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）