

The multiple group IRT measurement invariance analysis of the Self-Compassion Scale in ten international samples

Martin Kanovský¹, Júlia Halamová^{2}, Nicola Petrocchi³, Helena Moreira⁴, Eunjoo Yang⁵, Jan Benda⁶, Michael Barnett⁷, Elmar Brähler⁸, Xianlong Zeng⁹, Markus Zenger¹⁰*

Abstract

The purpose of this study was to examine the measurement invariance of the Self-Compassion Scale by IRT differential test functioning in ten distinct populations (n = 13623 participants) from ten different countries: Australia (n = 517), China (n = 321), Czech Republic (n = 5081), Germany (n = 2510), Italy (n = 384), Portugal (n = 512), Slovakia (n = 1181), South Korea (n = 1813), Turkey (n = 471), and USA (n = 833). We assessed differential test functioning for the two SCS subscales, Self-compassionate responding and Self-uncompassionate responding separately, because previous bifactor and two-tier analyses of the scale showed the best fit with two separate general factors, and not for the overall score. Only 13 of the 45 comparisons for Self-compassionate responding and 13 of the 45 comparisons for Self-uncompassionate responding (analyses of every pair) demonstrated measurement invariance (no differential test functioning). Generally, our results revealed that the two subscales of Self-compassionate responding and Self-uncompassionate responding were not equivalent among all countries and groups. Therefore, it is impossible to compare overall scores across all countries. Two subscales of the Self-Compassion Scale (Self-compassionate responding and Self-uncompassionate responding) are valid and reliable instruments with substantial potential of use cross-culturally, but results reveal significant cross-cultural differences in the way these two constructs are measured by the subscales of the SCS. Future analyses of the meanings and connotations of this construct across the world are necessary to develop a scale which allows cross-cultural comparisons of various treatment outcomes related to self-compassion.

¹ Institute of Social Anthropology, Faculty of Social and Economic Sciences, Comenius University in Bratislava, Bratislava, Slovakia

² Institute of Applied Psychology, Faculty of Social and Economic Sciences, Comenius University in Bratislava, Mlynské luhy 4, 821 05 Bratislava, Slovakia

³ Department of Economics and Social Sciences, John Cabot University, Rome, Italy

⁴ Cognitive and Behavioural Centre for Research and Intervention, University of Coimbra, Coimbra, Portugal

⁵ Department of Psychology, Korea University, Seoul, South Korea

⁶ Department of Psychology, Faculty of Arts, Charles University, Prague, Czech Republic

⁷ Department of Psychology and Counseling, University of Texas at Tyler, USA

⁸ Department of Medical Psychology and Medical Sociology, University of Leipzig, Leipzig, Germany and Department of Psychosomatic Medicine and Psychotherapy, University of Mainz, Mainz, Germany

⁹ Faculty of Psychology, Beijing Normal University, Beijing, China

¹⁰ Faculty of Applied Human Studies, University of Applied Sciences Magdeburg and Stendal, Germany and Integrated Research and Treatment Center AdiposityDiseases - Behavioral Medicine, Psychosomatic Medicine and Psychotherapy, University of Leipzig Medical Center, Germany

E-mail corresponding author: julia.halamova@gmail.com

**Keywords:**

Self-compassion; Measurement invariance; Differential test functioning; Cross-cultural studies.

Received: 12 October 2020

Accepted: 6 March 2021

Published: 29 April 2021

Citation: Kanovský, M., Halamová, J., Petrocchi, N., Moreira, E., Yang, E., Benda, J., Barnett, M., Brähler, E., Zeng, X., Zenger, M. (2021). The multiple group IRT measurement invariance analysis of the Self-Compassion Scale in ten international samples. *Mediterranean Journal of Clinical Psychology*, 9(1).

<https://doi.org/10.6092/2282-1619/mjcp-2682>

1. Introduction

Inner speech of patients is known to have huge impact on their responsiveness (Shahar et al., 2015) to medical as well as psychological treatment. Inner speech can be in the form of uncompassionate or even harsh self-critical inner speech, which is one of the main risk factors for various types of psychopathology. For example, self-critical inner speech is associated with social anxiety (Shahar et al., 2012), depression (Greenberg & Watson, 2005), posttraumatic stress disorder, self-harming behavior, suicidal tendencies (O'Connor & Noyce, 2008), bipolar affective disorder, schizophrenia, eating disorders, and borderline personality disorder (Mearns et al., 2011). Inner speech in the form of a self-compassionate voice can function as an antidote to the self-critical voice. According to Neff and Dahm (2014, p. 121), "Self-compassion is simply compassion directed inward, relating to ourselves as the object of care and concern when faced with the experience of suffering." Multiple intervention research studies (Conversano et al., 2020; Falconer et al., 2014; Neff & Germer, 2013) have shown that learning to have a self-compassionate inner voice is not only possible but also is an important resilience factor which improves responsiveness to different kinds of treatment (Merlo et al., 2020; Terry & Leary, 2011; Warren et al., 2016) and it is related to many behavioural factors of happy and healthy life too. Therefore, it is crucial to provide a reliable, valid and invariant instrument for measuring levels of self-compassion which can enable comparisons in health treatment across and within different countries.

Neff (2003) developed a scale to measure self-compassion, the Self-Compassion Scale – SCS, which is the most frequently used scale for measuring self-compassion in both research and clinical practice. Despite a wide range of translations of the SCS and its increasing popularity among researchers and clinicians, to the best of our knowledge, there is no evidence that the SCS is invariant across different countries and language translations. There is currently only one invariance study with the SCS, in which Montero-Marín et al. (2016) compared Brazilian and Spanish samples in the dimension of SCS, which consists of negatively-worded items and did

not find strong construct invariance between them. However, the original long format, English-language version of the SCS has now been translated into eighteen different languages. These are Chinese (Chen et al., 2011), Czech (Benda & Reichová, 2016), Dutch (López et al., 2015), French (Kotsou & Leys, 2016), German (Hupfield & Ruffieux, 2011), Greek (Mantzios et al., 2013), Hungarian (Tóth-Király et al., 2016), Iranian (Azizi et al., 2013), Italian (Petrocchi et al., 2014), Japanese (Arimitsu, 2014), Korean (Lee & Lee, 2010), Norwegian (Dundas et al., 2016), Portuguese (Castilho & Pinto-Gouveia, 2011), Slovak (Halamová et al., 2017), Spanish (García-Campayo et al., 2014), Thai (Pisitsungkagarn et al., 2014), and Turkish language (Deniz et al., 2008). In addition, Neff (2016) suggests that research findings obtained with translations should not be automatically generalized to the original version of the SCS, because of potential difficulties with the quality of translations and/or potential cultural factors (Behling & Law 2000).

In a previous study of the factor structure of the SCS, in different countries and samples, Halamová et al. (2021) found that the most reliable and appropriate use of the SCS is not computing the overall score, but rather calculating two dimensional scores for Self-compassionate responding and Self-uncompassionate responding separately. We note that this conclusion was not based on testing simple two-factor model, but rather more general two-tier model with two general factors and six specific factors. Considering that the SCS (Neff, 2003) is the most frequently used scale of self-compassion across cultures, it is crucial to inspect whether the use of the SCS in comparisons between countries and languages is justified. Therefore, it is vital to scrutinize the psychometric properties and measurement invariance of the SCS in various languages in addition to English (Neff et al., 2017).

1.1 The Self-compassion scale and culture

Research is scarce when it comes to exploring whether self-compassion levels differ across different cultures, and only one study has tested measurement invariance of the SCS between its different translations Montero-Marín et al. (2016). Other studies have compared the levels of self-compassion between different cultures and societies using the SCS, but they did not previously test whether the measure was invariant in the different contexts (e.g., Birkett, 2014; Khramtsova & Chuykova, 2016; Neff et al., 2008), producing various findings which we discuss below. Nevertheless, all these findings must be considered with caution since measurement invariance, a prerequisite for meaningful differences, was not previously tested in these studies. SCS scores have been shown to be significantly and positively associated with well-being in all cultures, suggesting that people may benefit from self-compassion despite cultural differences (Neff et al., 2008). However, Neff and Vonk (2009) proposed that various cultures provide different messages regarding the meaning and value of self-compassion or its counterpart, and that individual variation in SCS scores within a particular culture may be partially dependent on

personal acceptance or rejection of these cultural messages. Clearly, there are unique aspects of every society that must be considered if we want to explore the impact of culture on levels of Self-compassionate responding and Self-uncompassionate responding. For example, differences in political and historical backgrounds, religious beliefs and cultural worldviews, or variation in parenting practices may have an impact. Individuals may develop their way of relating towards one's self on the basis of their interactions with other people (mainly caregivers), so that experiences with them are internalized and expressed either as a self-compassionate or a self-critical (self-uncompassionate) inner voice (Gilbert & Irons, 2008).

Indeed, results of Neff et al. (2009) indicated that three cultures (USA, Thai, and Taiwanese) differed significantly in SCS scores from one another, although differences within cultures in SCS scores were as large as differences between cultures. Furthermore, a cross-cultural study examined differences in self-compassion between Chinese and American undergraduates and found a non-significant difference in the overall Self-Compassion score (Birkett, 2014). However, Chinese students reported significantly higher levels of both positive and negative aspects of self-compassion than American students, the authors interpreted this as Chinese students experiencing aspects of self-compassion in both the positive and negative ways. Additionally, SCS scores from both Chinese and American students in Birkett's study are similar to scores of American undergraduates in a cross-cultural study including Thai and Taiwanese students conducted in previous research (Neff et al., 2008). According to Khramtsova and Chuykova (2016), American participants were higher on self-compassion than Russian participants measured by the shorter version of SCS. Kwan et al. (2009) investigated cultural differences while exploring sources of self-esteem. They suggested that self-compassion was one of three major sources of self-esteem together with self-efficacy and narcissism. In their study, they found that the Chinese participants had significantly higher levels of self-compassion and narcissism than the American participants. However, narcissism was significantly correlated with levels of self-compassion among the Chinese participants but not the American participants.

It appears that the types of constructs that correlate with self-compassion vary across cultures. With regards to semantics, Kitayama and Karasawa (1997) found that in the East Asian culture of Japan, individuals tend to have positive feelings towards themselves while still being self-critical. Heine et al. (1999) argue that self-criticism is not a psychological problem for individuals living in interdependent and collectivist cultures. This suggests that although some people may display higher levels of self-uncompassionate responding than other people, they may not necessarily display lower levels of self-compassionate responding. Additionally, Zeng et al. (2016) reported that SCS cannot be validated among Chinese Buddhists, as Self-kindness and Common Humanity were neither correlated with their opposite dimensions nor linked to better

emotional outcomes. They further illustrated that the ideas of Self-kindness and Common Humanity reflected in SCS were different from Buddhist philosophy and culture, although the concept of self-compassion originated from Buddhism. In all, empirical studies showed many cultural issues may impact the understanding and function of self-compassion.

However, there are only a few and very recent extant studies dealing with measurement invariance of the SCS (Costa et al., 2016; Cunha et al., 2016; Montero-Marín et al., 2016), and only one of them deals with equivalence across various cultures. According to Costa et al. (2016) a weak measurement and structural configural invariance of the two-factor model of the SCS across clinical samples (diagnosed with borderline personality disorder, anxiety disorder, eating disorder) and general populations showed that both properties and interpretations of scores of SCS were equivalent. Results of Cunha et al. (2016) confirmed the measurement invariance across genders for Portuguese adolescents. Montero-Marín and colleagues (2016) found that configural invariance (i.e., common factors are associated with the same items across groups) and partial metric invariance (i.e., weak factorial - common factors have the same meanings across groups) were achieved but not scalar invariance (i.e., strong factorial – comparisons of latent groups means are meaningful). When comparing the negative dimension of the SCS between Brazilian and Spanish samples, the authors found that the first-order factor loadings were not the same in the two groups. Although both Brazilian and Spanish samples reported equivalence in their understanding of the positive factor of the SCS (Self-compassionate responding), the equivalence was not detected in the negative part of the SCS, and therefore the authors could not compare the mean levels of the latent variables between the groups. This means that, generally, the scale could not be considered invariant.

In summary, studies using the SCS around the world indicate that self-related processes captured by the SCS appear to be useful and meaningful constructs across cultures (Neff & Vonk, 2009). However, previous research on the links between SCS and culture suggest potential cultural differences in levels of self-compassion (Self-compassionate responding and Self-uncompassionate responding). To date, very few differences in SCS scores have been found according to cultural background, which appears to play an important role in levels of self-compassion. Certainly, more research on cultural differences in self-compassion and self-criticism is necessary (Voruz, 2013). To develop this area of research, more cross-cultural comparisons on measurement invariance of tools assessing self-compassion is needed, which is the aim of the present study.

1.2 Aim of the Study

Stemming from the existing literature on the SCS, the aim of this study was to investigate measurement invariance by IRT differential test functioning of the ten different samples and nine language versions of the two dimensions of the SCS, Self-compassionate responding and

Self-uncompassionate responding (Halamová et al., 2021), with the main goal of determining whether comparisons between total scores of the two dimensions of SCS across countries and languages are justified.

2. Research Methods

2.1 Measuring Instrument

The Self-Compassion Scale (Neff, 2003) measures six aspects which, according to the author of the scale, constitute a self-compassionate response to difficult circumstances. The scale includes 26 items rated on a 5-point Likert-type format of frequency (1 = almost never; 5 = almost always). The subscale Self-Kindness represents the ability of taking care and being warm towards oneself when encountering difficult situations. Self-Judgment reflects how critical the individuals behave towards themselves. Common Humanity reflects the personal understanding that suffering is part of the shared human experience. Isolation relates to the sense of loneliness and isolation after failure. Mindfulness is a non-judgmental state of mind in which individuals observe their thoughts and feelings as they are, without trying to suppress or deny them. Over-identification captures responses to challenging situations that involve becoming absorbed with negative thoughts and feelings. The scale as a whole measures the degree to which individuals display self-kindness against self-judgment, common humanity versus isolation, and mindfulness versus over-identification. Neff (2016) perceives these components as distinct and at the same time influencing each other (see Table 1 for items description). Negative and positive items are represented in roughly equal proportions. Self-Judgment, Isolation, and Over-identification are reverse coded. After reverse-coding of negative items, means are calculated for each subscale, and a grand mean is calculated that is considered to represent a global measure (Neff, 2016) of self-compassion: higher overall scores indicate greater levels of self-compassion (Neff, 2003). The reliability and validity of the overall SCS scale and of its six subscales appear to be quite good in various translations, as well as in different corresponding countries (Halamová et al., 2021). However, recent cross-cultural findings on the factorial structure of the SCS have failed to replicate the higher order factor of self-compassion originally proposed by Neff, suggesting that a total score of the scale may not be appropriate to use as an overall measure of self-compassion across different translations and cultures (Halamová et al., 2021). Findings by Halamová and colleagues (2021) justify the use of two separate dimensions of the SCS, namely Self-compassionate responding and Self-uncompassionate responding. Thus, in the present paper, the Self-compassionate responding subscale and the Self-uncompassionate responding subscale will be analysed separately. From now on, in order to avoid potential confusion with results pertaining the overall Self-compassion score, we will refer to the two dimensions of the SCS as Self-compassionate responding subscale and Self-uncompassionate responding subscale.

Table 1. Dimensions and scale items of The Self-compassion scale

Dimensions	Scale items
Self-compassionate responding	
Self-Kindness	5. I try to be loving towards myself when I'm feeling emotional pain. 12. When I'm going through a very hard time, I give myself the caring and tenderness I need. 19. I'm kind to myself when I'm experiencing suffering. 23. I'm tolerant of my own flaws and inadequacies. 26. I try to be understanding and patient towards those aspects of my personality I don't like.
Mindfulness	9. When something upsets me I try to keep my emotions in balance. 14. When something painful happens I try to take a balanced view of the situation. 17. When I fail at something important to me I try to keep things in perspective. 22. When I'm feeling down I try to approach my feelings with curiosity and openness.
Common Humanity	3. When things are going badly for me, I see the difficulties as part of life that everyone goes through. 7. When I'm down and out, I remind myself that there are lots of other people in the world feeling like I am. 10. When I feel inadequate in some way, I try to remind myself that feelings of inadequacy are shared by most people. 15. I try to see my failings as part of the human condition.
Self-uncompassionate responding	
Self-Judgment	1. I'm disapproving and judgmental about my own flaws and inadequacies. When times are really difficult, I tend to be tough on myself. 8. I'm intolerant and impatient towards those aspects of my personality I don't like. 11. When I see aspects of myself that I don't like, I get down on myself. 16. I can be a bit cold-hearted towards myself when I'm experiencing suffering. 21. suffering.
Over-identification	2. When I'm feeling down I tend to obsess and fixate on everything that's wrong. 6. When I fail at something important to me I become consumed by feelings of inadequacy. 20. When something upsets me I get carried away with my feelings. 24. When something painful happens I tend to blow the incident out of proportion.
Isolation	4. When I think about my inadequacies, it tends to make me feel more separate and cut off from the rest of the world. 13. When I'm feeling down, I tend to feel like most other people are probably happier than I am. 18. When I'm really struggling, I tend to feel like other people must be having an easier time of it. 25. When I fail at something that's important to me, I tend to feel alone in my failure.

2.2 Sampling procedure

As we used the same research samples as the study of Halamová et al. (2021) the sampling procedure was also the same. The first author of this study wrote around 130 emails inviting cooperation to all authors of already published research studies as found at www-self-compassion.com and from Google scholar using the following search terms, “self-compassion scale” and “neff”. Inclusion criteria were at least 300 adult participants in a sample (to allow convergence of IRT models) and using 26 items of SCS and a non-clinical population. We sent out 130 number of emails and 14 number of people both agreed to and used the data. After preliminary analysis, we excluded three of four Turkish data sets because the statistical models failed to converge. The Netherland sample was also excluded due to use of a 24-item scale while the rest of the samples used the 26-item scale. After all sampling procedures and established collaborations, we included 10 data sets for further analyses.

2.3 The research samples and procedures from different countries

Out of currently eighteen different language versions of SCS we obtained data from nine of them. Our research samples consist of two distinct English language samples from different countries (Australia (n = 517) and USA (n = 833) and eight different language translations samples eight different countries namely, China (n = 321), Czech Republic (n = 5081), Germany (n = 2510), Italy (n = 384), Portugal (n = 512), Slovakia (n = 1181), South Korea (n = 1813), and Turkey (n = 471). See Table 2 for more information about the samples. In total, we tested 10 distinct samples including 13623 participants. The data collection from all these samples was in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Table 2. Sample information and internal consistency coefficients for SCS in ten different samples and ten various countries

Country	N	Female %	M Age	SD	Language	Cronbach α		
						Total	SUCR/SCR	SUCR/SCR
Australia	517	100	53.64	5.49	English	.94	.92/.90	.59***
Czech Republic	5081	71.80	28.20	11.10	Czech	.91	.86/.87	.64***
Germany	2510	53.70	50.19	17.36	German	.88	.87/.88	.14***
China	321	53.28	25.45	4.10	Chinese	.90	.89/.88	.35***
Italy	384	67.40	33.38	10.55	Italian	.93	.91/.89	.54***
Portugal	512	66.20	42.48	6.11	Portuguese	.87	.87/.85	.26***
Slovakia	1181	66.00	30.30	12.40	Slovak	.86	.86/.84	.27***
South Korea	1813	48.80	39.28	11.27	Korean	.91	.93/.91	.23***
Turkey	471	83.40	20.65	1.35	Turkish	.88	.78/.79	.78***
USA	833	71.20	21.00	3.86	English	.89	.93/.89	.17***

Note. M = Mean. SD = standard deviation. SUCR = Self-uncompassionate responding. SCR = Self-compassionate responding.

Australia

The participants from Australia were originally enrolled from the electoral roll (Murray et al., 2004) and then selected for the further study (Brown et al., 2014).

China

An online sample service enrolled participants from China (Zeng et al., 2016).

Czech Republic

The participants from the Czech Republic (Benda & Reichová, 2016) were recruited through Facebook and respondents under 18 were excluded.

Germany

The participants from Germany are representative for the German general population and they were recruited by the random-route-technique (Körner et al., 2015).

Italy

An online survey company recruited the participants from Italy (Petrocchi et al., 2014).

Portugal

The Portuguese sample was collected in public schools and in the general community (Moreira et al., 2015, 2016).

Slovakia

Slovak participants were recruited mainly online by convenience sampling (Halamová et al., 2017).

South Korea

The participants from South Korea were recruited by an online survey company (Hwang et al., 2016).

Turkey

Turkish participants were undergraduate students enrolled at a university (Şenyuva et al., 2014).

USA

Participants were recruited through the university research website (Barnett & Sharp, 2016).

2.4 Data Analysis

In testing measurement invariance/equivalence, confirmatory factor analysis (CFA) is the prevalent approach (Vanderberg & Lance, 2000). Despite the obvious advantages of IRT, these models are not used frequently. In psychometric literature, there is an ongoing debate among researchers comparing these two approaches (Kankaraš et al., 2011; Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Raju et al., 2002; Reise et al., 1993). While CFA models assume that

the item responses are continuous and linear, IRT models assume the item responses are either nominal or ordinal. Unlike CFA models, IRT models are inherently non-linear. Furthermore, CFA models typically estimate a single intercept per item because they use a continuous assumption. On the contrary, IRT models typically compute multiple parameters (thresholds) analogous to item intercepts per item – IRT models the polytomous data as categorical, and as a consequence IRT models usually result in greater sensitivity to more-nuanced group differences such as in central tendency or the use of extreme scores. Recent research shows that IRT models can quite accurately detect inequivalence in the intercept and slope parameters both at the scale and the item level (Kankaraš et al., 2011). However, CFA performs well only when inequivalence is located in the slope parameters, but wrongfully indicates inequivalence in the slope parameters when inequivalence is located in the intercept parameters (Kankaraš et al., 2011).

IRT non-linear factor analyses were used for models, but we acknowledge that non-linear confirmatory factor analysis is available (via WLSMV estimator in Mplus software and R package lavaan). Difference between non-linear IRT and non-linear CFA is that IRT method is full-information and CFA non-linear method is not (DeMars, 2012). The main advantage of the full information IRT approach is that item modelling relationships are far more general in that non-monotonic, non-parametric, or just plain customized item response models are easier to cope with because we can avoid worries about the sufficiency of using the polychoric correlation matrix used in the CFA non-linear approach (WLSMV). However, we should note that performing robust Maximum Likelihood estimation (MLR) (e.g., in Mplus) based on categorical variables with the integration method in order to get IRT-parameters through CFA could avoid the limitations of polychoric correlations.

SCS subscales (Self-compassionate responding and Self-uncompassionate responding) are unidimensional, and one general latent factor explains a substantive part of the variance – as it was shown in previous research (Halamová et al., 2021) by means of bifactor models and non-parametric IRT Mokken scale analysis. However, we performed our analyses for each population separately. Our previous results provided no information about whether the test score was comparable across different populations. IRT models are better suited than CFA models to explore this issue. The CFA measurement invariance analyses provide insights regarding the relationship between latent factors, so their use is preferable when the goal is to answer questions of the invariance of a multifactorial framework and many other aspects, such as configural invariance, metric invariance, scalar invariance, invariant uniquenesses, invariant factor variances, invariant factor covariance, equal factor means, etc. (Vandenberg & Lance, 2000). Since we decided to test the invariance of single, unidimensional scales (Self-

compassionate responding and Self-uncompassionate responding, respectively), IRT analyses are more desirable.

However, the results from previous research of Halamová et al. (2021) recommending two separate scores do not argue in favour of the simple two-factor model – they conclude that the two-tier model has better fit than the bifactor model. The two-tier model of SCS scale has two general factors and six specific factors, and the bifactor model of SCS scale has only single general factor and six specific factors. The conclusion of the work of Halamová et al. (2021) is that the use of the single total (sum) score is not justified because the single general factor is not essentially unidimensional – therefore we should use two separate sum scores which sufficiently capture the variance. This conclusion does not entail that the two-factor model is feasible. We cannot use the two-factor model to test measurement invariance in the standard confirmatory way: the fit of such model with data is far below the acceptable values. To repeat, the argument in the previous manuscript (Halamová et al., 2021) was that only two separate scales of SCS are unidimensional enough to capture sufficient amount of variance – and present analysis is based on this conclusion assuming the essential unidimensionality of separate scales and using the IRT DTF procedure to detect differential test functioning. Secondly, we are unsure how to settle the issue of whether these two general factors are response style / method factors or really different constructs: some authors (e.g., Montero-Marín et al., 2016) argue that self-criticism (or self-uncompassionate responding) is in fact a different construct than self-compassion, but we believe that the possibility of response style / method factors cannot be ruled out. However, even if two factors are response style / method factors due to the reverse scoring of their items, this is a further argument in favour of testing them separately. Thirdly, fitting the nested (second-order / hierarchical) factor models does not allow for the calculation of some indices, such as Omega hierarchical or ECV (see Mansolf & Reise, 2017 for theoretical justification and Cucina & Byle, 2017 for meta-analysis).

In the context of IRT models, measurement equivalence is tested by inspecting the differential item functioning (DIF), and/or the differential test functioning (DTF). Differential item functioning (DIF) means that an item within the SCS questionnaire measures the construct (self-compassion) differently for one population than for another. As a consequence, the presence of DIF compromises test validity. If this item bias accumulates to the extent to produce biased overall test scores, a test displays differential test functioning (DTF). DTF is present when respondents who have the same standing on the latent ability, but belong to different groups, obtain different scores on the test.

DIF is routinely tested and when it is found, some method of purification is usually adopted: items with DIF are flagged and removed. However, if a test has many items (e.g., SCS has 26 items) and only some of them have DIF, then the impact of these DIFs on the overall test score

may be negligible (DeMars, 2011). We can consider yet another situation: there could be large DIF effects in favour of one population for some items, but these effects are simultaneously cancelled out by DIF for other items in favour of other populations. Therefore, the presence of DIF for some items does not necessarily imply that the overall test itself is biased. On the other hand, it is also possible to have DTF in a situation where little or no DIF has been detected. Nontrivial DTF can occur in cases when the parameters systematically favour one group over another. Consequently, the aggregate of these small, insignificant differences at the item level can become substantial at the test level (Chalmers et al., 2016). DTF has greater practical sense for our purpose than DIF: we do not intend to inspect particular items of SCS subscales, nor do we intend to improve/purify them. Our purpose is more practical: we intend to test the assumption that the (expected) total score of the SCS subscales is equivalent across different populations, therefore only the latent ability – and not the belonging to a particular group – has any impact on the (expected) total score.

Testing the DTF involves two statistical measures (Chalmers et al., 2016). The first statistic (the signed DTF) tests whether there is any systematic test scoring bias, indicating that some group consistently scored higher across a specified range of latent ability. The second statistic (the unsigned DTF) assesses whether the test curves have a large degree of overall separation on average, suggesting that there may be substantial DTF at particular levels of latent ability. The signed DTF values can range from $-TS$ to TS (TS stands for the highest possible test score). Negative values of the signed DTF indicate that the reference group scores systematically lower than the focal group on average, while positive values indicate that the reference group scores higher. The unsigned DTF ranges from 0 to TS because the area between the curves is zero when the test scoring functions have exactly the same functional form. The signed DTF values are always lower than or equal to the unsigned values: when the curves do not cross, the signed DTF is equal to the unsigned DTF. If there is a small value for the signed DTF and a large value for the unsigned DTF, test curves intersect at one or more locations to create a balanced overall score, but there is substantial bias at particular levels of latent ability.

To test for differences among all countries, the multilevel IRT models are appropriate. We did not use traditional and popular GLMM model (e.g., `lmer` function in package `lme4`; see De Boeck et al., 2011), because it did not allow graded-response polytomous models – such models would require multivariate logits and estimation of different discrimination parameters. Instead, we used the extended mixed-effects IRT model (Chalmers, 2015).

For all statistical analyses, the “mirt” package (Chalmers, 2012) in statistical software R (R Core Team, 2017) was used. Our analysis proceeded as follows.

(1) We performed pairwise tests of all samples, separately for Self-compassionate responding and Self-uncompassionate responding: the total number of tests was $2 * ((10 * 9) / 2) = 90$. We

reported only the signed DTF values because they are of more practical use than the unsigned DTF values, as we explained above. The fitted IRT models were Samejima graded response models, which are in fact the sequences of 2 PL models.

(2) For samples with insignificant signed DTF equivalence obtained, we also reported the latent means difference (for Self-uncompassionate responding subscale and Self-compassionate responding subscale). Latent means in the reference group were constrained to zero, and latent means in the focal group were estimated.

(3) We performed the multilevel IRT analysis of all samples, separately for Self-compassionate responding and Self-uncompassionate responding, to check for relative differences among countries. We reported random effects.

3. Results

To justify the use of the two subscales, we report here the fit of two-tier models from previous research (Halamová et al., 2021), see Table 3. Two-tier models included two general factors and six specific factors, and previous research demonstrated that single general factor is not essentially unidimensional to capture sufficient amount of common variance, and that two general factors are necessary. In Table 4, we present the fit of multiple-group IRT models used for the DTF analyses.

Table 3. Fit Indices of Two-Tier IRT Models of 26-Item SCS

Sample	IRT model			
	CFI	TLI	RMSEA	SRMR
AUS	.954	.941	.051	.060
CHI	.910	.887	.053	.098
CZK	.905	.880	.063	.057
GER	.953	.940	.053	.087
ITA	.907	.885	.062	.066
POR	.932	.913	.052	.108
SVK	.939	.922	.044	.090
KOR	.956	.944	.065	.103
TUR	NC	N/A	N/A	N/A
USA	.950	.936	.059	.092

Note. NC = model failed to converge. N/A = no information due to non-convergence of the model. AUS = Australia (n = 517). CHI = China (n = 321). CZK = Czech Republic (n = 5081). GER = Germany (n = 2510). ITA = Italy (n = 384). POR = Portugal (n = 512). SVK = Slovakia (n = 1181). KOR = South Korea (n = 1813). TUR = Turkey (n = 471). USA (n = 833).

Table 4. Fit Indices of multiple-group IRT Models

Samples	Self-uncompassionate				Self-compassionate			
	CFI	RMSEA	SRMSR 1	SRMSR 2	CFI	RMSEA	SRMSR 1	SRMSR 2
AUS/CZK	.911	.042	.036	.055	.968	.034	.049	.023
AUS/GER	.956	.025	.036	.066	.945	.028	.045	.021
AUS/CHI	.962	.033	.034	.031	.939	.051	.046	.081
AUS/ITA	.936	.043	.043	.028	.934	.055	.052	.077
AUS/KOR	.960	.055	.036	.066	.944	.041	.049	.060
AUS/POR	.917	.043	.036	.077	.923	.055	.044	.070
AUS/SVK	.934	.048	.036	.071	.958	.032	.048	.071
AUS/TUR	.967	.045	.036	.079	.935	.055	.042	.082
AUS/USA	.901	.052	.043	.057	.923	.046	.044	.079
CZK/GER	.938	.037	.037	.082	.917	.026	.036	.077
CZK/CHI	.900	.042	.037	.028	.903	.032	.036	.060
CZK/ITA	.901	.042	.036	.065	.905	.028	.036	.038
CZK/KOR	.903	.046	.037	.095	.904	.028	.036	.032
CZK/POR	.902	.042	.038	.065	.908	.028	.037	.063
CZK/SVK	.908	.044	.036	.061	.948	.023	.036	.079
CZK/TUR	.928	.041	.036	.077	.928	.030	.036	.066
CZK/USA	.908	.044	.038	.070	.909	.027	.037	.071
GER/CHI	.907	.045	.048	.068	.951	.031	.040	.054
GER/ITA	.907	.039	.047	.072	.951	.033	.040	.057
GER/KOR	.908	.048	.049	.078	.914	.034	.039	.071
GER/POR	.908	.047	.048	.077	.949	.034	.038	.077
GER/SVK	.907	.050	.050	.073	.969	.027	.038	.078
GER/TUR	.924	.046	.047	.078	.952	.041	.037	.075
GER/USA	.906	.052	.048	.074	.939	.035	.040	.082
CHI/ITA	.910	.034	.045	.082	.913	.056	.047	.079
CHI/KOR	.908	.056	.045	.077	.906	.041	.047	.077
CHI/POR	.916	.034	.046	.073	.904	.056	.047	.079
CHI/SVK	.909	.050	.045	.077	.957	.036	.051	.074
CHI/TUR	.945	.051	.045	.041	.926	.068	.048	.072
CHI/USA	.907	.048	.045	.079	.906	.059	.051	.076
ITA/KOR	.902	.057	.037	.071	.907	.036	.049	.075
ITA/POR	.909	.043	.036	.072	.903	.049	.048	.069
ITA/SVK	.910	.050	.040	.075	.964	.032	.051	.058
ITA/TUR	.962	.047	.034	.069	.937	.056	.050	.080
ITA/USA	.905	.045	.035	.077	.925	.044	.046	.076
KOR/POR	.913	.056	.035	.075	.924	.050	.040	.057
KOR/SVK	.917	.056	.036	.071	.910	.043	.040	.078
KOR/TUR	.925	.064	.034	.069	.908	.057	.040	.077
KOR/USA	.900	.052	.035	.079	.904	.060	.038	.078
POR/SVK	.925	.057	.055	.077	.974	.025	.060	.072
POR/TUR	.905	.068	.053	.082	.962	.042	.058	.079
POR/USA	.903	.051	.053	.080	.910	.051	.062	.081
SVK/TUR	.907	.058	.046	.071	.965	.032	.050	.071
SVK/USA	.905	.058	.045	.080	.921	.040	.049	.079
TUR/USA	.909	.060	.071	.083	.914	.057	.070	.080

Note. AUS = Australia (n = 517). CHI = China (n = 321). CZK = Czech Republic (n = 5081). GER = Germany (n = 2510). ITA = Italy (n = 384). POR = Portugal (n = 512). SVK = Slovakia (n = 1181). KOR = South Korea (n = 1813). TUR = Turkey (n = 471). USA (n = 833).

We obtained 13 measurement equivalencies out of 45 comparisons for the Self-uncompassionate responding subscale (see Table 5) and 13 measurement equivalencies for the Self-compassionate responding subscale (see Table 6). We have to note that we did not propose any overall null-hypothesis (e.g., that all samples will be invariant across the world): p-values in Tables 5 and 6 are based on multiple imputation estimate (1000 draws) of the expected test scores for each sample and we had the null-hypothesis for each particular sample (i.e., that the bias of total scores between two particular samples is zero).

For the Self-uncompassionate responding subscale, the Czech sample was equivalent to six of the other samples; the Turkish and Chinese samples to four other samples; the Slovak, German, Italian, Korean, and USA samples to two other samples; the Australian sample to one other sample; and finally Portugal sample was equivalent to no other sample. As for the Self-compassionate responding subscale, the Czech and Italian samples were equivalent to four other samples; the Slovak, German, Portugal and USA samples to three other samples; the Australian, Chinese and Korean samples to two other samples; and the Turkish sample to no other sample.

Table 5. Signed differential test statistics for the Self-uncompassionate responding subscale

sDTF	AUS	CZK	GER	CHI	ITA	KOR	POR	SVK	TUR
CZK	-0.45, p = .063	-	-	-	-	-	-	-	-
GER	3.62, p < .001	2.44, p < .001	-	-	-	-	-	-	-
CHI	-1.17, p < .007	-0.39, p = .279	-0.59, p = .143	-	-	-	-	-	-
ITA	-1.24, p < .002	-0.27, p = .432	-1.28, p < .001	-0.02, p = .973	-	-	-	-	-
KOR	2.02, p < .001	1.51, p < .001	-0.14, p = .532	4.39, p < .001	4.12, p < .001	-	-	-	-
POR	1.11, p < .006	0.72, p < .032	0.91, p < .008	2.72, p < .001	2.43, p < .001	1.13, p < .001	-	-	-
SVK	-0.87, p < .003	-0.02, p = .918	-1.73, p < .001	1.03, p < .004	0.85, p < .007	-0.92, p < .001	-1.74, p < .001	-	-
TUR	-1.68, p < .001	-0.20, p = .541	-1.33, p < .001	-0.36, p = .424	-0.51, p = .209	-0.69, p = .056	-2.26, p < .001	-0.72, p < .032	-
USA	1.18, p < .002	0.02, p = .939	-1.31, p < .001	0.91, p < .016	0.69, p < .047	-0.80, p < .001	-1.57, p < .001	-0.11, p = .678	1.17, p < .002

Note. Insignificant sDTFs (equivalence between expected total scores obtained) are highlighted in bold.

AUS = Australia (n = 517). CHI = China (n = 321). CZK = Czech Republic (n = 5081). GER = Germany (n = 2510). ITA = Italy (n = 384). POR = Portugal (n = 512). SVK = Slovakia (n = 1181). KOR = South Korea (n = 1813). TUR = Turkey (n = 471). USA (n = 833).

Table 6. Signed differential test statistics for the Self-compassionate responding subscale

sDTF	AUS	CZK	GER	CHI	ITA	KOR	POR	SVK	TUR
CAN	2.10, p < .001	-	-	-	-	-	-	-	-
CZK	-1.87, p < .001	-	-	-	-	-	-	-	-
GER	-1.00, p < .001	0.56, p < .001	-	-	-	-	-	-	-
CHI	2.86, p < .001	0.97, p = .052	0.04, p = .938	-	-	-	-	-	-
ITA	-0.85, p < .044	0.03, p = .931	-0.31, p = .392	-3.98, p < .001	-	-	-	-	-
KOR	-0.04, p = .889	0.79, p < .001	0.51, p < .005	-3.95, p < .001	1.55, p < .001	-	-	-	-
POR	0.77, p = .069	0.72, p < .032	0.62, p = .076	-1.98, p < .004	1.65, p < .001	0.73, p = .051	-	-	-
SVK	-1.59, p < .001	-0.10, p = .619	-0.63, p < .003	-5.61, p < .001	-0.39, p = .229	-0.89, p < .001	-2.15, p < .001	-	-
TUR	-2.59, p < .001	-1.02, p < .002	-1.51, p < .001	-6.09, p < .001	-1.90, p < .001	-1.43, p < .001	-3.04, p < .001	-1.37, p < .001	-
USA	1.14, p < .002	-0.01, p = .987	-0.54, p < .025	-5.10, p < .001	-0.33, p = .355	-0.62, p < .013	-1.77, p < .001	-0.13, p = .608	1.50, p < .001

Note. Insignificant sDTFs (equivalence between expected total scores obtained) are highlighted in bold. AUS = Australia (n = 517). CHI = China (n = 321). CZK = Czech Republic (n = 5081). GER = Germany (n = 2510). ITA = Italy (n = 384). POR = Portugal (n = 512). SVK = Slovakia (n = 1181). KOR = South Korea (n = 1813). TUR = Turkey (n = 471). USA (n = 833).

Note that no perfect transitivity was present: for example, as to the Self-uncompassionate responding subscale, both USA and Australian samples were equivalent to the Czech sample, but they were not equivalent one to another. Therefore, we could not create a single linear rank based on the differences in latent means of equivalent samples, but rather clusters of mutually comparable samples. For example, again in the case of the Self-uncompassionate responding subscale, we could fully compare Czech, Slovak, and USA samples because all of them were mutually equivalent. However, adding another sample – for instance, Italian – was not possible: it was equivalent with Czech sample, but not with USA and Slovak samples. Another possible cluster was composed of Czech, Chinese, Turkish, and Italian samples – all of them were mutually equivalent. However, we could create two hierarchical scales, based on differences in random effects from the multilevel models (Table 7): for the Self-uncompassionate responding subscale, the highest position belonged to the German sample, followed by Korean, Portugal, and Australian samples. Lower levels of Self-uncompassionate responding were displayed in Czech, USA, and Slovak samples. Very low levels of Self-uncompassionate responding were found in Chinese, and Italian samples, and the lowest level was in the Turkish sample. As regards to the Self-compassionate responding subscale, the highest level was reported in the Turkish sample followed by Slovak, Italian, and Czech samples. At the opposite end, we found Portuguese, Australian, and Chinese samples to have lower levels of Self-compassionate responding. We did not attempt to provide any systematic interpretation of these differences –

far more detailed research is required to do so. However, we could see that no discernible pattern emerged from mutually equivalent samples – there was no cultural, linguistic, or geographical continuum able to explain clusters of mutually equivalent countries.

Table 7. Random effects for countries from the multilevel IRT model

Self-compassionate responding	Self-uncompassionate responding
Turkey (0.20)	Germany (0.01)
Slovakia (-0.06)	South Korea (-0.47)
Italy (-0.15)	Portugal (-0.68)
Czech Republic (-0.23)	Australia (-0.96)
USA (-0.25)	Czech Republic (-1.07)
Germany (-0.26)	USA (-1.32)
South Korea (-0.78)	Slovakia (-1.34)
Portugal (-0.94)	Italy (-1.62)
Australia (-1.10)	China (-1.66)
China (-2.14)	Turkey (-1.98)

In the Appendix, we present all test score functions for all comparisons. It was clear after inspection that even very large differences at particular levels of θ might have a negligible effect on differences in expected total scores if they were compensated after the intersection of test score functions. If test score functions did not intersect, the unsigned DTF equals to the signed DTF: it means that the reference group scores were systematically lower (or higher) than the focal group across all the range of latent ability. In Appendix, we present more detailed information for each comparison: (1) absolute values of sDTF and uDTF, (2) their standardized (percentage) values, (3) 95% confidence intervals for (1) and (2) values, (4) the significance test of the sDTF (based on 1000 simulation samples), and (5) visualization of test score functions. This database was useful for subsequent interpretation and could serve as a starting point for further research. It was beyond the scope of this article to speculate either about the lack of equivalence between countries, or about the differences in Self-uncompassionate responding and Self-compassionate responding subscales when equivalence was obtained.

An interesting finding is that some countries showed more invariance with other countries than others. For example, the Turkish sample was equivalent to no other sample for the Self-compassion responding subscale. In contrast, the Czech sample was equivalent to six other samples for the Self-uncompassionate responding subscale. Hypothetically, there are some clusters of similar countries regarding Self-compassionate responding and Self-uncompassionate responding subscales, and there are some countries which seem to view self-compassion differently. As no perfect transitivity was present in our samples, we cannot create a single linear rank based on the differences in latent means of equivalent samples nor compare countries on levels of Self-uncompassionate responding and Self-compassionate responding subscales.

Although it was impossible to compare total score across all countries, and the differences in latent means are available only for samples displaying the invariance (Tables 8 and 9), we can check random effects (relative differences across countries) from the multilevel IRT model. We can see (Table 7) that for the Self-uncompassionate responding subscale, the highest scores were found in the German sample, followed by Korean, Portuguese, and Australian samples. Lower levels of the Self-uncompassionate responding are displayed in Czech, USA, and Slovak samples. Very low levels of the Self-uncompassionate responding are in Chinese and Italian samples, and the lowest level is in the Turkish sample. For the Self-compassionate responding subscale, the highest level is in the Turkish sample, followed by the Slovak, Italian, and Czech samples, and the lower levels we can find in Portuguese, Australian, and Chinese samples.

Table 8. Latent means differences of invariant samples for the Self-uncompassionate responding subscale

Mean	AUS	CZK	GER	CHI	ITA	KOR	SVK
CZK	0.188	-	-	-	-	-	-
CHI	-	0.461	1.309	-	-	-	-
ITA	-	0.428	-	0.042	-	-	-
KOR	-	-	0.326	-	-	-	-
SVK	-	0.116	-	-	-	-	-
TUR	-	0.584	-	0.589	0.167	1.032	-
USA	-	0.158	-	-	-	-	0.018

Note. Latent means estimations of populations in first row were constrained to zero. All items were reverse-scored, so the positive values express lower amount of self-criticism. Differences significant at 0.05 level are highlighted in bold. AUS = Australia (n = 517). CHI = China (n = 321). CZK = Czech Republic (n = 5081). GER = Germany (n = 2510). ITA = Italy (n = 384). POR = Portugal (n = 512). SVK = Slovakia (n = 1181). KOR = South Korea (n = 1813). TUR = Turkey (n = 471). USA (n = 833).

Table 9. Latent means differences of invariant samples for the Self-compassionate responding subscale

Mean	AUS	CZK	GER	ITA	KOR	SVK
CHI	-	-1.076	-1.139	-	-	-
ITA	-	-0.111	0.088	-	-	-
KOR	0.087	-	-	-	-	-
POR	-0.079	-	-0.456	-	-0.194	-
SVK	-	-0.008	-	0.085	-	-
USA	-	0.002	-	0.086	-	-0.024

Note. Latent means estimations of populations in first row were constrained to zero. All items were reverse scored, so the positive values express lower amount of self-criticism. Differences significant at 0.05 level are highlighted in bold. AUS = Australia (n = 517). CHI = China (n = 321). CZK = Czech Republic (n = 5081). GER = Germany (n = 2510). ITA = Italy (n = 384). POR = Portugal (n = 512). SVK = Slovakia (n = 1181). KOR = South Korea (n = 1813). TUR = Turkey (n = 471). USA (n = 833).

To conclude, there was a remarkable amount of measurement equivalence between different countries which demonstrates that Self-uncompassionate responding and Self-compassionate responding subscales are valid and reliable instruments with substantial cross-cultural potential. However, many comparisons resulted in a lack of measurement equivalence and therefore displayed differential test functioning. This lack of measurement equivalence may be the result of linguistic variation, real differences in levels self-compassion across countries, or by peculiarities in sampling procedures (our samples were far from being representative for respective populations – except for the German sample).

4. Discussion

The goal of this study was to examine measurement invariance by IRT differential test functioning across ten different samples and nine language versions of the two factors of the SCS (Self-compassionate responding subscale and Self-uncompassionate responding subscale). Specifically, we were interested in determining whether comparisons between overall scores of the two SCS subscales across countries and languages were justified.

The main strength of this study was the amount of countries analysed and thoroughness of the data design and study, which allowed us to assess the cross-cultural extent and implications of the evaluated constructs of Self-compassionate responding and Self-uncompassionate responding. We used a large sample size of at least 300 participants in every sample, recruited from ten different countries with diverse language and cultural backgrounds.

Considering the diversity of cultures, we were not surprised to find that only one third of the comparisons between two countries proved to be invariant. For the Self-uncompassionate responding subscale, only 13 of the 45 comparisons (analysis of every pair) demonstrated measurement invariance (no differential test functioning). Generally, the results revealed that the observed constructs Self-compassionate responding and Self-uncompassionate responding from the SCS were not equivalent among all countries and groups. Therefore, it is impossible to compare their total scores across all countries. Considering the possible culturally and linguistically different expressions of self-compassion, future testing around the meanings and connotations of the constructs across the world is necessary. In addition, it would be beneficial to look carefully at the items of SCS and test whether they are measuring what they should measure. This may aid in the development of a new scale allowing for cross-cultural comparisons of various treatment outcomes related to self-compassion.

The primary advantage of this study is that we could assess and compare, without bias, real differences in raw or latent scores between countries which do not display differential test

functioning. Therefore, the differences in raw and latent scores are not distorted by different functioning of tests in specific countries and languages.

Our research findings are consistent with Heine et al. (1999), who suggested that simultaneously higher levels of Self-uncompassionate responding may not necessarily be associated with lower levels of Self-compassionate responding. For example, in the current study, the Czech sample showed high levels of both Self-uncompassionate responding and Self-compassionate responding, and the Chinese sample showed low levels in both. Moreover, the German sample showed very high levels of Self-uncompassionate responding, but it did not show particularly low levels of Self-compassionate responding. Nevertheless, the Slovak, Italian, and Turkish samples showed the typical pattern (Neff, 2003) of being low in Self-uncompassionate responding and high in Self-compassionate responding. These results provide additional support for not using total SCS score but using scores separately for Self-uncompassionate responding and Self-compassionate responding subscales.

In addition, based on the results of this study we advise against the use of the unweighted total scores to compare the relative group responses, while ignoring differential weights of group membership: e.g., after performing independent t-test or Mann-Whitney non-parametric test on unweighted total scores, one could easily come to the conclusion that the two populations had essentially equal mean test scores. But DTF analyses could discover that one population had a lower latent mean compared with another population and at the same time the first population could be scored more favourably on the test obtaining positively biased test scores. These two situations jointly explain the observed equivalence in the total scores, but they have very different theoretical interpretations and completely different consequences for practical purposes.

Our results show that the differences in the levels of Self-compassionate responding and Self-uncompassionate responding subscales cannot be attributed to a simple cultural distinction between Western and Eastern cultures. Of the two Asian samples, the Chinese sample had low Self-compassionate and Self-uncompassionate responding, but the Korean sample had low Self-compassionate responding, but high Self-uncompassionate responding. Such results support that specific cultural values rather than West-East differences are crucial in understanding cross-cultural differences of overall self-compassion (Neff et al., 2008), and further indicate that these cultural values may work differently with the two factors of the Self-Compassion Scale.

The main limitation of our study was that values for the considered variables were self-reported, and therefore they may have been influenced by socially desirable responses. On the other hand, the sample was recruited mainly online but also in paper-pencil form. Despite studies that confirm the reliability of the data obtained from online source (Ritter et al., 2004), these samples

might be more biased than those obtained using traditional methods. Also results could be influenced by different forms of obtaining data.

As self-compassion is a construct of high clinical significance, improving understanding of cross-cultural similarities and differences in Self-uncompassionate responding and Self-compassionate responding, as measured by the two subscales of the SCS, would have great impact on practice. This is so, because negative relation to oneself in the form of excessive self-uncompassionate (or even self-critical) inner speech is one of the most important psychological processes that influence susceptibility to psychopathology, and also its persistence (Falconer et al., 2015). On the other hand, self-compassion is generally associated with better psychological health (Gilbert et al., 2004; Neff, 2003). In an applied context, understanding the differences of Self-uncompassionate responding and Self-compassionate responding subscales across countries can help to inform more effective practices.

5. Conclusion

The results of this study substantially contribute to the growing body of knowledge about similarities and differences among cultures with respect to the two factors of the Self-Compassion Scale, Self-uncompassionate responding and Self-compassionate responding. While suggesting that Self-uncompassionate responding and Self-compassionate responding may tentatively be reflected as general, universal constructs, the results of this research reveal significant cross-cultural differences in the way these constructs are measured by the subscales of the SCS.

Given that the expression of self-compassion may vary across different cultures and languages including Self-uncompassionate responding and Self-compassionate responding, future international testing of the meanings and connotations of the constructs is necessary. Generally, the results revealed that the observed constructs Self-uncompassionate responding and Self-compassionate responding from SCS were not equivalent among all countries and groups. However, Self-uncompassionate responding and Self-compassionate responding subscales are valid and reliable instruments with substantial potential of use cross-culturally which needs to be further explored to allow reliable comparisons of subscales scores across different countries.

Funding

Writing this work was supported by the Vedecká grantová agentúra VEGA under Grant 1/0578/15 and 1/0075/19.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any potential conflict of interest.

References

1. Arimitsu, K. (2014). Development and validation of the Japanese version of the Self-Compassion Scale. *The Japanese Journal of Psychology*, 85(1), 50–59. <https://doi.org/10.4992/jjpsy.85.50>
2. Azizi, A., Mohammadkhani, P., Lotfi, S., & Bahramkhani, M. (2013). The validity and reliability of the Iranian version of the Self-Compassion Scale. *Iranian Journal of Clinical Psychology*, 2(3), 17-23.
3. Barnett, M. D., & Sharp, K. J. (2016). Maladaptive perfectionism, body image satisfaction, and disordered eating behaviors among US college women: The mediating role of self-compassion. *Personality and Individual Differences*, 99, 225-234. <https://doi.org/10.1016/j.paid.2016.05.004>
4. Behling, O., & Law, K. S. (2000). Translating questionnaires and other research instruments: Problems and solutions. Thousand Oaks: Sage Publications. <https://doi.org/10.4135/9781412986373>
5. Benda, J., & Reichová, A. (2016). Psychometrické charakteristiky české verze Self-Compassion Scale (SCS-CZ). *Československá psychologie*, 60(2), 120-136.
6. Birkett, M. A. (2014). Self-compassion and empathy across cultures: Comparison of young adults in China and the United States. *International Journal of Research Studies in Psychology*, 3(3), 25-34. <https://doi.org/10.5861/ijrsp.2013.551>
7. Brown, L., Bryant, C., Brown, V. M., Bei, B., & Judd, F. K. (2014). Self-Compassion weakens the association between hot flushes and night sweats and daily life functioning and depression. *Maturitas*. 78(4), 298-303. <https://doi.org/10.1016/j.maturitas.2014.05.012>
8. Castilho, P., & Pinto-Gouveia, J. (2011). Self-Compassion: validation of the Portuguese version of the Self-Compassion Scale and its relation with early negative experiences, social comparison and psychopathology. *Psychologica*, 54, 203-231. https://doi.org/10.14195/1647-8606_54_8
9. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/jss.v048.i06>
10. Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement*, 52, 200-222. <https://doi.org/10.1111/jedm.12072>
11. Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 76(1), 114–140. <https://doi.org/10.1177/0013164415584576>
12. Chen, J., Yan, L., & Zhou, L. (2011). Reliability and validity of Chinese version of Self-compassion Scale. *Chinese Journal of Clinical Psychology*, 19(6), 734-736.
13. Conversano, C., Giacchini, R., Orrù, G., Di Giuseppe, M., Gemignani, A., & Poli, A. (2020). Mindfulness, Compassion, and Self-Compassion Among Health Care Professionals: What's New? A Systematic Review. *Frontiers in psychology*, 11, 1683. <https://doi.org/10.3389/fpsyg.2020.01683>
14. Costa, J., Marôco, J., Pinto-Gouveia, J., Ferreira, C., & Castilho, P. (2016). Validation of the psychometric properties of the Self-Compassion Scale. Testing the factorial validity and factorial invariance of the measure among borderline personality disorder, anxiety disorder, eating disorder and general populations. *Clinical Psychology & Psychotherapy*. 23(5), 460-468. <https://doi.org/10.1002/cpp.1974>

15. Cucina, J. & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence*, 5, 27.
<https://doi.org/10.3390/jintelligence5030027>
16. Cunha, M., Xavier, A., & Castilho, P. (2016). Understanding self-compassion in adolescents: Validation study of the Self-Compassion Scale. *Personality and Individual Differences*, 93(9), 56–62.
<https://doi.org/10.1016/j.paid.2015.09.023>
17. DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, 24, 189-209. <https://doi.org/10.1080/08957347.2011.580255>
18. DeMars, C. E. (2012). A comparison of limited-information and full-information methods in Mplus for estimating Item Response Theory parameters for non-normal populations. *Structural Equation Modeling: A Multidisciplinary Journal*, 19, 610-632. <https://doi.org/10.1080/10705511.2012.713272>
19. Deniz, M. E., Kesici, S., & Sumer, A. S. (2008): The validity and reliability of the Turkish version of the Self-Compassion Scale. *Social Behavior and Personality*, 36(9), 1151-1160.
<https://doi.org/10.2224/sbp.2008.36.9.1151>
20. Dundas, I., Svendsen, J. L., Wiker, A. S., Granli, K. V., & Schanche, E. (2016). Self-compassion and depressive symptoms in a Norwegian student sample. *Nordic Psychology*, 68(1), 58-72.
<https://doi.org/10.1080/19012276.2015.1071203>
21. Falconer, C.J., King, J.A., & Brewin, C.R. (2015). Demonstrating mood repair with a situation-based measure of self-compassion and self-criticism. *Psychology and psychotherapy*. 88(4), 351-65.
<https://doi.org/10.1111/papt.12056>
22. Garcia-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>
23. Gilbert, P., Clark, M., Hempel, S., Miles, J.N.V. & Irons, C. (2004). Criticising and reassuring oneself: An exploration of forms, styles and reasons in female students. *British Journal of Clinical Psychology*, 43, 31-50.
<https://doi.org/10.1348/014466504772812959>
24. Gilbert, P., & Irons, C. (2008). Shame, self-criticism and self-compassion in adolescence. In NB Allen & LB Sheeber (Eds.), *Adolescent Emotional Development and the Emergence of Depressive Disorders*. (pp. 195-214). Cambridge University Press. <https://doi.org/10.1017/CBO9780511551963.011>
25. Greenberg, L. S. & Watson, J. (2005). *Emotion-focused therapy of Depression*. Washington, D. C.: American Psychological Association Press.
26. Halamová, J., Kanovský, M., & Pacúchová, M. (2017). Self-Compassion Scale: IRT psychometric analysis validation and factor structure – Slovak translation. *Psychologica Belgica*. 57(4), 190–209.
<https://doi.org/10.5334/pb.398>
27. Halamová, J., Kanovský, M., Petrocchi, N., Moreira, H., López Angarita, A. et al. (2021). The factor structure of the Self-Compassion Scale in eleven distinct populations across the world. *Measurement and Evaluation in Counseling and Development*. 54(1), 1-23, <https://doi.org/10.1080/07481756.2020.1735203>

28. Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review*, 106, 766- 794. <https://doi.org/10.1037/0033-295X.106.4.766>
29. Heshmati, R., & Pellerone, M. (2018). Relationship of dispositional mindfulness, self-destruction and self-criticism in academic performance of boy students. *Mediterranean Journal of Clinical Psychology*, 6(3). <https://doi.org/10.6092/2282-1619/2018.6.1716>
30. Hupfield, J. & Ruffieux, N. (2011). Validierung einer deutschen version der Self-Compassion Scale (SCS-D). *Zeitschrift für Klinische Psychologie und Psychotherapie*, 40 (2), 115–123. <https://doi.org/10.1026/1616-3443/a000088>
31. Hwang, S., Kim, G., Yang, J. W., & Yang, E. (2016). The moderating effects of age on the relationships of self-compassion, self-esteem, and mental health. *Japanese Psychological Research*. 58(2), 194-205. <https://doi.org/10.1111/jpr.12109>
32. Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: a comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, 40(2), 279 – 310. <https://doi.org/10.1177/0049124111405301>
33. Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(2), 212-228. <https://doi.org/10.1080/10705511.2011.557337>
34. Kitayama, S., & Karasawa, M. (1997). Implicit self-esteem in Japan: Name letters and birthday numbers. *Personality and Social Psychology Bulletin*, 23, 736-742. <https://doi.org/10.1177/0146167297237006>
35. Kotsou, I., & Leys, C. (2016). Self-Compassion Scale (SCS): Psychometric properties of The French translation and its relations with psychological well-being, affect and depression. *PLoS one*, 11(4). <https://doi.org/10.1371/journal.pone.0152880>
36. Körner, A., Coroiu, A., Copeland, L., Gomez-Garibello, C., Albani, C., Zenger, M., & Brähler, E. (2015). The role of self-compassion in buffering symptoms of depression in the general population. *PLoS ONE*, 10(10). <https://doi.org/10.1371/journal.pone.0136598>
37. Khrantsova I.I., & Chuykova T.S. (2016). Mindfulness and self-compassion as predictors of humor styles in US and Russia. *Sotsial'naiia psikhologiiia i obshchestvo = Social Psychology and Society*, 7(2), 93—108. <https://doi.org/10.17759/sps.2016070207>
38. Kwan, V. S. Y., Kuang, L. L., & Hui, N. H. H. (2009). Identifying the sources of self-esteem: The mixed medley of benevolence, merit, and bias. *Self and Identity*, 8, 176-195. <https://doi.org/10.1080/15298860802504874>
39. Lee, W. K., & Lee, K. (2010). The validation study of the Korean version of Self-Compassion Scale with adult women in community [Korean]. *Journal of Korean Neuropsychiatric Association*, 49(2), 193-200.
40. López, A., Sanderman, R., Smink, A., Zhang, Y., van Sonderen, E., Ranchor, A., & Schroevers, M. J. (2015). A reconsideration of the Self-Compassion Scale's total score: self-compassion versus self-criticism. *PLoS One*, 10(7). <https://doi.org/10.1371/journal.pone.0132940>
41. Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence*, 61, 120-129. <https://doi.org/10.1016/j.intell.2017.01.012>

42. Mantzios, M., Wilson, J. C., & Giannou, K. (2013). Psychometric properties of the Greek versions of the self-compassion and mindful attention and awareness scales. *Mindfulness*, 6(1), 123-132.
<https://doi.org/10.1007/s12671-013-0237-3>
43. Meade, A. W. & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance, *Organizational Research Methods*, 7(4), 361-388. <https://doi.org/10.1177/1094428104268027>
44. Meares, R., Gerull, F., Stevenson, J., & Korner, A. (2011). Is self-disturbance the core of borderline personality disorder? An outcome study of borderline personality factors. *Australian and New Zealand Journal of Psychiatry*, 45, 214–222. <https://doi.org/10.3109/00048674.2010.551280>
45. Merlo, E. M., Stoian, A. P., Motofei, I. G., & Settineri, S. (2020). Clinical Psychological Figures in Healthcare Professionals: Resilience and Maladjustment as the “Cost of Care”. *Frontiers in Psychology*, 11.
<https://doi.org/10.3389/fpsyg.2020.607783>
46. Montero-Marín, J., Gaete, J., Demarzo, M., Rodero, B., Lopez, L.C., & García-Campayo, J. (2016). Self-Criticism: A measure of uncompassionate behaviors toward the self, based on the negative components of the Self-Compassion Scale. *Frontiers in Psychology*. 30(7), 1281. <https://doi.org/10.3389/fpsyg.2016.01281>
47. Moreira, H., Carona, C., Silva, N., Nunes, J., & Canavarro, M. C. (2016). Exploring the link between maternal attachment-related anxiety and avoidance and mindful parenting: The mediating role of self-compassion. *Psychology and Psychotherapy: Theory, Research and Practice*, 89(4), 369-384. <https://doi.org/10.1111/papt.12082>
48. Moreira, H., Gouveia, M. J., Carona, C., Silva, N., & Canavarro, M. C. (2015). Maternal attachment and children's quality of life: The mediating role of self-compassion and parenting stress. *Journal of Child and Family Studies*, 24(8), 2332-2344. <https://doi.org/10.1007/s10826-014-0036-z>
49. Murray, G., Judd, F., Jackson, H., Fraser, C., Komiti, A., Hodgins, G., Pattison, P., Humphreys, J., & Robins, G. (2004). Rurality and mental health: the role of accessibility. *Australian and New Zealand Journal of Psychiatry*. 38(8), 629-634. <https://doi.org/10.1111/j.1440-1614.2004.01426.x>
50. 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/10.1080/15298860390209035>
51. Neff, K. D. (2004). Self-compassion and psychological well-being. *Constructivism in the Human Sciences*, 9, 27-37. <https://doi.org/10.1037/e633942013-240>
52. Neff, K. D. (2016). The Self-Compassion Scale is a valid and theoretically coherent measure of self-compassion. *Mindfulness*, 7(1), 264-274. <https://doi.org/10.1007/s12671-015-0479-3>
53. Neff, K. D., & Dahm, K. A. (2014). Self-Compassion: What it is, what it does, and how it relates to mindfulness (pp. 121-140). In M. Robinson, B. Meier & B. Ostafin (Eds.) *Mindfulness and Self-Regulation*. New York: Springer. https://doi.org/10.1007/978-1-4939-2263-5_10
54. Neff, K. D., Pisitsungkagarn, K., & Hsieh, Y. (2008). Self-compassion and self-construal in the United States, Thailand, and Taiwan. *Journal of Cross-Cultural Psychology*, 39, 267-285.
<https://doi.org/10.1177/0022022108314544>
55. Neff, K. D. & Vonk, R. (2009). Self-compassion versus global self-esteem: Two different ways of relating to oneself. *Journal of Personality*, 77(1), 23-50. <https://doi.org/10.1111/j.1467-6494.2008.00537.x>

56. Neff, K. D., Whittakar, T. & Karl, A. (2017). Evaluating the factor structure of the Self-Compassion Scale in four distinct populations: Is the use of a total self-compassion score justified? *Journal of Personality Assessment*, 99, 596–607. <https://doi.org/1080/00223891.2016.1269334>
57. O'Connor, R. C. & Noyce, R. (2008). Personality and cognitive processes: Self-criticism and different types of rumination as predictors of suicidal ideation. *Behaviour Research and Therapy*, 46(3), 392–401. <https://doi.org/10.1016/j.brat.2008.01.007>
58. Petrocchi, N., Ottaviani, C., & Couyoumdjian, A. (2014). Dimensionality of self-compassion: translation and construct validation of the self-compassion scale in an Italian sample. *Journal of Mental Health*, 23(2), 72-77. <https://doi.org/10.3109/09638237.2013.841869>
59. Pisitsungkagarn, K., Taephant, N., & Attasaranya, P. (2014). Body image satisfaction and self-esteem in Thai female adolescents: the moderating role of self-compassion. *International journal of adolescent medicine and health*, 26(3), 333-338. <https://doi.org/10.1515/ijamh-2013-0307>.
60. Raju, N. S., Laffitte, L. J. & Byrne, B. B. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517–529. <https://doi.org/10.1037//0021-9010.87.3.517>
61. Reise, S. P., Widaman, K. H. & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance, *Psychological Bulletin*, 114(3), 552-566. <https://doi.org/10.1037/0033-2909.114.3.552>
62. Ritter, P., Lorig, K., Laurent, D., & Matthews, K. (2004). Internet versus mailed questionnaires: a randomized comparison. *Journal of Medical Internet Research*, 6(3), e29. <https://doi.org/10.2196/jmir.6.3.e29>
63. Shahar, B., Carlin, E.R., Engle, D.E., Hegde, J., Szepeswol, O., & Arkowitz, H. (2012). A pilot investigation of emotion-focused two-chair dialogue intervention for self-criticism. *Clinical Psychology and Psychotherapy*, 6, 496–507. <https://doi.org/10.1002/cpp.762>
64. Shahar, B., Szepeswol, O., Zilcha-Mano, S., Haim, N., Zamir, O., Levi-Yeshuvi, S., & Levit-Binnun, N. (2015). A wait-list randomized controlled trial of loving-kindness meditation programme for self-criticism. *Clinical psychology & psychotherapy*, 22(4), 346-356. <https://doi.org/10.1002/cpp.1893>
65. Şenyuva, E., Kaya, H., Işık, B., & Bodur, G. (2014). Relationship between self-compassion and emotional intelligence in nursing students. *International journal of nursing practice*, 20(6), 588-596. <https://doi.org/10.1111/ijn.12204>
66. Terry, M. L., & Leary, M. R. (2011). Self-compassion, self-regulation, and health. *Self and Identity*, 10(3), 352–362. <https://doi.org/10.1080/15298868.2011.558404>
67. Tóth-Király, I., Bőthe, B., & Orosz, G. (2016). Exploratory structural equation modeling analysis of the Self-Compassion Scale. *Mindfulness*, 7, 1–12. <https://doi.org/10.1007/s12671-016-0662-1>
68. Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70. <https://doi.org/10.1177/109442810031002>
69. Voruz, S. (2013). An examination of the psychometric properties of the Self-Compassion Scale – Short Form among a help-seeking clinical sample (Doctoral dissertation, Pacific University).

Retrieved from: <http://commons.pacificu.edu/spp/430>

70. Warren, R., Smeets, E. & Neff, K. D. (2016). Self-criticism and self-compassion: Risk and resilience for psychopathology. *Current Psychiatry*, 15(12), 18-32.
71. Zeng, X., Wei, J., Oei, T. P., & Liu, X. (2016). The self-compassion scale is not validated in a Buddhist sample. *Journal of religion and health*, 55(6), 1996-2009. <https://doi.org/10.1007/s10943-016-0205-z>



©2021 by the Author(s); licensee Mediterranean Journal of Clinical Psychology, Messina, Italy. This article is an open access article, licensed under a Creative Commons Attribution 4.0 Unported License. Mediterranean Journal of Clinical Psychology, Vol. 9, No. 1 (2021).

International License (<https://creativecommons.org/licenses/by/4.0/>).

DOI: 10.6092/2282-1619/mjcp-2682