



研究&方法

# 量表編製理論與應用

Scale Development : Theory and Applications(2e)

對於測驗發展沒有接觸的讀者，

這是一本很好的入門書，

對於已經有相當基礎的研究者，

這本書讀來還是有許多令人豁然開朗之處。

本書重點以清晰明瞭的方式來傳遞基本原理背後的訊息，

並使讀者能夠窺探看起來像是「黑箱」的各種方法。

Robert F. DeVellis 著

鄭皓政 審定

鄭勇剛、龍長權、宋武 譯

PDG

# 審定者序

《量表編製》是我剛開始在心理系教書的時候就開始愛引的一本參考書，當年用的是第一版（1991），目前這一本則是修訂版（2002）。新版書增加了一些新的章節（例如項目反應理論），但是一樣保有精簡扼要的特色，閱讀起來一樣輕鬆自在，對於測驗發展沒有接觸的讀者，這是一本很好的入門書，對於已經有相當基礎的研究者，這本書讀來還是有許多令人豁然開朗之處，尤其是作者簡明清晰的筆觸，豐富的量表編製經驗，讓這本書在歲月的流逝中還是保有他的啟發性，非常難得，我十分樂於向各位推薦它。

這幾年我轉赴企管系任教，重新涉獵人力資源管理領域的理論與實務議題，發現人事測評在資訊科技發達的今天，在組織管理工作中發揮了更具體的作用，尤其是統計技術的高度躍進，雲端技術的成熟備便，讓心理測驗不僅在工具本身乃至於後端的決策應用與管理實務部署，都有嶄新的面貌。最近又看到自己培養出來的學生都有好出路，心裡暗喜當初力排眾議投入冷門的心理計量領域不僅成就自己的專業發展，也間接影響了後進晚輩的前途，心裡很安慰，比發表了國際期刊論文更覺得開心。

這篇序是我從洛杉磯返臺時在航機上所落筆，漫漫長路與機上的昏暗燈光，尤其容易令人陷入深刻的沉思。從南加大畢業回臺轉眼已經十數二十年，對於心理計量始終興趣不減的我，

也真能體會所謂十年法則的真義（在一個領域要持續努力十年以上才能獲得創新突破）。這次會飛到洛杉磯開會，目的也正是討論如何將測驗評量與統計技術直接串聯，並導入雲端部署的有關議題，跟我合作的 Alex 從史丹福大學取得博士學位後並沒有留在學術界，卻隻身投入管理顧問界長達二十年，卻也過得充實自在。我問他擔不擔心老來會沒有保障，他笑著跟我說，如果到了實務領域看到自己研究的東西能夠真的對別人有幫助，我可能會捨不得回到校園，那並不是金錢或地位的問題，而是一種自我實現。我多少同意他的話，更可以明瞭他眼中流露的另一種滿足，與我剛剛描述的那種安慰心境相差不遠。

我相信這本書會有一些年輕人正在閱讀，在此且容我倚老賣老的說一句，在臺灣的你們真的是活在一個幸福的世代，擁有俯拾即是的豐富知識、高端科技、自由的心靈與富裕進步的社會。我與 Alex 在帕薩迪納（Pasadena）街頭漫步時談到東方國家的發展時，下了一個結論，未來的舞臺不會再是西方社會獨領風騷，他希望我能帶給他更多關於亞洲市場的訊息，多瞭解臺灣在學術與管理實務上的進展。當時，我踏在一樣天空長藍的帕薩迪納舊地，會議中談的雖然是測驗統計與雲端科技，但我的心思浮上在另一個雲端，體會更是深長。

邱皓政謹誌  
於北海道的雲端

2010/4/8

# 作者序

作為一本介紹測量概念和測量方法的入門教材，本書的第一版得到了廣泛的使用。我確信，其成功之處就在於它使複雜的概念變得通俗易懂，這也是我的目的所在。此書出版的一個極其重要的出發點，就在於為了幫助各個水平的學生從概念上（**Conceptually**）來理解測量問題。在教堂山（**Chapel Hill**）的北卡羅萊納州大學的公共健康學院，我給本科生開設的量表編製課程，吸引了許多不同背景的學生。在同一學期內，我的學生裡既有只學過一門本科統計課程的，也有攻讀量化心理學博士的。教授該課程的經驗發現，不同水平的學生都可以從這種以清晰、概念性和非數學化的形式所呈現的材料中獲得益處。儘管公式在此類課程中還是有必要，但我盡力以清晰明瞭的方式來解釋這些公式與概念，它們只不過是合理地簡化了運用於數據中的一系列操作。我盡可能在第一版中介紹一些已獲得顯著效果的教學方式，在此一修訂版中，我也做了此類嘗試。本書的重點在於，以清晰明瞭方式來傳遞基本原理背後的訊息，並使讀者能夠窺探看起來像是「黑箱」的各種方法。

此修訂版做了大量的修改。在修訂版中，我保留了學生們認為最清晰有用的部分，增加了自第一版問世以來更受重視的主題。每一章都有所修改，有幾章的內容原來就已經很豐富。本修訂版增加了三十多則參考文獻，也保留了許多經典著作，它們在此版中再次被引用。有幾章增加了圖表以使關鍵點更為

直觀。在第一章中，我新增了一些例子，闡明為什麼一些變數需要用很多題項來進行評估，而其他變數卻不需要如此大費周章，並且對不同的題項組合類型進行了更廣泛的討論。第二、三章的內容經過修改後已變得更加清楚。為了做這些調整，我在第四章增加了關於表面效度的討論；在第五章列出了量表編製的指導方針，並增加了幾個學生有用的實務技巧；第八章則以一個更廣闊的角度來看待測量，並且有所拓展，包括在何處尋找測量工具、高品質的程序如何作為量表編製的基礎，以及與不同題項功能相關的一些問題。剩下的兩章與前一版相比改動最大。為了使因素分析過程更加生動、更可理解，第六章因素分析在論述的內容範圍上有了相當大的擴充，並完全重寫。我運用了大量圖表來補充文字材料。最後，新增加的第七章介紹了一個在第一版中只簡要提及的主題—項目反應理論（**item response theory; IRT**）。我的目的並不在於教導讀者關於 IRT 的複雜操作性知識和學者正在進行研究的部分，而是協助他們提供一個概念基礎，以幫助他們理解在別處會碰到的更艱澀的材料。

儘管增加了第七章的內容，本書的重點仍然是古典測量方法。毫無疑問的，隨著分析所需要的數據處理以及電腦運算程序的運用，像 IRT 這樣的理論一定會備受歡迎，但古典的方法也不會消失。雖然存在某些理論上的缺陷，那些方法在多種情境中運作得出奇地好。它們的基礎和運用都很容易理解。在修訂版的不同部分，我強調了一些我認為 IRT 優於古典方法的幾個重要面向。但是，現存的大量研究證明古典方法仍運行得很好。當 IRT 處於優勢時，古典理論並不會隨之淪為陳腔濫調。

兩者將由於它們擁有各自的優缺點而相互補充。許多應用研究者並不會真的需要用到古典測量以外的技術。因此，會讓我擔心的，倒不是那些能夠掌握最新測量方法進展或還沒有能掌握最新發展的這些專家之間的差距，而是那些已經有和還沒有擁有任何測量概念或方法學的人在這個領域內的落差，我希望本書的出版能夠幫助讀者縮小這一差距。

**Robert F. DeVellis**

# 目次

審定者序 i

作者序 iii

## 1

概論 1

測量概述 5

測量在社會科學中的歷史淵源 7

測量的後期發展 10

測量在社會科學中的作用 12

總結與展望 22

## 2

瞭解潛在變數 25

構念與測量 27

潛在變數：題目數值的假設導因 29

路徑圖 30

測量模型的進一步闡述 36

平行測試 37

其他替代模型 42

### 3 信度 47

連續與二分題目 49

內部一致性 50

以量表間相關為基礎的信度 66

概化理論 74

結語 77

### 4 效度 79

內容效度 82

效標關聯效度 83

構念效度 87

何謂表面效度 93

### 5 量表編製程序 97

步驟 1：清楚地決定你要測量什麼 99

步驟 2：建立題庫 104

步驟 3：決定測量格式 116

步驟 4：延請專家評估初編題庫 140

步驟 5：效度驗證題的納入 142

步驟 6：預試樣本施測 144

步驟 7：評估題目品質 147



步驟 8：量表長度合理化 157

## 6 因素分析 165

因素分析概述 169

因素分析的概念說明 176

因素的解釋 200

主成分與共同因素 202

驗證性因素分析 207

量表編製中因素分析的使用 210

樣本規模 215

結論 216

## 7 項目反應理論概述 217

項目難度 221

項目鑑別度 224

偽陽性 225

項目特徵曲線 227

IRT 的複雜性 231

何時使用 IRT 233

結論 238

**8 廣義研究中的測量 241**

量表發展之前 243

量表施測之後 250

最後的思考 253

**索引 255**

**參考文獻 263**

# 第一章

## 概論

測量概述

測量在社會科學中的歷史淵源

測量的後期發展

測量在社會科學中的作用

總結與展望





在社會科學研究的眾多領域中，「測量」（measurement）是一個不可或缺的基本議題。讓我們看看下面的幾個例子：

1. 某位健康心理學家面臨一個困境：他所需要的測量工具並不存在，如果他的研究必須能夠掌握一個能夠區分那些去看病的病人們所需要（**want**）的東西和所期待（**expect**）的東西兩者之間差異的測量分數，但他從過去的研究文獻中可能沒有辦法找到能夠區辨這兩者之間差異的理論觀點，能精確區分兩者差異的測量方法更是付之闕如。儘管他可以自己編寫一些題目來測量，但是他所「編製」的題目可能沒有信度，也可能不是測量那些東西的有效指標。
2. 某位流行病學家正在構思如何進行他的工作，他從一個全國性的大型健康調查研究中獲得一批資料來進行次級資料分析（**secondary analysis**）。他想探討人們所知覺到的某些心理壓力和健康狀況之間的關係。儘管在原來的調查中並沒有納入與壓力測量直接有關的題目，但問卷中有幾個本來不是用來測量壓力的題目似乎可以跟壓力扯上關係。那麼，把這些題目組織成一個有信效度的心理壓力得分雖然有可能，但是如果這些

不太理想的題目所組成的分數無法反映一個人的壓力，那麼研究者可能會得到錯誤的研究結論。

3. 某行銷團隊在策劃一個關於高價嬰兒玩具的促銷活動時遇到瓶頸。他們從焦點團體（**focus groups**）的討論中發現，父母的消費決策強烈受到這種玩具是否對兒童具有明顯的教育意義所影響，行銷團隊於是認為那些對他們小孩有著高度教育與事業成就期望心理的父母最易受到這新出產的玩具所吸引。因而，行銷團隊必須能夠從事一項大型跨地理區域的抽樣研究來測量父母的期望心理，因為要利用更多的焦點團體來獲得足夠數目的消費者來進行研究並不切實際。

在上面的每一個情境中，不同領域的工作者在研究之初都不約而同的遇到了測量問題。其實他們對測量本身並不感興趣，但如果要順利完成研究目標，他們每一個人都必須找到一個能夠針對特定現象進行量化的方法。但是在每一個狀況中，「現成的」測量工具要不是沒有就是不合適。所有的研究者都瞭解，如果他們採用隨隨便便的測量方法，極可能只會獲得一些不精確的資料。自行發展自己適用的測量工具，是他們唯一的選擇。

許多社會科學研究者都面臨相同的難題。他們對這類問題的普遍反應仍然是去尋求既存的測量工具，要不然就



是自己新編一些「看起來」不錯的題目來進行測量工作。如果測量的結果很糟，他們的共同藉口是歸咎於對編製一套可靠有效的測量工具有其困難或者是不熟悉，或是很難獲得關於研究主題的有用資訊。對一般研究者來說，要試圖去學習量表編製的技巧，努力到最後的結果可能只是蒐集了一堆原本是測量專家所使用的一些資源與材料，要不然就是找到一堆很通俗但卻不適用的東西。本書的目的就是為這些工作提供一個解決方案。

## 測量概述

測量是一個基本的科學活動。我們藉由觀察人、物體、事件和過程而獲得某些知識。然而要從這些觀察結果中得到有意義的資訊則需要我們對其進行量化，亦即對於我們所關心的科學議題進行測量。測量的過程與科學問題的解決兩者之間交相作用，其間的界線往往無法察覺。這個現象尤其容易發生在一個新問題被提出，必須從測量的角度來釐清時，或者是新概念的產生在邏輯推理過程上必須給予一套量化的定義時。例如，Smith、Earp與DeVellis（1995）調查了婦女對受虐（battering）的感受，他們首先透過文獻與理論分析建立了一個前導的概念模型，整理出受虐知覺具有六種不同成分。然後在實徵研究的部分，他們發展一套量表來測量這些感受，研究結果得到了一個相對單純的模型，不論是對那些受虐和未受虐的婦女，

40個題目中的37個題目所測量的是同一種概念。此一結果顯示，一開始研究者所認定的複雜概念模型，在那些社區中的婦女所感受到的卻只不過是一個單一、一般性的現象。換言之，在執行這項探討婦女對於受虐知覺的研究過程中，研究者最後得到一個不同於最初預期的解釋架構與測量工具。

Duncan (1984) 認為，測量的根基是社會歷程 (social processes)，這些歷程以及它們的測量實際上都先於科學：「所有的測量……都是社會測量。物理測量也有其社會目的」(p.35)。Duncan 注意到，最早的社會測量程序，例如投票、人口普查以及工作改善體系等等，「最初似乎是為了滿足大眾的需要，而不僅僅是為了合乎科學好奇心而進行的實驗。」(p.106) 他進一步指出，同樣的程序「可以從物理學歷史中看到：古代的人在解決實際的日常社會生活問題的過程中，發展出對長度或距離、面積、數量、重量和時間的測量技術，物理科學就是建立在這些成就的基礎之上。」(p.106)

無論最初的動機是什麼，科學的每一個領域都有自己的一套測量程序。例如，物理學發展了特定的方法和器材來研究次原子微粒。在社會行為科學領域，心理計量學 (psychometrics) 則是涉及心理和社會現象測量方法與技術的一門專業領域，測量程序基本上多使用問卷，所使用的測量變數則是廣泛理論框架中的某一部分。





# 測量在社會科學中的歷史淵源

## 早期例子

常識和歷史文獻支持了Duncan的下列看法：社會的需要使得測量在科學出現以前就有所發展。毫無疑問，某些測量形式是先民所具有的技能中的一部分。最早的人們必須對物體、財產以及敵人做出評估，例如根據對手的某些特點（如體格）來對其做出判斷。Duncan（1984）引用聖經上的文字以說明其對測量的關注（例如：A false balance is an abomination to the Lord, but a just weight is a delight. 即一個虛假的天平是對上帝的蔑視，而一個公平的砝碼是一種快樂），並指出亞里斯多德的作品中出現了負責檢查重量和測量的官員。Anastasi（1968）指出，古希臘時所使用的蘇格拉底方法在某種程度上可以被看作是知識測驗，它涉及以一種什麼樣的方式來理解事物。迪布瓦博士（P. H. DuBois）在他1964年的論文中，描述了中國早在西元前2200年就有文官測驗。Wright（1999）引述了古代一些能夠精確測量的其他的一些重要例子，包括7世紀建立在「七權數」（weight of seven）基礎上的穆斯林稅制。他還指出法國革命在某種程度上，是由於農民已經受夠了不公正的測量制度而爆發。