Cognitive Abilities Explain Wording Effects in the Rosenberg Self-Esteem Scale

Timo Gnambs

Leibniz Institute for Educational Trajectories

Ulrich Schroeders

Psychological Assessment, University of Kassel

Author Note

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:7.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Correspondence concerning this article should be addressed to Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, E-mail: timo.gnambs@lifbi.de.

Accepted for publication in *Assessment*.

Abstract

There is consensus that the ten items of the *Rosenberg Self-Esteem Scale* (RSES) reflect wording effects resulting from positively and negatively keyed items. The present study examined the effects of cognitive abilities on the factor structure of the RSES with a novel, non-parametric latent variable technique called *Local Structural Equation Models* (LSEM). In a nationally representative German large-scale assessment including 12,437 students competing measurement models for the RSES were compared: a bifactor model with a common factor and a specific factor for all negatively worded items showed an optimal fit. LSEM showed that the unidimensionality of the scale increased with higher levels of reading competence and reasoning, while the proportion of variance attributed to the negatively keyed items declined. Wording effects on the factor structure of the RSES seem to represent a response style artifact associated with cognitive abilities.

Keywords: LSEM, self-esteem, cognitive abilities, factor structure

Cognitive Abilities Explain Wording Effects in the Rosenberg Self-Esteem Scale

The *Rosenberg Self-Esteem Scale* (RSES, 1965) is a popular self-report instrument measuring a respondent's global self-worth and self-respect with 10 items. Due to its brevity and face validity, the RSES has dominated the literature on self-esteem since its introduction (see Donnellan, Trzesniewski, & Robins, 2011; Zuckerman, Li, & Hall, 2016). The RSES has been widely used in clinical (Salerno, Ingoglia, & Coco, 2017) and educational contexts (Diseth, Meland, & Breidablik, 2014) as well as in large-scale social survey research (Marsh, Scalas, & Nagengast, 2010). The concomitant discussion about the dimensionality of the measure is almost as old as the measure itself (for recent summaries see Donnellan, Ackerman, & Brecheen, 2016, and Reise, Kim, Mansolf, & Widaman, 2016). In line with its original conceptualization that conceives self-esteem as a unitary concept describing the feeling that "one's good enough" (Rosenberg, 1965, p .31), many authors confirmed the unidimensionality of the RSES (e.g., Chao, Vidacovich, & Green, 2017; Franck, de Raedt, & Rossel, 2008; Pullman & Allik, 2000; Schmitt & Allik, 2005). For example, an international large-scale study that translated the RSES into 28 languages and administered the instrument to almost 17,000 participants across 53 nations found a single factor underlying the 10 items in most samples (Schmitt & Allik, 2005). However, because the RSES assesses positive self-appraisals (e.g., "I feel that I have a number of good qualities.") and negative self-appraisals (e.g., "At times, I think I am no good at all.") with opposing keyed items, other researchers identified some form of multidimensionality (e.g., DiStefano & Motl, 2009; Donnellan et al., 2016; Gnambs, Scharl, & Schroeders, 2018; Quilty, Oakman, & Risko, 2006; Reise et al., 2016). It seems that although the RSES items are dominated by a common factor, the negatively keyed items capture systematic residual variance over and above general self-esteem. The structural ambiguity of the RSES resulted in a series of factor analytical studies within the last decades that explored whether the RSES scores reflect a single trait or represent a composite of different latent traits. In this discussion, potential moderating influences that might explain the divergent findings regarding the RSES's dimensionality have been somewhat neglected. Therefore, the present study scrutinized individual differences in potentially relevant cognitive abilities (i.e., reasoning, reading competence, vocabulary) to explain the multidimensionality of the RSES. On a more general stance, *Local Structural Equation Models* (LSEM; Hildebrandt, Lüdtke, Robitzsch, Sommer, & Wilhelm, 2016) are applied as a method for moderator analyses of latent structures to study wording effects along continuous context variables.

Are Wording Effects in the RSES More Substance or Style?

The nature and interpretation of wording effects in the RSES is subject to an ongoing debate. On the one hand, some authors considered them mere noise without substantial meaning (Marsh, 1996; Tomás & Oliver, 1999). According to this view, the residual variance captured by negatively keyed items represents a methodological artifact that needs to be controlled for in empirical analyses because it contaminates the measurement of self-esteem. Findings from an experimental study (Greenberger, Chen, Dmitrieva, & Farruggia, 2003) that administered three versions of the RSES, one with all items rephrased in a positive direction, one with all items written in the negative direction, and the original version, provided support for this view: Whereas the original RSES was best represented by a two-dimensional model, the RSES versions including items keyed in only one direction were essentially unidimensional. On the other hand, some authors considered the existence of wording effects the result of systematic response styles such as acquiescence (DiStefano & Motl, 2006; Tomás, Oliver, Galiana, Sancho, & Lila, 2013). In this view, in addition to the focal construct of self-esteem negatively keyed items also capture a conceptually distinct trait representing a respondent's response consistency independent of the scales' content. In line with this assumption, wording effects in the RSES have been found to be stable across measurement occasions (Gana et al., 2013; Marsh et al., 2010; Michaelides, Koutsogiorgi, & Panayiotou, 2016; Motl & DiStefano, 2002) and subgroups (DiStefano & Motl, 2009; Lindwall et al., 2012; Michaelides, Zenger, et al., 2016; Salerno et al., 2017), they were identified in different language versions (Tomás et al. 2013; Wu, 2008; Wu, Zuo, Wen, & Yan, 2017), and have been replicated across similar instruments (DiStefano & Motl, 2006; Horan, DiStefano, & Motl, 2003). Furthermore, because criterion-related validity studies associated the RSES scores for positively and negatively keyed items with distinct personality traits and motivational tendencies (e.g., Donnellan et al., 2016; Quilty et al., 2006), some authors even argued that these subdimensions imply a substantive distinction between two unique, albeit correlated, personality traits, positive and negative self-esteem (e.g., Alessandri, Vecchione, Eisenberg, & Łaguna, 2015; Owens, 1994; Roth, Decker, Herzberg, & Brähler, 2008). According to this perspective, the negatively keyed items of the RSES measure self-derogation or self-deprecation, whereas the positive items capture self-competence. Hence, differently keyed items of the RSES form two subscales reflecting different forms of self-esteem.

Moderating Influences on the Structure of the RSES

Despite an abundance of research on the RSES, little is known about moderators that might explain why some studies supported an essentially unidimensional structure whereas others advocated for a multidimensional structure. Some authors attributed aberrant responses for negatively keyed items to respondents' cognitive abilities (Cordery & Sevastos, 1993; Marsh, 1996; Weems, Onwuegbuzie, & Collins, 2006; Williams & Swanson, 2001). For example, Marsh (1996) suggested that responses to the RSES might be affected by the verbal skills of the respondents because responding to negatively worded items requires more complex cognitive processes than responding to positively keyed items. Individuals lacking the necessary competencies to properly understand grammatical negations might perceive negatively worded items differently. Along this line, Sliter and Zickar (2014) demonstrated that negatively keyed items functioned differently and, on average, exhibited higher category thresholds (i.e., they were more difficult) than positively keyed items. Moreover, Marsh

(1996) showed—by dividing a sample into different ability groups—that wording effects decreased for students with higher reading competence. Thus, the RSES seems to be a relatively unidimensional scale among verbally competent respondents, whereas the negatively keyed items capture systematic residual variance among respondents with limited verbal skills. These results were replicated in some samples (Corwyn, 2000; Dunbar, Ford, Hunt, & Der, 2000), but not in others (von Collani & Herzberg, 2003b).

A fundamental problem with this sort of analysis is how moderating effects were modeled. Although reading competence was measured on a continuous scale, the variable was post-hoc classified into different categories to create artificial competence groups. However, this artificial categorization of a naturally continuous context variable is associated with several methodological problems (MacCallum, Zhang, Preacher, & Rucker, 2002; Preacher, Rucker, MacCallum, & Nicewander, 2005; Rucker, McShane, & Preacher, 2015): First, in the framework of *Multiple-Group Mean and Covariance Structure* (MGMCS) analyses that are widely used and accepted for investigating factorial invariance across categorical context variables (van de Schoot, Lugtig, & Hox, 2012; Wicherts & Dolan, 2010), creating artificial subgroups increases the risk of missing nonlinear trends and interaction effects. Unless a large number of groups are used, they do not allow the identification of the onset of a parameter change (Hildebrandt et al., 2009). Second, categorization leads to a loss in information on individual differences within a given group. When observations that differ across the range of a continuous variable are grouped, respondents within groups are assumed homogenous and potential variations within these groups are ignored. Third, when splitting a continuous distribution of a moderator into several distinct sections, the selection of cutpoints is frequently rather arbitrary. Thus, neither the number of groups nor their ranges along the context variables are unique. Critically, in case of nonlinear parameter changes the selected ranges can influence the results of MGCMS analyses and increase the likelihood of Type II errors (Hildebrandt et al., 2009; MacCallum et al., 2002).

Local Structural Equation Modeling

To overcome shortcomings of categorizing continuous context variables, the present study capitalizes on a recently developed non-parametric structural equation modeling (SEM) technique called *Local Structural Equation Models* (LSEM; Hildebrandt et al., 2016; Hildebrandt, Wilhelm, & Robitzsch, 2009) which allows studying variance-covariance structures contingent on a continuous context variable. In principle, LSEMs are traditional SEMs that weight observations around focal points (i.e., specific values of the continuous moderator variable) with a Gaussian kernel function (Gasser, Gervini, & Molinari, 2004). Thus, in contrast to grouping participants according to a moderator variable (as is the custom in MGMCS), in LSEM participants are weighted depending on their value of the moderator. The core idea is that observations near the focal point provide more information for the corresponding SEM than more distant observations. Figure 1 exemplifies three weight functions using the cognitive ability of the respondents (*z*-standardized) as moderator at focal points of $z = -\frac{1}{3}$, 0, and $\frac{1}{3}$. Observations exactly at the focal point receive a weight of 1; observations with moderator values higher or lower than the focal point receive smaller weights. For example, if the difference between the focal point and moderator is $\left|\frac{1}{3}\right|$, the weight is about .50 (see the gray dashed lines in Figure 1). For each focal value of the context variable, a separate SEM is estimated resulting in a series of models that provide gradients of model parameters. A more formal introduction into LSEM is given in the Appendix.

An advantage of LSEM is the opportunity to study any model parameter (e.g., means, factor loadings, or variances) across a continuous context variable. It is even conceivable to explore changes in model fit indices (e.g., comparative fit index) or other indices such as composite reliability that is derived from estimated factor loadings (Rodriguez, Reise, & Haviland, 2016). Because LSEM does not require an *a priori* function regarding change, the approach is also viable when there are no explicit assumptions regarding the onset or the trajectories of parameter change.

Generally, each local SEM will utilize more observations than a SEM that is limited to respondents with a specific value of the context variable. The weighting scheme used in LSEM results in an effective sample size at each focal point that depends on the available observations near the value of the context variable. In the case of a normally distributed context variable, focal points in the midrange will integrate the information of many respondents, resulting in a larger effective sample size. In contrast, focal points in the extremes of the moderator range will rely on less observations. As a consequence, the effective sample size will be smaller at the lower and upper ends of the moderator distribution and, thus, result in less precise parameters estimates and larger confidence intervals.

LSEMs allow for the visual inspection of gradients of SEM parameters. For example, examining the trajectories of factor loadings or latent means can help in identifying the onset of parameter changes or in describing developmental aspects (e.g., curvilinear trends). Traditional model fit indices or statistical likelihood-based tests to evaluate the effect of the moderator are not available since the moderator is no explicit parameter in the model, but influences the SEM only indirectly through the weighting function. Statistical inferences can be made using so-called permutation tests that evaluate if a SEM parameter is constant across different values of the context variables (initially described in Hülür, Wilhelm, & Robitzsch, 2011; see also Hildebrandt et al., 2016). For this test, a large number of datasets are generated from the observed data (e.g., 1,000 permutations) that each randomly assigns the observed values of the moderator to the individuals. This approach ensures that the data in the permuted data set are completely independent of the context variable and, thus, allows to test whether changes of the gradients in the real dataset are connected to the moderator variable. More specifically, through the random assignment, the results of the permutation test are adjusted for a main effect of the moderator. Therefore, the *shape* of the parameter estimates is compared between the observed and the permuted datasets rather than *absolute values* (see also Schroeders, Schipolowski, & Wilhelm, 2015).

Present Study

In order to shed light on the question whether the negatively worded items of the RSES capture trait-specific variance or if they simply add trait-irrelevant variance to an otherwise unidimensional scale, we studied the dimensionality of the RSES along a range of cognitive abilities in a representative sample of German students. If negatively keyed items formed a substantive trait independent of general self-esteem, the factor structure should remain invariant independent of the verbal ability of the students. In contrast, if negatively keyed items represented response artifacts associated with cognitive abilities, we would expect the method factor to account for a larger proportion of item variance among less competent respondents, whereas the proportion of explained variance should decline for more competent respondents. Thus, the goal of the study is the examination of the factor structure of the RSES across potentially relevant continuous moderators. More specifically, we study changes in selective model parameters across different levels of reading abilities, vocabulary, and reasoning by means of LSEM.

Method

Participants

The *N* = 12,437 respondents (50% girls) were part of a representative sample of German students in the National Educational Panel Study (see Blossfeld, Roßbach, & von Maurice, 2011) that attended ninth grade at various schools across rural and urban localities. To reach a diverse sample of students, all major school types were included (see Steinhauer, Aßmann, Zinn, Goßmann, & Rässler, 2015, for details on the sampling procedure): about 54% attended general or intermediate secondary schools, 39% went to higher secondary schools, and the remaining 7% encompassed students from several specialized school branches. Their mean age was *M* = 14.68 (*SD* = 0.69) years. Data collection was conducted in small groups at the students' respective schools by a professional survey institute (for details on the data collection process see the field reports provided at http://www.neps-data.de).

Instruments

Self-esteem was measured with a German translation (von Collani & Herzberg, 2003a) of the Rosenberg (1965) scale using ten items on 5-point response scales from 1 “strongly disagree” to 5 “strongly agree” (see Appendix). The negatively keyed items (2, 5, 6, 8, 9) were recoded so that higher scores indicate higher self-esteem. For each item between 1% and 3% of the respondents exhibited missing values. The means, standard deviations, and correlations between all items are summarized in Table S1 of the supplemental material. The average score of the 10 RSES items had a mean of $M = 3.94$ ($SD = 0.63$) and a reliability ω_{total} of .85 (McNeish, 2017). *Vocabulary* was measured with an adapted German version of the Peabody Picture Vocabulary Test (Dunn & Dunn, 2004) including 89 items. For each item, the respondents had to select one out of four pictures that corresponded to a spoken word. The sum score of correctly answered items ($M = 57.76$, $SD = 10.25$) had a reliability categorical ω_{total} of .87 (Green & Yang, 2009). *Reading competence* was measured with an achievement test ($M = 0.03$, $SD = 1.25$) including 31 items that required either multiple-choice or short constructed responses (see Haberkorn, Pohl, Hardt, & Wiegand, 2012). The test was scaled using a unidimensional logistic item response model (Rasch, 1960). Competence scores for each respondent were derived as weighted maximum likelihood estimates (Warm, 1989) with a reliability of .75. *Reasoning* was measured with a matrices test including 12 items. Matrices tests are good proxies for fluid intelligence, because the figural content is seen as prototypical for the construct (Wilhelm, 2005). For each item, respondents had to identify a missing element from several response options that logically completed a geometrical pattern. The number of correctly solved items ($M = 8.74$, $SD = 2.40$) had a reliability of categorical $\omega_{\text{total}} = .70$.

Statistical Analyses

The dimensionality of the RSES was examined by confirmatory factor analyses using a full information maximum likelihood (FIML) estimator with heteroskedasticity-consistent standard errors (Freedman, 2006) and a robust test statistic (Yuan & Bentler, 2000) in *lavaan* version 0.5-23.1097 (Rossell, 2012). Simulation studies indicate that linear factor analyses allow for valid inferences as long as all variables have at least 5 response categories (Beauducel & Herzberg, 2006; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Model fit was evaluated in line with conventional standards (see Schermelleh-Engel, Moosbrugger, & Müller, 2003) using the *Comparative Fit Index* (CFI), the *Root Mean Square Error of Approximation* (RMSEA), and the *Standardized Root Mean Square Residual* (SRMR). Models with CFI ≥ .95, RMSEA ≤ .08, and SRMR ≤ .10 are interpreted as “acceptable” and models with CFI ≥ .97, RMSEA ≤ .05, and SRMR ≤ .05 as “good” fitting. We tested four structural models for the RSES that have been frequently adopted in the literature (see Figure 2). In all models, factor loadings and residual variances were freely estimated. For identification purposes, the latent factor variances were fixed to 1. Moreover, the residual variances for all items were uncorrelated. Model 1 was strictly unidimensional and assumed a single general factor explaining the covariances between the RSES items. Model 2 specified a bifactor structure (see Brunner, Nagy, & Wilhelm, 2012; Reise, 2012) including a general factor for all items of the RSES and two specific factors for the positively (1, 3, 4, 7, 10) and negatively keyed items (2, 5, 6, 8, 9). In this model, the two method factors capture the residual variance that is attributed to the positively and negatively keyed items after accounting for the shared variance of all items. Trait and method factors were uncorrelated. Model 3 specified two correlated latent factors representing positive and negative self-esteem. The latter was indicated by the five negatively keyed items, whereas the former was specified by the positively keyed items. This model is mathematical equivalent to a bifactor model with proportional constraints on the factor loadings (Reise, 2012). Thus, model 3 is more parsimonious than model 2. Finally, model 4 specified a bifactor-(*S*-1) structure (see Eid, Geiser, Koch, & Heene, 2017) that included a general factor for all items and a single specific latent factor for the negatively keyed items. In the factor analytical literature, such models have previously been termed nested factor models (Schulze, 2005). In this model, the general factor can be understood as general self-esteem which is instantiated by the positively keyed items and orthogonal to a method factor capturing the residual variance of the negatively keyed items.

A recent simulation study (Gu, Wen, & Fan, 2017) highlighted that wording effects might have a detrimental effect on the homogeneity of a scale, that is, ignoring negative wording effects leads to biased estimates of reliability and criterion-based validity. Therefore, the focal parameter in our analyses pertained to the percentage of variance in total scores attributable to the general factor (i.e., general self-esteem) in terms of omega hierarchical (ω_H; Rodriguez et al., 2016). For a bifactor model (Model 2 in Figure 2) ω_H is given in [1] with λ_{gen} representing the standardized factor loadings on the general factor, λ_{pos} the respective loadings on the positive factor, λ_{neg} the respective loadings on the negative factor, and h^2 the explained item variance.

$$\omega_H = \frac{\left(\sum_{i=1}^{10} \lambda_{gen,i}\right)^2}{\left(\sum_{i=1}^{10} \lambda_{gen,i}\right)^2 + \left(\sum_{i=1}^{5} \lambda_{pos,i}\right)^2 + \left(\sum_{i=1}^{5} \lambda_{neg,i}\right)^2 + \sum_{i=1}^{10}\left(1-h_i^2\right)} \qquad [1]$$

Moreover, because some authors advocated the interpretation of subscales in the RSES (e.g., Alessandri et al., 2015; Owens, 1994), we also examined omega hierarchical subscale (see Rodriguez et al., 2016) for the negatively keyed items which reflects the proportion of unique variance in the subscale score after accounting for the general factor. For the five items of the negative self-esteem subscale $\omega_{HS.NEG}$ is given as:

$$\omega_{HS.NEG} = \frac{\left(\sum_{i=1}^{5} \lambda_{neg,i}\right)^2}{\left(\sum_{i=1}^{5} \lambda_{gen,i}\right)^2 + \left(\sum_{i=1}^{5} \lambda_{neg,i}\right)^2 + \sum_{i=1}^{5}\left(1-h_i^2\right)} \qquad [2]$$

Moderating effects of cognitive abilities on ω_H and $\omega_{HS.NEG}$ were studied using LSEMs with robust FIML estimation (Hildebrandt et al., 2016), implemented in the R package *sirt*, version 2.0-25 (Robitzsch, 2017). We selected 37 equally spaced focal points between -1.8 and 1.8 on the *z*-standardized scale of each moderator. Because fewer participants achieved extreme scores on the moderators (i.e., ≤ -2 or ≥ 2) and the robustness of estimated SEM parameters is affected by the sample size (Wolf, Harrington, Clark, & Miller, 2013), the analyses were limited to focal points resulting in an effective sample size of at least 400. This resulted in effective sample sizes across the 37 focal points between 421 and 3,969 (*Mdn* = 2,310 to 2,486 for the different values of the moderators). Gradients of ω_H and $\omega_{HS.NEG}$ were derived by reestimating a confirmatory factor model at different focal points of the moderator using appropriate sample weights. Permutation tests on the derived vectors of ω allow for statistical inferences on the variability and potential trends of ω across the continuous moderators (Hildebrandt et al., 2016; Hülür et al. 2011).

Open Data

The variance-covariance matrix between the 10 items of the RSES is provided in the supplemental material (Table S1). Moreover, researchers accepting the respective legal and confidentially agreement can download the complete data set analyzed in this study (http://www.neps-data.de). We also provide all R scripts (R Core Team, 2017) in an online repository of the *Open Science Framework* (https://osf.io/bkzjy) to make the present analyses as transparent and reproducible as possible (Nosek et al., 2015).

Results

All items of the RSES were moderately correlated (Table S1). An exploratory maximum likelihood factor analysis with oblimin rotation ($\delta = 0$; see Table 1) resulted in a clearly interpretable two factor solution (first four eigenvalues: 4.31, 1.20, 0.74, 0.72) explaining about 44 percent of the item variance. Negatively keyed items had average pattern coefficients on the first factor of *Mdn*(λ) =.71 (*Min* = .49, *Max* = .85); positively keyed items exhibited average pattern coefficients on the second factor of *Mdn*(λ) =.45 (*Min* = .43, *Max* = .78). The correlation between both factors amounted to *r* = .67. The three cognitive measures showed rather negligible associations with the RSES scores. Vocabulary, reading competence, and reasoning correlated ($p < .001$) with self-esteem at .11, .05, and .04, respectively.

Dimensionality of the Rosenberg Self-Esteem Scale

The goodness of fit indices for the competing factor models are summarized in Table 2, whereas the respective standardized parameter estimates are included in Figure 2. A unidimensional factor model (Model 1) for the RSES exhibited an unsatisfactory fit (CFI = .87, RMSEA = .09, SRMR = .06), although all items showed substantial factor loadings, *Mdn*(β) = .63 (*Min* = .47, *Max* = .71). In contrast, a bifactor model (Model 2) that also included specific factors for the positively and negatively keyed items resulted in a significantly ($p < .05$) better fit (CFI = .99, RMSEA = .04, SRMR = .02). Standardized loadings were larger than .40 on the general factor, *Mdn*(β) = .53 (*Min* = .41, *Max* = .79). In addition, the negatively keyed items had non-ignorable loadings on the method-specific factor, *Mdn*(β) = .40 (*Min* = .29, *Max* = .54). However, only two positively keyed items (3, 4) exhibited substantial loadings on the respective factor, *Mdn*(β) = .16 (*Min* = .01, *Max* = .55); in contrast, one item (7) had a significant ($p < .05$) but non-substantial loading and two items (1, 10) exhibited no significant ($p > .05$) factor loadings on the method factor. These results fall in line with previous findings (e.g., Donnellan et al., 2016; Marsh et al. 2010) that demonstrated more pronounced method effects for negatively keyed items and unclear loading patterns (i.e., non-significant or even negative) for the positively keyed items. This pattern of result matched with the respective reliability estimates: The negative factor accounted for about 10 percent of the test score variance, whereas only about 4 percent were attributable to the positive factor. However, the general factor accounted for most of the variance (ω_H = .79). Thus, the RSES was dominated by a single general factor.

We also examined whether more parsimonious models might adequately describe the data. A model with two correlated factors reflecting positive and negative self-esteem (Model 3) showed a notable decline in fit (see Table 2). Similarly, the bifactor-(*S*-1) model (Model 4) that included only a single specific factor for the negatively keyed items exhibited a worse fit than the full bifactor model. Finally, because the positive factor in model 2 was primarily defined by two items (3, 4), we extended the bifactor-(*S*-1) model and allowed the residuals between these items to correlate freely. The respective model (Model 5 in Table 2) showed a negligible decline in fit as compared to the full bifactor model. All goodness of fit indices indicated an excellent fit (CFI = .98, RMSEA = .04, SRMR = .02). Moreover, all standardized loadings on the general factor were substantial, *Mdn*(β) = .54 (*Min* = .40, *Max* = .78), which also holds true for the negative factor, *Mdn*(β) = .41 (*Min* = .30, *Max* = .55). Again, most of the test score variance was attributable to the general factor (ω_H = .79) as compared to the negative factor (11%). The two residuals correlated at *r* = .31 (*p* < .001)[1]. Therefore, all subsequent analyses were based on the bifactor-(*S*-1) model with correlated residuals for items 3 and 4. In addition, all analyses were also replicated using the full bifactor specification (Model 2). But, these yielded no significantly and substantially different results (see online supplement).

Moderating Effects of Cognitive Abilities

Neither the general factor nor the negative factor was substantially correlated with vocabulary, reading comprehension, or reasoning (*r*s between -.06 and .11; see Table S2 in the supplement material). To study potential moderating effects of individual differences in cognitive abilities on the dimensionality of the RSES and the homogeneity of the general factor, LSEMs were conducted. We investigated the influence of vocabulary, reading competence, or reasoning as a continuous moderator of the factor structure in three separate LSEMs. All models indicated good model fits: The average CFI was *Mdn* = .979 (*Min* = .962, *Max* = .982), the average RMSEA was *Mdn* = .049 (*Min* = .040, *Max* = .071), and the average

SRMR was *Mdn* = .022 (*Min* = .020, *Max* = .029). The variability of ω_H for the general factor along the *z*-standardized ability scores is summarized in Figure 3 (left column). Overall, ω_H gradually increased along the values of all three cognitive measures. Moreover, permutation tests (see Table 3) corroborated that ω_H for the general factor was not constant across different values of the moderators, but demonstrated significant variation along values of vocabulary (*SD* = 0.03, *p* < .001), reading competence (*SD* = 0.04, *p* = .01), and reasoning (*SD* = 0.04, *p* < .001). A regression of the estimated ω_H on the different values of the moderators identified significant linear trends for vocabulary (*B* = .04, *p* < .001), reading competence (*B* = .05, *p* < .001), and reasoning (*B* = .05, *p* < .001). Thus, a difference of one standard deviation on either cognitive measure corresponded to an increase of about 4 to 5 percent in variance explained by the general factor. In the same vein, the percentage of variance in subscale scores attributed to the negatively keyed items mirrored the results for the general factor (see right column of Figure 3): $\omega_{HS.NEG}$ gradually decreased along the values of all three cognitive measures. A one standard deviation difference on vocabulary, reading competence, or reasoning was accompanied by a decrease of about 5 to 10 percent in unique variance explained by the negative factor.

The three cognitive measures were substantially (*p* < .001) correlated. Vocabulary and reading competence correlated at *r* = .56 and reasoning correlated with the two former at *r* = .41 and *r* = .45, respectively. Therefore, we also studied the partial effects of each measure on ω_H, that is, vocabulary, reading competence, and reasoning were residualized to cancel out their shared variance. In the following, we report the results of analyses replicated with these residualized scores. The variability of ω_H for the general factor across the residualized cognitive measures is summarized in Figure 4 (see also Table 3). Vocabulary had no substantial partial effect on the variance explained by the general factor. Despite some variability of the estimated ω_H (*SD* = 0.01, *p* = .02) across different values of vocabulary, there was no systematic linear trend (*B* = .01, *p* = .22). In contrast, the residualized reading competence (*SD* = 0.02, *p* < .001) and reasoning scores (*SD* = 0.03, *p* < .001) both replicated the linear trends identified previously, *B* = .03 (*p* < .001) and *B* = .02 (*p* < .001), respectively. Similarly, whereas vocabulary showed no partial effects for $\omega_{HS.NEG}$ (*B* = .00, *p* = .84; see Table 3) it gradually decreased with higher residualized reading competence (*B* = -.02, *p* < .001) or reasoning scores (*B* = -.02, *p* < .001). Thus, the factor saturation in the subscale for the negatively keyed items was most pronounced for lower levels of reading competence and reasoning, whereas it gradually decreased for higher levels of these scores. Overall, for verbally more competent students the negatively worded items captured little unique variance beyond general self-esteem.

Discussion

In applied measurement, it is a common finding that psychological measures often have a dominant general factor capturing the commonality between all items, but also some evidence of multidimensionality. As a consequence, such "structural ambiguity leads to seemingly endless 'confirmatory' factor analytic studies, in which the research question is whether scale scores can be interpreted as reflecting variation on a single trait" (Reise, Moore, & Haviland, 2010, p. 544). More than fifty years of research on the dimensionality of the RSES, has not settled this dispute: Many authors concur that the RSES is not strictly unidimensional, but also captures wording effects from negatively keyed items (cf. Alessandri et al., 2015; Donnellan et al., 2016; Gnambs et al., 2018; Reise et al., 2016), while others found wording effects to be seemingly negligible and unlikely to distort the measurement of self-esteem (e.g., Chao et al., 2017; Franck, et al., 2008; Schmitt & Allik, 2005).

With the present study, we took a different approach arguing that the extent of wording effects depends on the verbal and cognitive abilities of the test-takers. More precisely, we explored potential moderating influences on the dimensionality of the RSES by means of LSEM. In line with previous research (Corwyn, 2000; Dunbar et al., 2000; Marsh, 1996) we showed that respondents with poor reading competences and reasoning abilities provided biased responses to negatively keyed items. Whereas the RSES was essentially unidimensional among skilled readers, the responses of students with difficulties in adequately understanding negatively keyed items also reflected a secondary trait beyond self-esteem. Such an interaction with test-takers' reading abilities seems particularly troublesome if RSES subscales are interpreted as representing different types of self-esteem (Alessandri et al., 2015; Owens, 1994). The unique variance among the five items forming a putative negative self-esteem scale gradually decreased with increasing reading abilities. As a consequence, the factor saturation has been halved for skilled readers as compared to less skilled readers. Thus, distinguishing between qualitatively different aspects of self-esteem (i.e., positive and negative self-esteem) is not warranted, because differently keyed items measure essentially the same construct among gifted readers. Given that the interpretation of scale scores depends on the respondent's cognitive abilities, it seems questionable to view positive and negative self-esteem as substantially different traits of personality.

Prospects of Local Structure Equation Modeling

On a more general note, this study presented LSEM as a versatile and powerful method to examine changes in the variance-covariance structure depending on continuous context variables. Until now, LSEM has been mainly applied to study age-related changes in cognitive abilities such as the differentiation-dedifferentiation of intelligence in childhood and adolescence (Hülür et al., 2011; Schroeders et al., 2015) or the development of face cognition abilities across the life span (Hildebrandt, Sommer, Herzmann, & Wilhelm, 2010; Hildebrandt, Wilhelm, Herzmann, & Sommer, 2013). However, LSEM has a greater potential, whenever moderating hypotheses pertain to continuous context variables. Unfortunately, all too often the set of available statistical methods shapes the way researchers conceptualize and study human behavior. For example, in cross-cultural research group differences between different ethnic groups are studied with MGMCS, when, more appropriately, the variable of interest should be cultural values or national identity. Thus, instead of analyzing the influence of membership to a salient group, it would be more informative to examine the underlying psychological mechanisms that are hypothetical and continuous in nature. Furthermore, the statistical methods and not the nature of the variables affect the way in which we analyze data. For instance, group designs are frequently used in research on aging, although the underlying context variable age is continuous (e.g., Gnambs, & Buntins, 2017; Marsh, Nagengast, & Morin, 2013). Similar limitations pertain to previous studies on reading abilities and wording effects (Corwyn, 2000; von Collani & Herzberg, 2003b; Dunbar et al., 2000; Marsh, 1996). In these studies, researchers artificially and arbitrarily created ability groups, albeit ability was measured on a continuous scale. Although most studies corroborate the present findings—since the observed effect was nearly linear—such split-design analyses always run the risk of masking potential nonlinear changes[2]. Thus, LSEMs offer a flexible opportunity to study parameter changes in the mean and the variance-covariance structure across a continuous context variable. In particular, LSEM can also be used as an exploratory approach for situations when little is known about the onset and precise form of moderating effects.

Limitations and Future Research

The present findings offer various avenues for future studies. For one, it is rather disconcerting that even in such a plain measure as the RSES, which is praised for its simple language and the brevity of its items, we were able to show that there is substantial construct-irrelevant variance associated with negatively keyed items. Most likely similar issues apply to the majority of measurement instruments including negative items. Therefore, it should be investigated to what degree the reported results also translate to other measurement instruments. In particular, it would be interesting to know whether the observed cognitive effects are more severe for instruments that include linguistically more complex items and if they are still identifiable in simplistic items including only one or two words (e.g., adjective lists; e.g., Watson, Clark, & Tellegen, 1988). Second, a number of studies observed that respondents with lower educational attainment were more prone to acquiescence than those with higher levels of education (e.g., Meisenberg & Williams, 2008; Rammstedt & Farmer, 2013; Rammstedt, Kemper, & Borg, 2013). Even cognitive abilities have been identified as a pivotal source of individual differences in acquiescence responding (Lechner & Rammstedt, 2015). Therefore, the identified moderating effects of cognitive differences might represent indirect effects of systematic response style: Respondents that are unable to properly understand and evaluate the content of an item might more frequently resort to acquiescent responding instead of processing the item and elaborating a response and, thus, introduce multidimensionality in an otherwise unidimensional scale. Thus, it could be informative to scrutinize potential mediating mechanisms between cognitive abilities and acquiescence responding for the study of dimensionality issues in self-report scales. Finally, our results pertain to a rather specific population in the form of teenaged students in Germany. Future research is encouraged to extend these results to other age groups and language versions. For example, stronger individual differences in acquiescent responding have been observed among younger age groups (Soto, John, Gosling, & Potter, 2008) and in societies emphasizing collectivistic values (Johnson, Kulesa, Cho, & Shavitt, 2005). Therefore, it might be fruitful to contrast wording effects in different age groups and study the effects of negatively keyed items across the life span. Because cognitive abilities typically show age-related changes (e.g., reading competences are likely to increase in childhood and adolescence, whereas reasoning abilities tend to decline in old age) it could even be worthwhile to consider moderating effects for both variables including their interactions simultaneously.

Conclusions

Many previous studies observed that the negatively keyed items in the RSES distorted its factor structure. The present study on a representative sample of German young adults showed that the structural ambiguity of the scale is subject to individual differences in cognitive abilities. Respondents with poor reading skills or reasoning abilities showed systematic response styles associated with the negatively keyed items, whereas good readers showed limited wording effects. Among others, these results highlight the need for taking into account acquiescence in latent variable modeling of the RSES. Conversely, we found no evidence for negative self-esteem as a substantive personality trait.

References

Alessandri, G., Vecchione, M., Eisenberg, N., & Łaguna, M. (2015). On the factor structure of the Rosenberg (1965) General Self-Esteem Scale. *Psychological Assessment, 27*, 621-635. doi:10.1037/pas0000073

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186-203. doi:10.1207/s15328007sem1302_2

Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J. (Eds.) (2011). Education as a lifelong process - The German National Educational Panel Study (NEPS). [Special Issue] *Zeitschrift für Erziehungswissenschaft, 14.*

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80*, 796-846. doi:10.1111/j.1467-6494.2011.00749.x

Chao, R. C. L., Vidacovich, C., & Green, K. E. (2017). Rasch analysis of the Rosenberg Self-Esteem Scale with African Americans. *Psychological Assessment, 29*, 329-342. doi:10.1037/pas0000347

von Collani, G., & Herzberg, P. Y. (2003a). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg [A revised version of the German adaptation of Rosenberg's self-esteem scale]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24*, 3-7. doi:10.1024//0170-1789.24.1.3

von Collani, G., & Herzberg, P. Y. (2003b). Zur internen Struktur des globalen Selbstwertgefühls nach Rosenberg [On the internal structure of global self-esteem (Rosenberg)]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24*, 9-22. doi:10.1024//0170-1789.24.1.9

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised job diagnostic survey: Is education a factor in responses to negatively-worded items? *Journal of Applied Psychology, 78*, 141-143. doi:10.1037/0021-9010.78.1.141

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*, 357-379. doi:10.1006/jrpe.2000.2291

Diseth, Å., Meland, E., & Breidablik, H. J. (2014). Self-beliefs among students: Grade level and gender differences in self-esteem, self-efficacy and implicit theories of intelligence. *Learning and Individual Differences, 35*, 1-8. doi:10.1016/j.lindif.2014.06.003

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). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex. *Structural Equation Modeling, 16*, 134-146. doi:10.1080/10705510802565403

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

Donnellan, M. B., Trzesniewski, K. H., & Robins, R. W. (2011). Self-esteem: Enduring issues and controversies. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell handbook of individual differences* (pp. 718-746). New York, NY: Wiley-Blackwell.

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

Dunn, L. M., & Dunn, D. M. (2004). *Peabody Picture Vocabulary Test* (PPVT) (German version). Göttingen, Germany: Hogrefe.

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods, 22*, 541-562. doi:10.1037/met0000083

Freedman, D. A. (2006). On the so-called “Huber sandwich estimator” and “robust standard errors”. *American Statistician, 60*, 299-302. doi:10.1198/000313006X152207

Franck, E., De Raedt, R., Barbez, C., & Rosseel, Y. (2008). Psychometric properties of the Dutch Rosenberg self-esteem scale. *Psychologica Belgica, 48*, 25-35. doi:10.5334/pb-48-1-25

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

Gasser, T., Gervini, D., & Molinari, L. (2004). Kernel estimation, shape-invariant modeling and structural analysis. In R. Hauspie, N. Cameron, & L. Molinari (Eds.), *Methods in human growth research* (pp. 179-204). Cambridge, UK: Cambridge University Press.

Gnambs, T., & Buntins, K. (2017). The measurement of variability and change in life satisfaction: A comparison of single-item and multi-item instruments. *European Journal of Psychological Assessment, 33*, 224-238.. doi:10.1027/1015-5759/a000414

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the Rosenberg Self-Esteem Scale: A cross-cultural meta-analysis. *Zeitschrift für Psychologie*. Accepted for publication.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155-167. doi:10.1007/s11336-008-9099-3

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

Gu, H., Wen, Z., & Fan, X. (2017). Examining and controlling for wording effect in a self-report measure: A Monte Carlo simulation study. *Structural Equation Modeling, 24*, 545-555. doi:10.1080/10705511.2017.1286228

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology, 16*, 87-102.

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research, 51*, 257-258. doi:10.1080/00273171.2016.1142856

Hildebrandt, A., Sommer, W., Herzmann, G., & Wilhelm, O. (2010). Structural invariance and age-related performance differences in face cognition. *Psychology and Aging*, *25*, 794-810. doi:10.1037/a0019774

Hildebrandt, A., Wilhelm, O., Herzmann, G., & Sommer, W. (2013). Face and object cognition across adult age. *Psychology and Aging, 28*, 243-248. doi:10.1037/a0031490

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

Hülür, G., Wilhelm, O., & Robitzsch, A. (2011). Intelligence differentiation in early childhood. *Journal of Individual Differences*, *32*, 170-179. doi:10.1027/1614-0001/a000049

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

Lechner, C. M., & Rammstedt, B. (2015). Cognitive ability, acquiescence, and the structure of personality in a sample of older adults. *Psychological Assessment, 27*, 1301-1311. doi:10.1037/pas0000151

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

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*, 810-9. doi:10.1037/0022-3514.70.4.810

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*, 1194-1218. doi:10.1037/a0026913

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40. doi:10.1037/1082-989X.7.1.19

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*. Advance online publication. doi:10.1037/met0000144

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539-1550. doi:10.1016/j.paid.2008.01.010

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

Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y., & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 97*, 13-18. doi:10.1016/j.paid.2016.03.011

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

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1420-1422. doi:10.1126/science.aab2374

Owens, T. J. (1994). Two dimensions of self-esteem: Reciprocal effects of positive self-worth and self-deprecation on adolescent problems. *American Sociological Review, 59*, 391-407. doi:10.2307/2095940

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods*, *10*, 178-192. doi:10.1037/1082-989X.10.2.178

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/s15328007sem1301_5

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*, 1137-1145. doi:10.1037/a0033323

Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five personality measurements for acquiescence: An 18-country cross-cultural study. *European Journal of Personality, 27*, 71-81. doi:10.1002/per.1894

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

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., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*, 818-838. doi:10.1080/00273171.2016.1243461

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

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354-373. doi:10.1037/a0029315

Robitzsch, A. (2017). *sirt: Supplementary item response theory models*. R package version 1.15-41. https://CRAN.R-project.org/package=sirt

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*, 137-150. doi:10.1037/met0000045.

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,* 1-36. doi:10.18637/jss.v048.i02

Roth, M., Decker, O., Herzberg, P. Y., & Brähler, E. (2008). Dimensionality and norms of the Rosenberg Self-Esteem Scale in a German general population sample. *European Journal of Psychological Assessment, 24*, 190-197. doi:10.1027/1015-5759.24.3.190

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher’s guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology, 25*, 666-678. doi:10.1016/j.jcps.2015.04.004

Salerno, L., Ingoglia, S., & Coco, G. L. (2017). Competing factor structures of the Rosenberg Self-Esteem Scale (RSES) and its measurement invariance across clinical and non-clinical samples. *Personality and Individual Differences, 113*, 13-19. doi:10.1016/j.paid.2017.02.063

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8,* 23-74.

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology, 89*, 623-643. doi:10.1037/0022-3514.89.4.623

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492. doi:10.1080/17405629.2012.686740

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2015). Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. *Intelligence, 48*, 15-29. doi:10.1016/j.intell.2014.10.006

Schulze, R. (2005). Modeling structures of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 241-263). Thousand Oaks, CA: Sage Publications.

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

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*, 718-737. doi:10.1037/0022-3514.94.4.718

Steinhauer, H. W., Aßmann, C., Zinn, S., Goßmann, S., & Rässler, S. (2015). Sampling and weighting cohort samples in institutional contexts. *AStA Wirtschafts-und Sozialstatistisches Archiv, 9*, 131-157. doi:10.1007/s11943-015-0162-0

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450. doi:10.1007/BF02294627

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063-1070. doi:10.1037/0022-3514.54.6.1063

Weems, G. H., Onwuegbuzie, A. J., & Collins, K. M. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation & Research in Education, 19*, 3-20. doi:10.1080/09500790608668322

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29,* 39-47. doi:10.1111/j.1745-3992.2010.00182.x

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm, & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373-392). Thousand Oaks, CA: Sage Publications.

Williams, S. A., & Swanson, M. S. (2001). The effect of reading ability and response formats on patients' abilities to respond to a patient satisfaction scale. *Journal of Continuing Education in Nursing, 32*, 60-67.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913-934. doi:10.1177/0013164413495237

Wu, C. H. (2008). An examination of the wording effect in the Rosenberg Self-Esteem Scale among culturally Chinese people. *Journal of Social Psychology, 148*, 535-552. doi:10.3200/SOCP.148.5.535-552

Wu, Y., Zuo, B., Wen, F., & Yan, L. (2017). Rosenberg Self-Esteem Scale: Method effects, factorial structure and scale invariance across migrant child and urban child populations in China. *Journal of Personality Assessment, 99*, 83-93. doi:10.1080/00223891.2016.1217420

Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 167-202. doi:10.1111/0081-1750.00078

Zuckerman, M., Li, C., & Hall, J. A. (2016). When men and women differ in self-esteem and when they don’t: A meta-analysis. *Journal of Research in Personality, 64*, 34-51. doi:10.1016/j.jrp.2016.07.007

Footnotes

[1] Because the correlated residual between item 3 (“I feel that I have a number of good qualities”) and item 4 (“I am able to do things as well as most other people”) was not *a priori* theorized, all interpretations of the observed local dependency have to remain speculative. Potentially, the two items more strongly capture ability-based self-views such as self-perceived skills and competences. In contrast, the other items (e.g., “I take a positive attitude toward myself”) might more strongly reflect attitudinal or affective self-perceptions.

[2] MGMCS mirrored the presented results of the LSEM analyses because the wording effect of cognitive abilities was approximately linear. We created three ability groups using cutscores at $M - 1$ *SD* and $M + 1$ *SD* for each cognitive measure and estimated the reliability within each group. The sample sizes for these groups fell between $N = 1{,}088$ and 9,182. For reading competence, the reliabilities were $\omega_H = [.69, .80, .86]$ and $\omega_{HS.NEG} = [.51, .32, .21]$, respectively. Thus, the reliability of the common factor increased with higher reading groups, whereas the reliability of the negative factor decreased. Similar patterns emerged for vocabulary, $\omega_H = [.70, .80, .83]$ and $\omega_{HS.NEG} = [.48, .33, .26]$, and reasoning, $\omega_H = [.71, .81, .85]$ and $\omega_{HS.NEG} = [.49, .32, .21]$.

Appendices

Appendix A. Rosenberg (1965) Self-Esteem Scale

To what extent do the following statements apply to you?

1. On the whole, I am satisfied with myself. (P)
2. At times, I think I am no good at all. (N)
3. I feel that I have a number of good qualities. (P)
4. I am able to do things as well as most other people. (P)
5. I feel I do not have much to be proud of. (N)
6. I certainly feel useless at times. (N)
7. I feel that I'm a person of worth, at least on an equal plane with others. (P)
8. I wish I could have more respect for myself. (N)
9. All in all, I am inclined to feel that I am a failure. (N)
10. I take a positive attitude toward myself. (P)

Response scale: 1 = strongly disagree, 2 = disagree, 3 = partly, 4 = agree, 5 = strongly agree

P = positively keyed, N = negatively keyed (reverse scored for creating a sum score)

Appendix B. Local Structural Equation Modeling

Formally, a LSEM is given as follows (see also Hildebrandt et al., 2016): Assume for each person a vector of variables (M, $Y_1\ldots,Y_\mathrm{i},\ldots,Y_\mathrm{I}$), where M denotes a moderator variable and Y_i ($i = 1,\ldots, I$) represents the person's responses to the I items of a test. At the population level, the conditional means $\mu_i(m) = \mathrm{E}(Y_\mathrm{i} \mid M = m)$ and the conditional covariances $\sigma_{ii'}(m) = \mathrm{Cov}(Y_\mathrm{i}\ , Y_{\mathrm{i'}}| \ M = m)$ of the items are studied, where m denotes a specific value of the continuous moderator variable. To exemplify LSEM for a common factor model, the conditional covariance matrix $\Sigma(m)$, including the conditional variances and covariances $\sigma_{ii'}(m)$, are represented by a unidimensional common factor model as follows:

$$\sum m = \Lambda(m)\Lambda(m)^T + \Psi(m) \qquad [3]$$

In [3], $\Lambda(m)$ is a column vector of loadings (at a specific point m of the moderator variable M) and $\psi(m)$ is an I x I matrix of error variances and covariances assumed to be diagonal and conditional on m. In the formalization of the commonly used factor model, the model for item i that is conditional on m can be written as

$$Y_{im} = v_i(m) + \lambda_i(m)\cdot\eta_m + \varepsilon_{im}\,. \qquad [4]$$

The intercepts v_i, factor loadings λ_i, and residual variances ε_{im} are all assumed to vary at specific values of M (see [4]). LSEM aims to estimate a factor model or SEM for each possible value of the continuous moderator variable M and to inspect the course of the model parameter estimates across M. Ideally, SEMs are fitted in steps that are as narrow as possible on the scale of the continuous variable. However, the grading depends on the sample size over the moderator and the applied weighting function. An often used and theoretically sound weighting function of observations around focal points is the Gaussian kernel function (see Gasser et al., 2004). Using the Gaussian kernel function, weights around each focal point of M are normally distributed. Since the normal density function is not restricted, all observations will enter all models at each focal point in LSEM, but observations that are far away from the focal points have very small values and will have no practical influence on the model parameter estimation at a given focal point.

$$bw = \frac{h \cdot SD_M}{\sqrt[5]{N}} \qquad [5]$$

The bandwidth (*bw*) of the weighting function is calculated using [5], where *h* denotes the bandwidth factor, SD_M is the standard deviation of the moderator variable *M* and, *N* represents the total sample size. Then, the bandwidth parameter *bw* is the standard deviation of the normal density function around the focal points. In the literature on nonparametric density estimation, the factor *h* = 1.1 has been proposed as being adequate for many context variables, which has been confirmed with a recent simulation study in the context of LSEM (Hildebrandt et al., 2016). In general, the larger the bandwidth *bw*, the smoother the resulting parameter function along the values of *M* will be. For each observation, *M* is standardized using the bandwidth according to [6] and weights ranging between 0 and 1 are derived using the Gaussian kernel function in [7].

$$z(m, m_0) = \frac{m - m_0}{bw} \qquad [6]$$

$$W(m, m_0) = \exp\left(-z(m, m_0)^2 / 2\right) \qquad [7]$$

From the above description, it becomes apparent that *Multiple-Group Mean and Covariance Structure* (MGMCS) analyses can be seen as a special case of LSEM. Basically, MGMCS could be described as employing a weighting scheme in which several focal points along the scale of the moderator (as many as included in one group defined for the analysis) are fully weighted, and all other observations are allocated a weight of 0.

Table 1.

Exploratory Factor Analysis of the Rosenberg Self-Esteem Scale

	Factor 1	Factor 2	h^2
Item 1	.30	.43	.44
Item 3	-.13	.78	.49
Item 4	-.12	.71	.40
Item 7	.13	.45	.30
Item 10	.34	.45	.52
Item 2	.75	-.07	.50
Item 5	.49	.15	.36
Item 6	.85	-.10	.62
Item 8	.53	-.04	.26
Item 9	.71	.02	.53
Eigenvalue	2.63	1.79	
Explained variance	26%	18%	

Note. *N* = 12,437. Full information maximum likelihood factor analysis with oblimin rotation (factor correlation: .67). Gray cells indicate salient pattern coefficients of positively keyed items (1, 3, 4, 7, 10) and negatively keyed items (2, 5, 6, 8, 9).

Table 2.

Fit Statistics for Different Factor Models for the Rosenberg Self-Esteem Scale.

						Model fit					Model comparison	
	Model	χ^2	*df*	*c*	CFI	SRMR	RMSEA	90% CI	BIC	Comp.	$\Delta\chi^2$	Δdf
M1	General factor model	3,765.00*	35	1.251	.871	.055	.093	[.090, .095]	302,055.41			
M2	Bifactor model	438.04*	25	1.205	.986	.017	.036	[.034, .039]	297,967.87	M1	3,059.90*	10
M3	Correlated factor model	1,528.73*	34	1.238	.948	.035	.059	[.057, .062]	299,247.54	M2	1,026.00*	9
M4	Bifactor-(*S*-1) model	1,409.90*	30	1.215	.952	.033	.061	[.058, .063]	299,106.68	M2	934.07*	5
M5	Bifactor-(*S*-1) model with correlated residuals	565.50*	29	1.216	.981	.021	.039	[.036, .041]	298,090.19	M4	856.87*	1

Note. *N* = 12,437. M5 includes correlated residuals between items 3 and 4. *c* = scale correction factor (Yuan & Bentler, 2000); CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residual; RMSEA = Root Mean Square Error of Approximation; BIC = Bayesian Information Criterion; Comp. = comparison model. Robust full information maximum likelihood estimation.

* $p < .05$

Table 3.

Results of Permutation Tests for LSEM on Factor Reliabilities.

	General factor (ω_H)					Negative factor ($\omega_{HS.NEG}$)				
	M	*SD*	*p*(*SD*)	*B*	*p*(*B*)	*M*	*SD*	*p*(*SD*)	*B*	*p*(*B*)
Bivariate effects										
Vocabulary	.800	.033	< .001	.037	< .001	.320	.049	< .001	-.055	< .001
Reading competence	.801	.044	< .001	.052	< .001	.318	.080	< .001	-.095	< .001
Reasoning	.805	.037	< .001	.045	< .001	.314	.078	< .001	-.094	< .001
Partial effects										
Vocabulary	.798	.014	.02	.006	.34	.325	.024	.14	.005	.52
Reading competence	.797	.023	< .001	.027	< .001	.328	.051	< .001	-.058	< .001
Reasoning	.802	.026	< .001	.023	< .001	.320	.052	< .001	-.052	< .001

Note. *N* = 12,437. Based on the bifactor-(*S*-1) model with correlated residuals between items 3 and 4. *M* = average ω across values of the moderator (range: [-1.8, 1.8]); *SD* = variation of ω across values of the moderator; *p*(*SD*) = *p*-value of the permutation test for *SD*; *B* = linear effect of the moderator on ω; *p*(*B*) = *p*-value of the permutation test for *B*. Partial effects are based on the residualized values of the moderators.

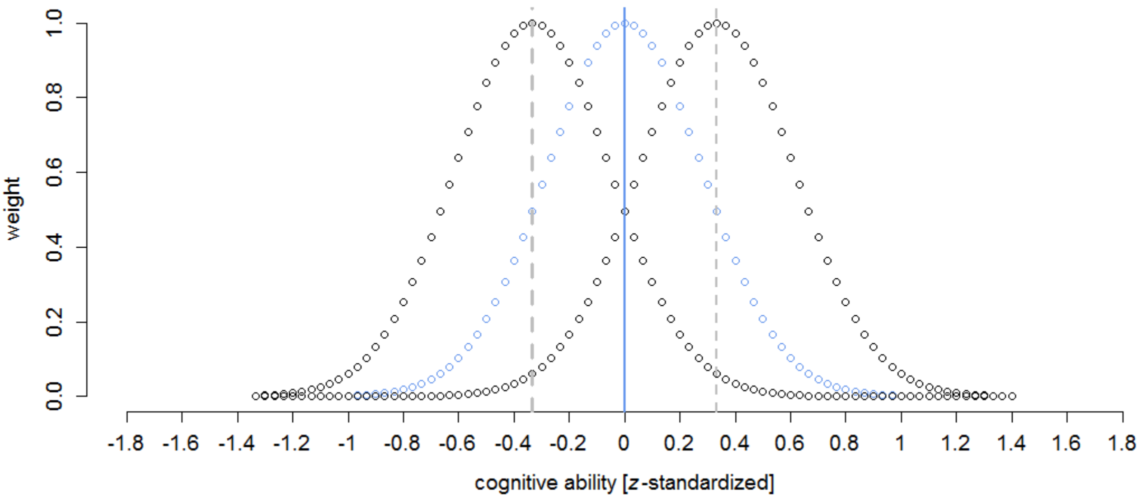

Figure 1. Weighting functions for cognitive ability as moderator. Focal moderator points are $-\frac{1}{3}$, 0, and $\frac{1}{3}$. Gray dashed lines indicate the value of the moderator at which an observation will receive a weight of .50 for focal point 0.

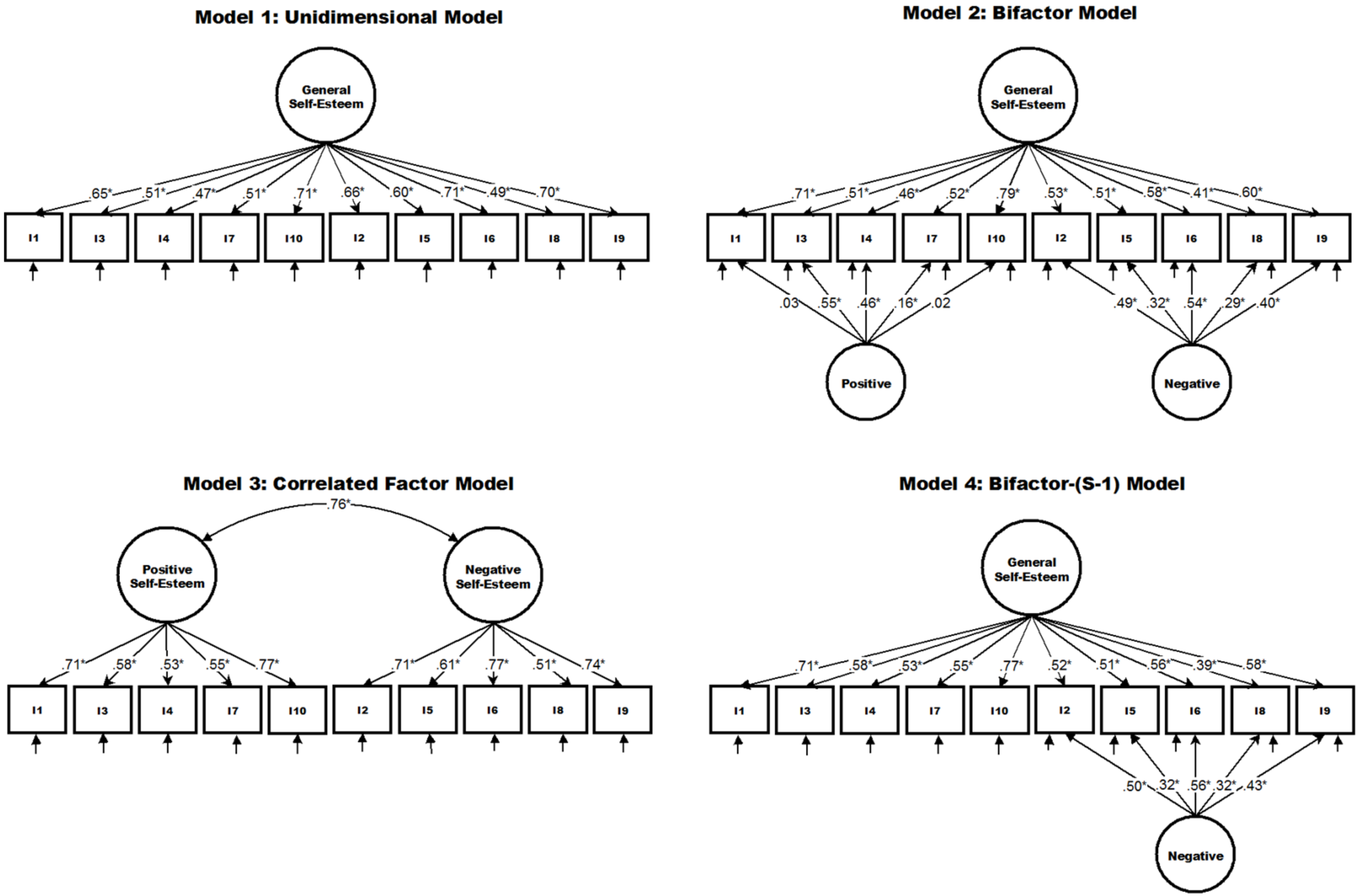

Figure 2. Factor models for the Rosenberg Self-Esteem Scale with standardized factor loadings (* p < .05).

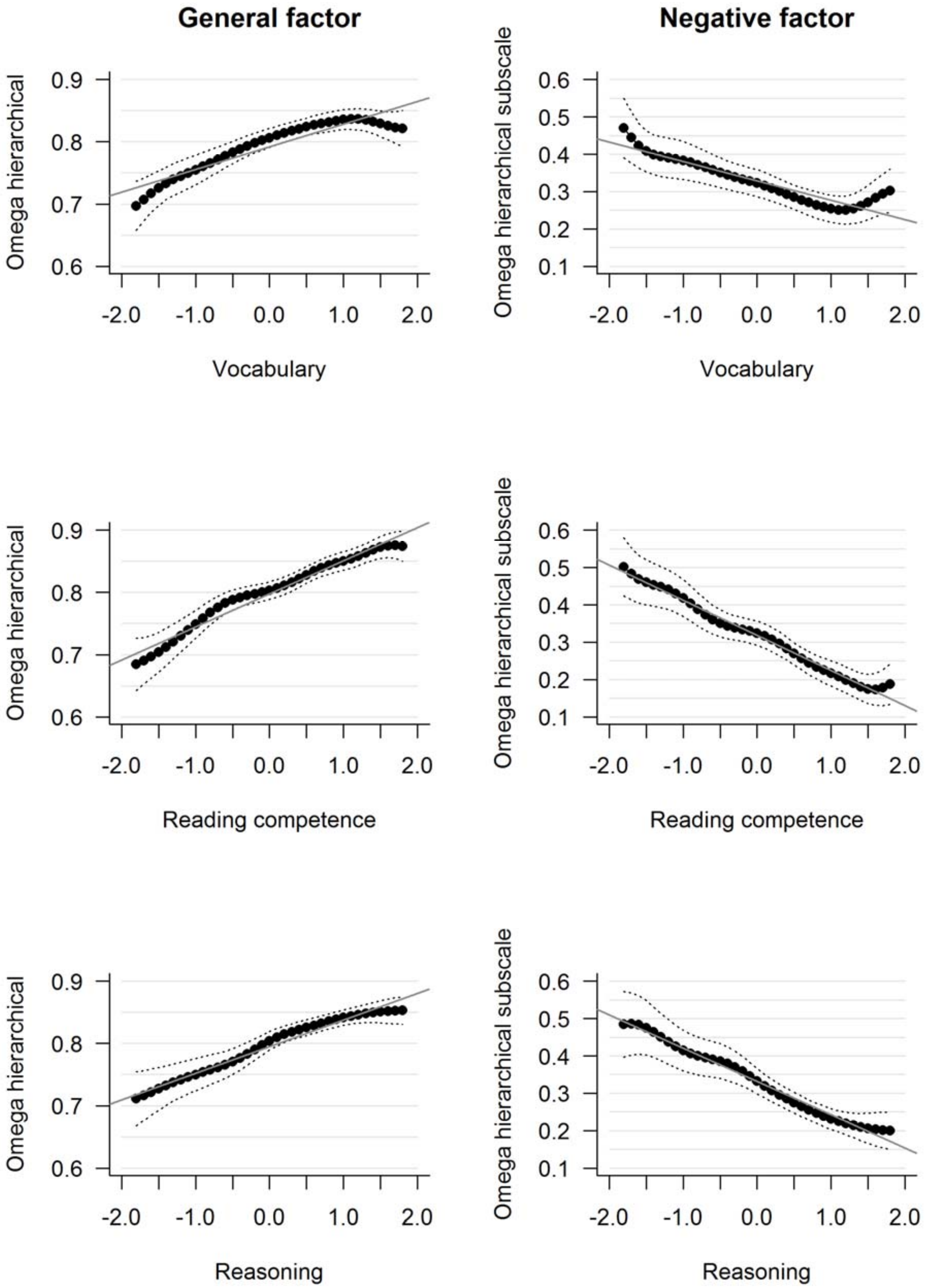

Figure 3. Reliability (black dots) for general and negative factors in the Rosenberg Self-Esteem Scale across *z*-standardized cognitive abilities with 95% confidence intervals (dashed lines) and regression line (gray solid line).

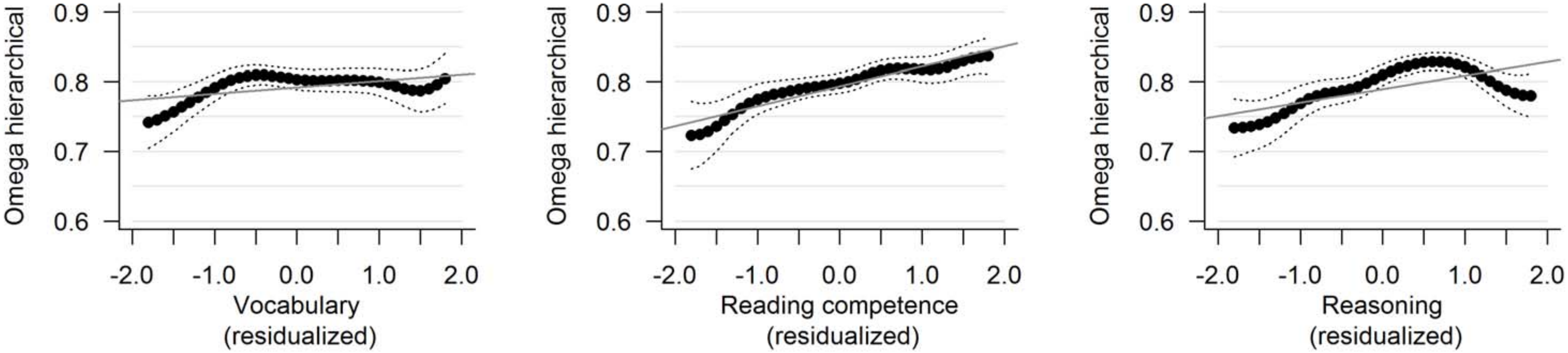

Figure 4. Reliabilities (black dots) for general factor in the Rosenberg Self-Esteem Scale across residualized cognitive abilities with 95% confidence intervals (dashed lines) and regression line (gray solid line).

Online Supplement for

"Cognitive Abilities Explain Wording Effects in the Rosenberg Self-Esteem Scale"

List of Tables

List of Figures

Table S1.

Means, Standard Deviations, and Correlations for the Items of the Rosenberg Self-Esteem Scale.

				Items of the Rosenberg Self-Esteem Scale												
	M	*SD*	MV	Item 1	Item 3	Item 4	Item 7	Item 10	Item 2	Item 5	Item 6	Item 8	Item 9	Vocab.	Read.	Reas.
Item 1	3.94	0.85	0.14		0.249	0.231	0.310	0.444	0.354	0.301	0.365	0.798	0.332	0.363	-0.009	0.037
Item 3	3.96	0.78	0.92	.380		0.299	0.276	0.295	0.213	0.247	0.213	0.158	0.217	0.724	0.024	0.027
Item 4	3.94	0.79	0.66	.345	.486		0.250	0.269	0.211	0.228	0.200	0.146	0.206	0.670	0.024	0.019
Item 7	3.99	1.00	0.98	.365	.354	.314		0.381	0.319	0.315	0.314	0.200	0.306	1.683	0.174	0.210
Item 10	3.93	0.92	1.00	.568	.413	.368	.411		0.412	0.345	0.428	0.348	0.436	0.470	-0.013	0.022
Item 2	3.69	1.08	0.94	.387	.254	.246	.294	.413		0.451	0.635	0.457	0.518	0.822	0.035	0.078
Item 5	3.97	1.00	0.85	.356	.319	.287	.314	.374	.417		0.481	0.359	0.413	1.061	0.114	0.127
Item 6	4.14	1.01	0.96	.425	.271	.249	.308	.457	.579	.474		0.423	0.555	-0.148	-0.045	-0.067
Item 8	3.55	1.16	2.89	.303	.177	.159	.172	.326	.366	.310	.361		0.450	1.421	0.122	0.135
Item 9	4.30	0.96	1.42	.408	.291	.269	.317	.491	.498	.430	.568	.405		0.058	-0.012	-0.015
Vocabulary	57.76	10.25	0.00	.042	.091	.082	.164	.050	.074	.104	-.014	.120	.006		7.188	10.099
Reading	0.03	1.25	0.00	-.008	.025	.025	.138	-.011	.026	.091	-.036	.084	-.010	.561		1.750
Reasoning	8.74	2.40	0.00	.018	.014	.010	.087	.010	.030	.053	-.028	.049	-.006	.411	.453	

Note. *N* = 12,437. MV = Percentage of missing values. Full information maximum likelihood estimation. Correlations are presented below the diagonal and covariances above. Gray cells indicate convergent correlations of positively keyed items (1, 3, 4, 7, 10) and negatively keyed items (2, 5, 6, 8, 9).

Table S2.

Correlations of Latent Factors with Cognitive Scores.

	Bifactor Model			Bifactor-(*S*-1) model		
Factor	Vocabulary	Reading	Reasoning	Vocabulary	Reading	Reasoning
General	.084*	.016	.035*	.113*	.036*	.040*
Positive	.116*	.062*	.007			
Negative	-.018	.003	-.027	-.056*	-.023	-.034*

Note. Bifactor-(*S*-1) model includes correlated residuals for items 3 and 4.

* *p* < .05

Table S3.

Results of Permutation Tests for LSEM on the Factor Reliabilities of the Bifactor Model.

	General factor (ω_H)					Negative factor ($\omega_{HS.NEG}$)					Positive factor ($\omega_{HS.POS}$)				
	M	*SD*	*p*(*SD*)	*B*	*p*(*B*)	*M*	*SD*	*p*(*SD*)	*B*	*p*(*B*)	*M*	*SD*	*p*(*SD*)	*B*	*p*(*B*)
Bivariate effects															
Vocabulary	.795	.036	< .001	.041	< .001	.316	.042	< .001	-.045	< .001	.140	.039	< .001	-.044	.02
Reading competence	.796	.053	< .001	.061	< .001	.314	.065	< .001	-.077	< .001	.140	.068	< .001	-.071	< .001
Reasoning	.799	.047	< .001	.056	< .001	.308	.060	< .001	-.071	< .001	.141	.070	< .001	-.082	< .001
Partial effects															
Vocabulary	.794	.014	.10	.006	.22	.321	.026	.10	.010	.56	.137	.016	.66	-.010	.44
Reading competence	.793	.026	< .001	.031	< .001	.325	.045	< .001	-.051	< .001	.137	.029	.02	-.031	.02
Reasoning	.797	.029	< .001	.028	< .001	.315	.048	< .001	-.044	.02	.139	.036	< .001	-.038	< .001

Note. *N* = 12,437. *M* = average ω_H across values of the moderator (range: [-1.8, 1.8]); *SD* = variation of ω_H across values of the moderator; *p*(*SD*) = *p*-value of the permutation test for *SD*; *B* = linear effect of the moderator on ω_H; *p*(*B*) = *p*-value of the permutation test for *B*. Partial effects are based on the residualized values of the moderators.

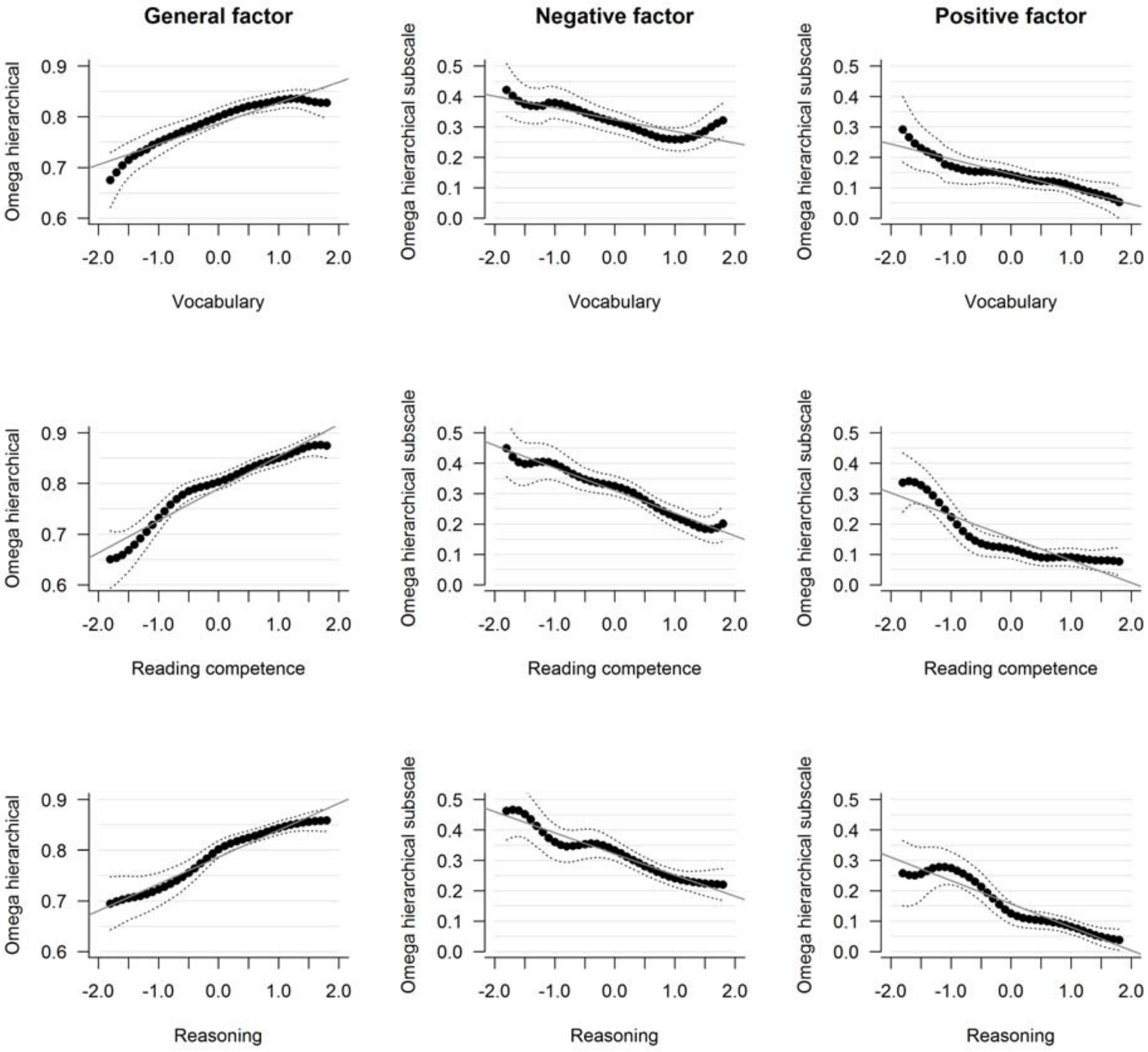

Figure S1. Reliability (black dots) for general and specific factors in the Rosenberg Self-Esteem Scale across *z*-standardized cognitive abilities with 95% confidence intervals (dashed lines) and regression line (gray solid line) for the bifactor model.

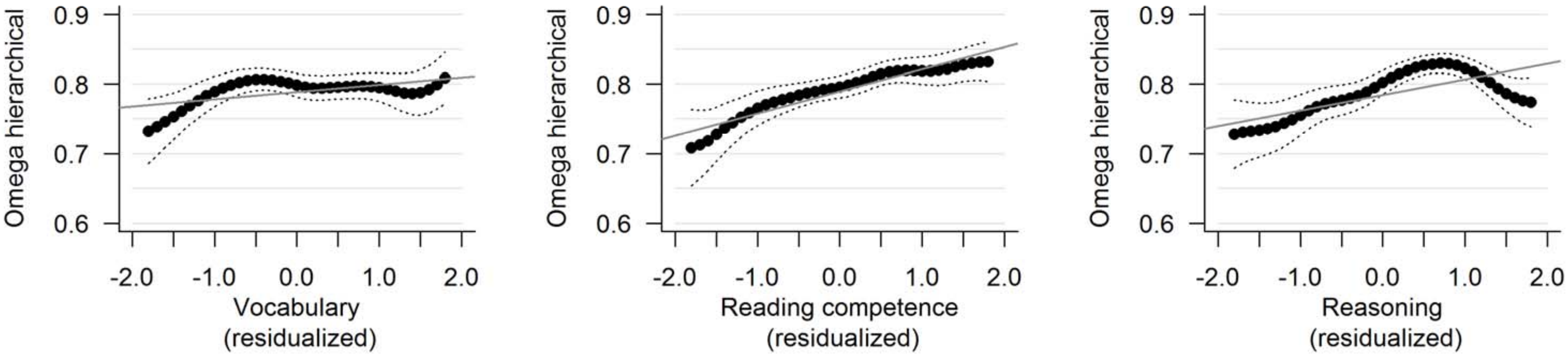

Figure S2. Reliabilities (black dots) for general factor in the Rosenberg Self-Esteem Scale across residualized cognitive abilities with 95% confidence intervals (dashed lines) and regression line (gray solid line) for the bifactor model.