Structural Equation Modeling: A Multidisciplinary Journal, 0: 1–11, 2017
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2017.1286228

Routledge
Taylor & Francis Group

Examining and Controlling for Wording Effect in a Self-Report Measure: A Monte Carlo Simulation Study

Honglei Gu,[1] Zhonglin Wen,[2] and Xitao Fan[3]
[1] *Xinyang Normal University*
[2] *South China Normal University*
[3] *University of Macau*

Wording effect refers to the systematic method variance caused by positive and negative item wordings on a self-report measure. This Monte Carlo simulation study investigated the impact of ignoring wording effect on the reliability and validity estimates of a self-report measure. Four factors were considered in the simulation design: (a) the number of positively and negatively worded items, (b) the loadings on the trait and the wording effect factors, (c) sample size, and (d) the magnitude of population validity coefficient. The findings suggest that the unidimensional model that ignores the negative wording effect would underestimate the composite reliability and criterion-related validity, but overestimate the homogeneity coefficient. The magnitude of relative bias of the composite reliability was generally small and acceptable, whereas the relative bias for the homogeneity coefficient and criterion-related validity coefficient was negatively correlated with the strength of the general trait factor.

Keywords: bifactor model, Monte Carlo simulation, reliability, validity, wording effect

The use of both positively and negatively worded items in personality and attitude surveys is commonly suggested for the purpose of reducing or minimizing the potential contamination caused by response set or style, such as acquiescence, affirmation, agreement bias, and so on (Weijters, Baumgartner, & Schillewaert, 2013). However, research has shown that positively and negatively worded items are not necessarily psychometrically interchangeable. Furthermore, greater cognitive efforts are needed to respond to negatively worded items (e.g., Marsh, 1996; Sliter & Zickar, 2014). In addition, the inclusion of negatively worded items might lead to some unintended method effect associated with the item wording (positive vs. negative wording effects), which might result in spurious covariances among the items.

The nature of such wording effect has been debated for a long time. Some researchers have considered such effect as irrelevant or "noise" variance that should be removed and controlled in analyses, or that could be modeled as correlated uniquenesses among the indicators (e.g., Marsh, 1996). Other researchers have conceptualized the wording effect as reflecting substantive personality trait or response style that could be modeled as method factors (DiStefano & Motl, 2006; Motl & DiStefano, 2002; Quilty, Oakman, & Risko, 2006; Tomás & Oliver, 1999), and have discussed the criterion-related validity of wording effect (DiStefano & Motl, 2009; Quilty et al., 2006); its convergent validity across different methods, cultures, and instruments (Michaelides, Koutsogiorgi, & Panayiotou, 2016; Tomás, Oliver, Galiana, Sancho, & Lila, 2013); its longitudinal stability (Gana et al., 2013; Marsh, Scalas, & Nagengast, 2010); its heritability (Alessandri et al., 2010), and even its neural mechanisms (Wang, Kong, Huang, & Liu, 2016).

Wording effect has been found in many personality measures, such as the Rosenberg Self-Esteem Scale (RSES; DiStefano & Motl, 2006; Marsh et al., 2010), the General Health Questionnaire (GHQ–12; Ye, 2009), the Erikson Psychosocial Stage Inventory (EPSI; Schwartz, Zamboanga, Wang, & Olthuis, 2009), the Social Physique Anxiety Scale (SPAS; Motl, Conroy, & Horan, 2000), the Inventory of Callous–Unemotional Traits (ICU; Paiva-Salisbury, Gill, & Stickle, 2016; Ray, Frick, Thornton,

Correspondence should be addressed to Professor Zhonglin Wen, School of Psychology, South China Normal University, Guangzhou, 510631, China. E-mail: wenzl@scnu.edu.cn

Steinberg, & Cauffman, 2016), the Occupational Personality Questionnaire (OPQ; McLarnon, Goffin, Schneider, & Johnston, 2016), the Social Dominance Orientation scale (SDO; Xin & Chi, 2010), the Loneliness Questionnaire (LQ; Ebesutani et al., 2012), and so on. In research practice, however, researchers often ignore the potential wording effect introduced by positive or negative items on a self-report measure. As a result, a researcher might either construct a unidimensional model (i.e., the target construct as the sole latent variable with the items or item parcels as its indicators), or calculate the total score (or average score) of the whole scale for further analyses (e.g., correlational analysis between the scale score and other variables). This practice of ignoring the potential wording effect of a self-report measure could have unintended measurement and statistical consequences. As shown in Gu, Wen, and Fan (2015), ignoring the possible wording effect of a self-report measure might not only lead to biased estimates of measurement reliability, but also to biased estimates of the relationships between the measured trait and other variables (e.g., criterion-related validity), resulting in misleading conclusions. At this time, however, there have been no studies that systematically examined the potential consequences introduced by ignoring wording effect in a self-report measure.

Accordingly, this study intended to fill in this gap in the research literature by examining the following questions:

1. What impact could such wording effect have on measurement reliability estimation?
2. What impact could such wording effect have on measurement validity estimation?

We designed a Monte Carlo simulation experiment to study the measurement and statistical issues introduced by positive or negative wording in a self-report measure. Here, we would like to clarify that the focus of this study was on the potential method effect introduced by negatively worded items, and the methodological consequences of such a method effect if it was ignored in research practice. As discussed in Dalal and Carter (2015), there are two types of negatively worded items: *polar opposite* items (e.g., "I am always on time" vs. "I am always late") and *negated regular* items (e.g., "I am always on time" vs. "I am not always on time"). In the former situation (i.e., polar opposite items), both items are positively worded, but the first is positively keyed, and the second is negatively keyed. The two items are designed to measure the opposite ends of a pole. In the latter situation (i.e., negated regular items), the first item is positively worded, whereas the second item is truly negatively worded (i.e., involving negative words of phrasing), and the second item is intended to be the opposite of the first item. In research practice, the distinction between these two types of items (polar opposite items vs. negated regular items) is not always clear, and both types are considered negatively worded items in a broad sense. In research practice, these two types of items can be used interchangeably (e.g., Lindwall et al., 2012). For our study, our intention was to focus on the second type of negatively worded items (i.e., negated regular items; see earlier discussion). Because our study is a simulation experiment, not a real survey, this issue is actually not critical.

This article is organized as follows. In the next section, we briefly introduced the statistical model of wording effects. We then present the design of the simulation, followed by the results of the study. Finally, for the benefit of applied researchers, we also provide a decision tree and a step-by-step procedure for examining and controlling this method effect.

STATISTICAL MODEL OF WORDING EFFECTS

From the modeling perspective, measures with potential wording effect introduced by positive versus negative wording in a self-report measure could be modeled by using a variation of the bifactor model. More specifically, it could be hypothesized that both the general factor and group factors exist to account for the covariation of the items (Chen, West, & Sousa, 2006; Reise, 2012). For measures with possible wording effect, the targeted trait could be considered as the general factor, which explains the variation among all the items. The positive and negative wording effect could be accommodated by one or two group factors accounting for the method variance introduced by the positively and negatively worded items (Donnellan, Ackerman, & Brecheen, 2016; McKay, Boduszek, & Harvey, 2014; Vecchione, Alessandri, Caprara, & Tisak, 2014). Empirical research (e.g., DiStefano & Motl, 2009; Gu et al., 2015; Tomás et al., 2013) indicates that, in a bifactor model, it is sufficient to use one group factor (e.g., a factor representing the negative wording effect) instead of two (i.e., both positive and negative wording factors).

Without loss of generalization, we assume that a self-report measure is composed of four negatively worded items (x_1–x_4) and four positively worded items (x_5–x_8), which measures a general trait factor (G) and a specific method factor associated with the negative item wording (S); as such, the measurement model can be expressed as the following equations:

$$\begin{aligned} &x_1 = \lambda_{g1}G + \lambda_{s1}S + \delta_1, \quad x_2 = \lambda_{g2}G + \lambda_{s2}S + \delta_2, \\ &x_3 = \lambda_{g3}G + \lambda_{s3}S + \delta_3, \quad x_4 = \lambda_{g4}G + \lambda_{s4}S + \delta_4, \\ &x_5 = \lambda_{g5}G + \delta_5, \quad x_6 = \lambda_{g6}G + \delta_6, \\ &x_7 = \lambda_{g7}G + \delta_7, \quad x_8 = \lambda_{g8}G + \delta_8 \end{aligned} \tag{1}$$

where λ_{g1} is the factor loading of x_1 on the general factor G, λ_{s1} is the factor loading of x_1 on the negative group factor S, δ_1 is the residual term of x_1, and so on. For identification of the latent variable model and easier interpretation, the general and group factors are also hypothesized to be orthogonal (Chen et al., 2006). Thus, the item variance can be decomposed into the variance explained by the general factor (trait variance), the variance explained by the group factor (method variance), and the residual variance (error variance). Two model-based reliability estimates could be defined. The homogeneity coefficient is the percentage of variance accounted for by the general factor (Revelle & Zinbarg, 2009), and the composite reliability (Brunner & Süb, 2005; Raykov & Grayson, 2003) is the percentage of variance accounted for by both the general and group factors (Bentler, 2009).

The homogeneity coefficient (also called coefficient omega hierarchical, ω_H) can be represented as:

$$\omega_H = \frac{(\sum_{i=1}^{8} \lambda_{gi})^2}{(\sum_{i=1}^{8} \lambda_{gi})^2 + (\sum_{i=1}^{4} \lambda_{si})^2 + \sum_{i=1}^{8} \mathrm{var}(\delta_i).} \qquad (2)$$

The composite reliability (also called internal consistency reliability or coefficient omega, ω) can be expressed as

$$\omega = \frac{(\sum_{i=1}^{8} \lambda_{gi})^2 + (\sum_{i=1}^{4} \lambda_{si})^2}{(\sum_{i=1}^{8} \lambda_{gi})^2 + (\sum_{i=1}^{4} \lambda_{si})^2 + \sum_{i=1}^{8} \mathrm{var}(\delta_i)} \qquad (3)$$

Moreover, the explained common variance (ECV), which provides the estimate for the relative strength of the general to group factors, is defined as follows (Reise, 2012; Rodriguez, Reise, & Haviland, 2016):

$$\mathrm{ECV} = \frac{\sum_{i=1}^{8} \lambda_{gi}^2}{\sum_{i=1}^{8} \lambda_{gi}^2 + \sum_{i=1}^{4} \lambda_{si}^2} \qquad (4)$$

Generally speaking, the higher the ECV, the stronger the general factor relative to the group factor, thus the less contamination the wording effect has on a self-report measure.

DESIGN OF THE SIMULATION STUDY

This simulation study was designed to examine the impact that wording effect could have on the estimation of measurement reliability and validity, if the method factor of wording was ignored. Previous studies (e.g., Dahal & Carter, 2015; Gu et al., 2015) indicated that wording effect was primarily reflected by negatively worded items. For the analytical model to be simulated, we assumed that a self-report measure was composed of both positively and negatively worded items. All items measured a general trait factor (*Trait*), and a specific method factor was associated with the negative-wording items (*NWE*). Moreover, for the purpose of assessing criterion-related validity, a criterion factor (*Criterion*) was specified and measured by three continuous and normally distributed indicators, and the loadings of the indicators on the criterion factor were 0.70. The basic simulated true model is shown in Figure 1 for easy understanding.

The simulation experiment had a 4 × 4 × 2 × 3 factorial design with four design conditions. The four design conditions were fully crossed, creating 96 (4 × 4 × 2 × 3 = 96) unique cell conditions. Within each of the 96 unique cell conditions, two models were fitted to each simulated sample

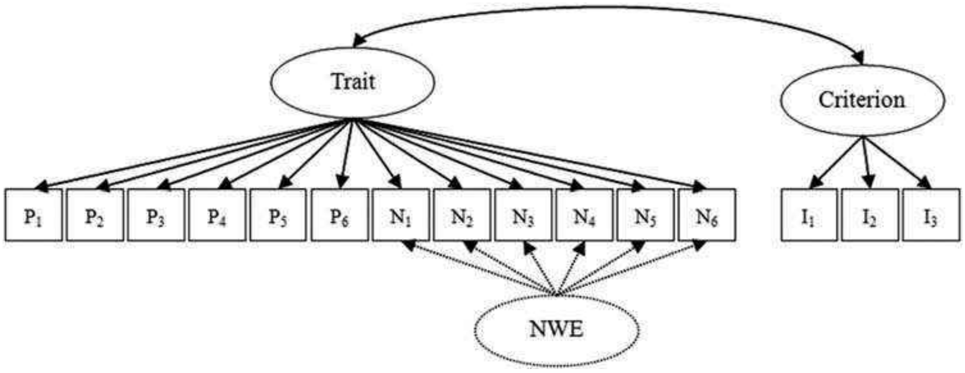

FIGURE 1 Simulated model. *Note*. P_1–P_6 = positively worded items; N_1–N_6 = negatively worded items; Trait = trait factor; NWE = factor for negative wording effect (method factor); Criterion = external criterion factor with three indicators (I_1–I_3). In the simulation experiment, the misspecified model does not include the negative wording effect factor (the component as indicated by dashed lines). For the sake of simplicity, all indicators have residuals that are not shown in the graph.

TABLE 1
Positive to Negative Items (N_p:N_n) on the Simulated Self-Report Measure

	Total Number of Items	
Ratio of N_p and N_n	*12*	*18*
1:1	6:6	9:9
2:1	8:4	12:6

data set, a true model (with negative wording effect correctly modeled) and a misspecified model (with negative wording effect ignored). The details of these design conditions and their levels are described next.

1. Numbers of positively and negatively worded items (N_p:N_n). The number of items on the "self-report" measure was either 12 or 18, and the proportion of positively to negatively worded items was either 1:1 or 2:1, so four different combinations were used as shown in Table 1.

2. Item loadings on the general trait factor and specific method factor (λ_g:λ_s) were specified as having the following sets of population values:

a. 0.6:0.6
b. 0.6:0.3
c. 0.3:0.6
d. 0.3:0.3

3. The simulated population validity coefficient (ρ) between the self-report measure and the external criterion variable (factor) were (a) 0.4 and (b) 0, respectively.

4. Simulated sample size conditions were (a) 200, (b) 500, and (c) 1,000.

Within each unique design condition, 500 sample data sets were generated based on a set of specified population parameters. Each simulated sample data set was fitted to two models: (a) the bifactor model (the true model) including both the trait factor and the method factor (i.e., negative wording factor) as shown in Figure 1; and (b) the unidimensional model (the misspecified model, as shown in Figure 1 without the components of the dashed lines) that only has the trait factor without the method factor (i.e., negative wording effect factor).

The factor loadings and the number of positively and negatively worded items might both have an effect on the strength of the general trait factor relative to the method factor, so the ECV value was calculated based on these two factors (see Table 2). In light of Equations 2 and 3, true values of the homogeneity coefficient and composite reliability would be obtained in each treatment.

For each unique combination of the four design conditions (1–4 as just outlined), M*plus* 6.11 was used to generate 500 sample data sets under the true model (see Figure 1, with all the components). For each sample data set generated under the true model, both the true model (i.e., bifactor measurement model with negative wording effect modeled) and the misspecified model (i.e., unidimensional measurement model with negative wording effect ignored) were fitted to the sample data. From both the fitted true and misspecified models, sample estimates of both the homogeneity coefficient (ω_H) and composite reliability (ω) were obtained. Because we were interested in whether the fit indexes could suggest the misspecified unidimensional model, comparative fit index (CFI) and root mean square error of approximation (RMSEA) were also reported. Finally, for criterion-related validity coefficient under both the true model (bifactor measurement model) and the misspecified model (unidimensional measurement model), the relationship between the trait factor of the "self-report" measure and the external factor representing the criterion was estimated for each sample replication.

RESULTS

First, we removed the sample data sets that did not converge or converged on improper solutions (i.e., estimated variance term was negative). We then compared the true and misspecified models in terms of model goodness of fit. Finally, we examined the relative bias of reliability and validity estimates, power, and Type I error rates under both the bifactor (correct) and unidimensional (misspecified) models.

Fully Proper Solutions

The proportion of samples with fully proper solutions for each cell of the design (the estimation procedure converged to a proper solution such that no estimated variance term was negative) is presented in Table 2. For all conditions, all analysis in the unidimensional model resulted in proper solutions. Within the framework of the bifactor model, except when $N = 200$ and loadings on the general and specific factors are 0.3:0.3, the proportions of fully proper solutions were close to 100%. In evaluating goodness of fit and parameter estimates, only samples with fully proper solutions were considered.

Model Fit

As depicted in Table 2, the bifactor model performed much better than the unidimensional model in terms of model fit as indicated by the widely used model fit indexes (i.e., CFI, RMSEA). More specifically, the bifactor model fitted the data well in all the conditions, with CFI being greater than 0.95, and RMSEA being lower than 0.02. In contrast, the unidimensional model performed considerably worse in 13 out of 48 conditions (nearly one fourth), with CFI being lower than 0.90, RMSEA being far above 0.08, or both. Because model misspecification was intentional, poorer model fit for the misspecified unidimensional model was expected, and thus not surprising. Our major research interest was on the potential bias such a misspecified model would cause for reliability and validity estimates, as discussed later.

TABLE 2
Percentage of Fully Proper Solutions and Goodness-of-Fit Indexes

				Bifactor Model			*Unidimensional Model*		
N	N_p:N_n	λ_g:λ_s	*ECV*	*% Fully Proper*	*RMSEA*	*CFI*	*% Fully Proper*	*RMSEA*	*CFI*
200	6:6	0.6:0.3	0.89	94.0	0.015	0.994	100.0	0.032	0.980
		0.6:0.6	0.67	100.0	0.016	0.996	100.0	0.090	0.930
		0.3:0.3	0.67	77.8	0.012	0.971	100.0	0.021	0.932
		0.3:0.6	0.33	94.6	0.014	0.989	100.0	0.034	0.959
	8:4	0.6:0.3	0.92	89.6	0.015	0.993	100.0	0.030	0.980
		0.6:0.6	0.75	100.0	0.016	0.995	100.0	0.107	0.874
		0.3:0.3	0.75	77.0	0.015	0.952	100.0	0.022	0.912
		0.3:0.6	0.43	100.0	0.015	0.980	100.0	0.044	0.900
	9:9	0.6:0.3	0.89	99.0	0.015	0.992	100.0	0.033	0.973
		0.6:0.6	0.67	100.0	0.016	0.995	100.0	0.082	0.917
		0.3:0.3	0.67	91.4	0.013	0.963	100.0	0.022	0.921
		0.3:0.6	0.33	99.8	0.014	0.987	100.0	0.034	0.951
	12:6	0.6:0.3	0.92	99.0	0.015	0.991	100.0	0.031	0.975
		0.6:0.6	0.75	100.0	0.016	0.994	100.0	0.093	0.864
		0.3:0.3	0.75	87.8	0.013	0.951	100.0	0.022	0.905
		0.3:0.6	0.43	100.0	0.015	0.978	100.0	0.042	0.889
500	6:6	0.6:0.3	0.89	100.0	0.009	0.998	100.0	0.033	0.983
		0.6:0.6	0.67	100.0	0.010	0.999	100.0	0.090	0.931
		0.3:0.3	0.67	100.0	0.009	0.986	100.0	0.018	0.955
		0.3:0.6	0.33	100.0	0.009	0.995	100.0	0.034	0.965
	8:4	0.6:0.3	0.92	100.0	0.009	0.997	100.0	0.031	0.983
		0.6:0.6	0.75	100.0	0.009	0.998	100.0	0.108	0.874
		0.3:0.3	0.75	100.0	0.009	0.981	100.0	0.019	0.939
		0.3:0.6	0.43	100.0	0.009	0.992	100.0	0.045	0.905
	9:9	0.6:0.3	0.89	100.0	0.007	0.998	100.0	0.031	0.979
		0.6:0.6	0.67	100.0	0.007	0.999	100.0	0.080	0.921
		0.3:0.3	0.67	99.2	0.007	0.987	100.0	0.017	0.953
		0.3:0.6	0.33	100.0	0.007	0.996	100.0	0.032	0.959
	12:6	0.6:0.3	0.92	100.0	0.007	0.998	100.0	0.029	0.980
		0.6:0.6	0.75	100.0	0.007	0.998	100.0	0.093	0.865
		0.3:0.3	0.75	98.8	0.007	0.984	100.0	0.018	0.938
		0.3:0.6	0.43	100.0	0.007	0.994	100.0	0.041	0.898
1,000	6:6	0.6:0.3	0.89	100.0	0.006	0.999	100.0	0.033	0.984
		0.6:0.6	0.67	100.0	0.006	0.999	100.0	0.091	0.931
		0.3:0.3	0.67	99.4	0.006	0.994	100.0	0.018	0.963
		0.3:0.6	0.33	100.0	0.006	0.998	100.0	0.035	0.966
	8:4	0.6:0.3	0.92	100.0	0.006	0.999	100.0	0.031	0.984
		0.6:0.6	0.75	100.0	0.006	0.999	100.0	0.107	0.875
		0.3:0.3	0.75	99.4	0.006	0.991	100.0	0.019	0.947
		0.3:0.6	0.43	100.0	0.006	0.997	100.0	0.045	0.908
	9:9	0.6:0.3	0.89	100.0	0.005	0.999	100.0	0.031	0.979
		0.6:0.6	0.67	100.0	0.005	0.999	100.0	0.080	0.921
		0.3:0.3	0.67	100.0	0.005	0.994	100.0	0.018	0.957
		0.3:0.6	0.33	100.0	0.005	0.998	100.0	0.032	0.960
	12:6	0.6:0.3	0.92	100.0	0.005	0.999	100.0	0.029	0.981
		0.6:0.6	0.75	100.0	0.005	0.999	100.0	0.093	0.865
		0.3:0.3	0.75	100.0	0.005	0.992	100.0	0.019	0.942
		0.3:0.6	0.43	100.0	0.005	0.997	100.0	0.041	0.899

Note. ECV = explained common variance; % fully proper = percentage of solutions that converged to fully proper solutions (other results are based on fully proper solutions); RMSEA = root mean square error of approximation; CFI = comparative fit index.

Relative Bias of Homogeneity Coefficient and Composite Reliability Estimates

The relative bias of estimation was calculated by subtracting the true value from the average of estimates and then dividing by the true value:

$$Bias(\hat{\theta}) = \frac{\overline{\hat{\theta}} - \theta}{\theta}, \tag{5}$$

where $\overline{\hat{\theta}}$ represents the average of the parameter estimates in each condition and θ is the population parameter. A relative bias (i.e., $Bias(\hat{\theta})$) less than 5% could be considered negligible (Hoogland & Boomsma, 1998), and bias less than 10% could be acceptable (Bandalos, 2002; Reise, Scheines, Widaman, & Haviland, 2013). Because reliability estimation was unrelated to the population validity coefficient value between the Trait factor and the Criterion factor, for presentation clarity and simplicity, in Table 3, we present the relative biases of the homogeneity coefficient and composite reliability for the condition of validity of 0.4.

The relative biases of homogeneity coefficient and composite reliability in the bifactor model were generally less than 5%. Moreover, there were only two cells for the homogeneity coefficient where the relative bias was greater than 5%, but less than 10%. In summary, the estimation of homogeneity coefficient and composite reliability in the bifactor model was accurate without noticeable bias.

As for the unidimensional model, all the relative biases of the homogeneity coefficient were positive, which indicated that the model overestimated the homogeneity of a self-report measure. For about 50% of conditions in Table 3, the relative bias of homogeneity coefficient was very substantial (e.g., considerably larger than 10%). The relative biases of the composite reliability were all negative, and their absolute values were slightly less than those in the bifactor model, but most of them were within 5%, which is generally considered negligible, as discussed earlier.

Sample size showed no observable effect on the relative bias of reliability estimation. As shown in Figure 2, the larger the ECV value, the less the relative bias of homogeneity coefficient in the unidimensional model when N = 200. With an ECV larger than 0.75, the relative bias was at acceptable levels.

Relative Bias of Validity Coefficient Estimation

The relative biases of validity estimation for the conditions of population validity of 0.4 are reported in Table 3. The relative bias of the validity estimates under the correct bifactor model was generally less than 5%. For the misspecified unidimensional model, all the relative biases of validity estimates were negative, indicating general underestimation for the validity coefficient when the wording effect was ignored. Sample size did not appear to have any systematic effect on the relative bias of validity estimation. The larger the ECV value, the less the relative bias of validity coefficient in the misspecified unidimensional model when N = 200 and ρ = 0.4 (see Figure 2). With an ECV less than 0.75, the unidimensional model would severely underestimate the criterion-related validity. When the population validity was 0, the unidimensional model performed almost the same as the correct bifactor model, and the absolute bias was minimal for most of the conditions. (Due to space limitations, we do not present these biases in our tables.)

Power and Type I Error Rate

We also evaluated the statistical power in statistically detecting the validity coefficient (when the population validity coefficient was nonzero) and the Type I error rate (when the population validity coefficient was zero) for both the correct bifactor model and the misspecified unidimensional model (see Table 4). As expected, statistical power increased with sample size. The bifactor model had higher levels of power than the unidimensional model, except when both the general and group factor loadings were 0.3. When the ECV value was less than 0.75, the levels of power of the bifactor model were much higher than those of the unidimensional model. Otherwise, the differences were negligible. Figure 3 showed that the ECV value had significant impact on the power levels of both models: The larger the ECV value, the higher the statistical power for detecting the validity coefficient.

The unidimensional model performed better than the bifactor model in terms of Type I error rates (see Table 4). In general, Type I error rate decreased as the sample size increased. For example, Type I error rates were all acceptable (less than 0.075 is acceptable according to the suggestion of Bradley, 1978; Wu, Wen, Marsh, & Hau, 2013) when N = 500, but those in the bifactor model were larger than 0.09 when N = 200 and ECV was less than 0.67. Type I error rates from the unidimensional model were generally lower than those from the bifactor model, regardless of the ECV value (see Figure 4). The reason for this finding is unclear.

DISCUSSION

In social and psychological measurement, it has been a common practice to use either the correlated-trait and correlated-method (CTCM) model or the correlated-trait and correlated-uniqueness (CTCU) model to model the wording effect, and to determine whether the wording effect was a meaningful and stable response style, or simply meaningless response noise (Weijters et al., 2013). However, many important issues have been ignored, such as the impact of the wording effect on the reliability and validity of a self-

TABLE 3
Relative Bias of Homogeneity Coefficient, Composite Reliability, and Validity

				Bifactor Model			Unidimensional Model		
N	N_p:N_n	λ_g:λ_s	*ECV*	*Bias_H*	*Bias_C*	*Bias_V*	*Bias_H*	*Bias_C*	*Bias_V*
200	6:6	0.6:0.3	0.89	0.1	0.0	0.3	5.6	−0.6	−3.0
		0.6:0.6	0.67	0.0	−0.1	0.5	22.5	−2.1	−22.5
		0.3:0.3	0.67	6.3	0.5	−2.8	29.7	−2.5	−14.5
		0.3:0.6	0.33	1.9	−0.3	−1.5	92.5	−3.6	−50.3
	8:4	0.6:0.3	0.92	−0.1	0.1	1.3	2.2	−0.6	−0.8
		0.6:0.6	0.75	−0.1	0.0	0.3	8.1	−2.7	−15.3
		0.3:0.3	0.75	−1.7	0.2	−0.8	8.1	−2.8	−6.0
		0.3:0.6	0.43	−0.7	−0.3	−0.5	31.5	−8.9	−44.5
	9:9	0.6:0.3	0.89	0.0	0.0	0.0	5.8	−0.4	−3.3
		0.6:0.6	0.67	−0.4	0.0	0.0	23.2	−1.4	−23.0
		0.3:0.3	0.67	1.1	−0.3	−1.8	22.6	−1.9	−15.3
		0.3:0.6	0.33	−0.5	−0.2	−1.8	94.6	−2.7	−50.8
	12:6	0.6:0.3	0.92	0.0	0.0	0.0	2.4	−0.4	−1.3
		0.6:0.6	0.75	0.1	−0.1	0.3	9.2	−1.8	−15.8
		0.3:0.3	0.75	−0.3	−0.3	−0.5	8.9	−2.1	−6.8
		0.3:0.6	0.43	−0.2	−0.5	0.0	34.4	−7.0	−44.5
500	6:6	0.6:0.3	0.89	0.1	0.0	0.5	5.6	−0.6	−3.3
		0.6:0.6	0.67	0.1	−0.1	0.5	22.7	−2.0	−22.8
		0.3:0.3	0.67	−1.3	0.3	0.0	30.1	−2.1	−15.3
		0.3:0.6	0.33	0.5	0.0	0.8	92.8	−3.5	−50.8
	8:4	0.6:0.3	0.92	0.0	0.1	0.5	2.2	−0.6	−1.0
		0.6:0.6	0.75	−0.1	0.0	0.3	8.1	−2.7	−15.8
		0.3:0.3	0.75	−0.6	0.2	1.0	8.7	−2.3	−6.8
		0.3:0.6	0.43	−0.4	−0.2	0.5	31.5	−8.9	−45.8
	9:9	0.6:0.3	0.89	0.0	0.0	1.0	5.9	−0.3	−2.5
		0.6:0.6	0.67	−0.1	0.0	1.0	23.4	−1.3	−22.0
		0.3:0.3	0.67	0.5	0.0	1.0	23.2	−1.4	−14.3
		0.3:0.6	0.33	0.0	0.0	1.3	95.1	−2.5	−49.8
	12:6	0.6:0.3	0.92	0.0	0.0	1.3	2.5	−0.3	−0.5
		0.6:0.6	0.75	0.0	−0.1	1.0	9.2	−1.8	−15.8
		0.3:0.3	0.75	0.0	−0.1	1.7	9.4	−1.6	−6.3
		0.3:0.6	0.43	−0.2	−0.3	1.7	34.7	−6.7	−45.3
1,000	6:6	0.6:0.3	0.89	0.0	0.0	0.0	5.6	−0.6	−3.8
		0.6:0.6	0.67	0.0	−0.1	0.0	22.7	−2.0	−23.5
		0.3:0.3	0.67	6.8	0.0	0.0	30.3	−2.0	−15.5
		0.3:0.6	0.33	0.0	0.0	0.0	92.8	−3.5	−51.3
	8:4	0.6:0.3	0.92	0.0	0.0	−0.2	2.3	−0.5	−1.5
		0.6:0.6	0.75	−0.1	0.0	−0.2	8.1	−2.7	−16.5
		0.3:0.3	0.75	−0.2	0.0	0.0	8.7	−2.3	−7.3
		0.3:0.6	0.43	−0.4	0.0	−0.2	31.7	−8.7	−46.0
	9:9	0.6:0.3	0.89	0.0	0.0	0.3	5.9	−0.3	−3.3
		0.6:0.6	0.67	−0.1	0.0	0.3	23.4	−1.3	−22.5
		0.3:0.3	0.67	0.0	0.0	0.8	23.2	−1.4	−15.0
		0.3:0.6	0.33	0.0	0.0	0.8	95.1	−2.5	−50.0
	12:6	0.6:0.3	0.92	0.0	0.0	0.5	2.5	−0.3	−1.0
		0.6:0.6	0.75	0.0	−0.1	0.5	9.2	−1.8	−16.0
		0.3:0.3	0.75	0.2	−0.1	0.8	9.6	−1.5	−6.8
		0.3:0.6	0.43	0.0	−0.1	0.8	34.7	−6.7	−45.3

Note. ECV = explained common variance; Bias_H = relative bias of homogeneity coefficient; Bias_C = relative bias of composite reliability; Bias_V = relative bias of validity.

report measure. Methodologically, the bifactor model is a special case of the CTCM model: When there is only one trait factor, the CTCM model would be equivalent to the bifactor model. From Equation 3, it is easy to see that the method effect of negative wording would be confounded with the random measurement error if correlated uniquenesses were used to represent the method effect, and that the internal consistency reliability would be significantly underestimated. Moreover, method factors explicitly estimate construct-irrelevant sources of variance, whereas correlated

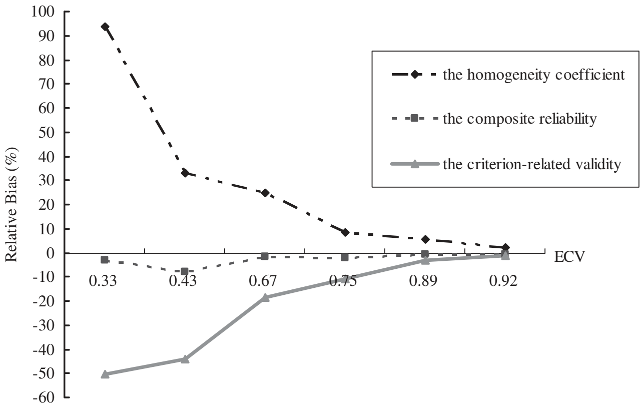

FIGURE 2 The impacts of the explained common variance (ECV) on the homogeneity coefficient, composite reliability, and criterion-related validity of the unidimensional model (N = 200).

TABLE 4
Power and Type I Error Rate

				Bifactor Model		*Unidimensional Model*	
N	N_p:N_n	λ_g:λ_s	*ECV*	*Power*	*Type I Error Rate*	*Power*	*Type I Error Rate*
200	6:6	0.6:0.3	0.89	0.994	0.080	0.996	0.064
		0.6:0.6	0.67	0.996	0.068	0.968	0.066
		0.3:0.3	0.67	0.850	0.094	0.902	0.070
		0.3:0.6	0.33	0.812	0.098	0.624	0.064
	8:4	0.6:0.3	0.92	0.993	0.065	0.998	0.056
		0.6:0.6	0.75	0.998	0.070	0.982	0.056
		0.3:0.3	0.75	0.891	0.083	0.936	0.082
		0.3:0.6	0.43	0.870	0.096	0.660	0.064
	9:9	0.6:0.3	0.89	0.994	0.062	0.996	0.050
		0.6:0.6	0.67	0.994	0.060	0.964	0.064
		0.3:0.3	0.67	0.920	0.084	0.948	0.058
		0.3:0.6	0.33	0.908	0.082	0.622	0.052
	12:6	0.6:0.3	0.92	0.994	0.059	0.996	0.054
		0.6:0.6	0.75	0.994	0.060	0.984	0.050
		0.3:0.3	0.75	0.952	0.078	0.972	0.054
		0.3:0.6	0.43	0.940	0.074	0.722	0.038
500	6:6	0.6:0.3	0.89	1.000	0.054	1.000	0.048
		0.6:0.6	0.67	1.000	0.056	1.000	0.048
		0.3:0.3	0.67	0.998	0.056	1.000	0.046
		0.3:0.6	0.33	0.998	0.046	0.944	0.064
	8:4	0.6:0.3	0.92	1.000	0.047	1.000	0.048
		0.6:0.6	0.75	1.000	0.048	1.000	0.048
		0.3:0.3	0.75	0.998	0.063	1.000	0.068
		0.3:0.6	0.43	1.000	0.052	0.954	0.058
	9:9	0.6:0.3	0.89	1.000	0.050	1.000	0.058
		0.6:0.6	0.67	1.000	0.054	1.000	0.050
		0.3:0.3	0.67	1.000	0.046	1.000	0.034
		0.3:0.6	0.33	1.000	0.030	0.968	0.046
	12:6	0.6:0.3	0.92	1.000	0.060	1.000	0.048
		0.6:0.6	0.75	1.000	0.060	1.000	0.042
		0.3:0.3	0.75	1.000	0.044	1.000	0.042
		0.3:0.6	0.43	1.000	0.044	0.974	0.036
1,000	6:6	0.6:0.3	0.89	1.000	0.050	1.000	0.046
		0.6:0.6	0.67	1.000	0.048	1.000	0.062
		0.3:0.3	0.67	1.000	0.064	1.000	0.058
		0.3:0.6	0.33	1.000	0.062	0.998	0.052
	8:4	0.6:0.3	0.92	1.000	0.058	1.000	0.054
		0.6:0.6	0.75	1.000	0.056	1.000	0.042
		0.3:0.3	0.75	1.000	0.064	1.000	0.060
		0.3:0.6	0.43	1.000	0.054	1.000	0.050
	9:9	0.6:0.3	0.89	1.000	0.042	1.000	0.036
		0.6:0.6	0.67	1.000	0.042	1.000	0.036
		0.3:0.3	0.67	1.000	0.048	1.000	0.046
		0.3:0.6	0.33	1.000	0.044	1.000	0.052
	12:6	0.6:0.3	0.92	1.000	0.038	1.000	0.032
		0.6:0.6	0.75	1.000	0.038	1.000	0.030
		0.3:0.3	0.75	1.000	0.038	1.000	0.042
		0.3:0.6	0.43	1.000	0.040	0.998	0.050

Note. ECV = explained common variance.

uniquenesses simply partial them out, thus bringing no new information to the model (Morin, Arens, & Marsh, 2016; Schweizer, 2012). Because of this, it would be inappropriate to use CTCU models to control the wording effect statistically.

In this study, we examined the impact of the method effect associated with the item wording on the reliability and validity estimates of a self-report measure under different conditions. Results showed that ignoring the wording effect in modeling had little impact on the composite reliability (e.g., the relative biases were less than 5% in general), but would lead to severe overestimation of the homogeneity coefficient (e.g., the relative biases could be substantial, and even as high as 95%). Because the homogeneity coefficient reflects the degree to which the total score is interpretable as a measure of a single common factor (Reise, 2012; Rodriguez et al., 2016), a very low homogeneity coefficient would suggest that scoring and reporting the total score (or average score) of the whole scale would be methodologically questionable. Thus, overestimation of the homogeneity coefficient would mislead researchers to calculate and use the total score of a whole scale for further analyses (e.g., the analysis of mediation and moderation effects involving the total scale score). For the issue of the wording effect on validity coefficient estimation, under certain conditions, ignoring the wording effect would severely underestimate the correlations between the trait variable and an external criterion variable (e.g., relative biases up to 50%), which would lead to increased Type II error rates.

The study also showed that the relative biases of homogeneity coefficient and criterion-related validity coefficient were negatively correlated with the strength of the general trait factor (i.e., ECV); that is, the larger the ECV value, the less impact the wording effect would have on a self-report measure. The ECV was determined by the factor loadings on the general and group factors, and the number of positively and negatively worded items. For a specific balanced scale, researchers could roughly evaluate the strength of the wording effect by comparing the item loadings on the general and group factors. If the negatively worded items' loadings on the general trait factor were far larger than

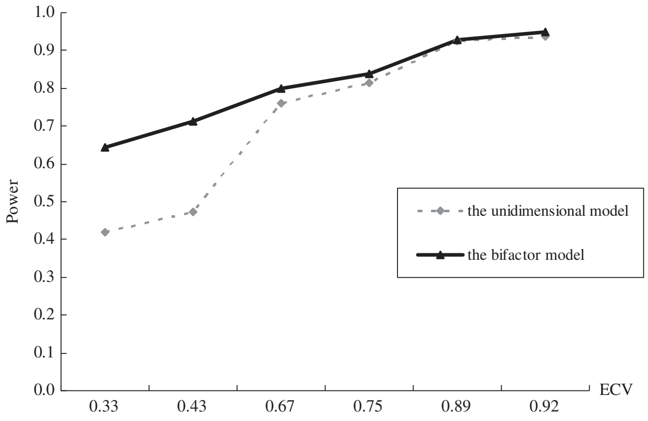

FIGURE 3 The impacts of the explained common variance (ECV) on the power levels of the unidimensional model and the bifactor model (ρ = 0.4, N = 200).

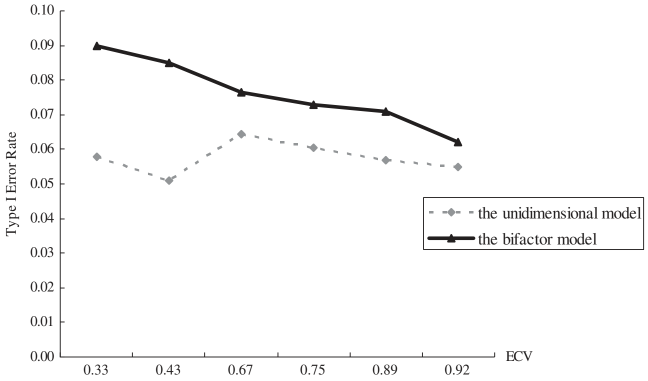

FIGURE 4 The impacts of the explained common variance (ECV) on the Type I error rates of the unidimensional model and the bifactor model (ρ = 0, N = 200).

those on the specific method factor, it would suggest that the negative wording effect could be sufficiently small to be negligible. Otherwise, it would be necessary to include the wording effect as a group factor in the model to obtain unbiased estimates. As noted by Stucky and Edelen (2014), ECV can also be computed at the item level to identify the percentage of item common variance attributable to a general dimension (I-ECV). The negatively worded items with I-ECV values above .80 or .85 should be retained for further analysis.

Because of the additional complexity of the bifactor model relative to the unidimensional model, many applied researchers might prefer to use the unidimensional model for a measure, even though the measure includes both positively and negatively worded items, and they might rely on model fit indexes to inform them about the appropriateness of the unidimensional model. In this situation, however, the information from the commonly used model fit indexes could be misleading. As shown and discussed earlier, under certain conditions (e.g., the ECV value was 0.33), the unidimensional model could fit the data well, with CFI being far above 0.90 and RMSEA being lower than 0.05, but the relative bias of the homogeneity coefficient could still be very large (e.g., up to 95%).

In this study, the simulation model only had one trait factor. If the target construct consists of several related dimensions, and there might be a hybrid of specific domain and method factors representing the group factors, the total covariance among the items can be accounted for by (a) a general factor underlying all the items, (b) a set of group factors where variance over and above the general factor is shared among subsets of items presumed to be highly similar in content, and (c) a group factor accounting for the method variance introduced by the negatively worded items (Gignac, 2010; Gignac, Palmer, & Stough, 2007; Gu, Wen, & Fan, 2017; Paiva-Salisbury et al., 2016). In this case, the ECV could be interpreted as the strength of trait factors (including both the general and specific trait factors) relative to method factors, and it is still reasonable to evaluate the impact of the wording effect based on the magnitude of ECV values.

SUGGESTIONS AND CONCLUDING REMARKS

Should researchers always control for the method effect associated with the item wording? The practice of ignoring the wording effect might overestimate the homogeneity coefficient and underestimate the criterion-related validity, and the degree of such estimation bias is negatively correlated with the ECV. The larger the ECV is, the smaller the bias will be for the reliability and validity estimates.

To facilitate researchers in making appropriate analytical decisions, we provide a decision tree (Figure 5) and the step-by-step procedure for examining and controlling for the wording effect in research. First, researchers who use self-report scales with both positively and negatively worded items should consider the wording effect and fit a bifactor model to the data. Second, the strength of the general trait factor relative to the specific method factor (i.e., the ECV) should be evaluated and reported. If the ECV value is large enough (e.g., > .75), researchers could ignore the impact of the wording effect and fit a unidimensional trait model. Otherwise, researchers should control for wording effect statistically and use the bifactor model for further analyses. On the other hand, researchers might consider reducing the use of negative-wording items, and find ways (e.g., to revise or construct items with stronger loadings on the trait; to select items with I-ECV values above .80) to enhance the dominance of the general trait factor. Third, the homogeneity coefficient (also called coefficient omega hierarchical, ω_H) should be reported based on the

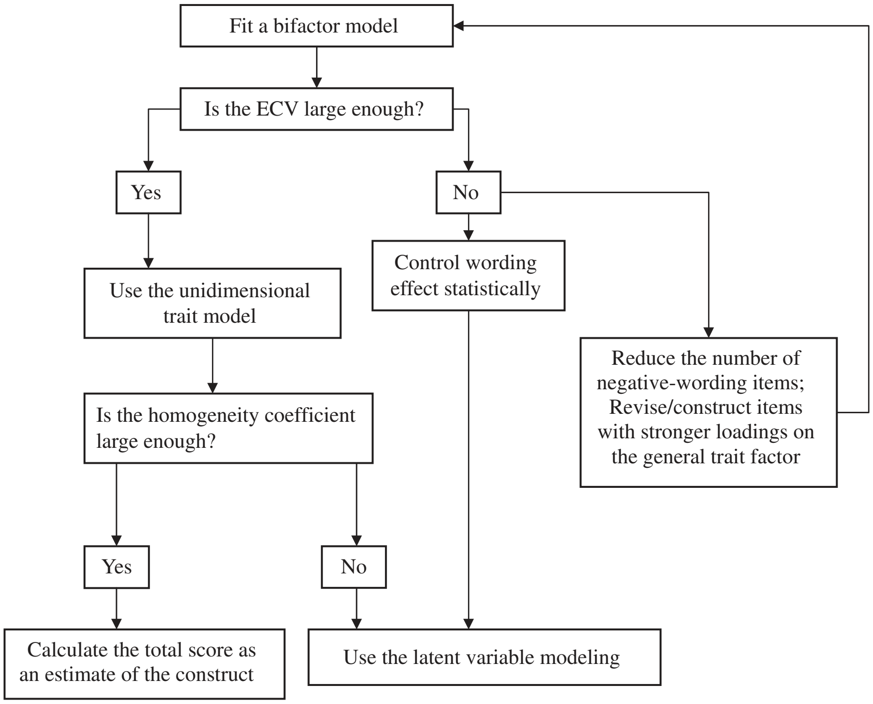
Fit a bifactor model
Is the ECV large enough?
Yes
No
Use the unidimensional trait model
Control wording effect statistically
Reduce the number of negative-wording items; Revise/construct items with stronger loadings on the general trait factor
Is the homogeneity coefficient large enough?
Yes
No
Calculate the total score as an estimate of the construct
Use the latent variable modeling

FIGURE 5 Decision tree for the examination and control of wording effect. *Note*. ECV = explained common variance.

unidimensional trait model. If ω_H is large enough, it is reasonable to consider the total score as an estimate of the construct of interest.

FUNDING

This research was funded by grants from the National Natural Science Foundation of China (31271116, 31400909).

REFERENCES

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P. M., Barbaranelli, C., Medda, E., … Caprara, G. V. (2010). Much more than model fitting? Evidence for the heritability of method effect associated with positively worded items of the Life Orientation Test Revised. *Structural Equation Modeling*, *17*, 642–653. doi:10.1080/10705511.2010.510064

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102. doi:10.1207/S15328007SEM0901_5

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137–143. doi:10.1007/s11336-008-9100-1

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. doi:10.1111/bmsp.1978.31.issue-2

Brunner, M., & Süb, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, *65*, 227–240. doi:10.1177/0013164404268669

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*, 189–225. doi:10.1207/s15327906mbr4102_5

Dalal, D. K., & Carter, N. T. (2015). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 112–132). New York, NY: Routledge.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*, 440–464. doi:10.1207/s15328007sem1303_6

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences*, *46*, 309–313. doi:10.1016/j.paid.2008.10.020

Donnellan, M. B., Ackerman, R. A., & Brecheen, C. (2016). Extending structural analyses of the Rosenberg Self-Esteem Scale to consider criterion-related validity: Can composite self-esteem scores be good enough? *Journal of Personality Assessment*, *98*, 169–177. doi:10.1080/00223891.2015.1058268

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., & Young, J. (2012). The importance of modeling method effects: Resolving the (uni)dimensionality of the Loneliness Questionnaire. *Journal of Personality Assessment*, *94*, 186–195. doi:10.1080/00223891.2011.627967

Gana, K., Saada, Y., Bailly, N., Joulain, M., Hervé, C., & Alaphilippe, D. (2013). Longitudinal factorial invariance of the Rosenberg Self-Esteem Scale: Determining the nature of method effects due to item wording. *Journal of Research in Personality*, *47*, 406–416. doi:10.1016/j.jrp.2013.03.011

Gignac, G. (2010). Seven-factor model of emotional intelligence as measured by Genos EI. *European Journal of Psychological Assessment*, *26*, 309–316. doi:10.1027/1015-5759/a000041

Gignac, G. E., Palmer, B. R., & Stough, C. (2007). A confirmatory factor analytic investigation of the TAS–20: Corroboration of a five-factor model and suggestions for improvement. *Journal of Personality Assessment*, *89*, 247–257. doi:10.1080/00223890701629730

Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the Core Self-Evaluation Scale (CSES): A bi-factor perspective. *Personality and Individual Differences*, *83*, 142–147. doi:10.1016/j.paid.2015.04.006

Gu, H., Wen, Z., & Fan, X. (2017). Structural validity of the Machiavellian Personality Scale: A bifactor exploratory structural equation modeling approach. *Personality and Individual Differences*, *105*, 116–123. doi:10.1016/j.paid.2016.09.042

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–368. doi:10.1177/0049124198026003003

Lindwall, M., Barkoukis, V., Grano, G., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*, 196–204. doi:10.1080/00223891.2011.645936

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. doi:10.1037/0022-3514.70.4.810

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*, 366–381. doi:10.1037/a0019225

McKay, M. T., Boduszek, D., & Harvey, S. A. (2014). The Rosenberg Self-Esteem scale: A bifactor answer to a two-factor question? *Journal of Personality Assessment*, *96*, 654–660. doi:10.1080/00223891.2014.923436

McLarnon, M. J., Goffin, R. D., Schneider, T. J., & Johnston, N. G. (2016). To be or not to be: Exploring the nature of positively and negatively keyed personality items in high-stakes testing. *Journal of Personality Assessment*, *98*, 480–490. doi:10.1080/00223891.2016.1170691

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016). Method effects on an adaptation of the Rosenberg Self-Esteem Scale in Greek and the role of personality traits. *Journal of Personality Assessment*, *98*, 178–188. doi:10.1080/00223891.2015.1089248

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*, 116–239. doi:10.1080/10705511.2014.961800

Motl, R. W., Conroy, D. E., & Horan, P. M. (2000). The Social Physique Anxiety Scale: An example of the potential consequences of negatively worded items in factorial validity studies. *Journal of Applied Measurement*, *1*, 327–345.

Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling*, *9*, 562–578. doi:10.1207/S15328007SEM0904_6

Paiva-Salisbury, M. L., Gill, A. D., & Stickle, T. R. (2016). Isolating trait and method variance in the measurement of callous and unemotional traits. *Assessment* Advanced online publication. doi:10.1177/1073191115624546

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling*, *13*, 99–117. doi:10.1207/s15328007sem1301_5

Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and negative item wording and its influence on the assessment of callous-unemotional traits. *Psychological Assessment*, *28*, 394–404. doi:10.1037/pas0000183

Raykov, T., & Grayson, D. (2003). A test for change of composite reliability in scale development. *Multivariate Behavioral Research*, *38*, 143–159. doi:10.1207/S15327906MBR3802_1

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi:10.1080/00273171.2012.715555

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *1*, 5–26. doi:10.1177/0013164412449831

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154. doi:10.1007/s11336-008-9102-z

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *98*, 223–237.

Schwartz, S. J., Zamboanga, B. L., Wang, W., & Olthuis, J. V. (2009). Measuring identity from an Eriksonian perspective: Two sides of the same coin? *Journal of Personality Assessment*, *91*, 143–154. doi:10.1080/00223890802634266

Schweizer, K. (2012). On correlated errors. *European Journal of Psychological Assessment*, *28*, 1–2. doi:10.1027/1015-5759/a000094

Sliter, K. A., & Zickar, M. J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement*, *74*, 214–226. doi:10.1177/0013164413504584

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183–206). New York, NY: Routledge/Taylor & Francis Group.

Tomás, J. M., & Oliver, A. (1999). Rosenberg's Self-Esteem Scale: Two factors or method effects. *Structural Equation Modeling*, *6*, 84–98. doi:10.1080/10705519909540120

Tomás, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining method effects associated with negatively worded items in trait and state global and domain-specific self-esteem scales. *Structural Equation Modeling*, *20*, 299–313. doi:10.1080/10705511.2013.769394

Vecchione, M., Alessandri, G., Caprara, G. V., & Tisak, J. (2014). Are method effects permanent or ephemeral in nature? The case of the revised life orientation test. *Structural Equation Modeling*, *21*, 117–130. doi:10.1080/10705511.2014.859511

Wang, Y., Kong, F., Huang, L., & Liu, J. (2016). Neural correlates of biased responses: The negative method effect in the Rosenberg Self-Esteem Scale is associated with right amygdala volume. *Journal of Personality*, *84*, 623–632. doi:10.1111/jopy.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*, 320–334. doi:10.1037/a0032121

Wu, Y., Wen, Z., Marsh, H. W., & Hau, K.-T. (2013). A comparison of strategies for forming product indicators for unequal numbers of items in structural equation models of latent interactions. *Structural Equation Modeling*, *20*, 551–567. doi:10.1080/10705511.2013.824772

Xin, Z., & Chi, L. (2010). Wording effect leads to a controversy over the construct of the social dominance orientation scale. *The Journal of Psychology*, *144*, 473–488. doi:10.1080/00223980.2010.496672

Ye, S. (2009). Factor structure of the General Health Questionnaire (GHQ–12): The role of wording effects. *Personality and Individual Differences*, *46*, 197–201. doi:10.1016/j.paid.2008.09.027