Method Effects: The Problem With Negatively Versus Positively Keyed Items

Lindwall, M; Barkoukis, V; Grano, C; Lucidi, F; Raudsepp, L; Liukkonen, J; Thogersen-Ntoumani, Cecilie

DOI:
10.1080/00223891.2011.645936

Running head: SELF-ESTEEM AND METHOD EFFECTS

Method effects: The problem with negatively versus positively keyed items

Magnus Lindwall,[1] Vassilis Barkoukis,[2] Caterina Grano,[3] Fabio Lucidi,[3] Lennart Raudsepp,[4] Jarmo Liukkonen,[5] and Cecilie Thøgersen-Ntoumani,[6]

[1]Department of Psychology, Department of Food and Nutrition, and Sport Science, University of Gothenburg, Gothenburg, Sweden

[2] Department of Physical Education and Sport Science, Aristotle University of Thessaloniki, Thessaloniki, Greece

[3] Department of Psychology, Sapienza University of Rome, Rome, Italy

[4] Institute of Sport Pedagogy and Coaching Science, University of Tartu, Tartu Estonia

[5] Department of Sport Sciences, University of Jyväskylä, Jyväskylä, Finland

[6] School of Sport and Exercise Sciences, University of Birmingham, Birmingham, UK

Address correspondence to:

Magnus Lindwall, Department of Psychology, University of Gothenburg, P.O. Box 500, SE-405 30 Gothenburg, Sweden

E-mail: Magnus.Lindwall@psy.gu.se

Method effects: The problem with negatively versus positively keyed items

Abstract

Using confirmatory factor analyses, we examined method effects in Rosenberg's Self-Esteem Scale (RSES; Rosenberg, 1965) in a sample of older European adults. Nine hundred forty nine community-dwelling adults 60 years of age or above from five European countries completed the RSES as well as measures of depression and life-satisfaction. The two models that made acceptable fit with the data included methods effects. The method effects were associated with both positively and negatively worded items. Method effects models were invariant across gender and age, but not across countries. Both depression and life-satisfaction predicted method effects. Individuals with higher depression scores and lower life-satisfaction scores were more likely to endorse negatively phrased items.

Keywords: Depression, Invariance, Life-satisfaction, Method Effects, Self-esteem

Self-esteem (SE) is often identified as a significant component of positive psychological health (Shiovitz-Ezra, Leitsch, Graber, & Karraker, 2009) and has been associated with functional health (Reitzes & Mutran, 2006) in older adults. The Rosenberg Self-Esteem Scale (RSES; (Rosenberg, 1965) is one of the most popular instruments used to measure SE (Blascovich & Tomaka, 1991), also in older adults (McAuley et al., 2005; Reitzes & Mutran, 2006).The RSES was originally designed as a unidimensional measure of global self-esteem and has been treated as such in the majority of studies that have utilized the scale. However, the factor structure of the RSES and its underlying dimensions has been a topic of debate among researchers for a long time (Carmines & Zeller, 1979; DiStefano & Motl, 2006; Marsh, 1996; Marsh, Scalas, & Nagengast, 2010). A general problem seems to be that the original uni-dimensional model of the RSES does not fit data well. Instead, a structure including two or more factors generally fit the data better in terms of factorial validity. Although some researchers have proposed that the RSES taps two substantively relevant underlying dimensions, for example self-liking and self-competence (Tafarodi & Swann, 1995), the majority of studies (Corwyn, 2000; DiStefano & Motl, 2009; Horan et al., 2003; Marsh, 1986; Marsh, et al., 2010; Tomas & Olivier, 1999; Wu, 2008) have found that method effects may explain the multidimensional factor structure of the RSES.

Method effects refer to tendencies to respond to questionnaires based on other criteria than their alleged content, resulting in systematic variance that is irrelevant to the study or the concept the researcher is attempting to measure (American Education Research Association, 1999; (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Common method variance is a general problem and reviews of more than 70 studies (Cote & Buckley, 1987) have found that approximately a quarter of the variance in a typical research measure may be due to systematic sources of measurement error, such as common method biases. More specifically, this indicates that, if measures include common method variance, the observed relationship

between the predictor and a criterion variable may be understated by approximately 25% (Podsakoff et al., 2003). However, method effects may inflate or suppress relations among variables and contribute to Type I or Type II errors (Bagozzi, 1993), and its consequences may therefore be hard to anticipate and highly problematic, in particular if researchers are not aware of their existence. Other potential consequences of method effects may be that models without method effects do not fit data and result in poor-fitting solutions which lead to the inaccurate conclusions that the construct has either poor, or good, discriminant validity (Brown, 2006).

One strategy to try to decrease these method effects has been to use both positively and negatively worded items, the basic idea being that reverse-coded items are like cognitive "speed bumps" that require respondents to engage in more controlled, as opposed to automatic, cognitive processing (Podsakoff et al., 2003). The RSES, for example, includes five positively and five negatively worded items. However, as pointed out by Marsh (1996), a critical assumption that underlies the strategy to use both positively and negatively phrased items is that positively and negatively worded items actually do measure the same underlying construct. When researchers identify two separate factors associated with the positively and negatively phrased items, the rationale for using both negatively and positively phrased items is called into question. A number of studies have found support for the notion that the use of both positively and negatively worded items instead may lead to method effects related to different wording of the items (Carmines & Zeller, 1979; DiStefano & Motl, 2006; Horan, DiStefano, & Motl, 2003; Marsh, 1996; Marsh et al., 2010; Tomas & Oliver, 1999). For example, early studies on the RSES using exploratory factor analysis found support for a two-dimensional structure of the RSES, consisting of positively and negatively worded items (Carmines & Zeller, 1979).

Due to the limitations of examining methods effects using exploratory factor analysis , researchers have started to use confirmatory factor analysis (CFA) in which they have adopted a multitrait-multimethod (MTMM) conceptual framework. In the development of this analytical framework, two types of models have been proposed and adopted (Bagozzi, 1993; Marsh & Grayson, 1995) to separate substantive content (e.g., self-esteem) from method effects. One type of model is the correlated trait, correlated uniqueness (CTCU) model. Another type is the correlated trait, correlated methods model (CTCM). As applied to RSES, the CTCU introduces correlations among the residuals or uniqueness (measurement errors) of the positively and negatively worded items. In contrast, the CTCM includes specific latent method effect factors underlying questionnaire items of the same method (i.e., positively or negatively worded format of items) along with a latent substantive factor (self-esteem). Thus, the method effects in CTCM models may be quantified and predicted by other factors or variables, something that is not possible when using the CTCU model. In using these two types of models to examine method effects, one seeks to establish if models including correlated measurement errors, or latent method factors, display a better fit with data compared with models that do not include them. If they do, support for the existence of method effects may be inferred. Using this strategy, a number of studies have found that models including method effects, either examined via CTCU or CTCM models or both, generally fit the data better compared with competing models without method effects (Corwyn, 2000; DiStefano & Motl, 2009; Horan et al., 2003; Marsh, 1986; Marsh, et al., 2010; Tomas & Olivier, 1999; Wu, 2008). Consequently, there is now fairly strong support for the proposition that the RSES is contaminated with method effects. Further, these studies show that the method effects are primarily associated with negatively worded items (Corwyn, 2000; DiStefano & Motl, 2006; Horan et al., 2003; Marsh, 1996; Tomas & Oliver, 1999). This is in contrast to other studies which have demonstrated that models including method

effects from both positively and negatively items result in the best fit to the data (Marsh et al., 2010; Quilty, Oakman, & Risko, 2006; Wu, 2008).

A general limitation with previous studies examining the method effects of the RSES is that the majority have used mainly young adults students and adolescents . As proposed in several papers, (DiStefano & Motl, 2006; Goldsmith, 1986; Quilty et al., 2006), method effects may vary across populations and may be more important for certain groups rather than others. To our knowledge, no previous study has examined method effects in the RSES in a sample of older adults. This is despite the fact that the RSES is widely used in elderly populations. According to the socio-emotional selectivity theory (SST; (Carstensen, Isaacowitz, & Charles, 1999) time horizons influence goals and consequently people's memories and attention. More specifically, as people age and time is perceived as more constrained, they will place increasing importance on emotionally meaningful goals and will be more likely to devote their memories and attention to the positive information that will enhance their current mood. Indeed, studies have found support for a positivity effect in older adults. Thus, older adults prefer positive information whereas for younger individuals negative information seem to be more salient (Carstensen & Mikels, 2005). Therefore, based on the SST and the positivity effect, it could be expected that stronger effect of method effects linked to positively worded items, rather than negatively worded items, would be found in samples of older adults.

Aside from age, gender could influence method effects in the RSES. Meta-analyses have found significant, albeit small, differences (d=.22) in self-esteem, favoring males (Kling, Hyde, Showers, & Buswell, 1999). A majority (135 of 218 effect sizes) of the effect sizes in this study was based on the RSES. Despite this, only one previous study has investigated if method effects in the RSES are similar across males and females. DiStefano and Motl (2009) found that the method effects associated with negatively worded items in the RSES did not

differ between males and females. However, the participants in this study were college students with a mean age of 22 years. Therefore, it is currently unknown if differences exist between older men and women in method effects in the RSES..

Another relevant question is if the potential method effects in the RSES is similar, or equivalent, across cultures and countries? For example, the results of a large study (D. P. Schmitt & Allik, 2005) examining differences in the RSES across 53 nations found that, although a one-dimensional factor structure of the RSES was largely invariant across cultures, negatively worded items were interpreted differently across nations. These findings suggest that in many cultures the answers to negatively worded items are systematically different from the answers to positively worded items. This may be termed a negative item bias.

The assumption that method effects of the RSES are not merely systematic measurement errors but may mirror underlying response styles, has also been suggested (DiStefano & Motl, 2006; Quilty et al., 2006). If the latent factor of method effects are related to, or predicted by, other constructs or variables, this would support the idea of method effects as a response style rather than systematic error. For example, DiStefano and Motl (2006) found that participants with higher apprehension of negative evaluation of others and higher levels of self-consciousness were less likely to demonstrate method effects associated with negatively worded items of the RSES. Examining the relation between personality constructs and method effect of negatively worded items, Quilty and colleagues (2006) found that the more conscientious and emotionally stable participantswere less likely to endorse negatively phrased items. Conversely, they also found that those withhigher avoidance motivation were more likely to endorse negatively worded items. These results suggest that method effects may be associated with particular response styles and therefore predicted by psychological factors and demographic variables.

One factor that has been suggested to affect people's self-report responses to questionnaire items is mood and positive and negative affectivity (Burke, Brief, & George, 1993; Podsakoff et al., 2003). For example, individuals with stable negative affectivity may use a response style which renders them more vulnerable for method effects associated with negatively keyed items. Similarly, individuals with high positive affectivity may be more prone to endorse positively worded items, regardless of content. Negative affectivity is one of the core symptoms of depression in older adults, that together with other core-symptoms such as feelings of worthlessness, self-critical cognitions, and cognitive distortions (Blazer, 2003; Fiske, Wetherell, & Gatz, 2009) may lead to halo effects and increased risk of method effects associated with negatively worded items. Positive affectivity, on the other hand, is strongly associated with life-satisfaction (DeNeve & Cooper, 1998). Thus, from a conceptual standpoint it seems likely that the two psychological factors of depression and life-satisfaction may predict method effects in the RSES in older adults.

The purpose of this study was to examine: (a) if method effects exist in RSES in a sample of older adults from five European countries; (b) if possible method effects in the RSES are linked primarily to positively or negatively worded items, or both; (c) if the level and nature (linked to positively or negatively worded items) of method effects in the RSES differ across gender, age and country; (d) if life-satisfaction and depression predict method effects in RSES.

METHOD

Participants

The sample consisted of 1.177 older community-dwelling adults 60 years of age or above (*M* age = 73.64; *SD* = 7.50) from five European countries; United Kingdom (*n* = 247), Sweden (*n* = 47), Finland (*n* = 159), Greece (*n* = 326), and Italy (*n* = 398). All participants resided in

urban settings. Females constituted 61.7% of the sample, and 58.70% of participants were married, while 29.10% were widowed. The majority of the sample reported primary (42.3%) or secondary (38.4%) education as their highest level of education.

Instruments

Self-esteem. Rosenberg's (1989) self-esteem scale (RSES) was used to measure older adults' global self-esteem. The scale is comprised of ten items measuring one factor; global self-esteem. An example item is "I feel that I'm a person of worth, at least on an equal plane with others". Responses were anchored on a four-point scale ranging from 1 (*strongly agree*) to 4 (*strongly disagree*). Previous studies with older populations have identified adequate internal reliability coefficients (Diehl, Hastings, & Stanton, 2001; McAuley et al., 2005). In the present study, the internal consistency coefficient of the scale was satisfactory (Cronbach α = .81).

Depressive symptoms. The Centre for Epidemiological Studies Depression scale (CESD; Radloff, 1977) was used to assess depressive symptoms during the past week. The scale is uni-dimensional and consists of twenty items. An example item includes "I felt tearful". Responses were provided on a four-point scale ranging from 1 (*rarely or none of the time-less than 1 day*) to 4 (*most or all of the time-5 to 7 days*) with overall scores ranging between 20 and 80. Previous research in older populations (Beekman, Deeg, Van Limbeek, Braam, & et al., 1997) provided support regarding the validity and reliability of the scale. The internal consistency of the scale was high (α = .85) in this study.

Life satisfaction. Global life satisfaction was measured using the Satisfaction With Life Scale (SWLS; (Diener, Emmons, Larsen, & Griffin, 1985). The scale comprises five items (e.g., "I am satisfied with my life"). Responses were provided on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The questionnaire has been widely

adopted, and high levels of reliability and validity have been reported (Diener et al., 1985). The internal consistency coefficient for the scale in the present study was .86.

Demographic characteristics. Apart from ticking a box representing their gender, the participants were asked to indicate their age by providing their date of birth. Further, a categorical variable was created in which the participants were asked to tick the response representing their highest level of education (primary, secondary, or further/higher education).

Translation Procedures

The scales were translated from English to the relevant languages by researchers within the research team in each participating country. Standardized back-translation procedures were used to develop the different language versions of the study measures using two independent bi-lingual translators for each language (Brislin, 1986). The back-translation procedure was repeated iteratively until the original and back-translated English versions of the questionnaires were virtually identical.

Procedure

Approval to conduct the study was obtained from the respective Ethics Committees of the universities involved in the study. The data was collected during the spring of 2008. Initially, the coordinator for each participating country drew up a list of places in the community they believed, based on experience, older adults would frequent (e.g., social clubs for older adults, community centres, libraries, supermarkets, cafés and post offices). The list differed slightly across the participating countries as it was acknowledged that the list should be culturally sensitive (for example, social clubs for older adults are common in Greece and Finland only). The investigators also made use of personal contacts they had from previous research conducted using older adults. Based on the list constructed, trained research assistants (RA) in each participating country sought out at least five different sites from each location identified over two weeks between 10 a.m. and 2 p.m. and approached older adults in

person. The RA introduced him/herself and explained the nature of the study. (S)he checked that each person approached fulfilled the inclusion criteria and only then asked them for their willingness to complete a questionnaire. All the participants provided written informed consent prior to taking part in the study. A small table was available for participants to use when completing the questionnaire and the completion was supervised by the RA. Thus, the participants had opportunities to ask questions. The ethical guidelines of psychological societies in each of the countries (similar to those produced by the British Psychological Society) were adhered to throughout. The completion of the questionnaires lasted approximately 20 minutes.

Data analysis

AMOS 18.0 (Arbuckle, 1995-2009) was used to analyze the data with the maximum likelihood (ML) estimator. The Full Information Maximum Likelihood estimation (FIML) was used to handle missing data. The following fit indices were used: (a) chi-square statistics; (b) Bentlers Comparative Fit Index (CFI; (Bentler, 1990); (c) the Root Mean Square of Approximation. (RMSEA;(Browne & Cudeck, 1993). In addition to these indices, the Akaike´s Information Criterion (AIC) was also used to allow for comparison between models that are not nested. For CFI, values close to .95 or greater indicates a well fitting model (Hu & Bentler, 1999). For RMSEA, values less than .05 indicate a good fit, whereas values up to .08 represent a reasonable fit (Browne & Cudeck, 1993). For the AIC, lower values represent a better fitting model

We tested how well our data fit eight different models that have been highlighted in previous studies (Marsh et al., 2010; Quilty et al., 2006). These eight models are described in Figure 1. Model 1 hypothesized one global self-esteem factor. Model 2 posited two oblique factors, one positive and one negative self-esteem factor. Model 3 included correlated

uniqueness (errors) between negatively worded items, whereas Model 4 included correlated uniqueness between positively worded items. Model 5 posited correlated uniqueness between both positively and negatively worded items. Models 3 to 5 are examples of CTCU models, and models 6-8 are CTCM models. Model 6 hypothesized a substantive self-esteem factor along with a method factor, in this model for negatively worded items. Model 7 was the same as model 6, except that the method factor included positively worded items. Finally, Model 8 included one self-esteem factor and both the method effect factors were included in models 6 and 7. Therefore, this model hypothesized one substantive self-esteem factor along with two method factors.

When examining the invariance of the best-fitting models across age and gender, we conducted multi-group invariance testing procedures according to recommendations of Byrne (2010). Consequently, we started with a baseline-model for all groups with no constraints. The fit of this model was then compared with the fit of models with increasing constraints. More specifically, based on the framework of Vandenberg and Lance (2000), we tested for: (a) configural invariance (same number of dimensions), that is, if the baseline models including all groups made acceptable fit to the data; (b) metric invariance (equal factor loadings); (c) equal residual covariances (correlated uniqueness); (d) scalar invariance (equal item intercepts). As we were interested in latent means differences, we also examined latent means differences (if the assumption of scalar invariance hold) by setting the latent means of the three factors to zero in one group (Byrne, 2010). If configural invariance does not hold, it means that the baseline model, in terms of patterns of free or fixed parameters, does not fit data equally well across groups. Lack of support for metric invariance suggests that the manifest variables (e.g., RSES items) fail to measure the same latent factor (e.g., self-esteem) in the same way (Meredith & Teresi, 2006). For example, some items may better mirror its latent factor in one group (e.g., men) than another (e.g., women). Finally, failure to find

support for scalar invariance essentially indicates that observed group differences in the factor means do not correspond to actual differences in factor means but are confounded by item-specific intercepts. In other words, given the same latent factor mean, different groups should have similar patterns of item-specific responses.

Interpretation of the invariance of the models was based on a non-significant drop in chi-square, taking differences in degrees of freedom into account, compared with the baseline model. However, as research also recommend interpreting a decline in CFI measure of less than .01 as an indication of invariance (Cheung & Rensvold, 2002), we also used this as a basis for our decision.

RESULTS

Due to incomplete answers, 165 participants were deleted, leaving a total of 1012 participants available for analyses. We used Mahalanobi's distances to identify and delete 62 multivariate outliers ($p<.001$), leaving 950 participants for further analyses. There were differences across countries in terms of age, self-esteem scores and distribution of gender and education[1]. Descriptive statistics for the ten RSES items are shown in Table 1. All items were normally distributed. However, the multivariate normality value and its critical ratio were 20.68 and 20.57 respectively, indicating nonnormality in the sample (Byrne, 2010). We therefore used maximum likelihood (ML) as the estimator but also ran all the analyses in Mplus using the robust ML estimator (MLR) and compared the results. As the results with the robust estimator did not differ substantially in terms of fit-indices and in particular in terms of which models fitted data best, we report only the results from AMOS and the ML estimator.

Fit of models

Fit indexes for the eight models for the whole sample are demonstrated in Table 2. Model 1 (positing a single factor) did not fit data well. Nor did Model 2, with one positive and

negative self-esteem factor fitting the data. Looking at the CTCU models (models 3-5), model 3 (correlated uniqueness among negative worded items), did not fit the data adequately (CFI<.95 and RMSEA>.08). However, the fit indexes for model 4 (correlated uniqueness for positively worded items) indicated a reasonable to good fit to the data. Hence, stronger support for the method effect of positively worded items (model 4), compared with negatively worded items (model 3) was found. Model 5, including correlated uniqueness among both positively and negatively worded items, showed overall the best fit to the data. Of the CTCM models (models 6-8) only model 8 which included two method factors along with a substantive self-esteem-factor made reasonable fit to the data. Standardized factor loadings in the two best fitting models (Models 5 and 8) are described in Table 3.

Examining more closely the nature of the method effects, we looked at the proposed correlated uniqueness between negatively and positively phrased items in model 5. All ten correlated uniqueness between positively worded items were significant (ps<.001), whereas only one of the 9 (the correlated uniqueness between item 3 and 8 was set to 0[2]) correlated uniqueness between negatively worded items was significant (between items 5 and 8). This indicates stronger support for the method effect associated primarily with positively worded items. In CTCM models, the factor loadings (see Table 3) of all negatively worded items with its method factor were significant. For the positively worded items, three out of five items loaded significantly on its method factor. Therefore, these results provide no clear cut support for neither positively nor negatively worded method effects.

Invariance testing

As models 5 and 8 were the best fitting models for the full sample, we chose to examine if these models were invariant across gender and age. Moreover, we tested latent differences between groups in the method factors and the substantive self-esteem factor in model 8. For

model 5, factor loadings and correlated uniqueness were invariant across males and females[3]. Although the model displayed a significant decrement in fit for the equal correlated uniqueness model, the CFI value did not decline more than .01. The same results were found for age groups; although the fit of the models constraining the factor loadings and correlated uniqueness dropped significantly compared with the baseline model, the decline in CFI was less than .01. To summarize, we found that factor loadings and correlated uniqueness were invariant across gender and age for model 5.

For model 8, factor loadings and item intercepts were invariant between males and females, displaying metric and scalar invariance. No significant differences were found in terms of latent means for the factor of the positively worded item method effect (PME) or the factor of the negatively worded item method effect (NME) between men and women. Moreover, the results supported equal factor variances for all three factors, indicating that men and women did not differ in range of scores in the latent factors. However, men had higher estimated latent self-esteem scores (mean estimate.080, z =2.95, p<.01) than women. We divided the difference in latent mean with the pooled standard deviation for men and women to compute a measure of Cohen's (Cohen, 1988) *d* effect size (d). The d value was .18, indicating a small difference using Cohen's guidelines.

For the invariance analyses, we divided the sample into two subgroups of younger (60-73, n=473) and older (74 and older, n=447) adults by mean split (mean 73.50). The model with equal factor loadings across age groups demonstrated a significant decline in chi square, but less than .01 in terms of CFI and was interpreted as invariant. Item intercepts however were not invariant, as the decline in CFI exceeded .01. The largest difference in item intercept was found on items 4 and 10. When the equality constraints on these two intercepts were released, the CFI decline was less than .01, supporting partial invariance (Byrne, Shavelson, & Muthen, 1989) of the item intercepts. Therefore we proceeded and examined

differences in latent means. The difference in latent mean of the self-esteem factor did not differ between age-groups. However, the younger group had higher estimated latent means on the NSE, .096, z =3.23, p<.01, Cohen's d=.17.

Model 5 was not invariant across countries. Although the baseline model did fit data well, indicating that Model 5 generally fit data well in all four[4] countries, both the chi-square value and the CFI dropped considerably in models constraining factor loadings and then correlated uniqueness to be invariant across countries. More specifically, the pattern of correlated uniqueness seemed to differ across countries. In the British and Finnish samples, almost all (9 out of 10) of the correlated uniqueness among the positively worded items were significant whereas only 1 out of 9 of the correlated uniqueness among negatively worded items were significant. In the Greek and Italian samples, however, the opposite trend was found; the majority of the correlated uniqueness for positively keyed items was not significant whereas the correlated uniqueness for the negatively keyed items was significant. Also Model 8 was not invariant across countries. The baseline model with no constraints displayed an adequate fit with the data, but in subsequent models, in which first factor loadings and then item intercepts were constrained to be equal, chi-square and CFI dropped considerably indicating lack of invariance.

Prediction of method effects

We created sum scores of life-satisfaction and depression, based on the five items in the Satisfaction with Life Scale and the 20 items in the CES-D scale, and used these two variables as predictors of latent means of the two method factors, NME and PME, in model 8 (see Table 4). For the full sample, this model displayed an acceptable fit to the data (χ^2 =159.95; df=38; p<0.01; CFI=.969; RMSEA=.058. Depression was negatively related to self-esteem (.-23) whereas life-satisfaction was positively associated with self-esteem (.36). Moreover, there was a weak negative relation between life-satisfaction and the PME factor (-.142). Given the

reversed coding of the RSES (1= strongly agree, 4= strongly disagree), this means that individuals with higher life-satisfaction scores are more likely to endorse a positively worded item. A higher life-satisfaction score was also weakly related to NSE (.112), showing that individuals with higher life-satisfaction scores are more likely to score high on the negatively worded items and therefore less likely to endorse these items. The strongest relation between the method factors and the predictors was found between NSE and depression (-.319), demonstrating that individuals with higher depression scores are more likely to score lower on the negatively phrased items and thus more likely to endorse these items.

DISCUSSION

This study was the first to examine method effects associated with the RSES in a sample of older people. The results show clear support for the existence of method effects in the RSES in older adults. Only models including method effects made acceptable fit with the data. Thus, the results are in line with a number of previous studies that have found support for method effects in the RSES in samples of adolescents and younger adults (Corwyn, 2000; DiStefano & Motl, 2009; Horan et al., 2003; Marsh, 1986; Marsh, et al.,2010; Tomas & Olivier, 1999; Quilty et al., 2009; Wu, 2008).

In the present study we also found that these method effects were associated with both positively and negatively worded items. A liberal interpretation of the results would point to stronger support for the method effect of positively phrased items, based on the closer fit of the CTCU model that included only positively phrased items (model 4) compared with the CTCU model with only negatively phrased items (model 3). All correlated uniqueness between positively worded items were significant whereas only one of the 9 correlated uniqueness between negatively worded items was significant. On the other hand, the CTCM models including only negatively or positively phrased items (models 6 and 7) did not differ in terms of fit with the data and both the CCTU and CTCM models that included both types

of method effects simultaneously provided the best fit with data. Thus, a more conservative interpretation would be that our results primarily are in line with previous studies (Marsh et al., 2010; Quilty et al., 2006; Wu, 2008) which support the notion that the RSES contains method effects both from positively and negatively phrased items. This is in contrast with the majority of previous studies with younger samples that have found strongest support for the method effects of negatively worded items (Corwyn, 2000; Horan et al., 2003; Marsh, 1996; Quilty et al., 2009; Tomas & Olivier, 1999).

The methods effects of negatively phrased items in instruments, such as the RSES, may seem more intuitive and be easier to interpret and understand. For example, it has been argued that the effects of negatively worded items may be the result of a process by which respondents first establish a pattern of responding and then fail to attend to the positive-negative wording of items (N. Schmitt & Stults, 1986). The method effect of positively phrased items may be harder to explain, particularly the potential interpretation of the results in the present study that the method effects actually may be stronger for positively worded items. The socio-emotional selectivity theory (SST; Carstensen et al., 1999) and the positivity effect (Carstensen & Mikels, 2005) may however afford relevant perspectives. As documented in a number of studies (for a review, see Carstensen & Mikels, 2005) older adults show stronger preference for positive information whereas for younger people negative information seems to be more salient. Thus, this perspective fits well with the stronger support for positively worded method effects in our sample of older adults and may explain, at least partially, why older adults appear to be more vulnerable to method effects of positively worded items. Thus, from a broader information-processing perspective, the greater focus on emotional regulation to optimize well-being and positive mood for older adults may result in a greater tendency to endorse positively worded items in self-report instruments, such as the RSES. Also, global evaluations of self-esteem (e.g., “On the whole I am satisfied with

myself”) may be more vague and cumbersome compared to evaluations of more specific components of the self (e.g., “I am good at most sports”). It has been proposed (Marsh & Yeung, 1999) that global evaluations, such as those made when completing the RSES, may be more influenced by immediate experiences and mood. The fact that global self-esteem has been found to be less stable across time than specific components of the self, support this view. Hence, the theory of Marsh and Yeung (1999) pertaining to the impact of immediate experiences and mood when evaluating global self-esteem may help to understand how the SET and positivity effect may render older adults more vulnerable to method effects of positively worded items in self-report instruments, such as the RSES.

Supporting previous research (Distefano & Motl, 2009), we found that the models were invariant across gender and that the estimated latent method effects means did not differ between men and women. However, the younger participant group (less than 74 years of age) had higher estimated latent factor scores on the negatively worded method effect factor compared with the older sub-group. This result may further strengthen the aforementioned reasoning linking the SET and positivity effect to the potential increased tendency to method effects of positively rather than negatively worded items as people get older and perceive time as more constrained. This would reflect a strongerfocus on optimizing mood and positive affect. Future studies should, however, test this hypothesis further.

The invariance analyses further demonstrated that neither of the two best fitting models that included method effects were invariant across countries. These results are in line with the findings of Schmitt and Allik (D. P. Schmitt & Allik, 2005) suggesting that method effects may display different patterns across countries as positively and negatively keyed items in the RSES in different cultures and languages may be interpreted differently. There may be several reasons to why the method effects may come out differently in different countries. As suggested in previous work (D. P. Schmitt & Allik, 2005) language may be one of the

moderating factors of the method effects associated with positively or negatively worded items. However, there were also marked differences in the present study across countries and samples in terms of age, gender distribution, level of education and self-esteem scores. These differences may also have contributed to the lack of invariance in patterns of method effects. In particular level of education and literacy has in previous work been linked to method effects, albeit in adolescents (Marsh, 1996). However, the results of the present study and the study by Schmitt and Allik suggest that researchers should be careful to directly compare RSES scores across countries unless the potential method effects have been taken into account.

A highly relevant question, both from a theoretical as well as practical viewpoint, is if these method effects may be predicted and if so, by which variables. Previous studies have shown that individuals with higher avoidance motivation and neuroticism are more likely to endorse negatively worded items (Quilty et al., 2006). In contrast, individuals who are apprehensive of negative evaluations and more self-consciousness are less likely to show presence of method effects associated with negatively worded items (Distefano & Motl, 2006). Our study adds to these results by demonstrating that older adults with higher depression scores on the CES-D are more likely to endorse negatively worded items and that older adults with higher life-satisfaction scores are less likely to endorse negatively worded items while more likely to endorse positively worded items. Thus, our results support the presumption (e.g., Distefano & Motl, 2006; Horan et al., 2003) that method effects may reflect a response style rather than an artifact and that it may be predicted by other psychological variables and concepts.

What are the implications of these method effects for researchers using the RSES or other scales with both positively and negatively worded items with older adults? As highlighted by several scholars (Bagozzi, 1993; Brown, 2006; Podsakoff et al., 2003), method

effects may cause a number of different problems that may have differential impact on the interpretation of the results. In the present study, depression was negatively related to self-esteem, but depression also predicted method effects associated with negatively worded items, as individuals with higher depression were more likely to endorse negatively worded items. Thus, the already low self-esteem score in individuals high in depression would be even more deflated due to the increased tendency also to endorse negatively worded items. A similar problem could have occurred also in previous studies that have used the RSES, but not separated potential method effects from the substantive factor of interest when examining relations between self-esteem and depression.

The potential problem of method effects associated with positive and negative items are likely to exist also in other instruments and rating scales aside from the RSES. Hence, this problem may have broad implications for researchers in social sciences. How then may researchers tackle this issue? First, a number of general recommendations regarding techniques and statistical remedies for controlling common method bias (of which method effects of positive and negative wording represents one of many sources), have been provided in a review paper by Podsakoff and collagues (2003). Also, several strategies to handle methods effects specifically associated with positively and negatively phrased items have been discussed in previously published papers (e.g., Marsh et al., 2010; Quilty et al., 2006). First, analytical approaches, such as CTCU or CTCM models to examine the potential existence of method effects should be adopted, particularly if instruments that involve a mixture of positively and negatively worded items are used. Perhaps the most important recommendation, at least for the RSES, is to include method effects factors into confirmatory factor analyses and use CTCM models to directly model the relation of the method effects and the substantive self-esteem factor on other variables of interest. Combining these CTCU and CTCM models with item response theory (Classen, Velozo, & Mann, 2007) could also be a

relevant strategy for gaining more complex information of the RSES items and its underlying structure.

However, these strategies may not be very realistic for many researchers or practitioners if they are not familiar with structural equation modeling or analyzing the types of models discussed in this paper. So what would be the best alternative? A less complex and thereby more user-friendly and realistic solution may be to use only positively worded items in the analyses, as the method effects in the majority of previous studies have been associated with negatively phrased items and as positively worded items generally have been found to be more accurate (Schriesheim & Hill, 1981). In particular, if the instrument is used in a younger population, this recommendation may be well balanced in terms of at least helping researchers partly address the problem without having to invest a lot of time and effort into new analytical methods. However, as method effects evidently have been associated also with positively phrased items (as in the present study), researchers should be aware that only excluding negatively items may not entirely solve the problem. In particular if researchers are working with older population groups, a general recommendation would be to conduct analyses separately with factors based on negatively and positively phrased items and comparing the results.

Limitations in the present study include the use of cross-sectional data, making longitudinal analyses of change or stability for the method effects impossible. Also, we used a convenience sample of older adults recruited from community centers, social clubs, and retirement unions that may not totally represent the general older population. On the other hand, one of the strengths of the study was the inclusion of a relatively large sample of older adults from five European countries, which enabled the examination of method effects in the RSES in five languages.

REFERENCES

American Education Research Association (1999). Standards for educational and psychological testing. Washington, DC: Author

Arbuckle, J. L. (1995-2009). *Amos 18 User's Guide*. Crawfordville, FL: Amos Development Corporation.

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality, 27*(1), 49-87.

Beekman, A. T. F., Deeg, D. J. H., Van Limbeek, J., Braam, A. W., & et al. (1997). Criterion validity of the Center for Epidemiologic Studies Depression scale (CES-D): Results from a community-based sample of older subjects in the Netherlands. *Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences, 27*(1), 231-235.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

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* (pp. 115–160). San Diego, CA, US: Academic Press.

Blazer, D. G. (2003). Depression in late life: Review and commentary. *The Journals of Gerontology: Series A: Biological Sciences and Medical Sciences, 58A*(3), 249-265.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Thousand Oaks, CA, US: Sage Publications, Inc.

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

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

Burke, M. J., Brief, A. P., & George, J. M. (1993). The role of negative affectivity in understanding relations between self-reports of stressors and strains: a comment on the applied psychology literature. *Journal of Applied Psychology, 78*(3), 402-412.

Byrne, B. M. (2010). Structural equation modeling with AMOS: Basic concepts, applications, and programming (2nd ed.). *Multivariate applications series. New York, NY, US: Routledge/Taylor & Francis Group, 2010*(396).

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

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

Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously: A theory of socioemotional selectivity. *American Psychologist, 54*(3), 165-181.

Carstensen, L. L., & Mikels, J. A. (2005). At the Intersection of Emotion and Cognition: Aging and the Positivity Effect. *Current Directions in Psychological Science, 14*(3), 117-121.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Classen, S., Velozo, C. A., & Mann, W. C. (2007). The Rosenberg Self-Esteem Scale as a measure of self-esteem for the noninstitutionalized elderly. *Clinical Gerontologist: The Journal of Aging and Mental Health, 31*(1), 77-93.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates Inc.

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*(4), 357-379.

Cote, J. A., & Buckley, R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24*, 315–318.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: a meta-analysis of 137 personality traits and subjective well-being. *Psychol Bull, 124*(2), 197-229.

Diehl, M., Hastings, C. T., & Stanton, J. M. (2001). Self-concept differentiation across the adult life span. *Psychology and Aging, 16*(4), 643-654.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49*(1), 71-75.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440-464.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 46*(3), 309-313.

Fiske, A., Wetherell, J. L., & Gatz, M. (2009). Depression in older adults. *Annual Review of Clinical Psychology, 5*, 363-389.

Goldsmith, R. E. (1986). Dimensionality of the Rosenberg Self-Esteem Scale. *Journal of Social Behavior & Personality, 1*(2), 253-264.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10*(3), 435-455.

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

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125*(4), 470-500.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810-819.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling* (pp. 177-198). Thousand Oaks, CA, US: Sage Publications, Inc.

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*(2), 366-381.

Marsh, H. W., & Yeung, A. S. (1999). The lability of psychological ratings: The chameleon effect in global self-esteem. *Personality and Social Psychology Bulletin, 25*(1), 49-64.

McAuley, E., Elavsky, S., Motl, R. W., Konopack, J. F., Hu, L., & Marquez, D. X. (2005). Physical Activity, Self-Efficacy, and Self-Esteem: Longitudinal Relationships in Older Adults. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences, 60B*(5), P268-P275.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11 Suppl 3), S69-77.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling, 13*(1), 99-117.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Reitzes, D. C., & Mutran, E. J. (2006). Self and Health: Factors That Encourage Self-Esteem and Functional Health. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences, 61B*(1), S44-S51.

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

Rosenberg, M. (1989). *Society and the adolescent self-image*. Middletown, CT: Wesleyan 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*(4), 623-642.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait–multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114.

Shiovitz-Ezra, S., Leitsch, S., Graber, J., & Karraker, A. (2009). Quality of life and psychological health indicators in the National Social Life, Health, and Aging Project. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences.*

Special Issue: The National Social Life, Health, and Aging Project, 64B(Suppl 1), I30-I37.

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*(2), 322-342.

Tomas, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling, 6*(1), 84-98.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Wu, C.-H. (2008). An examination of the wording effect in the Rosenberg Self-Esteem Scale among culturally Chinese people. *The Journal of Social Psychology, 148*(5), 535-551.

FOOTNOTES

[1] There were significant differences across countries in terms of age, $F(4,933)=6.52$, $p<.001$. Participants from UK were the oldest (M= 75.61; SD=7.71) and participants from Italy were the youngest (M=72.48; SD=7.71). There were also differences in terms of the gender distribution across countries, $\chi2$ (4, N=949) = 64.77, $p<.001$. There were a larger proportion of women in the British and Finnish samples (75%) compared with the Swedish (65%), Italian (60%) and Greek (42%) samples. Moreover, there were also differences in terms of highest level of education, $\chi2$ (8, N=932) = 287.99, $p<.001$. The largest proportion in the UK sample reported tertiary education as their highest level of education. In the Swedish and Finnish samples, however, the largest proportion reported secondary education, and in the Italian and Greek samples, most participants reported primary education as their highest level. Finally, there were significant differences in self-esteem scores across countries, $F(4,933)=6.52$, $p<.001$. The British sample reported significantly (ps<.05) higher self-esteem scores (M: 2.55; SD=.22) than the Italian (M: 2.48; SD=.26), Swedish (M: 2.46; SD=.26), Finnish (M: 2.30; SD=.25) and Greek (M: 2.25; SD=.28) samples, and the Greek and Finnish samples reported significantly (ps<.001) lower scores than the other countries.

[2] We found model 5 to be empirically underidentified. Hence, similar to the procedure in previous studies (e.g., Tomas & Olivier, 1999) we constrained the correlated uniqueness between the errors of items 3 and 8 to 0 as this correlation was non-significant

[3] The complete results from all the invariance analyses are available from the first author upon request.

[4] As the Swedish sample only comprised of 47 participants, we choose not to include this sample in the invariance analyses across countries. Thus, the invariance analyses across countries included the four samples from Finland, Italy, Greece and UK.

TABLE 1. Descriptive statistics (mean, standard deviation, skewness, and kurtosis) for the ten items in the Rosenberg Self-Esteem Scale (n=950)

Rosenberg Self-Esteem Scale items	M	SD	Skewness	Kurtosis
Positively phrased items				
1. I feel that I'm a person of worth, at least on an equal plane with others	3.38	.62	-.70	.67
2. I feel that I have a number of good qualities	3.33	.56	-.33	.73
4. I am able to do things as well as most other people	3.23	.70	-.71	.56
6. I take a positive attitude toward myself	3.22	.65	-.60	.78
7. On the whole, I am satisfied with myself	3.16	.67	-.55	.56
Negatively phrased items				
3. All in all, I am inclined to feel that I am a failure	3.45	.61	-.72	-.02
5. I feel I do not have much to be proud of	3.14	.83	-.74	-.04
8. I wish I could have more respect for myself	2.69	.92	-.08	-.90
9. I certainly feel useless at times	3.00	.91	-.46	-.77
10. At times I think I am no good at all	3.37	.75	-1.03	.50

Note: Scores ranging from 1-4 for all items. Higher values indicate higher self-esteem

TABLE 2. Model fit-indices for the different models in the full sample (n=950)

	χ^2	df	CFI	RMSEA (90% CI)	AIC
Model 1	701.85	35	.803	.142 (.133-.151)	741.85
Model 2	347.01	34	.907	.098 (.089-108)	389.01
Model 3	238.69	25	.937	.095 (.084-.106)	298.69
Model 4	162.09	25	.959	.076 (.065-.087)	222.09
Model 5	44.50	16	.992	.043(.028- .059)	122.49
Model 6	308.06	30	.918	.099 (.089-.109)	358.06
Model 7	320.83	30	.914	.101 (.091-.111)	370.83
Model 8	137.23	24	.967	.071 (.059-.082)	199.23

Model 1: single factor; Model 2; two factors (positive & negative); Model 3; correlated uniqueness, negative items; Model 4; correlated uniqueness, positive items; Model 5; correlated uniqueness, both positive and negative items; Model 6, two factors, global + negative method effect; Model 7, two factors, global + positive method effect; Model 8, three factors, global + both negative and positive method effects.

TABLE 3. Standardized factor loadings in the best fitting models (Models 5 and 8).

Items	Model 5	Model 8		
	GSE	GSE	PME	NME
Pos1	.42	.60	.57	
Pos2	.38	.55	.44	
Pos4	.51	.63	.19	
Pos6	.47	.75	.01 *ns*	
Pos7	.59	.89	-.14 *ns*	
Neg3	.82	.55		.33
Neg5	.45	.31		.35
Neg8	.28	.22		.29
Neg9	.77	.51		.59
Neg10	.76	.51		.65

Note: All factor loadings except the ones with *ns* is statistically significant at $p<.001$

TABLE 4. Standardized regression weights between predictors (depression and life-satisfaction) and RSES latent factors.

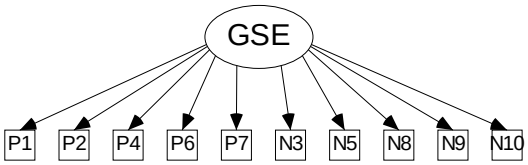
	SE	PME	NME
Depression[a]	-.231**	.093	-.319**
Life-Satisfaction[b]	.359**	-.142*	.112*

Note : *:p<.05; **:p<.01: SE=Global Self-Esteem; PME= method effect factor for positively worded items; NME= method effect factor for negatively worded items;

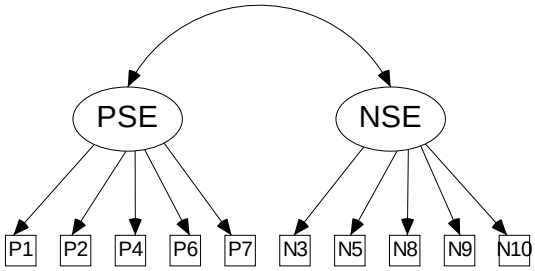
[a]Measured with 20 item CES-D (Radloff, 1977); [b]Measured with 5 item from Satisfaction with Life Scale (Diener et al., 1985)

FIGURE CAPTIONS

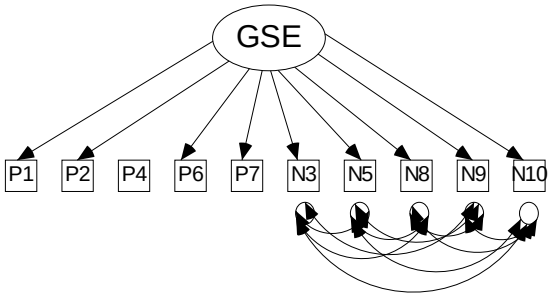
Figure 1. The eight factor structure models of the Rosenberg Self-Esteem Scale tested in the study.


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10

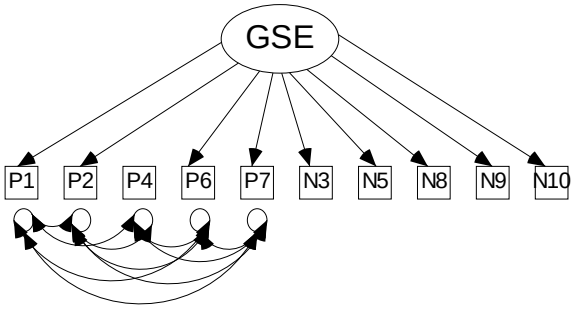
(1)


PSE
NSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10

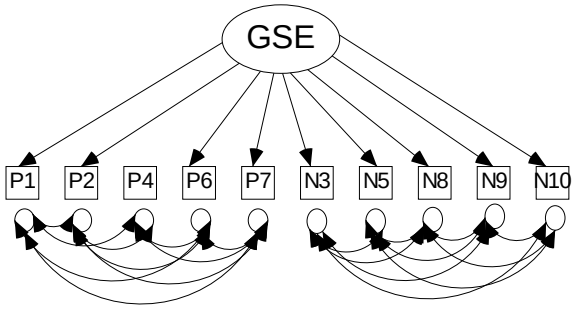
(2)


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10

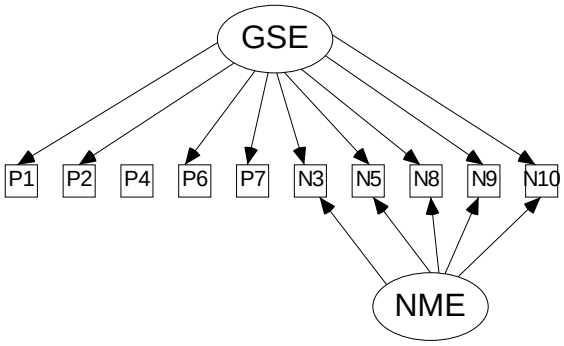
(3)


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10

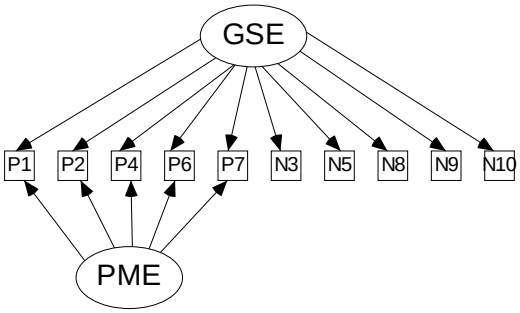
(4)


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10

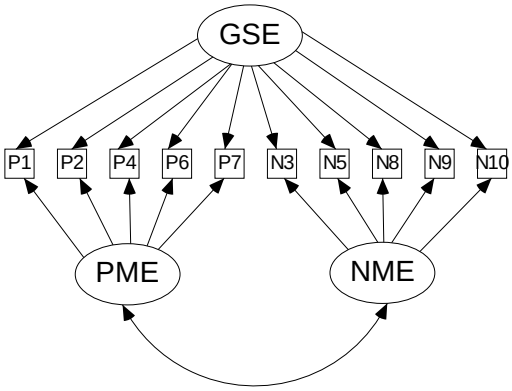
(5)


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10
NME

(6)


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10
PME

(7)


GSE
P1
P2
P4
P6
P7
N3
N5
N8
N9
N10
PME
NME

(8)