Educational and Psychological
Measurement
2018, Vol. 78(2) 253–271

DOI: 10.1177/0013164417690858
journals.sagepub.com/home/epm
SAGE

A Multilevel Bifactor Approach to Construct Validation of Mixed-Format Scales

Yan Wang[1], Eun Sook Kim[1], Robert F. Dedrick[1], John M. Ferron[1], and Tony Tan[1]

Abstract

Wording effects associated with positively and negatively worded items have been found in many scales. Such effects may threaten construct validity and introduce systematic bias in the interpretation of results. A variety of models have been applied to address wording effects, such as the correlated uniqueness model and the correlated traits and correlated methods model. This study presents the multilevel bifactor approach to handling wording effects of mixed-format scales used in a multilevel context. The Students Confident in Mathematics scale is used to illustrate this approach. Results from comparing a series of models showed that positive and negative wording effects were present at both the within and the between levels. When the wording effects were ignored, the within-level predictive validity of the Students Confident in Mathematics scale was close to that under the multilevel bifactor model. However, at the between level, a lower validity coefficient was observed when ignoring the wording effects. Implications for applied researchers are discussed.

Keywords

wording effects, multilevel bifactor model, mixed-format scales, construct validation

Scales have been widely used in research and evaluation to measure constructs of interest. A mixed-format scale is one that includes a combination of positively and negatively worded items. From an instrument design perspective, mixed-format

[1] University of South Florida, Tampa, FL, USA

Corresponding Author:
Yan Wang, Department of Educational and Psychological Studies, University of South Florida, 4202 E. Fowler Ave. EDU 105, Tampa, FL 33620-7750, USA.
Email: yanw@mail.usf.edu

scales have the potential to reduce bias caused by respondents' tendency to agree with items regardless of the content (i.e., acquiescence bias; DeVellis, 2016). Because negatively worded items tend to keep respondents more engaged in processing information conveyed by items, they can potentially enhance construct validity (Y. Chen, Rendina-Gobioff, & Dedrick, 2007; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

The underlying assumption of mixed-format scales is that both positively and negatively worded items measure a given construct in an equivalent way (Marsh, 1996). Two main concerns have been raised about this assumption. The first concern lies in distinctive response patterns between positively and negatively worded items (Wang, Chen, & Jin, 2015). For example, responses to positively worded items were found to be significantly higher than negatively worded items, indicating that respondents were more likely to agree with positively worded items than to disagree with negatively worded items (Weems, Onwuegbuzie, Schreiber, & Eggers, 2003). Such systematic bias might distort interpretations of results (Chessa & Holleman, 2007; Horan, DiStefano, & Motl, 2003).

The second concern lies in the factor structure of mixed-format scales. Mixed-format scales that have been designed to measure a single construct can sometimes have several alternative factor structures to the one-factor model. For instance, there has been disagreement over the factor structure of the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1989), which is one of the most widely used scales in psychology. A two-factor structure has been suggested by some studies, with positively and negatively worded items loading onto two separate factors, positive self-esteem and negative self-esteem (e.g., Boduszek, Hyland, Dhingra, & Mallett, 2013; Kaufman, Rasinski, Lee, & West, 1991). Alternative factor structures of the RSES have also been proposed and examined in the extant literature, such as the correlated uniqueness model where residuals of positively worded items and residuals of negatively worded items are correlated, respectively, and the correlated traits and correlated methods (CTCM) model where the positive wording factor and the negative wording factor are correlated (Tomás & Oliver, 1999). The correlated uniqueness and CTCM models with only the negative wording effect taken into account have been supported as well (Horan et al., 2003). In sum, there is sufficient evidence that positively and negatively worded items do not measure a given construct equivalently. In order to ensure unbiased parameter estimates and accurate interpretations of scores in mixed-format scales, wording effects should be taken into account.

Recently, bifactor models have been shown as the best-fitting model for mixed-format scales (e.g., Hyland, Boduszek, Dhingra, Shevlin, & Egan, 2014; McKay, Boduszek, & Harvey, 2014). That is, a substantive factor is measured by all items and two specific factors (i.e., positive and negative wording effects) are measured by positively and negatively worded items, respectively (see Figure 1). Bifactor structures have been used to model multidimensionality of measures when all items are expected to measure a common construct while a subset of items measure a subdomain construct in addition to the common trait (Reise, 2012). A bifactor structure is

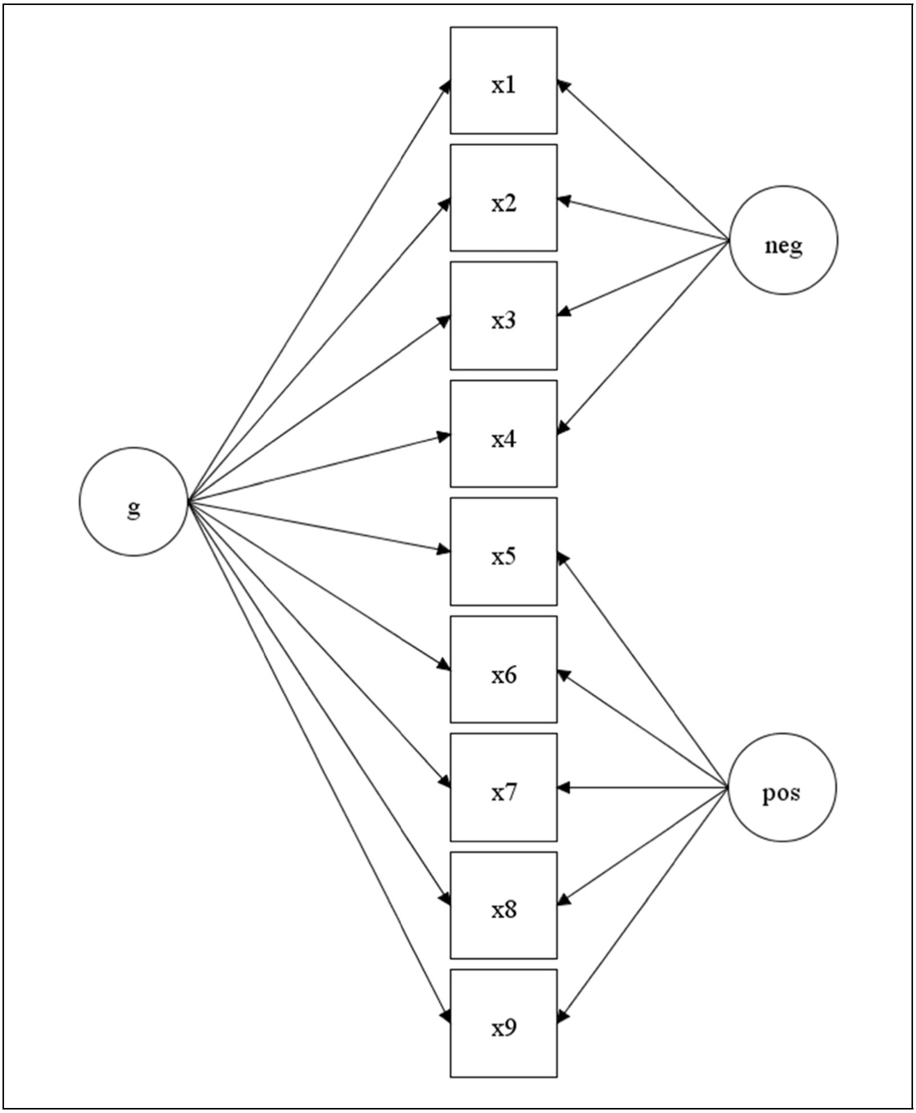

Figure 1. An example of bifactor modeling to address wording effects.
Note. g = the general factor; neg = the negative wording effect; pos = the positive wording effect. All factors are orthogonal to each other. For simplicity error terms associated with x1 to x9 are not shown on the diagram.

suitable for mixed-format scales because item variances can be easily partitioned into two sources, the substantive construct and wording effects. The independent contribution of each source can be evaluated based on the strength of standardized factor loadings (Reise, Moore, & Haviland, 2010). This feature of bifactor modeling can be valuable in the development of mixed-format scales and psychometric analyses to disentangle the contributions of the substantive construct and wording effects.

The current study will discuss how the bifactor approach handles wording effects within a multilevel context. Because data dependency can occur when individuals are nested within contexts, multilevel modeling is adopted to take such nested data into account. With nested data, the independence of observations assumption underlying confirmatory factor analysis (CFA) is violated. The single-level CFA analytic procedures might lead to biased results and incorrect conclusions, because ignoring the data dependency would underestimate standard errors of parameter estimates and inflate Type I error rates (e.g., Hox, 2010). Methods that are commonly used in the literature to take into account the nested data structure include a design-based approach that adjusts standard errors and a model-based approach that involves multilevel analysis. This study will apply the model-based multilevel CFA approach for two conceptual reasons.

First, between-level (e.g., between-classroom where students are nested within classrooms) variations in constructs can be of interest. Increasing evidence has shown individual and group differences in constructs such as self-esteem and self-efficacy, which calls for the need to further investigate the significant role of social contexts (e.g., Eccles & Roeser, 2011). Social context is also important because many items explicitly reference a social context. For example, the RSES scale includes the item ''I feel that I am a person of worth, at least on an equal plane with others,'' and the Students Confident in Mathematics (SCM) scale (TIMSS & PIRLS International Study Center, 2011) includes the item ''Mathematics is more difficult for me than for many of my classmates.'' Thus, individuals' responses to those items might be affected by the social environment they are immersed in (e.g., classroom) and various social comparisons they make. Examining the factor structure of psychological constructs at the between level enables researchers to address research questions related to between-level variations and the potential impact and/or causes of such variations, and to build multilevel theories (Zimprich, Perren, & Hornung, 2005).

Second, although positive and/or negative wording effect factors have been identified in mixed-format scales with independent observations, wording effects at the between level have not been investigated systematically. The exploration of the existence and interpretation of wording effects at the between level can potentially further our understanding of wording effects. For instance, wording effects might be interpreted as individual response styles rather than method artifacts, based on the relationship between wording effects and some individual characteristics (e.g., fear of negative evaluations by others, self-consciousness, behavioral inhibition, anxiety, and reading ability and verbal reasoning; DiStefano & Motl, 2006, 2009; Dunbar, Ford, Hunt, & Der, 2000; Tomás, Oliver, Galiana, Sancho, & Lila, 2013; Weems et al., 2003; Ye, 2009). However, group-level response styles associated with wording effects have not been studied. By contrast, other types of response styles (such as extreme response styles) have been found to have group-level differences, especially cross-cultural differences (e.g., Johnson, Kulesa, Cho, & Shavitt, 2005). For example, Scherer and Gustafsson (2015) modeled response style at both the student level and the classroom level. Their purpose was to take into account the differences in response styles due to individual and classroom/cultural variations, when modeling the factor structure of students' assessments of teaching quality across three countries. Results showed high loadings on the classroom-level response style, indicating that students responded to items differently across classrooms. Therefore, it is worthwhile exploring if wording effects are present at the between level. If present, researchers can further examine the relationship between wording effects and group-level characteristics and the possibility of conceptualizing response styles as a group characteristic. Of note is that such investigation into the substantive interpretations of wording effects at the between level is beyond the scope of this study. Instead, it focuses on demonstrating how the multilevel bifactor model can be applied to examine the presence of wording effects at the between level.

Bifactor Modeling

The bifactor model was first introduced by Holzinger and Swineford (1937) as an exploratory approach to factor analyze a large test battery of abilities. In a typical bifactor model, each item loads onto a general factor, which is the common trait measured by a scale (or scales), and at most one specific factor that a well-defined set of items aims to measure in addition to the common trait. No cross-loadings are allowed in bifactor models. It is assumed that all factors are uncorrelated (i.e., orthogonal) to each other. An example of a bifactor pattern matrix can be written as

$$\Lambda = \begin{pmatrix} \lambda_{1g} & \lambda_{11} & 0 & 0 & \cdots & 0 \\ \lambda_{2g} & \lambda_{21} & 0 & 0 & \cdots & 0 \\ \lambda_{3g} & 0 & \lambda_{32} & 0 & \cdots & 0 \\ \lambda_{4g} & 0 & \lambda_{42} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \lambda_{(J-1)g} & 0 & 0 & 0 & 0 & \lambda_{(J-1)K} \\ \lambda_{Jg} & 0 & 0 & 0 & 0 & \lambda_{JK} \end{pmatrix}.$$

Here each indicator j (j = 1 to J) has a nonzero loading on the general factor g and another loading on one of k (k = 1 to K) specific factors. For example, indicators 1 and 2 load on the first specific factor, whereas indicators 3 and 4 load on the second specific factor. Assuming that responses to all items are continuous, a linear bifactor model can be expressed as

$$x_i = \tau + \Lambda \xi_i + \delta_i,$$

where x_i is a $J \times 1$ response vector of person i, and τ is the intercept vector; Λ refers to the bifactor pattern matrix (example shown above); ξ_i is the vector of latent factor scores for person i, which consists of the general factor score (ξ_{ig}) and group factor scores ($\xi_{i1}, \xi_{i2}, \ldots, \xi_{iK}$); and δ_i is the vector of item residuals for person i. Due to the orthogonality of the general and the specific factors, a person's response vector (x_i) can be decomposed to two parts, attributable to the general factor score $\left(\xi_{ig}\right)$ and the group factor scores ($\xi_{i1}, \xi_{i2}, \ldots, \xi_{iK}$), respectively.

Recent years have witnessed a substantial increase in applications of bifactor models in the fields of education and psychology. For instance, bifactor models have been used to represent the underlying factor structure of testlet-based tests, where the testlet refers to a group of items that share the same stimulus and enhance the efficiency of test construction (Cho, Cohen, & Kim, 2014; Rijmen, 2010). For example, in a reading test with multiple items sharing the same passage as the prompt, each item measures both students' reading comprehension and their content knowledge. Additionally, many applications of bifactor models can be found in the field of psychology involving the psychometric analyses of scales, such as intelligence (Beaujean, Parkin, & Parker, 2014), quality of life (F. Chen, West, & Sousa, 2006; Deng, Guyer, & Ware, 2015), depression (Yang & Jones, 2008; Yang, Tommet, & Jones, 2009), attention-deficit hyperactivity disorder (Matte et al., 2015; Toplak et al., 2012), anxiety (Allan, Albanese, Short, Raines, & Schmidt, 2015), fatigue (Varni, Beaujean, & Limbers, 2013), and personality (MacAbee, Oswald, & Connelly, 2014).

Bifactor Modeling and Mixed-Format Scales

In addition to the applications of bifactor modeling discussed above, there has been growing interest in the use of bifactor models to address wording effects in mixed-format scales. For example, Marsh, Scalas, and Nagengast (2010) used longitudinal data to examine the factor structure of the RSES scale. The longitudinal approach provided sufficient evidence that supported the bifactor model with self-esteem as the primary construct and positively and negatively worded items loading onto the positive wording effect and negative wording effect factors, respectively. Longitudinal measurement invariance tests also showed the stability of method effects over time, including the bifactor structure, item loadings and intercepts, and means of method effects. The temporal stability of method effects supported the interpretation of method effects as response styles that were substantively meaningful. Similarly, McKay et al. (2014) examined the factor structure of the RSES with a sample of high school students by testing four competing solutions to wording effects, including a one-factor model, a two-factor model, a second-order model, and a bifactor model. The bifactor model was also supported. Additionally, based on comprehensive evaluations of item loadings, reliability, and correlations with other measures, the authors concluded that the RSES was a unidimensional measure of self-esteem and positive and negative wording effects served as two nuisance dimensions. The bifactor modeling approach was further confirmed by Hyland et al. (2014) who used an adult cross-sectional sample.

A slightly different bifactor conceptualization has been provided by other studies. Gu, Wen, and Fan (2015) modeled the wording effect in the Core Self-Evaluation Scale using a Chinese sample and suggested a bifactor model with only a negative wording effect factor associated with negatively worded items. They also found that when wording effects were ignored, the predictive validity coefficients of the scale were underestimated when correlations between core self-evaluation and the criterion variables were positive and overestimated when correlations were negative. The bifactor model with only one specific factor for negatively worded items has been supported within the item response theory framework as well by Wang et al. (2015). Ignoring wording effects was found to result in overestimated scale reliability and biased estimates of person measures (i.e., the general factor scores). Of note is that the bifactor model with only one wording effect factor associated with either positively or negatively worded items is equivalent to the CTCM framework when only the positive method effect or the negative method effect is included. The CTCM model with only the negative wording effect taken into account has been shown to have good fit to data from mixed-format scales, including the RSES, Attitude Toward School, and Locus of Control, which were all included in the National Educational Longitudinal Study (Horan et al., 2003).

Zimprich et al. (2005) examined the factor structure of a modified RSES within a multilevel context and identified a bifactor model with only the negative wording effect factor associated with negatively worded items at the within level, and a unidimensional model at the between level. In other words, the negative wording effect occurred only at the within level, whereas the self-esteem factor itself captured the variations across between-level units (i.e., school classes). Although this study extended the investigation into the factor structure of mixed-format scales to a multilevel context, only a limited number of models were examined in the study, compared with a variety of models examined within the single-level context in the extant literature.

Along this line of research, the major purpose of this study is to demonstrate the multilevel bifactor approach, in comparison to a series of alternative models, to handling wording effects with a nested data structure. Second, the presence of negative and/or positive wording effects at the within and between levels is also examined through model comparisons. Third, this study shows the impact of ignoring the wording effects on the predictive validity of the mixed-format scales at the within and between levels. These research purposes are achieved through an empirical example using multilevel data (students nested within classrooms) obtained from the administration of the SCM scale as part of the 2011 Trends in International Mathematics and Science Study (TIMSS).

Method

Data Source

Data were drawn from the 2011 TIMSS USA eighth-grade sample, with a total student population of 10,477. TIMSS assesses mathematics and science achievements of fourth- and eighth-graders across participating countries and educational systems. The 2011 TIMSS was the fifth assessment, and over 57 countries and 20 educational systems participated. Two-stage sampling was implemented with schools randomly selected at Stage 1, and one or more classes were then randomly selected from each school. The sample for this study included 10,416 students who were nested within 531 classrooms. Classroom was treated as the between-level unit of analysis. The average classroom size was 20 (*Max* = 38, *Min* = 5). Students were aged 14.40 years on average (*SD* = 0.56, *Max* = 18, *Min* = 12). Half (50.62%) were females.

Measures

The SCM scale consists of nine items, with four items negatively worded. It was administered as part of a student questionnaire that required 15 to 30 minutes to complete. All items were rated on a 4-point Likert-type scale ranging from 1 (*Agree a lot*) to 4 (*Disagree a lot*). Example items included ''I learn things quickly in mathematics'' and ''Mathematics makes me confused and nervous.'' It should be pointed out that based on the response options, a higher value of responses to positively worded items is associated with a lower confidence level in mathematics. Therefore, responses to positively worded items were reverse coded to ensure more straightforward interpretations of the scale. The revised 4-point Likert-type scale ranged from 1 (*Disagree a lot*) to 4 (*Agree a lot*). The criterion variable of this study was students' math achievement score. Students' math achievements were assessed with a 45-minute test, consisting of both multiple-choice and constructed-response questions. Scores were computed as the average of five plausible score values which were provided by TIMSS, using the TOTWGT sampling weight (Foy, Arora, & Stanco, 2013). Both the student questionnaire (including the SCM scale) and the achievement assessment were conducted toward the end of the school year across participating countries and educational systems.

Analytic Procedures

Statistical analyses were conducted using M*plus* 7.3 (Muthén & Muthén, 1998-2014). Because both maximum likelihood estimation with robust standard errors (MLR) and weighted least squares means and variances adjusted (WLSMV) have been used with ordinal data, preliminary analysis was conducted to compare the results of MLR and WLSMV. The same pattern of relative model fit was observed across these two estimation methods. Results reported in this study were based on the MLR estimation, in order to facilitate the interpretations of findings in comparison to the extant literature on wording effects, which was primarily based on treating responses as continuous.

The multilevel bifactor approach to handling wording effects was evaluated through a series of model comparisons. Nine models being compared are described below:

Model 1: A unidimensional model at the within and between levels, with all items loadings onto SCM.

Model 2: A bifactor model at the within level with the negative wording effect factor associated with negatively worded items; and a unidimensional model at the between level.

Model 3: a bifactor model at the within level with the positive wording effect factor associated with positively worded items; and a unidimensional model at the between level.

Model 4: A bifactor model at the within level with the positive and negative wording effect factors; and a unidimensional model at the between level.

Model 5: A bifactor model with the negative wording effect factor at both levels.

Model 6: A bifactor model with the positive wording effect factor at both levels.

Model 7: A bifactor model with the positive and negative wording effect factors at both levels.

Model 8: A two-factor model at the within level that has positively worded items loading onto confidence in mathematics and negatively worded items loading onto self-doubt, and the two factors, confidence and self-doubt, are correlated; a unidimensional model that has a single factor, SCM, at the between level.

Model 9: The same two-factor model described in Model 8 at both within and between levels.

It should be pointed out that originally multilevel correlated uniqueness models were included in the study (i.e., a correlated uniqueness model with errors of positively and negatively worded items correlated with each other, respectively, at the within level and a unidimensional model at the between level; and a correlated uniqueness model at both levels). However, these models were dropped due to inadmissible solutions. This might be caused by empirical underidentification when there are a substantially greater number of parameters to be estimated in the correlated uniqueness model than the other models (e.g., unidimensional, two-factor, and bifactor models; Marsh et al., 2010). Of note is that because determining the multilevel factor structure with the presence of wording effects is exploratory in nature, equal and distinctive factor structures across levels were specified and evaluated for each of the models being compared. Prior to conducting multilevel CFA, the four-step procedure (see Muthén, 1994) was employed to model factor structures at different levels of analysis. The first step was to perform the conventional factor analysis on the sample total covariance matrix; the second step was the estimation of between-level variation and the calculation of the intraclass correlation coefficient for each item; it was followed by factor analysis of the sample pooled-within covariance matrix (Step 3) and the sample between-level covariance matrix (Step 4). The four steps can provide justification of multilevel analysis and initial information regarding the fit of factor structures.

The nine models were evaluated based on model fit indices, including chi-square goodness of fit, comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean squared residual (SRMR within and between). Recommended cutoff values of those indices for good model fit are statistically nonsignificant chi-square ($p \geq .05$), CFI $\geq$ .95, RMSEA $\leq$ .06, and SRMR $\leq$ .08 (Hu & Bentler, 1999). As models evaluated in this study were not all nested within each other, they were compared based on Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample-size adjusted BIC (SaBIC). Smaller values of indices indicate better model fit. The adequacy of models was also evaluated by checking parameter estimates based on substantive and conceptual expectations (DiStefano & Motl, 2006; Marsh, 1996). Additionally, based on the magnitude of standardized factor loadings, the relative strength of the general factor to wording effects was evaluated, which was indexed by explained common variance (ECV). Larger ECV (closer to 1) suggests a stronger general factor. The contributions of the general factor and wording effects to item variances were also evaluated, indicated by coefficient omega hierarchical (ω_H) and omega subscale (ω_{POS} and ω_{NEG}), respectively (Reise, 2012; Reise et al., 2010). The higher coefficient omega is, the more item responses reflect the corresponding dimension.

After the best-fitting model was selected, the predictive validity of the SCM scale was examined by estimating the correlation between students' confidence in mathematics and their math achievement scores. These results were compared with the predictive validity of the scale when wording effects were ignored.

Prior to conducting statistical analyses, missing data were examined. For items of the SCM scale, missingness ranged from 1.42% to 2.33% (out of 10,416 cases), while no missing data were observed for the math achievement scores. The MLR estimation takes into account the missing data and uses all available data.

Results

Descriptive Statistics

Descriptive statistics of each item of the SCM scale and math achievement scores are presented in Table 1. Overall, eighth-graders in this sample had a relatively high confidence level in math. Item responses were approximately normally distributed. Item intraclass correlation coefficients (ICCs) ranged from .04 to .10, with an average ICC of .08. This indicates that on average, 92% of the total item variance could be attributed by the within-classroom individual differences in math confidence, while differences across classrooms accounted for 8% of the total item variance. By contrast, the ICC of the math achievement score was substantially higher (.65). This uncommonly high ICC might be explained by the sampling procedure of TIMSS that used eighth-grade classes that were tracked by students' performance. As expected, this would increase the between classroom variations in math achievement scores (Joncas & Foy, 2013).

Table 1. Descriptive Statistics of Variables Involved in the Study.

Variable	*N*	*M*	*SD*	Skewness	Kurtosis	ICC
Usually do well in math*	10,268	3.24	0.83	0.97	0.34	.12
Math is more difficult	10,263	2.83	1.03	−0.40	−0.10	.04
Math is not my strength	10,197	2.60	1.15	−0.12	−1.42	.10
Learn quickly in math*	10,173	2.88	0.95	0.42	−0.80	.07
Feel confused and nervous	10,219	2.88	0.99	−0.40	−0.95	.06
Good at working out problems*	10,222	2.66	0.98	0.23	−0.94	.07
Teacher thinks I can do well in math*	10,209	3.00	0.92	0.62	−0.48	.08
Teacher tells me I am good at math*	10,214	2.78	1.03	0.39	−0.99	.10
Math is harder for me	10,251	2.81	1.17	−0.41	−1.33	.08
Math achievements	10,416	509.81	73.98	−0.05	−0.20	.65

Note. ICC = intraclass correlation coefficient.

*Items were reversely coded to ensure that higher values were associated with higher confidence levels. Items of the Students Confident in Mathematics scale ranged from 1 (*Disagree a lot*) to 4 (*Agree a lot*). Math achievements ranged from 251.08 to 735.22.

Multilevel Confirmatory Factor Analyses

Prior to multilevel CFA, the four-step procedure proposed by Muthén (1994) was performed and the results supported the multilevel analysis (detailed results are omitted for brevity but are available on request). Table 2 presents the results of the nine multilevel models compared in this study to address wording effects in the SCM scale. Overall, Model 1 (the unidimensional model at both levels) did not fit the data as well as the models that took wording effects into account at either the within or the between level or both levels (Models 2-9, except Model 6 that did not converge). Comparing those models, Models 4 and 7 demonstrated better model fit, as indicated by higher CFI, lower RMSEA, lower SRMR_within, and lower AIC, BIC, and SaBIC values (see Table 2). Model 4 consisted of the bifactor model at the within level with the positive and negative wording effect factors and a unidimensional model at the between level. Model 7 refers to the bifactor model with positive and negative wording effect factors at both levels. Comparing Models 4 and 7, Model 7 which modeled wording effects at both levels, provided better model fit based on a slight increase in CFI, a relatively substantial decrease in SRMR_between, and smaller values of AIC, BIC, and SaBIC.

Parameter estimates of Model 7 were further examined to ensure appropriate interpretations of results (see Table 3). Overall, the general factor loadings were stronger than loadings of wording effects, as indicated by high ECV (.68 and .86 for the within and between levels, respectively). Item responses were dominated by the general factor ($\omega_H = .76, \omega_{POS} = .19, \omega_{NEG} = .30$ at the within level, and $\omega_H = .93, \omega_{POS} = .14, \omega_{NEG} = .07$ at the between level). Specifically, for the general factor of SCM, standardized factor loadings ranged from .43 to .82 at the within level and from .57 to .97 at the between level. Factor loadings of all negatively

Table 2. Confirmatory Factor Analyses Results for Each Model Compared in the Study.

Model	Chi-square	*df*	CFI	RMSEA	SRMR_ within	SRMR_ between	AIC	BIC	SaBIC
Model 1	6310.58	54	.83	.11	.08	.07	219,428	219,754	219,611
Model 2	2048.85	50	.95	.06	.05	.07	215,154	215,509	215,353
Model 3	1103.26	49	.97	.05	.02	.07	214,110	214,472	214,313
Model 4	403.41	45	.99	.03	.01	.08	213,386	213,777	213,606
Model 5	2022.50	46	.95	.07	.05	.05	215,137	215,521	215,353
Model 6					non-convergence				
Model 7	322.94	36	.99	.03	.01	.02	213,306	213,762	213,561
Model 8	2094.53	53	.95	.06	.05	.07	215,212	215,545	215,399
Model 9	2093.31	52	.95	.06	.05	.06	215,210	215,550	215,401
Model 10	616.78	44	.99	.04	.03	.03	213,634	214,032	213,858

Note. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; SaBIC = sample-size adjusted Bayesian information criterion. Model 1 = students' confidence in math (SCM) at both levels; Model 2 = bifactor (SCM and negative wording effect) at the within level and SCM at the between level; Model 3 = bifactor (SCM and positive wording effect) at the within level and SCM at the between level; Model 4 = bifactor (SCM, positive and negative wording effects) at the within level and SCM at the between level; Model 5 = bifactor (SCM and negative wording effect) at both levels; Model 6 = bifactor (SCM and positive wording effect) at both levels; Model 7 = bifactor (SCM, positive and negative wording effects) at both levels; Model 8 = correlated factors (math confidence measured by positively worded items, and math self-doubt measured by negative worded items) at the within level and SCM at the between level; Model 9 = correlated factors (math confidence and math self-doubt) at both levels; Model 10 = bifactor (SCM and negative wording effect) and correlated errors between Items 7 and 8 at both levels.

worded items at the within level were substantial (over .40), but the between-level factor loadings of two negative items became nonsignificant, that is, .03 and .41 for Items 3 and 5, respectively. In other words, at the within level, both students' confidence in math and the negative wording effect contributed substantially to the variance of the negatively worded items. However, at the between level, the general factor—math confidence was dominant in the aggregated responses of negatively worded items. Compared with general factor loading strength, factor loadings of positively worded items were not substantial at both levels, except Items 7 and 8 (.77 and .53 for Item 7 at the within and between levels, .47 and .71 for Item 8 at the within and the between levels, respectively). That is, both math confidence and positive wording effect contributed substantially to the variance of Items 7 and 8, whereas for the other positively worded items, math confidence was much more strongly associated with the item responses than positive wording effect.

Given that among positively worded items, only Items 7 and 8 had substantial factor loadings across levels, an additional model (Model 10) was constructed, with the factor structure of the positive wording effect modified based on Model 7. That is, Items 1, 4, and 5 were removed from the positive wording effect structure and

Table 3. Standardized Parameter Estimates of the Students Confident in Mathematics Scale With the Multilevel Bifactor Model (Model 7).

Item	SCM	Positive wording effect	Negative wording effect
Within level			
Usually do well in math	.72 (.01)	.14 (.02)	
Math is more difficult	.59 (.01)		.48 (.02)
Math is not my strength	.68 (.01)		.45 (.01)
Learn quickly in math	.82 (.01)	.11 (.02)	
Feel confused and nervous	.57 (.01)		.40 (.02)
Good at working out problems	.72 (.01)	.21 (.02)	
Teacher thinks I can do well in math	.43 (.01)	.77 (.04)	
Teacher tells me I am good at math	.48 (.01)	.47 (.03)	
Math is harder for me	.65 (.01)		.51 (.01)
Between level			
Usually do well in math	.95 (.02)	.12 (.04)	
Math is more difficult	.93 (.03)		.29 (.10)
Math is not my strength	.97 (.02)		.03[a] (.15)
Learn quickly in math	.95 (.02)	.21 (.09)	
Feel confused and nervous	.91 (.05)		.41[a] (.22)
Good at working out problems	.96 (.03)	.11[a] (.12)	
Teacher thinks I can do well in math	.75 (.05)	.53 (.15)	
Teacher tells me I am good at math	.57 (.05)	.71 (.17)	
Math is harder for me	.92 (.02)		.28 (.07)

Note. SCM = Students Confidence in Mathematics. Model was identified by fixing the latent factor variance to 1 at both the within and between levels. Standard errors of factor loadings are reported within parentheses.

[a]Factor loadings are not statistically significant at .05.

residuals of Items 7 and 8 were correlated. Compared to Model 7, Model 10 had a worse fit (see Table 2), based on larger values for the AIC, BIC, and SaBIC. Therefore, we concluded that Model 7 was the best-fitting model to address the wording effects of the SCM scale within the multilevel context.

Predictive Validity of the SCM Scale

Predictive validity of the SCM scale scores was examined by incorporating a criterion variable, math achievement scores. The impact of ignoring wording effects on the predictive validity was investigated by correlating SCM with math achievements under Model 7 (with wording effects taken into account) and Model 1 (with wording effects ignored). It should be noted that a transformation was applied by dividing math achievement raw scores by 100, to address the model nonconvergence issue caused by different scales of the SCM measure and the criterion variable. When fitting the multilevel bifactor model (Model 7), the correlation coefficients between the

SCM and math achievement scores were .56 and .68 for the within and between levels, respectively. Under the unidimensional model (Model 1), the correlation coefficients were .55 and .61 at the within and between levels, respectively. Although the predictive validity coefficients were close at the within level, the coefficients differed more substantially at the between level, with lower coefficient estimates when ignoring the wording effects in the SCM scale.

Discussion

Recently, the potential of using bifactor models to address wording effects has been studied (e.g., Hyland et al., 2014; Marsh et al., 2010; McKay et al., 2014). The bifactor model is specified as having one general factor measured by all items and two specific factors consisting of positively and negatively worded items, respectively (or only a single specific factor formed by negatively worded items), in which all factors are orthogonal. The present study extended the investigation into the factor structure of mixed-format scales to a multilevel context. The comparisons of different specifications of multilevel bifactor models and alternative models supported the multilevel bifactor model with one general factor and the positive and negative wording effect factors at both the within and between levels.

In the multilevel bifactor model, item variances can be partitioned into the within and between levels, and each variance component is then decomposed into the general factor and wording effects. The independent contributions of the general factor and wording effects can be evaluated at each level based on the strength of standardized factor loadings, because all factors are orthogonal (Reise, 2012; Reise et al., 2010). In the current analysis, although negatively worded items had lower loadings than the general factor loadings, the negative wording effect was not trivial, especially at the within level. Positively worded items reflect primarily the general factor (i.e., students' confidence in mathematics), with a few exceptions where the positive wording effect contributed to the item variances equally or more than the general factor. Results supported the use of multilevel bifactor models to take into account the wording effects.

Interestingly, when wording effects were ignored at both levels (fitting a two-level unidimensional model), the predictive validity coefficients were close at the within level, while the coefficients differed more substantially at the between level. That is, between-level coefficients were lower when ignoring the wording effects in the SCM scale. However, it is not clear whether such discrepancy in the validity coefficients at the between-level will be observed regardless of the strength of the wording effects. Or it might depend on several factors, such as the relative strength of wording effects across levels, the degree of data dependency (ICC), the sample size at each level, the factor structure, and so on. Also of note is that when fitting the two-level unidimensional model, both positive and negative wording effects were ignored. Thus, how ignoring each of the wording effects plays a role in the estimation of coefficients is another question. Future simulation studies are called for to fully examine the impact of ignoring wording effects on correlations or structural relationships among factors in the multilevel context.

The discovery of positive and negative wording effects at the between level provides a new perspective of investigating wording effects, which can potentially further our understanding of the nature of wording effects. Researchers have identified some potential explanations for wording effects (e.g., personality traits), supporting the interpretation of wording effects as individual response styles (DiStefano & Motl, 2006, 2009; Dunbar et al., 2000; Tomás et al., 2013; Weems et al., 2003; Ye, 2009). With the presence of wording effects at the aggregate level, wording effects might also reflect group response styles. Thus, aggregate-level characteristics, in addition to individual-level factors, can be taken into account, such as classroom culture, teacher–student relationships, school climate, social value orientation, and so on. Along with this line of research, multilevel bifactor models could be advantageous in that they allow wording effects to be examined as a distinctive factor that is separated from the general factor. This facilitates the investigation of relationships between wording effects and other variables, with the multilevel bifactor model extended to the multilevel structural equation model.

Another implication of this study is the possible presence of multiple wording effects for the same item. It was observed in this study that positive wording effects were not as strong as the general factor across levels, except for two items. Close examination of the item wording revealed that both items involved the perspective of the teacher, that is, the teacher's perception of students' ability in math. More precisely, those two items measure how students think their teachers perceive their ability in math, while the other items are about students' self-perceptions of their math ability. Therefore, it is possible that item responses were partially attributable to another type of wording effect due to the involvement of an additional subject besides respondents in item wording. Future research is called for to investigate this type of item wording and the potential wording effect it might cause.

Overall, this study demonstrated the application of the multilevel bifactor model to address wording effects across levels. There are several recommendations for applied researchers. First, applied researchers should be aware of potential wording effects caused by the presence of both positively and negatively worded items. Psychometric analyses can be conducted to model the underlying factor structure of mixed-format scales with wording effects taken into account. When multilevel constructs are present, multilevel analyses should be performed. The dimensionality and validity of mixed-format scales across the individual and the aggregate levels can thus be examined. Second, the multilevel bifactor approach has been demonstrated as a potential representation of the factor structure of mixed-format scales and is thus recommended to be included in the psychometric analyses. The evaluation of independent contributions of the primary construct and wording effects could help applied researchers tease out wording effects from the measurement of the primary construct. This would increase the likelihood of accurate interpretations of research findings.

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) received no financial support for the research, authorship, and/or publication of this article.

References

Allan, N. P., Albanese, B. J., Short, N. A., Raines, A. M., & Schmidt, N. B. (2015). Support for the general and specific bifactor model factors of anxiety sensitivity. *Personality and Individual Differences*, *74*, 78-83. doi:10.1016/j.paid.2014.10.003

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattel-Horn-Carroll factor models: Differences between bifactor and higher order factor models in predicting language achievement. *Psychological Assessment*, *26*, 789-805. doi:10.1037/a0036745

Boduszek, D., Hyland, P., Dhingra, K., & Mallett, J. (2013). The factor structure and composite reliability of the Rosenberg Self-Esteem scale among ex-prisoners. *Personality and Individual Differences*, *55*, 877-881. doi:10.1016/j.paid.2013.07.014

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

Chen, Y.-H., Rendina-Gobioff, G., & Dedrick, R. F. (2010). Factorial invariance of a Chinese self-esteem scale for third and sixth grade students: Evaluating method effects associated with the use of positively and negatively worded items. *International Journal of Psychological and Educational Assessment*, *6*(1), 21-35. Retrieved from https://sites.google.com/site/tijepa2012/articles/vol-6-1

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology*, *21*, 203-225. doi:10.1002/acp.1337

Cho, S. J., Cohen, A. S., & Kim, S. H. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling*, *21*, 375-395. doi:10.1080/10705511.2014.915371

Deng, N., Guyer, R., & Ware, J. E. (2015). Energy, fatigue, or both? A bifactor modeling approach to the conceptualization and measurement of vitality. *Quality of Life Research*, *24*, 81-93. doi:10.1007/s11136-014-0839-9

DeVellis, R. F. (2016). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

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

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). Question wording effects in the assessment of global self-esteem. *European Journal of Psychological Assessment*, *16*, 13-19. doi: 10.1027//1015-5759.16.1.13

Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, *21*, 225-241. doi:10.1111/j.1532-7795.2010.00725.x

Foy, P., Arora, A., & Stanco, G. M. (Eds.). (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill, MA: Boston College.

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41-54. doi: 10.1007/BF02287965

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, *10*, 435-455. doi: 10.1207/S15328007SEM1003_6

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

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

Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A bifactor approach to modelling the Rosenberg Self Esteem Scale. *Personality and Individual Differences*, *66*, 188-192. doi:10.1016/j.paid.2014.03.034

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264-277. doi:10.1177/0022022104272905

Joncas, M., & Foy, P. (2013). Sample design in TIMSS and PIRLS. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS* (pp. 1-21). Chestnut Hill, MA: Lynch School of Education, Boston College: TIMSS and PIRLS International Study Centre.

Kaufman, P., Rasinski, K. A., Lee, R., & West, J. (1991). *National Education Longitudinal Study of 1988. Quality of the responses of eighth-grade students in NELS88*. Washington, DC: US Department of Education.

MacAbee, S. T., Oswald, F. L., & Connelly, B. S. (2014). Bifactor models of personality and college student performance: A broad versus narrow view. *European Journal of Personality*, *28*, 604-619. doi:10.1002/per.1975

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

Matte, B., Anselmi, L., Salum, G. A., Kieling, C., Gonçalves, H., Menezes, A., & . . .Rohde, L. A. (2015). ADHD in DSM-5: A field trial in a large, representative sample of 18- to 19-year-old adults. *Psychological Medicine*, *45*, 361-373. doi:10.1017/S0033291714001470

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

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376-398. doi:10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (1998-2014). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Muthén & Muthén.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, M. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879-903. doi:10.1037/0021-9010.88.5.879

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., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544-559. doi:10.1080/00223891.2010.496477

Rijmen, F. (2010). Formal relations and an empirical comparison among the bifactor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361-372. doi:10.1111/j.1745-3984.2010.00118.x

Rosenberg, M. (1989). *Society and the adolescent self-image*. Middletown, CT: Wesleyan University Press.

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structure equation modeling. *Frontiers in Psychology*, *6*, 1-15. doi: 10.3389/fpsyg.2015.01550

TIMSS & PIRLS International Study Center. (2011). *TIMSS 2011 student questionnaire <Grade 8>*. Retrieved from http://timssandpirls.bc.edu/timss2011/downloads/T11_StuQ_8.pdf

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

Toplak, M. E., Sorge, G. B., Flora, D. B., Chen, W., Banaschewski, T., Buitelaar, J., & . . .Faraone, S. V. (2012). The hierarchical factor model of ADHD: Invariant across age and national groupings? *Journal of Child Psychology and Psychiatry*, *53*, 292-303. doi: 10.1111/j.1469-7610.2011.02500.x

Varni, J. W., Beaujean, A. A., & Limbers, C. A. (2013). Factorial invariance of pediatric patient self-reported fatigue across age and gender: A multigroup confirmatory factor analysis approach utilizing the PedsQLTM Multidimensional Fatigue Scale. *Quality of Life Research*, *22*, 2581-2594. doi:10.1007/s11136-013-0370-4

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75*, 157-178. doi:10.1177/0013164414528209

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, *28*, 587-606. doi: 10.1080/0260293032000130234

Yang, F. M., & Jones, R. N. (2008). Measurement differences in depression: Chronic health-related and sociodemographic effects in older Americans. *Psychological Medicine*, *70*, 993-1004. doi:10.1097/PSY.0b013e31818ce4fa

Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: An extension of the bi-factor item response theory model for use in differential item functioning. *Journal of Psychiatric Research*, *43*, 1025-1035. doi:10.1016/j.jpsychires.2008.12.007

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

Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement*, *65*, 465-481. doi:10.1177/0013164404272487