Mindfulness (2022) 13:1793–1803
https://doi.org/10.1007/s12671-022-01917-4

ORIGINAL PAPER

Should We Use a Total Score, Two Subscale Scores, or Six Subscale Scores for the Self-Compassion Scale? A Multi-faceted Assessment Beyond Model Fit Indices

José Buz[1] · Antonio Crego[2] · José R. Yela[2] · Elena Sánchez-Zaballos[2] · Antonio Ayuso[2]

Accepted: 21 May 2022 / Published online: 8 June 2022

Abstract

Objectives The aim of this study was to conduct a multi-faceted assessment of the psychometric properties of the Self-Compassion Scale (SCS). In addition to the goodness-of-fit, we aimed to assess the strength and replicability of three factorial solutions, and the quality and effectiveness of the three scoring schemes of the scale (i.e., total scale score, two subscale scores, and six subscale scores).

Methods Participants were 1508 Spanish-speaking community-dwelling adults ($M = 34.94$ years, $SD = 15.02$). Data were examined by means of a conjoint strategy using Rasch modeling, non-linear factor analysis, exploratory bifactor analysis, and parallel analysis. A procedure for selecting the optimal set of items that must be used to compute individual's scores was used.

Results The unidimensional solution showed a marginal model fit (RMSR = .089), and both the bifactor two-group and bifactor six-group solutions showed a good fit (RMSR = .043 and .019, respectively). However, only the unidimensional and the bifactor two-factor solutions showed interpretable and replicable factor structures, and high-quality and effective scores to be used for measurement purposes. Subscale scores derived from the six primary factors did not show adequate psychometric properties. It was observed that the information provided by 10 items was redundant and had already been provided by the other 16 items.

Conclusions Good model fit is neither sufficient nor necessary to justify the use of a scoring scheme. Goodness-of-fit statistics should be complemented by an assessment of the metric properties of the resulting scores when proposing SCS scoring schemes.

Keywords Self-Compassion Scale · Non-linear factor analysis · Determinacy · Marginal reliability

Compassion has been defined as a "multitextured response to pain, sorrow and anguish. It includes kindness, generosity, and acceptance" (Feldman & Kuyken, 2011, p. 144). Compassion can also adopt the form of an emotional positive attitude toward oneself in a way that "involves being open to and moved by one's own suffering, experiencing feelings of caring and kindness toward oneself, taking an understanding, nonjudgmental attitude toward one's inadequacies and failures, and recognizing that one's own experience is part of the common human experience" (Neff, 2003, p. 224). Meta-analyses (e.g., Ferrari et al., 2019) have positively related self-compassion to subjective well-being and negatively to depression, anxiety, psychological stress, and cognitive rumination. Self-compassion has been found to be a key psychological variable for clinical practice in mindfulness-based interventions (Yela et al., 2020). Neff (2003) considered three basic components of self-compassion, each consisting of two opposite dimensions: (a) self-kindness versus self-judgment refers to being benevolent, kind, and sympathetic toward oneself rather than harsh self-criticism or self-punishment; (b) common humanity versus isolation emphasizes that life experiences are widely shared by all of humanity, instead of experiencing feelings of separation and isolation; and (c) mindfulness versus over-identification refers to the ability to be aware of painful

✉ José Buz
buz@usal.es

[1] Faculty of Education, University of Salamanca, Paseo de Canalejas 169, 37008 Salamanca, Spain

[2] Faculty of Psychology, Pontifical University of Salamanca, Salamanca, Spain

thoughts and feelings without judging rather than over-identifying with them. Accordingly, the Self-Compassion Scale (SCS; Neff, 2003) was developed "to measure self-compassion as a single overarching construct" (p. 226) including (highly) intercorrelated subscales. In the original validation study, these components were examined by fitting three unidimensional models, but the goodness-of-fit statistics non-normed fit index (NNFI; also known as the Tucker-Lewis index) and comparative fit index (CFI) did not reach the established cutoff values for an adequate model fit. As a result, the scoring scheme with three subscale scores was (forever) abandoned. In subsequent analyses, a single higher-order solution and a six correlated factor solution achieved the expected fit and were used to justify the use of a total scale score and six subscale scores. This proposal stimulated a still ongoing passionate debate about the best scoring scheme for the scale (e.g., Muris & Otgaar, 2022; Neff, 2022). In the last 5 years, the methodological arsenal employed in validation studies testing competing models of the SCS has become increasingly varied and complex. However, it is not clear how much progress has been made, so researchers must rely on their instinct on how to use the scale or even if they should use it.

In psychometric terms, the definition of self-compassion proposed by Neff (2003) is a substantively complex construct. With a few exceptions, all the validation studies have examined whether unidimensional and multidimensional solutions fitted the SCS scores. Recently, authors have proposed bifactor solutions with an exploratory approach (e.g., Neff et al., 2021; Tóth-Király & Neff, 2021) because, in their opinion, confirmatory factor analysis might be too restrictive for the SCS considering its structure. Bifactor models are useful to (a) separate item response variance into general versus group factor sources, (b) determine the degree that item responses conform to a unidimensional versus multidimensional structure, and (c) assess the utility of subscale scores after variance due to the general factor controlled for (Reise, 2012). However, SCS studies based on factor analysis concur in their lack of attention to the scoring stage (i.e., when the item parameter estimates are used to compute the factor score estimates) after performing the calibration stage (i.e., when structural item parameters are estimated). This is surprising because researchers and practitioners frequently aim to use factor analysis–derived scores to measure individuals. In the case of the SCS, the superiority of one factorial solution over other competing solutions has been established by considering mainly goodness-of-fit statistics. Although an acceptable model fit is an important requirement, methodologists (e.g., Montoya & Edwards, 2021) are cautioning researchers against their overconfidence in goodness-of-fit statistics and its corresponding benchmarks, especially for selecting the correct number of factors. Among others, model fit is influenced by item distribution properties, item difficulties (i.e., the endorsement levels), item parceling, and the selection of the appropriate factor analysis (Sellbom & Tellegen, 2019). For instance, conducting linear factor analysis with non-normally distributed data and ordinal data yields bias in polychoric estimates and negatively affects goodness-of-fit indices (Foldnes & Grønneberg, 2019). Other sources, such as the residual matrix, can provide important information to evaluate sources of misspecification but are rarely reported in factorial analysis (Sellbom & Tellegen, 2019). On the other hand, even a weak factorial structure (e.g., including spurious factors) with a minimal degree of quality can show a good fit (Ferrando & Lorenzo-Seva, 2018). For instance, a six-factor model of the SCS with arbitrarily assigned items was found to achieve a good model fit (Coroiu et al., 2018). Therefore, relying solely on model fit statistics to choose the scoring scheme of any scale is a controversial and hazardous analytical strategy. As recommended, the strength and replicability of any factorial structure should be also examined to judge the appropriateness of any scoring scheme in terms of validity and generalizability (Calderón et al., 2019). In certain circumstances, unidimensional solutions may fail to reach the expected fit in favor of any other multidimensional solution. However, the superiority in goodness-of-fit may be insufficient empirical justification for specifying additional latent factors, and that multidimensionality does not justify the use of subscale scores because the interpretability of the factorial solution needs to be made (Reise, 2012). In this regard, the American Educational Research Association et al. (2014) highlighted the need for demonstrating the quality and distinctiveness of the scores resulting from a scale.

To date, no study has conducted a multi-faceted assessment of the psychometric properties of the SCS scoring schemes. This is a notable shortcoming, as the main objective of some studies has been "to determine whether or not the use of an overall SCS score (in addition to the six subscale scores) is justified" (Neff et al., 2017, p. 599). Thus, in addition to an assessment in terms of goodness of fit, the present study sought to examine (a) whether bifactor solutions with six-group factors and two-group factors, respectively, attained the expected standards of replicability, and whether the resulting factor score estimates also attained the standards of quality and effectiveness for measurement purposes; (b) whether the SCS scores conform to an essentially unidimensional structure; (c) whether the sum scores resulting from the total scale and the subscales were more appropriate than the corresponding factor score estimates; and (d) which SCS scoring scheme could be considered optimal.

Method

Participants

Data were gathered from a sample of 1508 individuals from the general population ($M = 34.94$ years, SD: 15.02; range: 18–70 years) from 15 Spanish-speaking countries: Venezuela (29.9%), Spain (11.6%), Nicaragua (11.3%), Bolivia (9.7%), Paraguay (7.2%), Argentina (7%), Dominican Republic (7%), and other Latin American countries (each less than 5%). Women represented 71.5% of the total sample ($M = 34.80$ years, SD: 14.89). Most of the sample (68.9%) had at least an undergraduate level of education. Nearly 40% of the participants were active workers, 33.8% were students, 15.7% were unemployed, and 10.2% were retirees. The majority of participants rated themselves as non-meditators (69.8%), while 30.2% affirmed practicing meditation (occasionally or regularly).

Procedure

We employed an online survey that allows recruitment via a wide range of social media sites. Websites contained a brief description of the study and a link to a battery of questionnaires about health-related variables that were included in the context of a broader study. Following a snowball strategy, the survey contained information to encourage participants to share the survey link among their contacts. We also requested participants to provide informed consent before sending data. All the responses were stored anonymously in a password-protected online database. Participants did not receive any compensation for filling out the questionnaire. The Pontifical University of Salamanca Research Ethics Board granted ethical approval.

Measures

Demographic Characteristics and Meditation Practice. Participants completed a demographic questionnaire that included country of residence, age, gender, educational attainment, job status, and the frequency of meditation using a single question with three options (1 = *non meditator*; 2 = *occasional meditator*; 3 = *regular meditator*).

Self-compassion. The SCS (Neff, 2003; Spanish version from García-Campayo et al., 2014) is an instrument widely used for the assessment of self-compassion in community and clinical populations. The scale is composed of 26 statements in six subscales. Three subscales are positively worded: self-kindness (e.g., "I'm kind to myself when I'm experiencing suffering"), common humanity (e.g., "I try to see my failings as part of the human condition"), and mindfulness (e.g., "When something upsets me, I try to keep my emotions in balance"). The remaining three subscales are negatively worded: self-judgment (e.g., "When times are really difficult, I tend to be tough on myself"), isolation (e.g., "When I fail at something that's important to me, I tend to feel alone in my failure"), and over-identification (e.g., "When something upsets me, I get carried away with my feelings"). Responses are rated using a Liker-type scale (1 = *almost never* to 5 = *almost always*). Subscale scores and the total SCS score can be calculated after reverse scoring negatively worded items.

Data Analyses

Authors such as Montero-Marín et al. (2018) suggested that the meaning of the self-compassion construct, as measured with the SCS, may be influenced by cultural values. These cultural differences may be a source of measurement bias. Therefore, before comparing factor solutions, we assessed the measurement invariance of the SCS across countries. For this purpose, we conducted a differential item analysis (DIF) based on the Rasch model. Invariance of person and item measures is a fundamental principle in Rasch measurement. In Rasch modeling, the item hierarchy should be invariant for individuals with the same ability regardless of their country membership. We used the DIF t statistic (with Bonferroni correction) that calculates the difference between the item difficulty for participants of each country and the expected score without DIF divided by the approximate standard error of difference. Then, we assessed the impact on person measures by observing whether the observed average of the scored responses was greater than expected. Once the measurement invariance was established, we examined the psychometric properties of the SCS by means of unrestricted factor analysis. In this exploratory factor analysis, the number of factors is fitted, so the factor solution is rotated to fit the proposed population model as closely as possible.

Bifactor solutions were examined using pure exploratory bifactor analysis (PEBI; Lorenzo-Seva & Ferrando, 2019) for oblique solutions. As recommended (Reise et al., 2010), after obtaining evidence of the presence of a strong latent variable, a unidimensional solution was modeled. Considering that the scale was not too long and the sample was large, we conducted a double cross-validation. Thus, the total sample was split to obtain fully independent subsamples by generating random samples. The coefficients of congruence for the results across samples were > 0.97 in all cases. So, the current study only presents the results from the entire sample. The person-fit statistic detected 7.4% of participants with inconsistent patterns of responses. The results of the analysis to test for negative impact on person measures, in particular on reliability, showed that it was negligible, so we retained them for the remaining analyses. Because

item scores were ordinal and some item distributions were skewed, a nonlinear factor analysis based on polychoric correlations was used to fit the data. To assess how many factors to retain, we applied optimal parallel analysis based on minimum rank factor analysis (PA-MRFA; Timmerman & Lorenzo-Seva, 2011) with bias-corrected and accelerated (BCa) bootstraps (500 samples). MRFA is the only factor analysis method that allows computing the percentage of explained common variance of each factor in the solution.

At the item calibration stage, we conducted a basic internal assessment of the scale by means of a set of indices (see Ferrando & Lorenzo-Seva, 2018) such as (a) the root mean square residuals (RMSR close to 0.08 for good fit); (b) the explained common variance (ECV); (c) the item explained common variance (I-ECV); (d) the item residual absolute loadings (IREAL); and (e) the mean of item residual absolute loadings (MIREAL). ECV is a measure of dimensionality and I-ECV indicates which items are the best to contribute to the essential unidimensionality of a scale. ECV and I-ECV values in the range from 0.70 to 0.85 are expected, but the value should be judged in the context of other indices. IREAL is a model-independent index based on the absolute loadings on the residual factor once the first canonical factor has been extracted. IREAL < 0.30 are expected for essential unidimensionality. To assess the construct replicability, we used the optimal-PA and the generalized H index (G-H) for multidimensional oblique solutions. The replicability largely determines the appropriateness of different scoring schemes. G-H values > 0.80 are expected.

At the scoring stage, we assessed the determinacy and reliability of the factor score estimates. The factor determinacy index (FDI) indicates whether the factor score estimates are good proxies for representing the latent factor scores. Values > 0.90 are desirable (Rodriguez et al., 2016). We assessed the ordering of individuals along the latent trait by means of the marginal reliability estimate (Brown & Croudace, 2015). Values > 0.80 are expected for predictive and clinical uses of SCS scores. The marginal reliability estimate does not inform about the accuracy of measurement along the latent trait, nor the magnitude of the differences that can be differentiated with the factor score estimates. Thus, we calculated the sensitivity ratio (SR) and the expected percentage of true differences (EPTD), respectively. EPTD values > 90% are expected to consistently differentiate individuals.

To empirically examine the resulting summative scores and the factor score estimates from these models, we used a procedure called Direct Item Addition of Non-Ahead (DIANA) proposed sets (Ferrando & Lorenzo-Seva, 2021). DIANA selects the optimal set of items that must be used to compute an individual's scores maximizing fidelity and correlational accuracy. Fidelity and accuracy can be interpreted as correlations between the sum scores and the true latent factors. Values ≥ 0.90 are expected for appropriate sum scores.

Descriptive statistics for the sample were obtained with IBM SPSS Statistics (Version 26), factor analysis was conducted with FACTOR (Version 11) (Lorenzo-Seva & Ferrando, 2021), and Rasch statistics were obtained with WINSTEPS (Version 5.0) (Linacre, 2021).

Results

Measurement invariance analysis revealed the existence of DIF in two items (items CH7 and CH10). In both items, the average effect of DIF on person measures was detected for both Bolivian and Paraguayan participants; for these participants, items CH7 and CH10 were more difficult than for the rest of the countries. As recommended (Linacre, 2021), when the purpose of the analysis is not refining the scale but to obtain evidence that items may have different meanings, items should be split and re-evaluated. In the case of DIF, they should be dropped. Our findings revealed no DIF across countries, so these items did not represent a threat for the construct.

In FA, both the KMO test = 0.95 and Bartlett's statistic = 17,240.6 ($df = 325$; $p < 0.001$) were excellent. As expected, due to its higher parameterization, the six-group solution obtained the best fit (Table 1). This solution accounted for 88.35% of the explained common variance. The general factor accounted for nearly 40% of the common variance, but the explained common variance accounted for by some of the six group factors was very low (e.g., 3.9% for Self-Kindness). Most of the items had substantive loadings (> 0.30) on the general factor and two or more group factors (Table 2). Moreover, we observed some weak factors (e.g., over-identification with $M_\lambda = 0.20$) that compromised the interpretability of the rotated matrix.

In the two-group solution, the percentage of explained common variance was 74.3%. Moreover, each of the two group factors added nearly half of the explained common variance accounted for by the general factor, and each item was only influenced by a general factor and a single group factor (λs above 0.50). Supporting this structure, the optimal-PA method advised the extraction of two factors ($\Phi = 0.60$). The correlation between the group factors, once the general factor was modeled, remained significant ($\Phi = 0.22$, 95% CI [0.149, 0.290]), but considering the size of the sample, the correlation can be considered relatively low and not substantial. Thus, they were considered distinct facets of the general latent trait.

The aforementioned findings were consistent with the presence of a substantial general factor, so we fitted a unidimensional solution. In terms of pure goodness-of-fit, the unidimensional solution reached a marginally acceptable fit. However, the majority of the loadings were quite high

Table 1 Basic internal assessment of the factorial solutions and quality and effectiveness of the factor score estimates

Model	RMSR		ECV	MIREAL		G-H latent		G-H observed		FDI	Marginal reliability	SR	EPTD
	Value	95% CI		Value	95% CI	Value	95% CI	Value	95% CI				
Single factor	.089	[.090, .104]	55.6%	.29	[.27, .30]	.95	[.94, .95]	.96	[.96, .97]	.98	.97	6.4	97.5%
Bifactor two-group factor	.043	[.043, .043]	74.3%	_									
General factor			39.2%			.86	[.84, .87]	.88	[.84, .89]	.95	.91	_	93.0%
Self-compassion			16.7%			.78	[.74, .79]	.80	[.78, .81]	.92	.85	2.4	90.3%
Self-criticism			18.4%			.79	[.77, .81]	.82	[.79, .83]	.93	.87	2.6	91.1%
Bifactor six-group factor	.019	[.019, .019]	88.3%	_									
General factor			39.6%			.92	[.87, .97]	.90	[.86, .94]	.96	.94	_	94.6%
Self-kindness			3.9%			.69	[.62, .92]	.64	[.58, .80]	.86	.74	1.7	87.0%
Common humanity			11.2%			.83	[.75, .89]	.77	[.71, .83]	.93	.87	2.6	91.2%
Mindfulness			5.9%			.71	[.55, .76]	.68	[.50, .71]	.87	.76	1.8	87.5%
Self-judgment			7.3%			.75	[.69, .95]	.72	[.66, .85]	.89	.80	2.0	88.9%
Isolation			13.8%			.77	[.71, .79]	.74	[.65, .75]	.90	.82	2.1	89.4%
Over-identification			6.5%			.72	[.69, .74]	.69	[.64, .70]	.88	.77	1.8	87.9%

RMSR, root mean square of residuals; *CI*, confidence interval; *ECV*, explained common variance (MRFA-based); *MIREAL*, mean of item residual absolute loadings; *G-H*, generalized H index; *FDI*, factor determinacy index (MRFA-based); *Reliability*, marginal reliability of estimates; *SR*, sensibility ratio; *EPTD*, expected percentage of true differences

($M_\lambda = 0.61$) and the inter-factor correlation was moderate, so this solution was further investigated. The inspection of the standardized correlated residuals revealed the presence of large residuals between three item pairs (CH10-CH7, CH7-CH3, and CH10-CH3). These three items were also responsible for the departure from essential unidimensionality, and four items (items MD9, MD22, SJ1, and IS18) were contributing little to the essential unidimensionality. The percentage of common variance explained for this solution was approximately 55%. This value can be considered adequate, considering the low percentage of large residuals, the number of factors, and that similar values are frequently obtained with bifactor models in personality scales (see Reise et al., 2015, for a review).

Regarding the construct replicability, the unidimensional solution and the two-group solution obtained good values. Conversely, none of the six-group factors showed a strong (i.e., high factor loadings) and clearly interpretable (i.e., no complex items) pattern solution to be replicable.

On assessing the quality and effectiveness of the factor score estimates, only those from the unidimensional model and the bifactor two-group solution attained an adequate level of determinacy (FDI values from 0.92 to 0.99). The suitability and accuracy of the SCS factor score estimates were optimal for the unidimensional model and the bifactor two-group solution as they reach marginal reliability values above 0.80 along all the effective measurement range (Fig. 1). Their factor score estimates were highly accurate for individual assessment and were able to effectively differentiate six trait levels and two trait levels (e.g., low–high self-compassion and low–high self-criticism), respectively. More than 90% of the differences were reflecting true differences. On the contrary, three of the factor score estimates derived from the bifactor six-group solution did not reach the expected value.

The bifactor six-group solution also attained, in general, adequate values of marginal reliability (range 0.74–0.87). Despite this, we observed two causes of concern. First, as can be seen graphically, in all the six factors, the reliability dropped dramatically at the lower and upper extremes, thus narrowing the effective range of measurement. Second, the factor score estimates from three out of the six subscales were unable to differentiate a minimum of two trait levels and nearly all the subscales were reflecting spurious differences rather than true differences.

Finally, the procedure to compute optimal sum scores (i.e., DIANA scores) showed that a total scale score and two subscale scores would be adequate (fidelity and accuracy ranged from 0.97 to 0.99 in all cases). However, these values were achieved by removing the following items: CH3, CH7, CH10, CH15, MI9, MI22, IS18, OI20, and OI24. DIANA revealed that these items were unproductive for measurement and then not very useful for differentiating persons along the latent variable. When using two subscale scores, only item OI20 was deleted.

Discussion

The aim of this study was to conduct a multi-faceted assessment of the psychometric properties of the SCS. Beyond comparing model fit indices to select the

Table 2 Item-level statistics and rotated loading matrices for the unidimensional solution and the bifactor solutions

	I-ECV	I-REAL	Single factor	Bifactor two-group factor			Bifactor six-group factor						
			GF	GF	SC	SCr	GF	SK	CH	MI	SJ	IS	OI
Self-kindness													
SK5	.81	.30	**.63**	**.62**	**.36**	*−.08*	**.62**	**.52**	*−.06*	−.17	.14	*−.01*	−.23
SK12	.92	.23	**.76**	**.66**	**.41**	*.08*	**.70**	**.60**	*−.06*	*−.04*	*.06*	.14	−.26
SK19	.88	.27	**.74**	**.68**	**.42**	*.01*	**.68**	*−.07*	**.56**	*−.01*	*.07*	*.06*	−.14
SK23	.97	.13	**.73**	**.73**	.24	*.07*	**.59**	.11	**.31**	*.03*	*.05*	*.09*	**.36**
SK26	.92	.21	**.70**	**.68**	**.31**	*.01*	**.60**	.12	**.36**	*.04*	*−.01*	*−.02*	.25
Common humanity													
CH3	.07	.53	.14	.09	**.52**	−.33	.16	*.05*	.12	**.52**	*.08*	−.27	−.28
CH7	.13	.53	.20	.24	**.44**	−.40	**.34**	−.19	−.14	**.67**	*−.00*	−.27	−.32
CH10	.27	.55	**.34**	**.36**	**.50**	−.37	**.51**	−.22	**.68**	−.12	*−.06*	−.30	−.23
CH15	.69	.39	**.59**	**.52**	**.49**	−.11	**.49**	.13	**.36**	.12	.26	*−.05*	*.04*
Mindfulness													
MI9	.80	.27	**.53**	.24	**.60**	.17	**.47**	.13	.22	**.50**	*−.06*	*−.04*	−.14
MI14	.86	.27	**.66**	**.40**	**.59**	.17	**.47**	.13	**.46**	**.47**	*.05*	.13	−.15
MI17	.86	.28	**.70**	**.55**	**.49**	*.08*	**.51**	*.02*	**.45**	.27	.21	.16	*−.03*
MI22	.77	.32	**.59**	**.48**	**.48**	*−.01*	**.52**	*.07*	**.44**	.14	*.08*	*−.04*	−.15
Self-judgment													
SJ1	.65	.31	**.42**	**.52**	−.28	.25	**.31**	−.12	*−.06*	−.13	*−.03*	.27	**.61**
SJ8	.89	.25	**.71**	**.66**	*−.03*	**.40**	**.53**	*.05*	.15	*−.04*	*.00*	**.51**	**.39**
SJ11	.87	.25	**.67**	**.64**	*−.07*	**.37**	**.53**	*.06*	*.07*	*.01*	**.38**	−.14	**.55**
SJ16	.80	.33	**.67**	**.66**	−.16	**.41**	**.56**	.10	*−.04*	−.12	**.43**	−.10	**.58**
SJ21	.94	.18	**.70**	**.60**	*.06*	**.39**	**.64**	*.04*	.14	*−.06*	**.37**	*−.08*	.19
Isolation													
IS4	.88	.27	**.74**	**.62**	*.02*	**.48**	**.53**	*−.01*	.17	*.01*	.10	**.64**	.24
IS13	.88	.26	**.71**	**.45**	.17	**.60**	**.53**	**.40**	.15	.11	*−.07*	**.55**	*.06*
IS18	.79	.27	**.51**	.25	.12	**.55**	**.42**	**.71**	*.05*	*.01*	−.18	**.42**	*−.08*
IS25	.86	.29	**.74**	**.59**	*.02*	**.52**	**.61**	*.05*	*.07*	*−.03*	*.04*	**.61**	.19
Over-identification													
OI2	.90	.25	**.75**	**.57**	*.09*	**.52**	**.54**	*.01*	.18	.14	*.03*	**.62**	**.31**
OI6	.88	.27	**.73**	**.62**	*.01*	**.48**	**.56**	*−.03*	.15	*.03*	*.01*	**.59**	**.38**
OI20	.92	.14	**.48**	.14	.29	**.52**	**.68**	*.04*	−.40	**.61**	−.49	*.05*	*.04*
OI24	.91	.19	**.62**	**.32**	.24	**.55**	**.43**	.12	.11	**.35**	*−.06*	**.50**	*.06*

N = 1508. Factor loadings above .30 are in bold. Non-significant loadings (based on 95% CI) are italicized. *I-ECV*, item explained common variance; *I-REAL*, item residual absolute loadings; *SC*, self-compassion; *SCr*, self-criticism

best-fitting model, we were interested in the interpretability and replicability of three different factor solutions of the SCS, and the quality and effectiveness of the resulting scoring schemes: a total scale score, two subscale scores, and six subscale scores. For this purpose, after assessing the measurement invariance of the scale, we employed a non-linear factorial approach and a set of statistics scarcely used in validation studies and never used in SCS studies. The analytical strategy we present is aimed to help the researcher and practitioner make decisions about the most appropriate scoring scheme for the scales they need to use.

Regarding our first research question, the best fit was obtained by the bifactor model with six primary factors. However, this model showed complex items and target loadings below 0.50, which is the minimum value recommended for computing subscale scores (Reise et al., 2010). In a study from Neff et al. (2019) examining the same model, we observed that items SK23, MI22, and OI6 had higher loadings on the non-target factor than on the target factor while other items (e.g., SK26) had nearly identical loadings on three factors. Similarly, in Tóth-Király et al. (2017), the presence of complex items was noticeable and some of the group factors did not have sufficient simple

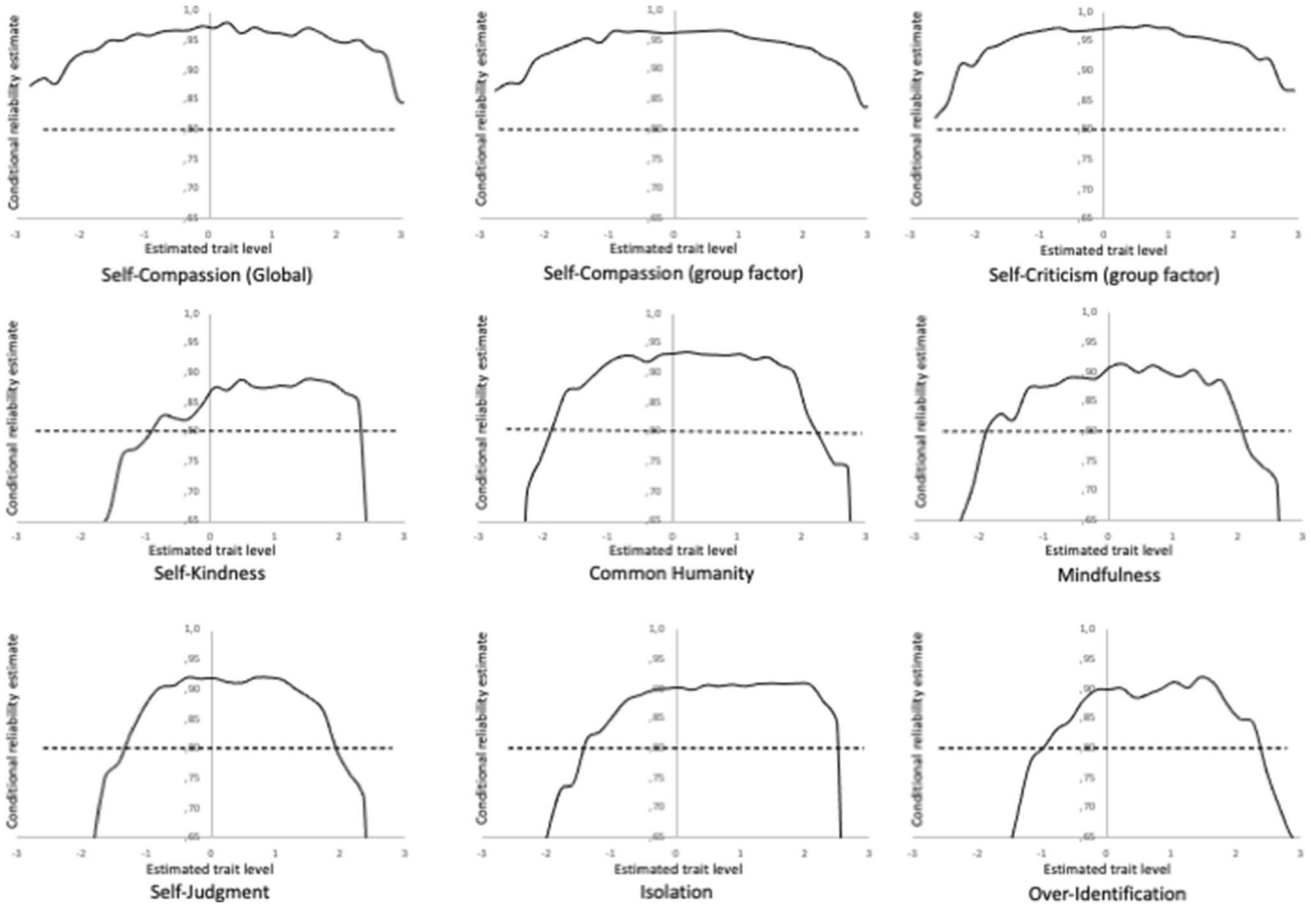

Fig. 1 Cross plots of marginal reliabilities for the total score, two subscale scores, and six subscale scores

indicators to yield identifiability (i.e., at least two items factorially simple in each correlated factor). These authors argued that the presence of non-target cross-loadings can be considered a problem of little relevance if the general factor is well defined, or as an indicator of conceptual overlap between construct facets. However, the presence of complex items and low target loadings negatively affects the interpretability of the factor solution and construct replicability. In our study, although the factor score estimates from three factors (common humanity, self-judgment, and isolation) were determinate and showed adequate reliability values, its metric usefulness was limited because of their narrow effective measurement range. In practical terms, the effective range for the subscales with a total range between 4 and 20 points (common humanity, mindfulness, over-identification, and isolation) was narrowed to between 10 and 14 points, while for the remaining scales with a total range between 5 and 25 points, the effective range was narrowed to scores between 12 and 19. The lack of previous evidence demonstrating the specific reliable variance of each subscale beyond that due to the general factor is an evident limitation to establish whether the six subscales are accurate indicators of unique components of self-compassion.

Our findings indicated that the two-group solution presented a well-defined and interpretable structure, and was replicable across studies. The quality and effectiveness of their factor score estimates indicated that they were (a) good proxies for assessing individual differences and highly correlated with the corresponding latent factors, (b) accurate all along the latent trait, and (c) able to consistently order individuals along the latent trait. Thus, the two subscales could be useful for research and clinical practice. Unfortunately, there are no comparable studies to assess the consistency of our results.

Regarding the second research question, we found evidence of the presence of a dominant latent trait. As in previous studies (Brenner et al., 2017; Zhang et al., 2019), the unidimensional solution only attained a marginally acceptable model fit. This resemblance may be due to the fact that, in general, unidimensional models do not fit perfectly to scales with several dimensions or facets (Rodriguez et al., 2016). An innovative aspect of our analytical approach revealed the good psychometric properties of the unidimensional solution

and the general factor from the bifactor solutions in terms of (a) the strength, quality, and replicability of the factor solution, and (b) the interpretability, accuracy, and determinacy of the factor score estimates derived from it.

Regarding the third research question, we demonstrated that the sum scores and the factor score estimates were appropriate for measurement purposes with the SCS. However, when using a total score, the variance of the factor score estimates from 10 items was already accounted for by other items in the model. That means that the remaining 16 items were providing nearly all the information. So, the researcher and the practitioner can choose to compute the total score of the SCS by (a) summing 16 items, (b) summing the original 26 items, or (c) using the factor score estimates. Options (b) and (c) will produce a slightly more reliable estimation of the latent trait. In practical terms, the difference in using any of them will be very small. To test this, we transformed the three aforementioned scoring schemes into *z*-scores; the differences were only observed in the third decimal place. Interestingly, our "unproductive" items had a remarkable coincidence with items with poor performance in previous studies (e.g., items CH3, CH7, CH10, and CH15 in Finaulahi et al., 2021; Neff et al., 2017; Tóth-Király et al., 2017; item SJ1 in Tóth-Király et al., 2017; and item OI24 in Zhang et al., 2019). Six of these items were not included in the Self-Compassion Scale Short Form (SCS-SF-12; Raes et al., 2011).

Regarding the last research question, this is our proposal. Supported by the existence of a strong and reproducible latent trait showing good psychometric properties, a total score may be suitable for individual measurement and for research with general population where a wide range of levels of self-compassion is expected. Its use may have different purposes (e.g., assessment, classification, and change).

The use of two subscale scores (self-compassion and self-criticism) may also be appropriate for individual assessment and particularly suitable to explore the relationship of each component with psychophysiological variables. We do not recommend the use of only a single subscale score (self-compassion or self-criticism) because its usefulness may be limited (e.g., they only can distinguish high and low self-compassionate individuals). From a psychometric perspective, their removal may reduce the breadth of the construct and jeopardizes the content validity.

The use of six subscale scores is desirable to study the relationship of each facet of self-compassion to external variables. At the empirical level, there is a strict requirement: the observed scores have to demonstrate a high correlation with the "true" score on the latent trait. Unfortunately, we did not find this outcome in our data. This, in addition to the poor psychometric properties of the subscales, raises the need for further studies with different populations before establishing a general recommendation on their use.

At this point, we would like to provide a word of caution. Our recommendations are based on "internal evidence" results in which the positive subscales of the SCS used separately performed poorly. We are aware that both positive and negative components of the SCS are subject to debate (see Montero-Marin et al., 2018; Muris & Otgaar, 2022; Muris et al., 2016). While some authors (e.g., Mantzios et al., 2020) demonstrated that both components were necessary for obtaining effective clinical outcomes, other authors found evidence to the contrary. For example, in a meta-analysis conducted by Muris and Petrocchi (2017), the authors found evidence showing a higher relationship of the negative components of the SCS with psychopathology than that of the positive components with well-being. Accordingly, they recommended not using the negative component of the scale (i.e., self-criticism). This recommendation that should not be ignored was based on findings from external validation procedures more suitable than ours for reaching a consensus on the *true* components of self-compassion. We believe that this consensus is at the end of a long road that will require a methodological effort not yet incorporated into the study of the SCS. For instance, instead of raw scores, interval scores should be used in external validation studies, especially when looking for relationships between instruments where there may be significant percentages of scores at the extremes. Alternatively, studies should test whether raw scores can be substituted by factor score estimates to obtain correct validity inferences. Second, in modern psychometrics, the existence of a scale straddling both unidimensional and multidimensional structures, that is, the construct as both unitary and divisible latent variable, is considered feasible. This is the case of the SCS. For this reason, other external validity strategies that include bias-and-error corrections should be added to the traditional correlation statistics. For example, differential validity procedures would allow testing whether subscale scores (e.g., isolation) are related to the criterion (e.g., depression) as expected when all validity relationships are mediated by the general factor (Ferrando & Lorenzo-Seva, 2019).

In sum, as demonstrated in our study, a marginally acceptable fitting solution may provide optimal scores, while the scores derived from an excellent model fit may be not useful for measuring individuals. Therefore, before choosing the scoring scheme, the researcher should check the quality of the scores in his/her sample and use the statistics only as a reference for making decisions. We also strongly recommend validation studies with alternative mathematical models (e.g., biplot models and sparse principal component analysis) that could provide better insight of the internal structure of the SCS. Bifactor analyses are promising, but

they are not a panacea because they are not free of problems. For instance, their propensity to show good model fit despite the existence of aberrant response patterns is well-known (Reise et al., 2015).

Limitations and Future Research

This study is not exempt from limitations that should be mentioned here. First, we gathered data by using an online survey and our sample consisted mainly of women. This clearly raises some uncertainty about the conditions under which participants responded and limits the generalizability of our results. From our perspective, the high percentage of women observed in the SCS validation studies and, in general, in studies on self-compassion (higher than 66% in a meta-analysis from Yarnell et al., 2015) raises different hypotheses of interest to the researcher. Some related to personality (e.g., differences in empathy), motivation, or, simply, ease of access for the researcher. The results could resolve doubts about the nature of compassion across genders. Although our percentage was very similar to that of previous validation studies of the SCS (e.g., Finaulahi et al., 2021; Montero-Marin et al., 2018; Neff et al., 2017; Tóth-Király & Neff, 2021), we conducted preliminary analyses, such as the assessment of person fit and searching for random or careless responses, to guarantee the quality of the data used. Second, despite the fact that we used some cross-validation procedures to make the results less sample-dependent, our findings should be interpreted with caution because, as in any validation study, they do not reveal an intrinsic property of the instrument, but rather the properties of the scores of a specific sample for specific purposes. Third, we have no evidence in favor or against the use of three subscale scores resulting from a bifactor model with three factors (one for each bipolar dimension). The reason is that our interest was focused on the scoring schemes most frequently used by practitioners and researchers and introducing a new scoring scheme may add confusion and not necessarily progress. Our findings only constitute the first empirical evidence about an unknown aspect of the SCS. We hope that future research will place more emphasis on the quality of scores for measuring self-compassion.

Author Contribution JB conceptualized the current research questions, carried out analyses, and wrote the first draft of the manuscript. AC and JRY designed and executed the study from which data were drawn. AC, JRY, ESZ, and AA collaborated in the writing and editing of the final manuscript. All the authors approved the final version of the manuscript for submission.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The data analyzed in the current study are available at https://osf.io/a98n7/.

Declarations

Ethics Approval The procedures of this study were in accordance with the 1964 Helsinki Declaration and its later amendments, and were approved by the Ethical Review Committee Psychology at Pontifical University of Salamanca, Spain.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Conflict of Interest The authors declare no competing interests.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Brenner, R. E., Heath, P. J., Vogel, D. L., & Credé, M. (2017). Two is more valid than one: Examining the factor structure of the Self-Compassion Scale (SCS). *Journal of Counseling Psychology, 64*(6), 696–707. https://doi.org/10.1037/cou0000211

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307–333). Routledge/Taylor & Francis Group.

Calderón, C., González, D. N., Seva, U. L., & Piera, P. J. F. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor analytic approach for assessing the dimensionality of tests and items. *Psicothema, 31*(4), 450–457. https://doi.org/10.7334/psicothema2019.153

Coroiu, A., Kwakkenbos, L., Moran, C., Thombs, B., Albani, C., Bourkas, S., Zenger, M., Brahler, E., & Körner, A. (2018). Structural validation of the Self-Compassion Scale with a German general population sample. *PLoS ONE, 13*(2), e0190771. https://doi.org/10.1371/journal.pone.0190771

Feldman, C., & Kuyken, W. (2011). Compassion in the landscape of suffering. *Contemporary Buddhism, 12*(1), 143–155. https://doi.org/10.1080/14639947.2011.564831

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762–780. https://doi.org/10.1177/0013164417719308

Ferrando, P. J., & Lorenzo-Seva, U. (2019). An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement, 79*, 437–461.

Ferrando, P. J., & Lorenzo-Seva, U. (2021). The appropriateness of sum scores as estimates of factor scores in the multiple factor analysis of ordered-categorical responses. *Educational and Psychological Measurement, 81*(2), 205–228. https://doi.org/10.1177/0013164420938108

Ferrari, M., Hunt, C., Harrysunker, A., Abbott, M. J., Beath, A. P., & Einstein, D. A. (2019). Self-compassion interventions and psychosocial outcomes: A meta-analysis of RCTs. *Mindfulness, 10*(8), 1455–1473. https://doi.org/10.1007/s12671-019-01134-6

Finaulahi, K. P., Sumich, A., Heym, N., & Medvedev, O. N. (2021). Investigating psychometric properties of the Self-Compassion Scale using Rasch methodology. *Mindfulness, 12*(3), 730–740. https://doi.org/10.1007/s12671-020-01539-8

Foldnes, N., & Grønneberg, S. (2019). On identification and non-normal simulation in ordinal covariance and item response models. *Psychometrika, 84*, 1000–1017. https://doi.org/10.1007/s11336-019-09688-z

García-Campayo, J., Navarro-Gil, M., Andrés, E., Montero-Marin, J., López-Artal, L., & Demarzo, M. M. (2014). Validation of the Spanish versions of the long (26 items) and short (12 items) forms of the Self-Compassion Scale (SCS). *Health and Quality of Life Outcomes, 12*(1), 4. https://doi.org/10.1186/1477-7525-12-4

Linacre, J. M. (2021). A user's guide to winsteps & ministeps: Rasch model computer programs. Retrieved February 3, 2021, from https://www.winsteps.com/index.htm

Lorenzo-Seva, U., & Ferrando, P. J. (2021). FACTOR (Version 11.05.01) [Computer software]. Universitat Rovira i Virgili. https://psico.fcep.urv.cat/utilitats/factor/Download.html/

Lorenzo-Seva, U., & Ferrando, P. J. (2019). A general approach for fitting pure exploratory bifactor models. *Multivariate Behavioral Research, 54*(1), 15–30. https://doi.org/10.1080/00273171.2018.1484339

Mantzios, M., Koneva, A., & Egan, H. (2020). When 'negativity' becomes obstructive: A novel exploration of the two-factor model of the Self-Compassion Scale and a comparison of self-compassion and self-criticism interventions. *Current Issues in Personality Psychology, 8*(4), 289–300. https://doi.org/10.5114/cipp.2020.100791

Montero-Marin, J., Kuyken, W., Crane, C., Gu, J., Baer, R., Al-Awamleh, A. A., Akutsu, S., Araya-Véliz, C., Ghorbani, N., Chen, Z. J., Kim, M. S., Mantzios, M., Rolim Dos Santos, D. N., Serramo López, L. C., Teleb, A. A., Watson, P. J., Yamaguchi, A., Yang, E., & García-Campayo, J. (2018). Self-compassion and cultural values: A cross-cultural study of self-compassion using a multitrait-multimethod (MTMM) analytical procedure. *Frontiers in Psychology, 9*, 2638. https://doi.org/10.3389/fpsyg.2018.02638

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement, 81*(3), 413–440. https://doi.org/10.1177/0013164420942899

Muris, P., & Otgaar, H. (2022). Deconstructing self-compassion: How the continued use of the total score of the self-compassion scale hinders studying a protective construct within the context of psychopathology and stress. *Mindfulness*. Advance online publication. https://doi.org/10.1007/s12671-022-01898-4

Muris, P., Otgaar, H., & Petrocchi, N. (2016). Protection as the mirror image of psychopathology: Further critical notes on the Self-Compassion Scale. *Mindfulness, 7*(3), 787–790. https://doi.org/10.1007/s12671-016-0509-9

Muris, P., & Petrocchi, N. (2017). Protection or vulnerability? A meta-analysis of the relations between the positive and negative components of self-compassion and psychopathology. *Clinical Psychology & Psychotherapy, 24*(2), 373–383. https://doi-org.jerome.stjohns.edu/https://doi.org/10.1002/cpp.2005

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and Identity, 2*(3), 223–250. https://doi.org/10.1080/15298860309027

Neff, K. D. (2022). The differential effects fallacy in the study of self-compassion: Misunderstanding the nature of bipolar continuums. *Mindfulness, 13*, 572–576. https://doi.org/10.1007/s12671-022-01832-8

Neff, K. D., Whittaker, T. A., & Karl, A. (2017). Examining the factor structure of the Self-Compassion Scale in four distinct populations: Is the use of a total scale score justified? *Journal of Personality Assessment, 99*(6), 596–607. https://doi.org/10.1080/00223891.2016.1269334

Neff, K. D., Tóth-Király, I., Yarnell, L. M., Arimitsu, K., Castilho, P., Ghorbani, N., Guo, H. X., Hirsch, J. K., Hupfeld, J., Hutz, C. S., Kotsou, I., Lee, W. K., Montero-Marin, J., Sirois, F. M., De Souza, L. K., Svendsen, J. L., Wilkinson, R. B., & Mantzios, M. (2019). Examining the factor structure of the Self-Compassion Scale in 20 diverse samples: Support for use of a total score and six subscale scores. *Psychological Assessment, 31*(1), 27–45. https://doi.org/10.1037/pas0000629

Neff, K. D., Bluth, K., Tóth-Király, I., Davidson, O., Knox, M. C., Williamson, Z., & Costigan, A. (2021). Development and validation of the Self-Compassion Scale for youth. *Journal of Personality Assessment, 103*(1), 92–105. https://doi.org/10.1080/00223891.2020.1729774

Raes, F., Pommier, E., Neff, K. D., & Van Gucht, D. (2011). Construction and factorial validation of a short form of the Self-Compassion Scale. *Clinical Psychology & Psychotherapy, 18*(3), 250–255. https://doi.org/10.1002/cpp.702

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477.Bifactor

Reise, S. R., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13–40). Routledge.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment, 31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*, 209–220. https://doi.org/10.1037/a0023353

Tóth-Király, I., & Neff, K. D. (2021). Is self-compassion universal? Support for the measurement invariance of the Self-Compassion Scale across populations. *Assessment, 28*(1), 169–185. https://doi.org/10.1177/1073191120926232

Tóth-Király, I., Bőthe, B., & Orosz, G. (2017). Exploratory structural equation modeling analysis of the Self-Compassion Scale. *Mindfulness, 8*(4), 881–892. https://doi.org/10.1007/s12671-016-0662-1

Yarnell, L. M., Stafford, R. E., Neff, K. D., Reilly, E. D., Knox, M. C., & Mullarkey, M. (2015). Meta-analysis of gender differences in self-compassion. *Self and Identity, 14*(5), 499–520. https://doi.org/10.1080/15298868.2015.1029966

Yela, J. R., Gómez-Martínez, M. A., Crego, A., & Jiménez, L. (2020). Effects of the mindful self-compassion program on clinical and health psychology trainees' well-being: A pilot study. *Clinical Psychologist, 24*, 41–54. https://doi.org/10.1111/cp.12204

Zhang, H., Dong, L., Watson-Singleton, N. N., Tarantino, N., Carr, E. R., Niles-Carnes, L. V., Patterson, B., & Kaslow, N. J. (2019). Psychometric properties of the Self-Compassion Scale (SCS) in an African American clinical sample. *Mindfulness, 10*(7), 1395–1405. https://doi.org/10.1007/s12671-019-01099-6