國立臺南大學
「教育研究學報」第 54 卷第 1 期（民國 109 年）：1~ 26
DOI： 10.3966/199044282020045401001

Alpha 係數及相關的信度估計方法探討

涂柏原

國立臺南大學教育學系 副教授

中文摘要

自從 Spearman 於 1904 年發表兩篇日後成為古典測驗理論和因素分析基礎的研究報告之後，古典測驗理論（classical test theory, CTT）逐漸成形，信度的概念和理論也被發展成熟。在眾多信度估計方法中，最負盛名的是 α 係數（Cronbach, 1951）。因為運用 α 係數來報導信度的學者中，未必人人皆都熟悉 α 的原理及優劣處，因此本文針對信度的理論與 α 係數的原理與計算，做了深入的說明；本文也介紹可以與 α 係數一起並用的信度估計方法，然後呈現學者針對這些方法的所進行的比較研究結果，主要結論是對於測驗真實信度的估計，在某些情境下，α 的表現不佳，最好在報告測驗資料的 α 係數時，能同時輔以表現更好的估計值，如 ω_t 或 λ_4 等。

關鍵字：Guttman 的下限、α 係數、ω_t、信度

通訊作者：涂柏原，email: bortwu@gmail.com

壹、緒論

自從 Spearman 於 1904 年發表兩篇日後成為古典測驗理論和因素分析基礎的研究報告之後，古典測驗理論（classical test theory, CTT）逐漸成形，信度的概念和理論也逐漸被發展起來。任何以測量結果為基礎的研究，都必須考慮測量之精確性、可靠性或是信度（Cronbach, 1951）。要知道一個研究所用的測量是有多精確，最理想的是進行兩個獨立的測量，然後比較由二者所得到的結果，然而，在現實上，心理學家和教育學者很難有機會讓他們的受試者進行第二次的測驗（Cronbach, 1951）。因此，在整個心理計量學（或是測驗理論）的發展歷史中，利用單次施測的結果來估計信度的方法屢被提出（例如，Kuder & Richardson, 1937; Guttman, 1945; McDonald, 1978 等）。

在二十世紀初期，當 Spearman 瞭解心理學家需要評估他們所使用測量工具之精確性時，依據他自己研究上的需要，Spearman（1904）提出了折半信度（split-halves reliability）的估計方法，將一個測驗分成奇、偶數題兩個部分之後，每一個受試者在某一個測驗上面的觀察分數就可分割成兩個，計算二者之間的相關，經過 Spearman-Brown 校正公式之後，即可得到折半信度（Cronbach, 2004）。這個折半信度係數的估計方法雖然簡單，但是受到一些批評，雖然題數長短對於信度的影響可以利用 Spearman-Brown 校正公式來加以解決，而且，Rulon（1939）和 Guttman（1945）也都發展出不受題數影響的折半信度係數的估計方法。然而，一個測驗可以用很多的方式來要拆成兩半，而且每一種折半的方式所得到的信度係數不一定相同，要用哪一個值來代表測驗的信度，這個問題卻是不易解決。

在討論兩個變項之間的關聯時，Spearman（1904）特別介紹了由 Galton 所提出來並由 Pearson 改良的相關係數，並討論了影響相關的一些誤差以及如何進行相關減縮校正（correction for attenuation）。在那個時候，英國的心理學研究者特別受到 Darwin 的自然選擇理論的影響，因此人與人之間的個別差異就成為心理學家研究的焦點。當個別差異被測量時，測量的精確度經常被檢視，因此結果的報導幾乎都是以信度係數的型態出現（Cronbach, 2004）；而這種信度係數與積差相關類似，其值在 0 ～ 1 之間。

在那時候，信度的定義為何，也引起一些反覆出現的爭辯（Cronbach, 2004, p. 394）。雖然大家都知道所關心的是由一個測量到另一個測量之間的一致性，當 Spearman(1904) 提出信度這個概念時，是利用折半的方式來提供一個代表信度的數值，Traub（1997）提到 Brown 在 1910 年使用信度係數這個名詞來指稱同一群人在同一個測驗兩次施測的分數之相關，對於 Brown 的做法，Kelly（1923）提出了批評，他認為信度是由兩個 “可加比較的” 測驗之分數之相關係數所定義的，因此其他的定義似乎都是不可接受的（引自 Traub, 1997）。甚至當 Kuder 和 Richardson（1937）著名的 KR20 公式被提出來時，Kelly（1942）仍然認為信度的概念的必要條件是對於一個心理功能必須有兩個或是更多的測量。Cronbach（2004）認為無論信度估計公式是如何被推導得到，只要計算的方式並未與定義對齊的話，那些信度的估計方法所得到的都僅是個近似值而已。

所謂的內部一致性信度估計方法中，除了最早見到的折半信度以外，Kuder 和 Richardson（1937）發展了 KR20 這個公式，讓研究者可以僅將測驗工具施測一次，即可估計信度係數。KR20 的公式形式限制它僅能應用在二元計分試題上面，隨後 Cronbach（1951）提出 α 係數將此限制解除，可以應用在多元計分試題的資料上面。在 α 係數被提出來之後，可能是因為不必施測兩次（同一份測驗或是不同的測驗），而且是幾乎每一本測驗或是教學評量教科書必定提到的，以致於 α 係數成為最受歡迎的信度估計方法。無論研究者熟悉或不熟悉其原理，幾乎在所有與測驗有關的研究中，α 係數皆被用來作為測驗信度的指標，Cronbach（1951）的研究報告，已經被引用超過 6500 次（Sijtsma, 2009a），幾乎每年被引用 325 次左右（Cronbach, 2004）。截至 2014 年 10 份，Cronbach 的文章在 Google Scholar 中成為最常被蒐集的文獻中，排名第 64 名，且在心理學這個領域中，排名第 3 名（McNeish, 2018）。

然而，α 係數被如此廣泛地應用著，並不代表使用 α 係數來報導測驗分數信度的研究者，都知道 α 係數的適用範圍與限制。就如 Cronbach（2004, p. 392）所說的，有引用該文的未必真正讀過它，甚至連看過都沒有，以致於時常有誤用的情形，這種情形也存在國內的研究者中。針對 α 係數的一些限制，許多學者提出了一些他們認為比 α 係數更好的信度估計方法，比如，分層的 α 係數（α_{strat}; Cronbach, Schoneman & Mckie, 1965）、β 係數（Revelle, 1979）、ω_t（McDonald, 1999）、ω_h（McDonald, 1978, 1999）、Θ 係數（Armor, 1974）；還有利用結構方程模式（structural equation modeling, SEM）估計信度的方法（例如，Bentler, 2009；Green & Yang, 2009b；McDonald, 1999；Raykov & Shrout, 2002）。這些信度估計的方法，雖然被提出來已有一段時日，然而實務研究中，似乎仍然未被廣泛採用。因此，本文的目的是希望對於 α 係數提供一個完整的描述，對於 α 的一些問題進行文獻探討，希望能夠加以澄清部分的問題；並對 Guttman（1945）所提出的六個信度估計的下限（lower bound）以及上面所提到的 α_{strat}、β、ω_t、ω_h、Θ 等這幾個信度估計方法進行理論的介紹，期望讓國內的研究者在報告他們所使用的工具之信度資料時，除了 α 係數以外，能夠選用更合適的方法來得到他們的信度證據。除了這些信度估計方法以外，還有最大信度（maximal reliability; Bentler, 2007）、H 係數（Coefficient H; Hancock & Muller, 2001）、最大下限（greatest lower bound; Jackson & Agunwamba, 1977）等一些有用的信度估計方法，限於篇幅，本文不將它們納入討論。

貳、信度的定義及估計方法

令 X 代表一位考生在某一個測驗上面的觀察分數，根據古典測驗理論，$X=T+E$，其中 T 是真分數（true score），E 是隨機誤差分數（error score）。真分數 T 是未知的，一般假設為同一份考卷讓同一群人作答 η 次，每一個人在這 η 個分數之期望值就是該受試者之真分數（亦即，$\varepsilon(X)=T$）。根據定義，T 和 E 被假設為是彼此獨立的（亦即，$\rho_{ET}=0$），觀察分數的變異數等於真分數變異數與誤差變異數之和（亦即 $\sigma_X^2=\sigma_T^2+\sigma_E^2$），而理論的信度係數 $\rho_{XX'}$ 就定義為真分數變異數與觀察

分數變異數之比值（σ_T^2/σ_X^2）。然而，如果缺乏一些適當的假定，古典測驗理論是沒有辦法提供我們更多的訊息。

如果 X_1 和 X_2 是根據相同的雙向細目表所編製而成的兩個測驗之原始分數，根據古典測驗理論，$X_1=T_1+E_1$ 且 $X_2=T_2+E_2$，因為 $\rho_{ET}=0$，所以在此進一步假定不同式測驗之間的真分數與誤差分數是獨立的，也就是 $\rho_{E_1T_2}=0$ 、$\rho_{E_2T_1}=0$ 和 $\rho_{E_1E_2}=0$。對 X_1 和 X_2 這兩個測驗來說，如果 $T_1=T_2$、$\sigma_{E_1}^2=\sigma_{E_2}^2$，則它們被稱為「平行測驗」或是「嚴格的平行測驗」（strictly parallel test; Lord, 1980）。這些假定（assumption）在 Allen 和 Yen（1979）與 Crocker 和 Algina（1986）中皆有詳細的描述。

以嚴格的平行測驗為例，研究者常以受試者在兩個平行測驗上得分（X_1 和 X_2）之相關作為測驗的信度估計值，

$$\rho_{X_1X_2}=\frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}=\frac{\sigma_{(T_1+E_1)(T_2+E_2)}}{\sigma_{X_1}\sigma_{X_2}}=\frac{\sigma_{(T_1T_2+E_1T_2+E_2T_1+E_1E_2)}}{\sigma_{X_1}\sigma_{X_2}}$$

$$=\frac{\sigma_{T_1T_2}+\sigma_{E_1T_2}+\sigma_{E_2T_1}+\sigma_{E_1E_2}}{\sigma_{X_1}\sigma_{X_2}}=\frac{\sigma_{T_1T_2}}{\sigma_{X_1}\sigma_{X_2}}=\frac{\sigma_T^2}{\sigma_X^2}=\rho_{XX'}$$

上述公式中的 $\sigma_{E_1T_2}=\sigma_{E_2T_1}=\sigma_{E_1E_2}=0$，$\sigma_{X_1}=\sigma_{X_2}=\sigma_X$，$\sigma_{T_1T_2}=\sigma_{T_1T_1}=\sigma_{T_2T_2}=\sigma_T^2$，因此最後的 σ_T^2/σ_X^2 可以得到；也就是，考生在平行測驗的得分之間的相關，等於理論上的信度係數 $\rho_{XX'}$ 的值，這是利用平行測驗的概念來估計信度的方法之基礎。如果將 X_1 和 X_2 視為是一個測驗之前後測分數，或是分成兩個半測驗之後各個半測驗的分數，則上述的公式亦可作為再測信度以及折半信度的理論基礎，只是所得到的若是折半信度係數，則該信度值不是原來測驗的長度應有的結果，得再用 Spearman-Brown 校正公式來加以校正。

採用複本信度估計方法，不但要施測兩次，也得編製兩份測驗；若是採用再測信度估計方法，雖然不需要編製兩份等值的測驗，仍然得施測兩次。Cronbach（1951）提到在現實上心理學家和教育學者很難有機會讓他們的受試者進行第二次的測驗，因此要如何僅根據一次測驗施測的結果來估計信度係數，成為學者們努力的目標。

Kuder 和 Richardson（1937）提出了著名的 KR20 公式，讓測驗實務中的研究者可以僅根據一次測量的結果，估計測驗的信度係數；利用這個公式所得到的信度，一般被認為是一種內部一致性（internal consistency）的信度係數：

$$KR20=\frac{k}{k-1}\left(\frac{\sigma_X^2-\sum p_iq_i}{\sigma_X^2}\right) \qquad (1)$$

其中，κ 是題數，σ_X^2 為全測驗分數之變異數，p_i 是所有的受試者答對第 i 題的人數

比率，而 q_i 是所有的受試者答錯第 i 題的人數比率（ $q_i = 1 - p_i$ ）。因為 KR20 的公式表徵方式僅適合二元計分試題，對於多元計分資料適用的公式，仍然有待發展；但是由 KR20 的公式可以理解，只要將 $\sum p_i q_i$ 改變成一般試題也可以接受的形式，應該就可以找到合適的公式表徵；所以在 1940 年代起，就有許多適用於多元計分資料的信度估計公式被提出來，比如 Guttman（1945）的六個信度下限。

參、不同的平行程度之定義

如前一節所陳述的，若 X_1 和 X_2 分別代表由相同的雙向細目表所編製而成的兩份測驗之原始分數，如果 $T_1 = T_2$、 $\sigma^2_{E_1} = \sigma^2_{E_2}$ ，因此 $\sigma^2_{X_1} = \sigma^2_{X_2}$ ，則 X_1 和 X_2 這二個測驗是「嚴格的平行測驗」（strictly parallel tests），或是「平行測驗」（Lord, 1980）。嚴格的平行測驗之假定，在測驗實務工作中，實在不容易被滿足，因此從 1950 年代之後，要求較寬鬆的平行測驗之定義（或概念）紛紛被提出來。

Feldt 和 Brennan（1989, pp. 110-111）介紹了五個不同的平行測量，包括古典平行模式（classical model）、τ 等值測驗（tau-equivalent forms; Lord & Novick, 1968）、本質的 τ 等值測驗（essentially tau-equivalent forms; Lord & Novick, 1968）、同因素測驗（congeneric forms; Jöreskog, 1971）及多因素的同因素測驗（multi-factor congeneric form）。表 1 摘要呈現常見的平行測驗之定義。

表 1

各種程度之平行測驗

I. 古典平行題本或部分（Classical Parallel Forms or Parts）

A. 內容相似性

B. T_i 在各個題本或部分中是一樣不變的

C. $\mu_{X_1} = \mu_{X_2} = \mu_{X_3} = \cdots *$

D. $\sigma^2_{X_1} = \sigma^2_{X_2} = \sigma^2_{X_3} = \cdots *$

E. $\sigma_{X_1X_2} = \sigma_{X_1X_3} = \sigma_{X_2X_3} = \cdots *$

F. $\sigma_{X_1Y} = \sigma_{X_2Y} = \sigma_{X_3Y} = \cdots *$

II. τ 等值題本或部分（Tau-equivalent Forms or Parts）

A. 內容相似性

B. T_i 在各個題本或部分中是一樣不變的

C. $\mu_{X_1} = \mu_{X_2} = \mu_{X_3} = \cdots *$

D. $\sigma^2_{X_1} \neq \sigma^2_{X_2} \neq \sigma^2_{X_3} \neq \cdots *$（因為 $\sigma^2_{E_g} \neq \sigma^2_{E_h}$ ）

E. $\sigma_{X_1X_2} = \sigma_{X_1X_3} = \sigma_{X_2X_3} = \cdots *$（因為 $\mathrm{T}_{ig} = \mathrm{T}_{ih}$ ）

F. $\sigma_{X_1Y} = \sigma_{X_2Y} = \sigma_{X_3Y} = \cdots *$（因為 $\mathrm{T}_{ig} = \mathrm{T}_{ih}$ ）

III. 本質地 τ 等值題本或部分（Essentially tau-equivalent Forms or Parts）

A. 內容相似性

B. $\mathrm{T}_{ig} = \mathrm{T}_{ih} + c_{gh}$（不是所有的 $c_{gh} = 0$ ）

C. $\mu_{X_1} \neq \mu_{X_2} \neq \mu_{X_3} \neq \cdots *$

D. $\sigma^2_{X_1} \neq \sigma^2_{X_2} \neq \sigma^2_{X_3} \neq \cdots *$（因為 $\sigma^2_{E_g} \neq \sigma^2_{E_h}$ ）

E. $\sigma_{X_1X_2} = \sigma_{X_1X_3} = \sigma_{X_2X_3} = \cdots *$（因為 c_{gh} 不影響 $\sigma_{X_gX_h}$ ）

F. $\sigma_{X_1Y} = \sigma_{X_2Y} = \sigma_{X_3Y} = \cdots *$（因為 c_{gh} 不影響 $\sigma_{X_gX_h}$ ）

IV. 同因素的題本或部分（Congeneric Parts or Forms

A. 內容相似性

B. $\mathrm{T}_{ig} = b_{gh}\mathrm{T}_{ih} + c_{gh}$（不是全部的 $b_{gh} = 1.0$, 不是全部的 $c_{gh} = 0$ ）

C. $\mu_{X_1} \neq \mu_{X_2} \neq \mu_{X_3} \neq \cdots *$

D. $\sigma^2_{X_1} \neq \sigma^2_{X_2} \neq \sigma^2_{X_3} \neq \cdots *$

E. $\sigma_{X_1X_2} \neq \sigma_{X_1X_3} \neq \sigma_{X_2X_3} \neq \cdots *$

F. $\sigma_{X_1Y} \neq \sigma_{X_2Y} \neq \sigma_{X_3Y} \neq \cdots *$

註：本表摘要整理自 “Reliability” by Feldt, L. S., & Brennan, R. L. (1989). In R. L. Linn (Ed.), Educational measurement(3rd ed.)(pp. 105-146). New York, NY: American Council on Education and Macmillan. * 表示可觀察得到的關係。

古典測驗理論（CTT）與因素分析（factor analysis, FA）有密切的關係存在（見 Bollen, 1989; Lord & Novick, 1968; McDonald, 1999; Raykov & Marcoulides, 2011），單一的觀察指標變項 X 在 CTT 中是真分數 T 與誤差分數 E 之和，在 FA 中，可以寫成 $\lambda F+\delta$，也就是 $X=T+E=\lambda F+\delta$；其中 F 為共同因素，δ 為誤差因素，λ 是因素負荷量。在圖 1 中，(a) 的部分為 CTT 的表示方式，e 即是誤差 E，(b) 的部分為 FA 的表示方式，其中 d 是 δ，所以 CTT 中的真分數 $T=\lambda F$，而 δ 即為 CTT 中的 E。因此，信度係數 $\rho_{XX'}=\sigma_T^2/\sigma_X^2=\lambda^2\sigma_F^2/\sigma_X^2$，亦即，信度係數可以利用因素分析所得到的結果來計算，在這裡的觀點是真分數的成分僅包括共同因素而已。

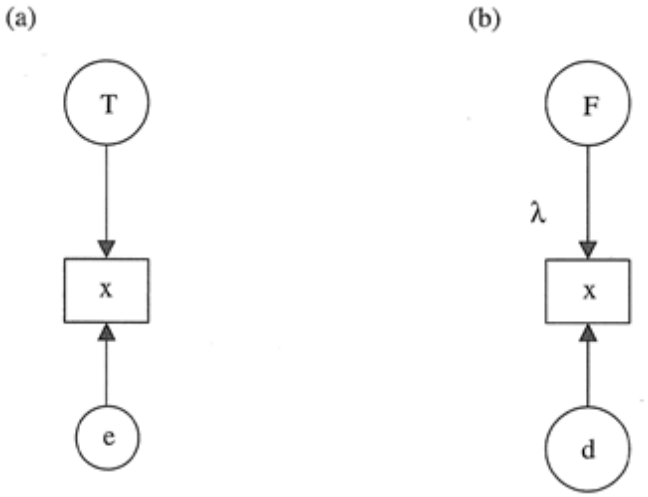

圖 1. CTT 與因素模式

資料來源：取自“Scaling procedures: Issues and applications” by Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). Thousand Oaks, CA: Sage Publications.

表 1 所介紹的五個不同的平行測量中，較常被討論的是古典平行測驗、本質的 τ 等值測驗及同因素測驗三者。雖然由表 1 的內容來看，τ 等值與本質的 τ 等值測驗是有差異的，但是有些學者將二者視為是一樣的（如 Osburn, 2000, p. 345），Graham（2006）提到本質的 τ 等值中之真分數比起 τ 等值多了一個加法常數（additive constant），僅影響了一個試題的平均數，但不會影響其變異數或是與其他試題之共變數，而信度是一個變異數被解釋比率的統計數，不受平均數影響，因此對於估計信度這個目的，本質的 τ 等值的 SEM 路徑圖與 τ 等值的是模式是相同的（p. 935）。由因素分析的觀點，這三種的平行測驗的內涵可以重新表述如下：
（蔡佩圜、冷柏原、吳裕益，2017, 2019；Osburn, 2000）

（一）平行測驗（parallel test）：是指測量同一個構念的 κ 個試題滿足：(1) 各試題因素負荷量相等；(2) 各試題誤差變異數相等；(3) 各試題截距項相等。在滿足平行測驗條件情況下，α 等於 $\rho_{XX'}$。

（二）τ 等值測驗（tau-equivalent test）：若測驗試題之因素負荷量均相等，但誤差變異數不相等即屬於 τ 等值測驗類型。由於因素負荷量相等，因此各觀察分數之變異中屬於真分數變異（因素負荷量之平方乘以真分數之變異數）之部分均相等，但由於誤差變異數不同，因此各個平行測驗觀察分數之總變異數也就不同，導致各測驗之信度也不同。

（三）同因素測驗（congeneric test）：當所有試題均測量同一個構念，但其因素負荷量不相等時，即屬於同因素測驗或同源測驗。

不同平行程度測驗的定義對於信度估計有何影響呢？實務工作者為何需要在意這些呢？ Feldt 和 Brennan（1989）用了一些篇幅介紹了根據不同平行定義所發展出來的信度係數（或信度估計方法），並整理在該文的表 3.1 中（p. 115）中。一般研究者比較熟悉的利用 Spearman-Brown 校正公式所得到的折半信度是以平行測驗為基礎的，KR20 和 α 係數的理論基礎是本質地 τ 等值測驗，以同因素測驗為基礎的 CTT 信度估計方法（例如，Bentler, 1985; Kristof, 1974; Raju, 1977; Gilman & Feldt, 1983; Jöreskog & Sörbom, 1985 等），則很少在實徵研究中看到的。而由各個碩、博士論文或各領域期刊所刊登的論文來看，在信度證據方面，大都僅報告 係數來看，可以瞭解一般實徵研究者在計算其測驗工具的信度時，可能很少檢查信度估計方法的假定是否被滿足。

肆、Alpha 係數以及分層 Alpha

根據文章被發表的時間以及其背後的原理來看，Cronbach（1951）的 α 是 KR20 的更一般化公式，但是一般的教科書常為了呈現的方便，會先介紹 α 係數，然後將 KR20 視為 α 的特例，這種觀點其實也沒有什麼不對之處，但是容易造成讀者對於信度係數的發展的歷史脈絡失去全觀。

一般來說，α 係數之公式如下：

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum\sigma_i^2}{\sigma_X^2}\right) \qquad (2)$$

其中 κ 是題數，σ_X^2 是受試者的觀察分數之變異數，σ_i^2 是第 i 題的變異數。這個公式其實不是 Cronbach 所發明創新的，Gulliksen (1950/1987, p. 223) 的公式 10 以及 Guttman（1945, p. 259）的 λ_3 等皆與 α 係數是相同的。在 Cronbach（2004, p. 397）的文章中，Cronbach 本人也清楚地提到這個部分，他甚至為 α 係數被稱為 Cronbach' s α 感到難為情。

因為公式（2）的計算是利用某一個測驗題本各個試題間的共變數矩陣之元素，如果將該共變數矩陣加以標準化，則共變數矩陣變成了相關矩陣。在這個情形下，相關矩陣主軸以外的元素共有 $\kappa(\kappa-1)$ 項，若這 $\kappa(\kappa-1)$ 個相關係數的平均數等於 $\bar{r}$，因為相關矩陣主對角線的元素值皆為 1，因此 $\sum\sigma_i^2 = k$，所以 $\sigma_X^2 = \sum\sigma_i^2 + \sum\sum_{i\neq j}\sigma_{ij} = k + k(k-1)\bar{r}$，因此公式（2）就變成

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum\sigma_i^2}{\sigma_X^2}\right) = \frac{k}{k-1}\left(1-\frac{k}{k+k(k-1)\bar{r}}\right) = \frac{k}{k-1}\left(\frac{k+k(k-1)\bar{r}-k}{k+k(k-1)\bar{r}}\right)$$

$$= \frac{k}{k-1}\left(\frac{k(k-1)\bar{r}}{k+k(k-1)\bar{r}}\right) = \frac{k}{k-1}\left(\frac{(k-1)\bar{r}}{1+(k-1)\bar{r}}\right) = \frac{k\bar{r}}{1+(k-1)\bar{r}} \quad (3)$$

這就是所謂的標準化試題 α 係數（standardized item alpha）（Cortina, 1993; Netemeyer, Bearden, & Sharma, 2003; Osburn, 2000），或簡稱為標準化 α 係數，其公式形式恰好與 Spearman-Brown 的校正公式相同。如果試題原始分數被加總作為一個量表的分數，Cortina（1993）認為在此情境中，標準化的 α 係數不適合作為信度的估計值，因為試題分數變異數的差異會影響了總分。

當一個測驗總分（或稱為組合分數 ; composite）之成分可以根據內容或是難度分成數個分測驗時，分層 α 係數（stratified alpha）可能是一個比 α 係數來得好的信度估計值（Cronbach, Schoneman, & McKie, 1965）。然而，在某些情形下，由同一筆資料所計算得到的分層 α 係數之值，可能會低於 α 係數的值（Osburn, 2000）。

分層 α 係數的計算公式為（Feldt & Brennan, 1989）：

$$\alpha_{stra} = 1-\sum\left[\sigma_i^2(1-\alpha_i)/\sigma_X^2\right] \quad (4)$$

其中 α_i 是第 i 個分測驗的 α 係數，σ_i^2 是第 i 個分測驗之變異數。基本上這個公式是利用 1 減去誤差變異數與觀察分數變異數之比值來得到的，而全測驗的誤差變異數等於各個分測驗誤差變異數之和，所以 $\sigma_E^2 = \sum\sigma_{E_i}^2$，而 $\sigma_{E_i}^2 = \sigma_i^2(1-\alpha_i)$。如果同一個分測驗內各個試題之間的相關是高的，而不同分測驗之試題間的相關是低的，則分層 α 係數是一個比 α 係數還要好的估計數（Osburn, 2000）。

伍、Alpha 的涵意

有許多作者試圖對於 α 係數作一個介紹，無論是很仔細的或是提供一個整體的宏觀。Cortina（1993）曾歸納了文獻中對於 α 係數的描述，得到下列五個陳述：(a) α 是所有折半信度的平均數（Cronbach, 1951）；(b) α 是一個測驗的信度係數之下限（Kristoff, 1974; Novick & Lewis, 1967; Ten Berge & Zegers, 1978）；(c) α 是第一個因素飽和度的一個測量（Crano & Brewer, 1973; Hattie, 1985）；(d) 當「本質的 τ 等值」（essentially τ -equivalent）被滿足時，亦即各個試題彼此之間是具有本質的 τ 等值之平行試題時，α 的值與真正的信度係數值是相等的（Kristoff, 1974; Novick & Lewis, 1967; Ten Berge & Zegers, 1978）；(e) α 是 KR20 等值係數之一般化版本（Cronbach, 1951; Fiske, 1966; Hakstian & Whalen, 1976）。Sijtsma（2009a）則是探討兩個主要的議題：(a) 因為 α 是真實信度係數之下限，而 $\alpha < \rho_{glb} < \rho_{XX'}$，其中 ρ_{glb} 是最大的信度下限（greatest lower bound），所以他建議使用 ρ_{glb} 而不是 α 來作為一個測驗分數之信度估計值；(b) α 是不是一個內部一致性的測量。

McNeish（2018）提到儘管很多文獻已指出 α 係數因一些不合實際的假定造成了一些問題，然而心理學中的實徵研究通常還是報導 α 係數做為內部一致性信度的一個測量。因此 McNeish 詳細介紹 α 係數的假定以及包括 ω_h、ω_t、ω_{RT}、H 係數、最大信度（maximal reliability）、和最大下限（greatest lower bound, glb）等信度估計方法，並以實徵資料說明對於同一筆測驗資料，這些信度係數估計方法所得到的值通常都高於 α 係數，因此 McNeish 建議研究者不再使用 α 係數，並提供 R 的語法，以協助研究者應用 ω_h、ω_t、ω_{RT}、H 數、最大信度、最大下限這些方法來估計信度係數。

雖然 McNeish（2018）的研究引發了一些回應，比如 Raykov 和 Marcoulides（2019），對筆者而言，McNeish 的論點基本上只是 Zinbarg, Revelle, Yovel 和 Li（2005）、Revelle 和 Zinbarg（2009）和 Sijtsma（2009a）等之延伸，因此，筆者在本節討論的內容，還是以 Cortina（1993）和 Sijtsma（2009a）兩人所探討的議題為主，底下分成數點進行討論：（1）α 是所有折半信度的平均數；（2）α 是一個測驗的信度係數之下限；（3）α 是第一個因素飽和度的一個測量；（4）α 是個內部一致性指標；（5）影響 α 係數的因素。

一、alpha 是所有折半信度的平均數

關於 α 是所有折半信度的平均數，Novick 和 Lewis（1967）曾經證明了這個陳述，由他們的證明可以看到當所有的折半信度皆是以 Flanagan 的公式計算時，則 α 的值等於所有可能的 Flanagan 折半信度之平均數。Cronbach（1951, pp. 302-305）也證明了這一點，雖然他提到是根據其文中表 1 的公式 2A 來計算折半信度的，但是在這二頁多的證明中，他僅在表 1 的下方註明公式 2A 是 Flanagan 的公式，並未在本文中明說。或許是這個原因，造成某些教科書提到 α 係數是所有折半信度係數之平均時，並未明確提到如果折半信度是利用 Spearman-Brown 校正的方式得到時，此定理未必成立的。

前面也提到，α 係數的計算方式主要有二，除了一般常見的利用共變數矩陣得到的以外，亦可利用相關矩陣來計算得到；然而，標準化的 α 係數並不等於所有折半信度係數之和。如前面有關不同平行測量程度那一節中所提到的，因為定二式測驗的 E_1 和 E_2 是被假定為無關的，所以，α 是所有折半信度的平均數是沒有問題的。然而，在驗證性因素分析（CFA）中，不同的觀察指標變項之誤差之間，被允許可以有相關存在，即 E_1 和 E_2 是可以有相關的，當誤差被允許是有相關存在時，α 不一定等於所有 Flanagan 折半信度的平均數。

二、Alpha 是一個測驗的信度係數之下限

Guttman（1945）認為 Spearman 的重要貢獻之一是將焦點放在一些變項之和的信度，因此他自己在該文中，也將重心放在一個加總分數的信度上面。他也強調信度係數一般來說沒有辦法由一次施測的資料估計得到，但是由於進行兩次獨立的施測在現實上有極大的困難，因此 Guttman 認為由一次施測的結果，我們可以計算信度的下限（lower bound）。

在 Guttman 的文章中，他討論了 λ_1、λ_2、λ_3、λ_4、λ_5 和 λ_6 等六個信度下限的計算公式（本文稍後會較詳細介紹），其中 λ_3 就是 α，而 Cronbach（1951, 2004）也指出 α 是信度係數的下限，如前一段所提到的，當 CTT 進一步假定不同式別的誤差分數是沒有相關的情形下，α 是 $\rho_{XX'}$ 下限的證明是 Novick 和 Lewis（1967）所提供的。

Novick 和 Lewis（1967）證明了如果資料不是本質上 τ 等值時，Cronbach 的 α 係數低估了信度 $\rho_{XX'}$。當測驗所有的成分皆是本質上 τ 等值時，α 的值等於真實的信度係數（p. 6, 定理 3.1）；必須要注意的是，標準化 α 並不是信度的下限，其值可能會大於理論上的信度（Osburn, 2000）。然而，當測驗試題的誤差分數部分彼此間有相關時，這些相關常是正的，會造成 α 係數高估了實際的信度（Bentler, 2009; Green & Hershberger, 2000; Green & Yang, 2009b; McNeish, 2018），也就是 $\alpha \geq \rho_{XX'}$。

對於 α 係數是個信度的下限，Cronbach 本人在 1951 年的那篇文章中其實就有探討這個問題，但是似乎他本人於這個問題不是很在意，而對照 Cronbach（2004），更可以看到其實他在意的不是 α 是否為一個下限，而是讀者從他的 1951 年文章中，是否嗅得到他所在意的，由隨機的平行測驗之概念上所得到的真分數之意義，以及相關連信度的意義，那些終究將 Cronbach 引領到類推性理論（generalizability theory）去。

三、Alpha 是第一個因素飽和度的一個測量

這個說法意指 α 是在一組試題中，一個普通因素（general factor）代表該組試題之程度，因此也代表了該組試題彼此之間相互關連的程度。Cortina（1993, p. 99）認為這個陳述與 Cronbach 在其原始的研究報告之說法相反，且已在後續的研究中被說明是錯誤的；然而，Cortina 也指出這個陳述並不是完全錯誤的，因為根據 Kaiser（1968）的研究，當標準化 α 被使用時，此說法是部分正確的。Kaiser 說明了如果所有試題間之交互相關皆等於相關之平均數（也就是整組試題只有一個主成分），則標準化的 α 係數與第一個未轉軸前的主成分之特徵值有直接的關連。這項關係之成立決定於單一向度性是否存在，因此當解釋一組試題之間的相關需要超過一個因素時，標準化的 α 是不合適的（Cortina, 1993, p. 99）。

為何研究者會有 α 是第一個因素飽和度之測量這樣的印象呢，除了 Kaiser 的研究以外，是否有其他的原因呢？在 Cronbach（1951）的文章中，在 α 係數是如何與測驗的同質性、內部一致性或飽和度（saturation）產生關聯的那一小節中（pp. 319-320），Cronbach 提到 α 估計了測驗變異數（test variance）為試題間所有共同因素所解釋的比例，在每一個試題為共同因素所解釋的變異數之平均等於試題間共變數之平均的假設之下，Cronbach 證明了總變異數（試題變異數與共變數之和）被共同因素所解釋的部分等於 $2(\kappa/\kappa\text{-}1)C_t$，其中 κ 為題數，C_t 是所有試題間共變數之和；而 $2(\kappa/\kappa\text{-}1)C_t$ 除以 σ_X^2 所得到的就是 α 係數。當測驗資料是單一向度時，似乎將 α 想成是一個普通因素所解釋的變異量也不見得有問題。另外，在討論測驗變異數的因素組成部分（pp. 312-316），Cronbach 以雙因素模式（bi-factor）的方式來將任何試題之變異數分解成三個部分：普通因素、群組因素以及獨特因素。

當假設一個測驗是由一個普通因素及五個群組分數所組成（f_1 為普通因素，f_2,.....,f_6 為群組因素），因此 m=6 為因素的個數，$\eta_1=\eta$，$\eta_2=\eta_3=\eta_4=\eta_5=\eta_6=(1/5)\eta$，若所有的題目有相等的變異數，且任何因素有相同的負荷量（f_m），則

$$f_{1t}^2=\frac{5n^2}{6n^2+5n}=\frac{5n}{6n+5}$$

$$\underset{n\to\infty}{Lim}\, f_{1t}^2=\frac{5}{6}=.83$$

$$f_{2t}^2=...=f_{6i}^2=\frac{n^2/5}{6n^2+5n}$$

$$\underset{n\to\infty}{Lim}\, f_{2t}^2=.03$$

$$\sum_i f_{u_it}^2=\frac{5}{6n+5}$$

$$\underset{n\to\infty}{Lim}\sum_i f_{U_it}^2=0$$

以上是 Cronbach（1951, p. 313）的公式 33-38 的部分，由這項結果可以看到如果測驗是依照上面所假設的方式編製的，測驗變異數有很大的部分是由普通因素所解釋（在此例中為 .83）。很可能 Cronbach（1951）文中的這些敘述，造成後來的學者會有 α 是第一個因素飽和度的一個測量那樣的想法。

四、Alpha 是個內部一致性指標

在文獻中，內部一致性（internal consistency）這個詞彙常與同質性（homogeneity）或是單一向度性（unidimensionality）產生混淆，比如說 Nunnally 和 Bernstein（1994, p. 243）將這兩個名詞視為是同義字，並未加以區別，而 Cronbach（1951, p. 320）似乎也將這二個名詞視為同義字而交換使用；但是 Schmitt（1996）將內部一致性與同質性這兩個名詞加以區分，他宣稱內部一致性指的是一組試題彼此間之交互相關，而同質性則是指一組試題具有單一向度性說的，也就是同質性與單一向度性是同義字。

Sijtsma（2009a）認為 Schmitt 的分法並未完全解決了這兩個名詞混淆之問題。對 Sijtsma 而言，單一向度性並非是一個單一的概念，而是一個模式特定的概念，比如說在 IRT 中，不同的模式（例如，1PL, 2PL, 3PL 等）似乎有其自己的單一向度性之定義；因此，在這種意義之下，很清楚的是在某個特殊的模式之下，單一向度性的意義可能有所不同。然而，Sijtsma（2009a, p. 114）認為內部一致性的定義從來就不像單一向度

性那麼明確。Cronbach （1951, p. 320）提到當時許多學者將他們的注意力轉移到一個被稱為同質性（homogeneity）、可量尺化性（scalability）、內部一致性（internal consistency）或是類似名稱的特性（property）上面，但是該特性並未被清楚地定義，由 Cronbach 接下去的陳述，可以清楚瞭解他對於內部一致性的觀點，其實就是測驗是單一向度的。他提到一個具有實質的內部一致性之測驗，該測驗是心理學地可解釋的（psychologically interpretable），而一個測驗要是可解釋的，所需要的是測驗的變異數中，有一大部分是由主要的因素所解釋即可。由此來看，Cronbach 的內部一致性與同質性（或單一向度的）是具有同等意思的。

Sijtsma（2009a）提到在測驗編製實務中，內部一致性經常指的是試題彼此間有交互相關，但是其他的解釋（例如，單一向度性）也經常被使用。「在實際的測驗編製中，關於測驗之內部一致性一個最流行但非正式的解釋是試題交互相關矩陣的第一個特徵值與第二特徵值相較之下大了很多，但是到底要多大才算是足夠的大，卻是不清楚的」（p. 114）。Sijtsma 認為這個解釋與將內部一致性和單因素解或是 IRT 的單一向度性劃上等號的確是不同的，因此留下了不同的試題有不同的因素負荷量組型之可能性。如果超過一個以上的因素被保留的話，這就接近了 Cronbach 的說法了，試題不需要具有相似的因素結構，測驗仍然是內部一致的。Sijtsma（2009a）認為以上的分析最重要的部分是指出內部一致性這個概念的模糊性，然而這個模糊性並沒有使 α 係數不成為內部一致性之代表性統計數（p. 114）。

對於學術界中依舊以內部一致性來解釋 α 的作法會如此固執地持續著，Sijtsma（2009a）認為有下列兩個理由：（1）雖然有一些研究清楚的解釋 α 與其他量數之關係（例如，Crontina, 1993; Green, Lissitz, & Mulaik, 1977; 也請參見 Cronbach, 1988），特別是測驗的因素結構；它們同樣也傳達了底下的想法，因為 α 也與測驗的因素結構有一些關係，因此其值一定表達了這個因素結構之特徵。雖然邏輯上這是不正確的，也不是那些作者想要的，然而對於部分的研究者來說，這種想法可能是無法抗拒的。（2）在 1950 年代之後，心理計量學已經發展成比較是數學和統計學的取向，然而心理學家主要仍然還是心理學家；這兩個世界彼此分離的程度超過眾人所期待的，倒是一個事實。雖然 Cronbach 的研究報告仍然是很多心理學家可以取得並閱讀的，然而，由 Lord、Novick 和 Lewis 和其他很多人所完成的著作，卻是未曾得到大部分心理學家們的注意（Sijtsma, 2009a, pp. 114-115）。

總結來說，Sijtsma（2009a）認為 α 不是一個內部一致性的測量，也不是一個接近單一向度性程度的測量；因此他建議，既然如此，為何不將內部一致性這個名詞一併捨去不用呢（Sijtsma, 2009b）。筆者同意他的看法，在一般的教科書所提到的信度類別之一的內部一致性這個名詞，某個程度是不易解釋的。因為內部既然是一致的，那麼測驗中所有的試題應當測量相同的構念，令人容易認為測驗是單一向度的。

五、影響 alpha 係數的因素

Green, Lissitz 和 Mulaik（1977）的研究結果指出即使是一個具有兩個向度之測驗，

α 值仍然可以達到 .80 以上，而且 α 的值似乎明顯受到題數的影響；為了進一步澄清這些議題，Cortina（1993）進行模擬研究，而得到如表 6 的結果。

表 6

不同向度、不同題數、和不同的試題間相關平均值之組合下之 alpha 值和精確度估計值

	試題交互相關的平均					
	r = .30		r = .50		r = .70	
題數	α	精確度	α	精確度	α	精確度
		單一向度				
6	.72		.86		.93	
12	.84		.92		.96	
18	.88		.95		.98	
		二向度				
6	.45	.04	.60	.07	.70	.09
12	.65	.02	.78	.03	.85	.04
18	.75	.01	.85	.02	.90	.03
		三向度				
6	.28	.03	.40	.05	.49	.08
12	.52	.02	.65	.03	.74	.04
18	.64	.01	.76	.02	.84	.02

註：r 為試題間相關係數之平均。取自 “What is coefficient alpha? An examination of theory and applications.” by Cortina, J. M. (1993). *Journal of Applied Psychology, 78*(1), 101.

綜合表6的結果，可以獲致下列的幾項結論：（1）無論試題間之相關的平均值為何，題數增加時，α 的值隨著變大，因此，題數對於 α 的值有深遠的影響。（2）在一個單一向度的量表中，無論題數為何，當試題間之相關的平均在 .50 時，所得到的 α 的值已在傳統可接受的水準以上（大於 .75），即使平均相關只有 .30，12 題以上的測驗之 α 也在傳統可接受的水準以上。在二向度中，當試題間平均相關在 .50 以上，題數在 12 以上的話，α 的值皆在 .75 以上，三向度資料的 α 雖然低一些，但是也只要 18 以上，即可具有傳統可接受的水準。（3）如果一個單向度的量表有足夠的試題（即多於 6 題），其 α 的值就會大於 .70，即使試題間的相關之平均是非常小；而二向度的測驗只有 12 題，三向度的測驗只要有 18 題，測驗的 α 值幾乎都可在 .65 以上（Cortina, 1993）。

因此，隨著試題間相關的值變大，或是題數的增加，α 的值都隨著變大，雖然向度性變得複雜會使得 α 係數的值變低，但是只要題數增加，α 仍然變大；所以由 α 係數值的大小是無法判斷測驗資料之向度性的。而且，α 的值是受到題數的影響最大，如由表 6 中所看到的，即使試題間的平均相關為 .30，且是個三向度的測驗資料，當題數為 18 時，α 等於 .64；對於一般的心理測驗來說，這個值已經不低了。而上面第二

個例子之平均相關僅有 .137（見表 5），但因有 60 題，α 值高達 .906，因此在解釋 α 時，必須將題數放在心上（Cortina, 1993）。

陸、Guttman 的六個下限

如果有一份測驗是由 κ 個試題所組成，σ_1^2 、...、σ_k^2 分別為每一個試題分數之變異數，而 σ_X^2 為每一個受試者觀察分數之變異數。Guttman（1945）提出了 λ_1、λ_2、λ_3、λ_4、λ_5 和 λ_6 等六個信度估計值之下限，Jackson 和 Agunwamba（1977）以及 Revelle 和 Zinbarg（2009）也從不同的角度介紹這六個下限。其中最簡單的下限是 λ_1，可以由下列的公式計算得到：

$$\lambda_1 = 1 - \frac{\sum_{j=1}^{k} \sigma_j^2}{\sigma_X^2} \tag{5}$$

若由 $\rho_{XX'} = \sigma_T^2 / \sigma_X^2 = 1 - \sigma_E^2 / \sigma_X^2$ 來看，可以看到 λ_1 基本上是把 $\sum_{j=1}^{k} \sigma_j^2$ 視為是 σ_E^2，亦即將測驗中每一個試題的變異數之和視為是整個測驗之誤差變異數 σ_E^2。如果這 κ 個試題間的共變數被平方，然後加總起來，並以 C_2 來表示，Guttman 認為另一個信度估計值的下限可以由下式得到，

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{k}{k-1} C_2}}{\sigma_X^2} \tag{6}$$

當初 Guttman 提出 λ_3 時，主要是要節省計算 C_2 所需要的時間，

$$\lambda_3 = \frac{k}{k-1}\lambda_1 = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} \sigma_j^2}{\sigma_X^2}\right) = \alpha \tag{7}$$

由公式（7）可以清楚看到，λ_3 與 α 是一模一樣的。就 λ_1、λ_2 和 λ_3 這三者而言，其值的大小關係為 $\lambda_1 < \lambda_2 < \lambda_3$。

Guttman 的第四個下限 λ_4 與經 Spearman-Brown 公式校正的折半信度類似，但是，λ_4 比用 Spearman-Brown 公式校正的折半信度來得容易計算。要計算 λ_4，測驗必須被分成兩個部分計分，如果這兩個部分的變異數為 σ_a^2 和 σ_b^2，則

$$\lambda_4 = 2\left(1 - \frac{\sigma_a^2 + \sigma_b^2}{\sigma_X^2}\right) \tag{8}$$

Guttman 提到，無論測驗是如何分成兩半的，λ_4 永遠是一個下限。如果 σ_a^2 的值與 σ_b^2 相等，則 λ_4 的值會等於經 Spearman-Brown 公式所校正的折半信度的值，但仍然是一個信度的下限。因為一個測驗有 κ 個試題，有 $k!/[2(\frac{k}{2}!)(\frac{k}{2}!)]$ 個可能的折半組合方式，不同折半方式得到的 λ_4 值是不相同的，所以 Jackson 和 Agunwamba（1977）將 λ_4 定義為具有最大值的那個折半信度；此作法是合理的，因此本文也依循他們的想法，將 λ_4 視為是最大的折半信度。一些作者會使用 λ_4(max) 這個符號，比如 Revelle & Zinbarg（2009）的表 1，但在本文中，直接以 λ_4 來表示。

如果令 C_{2j} 代表第 j 題與其他 κ -1 個題目之間的共變數之平方和，且 $\bar{C}_2$ 代表 κ 個 C_{2j} 中最大那一個的值，亦即 $\bar{C}_2 = \max_j(C_{2j})$ ，那麼 λ_5 可以由下式得到：

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\bar{C}_2}}{\sigma_X^2} \tag{9}$$

最後一個下限是以多元迴歸為基礎的，如果 e_j^2 是以其他 κ -1 個題目的分數來預測第 j 題的表現時所得到的誤差變異數，那麼 e_j^2 相當於 $1-r_{smc,j}^2$ ，$r_{smc,j}^2$ 是第 j 題的多元相關之平方（squared multiple correlation, SMC）。因此，

$$\lambda_6 = 1 - \frac{\sum_{j=1}^{k} e_j^2}{\sigma_X^2} \tag{10}$$

柒、ρ_{glb}、ω_{t}、ω_{h}、β 及 Θ 係數

在 ρ_{glb}、ω_{t}、ω_{h}、β 及 Θ 係數這五個係數中，除了 ρ_{glb}和 β 係數以外，其餘三者皆是由因素分析或主成分分析的結果中計算得到的。β 係數是由 Revelle (1979) 所提出的，其計算公式如下：

$$\beta = \frac{k^2 \bar{\sigma}_{ij'}}{\sigma_X^2} \tag{11}$$

其中 $\bar{\sigma}_{ij'}$ 是在最差的折半中不同半測驗中試題之間的共變數之平均；也就是，該折半使得某一個半測驗之試題與另一個半測驗試題之共變數的平均值最小，且不同半測驗之題數可以不等。Revelle（1979）認為在一個比 α 係數還要一般的情形下，β 的值等於量尺分數的變異數由一個普通因素所解釋的比率。α 和 β 之間的關係可以由 Cronbach (1951, p. 304) 的公式 16 看得到，

$$\alpha = \frac{k^2 \bar{\sigma}_{ij}}{\sigma_X^2} \quad (12)$$

其中 $\bar{\sigma}_{ij}$ 是試題間共變數之平均值，由公式（11）和（12）可以瞭解 $\beta \le \alpha$，因為 $\bar{\sigma}_{ij'} \le \bar{\sigma}_{ij}$（亦請參見 Zinbarg, Revelle, Yovel, & Li, 2005）。

當有普通因素和群組因素存在時，根據因素分析的模式，一個測驗分數的變異數 σ_X^2 可分解成四個部分：由一個普通因素（general factor）所造成的 g，由一組群組因素（group factors）所造成的 **f**，每一個試題所獨特擁有的特殊的因素 **s** 和隨機誤差 **e**。因為在因素分析中，一般來說特殊變異數與隨機誤差是無法加以區別的，除非這個測驗被施測兩次，因此 McDonald（1999）將這二者合併成為誤差。所以

$$\mathbf{x} = \mathbf{c}g + \mathbf{A}\mathbf{f} + \mathbf{D}\mathbf{s} + \mathbf{e} \quad (13)$$

根據普通因素和群組因素，試題 j 的共同性（communality）為

$$h_j^2 = c_j^2 + \sum f_{ij}^2 \quad (14)$$

其中，c_j^2 是第 j 題在普通因素上的負荷量，而該題的獨特變異數

$$u_j^2 = \sigma_j^2 (1 - h_j^2) \quad (15)$$

可以被用來估計測驗的信度，也就是，如果 h_j^2 是試題 根據普通因素以及群組因素所得到的共同性，那麼對標準化的試題來說，$e_j^2 = 1 - h_j^2$ ，且

$$\omega_t = \frac{\mathbf{1cc'1} + \mathbf{1AA'1'}}{\sigma_X^2} = 1 - \frac{\sum (1 - h_j^2)}{\sigma_X^2} = 1 - \frac{\sum u^2}{\sigma_X^2} \quad (16)$$

因此 ω_t 是根據所有的因素負荷量所得到的信度係數估計值。McDonald（1978）原先將此係數稱為 ω 係數，因為 McDonald（1999）在其書中的公式（6.20a）提到另外一個 ω 係數；因此 Revelle 和 Zinbarg (2009) 將前者稱為 ω_t 係數（即 ω_{total}），而將後者稱為 ω_h 係數（即 $\omega_{hierarchical}$），以避免混淆。仔細比較公式（16）與（10），可以發現 ω_t 與 λ_6 非常類似，因為 $h_j^2 \ge r_{smc,j}^2$，所以 $\omega_t \ge \lambda_6$。要注意的是，McNeish（2018）介紹了另一個 omega total，而將此處的 ω_t 稱為 Revelle 的 ω_t （Revelle’s omega total）；在本文中，僅考慮 Revelle 的 ω_t ，因此仍然以 ω_t 來表示。

如果僅以普通因素負荷量之平方和作為共同性的估計值，上式則變成

$$\omega_h = \frac{\mathbf{1cc'1}}{\sigma_X^2} \quad (17)$$

也就是一個測驗的量尺分數變異數為一個普通因素所解釋的比率（McDonald, 1978）。因此，Revelle 和 Zinbarg (2009) 認為 ω_h 是一個測驗測量一個共同因素（一個構念）之程度的重要指標，似乎他們意指 ω_h 可以作為資料是否為單一向度的指標。蔡佩圜、涂柏原和吳裕益（2018）建議可以 ω_h 與 ω_t 之比值作為評估測驗是否接近單一向度的指標。

Armor (1974) 介紹了一個信度的估計值 Θ 係數，它被發展用來解釋一個量表之多向度性，且是根據一個主成分模式。單因素解之 Θ 係數是利用下述公式來計算的：

$$\Theta = \frac{k}{k-1}\left(1-\frac{1}{\lambda_1}\right) \quad (18)$$

在此，λ_1 是利用相關矩陣進行主成分分析所得到最大的特徵值，而不是 Guttman 的信度下限中的 λ_1。Θ 係數似乎是在社會學領域中被使用的比較多，而在教育測驗這個領域中，較少被使用。

Jackson 與 Agunwamba（1977）及 Woodhouse 與 Jackson（1977）根據 CTT 導出最大信度下限 ρ_{glb}。要計算 ρ_{glb} 需將試題共變數矩陣（C_x）分解成真分數共變數矩陣（C_T）與誤差共變數矩陣（C_E），這兩個矩陣必須都是半正定矩陣，也就是矩陣所有特徵值都大於或等於 0，不能有負的特徵值。Bentler 與 Woodward（1980）將 ρ_{glb} 定義為：

$$\rho_{\text{glb}} = 1 - \frac{tr(C_E)}{\sigma_X^2} \quad (19)$$

捌、各種信度係數之比較

本文介紹了包括 α、ρ_{glb}、λ_2、λ_4、λ_6 及 ω_t 等信度估計的方法。對於這些信度係數指標，過去已有許多比較研究，在此針對本文的目的，簡單地摘要部分的研究發現。首先，Ten Berge 與 Sočan（2004）利用 De Leeuw（1983）所提供的相關係數矩陣作為母群真值，以進行資料模擬，該相關係數是由 119 位受試者在六個與政治有關的調查題目上的反應所得到的；因為這六個題目測量的是相同的特質，可以被視為單一向度的。Ten Berge 和 Sočan 模擬了 100、250、500 和 1000 人等不同的樣本人數，每一種人數產生 500 個資料，以探討 ρ_{glb} 估計信度時之偏誤情形，並比較 α、λ_4、ρ_{glb} 以及一個以因素分析為基礎的信度估計值等 4 種不同信度估計方法之差異，主要發現是在各個資料集所估計得到的 ρ_{glb} 均較 α 大；ρ_{glb} 的計算受到樣本人數影響的情形比 α

大；ρ_{glb} 的計算受到樣本人數影響的情形比 α 明顯；無論在何種情況下，λ_4 的值皆與 ρ_{glb} 相近。也就是，即使在單一向度的情境下，α 所得到的信度估計值均比 λ_4 和 ρ_{glb} 小。

Sijtsma（2009）使用 Cavalini（1992）的實證資料（828 人在 8 個 4 點量尺的評定量表上的反應資料）探討 α 、λ_2 和 ρ_{glb} 之估計值之大小關係，所得到的結論主要有以下四點：（1）$\alpha \le \lambda_2 \le \rho_{glb}$，$\alpha$ 與 ρ_{glb} 之間的差距給予 α 不正確程度的一個印象；（2）有一些信度指標所得到的估計值的大小介於 α 與 ρ_{glb} 之間（比如 λ_2），因此很難為繼續使用估計值最低的 α 來辯護；報導 α 的唯一理由可能是較頂級的期刊傾向接受使用已存在一段時日的統計方法（如，α）的文章，因此合理之道在於 α 以外，還要有其他較大的信度下限；（3）最佳的信度估計下限為 ρ_{glb}，然而當樣本數小於 1,000 人，且試題超過 10 題時，ρ_{glb} 可能會嚴重地正向的偏誤，因此需要更多的努力來校正 ρ_{glb} 之偏誤；（4）α 不是一個內部一致性的測量，也不是單一向度性程度的測量。

為了回應 Sijtsma（2009）的看法，Revelle 與 Zinbarg（2009）使用九筆實徵資料來比較 β 、ω_h、λ_1、α 、α_{PC}、λ_2、μ_2、μ_3、λ_5、λ_6 (smc)、λ_4、ρ_{glb}、ω_t 等 13 種不同的信度估計方法所得到的結果，其中六筆資料來自於 Sijtsma（2009），兩筆來自於 Bentler 和 Woodward（1980），另外一筆則取自於 Ten Berge 和 Sočan（2004）年所使用來自於 De Leeuw（1983）的資料。筆者將他們的研究發現複製在表 7 之中。從表 7 可以看到，在 9 筆資料中有 7 筆資料的最大信度估計值是 ω_t，2 筆資料的最大信度估計值是 λ_4，相較之下，ρ_{glb} 並非如其名是最大信度估計值下限；且在每一筆資料中，α 係數的值皆明顯低於 ω_t 或是 ρ_{glb}。因此 Revelle 與 Zinbarg 建議使用 ω_t 作為信度最佳估計值，並附議 Sijtsma（2009）的看法，建議研究人員在報告 α 係數時，應同時報告其他較佳的信度估計值。

McNeish（2018）發現無論是在明顯違反 τ 等值的測驗，或是一般的 Likert 量表中，α 係數所得到的信度估計值皆小於 omega total、glb、ω_t 和 H 係數。蔡佩圜、涂柏原、吳裕益（2017）探討三種測驗類型、兩種因素數目、四種因素題數及四種樣本數對於 ρ_{glb} 、λ_2、α 、λ_4、λ_6 及 ω_t 等六種信度估計方法之估計誤差的影響，發現這六種方法中， ω_t 及 λ_4 在各種情境組合下，估計誤差均極微，所以他們認為 ω_t 及 λ_4 是較佳的兩種信度估計方法；因此蔡佩圜等人建議當測驗資料之因素結構很明確時，以 ω_t 估計信度，但若因素結構不明確時以 λ_4 估計信度。筆者認為以上的建議是因為大多數的實務工作者很少在估計資料的信度之前，會先利用因素分析或結構方程模式檢驗其手上資料的結構，這是理論研究者才會做的。筆者自己也認為 Sijtsma（2009）的看法是值得實務工作者注意且採用的，也就是在 α 以外，應同時報告 ω_t 或是 λ_4 等表現較佳的信度估計值。一個商業發行的教育或心理測驗報告常會同時提供重測信度、複本信度及 α 係數（或 KR20），因此，一般的實徵研究在 α 以外再加上 ω_t 及 λ_4 應是好的作法。

表 7

13 種信度估計法的結果之比較

估計法	S-1	S-1a	S-1b	S-2a	S-2b	S-2c	B & W1	B & W2	TB & S
題數	8	4	4	6	6	6	4	6	6
β	.656	.651	.610	.000	.000	.437	.756	.854	.739
ω_h	.593	.643	.676	.049	.000	.532	.706	.921	.767
λ_1	.687	.561	.507	.444	.444	.444	.671	.785	.700
$\lambda_3(\alpha)$	.785	.749	.676	.533	.533	.533	.894	.942	.840
α_{pc}	.787	.749	.676	.553	.533	.553	.896	.943	.841
λ_2	.789	.753	.678	.643	.585	.533	.898	.943	.842
μ_2	.790	.755	.657	.663	.592	.533	.899	.943	.843
μ_3	.791	.755	.658	.666	.592	.533	.900	.943	.843
λ_5	.766	.738	.660	.593	.549	.511	.881	.911	.819
λ_6	.785	.713	.593	.800	.571	.488	.880	.960	.830
λ_4	.853	.820	.696	.889	.647	.533	.913	<u>.979</u>	.884
ρ_{glb}	<u>.852</u>	.820	.696	.889	.667	.533	.920	.976	.885
ω_t	.844	<u>.893</u>	<u>.859</u>	<u>.889</u>	<u>.669</u>	<u>.561</u>	<u>.951</u>	.972	<u>.900</u>

註：每一欄中加上底線的數字是最大的估計值。資料取自 “Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma” by Revelle, W., & Zinbarg, R. E. (2009). *Psychometrika, 74*(1), 150.

玖、結論與建議

本文的目的是對於古典測驗理論的信度估計理論及方法作一個詳細的說明，尤其是對於目前實徵研究中被運用最多的 α 係數；如何讓信度報告變成更有品質，是筆者撰寫此文的出發點。為何許多學者皆知道對於 $\rho_{xx'}$ 的估計，α 的表現實是不佳，然而絕大多數的社會科學研究者，依舊僅報告 α 作為其測驗工具的信度？或許是因為大多數的研究者在信度理論方面的訓練不是很充分，且以通用的統計軟體 SPSS 或 SAS 要得到 α 值是非常容易的。本文用了一個小節呈現測驗平行程度的定義，目的是在提醒讀者各個信度估計方法皆有其所需要的假定，就像一般學者常用的各種統計方法一樣。有趣的是，當大家在應用像是 t-test 或是 ANOVA 等統計方法時，總是會記得至少去檢驗變異數同質性這個假定有被滿足否，然而在利用 α 係數估計信度時，似乎沒有任何一個人會去檢查本質的 τ 等值這個假定是否被滿足。這應當是目前測驗工具信度報導中最主要的問題。筆者認為在許多情境中，由 α 所提供的測驗資料的信度估計值可能嚴重低估了真實的信度值，其主要理由是本質的 τ 等值假定未被滿足，Graham（2006）也持相同的觀點。

由 Revelle 與 Zinbarg（2009）的研究結果來看，可以知道 ρ_{glb}、λ_2、λ_4 及 ω_t 所得到的信度估計值皆不低於 α，其中表現最佳的是 ω_t；蔡佩圜、冷柏原、吳裕益（2017）的研究也發現 λ_4 及 ω_t 兩個估計方法最不會受到各種因素的影響。因此，對實徵研究者來說，如果不確定其資料的因素結構，則除了呈現 α 係數以外，遵行 Sijtsma（2009a）和 Revelle 及 Zinbarg（2009）的建議，再輔以 ω_t 及 λ_4 等估計值，應是值得大家去做的。

本文有一個主要的限制，就是所介紹的方法被侷限在以 CTT 為基礎的，過去多年來，以結構方程模式（structural equation modeling, SEM）為本的方法也被不少學者提出（例如，Bentler, 2009；Green & Yang, 2009b；McDonald, 1999；Raykov & Shrout, 2002）。對許多學者而言，以 SEM 為基礎的方法可能是比以 CTT 為基礎的更好，然而，限於篇幅，這個部分就留到以後了。

參考文獻

一、中文部分

蔡佩圜、冷柏原、吳裕益（2017）。測驗類型、因素數目、各因素題數及樣本數對六種信度估計法估計誤差之交互影響分析，**教育學刊，49**，35-77。

蔡佩圜、冷柏原、吳裕益（2018）。九種古典測驗理論信度指標精確性之研究，**測驗學刊，65**（2），217-240。

蔡佩圜、吳裕益、冷柏原（2020）。向度數、題數及樣本數分別與六種信度估計法估計誤差交互作用效果之探討，**教育學誌，43**，67-104。

二、西文部分

Allen, M. J. & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Armor, D. J. (1974). Theta reliability and factor scaling. In H. Costner (Ed.), *Sociological methodology* (pp. 17-50). San Francisco, CA: Jossey-Bass.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program.* Los Angeles: BMDP Statistical Software.

Bentler, P.M. (2009). Alpha, dimension-free, and model-based internal consistency *reliability. Psychometrika*, 74(1), 137-143.

Bentler, P. M., & Woodward, J. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika, 45*(2), 249-267.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on ondour annoyance and the dispersions of odorant concentrations from industries.* Ph. D. thesis, University of Groningen, The Netherlands.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Crano, W. D., & Brewer, M. B. (1973). *Principles of research in social psychology.* New York, NY: McGraw-Hill.

Crocker, L. & Algina, L. (1986). *Introduction to classical and modern test theory.* New York, NY: Harcourt Brace Jovanovich College.

Cronbach, L. J. (1951). Coefficient alpha and the internal structural of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418.

Cronbach, L. J., Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational and Psychological Measurement, 25,* 291-312.

De Leeuw, J. (1983). Models and methods for the analysis of correlation coefficients. *Journal of Econometrics, 22,* 113–137.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York, NY: Macmillan.

Fiske, D. W (1966). Some hypotheses concerning test adequacy. *Educational and Psychological Measurement, 26*, 69-88.

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika, 48,* 99-111.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930-944.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7,* 251–270.

Green, S. B., Lissitz, R. W, & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827-838.

Green, S. B. & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121-135.

Green, S. B. & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155-167.

Gulliksen, H. (1950/1987). *Theory of mental tests.* Hillsdale, NJ: Erlbaum.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282.

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219-232.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: 1: Algebraic lower bounds. *Psychometrika, 42*(4), 567-578.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133.

Jöreskog, K. G., & Sörbom, D. (1985). LISREL VI: Analysis of linear *structural relationships by the method of maximum likelihood. User's guide.* Uppsala, Sweden: University of Uppsala.

Kaiser, H. F. (1968). A measure of the average intercorrelation. *Educational and Psychological Measurement, 28,* 245-247.

Kelly, T. L. (1923). *Statistical method.* New York, NY: Macmillan.

Kelly, T. L. (1942). The reliability coefficient. *Psychometrika, 7,* 75-83.

Kristoff, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39,* 491-499.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151-160.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDonald, R. P. (1978). Generalizability in factorable domains: "Domain validity and generalizability." *Educational and Psychological Measurement, 38,* 75-79.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23,* 412-433.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* Thousand Oaks, CA: Sage.

Novick, M. R., & Lewis, C. L. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1-13.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficient. *Psychological Methods, 5*(3), 343-355.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis.* Thousand Oaks, CA: Sage.

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria (ISBN 3-9000051-07-0).

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549-565.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Routledge.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200-210.

Raykov, T., & Shrout, P. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9,* 195–212.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14,* 57-74.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145-154.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9,* 99-103.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*(1), 169-173.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72-101.

Ten Berge, J. M. F. & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613-625.

Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43*(4), 575-579.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.

Woodhouse, B., & Jackson, P. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42*(4), 579-591.

Zinbarg, R.E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1). Available at: http://digitalcommons.wayne.edu/jmasm/vol6/iss1/4

投稿日期：2019 年 09 月 19 日
修正日期：2020 年 02 月 04 日
接受日期：2020 年 03 月 27 日

A Note on Coefficient Alpha and Some Related Reliability Estimation Methods

Bor-Yaun Twu

Associate Professor, Department of Education, National University of Tainan

ABSTRACT

Since Spearman published two studies in 1904, classical test theory (CTT) had been developed and matured, and the concept of reliability had also been established. Among the methods proposed to estimate the reliability, the coefficient alpha proposed by Cronbach (1951) had emerged to be the most popular one. However, most of those who employed coefficient alpha to estimate the test reliability may not be familiar to the theory foundation. In this paper, the nature and limitation of coefficient alpha had been explicated; and some other reliability estimates, such as, ρ_{glb} ,λ_1,λ_2,λ_3,λ_4,λ_5,λ_6,β,ω_t,ω_h,and Θ were also introduced. It is suggested that when reporting the reliability for an empirical study, other indices should be presented together with coefficient alpha.

Keyword: Guttman's lower bounds, coefficient alpha, ω_t , reliability

Corresponding Author: Bor-Yaun Twu, e-mail: bortwu@gmail.com