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(II)

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

計畫編號：NSC 90-2413-H-002-022
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日

主持人：翁儷禎　　　國立台灣大學心理學系
計畫參與人員：鄭中平　國立台灣大學心理學系
E-Mail: ljweng@ccms.ntu.edu.tw

一、 中文摘要

本研究以模擬研究方法探討量表特徵與受試者特性對評定量表效標關聯效度的影響。操弄變項包括量表與效標各題真分數的相關、量表題數、同量表內題目間的共變數、量表點數、題目分配、受試者判斷的正確度(即個別題目之信度)。結果發現，題間共變、量表題數與受試者判斷正確度對效標關聯效度的影響較大，量表點數與題目分配則影響不大。當量表與效標各題真分數的相關偏低時，量表設計難以提高效標關聯效度。

關鍵詞：評定量尺、李克特式量尺、量尺點數、題目分配、效標關聯效度、相關係數

Abstract

The research examined the effects of scale properties and respondent characteristics on criterion-related validity of Likert-type rating scales by Monte Carlo method. The independent variables manipulated include correlation between true score of items from the scale and the criterion, the number of response categories, the inter-item covariance, the number of items, respondents’ judgment accuracy (or reliability of individual item), and the distribution of scores of the rating scales. The results suggested that inter-item covariance, number of items, and judgment accuracy affected criterion-related validity. Number of response categories and score distributions had little effect on validity coefficient. When the correlation between true scores of test and criterion is low, it is unlikely to increase criterion-related validity by manipulations of scale design.

Keywords：rating scales, Likert-type scales, number of response categories, item distributions, criterion-related validity, correlation coefficient

二、 緣由與目的

Likert-type rating scales have received great popularity among social science researchers since its introduction in 1932 (Likert, 1932). It is essential to understand factors that affect results from using Likert-type rating scales. This study is part of a two-year research project proposed to investigate the effects of scale properties and respondent characteristics on reliability and validity by Monte Carlo method. The effect of various factors on criterion-related validity is the focus of this second-year study.

Validity is the most critical property of any measure (Anastasi & Urbina, 1997). A measure of high reliability but low validity demonstrates little use in understanding human behaviors. This study attempts to investigate relevant scale properties and respondent characteristics that might affect criterion-related validity. Criterion-related validity is important in both theoretical and applied research. For example, predictive validity as one type of criterion-related validity is crucial in the prediction of applicant performance from test results taken at the time of applications.

A criterion-related validity is commonly evaluated by the correlation between a test and a criterion. Martin (1973, 1978) investigated the degree of attenuation that categorization caused on Pearson product-moment correlations. Continuous variables were generated from bivariate normal distributions of different degrees of correlations between two variables. The continuous variables were transformed into ordered discrete variables of different number of scale points. Pearson product-moment correlations were calculated on the transformed discrete variables. The results indicated that number of scale points had a substantial effect on correlations. Martin (1978) suggested the use of 10 points or more on a scale whenever possible.

Martin's (1973, 1978) studies can be extended in three aspects. First, single item is insufficient for measuring abstract psychological constructs. Multiple items are needed to fully understand the complexity of human traits and attitudes. A study design that includes multiple items is necessary to evaluate the effects of categorization on criterion-related validity. Second, Martin's study represented a population study taking no consideration of sampling errors. In most psychological research, researchers are facing sample data rather than population data. Monte Carlo study can be employed to simulate the effect of sampling errors on criterion-related validity as assessed by correlation. Third, measurement errors as assessed by test reliability set the upper bounds for criterion-related validity. The results from the first-year project showed differential effects of various scale properties and respondent characteristics on three types of reliability. Because of the influence of reliability on criterion-related validity, a Monte Carlo simulation on criterion-related validity should take into account factors that influence reliability.

Compared to the great number of studies addressed to the issue of factors affecting test reliability, there has been only limited research on factors influencing validity. Hancock and Klockars (1991) empirically evaluated the effect of scale manipulations on criterion-related validity in frequency domain. Three scales were constructed, a five-point balanced scale, a five-point packed scale, and a nine-point balanced scale. Validity was measured as the correlation between respondent's ratings and the actual performance of the target behaviors as shown on a video game. The nine-point scale was found to provide the highest correlation between respondent ratings and actual performance. Klockars and Hancock (1993) repeated the same design with evaluative rating scales and found that scale format showed little effect on validity. Although the optimal scale design for any measure is perhaps best decided by empirical research (Guilford, 1954), simulation study considering various sources of influences is likely to shed light on optimal scale design for reaching satisfactory criterion-related validity.

In short, the research continued the study design in the first year and systematically investigated the effects of various scale properties and respondent characteristics on criterion-related validity. Factors affecting reliability as well as the correlation between true scores of test and criterion were expected to influence criterion-related validity.

三、研究方法

Continuous responses on each item were simulated according to the assumption of classical test theory that each response is the sum of true score and error score. The composite scores of true and error scores were transformed to the discrete distributions that represented participant responses on Likert-type rating scales. Different numbers of responses were assigned to each response category to yield desired distributions.

The Monte Carlo study manipulated five variables in addition to the correlations between true scores of the test and the criterion. The first variable was the number of scale points, varying from two to fourteen. The design included the odd number of scale points previously studied (e.g., Enders & Bandalos, 1999; Jenkins & Taber, 1977; Lissitz & Green, 1975) and was expanded to cover even number of response categories to give a more complete illustration of the trend. The second variable was the inter-item covariance set at 0.2, 0.5, and 0.8 as in Lissitz and Green, and Jenkins and Taber. The third variable was the number of items including 2, 3, 5, 7, 9, 10, and 14 as in Jenkins and Taber. Test length of 12, 15, and 20 was also studied to cover tests of a longer length. The fourth variable was the respondent judgment accuracy as represented by individual item reliability. Following Jenkins and Taber, respondent judgment accuracy was defined as the ratio of true score variance to the observed score variance for each item and took the values of 0.50, 0.70, 0.85, and 1.00. The fifth variable was the distribution of the discrete observed variable. Six distributions of various degrees of departure from normality were compared. Skewness was expected to affect correlations (Dunlap, Burke, & Greer, 1995; Dunlap, Chen, & Greer, 1994; Hutchinson, 1997). In addition to uniform distribution, the skewness and kurtosis of the other five discrete distribution were (0, 0), (0.5, 0.5), (1.0, 1.5), (1.5, 2.25), (2.0, 4.0). It was expected that these distributions would represent the majority of distributions observed in empirical studies (Micceri, 1989). The correlations between true item scores of the test and the criterion varied with inter-item covariances. For inter-item covariance of 0.2, 0.5, and 0.8, this correlation was set at 0.1, 0.1 to 0.4, and 0.1 to 0.7, by an increment of 0.1, respectively.

One hundred observations were generated for each condition, and 100 samples were replicated for each condition. After the data were generated, the correlation between test and criterion was estimated for each sample. Following Jenkins and Taber (1977) the eta squares from the analysis of variance was employed to investigate the influences of each factor and their interaction terms on the validity estimates. Eta squares was computed by the ratio of the sum of squares of the effect to the sum of squares of total. An eta squares over 0.010 and 0.059 suggested a small and a medium effect size and a value greater than 0.138 indicated an effect of large size (Cohen, 1988).

四、 結果與討論

Table 1 presented the eta squares for each effect under three inter-item covariances. Only one interaction effect reached a small size effect. The effects of factors could be therefore treated as an additive model. Table 2 gave the average estimates of criterion-related validity and the associated standard errors for each condition of the manipulated variables.

The correlation between true scores of test items and true scores of criterion items had a very large effect on criterion-related validity as expected. The criterion-related validity increases as the true correlation increases. Number of items that had large effects on three types of reliability showed a large, medium, and small effect on criterion-related validity with inter-item covariance of 0.2, 0.5, and 0.8. Test length had less effect on criterion-related validity with high inter-item covariance. The criterion-related validity seems level off after seven items. Further analysis indicated that the effect of number of items increased as the correlation between true scores of test and criterion increased. Although test reliability increases as test length increases, increasing test length has little effect on criterion-related validity if true score correlation is low.

Number of scale points had a medium size effect on the squared correlation between true scores and observed scores and small effects on coefficient alpha and test-retest reliability. The number of scale points showed only a small effect on criterion-related validity with inter-item covariance of 0.2. The result of no substantial increase of validity with number of scale points is consistent with the findings by Klockars and Hancock (1993) but inconsistent with the findings by Hancock and Klockars (1991). Further investigation is needed to explain the discrepancy. The effects of discrete score distribution on criterion-related validity were similar to the number of scale points.

Respondent judgment accuracy had a large effect on test-retest reliability and squared correlation between true and observed scores and a medium-sized effect on coefficient alpha. This factor was shown to have a medium-sized effect with inter-item covariance of 0.2 and small effects with inter-item covariances of 0.5 and 0.8. Further analysis indicated that the influences of respondent judgment accuracy increased as the correlation between true scores of test and criterion increased. It is suggested that individual item reliability has little effect on criterion-related validity when true score correlation between test and criterion is small.

Although the effect size of inter-item covariance was not included in Table 1, it can be seen from Table 2 that the criterion-related validity increased as inter-item covariance increases.

The results indicated that when the correlation between true scores of test and criterion is small, it is impossible to manipulate scale design to increase criterion-related validity. A careful choice of relevant criterion is absolutely indispensable. Although the size of reliability affects observed criterion-related validity, factors affecting reliability does not systematically influence criterion-related validity as expected. Future research is needed to examine the reasons for the lack of impact of these factors.

五、 計畫成果自評

本研究執行內容與原計畫相符，並依審查人意見修正。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，對效標關聯效度之影響情形，並作為建構點量表時之參考，選擇適合的量尺點數，以提高測量工具的效度。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，應適合於國內外學術期刊發表。

六、 參考文獻

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

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

Dunlap, W. P., Burke, M. J., & Greer, T. (1995). The effect of skew on the magnitude of product-moment correlations. Journal of General Psychology, 122, 365-377.

Dunlap, W. P., Chen, R., Greer, T. (1994). Skew reduces test-retest reliability. Journal of Applied Psychology, 79, 310-313.

Enders, C. K., & Bandalos, D. L. (1999). The effects of heterogeneous item distributions on reliability. Applied Measurement in Education, 12, 133-150.

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

Hancock, G. R., & Klockars, A. J. (1991). The effect of scale manipulations on validity: Targetting frequency rating scales for anticipated performance levels. Applied Ergonomics, 22, 147-154.

Hutchinson, T. P. (1997). A comment on correlation in skewed distributions. Journal of General Psychology, 124, 211-215.

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.

Klockars, A. J., & Hancock, G. R. (1993). Manipulations of evaluative rating scales to increase validity. Psychological Reports, 73, 1059-1066.

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.

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A test of validity. Journal of Marketing Research, 10, 316-318.

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: Additional considerations. Journal of Marketing Research, 15, 304-308.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Table 1

Effect Size η^2 of Manipulated Factors on Criterion-Related Validity

	Inter-Item Covariance		
	.2	.5	.8
R		.66341	.80542
I	.37831	.09150	.02815
C	.01816	.00746	.00583
D	.01343	.00521	.00432
P	.05958	.02040	.01031
R*I		.01959	.00769
R*C		.00121	.00090
R*D		.00042	.00032
R*P		.00526	.00398
I*C	.00029	.00030	.00035
I*D	.00021	.00006	.00008
I*P	.00418	.00217	.00397
C*D	.00124	.00053	.00046
C*P	.00005	.00000	.00002
D*P	.00006	.00000	.00003
R*I*C		.00010	.00013
R*I*D		.00005	.00005
R*I*P		.00124	.00172
R*C*D		.00010	.00010
R*C*P		.00001	.00003
R*D*P		.00002	.00003
I*C*D	.00018	.00004	.00004
I*C*P	.00022	.00005	.00001
I*D*P	.00024	.00003	.00002
C*D*P	.00006	.00000	.00002
R*I*C*D		.00003	.00002
R*I*C*P		.00006	.00004
R*I*D*P		.00009	.00006
R*C*D*P		.00001	.00001
I*C*D*P	.00029	.00002	.00001
R*I*C*D*P		.00007	.00005
Residual	.52349	.18056	.12582

Note: All the effects are significant at $p < .001$. R = correlation between true scores of test and criterion; I = number of items; C = number of response categories; D = distribution of discrete scores; P = respondent judgment accuracy.

Table 2

Means and Standard Errors of Estimates of Criterion-Related Validity

Factor	Condition	Inter-Item Covariance		
		.2	.5	.8
R	.1	.256(.129)	.146(.104)	.100(.102)
	.2		.293(.111)	.206(.104)
	.3		.445(.124)	.307(.104)
	.4		.597(.138)	.415(.107)
	.5			.524(.108)
	.6			.634(.108)
	7			.748(.110)
I	2	.111(.103)	.238(.156)	.328(.206)
	3	.138(.106)	.282(.169)	.364(.216)
	5	.203(.103)	.336(.187)	.403(.230)
	7	.237(.104)	.369(.195)	.424(.238)
	9	.266(.108)	.391(.203)	.433(.241)
	10	.284(.104)	.393(.203)	.437(.242)
	12	.303(.099)	.409(.209)	.445(.246)
	14	.324(.101)	.420(.211)	.450(.246)
	15	.327(.097)	.424(.211)	.450(.247)
	20	.368(.095)	.440(.217)	.458(.252)
C	2	.213(.127)	.325(.197)	.372(.229)
	3	.224(.130)	.338(.201)	.386(.234)
	4	.245(.129)	.360(.204)	.409(.237)
	5	.255(.129)	.369(.206)	.418(.240)
	6	.260(.128)	.374(.207)	.424(.240)
	7	.263(.128)	.377(.207)	.426(.241)
	8	.265(.128)	.379(.207)	.428(.241)
	9	.266(.128)	.380(.208)	.430(.242)
	10	.268(.128)	.382(.208)	.431(.242)
	11	.267(.128)	.381(.208)	.430(.242)
	12	.268(.128)	.382(.208)	.431(.242)
	13	.269(.128)	.383(.208)	.432(.242)
	14	.269(.128)	.383(.208)	.432(.242)
D （skewness / kurtosis）	Uniform	.270(.126)	.383(.206)	.431(.240)
	Normal	.271(.127)	.385(.207)	.435(.240)
	0.5/0.50	.265(.128)	.380(.206)	.430(.239)
	1.0/1.50	.258(.129)	.373(.206)	.422(.240)
	1.5/2.25	.243(.130)	.357(.206)	.405(.240)
	2.0/4.00	.230(.132)	.344(.206)	.391(.240)
P	0.50	.209(.127)	.324(.193)	.381(.223)
	0.70	.251(.126)	.366(.202)	.415(.237)
	0.85	.268(.127)	.387(.207)	.434(.244)
	1.00	.297(.122)	.403(.215)	.446(.251)
Total		.256(.129)	.370(.207)	.419(.240)

Note: Standard errors are in parentheses. R = correlation between true scores of test and criterion; I = number of items; C = number of response categories; D = distribution of discrete scores; P = respondent judgment accuracy.