健康相關生活品質概念與測量原理之簡介

姚開屏
國立台灣大學理學院 心理系

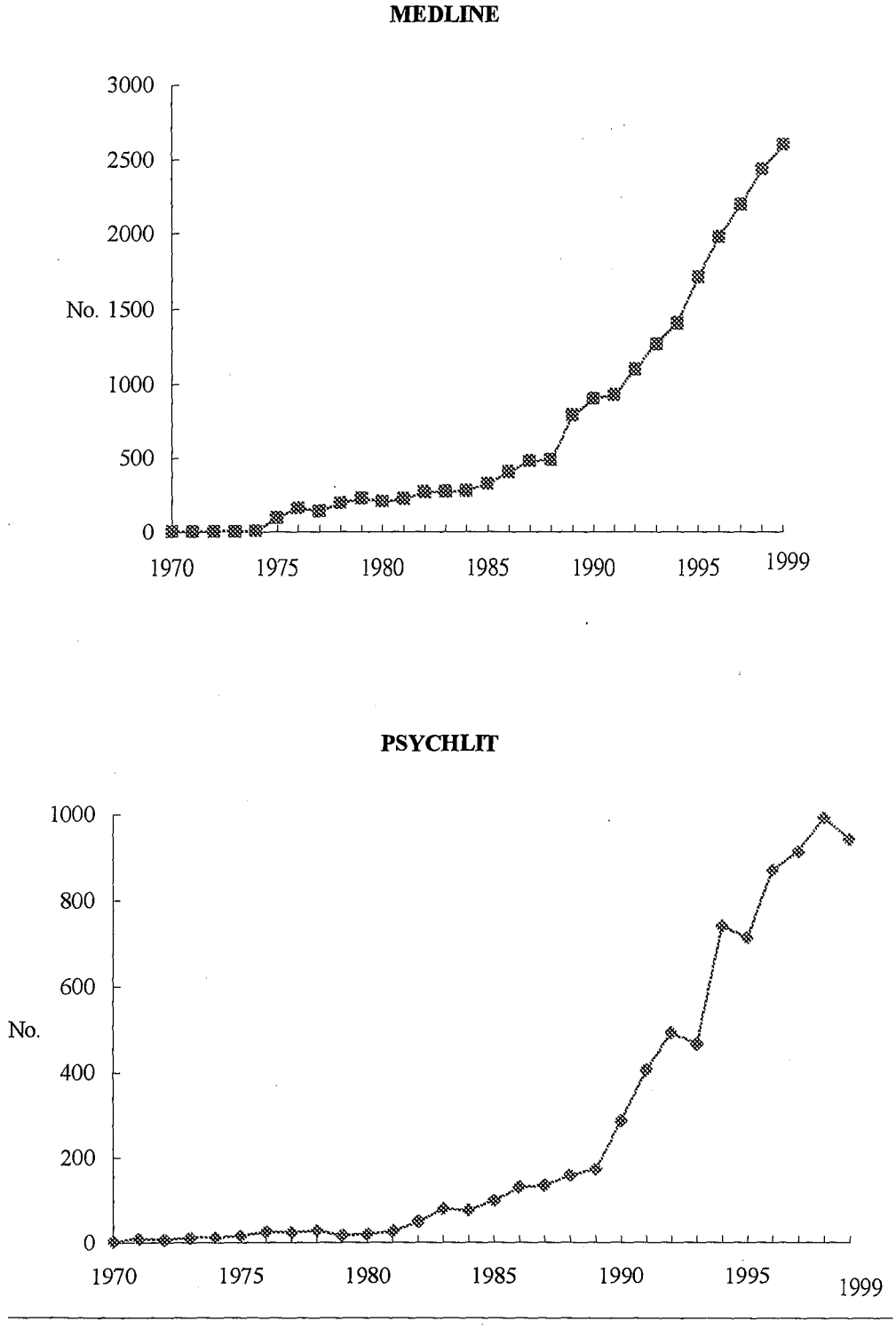
健康相關生活品質的概念

「生活品質」這個概念最早可說是由亞理斯多德(Aristotle)所提出來的，他是從「快樂(happiness)」的角度來看生活品質，認爲快樂是上帝所恩賜給人的，是一種貞潔的心靈活動，因此快樂的人可以活得好、事情也做得順利[1]。傳統華人對這方面的看法是從陰陽調和的角度來看，認爲生活中的各種事物可分爲陰與陽，若一個人能順天地之運行法則而生活、陰陽能調和，則就可長壽、就有好的生活品質[1]。「生活品質」一詞正式出現在美國的詞彙中是要到第二次世界大戰之後，當時所強調的生活品質是有好的生活，而不單只是物質上的滿足而已[2]。從研究的角度來看，心理學家及社會學家很早就已經涉入了相關的研究，他們早期所常使用的詞爲「幸福感(well-being)」、「主觀幸福感(subjective well-being)」、「心理幸福感(psychological well-being)」、「快樂(happiness)」、「生活滿意度(life satisfaction)」等。這些詞所隱含的意義不外乎是從個人正負向的情緒、主觀認知的層面、以及身心健康的角度來評估一個人整體的生活情形。隨著時代的變化、社會及經濟的發展，以及醫療水準的提升，「生活品質」這個詞越來越被用的多，且其相關的研究也有越來越被重視的趨勢。我們從醫學及心理學界所常用的資料搜尋庫MEDLINE及PSYCHLIT，以quality of life (QOL)爲關鍵字來統計過去三十年來提及此關鍵字的論文篇數，也可以看出與生活品質相關的研究始於1970年代中期，至1980年代末期則大幅攀升，且有越來越多的趨勢(圖一)。生活品質研究所涉及之範圍相當廣泛，不同領域的學者對生活品質有不同的定義或看法，例如：經濟學家、醫療學者、心理學家對生活品質的定義及其研究之著重點就有不同，本篇只打算將重點放在「與健康相關的生活品質」(health-related quality of life, HRQOL)上。簡而言之，在區分上若「生活品質」指的是個體對生活中自認對其重要部分的滿意程度[3]；則「健康相關生活品質」指的是個體對生活中受到健康而影響之重要部分的滿意程度[4]，這要與常被拿來取代作爲測量健康相關生活品質的「健康狀態(health status)」的好壞程度來做區別，後者指的是個體在生理失能、疾病症狀、功能損失方面之相對健康程度，也就是說後者只從生理狀況的角度來反映生活品質，在包含範圍上是比較狹隘的。本篇從以下五個「W」來談健康相關生活品質的概念。

WHY？：爲什麼重視健康相關生活品質的研究

這個議題被重視的原因，是與醫療水準的提升有關。Lohr 曾對醫療照護結果之測量分成五個等級，即所謂的 5D：死亡(death)、疾病(disease)、殘障(disability)、不適(discomfort)、不滿意(dissatisfaction)，其中死亡以死亡率(mortality)的計量爲代表，疾病、殘障、不適則可合爲以罹病率(morbidity)的計量爲代表[5]。由於醫藥衛生的發達，使得人類生命歲數延長，疾病型態由過去的傳染性疾病演變至今以慢性疾病爲主，因而死亡率(mortality)或罹病率(morbidity)的多寡不再成爲能代表生活品質好壞的指標；另外醫療成本日漸增加，醫療資源付出者期望所付出的是最具有醫療價值及效果的；並且目前的治療方式越來越重視病人個體的主觀感受(subjective perception)，致使醫療及經濟學者開始探討以病人爲中心，能測出健康相關生活品質療效的工

Title: Introduction of the EORTC Disease-Specific Quality of Life Questionnaires for Cancer Patients
Author: Kai-Ping Yao, Department of Psychology, National Taiwan University
Key Words: quality of life, concepts, measurement



圖一：以「quality of life」爲關鍵字查詢 MEDLINE 及 PSYCHLIT 中論文篇數

具，這種現象也促使過去以醫生測量病人生理狀況之變化爲主，卻未能真正反應病人的感受及實際功能之方式，演變至目前能從病人主觀的(subjective)感受及多向度的(multidimensional)的角度來測量生活品質。這也反映出來生理狀況之測量只能算是測中介的(intermediate)生活品質結果，健康相關生活品質的測量才是最終(final)生活品質結果的代表。

WHAT？：什麼是健康相關生活品質

不同領域的學者對「生活品質」有不同的看法及評量方式，例如經濟學家是從國民生產毛額(Gross National Product, GNP)的角度來看生活品質[6]；心理及社會學家認爲生活品質是個人對其婚姻、家庭生活、朋友關係、生活水準、財務及宗教等向度的滿意度，也是指個人與環境互動中，對壓力處理的能力、社會支持的程度，以及自我評價的程度[7,8]；臨床醫學專家認爲生活品質是指疾病患者接受治療前功能損害及治療後所獲得效益的程度[9]；護理學家則認爲生活品質乃個人對幸福的感受以及對生活中自覺重要的部分之滿意程度[10]。由此可知不同背景的人對生活品質會有不同的定義及看法，並沒有一個統一的標準。不管是何種定義，研究者宜根據自己的研究目標來下定義，例如：醫療公衛學者較會採用與健康有關的生活品質定義，經濟學家則會採用以經濟指標爲主的定義。世界衛生組織對生活品質的定義則爲「個人在所生活的文化價值體系中的感受程度，這種感受與個人的目標、期望、標準、關心等方面有關。它包括一個人在生理健康、心理狀態、獨立程度、社會關係、個人信念以及環境六大方面」[11-17]，此定義強調個人於自己文化下之主觀感受的重要性。

依循生活品質研究方面的歷史發展脈絡，當我們談到生活品質時，可用不同的分類方式來看此概念。我們可將「生活品質」分成一般生活品質(global QOL)及健康相關生活品質(health-related QOL)。一般生活品質強調個人在所處的環境中，對一般廣泛性的生活各方面之滿意度，此方面常是由對一般大眾個人的主觀感受來評斷；健康相關生活品質則強調因爲疾病、意外或治療所導致個人身體功能改變進而影響個體在心理、社會層面健康相關生活品質的改變，可由主觀判斷及客觀測量來評量，這方面的生活品質是本篇所要談的。「生活品質」也可分成單向度(unidimensional)生活品質及多向度(multidimensional)生活品質。單向度生活品質是指人們的生活品質只由單一個整體因素所影響，早期的生活品質研究比較是採這種主張；多向度生活品質則是從數個向度(例如：生理、心理、社會)來看人們的生活品質，目前的學者則比較贊同此種看法。世界衛生組織於其所發展的簡明版問卷，則是從四個向度來看生活品質，即生理健康、心理、社會關係及環境。另外，「生活品質」又可分成客觀性(objective)及主觀性(subjective)。客觀性的生活品質是從他人的角度來看生活品質，例如醫生測量病人的生理癥狀來表示病人的生活品質，政府官員使用客觀經濟指標來代表百姓的生活品質，早期生活品質的研究皆是採用此觀點；主觀性的生活品質則是從自己個人的角度來看生活品質，強調個人主觀的感受。隨著研究的發展，我們雖然不能完全否定客觀性生活品質的重要性，特別是在做公共議題的研究時(例如：資源的分配、經濟方面之跨國比較)，但研究者已經越來越重視主觀性的生活品質的重要性，從世界衛生組織對生活品質的定義即可看出。

WHEN？何時需做健康相關生活品質測量

我們在此提出數個使用生活品質測量的時機：(1)醫藥療效分析：從治療前後生活品質的變化情形，可以反映出這個治療對生活品質各方面的影響效果，因此治療前後生活品質變化情形的測量，已經成爲申請新藥上市時，會提供的研究報告參考內容之一部分。(2)臨床醫療決策分析：對採用不同治療方式後之生活品質結果之分析，可提供給病人及醫療人員做爲選擇適當的治療方式之參考依據。例如：對癌症的不同治療方式會帶給病人在生活品質不同方面之不同影響，據此病人可與醫療人員商量選擇較適當的治療方式。(3)衛生政策評估、分析、擬定：健康相關生活品質的評量可作爲健保及醫療資源分配之依據，例如：依照不同治療方式所帶給病人

不同生活品質的程度來訂出不同的治療給付價格，不過這方面所涉及的醫療倫理議題，仍需要專家學者們深入的研究與討論，以達到全民之共識。(4)健康風險評估與管理：公司或機構中依據對員工之生活品質測量的結果，來瞭解公司或機構內所潛藏的影響健康之因素，並進而改善及管理其環境與設備，使能增進員工之健康。

WHOSE？以誰的健康相關生活品質觀點爲主

要回答此問題乃視測量之目的而定。由於最終可能會影響健康醫療資源的分配，因此我們需要考慮以下幾種人的意見：(1)病人：個人主觀的感受是最真實且實際的，從這個角度來瞭解病人的感受是最直接的。對於年紀太小、無病識感(如精神病人)或意識喪失的病人，則可能要詢問其最近親屬或直接照顧者，以對他們的生活品質做近似之測量；(2)醫療專家：他們所提供的生理指標或數據，可作爲病人健康相關生活品質之佐證；(3)醫藥衛生政策制訂者：其成員常常具有醫療相關背景，且具有較宏觀的角度及制訂法令的權力，理想上可顧及社會公平正義性；(4)一般大眾：由於醫療資源是來自於大眾也用之於大眾，因此當涉及醫療資源分配時，必須要納入大眾的聲音。針對使用生活品質測量之不同目的，所要參酌的意見就不同，例如：若是只要瞭解病人對生活品質的感受，則只需要參考病人的意見即可；若是要找出一種適當的治療方式，則除了參考病人的感受外，醫療專家的意見也是不容忽視的；若是要分配醫療資源、對醫療照護做定價，則還需參酌醫藥衛生政策制訂者及一般大眾的意見。

HOW？如何測量健康相關生活品質

健康相關生活品質的測量從類別來分，可分成一般性(generic)測量及特定疾病性(disease-specific)測量。一般性測量乃測量大家所共通的生活品質部分，其測量之結果可用來比較不同族群(例如：不同種族、不同疾病)間的差異，但卻較不能用來瞭解特定族群所特別關注的生活品質；特定疾病性生活品質的測量則剛好相反，能深入瞭解特定族群的生活品質，但卻較難用來做跨族群間的比較。另外，健康相關生活品質的測量，可使用一個題目來含蓋所有生活品質的部分，例如：「您認爲您的生活品質如何？」，此方式的優點是非常簡單及方便，但卻無法測到生活品質的多面性；多題測量則是在同一量表內使用的多個題目來達到測量的目的，此方法可測生活品質的多面性，但此類量表需考量信、效度的檢驗、題數之多寡會影響回答者的回答意願等。再者，健康相關生活品質的測量從編製量表的方法來分，可區分成以經濟學(economics)爲基礎的總計指標法(aggregated index)及以心理計量學(psychometrics)爲基礎的健康剖面法(health profile)。前者通常將量表得分轉換成單一數值的指標分數來代表整體的健康相關生活品質情形，此方法是具有效用(utility)或偏好(preference-based)的測量方式；後者則通常對量表各向度有各個不同的數值產生，因此是從多向度的角度來描述整體健康相關生活品質情形。以經濟學上預期效用理論爲基礎的常見評量方法包括：standard gambling(SG)，time trade-off(TTO)，willingness -to-pay(WTP)等；以心理計量來測量效用者主要是用 rating scale(RS)，所發展出之量表包括：Rosser's Index of Disability, Kaplan's Quality of Well-being Scale(QWB), Health Utility Index(HUI), Healthy-Year Equivalent(HYE), EQ-5D, Index of Health-Related QOL (IHQOL)等。用心理計量學所發展出常見的評量量表則如：Sickness Impact Profile(SIP), Nottingham Health Profile(NHP), McMaster Health Index Questionnaire(MHIQ), MOS SF-36, WHOQOL 等。讀者可參考相關書籍來瞭解各種評量的內容介紹，本文作者也曾發表了一篇針對 SIP, NHP, QWB, MOS SF-36, EQ-5D, WHOQOL 六個量表之介紹及評論性的文章[18]。以上的這些評量，若是以經濟學爲基礎來編製，則通常可被用來作爲醫療決策的根據，但是它太籠統，以致於難以在個別病人診治時作決策之依據。若是以心理計量學爲基礎來編製，則可描述應答者生活品質各方面的狀態，並可協助病人做治療方法選擇之用。這兩種方式都各有利弊，例如前一種方式在施測上較難讓人瞭解，在計分上也較複雜難懂，但是因爲只有一個單一數值，因此較容易被用來比較不同個

體間整體性生活品質之差異。後一種方式因爲量表有數個向度而有數個數值，因此比較不同個體間生活品質之差異時較爲困難。雖然二者之間的相關性有一定的程度，但卻可能不能完全直接互相取代，這是因爲二者的理論基礎是不同的。

測量方法原理之簡介

所謂測量(measurement)，根據 Stevens 的說法乃是「依據法則而分派數字於物體或事件上」[19]。由此說明了測量的基本性質包括了三方面，即測量是依照一定的步驟(法則)、對個體(人、事、物)使用數值(分派數字)來表示個體的特性。若個體是以人爲主，我們常稱此種測量爲測驗(tests)。測量可分成直接測量及間接測量兩種，其中直接測量多用在當測量的對象有具體的特質(physical properties)時，例如測量物體的物理、化學、生物特性等，這方面的測量通常都有可靠而精準的儀器來計算測量對象的特質，如：用標準尺量長度、用磅秤量體重、用血壓計來量血壓。然而從另一方面來說，許多人類的社會、心理及行爲等特質並不是那麼的具體，反而是比較抽象的，因此此種特質較難被定義清楚，也就較難被直接測量，而需要以間接測量的方法來測定。這些特質如：態度、個性、智力、價值觀、成就等，而主觀性的生活品質也是屬於其中的一種。這些抽象特質我們稱做「潛在變項」(latent variables)，這是因爲這些變項無法被直接觀察收集到，他們的存在是因爲根據某些學說或定義，而後被驗證推導出來的。不同的學說或定義，可能會產生不同的測量結果出來，這也反映出此類型測量的困難與多變性。研究者應該在測量之初，即選定測量所依據的學說或定義，而後根據此來設計測量的工具。至於如何設計工具以收集關於該特質之具體的資料，則需經過有經驗的心理測量學者，根據一套嚴格的過程而設計量表來收集資料而得[20,21]，如此才能減少因間接測量而有的誤差。一個好的測量工具，至少需要考慮到：適切度(包括：代表性、客觀性、標準化)、信度(reliability)及效度(validity)，才能真正準確的測到我們所期望得知的。接下來作者先介紹古典測驗理論，從這個理論可導出我們傳統所熟知的各種信度與效度之概念，之後作者再談信度與效度之種類，最後作者將談及現代測驗理論，即所謂的項目反應理論(item response theory)。

古典測驗理論

「古典測驗理論」主要描述了測量誤差是如何的影響觀察值，它包括了七個假設：

1. $X = T + E$，即 觀察值 = 真實值 + 誤差值。我們實際上無法直接觀察受試者的真實值或真正能力，而只能由測量的方式去找出觀察值或觀察到的能力。這種觀察值含有誤差，而此誤差被假設爲一個隨機(random)變數，其分配是以零爲集中趨勢指標的常態分配。這種誤差有時大於真實值也有時小於真實值，但總平均起來誤差爲零。

2. $\varepsilon(X) = \varepsilon(T + E) = T$，觀察值的期望值=真實值。用相同的測量方式重覆測同一個人很多次所得觀察值分配的平均值(即觀察值的期望值)是受試者的真實值，而誤差值的期望值等於零($\varepsilon(E)=0$)。

3. $\rho_{ET} = 0$，誤差與真實值不相關。就是說一個人真實值的高低不會與其測量誤差的高低有關係。

4.假如有兩個測驗，如同(1)：$X_1 = T_1 + E_1$ 與 $X_2 = T_2 + E_2$，則 $\rho_{E_1E_2} = 0$，兩測驗間之誤差不相關。

5. $\rho_{E_1E_2} = 0$，一個測驗的誤差與另一個測驗的真實值不相關。因此測某種特質的測驗並不受另一種測驗的誤差影響。

6.平行測驗(parallel tests)：若兩測驗(X及 X'爲其相對應之觀察值)皆符合假說(1)至(5)，且兩測驗有相同真實值(T=T')以及相同誤差變異量($\sigma_E^2 = \sigma_{E'}^2$)，則此二測驗稱作平行測驗。

7.主要真實值相等測驗(essentially τ-equivalent tests)：若兩測驗(X_1 及 X_2 爲其相對應之觀察值)皆符合假說(1)至(5)，且兩測驗的真實值差一個常數($T_1 = T_2 + C_{12}$)，則此二測驗稱作主要真實值相等測驗。

信度

根據以上這些古典測驗理論的假設，我們可

導出測驗的信度(reliability)公式，其所導出之公式的過程請參見 Allen 與 Yen[22]所出版介紹測量理論的書籍。「信度」的同義字是可靠性(trustworthiness)、一致性(consistency)、穩定性(stability)、可信賴度(dependability)或精確性(accuracy of precision)。所謂「信度」是指用同一測驗重覆測量某項持久性特質時，得相同結果的程度；或指測驗前後兩次分數一致的情形；或指測驗內部試題間是否相互符合的程度。通常「信度」可分爲下列四種：

1.施測者間信度(inter-rater reliability)：

兩個或兩個以上的施測者在同一時間對同一施測對象施測結果的一致性，測量方式是以相關法爲主。

2.再測信度(test-retest reliability)：

用同一種測驗對同一群受試者前後施測結果的一致性。此種信度易受練習、記憶或身心成熟的影響，因此前後施測時間間隔必須適當。時間的間隔沒有一致的規定，端視測驗的性質及施測對象的特質而定。例如：對尚在變化過程中的中風病人或成長中的孩童施測時，施測前後時間間隔宜短，以減少病人因隨時間而成熟變化，然而時間間隔又不至於短到讓病人有記憶練習施測內容的機會，而對長期慢性精神病人，則施測時間間隔可較長些。另外，施測者內信度(intra-rater reliability)是指同一施測者對相同受測者前後施測的評分是否一致的程度。

3.折半信度(split-half reliability)：

再測信度或施測者內信度都是使用相同測驗兩次或兩次以上。然而在一種測驗沒有複本(alternative form)且只能施測一次的情況下則可採用折半信度法，以了解測驗本身內容是否相互符合。通常的作法是將測驗題分前後半或單雙號半，而後求兩半間之相關性。我們在文獻中所聽到的斯布氏公式 (Spearman-Brown formula)、克氏阿爾法(Cronbach alpha，α)、范氏公式(Flanagan formula)及盧氏公氏(Rulon formula)等皆是用來計算折半信度的公式。將折半信度的概念延伸，以求各題目間的相關性，即是求整個測驗的內部一致性(internal consistency)，以計算克氏阿爾法值(Cronbach alpha，α)爲最常用到的方法。

4.複本信度(alternative form reliability)：

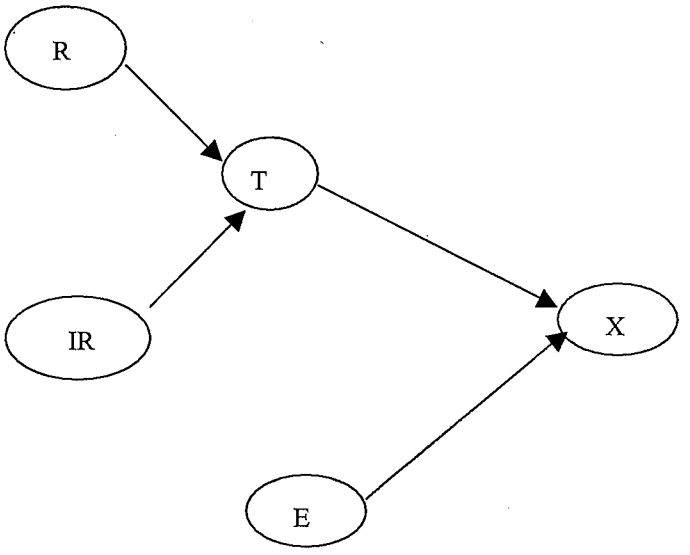
指兩個平行測驗間觀察值的相關($\rho_{XX'}$)。若一套測驗有兩種以上的複本，則複本間可交互使用以避免再測信度的缺點。

效度

在效度(Validity)方面，所謂「效度」是指正確性，即能正確測出所欲測量的特質之程度。每一個測量工具有其一定的適用範圍，例如我們若使用測量關節活動度的角度器(goniometer)來測量一個人的握力，或使用尺量手掌的大小以表示一個人的握力，則此種測量就「無效」。若一個測量工具不能測出所要測的特質，則有再好的信度、再優良的施測步驟也都沒有用，因此我們可說「效度」是科學測量工具最重要的必備條件，而信度只能算是效度之必要條件而非充分條件。我們以圖二來瞭解信度與效度之間的關係，由圖二可知道觀察分數(X)是真實分數(T)與測量誤差(E，或稱爲隨機誤差，random errors)之組合，而其中真實分數之來源又分成真正與所欲測量能力有關之分數(R)及與所欲測量能力無關之分數(IR)，後者又稱爲系統誤差(systematic errors)，指的是與所欲測之能力無關卻固定影響真實分數的因素。例如：一個想出國留學的學生考 GRE 測驗的計量分測驗，他所得到的分數並不能真的反映他的真實計量能力，因爲此分數是受到系統誤差(如：瞭解題目的英文能力)的影響。其他影響因素如：心情變化、天氣、不小心看錯等則都算是隨機誤差，因爲這些誤差不一定都固定存在的。從分數變異量(variance，S^2)的角度來看信度與效度的差別，則信度是指真實分數變異量占觀察分數變異量的百分比，而效度則是以信度值扣除掉系統誤差變異量之部分，亦即是真正所欲測量的部分占觀察分數變異量之百分比。通常「效度」可分爲下列三種類型：

1.內容效度(content validity)：

乃是指測驗內容適當的程度，包括了想研究的特質其測驗內容是否足以涵蓋各重要的特質元素，又測驗內容對各重要的特質元素分配比例是否適當。例如：若想測量中風病人的認知與知覺能力，一個好的施測工具需考慮到是否能適當



X=observed score
T=true score
E=random measurement error
R=relevant to purpose of test
IR=irrelevant to purpose of test (systematic measurement error)

信度：$S_T^2 / S_X^2 = 1 - (S_E^2 / S_X^2)$

效度：$S_R^2 / S_X^2 = (S_T^2 / S_X^2) - (S_{IR}^2 / S_X^2) = (1 - (S_E^2 / S_X^2)) - (S_{IR}^2 / S_X^2)$

圖二：效度分析的邏輯圖

的反映出一個人的認知與知覺能力的各層面。通常內容效度又分爲表面效度(face validity)及邏輯效度(logical validity)。所謂「表面效度」是指一個測驗在主觀認定上有沒有效的程度。而想達到「邏輯效度」則需對要被測的研究特質定義清楚的範圍，並且找出涵蓋所有重要的特質元素。在研究過程中也有「專家效度」之探討，此效度即是收集專家(各領域對此詞有不同的定義)對研究者所編製的量表題目內容之適切性、語句之流暢性等方面來評估，因此也算是內容效度之一種。

2.效標關聯效度(criterion-related validity)：

乃是指測驗的結果與效標(criterion)相關連的程度。而「效標」是指想用測驗來預測(predict)的某種特質或行爲。效標關聯效度又可分爲同時效度(concurrent validity)及預測效度(predictive validity)。所謂「同時效度」是指測驗結果與當前的效標相關連的程度，通常以相關法來計算。例如：已知測驗 A(效標)能有效的測出幼兒精細動作的發展，現在想發展一個新的施測工具 B，則施測者可同時將二測驗給予幼兒，而後求二測驗分數之相關以得同時效度。雖然讀者可能在某些文獻上發現一些作者將同時效度稱爲預測效度，但本質上這種效度仍是同時效度。所謂「預測效度」是指測驗結果與未來有關方面表現間之相關的程度。例如：想設計一種有「預測效度」的測量工具，以了解是否可用中風病人的出院前手部回復功能情形來預測他們出院後日常生活自理的能力，施測者需先測病人出院前手部回復的功能，並於病人出院後測其日常生活自理能力(效標)，以得知二者間的相關程度來判定此測驗的預測效度。

3.建構效度(construct validity)：

乃是指測驗能測量理論的概念、結構或特質之程度。「建構」(construct)是心理學理論所說的抽象而屬假設性的概念，例如：智力、焦慮、動機等，這些概念的建構效度並不容易且非單一之研究而能建立得完全，而是必須累積許多的研究結果才得以更臻健全。建構效度的建立通常由理論的架構而來，導出相關的假設，發展出適當的測驗，而後就施測結果的分析來看測驗題目是否符合理論，若否，則需修改測驗題目再施測，但有時也需考慮理論及假設的適當性，以決定是否需要對其做修正，經過如此這般來來回回反覆的過程後，而得到有建構效度的測驗。求建構效度所使用的方法並沒有絕對的依據，可用相關法、實驗法、因素分析(包括探索性及驗證性)、因徑分析等各種可能達到目的的方法。

在臨床測量時，研究者常常也希望測量工具能有好的反應性(responsiveness)，也就是說隨著時間的進行，測量工具所測得之生活品質分數能反映出病人真實生活品質之改變。這種反應性對縱貫性的研究(longitudinal studies)特別重要。以上是從發展近一百年的古典測驗理論的角度來談測量之原理，另外還有近五十年所發展出的項目反應理論(item response theory，IRT，或稱爲現代測驗理論)，由於此理論發展較新且涉及相當多的數學公式及電腦計算，因此下一段將只概略介紹此理論，其詳細的內容讀者可參考相關的書籍。

項目反應理論

由傳統古典測驗理論所發展出的測驗有一些缺點，例如：試題的統計量(難度及鑑別度、誤差等)可能因樣本之不同而有差異；理論上受試者不論程度如何，需完成整個測驗才能計算其得分，如此一來既費事也耗時；另外，使用古典測驗理論無法區別及預測每一個人的真實能力等， 因此而有現代測驗理論－「項目反應理論」(item response theory, IRT)的發展。IRT 根基於潛在特質模型(latent trait models)，它是 1950 年代被提出的。IRT 認爲受試者的表現受其潛在特質的決定，而其理論原型最基本的假設是「局部獨立」(local independence)及單一向度(unidimensionality)。前者是指受測者在別題上的表現並不影響此題的表現，並且別的受測者的表現並不影響此受試者的表現，後者則是指同一(分)測驗內的題目只有單一向度。此理論的目的是可估計受試者在連續的潛在特質上所站的位置(即該能力或特質的大小)，另外也可估計重要的題目參數，如試題難度(item difficulty)及鑑別力(discriminating power)等。題目參數的估計有助於對測驗試題的選擇，以致於所組成的測驗能比較正確的估計出受試者的潛在特質(即該特質或能力的大小)。相較於古典測驗理論，用現代測驗理論所發展出的測驗題目特質(或說題目參數值，如：難度、區辨度、猜測度等)具有不變性(invariance)，是不受所使用樣本之不同的影響(sample-free)及不受所使用的測驗內容之不同的影響(test-free)。究竟 IRT 對我們發展的測量工具有什麼功用呢？

1.選取測驗題目及對題目做加權：

從 IRT 模型的數學函數圖及參數估計值可知道該試題的好壞，也可知道對各題目的加權應是多少才合適。

2.測驗等質化(test equating)：

當測驗有複本(alternative form)時，我們通常會希望將正本與複本的測驗分數經過轉換，使二者分數有對等性的關係，若要如此做，則這兩個版本的測驗必須測量同一特質，且每一層次的特質需有相等的測量準確度。一旦測驗被成功的等質化後，不同版本的測驗可互通，受試者無論用那一種版本的測驗，所得結果應該相同，這種等質化被稱爲水平等質化(horizontal equating)。另外，經由此模型的操弄，即使受試者接受同性質但不同的測驗題內容，仍能比較彼此潛在特質能力的差異，這種等質化被稱爲垂直等質化(vertical equating)。目前這種方法已經應用於電腦化適性測驗(computer adaptive testing，CAT)，如：TOEFL、GRE 等，甚至在評估受測者的生活品質主觀感受上也有採用此方法的(如：SF-36 有電腦化適性測驗版)。

3.檢驗題目的偏差(bias)：

利用 IRT 的方法可用來檢驗出於不同組間(例如不同語言版本間)在各題目上的潛在偏差

情形，即所謂的題目差異功能(differential item function，DIF)，依此方法所得之結果能反映出哪些題目具有跨文化對等性，哪些題目沒有而需被修改或刪除，使得跨文化生活品質的題目得以選擇，使得跨文化的比較得以達成。

雖然 IRT 克服了許多古典測驗理論上的缺點，但在 IRT 的實際應用上需要非常小心，因爲 IRT 是一種大樣本模式，也就是說爲了要能準確的估計出參數值，必需要有足夠大的樣本才行，因此在實際的應用上，使用數百到數千個樣本人數是很常見的。IRT 的假設也局限了它的實用性，例如在使用 IRT 時需考慮模型的單向度(unidimensionality)假說，也就是說所有測量題需只能測量單一潛在特質。若題目的設計不良或題目是需計時完成(speed tests)，則這種測驗測量了兩個或兩個以上的潛在特質，因此違反了模型的單向度假說，並不適合使用 IRT 方法，不過近些年來已有學者從事多向度(multidimensionality) IRT 的研究。另外，模型的最基本假設「局部獨立」(local independence)使得試題的設計不能是連鎖(chained)形式，因爲受試者在別題上的表現是可能會影響此題的表現。再者，IRT 應用在電腦化適性測驗，除了需要有良好的測驗發展技術外，還必須建立一個心理計量特質含蓋範圍廣且精確，並且可隨資料的收集而不斷修正的題庫(item bank)，這方面是一個相當龐大的工程。因著以上的考慮，作者認爲 IRT 的用意及理想很好，其結果之推論也強，但需要發展技術足夠完善，且測量工具的設計非常優良、符合假說，否則恐怕在應用上仍不如古典測驗理論來得簡單容易。

推薦讀物

1. Zhan L: Quality of life: Conceptual and measurement issues. J Adv Nurs 1992;17: 795-800.
2. Meeberg GM: Quality of life. A concept analysis. J Adv Nurs 1993; 18: 32-8.
3. Oleson M: Subjectively perceived quality of life. Image 1990; 22: 187-90.
4. Wilson IB, Cleary PD: Linking clinical variables with health-related quality of life. JAMA 1995; 1995: 59-65.
5. Lohr KN: Outcome measurement: Concepts and questions. Inquiry 1988; 25: 37-50.
6. Stormberg MF: Instruments for clinical nursing research. Norwalk, CO: Appletion & Lange.
7. Abbey A, Andrews FM: Modeling the psychological determinants of life quality. Soc Indicator Res 1985; 16: 1-34.
8. Campbell A, Converse P, Rodgers W: The quality of American life. New York: Russel Sage Foundation, 1976.
9. Patrick DL, Erickson P: Assessing health-related quality of life for clinical decision-making. In: Walker SR, Rosser RM eds. Quality of Life Assessment: Key Issues in the 1990s. Netherlands: Kluwer Academic, 1993: 11-64.
10. Ferrans CE, Power MJ: The employment potential of hemodialysis patients. Nurs Res 1985; 34: 273-7.
11. Szabo S: The World Health Organization Quality of Life (WHOQOL) assessment instrument. In: Spilker B ed. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd ed. Philadelphia: Lippincott-Raven, 1996: 355- 62.
12. The WHOQOL Group. Development of the WHOQOL: Rationale and current status. Int J Ment Health 1994; 23: 24-56.
13. The WHOQOL Group: The World Health Organization Quality of Life assessment (WHOQOL): Position paper from the World Health Organization. Soc Sci Med 1995; 41: 1403-9.
14. The WHOQOL Group: The World Health Organization Quality of Life assessment (WHOQOL): Development and general psychometric properties. Soc Sci Med 1998; 46: 1569-85.
15. The WHOQOL Group: Development of the World Health Organization WHOQOL-BREF

quality of life assessment. Psychol Med 1998; 28: 551-8.
16. World Health Organization: WHOQOL study protocol. Geneva: WHO, 1993(MNH/PSF /93.9).
17. World Health Organization: Resources for new WHOQOL centers. Geneva: WHO, 1995 (MNH /PSF/95.3).
18. 姚開屏：簡介與評論常用的一般性健康相關生活品質量表兼談對未來研究的建議。測驗年刊 2000; 47: 111-38。
19. Stevens SS: On the theory of scales of measurement. Science 1946; 103: 667-80.
20. 姚開屏：從心理計量的觀點看測量工具的發展。職能治療學會雜誌 1996; 14: v-xxi。
21. 姚開屏、陳坤虎：如何編製一份問卷—以「健康相關生活品質」問卷爲例。職能治療學會雜誌 1998; 16: 1-24[特稿]。
22. Allen MJ, Yen WM: Introduction to measurement theory. CA: Brooks/Cole Publishing Co, 1986.