綜　論

測量工具的效度與信度

李中一

CHUNG-YI LI

輔仁大學醫學院公共衛生學系，台北縣新莊市242中正路510號
Department of Public Health, College of Medicine, Fu Jen Catholic University, No. 510 Chung-Cheng Rd., Hsin Chuang, Taipei Hsien 24205, Taiwan, R.O.C.

*通訊作者Correspondence author.　　E-mail: chungyi@mails.fju.edu.tw

目標：介紹測量工具的信度與效度概念與評估的方法，討論使用缺乏信度與效度的工具對研究結果的影響，並針對某些測量實務上的議題提出建議。**方法：**透過文獻探討，摘要其內容，列舉實例或假想之數據，說明測量工具的信度與效度概念及其評估方法。**結果：**測量工具效度的高低取決於測量所牽涉系統誤差大小，而信度大小則與隨機誤差有關；評估測量工具信效度的方法依測量資料的屬性而定，而流行病學研究將會因為所使用測量工具的信效度不完善而產生訊息偏差，此偏差的程度與方向則與測量誤差的本質有關。**結論：**研究者在使用測量工具前有必要選擇適當的方法評估其相關之效度與信度，以了解使用該測量工具對研究結果的可能影響，如此方能對研究結果作正確的闡釋。(台灣衛誌 2004；**23(4)**：272-281)

關鍵詞：信度、效度、流行病學、偏差、測量誤差

Validity and reliability of an instrument

Objectives: This paper illustrates the concept of validity and reliability associated with an instrument and how the validity and reliability are assessed. We also included a discussion on how an instrument with unsatisfactory validity and reliability may affect study results, and provide suggestions for certain practical problems encountered by investigators. **Methods:** We reviewed the literature and provided real-world or hypothetical examples. **Results:** The level of validity of an instrument is related to the magnitude of systematic errors associated with that instrument, while the level of magnitude of reliability is determined solely by the degree of random errors involved in the measurement. The choices of methods used for the assessment of validity and reliability depend on the attribute of research data. Results from epidemiological studies that used an instrument with non-perfect validity and reliability might entail certain degrees of bias, for which the direction and magnitude are associated with the nature of measurement errors. **Conclusions:** Researchers should assess, using appropriate methods, the validity and reliability of an instrument before it can be used. This would help to appreciate the potential effects on the study results caused by measurement errors, and lead to correct interpretations of the study results. (*Taiwan J Public Health. 2004;**23(4)**:272-281*)

Key Words:Reliability, Validity, Epidemiology, Bias, Measurement error

前　言

在進行測量的過程中，通常會遇到不同程度的測量誤差，而此測量誤差則往往會對研究結果的正確性產生影響，雖然研究者不太可能完全避免測量誤差的問題，但清楚瞭解測量誤差如何發生、評估測量誤差的方法，以及消除測量誤差的方法卻是促使研究者能夠提高研究結果正確性的重要因素[1]。

一個流行病學研究對於暴露因素、疾病或干擾因子測量過程中都可能發生測量誤差，而造成測量誤差的原因與使用測量工具者(可能是工具施測者或受測量者)或測量工具本身有關；測量誤差則會造成一個流行病學研究的結果發生訊息偏差 (information bias)。使用測量工具者常會有意或無意的提供錯誤的訊息，例如最近一個探討人工流產史與乳癌相關性的荷蘭病例對照研究顯示：人工流產史會造成乳癌發生率顯著增加，相對危險性估計值為14.6；但研究者後來比對相關資料後發現：由於對照組個案多來自荷蘭南部天主教區(民風較為保守)而傾向低報其人工流產史，因而高估了人工流產史的致乳癌危險性[2]，這類測量誤差通常必須透過訪視員訓練或佐以更客觀資料來降低其影響。

此外，測量誤差也可能導因於測量工具本身的設計問題。例如，早期探討居家極低頻電磁波暴露與小兒癌症的研究經常利用住家與鄰近高壓輸電線間的距離作為評估家戶內部極低頻電磁波強弱的方法，但後來的實測資料顯示，這種測量暴露的方法傾向低估家戶的實際暴露情形，而會造成研究結果的偏差[3,4]。要降低因為測量工具本身所造成的測量誤差問題，則需要由改進測量工具的設計著手。當然，上述兩種造成測量誤差的原因並非完全獨立，一個測量工具如果設計不良，它也經常會使得使用者無法透過它提供正確的測量訊息。

流行病學研究經常使用的測量工具包括問卷(questionnaire)、觀察(scheduled observations)、與實驗室檢查(laboratory check)等，其中問卷經常被用於蒐集個人生活習慣、心理計測等訊息；觀察法則常用於蒐集某些行為相關的訊息，例如老人失能、兒童發育行為等；至於生理檢查、環境污染物定量、生物標記等無法由上述兩種工具測量而得的資料，則常需借助實驗室的儀器進行定性或定量的分析。研究利用問卷進行資料蒐集，其正確性經常受限於受訪者有意或無意的不正確記憶，但某些研究儘管利用問卷以外的工具來獲得研究資料，例如生物檢測(biological assays)、醫院之診斷與生化檢查資料、及勞工健康檢查與受雇資料等，但其訊息的正確性也無法獲得保證。本報告將在以下各節中，介紹並比較各種評估測量工具信效度的方法，討論使用缺乏信效度工具對研究結果所將造成的影響，並針對某些測量實務上的議題提出建議，以提供研究者參考。

投稿日期：92年12月25日
接受日期：93年3月15日

從變異量解釋測量工具之效度與信度

測量誤差與工具本身的效度(validity)以及信度(reliability)有關。測量工具的效度是指一個測量工具能夠測量到真值(true value)的程度[5]。這　所指的真值可能是一個具體的物理量(例如，血壓)，也可能只是一個抽象的概念(例如，工作壓力)。試想一個包括10個題目的量表，研究者想透過此問卷測量職場工作者的壓力程度，研究者將此問卷施測於研究樣本而得到每個人的分數，這時研究者想知道的是：研究者的分數高低是否真能反應出每個人的壓力大小？究竟此工作壓力量表能量到它該量的東西嗎？這些問題都是與測量工具的效度有關。當測量所得數值與真值愈接近時，就表示此測量工具的效度愈高。信度(reliability)則是另外一種與測量誤差有關的概念，它也稱為再現性(reproducibility)，一個測量工具的信度是指當針對一個物理量或概念進行重複測量時，這些重複測量數值能夠重複出現的的程度[6]。當然，當我們在討論測量工具的信度問題時，必須假設此測量工具所針對測量的物理量或概念在重複測量的

期間維持恆定不變。對於某一測量工具，使用者最常提出的問題是：「它準不準？」以及「它穩不穩定？」，前者指的是效度問題，後者所指的即是信度問題。

信度與效度的概念可進一步由變異數分析的角度來說明[5,6]。試想10名成年人同時接受兩種不同廠牌血糖計測量其血糖濃度，如此共得到20個血糖的數值，而可以想像的是這20個數值不太可能都相等，也就是數值之間存有變異(variance)；而造成此變異的可能原因有三：一是因為這些血糖值分別來自10名病人，二是因為這些血糖值分別得自2種不同廠牌的血糖計(即血糖計本身可能有系統性的誤差(systematic error)，第三個可能的原因則是測量工具的隨機誤差(random error)上述三個原因所代表的變異量分別為：有效變異量(valid variance，以S_V^2表示)，無效但固定變異量(non-valid but stable variance，以S_I^2表示)，以及隨機變異量(error variance，以S_E^2表示)，而三者的和則是總變異量（total variance，以S_{total}^2表示)，上述變異量間相關性的數學運算式為：$S_{total}^2 + S_V^2 + S_I^2 + S_E^2$（公式一)，這些變異量中$S_I^2$與$S_E^2$兩者與測量工具的正確性有關，但$S_V^2$的大小則是與受測者個人因素有關(每名受測者可能因為年齡、飲食、遺傳等因素上的差異而有不同的血糖值)而與測量工具本身的設計無關。而效度與信度若以變異量來定義其表示方法如下：[5]

$$效度 = \frac{S_V^2}{S_{total}^2} \text{(公式二)}$$

$$信度 = \frac{S_V^2 + S_I^2}{S_{total}^2} \text{(公式三)}$$

公式二、三顯示，當一個測量工具的隨機誤差過大時，它的信度便不好(因為$S_V^2 + S_I^2$佔S_{total}^2的百分比就會降低)，而好的信度表現是允許儀器有系統誤差S_I^2的。至於好的效度表現，非但不能有過大的隨機誤差S_E^2，也不容許有過大的系統誤差S_I^2，只有當有效變異量以佔總變異量的百分比高的時候，此測量工具才會有好的效度。根據上述定義，吾人可以歸納以下有關信度與效度之敘述[1]：

1. 高信度(低隨機變異量)，是高效度的必要(necessary)但不是充分(sufficient)條件。
2. 針對同一測量工具而言，效度的數值必定不會大於信度的數值。

因此，一個測量工具具有高的信度但它不一定有高的效度，但一個信度不佳的測量工具其效度就不能算高，因此研究者可以有信度與效度同時都高或同時都低的測量工具、也有信度高但效度低的工具，但若說一個測量工具信度低但效度高[6]，此種說法並不合乎邏輯的；試想一個血糖測量機器不太穩定，針對同一血樣測量10次，測量值有時會比真值高很多，但有時會低很多（信度低)，而將10個測量值平均後，平均值便能等於真值，如此，此儀器僅能稱為無系統偏差[1]，但稱不上是高效度，因為依據公式，信度係數值必定大於效度係數值。因此，對於隨機誤差大但其平均值卻能正確估計真值的測量工具而言，它的信度是低的（因為S_E^2大)，並且相對於有系統性誤差的測量工具而言，它的效度是「相對」較高的。

研究者在決定使用某種測量工具之前必須事先評估此測量工具的信度與效度，其目的在於瞭解此工具的信效度係數值是否可以接受，另外也可事先評估此測量工具潛在測量誤差將會對研究結果產生何種影響。當然，選擇測量工具除了考量信效度係數大小外，研究者仍須考量成本、方便性、安全性、以及民眾可接受性等因素。針對測量變項不同的資料屬性，研究者必須採用不同的統計方法評估測量工具的效度與信度。

測量工具之效度指標

依評估方法不同，效度的指標可以區分為效標關聯效度(criterion related validity)、內容效度(content validity)、表面效度(face validity)、與建構效度(construct validity) [5, 7]，其個別的作法與範例分述如下，而研究者對其所採用測量工具的效度驗證工作，可採以下一種或多種效度的驗證。

當某個測量變項之效標(criterion)或稱黃金標準(gold standard)存在時，研究者因為有一客觀值或可供比較，因此可以針對測量工

具進行效標關聯效度之評估。效標關聯效度指標的使用依測量變項的屬性而定。若量測的變項屬於常態分布的連續性變項，則吾人通常會計算真值與測量值間的皮耳森積差相關係數(Pearson ’s product-moment correlation coefficient，以r_{x}X表示，其中x為測量值，X則為真值)來量測量的效度[5,7]。例如，某研究想了解丈夫回答有關其配偶懷孕史訊息的正確性，而以妻子的回答訊息為效標，其中，研究者計算皮耳森積差相關係數以評估丈夫回答從結婚到受孕時間的正確性，其值為0.84 [8]。另一個方法則是比較量測值與真值之平均值，如果兩者的平均值相同則表示測量工具並無系統性的偏差(systematic error caused bias)，若兩者的平均值不同，則顯示測量工具存在有某種程度的測量系統性誤差，若有系統性偏差，研究者通常會計算標準性偏差(standardized bias)來表示系統誤差相對於整體變異量的程度(公式四)[1]。例如某測量血壓的工具其系統偏差為4 mmHg而假設族群之血壓分佈標準差為15 mmHg，則標準偏差值即為4/15＝0.27，我們可以說此血壓計的誤差較真值高出0.27個標準差。

標準化偏差=(測量平均值−真值均值)/(真值分布的標準差)(公式四)

雖然研究者經常使用皮耳森積差相關係數作為評估連續性測量值與效標之間的一致性，但皮爾森積差相關卻有可能對效度的評估產生誤導；因為只要兩組數據的散佈圖(scattered plot)落在同一直線上，它們的皮耳森積差相關係數便會是1，但事實上當測量工具有系統誤差時，測量值與效標之散佈圖也會顯示一斜率不等於1或未通過原點(即截距不為0)之直線，但此兩組數據並不完全一樣，因此也有人建議應避免使用皮耳森積差相關係數作為效標效度的指標，而應使用組內相關係數(Intra-class Correlation Coefficient，簡稱ICC)，ICC可以避免上述利用皮爾森積差相關係數評估相關性所碰到的問題[5]。某研究利用儀器針對北台灣407個住家內部進行極低頻磁場的實測工作(此測量數據可視為效標)，該研究並同時根據住家鄰近電力設備的資料利用物理學公式計算磁場暴露的理論值，研究者計算兩種數據的ICC(其數值為0.90)，來評估利用理論值來替代實測值的可行性[3]。由於計算ICC的前提是兩個前後數值是隨機值(random readings) (即任一筆資料前後數值可以對調而不影響ICC數值)而非明確不同的數值(distinct readings)，而後者通常則是實際的情況，例如，前後兩個數值可能是同一工具前後不同時間的測量值，也有可能是來自兩種不同工具的測量值)，這是使用ICC的限制[9]，因此，有研究者建議計算Concordance Correlation Coefficient來避免ICC這項缺點[9,10]，也有研究認為應先以散佈圖(scatter plot)來呈現兩組數據相對於座標圖上斜率為1通過原點直線之散佈情形，然後再計算兩數據差之平均值及其信賴區間，用以呈現兩組數據之一致性[11]。

若測量變項屬於二分類別變項，敏感度(sensitivity)與特定度(specificity)則是常用的效標效度指標[1]。敏感度與特定度均為條件機率，從篩檢的角度而言，前者是指一個真正有病的人其篩檢結果亦呈現陽性的機率，後者則是指一個沒有病的人其篩檢結果也呈陰性的機率。例如，社區調查經常利用自訴症狀作為估計族群疾病率的基礎，為了解自訴疾病資訊的正確性，某研究以北台灣228名65歲以上老人為對象以面對面訪視方法請研究對象自訴其罹患高血壓、心臟病、以及糖尿病的情形，隨後此228名老人即被安排接受後續的臨床診斷(可視為效標)。研究數據顯示：高血壓、心臟病、以及糖尿病的敏感度(%)分別為20.5、49.3與66.7；特定度(%)則分別為90.0、82.8與95.2，顯示利用老人自訴上述症狀有無作為疾病率估算的基礎將會低估實際的疾病率[12]。在前述研究中，因為效標與測量同時存在，因此，此效標關聯效度又可稱為同時效標效度(concurrent criterion validity)，如果效標必須是在測量完成後的一段時間後才會出現或獲得，則所評估之效標關聯效度則是稱為預測效度(predictive validity)。

除了敏感度與特定度外，某些研究也利用kappa係數來進行類別性變項之效標關聯效度評估，kappa係數是一個校正隨機因素後用

以評估類別變項一致性的指標(chance adjusted agreement) [13,14]，例如前述比較北台灣407個住家內部極低頻磁場強度理論值與實測值的研究[3]，研究者將磁場強度區分為3組(＜1 mG、1-2 mG、與＞2 mG)，並計算kappa係數，用以評估理論值與實測值之間的一致性，然而，也有人認為kappa係數是是一種用以評估一致性或再現性的信度指標，並不適宜作為效度評估之指標[15]。在有些情況下效標並不存在，此時研究者可以採用內容效度或建構效度的方法進行效度評估。

內容效度(也稱為抽樣效度(sampling validity)通常是針對測量工具內容的適切性與代表性進行評定，內容效度的建立通常是從選擇適當的題目開始，例如，我們會很關心一份統計學考試試卷是否真能測驗出學生的統計學知識測驗，要了解此試卷是否具有內容效度，我們會徹底而有系統的去檢視此試卷是否涵蓋了課程中所有教授的主題，內容效度的評估通常藉由該領域專家來進行，稱為專家效度(expert validity)評估。例如某個調查餐飲服務從業人員肌肉骨骼疼痛(musculoskeletal pain)的研究設計了一份問卷量表，想利用此量表評估一名餐飲服務從業人員肌肉骨骼疼痛，該研究於發展量表之初便透過數次的專家會議，討論量表的內容的適切性與周延性包括：接受評估之身體部位、自覺疼痛嚴重程度之序位分數擬定、以及評估的時間等[16]。在內容效度的評估中，有時研究者會針對量表中的題目請專家就其題意的適當性與清楚性進行逐題評分，然後計算各題的平均得分，以評估每個題目包括於量表中的適切性。至於表面效度(face validity) (或稱為邏輯效度(logical validity)則是指對於受測者、使用測量工具者、或其他未曾接受的觀察者而言，測量工具是否能獲得正確的測量值[7]？在某些情況下，有些測量工具是否能從受訪者中獲得正確的訊息是淺而易見的，例如，利用醫院產科的紀錄蒐集而得之嬰兒出生體重其正確性會比由母親回憶小孩子的出生體重來的高；又如針對新生兒作神經生理檢查會比單純用新生兒出生體重更能判讀新生兒是否未盡成熟(premature)。

建構效度則是另一種當效標不存在時經常被使用的另一種效度指標，特別是在發展一個涵蓋多個抽象構面(construct)量表發展過程中，建構效度是一種經常被使用的效度指標。以國內發展中文版聯合國世界衛生組織生活品質量表(WHOQOL－100)的研究為例，研究者首先將WHOQOL－100原始問卷翻譯為本國文字後，然後利用探索性因素分析(exploratory factor analysis)與驗證性因素分析(confirmatory factor analysis)的方法，驗證原量表將96個核心題目區分為6個範疇(生理、心理、獨立程度、社會關係、環境、以及靈性／宗教／個人信念)、24個層面的架構是否合理[17,18]，分析結果發現：探索性因素分析總共抽取出21個因素，而這些因素也大致上反映出原先的架構並解釋了約67%變異量；而後續針對每個範疇內題目所作的探索性因素分析，其結果也大致反映出量表所設計的層面架構。驗證性因素分析的結果也顯示：6個範疇的模式均呼應了問卷所設計的潛在結構，其CFI值介於0.803與0.940之間，這些訊息顯示中文版WHOQOL－100具有不錯的建構效度。由於因素分析需要繁複的計算[19]，通常需要藉助統計軟體(如LISREL)來完成計算。

測量工具信度的指標

評估測量工具的信度可以是由兩位觀察者針對同一個被測量事物測量兩次，即所謂inter-rater reliability，也可以由同一個觀察者針對同一個被測量事物在不同的時間點上各測量一次稱為intra-rater reliability，由於同一個被測量的事物被測量了二次，因此此類的信度評估被稱為再測信度 (test-retest reliability)；但在某些情境下，研究者並無法執行再測信度的評估，因此只能執行一次測量，此時研究者則可採行另外一種信度評估的方式，即計算測量工具的內部一致性(internal consistency)[20,21]。

在test-retest reliability的評估中，若所測量的變項是屬於數值呈常態分佈的連續性變項，則皮爾森相關係數與ICC是常被用來評

估信度的指標，而若研究所牽涉的變項為類別性的變項，則kappa係數則是常被用來作為信度的指標。但如同前述，皮爾森相關係數並無法反映出兩組數據間平均值的差異，因此，有時大的皮爾森相關係數並不能代表高度的測量再現性。而ICC較皮爾森積差相關係數優越之處在於只有當兩組數列完全相同(即兩組數列的散佈圖成一斜率為1並通過原點的直線)時ICC的數值才會等於1.0[5]；不過當主要的變異量都來自隨機誤差(系統誤差很小)時，使用ICC或皮爾森積差相關係數並沒有很明顯的差別。

kappa統計值則是經常被使用於評估類別變項信度的方法[13,14]。某一研究探討我國主計處所訂定中華民國標準職業分類系統的信度[22]，該職業分類系統將職業區分為10大類，37中類，114小類，以及394細類。每一個職業以4碼表示依前後順序分別代表該職業的大中小細分類。該研究的兩位研究者利用此一分類系統獨立將145份問卷中載有描述受訪者工作內容的文字轉譯為4碼的職業代碼，並比較分析兩位研究者譯碼的一致性。要注意的是如果兩位研究者的分類結果一致，我們並不能排除是因為隨機造成，因此無法將兩者分類觀察所見的一致性(observed agreement percentage)完全歸因於研究者使用分類系統本身的信度，而應該將隨機造成的一致性從其中扣除，該研究分析發現：兩位研究者大分類譯碼一致者共110筆(觀察所見的一致性為75.8%)，而相對應的kappa係數為0.70。而大、中、小細分類譯碼均一致者則減少為98筆，觀察所見一致性百分比為68.6%，而kappa係數則為0.64。如果測量變項有3個或3個以上的序位層別，研究者也可以計算weighted kappa係數將相鄰的層別也計算部分的一致性(partial agreement)[14]，例如前述比較北台灣407個住家內部極低頻磁場強度理論值與實測值的研究[3]，研究者將磁場強度區分為3組(＜1 mG、1-2 mG、與＞2 mG)，而研究者將理論值與實測值分屬兩個相鄰層別(如一個屬於＜1 mG，另一個屬於1-2 mG)的觀察值給予0.5的加權量，並計算其weighted kappa係數，由於考慮了partial agreement，因此weighted kappa係數值將比kappa係數值要大一些。

執行再測信度的一個問題是，兩次測量的時間間隔要多久，時間太短受試者極可能還記得前次的回答，恐無法反映出量表本身的信度，但若時間太長，則恐怕測量變項的真值改變了，因此再測信度的實施，研究者要根據所測量的變項屬性謹慎選擇兩次測量的間隔時間。對於測量工具如果僅執行單次測量，則研究者可進行工具內部一致性信度的計算。

內部一致性信度係數可分為折半信度係數(split-half reliability coefficient)、Kuder-Richardson-20 (KR_{20})與Cronbach's alpha (Cronbach's α)。評估折半信度的作法是將量表中所有的題目(items)以隨機的方式分為兩部分，並計算此兩部分個別計分之皮爾森積差相關係數(r)；由於量表的信度會與量表所涵蓋題目的數目成正比，因此，這種方法所得到的相關係數將會是一個被低估的數據，因為實際應用量表時，量表題目會增加為原來的兩倍。為解決此問題，Spearman－Brown提供了校正的公式(公式五)[5]。

$$r_{S-B} = \frac{kr}{1+(k-1)r} \text{ (公式五)}$$

其中r為所觀察得到的相關係數，k為新量表項目數與原量表項目數的比值，若為折半信度，k值為2，所計算出來的信度稱為Spearman-Brown折半信度(Spearman-Brown split-half reliability coefficient)。利用Spearman－Brown公式，研究者也可以預估量表所需的題目數，例如，原始量表的信度為r，若研究者想將此量表的信度提高至R，則所增加題目的倍數k則是：

$$k = \frac{R(1-r)}{r(1-R)} \text{ (公式六)}$$

另一種折半信度係數稱為折半α值(split-half alpha(α))，其計算方法為：假設一個量表施測給n個人，我們可以計算此n個人整份量表得分分佈的變異數(variance)，以S^2表示，接下來，我們可以利用隨機的方式，將整份量表折半為兩部分，再分別統計並計算此n名受測者在這兩部份得分的變異數，分別

以S_{y1}^2和S_{y2}^2表示，而折半α值便等於：

$$Split-\alpha=\frac{2\left[S_x^2-\left(S_{y1}^2+S_{y2}^2\right)\right]}{S_x^2}\text{(公式七)}$$

不過，必須另外說明的是， Spearman-Brown折半信度與折半α值的使用時機多在當量表分數是連續變項的時候[5]。

KR_{20}與Cronbach's α的計算公式請見其他相關文獻[5,23]。前者適用於當量表題目提供之選項為二分變項資料（dichotomous data），例如「對與錯」或「是與非」，而後者則是適用於量表題目屬於序位選項時，例如李克氏五分或七分量表。某研究探討工作者的壓力與工作滿意度與非死亡意外事故危險性的關係，其中壓力量表涵蓋99個題目，而工作滿意度則包括12個題目，每個題目均採1-5分計分，測量壓力與工作滿意度的程度高低，研究者計算兩個量表內部一致性信度(以Cronbach's α表示)，分別為0.98與0.90，顯示兩量表都有不錯的內部一致性[24]。某些統計軟體提供了計算上述的計算功能，以上述工作壓力與工作滿意度的研究為例，研究者可以利用SPSS 11.5版軟體的Analyze選項，進入Scale之Reliability Analysis之中，然後在Model中選擇Alpha，後將量表各題目變項選入items視窗，此時軟體將計算出Cronbach's α (因為每個item為5分量數)，如果所選入的變項屬於二分變項資料，則所計算出來的數值則相當於是KR_{20})。

測量誤差對研究的影響

如果一個測量工具的信效度未達盡善盡美，則其測量結果便會有誤差，而此測量誤差(measurement error)可以區分為隨機誤差(random error)或非隨機誤差(non-random error)，後者也稱為系統性誤差(systematic error)或稱為偏差(bias)。隨機誤差常是由於測量工具的信度不佳所致，它常造成測量工具的不精確(low precision)，但在經過數次測量後其平均值並不會高估或低估真值（true value），因此透過多次測量並求取測量值平均數的作法可以解決低信度工具所造成隨機誤差的問題；但測量工具的偏差則會造成測量結果有偏差(即高估或低估真值)，且無法透過增加進行多次測量求取平均值的作法來解決問題，解決測量工具偏差的方法，會依工具本身的種類而有不同，例如，儀器(如血糖計、體重計等)可以透過校正的方法確保效度，而研究者通常會透過增加、刪除、或修改問卷部分內容來增進問卷的效度。就探討暴露與疾病相關性的流行病學研究而言，測量誤差會造成暴露或疾病的錯誤分組(misclassification)，而此錯誤分組又可以區分為差別性的(differential)或無差別性的(non-differential)；差別性的錯誤分組(differential misclassification)表示測量工具因為測量誤差而對於某一變項的歸類產生錯誤分組的方向與程度會隨著另一個變項的數值不同而有差異；而無差別性的錯誤分組(non-differential misclassification)則是指某一變項錯誤分組的方向與程度不會隨另一個變項的數值不同而有差異[1]。對於暴露與疾病均為二分類別變項的情況時(binary variable)，差別性的測量誤差將會使得相關性被高估或低估，而無差別性的測量誤差則通常只會讓相關性被低估而往虛值(null values)靠近[21,25,26]。但當世代追蹤研究中疾病的診斷被低估(即指有真正罹病個案未被發現，而沒有非罹病者被誤診為有病者)所造成的無差別性錯誤分組並不會影響相對危險性估計值[27]；而上述的情況也僅能應用於2×2表中，如果暴露或罹病為3層或3層以上則情況的變數更為複雜，此時無差別性的錯誤分組也不一定會低估實際的相對危險性[28]。一個病例對照研究針對老人探討研究個案先前居住與工作場所的電磁波暴露是否與日後發生認知功能障礙的危險性有關[29]，此研究最後並未發現兩者間存在有相關性，此結果所反映的可能事實有二：其一是職場或住家的電磁波暴露與老人是否發生認知功能障礙原本就無關，其二則是電磁波暴露與老人發生認知功能障礙的危險性有關，但因為研究者分別利用老人現住場所以及詢問老人家人有關老人過去所從事的主要職業來評估老人在居住與工作場所中所遭受的電磁波暴露狀況，而一個人通常會因為遷徙而改變居住場所，一天當中也不是都停留

在家中，而相同職業別的人也不一定就有相同的電磁波暴露狀況，因此，作者的這種暴露評估方法可能會將原本是高磁場暴露者誤歸為低暴露者，也有可能將原本是低暴露者誤歸為高暴露，並因此而造成暴露的無差別性錯誤分組(因為此錯誤分組發生在病例組與發生在對照組的情形是一樣的)，致使研究結果產生偏差，使得事實上存在的相關性被低估或甚至被掩蓋了。

前段討論所提及的暴露與疾病均為類別性變項，而若是暴露與疾病均為連續性變項，則其相關性是否會因為測量誤差而被低估則要視測量值與真值間的相關性而定，若是暴露或疾病變項的測量值與真值間的相關係數為1.0 (即百分之百正相關)，則研究的相關性便不受影響；但若測量值與真值間並未呈現百分之百的正相關，則暴露與疾病間的相關性將會被低估[1]。

上述討論只有在假定暴露與疾病兩者之一有測量誤差，而另一變項則無測量誤差發生，但在實際的情況下兩者均發生測量誤差的可能性很高(程度或有不同)，而此情況將使得研究結果產生偏差的程度更大。對於類別屬性的暴露與疾病而言，無差別性錯誤分組所造成研究結果偏差的程度除了與測量誤差的程度有關之外，也與暴露或疾病的盛行率有關，在固定的測量誤差下，無差別性錯誤分組所造成相關性被低估的程度將在暴露或疾病的盛行率遠離50%時達到最大；而對於連續性的暴露與疾病變項而言，無差別性測量誤差對暴露與疾病相關性降低的影響則隨測量值與真值間相關性愈低而降低愈大[1]。

一些實務問題的討論

在討論完測量誤差的來源、評估方法、與對研究結果可能的影響後，本報告最後要針對一些信效度評估相關的議題進行簡要的討論，其中包括：(一)何時應該進行測量工具信效度的評估？(二)信度與效度要多高方能為研究所用？(三)降低測量隨機誤差的方法有哪些？(四)如何考量信度與效度係數的估計誤差？以及(五)研究者應釐清測量工具的信效度與研究結果的信效度。

一、何時進行測量工具信效度的評估？

研究者必須盡可能在正式研究之前執行一個先驅研究(a pilot study)，並在此先驅研究中針對測量工具的信效度進行評估，此先驅研究的樣本數可由數人到數十人，依研究條件與可行性的考量，選擇一種或數種效度或信度的評估方法，若是研究屬於發展量表的研究，則其信效度的評估便需較為完備，例如前述有關發展中文化世界衛生組織生活品質量表之研究；但若屬於使用既有量表或因施測對象不同而稍為修改既有量表之研究，則僅需進行簡單信度與效度的評估。研究者應了解，同一種測量工具當它應用於不同族群或在不同情境之下應用於同一族群都可能會產生不同的效度與信度。

二、信度與效度要多高方能為研究所用？

目前文獻上並無一致性的意見說明何種程度的信度或效度是可以被接受的。就信度而言，Kelly建議一測量工具至少要有0.94的信度，而Weiner與Stewart則建議測量工具的信度應有0.85[5]。雖然對於可被接受的信度有不同的意見，但一般咸信，當測量工具要應用於個人的測量時，它需要有較高的信度係數；而當它應用於群體時，它的信度值可以較低，此論點的理由是當工具針對群體測量，研究者通常會計算平均值，而群體樣本數則可以在比較群體間平均值時降低測量誤差造成的影響[1]，例如與10人的樣本相較，1000人的樣本能容忍的隨機測量誤差則較大。至於效度值要多高才能為研究所用？也沒有大家都能同意的答案，但若以敏感度與特定度而言，試想，有人利用擲銅板的方式來進行疾病篩檢，如果此銅板是公正的，那幾乎是以隨機的方式來判讀篩檢的結果，如此，擲銅板此一「篩檢工具」的敏感度與特定度都將會是0.5，因此一個篩檢工具的敏感度與精確度的和至少應大於1.0，否則我們寧可選擇公正銅板來作為疾病篩檢的工具。

三、降低測量隨機誤差的方法

增進信度可從降低隨機變異（random variance)與增加真正變異(true variance)兩方面著手，在降低隨機差異部分，多數人認為可從測量者的訓練著手，具體的作法則是從訓練過程中找出某些經常會出現測量結果不一致的測量者，而將此測量者排除於研究之外[30]。此外，採用多次測量（multiple measurements)，也是降低隨機誤差增加測量信度的方法，對於連續性變項而言，計算多次測量的平均值可以有效降低因為單次測量所造成的隨機誤差；而對於類別性變項而言，則可以求取一致性的測量結果(concord)作為最後分類的依據[1]。至於在增加真正變異方面，可在題目的選項中增加回答者答案之變異性，例如可由原來的劣—中—優選項更改為劣—差—中—良—優，如此將可增加受訪者答案之差別性，另一種增加信度的方法則是將問卷施測於一異質性較高的族群，或增加問量或量表的題目，如此均能增加工具的信度。

四、信度與效度估計誤差

無論前趨或實際的研究都是基礎於一個樣本，因此所計算所得的信度與效度係數均應被視為一點估計值，而研究者宜提供母數信賴區間的估計，讓讀者能夠了解信效度估計誤差的大小。前述比較北台灣407個住家內部極低頻磁場強度理論值與實測值的研究[3]，研究者將磁場強度區分為3組(＜1 mG、1-2 mG、與＞2 mG)，並得到理論值與實測值間kappa係數為0.64，而其95%信賴區間則為0.50-0.78。

五、測量工具的信效度與研究結果的信效度

測量工具的信度與效度是有別於研究結果的信度與效度，一個研究結果的效度佳指的是該研究結果沒有因為選樣、測量、或干擾(confounding)問題而造成研究結果的偏差(bias)，因此測量工具的信度與效度未盡完善將會造成一個相關性研究結果產生偏差。而研究結果缺乏信度則是指該研究母數之信賴區間估計不夠精確(low precision)，這通常是因為研究樣本數不足所致，而與研究所使用測量工具的信效度程度無關。

結　論

研究者在使用測量工具前有必要選擇適當的方法評估其相關之效度與信度，以了解使用該測量工具對研究結果的可能影響，如此方能對研究結果作正確的闡釋。

致　謝

本研究部分研究經費由國科會計畫所支助(NSC-92-2320-B-030-010)。

參考文獻

1.Kelsey JL, Thompson WD, Evans AS. Methods in Occupational Epidemiology. New York: Oxford University Press, 1986: 285-308.
2.Rookus MA, van Leeuwen FE. Induced abortion and risk for breast cancer: reporting (recall) bias in a Duch case-control study. J Natl Cancer Inst 1996;**88**:1759-64.
3.Li CY, Thériault G, Lin RS. A validity analysis of residential magnetic fields estimated from high-voltage transmission lines. J Expo Anal Environ Epidemiol 1997;**7**:493-504.
4.Thériault G, Li CY. Risks of leukemia among residents close to high voltage transmission electric lines. Occup Environ Med 1997;**54**:625-8.
5.Streiner DL, Norman GR. Health Measurement Scales－A Practical Guide to their Development and Use. New York: Oxford University Press, 1989; 79-95, 106-25.
6. 王榮德：流行病學方法論－猜測與否證的研究。第2版。台北：國立台灣大學醫學院出版委員會，1990；67。
7. 危止芬：心理測驗。台北：雙葉書廊有限公司，1999；74-151。
8.Coughlin MT, LaPorte RE, O'Leary LA, Lee

PA. How accurate is male recall of reproductive information? Am J Epidemiol 1998; **148**:806-9.

9.Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;**45**:255-68.

10.Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues, and tools. J Am Stat Assoc 2002;**97**:257-70.

11.Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; **1**:307-10.

12.Wu SC, Li CY, Ke DS. The agreement between self-reporting and clinical diagnosis for selected medical conditions among the elderly in Taiwan. Public Health 2000;**114**:137-42.

13.Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; **20**:37-46.

14.Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;**70**:213-20.

15.Maclure M, Willett W. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;**126**:161-9.

16. 全中妤、杜宗禮、葉文裕、李中一：台灣旅館業餐飲人員工作動作特性與肌肉骨骼傷病之橫斷式研究。台灣衛誌 2002；**21**：140-9。

17. 台灣版世界衛生組織生活品質問卷發展小組：台灣版世界衛生組織生活品質問卷之發展簡介。中華衛誌2000；**19**：315-24。

18. 姚開屏：台灣版世界衛生組織生活品質問卷之發展及使用手冊。台北：世界衛生組織生活品質問卷發展小組， 2001；12-17。

19. 林清山：心理與教育統計學。第9版。台北：台灣東華書局股份有限公司, 1999；620-53。

20.Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975;**31**:651-9.

21.Newell D J. Errors in the interpretation of errors in epidemiology. Am J Public Health 1962;**52**:1925-28.

22. 李中一、張恭賀、馮兆康、吳淑瓊：職業別譯碼之一致性分析。中華衛誌 1999；**18**：255-61。

23.Kaplan RM, Saccuzzo P. Psychological Testing : Principles, Applications, and Issues. Pacific Grove, Calif. : Brooks/Cole Pub. Co., 1989;85-135.

24.Li CY, Chen KR, Wu CH, Sung FC. Job stress and dissatisfaction in relation to the development of nonfatal injuries on the job among a cross-sectional sample of petrochemical workers. Occup Med 2001;**51**:50-5.

25.Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 1977;**105**:488-95.

26.Gullen WH, Bearman JE, Johnson EA. Effects of misclassification in epidemiologic studies. Public Health Rep 1968;**83**:914-8.

27.Rothman KJ. Modern Epidemiology. Boston: Little, Brown and Company, 1986; 87.

28.Correa-Villaseñor A, Stewart WF, Franco-Marina F, Seacat H. Bias from non-differential misclassification in case-control studies with three exposure levels. Epidemiology 1995;**6**:276-81.

29.Li CY, Sung FC, Wu SC. Risk of cognitive impairment in relation to elevated exposure to electromagnetic fields. J Occup Environ Med 2002;**44**:66-72.

30.Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. Med Edu 1980;**4**:345-9.