職能治療學會雜誌
民國八十五年十四卷第一期

特稿

從心理計量的觀點看測量工具的發展

姚開屏

摘要

在職能治療的臨床及研究中，良好的評估工具是非常重要的。一個好的評估工具能引導職能治療人員了解病患的病情變化及進展情形，因而做出適當的判斷，並進而針對病患現況作最適當的處置。本文從心理計量的觀點來談測量工具該如何發展，文中將先從歷史的角度來看測量工具的進展，並復習一些基本的測量相關概念。另外，將從古典測驗理論的觀點來談現今最被廣而使用的測量工具發展之技巧，其中內容包括了信度及效度的介紹、測驗建立的步驟，及試題的量化分析方法等。最後，本文將介紹一個現代測驗理論——項目反應理論，以提供給大家在發展或修訂測量工具時的一個參考。（職能治療學會雜誌 1996; 14:v-xxi)

關鍵語：古典測驗理論、項目反應理論、信度、效度

前　言

受中華民國職能治療學會雜誌主編的邀請，本人將去年十月在學會主辦的精神科職能治療再教育的演講及十一月在台中市立復健醫院的演講整理成此篇，希望藉由此篇對職能治療從業人員在發展或修訂測量工具時有所幫助。本篇將省略許多理論部份，而著重在基本觀念的介紹。

本人自去年八月底回國至今的五個月期間，已被近十人次詢問協助發展或修訂適合國人的評估工具，我很高興看到國內醫療人員對這方面已展開具體的行動。我相信國內醫療人員可能早就了解到直接使用國外的評估工具或自己設計的評估工具，在信度及效度方面皆需進一步的考量，然而在過去因爲沒有足夠的知識及技術針對這個問題做更深入的探討，因此也就多半將就現有的工具使用，而不太管這樣的評估究竟是否有效又是否可信。我自己在國內任職能治療師時也早就了解評估工具的信度及效度的重要性，後來在美國唸書寫第一個碩士的論文時，也順便研究了當時台大醫院復健部成人職能治療所常用評估工具的信、效度，這些工具包括了測手部動作的布氏

國立台灣大學心理系講師

受文日期：85年2月1日
接受刊載日期：85年2月5日
索取抽印本聯絡人：姚開屏，台北市羅斯福路四段一號（國立台灣大學心理系）

手部動作恢復階段 (Brunnstroms Arm/Hand Recovery Stages);日常生活自理能力；動態及靜態站坐平衡；認知知覺功能測驗等，並且將結果寫成文章向學會的雜誌投稿，然而可能當時編審者不太了解我的目的而我人又在國外難以聯絡，因此在我不知情的情況下，最後文章被登出時只剩下對信、效度的文獻回顧部份。[1]事隔多年，我很高興大家對評估工具信度及效度的問題已開始重視，並且已展開實際的研究行動，希望此篇的整理能或多或少提供給各位幫助。

歷史回顧及基本觀念簡介

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」。[2]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。若個體以人爲主，我們常稱此種測量爲測驗 (tests)。測驗的實際用途主要包括了選擇 (selection)、分類 (classification)、評鑑(evaluation)及諮商(counseling)。例如:我們用測驗來選擇出適合做某種工作的人；用測驗將神經質者與憂鬱症者分別出來；我們也可用測驗來評鑑學生學習成果；另外，測驗也可被用在職業輔導及婚姻諮商。在臨床的使用上，測驗也常被用來做爲評定 (assessment)、診斷 (diagnosis)及預測(prediction)的工具，因而測驗必需具有相當的水準(例如:良好的信、效度)才能測到我們所期望得知的。倘若所使用的測驗不良，則我們無法就測驗結果對受測對象的情形做了解、下一定論或甚至預估受測對象的未來發展情況。

談到測驗(或說心理測驗)的發展史，[3]我們可從三類來源來看，即 (a) 出任公職的考試，(b) 學校測驗的使用，以及 (c) 對個別差異 (individual difference) 的研究。遠自公元2,200年前在古中國已有用考選的方式來評定適任公職人員，然而對此所知文獻有限；在公元1,000年前則已有清楚文獻記載當時的皇帝使用考試的方式來甄選適任政府部門的官員，當時所考的項目包括有馬術、音樂、寫作、儒家思想、法律、公私禮儀等；而在西元1905年中國的科舉考試廢除時，英美等國卻正大量採用考試的方式以選取公務人員。在學校測驗的使用方面，直至十二世紀後，歐洲的學校才在紙張發明後開使使用筆試，到了十六世紀，許多教會學校已使用測驗來評量學生。在對個別差異的研究方面， 起始於英國人 Galton (1822-1911)在倫敦的實驗室專門測量人們的感覺知覺及動做反應，每個人可以花三便士的錢到他的實驗室測出自己的反應情形並與他人做比較，在美國的 Cattell (1860-1944) 也有類似測量知覺與操作反應的個別差異研究。 法國的 Binet (1857-1911) 於1905發展出第一個個別智力測驗，而德國人Stern (1871-1938) 則發展出有名的智商分數 (IQ, intelligence quotient)是等於心智年齡(mental age)除以實際年齡 (actual age)。而在團體測驗方面的發展則是美國陸軍在第一次世界大戰時期，以集體施測的方式選取適合從事特定任務的人，之後各種如人格測驗、成就測驗、態度測驗、性向與職業測驗等也就如火如荼的接著發展出來，在近四十年間新的測驗理論及研究也一一出現，例如：項目反應理論(item response theory)、類化理論(generalizability theory)、電腦適性化測驗 (computer adaptive tests)等。綜觀心理測驗研究的進展，實際上是伴隨著統計技術的發展而成，如 K. Pearson(1857-1936) 的相關係數的發展對基本測量理論(basic measurement theory)的形成有不可磨滅的影響，而C. Spearman (1863-1945)的研究更對日後測

驗信度(test reliability)及因素分析法(factor analysis)開創了先河。

在我們繼續談測驗理論及實務之前，讓我們先復習一些重要的觀念。Stevens (1946)在談到測量的定義時也提出四種測量尺度(scale):類別或說名義(nominal)、等級或說順序(ordinal)、等距(interval)及等比或說比率(ratio)。所謂「類別」是指所用以代表個體的數值只有名義區別上或說分類上(classification)的意義，例如球員球衣號碼、性別等。所謂「順序」是指所用以代表個體的數值有次序大小之關係的意義，但數值間的差距不一定相等，例如考試名次、等第，第一名及第二名之間能力的差距與第二名及第三名之間能力的差距不一定相等，但這三個名次有大小順序的關係。所謂「等距」是指所用以代表個體的數值彼此間的差距是有意義且可做比較的，例如天氣溫度攝氏40度比攝氏30度高出10度，而這10度的差距是與攝氏30度比攝氏20度高出10度的差距是相等的。最後，所謂「等比」是指所用以代表個體的數值彼此間的比率是有意義的，且有絕對零值(absolute zero)的存在，例如對身高或體重的測量，零公分或零公斤就是眞的什麼都沒有，不像攝氏零度還是有溫度存在，只是人們定義在那種溫度之下叫攝氏零度。另外，20公分是10公分的兩倍長，五公斤只有拾公斤的一半重，而八秒鐘是二秒鐘的四倍長，但若說攝氏40度的水是攝氏20度的水的兩倍熱則就不合適，以上這些例子都指出等比尺度的數值彼此間的比率及絕對零值是存在的。表一就以上的四種尺度在四個特性方面(區別能力、幅度順序關係、等距關係、絕對零值的存在與否)做一總結，「o」者表示特性的存在，「x」者表示特性的不存在。

另一個重要的觀念是一二個變項之間的相關性(correlation)。表二總結了當這二個變項在使用不同的尺度下測量時所需使用的相關法。我們所熟知的皮爾森相關(Pearson correlation)只適合用

表一　測量的尺度

	等級 類別 (nominal)	順序 (ordinal)	等距 (interval)	等比 比率 (ratio)
區別度 (distinctiveness)	0	0	0	0
幅度順序(order in maganitude)	X	0	0	0
等距 (equal intervals)	X	X	0	0
絕對零值 (absolute zero)	X	X	X	0

表二　二個變數相關的形式

	相關	其中二分變數可假設有常態分配(normal dist)
二變數皆爲二分(dichotomous)	phi(ϕ)相關	四分相關 (tetrachoric)
一爲二分,一爲連續(continuous)	點值雙列相關 (point-biserial)	雙列相關 (biserial)
二變數皆爲連續	皮爾森相關 (Pearson)	
		假設二變數有常態分配
二變數皆爲等級	Spearman Coeff. Kendall τ	多序類相關 (polychoric)
一爲等級 一爲連續	多列相關 (polyserial)	
一爲等級 一爲二分	多序類相關 (polychoric)	

在當二個變項皆使用連續 (continuous) 的尺度 (即等距或等比尺度) 來測量時。 如果我們想求兩題使用李克式量表(Likert scale)題目之間的相關，因爲李克式量表是一種等級尺度， 由數點等級所組成(例如: 1－2－3－4－5；非常同意--同意--不同意 -- 非常不同意) ， 則考慮使用 Spearman coefficient或Kendall較恰當。如果我們想了解某班學生的性別與答某題目之對錯的關係， 因爲性別及答題的對或錯皆只有兩種可能(二者皆爲二分變項，即男或女； 對或錯) ，則使用的相關稱做Phi(Φ) 相關。 如果我們想了解某班學生答某題目之對錯與考試總分 (連續變項) 的關係，則使用點值雙列相關 (point-biserial correlation) 較合適。另外， 若對使用的二分變項或等級變項可假設其背後潛在的特性乃是常態分配， 則我們會用不同的相關如四分相關(tetrachoric correlation)、雙列相關(biserial correlation)及多序類相關(polychoric correlation)等。 例如 : 如果我們想了解某班學生答某題目之對錯與考試總分 (連續變項) 的關係，若我們可假設學生答此題目之能力乃一常態分配， 而能力超過某閾值 (threshold) 時會答對， 反之則答錯， 則這種答此題目之對錯與考試總分的關係是爲雙列相關。在使用多序類相關方面，乃是對於所使用的兩個等級變項背後做常態分配之假設， 而每一變項皆使用兩個或兩個以上的閾值以區分出所觀察到的等級情形。

復習了以上的基本觀念後， 接下來就從古典測驗理論(classical test theory)著手來談我們所關心的話題 -- 甚麼是信度、甚麼是效度、發展一個良好測驗的步驟， 及如何分析測驗內容之好壞而加以修正。 之後將介紹一個現代的測驗理論 -- 項目反應理論， 以提供給各位在考慮發展新的測驗時的另一種參考。

古典測驗理論

「古典測驗理論」 聽起來似乎是個古老又遙遠的東西， 然而各位所了解的測驗概念 (如信、效度) 或從文獻中所看到的測驗研究， 十之八九是從古典測驗理論而來的， 因此我們先談一下甚麼是古典測驗理論。 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: (3)

(1)$X = T + E$，即觀察值=眞實值+誤差值。我們實際上無法直接觀察受試者的眞實值或眞正能力， 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。這種誤差有時大於眞實值也有時小於眞實值，但總平均起來誤差爲零。由於此隨機誤差的存在，因此即使受試者的眞實值 T 是固定不變 (fixed) 的， 我們每一次的觀察值卻不一定都相等，不過觀察值的分配亦爲常態分配。

(2) $\varepsilon(X) = \varepsilon(T + E) = T$，觀察值的期望值=眞實值。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) 是受試者的眞實值， 而誤差值的期望值等於零 ($\varepsilon(E) = 0$)。因此雖然測量有誤差，但如果我們收集到足夠多次的觀察值，則這些觀察值的平均值可被視爲眞實值的指標。

(3) $\rho_{ET} = 0$，誤差與眞實值不相關。比較眞實值較高的人與眞實值較低的人， 測眞實值較高 (或測較低) 的人時不會有系統性的有較高 (或較低) 的誤差， 也就是說一個人眞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$\rho_{E1E2} = 0$，兩測驗間之誤差不相關。

一個人在一測驗上有較高的誤差，不一定在另一測驗上有較高(或較低)的誤差。這個假設只有在某些情形下才不成立，例如當受試者疲倦(fatigue)時、前二次測驗產生了練習效果(practice effects)時以及受試者受情緒或環境因素影響時。

(5) $\rho_{E1T2} = 0$，一個測驗的誤差與另一個測驗的眞實值不相關。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'爲其相對觀察值)皆符合假說(1)至(5)，且兩測驗有相同眞實值(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) 主要眞實值相等測驗 (essentially τ -equivalent tests):若兩測驗(X_1及X_2爲其相對觀察值)皆符合假說(1)至(5)，且兩測驗的眞實值差一個常數($T_1 = T_2 + C_{12}$)，則此二測驗稱作主要眞實值相等測驗。

根據這些假設，我們可導出測驗的信度(reliability)。「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。由於測量誤差愈小，信度就愈高，因此信度可視爲測驗結果受機遇影響的程度。通常我們用相關係數(correlation coefficient)來表示信度的大小，從心理計量學的觀點來看，信度即是指兩平行測驗間觀察值的相關($\rho_{XX'}$)或說是觀察值與眞實值間相關的平方(ρ_{XT}^2)。通常「信度」可分爲下列四種：[1-2,4]

(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。

(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施測時間宜短，以減少病人因隨時間而成熟變化，然而時間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。再測信度是指同一受測者前後表現的一致程度，而施測者內信度(intra-rater reliability)則是指同一施測者對相同受測者前後施測是否一致的程度。

(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可採用折半信度法，以了解測驗本身內容是否相互符合，因此此法又稱爲內部一致性(internal consistency)。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，這種方法只需施測一次即可得相關係數，而測驗題數愈長所得折半信度愈可靠。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。另外，測驗題不僅可被二等分以計算測驗的內部一致性，還甚至可被分成兩份以上，此時我們常用以計算內部一致性的公式包括了斯布氏公式、克氏阿爾法、庫李公氏-20 (Kuder-Richardson formula 20，KR-20)及庫李公氏- 21 (

Kuder-Richardson formula 21，KR-21)等。

(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。不過複本的產生並非容易，必須在題數、型式、難度、鑑別度等方面皆與原本一致。

在效度(Validity)方面，所謂「效度」是指正確性，即能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」。效度愈高，表示愈能測出受測者的特質，因此自行設計施測工具或使用標準化的工具，「效度」是最重要的條件。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。通常「效度」可分爲下列三種類型:[1-2,4]

(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：欲測量中風病人的認知與知覺能力，是否所使用的施測工具能適當的反映出一個人的認知與知覺能力，是否工具已適度的涵蓋了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」是指一個測驗主觀上有沒有效的程度。而想達到「邏輯效度」則需對要被測的研究特質細心的定義範圍並且經由邏輯的設計而找出涵蓋所有重要的特質元素。以上不論是那一種內容效度皆受主觀判斷的因素影響。

(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。例如:已知測驗A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。又若想設計測量工具以預測病人的日常生活自理能力，除了對病人施予此種測量外，還同時測量了效標--日常生活自理能力，將這二者作相連結以求得的效度是爲「同時效度」而非常被人誤稱的「預測效度」，這是因爲施測者在同一時間測量二者，雖然目的是想用某測量結果以預測另一結果，但進行的方式是在同一時間以相關或迴歸的方式進行。這好比做迴歸分析(regression analysis)時，同時收集自變項及依變項，而用自變項來「預測」依變項一樣。雖然你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，同時效度所謂的「預測」不應與接下來所要談的預測效度混爲一談。

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，以得知二者間的相關程度來判定此測驗的預測效度。在職能治療的文

獻中也不難發現對「預測效度」的研究，例如參考文獻五研究幼兒兩歲時在貝李氏嬰兒發展量表(Bayley Scales of Infant Development)上的得分以預測幼兒四歲半時在動作及認知方面的能力間之關係。由以上例子可知時間的間隔與否是區別「同時效度」與「預測效度」的最大因素。另外，在研究效標關聯效度時，所使用的效標水準很重要，因爲與一個不好的效標求相關所得的結果並不能使我們了解我們的測驗是否達到可被接受的效度。

(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的完全，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，而後就施測的結果來看是否符合理論，若否，則需修改測驗再施測，又有時也需考慮理論及假設的適當性是否需修正，經過如此這般來來回回重覆的過程後，而得到有建構效度的測驗。求建構效度所使用的方法沒有絕對的依據，你可用相關法、實驗法、因素分析、因徑分析等各種可能方法達到目的。參考文獻六至八提供了美國職能治療人員對中風病人測驗的建構效度之研究。

信度僅指測量結果是否一致、可靠、穩定，卻不涉及測量的東西是否正確；效度則針對測量的目的探討測驗能否測出所想要測量的特質。因此有可能一個測驗的多次結果皆非常的一致，但並不能測出我們所想要測的。我們可說信度是效度的必要而非充分條件，有信度的測驗不一定就有效，但有效度的測驗必定有信度。

測驗建立的步驟[9]

(1) 計劃:要發展一個新的測驗，首先需要有周詳的計劃，因此(a)先要確定測量的範圍與目的以及測量的對象，(b) 分析構成欲測量的特質或行爲的各元素，(c)搜集與這些特質或行爲的各元素相關的資料，(d) 設計測驗藍圖以做爲編製測驗的依據。例如:想編製一個測幼兒精細動作發展的測驗，第一步需先確定甚麼是精細動作及測量對象的年齡範圍，再來則要分析精細動作包含那些細項目，而後要搜集與這些項目相關的資料文獻，以便訂定測驗設計藍圖，作爲日後編製測驗的依據，如此的測驗才會具有代表性、涵蓋了測精細動作的各種細節，達到測量的目的。

(2) 編製:當設計測驗的藍圖確定後則可以著手開使編製測驗，測驗的編製有一般性的原則及視測驗的類型而不同的特殊原則。許多談編製測驗的書籍皆會談到這些原則，在此不贅述。

(3) 預試:測驗編製完成後需取樣做預試，以了解此種新編的測驗是否適當。取樣時需注意樣本的代表性，宜取自將來正式測驗擬用的群體中，另外，施測之狀況(如環境、指示語、施測方式)需與將來正式施測時的情況儘量相近，又應該記錄受測者的各種反應以作爲日後修改測驗的參考。

(4) 分析:可就預試時所收集到的資料做試題的分析，以作爲日後修改測驗的依據。分析的方法分質的分析 (qualitative analysis) 與量的分析 (quantitative analysis)。所謂質的分析是對試題的內容與形式，從涵蓋欲測特質的適合度及編製技術加以評量。量的分析方面則是分析試題

的難度(item difficulty)與試題的鑑別度(item discrimination)，這部份的細節將在下一節中討論。

(5) 修正：試題選取的標準主要是依照試題的難度及鑑別度分析的結果。通常我們會選取鑑別度高的題目，如此才可區分不同程度的受試者，另外，我們會選取難易度適中的題目，太容易的題目多半的受試者皆會或太難的題目多半的受試者皆不會，都不適合被選用。當一些題目不適合時，我們可修正它們，修正完後則回到步驟(3)再測試、步驟(4)再做分析、步驟(5)再修正，重覆此過程以得到最後可被接受的測驗。

(6)複核(cross validation):由於前面預試樣本可能有取樣的誤差，因此以上所做的分析不一定可靠，需再選取另一有適當代表性的樣本再測試、再做分析，以了解是否前後兩次試題的難度及鑑別度分析的結果一致，以複核試題的特性。

(7) 常模的建立：若是建立常模參照測驗 (normed-reference tests)，則需建立常模(norm)。常模的功用是可得知一個人在團體中的相對地位如何，常見的常模形式是百分等級 (percentile rank) 和標準分數 (standard score)，百分等級說明了一個人在群體中的表現勝過百分之多少的人，而標準分數則表示一個人的表現相距群體的平均值多少個標準差，在建立常模時樣本選取的代表性及常模的建立步驟都需要非常注意。

試題的量化分析[3,9,10]

(1) 難度(item difficulty)分析：試題的難度分析有許多方式，茲舉較常見的兩種方式。

(a) 試題的難度 (P) 被定義爲全體受試者答對或通過該題的百分比(percentage passing)

$$p = \frac{R}{n} \times 100\% \quad ,$$

R =通過該題的人數，

n =全體人數 。

從這種定義來看，所謂「試題的難度」可視爲試題的容易度(item easiness)，因爲P愈大表示試題愈容易。對某些測驗而言，並非用通不通過(答對或答錯)來計分，而是以有或沒有(存在或不存在)某項特質爲依據，因此對這類測驗，「試題的難度」即等於全體受試者擁有該特質的比例。

(b) 依受試者總分高低排列， 各取得分最高及得分最低的27%受試者，P_H表示高分組通過某題人數百分比；P_L表示低分組通過該題人數百分比，則試題難度(P)等於

$$P = \frac{P_H + P_L}{2} \quad 。$$

我們爲何取27%高、低分組爲標準呢？這是因爲當受試者總測驗分數的分配是常態時，取上下各27%的受試者會產生最好的P估計值。通常我們可取上下各10%到33%之間的受試者來計算P值。

從統計二項分配(binomial distribution)的原理來看，當P = 0.5會使題i的變異量$P_i(1-P_i)$達到最大，也就是使受試者之間的差異會變最大，這表示此題比較可能區別出不同程度的受試者，這也是爲什麼前面提過我們會選取難易度適中的題目，而太容易的題目(P趨近於1)多半的受試者皆會或太難的題目(P趨近於0)多半的受試者皆不會，都不適合被選用。然而更進一步的考慮到試題的類型及試題間相關性等因素，我們最好選取試題的難度範圍是在0.3到0.7之間。

(2)鑑別度(item discrimination)分析也有許多方式，茲舉較常見的兩種方式。

(a)內部一致性(internal consistency):乃是檢驗個

別試題與整個測驗的一致性。一個有鑑別度的試題應該與整個測驗的走向是一致的，也就是說測驗分數高的受試者要比測驗分數低的受試者有較高的可能答對某題目，否則此題目並不能反應出受試者的實力，因此並非是個良好的題目。通常我們可用相關法如雙列相關或點值雙列相關，以了解是否得總分高的受試者答對某題的比例就高，前種相關法與後者之差別在於前者有常態分配的假設存在(見本文前面觀念簡介的部份)。我們也可用下列方法求試題鑑別度(D)，

$$D = P_H - P_L$$

P_H與P_L的定義與前面相同

這種方法是比較高、低分組的受試者在個別試題上通過人數的百分比，D愈大表示試題愈能鑑別出高、低分組的受試者，並且個別試題與測驗總分的一致性愈高。下面是依照D的大小來選取試題的簡單原則:

$D \geqq 0.40$	此題非常優良
$0.30 \leqq D \leqq 0.39$	此題良好，修改更佳
$0.20 \leqq D \leqq 0.29$	此題尚可，仍需修改
$D \leqq 0.19$	此題不良，必需修改

(b)外在效度(external validity):即試題的效度分析。在這裡我們指的是個別試題的效度，做法與前面求試題與整個測驗的內部一致性很相似，只是所用的效標不同。在此是以外在效標(external validation criterion)而不是用整個測驗分數爲依據，來衡量試題反應與外在效標分數的相關程度。

(3)試題信度(item reliability)與試題效度(item validity):

假設我們要從所有n個試題中選取k個試題，使得這k個試題所組成的測驗有最大的內部一致信度$r_{XX'}$(internal consistency reliability)及效標關聯效度r_{XY}(criterion-related validity)，則該如何找出這k個試題?從另一個角度來看，假設我希望測驗的內部一致信度及效標關聯效度至少達到某一水準(desired minimal value)，則k應是多少?以上這兩個問題皆可從對試題信度與試題效度的分析得到答案。

用X*表示以n個試題爲準的測驗總分，X表示以k個試題爲準的測驗總分，Y表示我們有興趣的外在效標分數。假設試題i的困難度是P_i，則試題i的變異量是$s_i^2 = P_i(1-P_i)$，而試題信度被定義爲$s_i\, r_{iX'}$($r_{iX'}$是試題i的分數與n個試題測驗總分的點值雙列相關)，試題效度被定義爲$s_i\, r_{iY}$(r_{iY}是試題i的分數與外在效標分數的點值雙列相關)。我們可計算以這k個試題所組成的測驗的平均值、標準差、信度及效度估計值分別如下：

$$\bar{X} = \sum_{i=1}^{k} p_i \quad ,$$

$$\hat{s}_X = \sum_{i=1}^{k} s_i r_{iX'} \quad ,$$

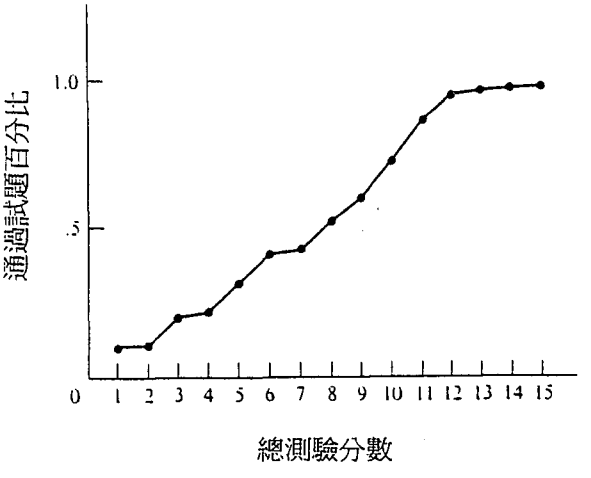
$$\hat{r}_{XX'} = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} s_i r_{iX'}\right)^2}\right] \quad ,$$

$$\hat{r}_{XY} = \frac{\sum_{i=1}^{k} s_i r_{iY}}{\sum_{i=1}^{k} s_i r_{iX'}} \quad 。$$

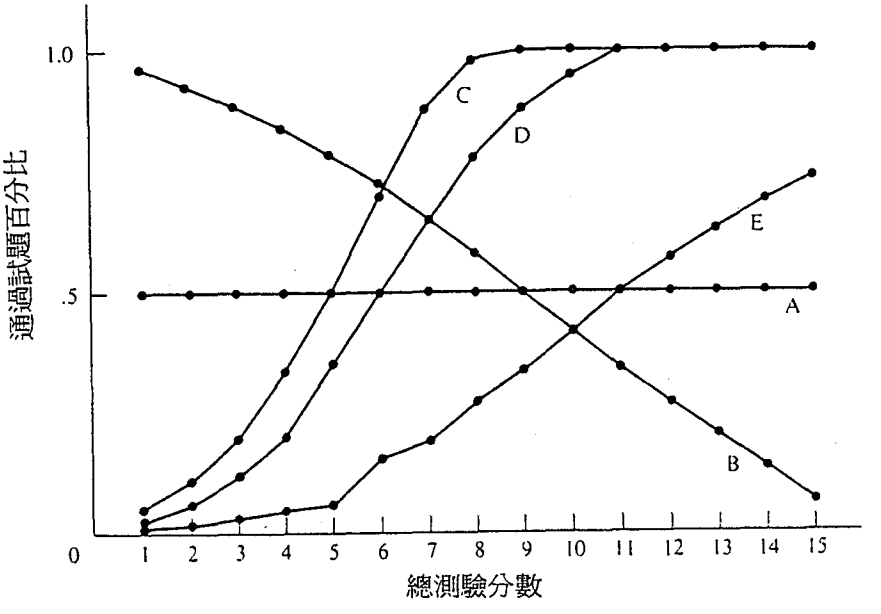
從上面第三個公式我們可知若想使得 k 個試題所組成的測驗求得最大的內部一致信度$r_{XX'}$，則 r_{iX}. 要愈大愈好，也就是說這些試題分數與n個試題爲準的測驗總分相關要愈大愈好。從上面第四個公式我們可知若想求得最大的效標關聯效度r_{XY}，由於r_{XY}一定在-1及+1之間，因此 r_{iY} 應與 r_{iX}. 愈接近愈好，也就是說試題效度應與試題信度愈相近愈好，我們即依照這些標準來選出k個試題以組成測驗。

(4) 因素分析法 (factor analysis) 分析試題間的關係且用較少的因素 (factors) 以解釋這些關係，因此可被用來做建構效度的分析。另外，因素分析法也可被用來選擇試題並發展測驗，幫助我們了解是否一組的試題間彼此是同質的 (homogeneous)，如果答案是肯定的，則表示某因素影響了此組試題，這些試題可被考慮編入測驗中。使用因素分析法需要許多細心的思考及實際的經驗才能夠應用分析並解釋結果得當，國內雖然有許多人用此法來發展測驗，但可能對因素分析法的本質了解不清，因此在使用及解釋上會有不適當的情形出現。[11]

(5) 項目特徵曲線(item-characteristic curves, ICC)是一種表示通過某特定試題的比率及受試者在測驗方面潛在特質(latent trait)情形之間關係的圖示。所謂「潛在特質」是指一種假設而且不能被直接觀察到的特質，例如：內外向、智力、動機等。由於不能直接觀察到潛在特質，通常我們用測驗的總分作爲對潛在特質能力的一種估計。圖一是某題的ICC，橫軸是整個測驗的總分，縱軸是通過某特定試題的比率，例如：測驗總分爲 8 的受試者約有 50% 答對此題。根據此ICC，這一測驗題是一合理的題目，因爲此圖形線的斜率(slope)是正的，它反應出能力愈好的受試者通過該試題的比例愈高。ICC 所提供的訊息與前面我們所談的試題難度與試題鑑別度有相似之處，我們可用ICC的斜率來表示試題鑑別度，而用相對於50%受試者通過某試題的測驗總分來表示試題的難度，因此圖一試題的難度是8。圖二是五題試題的ICC，就試題鑑別度而言，試題A是很不好的題目，因爲它無法區別出不同能力的人(斜率爲0)；試題B也是很不好的題目，因爲能力强的人反而無法通過該題(斜率爲負值)；試題C與D是不錯的試題(斜率爲正值且夠大)；試題E的ICC較試題C與D的ICC平，可知試題E的鑑別度較不如試題C與D，也就是說當受試者的測驗總分差距不大時，相較於試題C與D，試題E較無法有效的區別受試者的能力。另外，就試題難度而言，試題C、D與E的難度分別是5，6及11，因此我

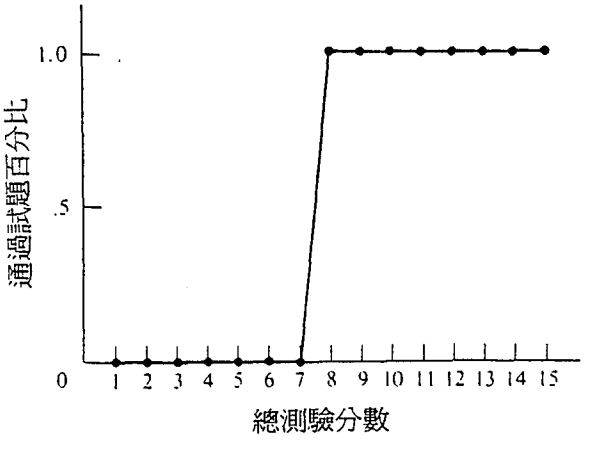


圖一　某題項目特徵曲線圖（ICC）

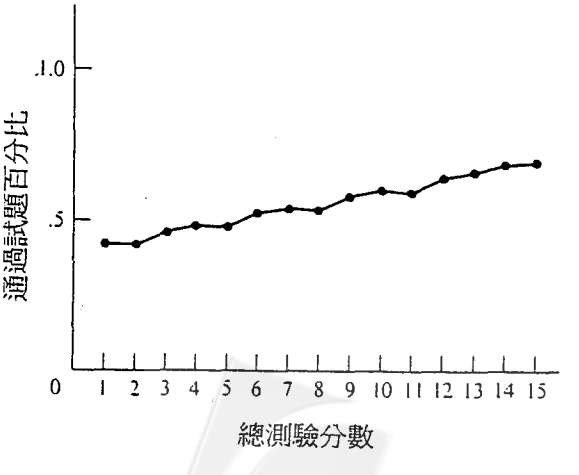


圖二　五題項目特徵曲線圖

試題C、D與E的難度分別是5，6及11，因此我們可知試題E要比試題C、D的難度高。圖三表示一個具有非常好鑑別度的試題的ICC，其斜率幾乎是1，此試題能區別測驗總分低於8分及等於或高於8分的人，也就是說測驗總分低於8分的人不能通過此題，但測驗總分等於或高於8分的人，會通過此試題。相對來說，圖四表示一個鑑別度不好的試題的ICC，因爲測驗總分高及測驗總分低的受試者皆有相近的通過率。一個好試題的ICC應該斜率爲正而且難度中等。



圖三 鑑別度好的項目特徵曲線圖



圖四　鑑別度不好的項目特徵曲線圖

綜合以上的討論，一個好的測驗必須有:[9]

(1)適切度(relevance):其中包括了

(a)代表性(representation):對欲測特質需有充分的代表性，而試題需有相當的數量且試題對欲測特質的內容分配比例適當。

(b)客觀性(objectivity):從試題的編選、施測的步驟到試題的分析 修正皆需有客觀的標準。

(c)標準化(standardization):對施測情境、施測指示語等應有清楚的規定。另外，依照測驗目的的不同而可能有常模(norm)的建立，在建立常模的步驟及方法上也需非常小心。

(2)信度(reliability)即指測驗是可靠、一致而穩定的，細節已在前面討論過。

(3)效度(validity)即能測出所想要測出的特質，細節已在前面討論過。

項目反應理論

前面所談的「古典測驗理論」有一些缺點，例如:試題的統計量(難度及鑑別度等)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；有可能需使用平行的複本測驗，但建構複本並不是一件容易的事；另外，使用古典測驗理論無法區別及預測每一個人的能力，因此有現代測驗理論--「項目反應理論」(item response theory, IRT)的發展。項目反應理論根基於潛在特質模型(latent trait models)，它是1950年代被提出的，相較於古典測驗理論，它可算是現代的測驗理論。IRT認爲受試者的表現受其潛在特質的決定，而其最基本的假設是「局部獨立」(local independence)，也就是說受試者在別題上的表現並不影響此題的表現，並且別的受試者的表現並不影響此受試者的表現。IRT是用數學函數的方式來說明不可直接觀察的潛在特質（如數學推理能力、智力等)與試題作答情形(如答對的機率或有此特質的程度)的關係。這種數學函數(稱爲項目特徵函數-- item characteristic function)的圖形如同前面談過的ICC，只是橫軸使用潛在特質，經標準化後的潛在特質數值的範圍可從-∞到+∞。此理論的目的是要估計受試者在連續的潛在特質上所站的位置(即能力大小)，另外也估計兩個重要的參數--試題難度(item difficulty)及鑑別力(discriminating power)。參數的估計有助於試題的選擇，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(即能力的大小)。IRT常使用的兩個數學函數$P_i(\theta)$(就試題i)是

(1)常態分配肩形曲線模型(normal ogive models):

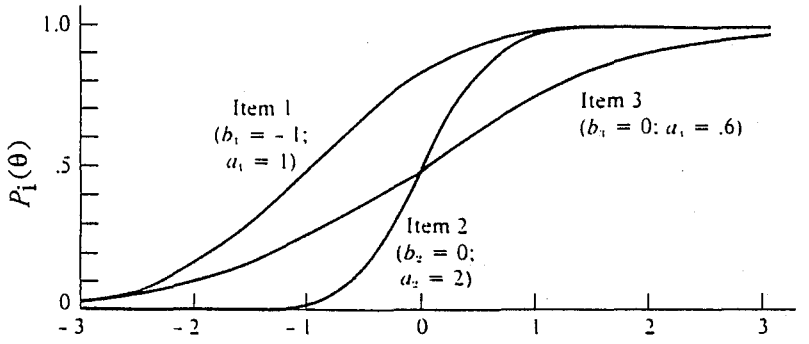
$$P_i(\theta)=\int_{-\infty}^{a_i(\theta-b_i)}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz$$

常態分配肩形曲線也可說是累積常態分配曲線(cumulative normal distribution)，見圖五爲例。

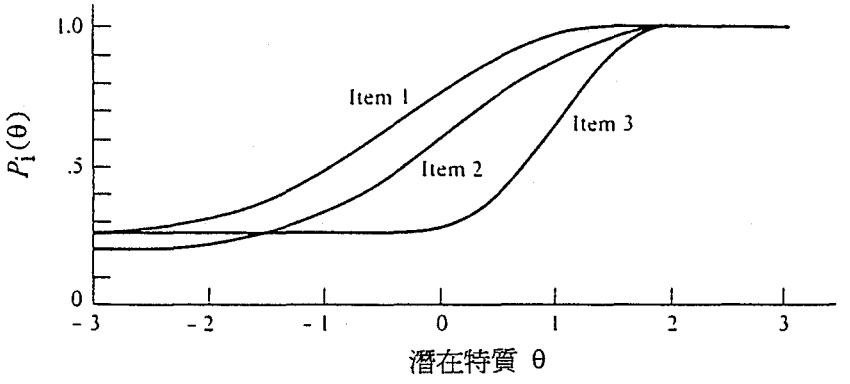
(2)洛基斯第(logistic)模型:

$$P_i(\theta)=\frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

以上兩種模型中表示受試者的潛在特質；a_i表示試題i的鑑別力，即曲線斜率；b_i表示試題i的難度，即50%的受試者通過此題的相對潛在特質的位置；Z是標準分數；D是常數1.7。此二模型的分配相似，而後者在數學算式上較簡潔，因此較常被使用。以上這兩種模型又統稱爲二參數模型(two-parameter models)，因爲模型中只有兩個試題參數a_i及b_i。圖五是三個試題的二參數模型圖，各題的鑑別力及難度皆標明在圖上。例如:試題二與試題三的難度相當，皆在能力爲平均值處，即經標準化後的潛在特質數值爲零處($b_2=b_3$



圖五　三個試題的二參數模型圖



圖六　三個試題的三參數模型圖

=0)，而三題中試題二的鑑別力最好(a_i最高)。

有些情況下，即使受試者的能力很低仍能正確的答對試題，例如：四選一的選擇題即使完全用猜的，也有0.25的機率答對，則我們可考慮使用三參數模型(three-parameter models):

$$P_i(\theta) = c_i + (1-c_i)\left[\int_{-\infty}^{a_i(\theta-b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz\right] \quad ,$$

$$P_i(\theta) = c_i + (1-c_i)\left[\frac{1}{1+e^{-Da_i(\theta-b_i)}}\right]$$

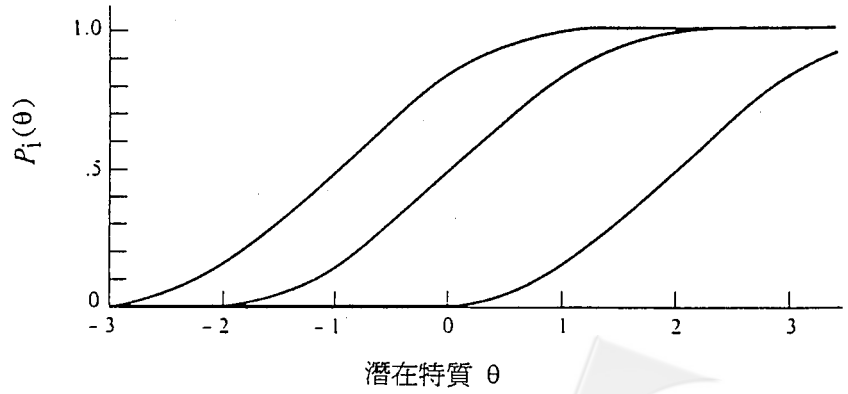
也就是多加了一個參數 c_i，稱爲猜測參數(guessing parameter)或漸近線參數(asymptote)。圖六是三個試題的三參數模型圖，每一題相對於最小潛在特質的試題通過率皆非從零開始，也就是說即使能力再差的受試者也會有一定比例通過此這三題。

另外，若我們假設沒有 c_i 的存在，並且每題的鑑別力皆相等(皆等於a)，則我們可得到一參數模型(one-parameter model)如下，此模型又稱爲羅序模型(Rasch's logistic models):

$$P_i(\theta)=\frac{1}{1+e^{-1.7a(\theta-b_i)}}$$

一參數模型只用一個位置 (location) 參數或說試題難度 b_i 來表達試題曲線。圖七表示三個試題的一參數模型圖，每題皆有相同的鑑別力，因此三條曲線平行，但因爲難度不同因此各題的位置不同，這三題的難度相對潛在特質的位置分別是 -1，0，1。一參數模型使用了一個很强的假設，也就是每一題的鑑別力皆相等且漸近線參數趨近於零，這種假設常被批評爲不切實際、不符合現實，因爲要設計出一個測驗的所有題目皆有相同的鑑別力是非常困難，爲了要得到符合一參數模型的試題，在一開使就必需準備非常多的預備試題以便最後挑選出足夠的試題使用。不過一參數模型的一個好處是此模式非常簡單，若果眞所有題目皆有相同的鑑別力，則受試者答對題數的總和是此受試者的測驗總得分，以此來估計潛在特質能力的大小。但在二參數模型或三參數模型時則不可如此計算測驗總分，因爲若題目的鑑別力彼此不同，則測驗總得分的算法需加權(weighting)各題以得知，加權的方式則與試題鑑別力的大小有關係。在以往我們所使用的測驗，往往就直接用答對題數的總和爲受試者的測驗總得分，而沒有考慮可能應該對各題做加權，這種作法有待商榷。

一般說來，一參數模型由於它的假設而較難符合資料，而二參數模型或三參數模型 (統稱爲多參數模型，multiparameter models) 比較沒有這個問題。主張一參數模型的學者認爲應該是資料來符合模型而非模型來符合資料，而主張多參數模型的學者則認爲剛好相反，應該是能符合資料的模式才是好模式。支持一參數模型 (以美國芝加哥大學及歐洲爲大本營) 及多參數模型的學者 (以美國的教育服務社 ETS 及美國大學測驗社 ACT 爲基地)各說各話，彼此難有溝通與交集。1992年美國教育研究學會 (AERA) 邀請了兩派學者舉行一場辯論，讓兩派學者發表已見，並開放給數百位聽眾提出問題及看法。[12]這個辨論本就無法提供結論，因爲兩派學者的思考方向完全不同，無法說誰的應該是正確的，而誰的應該是錯誤的。



圖七　三個試題的一參數模型圖

以上談了現代測驗理論 -- IRT，究竟它對我們發展測量工具有什麼功用呢?

(1) 選取試題及對試題加權 : 從模型的數學函數圖及參數估計值可知道試題的好壞，以作爲我們選試題的參考。另外，也可知道對各試題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，受試者的得分就不再受使用測驗的版本不同而有所差別，也就是說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，仍能比較彼此潛在特質能力的差異，因爲測驗試題間的彼此對等關係可經由此模型的操弄而連上，這種等質化被稱爲垂直等質化(verticalequating)。它在實用上有其價值，因爲不同程度的受試者可使用不同程度的測驗試題，因而程度差的人只要做較容易的試題而不致於做太難的試題感到挫折，而程度好的人只要做較困難的試題而不致於做太容易的試題感到無聊，最後可將他們放在同一度量 (scale) 上來考慮他們能力的差異。

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計算出參數值，必需要有足夠的樣本才行，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。通常的研究很少能有如此龐大的樣本群，目前只有在與教育相關的成就測驗方面比較有可能， 因而美國的教育服務社 (ETS) 及美國大學測驗社(ACT)成爲全美兩大研究及使用IRT的機構。另外， 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，並不適合用 IRT 。 近些年來已有學者從事多向度 (multidimensionality) IRT的研究，然而爲了模型能多向度化，又必需加上其他的限制及假設，使得使用起 IRT 來礙手礙腳。 而模型的最基本假設「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會影響此題的表現。因著以上的考慮，筆者認爲 IRT 的用意及理想很好，但除非測量工具的設計非常優良符合假說，否則恐怕目前還是在實驗及研究階段使用較佳， 與能很自如的被應用在各領域以發展測量工具還有一段距離。 筆者遂建議讀者在發展新的測量工具時， 可同時使用古典及現代測驗理論，以比較兩種方法的異同，如果結果相同，則對設計出的測驗工具品質更有信心，如果結果有差異，則可進一步的探討差異的原因，研究是否應修改所設計出的測驗。

參考文獻

1. 姚開屏 : 淺談信度與效度。 職能治療學會雜誌 1988；6: 51-54。
2. Stevens SS: On the theory of scales of measurement. Science 1946；103:667-680.
3. Allen MJ，Yen WM: Introduction to Measure-

ment Theory. California: Wadsworth，1979.

4. 簡茂發:信度與效度。楊國樞、文崇一、吳聰賢、李亦園編:社會及行爲科學研究法(上冊)。台北;東華，1978:323-351。
5. Crowe TK, Deitz JC, Bennett FC: The relationship between the Bayley Scales of infant Development and preschool gross motor and cognitive performance. Am J Occup Ther 1987; 41: 374-378.
6. Van Deusen J, Harlowe D: Construct validation of occupational therapy measure used in CVA evaluation: A beginning. Am J Occup Ther 1984; 38: 101-106.
7. Van Deusen J, Harlowe D: Continued construct validation of the St. Marys CVA evaluation: Brunnstrom arm and hand stage ratings. Am J Occup Ther 1986; 40: 561-563.
8. Van Deusen J, Harlowe D: Continued construct validation of the St. Marys CVA evaluation: Bilateral awareness scale. Am J Occup Ther 1987; 41: 242-245.
9. 簡茂發、郭生玉:測驗的編製。楊國樞、文崇一、吳聰賢、李亦園編:社會及行爲科學研究法(上冊)。台北;東華，1978:439-461。
10. Crocker L, Algina J: Introduction to Classical & Modern Test Theory. Holt, Rinehart and Winston, 1986.
11. 翁儷禎:因素分析應用之一覽。章英華、傅仰止、瞿海源編:社會調查與分析。台北:中央研究院民族學研究所，1995: 245-259。
12. 王文中:心理測驗與心理計量。測驗年刊(審稿中)。

J Occup Ther Assoc ROC
Volume 14, Number 1, 1996

Invited Article

The Development of Measurement Instruments from Psychometric Point of View

Kaiping Grace Yao

Abstract

This paper reviews the development of measurement instruments from psychometric point of view. In the beginning, the history of testing and measurement is discussed. Several measurement-related concepts such as measurement scales and various correlation coefficients are reviewed. In addition, this paper presents reliability, validity, the procedures of establishing measurement instruments, and quantitative analysis on test items from the perspectives of classical test theory. Finally, a modern test theory -- item response theory is introduced. (J Occup Ther ROC 1996; 14:v-xxi)

Key words: classical test theory , item response theory, reliability, validity

Lecturer
Department of Psychology
National Taiwan University
Taipei, Taiwan, ROC

Received: Feb. 1, 1996
Accepted for publication: Feb. 5, 1996

Address reprint requests to: Dr. Kaiping Grace Yao
Department of Psychology, National Taiwan University, Taipei, Taiwan, ROC