行政院國家科學委員會專題研究計劃成果報告

點數對評定量表信度與效度影響之模擬研究(I)

Effects of Number of Categories on Reliability and Validity: A Monte Carlo Study (I)

計畫編號：NSC 89-2413-H-002-048
執行期限：89 **年** 8 **月** 1 **日至** 90 **年** 7 **月** 31 **日**

主持人：翁儷禎　國立台灣大學心理學系
ljweng@ccms.ntu.edu.tw

一、 中文摘要

本計畫以模擬研究的方法探討影響評定量表信度的因素。模擬研究中操弄量表點數、題目間的共變數、量表題數、受試者判斷的正確度（即個別題目之信度），與點量表題目得分分配五個變項，以探究此五因素對內部一致性信度α，再測信度，真分數與觀察分數相關的平方等三種信度係數的影響。結果發現各交互作用項的影響相當小，各因素的影響可以加成模式解釋。內部一致性信度主要受題間共變數與量表題數影響；再測信度主要受受試者判斷的正確度影響，量表題數次之；真分數與觀察分數相關的平方主要受受試者判斷的正確度與量表題數影響，量尺點數與分配次之。此研究結果可作為評定量表設計之參考，以編製最能增進測量結果信度之量尺。

關鍵詞：評定量尺、李克特式量尺、量尺點數、題目分配、再測信度、內部一致性信度

Abstract

The Monte Carlo method was proposed to examine the factors that influence the reliability of Likert-type rating scales. The independent variables manipulated included number of response categories, inter-item covariance, number of items, respondents' judgment accuracy (or reliability of individual item), and item score distributions of the rating scales. Dependent variables included three estimates of test reliability: the internal consistency reliability α, the test-retest reliability, and the squared correlation between true scores and observed scores. The research is the first part of a two-year project. Results indicated that internal consistency reliability α was mainly affected by inter-item covariance and number of items. Respondent judgment accuracy and number of items affected test-retest reliability. The squared correlation between true scores and observed scores was affected by judgment accuracy, number of items, number of response categories, and item score distributions of the rating scales. It is anticipated that the results of the simulation study will assist developers of Likert-type rating scales in choosing appropriate design to achieve optimal reliability for measurement in their research.

Keywords：rating scales, Likert-type scales, number of response categories, item distributions, test-retest reliability, internal consistency reliability

二、 緣由與目的

此研究為一兩年研究計畫之第一部份，旨在探討評定量尺點數及其他因素對量表信度的影響。評定量尺為社會科學研究中常用的度量化方法之一，研究者常藉評定量表測量個人特質或反應（楊、趙，民 76)。建構評定量尺時，研究者一般會考慮量尺點數與量尺標示語的選擇兩個向度。自 Likert (1932) 提出此測量方法以來，已有許多學者探討此類量尺的特性，其中量尺點數對量表信度影響的研究最多，廣為研究者重視。然而過去有關此議題的研究結果並不一致，研究方法與設計亦是紛歧。大多數研究為以實際量表施測之實徵研究，另有一些則以模擬研究的方法進行。其中實徵研究使用的量表，討論的點數以及信度的類別不盡相同。大多數研究乃探討內部一致性信度係數α受點數影響的情形(如：柯，民 83;翁，民 88；黃，民 75; Aiken, 1983; Halpin, Halpin, & Arbet, 1994; Jenkins & Taber, 1977; Komorita & Graham, 1965; Lissitz & Green, 1975; Masters, 1974; Matell & Jacoby, 1971; Oaster, 1989; Wong, Chuen, & Fung, 1993)，點數對再測信度影響的研究則較少(翁, 民 88;黃, 民 75;Jenkins & Taber, 1977; Lissitz & Green, 1975; Matell & Jacoby, 1971; Oaster, 1989)。再測信度研究由於需要對同一群受試者施測至少兩次，較內部一致性信度研究耗力費時，因之與其相關的研究結果亦較少見。然而，再測信度受量表點數的影響乃必須研究的議題，因為內部一致性信度係數α與再測信度兩者所關切的測量誤差是不同的，內部一致性信度考慮的是題目間的相關一致性與量表測量特質的同質性，再測信度則探討不同時間測量結果的穩定性(Anastasi & Urbina, 1997)。一個具備高度內部一致性的量表，如果在不同時間測量的結果並不穩定，亦即再測信度低，則該量表將影響整體研究推論的可信度。

以往的實徵研究中，有的研究者認為點數與信度並無顯著關係 (如：Aiken, 1983; Wong, Chuen, & Fung, 1993)，有的研究認為點數與信度高低有關，但各研究建議的最優點數亦不盡相同，少至兩、三點 (如：Matell & Jacoby, 1971)，六、七點 (如：柯，民 83; Oaster, 1989; Symonds, 1924)，乃至多達 18 點或更多 (如：Champney & Marshall, 1939)。Komorita 與 Graham (1965) 認為點數對信度的影響乃視題目的同質性而定，對於同質的題目，兩點與六點的α並無差異，但對異質性的題目，六點量表的α值高於兩點量表。Masters(1974)則認為點數的影響乃需考慮量表總分的變異量。總分變異量高時，亦即受試者意見較不同時，量表點數與α無關；但總分的變異量較小，亦即受試者意見較接近時，增加量表點數可提高變異量與量表的α值。唯總分變異量同時受各題變異數與題間共變數影響，解釋上之意義或較模糊。Churchill 與 Peter(1984)整合分析 108 個與測量工具信度相關的市場行銷研究，結果發現量表點數愈多信度愈高。他們所分析的信度乃以內部一致性信度為主，再測信度為輔。

在回顧一系列與量表點數相關的文獻後，Guilford (1954)覺得多少點數最佳或乃隨情境而異，因而建議就一量表而言，多少點數最適當宜由實徵研究結果決定。此建議固然很好且值得嘗試，但許多研究可能無法在實際進行主要研究前，先進行一系列前導研究來探討將要採用或編製之量表的最佳點數。有鑑於此，Lissitz 與

Green (1975)乃率先以模擬研究的方法探討量表信度如何隨著點數與題間共變數而改變，盼望該研究結果能提供量表編製與使用者一些線索，以選擇適當的點數。

繼 Lissitz 與 Green (1975)之後，Jenkins 與 Taber (1977)以 Lissitz 與 Green 的研究為基本架構，再考慮可能與量尺點數交互影響量表信度的其他因素，進一步探討量表點數、題目間的共變數、量表題數、受試者判斷的正確度（亦即個別題目之信度）四個因素及其交互作用對信度係數的影響。吳（民 85）亦以模擬實驗探討問卷長度、項目間平均相關係數、各項目變異數變異程度三者對內部一致性信度係數α之影響，此研究乃評估各因素對模擬之連續資料α值的影響，未將連續資料轉換成間斷之評定量表式點資料，。Jenkins 與 Taber 的結果發現，操弄之四因素的交互作用並不顯著，因此可以加成的模式整合這四個因素對量表信度的影響。Lissitz 與 Green 及 Jenkins 與 Taber 此二模擬研究均發現量表點數達到五點之後，就算點數增加，信度亦不會隨之而增高，因而認為以五點量尺進行測量即已足夠。然而此二研究假設回答各點數的人數均等，亦即點量表的題目分數分配為均等分配 (uniform distribution)。此種分配在一般真實資料中並不多見，也因此其結論不一定能直接類推到實際資料上。Jenkins 與 Taber 亦指出，假設各點數的人數均相等乃該研究的限制，因為大多數實際資料的分配並非如此。因此，為考慮題目分配的影響，Bandalos 與 Enders (1996)乃進一步以十題量表為標的，研究不同題目分配對信度的影響，然其研究主題乃在探討連續真分數分配與點量表觀察分數分配兩者的相似性對信度的影響，且僅只研究內部一致性信度係數α，而未包含再測信度等其他信度估計值。因此，為增進模擬研究結果在實際資料分析時的參考性，本研究以模擬研究方法探討量尺點數等因素對量表信度的影響時，將對連續的觀察分數作不同的處理，以形成各式的點量表題目分配，而非僅假設各點的機率相同。

簡言之，本研究以模擬研究的方法探討量尺點數等五個變項對量表信度的影響。信度估計延續 Lissitz 與 Green (1975)及 Jenkins 與 Taber (1977)的作法，包含內部一致性信度α，再測信度，真分數與觀察分數相關的平方三者。模擬研究中除了操弄 Jenkins 與 Taber 所探討的量表點數、題目間的共變數、量表題數、受試者判斷的正確度（即個別題目之信度）四個因素外，亦操弄點量表的題目分配，使模擬情境更接近實徵資料的特性，以提高此模擬研究結果對量尺設計的參考價值。

三、研究方法

本研究共操弄五個獨變項，各獨變項包含之數述如下。(a)量尺點數：量尺點數涵蓋 2 至 14 點，此設計不僅包括 Lissitz 與 Green(1975)和 Jenkins 與 Taber (1977)所研究的 2、3、5、7、9、10、14 點，並且把其中的各數點量尺均包含在內。(b) 題目間的共變數：此變項依 Lissitz 與 Green 和 Jenkins 與 Taber 的作法，包含 0.2、0.5、0.8 三數值。(c) 量表題數:此變項依 Jenkins 與 Taber 的作法，包含 2、3、5、7、9、10、14 題，另加 12、15 與 20 題三種情形。(d) 受試者判斷正確度：此變項依 Jenkins 與 Taber 的作法，包含 0.50、0.70、0.85、1.00 四種情形。(e) 點量尺題目得分分配：本研究由檢視以往實徵資料中各點數量尺題目的頻率分配（例如翁，民 88），選取較常出現之分配，作為模擬研究中點量表題目

分配選取之依據。共選取六種分配情形，包含均等分配、常態分配，以及四種不同程度偏態之分配，此四分配之偏態係數與峰度係數分別為（0.5，0.5）、（1.0，1.5）、（1.5，2.25）與（2.0，4.0）。

此模擬研究之依變項包括內部一致性信度α，再測信度，真分數與觀察分數相關的平方等三種信度係數。資料產生步驟乃根據 Lissitz 與 Green(1975)之作法，先依古典測驗理論建構連續觀察資料，再轉換成各種題目得分分配之點量表反應資料。每一情境重覆 100 次，每次產生 100 個受試者資料。資料收集後，即計算各情境各信度係數的平均數與標準誤等基本統計量，並以變異數分析之 η^2 探究各因素及其交互作用對各信度係數的影響程度，$\eta^2 > .14$ 代表效果值高，$\eta^2 > .06$ 代表中等程度效果值(Cohen, 1988)。

四、 結果與討論

表一列出五個操弄變項主要效果的效果值大小 η^2。除了題數與判斷正確度對再測信度的交互作用達中等程度效果（η^2 ＝.08）外，所有的交互作用項均未有中等程度以上的效果，與 Jenkins 和 Taber(1977)的結果一致，故表一僅列出主要效果的效果值大小。結果發現影響各類信度數值的變項不一。內部一致性信度α主要受題間共變數與量表題數影響，受試者判斷正確度亦有影響。再測信度主要受受試者判斷正確度，另受量表題數與題間共變數影響。真分數與觀察分數相關的平方則同時受到受試者判斷正確度、量表題數、點量尺題目分配型態與量尺點數影響。此研究結果支持 Jenkins 和 Taber 結果，惟本研究進一步提供題目分配特性對信度的影響，題目分配主要影響真分數與觀察分數相關的平方，對內部一致性信度α與再測信度的影響小。

由於絕大多數的交互作用效果小，表二乃列出各操弄變項每一情境下三信度係數估計值的平均數與標準誤。整體而言，三種信度估計值的平均數以再測信度最高，真分數與觀察分數相關的平方次之，內部一致性信度α最低，與 Lissitz 與 Green(1975)及 Jenkins 和 Taber(1977)的結果一致。三者的標準誤則以內部一致性信度α高於其他兩者，再測信度與真分數與觀察分數相關的平方的標準誤則相去不大。表二中的信度平均數與標準誤乃多種不同情境組合之信度係數的統計量，因此顯示內部一致性信度α的數值容易因為抽樣誤差以及本研究所操弄得量表條件而變化，其餘兩者受到的影響相對而言則較小。

量尺點數主要影響真分數與觀察分數相關平方的估計值，由表二的結果可發現量尺點數超過六點後，真分數與觀察分數相關平方估計值的增加並不大，超過八點後增加量更少；此外，雖然量尺點數對其於兩種信度估計值的影響不大，但可發現點數超過五點後信度的增加量變小，七點之後增加量更是微少。雖然 Lissitz 與 Green(1975)及 Jenkins 和 Taber(1977)均建議五點的量表設計即可有令人滿意的量表信度，本研究的結果雖未與之相背，但若檢視信度的平均值與標準誤，可發現偶數點量尺採用六點，或是奇數點量尺採用七點的結果應會較為穩定。

題目間的共變數大小對內部一致性信度α影響極大，此乃與內部一致性信度α的定義一致，因其主要的誤差來源為題目同質性之高低，亦即題間共變數之大小。量表題數對三種信度係數均有影響，其中對內部一致性信度α影響最大。由表二可發現當量表題數達到五或七題後，題數增

多對再測信度與真分數和觀察分數相關平方兩種信度的增加量影響極微，內部一致性信度α則需在量表題數達到九或十題後信度估計值隨題數增加的程度方較趨緩。受試者判斷正確度主要影響再測信度，再者為真分數與觀察分數相關的平方。此結果亦與預期一致，因為受試者判斷正確度會影響受測者在同一份量表上反應的穩定程度，亦即影響觀察分數的穩定性，進而影響再測信度與真分數與觀察分數相關的平方。此結果亦顯示評量量表信度時，如果僅考慮內部一致性信度是不夠的，應再評估再測信度，如此方能知道受測者在量表反應上的穩定性。由於當量表編製完成後，以量表為測量變項的實徵研究通常僅收集一次量表資料，因此在量表發展歷程中實在需要評估其再測信度，如此將來的量表使用者方能知道一次測量所得的結果是否穩定，量表結果的穩定與否極可能影響整體研究的最終結論。

點量尺題目分配型態對信度係數的影響主要呈現在真分數與觀察分數相關的平方上，雖然點量尺分配型態偏離常態時信度係數會下降，但此下降程度在內部一致性信度α與再測信度上均極微，僅在真分數與觀察分數相關的平方上較大，信度係數平均數由常態分配時之.870 降至偏態係數 2.0，峰度係數 4.0 時之.719。此低影響度可能因為所操弄分配的型態並未極端地偏離常態，譬如偏態為 3.0 或 4.0，或是峰度高達 4.0 或 7.0 等。然而本研究所操弄得分配型態應為一般研究較常見者，顯示內部一致性信度α與再測信度對量表題目分數分配的強韌性。臨床診斷用心理測驗的題目分配可能異於一般研究，而有偏態與峰度較高的題目分配，未來亦可針對此類分配進行研究，以更全面性地瞭解題目分配對信度係數的影響。

五、 計畫成果自評

本研究執行內容與原計畫相符，由於點量表廣為研究者使用，本計畫之研究結果將具理論與實用價值，可供研究者了解各因素對信度之影響情形，並作為建構點量表時之參考，選擇適合的量尺點數，以提高測量工具的信度。研究成果在評定量尺使用廣泛的國內兼具學術及應用價值，且可與國外研究比較，應適合於國內外學術期刊發表。

六、 參考文獻

吳瑞屯（民85)。影響α內部一致性係數的因素。「中華心理學刊」，卷38：51-59。

柯永河(民83)。同一量尺，類似受試，不同作答方式會產生甚麼測驗結果？「測驗年刊」，卷41：55-72。

翁儷禎（民88)。「頻率及同意度副詞的心理量尺值研究」。國科會專題研究計劃報告：NSC 88-2423-H-002-010。

黃恆獎(民75)。「問卷調查量度方法之研究－以Likert量表為例」。國立台灣大學商學研究所碩士論文。

楊中芳、趙志裕(民76)。中國受試者所面臨的矛盾困境：對過分依賴西方評定量表的反省。「中華心理學刊」，卷29：113-132。

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. Educational and Psychological Measurement, 43, 397-401.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). London: Prentice-Hall International, Inc.

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of resonse categories on reliability. Applied Measurement in Education, 9, 151-160.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. Journal of Applied Psychology, 23, 323-331.

Churchill, G. A., Jr., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. Journal of Marketing Research, 21, 360-375.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Halpin, G., Halpin, G., & Arbet, S. (1994). Effects of number and type of response choices on internal consistency reliability. Perceptual and Motor Skills, 79, 928-930.

Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.

Jenkins, G. D., Jr., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. Journal of Applied Psychology, 62, 392-398.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. Educational and Psychological Measurement, 15, 987-995.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. Journal of Educational Measurement, 11, 49-53.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educational and Psychological Measurement, 31, 657-674.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. Perceptual and Motor Skills, 68, 549-550.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. Journal of Experimental Psychology, 7, 456-461.

Wong, C.-S., Chuen, K.-C., & Fung, M.-Y. (1993). Differences between odd and even number of response scale: Some empirical evidence. Chinese Journal of Psychology, 35, 75-86.

表一

各操弄變項在三信度指標之主要效果的效果值 η^2

操弄變項	信度		
	內部一致性信度 α	再測信度	真分數與觀察分數相關的平方
量尺點數	.019	.035	.153
題間共變數	.452	.070	.049
量表題數	.326	.158	.208
受試者判斷正確度	.087	.521	.270
點量尺題目分配	.008	.009	.163

註：所有效果值對應之 F 值皆顯著（$p<.001$）。

表二

操弄變項各情境之三信度係數估計值的平均數與標準誤

		信度		
		內部一致性係數α	再測信度	真分數與觀察分數相關的平方
操弄變項	變項水準	平均數(標準誤)	平均數(標準誤)	平均數(標準誤)
量尺點數	2	.632(.236)	.808(.177)	.683(.144)
	3	.653(.233)	.826(.167)	.719(.151)
	4	.688(.223)	.854(.146)	.786(.135)
	5	.706(.219)	.869(.136)	.813(.131)
	6	.714(.216)	.876(.131)	.830(.124)
	7	.720(.215)	.882(.127)	.835(.120)
	8	.723(.214)	.884(.125)	.844(.119)
	9	.726(.213)	.887(.122)	.846(.117)
	10	.728(.212)	.888(.122)	.851(.117)
	11	.727(.213)	.887(.122)	.849(.116)
	12	.728(.212)	.888(.122)	.852(.116)
	13	.731(.212)	.891(.120)	.855(.115)
	14	.730(.212)	.891(.121)	.854(.118)
題間共變數	.2	.510(.197)	.823(.160)	.774(.151)
	.5	.749(.163)	.882(.126)	.832(.127)
	.8	.865(.117)	.910(.106)	.844(.117)
量表題數	2	.448(.245)	.759(.192)	.688(.169)
	3	.531(.233)	.795(.173)	.729(.157)
	5	.637(.210)	.840(.145)	.781(.137)
	7	.698(.191)	.867(.126)	.812(.123)
	9	.742(.173)	.886(.113)	.834(.113)
	10	.759(.165)	.894(.108)	.842(.110)
	12	.787(.151)	.905(.098)	.855(.101)
	14	.809(.141)	.915(.092)	.866(.097)
	15	.817(.136)	.920(.085)	.871(.092)
	20	.854(.114)	.935(.072)	.889(.083)
受試者判斷正確度	0.50	.610(.225)	.730(.145)	.709(.144)
	0.70	.696(.213)	.843(.103)	.804(.116)
	0.85	.743(.205)	.914(.066)	.857(.101)
	1.00	.784(.197)	1.000(.000)	.897(.098)
點量尺題目分配（偏態/峰度）	均等分配	.727(.212)	.886(.124)	.861(.116)
	常態分配	.727(.211)	.883(.125)	.870(.121)
	0.5/0.50	.720(.214)	.878(.131)	.854(.124)
	1.0/1.50	.711(.217)	.873(.134)	.827(.126)
	1.5/2.25	.690(.226)	.859(.148)	.770(.131)
	2.0/4.00	.674(.233)	.850(.154)	.719(.129)