測驗信度的基本原理
——古典測驗理論的觀點

李　克　明

壹、前　言

對於社會及行爲科學方面的問題，有越來越多的人，採用實徵的研究來加以探討。而在收集實徵的資料之前，研究者必須先決定或設計測量的工具，以便對於研究的變項賦予相當的值。但是相對於測量工具使用的日趨普遍，仍然有部份的研究者，對於測驗信度（ reliability ）的重要性，並未予以應有的關切。有的是使用他人編製的測量工具，而未必了解它的信度與測量結果的關聯；甚至有些測量工具的編製者，在提出編製經過的報告時未提供信度的資料。換句話說，受試者在測驗上表現的結果，與他所應該反映出的眞實情形，兩者之間有多少程度的關係，在部分測量工具的編製者或使用者的心目中，遠不及測驗本身的重要。然而，如果欠缺信度方面的訊息，事實上很難對測驗的結果，作出合理的解釋。

另外，國內有關心理測驗的書籍，關於信度的部分，至少有兩個地方，容易使讀者在觀念上造成混淆的，一是將信度解釋爲兩個測驗結果相符合的程度；另一個是以測量誤差與變異數解釋信度時，直接將眞實分數（ true score ）的變異數與誤差分數（ error score ）的變異數兩項的和，與實得分數（ observed score ）的變異數，兩者之間劃上等號。事實上，以上兩點如純就統計學的觀點來看，並不必然成立，必須要先滿足測驗理論上的一些假設。本文是以古典測驗理論的觀點，介紹信度的涵義，以及估計信度的推演過程，讓讀者對於測驗信度的基本原理，能有完整的概念。

貳、古典測驗理論的信度原理

古典測驗理論大約是在1950年代開始被提倡，也是最早的測量理論（ Suen, 1990 ）。它一向被視爲古典信度理論（ classical reliability theory ），因爲其理論的重心，是在估計一份測驗實得分數的可信度。近二、三十年來，雖然有其它理論，像是類化度（ generalizability ）、試題反應理論（ item response theory, 即IRT ）等陸續被提出，但是以古典測驗理論的觀點，來解釋信度的原理，至今仍然有它適用的價值。

信度通常被認爲受試者在一項測量工具上，表現出來的結果，其被測得的數量，可信度有多大。然而依據古典測驗理論，信度眞正的定義，是指測驗的實得分數與眞實分數之間關係的強度。因爲眞實分數永遠無法知道，爲了求得兩者之間的相關，所以必須在測驗的基本原理之下，對於信度的估計，作一番推演。其推演的過程，是以一個眞實分數模式（ true score model ）的數學模式作爲基礎。

一、眞實分數模式

如果我們可以在所有可能的條件下，於不同的時間，使用不同的可能的題目，對相同的受試者重複進行施測，就能夠得到該受試者許許多多的實得分數，這所有實得分數的平均數，就是這位受試者的眞實分數，因爲它是對該受試者的表現最沒有偏差的估計值。但是這只是理論上理想的假定，實際上根本無法做到，僅能對相同的受試者進行有限的施測。而當一位受試者接受一份測驗時，會有很多因素影響他的表現，這些因素包括不同的題目，不同的時間，以及不同的施測情境與個人的狀況等。所以任何單一施測的實得分數，通常會與眞實分數之間有差距。因爲只有在所有施測條件以及測量工具都是十全十美的情形下，實得分數才有可能和眞實分數完全相等。因此任何實得分數，都是由兩方面的成份所組成，即眞實的部份與誤差的部份。這個眞實分數模式，可以用下面公式來表示：

$$\mathrm{x} = \mathrm{t} + \mathrm{e} \qquad \text{（公式 1）}$$

其中 X 代表實得分數，t 代表眞實分數，e 則代表誤差分數。最後一項的誤差分數，是除了眞實分數以外，任何隨機誤差的因素所造成。誤差越小，表示實得分數與眞實分數越接近，因此測驗的分數就越可靠、越正確，也就是可信度高；反過來說，誤差越大，表示實得分數中眞實分數的部份越小、所以測驗分數的可度就越低。隨機誤差分數，時大時小，時正時負，如果是在重複施測的情形之下，因相互抵消的結果，其平均數會是零。而從公式 1 就可以導出下面的結果：

$$E(\mathrm{x}) = E(\mathrm{t}) + E(\mathrm{e})$$

其中 $E(\mathrm{x})$ 是實得分數的期望值或平均數，依次類推。而因爲 $E(\mathrm{t})=\mathrm{t}$ ，$E(\mathrm{e})=0$ ，所以 $E(\mathrm{x})=E(\mathrm{t})=\mathrm{t}$ 。這也就是前面提過的，對相同受試者連續不斷施測，其無數多的所有實得分數的平均數，就是該受試者的眞實分數。

二、信度的估計（ reliability estimation ）

前面提到，根據古典測驗理論，對於信度所下的定義，是指實得分數與眞實分數之間關係的強度，以皮爾遜（ Pearson ）積差相關來表示，就是 ρ_{xt} ，這個相關值就被稱爲信度指數（ reliability index ）（ Crocker & Algina, 1986 ）。兩者的關係越大，那麼實得分數就越能反映眞實分數。問題是眞實分數永遠是未知數，如果有人能得知眞實分數，那就用不著測量了，所以我們無法從實際施測的結果，直接估計 ρ_{xt} 的值。雖然如此，我們卻可以經由一些測驗理論上的假設，來估計這個相關值的平方。

三、獨立性假設（ assumption of independence ）

如果我們以：

$x = \mathrm{x} - \overline{\mathrm{x}}$（即實得分數的離均差）

$t = \mathrm{t} - \overline{\mathrm{t}}$

$e = \mathrm{e} - \overline{\mathrm{e}}$

另外又以：

E_{x}^{2} ＝實得分數變異數

E_{t}^{2} ＝眞實分數變異數

E_{e}^{2} ＝誤差分數變異數

E_{xt} ＝ x 和 t 之間的共變數

E_{te} ＝ t 和 e 之間的共變數

因爲 x 與 y 兩個變數的皮爾遜積差相關的通式爲：

$$\rho_{xy} = \frac{\text{X 和 Y 之間的共變數}}{(\text{X 的標準差})(\text{Y 的標準差})}$$

$$= \frac{\mathrm{E}xy}{\sqrt{\mathrm{E}x^2\ \mathrm{E}y^2}}$$

同樣的對於信度指數 ρ_{xt}，我們也可以寫成：

$$\rho_{xt} = \frac{\mathrm{E}xt}{\sqrt{\mathrm{E}x^2\ \mathrm{E}t^2}} \text{（公式 2）}$$

從公式 1 的眞實分數模式，可以將公式 2 改寫成：

$$\rho_{xt} = \frac{\mathrm{E}(t+e)t}{\sqrt{\mathrm{E}x^2\ \mathrm{E}t^2}} + \frac{\mathrm{E}t^2 + \mathrm{E}te}{\sqrt{\mathrm{E}x^2\ \mathrm{E}t^2}} \text{（公式 3）}$$

獨立性假設的意義，是指誤差分數與眞實分數之間無關，它的意思是，既然誤差的量，是由於隨機因素所造成，那麼一位受試者在某一特定測驗的眞實分數，應該與接受該測驗時的機遇誤差，彼此相互獨立。而當兩個變數之間無關或彼此相互獨立時，這兩個變數就沒有共變數。現在所討論的例子來說 $\mathrm{E}te = 0$。根據這個假設，公式 3 如果以信度指數的平方來表示，就可以寫成：

$$\rho^2_{xt} = \frac{(\mathrm{E}t^2 + \mathrm{E}te)^2}{\mathrm{E}x^2\ \mathrm{E}t^2}$$

$$= \frac{(\mathrm{E}t^2)^2}{\mathrm{E}x^2\ \mathrm{E}t^2} \text{（因爲 } \mathrm{E}te = 0\text{）}$$

$$= \frac{\mathrm{E}t^2}{\mathrm{E}x^2}$$

$$= \frac{\text{眞實分數變異數}}{\text{實得分數變異數}} \text{（公式 4）}$$

這個信度指數的平方，也就是一般通稱的信度係數（reliability coefficient）。它所指的是實得分數變異數中，眞實分數變異數所佔的比例。

從公式 1，我們還可以將實得分數的變異數寫成：

$$\mathrm{E}x^2 = \mathrm{E}(t+e)^2$$

$$= \mathrm{E}t^2 + 2\mathrm{E}te + \mathrm{E}e^2$$

因爲 $\mathrm{E}te = 0$（獨立性假設），所以

$$\mathrm{E}x^2 = et^2 + \mathrm{E}e^2 \text{（公式 5）}$$

也就是：實得分數變異數＝眞實分數變異數＋誤差分數變異數。可知實得分數變異數，實際上由兩個變異的成份所構成，即眞實的成份與誤差的成份。

如果把公式 5 等號兩邊都除以 $\mathrm{E}x^2$，則成爲：

$$\frac{\mathrm{E}x^2}{\mathrm{E}x^2} = \frac{\mathrm{E}t^2}{\mathrm{E}x^2} + \frac{\mathrm{E}e^2}{\mathrm{E}x^2}$$

$$1 = \frac{\mathrm{E}t^2}{\mathrm{E}x^2} + \frac{\mathrm{E}e^2}{\mathrm{E}x^2}$$

又從公式 4 的 $\rho^2{}_{xt} = \frac{El^2}{Ex^2}$，可知信度係數

$$\rho^2{}_{xt} = 1 - \frac{Ee^2}{Ex^2} \quad \text{（公式 6）}$$

從公式 6，我們可以知道，如果把實得分數的變異數當作 1，則信度就是 1 減去誤差分數變異數在實得分數變異數中的比例。因此，當誤差越小時，信度越高；誤差越大時，信度越低。

可是到此爲止，我們仍然無法從公式 4 或公式 6 得到信度係數的大小，因爲不論眞實分數變異數或誤差分數變異數，都還是未知數。不過只要再經由平行測驗假設（ parallel tests assumptions ），我們便可估計出 $\rho^2{}_{xt}$ 的值。

四、平行測驗假設

如果我們能獲得相同受試者在測驗 A 和測驗 B 兩份測驗上的分數，同時又假定這兩份測驗都是用來測量相同的表現。那麼每位受試者在這兩份測驗上，應該有相同的眞實分數，也就是眞正的分數不會改變。而這兩份測驗實得分數的皮爾遜積差相關可寫成：

$$\gamma_{AB} = \frac{Ex_A x_B}{\sqrt{Ex_A{}^2 \; Ex_B{}^2}}$$

$$= \frac{E(l + e_A)(l + e_B)}{\sqrt{Ex_A{}^2 \; Ex_B{}^2}}$$

$$= \frac{El^2 + Ele_A + Ele_B + Ee_A e_B}{\sqrt{Ex_A{}^2 \; Ex_B{}^2}}$$

由於前面提過的獨立性假設，上面公式分子中的 Ele_A 和 Ele_B 都等於零，所以又可寫成：

$$\gamma_{AB} = \frac{El^2 + Ee_A e_B}{\sqrt{Ex_A{}^2 \; Ex_B{}^2}} \quad \text{（公式 7）}$$

如果測驗 A 和測驗 B 能符合平行測驗假設，上面這個相關便可用來估計信度係數。所謂平行測驗假設，主要包括兩點：(1)測驗 A 與測驗 B 的分數，有相同變異數，即 $Ex_A{}^2 = Ex_B{}^2$；(2)測驗 A 與測驗 B 的誤差，彼此相互獨立或無關，即 $Ee_A e_B = 0$。基於這個假設，公式 7 便可寫成：

$$\gamma_{AB} = \frac{El^2}{Ex_2} = \rho^2{}_{xt} = \text{信度係數（公式 8）}$$

換句話說，如果能確認兩份測驗確實符合平行測驗假設，那麼這兩個測驗實得分數的皮爾遜積差相關係數，就是實得分數與眞實分數之間相關係數的平方，也就等於信度係數。

五、獲得平行測驗的策略

爲了能夠從實得的測驗分數去估計信度，必須設法獲得對受試者測量相同表現的兩個測驗分數，也就是必須確保兩份測驗符合平行測驗假設。而所採用的方式，一般常見的有以下幾種：

(1)重測（ test-retest ）的方式，就是在兩次不同的時間，使用同一份測驗，對相同的受試者施測。由於這是假定同樣一份測驗，在不同時間得到的兩個測驗分數，仍舊包含相同的眞實分數，所以這種信度又稱穩定係數（ coefficient of stability ）。

(2)編製兩份同等版本（ equivalent versions ）的測驗，使這兩份版本儘可能相似到可被視爲同一份測驗，所以這種信度又稱爲等值係數（ coefficient of equivalence ）。

(3)上面的兩種方式都是要對相同的受試者施測兩次，較經濟的方式是採用折半（ split-half ）的方法，

就是只編製一份測驗，但是把題目折開成兩半，然後將兩半測驗的分數視同兩份同等版本測驗的分數。這是假定從一份測驗分開成同等的兩半，仍然符合平行測驗假設，所以這類信度又稱爲內部一致係數（coefficient of internal consistency）。

由於一份測驗分成兩半，而求得兩個測驗分數的相關係數，只是半個測驗的信度。在其他條件都相等的情況下，一份測驗的題數越多，其可信度就越大。爲了避免低估折半信度，從折半法所計算的相關係數，要再以斯布公式（Spearman Brown Formula）加以校正（Anastasi, 1988）。斯布公式的通則是，如果測驗的題數增加或減少爲原測驗的K倍，那麼從原測驗信度 ρ^2xt 估計新測驗信度 ρ^2xt* 是

$$\rho^2xt* = \frac{K\rho^2xt}{1+(K-1)\rho^2xt}$$

因爲折半法原測驗題數是折半後的兩倍，所以應用斯布公式來校正折半信度，便可簡化爲：

$$\rho^2xt* = \frac{2\rho^2xt}{1+\rho^2xt}$$

公式中的 ρ^2xt 是兩半測驗分數之間的相關係數。

採用折半的方式，另外面臨的一個問題是，將一份測驗分成兩半，可以有許多種分法。每種分法都會導致不同的折半信度，到底那一個才是信度係數最佳的估計？解決這個爭論的方式之一是，將一份測驗拆成以單一試題爲單位，也就是把一份有K個試題的測驗，視同有K份平行的單一試題的測驗。這些所有可能成對題目之間相關係數的平均值（$\overline{\rho^2xt}$），將會是單一試題信度的最佳估計值。這個相關係數的平均值，再以斯布公式校正，便可得到整份測驗的信度係數，也就是：

$$\rho^2xt* = \frac{K\overline{\rho^2xt}}{1+(K-1)\overline{\rho^2xt}}$$

其他還有以相同的原理爲基礎，將信度係數的估計加以簡化的方法，包括克隆巴赫 α（Cronbach Alpha）、庫李20（Kuder-Richardson Formula-20，或KR-20）、及庫李21（Kuder-Richardson Formula-21，或KR-21）等，讀者可自行參閱國內有關心理測驗的書籍，本文不加贅述。

參、結　語

以上的內容，可以歸納兩個要點：(1)信度眞正的定義，是指測驗的實得分數與眞實分數之間相關的程度；(2)這兩個分數的相關，以信度係數來表示時，事實上已經是其相關係數的平方。這也就是爲什麼以測量的標準誤（standard error of measurement, 即SEM）解釋個別的分數時，要用下面的公式（Anastasi, 1988）：

$$SEM = SD_l\sqrt{1-\gamma_{ll}}$$

根號裡的 γ_{ll} 即信度係數，具有平方的實質意義，所以要開根號。

另外還有兩點須要補充說明的：(1)信度雖然是藉由兩個測驗分數的相關加以推算，而統計上相關是可以有負數，但是因爲信度係數事實上得分數與眞實分數之間相關係數的平方，所以負的信係數是沒有意義的。如果信度係數出現負數的值，表示已經違反平行測驗假設，也就是兩個實得測驗分數，事實上不能被用來估計信度；(2)雖然獲得平行測驗的策略有好幾種方法，但是不論採用那一種，都要有理論上的依據。任何一種測驗，有其特定的功能及目的，因此其滿足平行測驗假設的策略，應該只有一種，所

以理論上，一個測量工具不能同時有兩個以上的信度係數。如果估計出兩個不同的信度係數，反而在解釋上無法自圓其說，而造成困擾。

參考書目：

Anatasi, A.(1988). Psychological testing (6th ed.). New York:Macmillian.

Crocker, L, & Algina.J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Suen H.K. (1990). Principles of test theories. Hillsdale, NJ: Lawrence Erlbaum Associates.