科學學習成就評量
II.　評量結果的統計分析

鄭湧涇
國立臺灣師範大學生物系

壹、前　言

科學教學的改革，應該是多方面的，至少在教材、設備、師資以及評量等方面，均應齊頭並進。多年來，我們在科學教育的改革上，尤其是課程發展，已經有了小小的成就。可是，在師資、設備以及評量等方面，卻仍有待積極努力，否則課程發展方面的成就，即將大打折扣。許多人在科學教育的革新活動中，稍遇挫折，便以一聲無奈，將一切歸之於"考試領導教學"，好像考試是諸惡之源。其實，就達成科學教育目標的過程而言，考試（我寧願稱之爲評量）本就應該領導教學；因爲，國內外許多科學教育學者[1,14]均認爲，正確的學習成就評量（所謂考試），其目的應該不只是爲了評定學生的等第而已，至少，它還應該具有鑑別學生的學習困難，引導學生的學習，發掘教師的教學缺失，以及激勵學生旺盛的學習意願，以達成學習目標等功能才對。因此，考試"領導"教學乃理所當然，又何庸感歎！所以，我們擔憂的應該不是考試領導教學，而是"低劣的評量"導致教學的偏離正軌，也因此，如何改進評量，使國內各級學校的學習成就評量臻於完美，才是關鍵。

欲改進評量，使評量的結果趨向更公正、公平和客觀，而且有助於引導學習，除了應在命題的準備和技巧上，特別斟酌之外[1,3,6,7]，在評量工具的評鑑與評量結果的統計分析方面，亦應多做探討，方能逐漸剔除不當命題，淘汰無效或低劣選目，提高評量工具之信度（Reliability）和效度（Validity）。就其功能來說，評量工具之評鑑與評量結果之分析正如一面明鏡，可以明確指出評量上的任何優劣。所謂人必「攬鏡以照，方見瑕疵」，在實施評量時，不要以爲閱完卷，給了分就算完事，接下來的分析和評鑑過程，在評量目標的達成和技術的改進上，才是更爲重要的。

貳、試題分析

試卷閱畢，應即進行「試題分析」（Item analysis），計算每一試題的「難度指數」（Difficulty index）和「鑑別指數」（Discrimination index）。難度指數代表該試題的難易程度，有些評量學者[11]以答對該題的人數，佔全部受試者的百分比來代表；而有些評量學者[10]則以答錯試題的人數，佔全部受試者的百分比來代表；本文擬採用後者的定義，因爲，事實上，前者應是指其「容

易度」而非「難度」；在本文中，難度指數將以 P 來代表。而鑑別指數則代表該一試題是否能精確鑑別高成就與低成就的學生，鑑別指數大的試題，較能區別高成就和低成就的學生，亦即試題的品質較高；在本文中，鑑別指數將以 D 來表示。

一、試題分析的步驟

我國目前仍缺乏專業的測驗機構，來進行評量結果的統計分析，因此，教師們應該熟悉各種試題分析的知識，方能在評量之後，藉試題分析來改進命題與評量。通常實施試題分析的步驟如下：

1. 將試卷依得分的高低排列。
2. 由最高分向下取全部試卷數的 27％ [10] 或三分之一 [11] ，稱爲「高分組」。
3. 再由最低分向上取與高分組相同份數的試卷，做爲「低分組」。
4. 分別計數高、低分組，選答各試題每一選目的人數，記錄在「試題卡」（ Test item card ）上，如圖一。
5. 計算各試題之「難度指數」，以百分比表示，其計算方法如下：

$$\text{難度指數}(P)=\frac{T-(R_U+R_L)}{T}\times 100$$

R_U ：高分組答對該題人數

R_L ：低分組答對該題人數

T ：全部取樣人數，即高、低分組試卷份數之和

6. 求取各試題之「鑑別指數」，其計算方式如下：

$$\text{鑑別指數}(D)=\frac{R_U-R_L}{\frac{1}{2}T}$$

7. 評鑑每一試題，「擾亂答案」（選目）之有效性。
8. 將所有試題，依其難度指數與鑑別指數值，製作綜合分析表，並求出其平均值；其綜合分析的方法，請參考圖三。

二、試題的評鑑與改進

試卷閱畢之後，最好能夠將試題分析結果與試題一起記錄下來，逐題加以評鑑，其P和D均適切者，蒐集起來，供自己將來命題之參考；而P或D不太適當的試題，則必須加以修飾，以改進其P或D值，無法改進或修訂者，則予以棄卻。爲了使試題的評鑑，不致消耗太多時間，且能明確鑑別並掌握每一試題之優劣點，「試題卡」的製作不失爲一有效的方法。

1. 試題卡的製作：

將試題連同試題分析結果，抄錄成卡，稱爲「試題卡」，試題卡可以因應各人的喜好與使用的方便，改變其格式，在本文中，擬介紹一簡單的格式，供讀者參考，別忘了它只是一種「格式」，不是「範例」，讀者可以有各種不同方式的變通。「試題卡」至少應包含三個部份（圖一）：

① 知識內容、階層和編號。

② 題目（試題和選目）。

③ 試題分析結果。

在第一部份裡，內容指學科知識內容；而階層則指回答該試題所需的思想操作，是屬於認知、德育（情意）和技能之中的那一領域（Domain）和認知階層[2]，[4]。第二部份則包括試題與標準答案，正確答案通常在選目前以＊號表示。第三部份是試題分析的資料和結果；這個部份以另行印製爲宜，因爲，試題於每使用一次，便有一次分析，同一試題之試題分析結果，可以浮貼在一起，供評鑑試題的參考。

編號：B-106

科　別：國　中　生　物

內　容：第9章生物的生殖

階　層：認　知　（綜合）

題　目：

食物儲存在冰箱內不會腐敗，是何原因？

A.冰箱太冷，將細菌殺死了。

＊B.在冰箱裡，細菌繁殖的速率很慢。

C.冰箱裡的食物很硬，細菌不易侵入。

D.冰箱裡的冷氣向下沈，清除了冰箱內的細菌。

※ 試題分析記錄 ※

可能答案	A	Ⓑ	C	D	空白
高分組(16人)	0	16	0	0	0
低分組(16人)	3	6	2	4	1

＊＊難度指數(P)：31％

＊＊鑑別指數(D)：0.63

備註：

圖一：試題卡的格式之一

類似此種試題卡，可以依編號或科別和內容來分類整理，以建立自用小型題庫，供將來命題之用。假若能更進一步以圖書編目的方式來處理，則在參考和應用上，將更爲便捷。不過，在使用時，必須特別注意的是，在每次測驗後，均應將試題全部收回，否則學生反覆演練，再好的試題，亦將失去其難度和鑑別度，使評量的信度降低。

2. 評鑑試題的另一種方式：

假若你覺得製作試題卡，耗時太多，或者，當試題尚在修飾階段，在試題卡上修改不太方便時，

可以先以這種方式來評鑑試題。其方法如下：

① 將空白試題剪下浮貼於硬紙或卡片上，或者直接就拿一份空白試卷來處理即可。

② 分別將高、低分組，選答每一「可能答案」（選目）的人數填在每一「可能答案」之前或後，例如：圖二所示。第一和第二個數字分別代表高、低分組的人數。

③ 再分別計算該試題之P和D，將結果記錄在試題編號前面。

	12. 就人類而言，水污染最嚴重的後果爲何？	
P＝57％	A.小溪中的魚、蝦死了。	（6－7）
D＝0.2	B.水溝很臭，蚊蠅很多，很不衞生。	（4－9）
	* C.水中生態系的平衡被破壞了。	（16－10）
	D.水很臭，不能飲用或灌溉了。	（3－2）
	（空白）	（1－2）

圖二：評鑑試題的另一種方式。

假若，有任何一「可能答案」（選目）沒有人選，也就是表示所有學生，不論是否具備該試題擬測定的知識，均可看出該「擾亂答案」是不合理或荒謬的；也就是說，該「擾亂答案」已失去其應有的功能。此時，必須另行設計一「擾亂答案」來取代，否則原設計爲四選一的選擇題，假若有一「擾亂答案」沒有人選，實際上，就學生來看，不過是三選一而已。許多難度與鑑別度均不太理想的試題，都是由於某一或兩個「擾亂答案」不十分有效所致，此時，只要將無效的「擾亂答案」另行設計取代，便可大幅度改進該試題的品質。

參、評量工具之評鑑

評量工具之品質如何，除了可藉試題分析略窺一、二之外，由基本的「描敘統計」（Descriptive statistics）資料以及信度和效度分析，亦可瞭解其是否適當。因此，評量完畢後，將評量結果稍做統計分析，以適當方法求取信度和效度，實爲從事科學教學者，必須具備的能力之一。

一、評量工具之綜合分析

分別以試題之P和D爲兩個向度座標，然後依各試題之P、D值，將題號填入座標內之空格，便成爲一綜合分析表，如圖三。 由這種綜合分析表，立刻可以瞭解在某一次評量中，共有多少試題，那些試題，其P、D值均達到理想，那些試題之P值或D值或兩者皆未達理想，而需要改進。 此外，還可知道該次評量中，所有試題P值（$\bar{P}$）和D值（$\bar{D}$）之平均（Mean）是否合乎要求。

1. 難度指數（P）之評鑑：

一般而言，試題之P值以在50％左右爲宜，亦即我們希望高分組的學生都答對，而低分組的學生都答錯，也只有在這種情形之下，試題的D值才能趨近於完美值（即1）。可是實際上，任何一試題，均有被學生盲目猜對的機會，其猜對的機會爲$\frac{1}{n}$（n爲「可能答案」的數目），例如：四選一的

D（鑑別指數）

P（難度指數）%	0以下	0－0.19	0.2－0.29	0.3－0.39	0.4－0.59	0.6以上
100 1 80				31	12,21	
79 1 60	50	18	37,39	1,14 25,29	27,35 44	3，7，16
59 1 40		8,40	22,43 47,48	6，9，20 24,46	13,15,19 28,36,45	10,23
39 1 20	32	2,26	11,33 42	4，5，38	17,30	34
19 1 0		49	41			

$\overline{P}=48\%$

$\overline{D}=0.37$

圖三：試題之難度指數與鑑別指數綜合分析（$\overline{P}$：難度指數之平均數；$\overline{D}$：鑑別指數之平均數）

選擇題，猜對的機會約爲$\frac{1}{4}$，因此，理想P值便會低些。因爲在實施學習成就評量時，我們通常都希望學生的平均得分，落在「滿分」（Maximum possible score；在國內，都常都以100分爲滿分）與「機遇得分」（Expected chance score）；就四選一之選擇題來說，即爲$\frac{100}{4}=25$分）的中間，若試題全部爲四選一之選擇題時，即爲$25+\frac{1}{2}(100-25)=62.5$分。因此，乃有人認爲，理想的P值，可以低至100減「期望平均得分」，就上例來說，就是$100-62.5$，亦即37.5%。

總而言之，假若我們實施學習成就評量的方式是一種「常模參考評量」（Norm-referenced test）的話，則我們的目標便是要將受試者的得分儘量分散開來，以便比較學習成就的優劣，此時，試題的P值，便應以在中等（50%左右）爲宜，因爲根據Cronbach和Warrington[9]的研究，P值愈集中於中等部份，則得分愈分散，P值愈分散，則得分反而愈集中。

2. 鑑別指數（D）之評鑑：

理想的試題應該是所有高分組的學生都答對，而低分組的學生都答錯，此時，D值爲1；假若相反的，高分組的學生都答錯，而低分組的學生都答對，則D值爲－1；因此，D值是介於－1與＋1之間。就「常模參考評量」而言，D值愈大，學生得分便愈分散，標準差（S.D.）也比較大；亦即試題愈有效，品質也較高。此外，一般而言，試題之平均D值愈大，評量工具之信度也愈大，因爲：

$$\frac{(\Sigma D)^2}{6}=\sigma^2$$

σ^2：學生得分的變方（Variance）

而 σ^2 愈大，則信度愈高。

那麼到底D值要多大才算好呢？完美的D值（ D＝1 ）通常不容易得到，因此，在學習成就評量上，我們通常以表一的標準來評鑑[10]。

表一：試題鑑別指數（ D ）的評鑑

D　值	評　鑑
0.40以上	極佳的試題
0.30－0.39	尚可的試題，可能需要稍加改進
0.20－0.29	不佳的試題，必須加以改進，或棄卻
0.19以下	極差的試題，應棄卻

二、評量結果之描述統計

描述統計的資料，亦是分析和描述評量工具，不可或缺的指標，現分述如下：

1. 平均數（ Mean ）：

平均數通常指算術平均數而言，以M表示；它的大小可以代表評量工具的難易。就學習成就評量來說，理想的平均數應是滿分與機遇得分之中間值；因此，在試卷閱畢後，必須求出其眞正的平均數，以便與理想值比較，看是偏高或偏低了。其求法如下：

$$M = \frac{\Sigma X}{N}$$

ΣX：每一試卷分數的總和

N　：試卷數

假若實際值比理想值低，代表評量工具太難，反之亦然。

2. 標準差（ Standard deviation ）：

標準差代表評量結果之分散度與變異性，通常以SD或 σ 表示 ，實際上，它是「變方」的平方根。標準差愈大代表分數之變異性愈大，當其他條件相同時，標準差愈大，評量工具之信度也愈大。理想的標準差，大約爲「滿分」與「機遇得分」之差的六分之一，評量結果的標準差，以約略與理想值相似爲宜。標準差的求法如下：

$$\sigma \text{ 或 } SD = \sqrt{\frac{\Sigma d^2}{N}}$$

d ：各分數與平均數之差

N：試卷總數

3. 相關研究（ Correlation Study ）：

於試卷閱畢後，將得分登記在成績記錄表上，然後與類似性質的評量結果，求取相關（ Correlation ），亦是分析評量結果的重要方式之一。相關的大小通常以「相關係數」（ Correlation coefficient ）即r表示，r值介於－1和＋1之間 ，當r值爲正數時，稱爲「正相關」，r值爲負數時，則稱爲「負相關」，r＝0時；則代表「零相關」，亦即沒有任何相關存在，r值愈大，代表兩組分組的相

關愈密切。

假若兩組分數之r值達0.8，表示兩者的每一分數間，均有密切關係存在，亦即第一次得分高者，第二次得分亦高，第一次得分低者，第二次得分也低。因此，相關係數（ r ）的大小，可以代表評量工具的信度。

在相關研究上，比較常用的是「積差相關係數」（ Pearson Product-moment correlation coefficient ），其計算的公式如下：

$$r=\frac{N\Sigma XY-\Sigma X\cdot\Sigma Y}{\sqrt{\left[N\Sigma X^2-(\Sigma X)^2\right]\left[N\Sigma Y^2-(\Sigma Y)^2\right]}}$$

N：學生數

$X\cdot Y$：各代表每位學生的兩個分數

爲了幫助讀者了解其計算方法，現舉一簡單計算實例來說明。

表二：某班學生生物科兩次月考分數的積差相關係數（ r ）之求法。

學生座號	姓名	第一次月考 X	第二次月考 Y	X^2	Y^2	XY
01	○○○	62	70	3844	4900	4340
02	○○○	78	81	6084	6561	6318
03	○○○	72	76	5184	5776	5472
04	○○○	96	92	9216	8464	8832
05	○○○	48	48	2304	2304	2304
06	○○○	76	70	5776	4900	5320
07	○○○	75	71	5625	5041	5325
08	○○○	90	88	8100	7744	7920
09	○○○	84	86	7056	7396	7224
10	○○○	56	60	3136	3600	3360
N=10	總和(Σ)	737	742	56325	56686	56415

$$\therefore r=\frac{10\times56415-737\times742}{\sqrt{\left[10\times56325-737^2\right]\left[10\times56686-742^2\right]}}=\frac{17296}{\sqrt{20081\times16296}}$$

$$=\frac{17296}{18089.8}=0.96$$

本例之r值爲0.96，表示生物科第一、二次月考之成績，有極明顯之正相關存在；也就是說，評量工具所測得的結果具有相當的一致性和可靠性，所以，信度也夠高。假若r值偏低，例如：r值爲0.2時，表示兩次評量所測得的結果不太一致，可靠性也較差，此時，評量工具便有詳加檢討的必

要了。

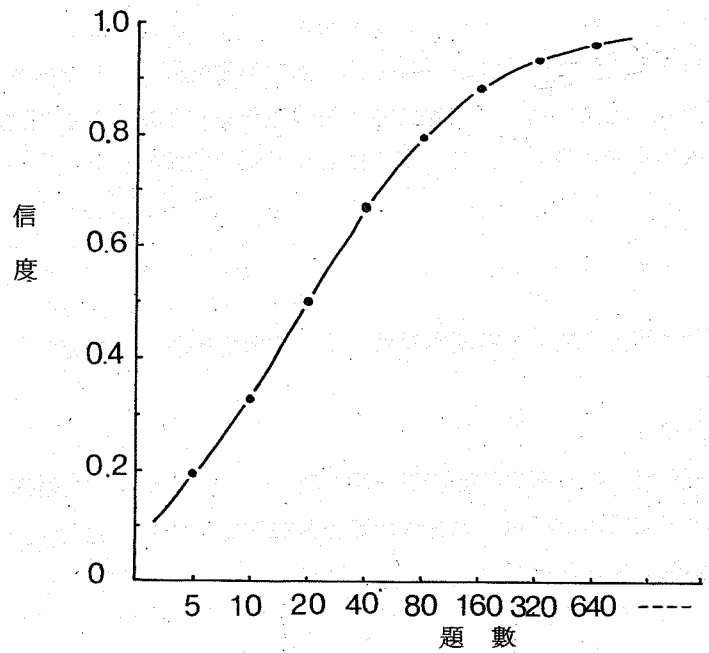
三、信度的意義及其改進之方

高的信度是優良的學習成就評量工具的特徵之一，其評估的資料來源，來自評量的結果而非評量工具本身，因此，可能因對象族群的不同而有異。所謂「信度」是指評量的結果（分數）與其擬測定的學習成就的一致性（Consistency）。由於學生的眞正學習成就頗不具體，我們只能以一些評量工具，經多次評估之後，給每位學生一個等第或得分，以資代表。因此，當我們擬評鑑某一評量工具所測得的結果（分數），可信度究有多大時，便往往藉觀察本次評量結果（分數）是否與其他類似目標的評量結果（即另一組分數）一致來評估；假若一致性頗高（即高分者仍得高分，低分者仍是低分），我們便認爲該評量結果（分數）應可測得學生的眞正學習成就。

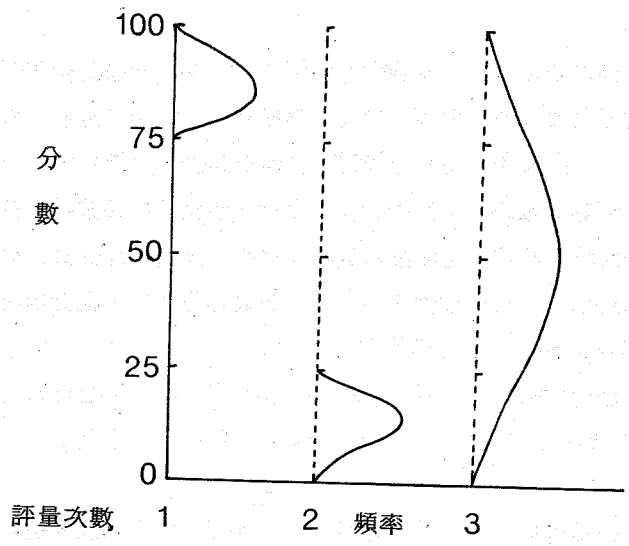
信度的評估可以讓命題者（教師）即時發掘評量工具的缺陷，加以改進，因此，它是改進命題和評量的另一重要指標。許多評量學者如Ebel.[10] Gronlund.[11] Popham[13] 都認爲，下列因素均會影響評量結果的信度，包括：

① 試題的多寡：通常題數愈多，信度也愈高；因爲，題數增加時，可以將學生以機遇（Chance）方式猜題，而僥倖答對的影響減低。但是，題數增加時，花在命題、實施評量、閱卷和統計等的時間也相對增加，因此，每次評量，應有多少試題才算適當，頗值斟酌，也無標準可言，教師應依個人的時間、能力、學科內容、進度以及學生的學習狀況來決定；圖四的「題數與信度關係表」或可提供些許參考資料。

② 分數的分散度：分數愈分散，信度愈高；因爲，分數分散時，表示每一得分的差異較大，於是，「測驗誤差」對學生名次（或得分）的影響便較小，由圖五的資料，便可略窺端倪。



圖四：試題數與信度的關係表



	難度太小	難度太大	適當難度
平均數(M)：	85	21	60
標準差(σ)：	3.9	4.2	12.4
信度(KR_{21})：	0.43	0.46	0.90

圖五：分數之分散度(Distribution)和難度(Difficulty)與信度之間的關係。

③ 試題的難度：就「常模參考評量」來說，試題太難或太容易均將導致信度降低(圖五)。

④ 試題的鑑別度(Discriminating power)：試題之鑑別度愈理想，信度愈大，已如前述。

⑤ 評量工具的客觀性(Objectivity)：當其他條件都相同時，通常客觀性評量工具所測得的結果，其信度要比主觀性評量工具所測得的結果爲大。但是請注意，筆者並無意給予讀者“主觀性評量工具不好，宜儘量少用”的印象，因爲兩者各有其長處和短處；假若爲了明確的達成我們的評量目標，有時犧牲一點信度是值得的。

四、信度係數的評估方法

目前至少有五種以上的方法來評估信度係數，由於篇幅關係，本文將僅就幾種常用的方法，加以說明。

1. 重測法(Test-retest method)：

一群(班)學生，以同一份試卷測驗兩次，兩次測驗相隔一段時間，於是每位學生各有兩個分數，然後求出兩組分數之間的相關係數，即爲信度係數。本方法實際上是在測定分數的「穩定性」(Stability)，因此，以之做爲信度係數，頗多爭議。

2. 對等法(Equivalent - forms method)：

又稱「並行法」(Parallel - forms method)，即根據相同的命題綱要，製作兩份在內容、難度

和方式上均儘可能類似的試卷，然後分別用這兩份試卷來測驗同一群（班）學生（可連續或相隔一段時間實施），每位學生亦各得兩個分數，再求出兩組分數之間的相關係數，即爲信度係數。此種信度旨在測定分數的「對等性」（Equivalence）或（和）「穩定性」。

3. 等分法（Split - half method）：

在實施測驗之後，將每一份試卷的奇數和偶數題分別計分，於是每一份試卷便可得到兩個分數。然後求出所有學生的奇數題分與偶數題分兩者之間的相關係數，以 r_{oe} 表示，再依照史皮曼—布朗公式（Spearman - Brown formula）計算信度係數。

$$r_{tt} = \frac{2\,r_{oe}}{1 + r_{oe}}$$

r_{tt}：整個測驗的信度。

假若依前述，r_{oe} 表示奇數組分數或偶數組分數的信度的話，顯然的，整個測驗的信度係數 r_{tt} 便要比 r_{oe} 大得多了，這個現象印證了增加試題數可提高信度的說法。

4. 庫李法（Kuder - Richardson method）：

庫－李二氏於 1937 年提出一些評估信度的公式，其中比較常用者有兩個公式，稱爲 KR_{20} 和 KR_{21}。其求法如下：

$$KR_{20} = \frac{K}{K-1}\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

$$KR_{21} = \frac{K}{K-1}\left(1 - \frac{M(K-M)}{K\sigma^2}\right)$$

K：試題數

P：答對某一試題的學生，所佔的比例

q：答錯某一試題的學生，所佔的比例（$q = 1 - p$）

σ：分數之標準差

M：分數之平均數

當評量工具的難度指數大體上均在 50% 左右時，用 KR_{21} 來計算信度比較簡便，不過當試題之難度不一，且變化很大時，使用 KR_{21} 往往會低估信度。庫一李法與等分法，就內容上來說，都在估定評量工具的「內在一致性」（Internal consistency），由於庫一李法基本上，假設所有試題都是均稱的（Homogenuous），因此，不適合用來求取「快速測驗」（Speeded tests）的信度，因爲在快速測驗中，有些學生無法做完所有題目，將導致 KR_{20} 或 KR_{21} 信度係數的混亂。

以上介紹的四種求取信度的方法之中，由於「等分法」與「庫一李法」均只需進行一次測驗，實施起來比較容易，因此，一般教師進行學習成就評量時，常用這兩種方法來評估信度。

五、測驗標準誤（Standard error of measurement）之估定

當我們以評量工具來評估學生的學習成就時，總是以某一特定的值來說明其學習成就；其實，假若我們能夠用同一評量工具，反覆測驗多次的話，便可發現就某一位學生來說，其每次的得分均不盡

相同；信度較差的工具，測得的諸分數之間，變異值（ Variations ）較大，信度較大的工具，所測得的諸分數間的變異值則較小。這種變異的估定值稱爲「測驗標準誤」；而多次測驗所得分數的平均數，則可代表「眞實分數」（ True scores ）。每次測驗所得分數與眞實分數之間的差，稱爲「測驗誤差」（ Error of measurement ），它們三者之間的關係可用下式來表示：

$$X = \bar{X} + e$$

X：某次評量之得分

$\bar{X}$：眞實分數

e：測驗誤差

所謂「測驗標準誤」，事實上，就是「測驗誤差」之標準差，具計算方法如下：

$$SE_m = \sigma \sqrt{1-r}$$

SE_m：測驗標準誤

σ：分數之標準差

r：信度係數

計算出SE_m可以讓教師明瞭分數的精確性，也提供學生學習成就的眞正指標，讓教師了解，學生的學習成就不可能用某一特定數字來代表，實際上，那個分數代表的是一個範圍，有人稱之爲「分數帶」（ Band of scores ），至於這個「分數帶」究竟有多寬，那就得看SE_m的大小了，因此，它也是分析和評鑑評量工具的重要指標之一。

六、效度的評估

就學習成就評量的範圍來說，所謂「效度」是指評量工具是否精確的測出了該工具應該測定的成就而言；通常效度是很難以具體的數值來定量的，因此也沒有一套計算效度的公式可資運用。事實上，效度的種類至爲繁多，目的不同，評估的方式也異，在此，不庸贅述。現僅就評鑑和改進「直接效度」（ Direct validity ）需要注意的地方，稍加說明；其實這些都與評量工具的內容有關，在本刊前幾期中，已有多人（毛松霖[1]，[2] 楊榮祥[4]、鄭湧涇[6]）提及，假若命題時能注意這些要點，效度便可無慮。

① 答題說明應明確、詳盡，使學生不致有任何混淆或誤會。

② 字彙和句子不可太艱澀深奧，以致學生因無法看懂而無從下筆答題。

③ 語意宜清楚明確，不可稍有模糊。

④ 試題數不可太少。

⑤ 不可無意的於題目上，提供任何與答題有關的線索，導致學生猜題。

⑥ 試題難度應適當。

⑦ 試題應儘可能測定重要的目標，概念，思考歷程，知識的理解、分析和綜合，以及較複雜的具體成就，而不宜故意佈設陷阱，測定一些瑣碎、零星的記憶性知識。

⑧ 試題的排列次序，宜先易後難，以免學生花太多時間於較難題目上，以致時間不足而放棄了一些容易的試題。

⑨ 信度是效度的必需條件，因此，效度高的評量，首先得信度要夠。

⑩ 某一學習成就評量工具若重複使用多次，亦會逐漸降低效度。

七、其他應注意的事項

除上述諸因素之外，尚有其他因素也會影響評量工具的品質，分述如下：

1. 內容的均稱性（ Balance ）：

命題的份量，是否涵蓋了所有擬評量的範圍？有無偏頗或無意的特別強調了某些內容（章節）？假若在命題時，能夠依照「命題綱要」（ Specification ）或「評量計畫」（ Test plan ）或「評量藍本」（ Test blueprint ）來製作試題[6]，則在評量的內容上，便可取得應有的平衡。

2. 內容的切合性（ Relevance ）：

試題擬測定的成就，是否爲學生必須具備而且可以引伸運用的知識？假若我們將試題依其構思答案的過程加以分類的話，約可分成「資料性」（ Information ）和「應用性」（ Application ）兩大類試題，在一評量中，這兩類試題應維持一“合理”的比例，而不可太偏重在「資料性」試題。因爲教學的極致是要學生能“轉移”（ Transfer ）其既得學習於日常生活之中，因此，在實施評量時，應適當的增加「應用性」的試題，以切合教學目標。

3. 內容的客觀性（ Objectivity ）

命題時，通常都難免於命題者主觀意識的影響，因此，有時題意模糊或答案並不明確，而命題者並不自知。假若在學校中，於實施測驗之後，同科目的教師間，會因某些試題的“標準答案”究竟是那一個而爭論的話，那就是內容不夠客觀所致。因此，於命題後要求幾位學科專家，例如：學校裡教同一科目的同事，“試考”看看，剔除或改良那些稍有爭議的試題，便可大爲改進評量之客觀性。就評量結果來說，也比較容易達到公正、公平的目標。

肆、結　　論

學習成就評量本質上是一項專業性的工程，其“施工”的方式應根據“成果”不斷的加以評鑑，方能有所改進而使“施工”的技術日臻完美。評鑑評量的工具則又是一項更專業化的過程，以上所陳述者，不過擇其重要而容易辦到者，略加介紹闡釋而已；這些專業化的技巧，並非學術研究者的專利，其實，嚴格說來，它應該是做爲一位各級學校的科學教師必須具備的「能力」（ Competency ）之一，因爲，我國各級學校的學習成就評量和專職性評量（如：聯考），已經不能再“盲人瞎馬”的繼續下去，也該是由岐途勒馬，重回正軌的時候了。因此，筆者在此要大聲呼籲，讓我們秉持職業良知和專業技能，共同努力來引導我國的科學學習成就評量走向正軌吧！

我們不要再用「考試領導教學」或「聯考領導考試」的嘆息，來嘲笑自己的無能了。至少，我們可以在每年、每次的聯考過後，仔細的“評鑑”一下聯考的工具是否恰當，以“逼迫”聯考的命題，不偏離正途。也應該隨時評鑑自己的評量方式和工具，以改進評量的品質，使其不再“領導”教學，誤入岐途。家長、學校和教師自己；也不會再爲70％的學生，自然科學科目成績不及格而嘆息、自責；

大部份學生也不會再因自然科學科目得分太低[5]而逐漸喪失其學習興趣了。 □

參考文獻：

1. 毛松霖，物理教學評量與命題設計(一)。 科學教育月刊12：35－40，民國66年。
2. 毛松霖，物理教學評量與命題設計(二)。 科學教育月刊13：42－47，28，民國66年。
3. 毛松霖，物理教學評量與命題設計(三)。 科學教育月刊14：41－49，民國66年。。
4. 楊榮祥，學習行爲目標分類與生物科測驗設計數例。 科學教育月刊10：17－23，民國66年。
5. 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。
6. 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。
7. 鄭湧涇，科學學習成就評量—— I、命題與閱卷。 科學教育月刊44：2－10，民國70年。
8. Cook, D.L., A Note on Relevance Categories and Item Statistics. Educational and Psychological Measurement 20(2): 321－331, 1960.
9. Cronbach, L.T., and W.G. Warrington, Efficiency of Multiple-choice Tests as a Function of Spread of Item Difficulties. Psychometrika 17: 127－147, 1952.
10. Ebel, R.L., Essentials of Educational Measurement. 2nd ed. 1972, Prentice-Hall, Inc., Englewood Cliffs, N.J.
11. Gronlund, N.E., Measurement and Evaluation in Teaching. 3rd ed. 1976, Macmillian Publishing Co. Inc., New York.
12. Kuder, G.F., and M.W. Richardson, The Theory of the Estimation of Test Reliability. Psychometrika 2: 151－160, 1937.
13. Popham, W.J., Criterion-Referenced Measurement. 1978, Prentice-Hall, Inc., Englewood Cliffs, N.J.
14. Trowbridge, L.W., Personal Final Report to National Science Council, Republic of China, p. 125－142, 1977.

編輯部小啓

「國中數學及自然科學課程目標及教材內容意見調查問卷之研究」調查報告尚存物理、地球科學兩篇，茲後仍將陸續刊出，以饗讀者。