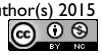
Article

Educational and Psychological Measurement
2016, Vol. 76(3) 357–386



DOI: 10.1177/0013164415596421
epm.sagepub.com


Improving the Factor Structure of Psychological Scales: The Expanded Format as an Alternative to the Likert Scale Format

Xijuan Zhang[1] and Victoria Savalei[1]

Abstract

Many psychological scales written in the Likert format include reverse worded (RW) items in order to control acquiescence bias. However, studies have shown that RW items often contaminate the factor structure of the scale by creating one or more method factors. The present study examines an alternative scale format, called the Expanded format, which replaces each response option in the Likert scale with a full sentence. We hypothesized that this format would result in a cleaner factor structure as compared with the Likert format. We tested this hypothesis on three popular psychological scales: the Rosenberg Self-Esteem scale, the Conscientiousness subscale of the Big Five Inventory, and the Beck Depression Inventory II. Scales in both formats showed comparable reliabilities. However, scales in the Expanded format had better (i.e., lower and more theoretically defensible) dimensionalities than scales in the Likert format, as assessed by both exploratory factor analyses and confirmatory factor analyses. We encourage further study and wider use of the Expanded format, particularly when a scale's dimensionality is of theoretical interest.

Keywords

reverse worded items, item wording, scale format, Likert format, acquiescence bias, method effects

[1]University of British Columbia, Vancouver, British Columbia, Canada

Corresponding Author:
Victoria Savalei, University of British Columbia, 2136 West Mall, Vancouver, British Columbia, Canada V6T 1Z4.
Email: v.savalei@ubc.ca

Many psychological scales in the Likert format[1] contain both positively worded (PW) and reverse worded (RW) items. In this article, we assume that PW items are phrased in the direction of the construct, while RW items are phrased in the opposite direction.[2] There are two main kinds of RW items: (1) negation RW items, created by adding negative particles such as *not* or *no* or by adding affixal negation such as *un-* or *-less* (e.g., *I'm unhappy* on a scale measuring happiness); and (2) polar opposite RW items, created by using words with an opposite meaning (e.g., *I'm sad* on a scale measuring happiness). The main rationale for including RW items on scales is to control acquiescence bias, or the tendency for respondents to endorse the item regardless of its content (Ray, 1983). This response style is to be distinguished from carelessness or confusion, in that the respondent is presumed to have understood the meaning of the item. When acquiescence bias is operating, it is assumed that for a balanced scale (i.e., a scale in which half of the items are PW and half are RW) the participant would agree with both PW and RW items, leading to the cancellation of the bias in the sum or average score. After the RW items have been reverse-coded, the total scale score is then presumed to be bias-free.

However, the wisdom of introducing RW items into scales has recently been questioned (e.g., Lindwall et al., 2012; Rodebaugh, Woods, & Heimberg, 2007; Roszkowski & Soven, 2010; Sonderen, Sanderman, & Coyne, 2013). First, to the extent that the tendency to acquiesce is an individual difference variable (Couch & Keniston, 1960; Ray, 1983), acquiescence bias will also contaminate the covariance structure of the data (Savalei & Falk, 2014). If the covariance structure of the data is of direct interest (e.g., if the researcher is interested in conducting factor analyses, reliability analyses, or structural equation modeling), the introduction of RW items into scales will bias the results of these analyses.

Second, the inclusion of negation RW items may cause confusion and lead to errors due to carelessness among some respondents. For example, a respondent may miss the presence of a negative particle (e.g., misread *I am not happy* as *I am happy*) or an affixal negation (e.g., misread *I am unhappy* as *I am happy*). Swain, Weathers, and Niedrich (2008) have also demonstrated that PW items that do not describe the respondent's actual state and negation RW that do describe the respondent's actual state will increase respondent's difficulty in interpreting the items. In addition, Schmitt and Stuits (1985) and Woods (2006) showed that the confusion and carelessness created by RW items might cause the emergence of a method factor common to all of the RW items. Specifically, if at least 10% of respondents are careless, a clearly identifiable method component emerges from principal component analysis (PCA; Schmitt & Stuits, 1985), and the fit of one-factor model through confirmatory factor analysis (CFA) is considerably worse (Woods, 2006).

Third, and finally, the inclusion of RW items may also create method effects that are not related to acquiescence bias nor confusion or carelessness but instead represent a consistent behavioral trait, such as fear of negative evaluation, self-consciousness, or approach and avoidance behavior tendencies (e.g., DiStefano & Motl, 2006; Quilty, Oakman, & Risko, 2009). That is, participants may be responding differently

to the sets of PW and RW items for idiosyncratic reasons. These reasons may or may not be related to the construct being measured.

The presence of method effects in Likert scales containing PW and RW items can influence the assessment of the factor structure of the scale and distort parameter estimates (DiStefano & Motl, 2006; Sonderen et al., 2013). For instance, Sonderen et al. (2013) showed that the correlations between pairs of items are often higher when they are worded in the same direction than when they measure the same construct. The inclusion of RW items in Likert scales can also result in lower validity (e.g., Rodebaugh et al., 2011), reduced reliability (e.g., Roszkowski & Soven, 2010), and the emergence of multiple factors for a scale that is supposed to measure one underlying construct (e.g., DiStefano & Molt, 2006; Rodebaugh, Woods, Heimberg, Liebowitz, & Schneier, 2006). For example, the emergence of a two-factor or two-component solution differentiating PW and RW items from exploratory factor analysis (EFA) or from PCA has been documented for the Rosenberg Self-Esteem Scale (RSES; e.g., Carmines & Zeller, 1979; Hensley & Roberts, 1976), Beck's Hopelessness Scale (e.g., Steed, 2001), and the trait scale in the State-Trait Anxiety Inventory (Bieling, Antony, & Swinson, 1998), to name a few.

Given the multitude of problems caused by the inclusion of RW items, some researchers have advocated against the continued use of RW items on Likert scales. These researchers have shown that for unbalanced Likert scales, removing RW items or replacing them with PW items often leads to better validity (e.g., Rodebaugh et al., 2007; Rodebaugh et al., 2011), reliability (e.g., Roszkowsky & Soven, 2010), and factor structure (e.g., Cordery & Sevastos, 1993; Greenberger, Chen, Dmitrieva, & Farruggia, 2003). However, this solution is suboptimal, because Likert scales consisting of only PW items will likely contain acquiescence bias (Jackson & Messick, 1961; Ray, 1983; Schuman & Presser, 1981; Swain et al., 2008).

One way to remove the impact of acquiescence bias or method effects on the covariance structure of a Likert scale with RW items is to explicitly model acquiescence bias (e.g., Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Savalei & Falk, 2014) or method factors (e.g., DiStefano & Motl, 2006; Lindwall et al., 2012; Marsh, 1996) using CFA. However, this modeling approach has several disadvantages. First, in a model that includes the acquiescence bias factor, the loadings for the acquiescence factor have to be fixed to one in order to identify the model (e.g., Savalei & Falk, 2014). Second, several competing models for modeling method factors have been proposed (e.g., DiStefano & Molt, 2006; Lindwall et al., 2012; Marsh, 1996). Some of these models contain one method factor while others contain two method factors. There is little consensus about which model is the best. Finally, a CFA model cannot contain both acquiescence factor and method factors because such a model will not be identified.

In order to simultaneously minimize acquiescence bias and to solve the problems caused by RW items, several authors (e.g., Brown & Maydeu-Olivares, 2011; Swain et al., 2008; White & Mackay, 1973) have advocated alternative scale formats. One such format is the forced-choice format (e.g., Javeline, 1999; Schuman & Presser,

1981). In the forced-choice format, the respondent is required to choose between two substantive response options rather than simply agreeing or disagreeing with a statement. For example, Schuman and Presser (1981) compared Likert format questions such as *Individuals are more to blame than social conditions for crime and lawlessness in this country; agree or disagree?*[3] to forced-choice format questions such as *Which in your opinion is more to blame for crime and lawlessness in this country—individuals or social conditions?* They argued that in the forced-choice format, acquiescence response bias should be eliminated because the response task requires a choice between two assertions rather than an endorsement of a single assertion. Additionally, because this format removes the very concept of PW and RW items, the factor structure of a scale should no longer be contaminated by method factors elicited by the direction of the item. Forced-choice format scales have shown good validity and reliability (Javeline, 1999; Schuman & Presser 1981). One disadvantage of the forced-choice format as originally defined, however, is that the items are dichotomous, making it hard to compare this format to the Likert format, which usually uses four or more response options.

In the present study, we examined an alternative scale format, which is an extension of the forced-choice format to include more response options. We refer to this format as the Expanded format. To our knowledge, this format has only been used in the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and in the Oxford Happiness Inventory (Argyle, Martin, & Crossland, 1989). To convert a Likert item to the Expanded format, each response option would be replaced by a full sentence. For instance, an item from the RSES that reads *On the whole, I am satisfied with myself* and that has four response options (with anchors *Strongly agree, Somewhat agree, Somewhat disagree*, and *Strongly disagree*) can be written in the Expanded format as follows:

- On the whole, I am very satisfied with myself.
- On the whole, I am satisfied with myself.
- On the whole, I am disappointed with myself.
- On the whole, I am very disappointed with myself.

Because both PW and RW items are presented for each item, acquiescence bias and unique method effects due to the direction of the item are theoretically eliminated. In addition, this format forces participants to pay more attention to the content of the item and to notice the subtle differences between options, thus potentially reducing confusion and careless responding. This format has the potential to effectively address the methodological issues associated with RW items in particular and with Likert scales more generally.

When the number of response options in the Expanded format and in the Likert format is the same, the factor structure of scales in the Expanded format and the Likert format can be compared. However, research comparing these two formats is virtually nonexistent. The only study is that of Hills and Argyle (2002), who

compared a happiness scale in the Likert and the Expanded formats. The scale in the Likert format contained both PW and RW items. But only PCA analyses were conducted and the outdated ''eigenvalue greater than one'' rule was used to judge dimensionality. Nonetheless, the dimensionality of the scale in the Expanded format was found to be one factor less, suggesting that the Likert format did in fact introduce additional variance contamination into the data.

The goal of the present study is to investigate the effectiveness of the Expanded format and to compare it to the Likert format. To conduct this comparison, we changed two popular Likert scales—the RSES (Rosenberg, 1965) and the Conscientiousness scale (CS) from the Big Five Inventory (John, Naumann, & Soto, 2008)—into the Expanded format. We also changed the BDI-II, which is already available in the Expanded format, into the Likert format. This was done in two different ways, as described in the Method section. For each scale, the factor structures of the data in the two formats were compared using both EFA and CFA. We hypothesized that scales in the Expanded format would have better (i.e., lower and more theoretically defensible) dimensionalities than scales in the Likert format when evaluated using EFA. We further hypothesized that when the one-factor CFA model is fit to data, scales in the Expanded format would have a better fit by the chi-square test and indices of approximate fit than scales in the Likert format. Furthermore, we expected that the fit of the Likert scales would improve substantially when two substantive factors were modeled or when a method factor was added to the one-factor model, whereas the fit of the scales in the Expanded format would stay relatively the same. Finally, we also examined the model-based reliabilities and convergent validities of the scales in both formats, although specific hypotheses were not formed because contamination due to item wording can cause higher or lower correlations as well as affect the variances of the observed variables.

Method

Participants

Participants were undergraduate students enrolled in psychology courses at the University of British Columbia. A between-subject design was used, so that each participant would only see one version of each scale. There were 641, 621, and 763 participants who completed one version of the RSES, CS, and BDI, respectively. The mean age of participants was 20 years (SD = 2.75), and 21% were male. The ethnic background of the participants was mainly Caucasian (27%) and East Asian (53%).

Procedure

Participants completed an anonymous survey online. The study took each participant about 20 minutes to complete. For two of the scales (RSES and CS), participants were randomly assigned to complete either the Likert version or the Expanded format

version. For the BDI scale, participants were randomly assigned to complete the Likert Version I, the Likert Version II, or the Expanded format version. Informed consent was obtained prior to the start of the survey.

The majority of participants (n = 472) were asked to complete one version for each of the RSES, CS, and BDI; however, some participants were asked to complete either one version of the BDI (n = 307) or one version of each of the RSES and CS[4] (n = 169). It should be noted that participants were also asked to complete several other psychological scales for other research projects. These scales included the Need for Cognition scale (Cacioppo & Petty, 1982), Subjective Happiness Scale (Lyubomirsky & Lepper, 1999), and Self-Competence and Self-Liking Scale (Tafarodi & Swann, 1995). The last portion of the survey consisted of questions relating to demographic information. After completing the survey, all participants were required to attend an oral debriefing session during which they received course credit for participation.

Measures

Rosenberg Self-Esteem Scale.[5] The 10-item Likert version of the RSES (n = 319) is the original RSES scale (Rosenberg, 1965) containing five PW items and five RW items, all measured on a 4-point scale, where 4 corresponds to *Strongly agree* and 1 corresponds to *Strongly disagree*. In the corresponding 10-item Expanded version of the RSES (n = 322), each item consists of four sentences to choose from, and all items are always arranged from the highest to the lowest self-esteem. For each item, participants were asked to select one of the four options that best describes them. The text of the corresponding Likert version item was used to create the four options for each item in the Expanded format (see Table 1 for examples). For most items in the Expanded version, the original Likert item was included as one of the four options (e.g., Sample Item II of RSES in Table 1); for the rest of the items, the options were created by adding a modifier to the original Likert item (e.g., Sample Item I of RSES in Table 1).

Conscientiousness Scale. The Likert version of the CS (n = 314) was taken from the Big Five Inventory (John et al., 2008). It contains five PW items and four RW items. The original items were on a 5-point scale; however, we used a 4-point scale from *Disagree strongly* to *Agree strongly*. This was done in order to match the number of response options to the four options in the Expanded format version. While it is possible to create the Expanded format version with five response options, it may be too many for the respondents to process. The comparison between Expanded formats with four versus five response options will be the subject of future research. The Expanded version (n = 307) contains nine items, with response options for each item ranging from the indication of the lowest conscientiousness to the highest conscientiousness (see Table 1 for examples). For each Expanded format item, the text of the corresponding Likert version item was used to create the four options. For eight of the nine items in the Expanded version, the original Likert item was included as one

Table 1. Sample Items for All Versions of the Three Scales.

	RSES	
	Likert version	Expanded version
Sample Item I	I feel that I have a number of good qualities.	• I feel that I have a great number of good qualities. • I feel that I have some good qualities. • I feel that I don't have many good qualities. • I feel that I have very few good qualities.
Sample Item II	I certainly feel useless at times.	• I feel useful most of the time. • I certainly feel useful at times. • I certainly feel useless at times. • I feel useless most of the time.
	CS	
	Likert version	Expanded version
Sample Item I	I am someone who does a thorough job.	• I am someone who does a sloppy job. • I am someone who does a somewhat sloppy job. • I am someone who does a somewhat thorough job. • I am someone who does a thorough job.
Sample Item II	I am someone who is a reliable worker.	• I am someone who is an unreliable worker. • I am someone who is a somewhat unreliable worker. • I am someone who is a somewhat reliable worker. • I am someone who is a reliable worker.
	BDI	
	Likert Version I / Likert Version II	Expanded version (BDI-II)
Sample Item I	I am so sad or unhappy that I can't stand it. / I feel sad much of the time.	• I do not feel sad. • I feel sad much of the time. • I am sad all the time. • I am so sad or unhappy that I can't stand it.
Sample Item II	I am not discouraged about my future. / I feel optimistic about the future.	• I am not discouraged about my future. • I feel more discouraged about my future than I used to be. • I do not expect things to work out for me. • I feel my future is hopeless and will only get worse.

Note. RSES = Rosenberg Self-Esteem Scale; CS = Conscientiousness Scale; BDI = Beck Depression Inventory. The response anchors for RSES and BDI Likert versions are "Strongly agree,""Somewhat agree,""Somewhat disagree," and "Strongly disagree." The response anchors for CS Likert scale are "Disagree strongly,""Disagree a little,""Agree a little," and "Agree strongly." For the items in the Expanded versions, participants were instructed to pick one of the four options that best describes themselves.

of the four options (see Table 1 for examples); for the remaining one item, the options were created by adding a different modifier for the original Likert item.

Beck Depression Inventory. The Expanded version (n = 256) is the original 21-item BDI-II (Beck, Steer, & Brown, 1996). Each item contains four options ranging in depression intensity (from absence of symptoms to frequent or intense symptoms), and participants were asked to pick the option that best described how they had been feeling in the last 2 weeks prior to study participation. The original BDI-II was designed to measure only the presence and the degree of depression. For every item on the scale, the first option is written to indicate that a particular symptom is absent (e.g., *I am not discouraged about the future*). Thus, the first option for each item could be viewed as a negation RW item for depression. However, a typical Likert scale usually contains both negation RW items and polar opposite RW items. For this reason, we created two Likert versions of the BDI. The Likert Version I (n = 254) was created by taking either the first or the last option in the original BDI-II as the corresponding Likert item. In other words, all RW items in the Likert Version I were negation RW items. The advantage of this version is that the wording of each item is identical to the wording of one of the response options for that item in the original BDI-II. The Likert Version II (n = 253) was created by replacing some of the negation RW items in Likert Version I with newly written polar opposite RW items (see Table 1 for examples). The advantage of this version is that it better resembles a typical Likert scale. For both Likert versions of the BDI, there were 11 PW items and 10 RW items, all measured on a 4-point scale from 4 = *Strongly agree* to 1 = *Strongly disagree*.

Analytic Method

For EFA, the *psych* package (Revelle, 2014) in R was used. Parallel analysis (Horn, 1965), available within the *psych* package, was used to determine the number of factors to be extracted. Parallel analysis is an improved version of the scree plot that incorporates sampling variability into the analysis (Zwick & Velicer, 1986). It does so by comparing the scree plot obtained from the data to an average scree plot obtained from simulated data sets of the same dimension as the original dataset but generated from a population where all variables are uncorrelated (e.g., Hayton, Allen, & Scarpello, 2004; Horn, 1965). Following this method, the recommended number of factors to be extracted is the number of original data's eigenvalues that are greater than the corresponding simulated data eigenvalues. Because all the items were measured on a 4-point scale, we treated the data as categorical (Rhemtulla, Brosseau-Liard, & Savalei, 2012). The polychoric correlation matrix was used in the parallel analyses and in the subsequent EFA analyses. The EFA extraction method was least squares (a.k.a. minres), followed by an oblimin rotation.

For CFA, the *lavaan* package (Rosseel, 2012) in R was used. Because the data were treated as categorical, the diagonally weighted least squares estimator with

robust corrections was used (i.e., estimator=''WLSMV'' in *lavaan*). Three different models were fit to the data from each scale (see Figure 1). These models are commonly used to study method effects in Likert scales (e.g., Lindwall et al., 2012; Marsh, Scalas, & Nagengast, 2010; Quilty et al., 2006). Model 1 posits a single substantive factor with no method effects. Model 2 posits two oblique factors with all PW items loading on one factor and all RW items loading on the other. Model 1 is nested within Model 2. Model 3 posits a single substantive factor and an orthogonal method factor for the RW items. In the Expanded format, there is no distinction between PW and RW items, and the factors for Models 2 and 3 were formed according to whether the corresponding item was RW or PW in the Likert format.[6] Model 1 is also nested within Model 3. Models 1 to 3 are fit to all data sets regardless of the number of factors suggested by the EFA. Thus, CFA analyses were treated as quite distinct from the EFA analyses: their main purpose was to investigate whether a unidimensional model (Model 1) would fit the data better under the Expanded format than under the Likert format, and whether the Likert format data would suggest more strongly than the Expanded format data that the construct being measured by the scale is multidimensional.

The following criteria were used to evaluate model fit: (1) the test of exact fit using the robust (mean-and-variance adjusted) chi-square statistic for categorical data; (2) the comparative fit index (CFI; Bentler, 1990), with the value of 0.95 or greater indicating a well-fitting model (Hu & Bentler, 1999); (3) the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), with the value of 0.08 or less indicating a reasonable fit (Browne & Cudeck, 1993); (4) the robust chi-square difference tests comparing Model 1 and Model 2, and Model 1 and Model 3—these tests are not simple differences of the robust chi-squares but involve complicated formulae (Asparouhov & Muthen, 2006; Satorra, 2000) and are implemented in *lavaan*; and (5) the tests of small differences in fit between Model 1 and Model 2, and between Model 1 and Model 3, using $RMSEA_A = 0.09$ and $RMSEA_B = 0.08$, testing the null hypothesis that the difference in fit between a pair of models is equal to or less than the difference in model fit that would change the RMSEA value from 0.08 to 0.09 (see MacCallum, Browne, & Cai, 2006, for more details).

We also examined reliabilities for each scale in each of the formats. Model-based reliabilities (e.g., Bentler, 2009; Raykov, 1997) were computed under Model 1 (the one-factor model) and Model 3 (the model with one substantive factor and one method factor; see Figure 1). For the model-based reliability computed under Model 3, the variance due to the method factor was treated as error. Finally, previous research showed that self-esteem and conscientiousness are moderately positively correlated (e.g., Pullmann & Allik, 2000) and that self-esteem and depression are largely inversely correlated (e.g., Greenberger et al., 2003). To see how scale format affects these relationships, correlations between different versions of the RSES and CS, and between versions of the RSES and BDI were examined.

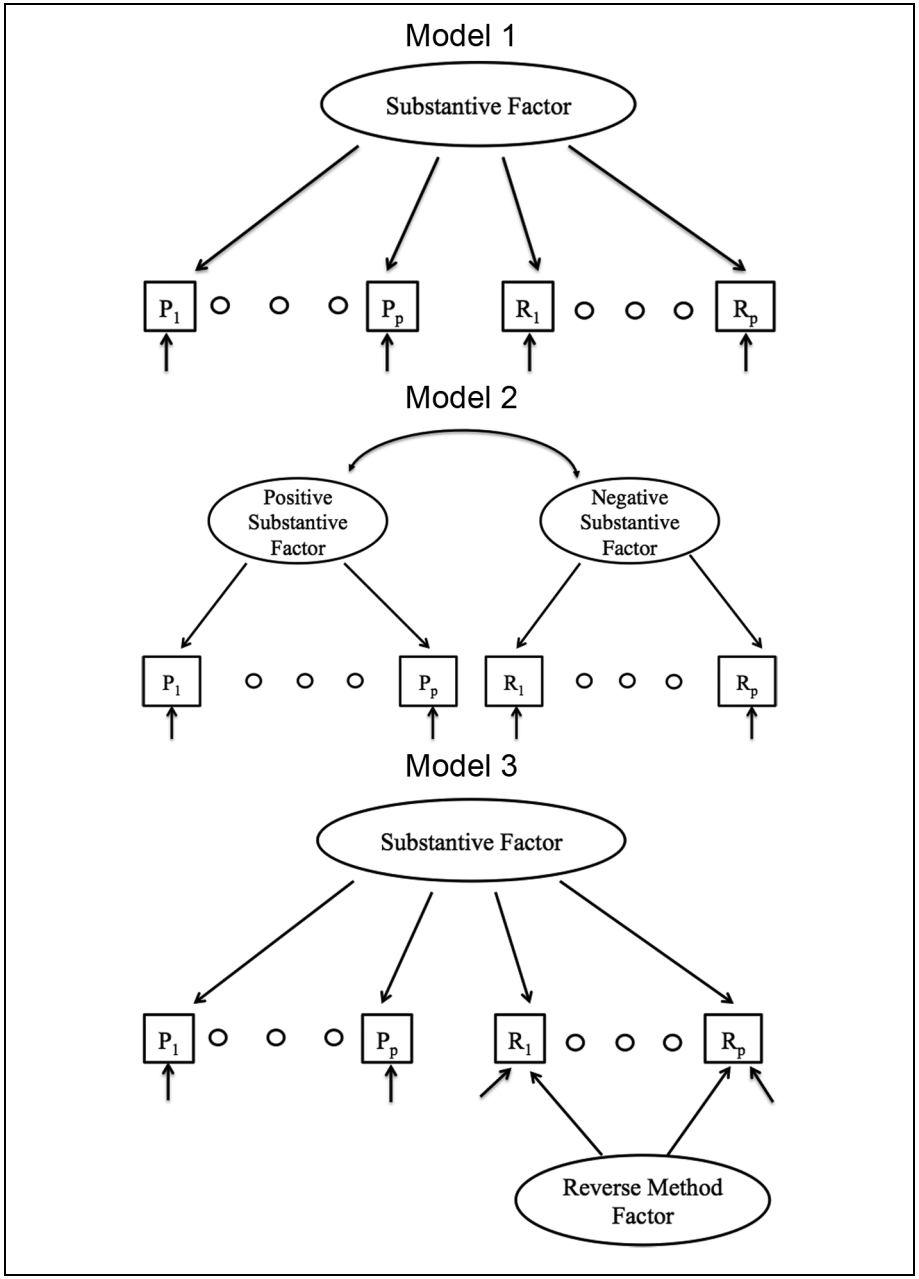


Figure 1. The three factor structure models tested in the study.
Note. P = Positively worded item; R = Reverse worded item.

Table 2. Average Item Mean and Standard Deviation for All Versions of the Three Scales.

	RSES		CS		BDI		
	Likert version	Expanded version	Likert version	Expanded version	Likert Version I	Likert Version II	Expanded version (BDI-II)
Average item mean	3.04	3.09	2.92	3.01	1.95	2.14	1.56
Average item standard deviation	0.77	0.62	0.81	0.62	0.85	0.86	0.72

Note. RSES = Rosenberg Self-Esteem Scale; CS = Conscientiousness Scale; BDI = Beck Depression Inventory. Scores ranging from 1 to 4 for all items. Higher values indicate higher endorsement of the construct the scale is measuring.

Results

Average item means and standard deviations for all three scales in different formats are shown in Table 2. Scales in the Expanded version consistently had smaller standard deviations than the corresponding scales in the Likert version, possibly due to the reduction in variance contamination due to the absence of RW method effects. For the RSES and CS, the average means across items were very similar with an average mean difference of 0.04 between the two versions of RSES and an average mean difference of 0.09 for those of CS. However for BDI, the average means were significantly different among the three versions, $F(2, 63) = 35.25$, $p < .001$; the Likert Version II had the highest average mean (average $M = 2.14$), followed by the Likert Version I (average $M = 1.95$), followed by the Expanded version (average $M = 1.56$). These results illustrate that it is relatively difficult to create a Likert version of the BDI that would map well onto the original BDI. Raw means, standard deviations, and frequency distributions of the responses for all items on the three versions of the BDI scale can be found in supplementary materials (available online at http://epm.sagepub.com/content/by/supplemental-data).

Exploratory Factor Analysis

Figures 2 and 3 show the results for the parallel analyses of all three scales in different formats. Overall, as expected, parallel analysis suggested fewer factors for the Expanded version of each scale than for the corresponding Likert version(s). Specifically, for the RSES and CS's Likert versions, the scree plots from parallel analysis clearly indicated two factors; however, for the Expanded version, the second factor was much weaker, and the corresponding eigenvalue was close to the cutoff line obtained from the simulated data, suggesting a single factor. For the two Likert versions of the BDI, the scree plots from parallel analysis indicated that four eigenvalues were above the sampling fluctuations observed in simulated data, suggesting as many as four factors, whereas the scree plot for the Expanded BDI indicated at

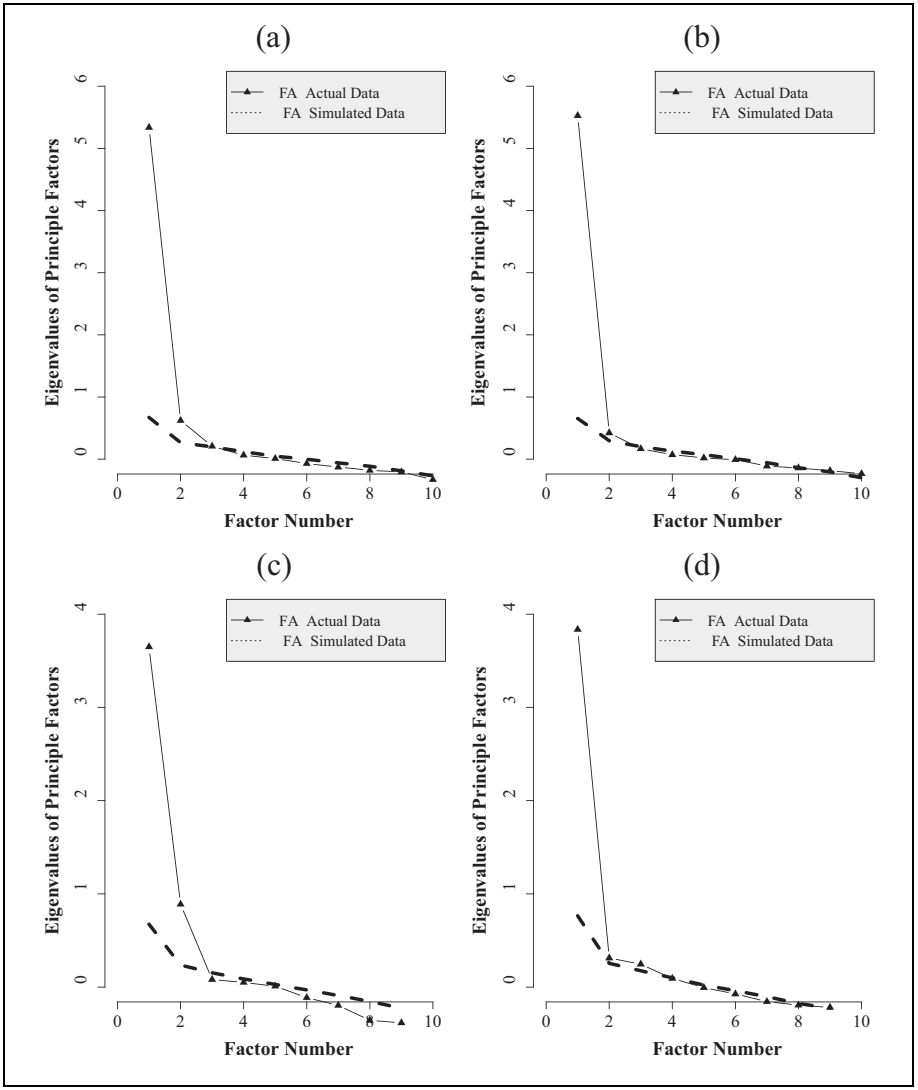


Figure 2. Parallel scree plots for the Rosenberg Self-Esteem Scale (RSES) and the Conscientiousness Scale (CS).

Note. Plot (a) is for the RSES Likert version; Plot (b) is for the RSES Expanded; Plot (c) is for the CS Likert version; Plot (d) is for the CS Expanded version.

most two factors, with the second factor being very close to the cutoff. These results were consistent with previous findings that the BDI is not a unidimensional scale (see Manian, Schmidt, Bornstein, & Martinez, 2013, for a comprehensive literature review; in particular see Table 1 of the article).

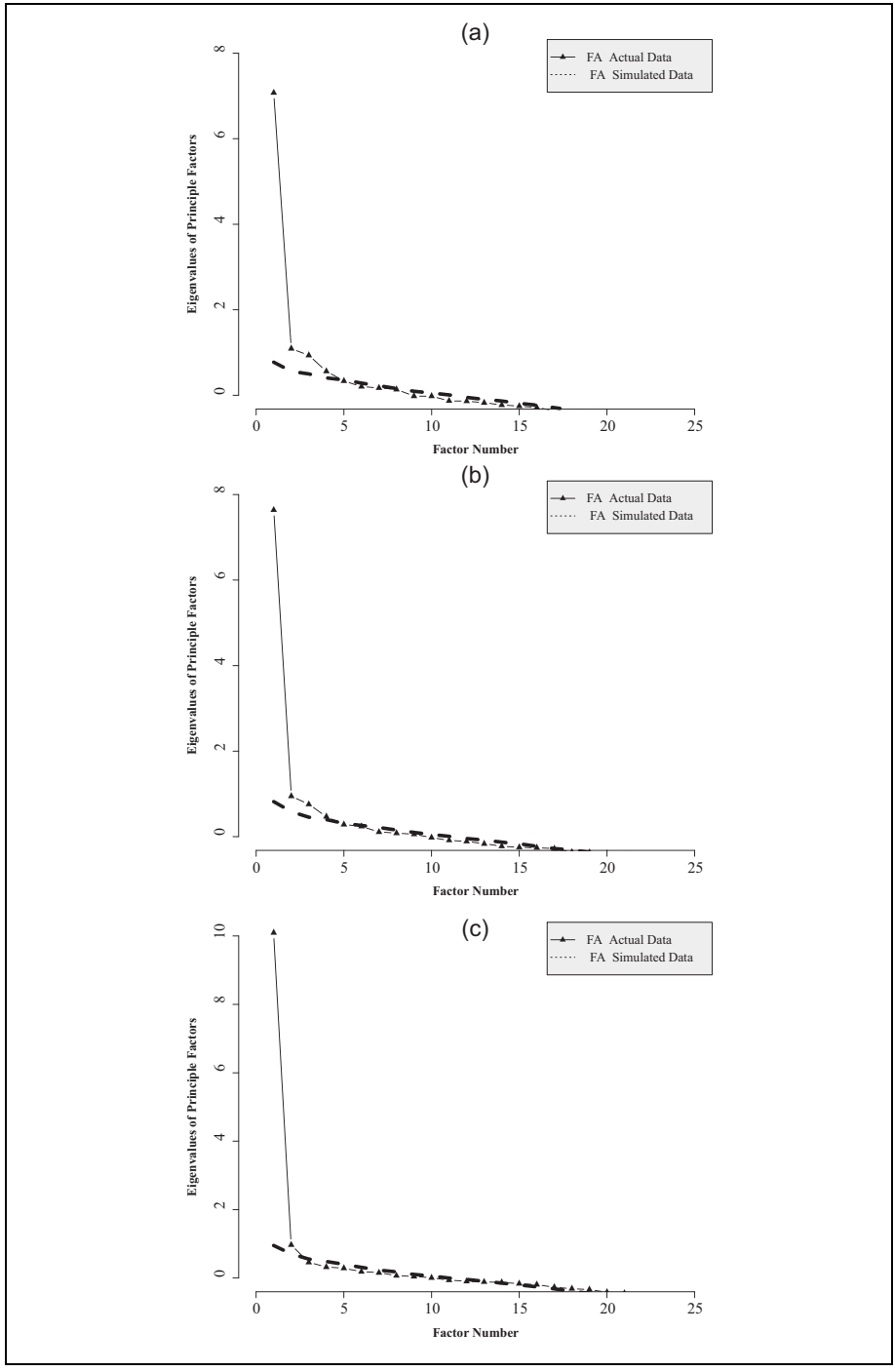


Figure 3. Parallel scree plots for the Beck Depression Inventory.
Note. Plot (a) is for the Likert Version I; Plot (b) is for Likert Version II; Plot (c) is for Expanded version.

EFA analyses were performed next (see supplementary materials for detailed results). For the RSES and CS scales in both formats, we extracted two factors. This was the number suggested by the parallel analyses for the Likert version of these scales. Even though the number of factors suggested for the Expanded versions was one, the same number of factors was extracted for both formats in order to compare the results across the formats. The initial extracted solution was obliquely rotated. As expected, for the Likert versions of the RSES and CS, the pattern of loadings indicated that the two factors should be formed based on whether the item was PW or RW. These results replicated previous findings that the inclusion of RW items in Likert scales causes the emergence of a two-factor solution based on the wording of the items (e.g., Carmines & Zeller, 1979; Hensley & Roberts, 1976). On the other hand, the RSES and CS scales in the Expanded version did not produce very clear two-factor solutions. Most items loaded highly on the first factor; the few items that loaded highly on the second factor had similar item content. For example, Items 1 and 4 in the RSES were both about social comparison between oneself and others. Items 1 and 3 in CS were both related to being a good worker. Thus, when a two-factor solution was forced on the data in the Expanded format, the factors formed based on the content and not a method effect due to the item wording. Additionally, the factor correlations in the Expanded RSES ($r = 0.72$) and CS ($r = 0.77$) were higher than those in the Likert RSES ($r = 0.63$) and Likert CS ($r = 0.49$), respectively, supporting the conclusions from parallel analysis that these scales in the Expanded format can be modeled using a single factor.

For the BDI, parallel analysis suggested a 2-factor solution for the Expanded version and a 4-factor solution for the Likert Versions I and II. To compare the versions on the same number of factors, we performed EFA analyses extracting two and four factors for all versions. For the 4-factor solutions of the two Likert versions, the factors were formed based on both item content and item wording direction. For the Likert Version I, all the PW items loaded highest on Factor 1 or 2 and most of the RW items loaded highest on Factor 3 or 4. In terms of item content, Factors 1 and 2 appeared to be the cognitive-affective factor and the somatic factor of depression among the PW items, respectively, whereas Factors 3 and 4 seemed to be the cognitive factor and somatic-affective factor of depression among the RW items, respectively. For the Likert Version II, all the PW items loaded highest on either Factor 2, 3, or 4, and 8 out of 10 the RW items loaded highest on Factor 1. In terms of item content, Factor 2 and Factor 3/4 corresponded to items tapping into the cognitive factor and somatic-affective factor among the PW items, respectively. Factor 1 seemed to tap into all aspects of depression among the RW items. For the Expanded version, the factors were formed based on the item content. Factor 1 was the cognitive factor for depression, whereas Factors 2 and 3 were most related to the somatic-affective aspect of depression. Factor 4 seemed to be a spurious factor with only two items. Consistent with the results from the parallel analysis, this solution resulted in over-extracted factors for the Expanded version of the scale.

Table 3. Summary of the Fit Statistics for the Rosenberg Self-Esteem Scale and Conscientiousness Scale.

	Likert version				Expanded version			
	χ^2	*df*	CFI	RMSEA	χ^2	*df*	CFI	RMSEA
Rosenberg Self-Esteem								
Model 1	296.73	35	0.93	0.15	106.12	35	0.98	0.08
Model 2	141.32	34	0.97	0.10	103.76	34	0.98	0.08
Model 3	95.08	30	0.98	0.08	95.25	30	0.98	0.08
$\Delta\chi^2_{m1\text{-}m2}$ (*df* = 1) (χ^2 critical, *p* value for test of small difference)	6.13 (*p* = .01) (38.72, 0.98)				0.11 (*p* = .74) (38.98, 1.00)			
$\Delta\chi^2_{m1\text{-}m3}$ (*df* = 5) (χ^2 critical, *p* value for test of small difference)	167.00 (*p* = .00) (54.12, 0.00)				13.46 (*p* = .02) (54.47, 0.99)			
Conscientiousness Scale								
Model 1	245.61	27	0.86	0.16	75.32	27	0.97	0.08
Model 2	86.06	26	0.96	0.09	63.81	26	0.98	0.07
Model 3	87.74	23	0.96	0.09	60.08	23	0.98	0.07
$\Delta\chi^2_{m1\text{-}m2}$ (*df* = 1) (χ^2 critical, *p* value for test of small difference)	11.33 (*p* = .00) (32.39, 0.75)				0.64 (*p* = .42) (31.87, 1.00)			
$\Delta\chi^2_{m1\text{-}m3}$ (*df* = 5) (χ^2 critical, *p* value for test of small difference)	137.89 (*p* = .00) (44.13, 0.00)				16.10 (*p* = .00) (43.45, 0.84)			

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation. Model 1: Single substantive factor; Model 2: Two oblique factors, one among positively worded (PW) item and one among reverse worded (RW) items; Model 3: Two factors, substantive and method effect among RW items. In Expanded version, there is no distinction between PW and RW; the factors were formed according to whether the corresponding item is RW or PW in Likert version I. $\Delta\chi^2$ was calculated according to Satorra's (2000) suggestion for calculating chi-square difference test using the robust chi-square; and $\Delta\chi^2$ was used to conduct both chi-square difference and test of small difference.

Table 4. Summary of the Fit Statistics for Beck Depression Inventory.

	Likert Version I				Likert Version II				Expanded version (BDI-II)			
	χ^2	*df*	CFI	RMSEA	χ^2	*df*	CFI	RMSEA	χ^2	*df*	CFI	RMSEA
Model 1 (M1)	775.46	189	0.84	0.11	561.57	189	0.91	0.09	412.01	189	0.95	0.07
Model 2 (M2)	672.54	188	0.87	0.10	481.45	188	0.93	0.08	410.84	188	0.95	0.06
Model 3 (M3)	584.42	179	0.89	0.09	462.52	179	0.93	0.08	397.13	179	0.95	0.06
$\Delta\chi^2_{m1\text{-}m2}$ (*df* = 1) (χ^2 critical, *p* value for test of small difference)	1.48 (*p* = .22) (115.57, 1.00)				0.87 (*p* = .35) (115.18, 1.00)				0.01 (*p* = .92) (116.34, 1.00)			
$\Delta\chi^2_{m1\text{-}m3}$ (*df* = 10) (χ^2 critical, *p* value for test of small difference)	182.41 (*p* = .00) (142.39, 0.00)				112.03 (*p* = .00) (141.94, 0.39)				0.01 (*p* = .91) (143.28, 1.00)			

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation. Model 1: Single substantive factor; Model 2: Two oblique factors, one among PW items and one among RW items; Model 3: Two factors, substantive and method effect among RW items. In Expanded version, there is no distinction between PW and RW; the factors were formed according to whether the corresponding item is RW or PW in Likert Version I. χ^2 was calculated according to Satorra's (2000) suggestion for calculating chi-square difference test using the robust chi-square; and χ^2 was used to conduct both chi-square difference and test of small difference.

When the two Likert scales were forced into a 2-factor solution, the Likert Version II solution was formed based on the item wording direction, with the PW items primarily loading on Factor 2, and the RW items primarily loading on Factor 1. Interestingly, the 2-factor solution in the Likert Version I was formed based on item content more so than on item wording: Factor 1 was most related to the somatic items and Factor 2 was most related to the cognitive-affective items. These results suggest that the substantive multidimensionality of the BDI (as captured by this scale) is stronger than the dimensionality due to method effects (item wording). However, this 2-factor solution was not very clear. Six items loaded poorly on both factors (standardized loadings below 0.4), and only five items had standardized loadings greater than 0.6 on either factor. In contrast, the 2-factor solution for the Expanded version of the BDI scale was formed based on item content and was much clearer than the solution produced by the Likert II format data. Factor 1 was again a cognitive-affective factor and Factor 2 was a somatic factor. Only two items loaded poorly on both factors (with loading sizes below 0.4), and 11 items had standardized loadings greater than 0.6 on one of the factors. This 2-factor solution was a good replication of previous EFA analyses from other studies using community samples (e.g., Beck et al., 1996; Steer & Clark, 1997). Consistent with the results of the RSES and CI, the factor correlation for the Expanded BDI ($r = 0.65$) was higher than the factor correlations for the Likert Version I ($r = 0.54$) and Version II ($r = 0.62$). However, BDI is not a theoretically unidimensional scale, and it is reasonable that the factor correlation for the Expanded BDI is not as high as those for the Expanded RSES and CI.

Confirmatory Factor Analysis

Model Fit. The fit indexes for the three CFA models tested (see Figure 1) for different versions of the three scales are shown in Tables 3 and 4. As expected, for the Likert versions of all three scales, Model 1, positing a single factor, did not fit the data well according to the CFI and the RMSEA. On the contrary, Model 1 had a relatively good fit for the Expanded versions of all three scales according to the CFI and the RMSEA. Specifically, the CFI values were all equal or above 0.95 and the RMSEA values were all equal or below 0.08.

Model 2, which is a CFA model with two oblique factors (one among PW items and one among RW items), produced significantly better fit relative to Model 1 for the RSES and CS Likert scales, according to the exact chi-square difference test but not the test of small difference (see Table 3 and 4). However, Model 2 did not produce significantly better fit than Model 1 for the two versions of BDI Likert scale according to both the exact chi-square difference best and the test of small difference (see Table 4). While the fit indices did improve for Model 2 relative to Model 1 for the Likert versions of the three scales, the fit indices did not always reach acceptable values, particularly for the BDI, with Likert Version I fitting worse than Likert Version II. This result is not unexpected given that parallel analyses revealed a 4-factor structure for the Likert versions of the BDI scale.

Table 5. Average Standardized Factor Loadings for All Versions of the Three Scales.

	Model 1						
	RSES		CS		BDI		
	Likert version	Expanded version	Likert version	Expanded version	Likert Version I	Likert Version II	Expanded version
	GSE	GSE	GC	GC	GD	GD	GD
Average loading	0.74	0.75	0.65	0.66	0.58	0.60	0.69

	Model 2													
	RSES				CS				BDI					
	Likert version		Expanded version		Likert version		Expanded version		Likert Version I		Likert Version II		Expanded version	
	PSE	NSE	PSE	NSE	PC	NC	PC	NC	PD	ND	PD	ND	PD	ND
Average loading	0.83	0.73	0.73	0.77	0.75	0.70	0.68	0.68	0.65	0.58	0.64	0.61	0.70	0.69
Factor correlation	0.78		0.97		0.57		0.89		0.76		0.83		0.98	

	Model 3													
	RSES				CS				BDI					
	Likert version		Expanded version		Likert version		Expanded version		Likert Version I		Likert Version II		Expanded version	
	GSE	RW	GSE	RW	GC	RW	GC	RW	GD	RW	GD	RW	GD	RW
Average loading	0.70	0.45	0.75	0.21	0.59	0.58	0.65	0.32	0.55	0.37	0.58	0.32	0.69	0.16

Note. Average loading sizes were calculated using absolute values of the loading values. GSE = global self-esteem; GC = global conscientiousness; GD = global depression; PSE = positive self-esteem; NSE = negative self-esteem; PC = positive conscientiousness; NC = negative conscientiousness; PD = positive depression; ND = negative depression; RW = reverse worded method factor.

Table 6. Model-Based Reliabilities for All Versions of the Three Scales.

	RSES		CS		BDI		
	Likert version	Expanded version	Likert version	Expanded version	Likert Version I	Likert Version II	Expanded version
Model-based reliability based on Model 1	0.93	0.93	0.88	0.87	0.92	0.92	0.95
Model-based reliability based on Model 3	0.84	0.92	0.75	0.84	0.84	0.87	0.94

Note. RSES = Rosenberg Self-Esteem Scale; CS = Conscientiousness Scale; BDI = Beck Depression Inventory. Model-based reliabilities were calculated based on Model 1 and Model 3 in Figure 1. Model-based reliabilities based on Model 3 were calculated by partialling out the variance due to the method factor as error.

Model 2 did not result in a significant improvement in fit over Model 1 for the Expanded versions of all three scales, according to both the chi-square difference test and the test of small difference. To understand this result for the BDI, it is important to remember that in the tested 2-factor CFA model, the factors were formed based on the item wording direction in the corresponding Likert versions (see Figure 1). The RMSEA, the CFI, and the test of close fit yielded very similar results for Models 1 and 2 for all three scales in the Expanded format.

Model 3, which posited two orthogonal factors, a substantive factor common to all items and a method factor for the RW items, produced very similar results to Model 2 for all scales in all versions. Model 3 generally produced significantly better fit for the Likert versions of the three scales in comparison with the fit of Model 1, although the BDI Likert Version II did not produce significantly better fit than Model 1 according to the test of small difference. In contrast, Model 3 did not produce significantly better fit relative to Model 1 for the Expanded versions of the three scales, with only the exact chi-square difference test of the CS Expanded scale being an exception.

In summary, for the Likert scales, Model 1, which posits unidimensionality, did not fit the data well, but Models 2 and 3, which allow for multidimensionality, resulted in better fit than Model 1 by multiple fit criteria. These results are consistent with our hypothesis that the RW items of the three scales in the Likert format contaminate the factor structure of the scales. On the other hand, Model 1 had acceptable approximate fit for all three scales in the Expanded version, and including additional factors did not improve fit much. These CFA results are also consistent with the EFA results, which demonstrated that the factor structures of the Expanded versions of the three scales had lower dimensionality. It is important to note that we did not test multidimensional models based on item content, only based on the direction of item wording. Allowing for multiple factors based on content could result in a better-fitting model even for the Expanded version of the scales. However, the goal of this

Table 7. Correlations Between the Rosenberg Self-Esteem Scale and the Conscientiousness Scale, and Between the Rosenberg Self-Esteem Scale and the Beck Depression Inventory.

		CS	
		Likert version (original scale)	Expanded version
RSES	Likert version (original scale)	0.33 (n = 160)	0.40 (n = 152)
	Expanded version	0.32 (n = 154)	0.32 (n = 155)
		RSES	
		Likert version (original scale)	Expanded version
BDI	Likert Version I	−0.70 (n = 74)	−0.55 (n = 81)
	Likert Version II	−0.82 (n = 67)	−0.60 (n = 87)
	Expanded version (original scale)	−0.60 (n = 88)	−0.76 (n = 66)

Note. All conditions were between-subject. Sample size for each condition is shown in parentheses. All correlations are significant at .0001 level. RSES = Rosenberg Self-Esteem Scale; CS = Conscientiousness Scale; BDI = Beck Depression Inventory.

article is to demonstrate the impact of scale format on fit. The scales' structure approximated unidimensionality to a much higher degree when contamination due to the direction of item wording was not an issue.

Standardized Solution. Average standardized factor loadings for each factor in all models of the three scales are shown in Table 5 (full results are available in supplementary materials). Several interesting patterns emerged in Table 5. First, for all three scales, average standardized factor loadings for Model 1 were similar for the Likert version(s) and the Expanded version. In other words, under both scale formats, the same percentage of reliable variance was attributed to the substantive factor when Model 1 was fit to data. Second, for all three scales, the factor correlation in Model 2 was much lower for the Likert version(s) than for the Expanded version; in fact, they were so high for the Expanded versions that most researchers would likely conclude from such data that the two factors should be collapsed into one. The two-dimensional solution sometimes found for these scales in the Likert format (e.g., ''positive self-esteem'' and ''negative self-esteem''; Marsh, 1996) was thus likely due to item wording and not the existence of two substantive dimensions. Finally, for each scale, the average item loading value on the RW method factor was much greater in the Likert versions than in the Expanded versions, consistent with the hypothesis that the modeled method effects were not present in the Expanded versions of the scales.

Reliability Estimates. Model-based reliability coefficients are shown in Table 6. When the model-based reliability was computed under Model 1, the Likert and Expanded versions of each scale had very similar reliability coefficients with a maximum

difference of 0.03. When the model-based reliability was computed under Model 3, the reliability of the Expanded version of each scale was very close to the one computed under Model 1; however, the reliability of the Likert version of each scale was less than the one computed under Model 1. As a result, for the reliabilities computed under Model 3, the Expanded version of each scale had consistently better reliabilities with an average difference of 0.085 between the Expanded version and the Likert version.

Scale Correlations

We also examined the correlations between different versions of the RSES and CS, and between the RSES and BDI to see whether they were significantly affected by the scale format (see Table 7). Self-esteem and conscientiousness are usually moderately positively correlated (e.g., Pullmann & Allik, 2000), whereas self-esteem and depression are usually inversely correlated (e.g., Greenberger et al., 2003). The correlations between the RSES and CS were very similar across formats, ranging from 0.32 to 0.40. Correlations between the RSES and BDI varied more across the formats, ranging from -0.55 to -0.82, but they all indicated a large inverse relationship between self-esteem and depression. The highest correlation was between the RSES Likert version and the BDI Likert Version II ($r = -0.82$). This high correlation was probably caused by the fact that some of newly created items on the BDI Likert Version II had considerable item content overlap with the items on the RSES Likert version.

Discussion

The purpose of the present study was to examine the impact of an alternative scale format, called the Expanded format, on the factor structure of three psychological scales. The advantage of the Expanded format is that it asks participants to choose among response options that are complete sentences, rather than simply asking them to indicate their level of agreement or disagreement with a statement. This more elaborate format eliminates the very concept of PW and RW items because each item contains sentence options of both types, and it also forces participants to pay more attention to the question. The performance of this format was compared to the performance of the traditional Likert scale format with PW and RW items. We hypothesized that scales in Expanded format would show a more parsimonious factor structure compared with the same scales in the Likert format, as assessed by EFA and CFA analyses, due to the elimination of method variance associated with item wording and the reduction of the number of careless or confused responses.

Our EFA results confirmed our predictions. The scales in the Likert format tended to produce 2-factor EFA solutions differentiating PW and RW items (for the BDI, a higher factor solution was necessary), and to require two factors to produce acceptable approximate fit in the CFA analyses. These findings are consistent with previous

studies (e.g., Carmines & Zeller, 1979; DiStefano & Molt, 2006). In contrast, the scales in the Expanded format did not show similar problems. The EFA analyses suggested fewer factors for the scales in the Expanded format. Specifically, for two out of the three scales, the suggested number of factors was one. The factors in the Expanded format scales were more consistent with the theoretical structure of the scales. These results are consistent with the findings of Hills and Argyle (2002), which is the only other study that compared the Expanded and the Likert format, as applied to a happiness scale, although they conducted PCA and not EFA analyses.

Our CFA results were also consistent with our predictions. The one-factor model fit the scales in the Expanded format substantially better than the scales in the Likert format. According to tests of small differences, adding a method factor for the RW items did not significantly improve fit for the scales in the Expanded format but significantly improved fit for the scales in the Likert format. This finding is quite impressive considering the similarities between the actual wording of the corresponding items in the two formats. For example, to create the Expanded version of CS, we created new response options for most items simply by adding an adjective modifier such as *somewhat* or by replacing a word with an antonym (e.g., changing *useful* to *useless*; see Table 1 as well as supplementary materials). Thus, the differences in the fit of one-factor CFA model can be argued to be entirely due to item format and not to item wording.

We believe that the main features of the Expanded format item that are responsible for this improvement in the factor structure are (1) that it is nondirectional; (2) that it does not require participants to agree or disagree with the item (i.e., participants only need to pick a response option); (3) that both PW and RW item wordings are included as response options in the item. The first feature eliminates method effects due to item keying. The second feature minimizes acquiescence bias, which theoretically only occurs when the scale requires respondents to agree or disagree with the items. We believe that the third feature reduces carelessness and confusion, although this point is difficult to demonstrate conclusively. When participants are presented with both PW and RW item wordings as response options to the same item, they might be more likely to notice the difference between the response options and thus less likely to misread the item.

Interestingly, scales in the Expanded format and the Likert format did not have substantively different reliability coefficients when these were estimated under the one-factor model (Model 1). However, Model 1 fit the Likert format data significantly more poorly than it fit the Expanded format data, and thus one-factor reliability coefficients may not be appropriate. When reliability was estimated under Model 3, which allows for a method factor, scales in the Likert format had much lower reliability, while the reliability estimate for the scales in the Expanded format was largely unchanged. This finding may indicate that when a Likert scale is modeled under the one-factor model, some of the variance due to the method factor may be forced to go through the substantive factor, thus resulting in an inflated reliability estimate.

Correlations between pairs of scales did not differ much across formats. This finding may seem to contradict some of the past research showing that removing RW items or replacing them with PW items resulted in better scale validity (e.g., Rodebaugh et al., 2007, 2011). However, these studies have examined only unbalanced scales. It is possible that for balanced scales (i.e., scales with an equal number of PW and RW items), the influence of method factors and acquiescence bias on the correlations may cancel out, as it does for sum scores (Ray, 1983; Savalei & Falk, 2014); however, this remains to be shown. The scales in this study are either perfectly balanced (RSES) or nearly balanced (the number of RW items in the Likert versions of the CS and the BDI is only one less that the number of PW items). Examining the impact of item format on the properties of sum scores of balanced and unbalanced scales will be subject of further research.

Writing Scales in the Expanded Format

Keeping in mind that ours was the first study to thoroughly examine the impact of the Expanded format on the factor structure of psychological scales, we offer a few tentative recommendations for researchers who want to create scales in the Expanded format. Further research will fine-tune these recommendations.

When creating items in the Expanded format, it may be prudent to minimize all other differences between the response options that do not pertain to the key difference. This can be achieved by varying just a few words as possible between the options. For example, for the second example of the CS in the Expanded format in Table 1, we created the options by switching between the words, *reliable* and *unreliable*, and by adding or deleting the word *somewhat*. Making the differences between the options obvious may reduce confusion for the respondents.

If the scale already exists in the Likert format, it may be prudent to try to keep the original wording of the Likert item by making it one of the response options in the Expanded format, in order to minimize changes in item content of a previously validated and popular scale. However, due to the ambiguity in the wording of some Likert items, this may not always be possible. For example, for the RSES item, *I wish I could have more respect for myself*, it is hard to predict what a respondent exactly means when he or she picks *Strongly Disagree* for this item. The respondent may mean that her current level of respect for herself is accurate and does not wish for it to change, or the respondent may mean that she does not respect herself but does not want to change this. To change the item into the Expanded format, we simplified the wording and created the response options: *I have a lot of respect for myself, I have some respect for myself, I have little respect for myself*, and *I have no respect for myself*.

The above example illustrates that an advantage of the Expanded format is that it forces researchers to resolve ambiguities inherent in the wording of many Likert items. But this advantage is also a challenge. For another example, consider the following item from the CS: *I am someone who can be somewhat careless*. If a

respondent strongly agrees with this item, it is unclear whether she means that she is very careless, she is careless most of the time, or she can be somewhat careless at times. In such a case, the researcher has to pick one particular interpretation to create the options for the corresponding Expanded item. By converting such ambiguous Likert items into Expanded items, both researchers and respondents will have a clearer understanding of the intended meaning of the response options of the items. In our future research, we will explore different ways of changing Likert items into Expanded items and examine how to create better Expanded items.

Limitations and Future Research

Despite the many advantages, the Expanded format may nonetheless have important limitations, and the impact of these limitations will be explored in future research. First, scales in the Expanded format may take longer to complete because they involve more reading and force participants to pay attention to the text of the item more carefully. This may, however, be the inevitable cost of obtaining higher quality data free of method effects and other types of contamination (e.g., confusion).

Second, while the method effects due to PW and RW items are theoretically eliminated in the Expanded format, there may still be order effects. Previous research on forced-choice format found both primacy and recency effects (e.g., Krosnick & Alwin, 1987; McClendon, 1984, 1991). Interestingly, McClendon (1991) found that when items on the RSES were translated into the forced-choice format, no order effects emerged. Since most research on the forced-choice format was conducted in political science and sociology, it is possible that participants are prone to order effects only when they do not have strong preexisting opinions on the question being asked and thus must construct an opinion at the time of the question; this problem would be reduced for questions about the self (McClendon, 1991). We also note that in the current study, the order of response options varied across scales. For instance, the RSES scale in the Expanded format had response options ordered from high self-esteem to low self-esteem, while the CS scale in the Expanded format had response options ordered from low conscientiousness to high conscientiousness. Yet both these scales produced item means that were very similar to the corresponding items in the Likert version (see Table 2 and supplementary materials), suggesting that neither order was superior. Nonetheless, future studies should investigate whether the Expanded format scales are affected by order effects in a more systematic fashion.

Finally, items on the Expanded format scales must necessarily have fewer response options than the typical Likert scale items. In this study, items in the Expanded format had four response options (and thus the data were treated as categorical). In contrast, Likert scales often have five to seven categories, and the data from such scales can be more safely treated as continuous (Rhemtulla et al., 2012). This difference may be viewed by some practitioners as a limitation; however, categorical methods for data analysis have become widely available, and it is straightforward to treat the data as categorical for EFA and CFA analyses, as was done in the present

study (e.g., Bock & Lieberman, 1970; Muthen, 1984; Muthen & Asparouhov, 2002). When it comes to average or sum scores, these can be safely treated as continuous even when the item data only have four categories. Nonetheless, future research will address whether the Expanded format can be extended to include more response options: that is, whether participants can choose meaningfully among five (or more) sentences. In a sense, the Expanded format makes explicit what is also true of the Likert format: that participants may have a hard time differentiating among too many response options.

In this research, only undergraduate student participants were used. This is a meaningful starting point as researchers in many areas of psychology often use undergraduate student samples. However, examining whether the same pattern of results would emerge in other samples is important and should be pursued in future research. In addition, a thorough investigation of the validity of the measures in the new format may be warranted in some cases.

Our results are encouraging enough that we recommend researchers consider creating and using scales in the Expanded format, particularly when the factor structure of the scale is of primary theoretical interest. Many theoretical debates about the dimensionality of a particular construct (i.e., the latent constructs measured by the RSES) may in fact be resolved by removing method factors or acquiescence bias factor that emerge due to item wording in Likert scales (e.g., Bieling et al., 1998; Furnham & Thorne, 2013; Hensley & Roberts, 1976; Steed, 2001). A typical Likert scale can be fairly easily converted to the Expanded format, although rewording challenges do arise. By reducing the impact of acquiescence response bias and method effects on the structure of the scale, the Expanded format may lead to better dimensionality and better fit under the researcher's theoretical model.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Social Sciences and Humanities Research Council of Canada (SSHRC) Standard Research Grant, 410-2011-0815.

Notes

1. Some disagreement about the use of the term *Likert scale* exists (Wuensch, 2015). In this article, we use the term *Likert scale* to refer to scales where respondents express their strength of agreement with each statement or item in the scale.
2. In particular, we are excluding from consideration items that contain negations but whose endorsement corresponds to being higher on the construct, for example, ''I'm not sad'' on a happiness scale.

3. This study was conducted over the phone, and thus item wording was adapted for the item to be delivered orally.
4. The data for the different versions of the RSES and CS were collected at the same time; most of data for the different versions of the BDI were collected together with the other two scales. Specifically, 641 participants were asked to complete one version of the RSES and one version of the CS. However, 20 participants' responses on one version of the CS were deleted due to missing data. Of the 641 participants, 472 were also asked completed one version of the BDI. An additional 307 participants completed only the BDI. Of the 779 participants who were asked to complete one version of the BDI, 16 participants were deleted due to missing data. Participants with missing data were deleted because the categorical estimator used in the analyses (i.e., WLSMV estimator; see Method section) does not allow for missing data handling. Fortunately, the amount of missing data was small (2% to 3%).
5. All versions of the RSES and CS can be found in supplementary materials. However, because the BDI-II is a licensed scale, different versions of the BDI used in this study are not included in supplementary materials.
6. For the Expanded format BDI, factors were formed in two different ways: according to whether the corresponding item is RW or PW in both Likert Version I versus in Version II. The results were very similar.

References

Argyle, M., Martin, M., & Crossland, J. (1989). Happiness as a function of personality and social encounters. In J. P. Forgas & J. M. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 189-203). Amsterdam, Netherlands: Elsevier.

Asparouhov, T., & Muthen, B. (2006). *Robust chi-square difference testing with mean and variance adjusted test statistics* (Mplus Web Notes: No. 10). Retrieved from http://www.statmodel.com/download/webnotes/webnot10.pdf

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561-571. doi:10.1001/archpsyc.1961.01710120031004

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246. doi:10.1037/0033-2909.107.2.238

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

Bieling, P. J., Antony, M. M., & Swinson, R. P. (1998). The State-Trait Anxiety Inventory, Trait version: Structure and content re-examined. *Behaviour Research and Therapy*, *36*, 777-788. doi:10.1016/S0005-7967(98)00023-0

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608-628. doi:10.1207/S15328007SEM0704_5

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179-197. doi:10.1007/BF02291262

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaire. *Educational and Psychological Measurement*, *71*, 460-502. doi:10.1177/0013164410375112

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Thousand Oaks, CA: Sage.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116-131. doi:10.1037/0022-3514.42.1.116

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Cordery, J. L., & Sevastos, P. P. (1993). Response to the original and revised Job Diagnostic Survey: Is education a factor in response to negatively worded item? *Journal of Applied Psychology*, *78*, 141-143. doi:10.1037/0021-9010.78.1.141

Couch, A., & Keniston, K. (1960). Yessayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal & Social Psychology*, *60*, 151-174. doi: 10.1037/h0040372

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

Furnham, A., & Thorne, J. D. (2013). Need for cognition: Its dimensionality and personality and intelligence correlates. *Journal of Individual Differences*, *34*, 230-240. doi: 10.1027/1614-0001/a000119

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. doi:10.1016/S0191-8869(02)00331-8

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191-205. doi:10.1177/1094428104263675

Hensley, W. E., & Roberts, M. K. (1976). Dimensions of Rosenburg's Self-Esteem Scale. *Psychological Reports*, *38*, 583-584. doi:10.2466/pr0.1976.28.2.582

Hills, P., & Argyle, M. (2002). The Oxford Happiness Questionnaire: A compact scale for the measurement of psychological well-being. *Personality and Individual Differences*, *33*, 1073-1082. doi:10.1016/S0191-8869(01)00213-6

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *32*, 179-185. doi:10.1007/BF02289447

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

Jackson, D. M., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement*, *21*, 771-790. doi: 10.1177/001316446102100402

Javeline, D. (1999). Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*, *63*, 1-28. doi:10.1086/297701

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201-219. doi:10.1086/269029

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thogersen-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

Lyubomirsky, S., & Lepper, H. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*, 137-155. doi: 10.1023/A:1006824100041

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19-33. doi:10.1037/1082-989X.11.1.19

Manian, N., Schmidt, E., Bornstein, M. H., & Martinez, P. (2013). Factor structure and clinical utility of BDI-II factor scores in postpartum women. *Journal of Affective Disorders*, *149*, 259-268. doi:10.1016/j.jad.2013.01.039

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantive 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*, 361-381. doi:10.1037/a0019225

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362. doi:10.1037/1082-989X.11.4.344

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, *67*, 205-211.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, *20*, 60-103. doi:10.1177/0049124191 020001003

Muthen, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132. doi: 10.1007/BF02294210

Muthen, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Notes: No. 4). Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Pullmann, H., & Allik, J. (2000). The Rosenberg Self-Esteem Scale: Its dimensionality, stability and personality correlates in Estonian. *Personality and Individual Differences*, *28*, 701-715. doi:10.1016/S0191-8869(99)00132-4

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/s15328007 sem1301_5

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology*, *121*, 81-96. doi:10.1080/00224545.1983.9924470

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184. doi:10.1177/01466216970212006

Revelle, W. (2014). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Rodebaugh, T. L., Heimberg, R. G., Brown, P.J., Fernandez, K. C., Blanco, C., Schneier, F. R., & Liebowitz, M. R. (2011). More reasons to be straightforward: Findings and norms for two scales relevant to social anxiety. *Journal of Anxiety Disorders*, *25*, 623-630. doi: 10.1016/j.janxdis.2011.02.002

Rodebaugh, T. L., Woods, C. M., & Heimberg, R. G. (2007). The reverse of social anxiety is not always the opposite: The reverse-scored items of the Social Interaction Anxiety Scale do not belong. *Behavior Therapy*, *38*, 192-206. doi:10.1016/j.beth/2006.08.001

Rodebaugh, T. L., Woods, C. M., Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (2006). The factor structure and screening utility of the Social Interaction Anxiety Scale. *Psychological Assessment*, *18*, 231-237. doi:10.1037/1040-3590.18.2.231

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rosseel, Y (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation method under range of non-ideal situations. *Psychological Methods*, *17*, 354-474. doi: 10.1037/a0029315

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*, 113-130. doi:10.1080/02602930802618344

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijman, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233-247). London, England: Kluwer Academic.

Savalei, V., & Falk, C. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407-424. doi:10.1080/00273171.2014.931800

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*, 367-373. doi: 10.1177/014662168500900405

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context*. Orlando, FL: Academic Press.

Sonderen, E. V., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS One*, *8*(7), e68967. doi: 10.1371/journal.pone.0068967

Steed, L. (2001). Further validity and reliability evidence for Beck Hopelessness scale scores in a nonclinical sample. *Educational and Psychological Measurement*, *61*, 303-316. doi: 10.1177/00131640121971121

Steer, R. A., & Clark, D. A. (1997). Psychometric characteristics of the Beck Depression Inventory-II with college students. *Measurement and Evaluation in Counseling and Development*, *30*, 128-136. doi:10.1080/13651500510014800

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

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marking Research*, *45*, 116-131. doi:10.1509/jmkr.45 .1.116

Tafarodi, R. W., & Swann, W. B., Jr. (1995). Self-liking and self-competence as dimensions of global self-esteem: initial validation of a measure. *Journal of Personality Assessment*, *65*, 322-342. doi:10.1207/s15327752jpa6502_8

White, R. T., & Mackay, L. D. (1973). A note on a possible alternative to Likert scales. *Research in Science Education*, *3*(1), 75-76. doi:10.1007/BF02558560

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 186-191. doi:10.1007/s10862-005-9004-7

Wuensch, K. L. (2015). *How do you pronounce ''Likert?'' What is a Likert Scale?* Retrieved from http://core.ecu.edu/psyc/wuenschk/StatHelp/Likert.htm

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442. doi:10.1007/BF02289447