A Psychometric Evaluation of 4-Point and 6-Point Likert-Type Scales in Relation to Reliability and Validity

Lei Chang

University of Central Florida

Reliability and validity of 4-point and 6-point scales were assessed using a new model-based approach to fit empirical data. Different measurement models were fit by confirmatory factor analyses of a multitrait-multimethod covariance matrix. 165 graduate students responded to nine items measuring three quantitative attitudes. Separation of method from trait variance led to greater reduction of reliability and heterotrait-monomethod coefficients for the 6-point scale than for the 4-point scale. Criterion-related validity was not affected by the number of scale points. The issue of selecting 4- versus 6-point scales may not be generally resolvable, but may rather depend on the empirical setting. Response conditions theorized to influence the use of scale options are discussed to provide directions for further research. *Index terms: Likert-type scales, multitrait-multimethod matrix, reliability, scale options, validity.*

Since Likert (1932) introduced the summative rating scale, now known as the Likert-type scale, researchers have attempted to find the number of scale points (i.e., item response options) that maximize reliability. Findings from these studies are contradictory. Some have claimed that reliability is independent of the number of scale points (Bendig, 1953; Boote, 1981; Brown, Widing, & Coulter, 1991; Komorita, 1963; Matell & Jacoby, 1971; Peabody, 1962; Remington, Tyrer, Newson-Smith, & Cicchetti, 1979). Others have maintained that reliability is maximized using 7-point (Cicchetti, Showalter, & Tyrer, 1985; Finn, 1972; Nunnally, 1967; Ramsay, 1973; Symonds, 1924), 5-point (Jenkins & Taber, 1977; Lissitz & Green, 1975; Remmers & Ewart, 1941), 4-point (Bendig, 1954b), or 3-point scales (Bendig, 1954a). Most of these studies investigated internal consistency reliability, except for Boote and Matell & Jacoby who used test-retest reliability, and Cicchetti et al. who examined interrater reliability.

One problem with these studies is that they did not distinguish between trait and method variance, both of which could be affected by the number of scale points. Method variance represents systematic error; if left unidentified, this component of variance would artificially increase reliability. Komorita & Graham (1965) speculated that additional scale points could sometimes raise reliability by evoking an extreme response set. Acting like halo error, such response set increases item homogeneity which is traditionally estimated as internal consistency reliability (Alliger & Williams, 1992). Part of the controversy surrounding these findings could be resolved by determining the extent to which scale points add to trait versus systematic error variance due to method.

There are three additional problems with existing reliability studies on the number of scale points. First, none of the studies used a model-fitting approach to determine which scale better fit the data. Simply comparing two reliability coefficients, as all existing studies have done, ignores other measurement considerations. For example, in the studies that found that fewer scale points resulted in higher reliability than more scale points [e.g., three scale points had higher reliability than five scale points (Bendig, 1954a); five points had higher reliability than six

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 3, September 1994, pp. 205–215

0146-6216/94/030205-11$1.80

(Matell & Jacoby, 1971) and seven points (McKelvie, 1978); and 17 points had higher reliability than 18 points (Matell & Jacoby)] it could be that the measurement model no longer fit the data obtained by using additional scale options. A second methodological limitation is that almost all of the studies (except Boote, 1981) used a nested design by comparing reliability coefficients computed from different groups of respondents. A repeated measures design would strengthen the statistical validity of this type of research. Third, researchers have compared even and odd numbers of scale points. Conclusions drawn from studies employing both even and odd numbers of scale points are indeterminate because the middle category in a scale with an odd number of points has been found to result in response sets (Cronbach, 1950; Goldberg, 1981; Nunnally, 1967). Comparing even numbers of scale options would eliminate this confound.

Apart from the contradictory reliability findings in relation to the number of scale points, little attention has been given to validity. Several studies have compared factor structures associated with 7-point versus binary scales (e.g., Comrey & Montag, 1982; Joe & Jahn, 1973; King, King, & Klockars, 1983; Oswald & Velicer, 1980; Velicer, DiClemente, & Corriveau, 1984; Velicer & Stevenson, 1978). These studies have not examined nomological or criterion-related validity involving variables not measured by the Likert-type scales. The possible systematic error due to number of scale points, such as response set and halo effect, would artificially increase reliability or monomethod correlations but not heteromethod or validity coefficients. Therefore, validity is a better criterion than reliability in evaluating the optimal number of scale points. Cronbach (1950) questioned the notion of adding scale points to increase reliability because the former may not lead to validity enhancement. He stated, "There is no merit in enhancing test reliability unless validity is enhanced at least proportionately" (Cronbach, p. 22). Studies of the optimal number of scale points, therefore, would be more meaningful if both reliability and validity were considered.

The present study compared 4-point with 6-point Likert-type scales in terms of internal consistency reliability and criterion-related validity. Systematic variations caused by the number of scale points that might spuriously increase reliability but not validity were identified. The purpose of the study was to investigate whether different numbers of scale points introduce no confounding to the latent relationship among a set of traits measured by the Likert scale, one common kind of confounding, or different kinds of confounding. Using a repeated measures design, the goodness-of-fit of different measurement models in relation to a multitrait-multimethod (MTMM) covariance matrix was examined.

Method

Instrument and Sample

Nine items taken from the Quantitative Attitudes Questionnaire (Chang, 1994) were used (see Table 1). The Quantitative Attitudes Questionnaire measures three quantitative traits—perceived quantitative ability, perceived utility of quantitative methodology for oneself, and values of quantitative methodology in social science research. Confirmatory factor analysis (CFA) was conducted on an initial sample of 112 people (Chang, 1993). A 3-factor structure was identified [$\chi^2(24) = 27, p = .32$]. The items also had been tested for τ-equivalence in relation to their respective traits; τ-equivalent items have equal true score variances (Jöreskog, 1971). τ-equivalence was tested by forcing each set of the three item loadings to be equal. Although having this restriction on the data increased the χ^2 value [$\chi^2(30) = 44$, $p = .05$], other goodness-of-fit measures [e.g., the χ^2 to degrees of freedom (df) ratio was 1.5] showed satisfactory fit to the data.

Respondents were 165 Master's students in education taking their first graduate quantitative methods course. They were enrolled in two sections of a statistics course or four sections of a research methods course. A composite score comprised of the students' midterm and final exams in either of these two courses was used as a criterion measure. Because of variable test length and item difficulty, z scores were used to form the composite. The nine items were administered twice at the beginning of the semester using 4-point and 6-point scales. The 4-point

Table 1
Nine Quantitative Attitudes Questionnaire Items: C = Perceived Quantitative Competence or Ability, U = Perceived Utility of Quantitative Methodology for Oneself, and V = Values of Quantitative Methodology in Social Science Research

Item Number	Item Content
C1	Compared to others I know, I'm very good in quantitative subjects.
U1	I need to know quantitative methodology in order to do my own research.
V1	Any theory "worth its salt" requires empirical testing.
U3	I need to know research methods to read research articles.
C2	I'm competent in quantitative research methodology.
U2	Quantitative research methodology is useful for my future career.
V2	Statistical tools are invaluable for understanding and interpreting one's data.
V3	A good researcher must have a strong background in quantitative methodology.
C3	I can learn statistics if I put in effort.

Likert scale was scored as 1=*disagree*, 2=*somewhat disagree*, 3=*somewhat agree*, and 4=*agree*. The 6-point scale was scored as 1=*strongly disagree*, 2=*disagree*, 3=*somewhat disagree*, 4=*somewhat agree*, 5=*agree*, and 6=*strongly agree*. The two administrations were one week apart. The order of the two administrations varied among the six classes. The resulting matrix (see Table 2) was a 19 × 19 MTMM variance-covariance matrix of responses to nine items measuring three quantitative traits obtained by two methods (4-point and 6-point scales) and one criterion variable, the composite exam score.

Maximum Likelihood Estimation

The 19 × 19 MTMM matrix was analyzed using maximum likelihood (ML) estimation by LISREL 7 (Jöreskog & Sörbom, 1988). Although other estimation methods have been proposed, such as weighted least squares (WLS) with a large sample asymptotic covariance matrix (Jöreskog & Sörbom) and the categorical variable methodology estimator (Muthén & Kaplan, 1985), studies by these same

Table 2
Variance-Covariance Matrix (CV is the Criterion Variable; U1–U3, V1–V3, and C1–C3 are the Items Shown in Table 1)

	4-Point Scale									6-Point Scale									
Scale	U1	U2	U3	V1	V2	V3	C1	C2	C3	U1	U2	U3	V1	V2	V3	C1	C2	C3	CV
4-Point																			
U1	.697																		
U2	.392	.825																	
U3	.413	.370	.621																
V1	.101	.074	.064	.683															
V2	.121	.095	.085	.213	.812														
V3	.204	.111	.154	.324	.350	.625													
C1	.023	.083	.048	.006	.086	.033	.849												
C2	.019	.174	.054	.011	.033	.047	.448	.778											
C3	.174	.151	.153	.042	.109	.093	.329	.209	.705										
6-Point																			
U1	.385	.302	.339	.164	.110	.164	.098	−.064	.143	1.197									
U2	.411	.638	.365	.132	.102	.207	.102	.126	.077	.653	1.584								
U3	.309	.355	.458	.073	.126	.170	.017	.042	.041	.545	.665	1.335							
V1	.088	.018	.043	.459	.168	.320	.076	.059	−.017	.372	.377	.248	1.424						
V2	.237	.181	.228	.292	.368	.322	−.077	−.101	−.017	.405	.435	.340	.586	1.556					
V3	.200	.177	.254	.204	.190	.246	−.011	−.028	.004	.478	.418	.466	.541	.603	1.074				
C1	.115	.193	.092	−.065	.087	−.018	.685	.451	.378	.187	.393	.175	.066	.032	.131	1.496			
C2	−.007	.091	.052	−.015	.021	−.007	.362	.533	.154	.105	.143	.103	.097	.079	.039	.608	1.219		
C3	.189	.248	.158	.139	.227	.129	.441	.326	.589	.407	.492	.258	.326	.270	.378	.682	.377	1.419	
CV	.187	.195	.175	.003	.202	.132	.377	.280	.239	.212	.181	.198	.154	.165	.239	.469	.277	.363	.941

authors have indicated the robustness of ML for ordinal or censored data. According to Jöreskog & Sörbom, "if the variables are highly non-normal, it is still an open question whether to use ML (or GLS) or WLS with a general weight matrix.... Previous studies have not given a clear-cut answer as to when it is necessary to use WLS rather than ML" (p. 205). Because ML has been used to analyze Likert-type data in CFA studies (e.g., Jöreskog & Sörbom), ML estimation was used here.

Goodness-of-Fit Indexes

The goodness-of-fit tests provided by LISREL 7 were used in this study. These include (1) the overall χ^2, which tests the difference of lack of fit between a hypothesized model and a saturated or just-identified model (a model is said to be just-, over-, or under-identified when there is an equal, larger, or smaller number of solutions to estimate the unknown parameters in the model, respectively; thus, a just-identified model has zero *df* and perfect fit to the data); (2) the goodness-of-fit index (GFI); (3) the adjusted goodness-of-fit index (AGFI), which adjusts for *df* (both the GFI and AGFI provide the relative amount of variance and covariance jointly explained by the model); and (4) the root mean square residual (RMR), which indicates the average discrepancy between the elements in the hypothesized and sample covariance matrices [see Jöreskog & Sörbom (1988, pp. 43-44) for a detailed explanation of these indexes]. Because χ^2 is sensitive to sample size, departure from multivariate normality, and model complexity, the ratio of χ^2 to *df* (which compensates for some of these "sensitivity" problems) also was used. A value below 2 is considered adequate fit (Bollen, 1989). Models specified in this study represented a parameter-nested sequence. The χ^2 difference test of the lack of fit between two adjacent models in a nested sequence was evaluated as the most important criterion for comparing different models.

Two subjective indexes of fit also were evaluated—the Bentler & Bonett (1980) normed fit index (BBI) and the Tucker & Lewis (1973) non-normed fit index (TLI). When the BBI is used to evaluate a hypothesized model against a null model, it represents the proportion of the maximum lack of fit that has been reduced by the hypothesized model. When it is used to compare two nested models, it represents the proportion of the maximum lack of fit that has been reduced by the relaxation of restrictions contributed by the less restricted of the two nested models. The BBI was selected because of its wide usage in the literature (Marsh, Balla, & McDonald, 1988; Mulaik, James, Alstine, Bennett, & Stilwell, 1989; Sternberg, 1992).

The TLI is similar to the BBI except that it has a penalty function on the number of parameters estimated. According to Marsh (1993; Marsh et al., 1988), the TLI is the only widely used index that compensates for the more restricted model and provides an unbiased estimate. Both the BBI and TLI range from 0.00, indicating total lack of fit, to 1.00 indicating perfect fit. Models were evaluated by examining the values of these goodness-of-fit indexes and, more importantly, by comparing the values of competing models (Marsh, 1989, 1993; Widaman, 1985).

Model Specifications

Nine a priori parameter-nested models representing different conceptions of the 4-point and 6-point scales were tested to determine which model best fit the data. This approach represents the most powerful use of structural equation modeling (Bentler & Bonnet, 1980; Jöreskog, 1971).

M0. M0 was a no-factor model, a commonly used null model in the CFA literature (Mulaik et al., 1989). Only 18 error/uniqueness variances were estimated.

Mla and Mlb. Mla was a simple CFA model. The estimated parameters included 18 factor loadings, three trait correlations, and 18 error/uniqueness variances. This model tested the hypothesis that covariation among observed variables was due only to trait factors and their intercorrelations. Acceptance of this model would lend support for the equivalence of the 4-point and 6-point Likert-type scales. In other words, the model implied that items measured by the two scale formats were congeneric indicators of the same traits. Mlb was a τ-equivalence model. It had the same specifications as Mla with the additional constraint that the factor loadings corresponding to the same traits had to be equal.

M1b was compared to M2b (discussed below).

M2a and M2b. Both were MTMM models that specified, in addition to three traits as in M1a and M1b, two method factors corresponding to the 4-point and 6-point scales. Method and trait factors were uncorrelated, which made trait, method, and error/uniqueness additive. Acceptance of M2a and M2b and rejection of M1a and M1b would indicate the presence of a method effect due to different numbers of scale points. Generally for an MTMM model to be identified there must be at least three traits and three methods (Marsh, 1989; Marsh & Hocevar, 1983). When there are fewer than three traits, the model can be identified by correlating the error/uniqueness corresponding to the same trait as a way of estimating the method variance (Kenny, 1979). When there are fewer than three methods, as was the case here, constraints are placed on the model, such as setting certain parameters equal to each other or setting them to fixed values (Hocevar, Zimmer, & Chen, 1990; Marsh & Hocevar, 1983).

Both types of parameter constraints were used in the present study. In M2b, a τ-equivalence constraint was imposed that set the factor loadings corresponding to the same trait by the same method to be equal. M2b was compared directly with M1b. In a separate analysis not reported here, the method loadings were fixed at the values obtained from M2b to estimate the trait loadings without the τ-equivalence constraints. Errors/uniquenesses obtained from this analysis were used as fixed values in M2a to estimate both trait and method factors. M2a was compared directly with M1a.

M3a, M3b, and M3c. These models estimated three traits and one method factor, instead of two method factors as was done in M2a and M2b. The same τ-equivalence constraint used in M2b was applied to these three models for identification. In M3a, one common method factor was parameterized as suggested by Widaman (1985). Comparing M2b with this model would determine whether reliability and validity were affected differently by the 4-point and 6-point scales or if the two scales had the same method contamination. In M3b, one method factor was estimated for items obtained only by the 4-point scale. In M3c, the method factor was estimated for the 6-point scale only. Comparing M3b with M3c answered the question of which scale format, 4-point or 6-point, had less method contamination.

M4. In M4, the nine items with the 4-point scale loaded onto three trait factors, whereas the nine items with the 6-point scale loaded onto another set of three trait factors. Within each set, the three traits were correlated. Intercorrelations between the two sets of three traits obtained by the two scales were not estimated. Under this model, items used with the 4-point and 6-point scales measured different traits.

Criterion-Related Validity

The nine models described above were tested again with the inclusion of the criterion variable. The criterion composite was treated as a single indicator variable with perfect reliability and 0.0 error/uniqueness. The only specification change was in the factor correlation matrix in which the criterion variable was allowed to correlate with trait but not method factors. Testing the nine measurement models with the inclusion of the criterion variable provided an opportunity to evaluate the stability of parameter estimates of the original measurement models. According to Widaman (1985), stability of common parameter estimates is an important criterion in assessing covariance structure models.

With the inclusion of the criterion variable, these models examined the nomological network relations among the three quantitative attitudes (as measured by the nine items) and quantitative performance (as measured by the composite score). Because these measurement models reflected different hypotheses regarding the behavior of scale options (namely, whether 4-point and 6-point scales introduce no method variance, one common kind or two different kinds of method contamination) the associated changes in the true network relations would provide construct and criterion-related validity evidence for or against each of the hypotheses. Similarly, internal consistency reliability also was evaluated within, and compared across, these different measurement models.

Results

Model Fit

Table 3 contains values of the goodness-of-fit in-

dexes for the nine models. Based on these indexes, the models at the two ends of the nested sequence—M1b, M1a, and M4—had the poorest fit. For example, the χ^2/df for these three models were 2.5, 2.5, and 3.6, respectively. For M1a and M1b, which had three trait factors and no method factor, it was assumed that the 4-point and 6-point scales resulted in congeneric measures of the same traits with no scale contamination. For M4, which had two sets of trait factors and no method factor, it was assumed that the 4-point and 6-point scales measured different traits (which had the same factor structure). The poor fit observed for these models indicated that the impact of the number of scale points was somewhere between two extremes—the unwanted scale confounding was neither totally absent, as shown by the poor fit of M1a and M1b, nor was the confounding to the extent that it changed what was being measured, as exemplified by the rejection of M4.

Among the remaining two sets of models—M2a and M2b, and M3a, M3b, and M3c—M2a and M2b indicated better fit. M3b, the 4-point-scale model, had the poorest fit of these four models. M3b and M3c had the same specifications, but the method factor was estimated for items using either the 4-point or 6-point scale. The difference in the χ^2 between the two models (302 – 274) was 28 (with 0 *df*) in favor of M3c. These results indicated that the 4-point scale contributed less to the method variance than did the 6-point scale.

This result was confirmed further by comparing M3a (the common-method model) with M3b (the 4-point-scale model) and with M3c (the 6-point-scale model). The reduction in the χ^2 values was substantially larger in the first comparison (M3a compared with M3b; from the χ^2 column in Table 3, 302 – 263 = 39 and from the *df* column, 135 – 126 = 9) than in the second comparison [M3a compared with M3c; $\chi^2(9) = 11$]. These results indicated that the method variance estimated in the common-method model was contributed mostly by the 6-point scale.

When two method factors were estimated to distinguish between the different numbers of scale points—M2a and M2b—the data were better explained (i.e., lower values of the fit indexes were obtained) than when one kind of scale factor was estimated using a single-method model, such as M3a, M3b, and M3c (see Table 3). M2b was directly comparable with M3a, M3b, and M3c because they all

Table 3
Goodness-of-Fit Indexes of Competing Models When the Criterion Variable Was Not Included and When the Criterion Variable Was Included

Model and Description	χ^2	*df*	χ^2/df	GFI	AGFI	RMR	BBI	TLI
Criterion Variable Not Included								
M0: null	1,128	153	7.4	.46	.39	.27	-	-
M1a: 3 traits only	339	132	2.5	.80	.75	.09	.70	.77
M1b: 3 traits only	359	144	2.5	.80	.76	.10	.68	.77
M2a: 3 traits 2 methods	223	131	1.7	.88	.84	.06	.80	.89
M2b: 3 traits 2 methods	224	125	1.8	.88	.83	.07	.80	.88
M3a: 3 traits common method	263	126	2.1	.85	.79	.07	.76	.83
M3b: 3 traits 4-pt. method	302	135	2.2	.83	.79	.10	.73	.81
M3c: 3 traits 6-pt. method	274	135	2.0	.85	.80	.08	.76	.84
M4: 6 traits	463	129	3.6	.77	.72	.18	.59	.59
Criterion Variable Included								
M0: null	1,194	172	6.9	.45	.39	.26	-	-
M1a: 3 traits only	352	148	2.4	.81	.76	.09	.70	.76
M1b: 3 traits only	359	141	2.5	.80	.75	.11	.69	.74
M2a: 3 traits 2 methods	227	147	1.6	.88	.84	.06	.81	.90
M2b: 3 traits 2 methods	243	141	1.7	.87	.83	.07	.80	.88
M3a: 3 traits common method	282	142	2.0	.85	.79	.07	.76	.83
M3b: 3 traits 4-pt. method	321	151	2.1	.83	.79	.10	.73	.81
M3c: 3 traits 6-pt. method	288	151	1.9	.84	.81	.08	.76	.85
M4: 6 traits	474	142	3.3	.79	.72	.17	.60	.61

had the τ-equivalence constraint. These models also were compared using χ^2 difference tests. For M2b versus M3a, the result was $\chi^2(1) = 39$; for M2b versus M3b, $\chi^2(10) = 78$; and for M2b versus M3c, $\chi^2(10) = 50$. These results supported the superiority of the less parsimonious M2b—the MTMM model. The lower portion of Table 3 also shows that a similar pattern of results was obtained for the various models when the nine models were tested again with the inclusion of the criterion variable.

Validity and Reliability Coefficients

Validity coefficients were the heterotrait-monomethod (HTMM) correlations among the three quantitative traits and the heterotrait-heteromethod (HTHM) correlations between the three quantitative traits and the criterion variable. These results are shown in Table 4. The three HTMM correlations estimated from the MTMM model in Table 4 (M2a) were .32, .21, and .06; these were uniformly lower in comparison with those estimated from the trait-only model (M1a; .48, .27, and .12). This difference represented the confounding method variance that inflated the trait correlations when the method components were not factored out from the trait or true score variance. However, there was almost no change across these two models in the HTHM correlations (.30, .24, and .50 in the MTMM model and .31, .25, and .51 in the trait-only model) because the criterion variable involved was measured by a different method.

When these validity coefficients were obtained from the 4-point and the 6-point scales separately (also shown in Table 4), the 6-point scale had much higher HTMM correlations (.69, .42, and .26) than the 4-point scale (.35, .15, and .09), whereas the HTHM correlations were approximately the same for the two scales (.26, .27, and .50 versus .30, .19, and .51). Apparently, the unidentified method variance inflated the HTMM correlations for the 6-point scale, producing the impression that the 6-point scale had higher HTMM validity than the 4-point scale. This method variance did not create such a difference between the two scales in the HTHM coefficients relating to the criterion variable not measured by Likert-type scales.

As is also shown in Table 4, reliability coefficients were similar between the two scales when these coefficients were estimated within the two scales respectively using separate CFAs. The reliability estimates were .67, .66, and .64 for the 6-point scale and .75, .66, and .66 for the 4-point scale. However, when they were estimated in the combined CFA MTMM model, the 6-point scale had much lower estimates (.51, .44, and .60) than the 4-point scale (.69, .63, and .69). Because method variance due to num-

Table 4
Reliability Coefficients (on the Diagonals) and Validity Coefficients (Off Diagonals) for the Criterion Variable (CV) and Traits 1–3 (T1, T2, T3)

Analysis and Variable	CV	T1	T2	T3	CV	T1	T2	T3
Combined CFA[a]	MTMM (M2a)				Trait Only (M1a)			
CV	—				—			
T1	.30	.69			.31	.51		
T2	.24	.32	.63		.25	.48	.44	
T3	.50	.21	.06	.69	.51	.27	.12	.60
Separate CFA	4-Point				6-Point			
CV	—				—			
T1	.30	.75			.26	.67		
T2	.19	.35	.66		.27	.69	.66	
T3	.51	.15	.09	.66	.50	.42	.26	.64

[a]Validity coefficients in the left matrix were estimated from the 4-point and 6-point scales using MTMM CFA which extracted 3 trait and 2 method factors. Validity coefficients in the trait only matrix were estimated from the 4- and 6-point scales using CFA which extracted 3 trait but no method factors. Reliability coefficients on the diagonals were estimated from the MTMM CFA for the 4-point (left matrix) and 6-point scales.

ber of scale points represents systematic rather than random error, when left unaccounted for in the separate CFA, it artificially raised the internal consistency reliability. This artifact affected the 6-point scale more than the 4-point scale because the former had more systematic method variance due to additional scale points.

Discussion

Simply comparing coefficients computed from 4-point and 6-point scales can give the false impression that the two scales were approximately the same in reliability and that the 6-point scale had higher HTMM coefficients. Results from this study showed that the 6-point scale added more to the systematic method variance. When this component was factored out from the trait variance, both the reliability and the HTMM correlations were substantially reduced for the 6-point scale. Within the MTMM framework, the 4-point scale had higher reliability than the 6-point scale.

One important finding was that the number of scale points in a Likert scale affects internal consistency reliability and HTMM validity but not HTHM validity. This finding is consistent with speculations made by Cronbach (1950) and Komorita & Graham (1965). Apparently, increasing the number of scale points creates opportunities for response sets to arise, which artificially raise correlations involving the same measurement method (reliability and HTMM validity). However, such artificial scale variance does not inflate correlations with variables measured by a different method. A practical implication of this finding is that test-retest reliability as well as concurrent validity between two similar Likert scales can be better evaluated if the two scales use different numbers of scale points.

The separation of method variance from internal consistency is important. Existing studies comparing reliability among different numbers of scale points have indiscriminately allocated two kinds of systematic variance—trait and method—as internal consistency and, in some cases, may have erroneously contributed to the belief that the number of scale options is positively associated with the internal consistency reliability of a Likert scale. The present study identified a method confound contributing to such a positive association. This finding indicates that additional scale points may not necessarily enhance reliability. On the contrary, they may lead to a systematic "abuse" of the scale. In the present study, for example, some respondents may have systematically skipped certain response categories associated with the 6-point scale; or they may have used, for example, *strongly disagree* interchangeably with *disagree* throughout the instrument. Both response behaviors contribute to systematic error but not trait variance.

There are two issues concerning the number of scale points that existing studies have failed to distinguish. The first is a measurement issue that concerns the consistency or stability of responses as a function of the number of scale points. The present study addressed this issue. The results showed that consistency or reliability as well as intertrait relations were enhanced by additional scale points if the latter added to trait but not systematic or random error variance.

The other issue is statistical. The reliabilities frequently used in scale investigations are coefficient alpha (Cicchetti et al., 1985; McKelvie, 1978) and test-retest reliability, both of which are related to the Pearson correlation. Restriction of range is a well-known problem that affects the magnitude of a correlation coefficient. Studies of the number of scale points demonstrate the same problem. Nunnally (1970) stated, "The numbers of steps on rating scales tend to place limits on the sizes of correlations between scales ... and [the restriction] tends to become less and less as the number of scale steps is increased" (p. 427). Cohen (1983) demonstrated significant reduction in the correlation coefficient when a continuous scale of measurement was dichotomized. Martin (1973, 1978) conducted monte carlo studies to examine the effects of different numbers of scale points on different levels of correlation and came to the same conclusion. He pointed out further that the correlation reduction due to collapsing scales was greater when the original variables were highly correlated.

It seems that additional scale options increase statistical correlations but, up to a certain point, tend

to reduce measurement consistency. Subsequently, reliability is influenced by two competing forces: When there is more statistical gain in comparison to measurement loss, reliability increases; otherwise, reliability suffers. The present study demonstrated a useful approach for clarifying this seeming paradox—decompose correlations into those representing true trait association and those representing systematic error association. Further research should investigate the conditions under which additional scale points are more likely to increase correlations as a result of systematic trait but not systematic error or method association.

Two such response conditions are hypothesized based on the current study. One is respondent knowledge with respect to what is being measured. The respondents in this study were Master's students taking their first research methods or statistics course. Because the Likert scale was administered at the beginning of the course, they had relatively limited knowledge of quantitative research methodology. The lack of stimulus knowledge may have contributed to an "abuse" of the additional scale points associated with the 6-point scale because the respondents were unable to apply them in making the finer stimulus distinctions of which they were not fully aware. However, had the respondents perhaps been more familiar with quantitative research procedures, the finer 6-point scale might have enabled them to sort out the items in a way closer to the structural pattern of the scale, resulting in higher reliability and validity. Future studies need to look beyond a simple relation between numbers of scale points and reliability or validity for possible interaction effects between scale points and other factors, such as respondent knowledge.

Another related factor to consider in future studies is the heterogeneity of the reference frames people employ when responding to a Likert scale. In the context of the present study, such reference frames would be respondents' experiences with research and quantitative methodology. When respondents are heterogeneous with respect to the knowledge and experience they use as references, increasing the number of response alternatives may add error by allowing the respondents to draw more freely on their divergent frames of reference. In this situation, additional scale alternatives enabled the scale to capture more individual differences not reflecting attitudes toward quantitative methodology but possibly different understandings about quantitative methodology. In such a situation, the same endorsement for *I agree that statistics is important for research* might mean different things for different people. Other response conditions, such as item heterogeneity (Komorita & Graham, 1965; Masters, 1974), also offer clues to a better understanding of the function of scale options.

In measuring attitude, a person responds to an item in a way that reflects the strength or valence of the item in relation to his/her position with respect to the latent attribute that is being measured (Torgerson, 1958). The two response conditions discussed above affect respondents' ratings of their own attitude positions as well as their simultaneous activity of scaling the items. That is, for example, if Item 1 is more positive than Item 2 (with respect to the attribute being measured), a good knowledge base and a similar reference frame concerning the content of the attribute being measured will help ensure that the two items be scaled as such by the respondents (independent of their own attitude standings). Simultaneously, respondents use the same scale to gauge their own standings on each of the items of varying valence. Thus, both the respondents' attitudes and the attitudes reflected by the items determine the responses on a Likert-type scale. Future research should examine these proposed response conditions as well as other influencing factors in relation to these two entwined rating activities. A possible study could investigate, under varying response conditions, how many Likert-type scale points will best enable respondents to express their attitudes in conjunction with their scaling the items.

References

Alliger, G. M., & Williams, K. J. (1992). Relating the internal consistency of scales to rater response tendencies. *Educational and Psychological Measurement, 52,* 337–343.

Bendig, A. W. (1953). Reliability of self-ratings as a function of the amount of verbal anchoring and of

the number of categories on the scale. *Journal of Applied Psychology, 37,* 38–41.

Bendig, A. W. (1954a). Reliability and the number of rating scale categories. *Journal of Applied Psychology, 38,* 38–40.

Bendig, A. W. (1954b). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38,* 167–170.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bollen, K. M. (1989). *Structural equations with latent variables.* New York: Wiley.

Boote, A. S. (1981). Reliability testing of psychographic scales: Five-point or seven-point? Anchored or labeled? *Journal of Advertising Research, 21,* 53–60.

Brown, G., Widing, R. E., II, & Coulter, R. L. (1991). Customer evaluation of retail salespeople utilizing the SOCO scale: A replication, extension, and application. *Journal of the Academy of Marketing Science, 9,* 347–351.

Chang, L. (1993, April). *Using confirmatory factor analysis of multitrait-multimethod data to assess the psychometric equivalence of 4-point and 6-point Likert-type scales.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

Chang, L. (1994). *Quantitative Attitudes Questionnaire: Instrument development and validation.* Manuscript submitted for publication.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A monte carlo investigation. *Applied Psychological Measurement, 9,* 31–36.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Comrey, A. L., & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement, 6,* 285–289.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10,* 3–31.

Finn, R. H. (1972). Effect of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 34,* 885–892.

Goldberg, L. R. (1981). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality and Social Psychology, 41,* 517–552.

Hocevar, D., Zimmer, J., & Chen, C. Y. (1990, April). *A multitrait-multimethod analysis of the worry/emotionality component in the measurement of test anxiety.* Paper presented at a joint session of the American Educational Research Association and the National Council on Measurement in Education, Boston.

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.

Joe, V. C., & Jahn, J. C. (1973). Factor structure of the Rotter I-E Scale. *Journal of Clinical Psychology, 29,* 66–68.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–132.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications* [Computer program manual]. Chicago: SPSS, Inc.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

King, L. A., King, D. W., & Klockars, A. J. (1983). Dichotomous and multipoint scales using bipolar adjectives. *Applied Psychological Measurement, 7,* 173–180.

Komorita, S. S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology, 61,* 327–334.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25,* 987–995.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5–55.

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.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor model. *Journal of Educational Measurement, 30,* 157–183.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. *Journal of Educational Measurement, 20,* 231–248.

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.

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.

McKelvie, S. J. (1978). Graphic rating scales—How many categories? *British Journal of Psychology, 69,* 185–202.

Mulaik, S. A., James, R. L., Alstine, J. V., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit for structural equation models. *Psychological Bulletin, 105,* 430–445.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

Oswald, W. T., & Velicer, W. F. (1980). Item format and the structure of the Eysenck Personality Inventory: A replication. *Journal of Personality Assessment, 44,* 283–288.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review, 69,* 65–73.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38,* 513–533.

Remington, M., Tyrer, P. J., Newson-Smith, J., & Cicchetti, D. V. (1979). Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychological Medicine, 9,* 765–770.

Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula. *Journal of Educational Psychology, 32,* 61–66.

Sternberg, R. J. (1992). *Psychological Bulletin's* top 10 "Hit Parade." *Psychological Bulletin, 112,* 387–388.

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

Torgerson, W. J. (1958). *Theory and methods of scaling.* New York: Wiley.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Velicer, W. F., DiClemente, C. C., & Corriveau, D. P. (1984). Item format and the structure of the personal orientation inventory. *Applied Psychological Measurement, 8,* 409–419.

Velicer, W. F., & Stevenson, J. F. (1978). The relation between item format and the structure of the Eysenck Personality Inventory. *Applied Psychological Measurement, 2,* 293–304.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Acknowledgments

The author thanks the editor, two anonymous reviewers, and Dennis Hocevar for their constructive and helpful suggestions. An earlier version of this paper was presented at the 1993 annual meeting of the National Council for Measurement in Education, Atlanta GA. Preparation of this paper was partially supported by a 1993 University of Central Florida Faculty In-House Grant.

Author's Address

Send requests for reprints or further information to Lei Chang, University of Central Florida, Department of Educational Foundations, Orlando FL 32816-1250, U.S.A.