Personality and Individual Differences 54 (2013) 153–157

Contents lists available at SciVerse ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Evaluation of the factor structure of the Rosenberg Self-Esteem Scale in older adults

Sean P. Mullen *, Neha P. Gothe, Edward McAuley

University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Article history:
Received 12 May 2012
Received in revised form 26 July 2012
Accepted 11 August 2012
Available online 16 September 2012

Keywords:
Global self-esteem
Psychometrics
Aging

ABSTRACT

The Rosenberg Self-Esteem Scale is the most utilized measure of global self-esteem. Although psychometric studies have generally supported the uni-dimensionality of this 10-item scale, more recently, a stable, response-bias has been associated with the wording of the items (Marsh, Scalas, & Nagengast, 2010). The purpose of this report was to replicate Marsh et al.'s findings in a sample of older adults and to test for invariance across time, gender and levels of education. Our results indicated that indeed a response-bias does exist in esteem responses. Researchers should investigate ways to meaningfully examine and practically overcome the methodological challenges associated with the RSE scale.

1. Introduction

Self-esteem has been an integral construct in the field of psychology for decades. The field's most commonly used measure of global self-esteem is the 10-item Rosenberg's Self-Esteem (RSE) scale (Rosenberg, 1965). This scale has been used extensively with samples of all ages, from adolescents to older adults. Although the psychometric properties of this self-report measure have been rigorously tested, researchers have questioned whether the positively and negatively worded items are interchangeable, i.e., assess the same construct (Corwyn, 2000; DiStefano & Motl, 2009; Marsh, 1996; Motl & DiStefano, 2002; Wang, Siegal, Falck, & Carlson, 2001). Few researchers understand the origins of method effects, however there is a strong case for controlling consistent bias associated with the RSE scale, despite criticism of the general post hoc approach to account for "common method variance" (Conway & Lance, 2010; Lance, Dawson, Birkelbach, & Hoffman, 2010; Richardson, Simmering, & Sturman, 2009). Overall, researchers agree that methodological decisions should be guided by substantive and theoretical arguments.

Recently, Marsh, Scalas, and Nagengast (2010) systematically tested multiple models of global self-esteem based on the RSE scale to determine the extent to which method effects were ephemeral or stable. They used two approaches, correlated uniquenesses and latent method factors, to account for the hypothesized method effects. Each approach has strengths and weaknesses. Most notably, the correlated uniquenesses approach assumes different types of method bias are uncorrelated with each other, and the latent method approach relaxes this assumption. However, the latter solutions are prone to producing inadmissible solutions and do not converge as easily as the former approaches. Overall, Marsh et al. found a consistent response-style bias associated with the item wording of the RSE. Furthermore, Marsh et al. claimed these method effects call into question the vast literature based on the RSE and that "failure to control for them will bias the interpretations of RSE responses" (p. 378). However, one of the limitations of their study was that their findings were based solely on the responses of adolescent males.

To our knowledge, only one study (Whiteside-Mansell & Corwyn, 2003) has tested the possibility of differences in global self-esteem measurement across age. However, Whiteside-Mansell and Corwyn only tested the invariance of the RSE scale's structure across a sample of 12–17 years old ($M_{age} = 14.8$) and a sample of 18–80 years old ($M_{age} = 33$). Thus, the "adult" sample contains young, middle, and older adults, and it cannot be determined whether the esteem construct has the same meaning for all age groups. It is reasonable to expect that older adults may interpret and respond to questionnaires differently, based on established age differences in "affective balance" during emotional self-report (Robinson & Clore, 2002) and memory for positive and negative events (Mather & Carstensen, 2005). Furthermore, no studies have explored potential individual differences in the interpretation of RSE, across gender or levels of education among older adults. The purpose of this study was to replicate Marsh et al.'s (2010) findings in a homogenous sample of older adults and to extend this work by exploring potential differences across subgroups. Given that there are considerable implications for older adults' self-esteem, it is important to verify the structural integrity of the scale in older populations.

* Corresponding author. Address: University of Illinois at Urbana Champaign, Department of Kinesiology and Community Health, 906 S. Goodwin Ave., Urbana, IL 61801, United States. Tel.: +1 (217) 244 4502; fax: +1 (217) 333 3124.

E-mail address: spmullen@illinois.edu (S.P. Mullen).

0191-8869/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.paid.2012.08.009

2. Method

2.1. Participants, procedure, and measures

Data were collected from sedentary older adults ($n = 603$) as part of a baseline questionnaire packet prior to participation in an exercise program; a smaller subsample ($n = 298$) completed the questionnaire packet 12 months later. The sample were community-dwelling older adults ($M_{age} = 69.94$, $SD = 5.66$; range = 60–95), mostly white (94.7%, vs. 3.7% Black/African–American, 1.3% Asian, .3% American Indian/Alaskan Native; .3% missing), female (72.5%, $n = 437$), married (59.4%), and earned an income of at least 40 K (54.1%). Approximately half of the sample graduated from college or attained higher education (47.3%). Participants completed the original 10-item RSE (Rosenberg, 1965) along with demographic information. The responses on the scale were measured on a 5-point Likert scale: 1 (strongly agree), 2 (agree), 3 (neutral), 4 (disagree), and 5 (strongly disagree). Five of the items are positively-worded (items 1, 2, 4, 6, and 7) whereas the remaining five are negatively-worded (3, 5, 8, 9, and 10); negative items were reverse-coded prior to data analysis.

All modeling was conducted using raw data with version 6.1 of *Mplus* (Muthén & Muthén, 1998–2012) and we used full information for missing data with robust maximum likelihood estimation (MLR). We elected to use multiple criteria for evaluating model misspecification, including the chi-square statistic (χ^2), root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker–Lewis Index (TLI). Accordingly, it is recommended that models should result in non-significant χ^2 values ($p \geqslant .05$), RMSEA of <.06, CFI values of $\geqslant .95$ (Hu & Bentler, 1999), and TLI values of $\geqslant .95$ (Marsh, Hau, & Grayson, 2005). Paralleling the procedures used by Marsh et al. (2010), we tested eight models, including a 1-factor structure with no additional parameter constraints (Model 1), a model involving a 2-factor latent structure, i.e., positive, negative (Model 2), a 1-factor model with correlated residuals among negative and among positive items (Model 3), a 1-factor model with only correlated negative residuals (Model 4), a 1-factor model with correlated positive residuals (Model 5), a 1-factor model with two method factors, i.e., representing systematic error in responses to positive and negatively-worded items (Model 6), a 1-factor model with only a negative method effect (Model 7), and a 1-factor with only a positive method effect (Model 8). Note that latent method factors were specified to covary a priori. Lastly, we also conducted invariance testing with the best-fitting models. The invariance routine we employed involved adding sequential restrictions to test equality of factor configurations (i.e., configural invariance), followed by loadings (i.e., metric invariance), intercepts (i.e., scalar invariance), residuals (i.e., strict invariance), and latent means and variances across measurement occasions. More restrictive models were deemed invariant from less restrictive models if the corrected Satorra–Bentler (S–B) χ^2 difference (Δ) test (Satorra & Bentler, 2001) was not significant ($p > .05$), ΔCFI < .01 (Cheung & Rensvold, 2002) and ΔRMSEA < .015 (Chen, 2007).

3. Results

3.1. Preliminary analyses

The majority of items were found to be negatively skewed. In addition, the variability in item 2 (i.e., "I feel that I have a number of good qualities") was restricted to just three responses (range = 1–3) at the second measurement occasion. We therefore employed MLR estimation.

3.2. Measurement models

3.2.1. Overall sample

Model 1 showed a poor fit to the data (e.g., RMSEA = .185, CFI = .758) and was the poorest fitting model tested. Model 2 was an improvement, but still failed to fit the data according to χ^2 and recommended cutoff values for RMSEA and TLI. Model 3 failed to converge, as was the case in some of the samples reported by Marsh et al. (2010). Similar to Model 2, Model 4 did not fit according to χ^2, RMSEA, and TLI. However, Model 5 (χ^2 p value >.05; RMSEA = .011, CFI = .999, TLI = .999) and Model 6 provided an excellent fit (χ^2 p value >.05, RMSEA = .023, CFI = .997, TLI = .995). For Model 5, overall factor loadings ranged from .024 to .776) and the correlated uniquenesses among positive items ranged from .790 to .960, indicating a method effect. For Model 6, factors loadings for the overall model were low and not significant (range = .053–.593), whereas significant loadings were found among positive items (range = .748–.977) and negative items (range = .535–.778), suggesting two method effects. Again, Model 7 failed to meet all recommended criteria (e.g., TLI = .919, RMSEA = .094), as did Model 8 (e.g., TLI = .917, RMSEA = .096). In sum, the best-fitting representations of RSE's factor structure were Models 5, reflecting correlations among positively worded items, and Model 6, reflecting two underlying uncorrelated method effects. Fit indices for all measurement models are included in Table 1, whereas conceptual diagrams of the best-fitting models, i.e., Model 5 and 6 are displayed in Fig. 1. See Table 2 for factor loadings and item uniquenesses associated with the best-fitting models.

In addition to examining cross-sectional measurement models, we attempted to replicate Marsh et al.'s findings showing temporal invariance which would indicate a stable response-style bias. Following conventional procedures, we first tested configural invariance (i.e., same items regressed on same constructs) for models 5 and 6. Model 5 provided an adequate fit to the model ($\chi^2 = 248.059$ (139), $p < .001$, RMSEA = .051, CFI = .940, TLI = .918), although the fit significantly worsened from the baseline model. Model 6 did not provide an admissible solution at this stage and invariance testing was terminated. For Model 5, relatively speaking, the metric invariance model, i.e., loadings constrained to equality across time ($\chi^2 = 270.618$ (148), $p < .001$, RMSEA = .053, CFI = .933, TLI = .913) did not significantly differ from the configural invariance model based on S–B χ^2 test, ΔCFI or ΔRMSEA. The scalar invariance model, i.e., intercepts constrained to be equal across time ($\chi^2 = 282.557$ (157), $p < .001$, RMSEA = .052, CFI = .931, TLI = .916) also did not significantly change the fit. The same was also true for the strict invariance model, i.e., residual variances and correlated uniquenesses (method effect) constrained to be equal across time ($\chi^2 = 314.968$ (172), $p < .001$, RMSEA = .053, CFI = .921, TLI = .913), latent mean invariance ($\chi^2 = 316.767$ (173), $p < .001$, RMSEA = .053, CFI = .921, TLI = .913) and latent variance invariance ($\chi^2 = 318.998$ (174), $p < .001$, RMSEA = .053, CFI = .920,

Table 1
Fit indices for all measurement models based on the entire sample ($N = 603$).

Model	χ^2	*df*	*p* Value	CFI	TLI	RMSEA (90% Interval)
1	756.283	35	<.001	.758	.689	.185 (.174–.196)
2	200.089	34	<.001	.944	.926	.090 (.078–.102)
4	179.140	25	<.001	.948	.907	.101 (.087–.115)
5	26.803	25	.368	.999	.999	.011 (.000–.035)
6	33.237	25	.125	.997	.995	.023 (.000–.043)
7	190.278	30	<.001	.946	.919	.094 (.082–.107)
8	195.239	30	<.001	.945	.917	.096 (.083–.109)

Note: Model 3 did not successfully converge.