

淺談信度與效度

姚開屏

職能治療是一門專業科學，它所涉及的領域不僅止於醫學，更涉及了社會及行為科學。一個科學工作者在從事研究時的一個重要方法，乃是用客觀、精確及數量化的方法，對其研究對象作描述⁽¹⁾。若科學工作者僅止於對研究對象作粗淺的觀察和分類描述，則很難對研究的對象作精確的推算或預測。在社會及行為科學中，研究人員常使用測量工具，以搜集研究對象的量化資料⁽²⁾。用以測定人或物之特質，此時量化資料的可靠度（信度，reliability）及正確度（效度，validity）則成為優良施測工具必備的條件。倘若研究者對施測工具測得結果之信度及效度沒有把握，則就難據此以分析問題，並從中獲得研究之結論，這門科學也將難有突破性的發展。由此可知，施測工具信度與效度的研究，是科學研究的初步，而信度與效益的增進，更是研究者的目標。

傳統的職能治療，並非以研究導向為目標的專業（research-oriented profession）。它過往乃從經驗的累積及環境的變化中成長。然而隨著時代的變遷，這門被規劃在醫療健康體系中的學門，也漸必需朝著更符合現代科學研究精神方向發展，也就是說——職能治療的成果必須精確可信；職能治療的方式必須合理有效。倘若不符合這些重點，則職能治療仍被視為不夠專業、不夠科學，甚至不為大眾所認同、接受及尊重。美國及世界各國的職能治療界，近些年也重視這個問題。為了讓大多數非職能治療界的人員了解職能治療價值的信度及效度，美國職能治療學會也大力的支持職能治療從業人員，在施測工具方面標準化的發展^(3,4)。因為唯有標準化、可信及正確之施測方式及工具，才能給予人足夠的信心，才能促進專業進一步的發展。一些目前被廣泛應用於職能治療或物理治療界的施

測工具，例如 Bayley Scales of Infant Development (BSID)⁽⁵⁾、Gesell Developmental Scales⁽⁶⁾、Southern California Sensory Intergration Tests (SCSIT)⁽⁷⁾，皆提出了施測工具信度或效度的研究結果。參考文獻 8—17 也提供了復健人員在施測工具上信、效度的研究方法及過程。

信度的意義及類型^(1,2,18-23)

信度（reliability）係指一測驗結果之一致性（consistency）、可靠性（trustworthiness）或穩定性（Stability）。對同一施測對象，不同的人使用同一施測工具，或同一個人在不同的時間使用同一工具，其所得的結果可能會不一致。若此差別太大，則工具的可信度就不好。通常用相關係數（correlation coefficient）來表示兩個不同施測者間施測分數一致性。同一個人前後施測分數之一致性，或同一測量內容前後一致性之高低。相關係數愈高者，表示一致性愈高，其工具之可信度也就愈高。通常測定信度有三種類型：

- (1) 施測者間信度（interrater reliability）：
二個或二個以上的施測者，在同一時間對同一施測對象評分標準異同之關係。一般說來，一個標準化的測驗工具，應有完好的施測評分標準及步驟。若每一個施測者皆遵循此標準，則應該有良好的施測者間信度。
- (2) 再測信度（test-retest reliability）：同一個測驗讓受測者前後二次不同時間受測，所得前後二次測驗分數一致之關係。這個信度關係反應

出測驗分數的穩定度。此信度有可能受到練習、記憶或身心成熟之影響，因此二次施測時間之間隔必須適當。適宜之時間間隔並沒有一致的規定，端視測驗之性質、施測對象之特色而定。例如：給予尚在變化中的中風病人施測間隔宜短，以減少病人因隨時間而成熟變化，然而間隔時間也不能短至讓病人有記憶練習施測內容的機會。對於長期慢性的精神病人，則施測間隔可較久。

- (3) 折半信度 (Split-half reliability) : 又稱為內部一致性 (internal consistency) 。有的測驗工具受限於只能使用施測一次，因此將測驗題目分為前後二半或單、雙號二半，求此二半間之相關係數。這方法只需施測一次即得相關係數。

效度的意義及類型

效度 (validity) 係指測驗工具能測出其所欲測量特質的程度。每一個測量工具有其一定的使用範圍，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，則就「無效」。效度愈高，表示愈能測出所要的施測對象特質。因此自行設計施測工具或使用標準化的工具時，「效度」是最重要的條件。若一個測量工具不能測出所想要的特質，則有再好的信度、再優良的常模也都沒有用。通常測定效度有四種類型：

- (1) 內容效度 (content validity) : 乃是指測驗內容適當的程度，包括對有興趣研究之特質，其測驗內容是否有足夠的涵蓋及適當的比例分配。例如：欲施測中風病人的知覺 (perception) 能力，是否所使用的施測工具能適當的反映出一個人的知覺能力，是否工具已適度的涵蓋了知覺的不同層面。
- (2) 同時效度 (concurrent validity) : 已知一施測工具能有效的測出施測對象的某一特質，現有另一新的工具亦測同一特質，則此一新的工具與原有工具測量該一特質間之相關關係，稱為「同時效度」。例如：已知測驗 A 能有效的測出一歲以內嬰兒精細動作的發展，現發展一新的施測工具 B，亦同樣的用來測一歲以內嬰兒精細動作的發展，則施測者可同時將二測驗給予同一嬰兒，而

得二測驗分數間之相關度。

- (3) 預測效度 (predictive validity) : 指測驗分數與未來有關方面表現間之關係。例如：中風病人入院時日常生活自理能力的好壞與出院後適應生活能力間之關係。另外，參考文獻 24 研究兩個常用以測量嬰幼兒動作發展評量表 (Bayley Motor Scale and Peabody Developmental Motor Scales) 間之同時效度及預測效度。而參考文獻 25 則試圖以幼兒 2 歲時在 Bayley Scales of Infant Development 上的得分，以預測其四歲半時在動作及認知方面的能力間之關係。
- (4) 建構效度 (construct validity) : 乃指測驗能測量理論的概念或結構之程度。這是一種抽象的觀念，困難且絕非單一之研究能建立的完全，而是必須累積許多的研究結果，才得以更臻健全。建構效度的考驗須由某一建構的理論出發，例如：欲測「智力」這個建構，依據一些理論所發展出來之假設來設計智力測驗，再就實施測驗後資料的分析，以看出是否符合原有之假設狀態，此研究之結果即為此一測驗之建構效度。參考文獻 26 - 28 提供研究的方法。

建議

作者對未來有興趣從事信、效度研究的職能治療人員，有下列數點的建議：

1. 就自行設計或修正國外的評量表作信、效度的分析，使這些施測工具更具說服力。當信、效度建立後，接下來的步驟就是常模 (norm) 的建立。
2. 修改現有不切實際的評量表。治療人員可研究如何挑選最具代表性及涵蓋性的項目成為新的施測工具，以便研究之進行。另外，「量化」目前現有的施測工具，亦是從事研究的重要過程之一。而施測的過程及評分的標準必須有一定的依據，最好能清楚的記錄下來。
3. 當一施測工具信、效度研究完成後，最好用於其他醫療機構，以不同的治療人員對不同的病人施測，求取此工具之普遍化 (generalization) 或稱外部效度 (external validity) 。

參考文獻

1. 劉英茂：普通心理學（下冊），大洋出版社，1978. pp 311-320。
2. 簡茂發：信度與效度，楊國樞等編：社會及行為科學研究法（上冊）。東華書局，1987. pp 321-339。
3. The Foundation: Research competencies for clinicians and educations. *Am J Occup Ther* 37: 44-46, 1983.
4. Jaffe E: Nationally speaking: Transition in health care--Critical planning for the 1990s: Part one. *Am J Occup Ther* 39:431-435, 1985.
5. Bayley N: Manual of Bayley Scales of Infant Development, New York: The Psychological Corp., 1969.
6. Knoblock H, Passaminick B: Gesell's manual of developmental diagnosis, revised, New York: Harper & Row, 1980.
7. Ayres AJ: Southern California Sensory Integration Tests: manual, Los Angeles: Western Psychological Services, 1972.
8. Goodkin R, Diller L: Reliability among physical therapists in diagnosis and treatment of gait deviations in hemiplegics. *Perceptual and Motor Skills* 37:727-734, 1973.
9. Mitchell S: Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychol Bull* 86:376-390, 1979.
10. Siegner C, Crowe T, Deitz J: Inter-reliability of the Southern California Postrotary Nystagmus Test. *Phys Occup Ther Pediatr* 2: 83-91, 1982.
11. Pandya S, Florence J, King W, Robison J, Oxman M, Province M: Reliability of goniometric measurements in patients with Duchenne muscular dystrophy. *Phys Ther* 65:1339-1342, 1985.
12. Krebs D, Edelstein J, Fishman S: Reliability of observational kinematic gait analysis. *Phys Ther* 65:1027-1033, 1985.
13. Gajdosik R, Simpson R, Smith R, Don Tigny R: Pelvic tilt -- Inter-rater reliability of measuring position and range of motion. *Phys Ther* 65:169-174, 1985.
14. Ottenbacher K, Dauck B, Grahn V, Gevelinger M, Hassett C: Reliability of the Behavioral Assessment Scale of Oral Function in Feeding. *Am J Occup Ther* 39:436-440, 1985.
15. Geveinger M, Ottenbacher KJ, Tiffany T: The reliability of the motor development checklist. *Am J Occup Ther* 42:81-86, 1988.
16. Hamrin E, Wohlin A: I. Evaluation of the functional capacity of stroke patients through an activity index. *Scand Rehab Med* 14:93-100, 1982.
17. Bhavnani G, Cockburn J, Whiting S, Lincoln N: The reliability of the Rivermead Perceptual Assessment. *Br J Occup Ther* 46:17-19, 1983.
18. American Educational Research Association, American Psychological Association, National Council on Measurement in Education: Standards for educational and psychological testing. Washington, DC: American Psychological Association, 1985.
19. Brown FG: Guidelines for test use: A Commentary of the standard for educational and psychological tests.

- National Council on Measurement in Education, 1980. pp. 42-70.
20. Aiken LR: Psychological testing and assessment, 6th ed. Boston: Allyn and Bacon, 1988. pp. 95-119.
 21. Anastasi A: Psychological testing, 6th ed. New York: Macmillan, 1988. pp. 109-138.
 22. Hartmann DP: Using observers to study behavior. In Fiske DW (ed): New directions for methodology of social and behavioral science series. CA: Jossey-Bass, 1982. pp. 51-63.
 23. Monette DR, Sullivan TJ, Dejong CR: Applied social research: Tool for the human services. New York: CBS College, 1986. pp 96-103.
 24. Palisano RJ: Concurrent and predictive validities of the Bayley Motor Scale and the Peabody Developmental Motor Scales. Phys Ther 66:1714-1719.
 25. 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 41: 374-378, 1987.
 26. Van Deusen J, Harlowe D: Construct validation of occupational therapy measure used in CVA evaluation: A beginning. Am J Occup Ther 38:101-106, 1984.
 27. Van Deusen J, Harlowe D: Continued construct validation of the St. Marys CVA evaluation: Brunstrom arm and hand stage ratings. Am J Occup Ther 40:561-563, 1986.
 28. Van Deusen J, Harlowe D: Continued construct validation of the St. Marys CVA evaluation: Bilateral awareness scale. Am J occup Ther 41:242-245, 1987.



*Kai-Ping Yao, MS, OTR
Graduate Student,
Department of Psychology
University of Illinois
at Urbana-Champaign
Champaign, Illinois,
U.S.A.*