Title

On the Factor Structure of the Rosenberg (1965) General Self-esteem Scale

Abstract

Since its introduction, the Rosenberg General Self-esteem scale (RGSE, Rosenberg, 1965) has been one of the most widely used measures of global self-esteem. We conducted four studies to investigate (a) the goodness of fit of a bifactor model positing a general self-esteem factor (GSE) and two specific factors grouping positive (MFP) and negative items (MFN), and (b) different kinds of validity of the GSE, MFN and MFP factors of the RSGE. In the first study (*n*= 11,028), the fit of the bifactor model was compared with those of nine alternative models proposed in literature for the RGSE. In Study 2 (*n* = 357), the external validities of GSE, MFP and MFN were evaluated using objective grade point average data and multi-method measures of prosociality, aggression, and depression. In Study 3 (*n* = 565), the across-rater robustness of the bifactor model was evaluated. In Study 4, measurement invariance of the RGSE was further supported across samples in three European countries, Serbia (*n*= 1,010), Poland (*n*= 699), and Italy (*n*= 707), and in the United States (*n* = 1,192). All in all, psychometric findings corroborate the value and the robustness of the bifactor structure and its substantive interpretation.

Keywords. Bifactor model; self-esteem; Rosenberg self-esteem scale; method effects; method factors.

Self-esteem reflects an overall subjective evaluation of personal worth (Marsh & O'Mara, 2008; Rosenberg, 1965). A considerable amount of research has investigated the nature of this construct(Baumeister, Campbell, Krueger, &Vohs, 2003), which represents one of the most popular individual differences constructs in psychology (see Donnellan, Trzesniewski, & Robins, 2011, for a review). Self-esteem, similar to any other psychological construct, is a latent variable that is not directly observable. Yet individuals' standing on the latent attribute can be inferred through their answers to statements intended to describe internal positive and negative states, such as feelings and emotions about the self (Borsboom, Mellenbergh& van Heerden, 2003).

Since its introduction, the Rosenberg General Self-esteem scale (RGSE; Rosenberg, 1965) has been one of the most popular and widely used measures of global self-esteem (Blascovich &Tomaka, 1991; Donnellan et al., 2011; Schmitt &Allik, 2005).According to *PsycInfo*, the instrument has been cited 3,016 times during the last five years (2010-2014).The scale assesses the "feeling that one is good enough" (Rosenberg, 1965, p. 31), and consists of 10 items with a high degree of face validity. A large body of empirical evidence supports the internal consistency of the instrument (Byrne, 1983), its predictive validity (Kaplan, 1980), and its equivalence over time (Marsh, Scalas, &Nagengast, 2010; Motl & DiStefano, 2002). The popularity of the 10-item RGSE has been due in part to its long history of use, its uncomplicated language, and its brevity (it takes only 1 or 2 minutes to be completed).

In addition to its privileged place in the literature, the RGSE offers other potential advantages. For example, it was developed in accordance with the recommended strategy of building instruments with a balanced number of positively and negatively worded items (Paulhus, 1991). This approach helps to address acquiescence response bias (Marsh, 1996). One perhaps unexpected drawback of this otherwise desirable property is that balanced scales can introduce complexities with regard to the dimensionality of the measure. Thus, it is not surprising that over the years, several authors have championed different structural models for the RGSE involving multiple factors. Recent studies have demonstrated that the deviation from the unidimensionality observed for the RGSE is mostly due to the effect of items' wordings (Marsh et al., 2010; Tomás &

Oliver, 1999). The upshot of this psychometrically-oriented debate is a lack of consensus on how the observed factors should be conceived in substantive terms (Tafarodi& Milne, 2002).

The Dimensionality of the Rosenberg (1965) Self-esteem Scale

Several researchers have acknowledged the need to consider one or more method factors, along with general self-esteem (GSE), in testing the dimensionality of the RSGE (Kuster & Orth, 2013; Orth, Robins, &Widaman, 2011). In the vast majority of studies, factors associated with positively and negatively worded items have been considered as methodological artifactsthat should be controlled to obtain a satisfactory model fit, and thus authors did not indulge in further speculations on the foundations of these factors (Kuster & Orth, 2013; Orth et al.,2011).

The present research was designed to provide additional insight into the nature of the latent factors associated with positively and negatively worded items of the RGSE. The starting point was the notion that the label "method effect" is inadequate to depict these factors. Method effects refer to defects in method for assessing constructs (Fiske, 1987). Typical method effects are the inflated correlations between unrelated traits due to the use of the same informant (e.g. self-report; see Kenny &Kashy, 1992). These kinds of measurement artifacts are unrelated to substantive construct variance. However, the so-called method effect factors for the RGSE have shown psychometric properties similar to the substantive GSE factor (see Motl & DiStefano, 2002).In this regard, a number of investigators who have analyzed method effects associated with the RGSE have been able to demonstrate: (a) the convergent validity across instruments for the factors viewed as reflecting method effects (Horan, DiStefano, &Motl, 2003); (b) their long-term stability (Motl& DiStefano, 2002; Saada, Bailly, Joulain, Hervé, & Alaphilippe, 2013); and (c) their criterion-related validity (Quilty, Oakman, &Risko, 2006). Other researchers using other instruments have further demonstrated that method factors are stable across observers (Alessandri, Vecchione, Tisak & Barbaranelli, 2011)and moderately heritable (Alessandri et al., 2009).

A Theoretical Interpretation

Alessandri, Vecchione, Donnellan and Tisak (2013) recently offered an overarching interpretation of the three factors assessed by the ten items of the RGSE in terms of a bifactor

model(Chen, West & Sousa, 2006; Reise, Morizot, Hays, 2007). Bifactor models can be considered when (a) there is a general factor accounting for the commonality of the items; (b) there are multiple domain specific factors; and (c) both the common factor and the domain specific factors are interesting for researchers (Chen et al., 2006). In the case of RGSE, the proposed bifactor model is comprised by a general self-esteem factor plus two substantive specific factors. Alessandri et al. (2013) suggested that the method factor associated with negatively worded items (MFN) shares similar characteristic as the self-derogation factor described by Kaplan and Pokorny (1969), who interpreted it as, "the expression of intense negative affect towards general self-conception" (p. 425). This interpretation has been repeatedly put forth by several other authors (e.g., Epstein, Griffin &Botvin, 2004; Kaplan, 1978; Kaplan, Martin & Robbins,1982), who referred to the negative items of the RGSE as "self-derogation" and found that these items predict adolescents' drug use, aggression, and violence, and are associated with perceived low levels of self-efficacy. The method factor associated with positively worded items (MFP), by contrast, seems to capture self-competence, an aspect of self-evaluation linked with the individual's appraisal of his or her own abilities (Diggory, 1966; Gecas, 1971). This interpretation is in line with the observation made by Tafarodi and Swann (1995) that, "high self-competence has an intrinsically positive and evaluative character" and "is cognitively characterized by the presence of a generalized expectancy for success" (p. 325).

An important caveat in embracing such an interpretation is that the work of Alessandri et al. (2013) addressed a daily version of the RGSE. In their measure, items and instructions were modified so that participants were instructed to give the response that best reflected how they felt at the moment they completed the measure. Thus, it is not entirely clear if their results translate to the general version of RGSE. Instead, an attractive feature of this interpretation is that it casts the two MFN and MFP specific factors in a manner consistent with previous empirical research that obtained self-derogation and self-enhancement measures from the RGSE item pool. However, if MFP and MFN assessed with the bifactor model represent measures of self-derogation and self-competence, they should be considered "purified" versions of those used in the past that were

obtained by merely summing either the positive items and the negative items.This procedure leaves room for contamination of the scale scores used in previous studies because variance due to GSE is not partialled-out of scales for MFP and MFN. This problem is avoided by the bifactor approach.

The Present Research

We conducted four studies with the aim to address several important psychometric issues pertaining to the use of the RGSE. The first study was designed to provide empirical support for the bifactor model described above. The goodness fit of this model was evaluated and comparedwith those of nine alternative models. These models were already identified and tested in previous studies (see, among others, Marsh, 1996, Marsh et al., 2010, Motl & DiStefano, 2002, and Tomás& Oliver, 1999). However, most researchers considered only a subset of models at a time (see Motl & Di Stefano, 2002), or used slightly different (Marsh et al., 2010) or shortened (i.e., Marsh et al. 1996) versions of the RGSE. Moreover, the appropriateness of the model proposed by Tafarodi and colleagues (Tafarodi & Milne, 2002; Tafarodi& Swann, 1995) was rarely investigated empirically (for exceptions, see Alessandri et al., 2013, and Marsh et al., 2010). Thus it seemed appropriate to compare all alternative models addressed in the literature, by using a standard version of the RGSE and a large representative sample (Study 1). The ten competing models are presented in detail in the first study. Subsequent studies were designed to corroborate the psychometric properties of the hypothesized model, along with the convergent and discriminant validity of the three factors (Study 2), their convergent validity across raters (Study 3),and their equivalence across cultures (Study 4). Findings from these studies were expected to support the aforementioned theoretical arguments regarding a bifactorial structure for the RGSE.

Study 1

This study was designed to compare the fit of the hypothesized model with that of nine alternative models presented in the literature for the RGSE. The ten competing models are presented in Figure 1. The first model is the one-factor model, which has been found to be inadequate in most prior works (see Byrne &Shavelson, 1986; Shevlin, Bunting, & Lewis, 1995, for exceptions). Model 2 posits two correlated factors, capturing positively and negatively worded

items, respectively. The emergence of these two factors has been interpreted as resulting from a methodological artifact (Carmines & Zeller, 1979), or as reflecting a substantive distinction between positive and negative self-esteem (Kaplan &Pokorny, 1969; Openshaw, Thomas, & Rollins, 1981; Owens, 1994). Model 3 posits two factors based on the distinction between transient (corresponding to a rather unstable evaluation of self-.esteem) and general (corresponding to general self-esteem) evaluations introduced by Kaufmann, Rasinski, Lee and West (1991). Model 4 represents the two components that were described by Kaplan and Pokorny (1969) as *defence of individual self-worth*, or *defensive self-enhancement*, and *self-derogation*. Models 5 through 9 include models that posit a GSE factor and use several methods to control for method effects due to item wording (Marsh et al., 2010; Tomás& Oliver, 1999). Following Marsh (1996), method effects are defined herein as the variance linked to the nature of the measurement procedures, namely, the use of negatively and positively worded items to assess general self-esteem. These variance components can be modelled either (1) as correlations among equally worded items uniqueness, or (2) as additional factors influencing items sharing the same wording. Accordingly, Model 5 posits correlated uniqueness among the residual variances of negatively worded items, whereas Model 6 posits correlated uniqueness among the residual variances of positively worded items[1]. Model 7 posits a trait factor representing self-esteem plus a method factor underlying negatively worded items. Model 8 posits a trait factor representing self-esteem plus a method factor underlying positively worded items (Tomás& Oliver, 1999). Psychometrically, Models 7 and 8 used a version of the Correlated Trait-Correlated Method Minus One (CT-C(M-1)) framework (Eid, Lischetzke, Nussbeck, &Trierweiler, 2003), where method refers to the direction of item wording. Both models contain one method factor less than the number of methods included. In Model 7, the positive wording method was chosen as the comparison standard and dropped from the model. Thus, the latent self-esteem factor represents true score variance of the positively worded items. In Model 8,

[1] We did not test a model with correlated uniqueness among both positively and negatively worded items because, as noted by Marsh (1996), this model "is not identified (this is an inherent limitation with the model and has nothing to do with the particular data being tested)" (p. 816).

by contrast, the negative wording method was chosen as the comparison standard. As a consequence, the latent self-esteem factor represents true score variance of the negatively worded items. Both models posited method factors as uncorrelated with the trait factor (i.e., self-esteem). In Model 9 (the hypothesized model), each item is explained by a trait factor (GSE) plus two method factors (i.e., MFP and MFN). This model is essentially a classical bifactor model (Chen et al., 2006; Reise, et al. 2007), which accounted for the covariation among RGSE items in terms of a broad general factor (self-esteem) reflecting the overlap across all ten items, and two correlated factors that capture item wording effects. Model 10 is the five-factor model introduced by Tafarodi and colleagues (Tafarodi & Milne, 2002; Tafarodi & Swann, 1995), which includes a general factor of self-esteem, the two factors of self-acceptance and assessment, plus MFN and MFP.

Method

Participants and Procedures

The present study included 11,028 participants (43.7% males) recruited within a cross-sectional nationwide survey conducted in Italy to examine the attitudes of children, parents, grandparents, and teachers with regard to lifelong learning. Participants were recruited by a team of researchers and agreed to complete a set of questionnaires at their homes. For their participation in the study, participants were offered feedback about their psychological profile. The age range was from 15 to 85 years, with a mean of 38.17 years (*SD*=18.95). Participants were residents in various geographic areas of the country: 28% represent northern Italy, 40% central Italy, and 32% southern Italy. They varied widely in demographic and socioeconomic backgrounds, although the sample was homogeneous in terms of ethnicity (all participants were Caucasian). Fifty-four percent were married, and 46% were either unmarried, divorced, or widows/widowers. Eleven percent were in professional or managerial ranks, 22% were merchants or operators of other businesses, 36% were skilled workers, 27% were unskilled workers, and 4% were retired. Years of education ranged from 5 to 18 years (*M*=10.32;*SD*=3.56); 34% completed only elementary or junior high school, 55% completed high school, and 11% earned a university degree.

Measures

Self-Esteem. The RGSE is comprised of 10 items (Rosenberg, 1965) that measure the extent to which participants feel they possess good qualities and have achieved personal success. This scale has been validated in Italian (

et al., 2012).Items were preceded with the following statement: "Below is a list of statements dealing with your general feelings about yourself."Each item is scored on a 4-point scale ranging from 1 “strongly disagree” to 4“strongly agree,” Alpha coefficients were not computed because they are potentially inappropriate given the proposed multidimensional structure of the RGSE (see Sijtsma, 2009). Estimates of reliability derived from the measurement model are presented in the results section.

Statistical Analysis

To investigate the fit of alternative models, Confirmatory Factor Analysis (CFA) was used, using the program M*plus*7.11 (Muthén & Muthén, 2012). Whereas there has been considerable debate in the literature concerning the use of maximum likelihood estimation (ML) with ordinally scaled variables treated as continuous (West, Finch, & Curran, 1995), different simulation studies have found ML to perform well with variables with four or more categories (Bentler & Chou, 1987), and under less than optimal analytical conditions (for example, in the presence of small samples sizes and moderate departures from normality). ML has been also the elective method suited in previous studies addressing the psychometric properties of the RGSE (e.g., Marsh, 1996; Marsh et al., 2010; Tomás& Oliver, 1999). Because the hypothesis of multivariate normality was non-tenable in the present sample (see Figure 2 and the Online Appendix A for more information about multivariate skewness and kurtosis across the four studies), we employed the Satorra-Bentler (1988) scaled chi-square statistic (SBχ^2) and standard errors, which takes into account the non-normal distribution of the data (Mplus estimator=MLM)[2]. The same software and estimation method were used in subsequent studies. Because the chi-square is highly sensitive to the size of the

[2] As a sensitivity test, we also ran some of the models using the WLS estimator. The parameter estimates were nearly identical, and thus we present results obtained using the MLM estimator.

sample,theχ^2likelihood ratio statistic was supplemented with other indices of model fit, such as the Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA) with associated 95% confidence intervals (CI). We accepted CFI values greater than .95 and RMSEA values lower than.08 (Kline, 2010). The Akaike Information Criterion (AIC, Burnham & Anderson, 1998) was used to compare the alternative non-nested models proposed for the RGSE. The lower the AIC, the better the fit of the model.

Results and Discussion

Table 1 reports goodness of fit indices for alternative models. Because Model 10 was not-identified, it was excluded from further consideration. As expected, Model 9 showed the best fit. Unstandardized and completely standardized loadings for Model are presented in Figure 2, Panel A. These loadings ranged (in a standardized metric) from .35 to .80 (*M*= .53; *SD*= .15) for the GSE, from-.14 to.59 (*M* = .37; *SD* = .19) for the MFP, and from .30 to .69 (*M* = .46; *SD* = .18) for the MFN. The positively worded item (henceforth labeled PO)number 7 (PO7),as well as PO6, did not loaded significantly on MFP(all other items were significant).Of interest, MFP and MFN were not significantly correlated (-.05, p = .21). Setting this correlation to zero did not degrade model fit (SB$\Delta\chi^2_{(1)}$ = .74, p < .39).

Deconstructing RGSE Variance and Estimating Overall Scale Reliability

In order to clarify the relative weight of each factor in explaining the covariance among the RGSE items, unstandardized estimates were used to perform a formal variance decomposition at both item and scale levels. Procedures used for variance decomposition are detailed in the Online Appendix C, and a detailed presentation of variance decomposition at the item level is given in Online Appendix B, Table 1. Following the common definition of scale reliability as the ratio between the “variance of the true” score and the “total score variance” (Lord &Novick, 1956),we computed the total scale reliability for the RSGE as the sum of the variances of the latent factors of GSE, MFP, and MFN divided by the total scale variance (see Online Appendix C). This index is more appropriate than the Cronbach’s alpha coefficient for multidimensional scales(see Sijtsma 2009, and Alessandri et al., 2013). Results showed that items PO6 and PO7 were almost completely

composed by GSE variance (plus measurement error). Items NE10 and NE9 were primarily measures of MFN, whereas item PO1 mostly captured MFP variance. All other items (i.e., PO2,PO4, NE3,NE5,and NE8) presented a higher proportion of GSE variance than of MFP or MFN variance, plus measurement error. At the scale level, the RGSE functioned mainly as a measure of GSE (about 67% of reliable variance); MFP (about 18%) and MFN (about 2%) explained a lower proportion of variance. Scale reliability was quite high, being .87, above the value of .80,which is a common result when Cronbach's alpha is applied to the RGSE (Blascovich & Tomaka, 1991).

In conclusion, this study supported Model 9 as the best fitting model for the RGSE items. Results from variance decomposition indicated that, at the scale level, the variance of the RGSE items was mostly captured by GSE, MFP and measurement error. MFN explained only a small proportion of variance. We discuss these results in detail in the General Discussion section.

Study 2

This study evaluated the construct validity of the bifactorial model including GSE, MFP, and MFN in predicting different constructs, such as grade point average (GPA), prosocial behavior, aggression, and depressive symptoms levels. These constructs were selected because they have been consistently linked to self-esteem in previous studies. Specifically, high levels of self-esteem have been related to high GPA (Baumeister et al., 2003), low levels of aggressive behavior (e.g., Donnellan, Trzesniewski, Robins, Moffitt &Caspi, 2005), high levels of prosocial orientation (Bartko & Eccles, 2003), and low levels of depressive symptoms (Donnellan, et al. 2011). Indeed, different authors have theorized that self-esteem might in fact represent one of the most important determinants of GPA, aggression, prosociality and depression (for a discussion, see Donnellan et al., 2011). These variables may, therefore, be considered as appropriate outcomes to test the construct validity of the general self-esteem factor.

Of interest, Alessandri et al. (2013) recently provided proof of differential associations of daily GSE, MFP and MFN with the above factors. In this study, the daily GSE latent construct was negatively associated with depressive symptoms levels; MFP was positively related to GPA; and MFN was negatively associated to implicit self-esteem (i.e., non-conscious, automatic, over-learned

self-evaluations affectively charged, see Alessandri, et al. 2013). The present aimed to assess whether Alessandri et al.'s(2013) findings generalize to the classical version of the RGSE, introduced by Rosenberg (1965) to assess trait self-esteem.

In light of the aforementioned results, and of the large literature linking self-esteem to GPA, aggression, prosociality and depression, we expected a negative relation between depressive symptoms and GSE. Furthermore, we expected a positive relation of GSE with prosociality and a negative relation with aggression, but no relations of MFP and MFN with prosociality and aggression. Whereas literature widely supports our statements about GSE, we speculated that the portion of variance captured by MFP and MFN taps very specific sets of core self-attitudes involving sensitive emotionally loaded, but more specific, aspects of self-evaluations. MFN, in particular, seems to tap the experience of a variety of negative emotions associated to a general negative view of oneself, as implied by lower scores on implicit self-esteem (Alessandri et al., 2013). Accordingly, we hypothesized that this factor would reveal a significant positive association with levels of depressive symptoms. Finally, MFP seems to tap a subjective specific evaluation of one's own competence and abilities. Indeed, this factor has been found to significantly and positively predict GPA in previous studies (Alessandri et al., 2013). Therefore, it seemed reasonable to expect a positive and significant association between MFP and GPA. To reduce the risk of common method bias, we obtained other-ratings along with self-reports for prosociality, aggression, and depressive symptoms levels. This multi-trait-multi-method design (Campbell and Fiske 1959) is likely to increase the validity of results.

Method

Participants and Procedures

The sample was composed of 357 university students (52% females) from Italy recruited by several psychology majors as a part of a course assignment. This is a different sample than in Study 1.Participants were aged from 19 to 22 years (M = 21.01;SD =.97), and were all Caucasian. They were contacted and administered the RGSE along with other measures of interest (i.e., GPA, depression, aggression, prosociality). Each participant was required to bring one peer rater who

knew them “very well” with them at the designated time for completing the questionnaire. The peer raters (n = 357) were described by participants as friends or colleagues. They responded to two items asking: (1) how well they knew the target participant, and (2) to what degree they felt emotionally close to him or her. Possible responses ranged from 1 (not at all) to 10 (very much). The mean scores were 9.01 (SD = 1.76) for the first item and 9.22 (SD = 1.98) for the second item. This suggests that, on average, raters felt close to the target and knew them very well. Each pair (i.e., participant plus peer rater) completed the questionnaires at the same time during specially scheduled sessions. They were separated to prevent sharing of information and informed that they would not be able to view the other’s questionnaires to prevent possible data manipulation. In this study, we did not ask informants to report on self-esteem of participants, but only on the outcomes of interest.

Grade point average. Grade point average was assessed using a single question asking participants to report their actual academic grade point average.

Depression. Participants rated their depressive symptoms using the CES-D(Radloff, 1977). This 20-item scale measures the symptoms that characterize depression, such as despondency, hopelessness, loss of appetite and interest in pleasurable activities, sleep disturbance, crying bouts, loss of initiative, and self-deprecation. For each symptom, respondents rated the frequency of occurrence during the past week, using a Likert scale that ranged from 1 = “rarely or none of the time (less than 1 day)” to 4 = “most or all of the time (5-7 days)”(α = .90).The same items, worded in third person, were completed by the informants in regard to the target participant (α = .91).

Aggression. Aggression was assessed using the 29-item Buss-Perry Aggression Questionnaire (AQ; Buss & Perry, 1992). Participants ranked each statement (e.g., “If I have to resort to violence to protect my rights, I will”, “I can’t help getting into arguments when people disagree with me,” and “When frustrated, I let my irritation show”) from "extremely uncharacteristic of me (1)" to "extremely characteristic of me (5)”(α = .89). Friends/colleagues rated participants on the same items worded in the third person(α = .93).

Prosociality. Participants rated (1 = "never/almost never true"; 5 = "almost always/always true")their prosociality on a 16-item scale that assesses the degree of engagement in actions aimed at sharing, helping, taking care of others' needs, and empathizing with their feelings (Caprara, Steca, Zelli, &Capanna, 2005; α= .92). These same items, worded in third person, were completed by the informants (α=.96).

Results and Discussion

Model 9 resulted in a good data fit: SBχ^2(24) = 42.87; CFI = .986, TLI =.975, RMSEA = .045(.020,.067). Loadings (presented in full detail in Figure 2, Panel B) ranged (in a completely standardized metric) from .44 to .87 (*M*= .63; *SD* = .14) for GSE, from -.13 to .51 (*M* = .36; *SD* = .21) for MFP, and from .24 to .65 (*M*= .44; *SD* = .17) for MFN (Table 2). As in Study 1, only itemsPO6 and PO7 loaded significantly only on GSE. Only item NE10showed a primary loading on MFN. Other items loaded primarily on GSE, although the primary loadings were quite similar in size than the secondary loadings, differently from the first study. A further difference from Study 1 was that the primary loadings of itemsNE9 and PO1 were on GSE rather than on MFN and MFP, respectively. MFP and MFN were not significantly correlated. As in Study 1, fixing this correlation to zero did not degrade model fit significantly (SB$\Delta\chi^2$(1) = 3.23, *p*= .07).

Variance decomposition and scale reliability

Results from variance decomposition replicated those for Study 1 with some caveats (see Online Appendix B, Table 1). First, GSE explained more variance for itemNE9 and item NE10 than in Study 1. Second, GSE explained a consistently higher proportion of variance for all items. At the scale level, GSE explained the higher proportion of variance (about 78%), MFP about 11%, and MFN about 2%. Overall scale reliability was.91.

Empirical correlates of GSE, MFP and MFN

After having established the best fitting model, aggression, prosociality, and depression were added as latent factors loaded by self- and other reports. Correlated uniquenesses were included among the same items when reported by different informants (i.e., self and other). These correlations allow for associations between the same item assessed with different reporter that are

due to the content shared by the couple of items (Kenny & Kashy, 1992). GPA was added as an observed variable. Then, GSE, MFP, and MFN were specified as predictors of the other variables. This model (Figure 3) resulted in a good fit,SBχ^2(89) = 114.65, p>.05, CFI = .97,TLI = .96, RMSEA = .042(.01,.06). Prosocial behavior was positively and significantly predicted by GSE (.20;p <.05), but not with MFP (.06;p <.05) and MFN (.03;p <.05). Aggression was significantly and negatively predicted by GSE (-.53;p <.05), but not with MFP (-.06; p <.05)or MFN (.04;p <.05). Depression was negatively predicted by GSE (-.63;p <.05), positively predicted by MFN (.32; p <.05), and unrelated with MFP (.04; p <.05). GPA was positively predicted by MFP (.21; p <.05) and not predicted by GSE (.05; p <.05) or MFN (.07; p <.05). Summarizing, this study provided further support for Model 9 as the best fitting model and corroborated the predictive and discriminant validity of GSE, MFN and MFP factors. We discuss in more detail these results in the "General Discussion" section.

Study 3

The third study investigated the robustness of the RGSE structure across self-and other-ratings. As stated above, previous studies commonly relied on self-reports, and were therefore unable to disentangle artifactual from substantive sources of covariance among RGSE items. A test of the substantive nature of GSE, MFP, and MFN can be performed by using different methods for assessing self-esteem, such as self- and other-ratings. This procedure allows computation of a multi-method matrix (Campbell & Fiske, 1959) in which the variance common to different informants would represent construct variance and the correlations between measures of traits obtained by self- and other-ratings would reflect the substantive nature of the trait. We hypothesized that the substantive nature of GSE, MPN and MFN would be further supported by both (1) the emergence of the same factors in both self- and other-ratings, and (2) substantial inter-rater agreement (i.e., high and significant correlations between self- and other ratings of the same factors).

Method

Participants and Procedures

Participants were 565Italian adults (56% females) ranging in age from 19 to 61 years

(*M*=38.51;*SD*=10.91),and were all Caucasian. This is a different sample than in Studies 1 and 2. Years of education ranged from 8 to 18; 18% completed junior high school, 60% completed high school and 22% earned a university degree. Each participant was required to bring with them one peer rater who knew them "very well" at the designated time for completing the questionnaire. Procedures were identical to those described for Study 2: The peer raters (*n* = 565) were described by the participants as friends or colleagues. They responded to the same two items as in Study 2, designed to assess their knowledge of the target participant and their feeling of closeness to her/him (using a scale from 1 to 10). On average, informants knew the target reasonably well (*M* = 8.54; *SD* = 1.12), and felt emotionally close to them (*M* = 9.01; *SD* =1.09).

Measures

Self-esteem. Participants completed the 10 items of the RSGE, as in previous studies. The same 10 items, worded in third person, were also completed by informants.

Results and Discussion

Model 9showed a good fit for both self-ratings: SBχ^2(24) = 32.93, *p*= .13; CFI = 1.00, TLI =.990, RMSEA = .025(<.01,.045), and other- ratings:χ^2(24) = 41.67, *p*> .05; CFI = .990, TLI =.983, RMSEA = .034(.01,.052). For self-report, the pattern of loadings was similar to that found in Studies 1 and 2.In particular, completely standardized loadings (presented in full detail in Figure 2, Panel C) ranged from .31 to .90 (*M* = .56, *SD* = .17) for GSE, from -.22 to .64 (*M* = .10, *SD* = .36) for MFP, and from .39 to .70 (*M* = .52, *SD* = .13) for MFN. PO6 and PO7 were markers of GSE and, along withPO4, did not load significantly on MFP (all other loadings were significant).For other-report data, loadings showed a quite similar pattern to that found for self-reported data. In particular, completely standardized loadings (presented in full detail in Figure 2, Panel D) ranged from .38 to .61 (*M* = .41, *SD* = .09) for GSE, from -.07 to .36 (*M* = .15, *SD* = .23) for MFP, and from .26 to .50 (*M* = .37, *SD* = .11) for MFN.PO6 and PO7 were markers of GSE and did not loaded significantly on MFP(all other loadings were significant).MFP and MFN were not significantly correlated in either the self-or the other-rated versions of the RGSE.Zeroing this correlation did not degrade model fit for either the self- (SB$\Delta\chi^2_{(1)}$ = 1.88, *p*= .17) or for the other-

rated(SB$\Delta\chi^2_{(1)}$ = .03, p= .86) versions of the scale. All in all, despite minor differences, the structure of loadings for the self- and the other-version of the RGSE was similar.

Variance decomposition and scale reliability

For self-report, results from variance decomposition replicated those for Study 1 (see Online Appendix B, Table 1). At the scale level, GSE explained the higher proportion of variance (about 70%), MFP about 17%, and MFN about 2%. Scale overall reliability was.90. These results were closely replicated for other-report at the item level. The exception was a higher proportion of GSE variance with respect to MFP variance for item PO1.At the scale level, GSE explained the higher proportion of variance (about 72%), MFP about 15%, and MFN about 2%. Scale reliability was .90.

Factors convergence across methods

The presence of a similar pattern of loadings in both methods of assessment suggests that the same model was obtained for self- and other-rated data. To assess the degree of correspondence across raters, we built a single-group Correlated-Uniqueness model (CT-CU; Kenny & Kashy, 1992) that included the hypothesized structure of the RSGE for both self- and other-ratings (see Figure 4). In this model, the correlations between the latent factors of GSE, MFP and MFN reported by the self and peers were freely estimated. The degree of inter-rater agreement was investigated by looking at these correlations. Correlations between item-specific residuals in self-evaluations and the corresponding residuals in other evaluations were also estimated. This allowed us to take into account the association between the same item assessed by a different rater that is not accounted for by the convergence in the underlying latent trait, but is due, for example, to the shared content of the items. Finally, the across-rater covariance of GSE with both MFP and MFN was set to zero.

As a preliminary step, we examined the measurement invariance of the bifactor model across raters.As the difference between two scaled chi-squares for nested models is not distributed as a chi-square, the tenability of the constraints imposed for testing measurement invariance was examined with the scaled difference chi-square (Satorra & Bentler, 2001). Moreover, as the SB$\Delta\chi^2$ test has substantial power in large samples (Kline, 2010), we supplemented this statistic with the *ΔCFI.* In this regard, Cheung and Rensvold (2002) wrote that "it makes no sense to argue against

the usefulness of the chi-square and rely on various goodness-of-fit indices (GFI) to evaluate the overall model fit, and then argue for the usefulness of the chi-square instead of various GFIs to test for measurement invariance" (p. 252). On the basis of their simulation study, the authors recommended that investigators consider a difference in CFI larger than .01 as indicative of a meaningful change in model fit. Although we present both $SB\Delta\chi^2$ and ΔCFI, we based our decisions on the equivalence of the models on the latter index, in accordance with the suggestion of Cheung and Rensvold (2002).

The configural invariance model[$\chi^2(133) = 182.20$, $p > .05$, CFI = .987, TLI = .981,RMSEA = .03(.02-.04)] showed a good fit to the data. We therefore proceeded with tests of measurement invariance, by constraining factor loadings to be equal across raters (metric invariance model). In this model, we freed the variances of the self-rated factors. This model showed a good fit [$\chi^2(150) = 241.72$, $p > .05$, CFI = .975, TLI =.969, RMSEA = .03(.03-.04)], although it was substantively different from the configural model(i.e., $\Delta\chi^2(17) = 55.51$, $p < .01$;ΔCFI = -.012). Partial metric invariance was established after allowing itemsPO4 andNE10 to have different loadings for self- and other-rated MFP and MFN[$\chi^2(147) = 198.61$, $p > .05$, CFI = .986; TLI = .982, RMSEA = .03(.02-.04); $\Delta\chi^2(14) = 16.64$, $p = .28$, ΔCFI =.001)].Next, we constrained item intercepts to be equal across raters (scalar invariance model). In this model, we freed the latent means of the self-rated factors, keeping the means of the other-rated factors fixed to zero. Accordingly, the estimated means of self-rated factors can be interpreted as the difference relative to other-ratings. Moreover, because the measurement unit corresponds to the standard deviation of self-rated factors, these scores correspond to standardized mean differences (SMD_{other}).[$\chi^2(152) = 217.67$, $p > .05$, CFI = .982, TLI =.978, RMSEA = .03(.02-.04); $\Delta\chi^2_{(5)} = 20.63$, $p < .01$,ΔCFI =-.004)]. At this point, correlations suggested a high degree of convergence among observers for GSE and a moderately high convergence for both MFP and MFN (Figure 3). Of interest, we found no mean-level differences between self- and other-ratings of GSE, but significant differences for MFP (SMD_{other}.42;$p < .05$) and MFN (SMD_{other}-.20;$p < .05$). These differences suggest that individuals

tend to overestimate their own competences, while underestimating their tendency to self-derogate in comparison to an external observer.

To summarize, results from this study supported (1) the robustness of Model 9 across methods of administration, (2) a good degree of convergence between self- and other-rated GSE, MFP, and MFN, (3), the existence of mean-level differences between self- and other-ratings for MFP and MFN (but not for GSE).

Study 4

The aim of this study was to investigate the cross-cultural invariance of the bifactorial structure of the RGSE and its generalizability across four different language versions of the instrument, (i.e., English[US], Italian, Polish and Serbian) and thus across three European and a non-European country such as the US. These samples were chosen due to an established collaboration among scientists from respective countries. These cultures are deeply different in terms of language, ways of living, and cultural traditions. For example, Italians score relatively higher than Poland and U.S. on values related to *egalitarianism*(Schwartz, 2006). In turn, Polish people score higher in values related to *social embeddedness* and to *respect of the hierarchy* than people in Serbia, Italy, and U.S. (Schwartz, 2006). The culture of the U.S. is, instead, especially high in affective *autonomy* and *mastery* compared with the rest of the countries (Schwartz, 2006). Different values on the cultural level are linked to different self-construals on the individual level (Schwartz, 2006).We predicted that factor loadings and intercepts would be equivalent across the samples, which would indicate that (1) Model 9 replicates across countries, and (2) mean scores on the RGSE can be reliably compared across countries. A previous study (Schmitt &Allik, 2005) provided information regarding plausible mean level differences in GSE, MFP and MFN across these four countries. Schmitt and Allik (2005) reported higher levels of GSE, MFP and MFN, in Serbia, followed by U.S., Italy and finally by Poland. In their study, measures of GSE, MFP and MFN were computed as the sum of the items, all positively scored. Thus, despite possible differences arising from the method used to compute measures of GSE, MFP and MFN, we expected to replicate these results.

Method

Participants and procedures

U.S. participants were 520 men and 672 women ranging in age from 18 to 28 years (*M* = 18.62;*SD* = 2.52).Serbian participants were 501 men and 509 women between 19 and 29 years of age (*M* = 23.12;*SD*= 4.63).Polish participants were 354women and 345 men ranging in age from 18 to 35 years (*M* = 21.55;*SD* = 2.13). In Italy, participants were 386 women and 321 men ranging in age from 18 to 28 years (*M* = 19.21;*SD* = 1.40).This is a different sample than in Studies 1, 2, and 3. All participants were college students and homogeneous in terms of ethnicity (all participants were Caucasian).Data were collected as part of a course assignment at the Arizona State University (United States), at the "Sapienza", University of Rome(Italy), and at the Catholic University of Lublin (Poland), and at the University of Novi Sad (Serbia). Students from Italy and U.S. received course credits for their participation in the study.In all countries, we administered well validated versions of the RGSE already available in English (Rosenberg, 1965), Polish (Łaguna, Lachowicz-Tabaczek, Dzwonkowska, 2007), Serbian (Opačić, 1993), and Italian (Caprara, et al. 2012) languages. Items were presented in the same order in all samples.

Results and Discussion

Results showed that Model 9 yielded a good fit within each of the four countries (Table 1).Factor loadings (factors loadings were presented in the Online Appendix A) on GSE were all significant (*M* = .55; *SD* = .13), ranging from .25 (Italy) to .87 (Poland). Factor loadings of items PO6 and PO7 on MFP were non-significant in U.S., Serbia, and significant (but negatively signed) in Italy and Poland. The remaining loadings on MFP were positive and significant (*M* = .32, *SD* = .12), ranging from .22 (U.S.) to .59 (Italy). Factor loadings on MFN are all positive and significant (*M* = .44, *SD* = .18), ranging from .15 (U.S.) to .72 (Poland). In all countries, only itemsPO6 and PO7 were pure markers of GSE, as in previous studies. However, four additional items emerged as relatively pure markers of GSE in one or two countries. These were items NE3 and NE5 (U.S. and Poland) and items PO1 and PO2 (U.S.). Item NE9 consistently showed a primary loading on MFN

in all countries. All other items loaded primarily on GSE (although the primary loadings are not apparently higher than the secondary).

Variance decomposition and scale reliability

Results from variance decomposition replicated those of previous studies (Online Appendix B, Table 2):items PO6 and PO7 were almost completely composed of GSE variance (plus measurement error) in all countries. The highest proportion of MFN variance was observed for NE10 in all countries, and for NE9 in U.S., Italy, and Poland. All other items were composed primarily of GSE variance, with exception of PO1 in Italy, which was composed primarily by MFP variance. At the scale level, GSE explained the highest proportion of variance (from 60% in Italy to 77% in U.S.), followed by MFP (from 10% in U.S. to 19% in Italy), and MFN (from 1% in U.S. to 8% in Italy). Scale overall reliability was high: .88 (U.S.), .87 (Italy), .86 (Poland), and .86 (Serbia).

Cross-cultural measurement invariance

As a next step, we used a multi-group CFA to assess the cross-cultural invariance of Model 9, following the same procedure as in Study3. The configural and the metric invariance models fitted the data well (Table 1). However, the addition of equality constraints on item loadings substantially worsened model fit (Table 1).Partial weak invariance was reached after releasing the equality constraints of (1) the factor loading of the negatively worded item NE3 on MFN in Serbia and Poland; (2) the factor loading of negatively worded item NE5 on MFN in Serbia and the U.S.; (3) the factor loading of the positively worded item PO7on MFP in Poland and Serbia; (4) the factor loading of NE10 on MFN in the U.S.; (5) the factor loading of NE8on MFN in Poland; (6) the factor loading of NE8 in Serbia on GSE; (7) the factor loading of PO7 in Poland on GSE; and (8) the factor loading of PO6 in Italy on GSE. This partial-metric-invariant model fit the data well, and was not appreciably different from the configural invariance model according to the ΔCFI index (Table 1).The scalar model invariance fit the data well, although significantly worse than the previous partial scalar model (Table 1). To obtain scalar invariance, we released constraints imposed on the intercepts of (1) NE9 in the U.S., (2) PO4 in Serbia, and (3) NE10 in Poland (Table 1).As a final step, we compared latent means for GSE, MFP and MFN. We selected the U.S. as the

reference group because previous findings have consistently reported higher levels of GSE for U.S. than for other countries (see Schmitt &Allik, 2005).Accordingly, the estimated latent means can be interpreted as standardized mean differences with respect to the U.S. sample (SMD_{usa}).Constraining latent means to equality resulted in a significant decrement of the fit, ΔCFI=.019, $\Delta\chi^2_{(9)}$= 154.64; p< .05. Eventually, equality constraints were maintained (1) for GSE between Serbia and Poland, and (2) for MFP in Italy and Serbia, ΔCFI=.00;$\Delta\chi^2_{(2)}$=.07, p =.97. The meaning of these constraints is that Serbia and Poland did not differ in mean levels of self-esteem, and Italy and Serbia did not differ in mean levels of MFP. As expected, the highest levels of self-esteem were found in the U.S. and Italy (SMD_{usa} = -.21, p = .14), followed by Serbia (SMD_{usa} =-.30, p< .05), and Poland (SMD_{usa} =-.31, p< .05). The highest levels of MFP were found in Italy (SMD_{usa} =.16, p< .05) and Serbia (SMD_{usa} =.14, p< .05), followed by the U.S. (SMD_{usa} =.00) and Poland (SMD_{usa} =-.07,p = .40). Higher levels of MFN were found in Poland (SMD_{usa} =.68,p< .05) and Serbia (SMD_{usa} =.30,p< .05), followed by the U.S (SMD_{usa} =.00) and Italy (SMD_{usa} =-.08,p = .30).In summary, this study supported the viability of Model 9 across four different countries, along with the robustness of the measurement properties of the RGSE, in terms of partial measurement invariance. Finally, this study revealed a substantive pattern of mean level differences in GSE, MFP and MFN mean levels. These results are discussed in full detail below.

General Discussion

The RSES is perhaps the most widely used instrument to assess self-esteem (Donnellan et al., 2011; Gray-Little, Williams, and Hancock, 1997; Schmitt & Allik, 2005). It has been included in important longitudinal studies (e.g., the Americans' Changing Lives, Longitudinal Study of Generations, or the National Longitudinal Survey of Youth) and probably has been subjected to more psychometric analyses and empirical validation than any other self-esteem measure (Wylie, 1989). The goal of this investigation was to address unresolved issues related to the dimensionality of the instrument by testing alternative models. Below we provide a detailed discussion of resultsand suggestions for future research.

Is a One-factor Model Appropriate to Describe the RGSE Structure?

We compared ten structural solutions to the RGSE supported by previous studies (Marsh et al., 2010). Our data clearly suggest that the RGSE does not have a one-factor structure. Across different samples, raters, and cultures, results suggested that there are three dominant latent factors assessed by the RGSE—a general self-esteem factor (GSE), plus two specific factors associated with negatively (MFN), and with positively worded items (MFP). When computed with an appropriate index, scale reliability was adequate, according to current standards (Kline, 2010).

These results have several implications for the use of the RGSE scale. First, the overall mean score on the RGSE does not take into account the presence of specific factors associated with positively and negatively worded items might generate biased results. As an example, we re-estimated the model reported in Study 2, using the poor fitting Model 1 (see Alessandri et al., 2013, for a similar approach). Of course, we are aware that a misspecified model may lead to biased parameter estimates (Kline, 2010), but we believe that these data can be useful to understand the effect of biases associated with collapsing scores. In this model, the significant relations of GSE with aggression, depression, and prosociality remained unaltered. Yet, unlike what we found by using a bifactorial model (i.e., Model 9), GPA was significantly predicted by GSE. All in all, it is likely that collapsing in a single factor the proportion of variance belonging to different constructs is responsible, for example, for some of the problems raised in the debate surrounding the relation between GSE and GPA (Donnellan et al., 2011). This relation might be significantly attenuated when measures of GSE and MFP are confounded. Anyway, it is possible that imposing a unidimensional model for the RGSE items can mistakenly generate spurious relations between GSE and other variables. A more nuanced approach in future studies might be the use of appropriate methodologies to clarify how GSE, MFP, and MFN are related to outcomes of interest. On this regard, even though the RGSE scale appears to be composed of three factors does not mean that researchers should consider each of them in future studies. Instead, it seems reasonable that researchers focus their attention on the factor that best matches their research interests. However, we recommend that researchers compute measures of GSE using the appropriate bifactor model.

In short, we obtained evidence that the three factors assessed by the RGSE reflect different psychological features. Most importantly, an interpretation of the general factor as "general self-esteem", of MFP as a measure of "self-competence," and of MFN as "self-derogation" seems sensible in light of the data, and in line with our initial expectations. From a general stance, these results align with earlier intuitions of Tafarodi and Swann (1995), and of Kaplan and Pokorny (1969), who surmised that positive and negative worded items from the RGSE assessed aspects of self-evaluations different from general self-esteem. Moreover, these data confirm previous studies suggesting that MFP and MFN reflect different psychological features (DiStefano & Motl, 2009; Quilty et al., 2006). In other words, these factors do not simply capture pure method variance that is unrelated with substantive variables. Seemingly stable individual tendencies, apart from GSE level, also influence individuals' responses to RGSE items. In sum, it seems safe to conclude at this point that individuals' answers on items of the RGSE are influenced by a combination of individual self-esteem, perceptions of self-confidence (positively worded items), and feelings of self-derogation (negatively worded items), plus, of course, measurement error.

Are MFP and MFN the Product of Biases Associated with the Method of Assessment?

We submitted the RGSE to a relatively strong test of the robustness of its structure. Based on the results, we can confidently assert that the RGSE structure is relatively stable across observers and cultures. Taxing tests of measurement invariance provided evidence of configural and (partial) metric and scalar invariance. The newest and perhaps most interesting feature of this study was the examination of cross-rater invariance of the RGSE. First, we found that psychometric properties were basically preserved when the RGSE was used to evaluate others' rather than one's own self-esteem. In addition, we found a moderately high degree of cross-rater convergence for GSE (see also Robins, Hendin & Trzesniewski,2001), and a moderate degree of convergence for both MFP and MFN. Whereas care always should be taken when evaluating the nature of convergence among psychometric factors (see Alessandri et al., 2010, for a discussion and a similar point), we believe that from a theoretical point of view, these findings, along with the data corroborating the cross-cultural generality of the model, suggest that it is highly debatable to consider MFP and MFN the

result of spurious method variance. We contend, instead, that the term *specific factor* should be considered more appropriate when describing these factors, in line with their psychometric status when considered under the lens of a classical bifactor model (see Chen et al., 1996).

Other interesting findings of this analysis concern the accuracy with which levels of GSE, MFP, and MFN can be reliably assessed by an external observer. According to our results, it seems that observers were quite accurate in evaluating GSE, and, in fact, we found non-significant mean differences between self- and other-rated GSE. Instead, it seems that individuals (i.e., those self reporting), in comparison to observers, generally hold a more positive attitude toward their perceived capacities (i.e., higher MFP), and have a less negative attitude toward themselves (i.e., lower MFN). These results accord nicely with previous studies showing that self-assessments are usually biased in the direction of positively distorted evaluations (e.g., Alicke & Govorun, 2005). However, these results require further empirical replication and validation.

What is the Relative Importance of Each Factor?

One important feature of our study is that it offers a decomposition of observed variance into GSE, MFP, MFN, and measurement error components. Across studies, we found that the RGSE scale reflects a preponderance of self-esteem variance (*M*= 70.06%;*min* = 66.54%; *max* = 77.22%). Although the MFP (*M* = 15.00%; *min* = 9.79%; *max* = 18.06%) and MFN (*M* = 3.19%; *min* = 1.45%; *max* = 4.17%) represent a small portion of variance in the items, they should not necessarily be considered less important factors because they were predictive of distinct outcomes. Instead, these results simply indicate that the RGSE provides a valid measure of GSE, and a less efficient measure of MFP and MFN. However, the presence of MFP and MFN should be acknowledged and explicitly modelled. Blending these variance components together has the potential to impact the research questions that psychologists pose and, as a result, the trustworthiness of the answers they obtain. Finally, looking at the item levels, we noted that only two items were consistently pure (although noisy, and in fact contaminated by a high proportion of measurement error) measures of global self-esteem. As it stands, these results seem to corroborate the idea that, at times, using short measures to assess widely acknowledged constructs like self-esteem may be advantageous and

possible (Burisch, 1984; Robins et al., 2001). Future studies might test the idea that these two items represent useful pure markers of self-esteem.

Is the Bifactor Structure Sufficiently Stable?

A significant contribution of this study is to the understanding of the degree of cross-cultural invariance of the RGSE items. All in all, we found a satisfactory degree of invariance, pointing to a satisfactory degree of robustness of the psychometric properties of the bifactor model. We also found a pattern of interesting mean-level differences in GSE, MFP and MFN. In part, our results accord nicely with what reported by Schmitt and Allik (2005). For example, participants from the U.S. reported higher levels of self-esteem than participants from European countries. However, our results were unique because we tested the ways in which MFP and MFN differed across countries. For example, people in Italy and Serbia reported higher scores on MFP than in the US, and those from Poland resulted in higher scores on MFN than in the U.S. Because scores on the MFP and MFN factors can be interpreted as a measures of self-competence and self-derogation, the pattern seems to mimic quite well differences across countries in values of *egalitarianism*, *social embeddedness*, and *autonomy*(Schwartz, 2006). However, more data and representative samples are necessary to put the above differences into a broader theoretical framework of reliable cultural differences in self-derogation and self-competence. Most importantly, the bifactor structure of the RGSE needs to be further evaluated for generalizability across a wider range of countries and cultures.

What are the Practical implications of the Model?

We believe that our research could have wide relevance for the field of general self-esteem research, which has classically relied on the RGSE scale. In the context of theory building, for example, researchers should be aware that the scale does not measure a single, general factor. It should be noted that present results do not necessarily question the practical usefulness of the RGSE in applied contexts, and do not call into question its usefulness for screening purposes, given that variance decomposition revealed that the GSE factor explained the lion's share of items' variance. However, we surmise that using a total score as a basis for screening individuals at risk for low self-

esteem might be suboptimal, as the total score may be contaminated by different sources of variance. Recognizing the different variance components tapped by the ten RGSE items may give practitioners additional useful information.

In addition to these practical and theoretical benefits, there are other potential conceptual advances that follow from carefully considering our bifactorial model. For instance, theories in developmental psychology might benefit from simultaneously investigating the individual differences in the development of each component of the RGSE to achieve a richer understanding of how the evaluative components of the self develop over time. Likewise, theories in clinical psychology could try to link each of the three latent factors composing the RGSE psychometric structure to the same clinical phenomena (i.e., depression, anxiety, etc.) to better understand the relations between different components of the evaluative self-concept.

There are also important psychometric considerations that follow from our results. For instance, the bifactor model can be viewed as incorporating hypotheses concerning the way that individuals answer the RGSE items. In this sense, variance decomposition at the item level represents a suitable way to understand the degree of precision with which each item does capture different components of self-evaluations. Results from this analysis might be further suited to build a more refined understanding of the different factors that shape one's evaluation of the self.

Conclusion

Even though several studies have investigated the dimensionality of RGSE, the issue seems at present to be far from being definitively settled. We hope to have provided new data supporting the view that this instrument is best represented by a bifactor model (Chen, et al. 2006), which is the only model available in the literature able to capture the multifaceted sources of influence reflected in these items. Each factor included in the model (i.e., GSE, MFP, MFN) has an apparently straightforward interpretation that, of course, could be further refined in future studies. In particular, the generalizability of observed results should be extended to countries with different cultural background than Serbia, Poland, Italy, and the US. We emphasize that our current results do not necessarily hamper the value of previous studies based on the RGSE total score. It is difficult on the

basis of this study to know to what extent the use of a more appropriate structure for the RGSE would determine change to results in studies using a one-factor model. However, one should bear in mind that the RGSE scale measures three distinct aspects of self-evaluations. Depending on the specific research goals, the results of our study might provide a basis for redefining the current scale to arrive at distinct measures of each of these factors.

References

Alessandri, G., Vecchione, M., Donnellan, M.B., &Tisak, J. (2013). An application of the LC-LSTM framework to the self-esteem instability case. *Psychometrika, 4,* 769-792.10.1007/s11336-013-9326-4

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

Alessandri, G., Vecchione, M., Tisak, J., & Barbaranelli, C. (2011). Investigating the nature of method factors through multiple informants: Evidence for a specific factor? *Multivariate Behavioral Research, 46,*625-642.doi:10.1080/00273171.2011.589272

Alicke, M.D., &Govorun, O. (2005). The better-than-average effect. In M.D. Alicke, D.A. Dunning, & J.I. Krueger, (Eds.), *The self in social judgment*. New York: Psychology Press.

Bartko, W.T., &Eccles, J.S. (2003). Adolescent participation in structured and unstructured activities: A person-oriented analysis. *Journal of Youth and Adolescence, 32,* 233-241.doi:10.1023/A:1023056425648

Baumeister, R.F., Campbell, J.D., Krueger, J.I., &Vohs, K.D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest, 4,* 1-44.doi:10.1111/1529-1006.01431

Bentler, P.M. & Chou, C.P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16,* 78-117

Blascovich, J., &Tomaka, J. (1991). Measures of self-esteem. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.). *Measures of personality and social psychological attitudes* (vol.1, pp. 115-160). San Diego, CA: Academic Press.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203-219.doi:10.1037/0033-295X.110.2.203.

Burisch, M. (1984). Approaches to personality inventory construction. A comparison of merits. *American Psychologist, 39,* 214-227.doi:10.1037/0003-066X.39.3.214

Burnham, K.P. & Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research, 33,* 261-304.doi:10.1177/0049124104268644

Buss, A.H., & Perry, M.P. (1992).The aggression questionnaire. *Journal of Personality and Social Psychology, 63,* 452-459.doi:10.1037/0022-3514.63.3.452

Byrne, B.M. (1983). Investigating measures of self-concept. *Measurement and Evaluation in Guidance, 16,* 115-2.

Byrne, B.M., &Shavelson, R. (1986). On the structure of adolescent self-concept. *Journal of Educational Psychology, 7,* 474-81.doi:10.1037/0022-0663.78.6.474

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Caprara, G.V., Alessandri, G., Trommsdorff, G., Heikamp, T., Yamaguchi, S., & Suzuki, F. (2012). Positive orientation across countries. *Journal of Cross Cultural Psychology, 43,* 77-83.doi:10.1177/002202211142225

Caprara, G.V., Steca, P., Zelli, A., & Capanna, C. (2005). A new scale for measuring adult's Prosociality. *European Journal of Psychological Assessment, 21,* 77-89.doi:10.1027/1015-5759.21.2.77.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and validity assessment.* Newbury Park, Sage.

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

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling,* 9, 235-255.

Diggory, J.C. (1966). *Self-evaluation: Concepts and studies.* New York: Wiley.

DiStefano, C. & Motl, R.W. (2009). The relationship between personality factors and item phrasing. *Personality and Individual Differences, 46,* 309-313.doi:10.1016/j.paid.2008.10.020

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

Donnellan, M.B., Trzesniewski, K.H., Robins, R.W., Moffitt, T.E., & Caspi, A. (2005). Low self-esteem is related to aggression, antisocial behavior, and delinquency. *Psychological Science, 16,* 328-335.doi:10.1111/j.0956-7976.2005.01535.x

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.doi: dx.doi.org/10.1037/1082-989X.8.1.38

Epstein, J.A., Griffin, K.W., &Botvin, G.J. (2004). Efficacy, self-derogation, and alcohol use among inner-city adolescents: Gender matters. *Journal of Youth & Adolescence, 33,* 159-166.doi:10.1023/B:JOYO.0000013427.31960.c6

Fiske, D.W. (1987). Construct invalidity comes from method effects. *Educational and Psychological Measurement, 47,* 285-336.doi:10.1177/0013164487472001

Gecas, V. (1971). Parental behavior and dimensions of adolescent self-evaluations. *Sociometry, 34,* 466-482.doi:10.2307/2786193

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

Kaplan, H., &Pokorny, A. (1969). Self-derogation and psychosocial adjustment. *Journal of Nervous and Mental Disease, 149,*421-434.doi:10.1097/00005053-196911000-00006

Kaplan, H.B. (1978). Deviant behavior and self-enhancement in adolescence. *Journal of Youth and Adolescence.7,* 253-277.doi:10.1007/BF01537977

Kaplan, H.B. (1980). *Self-attitudes and deviant behavior.* Santa Monica, CA: Goodyear.

Kaplan, H.B., Martin, S.S., & Robbins, C. (1982). Application of a general theory of deviant behavior: Self-derogation and adolescent drug use. *Journal of Health and Social Behavior, 23,* 274-294.doi:10.2307/2136487

Kaufman, E., 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.doi:10.1037/0033-2909.112.1.165

Kline, R.B. (2010). *Principles and practices of structural equation modeling.* New York: Guilford.

Kuster, F. &Orth, U. (2013). The long-term stability of self-esteem: Its time-dependent decay and nonzero asymptote. *Personality and Social psychology Bulletin,39,*677-690.doi:10.1177/0146167213480189

Łaguna, M., Lachowicz-Tabaczek, K., Dzwonkowska, I. (2007). Skalasamooceny SES MorrisaRosenberga – polskaadaptacjametody [The Rosenberg Self-Esteem Scale: Polish adaptation of the scale]. *PsychologiaSpołeczna* [Social Psychology], *2*, 164-176.

Lord, F.M. &Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Marsh, H.W. (1996). Positive and negative self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810-819.doi:10.1037/0022-3514.69.6.1151

Marsh, H.W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years. *Personality and Social Psychology Bulletin, 34,* 542-552.doi: 10.1177/0146167207312313

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

Motl, R.W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling, 9,*562-578.doi:10.1207/S15328007SEM1003_6

Muthén, L., & Muthén, B. (2004). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Opačić, G. (1993). *Porodične varijable i koncept o sebi kod adolescenata* [Family variables and self-concept of adolescents]. Research Report, University of Belgrade, Belgrade, Serbia.

Openshaw, D.K., Thomas, D.L., & Rollins, B.C. (1981). Adolescent self-esteem: A multidimensional perspective. *Journal of Early Adolescence, 1,* 273-282.doi:10.1177/027243168100100306

Orth, U., Robins, R.W., &Widaman, K.F. (2012). Life-span development of self-esteem and its effects on important life outcomes. *Journal of Personality and Social Psychology, 102,* 1271-1288.doi:0.1037/a0025558

Owens, T.J. (1994). Two dimensions of self-esteem. *American Sociological Review* 59:391-407.doi:10.2307/2095940

Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver & L.S. Wrightsman, (Eds.), *Measures of personality and social psychological attitudes,* (Vol. 1, pp. 17–59). Academic Press, San Diego.

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

Radloff, L.S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385-401.doi:10.1177/014662167700100306

Reise, S.P., Morizot, J., & Hays, R.D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16,* 19-31.doi:10.1007/s11136-007-9183-7

Robins, R.W., Hendin, H.M., & Trzesniewski, K.H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27,* 151-161.doi:0.1177/0146167201272002

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

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

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *ASA 1988 Proceedings of the Business and Economic StatisticsSection* (308-313). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P.M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.

Schmitt, D.P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations. *Journal of Personality and Social Psychology, 89*, 623-642.doi:10.1037/0022-3514.89.4.623

Schwartz, S.H. (2006). Value orientations: Measurement, antecedents and consequences across nations. In R. Jowell, C. Roberts, R., Fitzgerald & E. G (Eds.), *Measuring attitudes cross-nationally* (pp. 169-203). London: Sage

Shevlin, M.E., Bunting, B.P., & Lewis, C.L. (1995). Confirmatory factor analysis of the Rosenberg Self-Esteem Scale. *Psychological Reports, 76,* 707-710.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.doi:10.1007/s11336-008-9101-0

Tafarodi, R.W., & Milne, A.B. (2002). Decomposing global self-esteem. *Journal of Personality, 70,* 443-483.doi: 10.1111/1467-6494.05017

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

Tomás, J., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling, 6,* 84-98.doi:10.1080/10705519909540120

West, S.G., Finch, J.F., & Curran, P.J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 56-75). Thousand Oaks, CA: Sage.

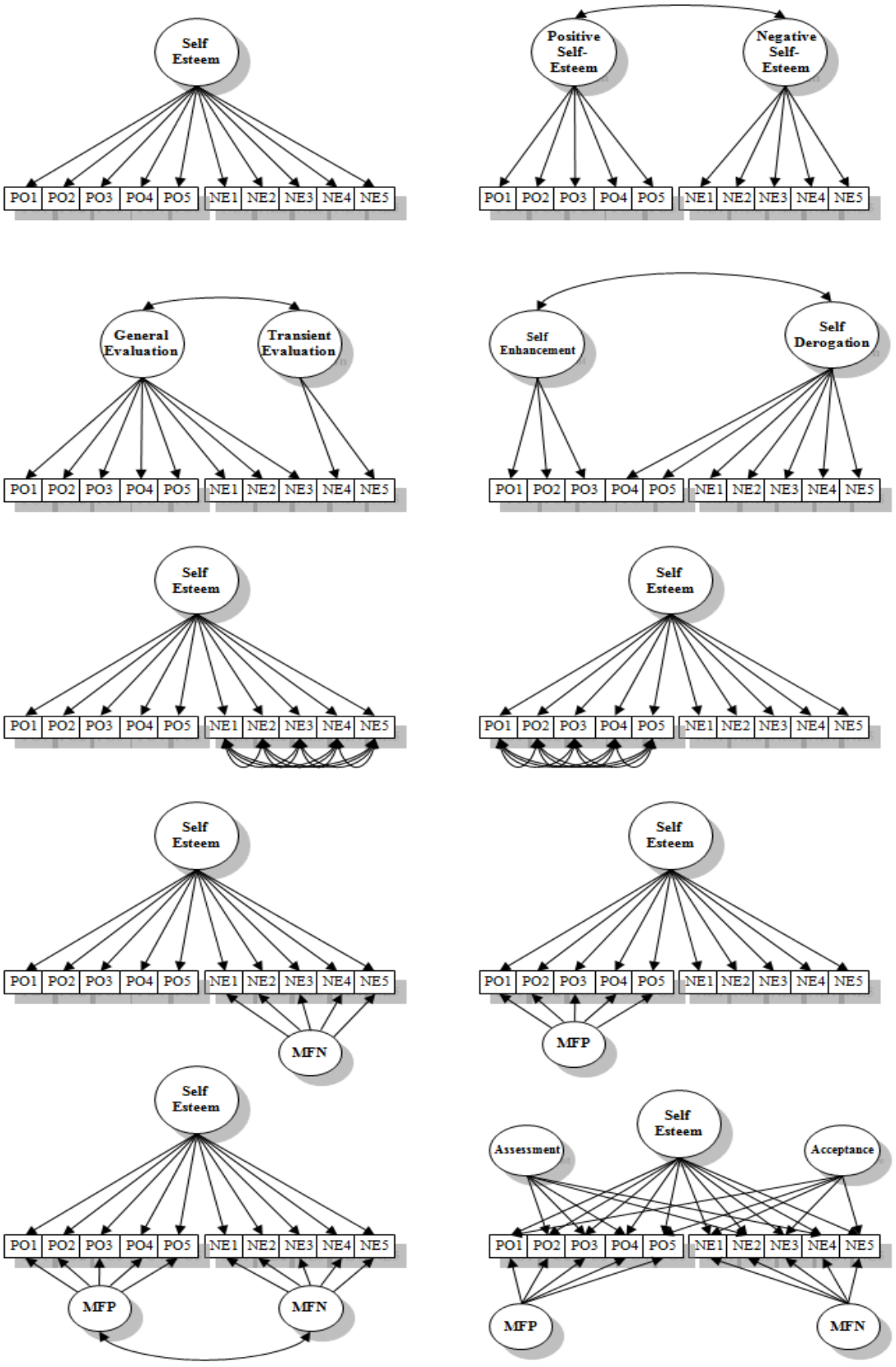
Wylie, R.C. (1989). *Measures of self-concept*. Lincoln: University of Nebraska Press.

Table 1. Goodness of fit of alternative models for the RGSE(study 1) and cross-cultural invariance of the best-fitting Model 9 (study 4).

	Study 1: Goodness of fit of alternative models							
	χ^2	*df*	CFI	TLI	RMSEA	AIC		
Model 1	7344.73*	35	.74	.67	.14(.13-.14)	200715.82		
Model 2	3977.51*	34	.86	.82	.10(.10-.11)	196095.73		
Model 3	9854.04*	34	.73	.65	.16(.16-.17)	200708.80		
Model 4	3625.19*	34	.87	.83	.10(.09-.10)	195656.56		
Model 5	1038.68*	25	.97	.95	.06(.05-.06)	191911.43		
Model 6	3318.52*	25	.88	.79	.11(.10-.11)	195299.37		
Model 7	2707.53*	30	.90	.86	.09(.08-.09)	194413.31		
Model 8	3625.47*	30	.87	.81	.10(.10-.11)	195642.32		
Model 9	**483.40***	**24**	**.98**	**.97**	**.04(.04-.05)**	**191493.54**		
	Study 4: Cross-Cultural Invariance of Model 9							
	$SB\chi^2$	*df*	CFI	TLI	RMSEA	$\Delta SB\chi^2$	Δdf	ΔCFI
U.S	13.52	24	1.00	1.02	.01(<.01-.01)	-	-	-
Serbia	26.02	24	1.00	1.00	.01(<.01-.04)	-	-	-
Poland	35.12	24	.99	.98	.04(<.01-.06)	-	-	-
Italy	37.34*	24	.98	.97	.04(.01-.06)	-	-	-
Configural	142.17*	96	.993	.986	.026(.016, .035)	-	-	-
Metric	414.77*	147	.963	.954	.051(.045, .056)	280.74*	51	-.030
$Metric_{Partial}$	268.74*	136	.984	.978	.037(.030, .044)	128.24*	40	-.009
Scalar	407.26*	147	.964	.955	.050(.044, .056)	178.83*	11	-.020
$Scalar_{Partial}$	302.89*	144	.980	.974	.039(.033, .046)	40.69*	8	-.004

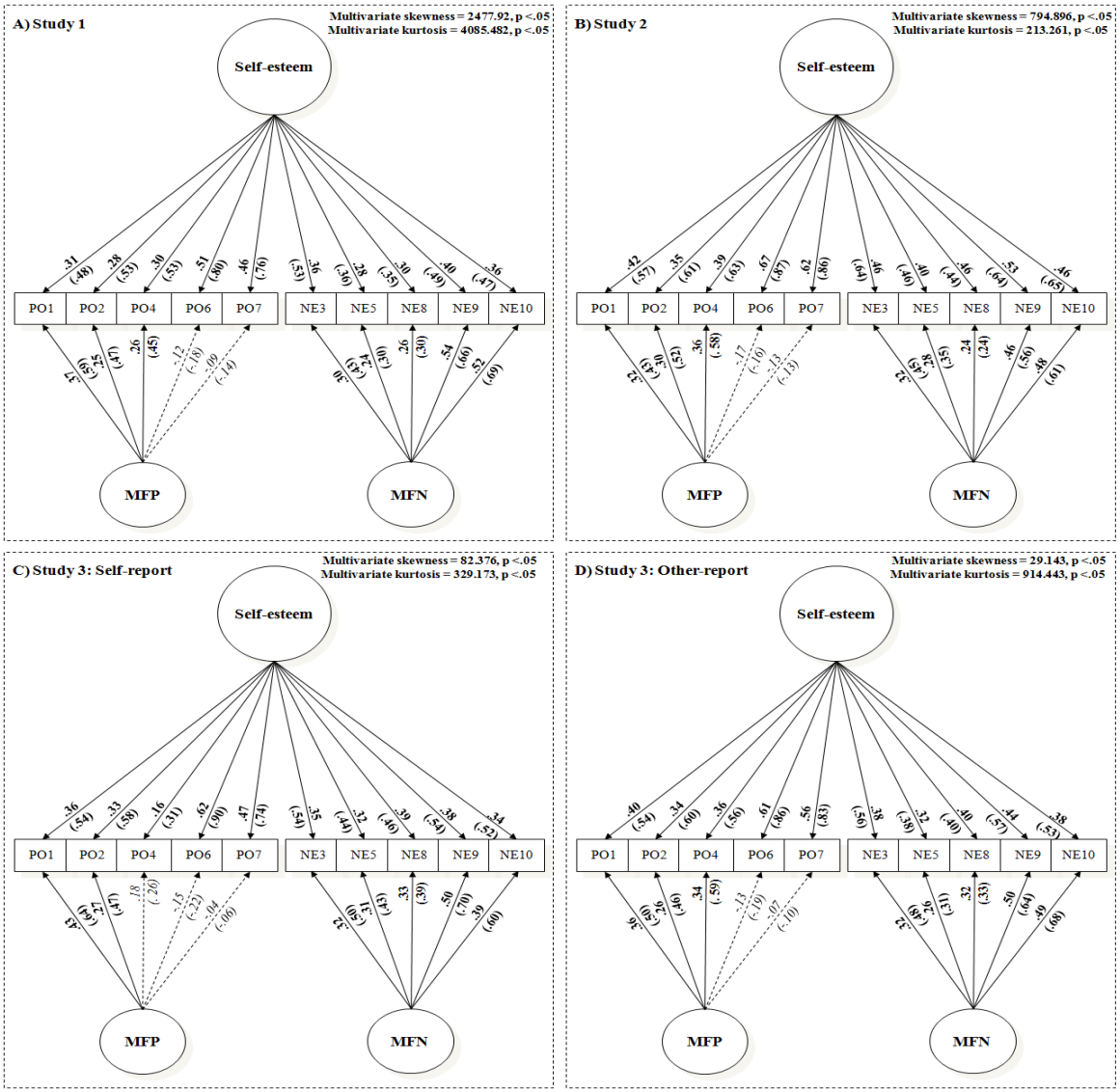
*Note.*Model 10 was not-identified, it was excluded for further consideration.
* p< .05.

Figure 1. Alternative factorial models for Rosenberg General Self-Esteem Scale.



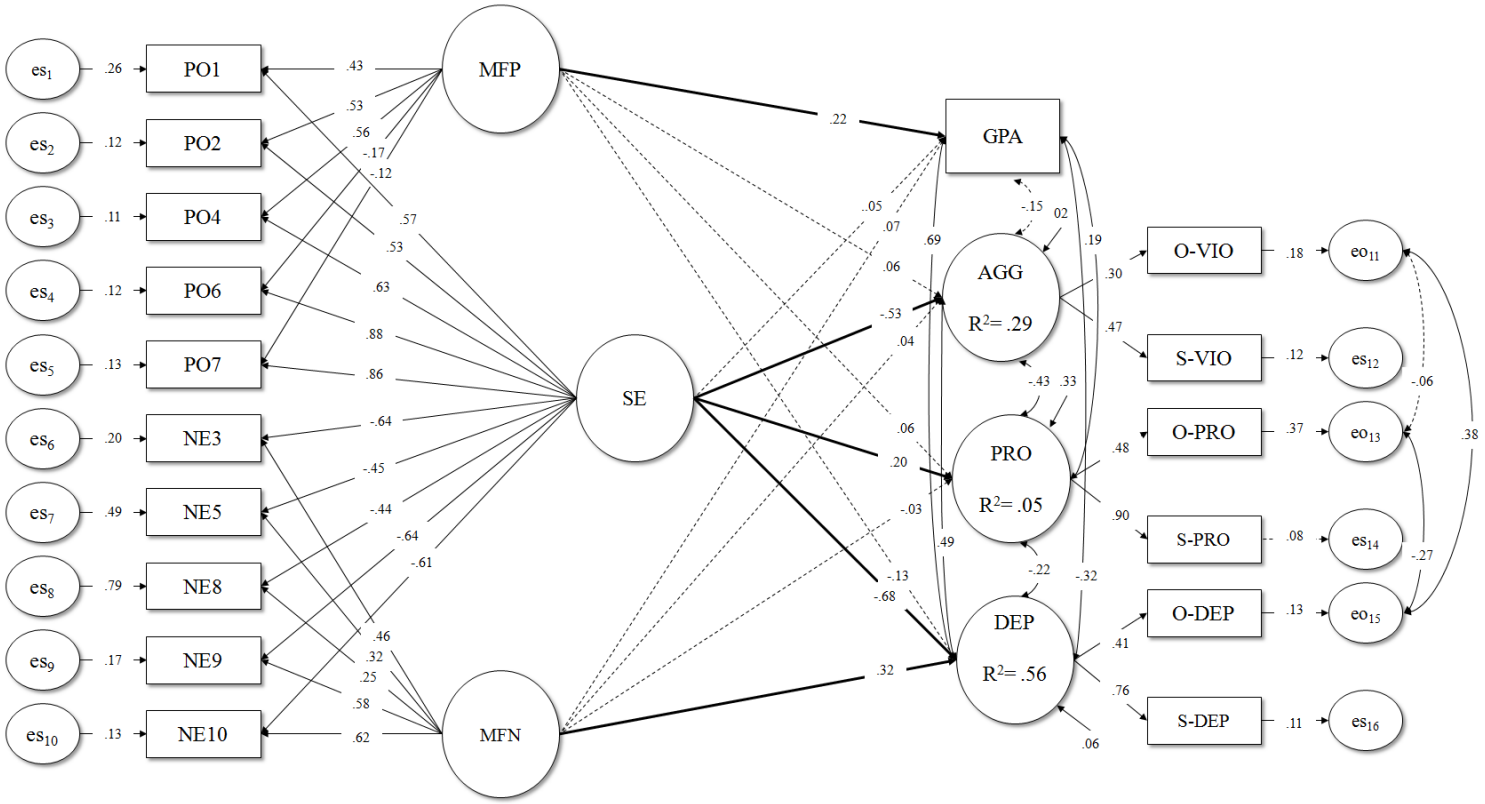
Note. PO = positively worded item; NE = negatively worded item; MFP = method factor associated with positively worded items; MFN = method factor associated with positively worded items.

Figure 2. Unstandardized (in parentheses) and completely standardized loadings for Model 9 in the first three studies



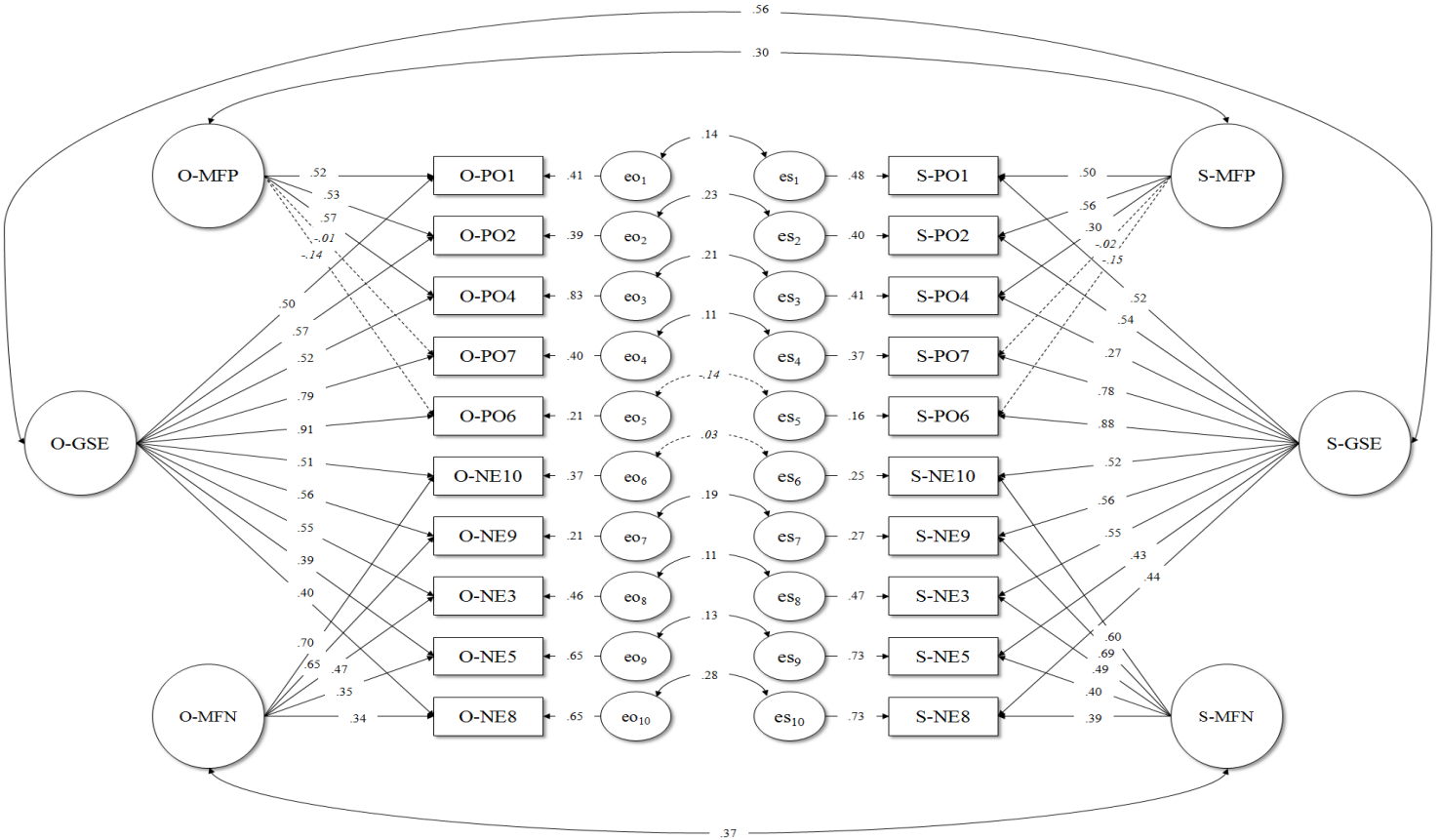
Note. Statistically significant (p< .05) coefficients are in bold and on solid lines; statistically not significant (p> .05) coefficients are in italics and on dotted lines. Coefficients outside rounded brackets are unstandardized; coefficients inside rounded brackets are completely standardized; PO1 =I feel that I'm a person of worth, or at least on an equal plane with others; PO2 = I feel that I have a number of good qualities; PO4 =I am able to do things as well as most other people; PO6 = On the whole, I am satisfied with myself; PO7 =I take a positive attitude toward myself; NE10 = At times I think I am no good at all; NE9 = I certainly feel useless at times; NE8 = I wish I could have more respect for myself; NE5 = I feel I do not have much to be proud of; NE3 = All in all, I am inclined to feel that I am a failure.

Figure 3. Predictive value of GSE, MFP, and MFN factors. All estimates presented in the figure are completely standardized



Note. Solid lines represent significant paths; dashed lines represent non-significant paths. S = self-rated; O = other rated; es-eo = error terms associated with self-rated (es) and other rated items (eo).PO = positively worded item (self-rated); NE = negatively worded item (self-rated).GSE = general self-esteem (self-rated); MFP = positive wording specific factor; MFN = negative wording specific factor. GPA = grade point average; AGG = aggression; PRO = prosociality; DEP = depression. To simplify interpretation of direct effect of MFN, negative items were not reversely scored. Items are indexed by their position in the scale.

Figure 4. Convergence of GSE, MFP and MFN across self- and other ratings. All estimates presented in the figure are completely standardized



Note. S = Self = self-rated; O= other rated; es = error terms associate with self rated items; eo = error terms associate with other rated items; PO = positively worded item; NE = negatively worded item. GSE = general self-esteem; MFP = the specific factor associated with positively worded items; MFN = the specific factor associated with negatively.