DOCUMENT RESUME

ED 358 119 TM 019 891

AUTHOR Chang, Lei
TITLE Using Confirmatory Factor Analysis of Multitrait-Multimethod Data To Assess the Psychometrical Equivalence of 4-Point and 6-Point Likert-Type Scales.
PUB DATE Apr 93
NOTE 47p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Comparative Testing; Education Majors; *Graduate Students; Higher Education; *Likert Scales; Measurement Techniques; Models; *Multitrait Multimethod Techniques; *Psychometrics; *Test Construction; Test Format; Test Reliability; Test Validity; True Scores
IDENTIFIERS *Confirmatory Factor Analysis; Rating Scale Analysis; *Test Equivalence

ABSTRACT

Equivalence in reliability and validity across 4-point and 6-point scales was assessed by fitting different measurement models through confirmatory factor analysis of a multitrait-multimethod covariance matrix. Responses to nine Likert-type items designed to measure perceived quantitative ability, self-perceived usefulness of quantitative methodology, and research values of quantitative methodology were obtained from 112 graduate education students at the University of Central Florida (Orlando). Systematic method variance due to numbers of scale points was identified. Separation of this systematic variance from true score variance resulted in greater reduction of reliability and validity for the 6-point scale than that for the 4-point scale. Overall, the 4-point scale was found to have better psychometric properties than the 6-point scale given the measurement conditions of the study. Factors that are speculated to intervene in the behavior of scale options are discussed to provide directions for further research. Nine tables present study findings. (Author/SLD)

Using Confirmatory Factor Analysis of Multitrait-Multimethod Data to Assess the Psychometrical Equivalence of 4-Point and 6-Point Likert-Type Scales

Lei Chang
College of Education
University of Central Florida
Orlando, FL 32816
(407) 823-2012

Paper Presented at the 1993 Annual Meeting of National Council on Measurement in Education, Atlanta

Abstract

Equivalence in reliability and validity across 4-point and 6-point scales was assessed by fitting different measurement models through confirmatory factor analysis of a multitrait-multimethod covariance matrix. Systematic method variance due to numbers of scale points was identified. Separation of this systematic variance from true score variance resulted in greater reduction of reliability and validity for the 6-point scale than that for the 4-point scale. Overall, the 4-point scale was shown to have better psychometric properties than the 6-point scale given the measurement conditions of the study. Factors that are speculated to intervene in the behavior of scale options were discussed to provide directions for further research.

Using Confirmatory Factor Analysis of Multitrait-Multimethod Data to Assess the Psychometrical Equivalence of 4-Point and 6-Point Likert-Type Scales

The Likert-type scale has become so much more widely used than when it was first developed over 50 years ago that renewed efforts to understand the psychometric properties of this scale type are in order. One frequently considered psychometric issue regarding the Likert-type scale is the number of points or steps used to define the scale. This issue has taken on added importance in that today's respondents either have little time for or are "fed up" with surveys and questionnaires. Thus, optimally-sized scales designed with both reliability and validity and time constraints in mind are essential for test users to continue to use Likert-type scales as an effective data collection method.

Many studies have been reported that examined the measurement properties of the Likert-type scale in terms of the number of points defining the scale. In these research efforts, the most frequently examined measurement property is reliability (Symonds, 1924; Remmers & Ewart, 1941; Bendig, 1953, 1954; Peabody, 1962; Komorita & Graham, 1965; Matell & Jacoby, 1971; Finn, 1972; Lisstz & Green, 1975; Cicchetti, Showalter, & Tyrer, 1985; Wyatt & Meyers, 1987). Findings from these studies represent an immense contradiction. Some investigators have claimed that reliability is independent from number of scale points (Bendig, 1953; Peabogy, 1962; Komorita & Graham, 1965;

Matell & Jacoby, 1971). Other researchers have maintained that reliability is maximized using 7 scale points (Symonds, 1924; Finn, 1972; Cicchetti et al., 1985), 5 scale points (Remmers & Ewart, 1941; Lisstz & Green, 1975), or 3 points (Bendig, 1954a; 1954b). In addition, employment of various test instruments and different research methods made it difficult or impossible to compare the conflicting findings. To date, there is no consensus on the number of scale points that is optimal for reliability.

Apart from the contradiction regarding the reliability of different Likert-type scale points, little attention has been given to the validity inference, especially nomological and criterion-related validity inference, in relation to scale points. Cronbach (1950) has long pointed out that the ultimate criterion for examining the number of scale points is validity and that, unless for validity, there is little merit in increasing reliability. The same point of view was shared by Komorita and Graham (1965), and Matell and Jacoby (1971). Komorita and Graham reasoned that the increase in reliability due to the addition of scale points could be spurious, an unwanted result of evoking an extreme response set. They further noted that such a reliability increase would not enhance validity if the criterion was not to be related to the response set component. Alliger and Williams (1992) also noted that halo, defined as the mean of all item intercorrelations, was a direct function of internal consistency. High reliability as a result of halo would not add to the validity of the test. Velicer and

Stevenson (1978) pointed out that the contradictory findings regarding the impact of number of scale points on test reliability may be due to the corresponding changes in factorial structure of the test items. Because of the clear relationship between internal consistency and eigenvalues and factor loadings, it is important to examine the impact of scale format on factorial validity before findings about reliability can be appropriately interpreted (Velicer & Stevenson, 1978).

Yet, no empirical studies have examined different scale formats in measuring a construct in terms of the expected nomological relations as defined by Cronbach and Meehl (1955) where the construct is a part. A few validity studies employing factor analysis have limited their focus on the influence of scale points on the internal structure of a test (Joe & Jahn, 1973, Oswald & Velicer, 1980, Velicer & Stevenson, 1978; Comrey & Montag, 1982; King, King, & Klockars, 1983; Velicer, Cherico, & Corriveau, 1984). The majority of these factor analysis studies have shown that Likert-type scales with more than two points (mostly 7-point) allow more meaningful and finer manifestations of individual differences on the latent construct than does the two-choice scale format. Thus, the former resulted in larger factor loadings and explained variance and different factorial structures than the latter (Joe & Jahn, 1973, Oswald & Velicer, 1980, Velicer & Stevenson, 1978; Comrey & Montag, 1982; Velicer, Cherico, & Corriveau, 1984). Contradictory findings were reported by King et al. (1983), who observed similar explained variance

and factorial invariance across 7-point and 2-point scales. Not only are these studies limited to examining only factorial validity, but the comparison across scale points is limited to that between 7-point and binary scales (except for one study by Velicer et al. (1984) where 6-point scale was compared with the binary format). Most of these studies also invite questions regarding the use of traditional factor analysis for binary data.

Also, most of the previous research has examined odd numbers of scale options, whereas the impact of even numbers of scale options on reliability and validity has not been studied. Studies where both odd and even numbers of scale points were employed confounded their investigations of the "number of scale points" issue with the separate issue of "odd vs. even" numbers of points. For example, when 4 scale points were found to have higher reliability than 3 points, it would be unclear whether the rise of reliability was due to an increase in scale points from 3 to 4 or to the absence of a "neutral" category in the 4-point scale. Cronbach (1950) and Nunnally (1967) advocated the use of even numbers of scale points because the "neutral" category in an odd number of scale points was believed to introduce room for response set.

As an exploratory study, the present investigation aimed at comparing internal consistency reliability and nomological validity between 4-point and 6-point Likert-type scales. Emphasis was given to the investigation of validity, which was distinguished between internal and external validity. In this

study, internal validity referred to the construct validity in a narrower sense or the internal structure of the traits as represented by the hypothesized relations between the test items and latent traits the items purported to measure as well as the inter-trait correlations. External validity referred to the construct validity in its broad sense or a hypothesized network of relations between the latent traits of interest and other constructs or variables. Both validities as well as internal consistency reliability were examined using confirmatory factor analysis of item-level multitrait-multimethod (MTMM) matrices. The focus of the study was to identify the internal structural relations among the latent traits that were being measured as well as the nomological relations between the latent traits and external traits or constructs by reproducing the MTMM matrix. Also, systematic variations due to the use of different numbers of scale points that could confound the true relations among the internal and external traits were to be identified. To the extent such true trait relations are confounded, the instrument is unreliable and invalid. In the existing studies assessing different scale formats separately, it is unknown whether or not and how much the trait relations are influenced by the specific scale formats wherein the items are written. Different numbers of scale points may introduce no confounding of the latent trait relations, one common kind of confounding, or different kinds of confounding. Determining and decomposing these potential sources of confounding can be achieved by analyzing MTMM matrices.

Method

Procedure

Responses to 9 Likert-type items designed to measure perceived quantitative ability, self-perceived usefulness of quantitative methodology, and research values of quantitative methodology were obtained from a convenient sample of 112 graduate Education students enrolled in 4 research methods and statistics courses at the University of Central Florida. These items were taken from the Quantitative Attitudes Questionnaire (QAQ) currently being developed by the author to measure different aspects of graduate students' attitudes toward quantitative research methods. The majority of the subjects were Masters students taking their first graduate-level quantitative courses. The students' responses were collected on two occasions using 4-point and 6-point Likert-type scales. The order of administration of the two scale forms varied among the four classes. The anchors for the 4-point scale were 1: Disagree, 2: Somewhat disagree, 3: Somewhat agree, 4: Agree. For the 6-point scale, the anchors were 1: Strongly disagree, 2: Disagree, 3: Somewhat disagree, 4: Somewhat agree, 5: Agree, 6: Strongly agree. In addition to the 9 Likert-type QAQ items, the students' midterm scores from the research methods and statistics courses where they were enrolled were used as an external validity variable. The midterms, for both kinds of courses, consisted of 75 multiple-choice items. Because of different item difficulties of the midterms, the original scores were converted into z-scores.

LISREL Analysis

LISREL-7 (Joreskog & Sorbom, 1989) was used to conduct the analyses. To follow the convention established in the literature on confirmatory factor analysis and MTMM covariance analysis, parameterizations of different models in this study were illustrated by LISREL notations. The input matrix was an 18 x 18 covariance matrix where there were 9 QAQ items measuring 3 components of quantitative attitudes (Multitrait) obtained by 4-point and 6-point Likert-type scales (multimethod). Standard parameterization (in contrast to the Rindskopf parameterization) (Marsh, 1989), which means fixing one factor loading for each factor or fixing the factor variance, was used. Although fixed-factor-loading and fixed-factor-variance parameterizations are equivalent, Marsh (1989) observed that the latter was more proper than the former. In the present study, the three QAQ trait factors were fixed at zero mean and unity variance so that the ϕ matrix in LISREL contained correlations among the 3 QAQ traits after correction for measurement attenuation. With such specifications as well as the use of correlation matrix as the input matrix, the LISREL factor loadings were also standardized; that is, the squared factor loadings, which were the lower bound of item reliability, and error/uniqueness added up to 1. Correlation rather than covariance matrices were used when determining the variance components associated with the quantitative traits (3 traits) and scale forms (2 methods).

The 4-point and 6-point Likert-type scales were treated as

interval data and were analyzed using maximum likelihood (ML) estimation in LISREL-7. Although some authors proposed other estimation methods, such as weighted least squares with a large sample asymptotic covariance matrix (WLS) (Joreskog & Sorbom, 1989) and the categorical variable methodology estimator (CVM) (Muthen & Kaplan, 1989) for treating censored data which Likert-type scales represent, Monte Carlo studies, often conducted by the same authors, mostly indicate the robustness of ML for ordinal or censored data (e.g., Muthen & Kaplan, 1985; Joreskog & Sorbom, 1989). "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." (Joreskog & Sorbom, 1989, p.205). Given that ML has been used to analyze Likert-type data in the published confirmatory factor analysis studies, the present study proceeded with ML without getting into the nuances of the estimation issue.

Goodness of fit

All the statistical tests provided by the LISREL-7 computer program were evaluated. These included the z-test associated with each parameter estimate which is the ratio between the estimate and its asymptotic standard error, the overall χ^2 which tests the difference in fit between a hypothesized model and a saturated model that perfectly reproduces data, goodness-of-fit-index (GFI) and adjusted goodness-of-fit-index (AGFI) adjusting for degrees

of freedom both of which give the relative amount of variance and covariance jointly explained by the model, the root mean square residual (RMR) which is the average fit residual between the hypothesized covariance structure and the observed covariance matrix.

Because χ^2 is sensitive to sample size and model complexity (Marsh & Hocevar, 1985), alternative indices of model fit were also evaluated as have been routinely done in the structural equation literature. These included the two incremental fit indices, the nonnormed fit index (ρ) and the normed fit index (Δ), which were developed by Bentler and Bonnet (1980) and have been commonly adopted by structural equation researchers as a non-statistical way of evaluating competing models. They are calculated in the following manner:

$$\rho = [(\chi_1^2/df_1) - (\chi_2^2/df_2)] / [(\chi_0^2/df_0) - 1]$$

$$\Delta = (\chi_1^2 - \chi_2^2) / \chi_0^2$$

In these two equations, χ_0^2, χ_1^2, and χ_2^2, df_0, df_1, and df_2 correspond to the null model, the more restrictive model, and the less restrictive model. The null model is always in the denominator to set a basis for the goodness-of-fit increments. When a model is compared with the null model, the null model also represents the more restrictive model in the numerator.

Since it is difficult to determine the values for ρ and Δ that represent various degrees of meaningfulness of models

(Bentler & Bonnet, 1980), emphasis was given in this study to the comparison of the two indices associated with different models without setting as a criterion a cut-off value.

The ratio of χ^2 and degrees of freedom (χ^2/df) was also used. No specific value was set to indicate goodness of fit whereas the purpose was to compare different models. However, as noted by Marsh and Hocevar (1985), cut-off values used by different researchers ranged from 2 to 5.

Because the difference in the χ^2 values of two nested models is itself distributed as χ^2 with degrees of freedom equal to the difference in degrees of freedom for the two nested models, the difference in fit between the two nested models can thus be tested statistically, in addition to evaluating other fit indices of each of the two models separately. All the models specified in this study represented nested models, and the χ^2 difference test was evaluated as the most important criterion for comparing different models.

In conjunction with the above fit indices, models were assessed according to parameter estimates, model parsimony, and comparisons of competing models as suggested by many authors (e.g., Widaman, 1985; Marsh, 1989), and by determining the proportion of variance explained uniquely by each set of parameters (Widaman, 1985). Models were considered "wrong" when they were underidentified, failed to converge, or produced out-of-range estimates (Marsh, 1989), or resulted in matrices that

were not positive definite when they were supposed to be so (Joreskog & Sorbom, 1989).

Initial Confirmatory Factor Analysis

Initial confirmatory factor analyses were conducted to test the a priori factorial structure of the 9 QAQ items within each of the two scale formats. In other words, if the hypothesized factorial pattern did not hold in either one of the two scale forms, no additional analyses were needed to determine validity and reliability equivalence; it would be concluded that the 4-point and 6-point scales measured different things. Two a priori models were specified based on the factorial structure of the QAQ. The first model, Model 1, was a general congeneric model (Joreskog, 1971; Linn & Werts, 1979) where there were 3 oblique factors with each factor being indicated by 3 items. The second, Model 2, was a tau-equivalent model discussed by Joreskog (1971) as a special kind of congeneric model having equal true score variance. In this model, the three factor loadings corresponding to the same factor were set to be equal. This model was derived based on the reasoning that if the items were indicators of the common trait, the true score variance of the items, which were the squared factor loadings in a congeneric model, should be equal (Joreskog, 1971). These two models were tested against a null model, Model 0, as well as an alternative single factor model, Model 3. These models were tested within the 9 items of the 4-point scale and the 6-point scale separately. For both the 4-point and 6-point scales, the a priori models fit data well

whereas the null and alternative models fit poorly. The tau-equivalent model did not fit the data as well as the general congeneric model. But it still indicated adequate fit when both parsimony and goodness of fit were considered. Goodness-of-fit indices of the two a priori and the null and alternative models are reported in Table 1. These two models established the trait factor structures for the subsequent multitrait-multimethod analyses where the impact of the number of scale points was examined.

Internal Validity

As stated earlier, the question of whether 4-point and 6-point Likert-type scales have equal reliability and validity is an exploratory question. Therefore, a series of a priori nested models representing different understanding of the scale format were tested to determine which one best fitted the data. This approach represents the most powerful use of structural equation modeling (Joreskog, 1971; Bentler & Bonnet, 1980). There are many different rationales for specifying nested models for multitrait-multimethod data (Widaman, 1985). One commonly used rationale is parsimony; other things being equal, simplicity of a model is preferred (Widaman, 1985). By the criterion of parsimony, 9 nested models of increasing complexity or decreasing parsimony were specified by adding successive sets of parameters to the previous more parsimonious model. Below are parameterizations of these models in reference to the standard confirmatory factor analysis (CFA) model of MTMM matrix. The general CFA MTMM model

for the present study was

$$\Sigma = [\Lambda_T | \Lambda_M] \begin{bmatrix} \phi_{TT} & \phi_{TM} \\ \phi_{MT} & \phi_{MM} \end{bmatrix} \begin{bmatrix} \Lambda_T' \\ \Lambda_M' \end{bmatrix} + \theta$$

where Σ was the observed 18 x 18 MTMM covariance matrix, $[\Lambda_T | \Lambda_M]$ partitioned the 18 x 5 factor loading matrix, Λ, Λ_T was an 18 x 3 submatrix of Λ containing loadings of the 18 items on the 3 traits, Λ_M was an 18 x 2 submatrix of Λ containing loadings of the 18 items on the 2 methods, ϕ_{TT} was a 3 x 3 symmetric submatrix of ϕ and contained correlations of the trait factors, ϕ_{TT} was a 2 x 2 symmetric submatrix of ϕ and contained correlations of the method factors, ϕ_{MT} was a 2 x 3 rectangular submatrix of ϕ and contained correlations of the 2 method with the 3 trait factors, Θ was an 18 x 18 diagonal matrix of error/unique variances of the 18 items.

Model 0 was a null model and the most restrictive of the series of nested models. The specifications were: Θ was free to estimate; Λ_T, Λ_M, and ϕ_{MT} were null; ϕ_{TT} and ϕ_{MM} were identity matrices.

Model 1a, the first alternative model, tested the hypothesis that covariation among observed variables was due only to trait factors and their intercorrelations. The specifications were: Θ,

Λ_T, and ϕ_{TT} were estimated; Λ_M and ϕ_{MT} were null; ϕ_{MM} was identity.

Acceptance of this model would lend support for the equivalence of 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. Rejection of Model 1a would indicate, among other possibilities, the presence of method effect or the effect due to different numbers of scale points.

Testing a model that estimated 2 method factors in addition to the 3 trait factors would answer the question whether variance due to scale format existed which could not be accounted for by the trait factors alone. This model, however, could not be identified without imposing restrictions. Without going into the details about model identification, the rule of thumb is that there have to be three traits and three methods for a MTMM confirmatory factor model to be identified (e.g., Werts, Linn, & Joreskog, 1974; Marsh & Hocevar, 1983; Hocevar, Zimmer, & Chen, 1990). Certain parameters have to be fixed to achieve model identification of MTMM CFA that have fewer than three traits and three methods. For instance, Marsh and Hocevar (1983) fixed the diagonal values of Θ at pre-computed values; Hocevar, Zimmer, and Chen (1990) fixed the method loadings when estimating trait loadings in a two-stage analysis. In the present study, such a MTMM model having 3 traits and 2 methods was identified by

imposing the tau-equivalent restrictions which set equal factor loadings corresponding to the same trait by the same method. These constraints were supported by the initial CFA's where the parsimonious tau-equivalent model indicated reasonable fit of the data.

The above described the second alternative model, Model 2a, which was specified as: Θ, Λ_T, Λ_M, ϕ_{TT}, ϕ_{TT} were estimated; ϕ_{MT} was null. By imposing tau-equivalence, 6 instead of 18 parameters were estimated in Λ_T. Making method and trait factors uncorrelated (ϕ_{MT} was null) resulted in additive trait, method, and error components (Marsh, 1989) and could answer the question of whether there was undesirable method variance due to numbers of scale points.

However, Model 2a could not be directly compared with Model 1a because of the tau-equivalence constraints in the latter that Model 1a did not have. To draw direct comparisons, Model 1b was specified to be the same as Model 1a except that tau-equivalence restrictions were introduced in Model 1b just as they were in Model 2a.

Model 2a was also tested against three alternative models, Model 2b, Model 2c, and Model 2d. In Model 2b, only one method factor was parameterized as was suggested by Widaman (1985). Comparing Model 2a with this model would determine whether there were two unique sources of method variance contributed by the 4-point and 6-point scales respectively or the same method variance

was shared by both the 4-point and 6-point scales. In the latter two alternative models, method variance was estimated for items obtained only by the 4-point scale (Model 2c) or the 6-point scale (Model 2d) but not both. These models would help answer the question of which scale format, 4-point or 6-point, had less method contamination.

Finally, Model 3 was specified where there were 2 sets of 3 trait factors and no method factor. The 9 items obtained by the 4-point scale loaded onto the three hypothesized trait factors whereas the other 9 items by the 6-point scale loaded onto another set of same three hypothesized trait factors. Within each set, the three factors were correlated. Intercorrelations between the two sets of 3 factors obtained by the two scales were not estimated. Under this model, items obtained by the two scale formats measured different traits.

External Validity

All the above internal validity models, except Model 3, were tested again with the inclusion of the external validity variable, midterm. Midterm was treated as a single indicator variable with perfect reliability and zero error/uniqueness. The only specification change involved was in the factor correlation matrix, ϕ, where for all the external validity models, midterm was allowed to correlate with trait but not method factors. Testing these external validity models provide an opportunity to evaluate the stability of parameter estimates from the internal validity models. According to Widaman (1985), stability of common

parameter estimates is an important criterion in assessing covariance structure models. These models examined the network relations among quantitative attitudes and quantitative performance when the former were specified under different measurement models. Since these measurement models reflected different hypotheses regarding the behavior of scale format, namely whether 4-point and 6-point scales introduced no method variance, one common kind of method variance, or two different kinds of variance, the associated changes in the true network relations would provide external validity evidence for or against equivalence of the scale formats.

Reliability

Internal consistency reliability corresponding to each of the three trait measures were computed for 4-point and 6-point scales using the following formula reviewed by Browne (1989):

$$\rho_c = (\Sigma\lambda_i)^2 / (\Sigma\lambda_i)^2 + \Sigma\theta_i$$

In this formula, λ^2 is the squared trait factor loading or true score variance, and θ is $(1 - \lambda^2)$, or error variance. Statistical comparison of reliability estimates from the two scales was achieved by applying Feldt (1980)'s procedure for testing differences in internal consistency reliability in the one sample case. Exposition of Feldt's rather complicated statistical test was beyond the scope of this paper. Interested readers can refer to Feldt (1969, 1980).

Results

Model fitness

Table 1 contains goodness-of-fit indices for all the internal and external validity models. Model 1a and Model 1b were equal in fitting data by almost all the goodness-of-fit indices (See Table 1). The intercorrelations among the 3 traits estimated in Models 1a and 1b were also very close in magnitudes. For Model 1a, $r_{12} = .34$, $r_{13} = .25$, $r_{23} = -.05$; for Model 1b, $r_{12} = .32$, $r_{13} = .27$, $r_{23} = -.04$. These results support the reasonableness of imposing the tau-equivalence constraints for the identification of method factors in addition to the trait factors.

Insert Table 1 about here

The models at the two ends of the nested restrictions, Model 1a, 1b, and Model 3, had the poorest fit by all the goodness-of-fit indices. Model 1a and 1b, which had 3 trait factors and no method factor, assumed that items written in 4-point and 6-point scales were congeneric measures of the same traits with zero scale contamination. Model 3, having two sets of within-scale trait factors and no method factor, made the assumption that 4-point and 6-point scales resulted in measures representing different traits (which have the same factor structures). The negative results related to these models indicate that, apparently, the impact of numbers of scale points lies somewhere

between the two extremes -- the unwanted scale confounding was neither totally absent as shown by the failure of Models 1a, 1b, nor was the confounding to the extent of changing what is being measured (failure of Model 3).

Among the remaining models, Models 2a, 2b, 2c, 2d, Model 2a, which had 3 trait and 2 method factors, indicated the best fit by all the goodness-of-fit indices. The poorest-fitting model was Model 2c, which had 3 trait factors and 1 method factor with loadings estimated only for the 4-point-scale items. In comparison with Model 2c, Model 2d, which had the same specifications and same degrees of freedom as Model 2c except that the method factor was related to items of the 6-point scale, showed contrastingly better fit. Difference in χ^2 between the two models was 37 with 0 degree of freedom in favor of Model 2d. These contrasting results indicated that the 4-point-scale contributed less to the method variance than did the 6-point scale. This point was also borne out by comparing Model 2b which specified one common method factor in addition to 3 trait factors with Models 2c and 2d which estimated scale variance in relation to either 4-point or 6-point scale but not both. When Model 2b (the common method model) was compared with Model 2c (4-pt method model), the former had an increment of fit of 7% of what could be improved over the null model (ρ =.08; Δ =.07). When the common-method model was compared with the "6-pt" model, the increment was about 2% (ρ =.01; Δ =.03). Reduction of χ^2 values was also

substantially larger in the first comparison (χ^2 = 59, df = 9) than that in the second comparison (χ^2 = 24, df = 9). This information is contained in Table 2.

Insert Table 2 about here

However, when two method factors were estimated to distinguish between the two numbers of scale points (Model 2a), the data were better explained than when one kind of scale factor was estimated (Models, 2b, 2c, 2d). When Model 2a, the 2-method model was compared with the latter three models, Model 2b having 1 common method factor, Model 2c estimating method variance for the 4-point scale, and Model 2d estimating method variance for the 6-point scale, χ^2 reduction was, respectively, 35 (df=1), 94 (df=10), and 59 (df=10). Nonnormed and normed fit indices also indicated increments in goodness-of-fit: $\rho_{2b.2a}$ = .07, $\Delta_{2b.2a}$ = .05; $\rho_{2c.2a}$ = .15, $\Delta_{2c.2a}$ = .12; $\rho_{2d.2a}$ = .08, $\Delta_{2d.2a}$ = .08.

Parameter Estimates

The parameter estimates which resulted from Model 2a are presented in Table 3. The Tau-equivalence constraints imposed in the analyses resulted in equal trait loadings corresponding to the same trait using the same method. In order to estimate the trait variance associated with each item without the tau-equivalence constraints, another analysis was conducted to test Model 2a where method factor loadings were fixed at the

previously estimated values to freely estimate all the trait loadings. Results from this new specification of Model 2a are reported in Table 4.

Insert Tables 3 and 4 about here

As can be seen from Tables 3 and 4, the factor loadings corresponding to the method factors were relatively small in comparison to the trait loadings. As a result, there were small changes in the inter-factor correlations which represented inter-trait and nomological relations free from measurement errors. Table 5 contains the correlations, estimated from different models, among the 3 traits and between the 3 traits and the external variable, midterm. In reference to Table 5, the internal validity coefficients or trait correlations for all the trait-and-method (T&M) models (Models 2a, 2b, 2c, 2d) have uniformly and substantially dropped from those estimated by the trait-only models (Models 1a, 1b, and Models 1 and 2 which have been estimated for 4-pt and 6-pt scales separately). This drop represented the confounding method variance that had inflated the trait relations when the method components were not factored out from the trait or true score variance. On the other hand, there was almost no change across these two sets of models (T&M vs. trait-only models) in the external validity coefficients. These findings are logical in that, since the external validity variable, midterm, was not measured by the Likert-type scales,

its correlation with the 3 QAQ trait factors was not expected to be influenced by the method variance due to the different numbers of Likert points. The relative stability in external validity and substantial change in internal validity associated with the two sets of models (T&M vs. trait-only models) further support the adequacy of the former models. The T&M models successfully factored out the variance due to numbers of scale points that inflated monomethod correlations but not heteromethod correlations.

Among the T&M models (Models 2a, 2b, 2c, and 2d), the magnitudes of the correlations were quite close (see Table 5). It seemed that, contrary to the goodness-of-fit indices which indicated the superiority of Model 2a, the 3-trait-2-method model, the different T&M models were the same as long as some method variance was factored out from the trait variance. However, small differences in correlation coefficients were not too surprising given that the total trait and method variance explained by any of the T&M as well as the trait-only models were not substantial. An important criterion in comparing different models is to examine the unique variance explained by the parameters specified in a model (Widaman, 1985). According to this guideline, the method variances explained by the four different T&M models were compared.

Insert Table 5 about here

Because an item loaded only on one trait and one method factor and because method and trait factors were orthogonal, squared trait and method factor loadings and error/uniqueness of an item formed additive variance components (Marsh, 1989; Joreskog & Sorbom, 1989), variance of an item was decomposed into these three sources. The method variance extracted by the different T&M models were thus calculated out. For Model 2a, method variance corresponding to the 6-point scale was 1.33, about 15% of total variance contributed by the nine 6-pt items; method variance due to the 4-point scale was .97 or 11% of the total variance of the nine 4-pt items. For Model 2b, 4-point-scale variance was .27 or 3%; 6-point scale variance was 1.18 or 13%. Model 2c which only extracted method variance due to the 4-point scale had it at .81 or 11%. Similarly, Model 2d which estimated method variance only for the 6-point scale had it at .99 or 11% of the total variance.

It is clear from above that Model 2a explained substantially more method variance than the rest of the models, corroborating its superiority as shown by the goodness-of-fit indices. Compared with Model 2b which specified one common method factor, Model 2a explained about twice as much variance at the loss of 1 degree of freedom. It was also clear that most of the common method variance in Model 2b was due to the 6-point scale format whereas the 4-point-scale items had only 3% of its variance in common with the 6-point scale. The 4-point scale had its unique scale variance at 9% of its total variance as was shown by Model 2c

which estimated method variance only for the 4-point scale. Similarly, the 6-point scale shared 2% of its variance with the 4-point scale and had unique scale variance at 13% of its total variance, as shown by Model 2d which estimated method variance only the 6-point scale. Adding up their respective unique and common variances resulted in the total variance due to the two numbers of scale points as estimated by Model 2a which specified two correlated method factors. The above information was summarized in Table 6.

Insert Table 6 about here

Reliability and Validity

Table 7 compares variance decomposition which resulted from Model 2a, the 3-trait-2-method model, and the a priori CFA models estimated within 4-point and 6-point items separately. From Table 7, it can be more clearly seen that the extraction of method variance resulted mostly in the decrease of true score variance, because method variance represented systematic rather than random variance. Normally, people would compute internal consistency reliability on the items without knowing that the computed reliability estimates had been inflated by the unwanted method variance. To demonstrate this point, coefficient α's were first computed from the separate confirmatory factor analyses within each of the two scales where method variance was not extracted. The same coefficients were then computed based on the MTMM model

which eliminated method contamination from trait variance. Within each model, the computed internal consistency coefficients were compared between the 4-point and 6-point scales using Feldt's (1980, 1969) procedures. This information is contained in Table 8. As can be seen, the 4-point scale had significantly higher reliability for only one trait measure when estimated by separate confirmatory factor analysis. When reliability was estimated by the MTMM analysis (which factored out the monomethod variance), the 4-point scale had higher reliabilities for all three trait measures.

Insert Tables 7 and 8 about here

Because of the direct relationship between true score variance and factor loadings, what was true about the reliability was true about the internal validity. From CFA estimation of trait factor loadings to the MTMM estimation, the 6-point scale showed a much bigger drop than the 4-point scale. For 6-point scale, the average factor loading for CFA was .62, and for MTMM, it was .51. The 4-point scale had .65 for CFA and .60 for MTMM.

When the method variance was not extracted, it inflated the trait variance resulting in confounded internal and external validities. This point can be seen clearly by comparing the internal and external validities estimated from Model 2a and the same correlations estimated within each scale format separately (See Table 5). The reason that correlations in the latter were

substantially larger than those in the former, was due to the 11% to 15% of the method variance that were not taken out in the separate analyses, causing the items to be more highly correlated. This impact was therefore less so in the external validity estimates (See Table 5) because the external validity variable was not supposed to share the method variance due to Likert-type scale points.

The above point is further demonstrated by information contained in Table 8. First, additional scale points added more room for method variance as was shown by the larger proportion of method variance for the 6-point scale (15%) than that of the 4-point scale (11%). Second, for the 6-point scale, the majority of the 15% method variance was factored out from true score variance (40% - 27% = 13%). For the 4-point scale, less method variance was extracted from true score variance (50% - 46% = 4%) than from error variance (50% - 43% = 7%).

Table 9 contains t-tests on validity estimates between the 4-point and 6-point scales. For all three heterotrait-monomethod correlations, or correlations among the 3 quantitative attitude components, those obtained by the 6-point scale were statistically higher than the ones based on the 4-point scale. Again, the larger amount of method variance associated with the 6-point scale inflated the true trait correlation estimates. In contrast, the difference between 4-point and 6-point scales on the heterotrait-heteromethod correlations, or the correlations between the quantitative attitudes and the external validity

variable, midterm, was much reduced because midterm did not share with the scale variance most of which was contained in the 6-point scale.

Insert Table 9 about here

Discussion

Summary and Conclusions

This study introduces a new approach to studying the traditional issue of number of scale options. This new approach is the model fitting approach to measurement. Existing studies have compared reliability and validity coefficients across different numbers of scale points as individual indices without too much concern over how different scale formats have contributed to the overall goodness-of-fit of the measurement model. As shown, simply comparing coefficients computed from 4-point and 6-point scales, an approach taken by all the existing studies, would give one the false impression that measures defined by the two numbers of scale points are basically the same.

By the goodness-of-fit indices, parameter estimates, and variance explained, the MTMM model that identified two method factors in addition to the trait factors indicate the best fit of the data. Acceptance of this model and rejection of other competing models suggest that, although measures using these two numbers of scale points converge on the same traits, the degree

of reliability and validity of the measures defined by 4-point and 6-point scales is not quite the same. Variance decomposition further demonstrates that the 6-point scale creates more method variance than does the 4-point scale resulting in inflated reliability, and internal validity or the heterotrait-monomethod correlations for the former when the method variance is not factored out from the trait variance and a reduction in these three coefficients when the method variance is separated out. Logically, the above observation is less pronounced when examining the external validity or the heterotrait-heteromethod correlations.

One important finding is that method variance contributed by numbers of scale options do not affect validity as much as they do reliability. This finding is consistent with Komorita and Graham (1965)'s speculations. Apparently, increasing scale options creates rooms for response set which artificially raises item intercorrelations. It is possible, for example, that certain scale options in the 6-point scale were systematically ignored by respondents, adding an artificial item association to the true internal consistency. This observation explains the substantial drop in internal consistency reliabilities from the general factor analysis estimates to the multitrait-multimethod analysis estimates. However, such artificial scale variance does not lead to higher correlations with variables not measured by the same scale format. The latter point is borne out by the rather stable external validity estimates from CFA to MTMM CFA. A practical

implication from this finding is that concurrent validity between two similar Likert-type tests can be better evaluated if the two tests use different numbers of scale points.

A contribution of this study is the separation of method variance from internal consistencies. Existing studies comparing scale points have indiscriminantly allocated the two kinds of systematic variance, trait variance and method variance, as internal consistencies, and, in some cases, may have erroneously contributed to the belief that number of scale points is positively associated with internal consistency reliability. Findings from the present study challenge the existence or persistence of such a positive association. As shown earlier, when the systematic method variance is accurately treated as inconsistency, the drop in reliability is significantly much more pronounced for the 6-point scale than for the 4-point scale. In addition, the 6-point scale is also found to contribute to more error variance than does the 4-point scale. This finding clarifies conclusions drawn from some existing studies supporting the other side of the controversy, namely, lower numbers of scale points have higher reliability and validity. Apparently, even when internal consistency is erroneously inflated by method variance, when such inflation is out-weighed by the error variance coming from the same source, reliability and internal validity will still suffer. However, it is premature to conclude that more scale points are bad or 6 scale points are worse than 4 scale points because of the likely interventions of other factors

that this study and most of the existing studies have not controlled. Speculations about two such intervening factors are discussed below.

Issues for Further Research

This study clearly shows that a 4-point scale has higher reliability than a 6-point scale. This finding, however, should not simply be viewed as supporting one side of a very controversial issue. Of significance is why the 4-point scale has better psychometric properties than the 6-point scale in the present study.

It is clear from this study that increase in scale options increase true score variance as well as error variance which include both systematic and random error. For example, the addition of two scale points from the 4-point to the 6-point scale added to method variance and random error variance, and added very little to the true score variance. The logic of this conclusion is apparent: It is simply harder and more error-prone to make a choice from ten options than to choose, say, from two choices. On the other hand, given that correct choices are made, ratings based on a finer scale (more scale points) have more precision than ratings based on a more coarse or crude scale. It also seems clear from this study that it is the ratio of these two sources of variance addition that determines reliability and validity (internal or heterotrait-monomethod validity); that is, if a relatively larger proportion of true score variance is added, reliability and validity will increase and visa versa.

Then the question becomes what are the conditions under which an increase in scale options leads to true score variance increase and conditions under which such an increase adds to error variance. It seems true to the author that, whatever these uncontrolled conditions are, they may provide an explanation for the existing controversial findings regarding an optimal number of scale points. The kind of findings obtained from the particular testing situations in this study lead to the speculation about two things that may contribute to the allocations of different variance additions. One is the stimuli whose differences are to be discerned. The other is respondents' knowledge of the stimuli to make the corresponding distinctions.

The three quantitative attitude components measured in the present study are quite independent as shown by the low intercorrelations among them. Quantitative confidence, perceived utility of the methodology for oneself, and value judgment of quantitative research methodology are themselves unique constructs. Relying on different reference frames, respondents could easily make the distinctions using the 4-point scale, having little need for finer scale points. In this situation, the two additional scale points contributed by the 6-point scale may have been systematically skipped by the respondents, or may have been wrongly used either in some systematic manner (e.g., "strongly disagree" is used interchangeably with "disagree") or in a random manner, adding either to method variance or error variance but not to true score variance. On the other hand, the

respondents, who were Masters students taking their first research methods course, had relatively limited knowledge of quantitative research methodology. The lack of stimuli knowledge thus contributed to the "abuse" of finer scale points because the respondents were unable to apply the finer scale points to making the stimuli distinctions of which they were not fully aware. On the other hand, one may speculate that, had the repondents been more familiar with the quantitative research procedures, the additional two scale points offered by the 6-point scale may have enabled them to sort out the items in a way closer to the structural patterns intended by the test, resulting in higher reliability and validity.[1] Future studies need to look beyond a simple linear relationship between numbers of scale points and reliability and validity for possible interaction effects between scale points and other factors, such as stimulus heterogeneity and respondents' stimulus knowledge.

A related factor the consideration of which is recommended for future study of scale points is what Masters (1974) called variability in respondents' opinion. Masters (1974) argued that where there was a lack of variability in respondents' opinion

[1] Target matter awareness or knowledge will also bring about normative change (e.g., positive vs. negative) in the attitudes towards the target matter that are being measured. The emphasis in this part of the discussion is on the impact respondents' knowledge or awareness of the target matter has on their attitudes or opinions about the target matter in terms of structural patterns and the strength of these structural relations intended by the measurement. Also, it is not knowledge but attitude that is being measured. However, there is a degree to which the knowledge or lack of it will strengthen or weaken the "true" attitude structure.

toward the content being measured, increase in the number of response categories provided opportunities for the homogeneous respondents to discriminate among themselves and therefore raise reliability. The present study seems to point in the opposite direction -- the 4-point scale had higher reliability and validity for a rather homogenous group of first-time methodology students. Studies are needed to clarify this question by more carefully determining the respondents' composition. However, it is difficult to assess homogeneity or heterogeneity of the respondents' opinion a priori when the purpose of the measurement is just that and when the assumption is that respondents are heterogeneous and normally distributed on the "opinion" that is being measured. So, stimuli knowledge is a more meaningful variable to consider in further examinations of numbers of scale options.

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(1), 38-40.

Bendig, A.W. (1954b). Reliability of short rating scales and the heterogeneity of the rated stimuli. Journal of Applied Psychology, 38(3), 167-170.

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88(3), 588-606.

Browne, R.L. (1989). Congeneric modeling of reliability using censored variables. Applied Psychological Measurement, 13(2), 151-159.

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(1), 31-36.

Comrey, A.L. & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. Applied Psychological Measurement, 6(3), 285-289.

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

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's Alpha or Kuder-Richardson Coefficient Twenty is the same for two tests. Psychometrika, 34, 363-373.

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Pschometrika, 45(1), 99-105.

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.

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

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

Joreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika, 36(2), 109-132.

Joreskog, K.G., & Sorbom, D. (1989). LISREL 7: A guide to the program and application, Spss Inc.

King, L.A., King, D.W., & Klockars, A.J. (1983). Dichotomous and multipoint scales using bipolar adjectives. Applied Psychological Measurement, 7(2), 173-180.

Komorita, S.S., & Graham, W.K. (1965). Number of scale points and the reliability of scales. Educational and Psychological Measurement, 25(4), 987-995.

Lin, R.L., & Werts, C.E. (1979). Covariance structures and their analysis, New Directions for Testing and Measurements, 4, 53-73.

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(1), 10-13.

Marsh, H.W. (1989). Confirmatory factor analyses of multrait-multimethod data: Many problems and a few solutions. Applied Psychological Measurement, 13(4), 335-361.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 20(3), 231-248.

Marsh, H.W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invarinace across groups. Psychological Bulletin, 97(3), 562-582.

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.

Muthen, B., & Kaplan, D. (1935). 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.

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

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.

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

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

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

Wyatt, R.C. & Meyers, L.S. (1987). Psychometric properties of four 5-point Likert-type response scales. Educational and Psychological Measurement, 47, 27-35.

Table 1
Goodness-of-Fit Indices for All the Models

Model	Specification	χ^2	df	χ^2/df	GFI	AGFI	RMR	ρ	Δ
Initial Confirmatory Factor Analysis									
(4-pt scale)									
0	Null	261*	36	7.3	.64	.55	.15	--	-
1	3 factors	27	24	1.1	.95	.91	.04	.98	.90
2	Tau-equivalent	44	30	1.5	.92	.87	.07	.92	.83
3	Single factor	143*	27	5.3	.77	.62	.10	.32	.45
(6-pt scale)									
0	Null	195*	36	5.4	.66	.58	.27	--	--
1	3 factors	32	24	1.3	.94	.89	.08	.93	.84
2	Tau-equivalent	52	30	1.7	.91	.80	.12	.84	.73
3	Single factor	78*	27	2.9	.86	.77	.13	.56	.61
Multitrait-Multimethod Analysis									
(Internal Validity)									
0	Null	768*	153	5.0	.49	.43	.22	--	--
1a	3 traits only	287*	132	2.2	.77	.71	.10	.70	.63
1b	3 Tau-equivalent traits	311*	144	2.2	.76	.71	.11	.70	.60
2a	3 traits 2 methods[a]	194*	125	1.5	.84	.78	.09	.87	.75
2b	3 traits 1 Method[a]	229*	126	1.8	.81	.74	.09	.80	.70
2c	3 traits 4-pt Method[a]	288*	135	2.1	.78	.72	.11	.73	.63
2d	3 traits 6-pt Method[a]	253*	135	1.9	.79	.74	.09	.78	.67
3	6 traits[b]	372*	129	2.9	.76	.68	.15	.53	.52
(External Validity)[c]									
0	Null	801*	172	4.6	.49	.43	.22	--	--
1a	3 traits only	297*	148	2.0	.78	.72	.09	.72	.63
1b	3 Tau-equivalent traits	321*	160	2.0	.77	.72	.11	.72	.60
2a	3 traits 2 methods[a]	225*	141	1.6	.83	.77	.08	.84	.72
2b	3 traits 1 Method[a]	238*	142	1.7	.81	.75	.09	.80	.70
2c	3 traits 4-pt Method[a]	299*	151	2.0	.78	.72	.10	.72	.63
2d	3 traits 6-pt Method[a]	262*	151	1.7	.80	.75	.09	.80	.67
3	6 traits	375*	142	2.6	.77	.69	.15	.56	.53

[a] Have Tau-equivalence restraints on the trait factors.
[b] The 2 sets of 3 trait factors obtained by 4-point and 6-point scales are uncorrelated while the 3 traits are correlated within each of the 2 scales.
[c] The 9 external validity models are same as the corresponding internal validity models with the inclusion of the external validity variable, midterm.

Table 2
Indices of Goodness-of-Fit Increments Between Nested Models

More vs. Less Restricted Models	Difference in χ^2	df	ρ	Δ
Internal Validity Analysis				
1b (3 traits, Tau-equivalent) vs. 1a (3 traits)	24	12	-.00	.03
1b vs. 2a (3 traits, 2 methods)	117*	19	.15	.15
2b (3 traits, 1 common method) vs. 2a	35*	1	.07	.05
2c (3 traits, 1 method for 4-pt items) vs. 2a	94*	10	.15	.12
2d (3 traits, 1 method for 6-pt items) vs. 2a	59*	10	.08	.08
3 (6 traits) vs. 2a	178*	4	.33	.23
2c vs. 2b	59*	9	.08	.07
2d vs. 2b	24*	9	.01	.03
2c vs. 2d	35*	0	.06	.05
External Validity Analysis[a]				
1b (3 traits, Tau-equivalent) vs. 1a (3 traits)	24	12	-.00	.03
1b vs. 2a (3 traits, 2 methods)	96*	19	.11	.12
2b (3 traits, 1 common method) vs. 2a	13*	1	.02	.02
2c (3 traits, 1 method for 4-pt items) vs. 2a	74*	10	.11	.09
2d (3 traits, 1 method for 6-pt items) vs. 2a	37*	10	.04	.05
3 (6 traits) vs. 2a	150*	1	.29	.19
2c vs. 2b	51*	9	.08	.06
2d vs. 2b	24*	9	.01	.03
2c vs. 2d	37*	0	.07	.05

[a] External validity models are same as the corresponding internal validity models with the inclusion of the external validity variable, midterm.
* $p<.01$.

Table 3
Estimates of Factor Loadings and Standard Errors with Tau-Equivalence Constraints

	3 Trait and 2 Method Factors				
	T_1	T_2	T_3	M_1	M_2
4-pt scale					
1	.71/.07	.00/.00	.00/.00	.49/.11	.00/.00
2	.71/.07	.00/.00	.00/.00	.20/.11	.00/.00
3	.71/.07	.00/.00	.00/.00	.39/.11	.00/.00
4	.00/.00	.68/.07	.00/.00	.10/.12	.00/.00
5	.00/.00	.68/.07	.00/.00	.12/.11	.00/.00
6	.00/.00	.68/.07	.00/.00	.15/.10	.00/.00
7	.00/.00	.00/.00	.68/.06	-.03/.10	.00/.00
8	.00/.00	.00/.00	.68/.06	-.16/.10	.00/.00
9	.00/.00	.00/.00	.68/.06	.70/.13	.00/.00
6-pt scale					
1	.52/.06	.00/.00	.00/.00	.00/.00	.30/.10
2	.52/.06	.00/.00	.00/.00	.00/.00	.37/.10
3	.52/.06	.00/.00	.00/.00	.00/.00	.22/.10
4	.00/.00	.45/.06	.00/.00	.00/.00	.20/.10
5	.00/.00	.45/.06	.00/.00	.00/.00	.37/.10
6	.00/.00	.45/.06	.00/.00	.00/.00	.44/.10
7	.00/.00	.00/.00	.60/.06	.00/.00	.19/.10
8	.00/.00	.00/.00	.60/.06	.00/.00	-.11/.10
9	.00/.00	.00/.00	.60/.06	.00/.00	.80/.10

Note. Zero's are fixed values.

Table 4
Estimates of Trait Factor Loadings and Standard Errors with Method Factor Loadings Fixed at Previous Estimates

	Estimated			Fixed	
	T_1	T_2	T_3	M_1	M_2
4-pt scale					
1	.72/.09	.00/.00	.00/.00	.49	.00
2	.71/.09	.00/.00	.00/.00	.20	.00
3	.67/.09	.00/.00	.00/.00	.39	.00
4	.00/.00	.56/.10	.00/.00	.10	.00
5	.00/.00	.65/.10	.00/.00	.12	.00
6	.00/.00	.82/.09	.00/.00	.15	.00
7	.00/.00	.00/.00	.80/.09	-.03	.00
8	.00/.00	.00/.00	.69/.09	-.16	.00
9	.00/.00	.00/.00	.57/.09	.70	.00
6-pt scale					
1	.45/.10	.00/.00	.00/.00	.00	.30
2	.63/.09	.00/.00	.00/.00	.00	.37
3	.45/.10	.00/.00	.00/.00	.00	.22
4	.00/.00	.42/.10	.00/.00	.00	.20
5	.00/.00	.48/.09	.00/.00	.00	.37
6	.00/.00	.40/.09	.00/.00	.00	.44
7	.00/.00	.00/.00	.67/.09	.00	.19
8	.00/.00	.00/.00	.49/.10	.00	-.11
9	.00/.00	.00/.00	.58/.09	.00	.80

Note. Zero's are fixed values.

Table 5
Internal and External Validity Estimates for Different Models

		r_{12}	r_{13}	r_{23}	r_{v1}	r_{v2}	r_{v3}
Internal Validity Models							
4.1	3 traits for 4-pt scale	.24	.14	-.10			
6.1	3 traits for 6-pt scale	.59	.57	.14			
1a	3 traits only	.34	.25	-.05			
1b	3 Tau-equivalent traits	.32	.27	-.04			
2a	3 traits 2 methods[a]	.29	.21	-.09			
2b	3 traits 1 Method[a]	.24	.23	-.07			
2c	3 traits 4-pt Method[a]	.38	.24	-.02			
2d	3 traits 6-pt Method[a]	.25	.24	-.11			
External Validity Models[b]							
4.1	3 traits for 4-pt scale	.24	.15	-.10	.31	.16	.32
6.1	3 traits for 6-pt scale	.57	.51	.19	.17	.19	.47
1a	3 traits only	.34	.26	-.05	.29	.20	.40
1b	3 Tau-equivalent traits	.32	.27	-.04	.29	.24	.42
2a	3 traits 2 methods[a]	.27	.22	-.09	.31	.19	.39
2b	3 traits 1 Method[a]	.25	.24	-.06	.29	.19	.41
2c	3 traits 4-pt Method[a]	.38	.25	.03	.29	.19	.40
2d	3 traits 6-pt Method[a]	.25	.24	-.11	.29	.19	.41

[a] Have Tau-equivalence restraints on the trait factors.
[b] The external validity models are same as the corresponding internal validity models with the inclusion of the external validity variable, midterm.

Table 6
Method Variance Estimates from Different Models

	4-Point Scale			6-Point Scale		
Item	2c	2b	2a	2d	2b	2a
1	.34	.06	.24	.10	.07	.09
2	.06	.01	.04	.15	.11	.14
3	.21	.02	.15	.07	.04	.05
4	.00	.00	.01	.07	.03	.04
5	.00	.00	.01	.25	.13	.14
6	.01	.00	.02	.27	.18	.19
7	.00	.02	.01	.05	.01	.03
8	.00	.06	.03	.00	.04	.01
9	.19	.10	.46	.25	.57	.64
Total	.81	.27	.97	1.21	1.18	1.33
	9%	3%	11%	13%	13%	15%

Note. Model 2c: one method factor corresponding to the 4-point scale was estimated; 2d: one method factor corresponding to the 6-point scale was estimated; 2b: one common method factor was estimated for both 4-point and 6-point scales; 2a: two correlated method factors were estimated for 4-point and 6-point scales.

Table 7
T-Test on Internal Consistency Coefficients Between 4-Point and 6-Point Scales

		CFA[a]			MTMM CFA[b]		
	$r_{4pt.6pt}$[c]	Coefficient α		t	Coefficient α		t
Trait		4-pt	6-pt		4-pt	6-pt	
T1	.58	.78	.63	3.37*	.72	.52	3.65*
T2	.55	.70	.64	1.15	.70	.41	4.33*
T3	.65	.68	.62	1.18	.72	.59	2.65*

[a] Reliability coefficients were estimated from onfirmatory factor analysis conducted separately for the 4-pt and 6-pt scales.
[b] Reliability coefficients were estimated from the multitrait-multimethod confirmatory factor analysis where two method factors were extracted.
[c] Inter-scale correlations for the corresponding trait.

* p<.01, two tailed, df=110.

Table 8
Variance Components

	4-Point Scale					6-Point Scale				
	CFA		MTMM CFA			CFA		MTMM CFA		
Item	T	U	T	M	U	T	U	T	M	U
1	.65	.35	.47	.24	.30	.33	.67	.21	.09	.70
2	.44	.56	.48	.04	.49	.53	.47	.40	.14	.46
3	.58	.43	.42	.15	.45	.27	.73	.20	.05	.75
4	.24	.76	.31	.01	.69	.23	.77	.18	.04	.78
5	.39	.61	.42	.01	.58	.52	.48	.23	.14	.62
6	.83	.17	.66	.02	.32	.46	.54	.16	.19	.64
7	.71	.29	.64	0	.36	.66	.34	.45	.03	.54
8	.42	.58	.47	.03	.50	.22	.78	.24	.01	.74
9	.24	.76	.32	.47	.22	.34	.66	.33	.64	.11
Total	4.5	4.5	4.2	.97	3.9	3.6	5.4	2.4	1.4	5.3
	50%	50%	46%	11%	43%	40%	60%	27%	15%	59%

Note. T: trait variance; M: method variance; U: error/unique variance.
CFA: Confirmatory factor analysis conducted separately for the 4-pt and 6-pt scales.
MTMM CFA: Multitrait-multimethod CFA where two correlated method factors were extracted.

Table 9
T-Test of Validity Coefficients Between 4-Point and 6-Point Scales

	Internal Validity					External Validity		
	4-pt	6-pt	t	$r_{4pt.6pt}$[a]		4-pt	6-pt	t
r_{12}	.24	.59	4.88**	.58	r_{v1}	.31	.17	1.67
r_{13}	.14	.57	5.95**	.55	r_{v2}	.16	.19	0.32
r_{23}	-.10	.14	3.15**	.65	r_{v3}	.32	.47	2.11*

[a] Inter-scale correlations for trait 1, 2, 3.
* p<.05, ** p<.01, two tailed, df=110.