Check for updates

Article

Self-Esteem in 60 Seconds: The Six-Item State Self-Esteem Scale (SSES-6)

Assessment
2022, Vol. 29(2) 152–168
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191120958059
journals.sagepub.com/home/asm
SAGE

Gregory D. Webster[1], Jennifer L. Howell[2], and James A. Shepperd[1]

Abstract

With 20 items, the State Self-Esteem Scale (SSES) can be cumbersome in settings that demand efficiency. The present research created an efficient six-item version of the SSES that preserves score reliability and validity and its three-dimensional structure: social, appearance, and performance self-esteem. Item response theory and confirmatory factor analyses identified the "best" six items—two from each dimension (Study 1). Participants completed the SSES four times at 2-week intervals (Studies 2 and 3). The six-item SSES' scores showed adequate test–retest reliability, explained substantial variance in trait-relevant measures, and showed convergent validity with related self-esteem measures. Participants completed the SSES and a laboratory experiment where they received negative feedback on an essay they had written and could retaliate against their evaluator by allocating hot sauce for them to consume (Study 4). The six-item SSES interacted with self-esteem instability in expected ways to predict hot sauce allocated.

Keywords

self-esteem, state self-esteem, item response theory, aggression, multilevel modeling

Researchers face a key problem when they must measure trait or state constructs quickly or efficiently—the measures must be brief while maintaining score reliability and validity. Several situations arise that demand efficient self-report measures ranging from daily-diary and experience-sampling studies—where participants complete the same measures repeatedly—to prescreening, mass-testing, and longitudinal studies—where participants complete a large suite of measures, and a premium is placed on the number of items.

With the advent of mobile technology and experience-sampling techniques, the past 20 years have witnessed a groundswell of support for brief, efficient versions of extant measures, including the Single-Item Self-Esteem Scale (SISES; Robins et al., 2001), the Ten-Item Personality Inventory (TIPI; Gosling et al., 2003), the Dark Triad Dirty Dozen (Jonason & Webster, 2010; Webster & Jonason, 2013), the Eight-Item Impusivity and Sensation Seeking Scale (Webster & Crysel, 2012), the Single-Item Need to Belong Scale (Nichols & Webster, 2013), and three-item scales for (a) social anxiety (Nichols & Webster, 2015) and (b) one's partner's alcohol consumption (Rodriguez & Webster, 2020). These and other measures are used in a wide range of settings and have allowed for great ease of data collection in situations where time or item space is valuable (e.g., field studies, daily-diary studies, round-robin designs).

One measure particularly suited for daily-dairy and experience-sampling studies is the State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991). Although it can be used as a global measure of state self-esteem, the SSES can also be scored as three related-but-distinct, factor-based subscales: social, appearance, and performance (see Table 1 for items and their subscales). The SSES has been widely adopted among researchers, garnering over 2,400 citations. Nevertheless, with 20 items, it can be unwieldy and cumbersome in situations that demand fewer items or repeated measurements. Thus, the key purposes of the present research were to (a) identify the psychometrically "best" items of the SSES' 20 items (i.e., items with differential difficulties and high discrimination); (b) test the three-factor structure; (c) test score reliability and validity; and (d) examine how the new measure performs in a laboratory aggression experiment. To these ends, we performed item response theory (IRT) analyses to identify the "best" SSES items (Study 1) and used confirmatory factor analyses

[1]University of Florida, Gainesville, FL, USA
[2]University of California, Merced, CA, USA

Corresponding Author:
Gregory D. Webster, Department of Psychology, University of Florida, P.O. Box 112250, Gainesville, FL 32611-2250, USA.
Email: gdwebs@ufl.edu

Table 1. Study 1: Item Response Theory Parameter Estimates: Three Separate Models, One for Each State Self-Esteem Subscale.

Item	Sub.	α	β_1	β_2	β_3	β_4	β_5	β_6
1. I feel confident about my abilities.	Perf.	0.89	−3.93	−3.42	−2.48	−1.38	−0.11	1.23
2. I am worried about whether I am regarded as a success or failure.[a]	Soc.	0.78	−2.16	−1.06	−0.13	0.44	0.90	1.85
3. **I feel satisfied with the way my body looks right now.**	App.	1.83	−1.70	−1.06	−0.47	0.16	0.89	1.87
4. **I feel frustrated or rattled about my performance.**[a]	Perf.	0.91	−2.69	−1.69	−0.67	0.08	0.69	1.79
5. I feel that I am having trouble understanding things that I read.[a]	Perf.	0.90	−3.08	−2.28	−1.23	−0.61	0.01	1.04
6. I feel that others respect and admire me.	App.	0.59	−4.16	−2.78	−1.90	−0.55	1.02	2.90
7. I am dissatisfied with my weight.[a]	App.	1.01	−1.61	−0.90	−0.21	0.18	0.55	1.34
8. I feel self-conscious.[a]	Soc.	1.19	−1.62	−0.98	−0.20	0.32	0.86	1.63
9. I feel as smart as others.	Perf.	1.05	−2.78	−2.06	−1.10	−0.36	0.71	1.72
10. I feel displeased with myself.[a]	Soc.	0.97	−2.80	−2.12	−1.21	−0.60	0.06	1.12
11. I feel good about myself.	App.	1.28	−2.94	−2.12	−1.49	−0.61	0.31	1.35
12. I am pleased with my appearance right now.	App.	2.75	−1.84	−1.25	−0.70	−0.14	0.57	1.42
13. **I am worried about what other people think of me.**[a]	Soc.	1.39	−1.74	−1.02	−0.22	0.37	0.92	1.73
14. I feel confident that I understand things.	Perf.	1.10	−3.25	−2.55	−1.79	−0.80	0.33	1.67
15. I feel inferior to others at this moment.[a]	Soc.	0.88	−3.34	−2.17	−1.31	−0.61	0.02	0.98
16. **I feel unattractive.**[a]	App.	1.43	−2.30	−1.63	−1.00	−0.45	0.04	0.85
17. I feel concerned about the impression I am making.[a]	Soc.	1.18	−1.78	−0.89	0.04	0.71	1.06	1.74
18. I feel that I have less scholastic ability right now than others.[a]	Perf.	1.71	−2.06	−1.34	−0.74	−0.20	0.28	1.07
19. **I feel like I'm not doing well.**[a]	Perf.	1.63	−2.16	−1.59	−0.92	−0.36	0.09	0.90
20. **I am worried about looking foolish.**[a]	Soc.	1.26	−1.83	−1.17	−0.45	0.06	0.49	1.17

Note. $N = 746$. Selected items are in bold. Sub. = subscale; Perf. = performance; Soc. = social; App. = appearance; α = discrimination parameter; β_{1-6} = difficulty parameters.
[a]Reverse-scored item.

(CFAs) to test and replicate its three-factor structure (Studies 1 and 3). We also had participants complete the SSES at 4 time points to decompose state- and trait-level variance using multilevel models, as well as examine test–retest reliability and convergence validity with related measures (Studies 2 and 3). We also conducted a laboratory aggression experiment in which participants completed the SSES, received negative feedback on a personal essay they had written, and then received the opportunity to aggress against their evaluator by preparing a sample of hot sauce for them to consume (Study 4). In all four studies, we report how sample size was determined, as well as all measures, manipulations, and data exclusions.

Study 1: Item Selection, Scale Structure, and Score Validity

In Study 1, we used IRT to select the "best" items from the 20-item SSES (i.e., SSES-20). To preserve its three-factor structure, we sought to identify the two best items from each state self-esteem subscale: performance, social, and appearance. Using the "best" six items, we then tested the new SSES-6's factor structure, comparing a single-factor with both a three-factor and a hierarchical factor structure. Consistent with the SSES-20's nature, we expected the SSES-6's three-factor structure to fit the data better than a one-factor model. As an exploratory exercise, we also measured the Big Five personality traits to assess correlations among the SSES, its subscales, and the Big Five. We made no a priori predictions about the SSES's relations with the Big Five, choosing instead to attempt to replicate any Study 1 findings later in Study 3. Finally, we assessed participants' body mass index (BMI) to examine its association with appearance state self-esteem, expecting a negative linear relation.

Method

Participants and Procedure. Participants were 746 undergraduates (210 men, 533 women, 3 did not report) enrolled in introductory psychology courses at the University of Florida who completed online prescreening questionnaires for course credit in the spring ($n = 384$) and fall ($n = 362$) semesters of 2012 (ages: 17–26 years, $Mdn = 18.0$, $M = 18.9$, $SD = 1.3$). Sample size was determined solely by the number of students who wished to complete the prescreening questionnaire, which consisted of several other measures submitted by multiple psychology labs. We had access to only the measures submitted by our lab.

Measures

Self-esteem. In both the spring and fall samples, we assed self-esteem with three measures: the SSES-20 (Heatherton & Polivy, 1991; Table 1), the SISES (Robins et al., 2001; "I have high self-esteem"), and the five-item Stability of

Self Scale (SSS; Rosenberg, 1965; Webster et al., 2017; e.g., "I change from a very good opinion from myself to a very poor opinion of myself") scored to reflect self-esteem *in*stability. The SSES used a response scale from 1 (*not at all*) to 7 (*extremely*) with the following stem (Heatherton & Polivy, 1991):

> This is a questionnaire designed to measure what you are thinking at this moment. There is, of course, no right answer for any statement. The best answer is what you feel is true of yourself at the moment. Be sure to answer all of the items, even if you are not certain of the best answer. Again, answer these questions as they are true for you RIGHT NOW.

Both the SISES and SSS used response scales from 1 (*strongly disagree*) to 7 (*strongly agree*).

Personality. In the fall sample, we used three Big Five personality measures: the TIPI (Gosling et al., 2003; e.g., "I see myself as extraverted, enthusiastic"), the 10-item short version of the Big Five Inventory (Rammstedt & John, 2007; e.g., "I see myself as someone who is outgoing, sociable"), and the Mini-IPIP (Donnellan et al., 2006; e.g., "I am the life of the party"). All three personality measures used response scales from 1 (*strongly disagree*) to 7 (*strongly agree*). We assessed the Big Five traits because (a) they are frequently used in scale-development studies, (b) represent a common framework for conceptualizing traits, and (c) we wished to see how state self-esteem related to the broader nomological network of personality. We chose to use three brief personality measures instead of one long one because any one assessment instrument can have its own idiosyncratic biases. Having three measures of the same construct can help guard against any systematic biases. Doing so also allowed us to assess how consistent correlations were across different personality measures.

Body mass index. In the fall sample, we also asked participants to self-report their height and weight, which were used to calculate their BMI (i.e., kg/m^2 or lbs. $\times$ $703/in^2$). In prior research, BMI correlated negatively with physical appearance self-esteem (French et al., 1996), and we expected to replicate this association.

Results and Discussion

Descriptive Statistics. The SSES-20 ($\alpha = .92$) and its subscales ($\alpha s \geq .85$) showed adequate internal consistency as did the SES-6 ($\alpha = .80$) and its subscales ($\alpha s \geq .72$), which is notable considering that Cronbach's α is a function of both the mean interitem correlation and the number of items (see Cortina, 1993). For details about the descriptive statistics of all measures, see supplementary Table S1 (available online).

Item Response Theory Models. We ran three independent two-parameter IRT analyses (i.e., graded-response models [GRMs]; see Samejima, 1969, 2016; see also Morizot et al., 2007)—one for each subscale with maximum likelihood estimation with robust standard errors—using M*plus* 6.1 (Muthén & Muthén, 1998/2011; see OSF link for code: https://osf.io/3cg5m). In choosing items for a brief measure, we sought to balance both item *discrimination* (α; higher numbers reflect greater discrimination) and item *difficulty* (βs; a.k.a. location or severity; the ease at which an item is endorsed across the latent trait, θ). Item discrimination describes how well an item differentiates among people with similar levels of the same latent trait (θ). Ideally, higher item discrimination parameters (αs) maximize the amount of information (discussed below). Item difficulty is the amount of the latent trait needed to have .5 probability of endorsing each item at each categorical response threshold (e.g., from a 1 to 2, 2 to 3, 3 to 4, etc.). Thus, there are *k*-1 difficulty parameters, where *k* is the number of response categories (e.g., a 7-point Likert-type scale produces 6 β parameters).

Rather than focus solely on content validity, we took a more empirical approach to item selection. Specifically, for each subscale, we sorted items by their mean difficulty scores (i.e., the average of each item's 6 βs), which allowed us to identify one item that was easy-to-moderate to endorse (smaller mean βs), and one item that was moderate-to-difficult to endorse (larger mean βs; see Nichols & Webster, 2015; Rodriguez & Webster, 2020; Webster & Crysel, 2012). We then chose the item with the highest discrimination score (α) within each region of difficultly (i.e., easy-to-moderate vs. moderate-to-difficult). Thus, for each subscale, neither of the chosen two items came from the same side of the latent trait spectrum (θ). For example, if the most difficult item was selected, an easy or easy-to-moderate item was also selected (e.g., see Table 1's Performance Items 4 and 19). In addition, five of the six chosen items had αs exceeding 1.0, indicating a high degree of item discrimination (Table 1).

In IRT, item or scale *information* reflects precision of measurement and is akin to reliability in classical test theory (CTT). Whereas CTT assumes reliability is invariant across the latent trait, IRT relaxes this assumption because reliability is often higher near a latent trait's center (where there is more information) than at its tails (where there is less information). Scale information per-item curves appear in Figure 1 for both versions of the SESS and its three subscales. Comparing the scale information per-item curves in Figure 1 (top panel) for the 6- and SSES-20 revealed that the SSES-6 was more efficient in recovering information from the latent trait (state self-esteem) than was the SSES-20. Specifically, information-per-item assessed via area under each information curve (Figure 1, top panel) was 3.79 for the SSES-20 and 4.66 for the SSES-6, yielding an

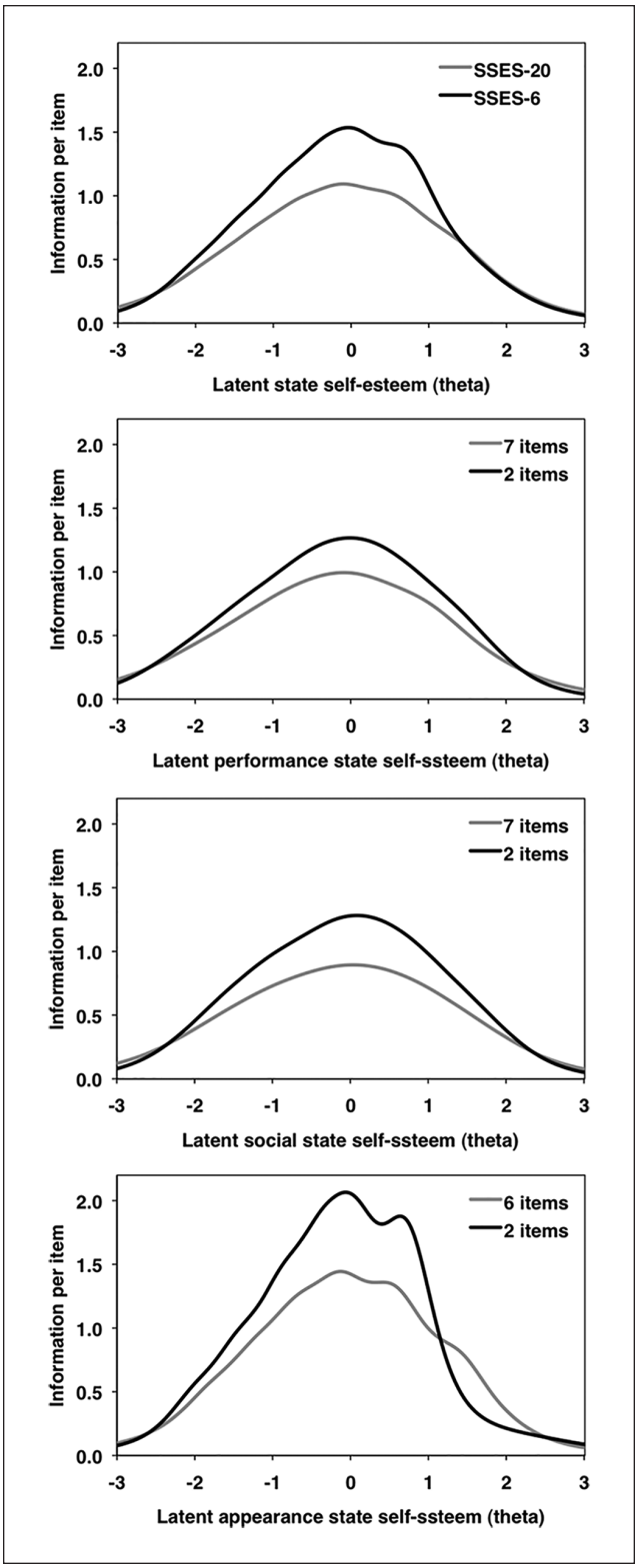

Figure 1. Study 1: Scale information per-item curves for the SSES-20 and SSES-6 (top) and their respective subscales (see Table 1): Performance (upper middle), Social (lower middle), and Appearance (bottom).
Note. SSES = State Self-Esteem Scale.

information efficiency improvement of 23.1%. This pattern was also true for the three SSES subscales. Specifically, areas under information curves (Figure 1, bottom 3 panels) were 3.58, 3.31, and 4.59 for the full performance, social, and appearance subscales, and 4.19, 4.21, and 5.58 for their respective two-item versions, yielding respective improvements of 17.2, 27.4, and 21.7%. Thus, in terms of per-item information, the SSES-6 and its subscales adequately and efficiently measure the same latent constructs as the SSES-20, but with far fewer items.

Confirmatory Factor Analyses. Because we sought to preserve the SSES-20's original three-subscale structure, we first tested a three-factor CFA for the SSES-6 (Table 2, top; Figure 2). In this specific case, a hierarchical or second-order-factor model yielded the same fit as a traditional three-factor one; the only difference was that the three latent subscale factors loaded onto a single, second-order latent state self-esteem factor, rather than simply correlating with one another. The three-factor/hierarchical model fit the data well; all fit indices were acceptable-to-good (Table 2, top; e.g., RMSEA = .064). We then fit a simpler, one-factor model to the data; this model fit the data poorly (e.g., RMSEA = .201) and significantly worse than the original three-factor model, $\chi^2(3) = 256.1$ (Table 2, top).

To examine the reproducibility of these measurement models, we ran independent CFAs on the spring and fall samples, which had adequate sample sizes (see MacCallum et al., 1999). Both CFAs yielded a pattern of results similar to the full sample; however, the three-factor model fit the data somewhat better in the fall sample than in the spring one (Table 2, top). Thus, the model fit replicated, and the predicted three-factor (or hierarchical) structure of the SSES-6 was supported.

Validity Correlations. The SSES-6's two-item subscales correlated highly with the SSES-20's respective six- or seven-item subscales for performance, social, and appearance (rs = .85, .87, and .94, respectively), and the SSES-6 correlated strongly with the SSES-20 (r = .94). The SISES correlated significantly with both the SSES and its subscales regardless of version (all rs ≥ .41); again, the SSES-6 and its subscales (rs = .41 to .61) were adequate reflections of their respective long-form measures (rs = .54 to .67).

Self-esteem instability, which often correlates negatively with self-esteem level (ρ = −.31; Okada, 2010), showed the expected pattern of negative correlations across both versions of the SSES and its subscales. Specifically, the trait measure of self-esteem instability (i.e., SSS) negatively correlated with the SSES-6 (r = −.57) and its three subscales (rs = −.43 to −.47), and correlated negatively with the SSES-20 (r = −.59) and its three subscales (rs = −.44 to −.58). Again, for details about the descriptive statistics of all measures, see supplementary Table S1 (available online).

Table 2. Studies 1 and 3: Confirmatory Factor Analysis Results by Sample and Time Point.

Models or differences	χ^2	*df*	CFI	TLI	RMSEA	90% CI LL	90% CI UL	p_{close}	SRMR
Study 1									
Full sample ($N = 746$)									
Three-factor/hierarchical	24.4	6	.986	.966	.064	.039	.092	.163	.020
One-factor	280.5	9	.797	.661	.201	.181	.22	.000	.073
Difference	256.1	3							
Spring ($n = 384$)									
Three-factor/hierarchical	23.2	6	.970	.926	.086	.051	.125	.046	.029
One-factor	136.2	9	.781	.635	.192	.164	.221	.000	.075
Difference	113.0	3							
Fall ($n = 362$)									
Three-factor/hierarchical	6.4	6	.999	.999	.014	.000	.071	.800	.015
One-factor	142.9	9	.822	.703	.203	.174	.233	.000	.071
Difference	136.5	3							
Study 3									
Time 1 ($n = 313$)									
Three-factor/hierarchical	16.7	6	.982	.956	.076	.034	.120	.138	.021
One-factor	92.8	9	.861	.769	.172	.142	.205	.000	.059
Difference	76.1	3							
Time 2 ($n = 269$)									
Three-factor/hierarchical	11.9	6	.989	.972	.060	.000	.111	.313	.020
One-factor	74.7	9	.874	.790	.165	.131	.200	.000	.058
Difference	62.8	3							
Time 3 ($n = 253$)									
Three-factor/hierarchical	10.7	6	.991	.979	.056	.000	.109	.368	.021
One-factor	105.5	9	.825	.708	.206	.172	.242	.000	.065
Difference	94.8	3							
Time 4 ($n = 254$)									
Three-factor/hierarchical	5.2	6	.999	.999	.000	.000	.075	.812	.014
One-factor	71.2	9	.885	.809	.165	.131	.202	.000	.057
Difference	66.0	3							

Note. Fit indexes (and suggested acceptable-fit cut-offs; see Browne & Cudeck, 1992; Hu & Bentler, 1999; see Kline, 2015, for a review and critique). All χ^2s were significant ($p < .05$) except for four three-factor models: Fall ($p = .38$), Time 2 ($p = .065$), Time 3 ($p = .097$), and Time 4 ($p = .52$). CFI = comparative fit index (≥.93); TLI = Tucker–Lewis index (nonnormed fit index or NNFI; ≥.90); RMSEA = root mean square error of approximation (≤.10); LL and UL = lower and upper limits; SRMR = standardized root mean square residual (≤.08); *df* = degrees of freedom.

Regarding personality, both extraversion and conscientiousness were positively correlated with state self-esteem, regardless of Big Five measure, SSES version, or SSES subscale (all *r*s ≥ .17); the same was true for neuroticism, albeit negatively correlated (all *r*s ≤ −.26). Patterns for agreeableness and openness were relatively less clear; however, the SSES-6 typically mirrored results for the SSES-20. Again, greater detail appears in supplementary Table S1 (available online). Overall, these correlations show an unsurprising pattern for personality and self-esteem, with people reporting higher extraversion, higher conscientiousness, or lower neuroticism also reporting higher state self-esteem. We revisit Big Five personality in Study 3.

BMI scores correlated significantly and negatively with the SSES-6 composite ($r = -.12$), but not the full SSES-20 ($r = -.08$). Regarding subscales, BMI only showed significant negative correlations with appearance state self-esteem, as expected, for both the six- ($r = -.20$) and two-item ($r = -.23$) versions. On an exploratory basis, we also tested for nonlinearity in the observed linear effects by regressing the SSES-6 and both versions of the appearance subscales onto linear and quadratic BMI (see Cohen et al., 2003). None of the quadratic effects was significant (*p*s > .18), suggesting that the three observed negative BMI effects were indeed linear.

We also examined differences in the criterion variables between the two SSES versions (and their respective subscales) using dependent correlations tests (Steiger, 1980) via an online calculator (Lee & Preacher, 2013). In supplementary Table S1 (available online), significant differences

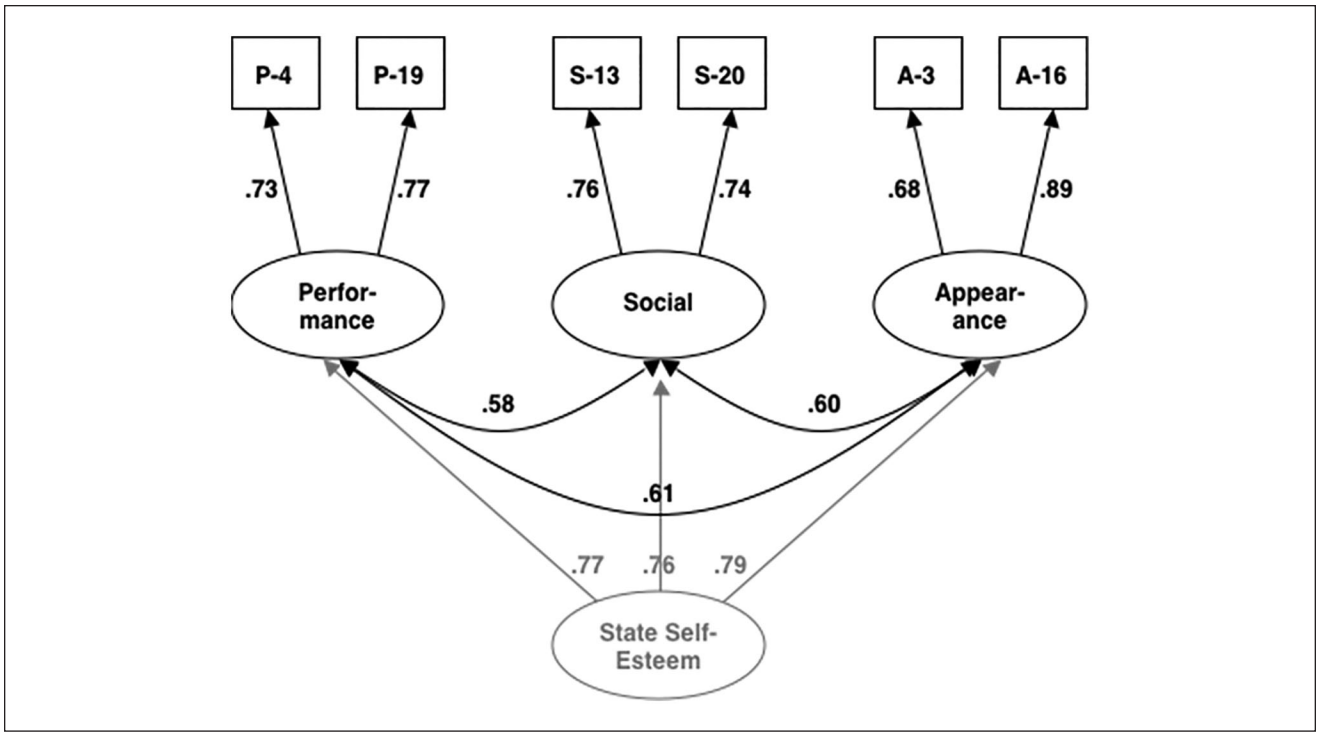

Figure 2. Study 1: Confirmatory factor analyses results for the SSES-6: Three-factor model in black and the hierarchical model in grey (equivalent fits; see Table 2).
Note. See Table 1 for corresponding indicator items (P = Performance, S = Social, A = Appearance, Number = item number). SSES = State Self-Esteem Scale.

between adjacent pairs of dependent correlations are bolded (see rows Single-Item Self-Esteem through BMI). In all, 25 of 72 (34.7%) of dependent correlations differed significantly. In particular, the SSES-6 and its subscales showed a consistent pattern of more modest (less negative) correlations with neuroticism than the SSES-20. Nevertheless, because of fairly large sample sizes, even small differences in correlations were significant. For example, the mean absolute difference in correlations between the two SSES versions was only .041 (SD = 0.032), whereas the mean absolute difference in variance explained (squared correlations) was only .021 (SD = 0.029). Thus, there was little evidence of substantial practical differences between the two SSES versions.

Study 2: Variance Partitioning, Validity, Test–Retest Reliability, and Instability

In Study 2, we compared the 6- and 20-item versions of SSES using a four-wave, repeated-measures study. Study 2 had four main purposes. First, using a series of multilevel measurement models, we examined variance decomposition (item-, day-, and person-level) for both versions of the SSES while using the SISES for comparison (see Nezlek & Gable, 2001; Nezlek & Plesko, 2001). Second, we also used these multilevel models to examine cross-level validity (i.e., "Do state-level self-esteem measures accurately reflect their respective trait-level measures?"). Third, we examined the test–retest reliabilities of the 6- and SSES-20 versions. Fourth, we created temporal measures of self-esteem instability (SD over time; see Kernis et al., 1989; Webster et al., 2017) and examined their convergent validity with a self-report measure of self-esteem instability (SSS). Given the promising results from Study 1, we expected scores from the SSES-6 to perform as well as scores from the SSES-20 across these tests of reliability and validity.

Method

Participants. Participants were 40 students enrolled in an upper-level undergraduate psychology class at the University of Florida in 2012 (12 men, 28 women; ages: 19–23 years, Mdn = 21.0, M = 20.9, SD = 1.0). Sample size was determined solely by the number of students who wished to participate, which consisted of most of the class. Thirty-nine of 40 participants had sufficient data for analyses (i.e., provided two or more waves of data).

Measures and Procedure. Using the same 7-point response scales as Study 1, participants completed state measures of

the SSES-20 and the SISES four times at 2-week intervals (e.g., "Today I felt . . . "). Thus, Times 1 and 4 were 6 weeks apart. Participants also completed one-time trait measure versions of these same constructs (e.g., "In general I feel . . . ") along with the 10-item Rosenberg Self-Esteem Scale (RSES; e.g., "On the whole, I am satisfied with myself") and the 5-item SSS (e.g., "I have noticed that my ideas about myself seem to change very quickly"; Rosenberg, 1965), all using 1 (*strongly disagree*) to 7 (*strongly agree*) response scales. Participants also completed the TIPI (Gosling et al., 2003), but because Studies 1 and 3 focused on personality correlations, and because Study 2 emphasized variance decomposition in the SSES, we left the TIPI data unanalyzed.

Participation was optional, voluntary, and without incentive. Because participants completed all measures in a classroom at the end of the class day, they could leave before the survey was administered by a research assistant who was not the instructor of record, thus avoiding any strong coercion to participate. Due to absenteeism or choices not to participate, there were some sporadic missing data. We used multilevel models with restricted maximum likelihood estimation, which can give accurate estimates with a limited amount of missing data (Nezlek, 2001). We used the program HLM 6.04 (Raudenbush et al., 2004) for all multilevel models (see Raudenbush & Bryk, 2002).

Results and Discussion

Variance Decomposition and Validity. Using three-level multilevel models to decompose item-, day-, and person-level variance (Levels 1, 2, and 3, respectively; see Nezlek & Gable, 2001; Nezlek & Plesko, 2001; Raudenbush & Bryk, 2002), we calculated the proportion of variance at each level for both SSES versions, along with the SISES (supplement Table S2 [available online], variance decomposition columns). But because the SISES has only one item, and thus no item-level variance, we used a two-level model for it. Of the SISES' total variance, 20% was at the day level and 80% was at the person level. Because it has a single item, the SISES confounds item- and day-level variance; its measurement error cannot be empirically separated from day-level variability. In contrast, variance *can* be portioned into three levels—item, day, and person—for multiple-item measures.

Next, we ran three-level "null," intercept-only models on the 6- and SSES-20 versions, where subscale differences were not modeled and no predictors were present (supplemental Table S2 [available online], variance decomposition columns). Because subscale differentiation is key for both SSES versions, 52% to 54% of the variance was at the item level, 3% to 5% was at the day level, and 41% to 45% was at the person level. In contrast, when we modeled the three-factor structure of SSES in both versions, only 18% to 21% of the variance was at the item level, about 8% was at the day level, and 71% to 73% was at the person level. Although the state- or day-level variance of just over 8% may seem comparatively small, it was at least marginally significant for both the SSES-20, $\chi^2(288) = 506.7$, $p < .001$ and the SSES-6, $\chi^2(288) = 323.9$, $p = .071$. We revisit this issue in Study 3 with a sample size over eight times larger. Overall, modeling the different subscales provided a clearer pattern of variance decomposition. Moreover, the state self-esteem subscale that should show the lowest day-to-day variability—appearance—indeed did, presumably because most people's appearance self-esteem fluctuates less than social or performance self-esteem. Importantly, the 6- and SSES-20 performed nearly identically. In addition, the person-level variances for both SSES versions were similar to the person-level variance for the SISES.

We also tested the cross-level (i.e., state–trait) validity of both SSES versions and their respective subscales (Table S2, validity correlations: zero-order; see Nezlek & Gable, 2001; Nezlek & Plesko, 2001). Both the 20- and 6-item SSES and their respective subscales showed high validity correlations (.84 to .98), and these correlations were similar to that of the SISES (.94). We also tested cross-level (state–trait) validity by conducting multiple regressions to see whether each state subscale related to its respective trait subscale and not the two others (i.e., controlling for the two other trait subscales). This was indeed true for both SSES versions (Supplemental Table S2 [available online], validity correlations: partial). For example, the state–trait validity correlations for the SSES-6's subscales were .56, .77, and .61 for performance, social, and appearance self-esteem, respectively. Thus, the day- or state-level measures explained a substantial amount of variance in their respective person- or trait-level measures.

We next examined convergent validity between the state-level SISES measure and both SSES versions and their subscales (Table 3, leftmost columns). All SSES measures explained substantial amounts of variance in the state SISES, and all but one were statistically significant (Table 3, leftmost columns). For example, the state–state validity correlations between the SISIE and SSES-6's three subscales were .67, .74, and .84 for performance, social, and appearance self-esteem, respectively.

We also examined state–trait convergent validity between both SSES versions and two trait measures self-esteem (i.e., SISES & RSES; Table 3, middle and rightmost columns). Both the 20- and 6-item SSES versions showed high validity correlations with both trait self-esteem measures. For example, the state–trait convergent validity correlation between the SSES-6 and the RSES was .89; its subscales—performance (.87), social (.75), and appearance (.84) state self-esteem—also correlated highly with the RSES. Overall, SSES-6 scores—both composite and subscales—showed a consistent pattern of convergent validity with scores from other self-esteem measures, and they did so at multiple levels of analysis.

Table 3. Study 2: Multilevel Modeling Results: Validity Correlations Among Three Self-Esteem Measures.

	State or day level (2)			Trait or person level (3)					
	Single-Item Self-Esteem			Single-Item Self-Esteem			Rosenberg Self-Esteem		
Item level (1)	Coef.	t_{38}	r/r_p	Coef.	t_{37}	r/r_p	Coef.	t_{37}	r/r_p
Twenty-item scale									
All 20 items	0.24	3.14	.56	0.54	10.12	.87	0.61	10.55	.90
Subscales									
Performance	0.24	3.02	.45	0.47	7.03	.79	0.54	8.32	.85
Social	0.28	3.41	.58	0.59	7.36	.78	0.68	7.31	.83
Appearance	0.20	3.33	.53	0.57	9.80	.83	0.61	9.41	.83
Six-item scale									
All six items	0.34	3.87	.999[a]	0.56	8.78	.84	0.64	9.48	.89
Subscales									
Performance	0.47	4.62	.67	0.59	9.02	.85	0.66	9.14	.87
Social	0.47	4.76	.74	0.53	5.01	.67	0.65	5.53	.75
Appearance	0.16	1.59*ns*	.84	0.56	8.51	.82	0.62	8.61	.84

Note. Although we report zero-order (r) and partial (r_p) correlations, these are simply the square-roots of pseudo-R^2s and should be interpreted with caution because such estimates of effect size in multilevel models, which are based on change in variance components between models, may not reflect those from ordinary least squares (OLS) regression (see Kreft & de Leeuw, 1998, p. 119).
[a]For example, the OLS r_p here would be .53 (vs. .999), which is a more reasonable estimate. All other estimates are within a reasonable range.
Coef. = Unstandardized regression coefficient. All t ratios were significant ($ps < .05$) except for *ns* (subscripted).

Test–Retest Reliability. The SESS-6 and its two-item subscales showed acceptable internal consistency reliabilities ($\alpha s > .75$; Supplemental Table S3 [available online], main diagonal). We assessed test–retest reliability by correlating Time 1 and Time 4 measures for both SSES versions (Supplemental Table S3 [available online], bold italics). For the SESS-20, the test–retest correlations were .90 for the composite scale score and in the .80s for its three subscale scores. For the SESS-6, the test–retest correlations were .81 for the composite scale scores and in the .70s for its three two-item subscale scores. Thus, although test–retest reliabilities diminished slightly when comparing the 20- and 6-item SSES versions (likely due in part to the fact that scales with fewer items have lower αs; see Cortina, 1993), the 6-item version and its subscales achieved acceptable levels of test–retest reliability, but with far fewer items. Finally, the correlation matrix also shows the "parent–child" or "long-form–short-form" convergent validity correlations; all were substantial ($rs > .90$; Supplemental Table S3 [available online], bold nonitalics).

Temporal Instability. Can the SSES-6 serve as an effective measure of self-esteem instability over time? We calculated the temporal self-esteem instability coefficients for the SISES and all SSES measures by taking their respective *SD*s over time (up to four waves) within each person (see Kernis et al., 1989; Webster et al., 2017). We then ran correlations between these *temporal* (*SD*-based) self-esteem instability measures for the SSES and a *self-report* self-esteem instability measure, Rosenberg's (1965) SSS, recoded to reflect self-esteem *in*stability (Supplemental Table S4 [available online]). Three findings were noteworthy. First, the temporal instability coefficients for both SSES versions were more highly correlated with the SSS ($rs = .31$ to .56) than was the temporal instability coefficient for the SISES with the SSS ($r = .30$). Second, the temporal instability coefficient for the SSES-20 composite had a higher correlation with the SSS ($r = .56$) than the same for the SSES-6 composite ($r = .37$), though both were significant. Third, the correlations varied widely for the SSES-20 subscales ($rs = .31$ to .52), but were more consistent for the SSES-6 subscales ($rs = .36$ to .39). Overall, these analyses showed that temporal self-esteem instability coefficients (i.e., *SD*s over time) based on either version of the SSES (and its subscales) appeared to be valid measures of trait self-esteem instability (i.e., each correlated positively with an established self-report measure; see Webster et al., 2017).

Study 3: Replicating Structure, Variance Partitioning, Validity, and Test–Retest Reliability

Although Study 2 showed that the SSES-6's scores were fairly valid and reliable, these finding were limited by Study 2's modest sample size ($n = 39$). In Study 3, we remedied this limitation by conducting a replication study with a larger sample ($n = 315$) using secondary data analyses from an unrelated project (see Howell & Shepperd, 2016, Study 1). In addition, we collected two multi-item state self-evaluation

measures to further assess the concurrent validity of the SSES-6's scores. We also examined "parent–child" convergent validity after removing items from the long form that appear in the short form. That is, we removed the six items from the SSES-20—hereafter the SSES-14—that appear in the SSES-6 to allow for a clearer assessment of convergent validity. Attempting to replicate aspects of Study 1, we again tested the SSES-6's structure with a series of CFAs at each time point. Finally, we examined the extent to which Big Five personality traits related to average state self-esteem levels. Based on Study 1's findings, we expected positive relations between the SSES-6 (and its subscales) and both (a) extraversion and (b) emotional stability, which is the inverse of neuroticism.

Method

We drew Study 3's data from Howell and Shepperd's (2016) Study 1, which was conducted to develop a measure of information avoidance tendencies. Howell and Shepperd used the SSES-20 as a whole (no subscales) and averaged across its 4 time points; they used the SSES-20 to assess convergent validity of scores from their information avoidance measure, along with 18 other measures. Because the SSES-20 was not a focus of Howell and Shepperd, the findings presented below are entirely new and reproduce none of the results from their work, even though the sample is the same as their Study 1. Please see Howell and Shepperd (2016) Study 1, for a description of all measures.

Participants. Participants were 315 students enrolled in introductory psychology classes at the University of Florida in 2011 who received research credit for participating (69 men, 244 women, 2 did not respond; ages: 17–23 years, $Mdn = 18.0$, $M = 18.34$, $SD = 0.79$). Regarding race and ethnicity, 56% were White/Caucasian, 14% Black/African American, 13% Hispanic/Latinx, 10% Asian, 1% Native American or Pacific Islander, 5% other or unreported. (Howell & Shepperd's [2016] Study 1 sample had 316 students, but we excluded one person who provided no SESS-20 data.)

Measures and Procedure. Using the same 7-point response scales as Studies 1 and 2, participants completed state measures of the SSES-20 at 2-week intervals (e.g., "Today I felt . . . "). Thus, Times 1 and 4 were 6 weeks apart. Participants also completed four biweekly versions of (a) the 10-item RSES (1965; e.g., "On the whole, I am satisfied with myself"), (b) the 25-item Coping Self-Efficacy Scale (CSE; e.g., "Make a plan of action and follow it when confronted with a problem"; Chesney et al., 2006), and (c) the TIPI (Gosling et al., 2003), using 1 (*strongly disagree*) to 7 (*strongly agree*) response scales. The TIPI uses the positively valenced term "emotional stability," whereas other Big Five measures use the negatively valenced term "neuroticism" to assess the same construct. Although we reverse-scored the TIPI's emotional stability scale to reflect neuroticism in Study 1 to be consistent with two other Big Five measures' neuroticism scales, in Study 3, we use the TIPI's originally intended term. Thus, higher scores on this dimension reflect greater emotional stability (or alternatively, less neuroticism).

Participants completed all measures online and order of measures was counterbalanced at each time point. Similar to Study 2, Study 3 had some sporadic missing data; we again used multilevel models with restricted maximum likelihood estimation, which can give accurate estimates with a limited amount of missing data (Nezlek, 2001). We again used HLM 6.04 (Raudenbush et al., 2004) for all multilevel models (Raudenbush & Bryk, 2002).

Results and Discussion

Confirmatory Factor Analyses. To be thorough, we sought to further replicate the SSES-6's three-factor structure based on its three two-item subscales (see Study 1) at each of 4 time points. Once again, we first tested a three-factor CFA for the SSES-6 (Table 2, bottom). Recall that in this case, a hierarchical or second-order-factor model yields the same fit as a traditional three-factor one. The three-factor/hierarchical model fit the data well; all fit indices were acceptable-to-good (Table 2, bottom). We then fit a simpler, one-factor model to the data at each wave; these models fit the data poorly and the differences between them and the original three-factor models were all significant (Table 2, bottom). Thus, the three-factor (or hierarchical) structure of the SSES-6 was again supported.

Variance Decomposition and Validity. Using three-level multilevel models to decompose item-, day-, and person-level variance (Levels 1, 2, and 3, respectively; see Nezlek & Gable, 2001; Nezlek & Plesko, 2001; Raudenbush & Bryk, 2002), we calculated the proportion of variance at each level for both SSES versions (Supplemental Table S5 [available online], variance decomposition columns). First, we ran three-level "null," intercept-only models on the 6- and SSES-20 versions, where subscale differences were not modeled and no predictors were present. Similar to Study 2, because subscale differentiation is key for the SSES in both versions, 60% to 64% of the variance was at the item level, 5% to 7% was at the day level, and 29% to 35% was at the person level. In contrast, when we modeled the three-factor structure of SSES in both versions, only 21% to 29% of the variance was at the item level, about 12% to 13% was at the day level, and 59% to 66% was at the person level. Although the day-level variance of about 12% to 13% may seem comparatively small, it was significant for both the SESS-20 and -6, $\chi^2(2322) = 4451.2$ and 3059.3, respectively; $ps < .001$), and likely reflects a more robust estimate of day-level variance than

Table 4. Study 3: Multilevel Modeling Results: Validity Correlations Between the State Self-Esteem Scales and Two Self-Evaluation Measures.

	State or day level (2)					
	Rosenberg Self-Esteem			Coping Self-Efficacy		
Item level (1)	Coef.	t_{314}	r/r_p	Coef.	t_{314}	r/r_p
Twenty-item scale						
All 20 items	0.44	6.46	.62	0.22	6.81	.51
Subscales						
Performance	0.55	7.12	.59	0.25	6.58	.44
Social	0.36	3.93	.59	0.19	4.21	.50
Appearance	0.41	6.18	.60	0.25	6.87	.55
Six-item scale						
All six items	0.41	4.98	.73	0.21	5.12	.68
Subscales						
Performance	0.60	5.40	.44	0.22	3.42	.43
Social	0.28	2.60	.63	0.18	3.22	.58
Appearance	0.38	4.20	.72	0.25	5.30	.68

Note. Although I report zero-order (r) and partial (r_p) correlations, these are simply the square-roots of pseudo-R^2s and should be interpreted with caution because such estimates of effect size in multilevel models, which are based on change in variance components between models, may not reflect those from ordinary least squares (OLS) regression (see Kreft & de Leeuw, 1998, p. 119). All t ratios were significant ($ps < .05$). Coef. = Unstandardized regression coefficient.

Study 2's small sample. Modeling the different subscales again provided a clearer pattern of variance decomposition. Moreover, the state self-esteem subscale that should show the lowest day-to-day variability—appearance—did, again presumably because physical appearance self-esteem is less likely than social or performance self-esteem to fluctuate greatly over short periods of time. Importantly, the 6- and SSES-20 performed nearly identically.

We next examined state–state convergent validity between the SSES (both versions and their subscales) and two other self-evaluation measures: RSES and CSE (Table 4). Both the RSES and the CSE explained substantial amounts of variance in both versions of the SSES and its subscales; all state–state convergent validity correlations were significant. For example, the RSES correlated .73 with the SSES-6 composite, and .44, .63, and .72 with its performance, social, and appearance subscales, respectively. Similarly, the CSE scale correlated .68 with the SSES-6 composite, and .43, .58, and .68 with its performance, social, and appearance subscales, respectively. Thus, the SSES-6 and its three subscales (a) covaried with two other self-evaluation measures that were also assessed biweekly across 4 time points (rs = .73 to .43) and (b) the SSES-6 showed nearly the same pattern of convergent validity correlations as the SSES-20 did (rs = .62 to .44). In sum, the SSES-6's scores showed evidence of state–state convergent validity.

Test–Retest Reliability. First, the SESS-6 and subscales showed acceptable internal consistency reliabilities for two-item scales (αs > .65; Supplemental Table S6 [available online], main diagonal). We assessed test–retest reliability by correlating Time 1 and Time 4 measures for both SSES versions (Supplemental Table S6 [available online], bold italics). For the SESS-20, the test–retest correlations were .66 for the composite scale score and averaged .63 for its three subscale scores: performance (.55), social (.62), and appearance (.70). For the SESS-6, the test–retest correlations were .63 for the composite scale scores and averaged .57 for its three two-item subscale scores: performance (.42), social (.59), and appearance (.67). Although test–retest reliabilities diminished slightly when comparing the 20- and 6-item SSES versions, the 6-item scale maintained an acceptable level of test–retest reliability with far fewer items (Cortina, 1993). In addition, the correlation matrix also showed the "parent–child" or "long-form–short-form" convergent validity correlations; all were substantial (rs ≥ .84; Supplemental Table S6 [available online], bold nonitalics).

"Parent–Child" Convergent Validity With Mutually Exclusive Items. Using Time 1's larger sample (n = 313), we also examined "parent–child" convergent validity correlations after removing items in the parent version also contained in the child version (i.e., correlating the SSES-14 with the SSES-6 along with their respective subscales). The "parent–child" convergent validity correlations were large for both the composite scale scores (r = .88) and for the performance, social, and appearance subscale scores (rs = .66, .77, and .84, respectively; $ps < .001$).

Big Five Personality. Attempting to replicate Study 1's findings, we examined the extent to which Big Five personality traits related to average scores on the SSES-6 and its subscales. According to Whole Trait Theory (Fleeson, 2001; Fleeson & Jayawickreme, 2015), traits can be conceived of as an aggregate of multiple states over time or across situations, and because personality is more often described as a trait than a state, we took this approach. Thus, we used four multilevel models to regress the SSES-6 and its three subscales onto the Big Five personality trait measures as simultaneous predictors (Table 5). In line with Whole Trait Theory, we created *trait* measures of personality for each person by averaging across their biweekly *state* measures of personality.

Replicating Study 1's results for neuroticism, trait emotional stability positively related to both the SSES-6 composite score (r_p = .47) and its performance (r_p = .38), social (r_p = .33), and appearance (r_p = .37) subscales. Also largely replicating Study 1's results, trait conscientiousness positively related to both the SSES-6 composite score (r_p = .18) and its performance (r_p = .19) and appearance (r_p = .15) subscales. Once again, these results stand to reason because more conscientious people may be especially concerned

Table 5. Study 3: Multilevel Modeling Results: Six-Item State Self-Esteem Scale and Subscales as Functions of Trait Personality.

	State Self-Esteem Scale: Item level (1)											
	All subscales			Performance			Social			Appearance		
Person level (3)	b	t_{309}	r_p	b	t_{309}	r_p	b	t_{309}	r_p	b	t_{309}	r_p
Extraversion	0.13	3.16*	.18	0.08	1.58	.09	0.24	3.76*	.21	0.08	1.51	.09
Agreeableness	−0.09	−1.60	−.09	−0.12	−1.71	−.10	−0.06	−0.75	−.04	−0.07	−0.86	−.05
Conscientiousness	0.16	3.20*	.18	0.21	3.34*	.19	0.09	1.19	.07	0.16	2.58*	.15
Emotional Stability	0.43	9.47*	.47	0.43	7.32*	.38	0.40	6.13*	.33	0.45	6.92*	.37
Openness	0.03	0.64	.04	0.08	1.19	.07	0.02	0.30	.02	−0.02	−0.21	−.01

Note. Results from two models: (1) All subscales and (2) Performance, Social, and Appearance. Personality traits were entered simultaneously (multiple regression) at the person level (Level 3).
*$p < .05$.

about (a) their day-to-day work, school, or professional performance and (b) managing their physical appearance.

In addition to these two anticipated findings, we also observed two novel-but-meaningful results for extraversion, which positively related to both the SSES-6 composite score ($r_p = .18$) and social self-esteem ($r_p = .21$), but neither performance nor appearance self-esteem (r_ps = .09), suggesting that extraverts tended to report higher state self-esteem in general, and social state self-esteem largely drove this effect. These results for extraversion help highlight the convergent and discriminant validity of scores from our two-item state self-esteem subscales. More generally, these findings further corroborate the validity of SSES-6's scores in the broader nomological network of the Big Five.

Study 4: Laboratory Aggression Experiment

Together, Studies 1 to 3 examined and supported the psychometic robustness of both versions of the SSES, while highlighting the efficacy and efficiency of the SSES-6. In Study 4, we extend the SSES-6 into the behavioral domain to see whether it can reproduce a self-esteem-level-by-self-esteem-instability interaction in predicting behavioral aggression in a laboratory setting, whereby there is a negative relationship between self-esteem level and aggression, but only among people with unstable self-esteem (see Webster et al., 2007; but see also Kernis et al., 1989). And because the participants were asked to write about—and then received critical feedback on—their global selves (details below; see also Webster & Kirkpatrick, 2006), we tested the expected interaction using the global or composite SSES-6 instead of its three domain-specific subscales.

Method

We performed secondary analyses on data originally collected to replicate another study (Kernis et al., 1989) and test a self-esteem-level-by-self-esteem-instability interaction in predicting behavioral aggression (Webster et al., 2007). Thus, although these data were not originally collected with the SSES-6 in mind, they do lend themselves to this purpose. These data were collected during the 2002–2003 academic year (three semesters), and sample size was not determined ahead of time; we merely collected as much data as we could before the lead author's second year of graduate school ended. In addition to the measures outlined below that are the focus of the present research, we also collected several other measures that were more germane to the original replication effort; these included mostly self-report measures described elsewhere (i.e., Kirkpatrick et al., 2002; Webster & Kirkpatrick, 2006).

Participants. Participants were 68 introductory psychology students at the University of Colorado Boulder who received credit toward a class research requirement in exchange for their participation. One participant expressed suspicion about the essay feedback manipulation during debriefing and was excluded from analyses. The remaining 67 participants were 34 men and 33 women ranging in age from 18 to 24 years ($Mdn = 19$, $M = 19.30$, $SD = 1.29$).

Measures. Self-esteem level and instability response scales ranged from 1 (*strongly disagree*) to 9 (*strongly agree*). Self-esteem level was measured using 18 items from the SSES-20 (Heatherton & Polivy, 1991); the six items with the highest factor loadings from each of three subscales (from Heatherton & Polivy, 1991)—performance, appearance, and social self-esteem—were chosen, and these included all SSES-6 items. Self-esteem instability was measured using a three-item version of the five-item SSS (Rosenberg, 1965; i.e., "I find that one day I have one opinion of myself and on another day I have a different opinion," "My opinion of myself tends to change a good deal instead of always remaining the same," and "I change from a very good opinion of myself to a very poor opinion of myself").

Procedure. Participants took part in an aggression experiment in groups of three to six. The procedures outlined below follow those used by Kirkpatrick et al. (2002), which used an essay feedback manipulation by Bushman and Baumeister (1998) in conjunction with the hot sauce aggression paradigm developed by Lieberman et al. (1999; see also Webster & Kirkpatrick, 2006). Students were told they were participating in a study about self-attitudes and taste preferences. On arriving at the laboratory, the experimenter led participants into separate rooms off of a central room to prevent social interaction.

For the supposed self-attitudes phase of the experiment, participants received a single sheet of paper with the following instructions:

> The essay that you write will be read and evaluated by two other students taking part in this study, and you will be asked to read and evaluate the essays of two other students. You will also get to view each other's evaluations. Using the space below, please write a brief, one-paragraph essay on what you would like to be doing with your life five years from now. Please take no more than about five or six minutes. When you are finished, please open your door slightly to let us know that you have completed your essay.

This essay topic was chosen due to its general self-relevance. The completed essays were then taken away, and the participants were led to believe that they were being partnered with two other participants of the same gender (who were in other rooms off of the same main room) and that they would be evaluating each other's essays in a round-robin fashion. Participants were also told that their evaluations of their partners' essays would be exchanged with their partners, so that each participant would see the feedback that their partners had provided on their essays.

Participants then received an essay supposedly written by one of their partners (which was prepared earlier by the experimenter) and evaluated this essay on six items—*organization*, *content*, *writing style*, *clarity of expression*, *thoughtfulness*, and *overall quality*—using a 7-point scale from –3 (*poor*) to 3 (*excellent*). Participants then saw another bogus essay from their other partner, and were asked to provide another set of ratings using the same items and response scales. Thus, each participant rated two essays, and later, each participant received feedback on their essay ostensibly from two other participants. Soon after, the experimenter returned and gave participants the two evaluations of their essays that their partners had presumably produced. One of these evaluations constituted a threat to the self (the essay feedback manipulation). Each participant received both a negative essay rating with a scale mean of −2.0 with a handwritten comment ("Weak essay. I didn't like it.") and a positive essay rating with a scale mean of +2.0 and a different handwritten comment ("Great essay. No suggestions.").

Next, in the supposed taste-preferences phase of the experiment, participants were asked to prepare, taste, and evaluate food samples. To give participants the illusion of realism, the experimenter asked each participant to say either "spicy" or "dry." Regardless of the participant's response, the experimenter replied with, "You have been randomly assigned to receive a dry food sample from your partners and you will be asked to prepare a spicy food sample for them." Participants were asked to prepare food samples for each other because the experimenter (ostensibly) needed to be blind to the type and quantity of food tasted. Participants were told that they were paired with the same two partners from the essay phase of the experiment, ostensibly to avoid experimenter confusion.

Participants next completed a taste preference inventory, on which they reported their liking for *salty*, *spicy*, *dry*, *sweet*, *sour*, and *creamy* foods on 21-point scales ranging from 1 (*extreme disliking*) to 21 (*extreme liking*). After a few minutes, the experimenter returned to collect the participants' taste preference inventories, presumably to deliver them to their partners. About 5 minutes later, the experimenter returned with a single saltine cracker in each of two envelopes (dry food samples), which were ostensibly prepared by each participant's respective partners. Participants were asked to consume the entire contents of each envelope and evaluate its taste using 9-point scales ranging from 1 (*complete disliking*) to 9 (*extreme liking*) for five items: *appearance*, *aroma*, *taste*, *texture*, and *overall satisfaction*.

About 5 minutes thereafter, the experimenter returned with a tray containing the hot sauce allocation materials, including a large container of hot sauce (prepared using the recipe specified by Lieberman et al., 1999), two 12-fluid-ounce (355-ml) Styrofoam bowls into which the hot sauce sample was allocated (one bowl for each partner), a small spoon for participants to taste-test the hot sauce themselves, a large spoon for transferring hot sauce from the large container into each sample container, and two bogus taste preference inventories, which ostensibly came from each participant's two partners. This form indicated both participant's fictional partners held a strong dislike for spicy foods, which were given a rating of 3 (out of 21) and were rated lowest among the six taste-preference items. Participants were given a checklist that outlined the procedure they were to follow.

Participants were first asked to taste small spoonful of the hot sauce so they would know the sauce's intensity. They then prepared separate samples of hot sauce for their partners to consume. Participants used a plastic spoon to transfer any amount of the hot sauce from the large container into each of the two Styrofoam bowls and sealed each with a lid. They were told any amount of hot sauce was useful to the experiment, and they should allocate as little or as much as they desired for each of their partners. It was made clear to participants that their partners would be asked to

consume the entire amount of hot sauce they allocated with some tortilla chips. Participants also wrote each partner's experiment identification number on their respective hot sauce sample containers. Soon after, the experimenter returned to collect and deliver the hot sauce samples to each participant's two partners. Instead, the hot sauce samples were weighed on a digital scale.

Minutes later, the experimenter returned to probe participants for suspicion by asking them whether they believed the essay feedback they had received had come from their two partners. Each participant was thoroughly debriefed on the purpose of the experiment, asked if they had any questions or concerns, and thanked for their participation.

Results and Discussion

Correlations and descriptive statistics appear in Table S7. Overall, internal consistency reliability was good (αs > .70) except for the two-item version of social state self-esteem (α = .52). Although not ideal, recall that alpha is a function of both the mean interitem correlation (which, at .36, was good) and the number of scale items (only 2), and thus shorter scales have smaller αs (see Cortina, 1993). But also recall that a goal of Study 1's IRT models was to choose items that represented the full breadth of the latent trait, and thus two-item scales (vs. longer ones) should have lower interitem correlations because they are attempting to capture a broad swath of the construct of interest without redundant items (Gosling et al., 2003).

The essay feedback experimental manipulation was successful: Participants allocated significantly more hot sauce (log grams) to their negative evaluator (M = 2.85, SD = 1.19) than to their positive evaluator (M = 2.38, SD = 1.07), paired $t(66) = 5.30$, $p < .001$, $d = 0.65$.

To test the predicted interaction effect, we ran a two-step hierarchical multiple regression analysis, regressing hot sauce weight (log grams) given to the negative evaluator onto the predictors, which were mean-centered (Table 6). The first step (Table 6, Model 1) revealed the predicted state self-esteem level × self-esteem instability interaction (Figure 3). In the second step, controlling for log hot sauce allocated to the positive evaluator—and its interaction with state self-esteem, instability, and their three-way interaction—did not affect the focal self-esteem × instability interaction (Table 6, Model 2; see Yzerbyt et al., 2004).

Using Model 1 and following established procedures for decomposing interactions (Cohen et al., 2003; Judd et al., 2017), we conducted simple effects tests at 1 SD above and below the mean instability score (see Aiken & West, 1991). For people with stable self-esteem (−1 SD), the simple relationship between self-esteem level and behavioral aggression was nonsignificant, $b = 0.14$, $t(63) = 0.76$, $p = .45$, $r_p = .09$; however, for people with unstable self-esteem (+1 SD), the simple relationship was significantly negative, $b = -0.34$, $t(63) = -2.33$, $p = .023$, $r_p = -.28$. These results not only supported the predicted spreading interaction between self-esteem level and instability, but also showed that the SSES-6 can be a viable measure in laboratory experiments.

Table 6. Study 4: Hot Sauce (Log Grams) Allocated to the Negative Evaluator as Function of State Self-Esteem (Six-Item) and Instability of Self-Esteem (Model 1) and Controlling for Hot Sauce Allocated to Positive Evaluator (Model 2).

Model or variable	*b*	*t*	r_p
Model 1			
(In)Stability of Self Scale (SSS)	−0.046	−0.63	−.08
Six-Item State Self-Esteem Scale (SSES-6)	−0.102	−0.84	−.11
SSS × SSES-6 Interaction	−0.113	−2.16*	−.26
Model 2			
(In)Stability of Self Scale (SSS)	0.018	0.37	.05
Six-Item State Self-Esteem Scale (SSES-6)	0.003	0.04	.00
SSS × SSES-6 Interaction	−0.076	−2.13*	−.27
Hot sauce (+)	0.899	9.79*	.79
Hot sauce (+) × SSS	−0.008	−0.17	−.02
Hot sauce (+) × SSES-6	0.019	0.25	.03
Hot sauce (+) × SSS × SSES-6	0.042	1.39	.18

Note. N = 67. (+) = Hot sauce weight (log grams) given to positive evaluator.
*$p < .05$.

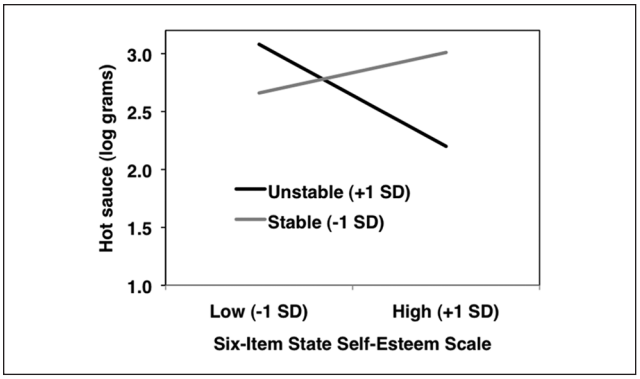

Figure 3. Study 4: Hot sauce (log grams) allocated to the negative evaluator as functions of the SSES-6 and self-esteem instability (see Table 6).
Note. SSES = State Self-Esteem Scale.

General Discussion

Across four studies, we set out to develop and test a brief six-item version of the SSES-20. Overall, the present studies showed that scores from the new SSES-6 showed both good reliability (both internally consistent and test–retest) and validity (via convergent validity). In addition, CFAs

suggested that the SSES-6 successfully reproduced the three-factor structure of its parent measure. The SSES-6 was also effective in a laboratory experiment, supporting the notion the people with high, stable self-esteem often report or exhibit the least aggression—or that people with low, unstable self-esteem report or show the most. The present findings support the efficacy of the SSES-6 as a viable substitute for the longer SSES-20 in situations that demand efficient assessment of state self-esteem.

Practical Applications

Because researchers are increasingly conducting studies that require brief measures, there is a parallel growing demand for efficient-yet-robust multi-item scales (Widaman et al., 2011). Study designs that often demand brief measures include daily-diary studies (Nezlek, 2001), experience-sampling studies (Christensen et al., 2003; Scollon et al., 2003), longitudinal studies, and large collaborative international or multilaboratory studies that often assess several constructs, and hence place a premium on the number of items per construct (Klein et al., 2014, Klein et al., 2018). Although the SSES-6 has several practical research applications, researchers should continue to use the original SSES-20 when time and resources permit, simply because the former's two-item subscales simply cannot match the broader assessment bandwidth offered by the latter. Indeed, the SSES-6 provides researchers with a new tool that fills a niche between the single-item, domain-general SISES (Robins et al., 2001) and the 20-item, domain-specific SSES (Heatherton & Polivy, 1991).

A second practical advantage of the SESS-6 is its flexibility because it can be scored as either one global measure of self-esteem or three separate domains of self-esteem: performance, social, and appearance. Recall that in Studies 1 and 3, a hierarchical or second-order factor structure (favoring a global approach) fit the six items just as well as a three-factor structure (favoring a domain-specific approach). But also recall that these models were empirically indistinguishable in terms of fit. Furthermore, findings from Studies 1 to 3 tended to favor the advantages of the SESS-6's domain specificity via its subscales, whereas findings from Study 4 showcased the SESS-6 as a composite measure general state self-esteem.

Finally, although Study 4 used a 9-point response scale for the SSES, we believe that the seven-point response scale (used in Studies 1 to 3) provides participants with sufficient response breadth and is our preferred response format for the SSES-6.

Strengths, Limitations, and Future Directions

Although the present findings make a convincing case for the SSES-6's efficacy and applicability, some words of caution are in order. First, the SSES-6 needs evaluation in broader non-self-report contexts. Many self-report measures suffer from acquiescence bias and socially desirable responding (Paulhus & Vazire, 2007); the SSES-6 is likely no exception. And because most people responding to self-esteem scales tend to see themselves in an especially positive light, responses to the SSES-6 likely show a better-than-average effect (Alicke et al., 1995), where people report their self-esteem as higher than it may be in truth. One way to address some of the limitations of self-reports is to supplement them with peer reports (e.g., roommates reporting on their own and each other's self-esteem). Another way would be to focus on behavioral correlates of self-esteem (see Baumeister et al., 2007), such as BMI and behavioral aggression (hot sauce), which we used in Study 4. Finally, we again suggest using the SSES's 20-item version in situations where researchers are unconcerned with time or space constraints, because the SSES-20's additional items provide construct bandwidth or content breadth that the SSES-6 simply cannot match.

Second, samples sizes of Studies 2 and 4 (ns = 39 and 67) were relatively modest. Although small, recall that both of these studies featured within-person, repeated-measures designs, which increases statistical power because residual error can be partitioned into within- and between-person sources. Specifically, Study 2 featured a four-wave longitudinal design, whereas Study 4 featured an experimental design whereby participants received either negative or positive feedback and then allocated hot sauce to both their negative and positive evaluator (the repeated-measures component). Consequently, with only modest samples sizes, Studies 2 and 4 nevertheless benefitted from increased power due to their within-person designs. But then again, statistically significant continuous-by-continuous-variable interactions are notoriously difficult to detect (vs. 2 x 2 categorical designs; McClelland & Judd, 1993), so Study 4's interaction findings should be interpreted with caution despite being in the expected direction. In addition, we largely replicated Study 2's results in Study 3 with a sample of over 300 people. Nevertheless, we caution readers about interpreting any "small" correlational effect sizes less than .10 in absolute magnitude (Cohen, 1992), because by definition, such correlations explain less than 1% of the variance in the outcome, which could be practically meaningless.

Third, the present findings were based on undergraduate samples drawn exclusively from large public universities in two regions of the United States. Consequently, the generalizability of the findings is limited (see Henrich et al., 2010). Future research should examine the SSES-6 in broader, more diverse populations, including people from multiple cultures, multiple countries, and broader socioeconomic backgrounds.

Additional research may be necessary to further examine the validity, reliability, and generalizability of the SSES-6's scores. For example, in Study 1, we chose to focus on one-dimensional IRT analyses for the sake of parsimony in

selecting the "best" two items from each subscale. Future research should also assess the SSES-6 in multidimensional IRT (MIRT) contexts using multidimensional GRMs (Ackerman et al., 2003; Manapat et al., 2019). Along these lines, future research should also examine differential item and scale functioning in an IRT or MIRT context (Morizot et al., 2007) to examine group differences in item or scale discrimination and difficulty. Similarly, future research could also use CTT and SEM approaches to examine group differences in measurement invariance. For example, do native and nonnative English speakers respond to the items in the same ways (Webster et al., 2014, 2015)? And do people with clinical depression show the same state self-esteem factor structure as do people who are not depressed?

Conclusions and Theoretical Implications

One theoretical implication of this research is that it supports a domain-specific view of self-esteem (Kirkpatrick & Ellis, 2001). From this perspective, self-esteem need not necessarily be thought of as a global construct, but rather consisting of multiple domains, such as social, performance, and appearance self-esteem. For example, a bad day at work or school might affect performance self-esteem, but perhaps not social or appearance self-esteem. Similarly, a positive first date may do wonders for one's social—and possibly appearance—self-esteem, but is unlikely to affect one's performance self-esteem. Indeed, when researchers threatened different domains of people's self-esteem (via targeted experimental manipulation), those threatened domains of self-esteem interacted with threat (vs. control) to predict increased aggression in a laboratory setting (Kirkpatrick et al., 2002). Thus, a domain-specific view of self-esteem (vs. a global one) offers a more nuanced and more diagnostic way to understand the causes, correlates, and potential consequences of changes in state or trait self-esteem as one interacts with their social environment. Thus, we encourage researchers to consider adopting this domain-specific view of self-esteem, and hope that the SSES-6 can provide a new and efficient tool for doing so.

Authors' Note

Presentations based on data in this article were given at the annual or biennial meetings of the Association for Research in Personality (June 2013, Charlotte, NC) and the Society of Experimental Social Psychology (September 2013, Berkeley, CA). Data used in Study 3 are from Howell and Shepperd's (2016) Study 1; however, the results in the present article's Study 3 are entirely new and reproduce none of those from Howell and Shepperd.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Gregory D. Webster https://orcid.org/0000-0001-7342-8444

Supplemental Material

Supplemental material for this article is available online.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*(5), 804-825. https://doi.org/10.1037/0022-3514.68.5.804

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: What happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396-403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology*, *75*(1), 219-229. https://doi.org/10.1037//0022-3514.75.1.219

Chesney, M. A., Neilands, T. B., Chambers, D. B., Taylor, J. M., & Folkman, S. (2006). A validity and reliability study of the coping self-efficacy scale. *British Journal of Health Psychology*, *11*(3), 421-437. https://doi.org/10.1348/135910705X53155

Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., & Kaschub, C. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, *4*(1), 53-78. https://doi.org/10.1023/A:1023609306024

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192-203. https://doi.org/10.1037/1040-3590.18.2.192

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011-1027. https://doi.org/10.1037/0022-3514.80.6.1011

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*(June), 82-92. https://doi.org/10.1016/j.jrp.2014.10.009

French, S. A., Perry, C. L., Leon, G. R., & Fulkerson, J. A. (1996). Self-esteem and change in body mass index over 3 years in a cohort of adolescents. *Obesity Research*, *4*(1), 27-33. https://doi.org/10.1002/j.1550-8528.1996.tb00509.x

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504-528. https://doi.org/10.1016/S0092-6566(03)00046-1

Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology*, *60*(6), 895-910. https://doi.org/10.1037/0022-3514.60.6.895

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Howell, J. L., & Shepperd, J. A. (2016). Establishing an Information Avoidance Scale. *Psychological Assessment*, *28*(12), 1695-1708. https://doi.org/10.1037/pas0000315

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), 1-55. https://doi.org/10.1080/10705519909540118

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the Dark Triad. *Psychological Assessment*, *22*(2), 420-432. https://doi.org/10.1037/a0019265

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach* (3rd ed.). Routledge.

Kernis, M. H., Grannemann, B. D., & Barclay, L. C. (1989). Stability and level of self-esteem as predictors of anger arousal and hostility. *Journal of Personality and Social Psychology*, *56*(6), 1013-1022. https://doi.org/10.1037/0022-3514.56.6.1013

Kirkpatrick, L. A., & Ellis, B. J. (2001). An evolutionary approach to self-esteem: Multiple domains and multiple functions. In G. J. O. Fletcher & M. S. Clark (Eds.), *The Blackwell handbook of social psychology: Vol. 2: Interpersonal processes* (pp. 411-436). Blackwell. [Reprinted (2004) in M. B. Brewer & M. Hewstone (Eds.), *Self and social identity* (pp. 52-77). Blackwell.] https://doi.org/10.1002/9780470998557.ch16

Kirkpatrick, L. A., Waugh, C. E., Valencia, A., & Webster, G. D. (2002). The functional domain-specificity of self-esteem and the differential prediction of aggression. *Journal of Personality and Social Psychology*, *82*(5), 756-767. https://doi.org/10.1037//0022-3514.82.5.756

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142-152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, S., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490. https://doi.org/10.1177/2515245918810225

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.

Lee, I. A., & Preacher, K. J. (2013, September). *Calculation for the test of the difference between two dependent correlations with one variable in common* [Computer software]. http://quantpsy.org

Lieberman, J. D., Solomon, S., Greenberg, J., & McGregor, H. A. (1999). A hot new way to measure aggression: Hot sauce allocation. *Aggressive Behavior*, *25*(5), 331-348. https://doi.org/10.1002/(SICI)1098-2337(1999)25:5<331::AID-AB2>3.0.CO;2-1

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99. https://doi.org/10.1037/1082-989X.4.1.84

Manapat, P. D., Edwards, M. C., MacKinnon, D. P., Poldrack, R. A., & Marsch, L. A. (2019). A psychometric analysis of the Brief Self-Control Scale. *Assessment*. Advance online publication. https://doi.org/10.1177/1073191119890021

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376-390. https://doi.org/10.1037/0033-2909.114.2.376

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). Guilford Press.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide* (6th ed.). Muthén & Muthén.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event-and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*(7), 771-785. https://doi.org/10.1177/0146167201277001

Nezlek, J. B., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day-to-day psychological adjustment. *Personality and Social Psychology Bulletin*, *27*(12), 1692-1704. https://doi.org/10.1177/01461672012712012

Nezlek, J. B., & Plesko, R. M. (2001). Day-to-day relationships among self-concept clarity, self-esteem, daily events, and mood. *Personality and Social Psychology Bulletin*, *27*(2), 201-211. https://doi.org/10.1177/0146167201272006

Nichols, A. L., & Webster, G. D. (2013). The Single-Item Need to Belong Scale. *Personality and Individual Differences*, *55*(2), 189-192. https://doi.org/10.1016/j.paid.2013.02.018

Nichols, A. L., & Webster, G. D. (2015). Designing a brief measure of social anxiety: Psychometric support for a three-item

version of the Interaction Anxiousness Scale (IAS-3). *Personality and Individual Differences*, *79*(June), 110-115. https://doi.org/10.1016/j.paid.2015.01.043

Okada, R. (2010). A meta-analytic review of the relation between self-esteem level and self-esteem instability. *Personality and Individual Differences*, *48*(2), 243-246. https://doi.org/10.1016/j.paid.2009.10.012

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224-239). Guilford Press.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203-212. https://doi.org/10.1016/j.jrp.2006.02.001

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows* [Computer software]. Scientific Software International.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151-161. https://doi.org/10.1177/0146167201272002

Rodriguez, L. M., & Webster, G. D. (2020). The Three-Item Thinking about Your Partner's Drinking Scale (TPD-3): Item response theory, reliability, and validity. *Journal of Marital and Family Therapy*, *46*(3), 471-488. https://doi.org/10.1111/jmft.12399

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1-97 (1969). https://doi.org/10.1007/BF03372160

Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 123-136). Chapman and Hall/CRC. https://doi.org/10.1201/9781315374512

Scollon, C. N., Prieto, C. K., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strength and weaknesses. *Journal of Happiness Studies*, *4*(1), 5-34. https://doi.org/10.1023/A:1023605205115

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245-251. https://doi.org/10.1037/0033-2909.87.2.245

Webster, G. D., & Crysel, L. C. (2012). "Hit me, maybe, one more time": Brief measures of impulsivity and sensation seeking and their prediction of blackjack bets and sexual promiscuity. *Journal of Research in Personality*, *46*(5), 591-598. https://doi.org/10.1016/j.jrp.2012.07.001

Webster, G. D., DeWall, C. N., Pond, R. S., Jr., Deckman, T., Jonason, P. K., Le, B. M., Nichols, A. L., Schember, T. O., Crysel, L. C., Crosier, B. S., Smith, C. V., Paddock, E. L., Nezlek, J. B., Kirkpatrick, L. A., Bryan, A. D., & Bator, R. J. (2014). The Brief Aggression Questionnaire: Psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior*, *40*(2), 120-139. https://doi.org/10.1002/ab.21507

Webster, G. D., DeWall, C. N., Pond, R. S., Jr., Deckman, T., Jonason, P. K., Le, B. M., Nichols, A. L., Schember, T. O., Crysel, L. C., Crosier, B. S., Smith, C. V., Paddock, E. L., Nezlek, J. B., Kirkpatrick, L. A., Bryan, A. D., & Bator, R. J. (2015). The Brief Aggression Questionnaire: Structure, validity, reliability, and generalizability. *Journal of Personality Assessment*, *97*(6), 638-649. https://doi.org/10.1080/00223891.2015.1044093

Webster, G. D., & Jonason, P. K. (2013). Putting the "IRT" in "dirty": Item response theory analyses of the Dark Triad Dirty Dozen—an efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences*, *54*(2), 302-306. https://doi.org/10.1016/j.paid.2012.08.027

Webster, G. D., & Kirkpatrick, L. A. (2006). Behavioral and self-reported aggression as a function of domain-specific self-esteem. *Aggressive Behavior*, *32*(1), 17-27. https://doi.org/10.1002/ab.20102

Webster, G. D., Kirkpatrick, L. A., Nezlek, J. B., Smith, C. V., & Paddock, E. L. (2007). Different slopes for different folks: Self-esteem instability and gender as moderators of the relationship between self-esteem and attitudinal aggression. *Self and Identity*, *6*(1), 74-94. https://doi.org/10.1080/15298860600920488

Webster, G. D., Smith, C. V., Brunell, A. B., Paddock, E. L., & Nezlek, J. B. (2017). Can Rosenberg's (1965) Stability of Self Scale capture within-person self-esteem variability? Meta-analytic validity and test–retest reliability. *Journal of Research in Personality*, *69*(August), 156-169. https://doi.org/10.1016/j.jrp.2016.06.005

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, M. B. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis: An introduction for psychologists* (pp. 39-61). American Psychological Association. https://doi.org/10.1037/12350-003

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, *40*(3), 424-431. https://doi.org/10.1016/j.jesp.2003.10.001