

101年公務人員特種考試身心障礙人員考試試題

等 別：四等考試

類 科：教育行政

科 目：教育測驗與統計概要

一、實作評量評分時常發生數種缺失，包括個人偏誤 (personal bias errors)、月暈效應 (halo effect) 和邏輯謬誤 (logical error)。請就此三者之定義加以說明，並簡單區分其差別。

【擬答】：

(一)個人偏誤 (personal bias errors) 偏見 (Bias)：教師常會因學生性別、種族、先前經驗等因素會對學生有先入為主的偏見觀念存在，且對不同族群的評分結果亦不同。此一偏見會造成對學生評量不具真實性，影響其效度與鑑別度。

1. 個人評分偏見：包括「嚴苛偏失」(Severity Error) (常將學生的表現一律評定在較低等級)、「寬容偏失」(Generosity Error) (常將學生的表現一律評定在較高等級)。

2. 個人偏見：觀察者過去經驗常會導致對情境、現象或行為產生不同知覺與解釋，觀察者具有某些偏見或態度，觀察結果更難以客觀正確。

(二)月暈效應 (halo effect)：又稱光環效應，教師根據單一特徵對某位學生的整體印象，是指評價過於實際表現，重點在於印象擴散作用。教師與學生經常相處，對學生瞭解會產生一般性印象，針對學生的實際表現進行評分即成系統性評分誤差；如學生性別、容貌、身份地位或族群意識的影響。月暈效應有以偏概全，先入為主概念，如匿名彌封、閱卷時受無關因素影響、評定量表、項目檢核表或後面題目評分受第一題影響；與月暈效應相反的為尖角效應 (Horn Effect)，亦即評分時傾向一致性低估。

(三)邏輯謬誤 (logical error)：評定者容易將二項特質或變項無關係卻視為直接關係，如評定者易將學業成績較佳學生評為智商較高、教師在評量學生的人際關係或學校適應能力時 (高分代表能力好)，若傾向於將學業成績較好的學生也給予較高的分數。

(四)一般個人偏誤與月暈效應大都必經過觀察後產生觀察的誤差，而邏輯謬誤則無須進行觀察就產生觀察的誤差。值得注意的是，邏輯偏誤和月暈效應並不相同，月暈效應是指評分者無法區別相同受試者內之所有特質；而邏輯偏誤則是由於評分者無法區別欲測量的主要特質和無關的特質。

二、教師在編製教學評量時，為呼應課程綱領的能力指標並提昇評量的信、效度，應遵循那些步驟？

【擬答】：

有關學生之學習評量，應參照學生成績評量準則之相關規定辦理，其辦法由教育部另定之。為了編製一份品質優良的教師自編測驗 (Teacher-made Test)，其步驟包括下列幾點：

(一)確立測驗的目的與目標：

使用的教學目的不同則編製測驗就應不同，一份測驗要能測量學生學習成果，教師必須事先確立所要測量教學目標為何，可以根據Bloom所提認知、情意、動作技能領域三大教學目標。有關教育評量之編製及施測事宜，應參照國民中小學課程綱要之能力指標及相關法令之規定。

(二)設計雙向細目表：

1. 編擬設計測驗藍圖主要是建立雙向細目表 (Two-way Specification Table)，此表考量二類資訊：①橫軸為教學目標；②縱軸為教材內容。以此繪製二向度分類表且平均分配好試題比重或題數於每一細目格中，儘量讓試題能充分涵蓋所要評量教學目標與教材內容。

2. 編製雙向細目表遵守原則：

(1)命題的內容範圍必須是有教學過的教學範圍，沒有教過的內容，教師不應該編製試

公職王歷屆試題 (101 身心障礙特考)

題，因為無法反應教學成效，也不公平與客觀。

(2)雙向細目表中邊緣細格中數字不可為0，應表示無恰當試題。

(3)教師可以根據自己命題與評分習慣，加以增減題目題數。

(三)決定測驗題型與命題技巧：

試題是測驗的主軸，不同測驗類型各有獨特的編製原則與方法，因此教師在選擇測驗題型，須瞭解其特性與技術，才能發揮其功能。試題題型可分為二大類：

- 1.客觀測驗（選擇型試題）（新法考試）：包括選擇題、是非題、配合題、填充題、解釋性習題、簡答題。
- 2.論文測驗（補充型試題）：包括申論題、限制反應題。

(四)編輯測驗：

- 1.測驗題數：測驗題數多寡無絕對標準，應依測驗目的、題目類型、信度高低、學生年齡、學生能力，而有不同題數標準。如高信度測驗，題數應增加；常模參照測驗題數應比標準參照測驗多；年齡愈長，題數應愈多；能力較低，學生題數應較少；速度測驗題數應較難度測驗多。
- 2.測驗難度：測驗難度決定於測驗目的，如標準參照測驗是教學後，評量學生精熟學習內容的程度預期均會得滿分，題目難度較低，其鑑別力低；常模參照測驗，題目難度很重要，因為目的在確定個人分數在團體中相對位置，題目以難易適中為主，其鑑別力高。
- 3.試題編排：編輯正式測驗試題，過程包括試題編排、實施指導說明與答案設計：
 - (1)試題編排：其方法有：①依據試題難度；②依據試題類型；③依據教材內容；④依據測驗能力。
 - (2)實施指導說明：說明中需包括測驗目的、作答方法、計分方式、時間限制、有無倒扣分、攜帶工具或教科書、答案紙與試題紙是否分開、考試中途是否可以舉手發問。
 - (3)答案設計：設計作答的答案紙，以便容易計分。

三、「常模」(norm)與「標準」(standard)有何不同？為何測驗常模不能當作良好表現的標準？請至少指出三項判斷常模適切性的要素，並分別舉例說明常模與標準的適用時機或情境。

【擬答】：

(一)常模 (Norm)：

常模 (Norm) 解釋測驗分數的依據，指特定參照團體或標準化樣本 (Standardization Sample) 在測驗上所得到的平均數水準典型表現，因此有時常模會等於平均數。根據對標準化樣本施測所得的結果，經統計分析，整理出一個系統的分數分配表，按照高低排列，所得的平均數，即是常模。

(二)常模與標準化測驗均為測驗編製過程，二者常混淆不清。常模並非標準，它是依據標準化樣本在測驗的實際表現的結果建立，而並非事先決定標準。常模不應為學生必須達到標準，而是學生的一種常態表現或典型表現。

(三)為何測驗常模不能當作良好表現的標準

- 1.過分強調相對地位高低、重視排名競賽，容易造成教育反效果。產生能力高者驕傲自大，能力低者長久打擊缺乏自信心。
- 2.無法指出學生真正成就，單從常模參照測驗分數，無法瞭解學生達到教學目標的程度，也無法實質反應學生學業成就；並不適合診斷學生學習困難，或評鑑某項教材和教法的優點。(陳英豪、吳裕益，民88)
- 3.經常省略大多數學生都會答對試題，因此無法提供教師認為最重要且有用訊息，教學與測驗常無法配合。測驗與教學不易配合，常模參照測驗結果解釋不夠明確，教師無法根據測驗結果設計教學計畫或改進教學缺失。
- 4.常因不同參照團體其測驗分數有所差異。例如：資優班與普通班同樣百分等級，其學業實際成就差異必很大。

(四)常模判斷標準

公職王歷屆試題 (101 身心障礙特考)

建立常模表時，不可能對整個母群體施測，必須只抽取部分樣本構成常模樣本 (Norm Sample) 或稱標準化樣本。測驗編製者所提供常模是否恰當，可由四方面價值判斷：

1. 新近性：測驗常模隨著時間，會愈來愈不適合用來解釋測驗分數。因為隨著時代演進，如教學方法更好、課程重點改變，常模即不再合適，甚至連測驗分數本身也不準確。因此，測驗常模應該採用最新的，但不能用測驗手冊的出版日期判斷。如九十年代人們價值取向和六十年代必有不同，隨著社會變遷、時代演進，課程內容不斷更新，傳統測驗與常模不再適用。
2. 代表性：代表性指選取常模樣本的人口特徵須與母群體特徵相吻合。測驗分數經常有測量誤差與抽樣誤差。其中抽樣誤差是由於抽樣不當所致，建立常模樣本其代表性比樣本大小重要，因為草率抽樣來的 20000 名樣本，比正確取樣過程 200 個樣本更不具代表性。隨機取樣所得樣本最具代表性。
3. 適切性：使用常模解釋分數，必須瞭解常模樣本性質是否和測驗對象相似，測驗對象背景條件如學業性向、教育經驗、文化背景、年齡是否和常模團體相近。總結任何測驗分數的正確解釋，均須顧及常模團體性質與常模類型。

四、(一)某測驗分數為常態分配，其百分等級50-60之間的原始分數差距，比百分等級80-90之間的差距，是大還是小？換句話說，接近平均數處的分數若稍作改變，較遠離平均數之分數稍作改變，其相對的百分等級之改變會較大或較小？這是什麼原因？請用次序量尺或等距量尺的觀念來說明。

(二)當某位學生之PR值為84時，請問其贏過多少% (percent) 的人？

【擬答】：

- (一) 1. 百分等級50-60之間的原始分數差距，比百分等級80-90之間的差距小
2. 接近平均數處的分數若稍作改變，較遠離平均數之分數稍作改變，其相對的百分等級之改變會較大。
3. 百分等級量尺單位並不相等，是一種次序量尺，靠近常態分配中央原始分數單位小，二端單位大；例如常態分配位於分配中央人數多，一兩分差距其百分等級差異很大，而在分配兩端因為人數較少，所以即使分數有很大差異，其百分等級變化卻不大。亦即接近中央原始分數差異會因為百分位數的轉換而擴大，但分配二端原始分數差異則會大幅地縮小。(危芷芬，民88)

(二) PR=84時，表示在團體中贏過多少84%的人。