研究效度

/ 周文欽

當我們評鑑一個研究的良窳時，常從兩個層面思索，第一是研究者使用的研究設計（方法）能否妥適的解釋其研究結果，其次為研究結果是否具有推論性，這兩個層面互相關聯。事實上，這兩者皆屬於研究效度的範疇。所謂研究效度（Validity of research），是指研究結果符合真實的程度，或是研究結果之正確性的程度；既然效度即是程度，所以只能說高研究效度或低研究效度，而不能說有效度或沒效度。效度在測驗中有另外的涵義，這種涵義請參閱教科書第8章第2節。庫克和坎貝爾（Cook & Campbell, 1976）將實驗法的效度分成4種：內在效度、外在效度、建構效度與統計結論效度。事實上，這4種效度也適用於一般的研究法。馬克伯尼（McBurney, 1998）即採取這一觀點，直接稱前述4種效度為研究效度。惟，建構效度和統計結論效度與內在效度和外在效度之間有很密切的關係，所以大部分的研究方法學者論及研究效度時，常只分成內在效度和外在效度兩種。甚至有學者認為，要區分內在效度和統計結論效度，或外在效度和建構效度有其困難之處（許擇基，民69）。本課程是討論研究方法的專書，所以將前述4種研究效度皆加以介紹和說明。

一、內在效度

(一)涵義

所謂內在效度（internal validity），是指一個研究之研究設計（research design）正確的說明研究結果，或呈現自變項與依變項之因果關係的程度（Judd, Smith, & Kidder, 1991）。就因研究設計包括研究對象、研究工具、實施程序和資料處理等4個層面（詳教科書第1章第3節），所以從研究對象的選取，研究工具（如測驗或問卷）的編製和使用，各種變項的安排和所蒐集資料的統計分析，都可評估該研究之內在效度的高低。

例如，有個研究者想瞭解各國民中、小學擁有電腦教室的情形如何；於是，該研究者在台北市展開問卷調查，研究結果顯示100％的國民中、小學都擁有電腦教室。試問這個研究結果，能正確的說明各國民中、小學擁有電腦教室的實際情形嗎？當然不能，主要是因該研究的研究對象涵蓋面不夠，致使其內在效度太低，所以研究結果就很難說明事物的真相。再如，你用實驗法來探討人們對藍色和綠色兩種燈光的反應時間是否有差異？現在你用成人測量藍色燈光的反應時間，用小孩測量綠色燈光的反應時間，實驗結果顯示二種燈光的反應時間有顯著差異。在這個實驗裡，燈光顏色是自變項，反應時間是依變項，則我們很難說這兩者存有很密切的因果關係。這主要是因依變項（反應時間）除了受到自變項（燈光顏色）的影響外，尚且受到受試者之年齡的干擾。就因年齡會影響實驗結果，而你未加以控制，所以這也是一個低內在效度的實驗研究。

(二)影響因素

簡要來講，凡是屬於研究設計的層面或範疇都是影響內在效應的因素。具體言之，影響內在效度的因素有歷史事件、受試者的成熟、研究工具、統計迴歸現象、受試者的選取、受試者的消失與霍桑效應等多種。

1.歷史事件

所謂歷史事件，是指在研究歷程中，會影響研究結果之一切與研究設計無關的經驗、活動或事件。例如，某研究者想瞭解某新式教學法對國中學生數學成績的影響，他找了某一班級的學生實施該教學法1學期，然後分析教學後的成績是否高於教學前的成績。但在實施新式教學法的期間，該班級的學生大部分都去參加校外的數學輔習。此時，研究結果發現，參加新式教學法的學生之數學成績確實進步了，然而卻不能下結論說，該教學法真的有效；原因很明確，學生參加補習這一歷史事件，降低了該研究的內在效應。再比方說，某醫藥研究者聲稱，只要連續服用新藥1個月，高血壓者的血壓都會恢復正常，然在實驗期間，所有受試者正巧都在民國88年的「921大地震」中受到重創。1個月後，受試者的血壓平均值全部都上升而沒下降，這時我們能說該新藥不具降血壓的療效嗎？很明顯的，「921大地震」這一歷史事件已經干擾了研究結果的正確性。

2.受試者的成熟

有時候研究結果並非研究設計或實驗處理所造成，而是肇因於受試者

自然成長或發展所產生的效應。例如，某藥廠聲稱吃了該廠的藥可幫助人長高，某國小學生集體吃了該藥3年，真的發現身高突飛猛進。我們能說該藥有助長身高的功效嗎？當然不能，對小學生而言，就算不吃該藥，在成熟的因素下，身高也會自然增高。成熟因素不只會影響生理功能，也會對心理功能產生效應。例如，在某個年齡之下，年齡愈大者，其智力會愈高，使用語言也會愈趨複雜，情緒也會愈穩定。準此，在挑選正發育或成長中之受試者參與研究時，就須特別控制有關變項，以提高內在效度。

3.研究工具

研究工具本身會使研究結果產生偏差的情況有2種。第一種是工具的使用對受試者造成了心理效應，致影響研究結果，心理效應的層面包括學習成效與情緒變化等。例如，在數學成就測驗前、後測的研究裡，由於已施測過一次的學習成效，後測成績常會高於前測。再如，筆者就讀大學時，曾參加一項有關壓力之下生理反應的實驗，該實驗是用心電圖（electrocardiogram, EKG）和皮膚電（Galranic Skin Response, GSR）來測量生理反應這個依變項。當研究者（筆者的心理學老師）將測量儀器安置於身上時，情緒即刻受到影響（產生莫名的恐懼）；以致於在還沒操弄壓力情境（自變項）前，生理反應就產生極大的變化，此種生理反應能歸因於壓力嗎？

第二種是因工具的特質，工具發生變化或研究者使用工具不當，所造成的研究結果偏差。現今有許多人不太相信民意調查的結果，其所持理由常是懷疑問卷的題目或題目是被委託調查機構「設計」過的。再如，前述測量EKG和GSR的儀器本身就會產生極大的誤差，或研究者操作不熟或不當，在在都會影響實驗結果，進而很難令人相信壓力與生理反應之間的因果關係。

研究工具的種類相當廣泛且複雜，舉凡可以拿來測量研究變項的都屬之，它可以是各種精密的科學儀器，甚至於隨手可得的紙、筆、電腦鍵盤或積木等，都可當作研究工具，在社會及行為科學的研究領域裡，最常使用的研究工具當推問卷與測驗。就因研究工具會影響研究的內在效應，所以研究者在選擇、設計、製作或使用它時，豈能不慎乎！

4.統計迴歸現象

在討論統計迴歸現象之前，容先說明迴歸效應。所謂迴歸效應（regression effect），是指兩極端的分數會有退回到平均數，或往平均數方向迴歸的現象（林清山，民81，頁148）。依據張春興（民78）的說法，統計迴歸

則是指：「在根據X變項預測Y變項時，若其相關係數不夠高，而且X變項的分散情形較大，極端分數較多，則Y變項的預測值有偏向Y變項平均值，使分散情形變小的現象（頁625）。」

依前文推知，極端好的人不可能一直好下去，他的各種特質只會往下走，很難再往上突破；極端差的人不可能再一直差下去，他的各種特質只會往上升，很難再往下探底（因已無底可探了），這就是所謂的統計迴歸現象。此種現象常會出現在教育情境中，例如，第1次月考數學考99或100分的學生，第2次月考的成績大都不會進步了；同理，第1次月考英文考1、2分的學生，第2次月考大都會進步。所謂的否極泰來或物極必反，即有類似的涵義。

職是之故，在抽取研究對象時，一定不能找特別好或特別差兩個極端的受試者，否則其內在效度必定低。例如，我們找來資賦優異學生參加某教學實驗計畫，結果發現學生的學業成就沒有進步，這時，可以說該教學實驗計畫沒有效嗎？現在如找來智能不足學生當受試者，結果發現他們的學業成就進步了，此時能說該計畫成功了嗎？上述兩個答案當然都是否定的，因為在統計迴歸現象下，資賦優異者很難再進步，智能不足者則只會進步；這兩個研究都是低內在效度，所以研究結果並不具代表性。為避免統計迴歸現象的干擾，當研究對象少時可用簡單隨機抽樣，研究對象多時則可用分層隨機抽樣或系統隨機抽樣來測試。

5.受試者的選取

進行研究所選取的各組受試者，本來就存在著差異，如研究結果顯示，在某個依變項上確實有顯著差異，我們可以說這是研究處理或實驗操弄造成的嗎？例如，以資賦優異者為實驗組並以智能不足者為控制組（實驗組或控制組的涵義詳參教科書第6章），進行提升IQ實驗計畫，實驗結果顯示，實驗組的平均IQ顯著的高於控制組；研究者可宣稱，他發現了提升IQ的方法嗎？為免受試者之選取降低研究的內在效度，應採取隨機抽樣和隨機分派（random assignment）的方法安排受試者，使各組受試者的各項特質都一樣或趨近於一樣。從抽樣的觀點來看，統計迴歸現象事實上是選取不當受試者造成的。

6.受試者的消失

就算受試者的選取沒有偏差，而且也安排各組受試者在各方面的特質都是相同的；但假如不是全部受試者都完成每一個研究活動，那麼你的研究結

果仍然是不可靠的。因為中途離開研究活動的受試者，很可能與全程參與者不同，所以受試者的消失就會影響到研究效度。假如比較特殊之受試者正好中途離開，那麼研究誤差的產生就無可避免了。

假設你現在研究兩組行為改變技術對體重控制的成效影響。第一組是遵照規定的食譜進食，而且記錄每次所吃的東西，秤全部食物的重量，並估計它們的所有卡路里（calory）；第二組則只是照著規定的食譜進食即可。我們可以想像得到，要求比較多那一組的受試者，他們中途離開研究活動者也會比較多。研究結束後，我們會發現，擁有高度動機受試者比率較高的那一組，比較可能會產生減肥的效果。於是，我們就可能會下一個錯誤的結論，聲稱第一組的方法比第二組有效。事實上，產生此種結果的原因，很可能是「無效的受試者」都離開了，剩下的則都是「有效的受試者」所造成的。因為受試者的消失而影響研究之內在效度者，最容易發生在縱貫研究裡。

7.霍桑效應

霍桑（Hawthorne）是位於美國芝加哥附近的一個地名，為美國西方電器公司的所在地。1927年該公司的一項研究發現，影響生產績效的主因是同仁間的人際關係，而非工作情境裡的物質條件。後來，管理學者和心理學者便將心理因素影響工作效率的現象，稱之為「霍桑效應（Hawthorne effect）」（張春興，民78）。在研究方法裡的霍桑效應，則是泛指受試者之心理因素干擾或影響研究結果的現象。例如，對中小企業管理階層的評鑑研究，只要是中央主管機關評鑑日，那些受評的管理幹部就會表現出特別的積極與敬業，其他時日則常會原形畢露，稍嫌散漫，甚至溜班觀看「大盤」走勢及「個股」漲跌。再者，有些受試者只要知道自己是「小白鼠」的話，則參與實驗或研究就常會漫不經心，甚至會抱持抗拒的心態。因此，為了降低霍桑效應，有些運用實驗法的研究者（特別是醫藥方面的實驗）會隱瞞研究的真相，有些運用觀察法的研究者則會採行非參與者觀察的方法進行觀察。

二、外在效度

(一)涵義

所謂外在效度（external validity），是指一個研究之研究結果能普遍推論到母群或其他相類似情境的程度。研究的終極目的常是在應用層面上，所以愈能根據研究結果來加以應用的研究，愈是有價值的研究，其外在效應也就愈

高。以已開發國家為研究對象所提出之經濟發展模式，常只適用於已開發國家，而不能推論到未開發（或開發中）國家；利用大學生為樣本所做的研究結果，常只適用於大學生，而不能推論到中、小學生；30年前所做有關民間信仰的研究結果，也常很難中肯的說明當下的民間信仰概況；日本在1995年針對阪神大地震所提出之各種復建計畫和措施，也無法完全移植到民國88年台灣的「921大地震」上。像前述的各種研究（或計畫），其外在效度常是偏低的，亦即前述的研究結果（或措施）不太容易應用到其他的地區、人們和年代上。當然，我們不能說，以特定樣本、地區或時代所做的研究結果，就不具有任何應用價值，事實上，或多或少還是有其使用功能；這也是本文緒言所提及的：只能說高研究效度或低研究效度，而不能說研究有效度或沒效度。

(二)影響因素

誠如前文所述，會影響研究結果之推論或研究之外在效度的主要因素有研究樣本、研究時代及研究地區與情境。

1.研究樣本

各種不同的物種或樣本都有其特性與差異性，所以用動物所做的研究結果，並不能完全推論到人類身上（因此醫藥的動物實驗有了成果後，還是要再繼續進行人體試驗）；用大學生做研究所得之有關行為現象的結論，也難一體適用於其他不同群體的學生身上。簡言之，研究樣本會影響到外在效應，主要是研究樣本代表性的問題。雖然，利用抽樣原則來選取研究樣本，能增進樣本的代表性，可是推論到其他母群時仍有其盲點存在。例如，完全用隨機抽樣抽取台北市成人進行研究，其結果能完全推論到南投縣成人身上嗎？當然不能，因為台北市成人與南投縣成人這兩個母群是截然不同的。

準此言之，周文欽（民80）以台北市高中生為研究對象進行越區就讀學生之研究，所得越區就讀的原因之結論，就不能用來說明台北市國中學生越區就讀的現象，更無法去解釋大學生為何要捨近求遠選擇離鄉700里（甚至飄洋過海出國唸大學）的大學就讀。由此觀點論之，任何研究要維持高的外在效度是一個不太容易達成的目標；所以研究者在下結論時，就要特別謹慎小心，不宜做過度的推論。

2.研究時代

不同的時代有不同的特色和背景，再者，隨著時間的流動，許多的產物、思維、發明和發現也不間斷的問世或出現，以致於不同時代的研究結果

或結論，很難「歷久彌新」或「一路走來，始終如一」。其中的顯例是考古學方面的研究，常會因新文物、新證據的出土與發現，而推翻了先前的結論。例如，人類的起源年代就不斷的被更新與修正，許多歷史事件的原因也常被改寫。同理，30年前的經濟發展理論、國富調查或流行性病學研究報告，大都不適用當今的年代。自然科學也有相同的情況，大者如「古典物理」（以牛頓的理論為基礎）與「近代物理」（以量子物理為中心概念）之分，小者如收音機的最佳元件隨著時代的不同，由真空管到電晶體，再由電晶體至現在的積體電路，其進步與精緻何止以千里計。

就因研究時代深深的影響到研究的外在效度，所以在研讀研究報告或應用研究成果時，務必要考慮到研究報告所提出的時間問題，也就是要在時間的架構下來思考研究結果。

3.研究地區與情境

不同的地區或地方常有其特殊的文化背景、意識形態、生活方式和物質條件，所以在某地區所得的研究發現，就不太容易推論或應用到其他地區。例如，西方社會的性觀念和性態度相異於國內，假如現在西方國家有一研究顯示，有婚前性行為的比率高於50％，有婚外性伴侶者也超過1/3；我們能推論說，國內有婚前性行為者高達50％？或國內有婚外性伴侶者也逾1/3嗎？再如，在台北市的問卷調查顯示，國中學生每日平均零用錢是120元；同理，我們也不能概括的說，台灣地區國中學生每日平均零用錢是120元。

另者，研究的情境常是人工化或控制化，所以在實驗室或臨床上所得的研究結果，如拿來解釋日常生活中的各種行為和現象，也會產生極大的誤差。因此，為提高外在效度，進行研究的情境要儘量與自然情境趨於一致，不要有太大的落差。

三、建構效度

所謂建構（construct），是指一種理論上的構想或概念，它是看不見也摸不著，甚至可說是存不存在都有待求證，但為了研究或實務上的需要，我們假設它是存在，而且可以加以探究的。所有的建構都包含兩種特質：㈠在本質上它們是某些規則的抽象摘述（abstract summaries of some regularity）；㈡它們與具體的、可觀察的實體或事件有關或相關聯。地心引力是建構的一個很好的例子，當蘋果落到地上時，可以用地心引力這個建構來解釋和預測

蘋果落地這個現象。地心引力是無法看到的，看到的只是落地的蘋果。然而，我們卻可用它來測量地心引力，並運用地心引力這個建構來發展相關的理論（周文欽，民85；Murphy & Davidshofer, 1994）。依此推之，社會及行為科學常提及的名詞，諸如經濟成長、失業率、態度、焦慮、意識型態和社經水準等都是建構。簡言之，我們可將建構界定為：解釋某個事象的變項。

基此論之，所謂建構效度（construct validity），是指在研究歷程中所涉及之變項成功操作化（operationalized）的程度；換言之，妥切賦予變項操作型定義的程度就是建構效度（Judd, Smith, & Kidder, 1991）。為了能成功的操作變項或妥切的定義變項，其所用之方法都須有理論基礎；具體而言，所用之方法主要是指變項的測量。假如，要探討焦慮與學習的關係，現在研究者以受試者咬指甲的次數代表焦慮，並以用腳趾頭夾筆寫字的速度代表學習成果，研究顯示這兩者之間沒有顯著的相關，我們能下結論說，焦慮和學習之間是沒有關係的嗎？當然不能，其主要原因是焦慮與學習的操作或測量，很明顯的並不符合焦慮與學習的理論。建構效度和內在效度很類似，兩者都是在探討研究中的變項，惟其區別是：內在效度是指正確解釋變項間關係的程度，建構效度則是指正確界定變項的程度。因此，要有高內在效度，就須先有高的建構效度。

職是之故，影響建構效度最主要的因素，是操作、界定或測量變項的方法或工具（如問卷或測驗）。所以，凡在研究歷程中，能以符合理論方法或工具來界定變項，則該研究就會有理想的建構效度。

四、統計結論效度

所謂統計結論效度（statistical conclusion validity），是指正確運用統計方法解釋研究結果的程度。至於如何運用統計方法來分析資料，俾據以解釋研究結果，因涉及到統計學的專業知能，所以本文不擬具體說明運用統計的方法，僅扼要的說明影響統計結論效度的主要因素：不當的統計考驗方法和顯著水準的高低。不當的統計考驗方法，包括下述諸項：

(一)將相依樣本的資料，誤用獨立樣本的統計方法。

(二)分析前、後測資料之差異考驗應使用共變數分析，卻使用t考驗或變異數分析。

(三)統計結果須加以校正卻沒校正，如：使用卡方考驗時，當df=1且理論

次數小於5時，就必須進行耶茲氏校正（Yate's correction for continuity）（林清山，民81）。

(四)應使用無母數統計考驗（nonparametric statistical test），卻使用母數統計考驗（parametric statistical test）。

再者，顯著水準（level of significance）的不同，其統計結果亦會產生極大的變化，例如，有差異變成無差異，有相關變成無相關。因此，研究者在進行統計分析之前，就須依研究性質，決定最適當的顯著水準。

參考書目

1.周文欽（民80） 台北市外來高中學生的就學成因、生活適應及其相關因素研究，國立台灣師範大學教育研究所博士論文，未出版，台北市。

2.周文欽（民85） 效度，載於周文欽、歐滄和、許擇基、盧欽銘、金樹人、范德鑫（著）：心理與教育測驗，頁177至頁233，台北：心理出版社。

3.林清山（民81） 心理與教育統計學，台北：東華書局。

4.許擇基（民69） 教育研究的統計結論效度與構想效度，測驗年刊，12，頁15至頁32。

5.張春興（民78） 張氏心理學辭典，台北：東華書局。

6.Cook, T.D., & Campbell, D.T.（1976） The design and conduct of quasi-experiments and true experi-ments in field setting. In M. D. Dunette（Ed.）, Handbook of industrial and organizational psychology（pp.223-326）. Chicago: Rand McNally.

7.Judd, C. M., Smith, E. R., & Kidder, L. H.（1991） Research methods in social relations. Fort Worth, TX: Halt, Rinehart and Winston.

8.McBurney, D. H.（1998） Research methods（4th ed.）. Pacific Grove, CA: Brooks/Cole.

9.Murphy, K.D., & Davidshofer, C.O.（1994） Psychological testing:Principles and applications（3rd ed.）. Englewood Cliff, NJ: Prentice-Hall.

（作者為本科目學科委員兼召集人）