Behavior Research Methods
2009, 41 (2), 295-308
doi:10.3758/BRM.41.2.295

The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects

Alberto Maydeu-Olivares
University of Barcelona, Barcelona, Spain

Uwe Kramp
University of Chile, Santiago, Chile

Carlos García-Forero and David Gallardo-Pujol
University of Barcelona, Barcelona, Spain

and

Donna Coffman
Pennsylvania State University, University Park, Pennsylvania

Despite a hundred years of questionnaire testing, no consensus has been reached on the optimal number of response alternatives in rating scales. Differences in prior research may have been due to the use of various psychometric models (classical test theory, item factor analysis, and item response theory) and different performance criteria (reliability, convergent/discriminant validity, and internal structure of the questionnaire). Furthermore, previous empirical studies on this issue have tackled the experimental design from a between-subjects perspective, thus ignoring intra-individual effects. In contrast with this approach, we propose a within-subjects experimental design and a comprehensive statistical methodology using structural equation models for studying all of these aspects simultaneously, therefore increasing statistical power. To illustrate the method, two personality questionnaires were examined using a repeated measures design. Results indicated that as the number of response alternatives increased, (1) internal consistency increased, (2) there was no effect on convergent validity, and (3) goodness of fit worsened. Finally, the article assesses the practical consequences of this research for the design of future personality questionnaires.

The development of rating scales for measuring psychological constructs, particularly those for the assessment of personality and attitudinal constructs, is an integral part of behavioral science research. Any researcher who is faced with the development of a new rating scale is confronted with the question of how many response alternatives to use for the measurement instrument in order to achieve optimal psychometric properties. No consensus has emerged, however, despite numerous studies that have attempted to answer the question of what the optimal number of response alternatives is for attitudinal and personality questionnaires (for an updated review of the literature, see Kramp, 2006; see also Churchill & Peter, 1984; Cox, 1980; Peter, 1979).

One reason for this lack might be that different criteria have been used to define "optimal." A certain number of response alternatives might maximize the reliability of the questionnaire, whereas a different number might maximize its validity (see, e.g., Chang, 1994; Preston & Colman, 2000; Sancerni, Meliá, & González Roma, 1990). Furthermore, the answer to this question may also depend on the psychometric model that is used. Researchers may use classical test theory (CTT), item factor analysis (IFA), or item response theory (IRT) as a psychometric model (see, e.g., McCallum, Keith, & Wiebe, 1988; Preston & Colman, 2000; Weng, 2004). As a result, it is necessary to investigate the effect of the number of response alternatives on the reliability and validity of the questionnaire under different psychometric models.

A researcher who is truly concerned about determining the optimal number of response alternatives, therefore, has no option but to perform a pilot study for his or her application. This pilot study can be performed using a randomized (one-way) design or a repeated measures design. In a randomized design, different groups of respondents receive a different response format. In a repeated measures design, however, the same questionnaire is administered repeatedly to the same respondents, with a different number of response alternatives each time. The

A. Maydeu-Olivares, amaydeu@ub.edu

use of a repeated measures design enables researchers to capture intra-individual effects that are due to changes in the number of response alternatives. Increased precision, and therefore higher power, moreover, can be obtained by using this design instead of a randomized design. The repeated measures design is harder to implement, however, and its data are more difficult to analyze.

The purpose of this article is to describe how repeated measures designs can be implemented to assess the effects of varying the number of response alternatives in rating scales. In particular, we describe how to assess differences using different criteria: internal consistency, evidence that is based on the relationship with other variables, and evidence that is based on internal structure (APA, AERA, & NCME, 1999). These criteria, arguably the most important when it comes to the quality of the instruments, were examined using the three different psychometric models that were described above: CTT, IFA, and IRT.

We report two examples in which a repeated measures design was used to assess the effect of the number of response alternatives. The data and software code that were used in these examples are available from the authors' Web page (www.ub.edu/gdne/amaydeusp.html).

Researchers usually choose a psychometric method first. In the following sections, therefore, we describe how to analyze the effect of varying the number of response alternatives for each of the psychometric models that is under consideration.

CTT

Given p items X_i intended to measure a psychological construct, CTT (see, e.g., Allen & Yen, 1979; Gulliksen, 1987; Lord & Novick, 1968) focuses on the respondent's observed test score, $Y_j = X_{1j} + \cdots + X_{pj}$. In CTT, Y_j is assumed to consist of two parts: the true score, η_j, on the psychological construct being measured, plus measurement error, ε_j; thus, the basic equation in this psychometric framework is

$$Y_j = \eta_j + \varepsilon_j. \tag{1}$$

The observed score's precision in measuring the psychological construct is assessed using the test reliability (i.e., the variance of the true score, divided by the variance of the observed score). A lower bound to the test reliability within CTT is obtained using coefficient alpha (Cronbach, 1951):

$$\alpha = \frac{p}{p-1}\left(\frac{\sum_{i=1}^{p}\sigma_i^2}{\sum_{i=1}^{p}\sigma_i^2 + 2\sum_{i}^{p}\sum_{i'}^{p}\sigma_{ii'}}\right), \tag{2}$$

where σ_i^2 denotes the variance of item i in the population, and $\sigma_{ii'}$ denotes the covariance between items i and i' $(i < i')$.

The evidence that is based on the internal structure of a questionnaire cannot be assessed in CTT, since Equation 1 cannot be verified in applications (see Lord, 1980, p. 5). Only the evidence that is based on the relationship with other variables (convergent and discriminant validity) can be assessed (Bollen, 1989; Jöreskog & Sörbom, 1979; McDonald, 1999). This assessment can be performed simply by computing the correlation between the test score Y and a set of relevant external criteria $C_1, \ldots, C_K$ (Gulliksen, 1987; Lord & Novick, 1968; McDonald, 1999).

To assess whether reliability is invariant across response formats, coefficient alpha may be computed as well as the 99% confidence intervals, and the overlap among the intervals can be examined. Because a normal distribution need not be a good approximation to the distribution of the item scores, asymptotically distribution-free intervals were computed for coefficient alpha (Maydeu-Olivares, Coffman, & Hartmann, 2007).

The extent to which correlations with convergent measures are invariant across response formats can also be assessed. To do so, the elements of the correlation matrix between scale scores and measures that are designed to assess related constructs must be constrained. Let Y_2, Y_3, and so on be the scale scores that are obtained when two, three, and so on response alternatives are used. The correlations between the scores for each of the external criteria $(C_1, \ldots, C_K)$ and Y_2 are then constrained to be equal to the correlations between $C_1, \ldots, C_K$ and Y_3, and so on.

IFA

In the IFA model, the observed item scores, rather than the test scores, are modeled. In particular, the observed item scores are assumed to be a linear function of a latent trait, η (a factor), representing the psychological construct that is being measured:

$$X_{ij} = \mu_i + \lambda_i\eta_j + \varepsilon_{ij}. \tag{3}$$

In Equation 3, μ_i is an intercept that changes from item to item, λ_i is the factor loading (i.e., a slope for regressing the latent factor on the observed item i), and ε_{ij} is a term containing both specification and measurement errors.

When the IFA model holds, the reliability of the test score $Y_j = X_{1j} + \cdots + X_{pj}$ is obtained using coefficient omega (McDonald, 1999):

$$\omega = \frac{\left(\sum_{i=1}^{p}\lambda_i\right)^2}{\left(\sum_{i=1}^{p}\lambda_i\right)^2 + \sum_{i=1}^{p}\psi_i^2}, \tag{4}$$

where ψ_i^2 is the unique variance for item i (i.e., the variance of ε_i in the population of respondents). Also, the evidence that is based on other variables (convergent and discriminant validity) can be assessed by estimating the correlation between the factor and the set of relevant external criteria (McDonald, 1999). Unlike in CTT, finally, the internal structure of the model can be evaluated by assessing its overall goodness of fit (McDonald, 1999; Mulaik, 1972).

IRT

In the IFA model, the categorical nature of the observed ratings is ignored. However, a suitable model for ordered categorical data can be obtained as follows. As in Equation 3, we assume that $X_{ij}^* = \mu_i + \lambda_i\eta_j + \varepsilon_{ij}$ holds. Now X_{ij}^*

is not directly observed, however; only its categorization, X_{ij}, is observed, and X_{ij} and X_{ij}^* are related via a threshold relationship. For instance, when each of the items consists of K categories (0, 1, . . . , k, . . . , $K - 1$), the threshold relationship is

$$X_i = \begin{cases} 0 & \text{if } X^* \le \tau_{i,1}, \\ k & \text{if } \tau_{i,k} < X^* \le \tau_{i,k+1}, \\ K-1 & \text{if } X^* > \tau_{i,K}. \end{cases} \quad (5)$$

Thus, the model specifies a set of $K - 1$ thresholds, τ, which change from item to item. To identify this model, it is necessary to set the intercepts μ_i to zero, and to set $\psi_i^2 = \sqrt{1 - \lambda_i^2}$.

This is the model that is estimated in structural equation modeling (SEM) when the observed variables are declared as categorical. Model estimation in SEM proceeds as follows: First, the thresholds and polychoric correlations are estimated from the data. A polychoric correlation is the correlation between two unobserved normal variables X^* that have been categorized using Equation 5. The remaining parameters of the model are then estimated from the polychorics.

The categorical IFA model that has just been introduced is a member of a larger class of latent-trait models for categorical data that are commonly referred to as *item-response models*. In the item-response literature, the model just described is referred to as Samejima's (1969) graded response model. Notice that here, the normal ogive is used as the link function.

In order to compare more easily the results for CTT, IFA, and IRT when the number of response alternatives in rating scales increases, we estimated Samejima's (1969) graded response model using an SEM approach. Thus, the estimation was performed using polychoric correlations. In so doing, we were able to assess the precision (reliability) of measurement by computing coefficient omega using Equation 4. Also, we were able to assess the convergent and discriminant validity (evidence that is based on the relationship with other variables) by estimating the correlation between the latent trait η and a set of relevant external measures (McDonald, 1999). Finally, we were able to evaluate the model's internal structure by assessing its overall goodness of fit to the polychoric correlations (Muthén, 1993).

In IRT, the precision of measurement of the psychological construct is customarily assessed using the test information function (Lord, 1980; Samejima, 1969). This is a nonlinear function of the item parameters; thus, in IRT, unlike in CTT or IFA, the precision of measurement is not constant for all levels of the latent trait. Nevertheless, it is possible to compare the models by computing coefficient omega from the estimated IRT parameters; thus, results across psychometric models (CTT, IFA, and IRT) can be compared more easily. McDonald (1999) has discussed the relationship between coefficient omega and the test information function.

The Present Studies

We now turn to the empirical studies that we performed. In each study, we investigated the effects of varying the number of response alternatives in a different target questionnaire. The samples that were used in each study were completely independent (i.e., no individual participated in both studies). All participants in a sample responded to the same questionnaire, administered with two, three, and five response alternatives in the same session. These studies addressed the issue of the optimal number of response alternatives for personality scales, using both precision (reliability) and validity as criteria. In particular, internal consistency (coefficient alpha and omega), internal structure, and convergent and discriminant validity were used to judge optimality. Prior investigations have used different criteria to judge the optimality of different numbers of response alternatives. In some cases, reliability was chosen as the optimal criterion, and in other cases, validity was chosen as the optimal criterion. Finally, previous investigations have used either CTT or IFA to examine the optimal number of response alternatives, but few studies have used IRT. In the present study, all three psychometric models were used.

METHOD

Participants

The participants were 1,172 undergraduate students at the University of Barcelona who volunteered to participate in these studies. The 1,172 students were divided into two samples, *Study A* and *Study B*. Study A consisted of 746 students who were enrolled in introductory psychology courses. Sample B consisted of 426 students who were enrolled in introductory psychology and pedagogy courses. For participants in Study A, age ranged from 17 to 53 years ($M = 20.42$, $SD = 4.14$), and for participants in Study B, age ranged from 18 to 57 years ($M = 21.33$, $SD = 4.21$). Among participants in Study A, 84.7% were female, and 84% of participants in Study B were female. Each study received a different test battery, as is described below.

Instruments

In these studies, the following questionnaires were used: the NEO Five-Factor Inventory (NEO–FFI; Costa & McCrae, 1992, 1999), the Positive Affect and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), the Social Problem-Solving Inventory–Revised (SPSI–R; D'Zurilla, Nezu, & Maydeu-Olivares, 2002), and the Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985).

NEO–FFI

The NEO–FFI is a reduced version of the Revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992, 1999). The NEO–FFI was designed to evaluate five personality dimensions: *neuroticism* (N), *extraversion* (E), *openness to experience* (O), *agreeableness* (A), and *conscientiousness* (C). Each scale consists of 12 items, and subjects use a 5-point rating scale to indicate their degree of agreement with each item (0 = *strongly disagree*, 4 = *strongly agree*). Additional information about the reliability and validity of the NEO–FFI for the Spanish form is reported in Costa and McCrae (1999).

PANAS

The PANAS consists of two 10-item scales that use a 5-point scale (0 = *very slightly or not at all*, 4 = *extremely*) that measure two broad mood types: *positive affect* (PA) and *negative affect* (NA). PA reflects the extent to which a person feels enthusiastic, active, and alert, whereas NA represents a general dimension of subjective distress (Watson et al., 1988). The PANAS can be used to measure either state affect or trait affectivity. In our case, the trait format was used. Further support for the reliability and validity of the PANAS

for the American form is reported in Watson et al., and that for the Spanish form is reported in Sandin et al. (1999).

SPSI–R

The SPSI–R consists of five scales that measure two problem-orientation dimensions—positive problem orientation (PPO) and negative problem orientation (NPO)—and three problem-solving dimensions—rational problem solving (RPS), impulsivity–carelessness style (ICS), and avoidance style (AS). The SPSI–R has two formats: the long form and the short form (SPSI–RL and SPSI–RS, respectively). The SPSI–RL was used in Study A, and the SPSI–RS was used in Study B. The SPSI–RL consists of 52 items, distributed in the following manner: PPO, 5 items; NPO, 10 items; RPS, 20 items; ICS, 10 items; AS, 7 items. The SPSI–RS is a reduced version of the SPSI–RL. Each of its scales consists of 5 items. Both the SPSI–RL and the SPSI–RS use a 5-point rating scale in which subjects indicate their agreement with statements that reflect their behavior when dealing with daily life problems (0 = *not at all true of me*, 4 = *extremely true of me*). The response format for the NPO scale was adapted for the present study to two and three response alternatives. Additional data supporting the reliability and validity of the SPSI–R for the American and Spanish forms are reported in D'Zurilla et al. (2002; see also Maydeu-Olivares, Rodríguez-Fornells, Gómez-Benito, & D'Zurilla, 2000).

SWLS

The SWLS consists of five items that were designed to evaluate individuals' subjective perception of their present life affairs (Diener et al., 1985). Respondents indicate their degree of agreement using a 7-point rating scale (1 = *strongly disagree*, 7 = *strongly agree*). For this study, the original response format was adapted to two, three, and five response alternatives. Further support for the reliability and validity of the SWLS for the American form is reported in Diener et al. and in Pavot, Diener, Colvin, and Sandvik (1991).

Procedure

Studies A and B each used a specific test battery.

Study A

The battery for Study A was composed of the NEO–FFI, the SPSI–RL (not including the NPO), the PANAS, and the NPO. The target questionnaire was the NPO (10 items), which was administered three times within the test battery, with two, three, and five response alternatives.

Study B

The test battery for Study B consisted of the NEO–FFI, the SPSI–RS (not including the NPO), the PANAS, and the SWLS. The target questionnaire was the SWLS, which was administered four times within the test battery. The first three administrations varied among two-, three-, and five-point response alternatives. The fourth administration was used to evaluate the intrasession temporal consistency of the SWLS scale (see Table 1). The NPO was not used within the SPSI–R, in order to equate both test batteries.

Different labels were used for the two-response-alternatives condition (0 = *yes*, 1 = *no*), the three-response-alternatives condition (0 = *false*, 1 = *sometimes true*, 2 = *true*), and the five-response-alternatives condition (0 = *very false*, 1 = *false*, 2 = *moderately true*, 3 = *true*, 4 = *very true*). Target questionnaires were intercalated within the longer, full test batteries. For Study A, the administration order was NPO, NEO–FFI, NPO, SPSI–R, NPO, and PANAS. For Study B, the same order was used (with the SWLS in place of the NPO scale), but with a fourth administration of the SWLS after PANAS. To control the effect of the different response formats, six counterbalanced test batteries were constructed—three for Study A and three for Study B, varying only the target instrument. Table 2 shows the distribution of the different response formats within each test battery, which kept approximately the same number of respondents for each form.

For Study A, therefore, we examined the optimal number of response alternatives for the NPO, and for Study B, we examined the optimal number of response alternatives for the SWLS. Using two instruments ensured that the optimal number of response alternatives did not depend on a particular instrument. Also, the number of items was different in each target instrument, which permitted examination of the effect of the number of items on differential reliability and validity that was due to the different number of response alternatives. In addition, we analyzed the temporal consistency for SWLS in Study B.

Table 1
Correlations Between Different Types of Response Format Options for Negative Problem Orientation (NPO) and Satisfaction With Life Scale (SWLS) Under Different Models (CTT, IFA, and IRT)

Response Format	NPO		SWLS	
	ρ	CI	ρ	CI
Classical Test Theory (CTT)				
2 & 3	.85	.82–.88	.79	.74–.84
2 & 5	.84	.82–.87	.81	.77–.85
3 & 5	.87	.85–.89	.86	.83–.89
Item Factor Analysis (IFA)				
2 & 3	.94	.91–.97	.94	.88–1.00
2 & 5	.92	.90–.94	.93	.88–.98
3 & 5	.93	.92–.95	.95	.92–.99
Item Response Theory (IRT)				
2 & 3	.94	.91–.97	.95	.89–1.00
2 & 5	.91	.88–.93	.95	.90–1.00
3 & 5	.92	.90–.94	.95	.91–.99

Note—N = 746 for NPO; N = 426 for SWLS. ρ, Pearson correlation coefficient; CI, 99% confidence interval for correlation coefficient.

The NEO–FFI, PANAS, and SPSI–R were used as variables for analyzing convergent and discriminant validity of the target instruments. These instruments were selected because of their strong theoretical background. The NPO and SWLS were selected as target instruments because they can be administered multiple times during the same session easily and because they both have a unidimensional structure, according to their respective authors. Finally, participants completed the test batteries in one session that lasted 1 h.

To summarize, the effect of the number of response alternatives on the reliability and validity of two target personality scales, the NPO and the SWLS, was investigated under three different psychometrics models: CTT, IFA, and IRT. For IRT, we used Samejima's (1969) Graded Response Model, because there is some evidence that it provides the best fit among classical parametric IRT models (Maydeu-Olivares, 2005), and also because of its relation to the IFA model. Finally, three experimental conditions were considered for each target instrument: two, three, and five response alternatives.

Statistical Analyses

Data analysis for both studies was carried out in two steps. The first step examined changes in the reliability of the test score when a different number of response alternatives were used under all

Table 2
Experimental Design That Was Used in Each Empirical Study

Test Battery	Response Alternatives Order	Test–Retest Order	Sample A		Sample B	
			N	%	N	%
A	5–3–2	5	249	33.34	140	32.86
B	3–2–5	3	251	33.65	147	34.51
C	2–5–3	2	246	33.01	139	32.63
Total			746		426	

Note—The target questionnaires that were used in Samples A and B were the NPO and SWLS, respectively.

three models. The second step examined changes in convergent and discriminant validity and in the internal structure of the model. Recall that evidence from internal structure cannot be assessed in CTT.

Analyses were performed using Mplus version 4.1. For CTT and IFA, maximum likelihood estimation was used with standard errors that were robust to nonnormality and with Satorra–Bentler mean corrections to the goodness-of-fit test statistics (see Satorra & Bentler, 1994). For IRT, diagonally weighted least squares estimation using polychoric correlations was used (see Muthén, 1993; Muthén, du Toit, & Spisic, 1997). For comparing nested models, we used the test statistic that was described in Satorra and Bentler (2001). Root mean square error of approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980), the Tucker–Lewis Index (TLI; Tucker & Lewis, 1973), and the comparative fit index (CFI; Bentler, 1990) were used as fit criteria. The 99% confidence intervals for the RMSEA were computed using FITMOD, and robust 99% confidence intervals for coefficient omega were computed using Mplus. Finally, asymptotically distribution-free 99% confidence intervals for coefficient alpha were computed in the same manner that was used in Maydeu-Olivares et al. (2007). Other SEM software can be used, such as AMOS 7.0, EQS, or LISREL.

We now describe in detail how the two steps were performed for each of the psychometric models under consideration. From this point on, *NPOx scale score* refers to the sum of NPO items with *x* response alternatives, *NPOx factor* refers to the common factor underlying those same items under the IFA model, and *NPOx latent trait* denotes the latent trait underlying the same items under Samejima's (1969) graded response model. The same notation is used for SWLS items.

CTT

First step. Coefficient alpha was computed for the NPO2, NPO3, and NPO5 scale scores and the SWLS2, SWLS3, and SWLS5 scale scores. We examined the overlap of the asymptotically distribution-free 99% confidence intervals when a different number of response alternatives was used.

Second step. For Study A, we investigated the extent to which correlations with other measures were invariant across response formats by constraining elements of the correlation matrix of scale scores and scores in external criteria. In this matrix, the correlations between each of the criteria and NPO2, NPO3, and NPO5 scale scores were constrained to be equal. The model that was used is depicted in Figure 1.

For Study B, the same procedure was used, with SWLS in place of NPO.

IFA

First step. As can be seen in Figure 2, three factors were used to model the correlations among the 30 NPO items. In this model, the 10 items of NPO with two response alternatives were indicators of an NPO2 factor, the 10 items of NPO with three response alternatives were indicators of an NPO3 factor, and the 10 items with five response alternatives were indicators of an NPO5 factor. The unique errors of items that had the same item stems were correlated. Coefficient omega was computed for NPO2, NPO3, and NPO5, and for SWLS2, SWLS3, and SWLS5. We examined the overlap of the robust 99% confidence intervals when different numbers of response alternatives were used.

Second step. In this step, we examined differences that resulted from increasing the number of response alternatives on convergent and discriminant validity and from evidence regarding the internal structure of the target questionnaires.

Differences in convergent and discriminant validity were assessed as follows. Two SEM models were fitted in each of the studies. One of the models, the *restricted* model, was a combination of the patterned correlation structure with the criteria that are in the model in Figure 1 and the latent factor model that is indicated by the items of each target scale (NPO and SWLS) in Figure 2. The structure of this restricted model is shown in Figure 3. The competing model, the *unrestricted* model, had the same specifications except that the correlations between factors and the external questionnaire scores were not constrained to be equal across factors. Since the restricted model was nested within the unrestricted model, a test could be performed to investigate the null hypothesis of equal convergent and discriminant validity.

For assessing differential internal-structure results when the number of response alternatives has been increased, a series of one-factor models was applied separately to the questionnaires with different numbers of response alternatives, and the fits of the models were compared. For Study A, therefore, a one-factor model was fit to the 10 dichotomous NPO items. The same model was fit to the 10 three-response-alternative items and to the 10 NPO items with five response alternatives. For Study B, the same procedure was used, with SWLS in place of NPO.

IRT

For IRT, we used the same procedures that we used with IFA. The only difference between both estimation methods was that in Mplus the items were declared as categorical. In so doing, the program estimated Samejima's (1969) graded response model using the sequential procedure that was described in Muthén (1984), except that diagonally weighted least squares was used instead of weighted least squares in the last stage of the estimation.

RESULTS

Classical Test Theory

Preliminary Checks

The correlations (and 99% confidence intervals) among the NPO and SWLS scale scores using different numbers of response alternatives in Study A are displayed in Table 1. The correlations ranged from .84 to .87 for NPO, and from .79 to .86 for SWLS. In neither case did the upper end of the 99% confidence intervals include 1. The lack of a perfect correlation between scale scores when different numbers of response alternatives were used can also be attributed to within-session temporal unreliability. We computed the test–retest reliabilities and their 99% confidence intervals for the Study B target questionnaire. Test–retest reliabilities were .88 (confidence interval, .82–.94) for SWLS2, .89 (confidence interval, .85–.94) for SWLS3, and .91 (confidence interval, .88–.95) for SWLS5. As can be seen, within-session test–retest reliabilities were far from 1; thus, the lack of a perfect correlation between forms with two, three, and five-response alternatives may be due to both within-session unreliability and the use of a different number of response categories.

Reliability Analysis

Coefficient alpha estimates for both studies (and their 99% confidence intervals) are presented in Table 3. NPO coefficient alpha increased from .78 for two-response-alternative categories to .88 for five-response-alternative categories. Similar results were obtained with the SWLS in Study B. As Table 3 shows, SWLS coefficient alpha increased from .62 to .76. Furthermore, inspection of the confidence intervals reveals that the increases in the internal consistency estimates from two to five response alternatives and from three to five response alternatives were statistically significant.

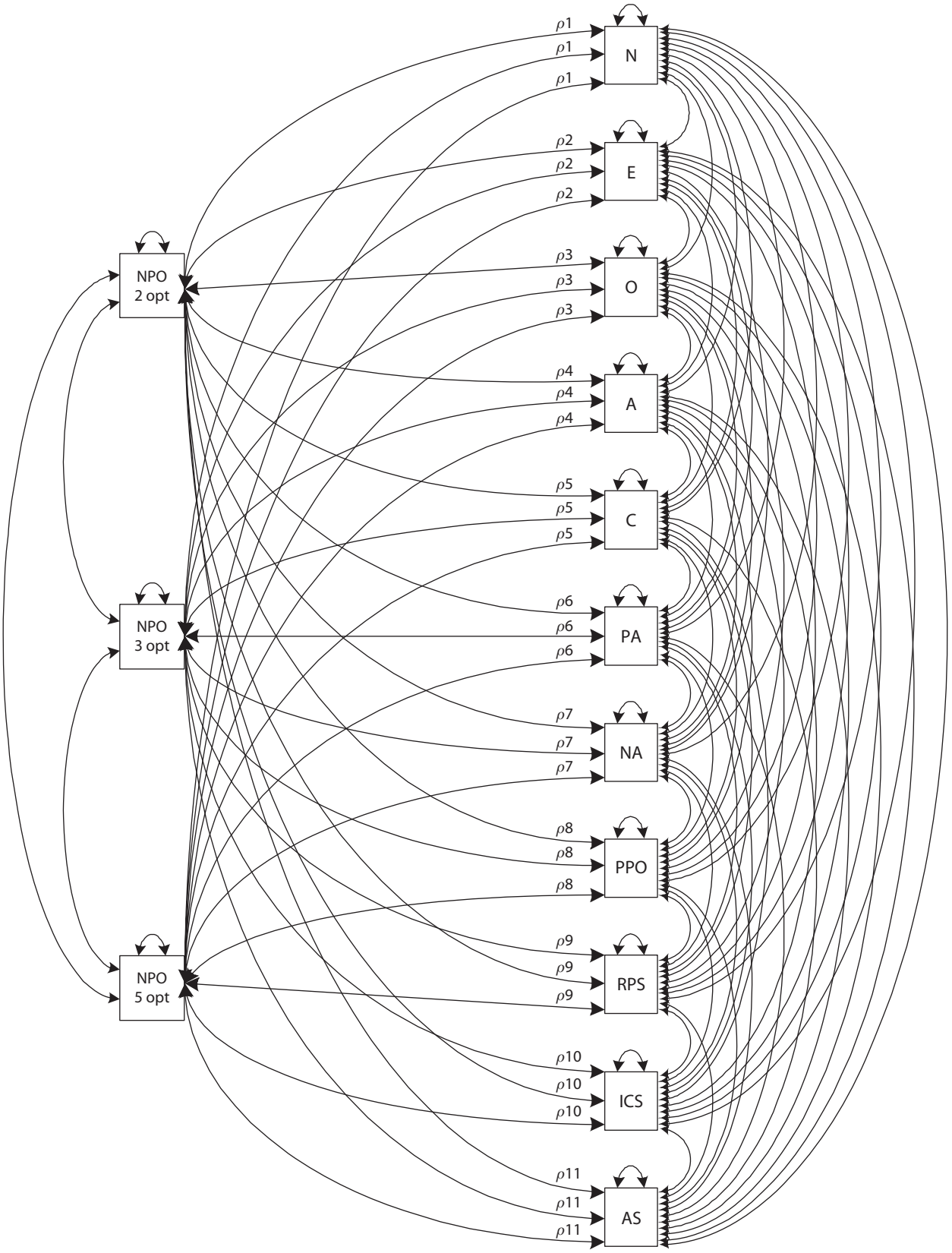

Figure 1. SEM model to assess differential predictive validity across response formats for NPO under classical test theory (Sample A). The same model was used in Sample B, with the SWLS as the target questionnaire.

Table 3
Reliability Coefficients (Internal Consistency) for Negative Problem Orientation (NPO) and Satisfaction With Life Scale (SWLS) Under CTT, IFA, and IRT Models

Response Format	Classical Test Theory (CTT)				Item Factor Analysis (IFA)				Item Response Theory (IRT)			
	NPO		SWLS		NPO		SWLS		NPO		SWLS	
	α	CI	α	CI	ω	CI	ω	CI	ω	CI	ω	CI
2	.78	.75–.81	.62	.56–.69	.78	.71–.86	.62	.55–.69	.89	.87–.91	.80	.75–.86
3	.84	.81–.86	.71	.65–.76	.84	.80–.87	.71	.66–.77	.89	.87–.91	.81	.76–.85
5	.88	.86–.90	.76	.71–.81	.88	.86–.90	.76	.71–.81	.92	.91–.93	.81	.78–.85

Note—N = 746 for NPO; N = 426 for SWLS. α, coefficient alpha; ω, coefficient omega; CI, 99% confidence interval.

Validity Analysis

The goodness of fit of the model for Study A, which is displayed in Figure 2, yielded $\chi^2(22) = 76.87$, $p < .01$ (CFI = .99, TLI = .96, RMSEA = .058). The 99% confidence interval for RMSEA ranged from .03 to .08. The hypothesis that correlations between the target scales and the 11 external questionnaire scores will be equal regardless of the number of response alternatives did not hold exactly, but it held approximately, as assessed by the RMSEA. There was no evident pattern of change in the correlation between the target scales and external scales on the basis of the number of response alternatives, since confidence intervals overlapped for each criterion in all three instances (two, three, and five response alternatives). It is worth noting, however, that there was a slight increase in correlation on average across all criteria as the number of options increased. Average correlations across criteria were computed using Fisher's Z transformation on the absolute correlation value ($\bar{\rho}_{NP02} = .36$, $\bar{\rho}_{NP03} = .37$, and $\bar{\rho}_{NP05} = .40$).

In the case of Study B, the goodness of fit of the model yielded $\chi^2(22) = 38.62$, $p = .02$ (CFI = .99, TLI = .97, RMSEA = .042). The 99% confidence interval for RMSEA ranged from .00 to .07. The hypothesis that correlations between the target scales and the external scores will be equal regardless of the number of response alternatives could not be rejected at the 1% significance level. Again, confidence intervals for the correlations between all of the investigated external scores and forms overlapped. Also, average correlations increased slightly with the number of response alternatives ($\bar{\rho}_{SWLS2} = .24$, $\bar{\rho}_{SWLS3} = .25$, and $\bar{\rho}_{SWLS5} = .27$).

Item Factor Analysis

Preliminary Checks

Table 1 shows the correlations among the NPO2, NPO3, and NPO5 factors using the model that is depicted in Figure 2. They ranged from .92 to .94, and their 99% confidence intervals did not include 1; hence, the use of different response-alternative conditions *might* have had an effect on the reliability and validity of the questionnaires that were under investigation. One would expect the effects to be very small, however, given the magnitude of these correlations. They were indeed higher than those that were computed between the NPO2, NPO3, and NPO5 scales, since the correlations between the factors took into account the unreliability of the measures (McDonald, 1999). Similar results were obtained for the SWLS factors. Their intercorrelations, which are also shown in Table 1, ranged from .93 to .95, and in this case one of the confidence intervals includes 1. As in the case of the NPO, given the magnitude of these correlations, one would expect little differential reliability and validity across forms.

Reliability Analysis

Coefficient omega and its 99% confidence intervals for all NPO and SWLS versions in Studies A and B are displayed in Table 3. McDonald (1999) points out that estimated coefficient omegas and coefficient alphas that are computed from the same data are almost invariably very similar. This was indeed the case here. Confidence intervals for alpha and omega were also very similar. As in the case of CTT, therefore, we observed a significant increase in internal consistency when more response alternatives were employed.

Validity Analysis

The goodness-of-fit results for the restricted model that is depicted in Figure 3 are presented in Table 4. Also shown are the results for its *unrestricted* counterpart where correlations with criteria were not constrained to be equal across forms. A test for comparing these nested models was used to evaluate these constraints across forms [NPO, $\chi^2(22) = 60.95$, $p < .01$, RMSEA = .049]. The 99% confidence interval for RMSEA ranged from .03 to .07. The hypothesis that correlations between the forms and the 11 external scores would be equal across forms did not hold exactly, but it held approximately, as assessed by the RMSEA. For SWLS, however, the hypothesis that correlations with external scores will be equal across forms could not be rejected at the 5% significance level [$\chi^2(22) = 32.37$, $p = .07$]. The point estimate for the RMSEA was .033, and its 99% confidence interval ranged from .00 to .07.

Regarding the evidence from the internal structure when different numbers of response alternatives were employed, Table 5 provides the goodness-of-fit results of applying a one-factor model to the NPO and SWLS questionnaires. For NPO, the one-factor model did not hold even approximately, regardless of the number of response alternatives that were employed. The 99% confidence interval for RMSEA did not include .05 in any case. Furthermore, the fit worsened as the number of response alternatives increased. For SWLS, fit also worsened as the number of response alternatives increased. When two

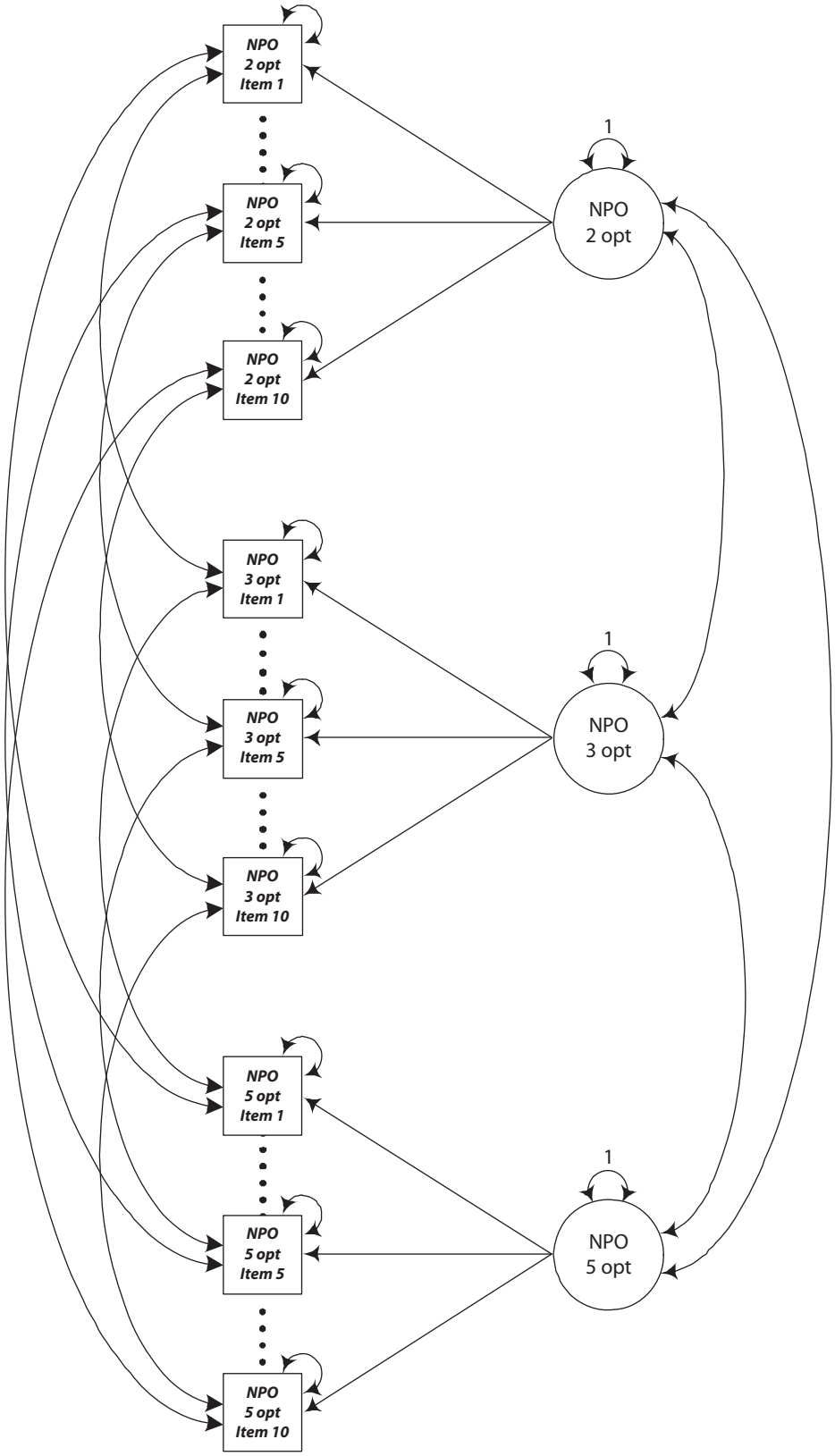

Figure 2. SEM model to assess equality of factors across response formats for NPO under item factor analysis (Sample A). Items were treated as continuous. The same model was used in Sample B, with the SWLS as the target questionnaire. Item response theory models were identical, but items were treated as categorical.

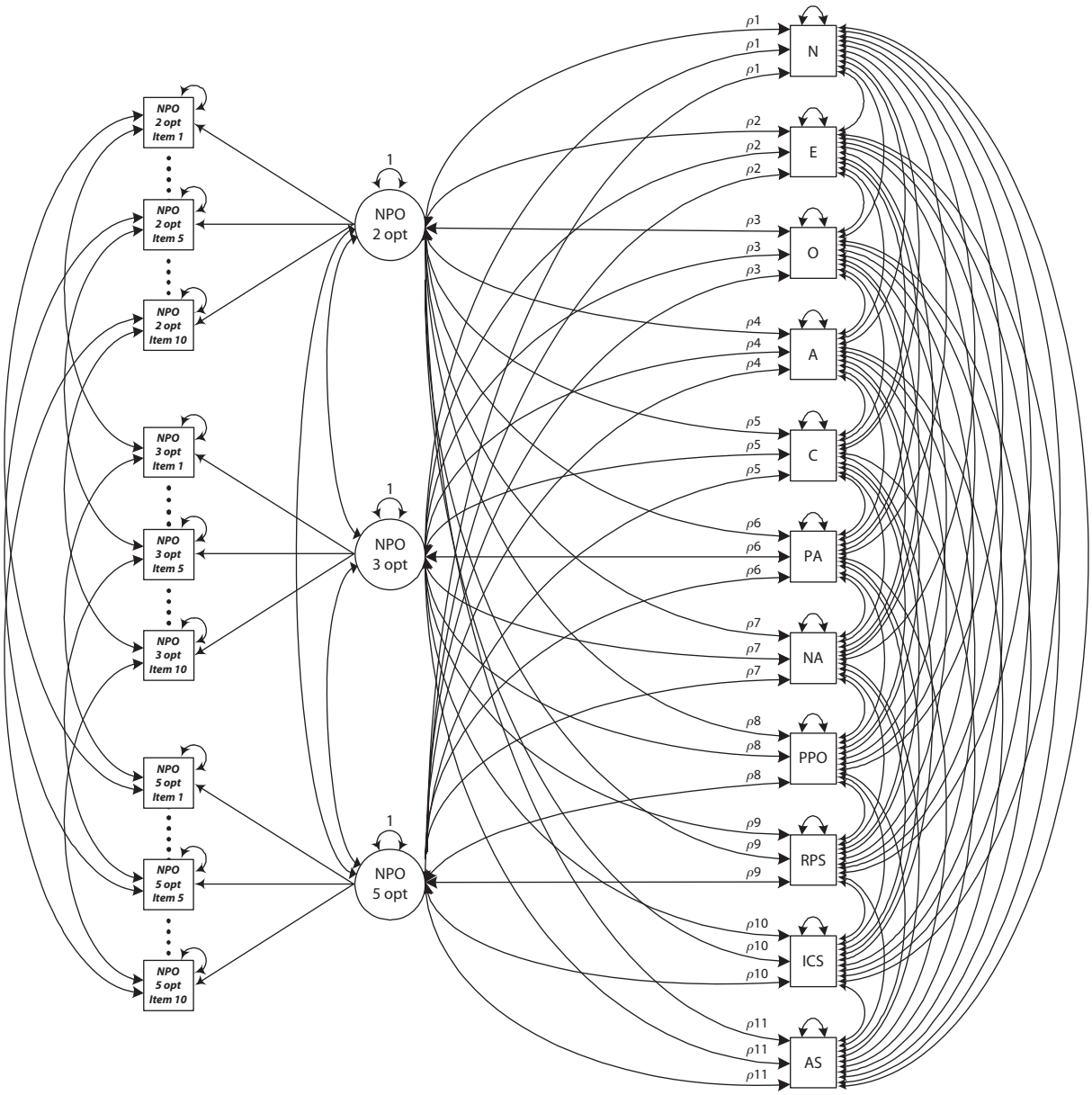

Figure 3. Constrained SEM model to assess predictive validity across response formats for NPO under item factor analysis (Sample A). Items were treated as continuous. The same model was used in Sample B, with the SWLS as the target questionnaire. Item response theory models were identical, but items were treated as categorical.

Table 4
Goodness-of-Fit Tests for Models That Were Used to Assess Differential Predictive Validity Across Forms With Different Types of Response Format Options for Negative Problem Orientation (NPO) and Satisfaction With Life Scale (SWLS)

	Item Factor Analysis (IFA)				Item Response Theory (IRT)			
	NPO		SWLS		NPO		SWLS	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
χ^2	1,574.84	1,637.75	279.12	312.37	6,453.34	6,544.19	382.20	430.27
df	669	691	204	226	669	691	204	226
p	<.01	<.01	<.01	<.01	<.01	<.01	<.01	<.01
RMSEA	.04	.04	.03	.03	.11	.11	.05	.05
CI	.04–.05	.04–.05	.01–.04	.02–.04	.10–.11	.10–.11	.03–.06	.04–.06
CFI	.94	.94	.99	.98	.92	.95	.99	.99
TLI	.93	.93	.98	.98	.90	.94	.98	.97

Note—N = 746 for NPO; N = 426 for SWLS. RMSEA, root mean square error of approximation estimate; CI, 99% confidence interval; CFI, comparative fit index; TLI, Tucker–Lewis index.

Table 5
Structural Validity for Different Types of Response Format Options for Negative Problem Orientation (NPO) and Satisfaction With Life Scale (SWLS) Under Item Factor Analysis (IFA) and Item Response Theory (IRT)

	NPO			SWLS		
	Two Options	Three Options	Five Options	Two Options	Three Options	Five Options
			Item Factor Analysis (IFA)			
χ^2	185.09	223.05	408.08	13.29	12.95	17.25
df	35	35	35	5	5	5
p	<.01	<.01	<.01	.02	.02	<.01
RMSEA	.08	.09	.12	.06	.06	.08
CI	.06–.09	.07–.10	.10–.14	.00–.13	.00–.12	.01–.14
CFI	.89	.90	.86	.96	.98	.97
TLI	.86	.88	.83	.93	.96	.95
			Item Response Theory (IRT)			
χ^2	351.05	363.23	1,784.19	9.21	17.24	24.41
df	35	35	35	5	5	5
p	<.01	<.01	<.01	.10	<.01	<.01
RMSEA	.11	.11	.26	.05	.08	.10
CI	.09–.13	.10–.13	.24–.28	.00–.11	.01–.14	.04–.16
CFI	.92	.95	.92	.99	.99	.99
TLI	.90	.94	.90	.98	.97	.98

Note—N = 746 for NPO; N = 426 for SWLS. RMSEA, root mean square error of approximation estimate; CI, 99% confidence interval; CFI, comparative fit index; TLI, Tucker–Lewis index.

or three response alternatives were employed, the null hypothesis that the one-factor model will hold exactly could not be rejected at the 1% significance level. When five response alternatives were employed, the model held only approximately, as indicated by the confidence interval for RMSEA.

Item Response Theory

Preliminary Checks

The correlations among the NPO2, NPO3, and NPO5 latent traits that were estimated under Samejima's (1969) Graded Response Model, depicted in Figure 2, are shown in Table 3. They ranged from .91 to .94, and their 99% confidence intervals did not include 1; hence, the use of different numbers of response alternatives might have had an effect on the reliability and validity of the questionnaires that were under investigation. One would expect the effects to be very small, however, given the magnitude of these correlations. They were not higher, however, than those that were computed between the NPO2, NPO3, and NPO5 factors in the IFA model.

Similar results were obtained for the SWLS latent traits. Their intercorrelations are also shown in Table 3. They were all .95, and in this case two of the confidence intervals included 1. Given these high correlations, one would expect little, if any, differential reliability and validity across forms.

Reliability Analysis

Coefficient omega and its 99% confidence intervals for the NPO2, NPO3, and NPO5 latent traits in Study A are shown in Table 3. Results from SWLS2, SWLS3, and SWLS5 in Study B are also shown in this table. For NPO, the point estimates for the two- and three-response-alternative conditions were the same (ω = .89). That for the five-response-alternative condition was somewhat higher than, but not substantially different from, those for the other conditions (ω = .92). For NPO, the confidence intervals for the two- and three-response-alternative conditions were the same, but neither overlapped with the confidence interval for the five-response-alternative condition. The results for the SWLS were similar to those for the NPO, but in this case the point estimates for the three- and five-response-alternative conditions were the same (ω = .81). The omega estimate for the two-response-alternative condition was slightly smaller than that for the other conditions (ω = .80), although not substantially. The omega coefficient confidence intervals overlapped for all conditions. Notice that the omega values in IRT were higher than those in the factor analysis model, a result that was to be expected, because the factor model for ordinal variables leads to higher factor loadings than does the linear factor analysis model (Olsson, 1979). As a whole, the results for coefficient omega suggest that the number of response alternatives had no substantial effect on the precision of measurement (as assessed using coefficient omega) for both target instruments under Samejima's (1969) Graded Response Model.

McDonald (1999) points out, however, that coefficient omega in this setting is just an approximation to the value of the information function for the *test score* at the average of the latent-trait distribution (i.e., the point at which the standardized latent trait takes the value of 0). Nevertheless, within IRT, precision of measurement depends on the value of the latent trait. The precision of the maximum likelihood estimator is given by the test information function. Figure 4 depicts the test information functions for both Study A and Study B. These functions provide a different picture of the effects of increasing the number of response alternatives on precision of measurement. Indeed, nonnegligible increments in measurement precision were observed when the number of response alternatives was

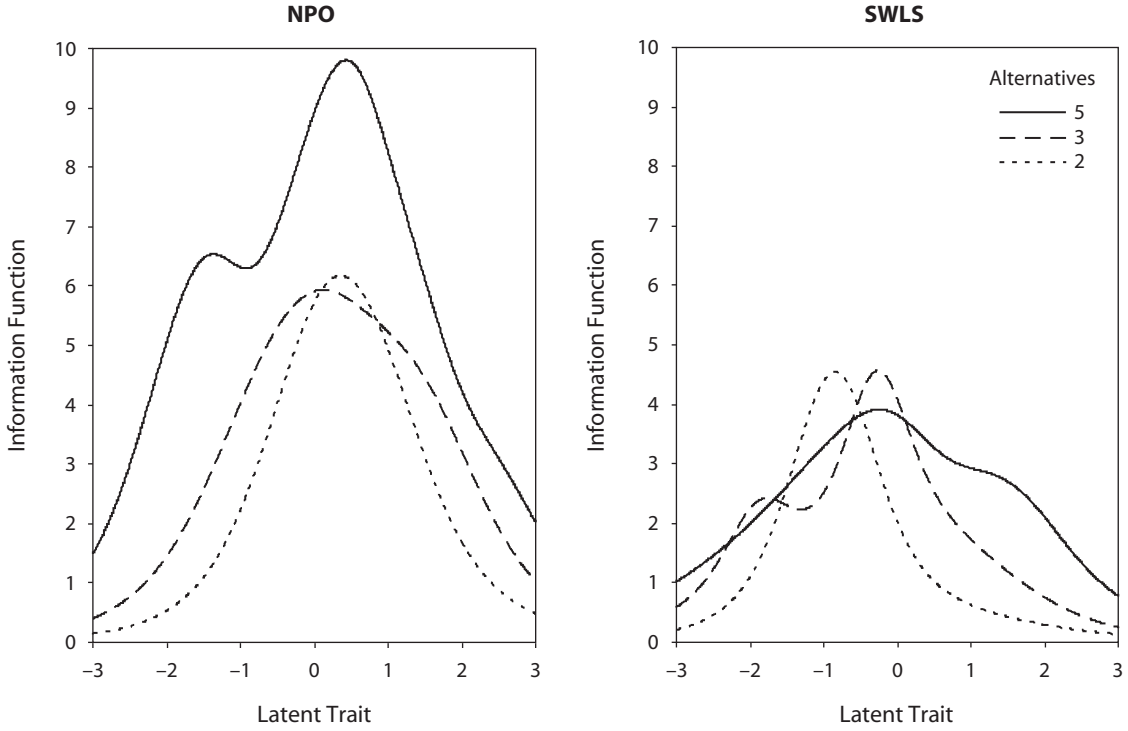

Figure 4. Test information functions for NPO and SWLS, by number of response alternatives.

increased. For NPO, the area under the information function in the latent trait range (−3 to 3) increased by 33% when the number of alternatives was increased from two to three, and by 70% when the number of alternatives was increased from three to five. For SWLS, the area under the information function in the latent trait interval increased by 37% when the number of alternatives was increased from two to three, and by 29% when the number of alternatives was increased from three to five.

Second Step: Validity Analysis

The goodness-of-fit results for the restricted model that is depicted in Figure 3 are presented in Table 4. Also shown in this table are the results for its unrestricted counterpart where correlations with criteria were not constrained to be equal across forms. A test for comparing these nested models was used to evaluate these constraints across forms. For NPO, we obtained $\chi^2(22) = 193.30, p < .01$, RMSEA = .010. The 99% confidence interval for RMSEA ranged from .08 to .12. For SWLS, we obtained $\chi^2(22) = 129.91$, $p = .07$, RMSEA = .011. The 99% confidence interval for RMSEA ranged from .08 to .13. Results suggest that the hypothesis that the correlations between the different forms and the 11 external scores will be equal across forms did not hold even approximately. Yuan and Bentler (2004) cautioned against using nested tests when the base model (i.e., the unrestricted model) is misspecified, because the results of the nested tests may be misleading in these cases. The unrestricted model here was severely misspecified for NPO.

Indeed, when we examined the estimated correlations between the NPO2, NPO3, and NPO5 latent traits and external criteria, we observed that the confidence intervals overlapped in all cases. The same result was observed when examining the correlations between the SWLS2, SWLS3, and SWLS5 latent traits and external criteria. There was no effect of increasing the number of response alternatives on the relations between the latent trait and external scores.

Regarding the internal structure when different numbers of response alternatives were employed, Table 5 provides the goodness-of-fit results of fitting a one-dimensional graded-response model to the NPO and SWLS questionnaires. For NPO, the one-factor model did not hold even approximately, regardless of the number of response alternatives that were employed. The 99% confidence interval for RMSEA did not include .05 in any case. Furthermore, the fit worsened dramatically as the number of response alternatives increased from three to five. For SWLS, fit also worsened as the number of response alternatives increased. When two response alternatives were employed, the null hypothesis that Samejima's (1969) model holds exactly could not be rejected at the 10% significance level. When three or five response alternatives were employed, the model held only approximately, as is indicated by the confidence interval for RMSEA.

DISCUSSION AND CONCLUSIONS

After over 80 years of research, there is no consensus on the optimum number of response alternatives in rating scales. Discrepancies in the results of previous studies can be attributed to the use of different optimality criteria (reliability or validity) as well as different psychometric

models (CTT, IFA, or IRT). We performed two empirical studies investigating the effects of increasing the number of response alternatives on both reliability and validity (convergent and discriminant, as well as evidence that is based on internal structure) using three psychometric models. In each of the studies, a different target instrument was used, varying the number of response alternatives from two to five. Also, the number of items in the target questionnaires was varied to investigate the effects of number of items. Finally, instead of a randomized design, we used a repeated measures design. This was an improvement over previous between-subjects designs, because (1) it enabled us to capture the effects of within-individual variability, and (2) it led to increased precision and higher power. A unified analysis framework was devised to handle the models from a within-subjects design.

The results are summarized in Table 6. Within a CTT framework, our results strongly suggest that increasing the number of response alternatives from two to five resulted in a significant increase in test reliability. The reliability increments were substantial. A 22% increase in reliability was obtained for the shorter test (5-item SWLS) when the number of response alternatives was increased from two to five, whereas a 12% increase was obtained for the longer test (10-item NPO). We would expect higher reliability increments when the test score reliability is small with few response alternatives. Because shorter tests are generally less reliable, increasing the number of response options may be the most beneficial when the questionnaire is shorter.

Increasing the number of response alternatives did not generally increase the convergent or discriminant validity of the test, however. For the longer test, correlations between test scores and external criteria were different in only 2 out of 11 criteria that we examined, and in only one of them was there a substantial (18%) increment in validity. We would expect higher convergent validity increments when the test score variances differ the most between tests that consist of few versus many response alternatives. Because this will occur in longer tests, we would expect to find some validity increases only in long tests.

The reliability estimates and their confidence intervals were identical, for all practical purposes, under CTT and IFA. As a result, the conclusions regarding reliability increments for increasing number of response alternatives that were drawn under CTT can also be drawn under IFA. For convergent validity, we found using IFA, as we did within CTT, that there were no significant differences across response formats in the short test. In the long test, there were significant differences in the relations in only one of the 11 external measures that were investigated. Within an IFA framework, unlike within a CTT framework, it was possible to assess the internal structure of the model (i.e., the goodness of fit of the postulated factor model). We found that increasing the number of response alternatives resulted in a worse fit of the postulated model. Because model data fit is generally worse in longer tests (there are more variables to be modeled), increasing the number of response options may be the most beneficial when the questionnaire is shorter.

Within an IRT framework, we found that increasing the number of response alternatives led to substantial improvement in precision of measurement, defined as the area under the test information function. Increasing the number of response alternatives did not lead to increasing convergent validity, however. Also, increasing the number of response alternatives resulted in a worse fit of the postulated model. This effect was more accentuated in the longer questionnaire. Interestingly, when reliability was assessed using coefficient omega, increasing the number of response alternatives led to only negligible increments in reliability. This probably resulted from the fact that even when two alternatives are used, coefficient omega is rather high; as a result, no substantial improvement can be obtained by increasing the number of alternatives.

Recommendations for Applied Researchers

Researchers usually begin by choosing a psychometric method. Consequently, our recommendations are divided by psychometric model. For applied researchers employing a CTT framework, increasing the number of response alternatives is clearly beneficial. It will lead to reliability increases with no trade-offs. Increasing the number of response alternatives within an IFA framework, however, results in a trade-off: increased reliability but poorer goodness of fit. Thus, researchers need to weigh both aspects when choosing the number of response alternatives to use. If they expect their test scores to be highly reliable (e.g., because the number of items is large) and the goodness of

Table 6
Effects of Increasing the Number of Response Categories From Two to Five on Reliability (Internal Consistency), Predictive Validity, and Structural Validity (Goodness of Fit of the Underlying Model) Under Different Psychometric Models

	Reliability	Predictive Validity	Structural Validity
Classical test theory	Increases, particularly for unreliable tests (e.g., short tests)	Small, almost negligible, effect	Not applicable
Item factor analysis	Increases, particularly for unreliable tests (e.g., short tests)	Small, almost negligible, effect	Worsens, particularly for tests in which the model fits poorly (e.g., long tests)
Item response theory	Area under information function increases (omega does not increase)	Negligible effect	Worsens, particularly for tests in which the model fits poorly (e.g., long tests)

fit of their model is of concern, they might consider using fewer response alternatives. If they expect their model to fit well and they are concerned about the possibly poor reliability of the test score (e.g., because the number of items is small), however, they should use more response alternatives. For applied researchers employing an IRT framework, increasing the number of response alternatives also results in a trade-off between overall precision of measurement and goodness of fit. If they expect their latent trait estimates to be highly reliable (e.g., because the number of items is large, or their items are highly discriminating), but the goodness of fit of their model is of concern, they might consider using fewer response alternatives. If they expect their model to fit well and they are concerned about the possibly poor precision of measurement (e.g., because the number of items is small, or their items show low discrimination), however, they should use more response alternatives.

Note that if a researcher chooses the psychometric framework, it might be beneficial to choose an IRT framework. IRT models simply extract more information from the data. Omega estimates that were computed under the IRT model that was employed here with two response alternatives were larger than the reliability estimates that were computed under IFA with five response alternatives. IRT results must be taken with caution, however, when the model does not fit the data. For instance, the area under the information function for NPO is much larger for five response alternatives than for two and three response alternatives. Goodness of fit that was obtained with five response alternatives (RMSEA = .26) was much worse than that obtained with two and three response alternatives (RMSEA = .11). It is not clear how robust IRT results (such as the test information function) are to model misspecification.

Limitations of the Present Study and Directions for Future Research

The present research is limited in several ways. First, strictly speaking, the conclusions that have been drawn from this research cannot be generalized beyond the scales that were investigated (NPO and SWLS). In particular, further research is needed to investigate the effect of increasing the number of response alternatives in longer tests (e.g., 20 items). Second, only response formats ranging from two to five response alternatives have been investigated. Further research should consider the effects of increasing the number of response alternatives further (e.g., five to nine response alternatives).

Nevertheless, the present study has enabled us to draw some conclusions regarding the optimal number of response alternatives in rating scales. Clearly, the optimal number of response alternatives depends on the researcher's measurement needs. There is a trade-off between reliability (precision) and model fit, which is based on the internal structure, rather than a trade-off between reliability (precision) and convergent and discriminant validity. Convergent and discriminant validity measures were relatively unaffected by the number of response alternatives. If one wishes to maximize reliability (precision), then more response alternatives should be used, and if one wishes to maximize good values for indices that are designed to evaluate internal structure, then fewer response alternatives should be used.

Future research is needed to examine why the evidence that was based on internal structure worsened as the number of response alternatives increased. In particular, for the five-response-alternative IRT model, the internal structure evidence from goodness-of-fit tests was so bad that it was troubling. An interesting finding of our study was that the lack of a perfect correlation among the response-alternative conditions may have resulted not only from the effect of the number of response alternatives, but also from the inconsistency of participants' responses over time. For the SWLS, we assessed within-session test–retest reliability and found less-than-perfect correlations for all three conditions. The small differences we found that were due to the number of response alternatives were, in fact, even smaller because of unreliability over time. Further investigation to separate unreliability over time and the effects of the number of response alternatives would be desirable. Other promising research lines are related to the odd or even number of response alternatives. In our design, all item formats used a central category—except for the dichotomous format, which is arguably a particular item format itself. Further research is needed to investigate the intrasubject effect of a central category.

In sum, we are able to offer some tentative guidelines to applied researchers in choosing the number of response alternatives for rating scales. Our conclusions are limited by the design of our study and by the number of instruments that we investigated. Clearly, additional research is needed to investigate what happens in longer tests, when more than five response alternatives are used, and when the number of response alternatives is even in polytomous items. The test–retest time might also influence the results. On the one hand, the short time between applications was chosen because it would lead to purer effects of number of response alternatives rather than unstableness in responses. On the other hand, the effect of varying the number of response alternatives in longer spaces and their interaction can be of great interest.

One important condition that was not considered in this research concerns the effect of the central category. Extant research has suggested that individuals may respond differently when the number of response alternatives is odd or even (see Weng, 2004). Nevertheless, the usual analysis framework for such a comparison from an SEM perspective differs notably from the method that was used in the present research (see Moustaki & Muircheartaigh, 2002). In Moustaki and Muircheartaigh's approach, the central option of an odd rating scale is expressed either as a function of the main latent trait being assessed, reflecting ambivalence toward the item stem, or as an indicator of a second latent trait (or in some instances, as a latent class), reflecting a "don't know" response that is almost equivalent to a missing response. Because these approaches would have further complicated the analysis, we left them out of the scope of our study, but they represent an important line of research.

Finally, future research using additional scales could add further support to these tentative conclusions. We be-

lieve, however, that the research design that was employed here shows promise of yielding in the near future a definitive answer to the question of what the optimal number of response alternatives is.

AUTHOR NOTE

This research was supported in part by Grant SEJ2006-08204/PSIC from the Spanish Ministry of Education to A.M.-O. Correspondence concerning this article should be addressed to A. Maydeu-Olivares, Faculty of Psychology, University of Barcelona, Passeig de la Vall d'Hebron, 171, 08035 Barcelona, Spain (e-mail: amaydeu@ub.edu).

REFERENCES

ALLEN, M. J., & YEN, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (APA, AERA, & NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

BENTLER, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, **107**, 238-246.

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

CHANG, L. (1994). A psychometric evaluation of four-point and six-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, **18**, 205-215.

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

COSTA, P. T., & MCCRAE, R. R. (1992). *Revised NEO Personality Inventory and NEO Five-Factor Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

COSTA, P. T., & MCCRAE, R. R. (1999). *Inventario de Personalidad NEO revisado* (*NEO PI–R) e Inventario NEO reducido de Cinco Factores (NEO–FFI): Manual profesional*. Madrid: Tea Ediciones.

COX, E. P., III (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, **17**, 407-422.

CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297-334.

DIENER, E., EMMONS, R. A., LARSEN, R. J., & GRIFFIN, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, **49**, 71-75.

D'ZURILLA, T. J., NEZU, A. M., & MAYDEU-OLIVARES, A. (2002). *The Social Problem-Solving Inventory–Revised (SPSI–R): Technical manual*. North Tonawanda, NY: Multi-Health Systems.

GULLIKSEN, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.

JÖRESKOG, K. G., & SÖRBOM, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

KRAMP, U. (2006). *Efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad*. Unpublished doctoral dissertation, University of Barcelona.

LORD, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

LORD, F. M., & NOVICK, M. R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.

MAYDEU-OLIVARES, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, **40**, 261-279.

MAYDEU-OLIVARES, A., COFFMAN, D. L., & HARTMANN, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, **12**, 157-176.

MAYDEU-OLIVARES, A., RODRÍGUEZ-FORNELLS, A., GÓMEZ-BENITO, J., & D'ZURILLA, T. J. (2000). Psychometric properties of the Spanish adaptation of the Social Problem-Solving Inventory–Revised (SPSI–R). *Personality & Individual Differences*, **29**, 699-708.

MCCALLUM, D. M., KEITH, B. R., & WIEBE, D. J. (1988). Comparison of response formats for Multidimensional Health Locus of Control Scales: Six levels versus two levels. *Journal of Personality Assessment*, **52**, 732-736.

MCDONALD, R. P. (1999). *Test theory. A unified treatment*. Mahwah, NJ: Erlbaum.

MOUSTAKI, I., & MUIRCHEARTAIGH, C. (2002). Locating "don't know," "no answer" and middle alternatives on an attitude scale: A latent variable approach. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 15-40). Mahwah, NJ: Erlbaum.

MULAIK, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

MUTHÉN, B. [O.] (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**, 115-132.

MUTHÉN, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Newbury Park, CA: Sage.

MUTHÉN, B. [O.], DU TOIT, S. H. C., & SPISIC, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript, University of California, Los Angeles.

OLSSON, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, **14**, 485-500.

PAVOT, W. G., DIENER, E., COLVIN, C. R., & SANDVIK, E. (1991). Further validation of the Satisfaction With Life Scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, **57**, 149-161.

PETER, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, **16**, 6-17.

PRESTON, C. C., & COLMAN, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, **104**, 1-15.

SAMEJIMA, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, **34**, 100.

SANCERNI, M. D., MELIÁ, J. L., & GONZÁLEZ ROMA, V. (1990). Formato de respuesta, fiabilidad y validez, en la medición del conflicto de rol [Response format, reliability, and validity in the measurement of role conflict]. *Psicológica*, **11**, 167-175.

SANDIN, B., CHOROT, P., LOSTAO, L., JOINER, T. E., SANTED, M. A., & VALIENTE, R. M. (1999). Escalas PANAS de afecto positivo y negativo: Validación factorial y convergencia transcultural [The PANAS Scales of Positive and Negative Affect: Factor analytic validation and cross-cultural convergence]. *Psicothema*, **11**, 37-51.

SATORRA, A., & BENTLER, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

SATORRA, A., & BENTLER, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, **66**, 507-514.

STEIGER, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, **25**, 173-180.

STEIGER, J. H., & LIND, J. M. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

TUCKER, L. R., & LEWIS, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, **38**, 1-10.

WATSON, D., CLARK, L. A., & TELLEGEN, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality & Social Psychology*, **54**, 1063-1070.

WENG, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. *Educational & Psychological Measurement*, **64**, 956-972.

YUAN, K.-H., & BENTLER, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational & Psychological Measurement*, **64**, 737-757.

(Manuscript received May 28, 2008;
revision accepted for publication October 6, 2008.)