【教育】

《教育測驗與統計》

試題評析

教育測驗部分著重測驗編製的基礎理論，並且重視測驗的解釋與實際應用及應注意的事項，教育統計部分則著重統計方法的使用與解釋，統計計算部分極少，注重研究方法或實驗設計與統計分析方法的連結，因此未來，對於統計方法的使用與解釋應多下功夫，並非會演算統計結果即可，應加強統計分析實例的分析。

一、某研究者想探討EQ訓練是否會對企業主管的管理能力有改善效果。該研究者以隨機分派的方式，把25名主管分配到實驗組（接受EQ訓練），25名分配到控制組（未接受任何訓練）。所有參與者都接受EQ測驗的前測和後測。經過EQ測驗取得前測（實驗組平均數7.75；控制組為7.26）和後測（實驗組平均數11.24；控制組為8.11）成績。統計分析的結果如附表（有一名參與者因故流失），請問：

(一)何謂「隨機分派」？其作用為何？（5分）

(二)從研究情境和附表判斷，該研究者使用何種研究設計和統計分析方法？（5分）

(三)附表中「交互作用」的意義為何？在本研究結果中最可能代表的意義為何？（10分）

(四)根據附表，研究者宣稱：由於組別差異未達顯著水準，因此EQ訓練對管理能力沒有幫助。此種結論是否合理？為什麼？（10分）

附表：EQ訓練的統計分析摘要表

變異來源	SS	自由度 (df)	均方 (MS)	F值
受試者問	581.27	48	75.96	
組別	62.97	1	62.97	5.73
群內受試	518.30	47	10.99	
受試者內	107.10	49	28.58	
前後測	20.78	1	20.78	12.30*
交互作用	7.11	1	7.11	4.22*
誤差（前後測x群內受試）	79.21	47	1.65	

*P<.01

答:

【參考資料：林清山(心理與教育統計學)p385，江金色(教育研究法)p10-11】

等組前後測設計：

(1)模式：

實驗組： R O_1 X O_2

控制組： R O_3 O_4

(2)實施步驟：

A.利用隨機抽樣選取受試者，並以隨機分派將受試者分派至實驗組和控制組。

B.實驗處理前，兩組皆接受前測。

C.實驗組接受實驗處理，而控制組則無。

D.實驗處理後，兩組皆接受後測。

E.使用適當的統計考驗。

(3)優點和缺點：

A.優點：等組前後測設計能完全控制所有影響內在效度的因素，研究者可以有信心說明，實驗結果是由實驗處理所產生的效果。

B.缺點：等組前後測設計使用前測，因此，就外在效度而言，此設計無法控制「測驗的反作用或交互作用效果」，無法推論到沒有前測經驗的情境，而且「實驗安排的反作用效果」亦不易控制。

(4)考驗方法：可以採用獨立樣本t檢定，比較後測平均數，如果組數在兩組以上，需採用變異數分析。但是，最適當的統計分析方法應是「共變數分析」，即以兩組的前測分數作爲共變量進行共變數分析。

(一)隨機分派(Random Assignment)：主要是在確保實驗的「內在效度」。

經由隨機抽樣選出的受試者，必須分派到不同組別進行研究，則須靠隨機分派，否則會造成兩組的條件不同，致使無法正確瞭解研究結果是否真正源於自變項的操弄。

(二)依研究情境和附表判斷，研究者是使用「真實驗設計」(True Experimental Design)中的「等組前後測設計」－有實驗組、控制組及前測、後測之設計，統計分析方法則爲變異數分析方法中之「混和設計二因子變異數分析」－受試者間（A）：實驗組、控制組（df=2-1）；受試者內（B）：前測、後測（df=2-1）；及交互作用（AxB）（df=1）。

(三)交互作用（AxB）的意義即是在看A與B兩個因子同時對測量變項影響的情形，若交互作用達顯著水準，則A與B兩個因子同時對測量變項影響較A或B單純影響測量變項來的有意義。就本研究結果而言，交互作用達顯著水準（4.22*，p<.01），顯示A（實驗組、控制組）與B（前測、後測）兩個因子同時對測量變項有影響。

(四)「研究者宣稱：由於組別差異未達顯著水準，因此EQ訓練對管理能力沒有幫助。」，此種結論是不合理的。因爲，從變異數分析摘要表得知，交互作用達顯著，顯示A（實驗組、控制組）與B（前測、後測）兩個因子同時對測量變項有影響，再從各組前後測的平均數及交互作用分析可得知，此研究的交互作用屬於「次序性交互作用」，前測時各組平均數差異不大（7.75與7.26），而後側時兩組平均數（11.24與8.11）有拉大差距的趨勢。若以受試者間分析：前後測之平均數合併，自然不易顯現兩組的差異，惟以前後測分開比較，自然可看出「實驗組後測」優於「控制組後測」，故可解釋EQ訓練對管理能力是有幫助的。（林清山p384）

二、卡方檢定(χ^2)的主要作法有那幾類？各適合應用在那些研究問題？（25分）

答：【參考資料：林清山(心理與教育統計學)p281】

(一)卡方檢定（X^2）統計法的用途：

1.適合度考驗(test of goodness of fit)：就某單一變項考驗其觀察次數與相對應的期望次數是否相同。

2.百分比同質性考驗(test of homogeneity of proportions)：考驗研究者所感到興趣的J個群體在I個反應方面的百分比是否都是一樣，亦即這些群體的反應是否爲同質。在I × J交叉表(crosstabulation table)上，只有一個變項是設計變項(design variable)，分爲J個群體；其餘一個則爲反應變項，分爲I種反應。

3.獨立性考驗(test of independence)：考驗研究者感到興趣的兩個自變項是否互爲獨立；如果不是互爲獨立事件，則繼續進行「關連性考驗」(test of association)，以瞭解二者之關連的性質和程度。在這種用途裏，I× J交叉表上的兩個變項均爲「設計變項」。

4.改變的顯著性考驗(test of significance of “change”)：用來考驗同一群受試者對一件事情的前後兩次反應之間的差異情形。在I× J交叉表上的兩個變項均爲反應變項。

(二)適合使用的研究問題：

1.適合度考驗：研究者可以根據一個自變項，例如性別、年齡、社經水準、身心特徵等「屬性變項」，或色光、教學法等「處理變項」來將自變項加以分爲幾個類別(categories)或層次(levels)。只根據一個自變項，然後將它分爲幾個類別或層次所搜集到的資料稱爲單因子分類(one-way classification)的資料。

2.百分比同質性考驗：進行X^2考驗時所搜集到的資料常安排成I個橫列和J個縱行的方格形狀的表，稱爲「交叉表」或「列聯表」(contingency table)。使用X2考驗進行百分比同質性考驗的主要目的在於考驗被調查的J組受試者在I個反應中選擇某一選項的百分比是否有顯著差異。例如：研究者是想

要瞭解家長、教師、心理學家、和學生四組受試者對「成績退步的學生應該受到老師的懲罰」這一問卷項目的三個選項中選擇「贊成」者之百分比是否有所不同。由此可見，在進行百分比同質性考驗時，交叉表的兩個變項中，事實上只有一個變項是「設計變項」(design variable)，亦即研究者所操弄的處理變項，或研究者找來比較的類別變項；至於交叉表的另外一個變項則爲反應變項，不算是設計變項。

3.獨立性考驗：獨立性考驗的目的在瞭解自母群中取樣而來的一組受試者的兩個設計變項之間是否互爲獨立？如果不是互爲獨立，則二者的關連性的性質和程度如何？因此，在進行獨立性考驗時，I×J交叉表的兩個變項均爲設計變項。交叉表中，只有總人數N事先知道，其他細格人數或邊緣人數由調查決定。

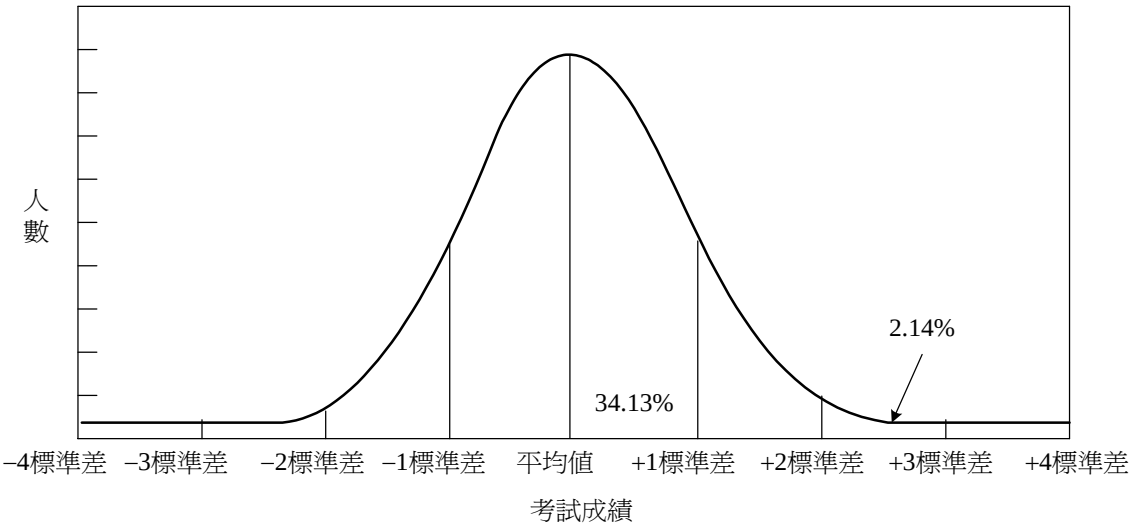
4.改變的顯著性考驗：使用交叉表的資料，除了百分比同質性考驗和獨立性考驗之外，還可用來進行「改變的顯著性考驗」。所謂改變的顯著性考驗的目的是檢定同一群受試者在同一個變項的前後兩次反應之差異是否達到顯著水準。因同一群受試者每人均需前後被重複測量兩次，所以相當於「重複量數」(repeated measures)的設計。進行改變的顯著性考驗之前，研究者事先只知道被調查的總人數N。

三、某心理學家設計一系列的性向測驗。在制訂空間語文性向測驗的常模時，用100作為平均數，15作為標準差。假定該心理學家施測的樣本在其測驗表現上呈常態分配。請問：

(一)參考附圖，當甲得分為115時，其百分等級為多少？此百分等級的意義為何？（5分）

(二)甲生的T分數為何？（5分）

(三)若該生另接受同一系列的標準化空間能力測驗（平均數80，標準差10），得分為100。請問該生空間性向和語文性向何者較高？為什麼？（5分）



附圖：常態分配圖

答：【參考資料：郭生玉（心理與教育測驗）p138】

(一)甲生的百分等級爲：Z=(115-100)/15=1＝＞標準差（σ）爲的位置，依附圖得知甲生的百分等級（PR）爲84；PR=84 代表甲生的測驗分數勝過84%的人。

(二)甲生的T分數爲 T＝10*Z+50 ＝＞ T=10*1+50=60。

(三)空間語文性向 Z＝1.0

另一空間性向 Z＝(100-80)/10=2.0

另一空間性向Z（2.0）大於 空間語文性向 Z(1.0)

由附圖得知：甲生另一空間性向的PR＝98 ＝＞勝過98%的人，

而甲生空間語文性向的PR＝86 ＝＞ 勝過84%的人，

因此，甲生的另一系列標準化空間能力性向（平均數80，標準差10，PR=98）較高。（註題意敘述不

清，有些矛盾）

四、解釋名詞：（每小題5分，共30分）
（一）雙向細目表（two-way specification table）
（二）實作評量（performance assessment）
（三）構念效度（construct validity）
（四）測量標準誤（standard error of measurement）
（五）系統性測量誤差（systematic measurement error）
（六）動態評量（dynamic assessment）

答:

(一)雙向細目表(two-way specification table)：內容效度尚沒有一種數量的表示方法，它的確定主要是採用邏輯的分析方法，仔細判斷每一個題目是否符合教材內容與教學目標，如果測驗的題目很能代表教材內容的樣本，及所預期的行爲改變，而沒有其他無關因素(如閱讀能力、指導語不清楚)的影響，則表示測驗有良好的內容效度。因爲此種分析是屬於邏輯的分析與合理的判斷，故又稱爲合理或邏輯的效度(rational or locical validity)。就教育測驗而言，在編製題目之前，均依據各種相關資料設計編製的說明書，詳細說明教材內容的主題、範圍、教學目標、及其相對的重要性。並根據教材內容與教學目標(預期的行爲改變)所建立的雙向細目表(two-way specification table)，可用來判斷測驗的內容效度。

(二)實作評量（performance assessment）又稱非紙筆測驗，係指根據學生實際完成一項特定任務或工作表現所作的評量。以觀察和專業判斷來評量學生學習成就的評量方式都可以稱爲實作評量（Stiggins，1987），其型式非常的多元化，例如建構反應題、書面報告、作文、演說、操作、實驗、資料蒐集、作品展示等，都是實作評量的例子，案卷評量也是實作評量的一種型式。實作評量具有下列幾點特徵（Herman， Aschbacher， & Winters， 1990）：1.要求學生執行或製作一些需要高層思考或問題解決技能的事或物；2.評量的作業（tasks）是具有意義性、挑戰性且與教學活動相結合；3.評量的作業能與真實生活產生關聯；歷程（process）和作品（product）通常是評量的重點；4.表現的規準（criteria）和標準（standards）—也就是評量的重要層面與給分標準，要事先確定。實作評量有時也被稱爲真實性評量（authentic assessment）

(三)構念效度(Construct validity)：指測驗能夠測量到理論上的構念或特質的程度(Anastasi， 1982，p.144)。易言之，就是指測驗分數能夠依據某種心理學的理論構念加以解釋的程度。因此，凡是根據心理學的構念，對測驗分數的意義所做的分析和解釋，即爲構念效度。構念(construct)心理學上的一種理論構想或特質，它是觀察不到的，但心理學假設它是存在的，以便能解釋一些個人行爲。像智力、性向、動機、焦慮、批判思考、社會性、內向性、外向性或機械性向等，均爲心理學上的理論構念，或假設性的概念。這些構念都有其心理學上的理論基礎，依據其理論可以預測人類的行爲，而提出行爲上的假設，然後加以驗證。例如，從焦慮的理論中，我們可以預測在競爭的情境裡，個人的焦慮較平常的情境爲高，或焦慮高的人，其抱負水準較高等。這些現象都可收集實際的證據，而予以驗證。一般而言，要發展測量心理學構念的工具和建立其效度，可根據下列幾個主要的步驟(Stanley & Hopking， 1981 p.106)：

1.根據構念的理論分析，發展一套測量的題目，亦即從構念的有關理論中，預測可能的行爲，並據此設計測驗的題目。
2.提出可考驗構念與其他變項間關係的預測。例如，預測焦慮測驗上的分數和臨床評定的焦慮程度有關，或預測焦慮分數和抱負水準有關等。
3.從事實證性的研究，以驗證上述的預測(假設)，亦即採用各種方法收集實際的資料，考驗第二步驟所提出的預測。
4.淘汰和理論的構念相反的題目(或修正理論)，並從第二和第三步驟再開始。如果上述的預測成立，測驗的構念效度就獲得支持。相反的。如預測不成立，則不是效度有問題，就是理論，或者是兩者。

(四)測量標準誤(standard error of measurement，$SE_{meas.}$)：是用來表示測驗信度的方法之一。信度係數較適合於比較不同測驗的信度，而測量標準誤則較適合於解釋個人的分數(Anastasim， 1982， p.127)。假

如能夠在標準情境下，使用相同測驗或複本測驗測量同一個受試者相當多次(假定不受練習、疲勞等影響)，所得分數的平均數稱爲個人的真實分數(true score)。例如，倘以比西量表重複測量某生一百次，將會得到100個智商，但由於機遇誤差的影響，這些智商未必相同，可能有的高於其真實分數，而有的卻低於其真實分數，其變動情形不一。不過，就一般而言，這些智商會以該生的真實分數爲中心而構成常態分配。這個分配的標準差，就是所謂的測量標準誤。換另一種方式來說，該生100次的智商和其真實分數會有一個差(也許是正的、負的或零)，此差稱爲測量誤差，這些測量誤差分配的標準差，就是測量標準誤。

因爲重複測量同一個人相當多次，而不會改變所測量的特質，實際上是不可能的，所以，測量標準誤通常是從團體的資料，加以估計，最簡易的方法是直接由測驗的信度來計算，其公式如下：

$$SE_{MEAS.} = S_X \sqrt{1 - r_{XX}}$$

$SE_{meas.}$：測量標準誤

S_x：測量的標準差

r_{xx}：測量的信度係數

(五)系統性測量誤差(systematic measurement error)：是指能使測驗分數產生變動，但是，它是在一種固定、一致的方式下高估或低估分數。易言之，在不同情境中，它對一位受試者的影響是一樣的，對所有受試者在相同情境中的影響，也是一樣的。例如，實施速度測驗未遵守指導語的時間限制，多給受試者5分鐘的時間做答，這對分數的影響是一致性的高估，因此，它不會影響信度，但會影響效度。一個體重計對每一個人體重的測量總是多一公斤，這就是一種系統性測量誤差。此種誤差又稱爲常誤(constant error)或偏誤(biased)。導致系統誤差的因素，主要有學習、訓練、遺忘與生長等。

(六)動態評量(dynamic assessment)：是由Feuerstein（1979年）首先使用的，Feuerstein在「表現遲緩的動態評量－學習潛能評量的設計理論工具與技術」一書首先使用「動態評量」一詞，而後研究者便相繼使用。

「動態評量」最初用來研究：

A.具有認知或是情緒功能損傷成人的認知歷程。

B.作爲實施標準化常模測驗之後的臨床參考。

C.特殊教育上的大量研究，如：「智能不足學生之鑑別」、「學前學生能力之探測」、「學習障礙」、「聽障成人及學生」、「文化不利學生及資優學生學習能力之探測」。

「動態評量」是指：教師以「測驗－介入－再測驗」（test-intervene-retest）的型式，對學生一般認知能力或特定學科領域進行持續性學習歷程的評量。藉此了解「教師介入」與「學生認知」之間的關係，以及學生認知發展的可修正程度，確認學生所能發展的最大學習潛能。並診斷學生學習錯誤原因，提供處方性訊息，以進行適當的補救教學措施。動態評量突破了傳統靜態測驗時評量情境標準化的要求，主動變化測驗情境，來比較個體認知學習能力的差異，並檢視評量過程，以找尋增進個體學習的有效策略，動態評量主張排除環境不利因素的影響，允許積極性的協助，允許主受試之間的互動關係， 評量目標不在於學生過去已有的知識、技巧與能力，而是在於評量其身心成長、認知改變與未來學習潛能的程度。（吳國銘、洪碧霞、邱上真，民84）。