Journal of Psychopathology and Behavioral Assessment, Vol. 28, No. 3, September 2006 (© 2006)
DOI: 10.1007/s10862-005-9004-7

Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis

Carol M. Woods[1]

Published online: 6 July 2006

Many self-report measures include some items worded in the direction opposite to that of other items. These so-called reverse-worded (RW) items can reduce the reliability and validity of a scale, and frequently form a separate method factor that does not appear to be substantively meaningful. One possible explanation for factors defined by RW items is respondent carelessness. The purpose of the present study was to evaluate whether relatively few careless responders to RW items can influence confirmatory-factor-analysis model fit enough that researchers would likely reject a one-factor model for a unidimensional scale. Results based on simulations indicated that if at least about 10% of participants respond to RW items carelessly, researchers are likely to reject a one-factor model for a unidimensional scale.

KEY WORDS Reverse-scored; reverse-worded; item wording; factor analysis; careless responding .

Many self-report measures of attitudes, beliefs, personality, and pathology include some items worded opposite to that of others. For example, 17 of the true–false items on the Fear of Negative Evaluation scale (FNE; Watson & Friend, 1969) are worded straightforwardly such as, "I am afraid that people will find fault with me," and "I am afraid that others will not approve of me," whereas 13 items are worded in the opposite direction; for example, "I rarely worry about seeming foolish to others," and "I am often indifferent to the opinions others have of me." Usually, it is expected that reverse-worded (RW) items measure the same construct as straightforwardly worded (SFW) items, with their purported value being to reduce or detect the tendency for respondents to agree more than disagree (acquiescence bias), or respond according to their general feeling about the topic rather than the specific content of the items (a response set).

However, RW items can be problematic. They can reduce the internal consistency, reliability, and validity of a scale, and frequently form a separate "method factor" that does not appear to be substantively meaningful (Barnette, 2000; Benson, 1987; Conrad et al., 2004; Greenberger, Chen, Dmitrieva, & Farruggia, 2003; Knight, Chisholm, Marsh, & Godfrey, 1988; Lai, 1994; Marsh, 1986, 1996; Motl, Conroy, & Horan, 2000; Pilotte & Gable, 1990; Rodebaugh et al., 2004; Rodebaugh, Woods, Heimberg, Liebowitz, & Schneier, in press; Schriesheim & Eisenbach, 1995; Schriesheim, Eisenbach & Hill, 1991; Schriesheim & Hill, 1981; Spector, van Katwyk, Brannick & Chen, 1997; Tomas & Oliver, 1999; Woods & Rodebaugh, 2005).

In studies employing confirmatory factor analysis (CFA), scales with RW items that are expected to be unidimensional are often represented poorly by a one-factor model. Instead, a model with two factors defined by item wording, method factors in addition to trait factors, or correlated errors among items with the same wording type have been preferred (Greenberger et al., 2003; Marsh, 1986, 1996; Motl et al., 2000; Rodebaugh et al., 2004; Rodebaugh et al., in press; Schriesheim & Eisenbach, 1995; Woods & Rodebaugh, 2005). As an example, Rodebaugh et al. (2004) found that a CFA model with two factors corresponding to item wording fit FNE data better than a one-factor model (see also Woods & Rodebaugh, 2005). The same was true for the Brief-FNE (Rodebaugh et al., 2004) and the Social Interaction Anxiety Inventory (Rodebaugh et al., in press). For all 3 measures, validity analyses indicated that scores on SFW items were more related than scores on RW items to other measures of social anxiety.

[1]Psychology Department, Campus Box 1125, Washington University in St. Louis, St. Louis, MO, 63130; e-mail: cwoods@artsci.wustl.edu.

One possible explanation for factors defined by RW items is respondent carelessness. Schmitt and Stults (1985) suggested that a careless respondent may read a few items on a scale, infer what is being measured, and respond the same way to all items without noticing that some are worded in the opposite direction. These authors used exploratory principle components analysis (PCA) with varimax rotation to evaluate whether the presence of careless responding can create a RW "factor". (The term factor usually refers to a latent variable, but PCA is a variance reduction technique that does not involve latent variables.) Schmitt and Stults generated responses to 30 ordinal items for 400 simulees with 4, 8, or 12 items treated as RW. Data sets were created wherein 0, 5, 10, 15, or 20% of simulees responded carelessly to RW items: The intended response was reversed to the other end of the response scale. Results showed that a principle component composed of only RW items was created when as few as 10% of respondents were careless, and as few as 4 in 30 items were RW. The size and strength of the RW component increased as the number of RW items, or the number of careless respondents, increased.

There is ample evidence that researchers employing CFA find poor-fitting one-factor models that improve when method variance among RW items is modeled in some way. These results could be created by the presence of a small number of careless respondents. The findings of Schmitt and Stults (1985) are based on exploratory PCA which does not involve latent variables or prior-hypothesized models, and is designed for continuous rather than ordinal item data. The general finding of Schmitt and Stults (1985) may generalize to factor analysis (wherein a "factor" refers to a latent variable), a confirmatory context, and also to factoring methods appropriate for categorical items. The purpose of the present study is to evaluate whether relatively few careless responders to RW items can influence CFA model fit enough that researchers would likely reject a one-factor model for a unidimensional scale.

METHOD

Simulation Design

A simulation study was carried out so that item response patterns evincing carelessness on RW items could be created at a known rate. A C++ program was written to generate responses to 23 binary items on the basis of the unidimensional two-parameter logistic (2PL, Birnbaum, 1968) item response theory model. True item parameters were randomly drawn from a normal distribution for each of 1,000 replications (μ=1.7, σ=0.8 for discrimination parameters, μ=0, σ=1 for threshold parameters), and values of the latent variable were drawn from a standard normal distribution. In all conditions, 10 items were considered RW and the remaining 13 were considered SFW. A careless respondent was simulated by switching 0–1 and 1–0 for the 10 RW items. Careless respondents were created at a frequency of 0, 5, 10, 20, or 30% for three different sample sizes: 250, 500, or 1,000.

For each of 15 conditions (five percentages × three sample sizes), a one-factor CFA model was fitted to each of the 1,000 data sets using robust weighted least-squares estimation based on a matrix of tetrachoric correlations ("WLSVM," see Muthén, 1998–2004 and Muthén & Muthén, 1998–2004). Additionally, a two-factor CFA model, with factors defined by item wording, was fitted to each data set. The two-factor model is a possible alternative a researcher might hypothesize in advance, or turn to if the one-factor model fit the data poorly. The external Monte Carlo facility of the Mplus program (version 3.12, Muthén & Muthén, 1998–2004) was used for the CFA. Of primary interest in this research are the global model fit indices, which were averaged over the 1,000 replications for each of the 15 conditions.

Factor-Model Fit

Fit indices used were the: (a) Tucker–Lewis Incremental Fit Index (TLI; Tucker–Lewis, 1973), (b) comparative fit index (CFI, Bentler, 1990), (c) root mean-square error of approximation (RMSEA; Browne & Cudeck, 1992; Steiger & Lind, 1980), (d) standardized root mean-square residual (SRMR; Bentler, 1995; Jöreskog & Sörbom, 1981), and the (e) weighted root mean square residual (WRMR; Muthén, 1998–2004).

Guidelines for interpretation of these indices have been discussed by Hu and Bentler (1999) and evaluated for categorical outcomes by Yu and Muthén (2001, as cited in Muthén, 1998–2004). The TLI, also known as the non-normed fit index (Bentler & Bonett, 1980), typically ranges from 0 to 1 (values above 1 are possible but rare), with larger values indicating better fit. Conceptually, the TLI indicates where the fitted model lies on a continuum between two hypothetical models: A baseline model with observed variables unrelated, and an ideal model that fits perfectly. The CFI is conceptually a comparison between the fitted model and the baseline model. It ranges from 0 to 1, with fit improving as CFI approaches 1. Hu and Bentler's (1999) suggestion that TLI and CFI should be at least about .95 for good fit has been found reasonable also for categorical outcomes (Yu & Muthén, as cited in Muthén, 1998–2004).

Table I. Indices of Model Fit (Averaged Over 1,000 Replications) for the One-Factor Model

	Fit index				
% Careless respondents	CFI	TLI	RMSEA	SRMR	WRMR
0% (n=1,000)	1.00 (<.01)	1.00 (<.01)	.01 (.01)	.04 (<.01)	0.77 (.04)
(n=500)	1.00 (<.01)	1.00 (<.01)	.01 (.01)	.06 (.01)	0.77 (.04)
(n=250)	1.00 (.01)	1.00 (.01)	.01 (.01)	.08 (.01)	0.78 (.05)
5%Ω(n=1,000)	0.96 (.02)	0.98 (.01)	.04 (.01)	.06 (.01)	1.26 (.21)
(n=500)	0.97 (.02)	0.98 (.01)	.04 (.01)	.07 (.01)	1.05 (.13)
(n=250)	0.97 (.02)	0.98 (.01)	.04 (.02)	.09 (.01)	0.95 (.10)
10%Ω(n=1,000)	0.91 (.02)	0.95 (.02)	.07 (.01)	.10 (.02)	1.90 (.32)
(n=500)	0.92 (.03)	0.95 (.02)	.07 (.02)	.10 (.02)	1.47 (.24)
(n=250)	0.93 (.03)	0.95 (.02)	.07 (.02)	.12 (.02)	1.18 (.16)
20%Ωn=1,000	0.82 (.03)	0.87 (.02)	.11 (.02)	.14 (.02)	2.95 (.43)
(n=500)	0.83 (.03)	0.87 (.03)	.11 (.02)	.15 (.02)	2.15 (.32)
(n=250)	0.85 (.04)	0.88 (.03)	.11 (.02)	.16 (.03)	1.66 (.24)
30% (n=1,000)	0.76 (.05)	0.81 (.05)	.14 (.02)	.19 (.03)	3.80 (.51)
(n=500)	0.78 (.05)	0.81 (.05)	.14 (.02)	.19 (.03)	2.77 (.36)
(n=250)	0.79 (.06)	0.82 (.05)	.14 (.02)	.20 (.03)	2.05 (.28)

Note. CFI : Comparative fit index, TLI: Tucker–Lewis incremental fit index, RMSEA: root mean square error of approximation, SRMR: standardized root mean-square residual, WRMR: weighted root mean square residual. The value in parentheses is the empirical standard error from the simulation.

The RMSEA indicates the degree of discrepancy between the model and the data per degree of freedom. It ranges from 0 to very large, with smaller values preferred. Rough guidelines for its interpretation are as follows: Values less than .05 indicate close fit, values between .05 and .08 indicate reasonably good fit, values between .08 and .10 indicate mediocre fit, and values above .10 indicate unacceptable fit (Browne & Cudeck, 1992). Hu and Bentler (1999) recommended that RMSEA no larger than around .06 indicates relatively good fit.

The residuals-based measures, *SRMR* and WRMR, also range from 0 to very large, with smaller values preferred. SRMR values less than about .08 indicate good fit for categorical outcomes provided the sample size is at least 250, and the WRMR should be less than about .90 for good fitting models (Muthén, 1998–2004).

RESULTS

Tables I and II give model fit statistics, averaged over 1,000 replications, for the one-factor and two-factor models, respectively. In conditions with zero careless respondents, both models fit nearly identically and perfectly for all three sample sizes. The two-factor model fits as well as the one-factor model because the estimated correlation between factors is 1.

With 5% of respondents careless, the one-factor model still fits fairly well for all sample sizes, although values for the WRMR are somewhat larger than ideal. Researchers would probably accept the one-factor model as providing satisfactory fit to the data. Perhaps the two-factor model would be fitted only if the researchers hypothesized it in advance, rather than because they were dissatisfied with the one-factor model. Fit is superior for the two-factor model; however, the mean correlation between the factors is large:.83 (for each sample size); thus, the one-factor model would likely be preferred for the sake of parsimony.

With 10% of respondents careless, there is a noticeable decline in fit of the one-factor model for each sample size, which would likely provoke researchers to evaluate alternative models. In contrast, the two-factor model fits quite well at each sample size. The mean correlation between factors is fairly large (.70 for $n = 1{,}000$ or 500; .71 for $n = 250$), but not as large as for the 5% conditions.

With 20% of respondents careless, fit is poor for the one-factor model, but excellent for the two-factor model (mean correlation between factors:.48 for $n = 1{,}000$; .49 for $n = 500$ or 250). With 30% of respondents careless, fit is abysmal for the one-factor model but excellent for the two-factor model (mean correlation between factors:.30 for $n = 1{,}000$; .31 for $n = 500$ or 250). Researchers choosing between the two models would undoubtedly select the two-factor model for these conditions.

Influence of Reducing the Percentage of RW Items

The percentage of RW items was held constant across all simulation conditions (10 RW items in $23 \approx 43\%$), but it was larger than in the simulations by Schmitt and Stults.

Table II. Indices of Model Fit (Averaged Over 1,000 Replications) for the Two-Factor Model

	Fit index				
% Careless respondents	CFI	TLI	RMSEA	SRMR	WRMR
0% (*n*=1,000)	1.00 (<.01)	1.00 (<.01)	.01 (.01)	.04 (<.01)	0.76 (.04)
(*n*=500)	1.00 (<.01)	1.00 (<.01)	.01 (.01)	.06 (.01)	0.77 (.04)
(*n*=250)	1.00 (.01)	1.00 (.01)	.01 (.01)	.08 (.01)	0.78 (.05)
5% (*n*=1,000)	0.99 (.01)	0.99 (.01)	.02 (.01)	.05 (.01)	0.94 (.14)
(*n*=500)	0.99 (.01)	0.99 (.01)	.02 (.01)	.06 (.01)	0.87 (.09)
(*n*=250)	0.99 (.01)	0.99 (.01)	.02 (.01)	.08 (.01)	0.83 (.07)
10% (*n*=1,000)	0.97 (.02)	0.99 (.01)	.03 (.01)	.05 (.01)	1.10 (.20)
(*n*=500)	0.98 (.02)	0.99 (.01)	.03 (.01)	.07 (.01)	0.96 (.13)
(*n*=250)	0.98 (.02)	0.99 (.01)	.03 (.02)	.09 (.01)	0.88 (.09)
20% (*n*=1,000)	0.97 (.02)	0.99 (.01)	.03 (.01)	.06 (.01)	1.19 (.23)
(*n*=500)	0.98 (.02)	0.99 (.01)	.04 (.02)	.07 (.01)	1.02 (.15)
(*n*=250)	0.98 (.02)	0.99 (.01)	.04 (.02)	.09 (.01)	0.93 (.10)
30% (*n*=1,000)	0.99 (.01)	0.99 (.01)	.03 (.01)	.05 (.01)	1.11 (.20)
n=500	0.99 (.01)	0.99 (.01)	.03 (.01)	.07 (.01)	0.98 (.13)
n=250	0.99 (.01)	0.99 (.01)	.03 (.02)	.09 (.01)	0.92 (.09)

Note. CFI: Comparative fit index, TLI: Tucker–Lewis incremental fit index, RMSEA: root mean square error of approximation, SRMR: standardized root mean-square residual, WRMR: weighted root mean square residual. The value in parentheses is the empirical standard error from the simulation.

(The largest percentage they used was 40%.) To evaluate the extent to which CFA results may be influenced by careless responding with fewer than 43% of RW items on the scale, the present simulations were repeated with 6 RW items out of 23 (≈26%), and 3 RW items out of 23 (≈13%).

Results were consistent with those of Schmitt and Stults (1985) and showed what might be expected intuitively. The pattern of results for 10 RW items that was observed in the present research was replicated with six and three RW items, but the decrement in fit of the one-factor model was less extreme as the percentage of RW items decreased. As an example, when 30% of 1,000 respondents were careless, fit of the one-factor model was fairly poor with three RW items [mean index over replications (empirical *SE*) = CFI: .86 (.10), TLI: .93 (.05), RMSEA: .07 (.03), SRMR: .08 (.03), WRMR: 1.92 (0.73)], but not as poor as it was with six RW items [CFI: .72 (.05), TLI: .81 (.05), RMSEA: .12 (.03), SRMR: .15 (.04), WRMR: 3.30 (0.66)]. Fit of the corresponding two-factor models remained excellent (results were similar to those for 10 RW items). Complete results of these simulations are available upon request.

DISCUSSION

The purpose of this study was to evaluate whether relatively few careless responders to RW items can influence CFA model fit enough that researchers would likely reject a one-factor model for a unidimensional scale. Results indicated that, for three different sample sizes, if at least about 10% of participants respond to 10 RW items on a 23-item scale carelessly, researchers are likely to reject the one-factor model. This is the same percentage of careless responders that Schmitt and Stults (1985) concluded could create a RW principle component. In both studies, the cut-off is approximate because percentages between 5 and 10% were not examined. Results also indicated a trend for fit of the one-factor model to worsen with increases in either the percentage of careless responders or the percentage of RW items on the scale.

An alternative two-factor model with separate SFW and RW factors fit the data closely regardless of the percentage of careless responders or the percentage of RW items, with the correlation between factors decreasing as the percentage of careless respondents increased. Thus, although SFW and RW items were generated to measure a single underlying construct, the relationship between the constructs underlying the two types of items became progressively lower in the presence of careless responding. With 30% of respondents careless, the correlation between the two constructs, which would be perfectly correlated in the absence of careless responding, was only .3.

These results imply that if more than a few participants in real life respond carelessly as in this simulation, CFA results are likely to be detrimentally affected by RW items. In fact, a method factor would presumably be detectable any time enough people (perhaps

10% or more) respond systematically aberrantly to enough items (perhaps 13% or more) that differ syntactically from other items on a scale. Thus, the present results are not unique to RW items. The focus of this research is on RW items because there is evidence that these types of items form method factors. Exactly why and how these items differ from other items is more difficult to determine, and probably depends on the particular RW items on a scale.

Some researchers have pointed out distinctions among types of RW items (Schriesheim et al., 1991; Schriesheim & Eisenbach, 1995), describing polar opposite, negated regular, and negated polar opposite variants. For example, the regularly worded item, "He is active in scheduling the work to be done," can be changed to polar opposite: "He is passive in scheduling the work to be done," negated regular: "He is not active in scheduling the work to be done," or negated polar opposite: "He is not passive in scheduling the work to be done" (Schriesheim & Eisenbach, 1995, p. 1182). Though regularly-worded items performed the best in these studies, there were differences among the types of RW items, with negated polar opposite items appearing the least valid.

Regardless of the psychological cause of the method variance, if a factor-analyst fails to model it, alternative models might be arrived at that confuse the substantive issues. Even if method variance is modeled, scoring could become needlessly complicated for otherwise unidimensional scales when RW items are included. For example, without careless responders, the scale simulated in this study could be scored as a sum (or more complicated function) of all the item scores. But with evidence that a one-factor model fits poorly and a two-factor model fits well, two scores (one based on SFW items and another based on RW items) would be needed for each person.

Authors of research on RW items often suggest abandoning RW items altogether (e.g., Barnette, 2000; Marsh, 1996; Schriesheim & Eisenbach, 1995). Barnette (2000) suggested a possibly more valid alternative for deterring or detecting response sets or acquiescence bias: To word all questions in the same fashion, but to reverse the order of the response scale for some items. For example, options for item 1 would range from strongly disagree to strongly agree and options for item 2 would range from strongly agree to strongly disagree. Barnette found that these types of items performed better than RW items; thus, continued evaluation of them is warranted.

A limitation of the present research is that only one type of aberrant responding to RW items was simulated. It is easy to imagine alternative and plausible types of aberrant responding wherein participants are careless to varying degrees, or respond in other haphazard ways because of the often awkward and confusing phrasing of RW items. The identification of a method factor composed of RW items could alert researchers to the presence of some unspecified bias in the data related to RW items. This is one argument for the utility of RW items, but requires that researchers actually use RW items for bias detection.

It may also be useful to identify particular individuals who exhibit aberrant responding. Various indices aimed at detecting aberrant or unusual item response patterns (Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1982; Meijer & Sijtsma, 1995, 2001; Reise, 1995; Reise & Flannery, 1996) could be applied. Perhaps a practical approach to dealing with suspected careless or other systematically biased responding to RW items would be to find an aberrancy-detection tool that identifies enough of the "right" respondents so that when they are excluded, CFA results are simplified.

In future research on RW items, scale construction methods for reducing acquiescence and response sets should be further evaluated, and types of aberrant responding to RW items other than the uniform carelessness simulated here should be evaluated. Though more difficult, it would also be useful to identify which types of aberrancy actually occur in real data, and why people tend to respond differently to (at least some types of) RW items. Prudent researchers using CFA for scales having RW items should be alert to the fact that careless or haphazard responding can influence results, test models that allow for factor structures based on item-wording (see e.g., Marsh, 1996; Motl et al., 2000), and pursue methods for identifying and eliminating deviant responders.

ACKNOWLEDGMENT

I am grateful to Adam R. Hafdahl for insightful comments on a draft of this manuscript.

REFERENCES

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*, 361–370.

Benson, J. (1987). Detecting item bias in affective scales. *Educational and Psychological Measurement, 47*, 55–67.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (eds), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison & Wesley.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230–258.

Conrad, K. J., Wright, B. D., McKnight, P., McFall, A., Fontana, A., & Rosenheck, R. (2004). Comparing traditional and Rasch analyses of the Mississippi PTSD scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement*, *5*, 15–30.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg self-esteem scale: do they matter? *Personality and Individual Differences*, *35*, 1241–1254.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources.

Knight, R. G., Chisholm, B. J., Marsh, N. V., & Godfrey, H. P. (1988). Some normative, reliability, and factor analytic data for the revised UCLA Loneliness Scale. *Journal of Clinical Psychology*, *44*, 203–206.

Lai, J. C. L. (1994). Differential predictive power of the positively versus the negatively worded items of the life orientation test. *Psychological Reports, 75*, 1507–1515.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. British *Journal of Mathematical and Statistical Psychology, 35*, 42–56.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*, 37–49.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810–819.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*, 261–272.

Meijer, R. R., & Sijtsma, K. (2001). Methodological review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.

Motl, R. W., Conroy, D. E., & Horan, P. M. (2000). The social physique anxiety scale: An example of the potential consequence of negatively worded items in factorial validity studies. *Journal of Applied Measurement Special Issue: Constructing variables, 1*, 327–345.

Muthén, B. O. (1998–2004). *Mplus Technical Appendices*. Los Angeles, CA: Muthén & Muthén. Downloaded from www.statmodel.com on June, 2005.

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide,* 3rd ed. Los Angeles, CA: Muthén & Muthén.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*, 603–610.

Reise, S. P. (1995). Scoring method and detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213–229.

Reise, S. P., & Flannery, P. W. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*, 9–26.

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment*, *16*, 169–181. (Contributions of the first two authors are equal.)

Rodebaugh, T. L., Woods, C. M., Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (in press). *The Factor Structure, Item Properties, and Screening Utility of the Social Interaction Anxiety Scale*. Psychological Assessment.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*, 367–373.

Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor- analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management*, *21*, 1177–1193.

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, *51*, 67–78.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*, 1101–1114.

Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, *23*, 659–677.

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

Tomas, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling*, *6*, 84–98.

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

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, *33*, 448–457.

Woods, C. M., & Rodebaugh, T. L. (2005). Factor Structures of the Original (FNE) and Brief (BFNE) Fear of Negative Evaluation Scales: Correction to an Erroneous Footnote. *Psychological Assessment*, *17*, 385–386.