Practical Assessment, Research, and Evaluation

Volume 25 Article 6

2020

Collapsing Scale Categories: Comparing the Psychometric Properties of Resulting Scales

Kimberly F. Colvin
University at Albany, SUNY

Guher Gorgun
University at Albany, SUNY

Follow this and additional works at: https://scholarworks.umass.edu/pare

 Part of the Educational Assessment, Evaluation, and Research Commons, and the Other Psychology Commons

Recommended Citation

Colvin, Kimberly F. and Gorgun, Guher (2020) "Collapsing Scale Categories: Comparing the Psychometric Properties of Resulting Scales," *Practical Assessment, Research, and Evaluation*: Vol. 25 , Article 6.
Available at: https://scholarworks.umass.edu/pare/vol25/iss1/6

This Article is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Practical Assessment, Research, and Evaluation by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 25 Number 6, October 2020 ISSN 1531-7714

Collapsing Scale Categories: Comparing the Psychometric Properties of Resulting Scales

Kimberly F. Colvin, *University at Albany, SUNY*
Guher Gorgun, *University at Albany, SUNY*

This study compares a scale, the Rosenberg Self-Esteem Scale, that was administered with four response categories to versions of the same scale that were administered with six and eight response categories. Respondents were randomly assigned to take one of the three versions of RSES. A rating scale utility analysis was conducted on all three versions creating two new four-category versions after collapsing (or combining) adjacent categories. The three different four-category versions were compared on such properties as average scores, correlations with external variables, and factor structure. While most of the psychometric properties were similar across all versions, there were moderate differences related to criterion validity: the scale that did not need to be collapsed had the strongest relationship with external variables, even though there were slightly stronger correlations for the collapsed versions compared to their original scales. A recommendation is made that if it is found that too many categories are being used for a scale then new administrations of the scale should also use the new format, however, an argument could be made to continue administering the survey in its original format, but then collapse responses before analysis.

Once a scale is administered, it may be apparent that the scale had too many rating scale categories; some categories may have been rarely, if ever, used, for example. In such cases, it may be necessary to collapse adjacent categories, which are not, in fact, reflective of unique positions along the latent trait scale, so keeping them as standalone categories could provide misleading information. In addition, polytomous item response theory models may be hard to accurately estimate without an adequate number of respondents in each category. Linacre (1999) provided guidelines to evaluate how rating scale categories were used by respondents to determine whether their use contributed to meaningful measurement of the latent trait.

The purpose of this study was to compare the psychometric properties of a scale that had adjacent categories collapsed resulting in four categories and the same scale that had been administered with four categories originally. In other words: If a scale was administered with eight categories and after the rating scale utility analysis it was determined that several adjacent categories should be collapsed, resulting in four categories, would its psychometric properties be equivalent to responses from the same scale had it been administered with four categories from the beginning?

The results of this study will be informative for a researcher who may have many years' worth of data from a 9-point scale, for example, but then determines that the rating scale categories were not being used adequately and that a 5-point scale is more appropriate. This study aims to answer the question: "If the researcher now starts administering the scale with five categories, can the new results be considered comparable to the old results that have been recoded from a scale from 1 to 9 to a scale from 1 to 5."

Rating Scale Utility Analysis

Linacre's eight guidelines (1999) for the rating scale analysis focus on adequate use of response scale categories. The first two guidelines recommend at least 10 observations in each category and a regular distribution of observations across the categories. A regular distribution includes uniform or unimodal distributions, however highly skewed distributions should be avoided. However, if category use is not uniform, as many as 25 or even 100 observations per category may be required for stable estimates.

Next, the average location of each subsequent category must increase. This guideline addresses the notion that higher categories correspond to higher levels of the latent trait. A related guideline ensures that with increasing levels of the latent trait, each category is at some point most probable. In other words, it should not be possible when moving along the latent trait scale that the most likely response category changes from 2 to 4, with category 3 never being the most likely response.

To evaluate fit, the Rasch outfit mean square statistic for each category must be smaller than 2.0; larger values indicate "unexpected use" of the category (Linacre, 1999). In addition, an evaluation is conducted to ensure that expected scores map through the Rasch model to observed scores, and vice-versa.

The last two guidelines concern the width of each category, ensuring that it is neither too narrow nor too wide to meaningfully reflect the latent trait. The language of Linacre (1999) refers to change in *step difficulties*, the width of a category from where it intersects the response category below it to the response category above. The changes in step difficulties, measured in logits, should be greater than 1.4 and less than 5.0 logits. These guidelines are related to the interpretability of the scale; with appropriate category width, respondents use the categories appropriately to relate to distinct regions on the latent trait scale.

There are many reasons why these guidelines might not be met; it could be because the items themselves were worded poorly, the response categories were confusing, or, in some cases, the scale was administered with too many response categories, which could lead to possible issues with misinterpretation of category meaning. To alleviate these issues, adjacent categories, which may be too narrow or overlap in such a way to confuse meaning, could be collapsed together. Collapsing adjacent categories mean that what were two distinct responses are now coded as the same response. For example if category 2 and 3 responses are to be collapsed, then all category 3 responses are recoded as 2s.

Once, categories have been collapsed by recoding, the rating scale utility analysis is repeated to determine whether more categories need to be collapsed or if the collapsing of categories created a less meaningful scale. The guidelines are just guidelines; they provide considerations when evaluating a scale. An improved scale could lead to improved reliability, but more importantly, a meaningful scale. The improved scale would avoid idiosyncratic category use and lead to consistent interpretation of the categories.

Comparing Response Scale Categories

There are numerous recommendations about the number of response categories to use in scales with Likert items. See Preston and Colman (2000) for an overview of the studies. They document a number of studies that considered the reliability, validity, and information obtained from scales with different numbers of response categories. In general, 5- and 7-point scales were most commonly recommended. Some more recent studies addressed similar issues. Dawes (2008) administered the same "price-consciousness" scale with 5-, 7-, and 10-point scales, then rescaled scores so they were on the same metric. The standard deviations, skewness, and kurtosis were the same, but the mean for the 10-point scale was significantly lower than that of the 5- and 7-point scales. Likewise, Steinberg and Holtzman (2013) compared the properties of the same scale when administered with four and six response categories to non-equivalent groups. They found that about half of correlations of pairs of the six subscales were within one standard deviation across the two versions and the correlations followed roughly the same pattern; in addition, reliability and measurement invariance held.

Lee and Paek (2014) conducted a simulation study to compare psychometric properties of scales with two to six response categories. There were no meaningful differences in reliability, convergent validity, divergent validity, and interitem correlation for scales with four to six response categories, however,

they noticed a marked deterioration when only two or three response categories were used.

More directly related to the current study, Smith, Wakely, de Kruif, and Swartz (2002) found that their 10-point self-efficacy instrument met Linacre's guidelines when adjacent categories were collapsed resulting in a 4-point scale. After administering the 4-point version to a new sample, the authors found a similar factor structure and same item fit statistics for the 4-point scale and the original 10-point collapsed down to 4-point scale. This study indicates that collapsing categories can lead to scales with roughly equivalent properties. However, using the last example, should a researcher administer future versions with a 10-point scale and recode all responses or a 4-point scale? Will conclusions, such as those related to criterion validity, be comparable?

Purpose

The purpose of this study was to compare the psychometric properties of response data from an instrument originally administered on a scale from 1 to 4 to the response data from the same instrument, but administered with scales from 1 to 6 and 1 to 8, which then had adjacent response categories collapsed according to Linacre's guidelines resulting in a 4-point scale. The psychometric properties to be compared include reliability, criterion validity, and factor structure. To compare the response data, this study used empirical data from administrations of the Rosenberg Self-Esteem Scale (RSES; 1965) with four, six, and eight response categories (RSES4, RSES6, and RSES8, respectively). The response data from the collapsed versions are referred to as RSES6.4 and RSES8.4. To date, most studies focused on the psychometric properties of scales when collapsed categories were used instead of the original response categories. In this study, we manipulated the response categories so that we can compare the functioning of a 4-point response scale to the response scales that we collapsed the response categories to be four. Hence, we have a point of reference in order to evaluate the psychometric properties of the collapsed scales. This manipulation of a well-established scale to answer questions about the effects of collapsing categories is novel.

The initial motivation for this study was to answer a question from a colleague: Once a survey has been administered with a given number of categories and found that collapsing categories is appropriate, what should be done the next time the survey is administered? The results of this study can be used to help a researcher, who may have collected a considerable number of responses to an instrument with "too many response categories," determine how to proceed for future data collection. Should the existing instrument be administered with the "too many" response options, then collapsed, so that responses are comparable over time, or can the researcher start administering the instrument on the scale with fewer response options? So, the question is even though collapsing categories has been shown to improve reliability (and other psychometric characteristics), can data from both administrations be interpreted together? Can data be considered comparable for analyses and conclusions if some was administered using eight categories then collapsed to four with results from surveys administered with four response categories?

Research Questions

Broadly, this study addressed how the psychometric properties change when scale categories are collapsed and whether a researcher should start administering a scale with the reduced number of categories after finding that it had been administered with too many categories? Specifically, the comparison of the psychometric properties of the responses from three versions of the RSES (RSES4, RSES6.4, and RSES8.4) addressed the following research questions:

1. How do the average scores compare across the three versions of the scale?
2. Are the estimates of internal consistency reliability and item total correlations consistent across the three versions?
3. Are the relationships with external variables, as a measure of criterion validity, consistent across the three versions?
4. Are the factor structure and factor loadings consistent across the three versions?

Methods

Instruments

The RSES was selected as the scale in which to manipulate the number of categories, because it has

been used in many studies and validity studies across many populations and demographics. The RSES is most commonly used with four response categories. We selected an instrument that has undergone many validity studies with a small number of response categories, so that when administered with more than the four categories, we expected that adjacent categories would need collapsing after conducting the rating scale utility analysis.

For the purposes of evaluating criterion validity, we selected another self-esteem scale, so that we can inspect the relationship of scores between the three versions of RSES with the three subscores of another self-esteem scale, the State Self-Esteem Scale (SSES). The SSES was selected because prior research suggested that its subscales had moderate to strong correlations with the RSES.

Rosenberg Self-Esteem Scale

Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) was originally developed to assess global feelings of self-acceptance and self-worth of adolescents. Rosenberg defined self-esteem as "a favorable or unfavorable attitude toward oneself" (Rosenberg, 1965, p. 15). Although primarily developed for use with adolescents, it is commonly administered to adults. RSES is seen as a standard measure of self-esteem and many researchers use it to evaluate convergent validity when developing new measures (Blascovich & Tomaka, 2013). The scale consists of 10 items and is typically administered on a scale from 1 to 4, where 1 is *strongly agree* and 4 is *strongly disagree*, with total scores ranging from 10 to 40 (see Appendix A). Lower scores represent higher self-esteem. Previous studies found coefficient alpha ranging from α=.72 to α=.82 (Gray-Little, Williams, & Hancock, 1997). In addition, test-retest reliabilities of r = .85 (Silber & Tippett, 1965) and r = .82 (Fleming & Courtney, 1984) were found in studies with college students. For the three versions of RSES that were used in this study, RSES4, RSES6, and RSES8, category 1 was labeled as *strongly agree*, and the highest category was labeled as *strongly disagree*; none of the intermediate categories were labeled.

State Self-Esteem Scale

The SSES was adapted by Heatherton and Polivy (1991) from the commonly employed Janis-Field Feelings of Inadequacy scale (Janis & Field, 1959). The scale consists of 20 situational (i.e., self-concept is not stable and momentary changes are possible) self-esteem items (see Appendix B). Factor analysis revealed that there were three correlated domain specific subscales: performance, social, and appearance (Heatherton & Polivy, 1991). Internal consistency was found to be α =.92; correlations between RSES and the subscales of performance, social, and appearance were r = -.57, r = -.58, and r = -.68, respectively (Heatherton & Polivy, 1991). Using a 5-point Likert scale, the total score ranges from 20 to 100, with higher scores indicating higher levels of self-esteem. Note: the directions of scores are reversed for RSES and SSES, so negative correlations between the two scales were expected.

Participants

The 991 participants were solicited by Qualtrics and responded to the scales and demographic questions on the Qualtrics survey platform. For $3 per participant, Qualtrics recruited and screened English speakers who were part of the general U.S. population. The participants were mostly female, 70.6%, and ranged in age from 18 to 86, with a mean age of 44.5 and standard deviation of 18.1 years. Most participants had at least some level of college education: 24.2% had a high school diploma or equivalent, 42.1% had some college, while 30.0% had a four-year college degree or higher. Participants were presented with an institutional review board-approved informed consent, to which they assented by advancing to the next web page and beginning the survey. Participants were randomly assigned to one of the three versions of the RSES (RSES4, RSES6, or RSES8). All participants were administered the same version of the SSES and five demographic questions.

Results

Rating Scale Utility Analyses

RSES6

The rating scale utility analysis of RSES6 revealed that there were too many categories because several categories overlapped each other, even though the distribution of responses was regular and only moderately skewed, and the average measure (i.e., the average of the scale locations of the observations in the given category) of subsequent categories was increasing. One problem with the scale was that

categories 2 and 3 essentially overlapped each other, as seen by the narrow category widths (changes in step difficulty; Table 1), which were all less than 1.4. This can also be seen in the probability curves shown in Figure 1; in addition, category 4 was never the most probable category at any point along the latent trait scale.

The location of the probability curves for each category indicated that respondents were not distinguishing between categories 2 and 3 nor categories 4 and 5, and were, therefore, good candidates of pairs of categories to collapse.

The response data were recoded to create the collapsed categories; see Table 2 for the details on how the responses were recoded. As seen in Table 3 and Figure 2, the rating scale utility analysis on the recoded responses revealed distinct categories of adequate width. The distribution of observations was unimodal, average category measures increased along the scale, and there was a symmetric correspondence between observed and expected responses.

Table 1. Rating Scale Utility Analysis of RSES 6

Category	Category Count	Category %	Average Measure	OUTFIT Mean Square	Threshold	Change in Step Difficulty	Most Probable From
1	311	9.9	-0.74	1.5	-	-	-
2	324	10.3	-0.44	1.0	-0.64	0.23	-0.64
3	394	12.6	-0.06	1.0	-0.41	0.43	-0.41
4	439	14.0	0.27	0.8	0.02	-0.05	Never
5	784	25.0	0.85	0.9	-0.03	1.09	0.00
6	886	28.2	1.58	1.1	1.06	-	1.06

Figure 1. Probability Curves of RSES 6

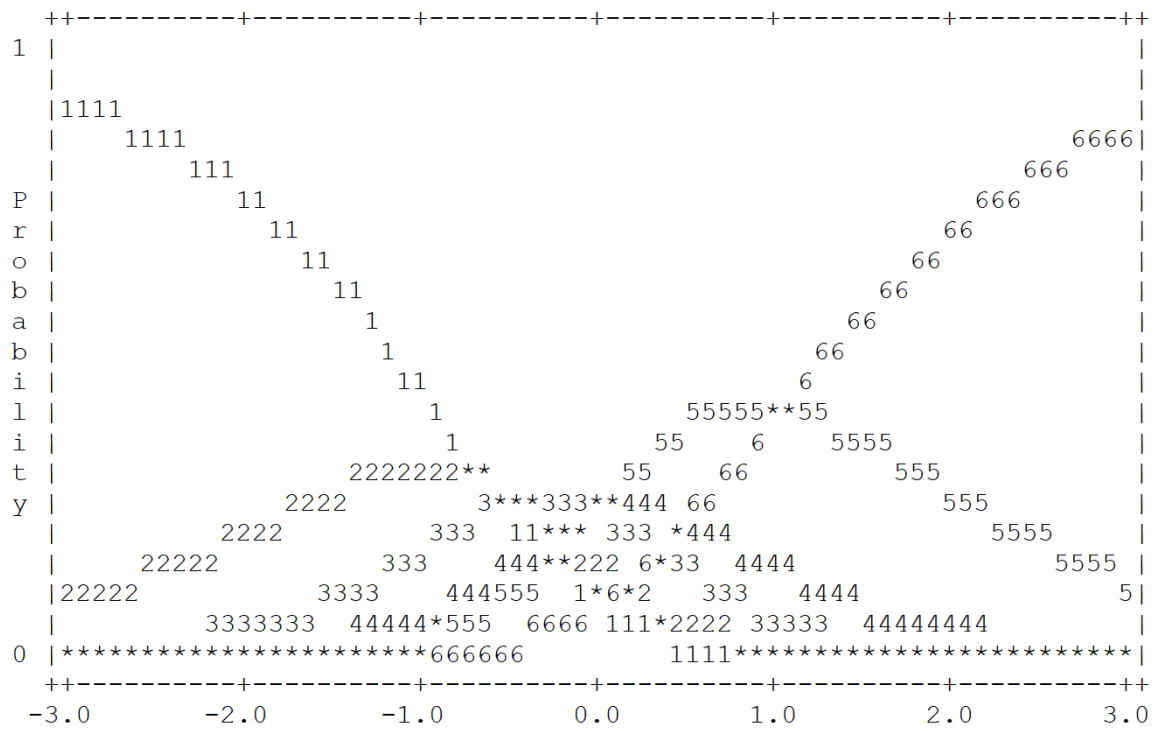


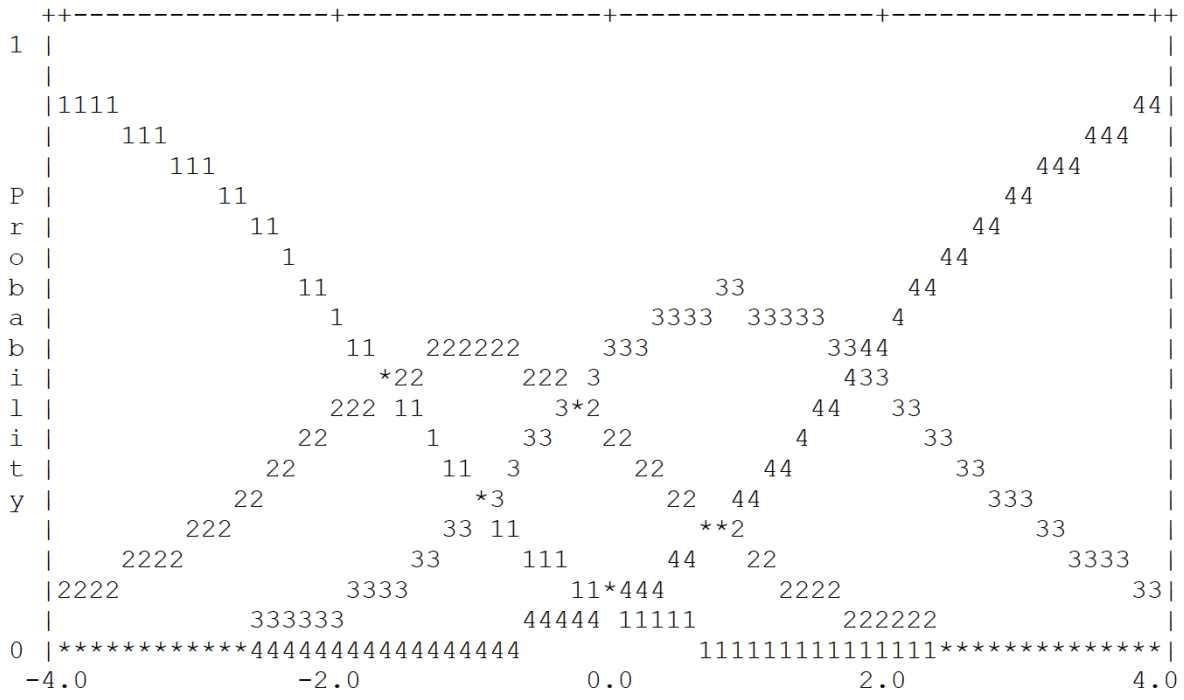
Table 2. Recoding of Item Responses for RSES6 to RSES6.4

RSES6	RSES6.4
1	1
2	2
3	
4	3
5	
6	4

Table 3. Rating Scale Utility Analysis of RSES 6.4

Category	Category Count	Category %	Average Measure	OUTFIT Mean Square	Threshold	Change in Step Difficulty	Most Probable From
1	311	9.9	-1.19	1.3	-	-	-
2 & 3	718	22.9	-0.37	0.9	-1.60	1.36	-1.60
4 & 5	1223	39.0	0.90	0.9	-0.24	2.09	-0.16
6	886	28.2	2.22	1.1	1.85		1.94

Figure 2. Probability Curves of RSES 6.4



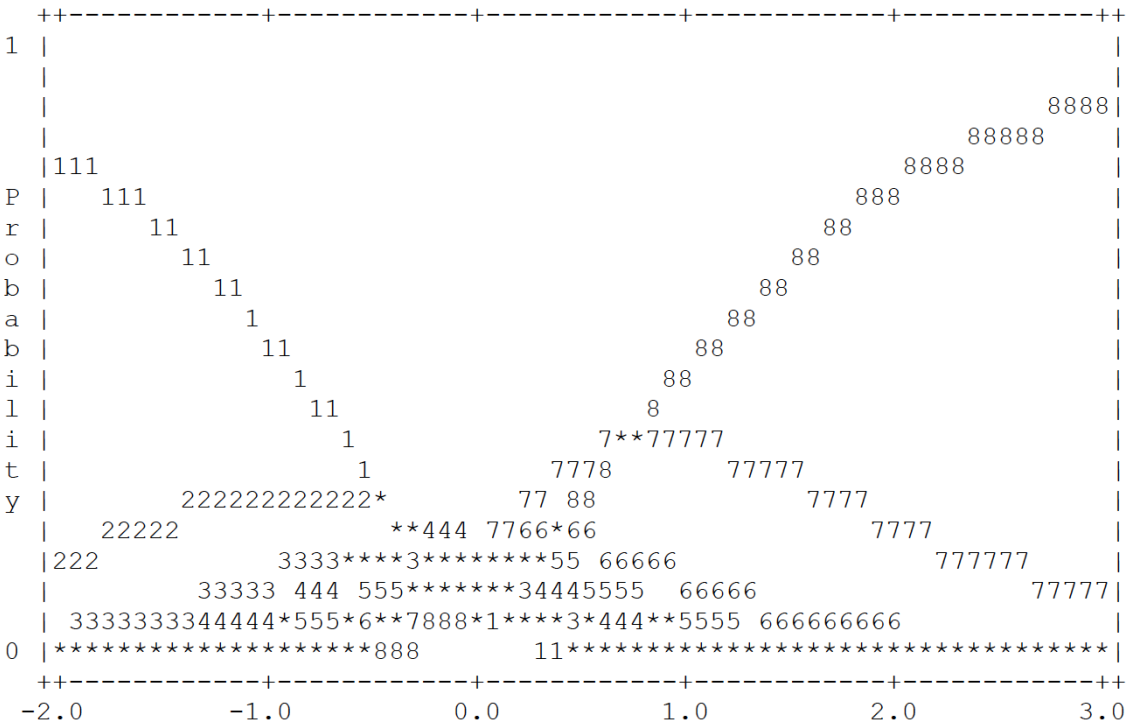
RSES8

The rating scale utility analysis of RSES8 and its possible recodings, took several attempts before reaching a satisfactory recoding of responses. The initial analysis revealed that eight categories for the RSES were too many; specifically, the distribution of observations was slightly irregular and the middle categories were narrow and not separated enough to represent distinct categories. The category widths were all less than 1.0, not greater than 1.4 (see Table 4) and the probability curves made it clear that categories did not correspond to distinct locations along the latent trait scale (Figure 3). Before collapsing categories, the meaning of the categories needs to be considered. Because there was originally an even number of options, we might expect respondents to consider original categories 1 through 4 to be some level of *agreement* and categories 5 through 8 to be some level of *disagreement*, so we would want to think carefully about whether it made sense to collapse categories 4 and 5 together, conceptually.

Table 4. Rating Scale Utility Analysis of RSES 8

Category	Category Count	Category %	Average Measure	OUTFIT Mean Square	Threshold	Change in Step Difficulty	Most Probable From
1	277	9.1	-0.54	1.7	-	-	-
2	246	8.0	-0.36	1.1	-0.33	0.14	-0.33
3	243	8.0	-0.10	1.0	-0.19	-0.15	Never
4	325	10.6	-0.03	0.9	-0.34	0.41	-0.27
5	325	10.6	0.06	0.6	0.07	0.06	Never
6	357	11.7	0.29	0.7	0.13	-0.17	Never
7	580	19.0	0.68	0.9	-0.04	0.73	0.06
8	703	23.0	1.23	1.1	0.69	-	0.69

Figure 3. Probability Curves of RSES 8



Based on the average measures and the width of the categories, it was apparent that categories 3 and 4 should be combined, as well as categories 5 and 6. (See Table 5, RSES8 #1, for the recoding scheme for this first attempt.) However, the analysis of the recoded responses indicated that there were still too many categories. In particular, the change in step difficulties for the 2nd and 5th categories were well below 1.4 (see Table 6) and did not represent distinct regions along the latent trait scale. This can also be seen in the probability curves (in Figure 4) with the overlap of the 2nd and 3rd categories and the 4th and 5th categories.

Table 5. Recoding of Item Responses for RSES8 to RSES8.4

RSES8	RSES8 #1	RSES8 #2	RSES8.4
1	1	1	1
2	2		2
3	3	2	
4			
5	4	3	3
6			
7	5	4	
8	6		4

Table 6. Rating Scale Utility Analysis of RSES 8 #1

Category	Category Count	Category %	Average Measure	OUTFIT Mean Square	Threshold	Change in Step Difficulty	Most Probable From
1	277	9.1	-0.71	1.5	-	-	-
2	246	8.0	-0.47	1.0	-0.45	-0.57	Never
3&4	568	18.6	-0.08	0.8	-1.02	0.95	-0.74
5&6	682	22.3	0.22	0.7	-0.07	0.72	-0.07
7	580	19.0	0.85	0.8	0.65	0.25	0.65
8	703	23.0	1.50	1.1	0.90	-	0.90

These categories, 2 and 5, could either be combined with the extreme categories, 1 and 6, or with the more intermediate categories: 3 and 4 (see Table 5, RSES8 #2 and RSES8.4, respectively). Both versions had regular distributions of observations, increasing average category measures, and reasonable model fit. However, categories 2 and 3 in the second attempt at regrouping, RSES8 #2, were too narrow, according to the guideline that change in step difficulty should be at least 1.4 logits (see Table 7), and thus categories 1 and 4 dominated the scale, meaning that the items were almost dichotomous, rather than having four distinct categories, as the graph depicts in Figure 5.

The regrouping that worked best, RSES8.4, kept the two extreme categories by themselves and combined the intermediate categories on either side of the midpoint (see Table 1, RSES8.4). This version resulted in evenly spaced categories (seen in Figure 6), corresponding to distinct regions along the latent trait scale, both changes in step difficulty were greater than 1.4 (see Table 8) and there was a symmetric correspondence between observed and expected responses.

Figure 4. Probability Curves of RSES 8#1

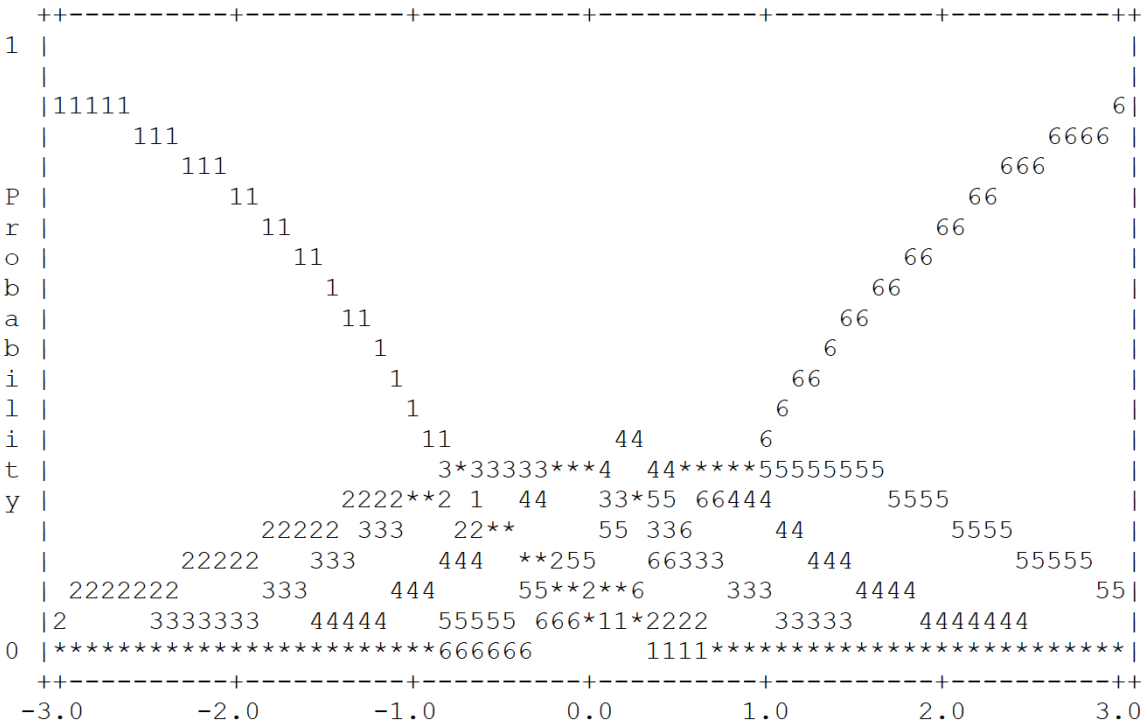


Table 7. Rating Scale Utility Analysis of RSES 8 #2

Category	Category Count	Category %	Average Measure	OUTFIT Mean Square	Threshold	Change in Step Difficulty	Most Probable From
1&2	474	17.6	-0.64	1.3	-	-	-
3&4	568	21.1	-0.05	0.9	-0.49	-0.53	-0.49
5&6	682	25.3	0.43	0.9	0.04	0.49	0.04
7&8	973	36.1	1.22	1.0	0.45	-	0.45

Figure 5. Probability Curves of RSES 8#2

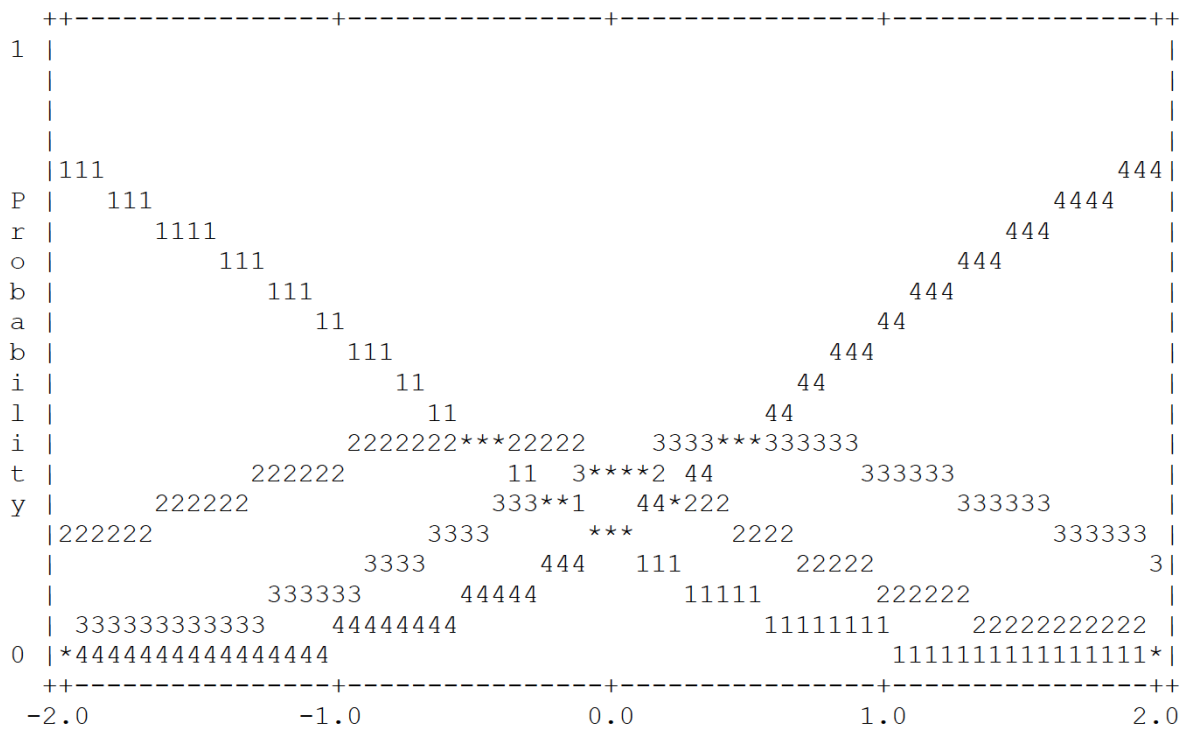
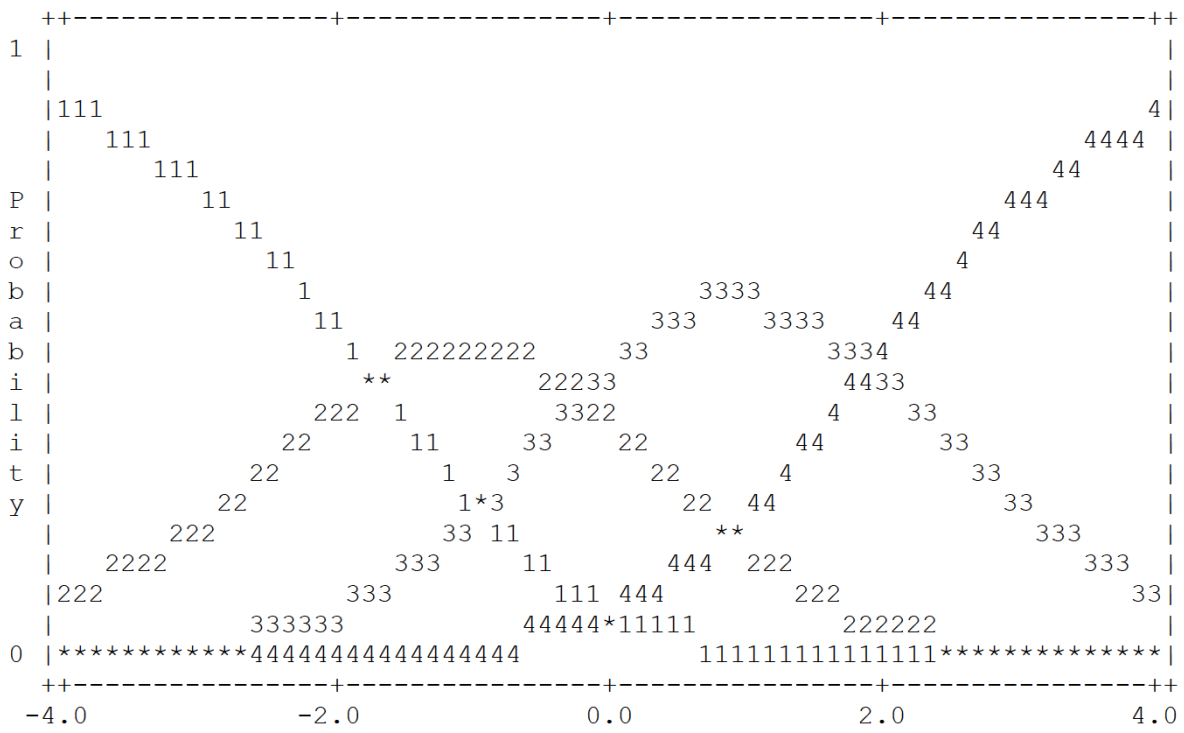


Table 8. Rating Scale Utility Analysis of RSES 8.4

Category	Category Count	Category %	Average Measure	OUTFIT Mean Square	Threshold	Change in Step Difficulty	Most Probable From
1	277	9.1	-1.08	1.4	-	-	-
2&3&4	814	26.6	-0.30	0.8	-1.72	1.52	-1.72
5&6&7	1262	41.3	0.72	0.9	-0.20	2.12	-0.20
8	703	23.0	2.12	1.0	1.92	-	1.92

Figure 6. Probability Curves of RSES 8.4



Research Question 1: Descriptive Statistics

The means and standard deviations of the average RSES scores across the three versions are presented in Table 9. The first line, *original*, corresponds to the average scores based on the three different scales: 1 to 4, 1 to 6, and 1 to 8. The second line, *scaled*, corresponds to the average scores after rescaling, for easier comparison of the average scores. To rescale RSES6, for example, we subtracted 1 from each score, divided by 5, then multiplied by 3, and finally added 1, transforming the 1 to 6 scale to a 1 to 4 scale. To rescale RSES8, we went through the same procedure, except dividing by 7, rather than 5. The purpose of this rescaling was to more easily compare the mean scores across RSES4, RSES6, and RSES 8. The rescaled average scores were not significantly different from each other ($F(2,988) = 0.74$, $p = .48$). This result demonstrated that the three randomly created groups were equivalent with respect to their RSES scores. The third line, *Collapsed*, has the mean scores of the recoded scales, RSES6.4 and RSEES8.4, resulting from the rating scale utility analysis.

Table 9. Means and Standard Deviations of RSES Scores

	RSES4	RSES6	RSES8
Original	2.01	2.77	3.53
	(0.71)	(1.31)	(1.81)
Rescaled	-	2.06	2.08
		(0.79)	(0.77)
Collapsed	-	2.12	2.04
(RSESX.4)		(0.73)	(0.84)

Research Question 2: Reliability

Internal consistency reliability and item total correlations are provided in Table 10. The values were calculated with the R package *psych* (Revelle, 2019). Confidence intervals around coefficient alpha were created based on Feldt, Woodruff, and Salih (1987). Coefficient alpha was similar across the different versions of RSES, as was the pattern of correlations between each item and the total score based on the other items.

Table 10. Item Total Correlations and Coefficient *alpha*

	RSES 4	RSES 6	RSES 6.4	RSES 8	RSES 8.4.2
Item 1	.66	.71	.65	.73	.65
Item 2	.69	.76	.75	.69	.67
Item 3	.62	.69	.66	.65	.63
Item 4	.43	.62	.58	.65	.64
Item 5	.67	.70	.68	.69	.67
Item 6	.73	.79	.79	.76	.75
Item 7	.59	.72	.67	.70	.67
Item 8	.54	.59	.58	.58	.59
Item 9	.74	.79	.77	.80	.78
Item 10	.70	.81	.78	.77	.75
alpha	.89	.93	.92	.92	.91
[95% CI]	[.87, .91]	[.92, .94]	[.91, .93]	[.91, .93]	[.89, .92]

Research Question 3: Criterion Validity

See Table 11 for correlations between scores on the original scales with external variables. Confidence intervals for correlations were computed using the R package, *psychometric* (Fletcher, 2010), based on Cohen, Cohen, West, and Aiken (2003). There is very little to no change in correlation with external variables after collapsing categories.

In addition, the relationship among non-RSES variables were similar for the three groups, demonstrating that groups were similar (see Table 12). The patterns across the correlations were the same for the three groups. For example, the correlation between the SSES Performance and SSES Social subscales were strongest in all three groups, followed by SSES Performance and SSES Appearance, with SSES Social and SSES Appearance having the lowest correlation.

Table 11. Correlations [95% CI] between RSES and SSES subscales.

	SSES.Perf	SSES.Soc	SSES.App
RSES 4	-.75 [-.79, -.70]	-.70 [-.75, -.64]	-.68 [-.73, -.62]
RSES 6	-.63 [-.69, -.56]	-.62 [-.68, -.55]	-.53 [-.60, -.45]
RSES 6.4	-.64 [-.70, -.57]	-.63 [-.69, -.56]	-.53 [-.60, -.45]
RSES 8	-.70 [-.75, -.64]	-.65 [-.71, -.58]	-.63 [-.69, -.56]
RSES 8.4	-.68 [-.73, -.62]	-.64 [-.70, -.57]	-.61 [-.67, -.54]

Note. SSES.Perf = Performance subscale of state self-esteem; SSES.Soc = Social subscale of state self-esteem; SSES.App = Appearance subscale of state self-esteem. The directions of scores are reversed for RSES and SSES, so negative correlations between the two scales were expected.

Table 12. Correlations between State Self-Esteem Subscales by Group

Group		SSES.Soc	SSES.App
RSES 4	SSES.Perf	.74 [.69, .79]	.63 [.56, .69]
	SSES.Soc	-	.59 [.51, .66]
RSES 6	SSES.Perf	.80 [.76, .84]	.66 [.59, .72]
	SSES.Soc	-	.62 [.55, .68]
RSES 8	SSES.Perf	.76 [.71, .80]	.60 [.53, .66]
	SSES.Soc	-	.54 [.46, .61]

Research Question 4: Factor Structure

There is conflicting research on whether the RSES is unidimensional or has a two-factor structure, where the positively worded items load on one factor and the negatively worded items on the other. More studies seem to come down on the side of two highly correlated factors. See Zimprich, Perren, and Hornung, (2005) for a discussion of the prior research. So a confirmatory factor analysis with two factors was conducted on the original RSES (RSES4) and the two versions that had been collapsed to a 4-point scale (RSES6.4 and RSES8.4), as well as the original RSES6 and RSES8. The ranges for good fit were used as follows: root mean square error of approximation (RMSEA) less than .08, standardized root mean square residual (SRMR) less than .10, and a comparative fit index (CFI) greater than .90 (Kline, 2005).

As seen in Table 13, the fit based on RMSEA was less than .08 only for RSES4 and there was a slight decrease in RMSEA after collapsing categories from RSES6 to RSES6.4 and RSES8 to RSES8.4, but the improvement in RMSEA was not significant. The SRMR and CFI values indicated good fit for all models tested.

Table 13. Single-Group CFA Fit Indices for 2-factor Structure

Version	df	χ^2	RMSEA	90% CI for RMSEA	SRMR	CFI
RSES 4	34	92.626*	.076	[.057, .094]	.049	.964
RSES 6	34	155.477*	.107	[.090, .124]	.044	.944
RSES 6.4	34	126.533*	.093	[.076, .111]	.044	.951
RSES 8	34	159.462*	.107	[.090, .124]	.044	.946
RSES 8.4	34	134.054*	.095	[.079, .112]	.042	.952

* $p < .001$

Table 14. Factor loadings with a 2-factor CFA.

		RSES 4	RSES 6	RSES 6.4	RSES 8	RSES 8.4.2
Positively Worded Items	Item 1	1.00	1.00	1.00	1.00	1.00
	Item 3	0.78	0.90	0.95	0.94	0.94
	Item 4	0.75	0.89	0.93	0.97	0.96
	Item 7	0.92	1.01	1.05	0.99	0.99
	Item 10	1.08	1.17	1.26	1.14	1.13
Negatively Worded Items	Item 2	1.00	1.00	1.00	1.00	1.00
	Item 5	0.95	0.83	0.83	0.93	0.94
	Item 6	1.06	1.08	1.09	1.11	1.08
	Item 8	0.88	0.78	0.78	0.88	0.83
	Item 9	1.01	0.97	1.00	1.04	1.01

Discussion

Participants were randomly assigned to one of three groups and took the RSES with either four, six, or eight response categories; these sets of responses were subjected to a rating scale utility analysis based on Linacre's (1999) guidelines. As expected, RSES6 and RSES8 were found to have more categories than respondents could meaningfully use and after collapsing some adjacent categories, both scales were found to be optimal with four categories, RSES6.4 and RSES8.4. The resulting 4-point scales were compared to the original RSES4. Most psychometric properties were comparable across all versions. The means and standard deviations were not significantly different from each other, nor were the reliabilities and item total correlations. This was consistent with prior research that found that basic properties were consistent when the same scale was administered with different numbers of response categories. Smith, et al. (2002) found that factor structure and model fit of a scale, with four response categories, and that of one with ten response categories collapsed down to four were the same. However, this study differs from previous research in that we were most interested in comparing the properties of a scale when administered with its usual four categories and the properties of the same scale with more categories, but collapsed down to four.

In considering the comparability of the different four-category versions, one difference was related to criterion validity. The relationship with external variables is the strongest when the scale was administered with only four categories, rather than being collapsed down to four. The correlations with external variables were almost identical for the original versions of a scale, RSES6 and RSES8, and its respective collapsed counterpart, RSES6.4 or RSES8.4.

The concern is that a scale that has been collapsed, such as RSES6.4 and RSES8.4, does not have the same correlation with external variables as RSES4. This was not due to group differences. The correlations among pairs of the external variables, the subscores of the SSE, were of approximately the same for the three groups of participants (see Table 6). In fact, the correlations between pairs of SSE scores were strongest for the RSES6 group, not the RSES4 group. This demonstrates that the stronger relationship between RSES scores and the SSE scores for the RSES4 group was not related to group membership, but because of the RSES version administered.

The other difference is that model fit, for the two-factor structure, was slightly better for the collapsed versions, RSES6.4 and RSES8.4, than their original counterparts. However, the fit of the collapsed versions was not as good as that of RSES4. Even though there were slight differences, the test of metric invariance demonstrated that the factor structure was the same in the three 4-category versions. While collapsing categories did result in improved model fit, the original 4-category version was best.

Limitations

One limitation to this study is that only the endpoints of the scale were labeled, following the typical way the RSES is administered. Spratto, Leventhal, and Bandalos (2020) found that endpoints were used with more frequency when only the endpoints were labeled. In terms of this study, respondents did not seem to shy away from using the endpoints, in fact, the analyses of RSES6 and RSES 8 found that the two endpoints should be treated as their own categories while collapsing the interior categories. While the labeling of only the endpoints may have played a role in the particular rating scale analyses, it is not clear whether the overall conclusions about the comparison of the original versus collapsed versions of the scale would have changed. This could be an area of future research.

Another limitation of this study is that only one scale, RSES, was used, and that it may be difficult to generalize to all scales. In addition, the findings were only evaluated for those researchers using classical test theory, or summed scores. We did not evaluate the differences in latent trait and parameter estimation between the scale originally administered with only four response categories and when administered with more, but collapsed down to four. The primary reason for this is the intended audience. The goal of this paper was to provide guidance to researchers who develop their own instruments or use existing instruments with nine categories, say, and decide to collapse categories.

Practical Implications

After a rating scale utility analysis, if a researcher finds that the scale should be administered with four or five response categories, then it does not

make sense to continue to administer the scale with more categories, particularly, because this means that the respondents will be unable to adequately distinguish among the responses. Even though the psychometric properties of the collapsed scales were similar to those of the original RSES4, there were some important distinctions. First, correlations with external variables was strongest with the original scale, rather than the collapsed versions. In addition, even though there was metric invariance, the model fit was slightly better for RSES4 than the scales that were collapsed down to four categories, RSES6.4 and RSES 8.4. While we would recommend that new versions of the scale be administered with the reduced number of response categories, it may not be that detrimental to continue collecting data with the original version. As with previous research, the psychometric properties were roughly equivalent whether administering a scale with more categories and collapsing down to the optimal number, or just administering with the optimal number. For a researcher who has years of data from a scale with nine response categories, but now knows that the scale should be collapsed down to five response categories, a justification could be made to either continuing to collect using the 9-point version and then collapsing and recoding responses for analysis or just administering the 5-point scale going forward. However, it seems wise to avoid the situation altogether. The rating scale utility analysis on the proposed scale could be conducted during the pilot stage and determine the appropriate number of response categories before administering the instrument on a large scale.

References

Blascovich, J. & Tomaka, J. (2013). Measures of self-esteem. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman, (Eds.), *Measures of personality and social psychological attitudes: measures of social psychological attitudes (Vol. 1)* (pp. 115- 160). San Diego, California: Academic Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61-77.

Feldt, L., Woodruff, D., & Salih, F. (1987). Statistical Inference for Coefficient Alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Fleming, J. S., & Courtney, B. E. (1984). The dimensionality of self-esteem: II. Hierarchical facet model for revised measurement scales. *Journal of Personality and Social psychology*, *46*(2), 404-421.

Fletcher, T. D. (2010). psychometric: Applied psychometric theory [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=psychometric (R package version 2.2)

Gray-Little, B., Williams, V. S., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*(5), 443-451.

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.

Janis, I. L., & Field, P. B. (1959). Sex differences and factors related to persuasibility. In C. I. Hovland & I. L. Janis (Eds.), *Personality and persuasibility* (pp. 55-68). New Haven, CT: Yale University Press.

Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment*, *32*(7), 663–673. https://doi.org/10.1177/0734282914522200

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103-122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Preston, C., & Colman, A. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Revelle, W (2019). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 1.9.12, https://CRAN.R-project.org/package=psych.

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

Rosseel, Y (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/.

Silber, E., & Tippett, J. S. (1965). Self-esteem: Clinical assessment and measurement validation. *Psychological Reports*, *16*, 1017-1071.

Smith, E. V., Jr., Wakely, M. B., de Kruif, R. E. L., & Swartz, C. W. (2002). Optimizing rating scales for self-efficacy (and other) Research. *Educational and Psychological Measurement*, *63*(3), 369–391.

Spratto, E., Leventhal, B., & Bandalos, D. (2020). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option Labeling. *Educational and Psychological Measurement*. Advance online publication.

Steinberg, J. & Holtzman, S. (2013, October). *The empirical similarities of a time management assessment based on varying response scales*. A paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement*, *65*(3), 465–481. https://doi.org/10.1177/0013164404272487

Citation:

Colvin, K. F., and Gorgun, G. (2020). Collapsing Scale Categories: Comparing the Psychometric Properties of Resulting Scales. *Practical Assessment, Research & Evaluation*, 25(6). Available online: https://scholarworks.umass.edu/pare/vol25/iss1/6/

Corresponding Author

Kimberly F. Colvin, Ed.D.
University at Albany, SUNY
Catskill 251
Albany, NY, 12222 USA

email: kcolvin [at] albany.edu

Appendix A

Rosenberg Self-Esteem Scale (Rosenberg, 1965)

Responses ranged from 1, *strongly agree,* to 4, 6, or 8, *strongly disagree*, depending on the version randomly administered to each participant.

1. On the whole, I am satisfied with myself.
2. At times I think I am no good at all.
3. I feel that I have a number of good qualities.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I certainly feel useless at times.
7. I feel that I'm a person of worth, at least on an equal plane with others.
8. I wish I could have more respect for myself.
9. All in all, I am inclined to feel that I am a failure.
10. I take a positive attitude toward myself.

Appendix B

State Self-Esteem Scale (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.

The response options: Not at all, A little bit, Somewhat, Very much, Extremely, and Prefer not to answer.

1. I feel confident about my abilities. *Performance*
2. I am worried about whether I am regarded as a success or failure. *Social*
3. I feel satisfied with the way my body looks right now. *Appearance*
4. I feel frustrated or rattled about my performance. *Performance*
5. I feel that I am having trouble understanding things that I read. *Performance*
6. I feel that others respect and admire me. *Appearance*
7. I am dissatisfied with my weight. *Appearance*
8. I feel self-conscious. *Social*
9. I feel as smart as others. *Performance*
10. I feel displeased with myself. *Social*
11. I feel good about myself. *Appearance*
12. I am pleased with my appearance right now. *Appearance*
13. I am worried about what other people think of me. *Social*
14. I feel confident that I understand things. *Performance*
15. I feel inferior to others at this moment. *Social*
16. I feel unattractive. *Appearance*
17. I feel concerned about the impression I am making. *Social*
18. I feel that I have less intellectual ability right now than others. *Performance*

19. I feel like I'm not doing well. *Performance*

20. I am worried about looking foolish. *Social*

Note: The italicized word after each item reflects the subscore to which the item belongs.