Journal of Personality and Social Psychology
2016, Vol. 110, No. 1, e16–e22

0022-3514/16/$12.00 http://dx.doi.org/10.1037/pspp0000037

Low Self-Esteem Prospectively Predicts Depression in the Transition to Young Adulthood: A Replication of Orth, Robins, and Roberts (2008)

Sven Rieger, Richard Göllner, and Ulrich Trautwein
University of Tübingen

Brent W. Roberts
University of Illinois, Urbana-Champaign

The present study is a close replication of the work of Orth, Robins, and Roberts (2008). Orth et al. (2008) tested three theoretical models of the relation between self-esteem and depression—the vulnerability model, the scar model, and the common factor model—using longitudinal, cross-lagged panel designs. The authors concluded that depression and self-esteem were not the same construct (contrary to the common-factor model), and furthermore, the results were clearly in line with the vulnerability model and not with the scar model (low self-esteem predicts subsequent levels of depression and not vice versa). In addition, the results held for both men and women. To conduct a very close replication of the work of Orth et al. (2008), we used data from another large longitudinal study (N = 2,512), which is highly similar in study design and that contains the same measures (self-esteem and depression). The present study replicated the results of the Orth et al. (2008) study in a notable manner, in regard to the comparability of the coefficients, and therefore, corroborates the vulnerability model (and not the scar- or the common-factor model).

Keywords: replication, self-esteem, depression, young adulthood

Does low self-esteem lead to depression (the vulnerability model) or does depression lead to low self-esteem (the scar model)? This question is of considerable theoretical interest, but also has important real-world implications. Perhaps the most influential article published in the *Journal of Personality and Social Psychology* (JPSP) that addressed this question is the study by Orth, Robins, and Roberts (2008; cited 209 times on Google Scholar to date). Orth et al. (2008) used two large longitudinal data sets (both with four repeated assessments) and examined the relationship between low self-esteem and depression in adolescence and young adulthood (age ranges from 15 to 21 and 18 to 21). In light of the policy change at JPSP to publish more replications (Smith, Simpson, & King, 2014), combined with the importance of the topic and the influence that the Orth et al. (2008) article has had to date, we sought to replicate the core findings of the study as closely as one can replicate a longitudinal study.

This article was published Online First April 27, 2015.

Sven Rieger, Richard Göllner, and Ulrich Trautwein, Hector Research Institute of Education Sciences and Psychology, University of Tübingen; Brent W. Roberts, Department of Psychology, University of Illinois, Urbana–Champaign.

Sven Rieger is a doctoral student of the LEAD Graduate School [GSC1028], funded by the Excellence Initiative of the German federal and state governments.

This work was supported by a grant from the Ministry of Science, Research and the Arts of the state of Baden-Württemberg (Az: 33-7532.20/735) to Ulrich Trautwein.

Correspondence concerning this article should be addressed to Sven Rieger, Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastr. 6, 72072 Tübingen, Germany. E-mail: sven.rieger@uni-tuebingen.de

The Orth et al. (2008) article tested three theoretical models of the relation between self-esteem and depression using longitudinal, cross-lagged panel designs: the vulnerability model, the scar model, and the common factor model. The vulnerability model states that low self-esteem is a risk factor for future depression (e.g., Beck, 1967; Metalsky, Joiner, Hardin, & Abramson, 1993). According to the cognitive theory of depression by Beck (1967), negative self-view—which can be understood as low self-esteem (Beck, Steer, Epstein, & Brown, 1990)—is a diathesis exerting causal influence in the onset and maintenance of depression (Beck, 1967; Butler, Hokanson, & Flynn, 1994). In contrast, the assumption underlying the scar model is that low self-esteem might be a consequence of depression rather than a causal factor (Joiner, 2000; Lewinsohn, Steinmetz, Larson, & Franklin, 1981). Indeed, it is conceivable that the experience of depression decreases individuals' self-esteem because depressive symptoms are connected with impaired functioning and negative attitudes toward the self (Rohde, Lewinsohn, & Seeley, 1990; Shahar & Davidson, 2003). The common factor model (Watson, Suls, & Haig, 2002; also called continuum/spectrum model; see Klein, Kotov, & Bufferd, 2011) notes that there is a large proportion of shared variance between self-esteem and depression. Therefore, researchers should test whether self-esteem and depression can be accounted for by a single factor before they test models that dictate an interplay between these two constructs (Watson et al., 2002).

Orth et al. (2008) investigated these three models by means of structural equation models. The results of the Orth et al. (2008) study were clearly in line with the vulnerability model (low self-esteem predicts subsequent levels of depression and not vice versa), but not with the scar and common factor model. Furthermore, the results held for both men and women.

In recent years a number of longitudinal studies were published, which supported the main findings of Orth et al. (2008): the

vulnerability effect of low self-esteem on depression (see, e.g., Orth, Robins, & Meier, 2009; Orth, Robins, Trzesniewski, Maes, & Schmitt, 2009). However, these studies were actually not direct replications of Orth et al. (2008), but rather studies that provide additional insight into the relationship of self-esteem and depression (e.g., the relation among self-esteem, stressful events, and depression; Orth et al., 2009). Moreover, the results of the meta-analysis of Sowislo and Orth (2013), which include cross-sectional as well as longitudinal studies, yielded support for both the vulnerability model and the scar model. Whereas the effect of self-esteem on depression ($\beta = -.16$) was significantly stronger than the effect of depression on self-esteem ($\beta = -.08$), the latter was large enough to warrant further investigation.

The Present Research

The objective of the present research was to replicate the Orth et al. (2008) article with a German longitudinal data set. When it comes to the need to replicate major research studies, longitudinal studies such as the Orth et al. (2008) study pose major challenges for several reasons: The financial costs are high, it takes a long time before data become available, and an enormous time and effort have to be invested in all steps of data collection and analyses. Therefore, it is quite uncommon for longitudinal studies, such as Orth et al. (2008), to be subject to either direct or conceptual replications. The present study benefited from the fact that another large-scale, longitudinal data set that is very similar to the Orth et al. (2008) study design and that contains the same measures (self-esteem and depression) was collected by our research team.

In terms of the overlap, our study used the same measures of self-esteem and depression, a highly similar sample, and a similar time line to the Orth et al. (2008) study (Study 2).[1] Our study differed slightly in that we had three rather than four assessments and the time lag between assessments was 2 years instead of 1 year as in the Orth et al. (2008) study. In terms of the analyses, we reproduced as closely as possible the approach taken by Orth et al. (2008). More specifically, we first tested whether self-esteem and depression were empirically separable or whether they should be modeled as two endpoints of one continuum (as proposed by the common factor model; Watson et al., 2002). Second, we investigated the reciprocal effects of self-esteem and depression, by means of cross-lagged-regression models (Biesanz, 2012). Thus, paralleling Orth et al. (2008), we simultaneously tested the vulnerability and the scar model. Furthermore, we tested for gender differences in the measurement as well as in the structural model as was done by Orth et al. (2008).

Method

The data come from a large, ongoing longitudinal German study (Transformation of the Secondary School System and Academic Careers; TOSCA; for a detailed overview see Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010). The TOSCA study currently encompasses six time points. Data for self-esteem as well as for depression are available for three waves (for the present study: T1, T2 and T3). T1 is 2 years after graduation from high school (February to May, 2004), when participants completed an extensive questionnaire taking about 2 hr in exchange for a financial reward of €10. The second (T2) and third (T3) assessment took place from February to May, 2006 and from February to May, 2008, respectively. Again, participants completed an extensive questionnaire taking about 2 hr in exchange for a financial reward of €10.

Participants

Data were available for $n = 2{,}318$ (62.1% female) individuals at T1, $n = 1{,}912$ (64.0% female) individuals at T2, $n = 1{,}871$ (63.0% female) individuals at T3. The sample size of the pooled data set is $N = 2{,}512$. Mean age of participants was $M = 21.5$ years ($SD = 0.8$) at T1, $M = 23.4$ years ($SD = 0.6$) at T2, and $M = 25.4$ years ($SD = 0.7$) at T3. In line with recommendations concerning replications (Simonsohn, 2014), the sample size is about seven times the size of the original longitudinal study (Study 2)[2] reported by Orth et al. (2008).

For attrition analyses, again paralleling the procedure chosen by Orth et al. (2008), we compared continuers, who completed all three time points, with dropouts, who participated only in the first wave. There were no significant differences on the study variables (self-esteem and depression). More detailed results about attrition analyses are reported in the Appendix.

Measures

Self-esteem. Like Orth et al. (2008) we used the Rosenberg Self-Esteem Scale (RSE; Rosenberg, 1965) to assess self-esteem: Three items were administered: (a) "*At times, I think I am no good at all.*" (b) "*All in all, I am inclined to feel that I am a failure.*" and (c) "*I wish I could have more respect for myself.*" These items were translated for the TOSCA project into German. A slight deviation was in measuring the responses. Orth et al. (2008) used a 5-point scale (1 = *not very true of me* to 5 = *very true of me*), whereas we used a 4-point scale, ranging from 1 (applies not at all) to 4 (applies totally). However, internal consistency was good across all three assessments (Cronbach's α .84 at T1, .84 at T2 and .86 at T3). These coefficients are close to the alpha reliabilities in the Orth et al. (2008) study (.89 to .91), which were able to use a total of 10 items.

Depression. Depressive symptoms were assessed with the 15-item German version ("Allgemeine Depressionskala"; ADS-K; Hautzinger & Bailer, 1993) of the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977). By comparison, Orth et al. (2008) were able to use the full 20-item scale of the CES-D in all assessments. The response scales were similar in both studies. For each item, participants reported how frequently they experienced the symptom within the past week using a 4-point scale (0 = *rarely or none of the time*, 1 = *sometimes*, 2 = *frequently*, 3 = *most of the time*). In the present study internal consistency was good across all three assessments (Cronbach's α .90 at T1, .90 at T2 and .91 at T3), and again, the alpha reliabilities are nearly the same as in the Orth et al. (2008) article ($\geq$.90).

[1] We mainly focus on Study 2 of Orth et al. (2008) because of the similar age range. However, the vulnerability effect of low self-esteem on depression seems to hold from childhood to old age (see Sowislo & Orth, 2013).

[2] It should be noted that the sample size of Study 1 of Orth et al. (2008) that focused on adolescents was $N = 2{,}403$.

Procedure for the Statistical Analysis

To reach the highest-possible degree of similarity to Orth et al. (2008) we followed their statistical analyses as closely as possible. To address the first research question (the common factor model), we used confirmatory factor[3] analyses. To this end, we compared, separately for each time point, a one-factor and a two-factor model. In the one-factor model all indicators (the three self-esteem indicators and the three item parcels for depression) were modeled to load on one factor. In a next step, we specified the two-factor model, again separately for each time point and compared the model fit to that of the one-factor model. This two-factor model simultaneously established the same factor structure over time. Following this, we constrained the factor loadings of each indicator over time and compared the models to one another. This model is at the same time our starting point for the second research question.

In terms of the research questions regarding the relationship between self-esteem and depression (vulnerability and scar model), we specified two cross-lagged-regression models (Biesanz, 2012; Jöreskog, 1978). In the first cross-lagged-regression model, all structural coefficients were freely estimated. In the second cross-lagged-regression model, we constrained the structural parameters (stability coefficients and cross-lagged coefficients) to be equal across the two time intervals. The models were identified by fixing the first factor loadings of each latent variable (construct) to 1 (also see Orth et al., 2008, p. 700). We used multigroup models to test for gender differences in the measurement as well as in the structural model by comparing a model with freely estimated factor loadings (resp. structural coefficients) with a model with constrained factor loadings (resp. structural coefficients) across gender.

To ensure comparability with Orth et al. (2008), we used item parcels to create the latent depression construct. Little, Cunningham, Shahar, and Widaman (2002) pointed out that item parcels produce more reliable latent variables than individual items (but see Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013). Therefore, we randomly aggregated the 15 items into three parcels. For self-esteem it was not possible to build item parcels, because there were only three item indicators in the TOSCA study.

Goodness of fit indices. Model fit evaluation was based on the common fit indices[4]: The Tucker-Lewis-Index (TLI), the comparative fit index (CFI), and the root-mean-square error of approximation (RMSEA; see Hu & Bentler, 1998). Based on recommendations of Hu and Bentler (1999), a good fit is indicated by values equal to or greater than .95 for CFI and TLI and equal to or less than .05 for RMSEA.

Comparisons of the model fit are typically done using the χ^2 statistics. However, the χ^2 test of overall model fit and the (corrected) χ^2-based likelihood ratio test have been shown to be sensitive to sample size (Browne & Cudeck, 1992). Thus, in large samples, the power to detect even trivial differences in model fit is extremely high. Hence, all of the above-described fit indices can also be used to compare nested models. A variety of simulation studies have shown that the CFI and RMSEA were more appropriate for detecting noninvariance in measurement models than χ^2 difference tests when sample sizes were large (Chen, 2007; Cheung & Rensvold, 2002). Their results showed that for testing for invariance in factor loadings, a change of $\geq .01$ in the CFI, supplemented by a change of $\geq .02$ in the RMSEA, would indicate noninvariance.

Missing data. As a consequence of study attrition or nonresponses at single time points some of the variables had missing values. To deal with these we used, again paralleling the approach chosen by Orth et al. (2008), the full information maximum likelihood procedure (see, e.g., Enders, 2001). Because of less biased parameter estimates, this procedure is believed to be superior to conventional methods such as listwise or pairwise deletion (Graham, 2009).

The analyses were conducted using Mplus 7 (Muthén & Muthén, 1998–2014). All statistical tests were performed two-sided at a level of significance of 5%.

Results

We structured the results following the approach chosen by Orth et al. (2008) and, to allow for easy comparisons, also report their findings. Means and *SD*s of all measures are reported in Table 1. The first four time points are the means and *SD*s of the Orth et al. (2008) study (Study 2). Following this, the means and *SD*s of our study (three time points) are shown. The means of self-esteem were slightly higher in our study compared with the means of self-esteem reported by Orth et al. (2008). Similarly, the means of depression differed slightly from those reported by Orth et al. (2008): they were somewhat lower in our study compared with the means of depression reported by Orth et al. (2008)[5] (see Table 1).

With regard to the first research question (to test the common factor model; Watson et al., 2002), we tested a one-factor model against a two-factor model of self-esteem and depression, separately for each time point (see also Orth et al., 2008, p. 701). The fit for the one-factor model was low: $\chi^2(114, N = 2{,}506) = 3739{,}20$, $p < .001$, TLI = .756, CFI = .818, RMSEA = .113, 90% CI of RMSEA [.110, .116] (see also Model 0, Table 2). The model fit for the two-factor solution was considerably better (see Model 1, Table 2). Hence, our results indicate that self-esteem and depression should be modeled separately, which was the same conclusion as drawn by Orth et al. (2008, p. 701).

Measurement Models for the Association Between Self-Esteem and Depression

With regard to the second research question (vulnerability and the scar model), we first estimated two measurement models and afterward two structural models, implementing further

[3] Because of the nonnormality distributions of the item indicators, we used the robust maximum likelihood estimator for all analyses (MLR; Muthén & Muthén, 1998–2014).

[4] Orth et al. (2008) did not report the standardized root-mean-square residual (SRMR) that has become a widely accepted index in recent years. Our analyses showed acceptable fit for all central models when using the SRMR.

[5] It should be noted, that the direct comparison of the means of self-esteem and depression of our study with the means of Orth et al. (2008) is compromised slightly because they are based on a different number of items (self-esteem: 3 vs. 10 items and depression: 15 vs. 20 items).

Table 1
Means and Standard Deviations (SD) of Measures

	Orth et al. (2008), Study 2								The present study			
	Time 1 (18 years)		Time 2 (19 years)		Time 3 (20 years)		Time 4/ Time 1 (21 years)		Time 2 (23 years)		Time 3 (25 years)	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Rosenberg Self-Esteem Scale	3.82	0.77	3.88	0.82	3.86	0.77	4.06	0.72				
Center for Epidemiological Studies Depression Scale	0.98	0.58	0.94	0.57	0.82	0.53	0.74	0.52				
Rosenberg Self-Esteem Scale							3.29	0.62	3.32	0.64	3.44	0.62
Center for Epidemiological Studies Depression Scale							0.70	0.52	0.64	0.51	0.59	0.50

Note. Response scales ranged from 1 to 4 for the Rosenberg Self-Esteem Scale (RSE) and from 0 to 3 for Center for Epidemiological Studies Depression Scale (CES-D).

model constraints. The procedure is the same as in Orth et al. (2008). The model fit for Model 1 was good (see Table 2). To establish the same factor structure with equal factor loadings over time, we estimated in the next step a measurement model with constrained factor loadings of each indicator over time (Model 2). The model fit was also good (see Table 2). We then compared the model fit of the freely estimated model to that of the model with equal factor loadings of each indicator over time. Using the corrected χ^2 difference test, the model comparison yielded no significant difference: $\Delta\chi^2 = 18.91$, $\Delta df = 12$, $\Delta p = .09$; ΔTLI = .000, ΔCFI = .000 and ΔRMSEA = .000. The cross-sectional latent correlations between self-esteem and depression were −.61 at Time 1, −.62 at Time 2 and −.63 at Time 3 (all ps $< .001$). These are comparable to the correlations of Orth et al. (2008): −.58 at Time 1, −.63 at Time 2, −.51 at Time 3, and −.62 at Time 4.

Gender Differences in the Measurement Model

Consistent with Orth et al. (2008) we tested for gender differences in the measurement model. We used a multigroup analysis and compared a model with freely estimated factor loadings ($\chi^2(240, N = 2{,}500) = 453.23$, $p < .001$, TLI = .986, CFI = .989, RMSEA = .027, 90% CI of RMSEA [.023, .030]) to a model with constrained factor loadings across gender ($\chi^2(246, N = 2{,}500) = 501.01$, $p < .001$, TLI = .984, CFI = .987, RMSEA = .029, 90% CI of RMSEA [.025, .032]). Although the corrected χ^2 difference test showed that the model fit of the more restrictive model was significantly worse than that of the less restrictive model ($\Delta\chi^2 = 39.09$, $\Delta df = 6$, $\Delta p < .001$), the other fit indices did not differ notably ΔTLI = .002, ΔCFI = .002 and ΔRMSEA = .002. This is in line with the results of Orth et al. (2008) and their conclusion that there are no meaningful gender differences in the model.

Table 2
Fit Indices of the Models Tested

Model	χ^2	*df*	TLI	CFI	RMSEA	90% CI of RMSEA
0. One-factor model	3739,20***	114	.756	.818	.113	[.110, .116]
Measurement models[a]						
1. Free loadings	197.09***	102	.993	.995	.019	[.015, .023]
2. Longitudinal constraints on loadings	215.63***	114	.993	.995	.019	[.015, .023]
2.1 Multigroup model for gender (free loadings)	453.23***	240	.986	.989	.027	[.023, .030]
2.2 Multigroup model for gender (constraint loadings)	501.01***	246	.984	.987	.029	[.025, .032]
Structural models[a]						
3. Free structural coefficients	268.80***	114	.990	.992	.023	[.020, .027]
4. Longitudinal constraints on structural coefficients	270.06***	118	.990	.992	.023	[.019, .026]
4.1 Multigroup model for gender (free coefficients)	538.83***	255	.983	.986	.030	[.026, .033]
4.2 Multigroup model for gender (constraint coefficients)	544.58***	259	.983	.986	.030	[.026, .033]
Fit indices of the models tested of Orth et al. (2008) Study 2						
Measurement models[b]						
1. Free loadings	226.8*	188	.99	.99	.024	[.009, .035]
2. Longitudinal constraints on loadings	257.1**	206	.98	.99	.026	[.014, .036]
Structural models[b]						
3. Free structural coefficients	284.0**	212	.98	.99	.031	[.021, .040]
4. Longitudinal constraints on structural coefficients	288.3**	220	.98	.99	.029	[.019, .038]

Note. TLI = Tucker-Lewis Index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval.
[a] $N = 2{,}506$. [b] $N = 359$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

The Longitudinal Relationship Between Self-Esteem and Depression

Following the analytical procedure of Orth et al. (2008), we used for the cross-lagged regression models the measurement models of Model 2 (equal factor loadings). In the first cross-lagged regression model (Model 3), all structural coefficients were freely estimated. Model fit was good (see Table 2). In the second cross-lagged model (Model 4), we constrained the structural parameters (stability coefficients and cross-lagged coefficients) to be equal across the two intervals (cf. Orth et al., 2008, p. 700). Again, the model fit was good. The difference in fit between Models 3 and 4 was nonsignificant: $\Delta\chi^2 = 2.23$, $\Delta df = 4$, $\Delta p = .69$; ΔTLI = .000, ΔCFI = .000 and ΔRMSEA = .000. Therefore, we favored the more parsimonious model. These results are also similar to Orth et al. (2008) (see Table 2, Model 3 and 4 of the Orth et al. (2008) results).

In Figure 1, the path diagram for the constrained model (Model 4) is depicted. The values shown are standardized. The averaged values from T1 to T4 of Orth et al. (2008, Study 2) are depicted in parentheses. The cross-lagged paths from self-esteem to depression were both statistically significant (−.23 and −.24, $ps < .001$) and are comparable with the coefficients reported by Orth et al. (2008) (averaged coefficient: −.21). By contrast, the cross-lagged paths from depression to self-esteem were both nonsignificant (−.04, $p = .11$), which is also in line with the results of Orth et al. (2008) and in support of the vulnerability model.

Not surprisingly, given the longer time intervals between the assessments (2 years vs. 1 year), the stability coefficients of self-esteem were somewhat lower in the present study (.72 and .73 vs. .83; averaged stability coefficient). For depression, the stability coefficients across both studies were indistinguishable (.33 for both paths vs. .35 in Study 2 of Orth et al., (2008)).

Gender Differences for the Longitudinal Relationship Between Self-Esteem and Depression

As a last step, we tested for gender differences just as was done in Orth et al. (2008). To this end, we again used a multigroup analysis and compared a model with freely estimated structural coefficients ($\chi^2(255, N = 2{,}500) = 538.83$, $p < .001$, TLI = .983, CFI = .986, RMSEA = .030, 90% CI of RMSEA [.026, .033]) to a model with equal structural coefficients across gender ($\chi^2(259, N = 2{,}500) = 544.58$, $p < .001$, TLI = .983, CFI = .986, RMSEA = .030, 90% CI of RMSEA [.026, .033]). The models did not differ significantly from each other (comparable with Orth et al., 2008, p. 703).

Discussion

The aim of this study was to conduct a unique replication in which we attempted to reproduce results from a previously published longitudinal study (Orth et al., 2008). Conducting replications of high impact articles is highly important for building a reliable body of scientific knowledge (Asendorpf et al., 2013). To date, most of the replication attempts have been focused on experimental studies that are more easily replicated (Johnson, Cheung, & Donnellan, 2014). Longitudinal studies are, by their

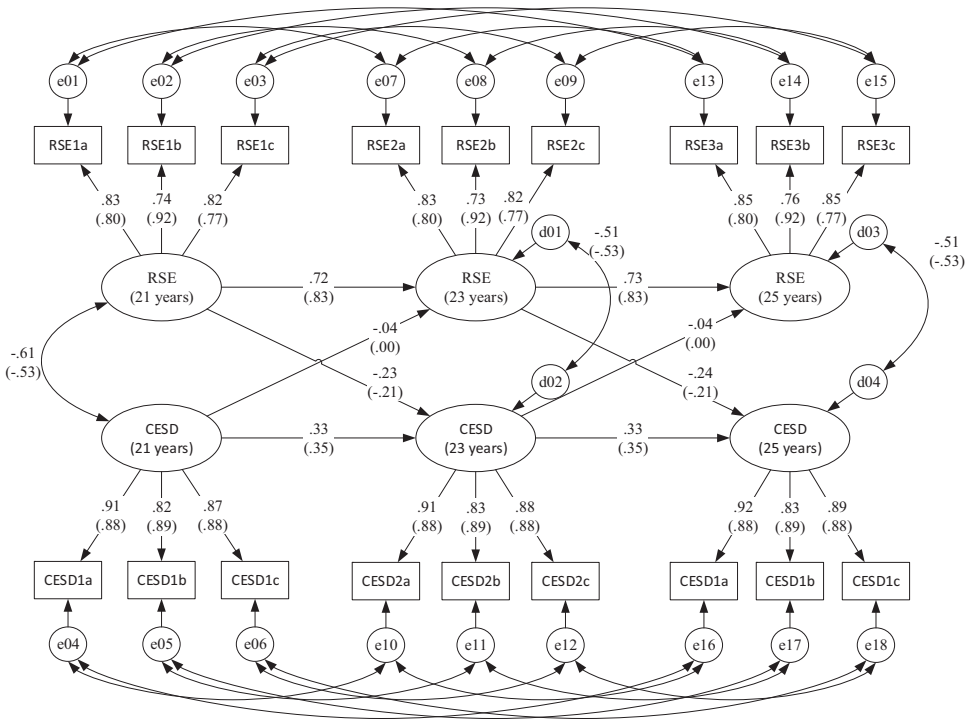

Figure 1. Cross-lagged regression model of self-esteem and depression with longitudinal constraints on factor loadings and structural coefficients (Model 4). Values in parentheses are the averaged coefficients (from T1 to T4) from Study 2 of Orth et al. (2008). All values shown are standardized. RSE = Rosenberg Self-Esteem Scale; CES-D = Center for Epidemiological Depression Scale.

very nature, almost impossible to directly replicate given the vagaries of sampling and history. Nonetheless, it is no less important to replicate longitudinal research when the opportunity arises.

In the current study, we took advantage of a highly similar data and sampling structure in the TOSCA longitudinal study to conduct a very close replication of the work of Orth et al. (2008). In their original contribution, Orth et al. (2008) concluded that (a) depression and self-esteem were not the same construct (the common-factor model), and (b) that the vulnerability model garnered more support than the scar model. The present study replicated the Orth et al. (2008) study in a notable manner (in regard to the comparability of the coefficients) and corroborates the results pattern in favor of the vulnerability model.

Moreover, the fact that the results of Orth et al. (2008) and our study are nearly indistinguishable, indicates the generalizability of the vulnerability effect of self-esteem on depression across two western countries (United States and Germany). Despite these similarities, future research should focus on more formal tests of cross-cultural differences of the scar and vulnerability model with the goal of determining if there are cultures that might be more prone to the scar model than the groups compared here.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. http://dx.doi.org/10.1002/per.1919

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York, NY: Harper & Row.

Beck, A. T., Steer, R. A., Epstein, N., & Brown, G. (1990). Beck Self-Concept Test. *Psychological Assessment, 2,* 191–197. http://dx.doi.org/10.1037/1040-3590.2.2.191

Biesanz, J. C. (2012). Autoregressive Longitudinal Models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 459–471). New York, NY: Guilford Press Pubn.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258. http://dx.doi.org/10.1177/0049124192021002005

Butler, A. C., Hokanson, J. E., & Flynn, H. A. (1994). A comparison of self-esteem lability and low trait self-esteem as vulnerability factors for depression. *Journal of Personality and Social Psychology, 66,* 166–177. http://dx.doi.org/10.1037/0022-3514.66.1.166

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504. http://dx.doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352–370. http://dx.doi.org/10.1037/1082-989X.6.4.352

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Hautzinger, M., & Bailer, M. (1993). *Allgemeine Depressions Skala. Manual* [German version of the Center for Epidemiological Studies Depression Scale]. Göttingen: Beltz Test GmbH.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeits test für 4. bis 12.* Klassen (KFT 4–12+R) [Test of Cognitive Functioning for Grades 4 to 12]. Göttingen, Germany: Beltz-Test.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453. http://dx.doi.org/10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvery (2008). *Social Psychology, 45,* 209–215. http://dx.doi.org/10.1027/1864-9335/a000186

Joiner, T. E., Jr. (2000). Depression's vicious scree: Self-propagating and erosive processes in depression chronicity. *Clinical Psychology: Science and Practice, 7,* 203–218. http://dx.doi.org/10.1093/clipsy.7.2.203

Jöreskog, K. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–477. http://dx.doi.org/10.1007/BF02293808

Klein, D. N., Kotov, R., & Bufferd, S. J. (2011). Personality and depression: Explanatory models and review of the evidence. *Annual Review of Clinical Psychology, 7,* 269–295. http://dx.doi.org/10.1146/annurev-clinpsy-032210-104540

Lewinsohn, P. M., Steinmetz, J. L., Larson, D. W., & Franklin, J. (1981). Depression-related cognitions: Antecedent or consequence? *Journal of Abnormal Psychology, 90,* 213–219. http://dx.doi.org/10.1037/0021-843X.90.3.213

Little, T. D., Cunningham, W., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173. http://dx.doi.org/10.1207/S15328007SEM0902_1

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18,* 257–284. http://dx.doi.org/10.1037/a0032773

Metalsky, G. I., Joiner, T. E., Jr., Hardin, T. S., & Abramson, L. Y. (1993). Depressive reactions to failure in a naturalistic setting: A test of the hopelessness and self-esteem theories of depression. *Journal of Abnormal Psychology, 102,* 101–109. http://dx.doi.org/10.1037/0021-843X.102.1.101

Muthén, L. K., & Muthén, B. O. (1998–2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén

Orth, U., Robins, R. W., & Meier, L. L. (2009). Disentangling the effects of low self-esteem and stressful events on depression: Findings from three longitudinal studies. *Journal of Personality and Social Psychology, 97,* 307–321. http://dx.doi.org/10.1037/a0015645

Orth, U., Robins, R. W., & Roberts, B. W. (2008). Low self-esteem prospectively predicts depression in adolescence and young adulthood. *Journal of Personality and Social Psychology, 95,* 695–708. http://dx.doi.org/10.1037/0022-3514.95.3.695

Orth, U., Robins, R. W., Trzesniewski, K. H., Maes, J., & Schmitt, M. (2009). Low self-esteem is a risk factor for depressive symptoms from young adulthood to old age. *Journal of Abnormal Psychology, 118,* 472–478. http://dx.doi.org/10.1037/a0015922

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. http://dx.doi.org/10.1177/014662167700100306

Rohde, P., Lewinsohn, P. M., & Seeley, J. R. (1990). Are people changed by the experience of having an episode of depression? A further test of the scar hypothesis. *Journal of Abnormal Psychology, 99,* 264–271. http://dx.doi.org/10.1037/0021-843X.99.3.264

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

Shahar, G., & Davidson, L. (2003). Depressive symptoms erode self-esteem in severe mental illness: A three-wave, cross-lagged study.

Journal of Consulting and Clinical Psychology, 71, 890–900. http://dx.doi.org/10.1037/0022-006X.71.5.890

Simonsohn, U. (2014). Small telescopes: Detectability and the evaluation of replication results. Available at SSRN: http://ssrn.com/abstract=2259879 or http://dx.doi.org/10.2139/ssrn.2259879

Smith, E. R., Simpson, J. A., & King, L. A. (Eds.). (2014). Call for replication studies papers. *Journal of Personality and Social Psychology, 106,* 1052.

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin, 139,* 213–240. http://dx.doi.org/10.1037/a0028931

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Eds.). (2010). *Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* [School achievement at the end of high school: Effects of the reform of upper secondary education]. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. http://dx.doi.org/10.1007/978-3-531-92037-5

Watson, D., Suls, J., & Haig, J. (2002). Global self-esteem in relation to structural models of personality and affectivity. *Journal of Personality and Social Psychology, 83,* 185–197. http://dx.doi.org/10.1037/0022-3514.83.1.185

Appendix

Attrition Analyses

One difficulty in multiwave studies involves cases with missing data. Incomplete information from participants may occur if participants are unavailable for one or more waves of data collection, or did not agree to give information at some time points. Thus, we investigated potential attrition impact. To this end, we compared continuers, who completed all three time points, with dropouts, who participated only in the first wave. There were no significant differences on the study variables (self-esteem: $t(2301) = -0.15$, $p = .99$, $d = 0.00$ and depression: $t(2295) = 0.06$, $p = .95$, $d = 0.00$). However, continuers were more likely to be female ($\chi^2(1) = 4.64$, $p = 0.03$), had better grade point averages[6] ($t(2377) = -3.28$, $p < .001$, $d = -0.20$) and performed better on a reasoning ability test (Heller & Perleth, 2000; $t(2499) = 2.179$, $p = .03$, $d = 0.13$). Overall, the differences between continuers and dropouts were rather small ($ds \leq |.20|$).

[6] In the German education system, lower grade point averages indicate higher achievement.

Received November 24, 2014
Revision received January 27, 2015
Accepted February 4, 2015 ■