避免評分的偏誤

洪素蘋

在電視機裡的歌唱選秀節目中，歌手的表現交由現場觀眾評分。
這些觀眾如何評分？他們是否完全遵照選手的歌唱表現來評分？
在評分過程中又會受到什麼因素影響？

無所不在的人為評分

評量一個人表現優劣的方式很多，常見的如紙筆評量，雖然也經常仰賴人工方式評分，但如果題目有標準答案，則儘管評分者不同，評分結果會大同小異。如果是開放性的申論題型，就可能因為涉及到每位評分者的主觀判斷而造成評分結果不一。生活中出現需要人為評分的場合甚多，舉凡歌唱選秀節目，跳水、體操、美術、作文、創作等比賽，升學與工作面試，甚或校園裡常見的整潔、秩序比賽等，都仰賴評分者針對受評者的表現評分。

評分者不考慮歌手的歌唱表現，僅依參賽者的服裝儀容評分。

這些比賽或考試是以邀請專家依據事先擬好的評分準則評分。如升學考試中的中文或英文寫作評分，主試機構事先提供評分者一套評分準則，以及不同評分等級的作品範本，評分者則依據這一評分準則逐一評定每位考生的作文表現。維持這套評分機制的公平性，主要建立在評分者能夠了解評分的準則、能夠分辨出不同評分等級作品之間的差異，並且能夠秉持客觀與公正的態度審視每一件作品。

**月暈效應指的是當評分者在評分時，
只根據受評者的某部分表現（好的或壞的）類推做為全面評分的依據。**

趨中傾向是指評分者給受評者的分數集中在某一固定的範圍內變動，而沒有顯著的好壞之別。

以口說、作品、歌唱或體能比賽為例，決定受評者的表現主要來自於受評者本身所展現出來的能力、這項比賽或面試問題的難度，以及評分者的評分。然而事實上，評分者在評分過程中可能摻雜其他因素而影響評分的客觀與公平。例如，評分者是否能夠從一而終地遵照評分準則評分呢？抑或是面對眾多等待評閱的試卷或作品，因為疲倦而胡亂評分？這也正是為什麼有些仰賴人為評分的場合會引發評分不公的爭議。

生活中常見的評分者偏誤有月暈效應、趨中傾向、過寬或過嚴傾向、刻板印象等。

月暈效應

月暈效應指的是當評分者在評分時，只根據受評者的某部分表現（好的或壞的）類推做為全面評分的依據。例如，升學面試時，評分者僅依應考者的履歷檔案就有了定見，忽略應考者在口試當下其他方面的表現，使部分的印象影響到全體。

在這種效應下，評分者對於受評者的評分可能會有所偏頗，陷入若受評者的某部分迎合自己的偏好，就認定這位受評者的其他方面也都表現優良，因而給予較高的評價。反之，若這位受評者的某部分表現不受評分者喜愛，則縱使其他部分的表現良好，評分者也可能忽略而給予較低的評價。也就是說，評分者因為自身偏頗的迷思而以偏概全，造成評分偏誤。

要克服這種偏誤，最主要就是要消除評分者的偏見。因此，可以採用分析式評分，事先設定各種評分項目，對受評者的各個項目分別評分，而非採用整體式評分，即依據受評者的表現給予單一的分數。這也是為何各類比賽或面試時，會事先擬訂計分項目與比重，並要求評分者分別依據各個項目評分，這對消除這種誤差有一定作用。此外，增加評分者的人數以及採用多階段的面試程序，並設計多重關卡，不僅可以在面試過程中多次觀察被評分者的表現，也可以避免月暈效應的影響。

趨中傾向

趨中傾向是指有些評分者可能不太熟悉評分準則，也可能因為受評者如考生或作品太多，無法精確判斷受評者的表現，因而給受評者的分數都集中在某一固定的範圍內變動，比較常見的是大多數的分數都集中在中間等級（平均值），而沒有顯著的好壞之別。以評等尺度 1 ～ 5 等級為例，趨中傾向的評分者所採用的評分策略，就是無論受評者的表現優劣，其評分都會落在中間值（3 級），而避免給予太低等級（1 ～ 2 級）或太高等級（4 ～ 5 級）。理論上，趨中傾向的謬誤又稱為「分數局限」。

克服這類偏誤，除了對評分者加強評分訓練外，也應該提供每個對應評分準則的說明與作品範本給評分者參考，以便了解與分辨不同準則之間的差異。此外，也可以結合定錨量表的使用，每一個分數旁邊有一些具體的表現水準說明，評分者的評分會較有依據，也能減少不同評分者之間的差異。

評分出現過寬或過嚴偏誤的原因，主要是由於評分者根據自己的經驗和能力，採取主觀的標準評估。

過寬或過嚴傾向

過寬是指有些評分者傾向給與受評者高估的成績，也就是不論受評者的表現如何，至少是 B 或 80 分以上，這類偏誤稱作慷慨的錯誤。過嚴則指有些評分者把受評者評定在較低的分數，表現再好最多也只給予 B 或 80 分。

要了解評分者對其所有受評者是否有過寬或過嚴的傾向，可以採用所有受評者被評分數的平均值，若這平均值遠高於中位數，則有過寬的問題，若其平均值遠低於中位數，則有過嚴的問題。

出現過寬或過嚴偏誤的原因，主要是由於評分者根據自己的經驗和能力，採取主觀的標準評估。克服這類評分誤差，除了對評估者建立其自信心，或給予角色互換的培訓外，還可以採用強制分配法，按照常態分配的比例來評估。

刻板印象

刻板印象是指個人對他人的看法往往受到其所屬社會團體的影響。其來自於對某一特定團體如民族、種族、省籍、宗教或性別等成員所形成的認知架構，由這架構對團體的其他成員形成過度簡化或以偏概全的意象，以致產生扭曲事實的認知。

常見的刻板印象例如身分地位高的人較溫文，身分地位低的人較粗野，戴眼鏡的人學識較淵博等。這樣的刻板印象也可能影響升學或職場面試，因此評分者必須小心謹慎，避免讓自己的偏好影響到對受評者表現的評分。

人為評分歷程建模

前述的人為評分偏誤都可能在評分歷程中發生，當然或許可以多增加評量的次數、增加評分者人數、加強評分者的訓練等，來減少這些偏誤的發生。對於一些非關鍵性的比賽或考試，或對公平性並沒有那麼嚴格要求的，這樣的處理方式並無不妥。

然而，如果這場考試或比賽結果攸關晉級或對學習、職涯有重大影響，那麼即使是些微的分數差距也可能影響重大。站在考生、參賽者或家長的角度，自然會想：自己（自己的孩子）會不會遇到給分較為嚴苛的評分者？或者是，即使所有考生都交由同一評分者評比，評分者是不是會維持一貫的評分準則評分？當評分過程出現了前面所提及的評分者偏誤時，對於受評者而言影響深遠。

心理計量專家 Linacre 於 1989 年提出的多相式模型，就是把考生能力、考題難度、評分者等因素同時考量所建構的統計模型。藉由這統計模型，分析並且偵測出評分者在評分歷程中，是否出現任何評分者偏誤以及所出現的評分偏誤類型。

目前，國內許多重要的考試或比賽已廣泛使用這個模型，如國民中學學生基本學力測驗寫作測驗、華語文口語能力測驗，以及運動競技比賽、創造力與想像力作品

刻板印象是指個人對他人的看法往往受到其所屬社會團體的影響，以致產生扭曲事實的認知。

不同評分者因為評分寬嚴程度不同造成評分結果的差異

比賽等。這些應用除了可以確保評分者的評分品質外，對於受試者實際能力的評估，也因為同步考慮了評分者因素的影響而使得評分更為精準。

再者，如果偵測出不容忽視的評分者偏誤時，也可了解造成評分者偏誤的原因。例如，若發現評分者出現趨中的評分傾向，則可能是評分者不清楚評分準則所致，這也顯示評分者的訓練不足。而透過這些分析所得的結果，可以使評分者的訓練更具效率。另外，當偵測出過嚴或過寬的評分傾向時，可以透過統計分析校正受評者的分數。

不論學校生活或就業職場，仰賴人為評分的時機甚多。隨著心理計量學者的努力，在評分過程中，評分者的反應與心理歷程也能透過建模的方式建構出來，並且運用在各種仰賴人為評分的場合。

洪素蘋
成功大學師資培育中心