Psychological Assessment
2010, Vol. 22, No. 2, 366–381

1040-3590/10/$12.00 DOI: 10.1037/a0019225

Longitudinal Tests of Competing Factor Structures for the Rosenberg Self-Esteem Scale: Traits, Ephemeral Artifacts, and Stable Response Styles

Herbert W. Marsh
University of Oxford

L. Francesca Scalas
University of Oxford and University of Cagliari

Benjamin Nagengast
University of Oxford

Self-esteem, typically measured by the Rosenberg Self-Esteem Scale (RSE), is one of the most widely studied constructs in psychology. Nevertheless, there is broad agreement that a simple unidimensional factor model, consistent with the original design and typical application in applied research, does not provide an adequate explanation of RSE responses. However, there is no clear agreement about what alternative model is most appropriate—or even a clear rationale for how to test competing interpretations. Three alternative interpretations exist: (a) 2 substantively important trait factors (positive and negative self-esteem), (b) 1 trait factor and ephemeral method artifacts associated with positively or negatively worded items, or (c) 1 trait factor and stable response-style method factors associated with item wording. We have posited 8 alternative models and structural equation model tests based on longitudinal data (4 waves of data across 8 years with a large, representative sample of adolescents). Longitudinal models provide no support for the unidimensional model, undermine support for the 2-factor model, and clearly refute claims that wording effects are ephemeral, but they provide good support for models positing 1 substantive (self-esteem) factor and response-style method factors that are stable over time. This longitudinal methodological approach has not only resolved these long-standing issues in self-esteem research but also has broad applicability to most psychological assessments based on self-reports with a mix of positively and negatively worded items.

Keywords: self-esteem, Rosenberg Self-Esteem Scale, method effects, longitudinal approach, structural equation models

Supplemental materials: http://dx.doi.org/10.1037/a0019225.supp

As quantitative psychologists in the area of psychological assessment, we live in exciting times. We have a range of new and evolving quantitative tools at our disposal to address a wide variety of substantive questions, with statistical power and flexibility that was previously unimaginable. However, this power comes at a cost. In order to make best use of these new tools, we must pursue research that is at the cutting edge of both the latest methodological developments and substantive issues: methodological–substantive synergies (Marsh & Hau, 2007). In the present investigation we illustrate the importance of this synergy, applying evolving methodological approaches to psychological assessment to resolve competing claims about the factor structure of one of psychology's most widely used instruments: the Rosenberg Self-Esteem Scale (RSE; Rosenberg, 1965).

Global self-esteem (GSE) is one of the most important constructs in psychology and is the basis of considerable theoretical and applied research. The RSE is clearly the most widely used scale to assess GSE and one of psychology's most widely used measures (Blascovich & Tomaka, 1991). Remarkably, however, there is a heated, ongoing, and as yet unresolved debate about the factor structure underlying responses to this instrument that fundamentally affects the interpretation of responses to it and the very meaning of GSE.

Typical applications of the RSE generally assume—at least implicitly—that both positively and negatively worded items are interchangeable and assess the same construct (DiStefano & Motl, 2006). Nevertheless, this assumption has been questioned for the RSE (e.g., Corwyn, 2000; DiStefano & Motl, 2006; Quilty, Oakman, & Risko, 2006; Wang, Siegal, Falck, & Carlson, 2001) and other self-esteem measures derived from it (e.g., Horan, DiStefano, & Motl, 2003; Marsh, 1986, 1996; Marsh & Grayson, 1994; Motl & DiStefano, 2002) or translated into non-English languages (Gana, Alaphilippe, & Bailly, 2005; Tomás & Oliver, 1999). A review of this literature on the structure of the RSE identifies four different perspectives: (a) one unidimensional GSE factor that is

Herbert W. Marsh and Benjamin Nagengast, Department of Education, University of Oxford, Oxford, England; L. Francesca Scalas, Department of Education, University of Oxford, and Department of Psychology, University of Cagliari, Cagliari, Italy.

This research was supported in part by a grant to Herbert W. Marsh from the United Kingdom Economic and Social Research Council. The data used in this article were made available by the Inter-University Consortium for Political and Social Research and were originally collected by Jerald Bachman.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, England. E-mail: herb.marsh@education.ox.ac.uk

consistent with Rosenberg's original design and its typical application; (b) two oblique GSE factors that represent different psychological constructs that should be interpreted as relatively distinct components of self-esteem—one based on positively worded items, the other based on negatively worded items; (c) one GSE trait factor and method effects associated with positively or negatively worded items that are ephemeral artifacts due to wording effects that have no substantive relevance; and (d) one GSE trait factor and method effects associated with item-wording effects that represent stable response styles. Although (d) is apparently new to RSE research and needs further investigation, the critical assumption is that these response-style effects are stable across time.

A Simple Unidimensional Self-Esteem Model

According to Rosenberg (1965), self-esteem is a unidimensional construct reflecting positive or negative attitudes toward the self, and in this sense it transcends evaluations of specific areas of functioning (Corwyn, 2000). For this reason, the RSE was originally designed to assess global self-esteem as one factor based on 10 items—a mixture of positively and negatively worded items. Hence, according to this perspective, GSE as measured by the RSE is a unidimensional construct that is consistent with Rosenberg's original design and its typical application in practice (see Model 1 in Figure 1).

Nevertheless, RSE studies using both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) have not supported the unidimensional structure. These results call into question the theoretical basis of the unidimensional structure of the RSE and the common use of its scale score in psychological practice and research. As highlighted by Corwyn (2000), if the scale taps more than one dimension, the use of a single scale might result in incomparable scores across participants, misinterpretation, and systematic bias in the interpretation of RSE responses. Therefore, it is particularly important to understand the structure underlying the RSE and derived measures of self-esteem.

A Bidimensional Model of Self-Esteem: Debate on the RSE and Derived Scales

EFAs of RSE responses often result in two oblique trait factors (see Model 2 in Figure 1)—one based on positively worded items, the other based on negatively worded items. Some researchers have interpreted these as substantially meaningful (Kaplan & Pokorny, 1969; Kaufman, Rasinski, Lee, & West, 1991; Owens, 1993, 1994; Prezza, Trombaccia, & Armento, 1997; Shahani, Dipboye, & Philips, 1990; Tafarodi & Milne, 2002;[1] Tafarodi & Swann, 1995); others have interpreted them as irrelevant method effects related to the wording of the items (Carmines & Zeller, 1979).

Carmines and Zeller (1979), for example, claimed that if the two factors were substantially meaningful, they should be differentially related to external criteria. Results showed no significant differentiation between the correlations based on the positive and negative factors with the 16 criteria used in the study, supporting a unidimensional structure of GSE with method effects. Subsequently, Owens (1993, 1994) used a CFA approach to compare a unidimensional model of self-esteem with positive and negative trait models of self-esteem in a two-wave sample. Owens (1994) argued for a bidimensional view of global self-esteem with a positive component that can be addressed as general self-confirmation (or positive self-worth) and a negative component addressed as self-deprecation. The major issue of this area of research is that the simplistic two-oblique-factor model does not allow the researcher to clearly distinguish between trait or substantive components and method components (if they actually exist).

A Unidimensional Self-Esteem Model With Ephemeral Method Effects: Wording Effects in Self-Esteem Scales as Methodological Artifacts

Several researchers introduced methodologically more sophisticated strategies to investigate RSE factor structure.

Method Effects and Self-Esteem Scales

Questionnaires widely used in psychology and the social sciences more generally can be affected by distortions associated with method effects. As stated by Bagozzi (1993) and others, a method effect is the variance linked to measurement procedures instead of the construct under investigation. Such method effects can lead to biased interpretations by suppressing or inflating links between variables. Response bias is a major problem in particular when self-report measures are used. In order to counteract response biases such as acquiescence, many researchers use positively and negatively worded items to address the same underlying construct (e.g., Anastasi, 1982; Billiet & McClendon, 2000; Paulhus, 1991), but this strategy introduces new problems.

Studies of method effects generally, not only with the RSE, have typically used one of two approaches: the correlated uniqueness (CU) strategy and the latent method factor (LMF) strategy (Bagozzi, 1993; Marsh & Grayson, 1995). The first approach resolves the issue of

[1] Tafarodi and colleagues (Tafarodi & Milne, 2002; Tafarodi & Swann, 1995) used a different approach, proposing that self-esteem can be considered a multifaceted construct formed by two substantive dimensions: self-competence and self-liking. In the RSE these two components are represented by items reflecting assessment of qualities and self-acceptance. Tafarodi and Milne (2002) contrasted three models: (a) one-factor structure, unidimensional self-esteem; (b) two-factor structure, positive and negative self-esteem; (c) two-factor structure, assessment and acceptance. Whereas the third model had the best fit to the data, they suggested that a combined model including all the five components should be examined. Indeed, according to the authors, the RSE items reflect to different extents a general common factor of self-esteem, self-acceptance and assessment, and positive and negative self-esteem. They found that their general five-factor model when applied to RSE responses resulted in a better fit than the three models focusing on separate components. Following this suggestion, in Wave 1 we tested the three models with separate components as well as the five-factor combined model with global self-esteem, positive and negative self-esteem, and acceptance and assessment (acceptance: Items 1, 5, and 6; assessment: Items 2, 3, 4, 7, 8, 9, and 10). The combined model presented the best fit indices (similar to those reported by Tafarodi & Milne, 2002), apparently supporting the model. However, inspection of parameter estimates showed nonsignificant factor loadings and nonsignificant variances for assessment and acceptance factors. Therefore, the model was not considered further in the present investigation.

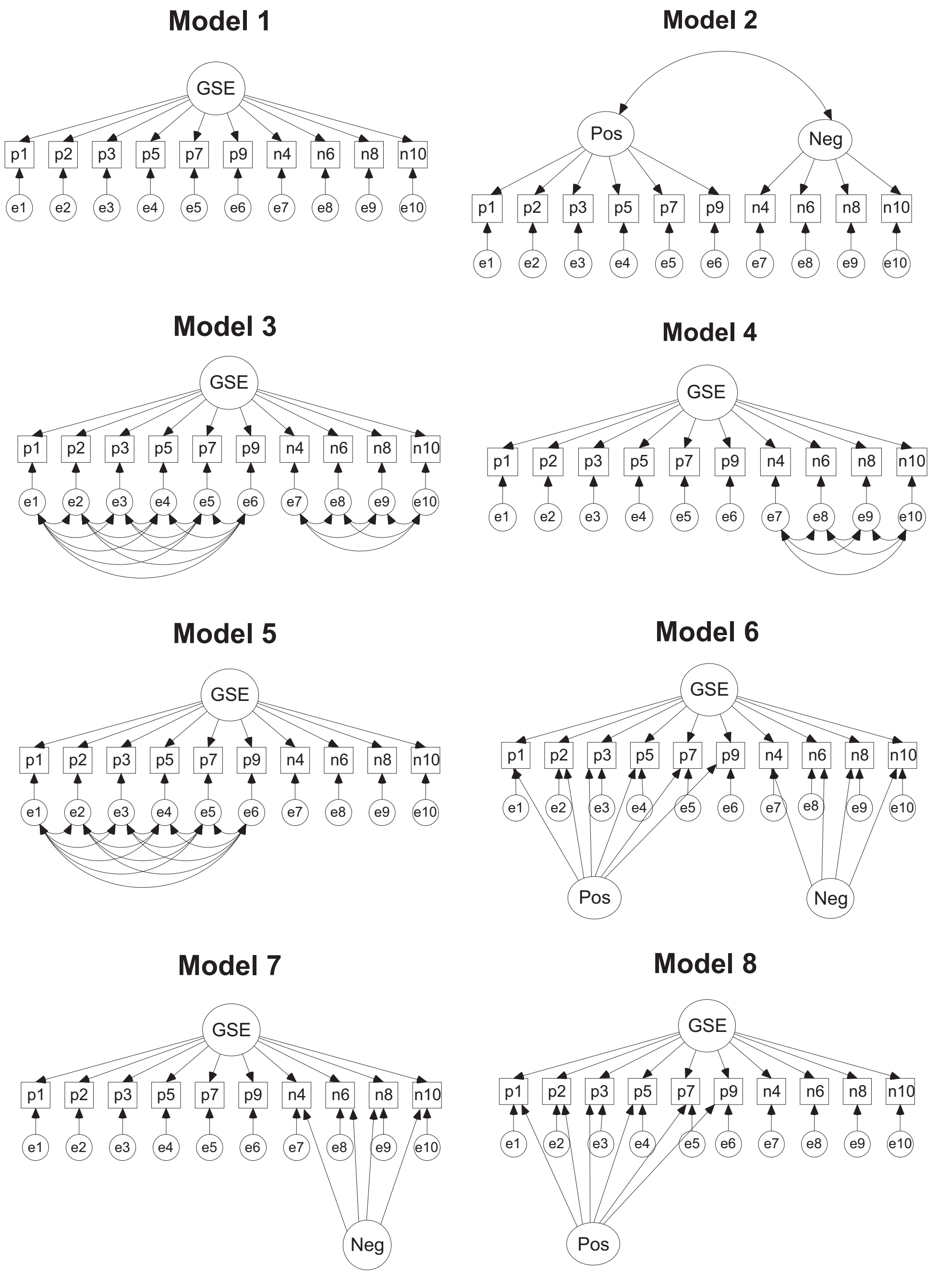
Model 1
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Model 2
Pos
Neg
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Model 3
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Model 4
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Model 5
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Model 6
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Pos
Neg
Model 7
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Neg
Model 8
GSE
p1 p2 p3 p5 p7 p9 n4 n6 n8 n10
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10
Pos

Figure 1. Eight structural equation models of self-esteem for single-wave data. Model 1 = one trait factor, no correlated uniqueness; Model 2 = two trait factors: correlated positive and negative trait factors; Model 3 = one trait factor with correlated uniqueness among both positive and negative items; Model 4 = one trait factor with correlated uniqueness among negative items; Model 5 = one trait factor with correlated uniqueness among positive items; Model 6 = one trait factor plus positive and negative latent method factors; Model 7 = one trait factor plus a negative latent method factor; Model 8 = one trait factor plus a positive latent method factor; p = positive items; n = negative items; e = error.

method effects by introducing correlations among the positively worded items and/or among the negatively worded items (see Models 3, 4, and 5 in Figure 1; e.g., Bachman & O'Malley, 1986; Marsh & Grayson, 1994). The second strategy introduces specific LMFs that capture the variance between the items with the same method (see Models 6, 7, and 8 in Figure 1). Both are based in part on the logic of multitrait–multimethod (MTMM) paradigms. This MTMM literature highlights strengths and weaknesses of both the LMF and CU approaches (Horan et al., 2003; see also Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Eid et al., 2008). In the

CU model, method effects are represented as correlated uniquenesses (e.g., Kenny & Kashy, 1992; Marsh, 1989; Marsh & Bailey, 1991; Marsh & Grayson, 1995; see also Horan et al., 2003; Lance, Noble, & Scullen, 2002). As adapted to studies of the RSE (Marsh, 1996), the CU model posits one GSE trait factor and CUs among positively worded and/or negatively worded items (see Models 3, 4, and 5 in Figure 1). As in the traditional MTMM-CU model, method effects associated with one method (e.g., positively worded items) are assumed to be uncorrelated with method effects associated with another method (e.g., negatively worded items). The strength of the CU model in MTMM studies is that it almost always converges to a proper solution, whereas models representing method effects as latent factors typically do not (Marsh & Bailey, 1991). However, because method effects are represented as a set of CUs rather than a separate factor, it is not so easy to summarize the size of the methods effects and relate them to other variables.

Using the LMF strategy, it is possible to directly estimate trait and method effects and to separate method variance from error variance, an elegant decomposition of variance associated with trait and method effects. However, nonconvergence, improper solutions (e.g., models with out-of-range parameter estimates such as negative variance estimates or factor correlations greater than 1.0), and admissibility problems (due to empirical underidentification) are typical in MTMM studies, particularly when the multiple methods factors are correlated (Lance et al., 2002; Marsh & Bailey, 1991; Marsh & Grayson, 1995; Quilty et al., 2006). For these reasons, it has been suggested that both approaches be used (e.g., Byrne & Goffin, 1993; Marsh & Grayson, 1995), although Lance et al. (2002) claimed that when convergent and admissible solutions are found, the LMF approach should be preferred. However, it is important to note that this area is still debated in the literature (cf. Brown, 2006) and that some authors supported the use of CUs (e.g., Marsh, 1996; Marsh & Grayson, 1994; Whiteside-Mansell & Corwyn, 2003).

Research Outcomes of Wording Effects as Artifacts

There are many studies pointing to the artifact nature of wording effects (e.g., Corwyn, 2000; Gray-Little, Williams, & Hancock, 1997; Greenberger, Chen, Dmitrieva, & Farruggia, 2003; Marsh, 1986, 1996). Some authors found that the negative item effects are stronger than the positive ones, showing better fit for the models that include negative method effects based on CUs or LMFs compared to models including only positive effects (e.g., Corwyn, 2000; DiStefano & Motl, 2006; Horan et al., 2003; Marsh, 1986, 1996; Motl & DiStefano, 2002; Schmitt & Allik, 2005). Schmitt and Allik (2005) translated the RSE into 28 languages and administered it to almost 17,000 participants from 53 countries. They found that "in many cultures the answers to negatively worded items are systematically different from the answers to positively worded items" (Schmitt & Allik, 2005, p. 638). More generally, in relation to negatively worded items, some authors argued that the negative method effect might be related to age and verbal ability (e.g., Corwyn, 2000; Marsh, 1996).

Although the different functioning of positive and negative item-wording effects might be a relevant issue, it is important to note that most authors have considered only a limited set of models. For example, Marsh (1996) did not test models with LMFs; DiStefano and colleagues (DiStefano & Motl, 2006; Horan et al., 2003) examined positive and negative factors separately (for both method factors and correlated uniqueness); and Corwyn (2000) examined positive and negative LMFs jointly but not separately. Furthermore, some researchers argued for models including only positive method effects (Aluja, Rolland, García, & Rossier, 2007; Dunbar, Ford, Hunt, & Der, 2000), although they tested only a few of the structural models considered here.

In summary, few of the studies have contrasted models with only one method effect with models including both method effects (or vice versa). Studies where positive and negative factors were assessed jointly, as well as separately, typically showed that the models including both positive and negative factors were preferable (Gana et al., 2005; Quilty et al., 2006; Tomás & Oliver, 1999).

A Unidimensional Self-Esteem Model With Substantive Method Effects: Method Factors as Response Styles

Implicit in many studies that treat method effects as artifacts is the assumption that method effects associated with item wording are ephemeral and substantively irrelevant. They must be included in the model, but their purpose is mainly to purge the GSE trait factor of contaminating method effects. However, some researchers have claimed that these method effects can be interpreted as stable response styles (DiStefano & Motl, 2006; Jackson & Messick, 1962; Quilty et al., 2006). Although different researchers offer alternative interpretations of these response styles, their critical characteristic is stability over time. According to Bentler, Jackson, and Messick (1971), response styles can be operationalized as latent constructs which are associated with stable response tendencies. Thus, although stability of LMFs over time does not constitute a definitive proof that method effects represent more than an ephemeral artifact, it is an essential requirement. Stability over time is one of the criteria to identify response styles (Bentler et al., 1971; see DiStefano & Motl, 2006; Horan et al., 2003). Furthermore, the importance of investigating the stability of method factors over time has been addressed several times in the literature (Bentler et al., 1971; Billiet & McClendon, 2000; Horan et al., 2003; Motl & DiStefano, 2002; Tomás & Oliver, 1999). Nevertheless, previous research has not explored this issue in depth—particularly in relation to the factor structure underlying responses to the RSE. Hence, an important focus of the present investigation is to evaluate whether method effects in RSE responses are fleeting or stable and to evaluate the effectiveness of alternative models designed to control for these effects.

Measurement Invariance as a Prerequisite of Construct Stability

A critical prerequisite to test the stability of latent constructs is to establish factorial invariance of responses to the same instruments over time, but this has rarely been evaluated in research based on the RSE. Depending on the model and hypotheses to be investigated, several sets of parameters are analyzed "in a logically ordered and increasingly restrictive fashion" (Byrne, 2004, p. 273), from the most unrestrictive hypothesis (configural invariance or unconstrained model) to the more restrictive ones (e.g., latent means).

Marsh and Grayson (1994; Marsh, 2007; Marsh et al., 2009) emphasized that different aspects of longitudinal variance are relevant, depending on the focus of an investigation. Tests based on covariance matrices are useful in testing the invariance of the factor structure over time (covariance stability). However, if researchers want to test the stability of latent means, analyses must include item means as well as the covariance matrix. In relation to covariance stability, equivalence of factor loadings over time is the most commonly studied invariance test and is the starting point for other more demanding tests of invariance (Bollen, 1989; Byrne, 2004; Cheung & Rensvold, 1999, 2002; Marsh & Grayson, 1994). Another invariance test is for the latent variance–covariance matrix that is the basis for contrasting correlations of latent variables in different groups or waves (Cheung & Rensvold, 1999; Jöreskog & Sörbom, 1996–2001). Finally, the invariance of error variances and covariances is relevant to examining the reliability across groups or waves (Byrne, 2004; Cheung & Rensvold, 1999; Jöreskog & Sörbom, 1996–2001). In multiwave studies in which the same measures are administered on more than one occasion, tests of invariance over time are an important prerequisite for making valid comparisons across constructs (e.g., Anderson & Gerbing, 1988; Leite, 2007; Muthén & Muthén, 1998–2006). This is particularly true when the focus is on tests of latent mean differences over time, which require the invariance of the item intercepts over time as well as the invariance of the factor loadings. Finally, tests based on manifest means also require the invariance of item uniquenesses (measurement error) over time. As noted by Marsh and Grayson (1994; Marsh et al., 2009), these prerequisites are in line with item response theory, in which factor loadings represent the discrimination parameters (slopes) and intercepts represent the difficulty parameters. Therefore, if slope and difficulty parameters remain invariant over time, changes in the means can reasonably be interpreted as changes in the constructs.

Outcomes From Previous Research on Construct Stability for the RSE

Marsh and Grayson (1994) considered the problem of both covariance and latent mean stability for RSE responses in a pioneering study that evaluated the implications of different levels of invariance. However, they focused only on CU models for negatively worded items; they did not consider LMF models or any models of method effects for positively worded items. Similarly, comparing samples of adolescents and adults, Whiteside-Mansell and Corwyn (2003) tested invariance for latent means only for a CU model for positive and negative items, whilst Motl and DiStefano (2002) examined invariance for an LMF model with a negative latent factor. Using two waves of data, Corwyn (2000) and Marsh (1996) reported CFAs in relation to several models. However, neither study examined longitudinal CFAs in which different waves of the same sample are related in one single model; moreover, they did not consider invariance across time. Similarly, Horan et al. (2003), using a simplex model (see Marsh & Grayson, 1994), investigated stability of wording effects over time (three occasions) for a model including GSE and a negative LMF. Nevertheless, they did not test measurement invariance as a precondition of construct stability.

The Present Investigation

Research Question: Eight Alternative Models

In the present investigation we extend the debate on the RSE structure that has plagued research for more than 30 years. We employ a methodological–substantive synergy based on stronger statistical methodology to address important complex substantive issues. Our overarching research question is how the RSE factor structure is best characterized in relation to the four perspectives reviewed earlier. We pursue this question on the basis of a comprehensive set of eight alternative models (see Figure 1) based on taxonomies of MTMM models that have been developed to unconfound trait effects from methods effects and a combination of cross-sectional (single-wave) and longitudinal (multiwave) data. We consider these models in two separate studies.

Model 1 posits a single GSE factor with no method effects (i.e., no CUs or LMF factors), consistent with the original design of the RSE and most applied research.

Model 2 posits two oblique factors; latent factors defined by positively and negatively worded items are purported to have a psychological meaning, and there is no overarching GSE trait factor. Therefore, if method effects exist, they are confounded with the trait factors. This model is based on EFA studies and is the bidimensional self-esteem model.

Models 3–8 examine the issue of method effects; they take into account method effects but use different strategies to do so.

Models 3, 4, and 5 are based on the CU approach. Model 3 posits a single GSE factor with separate sets of method effects for negatively worded items (CUs among negatively worded items) and positively worded items (CUs among positively worded items). Each set of method effects is uncorrelated with the GSE trait factor and uncorrelated with the other. Models 4 and 5 are submodels of the more general Model 3 in which method effects are posited only for negatively worded items (Model 4) or only for positively worded items (Model 5). The juxtaposition among the three models is important in establishing the relative importance and substantive nature of the method effects. Indeed, some research (e.g., Marsh, 1996) suggests that method effects are primarily associated with negatively worded items so that a model like our Model 4 should be preferred.

Models 6, 7, and 8 resolve the method issue by including LMFs correlated neither with the GSE trait factor nor with each other—the LMF approach. In Model 6, both positive and negative LMFs are specified. In Model 7 only a negative LMF is included. In Model 8, only a positive LMF is considered. Again, the juxtaposition among the models is substantively important in terms of assessing the relative importance of method effects associated with positively and negatively worded items.

Study 1. Before exploring the alternative models and the apparent inconsistency based on previous research, we test the robustness of the alternative models using simulated data. Model robustness, as used here, refers to the ability of a given model to consistently converge to a fully proper solution in which parameter estimates provide reasonable approximations to population parameters. This is an important issue, as demonstrated by nonconvergence and instability, that is a well-known problem in the CFA approach to MTMM data (see Marsh & Bailey, 1991) and the basis of many of the models considered here. Although the RSE factor structure must ultimately be tested with real data, model instability can be best evaluated in relation to simulated data in which the true population parameters are known. To the extent that any of the eight posited models is not able to provide accurate estimates of known population parameters based on the matching population-generating model, it would provide a dubious basis for estimating parameters from real data.

Study 2. In Study 2 we compare results from the eight models (see Figure 1) on the basis of fit indices and parameter estimates from both cross-sectional (single-wave) and longitudinal (multiwave) data. Although previous research is not completely consistent, it suggests two things: (a) Whilst it is preferable to test both CU and LMF methods, LMF approaches have important advantages—particularly for longitudinal data; and (b) at least some previous research comparing models including one and two method effects suggests that models with both positive- and negative-item method effects fit better than models incorporating only one method effect. Therefore, we expect that Model 6, based on LMF and including both positive and negative method effects, would represent the best factor structure of the eight proposed here—particularly for longitudinal data and so long as it results in a fully proper solution.

We suggest that the ambiguous results from previous research might be due in part to overreliance on a simple single-wave perspective. In particular, the distinction between method effects as ephemeral artifacts and method effects as stable response-style effects cannot be adequately tested with cross-sectional (single-wave) data. In this respect, the extension of the application of the eight models to incorporate longitudinal data is essential in testing this substantively important distinction and the associated theoretical models. Unfortunately, multiwave RSE studies are not common and have typically tested only a few of the model structures considered here or have not adopted a longitudinal perspective.

The use of longitudinal data also provides important tests of the invariance of parameter estimates over time, that is, their stability over time. In particular, to effectively test the stability of method effects associated with the RSE, it is important not only to inspect the RSE factor structure within each wave but also its stability over time. As a result, particular attention is given to the measurement invariance over time of factor loadings, item intercepts, and latent factor means, providing a test of whether the meaning of the factors has changed over time. Once measurement invariance has been established, it is possible to investigate the stability of constructs, and particularly the interpretation of method effects as short-term ephemeral effects or response-style effects that are stable over time.

Study 1

Method

Design and population. Using the Mplus 5.1 program (Muthén & Muthén, 1998–2006), we generated a simulated population according to each one of the eight models, and then we tested the true model with 500 replications of 2,000 cases each. The population parameters common to the models were as follows: *factor loadings* = .5, *error variances* = .2, *latent variances* = 1. For CU models, population correlations among uniquenesses for positively worded items and population correlations among uniquenesses for negatively worded items were .1; correlations among uniquenesses between positively and negatively worded items were zero. For LMF models, the LMFs were simulated to be uncorrelated with the GSE factor and uncorrelated with each other in Model 6. For the simulated data, there were five positively worded items and five negatively worded items. However, because the data were randomly generated, there was no difference between the positively and negatively worded items such that models for the positive and negative method effects were interchangeable. For this reason, models including only one method effect (for both CU and LMF) required only one simulation. Therefore only six models were examined: (a) Model 1 with one trait factor for self-esteem and no method effects, (b) Model 2 with two trait factors, (c) Model 3 with CU for both method effects, (d) Model 4/5 with CU for one method effect, (e) Model 6 with LMF for both method effects, and (f) Model 7/8 with LMF for one method effect.

The main purpose of Study 1 was to explore the stability of the models and the appropriate convergence of parameter estimates to the known population parameters when the true model was specified (thus no misspecification was included). The main criterion to evaluate model stability was the number of samples that converged to a proper solution. However, we also considered fit indices (χ^2, Akaike information criterion [AIC], root-mean-square error of approximation [RMSEA]; see Table 1), parameter estimates, and the variability of parameter estimates (see Table 2). In particular, for parameter estimates the coverage rates were considered. The coverage rate gives, for each parameter, the proportion of replications for which the 95% confidence interval contains the true population value (Muthén & Muthén, 1998–2006).

Results and Discussion

For each model, a summary of the average fit index values across the replications is provided (see Table 1). The means and

Table 1
Number of Completed Replications and Average Fit Indices for the Simulated Samples

Variable	Model 1	Model 2	Model 3	Model 4/5	Model 6	Model 7/8
Completed replications	500	500	324	500	500	500
χ^2 (*df*)	35.36 (35)	34.54 (34)	16.30 (15)	25.25 (25)	25.36 (25)	30.46 (30)
χ^2 *SD*	8.68	8.41	5.62	7.13	7.31	8.27
RMSEA	.005	.005	.007	.005	.005	.005
RMSEA SD	.006	.006	.007	.006	.006	.006
AIC	30,289	33,906	22,101	26,269	37,483	34,054
AIC SD	202	191	198	191	201	194

Note. See Figure 1 for a description of the various models. *SD* = empirical standard deviation over the completed replications; *RMSEA* = root-mean-square error of approximation; *AIC* = Akaike information criterion.

Table 2
Parameter Estimates Averaged Over 500 Simulated Samples

Model	Population parameter	Estimates average	*SD*	*SE* average	95% coverage
Model 1					
Factor loadings	.50	.5004	.0116	0.0117	.950
Variance	1.00	.9980	.0389	0.0379	.946
Residual variances	.20	.1998	.0071	0.0071	.948
Model 2					
Factor loadings	.50	.5000	.0122	0.0123	.954
Factor correlation	.40	.3985	.0267	0.0265	.950
Variances	1.00	.9994	.0389	0.0391	.942
Residual variances	.20	.1997	.0084	0.0084	.947
Model 3					
Factor loadings	.50	.5005	.0105	1.9099	.979
Correlated uniquenesses	.10	.0994	.0066	2.0519	1.000
Variances	1.00	.9956	.0291	6.8290	1.000
Residual variances	.20	.1993	.0081	2.2224	1.000
Model 4/5					
Factor loadings	.50	.5002	.0108	0.0109	.950
Correlated uniquenesses	.10	.0997	.0071	0.0068	.944
Variances	1.00	.9968	.0374	0.0386	.950
Residual variances	.20	.1997	.0079	0.0078	.948
Model 6					
Factor loadings	.50	.5001	.0213	0.0222	.959
Variances	1.00	.9978	.0528	0.0526	.952
Residual variances	.20	.1996	.0115	0.0115	.949
Model 7/8					
Factor loadings	.50	.4995	.0141	0.0143	.954
Variances	1.00	.9996	.0436	0.0436	.945
Residual variances	.20	.1996	.0083	0.0083	.951

Note. See Figure 1 for a description of the various models. Five hundred simulated data sets were generated for each model with known population parameters and then estimated with the same model. Shown are the population parameters, the average of parameter estimates, the (empirical) standard deviation of parameter estimates, the average standard error for each parameter reported in the analysis of each simulated data set, and the 95% coverage rate (the percentage of the 500 solutions that contain the true population parameter in the 95% confidence interval).

standard deviations of the fit indices are reasonable for all the models and do not highlight particular problems. Parameter estimates were able to capture the known population parameters (see Table 2), and coverage ratings closely approximated the expected value of 95%. The evaluation of the convergence behavior for the different models is more interesting. In particular, all but one of the models resulted in fully proper solutions for all 500 replications. The only exception was Model 3, which included two sets of CUs, representing method effects associated with positively and negatively worded items. For Model 3, only 324 samples converged to proper solutions, thus calling into question the usefulness of the model. Although clearly the model needs further research, we suggest that this problem is associated with empirical underidentification due to overparameterization. It is relevant that the number of estimated parameters in Model 3 (41) is substantially greater than any of the other models considered here (Model 1: 21; Model 2: 23; Model 4/5: 31; Model 6: 36; Model 7/8: 28).

Study 2

Method

Participants. Data were drawn from the Youth in Transition (YIT) study (Bachman, 2002) that included only boys. A two-stage sampling scheme was used. In Stage 1, a random sample of 87 U.S. public high schools was selected, and in Stage 2 a group of approximately 25 students was selected from each school. The database composition in the present investigation was as follows: Wave 1: early 10th grade (N = 2,213); Wave 2: late 11th grade (N = 1,886); Wave 3: late 12th grade (N = 1,799); Wave 4: 1 year after normal high school graduation (N = 1,620). Overall, the percentage of missing values across the four waves was 15.94 (Wave 1 = 1.15%; Wave 2 = 15.62%; Wave 3 = 19.24%; Wave 4 = 27.75%).

Instruments. A 10-item scale derived from the RSE was used to assess self-esteem in the YIT study. Six items are positively worded (e.g., "I feel that I'm a person of worth, at least on an equal plane with others"), and four items are negatively worded (e.g., "I feel that I can't do anything right"). A 5-point scale ranging from *almost always true* (1) to *never true* (5) was used.

Analyses. Analyses were based on structural equation models (SEMs) for both single-wave and longitudinal CFAs, as well as on tests of invariance and investigation of structural correlations over time to test the stability of the constructs. We used the full information maximum likelihood estimator (FIML) to deal with missing data (Enders & Bandalos, 2001; Muthén & Muthén, 1998–2006).

1. We tested the eight CFA models (see Figure 1) separately for each wave, on the basis of both fit indices and substantive interpretations of parameter estimates.

2. We performed longitudinal CFAs across all four waves in relation to each of the eight models. In longitudinal CFAs we included correlations across time within the same latent factor (e.g., GSE1, GSE2, GSE3, and GSE4). In these longitudinal models, correlations among the same items across the four waves (i.e., CUs) were posited to control for error measurement due to the use of repeated measures (for further discussion, see Jöreskog, 1979; Marsh, 2007; Marsh & Hau, 1996).

3. Subsequently, we extended the models to include tests of invariance of parameter estimates over time. First, we tested configural invariance in which none of the parameter estimates was constrained to be the same on different occasions. The correlations among latent trait factors were also examined, as changes in the correlations for trait factors when method effects are included in the model would support the need to take account of method effects when exploring the structure of self-esteem. Next, we tested the invariance of the factor loadings over time, followed by tests of the invariance of the latent factor variances. Subsequently, although not central for our purposes, measurement errors were set to be equal across time. In order to examine mean differences, we then tested invariance of intercepts.

4. Finally, we examined the stability of LMFs over time. Completely unstable LMFs would support the interpretation of method effects as ephemeral artifacts, whereas stable LMFs would support their interpretation as stable response styles.

Fit indices. Following Marsh, Balla, and Hau (1996; see also Marsh, 2007; Marsh, Balla, & McDonald, 1988; Marsh, Hau, & Wen, 2004), we considered the Tucker–Lewis index (TLI), the comparative fit index (CFI), and the RMSEA to evaluate goodness of fit in SEMs as well as the χ^2 test statistic (recognizing that it is sensitive to the number of parameters in the model and to sample size) and an evaluation of parameter estimates. The TLI and CFI vary along a 0-to-1 continuum in which values greater than .90 and .95 are typically taken to reflect acceptable and excellent fits to the data, respectively. RMSEA values less than .06 are taken to reflect a reasonable fit, whereas RMSEA values greater than .10 are unacceptable, although no golden rules exist (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh et al., 2004). The CFI contains no penalty for a lack of parsimony, so that improved fit due to the introduction of additional parameters may reflect capitalization on chance, whereas the TLI and RMSEA contain penalties for a lack of parsimony (for further discussion, see Cheung & Rensvold, 2002; Hu & Bentler, 1999; Marsh et al., 2004). In addition, for tests of invariance, we also considered χ^2 difference tests ($\Delta\chi^2 = \chi^2_2 - \chi^2_1$, based on the correction factors for robust χ^2 estimates; see http://www.statmodel.com/chidiff.shtml) and TLI differences (ΔTLI).

Results

Self-esteem structure in separate analyses of responses from Waves 1–4. In this section we examine the eight models (see Figure 1) based on CFAs of data from Waves 1 to 4 separately. Models[2] were compared in relation to fit indices (see Table 3) and parameter estimates (see Table 4). These results provide a basis of comparison for subsequent results—more complex longitudinal models based on data from all four waves. Also, it is important to emphasize that most studies of RSE structure are based on a single wave of data like those presented in this section.

Model 1 (see Figure 1), as expected from previous research, was not able to fit the data adequately (e.g., *TLI* = .700–.753; *RMSEA* = .089–.095) and is clearly the worst fitting model. The two-trait model of self-esteem (Model 2) fit the data better than Model 1 (e.g., *TLI* = .936–.961; *RMSEA* = .032–.046) but worse than all other models including method effects associated with positively and negatively worded items (Models 3–8). Among the latter, Model 3 results were problematic in Wave 1, as the solution was improper, consistent with our simulation results in Study 1 showing that Model 3 was the only model that frequently did not converge to a fully proper solution. However, Model 3 solutions for Waves 2–4 converged to fully proper solutions, showing good fit indices (*TLI* = .987–.994; *RMSEA* = .015–.020). Fit indices were good for Model 4 (*TLI* = .934–.959; *RMSEA* = .033–.046) and particularly good for Model 5 (*TLI* = .989–.995; *RMSEA* = .017–.026). However, factor loadings (Model 4 range: .22–.66; Model 5 range: .19–.63) and uniquenesses (Model 4 range: .56–.95; Model 5 range: .60–.96) were not completely satisfactory. Particularly, the low factor loadings for GSE might indicate substantial method effects, especially for Model 5. Model 6, including both positive and negative LMFs, showed good fit indices (e.g., *TLI* = .956–.970; *RMSEA* = .028–.037) and generally performed well for all four waves of data. Model 7, including only the negative LMF, had adequate fit indices (e.g., *TLI* = .932–.957; *RMSEA* = .034–.047). Finally, for Model 8 fit indices were as good as for Model 6 (e.g., *TLI* = .970–.976; *RMSEA* = .028–.039). (See footnote 2 regarding the results of other models.)

On the basis of separate analyses from each wave, it was hard to select a best model. This ambiguity as to which model is best is consistent with previous research, although few previous studies have evaluated such an extensive set of models—even for a single wave of data. Across all four waves Model 6 was apparently the best model overall. However, even though Model 3 resulted in an improper solution in Wave 1, it behaved well in Waves 2–4, and it provided the best fit of any of the models in these waves. We note, however, that this pattern of results is consistent with findings from the simulation results (Study 1) in which more than 1/3 of the solutions based on Model 3 were improper, but nearly 2/3 of the solutions were proper. Hence, based on these results, reliance on Model 3 seems to be problematic. In the next section we adopt a longitudinal perspective in order to more fully evaluate this set of models.

Individual differences stability: Factor structures across multiple waves. In this section, we begin with longitudinal models that do not require any of the parameter estimates to be the same across the four waves of data. These are referred to as configural invariance (unconstrained) models in which only the pattern of parameter estimates is assumed to be consistent across waves—not the actual values of the estimated parameters. Of interest in their own right, these models also provide a baseline for

[2] In order to conserve space, we summarize the results for all eight models but report in detail only the results of Models 1, 2, 5, and 6. However, results for Models 4 and 5 (which are submodels of Model 3) and for Models 7 and 8 (which are submodels of Model 6) are presented in the supplemental materials at http://dx.doi.org/10.1037/a0019225.supp

Table 3
Confirmatory Factor Analyses and Invariance Tests

Model	χ^2	*df*	*cf*	*TLI*	*CFI*	*RMSEA*
Model 1 (one trait factor, no correlated uniqueness)						
Single-wave CFAs						
1.1 wave 1	651.57**	35		.700	.767	.089
1.2 wave 2	624.03**	35		.710	.775	.095
1.3 wave 3	539.94**	35		.752	.807	.090
1.4 wave 4	542.04**	35		.753	.808	.095
Longitudinal CFAs (Model 1.5)						
1.5a Unconstrained model (UM)	2,930.10**	674	1.203	.838	.860	.039
1.5b Factor loadings (FL)	2,971.55**	701	1.200	.843	.859	.038
1.5c FL & Variances (Var)	2,975.17**	704	1.199	.844	.859	.038
1.5d FL-Var-Uniquenesses (Uniq)	3,134.95**	724	1.203	.839	.850	.039
Model 2 (two trait factors: positive and negative correlated factors)						
Single-wave CFAs						
2.1 wave 1	111.70**	34		.961	.971	.032
2.2 wave 2	120.16**	34		.956	.967	.037
2.3 wave 3	161.46**	34		.936	.951	.046
2.4 wave 4	133.00**	34		.950	.962	.043
Longitudinal CFAs (Model 2.5)						
2.5a UM	1,116.35**	652	1.199	.965	.971	.018
2.5b FL	1,152.80**	676	1.194	.966	.970	.018
2.5c FL & Var	1,161.67**	682	1.194	.966	.970	.018
2.5d FL-Var-Uniq	1,313.80**	702	1.200	.958	.962	.020
Model 3 (one trait factor with correlated uniqueness among both positive and negative items)						
Single-wave CFAs						
3.1 wave 1[a]	—	—	—	—	—	—
3.2 wave 2	24.23*	14		.987	.996	.020
3.3 wave 3	22.74	14		.989	.997	.019
3.4 wave 4	19.05	14		.994	.998	.015
Longitudinal CFAs (Model 3.5)						
3.5a UM	1,337.64**	590	1.180	.939	.954	.024
3.5b FL	1,365.49**	617	1.175	.941	.954	.023
3.5c FL & Var	1,373.87**	620	1.175	.941	.953	.023
3.5d FL-Var-Uniq	1,511.19**	640	1.176	.934	.946	.025
Model 6 (one trait factor plus positive and negative latent method factors)						
Single-wave CFAs						
6.1 wave 1	69.62**	25		.970	.983	.028
6.2 wave 2	70.62**	25		.969	.983	.031
6.3 wave 3	88.48**	25		.956	.976	.038
6.4 wave 4	78.83**	25		.963	.980	.037
Longitudinal CFAs (Model 6.5)						
6.5–0 No correlations for the same method factor over time	1,483.37**	634	1.186	.935	.947	.025
6.5a UM	916.52**	622	1.182	.977	.982	.015
6.5b FL	962.06**	673	1.188	.979	.982	.014
6.5c FL & Var	1,000.90**	682	1.189	.977	.980	.015
6.5d FL-Var-Uniq	1,161.90**	702	1.196	.968	.971	.017

Note. See Figure 1 for a description of the various models. In the FL-Var-Uniq models, uniqueness across Waves 2–4 were constrained to be invariant, and error variances for Wave 1 have been released following the Marsh & Grayson (1994) procedure used on the same data. In order to conserve space and facilitate presentation, results for Models 4 and 5 (submodels of Model 3) and Models 7 and 8 (submodels of Model 6) are presented in the supplemental materials. χ^2 = chi-square test statistic; *df* = degrees of freedom; *cf* = robust maximum likelihood (MLR) correction factor; *TLI* = Tucker–Lewis index; *CFI* = comparative fit index; *RMSEA* = root-mean-square error of approximation; CFA = confirmatory factor analysis.
[a] Model 3 when applied to data from Wave 1 resulted in an improper solution that was not considered further.
$^{*}p < .05$. $^{**}p < .001$.

comparison of increasingly restrictive models that do require different parameter estimates (e.g., factor loadings) to be the same across the different waves.

Unconstrained models. For these longitudinal analyses (see Table 3), there was an improvement in RMSEA for Model 1 (.039), although TLI remained poor (.838) and fit indices were generally much worse than in subsequent models. Fit indices were much better for Model 2 (e.g., *TLI* = .965; *RMSEA* = .018). Overall, models based on LMF factors (Models 6, 7, and 8) did better than models based on CU (Models 3, 4, and 5). Importantly, all three of the LMF models performed better than Model 2. This was the case particularly when both positive and negative aspects

Table 4
Factor Loadings and Uniquenesses for Models in Wave 1

	Model 1		Model 2			Model 3		Model 6			
Item	GSE	Uniqueness	POS	NEG	Uniqueness	GSE	Uniqueness	GSE	Pos	Neg	Uniqueness
1	.59	.66	.62		.62	.40	.84	.57	.28		.60
2	.62	.61	.66		.56	.40	.84	.61	.32		.52
3	.60	.64	.62		.62	.45	.80	.58	.22		.62
5	.54	.71	.53		.72	.54	.71	.51	*ns*		.73
7	.48	.77	.49		.76	.32	.90	.52	*ns*		.74
9	.47	.78	.47		.78	.41	.83	.60	*ns*		.55
4	.32	.90		.52	.73	.30	.91	.22		.47	.73
6	.35	.88		.59	.65	.35	.88	.24		.55	.64
8	.33	.89		.53	.72	.32	.90	.22		.49	.71
10	.41	.84		.63	.60	.42	.82	.32		.53	.62
M	.47		.56	.57		.39		.44	.27	.51	

Note. See Figure 1 for a description of the various models. In order to conserve space and facilitate presentation, results for Models 4 and 5 (submodels of Model 3) and Models 7 and 8 (submodels of Model 6) are presented in the supplemental materials. GSE = global self-esteem trait factor; POS = positive self-esteem trait factor; NEG = negative self-esteem trait factor; Pos = positive method factor; Neg = negative method factor.

were taken into account, pointing to Model 6 as the best model (e.g., $TLI = .977$; $RMSEA = .015$). Interestingly there were no admissibility problems with Model 3 for the longitudinal data (i.e., the solution was fully proper), but its fit indices (e.g., $TLI = .954$; $RMSEA = .024$) were systematically poorer than for the corresponding cross-sectional models or Model 6 based on the longitudinal data.

Models including both positive and negative effects performed better than models with only one method effect (see footnote 2). More generally, for models with CUs, fit indices of longitudinal CFAs performed worse than single-wave models. In contrast, fit indices for LMF models remained good or improved slightly compared to the corresponding single-wave models and were better than any of the other longitudinal models considered in this section. Among LMF models, Model 6 was the best. In marked contrast to results based on single waves of data, the results for the longitudinal models clearly show that Model 6 performed better than the others. These results support not only our a priori predictions but also the importance of extending traditional tests based on cross-sectional (single-wave) data to longitudinal data.

Tests of invariance: Comparison across nested models. In all subsequent tests of invariance for the eight SEM models, fit indices remained consistent when moving from the unconstrained models (configural invariance) with no invariance constraints to the ones including invariance for factor loadings and even variances of latent factors (trait and/or method latent factors according to the specific model under consideration; see Table 3). Based on fit indices—particularly those that include controls for parsimony (TLI and RMSEA)—there is good support for the invariance of factor loadings and the latent factor variances for all the models.

In relation to error variance equivalence tests, preliminary analyses did not support full invariance. Instead, we used the partial invariance strategy proposed by Marsh and Grayson (1994). More specifically, we freed the uniquenesses for Wave 1 but constrained the uniquenesses to be invariant over Waves 2–4. However, these partially invariant models still resulted in slightly poorer fits to the data, although the differences were small. We note, however, that the invariance of measurement error is not necessary in latent models but that tests based on manifest scale scores like those used in most applied studies would not be valid, as measurement error is not consistent over time.[3]

Finally, in relation to each SEM model, results were in line with what was highlighted in the previous sections. Model 1, not including method effects, performed poorly; Model 2 performed reasonably well, although it was impossible to disentangle from this model the unique contributions of trait and method factors. For structures based on CUs (Models 3, 4, and 5; see also footnote 2), the configural models were poorer than the single waves; thus, results from more stringent invariance tests were not as good (although reasonable stability was found). Structures based on LMFs (Models 6, 7, and 8) were more adequate than the others. Model 6 in particular, including both positive and negative LMFs, provided the best results. Fit indices for Models 6 and 8 were not substantially different, but a $\Delta\chi^2$ test showed that the inclusion of the negative latent factor was significant, $\Delta\chi^2(22) = 49.22$, $p < .001$.

These results confirm once again that the simple GSE Model 1 is too simplistic. Furthermore, although results for the bidimensional structure are apparently not bad, the outcomes from Models 3–8 are better and show that method effects do exist. Results also suggest that considering both positive- and negative-item method effects is a good strategy and that CU and LMF strategies are both reasonable. However, from the longitudinal perspective considered here—in contrast to tests based on each wave considered separately—the LMF approach is clearly preferable.

Stability of latent means over time. Tests of latent mean invariance are particularly useful when longitudinal data are considered. The main issues to consider are invariance of factor

[3] In relation to models based on CU (Models 3, 4, and 5), we also tested invariance for the patterns of correlated uniquenesses. Models with invariance for CU patterns seemed to perform a little bit better than those not including them. In tests of the invariance of CUs over time in Model 3 that included CUs for both positive and negative items, the χ^2 difference was not significant, $\Delta\chi^2(63) = 80.84$; $TLI = .937$; $RMSEA = .024$.

loadings and invariance of intercepts for each item. Only if the meaning of the item and its relation to the latent construct remain stable over time is it appropriate to interpret mean changes as changes in the latent construct—for example, increments or decrements in the mean levels of the latent construct (Marsh & Grayson, 1994). The typical procedure to compare latent means is to fix factor loadings and item intercepts to be equal across time: to fix the latent mean to be zero in the first wave and to free latent means in all other waves. In our case, this would imply that latent means in Waves 2–4 are scaled in relation to Wave 1.

Tests of invariance did not result in major decrements in fit indices (see Table 5) for any of the eight models. Hence, we can confirm that the properties of the RSE responses do not change over time. Once again, according to fit indices the best model is Model 6 ($TLI = .967$; $RMSEA = .017$). In relation to trait factors, it is notable that there is an increase in trait self-esteem when moving from Wave 1 to Wave 4 (see Table 6). Although this general trend is evident in each of the models, the size of these increases varies somewhat for different models and sometimes is not strictly monotonic. Comparing the different models, trait self-esteem increases if method effects (CU or LMF) are included in the model, and it increases more when both positive and negative wording effects are considered (e.g., Models 3 and 6 compared to Model 1).

In LMF models (see Table 6), the means of the LMF factors increase over time when only one method effect is included (Models 7 and 8), but not for Model 6 which included LMFs for both positively and negatively worded items. In Model 6, the means of the LMFs do not vary over time. These results further suggest that considering only one method effect might result in models in which method and trait variance are confounded. Finally, and importantly for the purposes of our investigation, the fact that the means of the LMFs for Model 6 are nonsignificant suggests that these latent means do not increase over time but, instead, remain stable. Thus, this result does not support the artifact hypothesis but is consistent with the hypothesis of stable response styles.

These findings are substantively important, emphasizing that biased results are likely to result in misinterpretations and invalid conclusions when method effects are not taken into account, but that sophisticated statistical models are needed to do so.

Stability of latent method-effect factors over time: Disentangling the nature of wording effects in Model 6. In the present section we examine the stability of method effects across time for Model 6. Results thus far have demonstrated that Model 6 is preferable to the other factor structures. Although single-wave analyses based on fit indices and parameter estimates were not definitive, outcomes from the longitudinal approach (longitudinal CFA, invariance tests) clearly supported Model 6. Moreover, Model 6 is based on LMFs. Unlike models based on CUs, Model 6 provides a test of the stability of method factors that is key to understanding the nature of wording effects (e.g., Bentler et al., 1971).

In Model 6 it is also possible to constrain the correlations between method effects over time to be zero (consistent with ephemeral method effects and assumptions implicit in the CU approach) or to allow them to be freely estimated (consistent with the stable, response-style explanation). Therefore, in relation to a longitudinal CFA for Model 6, we constrained stability coefficients for method-effect factors to be zero (Model 6.5–0 in Table 3) and contrasted this model with a model in which these correlations are freely estimated (Model 6.5a in Table 3). Fixing all cross-wave correlations to zero for the same latent method factor resulted in poor fit indices (e.g., $\Delta TLI = .042$) and a significant χ^2 difference statistic, $\Delta\chi^2(12) = 406.83$, $p < .001$, in comparison to those discussed above for the unconstrained model (Model 6.5a). These results argue against the ephemeral artifact interpretation of the method effects. In contrast, the substantial test–retest correlations for method factors (see Table 7) provide support for the stability of method effects. Indeed, the range of correlations for the positive method factor across waves was from .43–.60 and for the negative method factor was from .39–.65.

Discussion

The factor structure of self-esteem and the longitudinal approach. In Study 2 we used a longitudinal perspective to examine different hypotheses of the RSE factor structure. There is an active, ongoing debate regarding the latent structure of GSE based on responses to the RSE and related instruments, and the nature of associated item-wording method effects. We based our study on four theoretical perspectives on self-esteem (see earlier discussion) applied to single and multiple waves of data. However, in line with our predictions, results based on single waves were inconclusive, but results from the longitudinal approach clearly supported Model 6, which included both positive and negative LMFs, as the best model.

Table 5
Mean Invariance Tests: Fit Indices for the Eight Structural Equation Models

Model	χ^2	*df*	*cf*	*TLI*	*CFI*	*RMSEA*
Model 1	3,174.972	728	1.193	.837	.848	.039
Model 2	1,362.072	700	1.188	.954	.959	.021
Model 3	1,569.945	644	1.169	.930	.942	.025
Model 4	1,993.571	704	1.182	.911	.920	.029
Model 5	2,025.244	668	1.177	.902	.916	.030
Model 6	1,160.112	694	1.176	.967	.971	.017
Model 7	1,426.148	712	1.186	.951	.956	.021
Model 8	1,261.071	710	1.187	.962	.966	.019

Note. See Figure 1 for a description of the various models. χ^2 = chi-square test statistic; *df* = degrees of freedom; *cf* = maximum likelihood ratio (MLR) correction factor; *TLI* = Tucker–Lewis index; *CFI* = comparative fit index; *RMSEA* = root-mean-square error of approximation.

Table 6
Latent Mean Values for the Eight Structural Equation Models

	Wave 2					Wave 3					Wave 4				
Model	GSE	POS	NEG	Pos	Neg	GSE	POS	NEG	Pos	Neg	GSE	POS	NEG	Pos	Neg
Model 1	$.17^{**}$					$.30^{**}$					$.34^{**}$				
Model 2		$.15^{**}$	$.16^{**}$				$.29^{**}$	$.21^{**}$				$.39^{**}$	$.26^{**}$		
Model 3	$.22^{**}$					$.33^{**}$					$.39^{**}$				
Model 4	$.16^{**}$					$.30^{**}$					$.34^{**}$				
Model 5	$.19^{**}$					$.26^{**}$					$.31^{**}$				
Model 6	$.32^{*}$			$-.14$	$-.09$	$.28^{\dagger}$			.12	.01	$.45^{*}$			$-.05$	$-.10$
Model 7	$.15^{**}$				$.10^{*}$	$.29^{**}$				$.08^{\dagger}$	$.33^{**}$				$.12^{*}$
Model 8	$.17^{**}$			$.07^{*}$		$.21^{**}$			$.24^{**}$		$.27^{**}$			$.24^{**}$	

Note. See Figure 1 for a description of the various models. GSE = global self-esteem trait factor; POS = positive self-esteem trait factor; NEG = negative self-esteem trait factor; Pos = positive method factor; Neg = negative method factor.
† $p < .06$. * $p < .05$. ** $p < .001$.

The use of a longitudinal perspective was particularly relevant in this investigation, because it provided clear and stable results in relation to the best model. The longitudinal approach allowed us to perform tests of invariance over time that were a fundamental prerequisite of studying the stability of method effects over time and, thus, to understanding the nature of wording effects associated with the present measure of self-esteem. Moreover, the longitudinal approach used here has broad applicability to the study of other psychological constructs inferred on the basis of positively and negatively worded items and, more generally, a wide variety of method effects that might be idiosyncratic to particular measures.

Importance of considering the full set of models with both wording effects. An important limitation of previous research is that most studies have not directly compared models including only one method effect (associated with positively *or* negatively worded items) with models including both method effects (associated with positively *and* negatively worded items). In the application of CFA models, it is common to juxtapose a set of fully or partially nested models. Important examples of this type of taxonomic strategy include the set of CFA models used to evaluate MTMM models (which were one basis of the models considered here) and the set of CFA models used to evaluate factorial and measurement invariance. In each case, the comparison of the different models is much more important than the evaluation of any one model. In this respect, the juxtaposition of results from the entire set of eight models (see Figure 1) is an apparently important contribution to RSE research and related studies of method effects. Whereas all the proposed models have been considered before, previous RSE research has not considered a taxonomic approach based on the comparison of results from such a comprehensive set of models. Indeed, the juxtaposition of models positing no method effects, positive-item method effects only, negative-item method effects only, and both positive- and negative-item method effects is particularly important in understanding the nature and relative sizes of these method effects. We extended this taxonomic strategy by taking a longitudinal perspective. In fact, we found that the inclusion of method effects resulted in increasingly stable test–retest GSE correlations over time (for both CU and LMF models), particularly when both wording effects were considered. Moreover, for LMF models it emerged that when only one method effect is considered (e.g., Model 7 and Model 8), LMFs seem to be confounded with latent trait factors. Therefore, our results support the need to take into account both wording effects.

Invariance tests as a prerequisite for stability. One of the aims of our investigation was to disentangle the nature of wording effects as methodological artifacts or response-style factors. The main difference between these two hypotheses is stability over time. Methodological artifacts are typically posited to be inherently ephemeral and unstable over time, whereas response-style effects are typically posited to be stable over time (e.g., Bentler et al., 1971). Although some attempts to measure stability of method effects have been pursued, apparently only Marsh and Grayson (1994) considered the fundamental precondition of testing the stability of means over time—the invariance of factor loadings and item intercepts. Unfortunately, they considered only CU models that were apparently weaker and more limited than the LMF models considered here. Our new results confirm the usefulness of Marsh and Grayson's approach to studying the stability of structures and latent means.

Stability and nature of method effects. The present investigation is apparently the first to evaluate measurement invariance as a prerequisite to investigating the stability of response-style wording effects in self-esteem. These results showed that the meaning of the items—and thus the constructs based upon them—remained consistent over time and that the responses were reasonably reliable. In particular, factor loadings and intercepts were reasonably invariant over time, providing a basis for the examination of mean stability for all the models. Fit indices for Model 6 were the best for all the models considered. Means for LMFs in Model 6 did not change significantly over time, demonstrating the mean stability of method effects. Because stability of method factors was a crucial distinction between the two perspectives on the nature of method effects (artifacts or response styles), we adopted a longitudinal approach to test stability over time of the LMFs in Model 6. The results showed that LMFs in Model 6 had substantial test–retest stability. This finding is critical in providing support for the response-style hypothesis and undermining the ephemeral artifact hypothesis.

Limitations and future research. Despite the richness of the YIT database, it is based on responses by only adolescent males. Whereas gender differences have been found in GSE responses (e.g., Kling, Hyde, Showers, & Buswell, 1999), further research is needed to evaluate whether gender affects method factors.

As our study is based on longitudinal data for participants in middle adolescence to early adulthood, there is some question as to the generalizability of our results to other ages. The issue of age and cognitive development is a potentially important one—particularly in

Table 7
Correlations for Latent Factors Across Waves for Selected Factors

Variable	GSE1	GSE2	GSE3	GSE4	POS1	POS2	POS3	POS4	NEG1	NEG2	NEG3	NEG4
Model 1												
GSE1	—											
GSE2	.65	—										
GSE3	.59	.76	—									
GSE4	.51	.66	.76	—								
Model 2												
POS1					—							
POS2					.62	—						
POS3					.56	.73	—					
POS4					.49	.66	.73	—				
NEG1					.44	.34	.39	.37	—			
NEG2					.34	.49	.47	.42	.59	—		
NEG3					.33	.44	.58	.51	.52	.66	—	
NEG4					.22	.29	.43	.58	.46	.66	.70	—
Model 3												
GSE1	—											
GSE2	.89	—										
GSE3	.78	.95	—									
GSE4	.67	.82	.90	—								
	GSE1	GSE2	GSE3	GSE4	Pos1	Pos2	Pos3	Pos4	Neg1	Neg2	Neg3	Neg4
Model 6												
GSE1	—											
GSE2	.71	—										
GSE3	.64	.82	—									
GSE4	.55	.68	.80	—								
Pos1					—							
Pos2					.52	—						
Pos3					.48	.60	—					
Pos4					.43	.62	.60	—				
Neg1									—			
Neg2									.49	—		
Neg3									.39	.49	—	
Neg4									.39	.65	.60	—

Note. See Figure 1 for a description of the various models. In order to conserve space and facilitate presentation, results for Models 4 and 5 (submodels of Model 3) and Models 7 and 8 (submodels of Model 6) are presented in the supplemental materials. GSE = global self-esteem trait factor; POS = positive self-esteem trait factor; NEG = negative self-esteem trait factor; Pos = positive method factor; Neg = negative method factor.

relation to method effects associated with responses to negatively worded items for children, as already demonstrated by Marsh (1986), but it also might be relevant for early adolescence. Marsh found huge shifts with age (for children 7–11 years of age) in the ability to respond appropriately to negatively worded items. Consistent with a cognitive development model, he found that positively and negatively worded items designed to measure the same self-concept constructs were almost uncorrelated for the youngest children but were substantially correlated (*r*s of about .6) for the oldest children in this age range. Furthermore, even within each school-year group, children with better verbal abilities were better able to handle the negatively worded items. Although we suspect that this problem would generalize to responses by young children to the RSE, we note that the RSE is typically used for adolescents and adults but is rarely used with young children. Nevertheless, there remains a gap in our knowledge about age-related method effects for early adolescents. Therefore, further research is needed to fully evaluate the generalizability of our results in light of these limitations.

The results provide a clear demonstration that method effects were stable over time and not fleeting, in contrast to many accounts of the method effects associated with the RSE—including some of our own work (e.g., Marsh, 1996). The existence of method effects necessarily detracts from the construct validity of interpretations of the RSE that do not control for them. For us, this was the main focus of our study. Although beyond the scope of the present investigation, it is relevant for further research to use these approaches to further evaluate the meaning of stable response variables—the psychological processes and individual difference characteristics that are associated with them, how they are related to other self-report and non-self-report outcomes, and how generalizable they are across different constructs. Some researchers have found associations between the negative LMF and some personality characteristics (e.g., fear of evaluation and self-consciousness, DiStefano & Motl, 2006; avoidance motivation, consciousness, and emotional stability, Quilty et al., 2006). It is important to emphasize, however, that applied researchers need not fully understand the meaning of method effects to control them, and that failure to control them will bias the interpretations of RSE responses—whether or not their meaning is understood. Nevertheless, models developed here might provide a useful starting point for further such research into the meaning of method effects associated with stable response styles.

It would also be interesting to examine the relation between different structural models of self-esteem (trait and method factors) and implicit measures of self-esteem (e.g., Greenwald & Banaji, 1995). Implicit measures of several important constructs have been developed in the social sciences (Fazio & Olson, 2003), and it is hypothesized that these measures avoid response biases. To the best of our knowledge, research conducted until now has considered only scale scores of the RSE to compare explicit and implicit measures of self-esteem (e.g., Fazio & Olson, 2003; Greenwald et al., 2002; Zeigler-Hill, 2006), without taking into account the wording method effects associated with explicit measures of self-esteem.

Overall Conclusions and Implications for Applied Research

The present investigation contributes to the debate on the nature of wording effects associated with self-esteem instruments. We suggested that apparently conflicting results from previous literature might be due to an overreliance on single waves of data and structural models that were not robust, as shown in our Study 1 based on simulated data. We began by reviewing literature in which there was little agreement about the appropriate interpretation of responses to GSE based on the RSE—one of the most widely used instruments in the history of psychology. The only clear evidence was that the model implicitly used as the basis of almost all applied research (our Model 1) was clearly inappropriate, calling into question the vast literature based on the RSE. Previous research was clearly ambiguous as to whether responses to the RSE and related instruments should be interpreted as two trait factors or as one GSE trait factor and method effects. There was ambiguity as to whether there should be method effects for positively worded items, negatively worded items, or both. There was ambiguity as to whether method effects should be represented as CUs or LMFs. There was ambiguity as to whether method effects were ephemeral artifacts or stable response-style tendencies. These ambiguities, we speculated, were due to two limitations; many studies considered only a few of the taxonomy of models considered here (and the underlying assumptions upon which they are based), and most studies considered only a single wave of data that precluded evaluations of stability over time. In the present investigation, even when we considered the entire taxonomy of models, the ambiguities about the most appropriate model remained when we considered only single waves of data. However, when we took a longitudinal perspective, there was clear support for one model as most appropriate (Model 6) and clear evidence that item-wording effects reflect stable response-style effects rather than ephemeral method artifacts. In this respect, the longitudinal approach in combination with the taxonomy of models advocated here has apparently resolved the ambiguities associated with the most appropriate model to represent RSE responses and the appropriate interpretation of item-wording effects.

Consistent with our emphasis on methodological–substantive synergy, we argue that psychological assessment researchers should be familiar with a range of relevant quantitative methodologies so that they are able to apply the most appropriate methodologies to evaluate complex substantive issues. In this respect, the present investigation provides a case study for the importance of methodological–substantive synergy and multimethod research to better understand psychological assessment. However, there are costs to this approach in terms of suitable quantitative skills, the complexity of the analysis, and data collection. Our study, for example, was based on a large longitudinal database that is not always available to applied researchers. What are the implications of this for applied researchers?

First, it is important to emphasize that the application of our preferred model (Model 6) does not require longitudinal data. We had to use longitudinal data to demonstrate that Model 6 was superior, but now that we have shown this to be the case applied researchers can use Model 6 on this basis even if they do not have longitudinal data. Whereas the extension of Model 6 to include longitudinal data was substantively and methodologically important in showing that the method effects were stable over time, it is possible to control for method effects with a single wave of data. Failure to control method effects will result in biased interpretations. This is clearly evident from the rejection of Model 1, which is the implicit basis of most applied RSE research, as well as from the substantively important sizes of method effects based on parameter estimates (factor loadings on method factors in Models 6–8 and CUs in Models 3–5).

The second issue, relating to sample size, is more problematic. The only unambiguous conclusions that can be drawn about sample size are that more is better and that more is never too much (e.g., Marsh, Hau, Balla, & Grayson, 1998). Historically, there have been many ad hoc guidelines about sample size that relate to the number of factors, the number of items, and the number of estimated parameters. However, subsequent research has shown that the logic upon which these are based is often flawed and that empirical support for them is generally lacking (e.g., Marsh et al., 1998, but see also Gagné & Hancock, 2006). However, even when the sample size is less than desirable (e.g., less than 200) there are strategies such as imposing equality constraints that may result in a well-defined solution. Nevertheless, the smaller the sample size, the more problematic the application of CFA/SEM models is. Ultimately, however, this issue is part of the concern of methodological–substantive synergies that we have emphasized here. The resolution of complex substantive issues often requires the application of strong methodological approaches. If the strongest methodological approaches are not applied—because of small sample sizes or whatever other reason—then the substantive interpretations of the results are likely to be compromised—as illustrated in the present investigation.

The results of the present investigation are clearly relevant to the study of GSE responses to the RSE and related instruments. However, the issues considered, the methodological–substantive synergy perspective, and the taxonomic approach demonstrated here have broad generalizability to all studies that evaluate any psychological construct on the basis of self-report measures containing a mixture of positively and negatively worded items. We speculate that most psychological constructs based on responses to positively and negatively worded items—if appropriately evaluated using procedures in the present investigation at the level of the individual item—would fail to support a simple unidimensional model like our Model 1. In applied research most researchers implicitly assume Model 1 without explicitly testing it in relation to other models considered here. However, in some instances researchers assume that factors based on positively and negatively worded items measure substantively distinct factors (like our Model 2) without systematically evaluating this supposition in relation to alternative models like those considered here. Even if interpretations consistent with our Models 1 or 2 are appropriate, it is incumbent upon the developers, advocates, and users of such measures to defend the interpretation of their measures in relation to interpretations of competing models like those considered here. In summary, the approach used here should become part of the standard arsenal of tools that psychological assessment researchers routinely use to evaluate the construct validity of their assessment instruments—particularly self-report instruments based on a mixture of positively and negatively worded items.

References

Aluja, A., Rolland, J., García, L. F., & Rossier, J. (2007). Dimensionality of the Rosenberg Self-Esteem Scale and its relationships with the three- and the five-factor personality models. *Journal of Personality Assessment, 88,* 246–249.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York, NY: Macmillan.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bachman, J. G. (2002). *Volume I of the documentation manual.* Ann Arbor, MI: Interuniversity Consortium for Political and Social Research.

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frog pond revisited (again). *Journal of Personality and Social Psychology, 50,* 35–46.

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality, 27,* 49–87.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76,* 186–204.

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

Blascovich, J., & Tomaka, J. (1991). The Self-Esteem Scale. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes, Vol. I* (pp. 115–160). San Diego, CA: Academic Press.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling, 11,* 272–300.

Byrne, B. M., & Goffin, R. D. (1993). Modeling MTMM data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research, 28,* 67–96.

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

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research, 36,* 462–494.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

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

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.

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.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8,* 38–60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait–multimethod data: Different models for different types of methods. *Psychological Methods, 13,* 230–253.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology, 54,* 297–327.

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41,* 65–83.

Gana, K., Alaphilippe, D., & Bailly, N. (2005). Factorial structure of the French version of the Rosenberg Self-Esteem Scale among the elderly. *International Journal of Testing, 5,* 169–176.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23,* 443–451.

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.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellot, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3–25.

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.

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

Jackson, D. N., & Messick, S. (1962). Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition* (pp. 129–155). New York, NY: Wiley.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal investigations. In J. R. Nesselroade & B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York, NY: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1996–2001). *LISREL 8: User's reference guide.* Lincolnwood, IL: Scientific Software International.

Kaplan, H. B., & Pokorny, M. D. (1969). Self-derogation and psychological adjustment. *The Journal of Nervous and Mental Disease, 149,* 421–434.

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: U.S. Department of Education.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–

multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125,* 470–500.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait–correlated method and correlated uniqueness models of multitrait–multimethod data. *Psychological Methods, 7,* 228–244.

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling, 14,* 581–610.

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive–developmental phenomenon. *Developmental Psychology, 22,* 37–49.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

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

Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 774–798). Hoboken, NJ: Wiley.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analysis of multitrait–multimethod data: A comparison of the behavior of alternative models. *Applied Psychological Measurement, 15,* 47–70.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Mahwah, NJ: Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 102,* 391–410.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1,* 317–359.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait–multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and application* (pp. 177–198). Thousand Oaks, CA: Sage.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390.

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological–substantive synergies. *Contemporary Educational Psychology, 32,* 151–171.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16,* 439–476.

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.

Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Owens, T. J. (1993). Accentuate the positive—and the negative: Rethinking the use of self-esteem, self-deprecation, and self-confidence. *Social Psychology Quarterly, 56,* 288–299.

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.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson & P. R. Shaver (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Prezza, M., Trombaccia, F. R., & Armento, L. (1997). La Scala dell'Autostima di Rosenberg: Traduzione e validazione Italiana [The Rosenberg Self-Esteem Scale: Italian translation and validation]. *Bollettino di Psicologia Applicata, 223,* 35–44.

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

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

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,* 623–642.

Shahani, C., Dipboye, R. L., & Philips, A. P. (1990). Global self-esteem as a correlate of work-related attitudes: A question of dimensionality. *Journal of Personality Assessment, 54,* 276–288.

Tafarodi, R. W., & Milne, A. B. (2002). Decomposing global self-esteem. *Journal of Personality, 70,* 443–483.

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

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

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg's Self-Esteem Scale among crack-cocaine drug users. *Structural Equation Modeling, 8,* 275–286.

Whiteside-Mansell, L., & Corwyn, R. F. (2003). Mean and covariance structures analyses: An examination of the Rosenberg Self-Esteem Scale among adolescents and adults. *Educational and Psychological Measurement, 63,* 163–173.

Zeigler-Hill, V. (2006). Discrepancies between implicit and explicit self-esteem: Implications for narcissism and self-esteem instability. *Journal of Personality, 74,* 119–144.

Received May 7, 2009
Revision received January 7, 2010
Accepted January 10, 2010 ■