態度與行為研究的信度與效度：理論、應用、反省
（網路 2018 更新版）

吳統雄

第一版原刊：〈民意學術專刊〉，民74 夏：29-53; 1985f

一、信度與效度的基本概念

(一)定義

信度與效度均特指測量工具（如問卷、態度行為量表）減除可能影響測量結果因素後的準確程度。

信度與效度既然稱為「度」，就是一個可以度量的具體「數字」，不是抽象的感覺、感受，也不是廣泛的調查準確性。

為信度與效度下定義的學者很多，但歸納言之，且簡言之，可為：

1、信度(reliability)

測量工具本身的準確程度—是否有區別能力？測量的結果是否穩定一致？穩定一致的程度如何？

譬如一把捲尺昨天量一個人的身高是一百七十公分，今天再量卻變成一百六十五公分，一個人斷不可能一天矮了五公分，顯然這把尺可能受熱脹冷縮的影響很厲害，也就是「信度」不高。

又如這把尺從尺端開使量一個人的身高是一百七十公分，但從一公尺的地方開始量，同一人的身高卻變成一百七十五公分，就顯示這把尺刻度之間的距離不準確，「信度」自然就低了。

2、效度(validity)

測量工具是否可以測出研究者想要了解的某種特質？在人類行為研究上，尤指存在而看不見的抽象特質，亦即「構念(construct)」，能夠測出的程度為何？

舉一個較具體的例子來說譬如我們一把刻度很精確、不會熱脹冷縮也就是「信度」很高的尺，但如果用這把尺來量一群人，以判斷誰輕誰重，就可能不大準確，不很「有效」。因為尺並不擅於測量「體重」這個特質。尺對「體重」這個特質而言，就是一個「效度」不佳的測量工具。

但是判斷測量工具是否「可信」？「有效」？並不經常像上面的例子那麼顯而易見。譬如理論上尺是測量「長度」的最佳工具，但經

銷金屬線的大盤商，卻寧可用測量「重量」的秤，來計算金屬線的進貨、出貨的數量，否則用尺來計算一貨櫃規格不等的金屬線，很可能吃力不討好。

尤其在測量抽象特質時，往往不僅是「吃力不討好」而已，許多研究者擬了一份問卷，想要探討一些問題，收回資料後，經過信度與效度的分析，發現資料幾乎與研究問題無涉，這種事例以往屢見不鮮。

(二)計量方法與計量思想的意義

1、統計模式

信度與效度可以用邏輯方法辨識，也可以用計量方法分析。而在信度與效度計量研究之初，是針對「學習成就測驗」、也就是考試與考卷的形式而設計，日後才發展至其他心理、態度、行為測量工具的研究。

信度與效度的發軔期，與統計線性分析的學者分頭進行，且參與者眾多，使用的符號、名詞並不完全相同。

為避免閱讀與理解的困難，本研究者以下的介述，已經過整合，並加入本研究者的評論、研究發現與詮釋。

其統計模式如下(Peter [274])：

(1)信度統計模式

假設一個測量工具所測得的值為Ｘo（通常以平均數代表）則Ｘo可分解為：

Ｘo ＝Ｘt ＋Ｘe
Ｘo （observed Ｘ）：觀察值
Ｘt （true Ｘ）：真實值
Ｘe （error Ｘ）：誤差值

就「統計思想」言，經常「以面積（即變異量）解決線性問題」，以上的測量值為線性，故以變異量分析如下。

假設測量所得的變異量為Ｖo ，則Ｖo 可分解為：

Ｖo ＝Ｖt ＋Ｖe

其中真實變異量與觀察變異量之比，即為信度：

r_{tt}（信度）＝Ｖt／Ｖo

但Parameswaran et al. [266], Tryon [350]指出從統計的角度來看，Ｖt 很難直接估計，因此上式常移項改為：

$$r_{tt} = (Vo - Ve) / Vo = 1 - (Ve / Vo)$$

信度即為１減去「誤差變異量與觀察變異量之比」。

(2)效度統計模式

如果把Ｖt 再分解：

Ｖo ＝Ｖco ＋Ｖsp ＋Ｖe
Ｖco（correlated Ｖ）：與測量特質相關的共同變異量
Ｖsp（specific Ｖ）：與測量特質無關的其他變異量

則效度（val.）為：

$$Val. = Vco / Vo$$

譬如說，「體重」這項特質大致和「身高」這項特質成正比，尺大致也可以區別體重，但是體重又和「體型」很有關，尺卻無法測量「體型」，即「體型」並不和「身高」成正比，因此用尺來測量體重時，其中有存在測不出來的「體型(Vsp)」與誤差(Ve)，效度自然不夠高。

展開以上公式，可知效度與信度的關係為：

$$Val. = (Vt - Vsp) / Vo = r_{tt}（信度）- Vsp / Vo$$

故效度應不大於信度。

2、誤差：影響信度與效度的因素

從以上統計模式可以看出，信度與效度的大小與誤差的大小成反比。

測量中為何會產生誤差，學者論之甚詳，主要可以歸納為五方面：

a.受訪者的答復

受訪者可能因為個性、情緒、注意力、年齡、性別、反應力、知識背景、社會地位以及其他心理、生理因素，影響答復的正確性。

b.測量內容

遣詞用字、問題形式、以及內容是否敏感等。

c.情境

訪問時間長短、訪問當時的氣氛、及開頭的引導說明等。

d.研究者本身

訪員是否盡責，事前研究設計是否妥善，事後研究分析解釋是否合理。

e.疏忽

如聽錯、記錯、轉錄錯誤等等。

因此，產生誤差的原因是多方面，研究者必需面面俱到，才能提高信度與效度。

3、信度與效度的關係

a.效度是信度的充分條件

已經證明效度很高的量表，信度一定很高。效度的大小反映研究者的理論架構是否正確，因此追求合乎理想的效度是研究者最終的目標。

b.信度是效度的必要條件

信度很高的量表，效度不一定夠高；信度很低的量表，效度一定不符合要求。因此，即使限於研究資源，一項調查訪問做不到分析效度的階段，至少要達到分析信度，若是信度太低，就要及時修正，以免往後分析資料的步驟變成白費力氣。追求合乎理想的信度是研究者最起碼的目標。

學者已推算出：效度應不大於信度的平方根（林邦傑 [41], McCullough et al. [239]）。

4、信度與效度分析的功能

分析信度與效度可以了解測良工具是否優良，從而改善測量的內容或方法，更重要的是可以避免作錯誤的判斷及因錯誤導致的損失。

在純學術研究方面，學者曾分析了總加量表（Munson et al. [253]）、語義量表（Menezes et al. [244]）及其他特殊研究方法（ Zdep et al.[378]）的效度，證明他們是有效的測量工具。

在應用方面，美國學者曾發現美國空軍用來評估軍隊作戰能力的測驗，效度非常低，引起了整個制度的大改革；Best [98]用來分析工商行銷界用來開發市場的量表有沒有效；Jacoby [203]則分析「意見領袖」的現象是否存在。

在臺灣，2016年吳統雄所指導李玉淑的論文，證明國內醫學界，長期由醫學會所制訂的「醫院病患滿意度」問卷公版，並不具適當效度。

二、信度的類型

Peter[274]指出信度傳統上主要分為三類：「再測信度」（test-retest reliability）、「內在一致信度」（internal consistency reliability）及「複本信度」（alternative form reliability）；後來，Cronbach et al. [141]則推廣「綜合信度」（reliability as generalizability theory）。

1、再測信度（又稱「外在信度」 external reliability）

用同一組量表對同一群受訪者，隔一段時間先後訪問兩次，前後答復的相關係數就是再測信度。間隔的期間通常是兩周左右。不過，再測信度有三個難題：

a. 問隔的時間愈長，信度愈低。

b. 如果在再次訪問之前，有重大事故改變了受訪者的態度，研究者無法區別到底是「發生事故」或是「量表信度低」造成了改變。

c. Nunnally [258]認為再測信度常有高估的趨勢。

再測信度使用時期很長，技術也有很多革新（Burns et al. [114],Parameswarn et al. [266], Silk, [313]），但由於測量不方便，無法一次完成，受訪者也容易厭煩，因此這種分析方式並不十分理想。

2、內在一致信度（又稱「內在信度」 internal reliability）

「內在一致信度」由「折半信度」（split-half reliability）衍生而出，後者是將量表的項目分成兩半，各別計分，再算出這兩半的相關係數，即為折半係數。通常是按照項目編號的單、雙數來折成兩半。

但是按單、雙數折半，缺乏嚴謹的理論，因為不同的折半方式，會產生不同的信度係數。解決的方法就是算出所有折半信度的平均數，訂為「內在一致信度」。訪問的資料如果是「二元資料」，就用「庫李20號公式」（KR-20,Kuder et al. [224]）計算；如果是連續性資料，就用「α係數」（Cronbach [142])計算。這兩種公式，均已有電腦軟體程式可代勞計算。

在長期行為科學調查中，它是分析信度最常見的途徑。

唯在2008年後，本方法已遭遇質疑，將再後續專節說明。

3、複本信度（又稱「穩定等值信度」 stability and equivalence）

利用兩份內容類似的量表，訪問同一群受訪者，也就是在各種考試中，常見的「AB卷」。

簡茂發[80]認為這種分析法，可以改正再測信度很多的缺點，對教育成就測驗（即學科考試），或是一般心理實驗可能是最好的方法。

但它主要的難題是除了教育成就測驗等少數領域外，大多數行為科學想要研究的問題，很難找到符合理想的所謂「等值複本」。同時，複本信度的值通常和內在一致信度十分接近（Nunnally [258]）。

因此，除了教育、心理學科之外並不多用。

4、綜合信度

歸納傳統分析信度的力法，不外考慮兩個因素：

a. 不同的期間，是否會影響測量結果？（如再測信度）
b. 不同的項目，會不會影響測量的結果？（如內在一致信度）

這樣的考慮卻可能有兩種缺陷：

a. 相同的訪問測量，如果考慮的因素不同，分析的方法不同，會得到不同的信度係數，信度的意義及其進一步的解釋也會迥然不同。
b. 影響測量結果的來源，除了各因素獨立的效果之外，應該還有因素間「交互作用」的效果，後者在傳統的分析信度並沒有被考慮到。

為了解決以上兩個問題，以及把信度的意義一以貫之，Cronbach et al. [141]提出了「綜合理論」（generalizability theory）的概念：在不同因素的影響下，「觀察值」中「真實值」的代表性有多大？「真實值」和「觀察值」之比即為「綜合信度」（或稱綜合係數）。

譬如：再測信度可視為以期間為單因素的綜合信度；內在一致信度可視為以測量項目為單因素的綜合信度；而以期間、項目為雙因素的綜合信度，可將測量的總變異量根據統計模式分解為：

$$V = Vt + Vi + Vj + Vti + Vtj + Vij + Ve$$

V ：總變異量
Vt：真實變異量
Vi：項目變異量
Vj：期間變異量
Vti：真實與項目之間的「交互作用」變異量
Vtj：真實與期間之間的「交互作用」變異量

Vij：項目與期間之間的「交互作用」變異量
Ve：誤差

而綜合信度定義為：

G（綜合信度）＝ Vt ／（Vt＋Vti＋Vtj＋Ve）

同理，可以定義三因素以上的多因素綜合信度。分析信度係數的方式，則借助於「變異數分析」（ANOVA）。

綜合信度的優點，是統一了各種信度的概念，同時考慮了測量受到多種因素影響的可能，在理論上比較周延。但是，採用綜合信度的研究還不很多，它的潛力猶待更多的實證來啟發。

5、其他信度

a.觀察者信度（observer reliability）

就是在田野調查、觀察研究、或是問卷設計很粗放且以開放式問題為主的抽樣訪問時，訪員或是觀察員之間，對相同的事實記載是否一致的程度（Orenstein, [260]）

不過，在設計很精緻的抽樣訪問中，很少考慮這個問題。

b.評分者信度（Scorer reliability/ Inter-rater Agreement）

用在問答題類型的測驗，此較不同的評分者之間，對答復的正負、強弱看法是否一致。根據資料的型態、評分者人數，有多種檢定方法。

有些統計不用「信度」這個詞，內涵還是信度檢定，如使用Inter-rater Agreement 一詞的Cohen's kappa coefficient（κ），就是適用2位評分者時的信度。

c.內容分析信度

這類信度很多（如:composite reliability, coding reliability, Scott's Pi reliability）計算公式不同，但功能相若，均為比較不同的研究者，對開放式的答復，歸類彼此是否一致的程度（Holsti [198]）。

在抽樣調查中，除了以開放式問題為主的研究外，很少會考慮以上三種信度。

三、重要信度分析法：內在一致信度

1、推求方法

內在一致信度是調查訪問最有用、也是最常用的信度。推求的公式很多，包括：係數、KR-20 公式、KR-21 公式、斯布公式（Spearmen-Brown formula）、范氏公式（Flanagan formula），盧氏公式（Rulon formula） 變異數分析法之 rH（ 分別參見Hoyt [199], Stanley et al. [329])、α 係數，其中比較有實用價值的有兩種：

(1) α 係數：適於連續性資料

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_t^2}\right)$$

k：量表的項目數。

σ_i^2：量表第 i 題，全體受測對象的變異數。

其累積和，對變異數分析言為k組的「組內離均差」和，對迴歸言即誤差項變異數。

σ_t^2：量表總分，全體受測對象的變異數，即總變異數。

從以上定義公式之最右項可知，即為「相關係數」之變異數分析定義。

如果各組受測數據相近，即測量結果相近、則變異數小，各項目的累積和也會小，亦即分子小。

以「1」減去變異值，即穩定值、亦即「內在一致」的程度。

左項的分母減1，為題數少時（一般為少於30），校正偏誤(bias)的觀念與應用。

注意：α 係數並非相關分析中的「相關係數 r」，而是類同「判定係數 r2 」。

許多文獻中、以及統計報表中出現的「項目-總分相關係數」：

r_{it}

其中 i 表示項目，t 表示總分，r 表示各項目各自與總分的相關係數，要特別避免產生誤會。

(2)KR-20 公式：適於二元性資料

$$KR-20=\left(\frac{K}{K-1}\right)\left(1-\frac{\sum_{i=1}^{K}pq}{S_t^2}\right)$$

K ：量表中的項目數
p ：答覆為正面的項目數
q ：答覆為反面的項目數
S_t^2 ：總變異數

與 α 係數公式在實質上完全相同，因為對二元資料的變異數值，就是p*q。

同時，二元資料可視為兩個刻度的連續性資料，一併可用 α 係數推算。

2、分析的標準與詮釋

信度的高低通常都用一個係數表示，係數的數值要多大才可信呢？就純粹統計理論觀點，當然限制愈高愈好，但卻容易使研究通不過檢定；有的學者為了遷就現實，把標準又定得似乎太寬鬆了。

α 在性質上等同迴歸分析的「判定係數」，故其意義為：

.9< α 高信度
.7< α <.9 中上信度
.5< α <.7 中下信度
α<.5 沒有信度意義

對研究問題相當了解，已有相當多文獻可以參考的研究，至少要超過「.9」以上的水準；探索性，有關案例很少的研究，「.7」亦可通過；對研究問題的真象一無所知，欠缺可以參考的文獻，初步的探索研究至少也應該達到「.5」的水準。

3、α 信度的質疑、辯難與未來

唯 α 的適用性，在2008年後陸續遭遇質疑，主要批判者，如 Klaas Sijtsma 指出的原因包括：

●減少樣本、增加題數，可虛增α 值。
●α 只是信度的最低限度值，不如新發展之 GLB (Greatest Lower Bound) 方法，如 Trizano-Hermosilla et al.(2016)。

不過，這項爭議學界尚無共識，一般研究仍普遍採用α。

其實任何一次性測量，都無法衡量最正確的信度，包括 GLB 法。

統雄老師依據「統計實務」經驗，建議解決方案如下：

任何「『態度』量表」，組成項目應大於或等於 5。

唯當前「統計理論與統計應用脫節」的生態，造成研究統計者不作實證，實證研究者不理解統計原理，許多國際頂級期刊、甚至SPSS自己的範例，常見「『態度』量表」項目不到 5，是沒有信度者意義的。

四、效度的類型

效度的名目繁複，名異實同的情況尤多，概括而言可分為三大類：內容效度（content validity）、效標關聯效度（criterion-related validity）及構念效度（construct validity）(注[1])。

1、內容效度

林邦傑[41]指出，內容效度主要用在學科考試前，分析出題的分配是否得當，其他的行為科學很少用。

這種效度嚴格而言不是個「客觀量化的『度』」，因為是由出題者主觀評定的「程度」。

2、效標關聯效度/預測力檢定

效標關聯效度Criterion-related Validity 、簡稱為效標效度 Criterion Validity，即「預測效度 Predictive Validity (Burns**錯誤！找不到參照來源。**)」或「預測力檢定 Prediction Model Validation」，是科學追求「可複製 Replica」的最基本方法之一。其與「外在效度 External Validity」、「交互驗證 Cross Validation」、「概化 Generalization」在理論定位與檢定方法上，幾乎相同、或十分近似。

效標關聯效度最早用在的測量工具，是學科考試或各種心理測驗，目的是分析某一項考試工具，是否能夠預測樣本未來真實的學習

(注[1])國內曾將 construct validity 直譯為「建構效度」。但「建構」在中文語意中頗不能達意。

本研究者曾經鑑於 construct在國外文獻中常與trait（特質）互用，trait即指研究者所關注的抽象性質，故曾將其譯為「特質效度」

另外，楊國樞也覺得譯為「建構」不妥，主張譯作「構念」。本研究者經深思後：決定從楊說。

發展成就(Lissitz**錯誤！找不到參照來源。**)；或是某項測驗工具，可以預測到多少研究者想探知的特質（Hills [193]，Tyebjee [352]）。亦即分析單一樣本集、2個以上測量工具之間的關聯數值。

而現在已可使用在各種心理、態度、行為測量工具上，也可以比較2個以上樣本集，所以稱為「預測力檢定」將更具廣義性。

有些文獻將效標關聯效度又分為三型：預測效度（predictive validity，或預測力prediction），同時效度（concurrent validity）及事後效度（postdiction），分別指預測工具實施、與被測應變項產生，兩者前後的不同時機，詳如以下定義(1)的範例。

吳統雄將以上相關理論與方法，整合為以下定義：

(1)2 個以上測量工具/單一樣本集

統計模式的效標 criteria，即應變項Y；甲測量工具即預測變項predictor。

甲測量工具對特定樣本集預測，獲得各樣本之效標值為 Y_p。

乙測量工具對相同特定樣本集，預測相同、或同類效標，獲得各樣本之效標值為 Y_t。

則，效標關聯效度/預測力檢定值，為 Y_p 與 Y_t 之相關係數。

亦即其稱為「關聯」效度的原因。

範例：

應變項為某高中學生升學模擬考排名，甲測量工具為模擬考之考卷。

而乙測量工具為大學聯考的真正考卷。

該高中學生在大學聯考的真正排名，與其模擬考排名之相關係數，即為其效標關聯效度/預測力檢定值。

此檢定值為甲測量工具（模擬考）之預測效度（predictive validity，或預測力prediction），亦為乙測量工具（大學聯考真正考卷）的事後效度（postdiction）。

如，該高中學生，同時參加了大學聯考、與甄試，則大學聯考排名、與甄試排名的相關係數，即為兩者的同時效度（concurrent validity）。

(2)單一測量工具/2 個以上樣本集

統計模式的效標 criteria，即應變項Y；測量工具即預測變項predictor、即自變項X。

設：在甲樣本集（有些文獻稱為訓練樣本集）中，以迴歸分析方法，獲得以下預測模式：

$$\hat{Y} = b_0 + bX$$

對乙樣本集（有些文獻稱為實驗樣本集），以此預測模式獲得各樣本之預測效標值 Y_p。

而乙樣本集中，Y之真實觀察值效標值為 Y_t。

則，效標關聯效度/預測力檢定值，為 Y_p 與 Y_t 之相關係數。

(3) 多測量工具/2 個以上樣本集

則在甲樣本集中，以多元迴歸分析方法，獲得以下預測模式：

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

後續步驟相同。

(4) 重複驗證

可以多個預測值，分析其相關係數矩陣，再以「統合分析法」進一步驗證。

實作範例，請參考吳統雄統計神掌：「效標關聯效度/預測力檢定」。

(5) 檢定的標準與詮釋

注意：雖然效標關聯效度即相關係數 r，但在邏輯上，效度的有效程度，並不等於相關係數的重要性。

因為後者只是兩個變項的相關程度，而前者是測量工具能不能夠測量到所要研究的構念、存在而看不見的特質，要求的程度應更高。

一般的r 之重要性實取決於判定係數 r2 所反映可估計範圍之百分比，如果2變項間相互影響近50%，「一半」在邏輯上就是近於高影響了，所以其重要性可判定如下：

.9 < r　　相關性具近決定性高重要性

.7< r <.9　相關性具近高至很高重要性

.3 < r <.7　相關性具低至近高重要性，視個案而定。

r < .3　　相關性不重要

但對效度而言，測量工具應至少測量到50%以上的欲可測構念、存在而看不見的特質；「一半」在邏輯上，應是最低要求，所以其有效程度應判定如下：

.9 < r　　效度很高
.7< r <.9　效度中至高
.3 < r <.7　效度低至中
r < .3　　效度低

3、構念效度

「構念效度（Construct Validity）」就是分析：研究者是否測量到了他想研究的「構念」。

「構念」（construct）又稱為「潛在變項」，是一種存在而看不見的特質，包括：

第一，一般人共有的抽象概念，如健康。人們雖然不能明顯而具體地「看到」健康，但都能肯定「健康」是一種實際存在的概念，同時能夠用身高、體重、血壓、運動量、飲食量……等指標，間接地測量出「健康」的好壞。

第二，學者經由思考、觀察、歸納，所領悟出來、構想出來、並且創造出來的一種概念（楊國樞[66]）。譬如若是有一位學者發現一個國家的政治體系和一個個人的生理體系有很多相通之處，於是他構想出一個「政治健康」的概念，用生理現象說明政治現象，用影響健康的因素解釋影響政治的因素。如果他能夠說得通，能夠證明「政治健康」可以以簡馭繁地反映許多複雜的政治行為，同時能夠穩定地測量出「政治健康」的好壞。那麼「政治健康」便可以在學理上成為一種「構念」。

以上的2個層次，其間界限並不是截然分明的，它也可以是部分具體而又部分抽象的，可以是已經為人所熟知的，也可以是由於社會變遷，人們為了解釋新問題所提出來的新名詞。

譬如「動機」、「冒險性」、「滿意度」...是長期存在的構念；而「無力感」、「新人類」、「女強人」、「資訊社會」、「企業再造」…均為近２０年外界環境巨變，所新衍生出來，也新被偵測出來的構念。

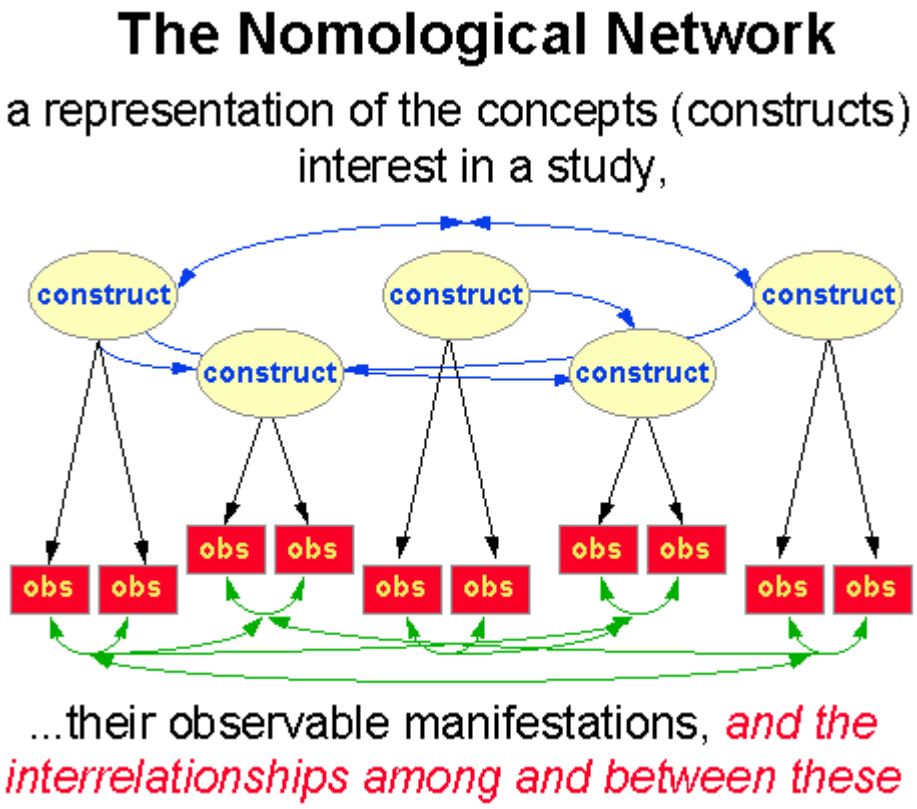
「構念　」是由枝節的行為所構成的；但我們在進行調查時，通常想要知道的是具備大方向感的「構念」，而不是枝微末節。

調查問卷經過淨化測量後產生的因素，正相當於「構念」；而淨化後的剩餘高相關項目，便可提供建立「構念」。

譬如，如果研究者想了解某班學生的「數學」程度，「數學」就是一種「構念」。他出了一張５０個題目的考卷請學生做答，內容包括加、減、乘、除、三角、幾何…微積分、與成語測驗。評分後，他使用淨化測量技術，發現成語能力與數學這項「構念」相關不大，於是便將「成語」從考卷中排除。但是，他也不需要知道誰比誰的加法好、減法好…，他把剩下所有的項目總加起來，再比較總分，就可以知道，到底誰的「數學」好。

(1)構念效度的興起

Cronbach et al. 在1955首先提出「Construct Validity」1詞，認為「構念」會以「法則關聯(Nomological Network)」（如圖）的形式存在，意義為：甲量表可測甲構念；乙量表可測乙構念，若甲構念和乙構念在理論上有相當的關聯，則甲構念之觀測變項和乙構念之觀測變項，在實證上應有相當程度的一致，形成一種關聯路徑。當時僅是一個觀念，並沒明確的分析方法，但啟發了後來的MTMM與CFA檢定方法。有些文獻引述，稱為「法則效度」可能易引起混淆，（Allison [82], Peter[273]）兩者其實合而為一。



(2)構念效度的條件

構念效度的討論在1980年後才大放異采，對人類行為研究啟發很深。

Kaplan [211]認為：堪以研究的「構念」至少應該具備兩個條件：

a.理論定位（systemic 若直譯則為系統性）

即「某種構念」在研究理論中明確的定義為何？在科學理論中擔負何種功能？譬如一組以某種態度為研究構念的量表，其中的「態度」到底何指？因為依據「態度理論」態度又包含三個層次：認知、態度和行為（Fishbein et al. [159]）研究者想探討的是廣義的態度（tripartite）還是狹義的態度（single component）？必需確切定義。

b.可測量性（observational）

構念必需可以直接測量或是經過「可測量化」（operationalized，中文常直譯為「操作化」）去間接測量（Torgerson，[345]）它的大小或強弱。研究的對象如果沒有可測量性，只是一個空洞的名詞，不能稱之為構念；而雖可以測量卻缺乏理論定位的性質，也只能算是一種物理現象，尚不足構成「構念」。

(3)構念效度的定義

「構念效度」學術上有兩種定義方法，第一種：具有構念效度量表可以：a.測出所有研究者想要探討的構念。b.只有這樣的構念被測量出來。Blalock [102]稱這種量表和構念之間為具有「知識論上的相關」（epistemic correlation）。

第二種定義是：一組量表的構念效度為：a.在一組具有代表性的合格樣本中，所有研究構念可以測出來的程度，b.在測量結果中，不包括其他構念和誤差的程度。

以上的定義都有一點「拗」，吳統雄建議一個簡單的「白話」定義：

甲工具測量甲構念準，乙工具測量乙構念準；

同時，

甲工具測量乙構念不準，乙工具測量甲構念不準；

則：甲工具對測量甲構念有效，乙工具對測量乙構念有效。

構念效度包含三個成分：a.信度：量表本身可信的程度；b.輻合效度（convergent validity）：相似的測量工具可以測出相同構念的程度；c.區別效度（discriminant validity）：不同的測量工具不會測到相同構念的程度。這三種成分必需靠合乎邏輯的推理，使它們合為一體的概念。

分析構念效度至少要經過兩次，最好是一連串的測量訪問後才能進行。

構念效度的功能除了分析量表是否有效之外，更可以驗證想要經過測量以建立的理論是否正確。如果一項調查研究的效度很低，可以提供研究者反省：a.是否研究理論不正確？b.是否研究構念的定位不正確？c.是否量表設計不正確？（Peter [273], Schwab [305]）

4、其他的效度

有些文獻上出現的「效度」和前述所談的效度不盡相同。

a.表面效度

「看起來」像不像有效的樣子，並沒有嚴密的數理推算力式。

b.內在效度與外在效度

內在效度原是實驗研究用來在概念上討論效度的8個問題（History, Maturation, Testing, Instrumentation, Selection, Experimental mortality, Selection-maturation interaction），它們的定義本不盡適用於抽樣調查研究（Babbie [87]）。但有些學者借用這個名詞以指稱「構念效度」，這樣的作法並不值得推廣。

「外在效度」即Generalization指研究工具所測量的構念，可以「一般化」的程度（Lin [235]）

c.診斷工具「效度」 (Assessment of Diagnostic Tests)

也有些統計方法沒有用「效度」這個詞，但其目的與方法，卻與效度檢定意義相同或接近，譬如生理統計對診斷工具的檢定，有時也被引用到行為研究中。

假設新發展一種希望診斷測量A疾病的試劑，我們希望知道這種試劑有沒有效，就把它使用在已知2群的受試者上，1群是已知有病者，2群是已知沒有這種病者。再比較新試劑所測的陽性/陰性結果，形成一個2*2的列聯表，再進行診斷工具檢定，包括以下幾個步驟：

-計算 Sensitivity and Specificity

-計算 Likelihood Ratios

這個方法也可以作為「篩檢」信度。就是在一群未知對象中，要篩檢出得A疾病的病患。

因為可能沒有一種百分之百有效的診斷工具，所以我們就使用2種診斷工具來測量同一群對象，而作成2*2的列聯表，再進行以上程序。

-這時會增加計算 Kappa 信度

這種方法原則上只適用二元資料，即測量工具僅測量「有無」。

如果用作行為上的態度量表檢定，因為不像「生病」有「確診病患」的對照組，只宜視為信度檢定，而非效度檢定。

五、重要效度分析法：構念效度

1、檢定方法類型

「構念效度」尚屬在發展中的概念，因此推求的力法相當地分歧，歸納言之，吳統雄[32]將其分為2大途徑：

a.兩個以上測量工具、兩個以上受測樣本的比較途徑

標準方法就是：多元特質--多重方法矩陣法（multitrait-multimethod matrix，簡稱MTMM）

這是由Campbell & Fiske [117]所倡導，奠定了構念效度檢定的里程碑。其他學者發展出來的分析方法，也經常採用MTMM一同考驗、相互印證，因此，MTMM可稱為檢驗構念效度的主流。另外還有「多元特質--多重側影矩陣」（multitrait-multiprofile matrix，簡稱MTMP）、「多重方法--多項活動矩陣」（multimethod-multiactivity matrix，簡稱MMMA）等，理論與MTMM完全一致，只是矩陣中「行」或「列」的性質略有變更。

有些時候測量工具不一定是量表，譬如大慧調查中的效度檢定，使用的就是2組「不同專家」作為測量工具。

b. 兩個以上測量工具、一個受測樣本的檢定途徑

這類方法與「因素分析」密切相關，又可稱「因素效度」，而主要又指驗證式因素分析(confirmatory factor analysis, CFA)，其應用常與因徑分析（path analysis/ causal model）、或稱結構方程模式建構(Structural Equation Modeling, SEM)有關，分析步驟包括相關分析（correlation）、多元迴歸（multiple regression）、共變數分析（analysis of covariance, ANCOVA）等。

2、MTMM分析

MTMM是最完整分析構念效度的方法（Lehnen[233]），故將其實施方式介紹如下：

(1)分析檢定程序

a.決定研究問題

假設研究問題為「臺灣地區選民的政黨、省籍和教育程度對政治態度的影響」。研究者在分析這個問題之前，必須確知他的測量工具

能有效的測出受訪者的「政治態度」，否則，任何分析推論都有可能是錯誤的。

b.決定研究構念

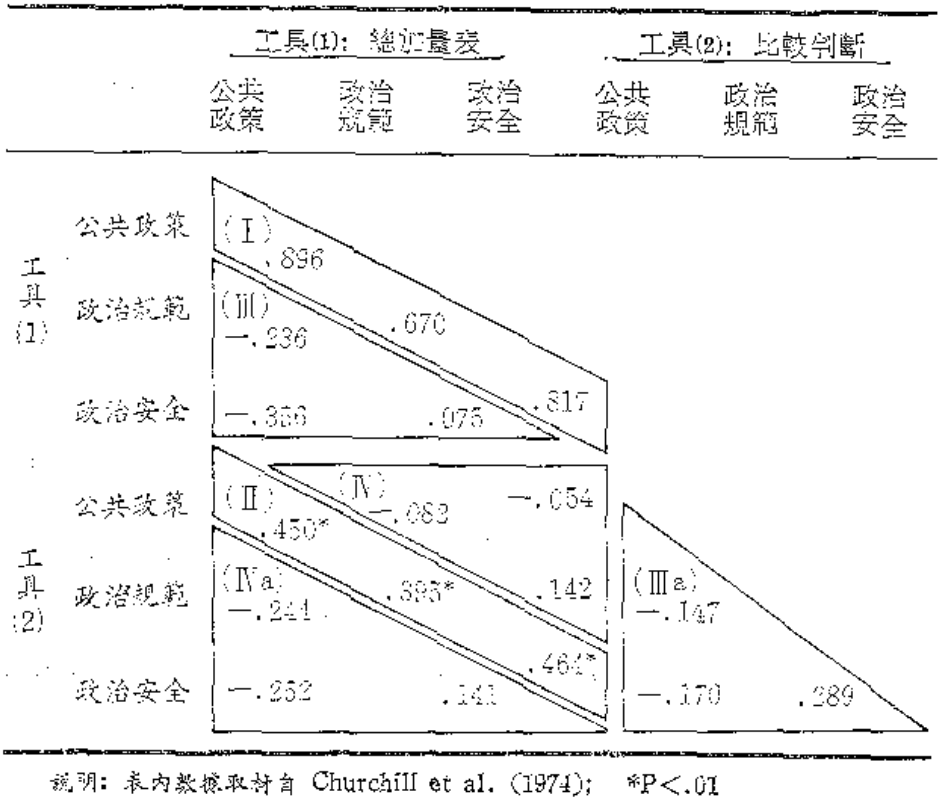
根據研究文獻，臺灣地區選民的政治態度，可以劃分為三個主要的層次：對「公共政策」的態度、對「政治規範」的態度及對「政治安全」的態度（吳統雄[38]）。因此，決定在訪問問卷中包含三組量表，分別測量這三種態度，每一種態度即為一項研究構念。分析構念效度所需要的「構念」必需在兩個以上。

c.決定測量工具

假設分別使用兩種工具：總加量表（summated scale）和比較判斷法（comparative judgment）測量相同的「構念」。

分析構念效度所需要的工具，必需在兩種以上。

d.列出MTMM矩陣



說明：表內數據取材自 Churchill et al. (1974)； *P<.01

（I）信度對角線（reliability diagonal）
（II）效度對角線（validity diagonal）
（III）（IIIa）異質同方域（heterotrait-monomethod triangles）
（IV）（IVa）異質異方域（heterotrait-heteromethod block）

(2)檢定的標準與詮釋

分析MTMM的時候，不見得都會像上例運氣那麼好，有時輻合效度不能確定是否充分大於零，有時效度係數不一定比所有的異質相關係數大，有時更不能用視覺判定某些區域內的組成形式是否一致。尤其在使用三種以上的工具，分析的構念又很多時，矩陣內的小區域更多，問題也就愈大。

因此，MTMM也曾備受Menezes et al. [244]批評。學者面對這個問題，經常是根據研究經驗及相關的文獻資料，來判斷檢驗是否「通過了」，迄今並未發展出一套固定的檢定標準。

本研究者建議：在效度檢定資訊系統中建立「效度係數資料庫」，每一項經過本資訊系統處理過的研究，自動將信度係數載入資料庫中，累積相當資料後，自然可提供比較信度高低的標準。

本研究者並根據以往的研究報告，建議以下參考標準作為系統預設值：

a.輻合效度(convergent validity)

即效度對角線（validity diagonal）之區域，相當於相關係數，故：

.9 < r　　效度很高
.7< r <.9　效度中至高
.3 < r <.7　效度低至中
r < .3　　效度低

b. 區別效度（discriminant validity）

（a）效度三角形與異質異方域

設效度係數大於異質相關係數的百分比為P，則：

P≦50%：無效
50%＜P≦70%：應據相關研究斟酌
70%＜P≦90%：有效
90%＜P：很有效

（b）效度三角形與異質同方域

設效度係數大於異質相關係數的百分比為Q，則：

Q≦50%：無效
50%＜Q≦70%：有效
70%＜Q≦90%：很有效
90%＜Q：十分有效

（c）同類區域的組成形式

學者通常用和諧係數（coefficient of concordance: Siegel [312]）檢定其間組成形式是否充分且顯著相關，可由本資訊系統呼叫軟體程式以供檢定。或者採用變異數分析、區別分析作輔助判斷。

3、驗證式因素效度分析

在因素分析中，各變項中可粹取出抽象的構念因素，Garson [171]認為因素變異量占總變異量的程度就是因素效度。但這種關係和前文所談的效度概念並不完全一致，Nunnally [258]便建議宜正名為「因素組合」（factorial composition）比較妥當。

但後來學者將因素分析又分為2類：上項稱為探索式因素分析（exploratory factor analysis, EFA）；另外又有驗證式(confirmatory factor analysis, CFA)，則被學者認為是檢驗構念的方法。

EFA分析是事前不知因素為何，經由題庫中的項目萃取而得。

CFA係依據理論建構事前已假設因素之存在與其所包含的項目，而後驗證其符合的程度。

分析因素與各項目間的相關，就是驗證輻合效度(convergent validity)，而分析各因素之間的的相關，就是驗證區別效度（discriminant validity），所以CFA也是構念效度的分析工具之一。

另外，CFA也經常配合因徑分析(Path Analysis)/結構方程模型(Structural Equation Modeling, SEM)，並使用其統計軟體進行分析。

(1) 分析檢定程序

a.決定研究問題

採用Jeremy J. Albright and Hun Myoung Park所介紹的美國政治態度研究為範例，研究問題為「左派與右派政治態度的差異」。

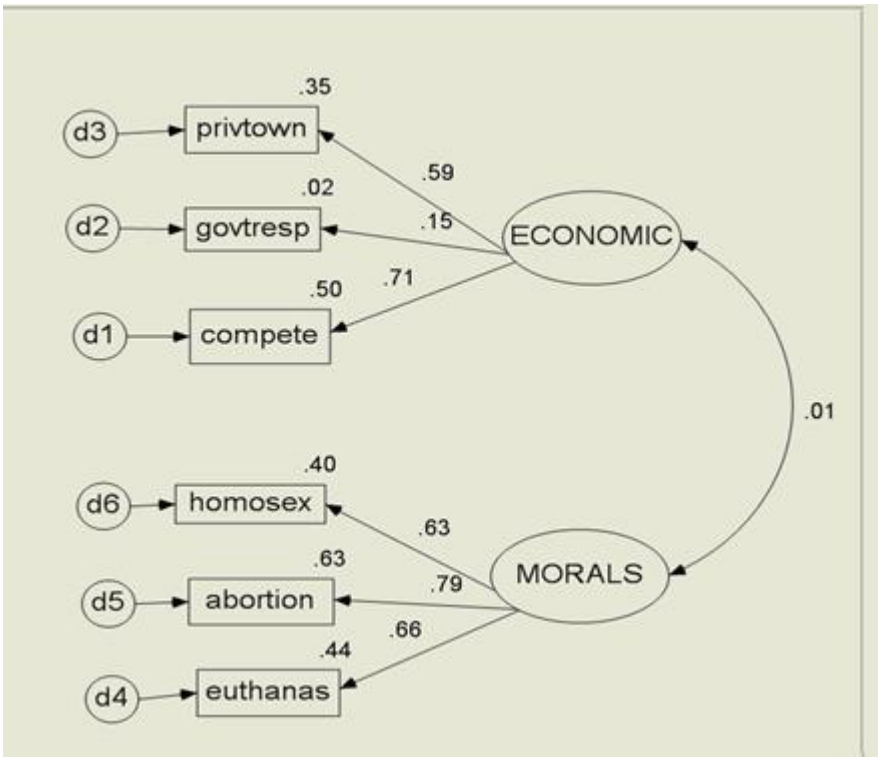
b.決定研究構念

根據研究文獻，研究者假設左派與右派政治態度的差異，可以劃分為二個主要的層次：對「社經」的態度、及對「道德」的態度。每一種態度即為一項研究「構念」，每個構念包含三個測量表項目。

c.選擇統計軟體

假設使用SPSS的Amos。

d.跑出CFA因徑圖



(2)檢定的標準與詮釋

橢圓形的是先驗假設的構念：社經、道德
矩形是項目
圓形是誤差
線條上的數字是相關係數
矩形右上方是決定係數，即可解釋變異量百分比。

a.輻合效度(convergent validity)
即構念與項目的相關係數，詮釋如下：

.9 < r　　效度很高
.7< r <.9　效度中至高
.3 < r <.7　效度低至中
r < .3　　效度低

b. 區別效度（discriminant validity）
即各因素之間的的相關係數，要愈趨近於0愈好。
本例是0.01，可以通過檢定。

CFA的進一步檢驗，是比較「觀察變項之共變項矩陣」與「理論模式中的共變項矩陣」的卡方分析。卡方分析有2型，最常見的是比較觀察值和期望值是否有顯著差異；而第二型「適合度分析 (Fit of Goodness)」剛好反過來，分析觀察值是否符合「理論值」（這時可能不是純隨機的期望值）的分配。所以，「適合」的指標是：

(a)卡方值必須接近於0。

(b)同時，檢定不得到達差異顯著水準。（因未達顯著水準經常是樣本不足造成的，所以「同時」的觀念非常重要。）

另外，也有一些學者提供了其他的檢定指標。

D:\TxData\LW\SKMS\@參考文獻.docx