貳、評量的基本觀念與方法

評量一詞，實包含「評鑑」(evaluation)與「測量」(measurement)兩個概念，都是了解事物特徵的程序。依據測量學家史蒂芬斯(S. S. Stevens)的解釋，測量乃是「依照規則賦予事物特徵特定數字的程序」。[①]據此可知，測量乃是將事物的特徵加以「量化」的過程，也是「量化研究」(quantitative research)不可缺少的程序。換言之，從事量的研究時，必須使用適當的工具，針對所欲探討的事物特徵加以測量。至於評鑑的程序，雖可包含研究者主觀的判斷與評價，但在實徵研究中仍以量化的特徵為基礎，故測量的相關概念與方法大體也可應用於評鑑。基於此種認識，本文說明評量的概念、方法、與應用時，並未將評鑑與測量作嚴格的區分。再者，基於本文旨在提示讀者一個簡潔的輪廓，俾能掌握評量的基本觀念，以增進選擇並運用各種評量工具的適切性，故本文有關評量的概念與方法、說明、力求簡明扼要，因而難免有過份簡化或掛一漏萬之疏失，尚祈讀者參閱文末所附參考書目之相關書籍與文獻。以下先提示評量的基本概念，然後再說明評量的主要方法及其應用。

一、評量的基本概念

欲了解評量的性質、並適當使用評量工具進行研究時，必須先了解「尺度」(scale)、「信度」(reliability)、效度(validity)、以及「常模」(norm)四個基本概念。以下分述之。

㈠尺度(scale)

任何測量必須有測量的準則和依據，例如量桌子長度時，或用掌距來量、或用台尺來量、或用米達尺來量，總要有個依據才行。這個作為測量的準則或依據，也就是測量的「尺度」。在測量時如採用不同的尺度，則對事物或變項特徵的描述和說明就提供不同的信息。以上述量桌子長度之例而言，採用不同尺度測量的結果，可能是五個掌握，也可能是三尺三寸、或一百公分。雖然桌子的長度不變，但因測量尺度不同，故對桌子長度的說明也不一樣。

尺度的種類大致有四種：一是「名義尺度」(nominal scale)、二是「順序尺度」(ordinal scale)、三是等距尺度(interval scale)、四是比例尺度(ratio scale)。[②]這四種尺度具有不同的特徵，也有不同的功用。研究者必須了解這四種尺度的性質，才能選擇適當的尺度，用來編製測量工具。以下分述四種尺度的性質：

1. 名義尺度(nominal scale)

名義尺度係依事物的特徵或屬性之不同，賦予不同名稱，作為一種標記，進而可將特徵或屬性相同者歸為類別，故也稱為「類別尺度」(categorical scale)。換言之，名義尺度的主要功用是在區分類別，給每一個類別適當名稱，藉以辨識。譬如：人之「性

別」可區分為「男性」與「女性」；婚姻狀況可區分為「已婚」與「未婚」；家長職業可區分為「專業」、「半專業」、以及「非專業」三類；而「休閒嗜好」可區分為「戶外休閒」與「室內休閒」，或區分為「益智性活動」、「運動性活動」、以及「娛樂性活動」三類等，都是應用名義尺度來分類。因此，應用名義尺度測量或描述事物的特徵時，就要設法將該事物依其特徵加以分類，並標示類別的名稱，然後給它一個代碼(code)。

2.順序尺度(ordinal scale)

順序尺度是將事物依其特徵或屬性的大小、或多少的程度，排成順序或等級。譬如，將十個參加演講比賽的學生依其成績高低自1排至10，這就是順序尺度的應用。換個方式來看，如果以順序尺度測量四年乙班五十名學生的成績，請問小明的成績如何？答案可能是「小明是第五名」，而不是「小明的成績是80分」。同樣的方式，我們可以採用順序尺度來測量一個人的價值觀念，了解忠、孝、仁、愛、信、義、和、平八德，何者最重要？何者次之？何者殿後？換言之，順序尺度可以排列八德重要性的等級，顯示一個人的價值觀念。由此看來，順序尺度的主要功用是排列等級，比較順序。在等級或順序的排列中，可以比較個體之間的地位，可說明「大於」或「小於」的關係和差異，但個體之間的差異並無相同的單位。故全班第一名的成績與第二名成績的差異，未必等於第二名成績與第三名成績的差異。這個特徵要特別留意。

3.等距尺度(interval scale)

等距尺度是一組具有連續性、單位又相等的數值。如果應用等距尺度來測量變項，乃是依其特徵或屬性之不同賦予不同的數值，使這些數值不僅顯示大小的順序，而且數值之間具有相等的距離。例如，以等距尺度測量學生的國語科成績，乃在0分至100分的範圍內，依學生的學習表現給予一定分數。從學生的分數既可看出學生成績高低的順序，也可以了解學生之間成績的差距。在教育研究中，許多變項都是採用等距尺度來測量的，如智力，以智商表示；性向，以性向測驗的分數表示；學業成就，以成就測驗的分數表示，都是等距尺度的應用。

由上述的說明可知，等距尺度的主要特徵在於：(1)分數、(2)連續性、與(3)等距；而其主要功用則在於採用連續且等距的分數說明變項特徵或屬性的差異情形。但是，等距尺度所採用的分數，雖然可以有「0」，卻非「真正的零點」(true zero)。試想學生的國語科測驗成績，範圍是0分至100分。假如大中的成績是0分，難道表示大中的國語科能力是空白嗎？當然不是。分數上的「0」是人為的零點，是研究者決定的一個點。因此，大中考了0分，只能解釋說，大中在這次測驗中，全部題目都答錯了，而不能說大中的國語能力是0。同樣的道理，溫度計上的刻度，也是一種等距尺度，但溫度計上的零度，也不是真正的零點。因此，攝氏零度並非表示沒有溫度。我們使用的「年代」，也是等距尺度，也沒有真正的零點。紀元元年只是以耶穌基督的誕生為起算點，並不意味著人類的歷史從那一年開始。由於等距尺度沒有真正零點，所以在比較差異時，只能就分數作加減運算，而不能以乘除倍比的關係來說明。昨天的氣溫是15°C，今天的氣溫

是30°C，我們只能說，今天比昨天熱了15°C，而不能說今天的熱度是昨天的二倍。同理，國語科測驗小明考了90分，小華只有45分，那並不意味著小明的國語能力是小華的兩倍。這一個解釋上的限制要特別予以留意。

4.比例尺度(ratio scale)

比例尺度具有等距尺度的全部特徵，而且有「真正零點」。因此比例尺度的數值之間有相等的比例(ratio)，不僅可以加減，也可以作乘除的運算。例如。人的身高，可以採用比例尺度來測量，以0代表沒有高度，0以上的不同數值代表實際高度，而身高200公分即為身高100公分的兩倍。體重的測量也是如此。又如年齡也可以採用比例尺度測量，因為零歲是真正的零點。據此可知，比例尺度所提供的信息最多，功用最大，但在實際測量的應用上卻不多見。一般說來，物理特徵的測量（如重量，長度等）比較可能採用比例尺度，但心理特徵的測量大體以等距尺度為主，因為人類的心理特質很難找到真正零點。

以上分別說明四種尺度的特徵與功用，也有一些舉例，綜合起來，在實際應用尺度測量事物特徵時，須特別注意下列幾點：(1)四種尺度的層次不同，名義尺度最低，然後是順序尺度、等距尺度、比例尺度，依次遞升。(2)尺度的層次不同，所提供的信息也不同；高層次的尺度提供較多信息；凡較低層次之尺度能提供的信息，其較高一層尺度均能提供。(3)尺度的應用可以轉換，但限於由較高層次尺度轉變為較低層次尺度使用。譬如學科成績，本可使用等距尺度測量，而以分數表示；但亦可改以順序尺度測量，依成績排成名次；也可以採取名義尺度測量，依成績將學生區分為及格與不及格兩類。(4)尺度的使用應依照變項性質與研究目的來選擇。有些變項只限於使用某種尺度始能測量，如性別只能用名義尺度，那就必須依據變項性質選用尺度；如變項可用多種尺度測量者，如前舉學科成績之例，則以研究目的來決定。(5)尺度的應用與資料的統計分析有關；以不同尺度測量的變項，各有其不同的、適用的統計方法。總括言之，研究者在應用各種尺度時，必須考慮變項的性質、測量的目的、以及未來統計分析的方法。

㈡效度(validity)

效度是指根據測量結果推論變項特徵的適切性(appropriateness)。譬如說，我們想了解學生的學習動機，因此採用一個學習動機量表加以測量，每一個學生都得到一個「分數」，我們要根據這個分數來推論學生的學習動機。在此種情況下，我們要先確定，依據這個分數來推論學生的學習動機是否適切？是否有意義？是否有用？換言之，這個測量所得的分數能否真正解釋學習動機？如果答案是肯定的，那麼依據這個測量結果所作的推論就有效；如果答案是否定的，那麼推論就無效。不過，效度並非「全有」或「全無」的概念，而是程度高低之分。由此可知，效度乃是測量的必要條件，缺乏效度則推論與解釋都不適切，這個測量就沒有意義，也沒有用處，因為它不能解釋真正想解釋的特徵或屬性。

習慣上，測量學者常將「效度」分為三類，即：內容效度(content validity)、構

念效度(construct validity)、與效標關聯效度(criterion related validity)。但美國心理學會(American Psychological Association)於一九八五年修訂出版的「教育與心理測驗標準」(Standards for Educational and Psychological Testing)中，一反傳統的觀點，而以效度證據(evidences of validity)來代替效度的分類。[3]換言之，如欲確定測量工具的效度，那就必須蒐集足以說明效度的證據（即資料）。因此，習慣上稱為內容效度者，宜改稱為「內容關聯的效度證據」(content-related evidence of validity)，而構念效度及效標關聯效度兩者，亦宜改稱為「構念關聯的效度證據」與「效標關聯的效度證據」。以下即以這三種效度證據的性質與蒐集方法加以說明：

1. 內容關聯的效度證據

這一類效度證據係從測量工具的內容來檢查，看看是否符合測量目標所預期的內容。譬如說，一個學科成就測驗的預期內容是一學期的教材，但測驗題目所涵蓋的範圍卻只有第一課至第五課，其餘十五課的教材在測驗題目中都付之闕如。這樣的測驗顯然缺乏效度，因為測量的內容未盡周延完整，故測量的結果無法有效推論全學期的學習成就。

2. 構念關聯的效度證據

一般而言，構念(construct)是一種假設性的實體，是學者或研究者基於學術的目的，為說明一個假設存在的屬性或特徵，而精心創造或借用的名稱。譬如「智力」、「焦慮」、「動機」等都是心理學的構念；而「地心引力」則是物理學的構念。因此，構念可視同一般概念，但卻是抽象的、假設性的存在，無法直接觀察或測量，而必須藉間接的指標來推論。譬如上述「地心引力」是由蘋果（或其他物體）落地的現象推論其存在；而「智力」則藉個體的學習行為與表現來推論其存在。

通常，學者提出一個「構念」時，都有一套相關的理論或原理來支持，因此，如果我們研究的變項或特徵是一個構念，則在應用測量時，須將測量的內涵與結果，與此一構念的相關理論及其衍生的現象相比較，藉以推論測量結果能否適切有效的解釋此一構念的性質與特徵。譬如，針對「智力」這個構念進行測量時，因智力理論提示智力隨年齡而發展的原則，故智力測量的結果應顯示測量分數隨年齡遞增的現象，始符合智力的理論。唯其如此，才能確定此一智力測驗適切有效。換言之，欲從構念的分析來考驗測量工具的效度時，須以相關的理論為分析檢驗的架構和依據。

3. 效標關聯的效度證據

此類效度證據之蒐集係以其他測量為標準（習稱效標），將測量結果與效標作一比較，若彼此相關程度愈大，顯示效度愈高，反之亦反。如果這種比較係以受試者受測一段期間後的實際行為表現為效標，則稱為「預測性效度證據」(predictive evidence of validity)；如果以受測時的其他資料（含測驗）為效標，則稱為「同時性效度證據」(concurrent evidence of validity)。舉例來說，對一群兒童實施創造力測驗，測定每名兒童創造力的高低，經過一段期間（也許是幾個月，甚至是好幾年），看看這些受試的兒童有何具體的創造性行為表現。如果兒童的測驗分數與創造行為表現有密切相關，顯

示測驗結果足以預測創造行為，因此效度高；若測驗分數與創造行為毫不相關或相關很小，那麼效度就低。這樣的資料就是「預測性效度證據」。至於「同時性效度證據」，通常係以受試兒童的其他測驗分數、學業成績、教師評定等當前資料為效標。譬如編製一份學業成就測驗後，將測驗結果與受試者當時之學業成績來比較，求得兩者的相關，即為「同時性效度證據」。若相關程度高，當然成就測驗結果的推論就適切有效。

以上三類效度證據並無優劣好壞之分，在研究過程中應用測量工具時，須視研究問題的性質與研究目的而決定採用何種效度證據。研究者在選擇現成的測量工具時，必須檢視該種測量工具的效度，如係自行編製測量工具，則須考驗並提示該項測量工具的效度。一個未曾提示或說明效度的測量工具，即難確定其測量結果的適切性，因此不可冒然使用。

㈢信度(reliability)

依通俗的說法，信度是指測驗結果的可靠性。如果照美國教育與心理測驗標準之定義，信度指的是測驗分數未受測量誤差(errors of measurement)影響的程度。④這兩種解釋並不衝突，蓋測量誤差愈小，測量結果愈可靠。換言之，如果測量的結果能反應受試者真實的特徵，而不因其他因素（如測驗情境、受試者心理情緒狀態、測驗題目的性質等）而影響其測驗分數，那麼這個測驗所測量的結果是可靠的。

信度也是測量的基本要素之一，缺乏信度的測量就不具意義，也不能使用。因此，研究者在使用測量工具時，一定要知道測量的信度。然而如何估量測量的可靠程度呢？大體有兩個途徑可循，一是估量測驗結果的穩定性(stability)，一是估量測驗題目的內部一致性(internal consistency)。兩種途徑各有數種方法可以使用，以下簡述之：

1. 穩定性之估量

測量結果的穩定性，係以同一測量工具實施二次測量結果的相關程度（即相關係數）來估量，相關程度愈高，表示測量結果愈穩定，亦即信度愈高；反之則反。如欲了解測量結果的穩定性，可採用「再測法」(test-retest method)與「複本法」(alternate forms method)，其程序如下：

⑴再測法

先選擇適當對象定期實施測量，經一段期間（通常是二至三週、或三至四週）後，以同一測量工具；對先前受測之對象實施第二次測量，求得兩次測量結果（通常是分數），計算其相關程度，即可說明此一測量工具的再測信度（test-retest reliability)。

⑵複本法

再測法費時較長，且兩次測量結果的同異易受記憶與成長的影響，故有時不易估量測量結果的穩定性。在此情形下，即可採用複本法，先編製乙份測量工具，稱為正本，然後另行編製乙份性質、內容、難度均相同、但文字不同的題目，作為複本，並以正本與複本針對相同對象實施測量，求得兩份測量結果，計算其相關程度，即可據以估量測量結果的穩定性，了解測量工具的信度。

2.內部一致性之估量

以再測法估量測量工具的信度，固有其可能的缺失與限制，但如採用複本法，因必須編製測量工具的複本，故也有其困難與不便。因此，如以一次測量結果來檢視測量題目的內部一致性，並據以估量測量結果的可靠程度，即可免除再測法與複本法之缺失與困難。以下三種方法可以估量測量題目的內部一致性：

(1)折半法(split-half method)

針對一群受試者實施測量之後，將題目平均分為兩組（通常以題號為準，單號題一組，雙號題一組），分別計算受試者在各組的得分，並進一步求得這兩組分數的相關程度，然後依據「斯－布公式」(Spearman - Brown Formula)計算，所得結果即為信度係數。斯布公式如下：

$$\text{測驗信度}=\frac{2\times(\text{折半測驗分數之相關})}{1+(\text{折半測驗分數之相關})}$$

(2)庫李公式(Kuder－Richardson Formula)

如果顧及測驗的完整性，或因折半計分有所不便，而不採用折半法時，則可改用庫李公式計算測量題目的內部一致性，作為信度的指標。庫李公式有兩種，分別是K-R 20與K-R 21，前一個公式計算的結果較精確，後一個公式的計算程序則較為簡便。公式如下

K－R21　信度係數$=\frac{K}{K-1}\left(1-\frac{M(K-M)}{KS^2}\right)$

K：測驗題數

M：測驗分數之平均數

S：測驗分數之標準差

K－R20　信度係數$=\frac{K}{K-1}\left(1-\frac{\Sigma pq}{S^2}\right)$

K：測驗題數

p：每一題目答對人數佔總人數之比率

q：每一題目答錯人數佔總人數之比率，即1－p

Σ：表示總和

S：測驗分數的標準差

(3)α係數

庫李公式適用於答題有對錯性質之測驗，但一般態度或意見量表均無對或錯的答案，故不適用。在此情況下，可採用史丹福大學(Stanford University)柯隆巴克(Lee J. Cronbach)教授所發展的α係數，依一定公式估量測驗的內部一致性，作為信度的指標。其公式如下：

$$\alpha=\frac{K}{K-1}\left(1-\Sigma\frac{Si^2}{Sx^2}\right)$$

K：測驗題數

Sx：測驗分數之標準差

Si：個別題目分數之標準差

綜合以上有關信度的種類及其計算方法之說明可知，一種測量工具的可靠性（即信度）可以藉不同的方法來了解，也都可以採用一個「係數」來表示其程度的高低。當然，不同性質的測量工具所要求的信度可能不一樣，標準化的成就測驗比普通智力測驗要求較高的信度，智力測驗又比人格測驗要求較高的信度，但一般而言，一個測量工具的信度至少應在.70以上，始稱得上可靠。須知信度是效度的必要條件，信度太低的測量工具，就不可能具有適當的效度。因此研究者在編製或選用測量工具時，一定要愼重考慮該種測量工具的信度。

㈣常模(norm)

一般說來，測量的目的有二，一是了解個體的特徵，二是探討群體的趨勢，而個體特徵的了解常須安置在群體趨勢中來比較始具有意義。因此測量的結果必須提示一個說明群體內差異情形的分數架構，作為解釋個別分數的標準與依據。這個群體的分數架構就是俗稱的「常模」(norm)。在教育研究相關的測量中，常用的常模有年齡常模。如依據年齡常模，即可知道各個年齡組受試者在某一種測量結果的平均分數，進而可以了解某一個受試者的測驗分數在群體中的位置。例如一個語文能力測驗的年齡常模，八歲組的平均分數是60，九歲組的平均分數是65分，而十歲組的平均分數是70分。那麼一個八歲兒童在該項測驗得了66分，即可依據上述常模，推斷他的語文能力不僅高於八歲兒童的平均程度，而且已具有九歲組兒童的平均程度。

在上述年齡常模中，也可以採用各種「標準分數」(standard scores)來解釋個別分數在其年齡組所居的地位。較常用的標準分數有：百分等級(percentile rank)、Z分數、與T分數。百分等級係指個體在特定群體中，分數在其下的人數百分比率。如以前段所舉之例而言，該名八歲兒童的語文能力測驗得分是66分，若計算其百分等級為80，意即在八歲組兒童中，有80%的兒童語文能力測驗分數不及得66分的該名兒童。換言之，百分等級愈高，表示在群體中的地位愈高。至於Z分數，實即以標準差為單位的相對分數。當Z＝0時，表示該個體的分數恰在群體的平均分數上；若Z＞1，表示在平均數之上；若Z＜1（亦即負號），表示低於平均數。概括而言，如個別分數轉換為Z分數、且Z＝1時，意即該個別分數大約勝過84%的人數；Z＝2時，大約勝過97%的人數；Z＝－1時，大約勝過15%的人數；Z＝－2時，則僅勝過2%的人數。通常資賦優異學生的智商，Z分數都大於2。至於T分數，則由Z分數轉換而來，若T＝50，表示恰好在平均數位置上，T＞50表示高於平均數，T＜50表示低於平均數。

除年齡常模外，教育測量也常使用年級常模，藉以了解各年級的平均程度，並解釋個別學生在該年級中的相對地位。但無論採用年級常模或年齡常模，通常都按地區、性別分別列出。換言之，男女生各有其年齡或年級常模，而城鄉不同地區也各有其常模。

以上四節分別說明尺度、效度、信度、與常模四個概念及其相關的方法與程序。這四個概念都直接關聯到測量工具的編製與使用。尺度是編製測量題目的依據，效度與信度是保證測量結果之可靠性與適切性的指標，而常模則是解釋測量結果的架構，四者缺一不可。因此在教育研究中使用評量工具時，必須確實了解這四個概念的意義、性質、以及應用的方法。

二、評量的主要方法

從研究的角度來看，評量工具的應用，旨在蒐集資料，藉以了解事物或變項的特徵。在教育研究中，最常使用的評量工具有問卷(questionnaire)、量表(scales)[5]、以及測驗(test)。換言之，研究者常採用問卷、量表、與測驗來評量事物或變項的特徵。因此，從事教育研究工作時，必須熟悉問卷、量表、與測驗的編製方法與使用要領。以下分節說明，供讀者參考。

㈠問卷的編製與應用

「問卷」是法文questionnaire一字的中譯名稱，原意是「一種為了統計或調查用的問題表格」，若直譯的話，可譯成「問題表格」。目前大家習慣用「問卷」一詞，可說是相當達意的譯法。在應用調查法(survey research)從事教育研究時，問卷是蒐集實徵資料的主要工具，因此研究者必須熟悉問卷的編製與實施方法。

乙份問卷是由許多題目(items)組成的，每一個題大都包含「問題」(questions)與「答題」(answer)兩部分，因此在應用問卷從事調查研究時，必須先會編寫題目。以下先就「問題」的型式與「答題」的型式略加說明，然後再分別提示問卷編製及實施的程序與要領。

1.問題的型式[6]

問題是一個題目的骨幹，故也稱為「題幹」(Stem)，一般而言，問卷所採用的問題，因其性質或內容不同，大致可作幾種分類，茲舉例如下：

(1)直接問題與間接問題：

①您對學校功課感到興趣嗎？（直接）

②您覺得學校的活動中什麼事令您愉快？（間接）

③您覺得學校功課對未來生活有益嗎？（間接）

(2)特定問題與普通問題：

①您喜歡爬山嗎？（特定）

②您喜歡從事休閑活動嗎？（普通）

(3)事實問題與意見問題：

①您家裡訂了幾份報紙？（事實）

②您認為一個家庭應該訂幾份報紙？（意見）

③您平時在家陪伴子女做功課嗎？（事實）

④您認為家長有必要在家陪伴子女做功課嗎？（意見）

(4)行為問題、知識問題，與態度問題：

①您自本學期以來打過電動玩具嗎？（行為）

②您知道本校所屬社區有幾家電動玩具店嗎？（知識）

③您贊成在校園裡放置電動玩具嗎？（態度）

(5)威脅性問題與非威脅性問題：

①您這學期來考試作弊過嗎？（行為、威脅性）

②您知道成人的正常血壓是多少？（知識、威脅性）

③您對婚前的性行為有何看法？（態度、威脅性）

上述三個例句中，第①個問題涉及道德規範；第②個問題是現代人的基本知識，不知道就不好意思；第③個問題也涉及社會價值與期望。當答題者回答這三個問題，心中都可能有所顧忌。因此這幾個問題都是威脅性的問題，在編寫時要小心處理，設法在文字及形式上避免引起答題者的心理防衛。至於非威脅性問題係指與道德規範、社會價值、及個人隱私無關的問題，不致引起填答者的心理防衛。

(6)獨立問題與關聯問題：

①您喜歡戶外運動嗎？

②您喜歡那一種戶外運動？（關聯問題）

從上述兩個問題可以看出，第②個問題是依附在第①個問題之下，只有在第①個問題回答「喜歡」時，才有機會回答第②個問題。所以這是一個關聯性問題。

(7)疑問句與敘述句：

①您認為中學男女分校應予取消嗎？（疑問句）

□應予取消　□不應取消　□無意見

②中學應取消男女分校：（敘述句）

□贊成□　不贊成　□無意見

2.答題的型式

一般常用的答題型式可以大分為兩類：一是「開放式答題」(open-ended response)，一是「封閉式答題」(closed response)。所謂開放式答題，就是不限定答題的方法，允許填答者在問題有關的範圍內，用各種方式自由回答；而封閉式答題則不僅限定答題範圍，也限制答題的方法。請看下述幾個例子：

①您為什麼要繼續升學？

②您為什麼要繼續升學？（可以複選）

□(1)為了充實自己

□(2)為了習得一技之長

□(3)父母的期望

□(4)其他

③您父親的職業是：_____

④您父親的職業是：_____

□(1)軍□(2)公□(3)教□(4)農□(5)工□(6)商□(7)其他

由上述的例子可以看出，例①是完全無結構的、開放式答題，答題者可以從各個角度充分說明升學的理由。例③也算是開放式答題，但範圍比較明確，填答者只要填寫幾個字即可，故稱為「填充式」(fill-in)答題。至於例②與例④則都是封閉式答題，填答者只要依據自己情形，在適當的空格裡勾選即可。

一般說來，開放式答題型式採用「問答」或「填充」；至於閉鎖式的答題型式則以「是非」、「選擇」為主，包括「檢核法」(checklist)、「類別法」(categories)、「評定法」(rating)以及「等級法」(ranking)。以下分別舉例說明封閉式答題的基本型式：

(1)檢核法：通常用在初步了解有關行為與知識的事實。

您平時做什麼戶外運動？（可以複選）

□①爬山	□⑤打籃球	□⑨打棒球
□②遊泳	□⑥打網球	□⑩慢跑
□③騎馬	□⑦打羽毛球	□⑪其他
□④打排球	□⑧騎單車	

(2)類別法：通常用在行為或知識已有明確類別時；但若態度或意見可以明確分類時，亦可採用。

您目前就讀的學校是：

□①公立學校　　□②私立學校

您贊成中學男女合班嗎？

□①贊成　　□②無意見　　□③不贊成

(3)評定法：通常用來評量態度的程度。

您認為下列因素對一個人事業成功的重要性如何？（請圈選適當的號碼）

重要性	低				高
教育	1	2	3	4	5
智力	1	2	3	4	5
人格	1	2	3	4	5
家世	1	2	3	4	5
運氣	1	2	3	4	5

(4)等級法：通常用來了解態度的相對程度。

請將下列因素依其對一個人事業成功的重要性，依序排列等級（等級愈低，表示重要性愈低；等級愈高，表示重要性愈高） ___教育　　___家世 ___智力　　___運氣 ___人格

3.編製問卷的程序

一份良好的問卷，必須能幫助研究者蒐集到預期蒐集的資料，藉以達到研究目的。因此編製問卷的時候，必須確實掌握研究的目的，以研究目的為起點構思問卷的題目，其程序大體如下：

(1)確定所欲探討的變項：在編製問卷之初，須將研究中有關的變項明確列出，然後根據變項的性質，逐一編寫題目。舉例言之，若想採用問卷調查國民小學的「學習環境」時，須先列出研究變項，包括物質環境與心理環境兩類變項。前者有「校園美化程度」、「學習空間大小」、與「專科教室之設置」；後者則有「班級氣氛」、「師生關係」、與「同儕關係」。確定這六個變項之後，再逐一構思題目。

(2)選定問題的型式：問題的型式很多，但每一種型式均有利弊，故如何決定問題型式，難有定則。基本的要領是：依據變項的性質、並考慮統計分析的方法，選擇一種較能引發填答者真實反應的題型。

(3)選定答題的型式：相同的問題可以採用不同的答題型式，前已述及。如何選擇答題型式，也無固定法則可資遵循。通常，可以從幾方面來考慮：其一，填答者是否容易填答？其二，要花費多少時間填答？其三，填答結果如何統計分析？所以，無論選定那一種答題型式，一定是填答者會填、也願意填，而且有適當的分析方法。

(4)撰寫問卷題目：在選定問題型式與答題型式之後，就要一題一題地撰寫題目。這是編製問卷最難的部分，也是最重要的部分。撰寫題目時，必須以變項為依據，並兼顧文字的適當運用，同時考慮題目的順序。

(5)撰寫指導語：在一份問卷裡，除了問卷的標題、編製者資料，以及題目之外，還要有一段指導語，置於題目之前。指導語一方面說明問卷的目的，一方面說明填答的方法，使填答者因了解這份問卷的意義、價值、與重要性，而能熱心據實填答，也使填答者因了解填答方法而能以適當的形式填答，增加問卷資料的效度。若問卷中有特殊題型及填答方法者，則在每個題目之前說明填答方法。

(6)預試與修正：問卷編擬完成後，一定要經過預試，以確定題目型式、內容，以及文字的適切性。預試的對象必須在研究的母群體中選取，也要略具代表性，人數不定，五至十人亦可。經過預試並加以修正後的問卷才能使用。

(7)建立效度與信度：問卷的信度大都以「重測信度」(test-retest reliability)為主。在問卷經過預試修正後，在研究的母群體中選取適當人數填答，大約二週後，同樣的

人再填答一次，比較兩次填答結果的一致程度而決定信度。若爭取時間，在第一次填答之後，改以訪問方式進行，然後比較填答與訪問結果的一致性而定信度的高低。至於效度，可請專家分析每個題目的適切性，建立「專家效度」。若能蒐集先前相關研究的調查資料作為效標，也可建立效標關聯的效度證據。

4.撰寫問卷題目的原則

在上述編製問卷程序中，撰寫研究題目可說是最難的部分，也是最重要的部分。撰寫問卷題目時，必須以研究的變項為基礎，並兼顧文字的簡潔流暢，同時也要考慮題目的順序，以利填答者填寫。以下幾個原則可供參考。

⑴要運用填答者能了解的文字、名詞、或概念；敘述要扼要、通順；避免使用學術性專門名詞。

⑵問題所傳達的信息要明確，不可語意含糊、目標曖昧；如涉及抽象概念，必須提示具體指標。

⑶避免「雙管」(double-barreled)的問題：原則上，一個題目只能問一件事、一種態度、或一個概念。如果在一個問題中包含並行的兩件事或兩個概念，就成為雙管的問題，填答者無從填答。例如，「您的教育抱負與職業抱負受誰的影響？」、「您贊成國民小學取消早自修與午休嗎？」都是雙管的問題，必須各拆成兩個問題來問。

⑷避免在問題中暗示或引導填答的方向。譬如，「爬山是一種很好的戶外活動，您喜歡爬山嗎？」這個問題的前半句就具有引導作用，容易影響填答者的反應，而造成偏差。

⑸對於敏感的、有威脅性的問題，要妥用文字。通常，使用較長的句子、或提供合理化的線索，有助於減低敏感性與威脅性。譬如，若開門見山就問「您有過自慰的行為嗎？」填答者可能有所顧忌而不敢據實回答：若改為「根據醫學報導，自慰並非病態行為，而且大部分身體健康的人都有過自慰的經驗，請問您有過自慰的行為嗎？」，不但句子加長，也有合理化的線索，可以減輕或紓解填答者的心理威脅而據實填答。

⑹在題目的順序方面，通常是容易回答的問題在先，複雜的問題在後；一般性的問題在先，特定的問題在後，呈漏斗狀；普通的問題在先，敏感的、涉及社會期望的問題在後；封閉式問題在先，開放式問題在後。同時，問題的排列要有邏輯順序。

⑺題目的長度、型式要有變化，藉以維持填答者的興趣；同時，要避免採用固定的答題型式，以免填答者受填答趨向的影響，而不仔細閱讀題目內容。

⑻問卷必須匿名填答；如需要填答者的背景資料，則置於題目的最後來問。

5.問卷的實施要領

問卷的實施有兩種方式，一是郵寄，一是當面實施。郵寄問卷可以簡省經費與人力，但無法控制填答情境，故回收率(return rate)低；當面實施雖然花費較大，但能掌握填答狀況，故可以有相當高的回收率。然而，當面實施只適合於團體施行，所以必須填答者有團體組織時才方便實施，譬如，以學校學生為對象時，若經校方同意，即可安排團體實施問卷。

郵寄問卷的最大困難是無法掌握回收率。一般的問卷調查，應有50％以上的回收率，資料的分析才有意義，若能得到70％以上的回收率，才算理想。不過，當樣本愈小時，要有較高的回收率才具有代表性。

問卷的回收率固然與問卷本身的內容及形式有關，如內容過於複雜、填答不易，就會減低填答的動機，但若在實施過程中予以適當留意，也有助於提高回收率。首先，在郵寄問卷時，一定要附上書明地址的回郵信封，讓填答者在填答完畢之後方便寄回。其次，要做追蹤催促的工作。雖然問卷係採匿名填答，但研究者可事先在每一份問卷上編號，作為識別填答者的依據。通常在問卷郵寄二、三週後，須查明未填回者姓名資料，再補寄一份問卷，請求填答。若問卷之實施要求完全保密，故不但匿名，而且問卷上不能編號時，則在郵寄問卷二、三週後，再寄一份問卷給全部填答者，並附一封「提醒信」(reminder letter)，說明如已填妥寄回者可不必理會此信，若因事忙碌而未克填回者，則請費神填答後寄回。在第一次催促後二、三週，如覺回收率仍嫌不足，則可再一次以同樣方式催促。總之，在問卷實施過程中要盡一切努力提高回收率。若最後結果回收情況不甚理想時，在研究報告中要據實說明。

㈡量表的編製與應用

在教育研究中最常用的量表是「態度量表」。一般常用的態度量表有三種型式：一是塞斯通式量表(Thurstone scale)，二是李克特式量表(Likert scale)，三是語意區分法(Semantic differential technique)。以下分述各種型式的特徵：

1.塞斯通式量表(Thurstone scale)

塞斯通量表係心理學家塞斯通(L. L. Thurstone)所創，採用「相等出現間隔法」(method of equal-appearing interval)來編製態度量表，其步驟如下：⑦

⑴針對特定的態度對象（人、事、物），編寫大約200條有關的積極性及消極性的「敘述」。

⑵請專家（大約50人）判斷這200條「敘述」。判斷的方法是將這200條「敘述」依其積極性（即贊成程度）平均分為11組（堆），每一組各給一個代碼，積極性最高者為11，最低者為1。這個代碼也就是每個評判者給予每一條「敘述」的分數。換言之，每一條「敘述」都得到50個分數（最小是1，最大是11）。

⑶分別將每一條「敘述」的得分作成次數分配，取其中數(median)（即次數居於50％位置的分數），作為「量表值」(scale value)；並計算其「四分差」(semi-interquatile)（即次數居於75％之分數與次數居於25％之分數的差之一半，亦即$(Q_3-Q_1)/2$作為「模糊指數」(ambiguity index)。模糊指數愈大，表示評判者意見愈不一致。

⑷將「四分差」太大之「敘述」刪除，保留20條「敘述」作為題目，即可構成一個態度量表。下圖是一個「教學態度量表」的部分題目，每一個題目均註明「量表值」(scale value)。

教 學 態 度

以下各題敘述，如果您同意，就請在題號前畫✓，若您不同意，則請在題號前畫×

量表值	
10.2	1.教學是社會上最迫切需要的一種專業。
10.0	2.教師是國家的領導者。
8.7	3.教師是社會的工程師。
5.0	4.如果教師嚴格依照教科書教學，學生會學得更多。
2.6	5.就教學方法言，教學專業大約落後20年。

但在實施塞斯通量表時，量表值不可以顯示出來，以免受試者在表示「贊成」或「不贊成」的意見時，受到量表值的暗示。俟受試者填答之後，計算其「贊成」的題目之量表值，並以這些題目的量表值的中數作為該受試者的態度分數。

2.李克特氏量表(Likert scale)

李克特式量表就是大家熟悉的評定量表(rating scale)，係由社會心理學家李克特(R. Likert)所創。[8] 下圖是李克特育兒態度量表的一部分題目：

下列各題的敘述，請依您贊成或不贊成的程度，在適當位置作✓號。

	非常贊成	贊成	不一定	不贊成	非常不贊成
1.家裡有小孩，使夫婦更為親密	□	□	□	□	□
2.教小孩子學習各種事物是件愉快的事	□	□	□	□	□
3.小孩子本質上有許多惡性，所以要嚴格	□	□	□	□	□
4.照顧小孩使做母親的失去許多社交機會	□	□	□	□	□
5當遭遇生活困境時，小孩是鼓勵與希望	□	□	□	□	□

李克特量表的編製方法比塞斯通量表簡單，故在應用上相當普遍。其編製程序如下：

(1)針對態度對象編寫題目。題目中要包含正向（如上圖例題中的第1、2、5題）與反向（如第2與4題）的題目，同時儘量避免太多中性或極端傾向的題目。

(2)決定評定的等級（通常採5點或7點評定量表）。

(3)選取一群受試者（大約100人左右）填答上述題目。

(4)決定計分方式後記分。通常以正向題目選答「非常不贊成」者給1分，依次往上，填答「非常贊成」者則給5分；若是反向的題目，則反之，亦即填答「非常贊成」者則給1分，填答「非常不贊成」者給5分。因此，得分愈高，表示態度愈積極，得分愈低表示態度愈消極。以上圖的題目為例，得分愈高，顯示對「育兒」的態度愈積極、愈有責任感；得分低則顯示消極的育兒態度，不喜歡、不願負責等。

(5)計算每一名受試者的總分（亦即將每一個題目所得的分數加起來），並進一步分別計算每一個題目的得分與總分之相關係數(correlation coefficient)。相關係數太低的題目即予刪除。這個過程也就是題目分析(item analysis)。

(6)經題目分析後保留的題目就構成一份態度量表。實施時，只要依據受試者填答情形換算分數，計算總分即可了解該名受試者態度傾向。必要時也可依據個別受試者的總分計算群體的平均數，作為比較群體間態度傾向的指標。

3.語意區分法(semantic differential technique)

語意區分法係奧斯古(C. E. Osgood)等人所發展的測量技術，旨在區分持定概念(concept)的意義，也可以用來測量態度。[9]應用語意區分法時，須以兩極化(bipolar)的形容詞（或副詞）為題目，讓受試者依自己的想法與感受，在兩極之間標定位置，藉以表達對特定概念的看法。其基本題型如下圖所示：

	教　師							
①快　樂	＿＿	＿＿	×	＿＿	＿＿	＿＿	＿＿	悲哀
②不公平	＿＿	＿＿	＿＿	＿＿	＿＿	×	＿＿	公平
③和　平	＿＿	×	＿＿	＿＿	＿＿	＿＿	＿＿	兇猛

在上圖中，「教師」係受評的概念；而每一對兩極化的形容詞即為題目，每一題目區分為七個間隔，代表趨向兩極的不同程度；間隔上的「×」號，就是要受試者標記的符號。計分的方法係從1分至7分，依題目的積極性增加。以上圖的題目為例，第①題靠「悲哀」的一端可得1分，靠「快樂」的一端可得7分，其中依次為2，3，4，5，6分；第②題則方向相反，靠「不公平」的一端為1分，往右遞增分數，至「公平」一端則為7分。計算每一名受試者的總分，即可知其在積極為消極之間的偏向。

奧斯古等人在其原著中提出50對形容詞，經因素分析(factor analysis)後歸為三類：一是「評價」(evaluation)，如「誠實——不誠實」、「勇敢——懦弱」、「美——醜」等；二是「潛力」(Potency)，如「大——小」、「輕——重」、「深——淺」等；三是「活動」(activity)，如「主動——被動」、「快——慢」、「和平——兇猛」等。其中也有些形容詞可以同時歸在不同類別。不過，在應用語意區分法測量態度時，儘量選用評價性形容詞作為題目。至於填答方式與計分方法則與上述一般語意區分法相同。根據受試者的填答結果，可以了解個別受試者對特定概念（如教室、教學、校園、學校風氣等）的態度傾向與強度，也可以比較個別受試者對不同概念的態度差異。同時，也可以作團體間的比較，或比較同一團體對不同概念的態度。總之，語意區分法編題簡易，施行方便，又能作多種分析，可以作為蒐集教育研究資料的方法。

㈢測驗的編製與應用

在教育研究中，常須應用測驗(test)來了解研究對象的特質與屬性。大體而言，測驗有兩類：一是心理測驗，一是教育測驗。心理測驗包括人格測驗、智力測驗等；教育測驗則以學科學習性向測驗與成就測驗為主。由於心理測驗的編製是一項艱鉅的工作，不僅程序繁瑣，而且費時費力甚多，故常非研究者獨力所能負擔；況且目前學術界及民間均有專門機構編製各類心理測驗出售應用，故通常研究者只要依據研究目的選擇適用

的測驗即可，不一定要自行編製。因此，研究者須具備選擇與使用心理測驗的能力與技巧。至於教育測驗之應用，因其內容較具特殊性，現成的測驗未必盡符研究之需要，而常須研究者自行編製，故研究者應熟悉編製的方法與程序，以下分就心理測驗的選擇與實施以及教育成就測驗的編製程序加以說明。

1.心理測驗的選擇與實施

就基本性質來看，一般心理測驗大都是「標準化測驗」(standardized test)。標準化測驗具有三個基本特徵：其一，它是經由標準化的程序編製而成的；其二，它有標準化的實施程序與計分方法；其三，它有解釋測驗結果的標準。就編製程序言，標準化測驗係經由客觀、嚴謹的編寫測驗題目，並經「預試」加以修正，然後定式，故具有相當程度的效度及信度。就實施程序及記分方法言，標準化測驗均明確規定實施程序與記分方法，故任何人主持測驗及評計測驗結果均無二致。就測驗結果之解釋言，標準化測驗均有「常模」(norm)作為解釋的標準。這些特徵就是心理測驗的基本性質，也是「好測驗」應具備的條件。

選擇測驗時，要從兩個問題來考慮：其一，這個測驗的性能好不好？其二，這個測驗合不合本研究之用？蓋測驗的性能係從測驗的信度、效度、實施程序、計分方法、以及常模幾方面來檢查。性能好的測驗，一定要有相當程度的信度與效度，一定有明確簡便的實施程序與計分方法，也一定有適當的常模。通常，這些資料都會在測驗手冊(manual)中詳細說明。因此，在選擇測驗時，一定要詳細閱讀測驗手冊，了解測驗的性能。同時，要看看文獻中有關該測驗的使用效果及評論，作為選擇應用的參考。

其次，從實用的觀點來看，一個測驗是否適用，必須從研究目的、研究對象，以及研究資源來考量。我們要考慮，測驗的性質與功用是否符合研究需要？測驗的適用對象是否符合研究對象？測驗所需的時間、人力、經費是否符合研究的條件？如能在這幾方面充份檢討，才能選到一個性能好的，而且適用的測驗。

至於心理測驗的實施要領，基本上要依照測驗手冊上規定的標準化程序執行。無論是個別施測抑或團體施測，都要安排一個適當的施測場所，避免外界的干擾；然後依照測驗手冊上的指導語說明作答的方法，以及注意事項，並鼓勵受試者認真作答。施測過程中，不可自作主張的補充解釋題目，或暗示作答方法，否則將影響測驗的客觀性與標準化；再者，也要確實遵照規定，控制作答時間。時間一到，務必收卷，不可提前、也不可延後。同時，測驗結果之整理與計分，也要依據測驗手冊上提示的方法來執行。唯有依據標準化的程序實施測驗，測驗結果才可靠，作為研究資料才有意義。

2.教育成就測驗的編製程序

教育成就測驗之編製程序，大致可以分成五個主要步驟：(1)界定測量目標、(2)發展題庫、(3)預試、(4)分析題目、以及(5)建立信度與效度。以下逐步說明：

(1)界定測量目標

測量什麼？這是編製成就測驗所要考慮的第一個問題，由於教育成就測驗必須以學習的材料、以及預期的學習目標為依據，因此在編製時，先要熟悉學習材料的內容，並

確定學習的目標，進而依據內容與目標製成一個「雙向細目表」(two-way specification)，如下圖，作為選題及分配題數的架構。

內容 目標 題數	第1單元	第2單元	第3單元	第4單元	第5單元
1.了解名詞與字彙 2.了解事實與信息 3.了解普遍原理 4.解釋關係 5.計算與解題 6.預測與推論 7.評鑑與判斷					

在上述雙向細目表中，橫排代表各項學習目標，直行代表學習內容（可直接填上單元名稱）。據此就可周全且均衡的分配測驗題數，填在「目標」與「單元」交叉的適當位置上，作為編選測驗題目(item)的依據。同時，也要考慮測驗的題型與難度，一併在雙向表上適當位置註明。

⑵發展題庫

在確定測量目標，決定測量項目之後，就要編寫測驗題目，建立題庫(item pool)。一般而言，研究用的成就測驗大都以「客觀測驗題」(objective-test item)為主，而較少採用申論的題目(essay)，除非測量的目的與寫作能力有關。常用的客觀測驗題包括：是非題，選擇題，配合題，以及簡答題（類似填充題的簡答）。編寫這類題目時，必須講求命題的技術，始能發揮題目的測量作用。以下列舉一般性的命題原則以供參考，至於各類題目的命題方法與要領，務請參閱教育測驗的專門書籍。編寫成就測驗題目的一般原則如下：⑩

①試題之取材宜均勻分佈，且應包括教材的重要部分。

②試題文字力求淺顯簡短，題意須明確，但不可遺漏解題所依據的必要條件。

③各個題目須彼此獨立，不可相互牽涉。

④試題應有不致引起爭論的確定答案。

⑤試題之中不可含有暗示本題或其他題目之正確答案的線索。

⑥試題文句須重新組織，避免直接抄襲課文或原來材料。

⑦試題宜注重基本原理之了解與活用，不可偏重零碎知識的記憶。

⑶預試

初步編寫完成的題目，必須經過預試，鑑別題目品質後，始能採用。預試的對象必須從研究的母群體中選取，也要有代表性。實施時，作答時間可以稍為加長，讓受試者均有完卷機會，但要紀錄完卷時間，以供決定測驗長度之參考。在受試者作答過程中，應容許

適當的發問（尤其是與題意有關的問題），藉以確定受試者是否完全了解題目文字意義。總之，預試時要儘量保持正式測驗時的情境與條件，但也要注意施測過程中受試者的反應，作為修訂測驗題目或施測程序之參考。

(4)分析題目

分析題目的目的，是要了解每一個題目的「難度」(difficulty)與「鑑別度」(discriminability)，作為取捨的依據。通常，這個步驟稱為「題目分析」或「項目分析」(item analysis)。所謂「難度」，是指題目的難易程度。如果有一個題目，全部受試者都答對了，顯然這個題目太容易；如果全部受試者都答錯了，顯然題目太難了。因此，由答對人數的比率可以推論題目的難度，難度太高或太低的題目都不是好題目。至於「鑑別度」則是指測驗題目能鑑別出受試者能力的程度。理論上，能力高的受試者比較可能做對題目，能力低的受試者比較可能做錯題目。因此，若有一個題目能力高的受試者答對了，能力低的受試者答錯了，那麼這個題目就有鑑別力；相反的，如果能力高的受試者答錯了，但能力低的受試者卻答對了，那麼這個題目就缺少鑑別力。鑑別力低的題目就不是好題目。

在分析題目難度與鑑別度時，係以預試的結果為依據。先計算每一受試者的測驗總分，然後按分數高低排列。分數在前27%者，稱為「高分組」，分數在後27%者，稱為「低分組」。然後依據下列公式分別計算每一個題目的難度：

$$\text{題目難度(P)} = \frac{\text{高分組答對人數} + \text{低分組答對人數}}{\text{高分組總人數} + \text{低分組總人數}}$$

至於題目的鑑別度，則依下列公式計算：

$$\text{題目鑑別度(D)} = \frac{\text{高分組答對人數}}{\text{高分組總人數}} - \frac{\text{低分組答對人數}}{\text{低分組總人數}}$$

通常，題目難度以.50最為適中，而鑑別度則愈大愈佳。不過，在實際選擇題目時，均定一個範圍，難度以.33至.67之間，鑑別度則要大於.30，凡逾越此範圍的題目，須予以刪除。刪除後留下的題目則按難度排列，自易而難。

預試的功用是在測試題目的適當性，故作為預試用的題目數量要多，以免在刪除修正後，所遺可用題目不足。若經預試修正後的題目確實不夠用，則須增編題目，再經預試，並作題目分析後，始可應用。若勉強選用難度及鑑別度均不適當的題目，將減低測驗結果的有效程度。

(5)建立信度、效度，並編寫手冊

標準化測驗必須具備相當程度的信度(reliability)及效度(validity)，故在編製成就測驗時，也要提示有關信度及效度的資料。有關信度及效度的性質，以及建立信度及效度的方法、本文第一部分已有說明，請自行參閱。

在建立信度與效度之後，必須編寫測驗手冊，作為未來正式實施測驗，以及計算分數的依據。通常，測驗手冊的內容，除了說明測驗的性質、目標、內容、編製經過、以及信度與效度資料外，要明定實施程序與計分方法。由於研究者應用成就測驗蒐集研究資料時，未必自己主持測驗之實施與計分，故這兩項一定要在手冊中說明清楚，以免影響測驗實

施與結果計分的可靠性。至於常模，並非每一項研究都需要。因為應用成就測驗所蒐集的資料，常作為團體的比較與分析，而較不需要了解個別分數的團體地位，故研究者可斟酌實際情形決定是否建立測驗常模。一旦測驗編製完成，即可藉之蒐集研究所需資料。至於實施成就測驗的要領，大體與一般心理測驗的實施相同，力求客觀、標準化。

以上各節分別說明評量的基本概念與主要方法，旨在增進有意從事教育研究者有關評量的了解，俾在選擇與應用各類評量工具時，能做最適切的決定與表現，且在自行編製評量工具時，能做最周密的考慮，藉以提昇教育研究的品質，唯以篇幅所限，本文之說明大都提綱契領，故難免不夠詳盡，如果讀者覺得本文有所不足，務請自行參閱參考書目中所列有關測驗與評量的專書。再者，本文所示觀點與資料，部分摘自拙著「教育研究：基本觀念與方法之分析」乙書，特此註明。（吳明清撰）

附註：

①參見Stevens, S. S.(1959). Measurement, Psychophysics, and Utility. In C.W. Churchman & P. Ratoosh(Eds). Measurement: Definition and Theories. New York: Wiley, P.19

②這四種測量尺度自Stevens於1959年提出後，沿用迄今。參見上註PP.18-63。

③美國心理學會(American Psychological Association). 教育研究學會(American Educational Research Association)與全國教育測量委員會(National Council on Measurement in Education)共同於1985年修正出版「教育與心理測驗標準」(Standards for Educational and Psychological Testing)，對於效度的解釋，強調效度的「單一概念」(Unitary Concept)，故捨棄傳統有關效度的分類方式，改以「效度證據」代之。詳見：American Psychological Association(1985). Standards for Educational and Psychological Testing. Washington, D.C.: American Psychological Association, P. 19。

④測量誤差係指測量時實得分數（或稱觀察值）與真實分數（或稱真值）的差距。通常以實得分數的變異量中所占比率來說明。測量誤差愈小，測量的信度就愈高；反之則反。參見：郭生玉（民78），心理與教育測驗（4版）。台北：精華書局，68頁。

⑤英文的scale一字具有「尺度」與「量表」兩種意義。尺度是測量的準則，而量表則指由尺度構成的測量工具。

⑥有關問卷題目的分類、型式、與設計，請參閱S. Sudman & N.M. Bradborn(1982). Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass.

⑦參見L. L. Thurstone & E. J. Chave(1929). The Measurement of Attitudes. Chicago:University of Chicago Press.

⑧參見R. Likert(1932), A Technique for Measurement of Attitude. Archives of Psychology, 22,140,PP.1-55.

⑨參見C. E. Osgood, et al.(1957). The Measurement of Meaning. Urbana, IL:University of Illinois Press.

⑩引自簡茂發與郭生玉撰（民67），編製測驗題目的技術。刊於楊國樞等編著，社會及行為科學研究法。台北：東華書局，頁442- 444。