評量的兩項主要工作，客觀型的試題，閱卷時至為簡便快捷；至於主觀型試題，則需要較長的時間以及特殊的訓練方能評閱。由於教師與學生個人的觀點和愛好不同，因此，同一分試卷，不但不同的評閱者給分不同，有時甚至同一個評閱者，評閱兩次所給的分數亦異，可見評閱者在評閱標準的把握上，頗有困難。因此，主觀型試題之評閱，必須要有一定的原則來遵循，以期待人為的誤差減至最低。

一、評閱主觀型試題應注意的原則：

1. 明確訂出該題期望答案的項目以及答對每一項目的給分，然後計項給分。
2. 應用適當的給分系統。

在給分時，常用的給分系統包括：

① 計點法 (Point method)：將試卷與期望答案比較，分項計點，例如假若該題有8分，期望答案有三項，答對第1、2項各給3分，答對第3項給2分。

② 分級法 (Rating method)：又稱為「三次評閱法」(Three-Stage Process)，即第一次評閱時，先將所有的試卷，依據答案的品質，先分成三級，第二次評閱時，再將每一級更細分為三級，此時所有試卷已分成九級，第三次評閱時，則比較第三和第四以及第六和第七級的試卷，看是否第三級之試卷品質確實比第四級的試卷高。確定無誤之後，再依級給分；當然，必要時，第一和第九級試卷可以再細分之後再給分。理論上，級數愈多，給分愈公平，但是限於評閱者能完全把握並鑑別試卷的優劣才行。

3. 事先決定如何處理與期望答案無關的答案，給多少分、不給分或者扣分。
4. 一次評閱一題，等所有試卷的該題都閱畢後，再評閱下一題。
5. 閱卷時，不要看學生姓名，以免給分受印象影響。
6. 若有兩位以上的評閱者，應分題評閱，不可分卷評閱，所有試卷之某一題，以由同一人評閱

為原則。

7. 可能的話，每一分試卷均能由兩個人重覆評閱，並計算兩人所給分數之間的相關，以求取給分的信度。

二、評閱客觀型試題應注意的原則：

客觀型試題之評閱至為簡單，無庸贅述。不過有關「倒扣」的問題，則值得探討。

1. 倒扣的方法：

是否倒扣以及擬如何倒扣，均應於測驗以前，在試題上說明清楚，或向學生宣布，讓學生可以自己決定要不要猜答案，一般人採用的倒扣方式，均是以下列公式為依據：

$$S = R - \frac{W}{N-1}$$

S ：實得分數

R ：答對題數

W ：答錯題數

N ：可能答案數

因此，是非題倒扣適用的公式，便是 $S = R - W$ 了。

2. 有必要倒扣嗎？

倒扣雖有公式和原則可以遵循，不過在決定實施倒扣之前，我們似乎應該先考慮，是否有必要倒扣？倒扣之後有何益處？倒扣的惟一好處大概只是在嚇阻學生以僥倖猜選答案來取分，問題是，猜選答案這種行為是否有必要予以禁止呢？更何況，倒扣的公式本身，就不合理，因為它是基於“所有答錯的答案都是猜錯造成的”這種假設上，因此，必須由對的答案中，將“猜對”的部分扣除。因此，在學習成就評量方面，為了鼓勵學習（評量也應是一種學習的機會），筆者認為還是以不倒扣為宜，甚至在適當的機會，還應該鼓勵學生“合理的”猜猜看，因為“合理的猜”（Rational guessing）本身已經包括了推理、分析、綜合和評鑑等心路歷程，而這些心智運作在科學學習上是有其價值的。

此外，還有下列考慮：

① 倒扣之後，通常並不會改變學生個人在班上的排名。

② 假若試題出得夠好，學生將很難由猜選答案上獲益。

③ 很少有學生是真正的盲目猜選答案。

④ 學習並非「有或無」（All-or-none）的過程，因此，就學習的分量上來說，適當的鼓勵學生運用“合理的推理”來猜選答案，也將有利於學習。

伍、教師們常忽略的事項：

學習成就評量是教學的回饋，學習的指標，其意義於教師和學生均極重要，因此，教師必須慎重處理，尤其是在命題、閱卷和結果的統計方面，更得具備相當的專業知識方能圓滿勝任。最後筆者願意就一般科學教師，在進行學習成就評量時，容易疏忽的地方，加以討論，與各位教師共勉，以改進我們的學習成就評量技術。

1. 在命題和閱卷時，太相信主觀判斷，以致命題不均或閱卷時產生偏差，導致評量結果的信度

偏低。欲糾正本項缺點，宜於命題時先設計「命題綱要」或「評量計畫」；閱卷時，遵守各種試題之閱卷原則。

2. 臨陣磨槍，往往在必須交出試題時，才“趕緊”命題，因此造成部份命題，品質低下。

3. 給學生答題的時間太短，學生“趕”著做題，無暇思考，造成評量結果，信度偏低。通常在沒有作弊的顧忌時，不應使用「快速測驗」(Speeded test)。

4. 評量專業知識不足，以致命題，實施評量和閱卷等均有瑕疵。固持個人的絕對標準，以致令大部分學生，表現之成就偏低，終而扼殺學生之學習興趣。

5. 缺乏做「試題分析」(Item analysis)的技巧或習慣，以致無從建立題庫，每次評量都得重新命題，事倍而功半。

6. 評量結果未做「統計分析」(Statistical analysis)，「平均數」(Mean)、「標準差」(Standard deviation)、「測驗標準誤」(Standard error of Measurement)以及「信度」(Reliability)和效度(Validity)等無法測知，以致評量技巧之改進無方。

假若筆者之臆測是真，則顯然的，欲改進我國各級學校的科學學習成就評量，實有加強科學教師的評量知識和技術的必要，因此，本文的下半將就「試題分析」和評量結果的「統計分析」兩方面，再做粗淺的介紹。我們的真切希望是，在不遠的將來，學生們將不致因科學科目的成績不及格而喪失學習興趣，教師、教育行政單位和學生家長也不會為絕大部分的學生，科學科目不及格而煩惱、自責；因為，科學教育的發展應是全面的，除了在教材、教法上必須力求革新、改進之外，至少，我們還需要更合理、更公平和更人性化的學習成就評量才對。拋磚之言，企望各位先進、同好的玉石之聲和指教。

□

陸、參考文獻：

1. 潘音利，科學教學評鑑的探討。科學教育月刊，42：31—35，民國70年。
2. 鄭湧涇，科學教學評鑑的理論與實際運用。科學教育雙月刊，36：2—10，民國69年。
3. Ebel, R. L., *Essentials of Educational Measurement*, 2nd ed. 1972, Prentice-Hall, Inc., Englewood Cliffs, N. J.
4. Gronlund, N. E., *Measurement and Evaluation in Teaching*, 3rd. ed. 1976, Macmillan Publishing Co., Inc., New York.
5. Lombard, J. W., *Preparing Better Classroom Tests*, The Science Teacher, Oct. 1965, p. 33—38.