Journal of Rational-Emotive & Cognitive-Behavior Therapy (2019) 37:411–444
https://doi.org/10.1007/s10942-019-00319-1

Multiple Group IRT Measurement Invariance Analysis of the Forms of Self-Criticising/Attacking and Self-Reassuring Scale in Thirteen International Samples

Júlia Halamová[1] · Martin Kanovský[2] · Paul Gilbert[3] · Nicholas A. Troop[4] · David C. Zuroff[5] · Nicola Petrocchi[6] · Nicola Hermanto[5] · Tobias Krieger[7] · James N. Kirby[8] · Kenichi Asano[9] · Marcela Matos[10] · FuYa Yu[11] · Marion Sommers-Spijkerman[12] · Ben Shahar[13] · Jaskaran Basran[3] · Nuriye Kupeli[14]

Published online: 30 April 2019

Abstract

The purpose of this study was to examine the measurement invariance of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale (FSCRS) in terms of Item Response Theory differential test functioning in thirteen distinct samples ($N = 7714$) from twelve different countries. We assessed differential test functioning for the three FSCRS subscales, Inadequate-Self, Hated-Self and Reassured-Self separately. 32 of the 78 pairwise comparisons between samples for Inadequate-Self, 42 of the 78 pairwise comparisons for Reassured-Self and 54 of the 78 pairwise comparisons for Hated-Self demonstrated no differential test functioning, i.e. measurement invariance. Hated-Self was the most invariant of the three subscales, suggesting that self-hatred is similarly perceived across different cultures. Nonetheless, all three subscales of FSCRS are sensitive to cross-cultural differences. Considering the possible cultural and linguistic differences in the expression of self-criticism and self-reassurance, future analyses of the meanings and connotations of these constructs across the world are necessary in order to develop or tailor a scale which allows cross-cultural comparisons of various treatment outcomes related to self-criticism.

Keywords Self-criticism · Self-reassurance · Measurement invariance · Differential test functioning · Cross-cultural studies

✉ Júlia Halamová
julia.halamova@gmail.com

✉ Nuriye Kupeli
n.kupeli@ucl.ac.uk

Extended author information available on the last page of the article

Introduction

Excessive self-criticism is a personality vulnerability factor that can cause and sustain various psychological difficulties and disorders (e.g. Blatt and Shichman 1983; Blatt 2004; Falconer et al. 2015; Shahar et al. 2012). Blatt recognized the clinical significance of self-criticism and his work has influenced current understandings of the forms and impacts of self-criticism on mental health (Blatt et al. 1979). Self-criticism is generally viewed as a relatively stable and intractable personality style (Hermanto et al. 2016; Zuroff et al. 2004). Zuroff et al. (2016) demonstrated that self-criticism displays both trait-like stability over time and a degree of variability over time reflecting state influences. In addition, self-criticism is responsible for poor response to psychological treatment (Blatt et al. 1995; Blatt and Zuroff 2005; Horvath and Symonds 1991; Stinckens et al. 2013a, b). Several authors suggest there is a need for closer examination of self-criticism across cultures (Lau et al. 2010; Luyten and Blatt 2013). Developing a thorough understanding of self-criticism and designing sensitive tools to measure it will provide methods to evaluate tailored interventions across cultures.

Self-Criticism and Culture

Although the concept of mutual definition, influence and constitution between culture and self is old and pervasive (Kitayama 2016), research on cross-cultural comparisons of self-criticism is scarce. The majority of cross-cultural research operates within the self-construal theory of Markus and Kitayama (1991). A comparison of Western and Eastern conceptualizations of the self revealed that Western cultures tend to have independent self-construal because they construe the self as separate from its social context, emphasizing autonomy and independence. Importantly, it has been suggested that individuals in Western cultures focus on their abilities, traits, and needs, and they tend to prioritise their individual goals over those of in-groups. In contrast, individuals in Eastern cultures tend to have interdependent self-construal as they usually construe the self as an integral part of a broader social context and their concept of the self involves characteristics of their social environment. They are also suggested to have a sense of connectedness with others and to focus on their role in in-groups, while prioritising group goals over individual goals (Markus and Kitayama 1991).

According to Heine and Hamamura (2007), independent cultures might facilitate more self-enhancement by promoting a focus on inner attributes, while among interdependent cultures, reflection on the same inner attributes may foster self-criticism. A meta-analysis supports this view, with a large cross-cultural effect ($d = .84$) between East Asians and Westerners (Heine and Hamamura 2007). This cross-cultural difference was most prominent between the USA and Japan, with North Americans presenting as more self-enhancing whilst the Japanese as more self-critical (Heine et al. 2000).

These findings are supplemented by research on topics explaining specific conditions in which the self-construal paradigm works. According to Kitayama et al. (1997), self-enhancement is defined as a general sensitivity to positive self-relevant information and self-criticism as a general sensitivity to negative self-relevant information. However, this definition of self-criticism is broad and vague. In comparison, a more precise definition of self-criticism has been offered by Blatt and Zuroff (1992) who characterized it as constant and harsh self-scrutiny and evaluation and feelings of unworthiness, inferiority, failure, and guilt.

The majority of cross-cultural research uses the definition of self-criticism provided from Kitayama et al. (1997), while those studies examining psychopathology generally use the definition offered by Blatt and Zuroff (1992). For example, Yamaguchi et al. (2014), found that in a sample of American students, independent self-perception was related to self-criticism but, in a sample of Japanese students, only interdependent self-perception was associated with high levels of self-criticism. Based on these findings, the authors argue that dominance of cultural self-perception is associated with self-criticism (Yamaguchi et al. 2014).

Another example of research on psychopathology is from Hermanto et al. (2016), who investigated the moderating effect of fear of receiving compassion on the association between self-criticism and depression. This large international study included a large multi-cultural city in Canada and midsized cities in Canada, England, and Portugal. There was a positive association between self-criticism and depression but the effect was more prominent for individuals who reported high rather than low levels of fear of receiving compassion from others.

Nonetheless, these findings must be considered with caution since measurement invariance of the tools used to assess self-criticism has generally not been tested. Measurement invariance means that a construct measures same property in different groups, and it is a prerequisite for identifying meaningful cultural differences since it is an indication of the degree to which participants from different cultures interpret constructs in the same way. Lack of measurement invariance means that a test is biased: respondents with some level of a latent trait from one group provide systematically lower or higher responses than respondents with the same level of latent trait from another group, and this bias is induced by the test and does not express real differences.

Several different tools measuring self-criticism have been developed including the Depressive Experiences Questionnaire which assesses self-criticism, dependency, and self-efficacy (DEQ; Blatt et al. 1979), the Levels of Self-Criticism Scale (LOSC; Thompson and Zuroff 2004), the Forms of Self-criticising/Attacking & Self-Reassuring Scale (FSCRS; Gilbert et al. 2004), The Self-Critical Rumination Scale (Smart et al. 2016), and a situational measure labelled as The Self-Compassion and Self-Criticism Scales (SCCS; Falconer et al. 2015). Among the listed scales, to date only one study has reported measurement invariance in the LOSC (Thompson and Zuroff 2004) between Japanese and USA students (Yamaguchi et al. 2014). Although the psychometric features of the FSCRS have been thoroughly explored in different languages demonstrating good validity and reliability as well as consistent factor structure (Halamová et al. 2018), to our knowledge, no study to date has tested the cross-cultural measurement invariance of the FSCRS.

Aim of the Current Study

The present study investigates the measurement invariance of the dimensions of the FSCRS using Item Response Theory (IRT) differential test functioning using 13 samples from 12 different countries and eight language versions. The main objective of this study is to determine whether comparisons between total scores of the three dimensions of the FSCRS across countries and languages are appropriate and whether these findings about measurement invariance allow further cross-cultural research using the FSCRS.

Methods

Measure

The Forms of Self-criticising/Attacking & Self-Reassuring Scale (FSCRS; Gilbert et al. 2004) is a 22-item instrument, which was developed to assess levels of self-criticism and the ability to self-reassure when one faces setbacks and failure. Participants use a 5-point Likert scale to rate the extent to which various statements are true about them (1 = not at all like me; 5 = extremely like me). The scale comprises three subscales: Inadequate Self, which focuses on feelings of personal inadequacy, Hated Self measuring the desire to hurt or punish oneself, and Reassured Self which is an ability to reassure and support the self. Items for the three subscales are given in Table 1.

Originally developed in English in the UK, the FSCRS has been translated into different languages including Chinese (Yu, personal communication), Dutch (Sommers-Spijkerman et al. 2018), French (Gheysen et al. 2015), German (Wiencke, personal communication), Hebrew (Shahar et al. 2015), Italian (Petrocchi and Couyoumdjian 2016), Japanese (Kenichi, personal communication), Portuguese (Castilho et al. 2015), Slovak (Halamová et al. 2017) and Swedish (Lekberg and Wester 2012). Previous studies revealed that the FSCRS has high internal consistency (Baião et al. 2015; Gilbert et al. 2004; Halamová et al. 2017; Kupeli et al. 2013) and good test–retest reliability (Castilho et al. 2015), even when translated into different languages.

The construct validity of the FSCRS is evident when it is correlated with a one-dimensional self-criticism measure like the DEQ (Blatt et al. 1979) and multidimensional measure like the LOSC (Thompson and Zuroff 2004). Correlations are in line with theoretical expectations, which indicate that all subscales of the FSCRS have good validity (Castilho et al. 2015; Gilbert et al. 2004; Halamová et al. 2017).

Some studies have demonstrated structural validity for the original three-factor solution of the FSCRS consisting of Hated self (HS), Inadequate self (IS) and Reassured self (RS) (Baião et al. 2015; Castilho et al. 2015; Kupeli et al. 2013). However, in more recent years research has favoured a two-factor solution consisting of self-criticism (IS + HS) and self-reassurance (RS), suggested merging the IS and HS subscales as a global measure of self-criticism in non-clinical populations (Gilbert

Table 1 Dimensions and scale items of The Forms of Self-Criticising/Attacking & Self-Reassuring Scale (Gilbert et al. 2004)

Dimensions	Scale items	
Self-criticism		
Inadequate self	1.	I am easily disappointed with myself.
	2.	There is a part of me that puts me down.
	4.	I find it difficult to control my anger and frustration at myself.
	6.	There is a part of me that feels I am not good enough.
	7.	I feel beaten down by my own self-critical thoughts.
	14.	I remember and dwell on my failings.
	17.	I can't accept failures and setbacks without feeling inadequate.
	18.	I think I deserve my self-criticism.
	20.	There is a part of me that wants to get rid of the bits I don't like.
Hated self	9.	I have become so angry with myself that I want to hurt or injure myself.
	10.	I have a sense of disgust with myself.
	12.	I stop caring about myself.
	15.	I call myself names.
	22.	I do not like being me.
Self-reassurance		
Reassured self	3.	I am able to remind myself of positive things about myself.
	5.	I find it easy to forgive myself.
	8.	I still like being me.
	11.	I can still feel lovable and acceptable.
	13.	I find it easy to like myself.
	16.	I am gentle and supportive with myself.
	19.	I am able to care and look after myself.
	21.	I encourage myself for the future.

et al. 2006a, b; Halamová et al. 2018; Halamová et al. 2017; Richter et al. 2009; Rockliff et al. 2011).

Sampling Procedure

To collate data from a variety of countries and cultures we used Google Scholar to identify publications which used the terms "the forms of self-criticising/attacking & self-reassuring scale" or "fscrs". We contacted the authors of all relevant publications which reported on samples of at least 220 non-clinical participants so as to enable the planned statistical methods. The planned statistical approach requires at least ten participants per item (Velicer and Fava 1998) and thus for the 22-item FSCRS, we required data from a sample of 220 participants. In addition, we found planned and not yet published research projects from the Compassionate Mind Foundation website (https://compassionatemind.co.uk/uploads/files/research-regis

ter-for-website.pdf). Approximately 40 emails with requests for data were sent, from which thirteen data sets were received and included in the current analyses.

Sample Characteristics and Procedures

Out of eleven existing language versions of FSCRS currently available, this study includes data from eight (Halamová et al. 2018). The complete data set consists of five distinct English language samples from four different countries including Australia ($N=319$), Canada ($N=380$), the United Kingdom (sample 1 $N=1570$ and sample 2 $N=883$) and USA ($N=331$). There were also samples from seven other language translations namely Chinese ($N=417$), Dutch ($N=363$), German ($N=230$), Hebrew ($N=475$), Italian ($N=393$), Japanese ($N=263$), Portuguese ($N=764$), and Slovak ($N=1326$). In total, we tested thirteen distinct samples with an overall sample size of 7714. Sample characteristics for each of the samples are reported in Table 2. The data collected from 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.

Australia Sample

The participants were Australians selected from a larger sample of general population participants from several provinces (Kirby, personal communication). Convenience sampling was used to recruit participants to an online survey.

Table 2 Sample characteristics and internal consistency coefficients for FSCRS for the thirteen samples

Country	N	Female (%)	M Age	SD	Language	Cronbach α
Australia	319	85.3	41.34	14.23	English	0.83–0.93
Canada	380	62.5	21.09	3.36	English	0.77–0.90
Israel	475	58.1	30.59	11.80	Hebrew	0.79–0.89
Italy	393	71.5	33.15	10.8	Italian	0.75–0.91
Japan	263	81.1	18.84	1.08	Japanese	0.80–0.88
Netherland	363	64.4	30.79	13.38	Dutch	0.80–0.89
Portugal	764	78.5	27.93	11.20	Portuguese	0.81–0.91
Slovakia	1326	68.2	29.61	12.06	Slovak	0.75–0.88
Switzerland	230	71	38.92	14.34	German	0.80–0.92
Taiwan	417	56.1	22.67	4.27	Chinese	0.85–0.90
UK 1	1570	82.5	28.47	10.65	English	0.86–0.93
UK 2	883	76.1	24.14	7.8	English	0.85–0.92
USA	331	73.1	20.77	5.25	English	0.85–0.92

M Mean, *SD* standard deviation

Canada Sample

Participants were students, who were recruited online through various university advertisements and the university pool of psychology research participants. Participants were required to be fluent in written English, and they received a small financial incentive or credit toward a course. The dataset comprised of data collected from various research studies (Hermanto and Zuroff 2016, 2017; Zuroff et al. 2016).

Netherlands Sample

A convenience sample of participants was recruited by various undergraduate students in an online cross-sectional survey conducted by a university in The Netherlands (Sommers-Spijkerman et al. 2018). The accuracy of the Dutch version of FSCRS was verified using back translation.

Israel Sample

The Israeli sample consisted of participants from the general population who were recruited via an online survey platform and by undergraduate students from a private college (Shahar et al. 2015; Shahar, personal communication). The Hebrew version of FSCRS was not back translated.

Italy Sample

This study (Petrocchi and Couyoumdjian 2016) was conducted through an online survey and participants were recruited via both an Italian university students mailing list, and other professional mailing lists and web advertising. The Italian version of FSCRS was back translated.

Japan Sample

The research sample from Japan consisted of students undertaking a degree in Psychology at University (Kenichi, personal communication). The Japanese version of FSCRS was not back translated.

Portugal Sample

The research sample from Portugal included participants recruited through convenience sampling using an online platform from a university setting and from the

general community (Gilbert et al. 2017). The Portuguese version of FSCRS was back translated.

Slovakia Sample

Data were collected gradually over 2 years within a research grant focused on self-criticism and self-compassion (Halamová et al. 2017). Data were obtained by convenience sampling; questionnaires were distributed on paper and in an online form via social networks. The Slovak version of FSCRS was back translated.

Switzerland Sample

Participants were recruited in the German-speaking part of Switzerland through a study website and postings on internet forums (Krieger et al. 2016; Krieger, personal communication). The German version of FSCRS was back translated (Wiencke, personal communication).

Taiwan Sample

Participants from Taiwan were recruited from universities through social media and through word of mouth between students; they completed either an online survey or a paper and pencil version (Yu 2013). The Chinese version of FSCRS was back translated.

United Kingdom Sample 1

Participants from the first UK sample were recruited online from a university and the general population through social networking sites and health and well-being forums (Kupeli et al. 2013).

United Kingdom Sample 2

The second UK sample was recruited from an undergraduate course at a university. Participants completed pen and paper questionnaires. The dataset included data collected from various research studies (Baião et al. 2015; Gilbert et al. 2002, 2004, 2005, 2006a, b, 2012; Gilbert and Miles 2000).

USA Sample

The USA sample were students attending university (Gilbert et al. 2017). Participants were recruited via online participant management software. Psychology students received credits for their participation in the research study.

Data Analysis

In testing measurement invariance/equivalence, linear confirmatory factor analysis (CFA) is the common approach (Vandenberg and Lance 2000) in which some parameters (factor loadings, intercepts, residual variances) are constrained and subsequent loss of fit compared. Despite the advantages of IRT methods, these models are not used frequently to test measurement invariance. In the psychometric literature, there is an ongoing debate comparing these two approaches (Kankaraš et al. 2011; Kim and Yoon 2011; Meade and 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 with a logistic method of estimation. Furthermore, CFA models typically estimate a single intercept per item because they work on the assumption that the data are continuous. In contrast, IRT models typically compute multiple parameters (thresholds) analogous to item intercepts per item—for IRT models, the polychotomous data are categorical, and as a consequence IRT models usually result in greater sensitivity to more-nuanced group differences such as in central tendency or the presence of extreme scores. Recent research shows that the IRT models can detect nonequivalence in the intercept (thresholds) and slope parameters both at the scale and the item level relatively accurately (Kankaraš et al. 2011). On the other hand, CFA performs well only when nonequivalence is located in the slope parameters, but wrongly indicates nonequivalence in the slope parameters when nonequivalence is located in the intercept parameters (Kankaraš et al. 2011). Some more advanced methods are available in CFA, especially the WLSMV estimator (weighted least squares means and variance adjusted), which estimates several thresholds instead of single intercept (Muthén 1993; Beauducel and Herzberg 2006), but comparisons of this method to the IRT approach are sparse (e.g. Kim and Yoon 2011). Recently, a new and promising method for testing measurement invariance has been proposed—the alignment method (Asparouhov and Muthen 2014), and we will use this approach to compare latent means across cultures.

FSCRS subscales (Inadequate Self, Hated Self, and Reassured Self) considered individually are unidimensional and moreover they share considerable variance, as shown in previous research by means of non-parametric IRT Mokken scale analysis (Halamová et al. 2018). Previously we performed the analyses for each population separately and therefore these results provide no information about whether the test scores are comparable across different populations. IRT models are better equipped than linear 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 on the invariance of a multifactorial framework. IRT analyses are suitable when testing the invariance of single, unidimensional scales such as Inadequate Self, Hated Self, and Reassured Self.

In the context of IRT models, measurement equivalence is tested by inspecting differential item functioning (DIF), and/or differential test functioning (DTF). Differential item functioning (DIF) means that an item within the FSCRS questionnaire measures the constructs (Inadequate Self, Hated Self, and Reassured Self) differently

for one population when compared with another. As a consequence, the presence of DIF compromises test validity. If this item bias accumulates to the extent that it produces biased overall test scores, a test will also display differential test functioning (DTF). DTF is present when respondents who have the same level of the latent construct, but belong to different groups, obtain different scores on the test.

DIF is routinely tested during scale construction and usually some method of purification is adopted; items with DIF are flagged and removed. However, if a test has many items (e.g., FSCRS has 22 items) and only some of them have DIF (see DeMars 2011), then the impact of these DIFs on the overall test score may be negligible. Moreover, there could be large DIF effects in favour of one population for some items, but these effects could 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 the case when the parameters systematically favour one group over another. Consequently, the aggregate of these small, nonsignificant differences at the item-level can become substantial at the test level (Chalmers et al. 2016). DTF is more relevant for our purpose than DIF; we do not intend to inspect particular items on FSCRS subscales nor do we intend to improve them. Rather, we intend to test the assumption that the (expected) total score of the FSCRS subscales is equivalent across different populations, and therefore only the latent trait—and not belonging to a particular group—has any impact on the (expected) total score. IRT methods are usually used to detect item bias (DIF), but for practical purposes, detecting the construct bias (DTF) is more useful; item bias could be large, with many items with DIF detected, but construct bias could be still negligible, with no DTF detected.

Testing the DTF involves two statistical measures (Chalmers et al. 2016). The first, the signed DTF tests whether there is any systematic scoring bias indicating that some groups consistently score higher across a specified range of the latent trait, and the second, the unsigned DTF, assess whether the test curves (plots of expected total score against a latent trait) have a large degree of overall separation on average, suggesting that there may be substantial DTF at particular levels of latent trait. 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 two 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, because 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 scoring, but there is substantial bias at particular levels of latent trait.

If there is substantial (significant) bias in the signed DTF, a FSCRS subscale is not invariant across countries; we cannot meaningfully compare test scores obtained from different countries, since the same values of test scores from

different countries correspond to different levels of latent trait. This has many practical consequences, but the most important lesson is that that it is misleading to compare naively test scores from countries where the DTF was detected.

The alignment method (Asparouhov and Muthen 2014) tries to search for invariant item loadings and intercepts and in turn latent means and standard deviations using an alignment optimization function (e.g., a quadratic loss function). The advantage of this procedure is that all groups can be compared simultaneously, and it allows aligning and comparing latent means even if some loadings and intercepts are severely non-invariant. Its logic is similar to factor rotation; the function minimizes some non-invariances while leaving some of them large. A configural invariance CFA model is fitted, and its parameter estimates (factor loadings and intercepts) are used as input for the alignment procedure. Asparouhov and Muthen (2014) provide effect sizes of approximate invariance based on R^2, and also the average correlation of aligned item parameters among groups. All aligned item factor loadings are approximately invariant (metric invariance) if the R^2 for factor loadings is close to 1 and the average correlation of aligned factor loadings is large. All aligned item intercepts are approximately invariant (scalar invariance) if the R^2 for intercepts is close to 1 and the average correlation of aligned intercepts is large.

Our analysis proceeded as follows:

1. Our procedure started with the identification of DIF, following which two randomly selected items with no DIF were used as anchors for DTF. If all items displayed DIF, two items were randomly selected as anchors for DTF (see Tables 3, 4 and 5). For DIF, we used the statistical program R (R Core Team 2017), package "lordif" (Choi et al. 2011).
2. We performed pairwise tests of DTF for all samples, separately for Inadequate Self, Hated Self, and Reassured Self. The total number of tests was 3*((13*12)/2) = 234 (see Tables 6, 7 and 8). We used the statistical program R (R Core Team 2017), package "mirt" (Chalmers 2012).
3. For samples with nonsignificant sDTF, we also report latent mean differences and their confidence intervals. Latent means in the reference group (first row) were constrained to zero, and latent means in the focal group (first column) were estimated (slopes and thresholds of items were constrained to be equal across countries). It must be highlighted that the DTF provides no information concerning the differences between countries in total scores; DTF only tests the assumption that these groups could be meaningfully compared, i.e. that their comparison would not be distorted. Only invariant samples, with no DTF present, can be meaningfully compared.
4. For all groups, we performed the alignment method proposed by Asparouhov and Muthen (2014), implemented in the R package "sirt" (Robitzsch 2018). A configural invariance CFA model was fitted, and its parameters (factor loadings and intercepts for each group) were used as input for the alignment procedure. Effect sizes R^2 for aligned factor loadings and intercepts are reported, as well as average correlations of aligned factor loadings and intercepts. Latent means and standard deviations for each subdimension and country are reported.

Table 3 Items with DIF for the Inadequate Self subscale (in parentheses), and items selected as anchors for DTF

DIF/IS	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	(1,2,8) 4,5	–	–	–	–	–	–	–	–	–	–	–
CH	(1,3,5,8) 6,7	(1–5,9) 6,7	–	–	–	–	–	–	–	–	–	–
ISR	(3,6–8) 6,7	(2,3,6,7) 4,5	(2–5,8,9) 6,7	–	–	–	–	–	–	–	–	–
ITA	(1–4,6–8) 5,9	(2–4,6–9) 1,5	(1–6,8) 7,9	(1–4,6,8) 5,7	–	–	–	–	–	–	–	–
JAP	(1–9) 4,5	(1–9) 4,5	(1–9) 4,5	(1–9) 4,5	(1–9) 4,5	–	–	–	–	–	–	–
NDL	(1,2,5,6,8) 3,4	(5,7) 2,3	(1–3,5,6) 7,8	(3–8) 1,2	(2,4–6) 7,8	(1–9) 4,5	–	–	–	–	–	–
POR	(1,2,6–9) 3,4	(2,3,7–9) 4,5	(1,4–6,9) 7,8	(2–9) 1,2	(2,6) 4,5	(1–9) 4,5	(2,4,5,8,9) 6,7	–	–	–	–	–
SVK	(1–3,6,8,9) 4,5	(1–3,9) 4,5	(1,2,4–8) 3,9	(1–8) 8,9	(2–4,6–9) 1,5	(1–9) 4,5	(2–4,7–9) 1,6	(1–3,7–9) 4,5	–	–	–	–
TWN	(1–9) 4,5	(3–7,9) 1,2	(2,4,5,7–9) 3,6	(3–5,7,9) 6,8	(2,4,5,8,9) 6,7	(1–9) 4,5	(1–9) 4,5	(1–9) 4,5	(1–2,4–9) 3,4	–	–	–
UK1	None 4,5	(1–3,8,9) 4,5	(1,3,5,8) 6,7	(3,6–8) 4,5	(1,2,4,6–8) 3,5	(1–9) 4,5	(1–8) 8,9	(1–9) 4,5	(1–9) 4,5	(1–9) 4,5	–	–
UK2	(1,4,8) 6,7	(1–4,6,8) 5,7	(2–5,8) 6,7	(1–9) 4,5	(1,2,4,6) 3,5	(1–9) 4,5	(1–3,6–8) 4,5	(1–3,5–9) 4,5	(1–9) 4,5	(1–6,8,9) 7,8	(1,2,4,5) 6,7	–
USA	(1,2,4,5,8) 6,7	(4,5,8) 6,7	(5–8) 3,4	(5–8) 3,4	(2,4,5–7,9) 1,8	(1–9) 4,5	(4,6,7) 2,3	(2,4,5,7,9) 6,8	(2–5,8–9) 6,7	(1–6,8,9) 7,8	(1,2,4,5) 6,7	(1,2,5) 3,4

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Table 4 Items with DIF for the Reassured Self subscale (in parentheses), and items selected as anchors for DTF

DIF/RE	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	(2) 4,5	–	–	–	–	–	–	–	–	–	–	–
CH	(2,5–7) 1,3	(3–5,7,8) 1,2	–	–	–	–	–	–	–	–	–	–
ISR	(1–3,6) 4,5	(1,3,6,7) 4,5	(2,3,5,7) 4,6	–	–	–	–	–	–	–	–	–
ITA	(2–7) 1,8	(3–7) 1,8	(3,4,8) 6,7	(1,4,5,7) 6,8	–	–	–	–	–	–	–	–
JAP	(1–8) 6,7	(1–8) 6,7	(1–8) 6,7	(2,4,6,7) 5,8	(3,4,6–8) 1,2	–	–	–	–	–	–	–
NDL	(1–3,6,7) 4,5	(1,3,5–7) 2,4	(3,5,7,8) 2,6	(1,3,6,7) 4,5	(1,3–7) 2,8	(1,3–8) 2,3	–	–	–	–	–	–
POR	(2,4–7) 1,8	(1,5–8) 3,4	(1,4–6,8) 2,3	(1,3–5,7) 6,8	(1,3–5,7,8) 2,6	(1–6) 7,8	(1,3,5–7) 2,8	–	–	–	–	–
SVK	(2,5,6) 4,7	(1,2,5–8) 3,4	(2,7) 4,5	(1,3–6) 7,8	(4,7,8) 5,6	(2–4,7,8) 5,6	(3,5,7,8) 4,6	(1,2,4–7) 3,8	–	–	–	–
TWN	(2–7) 1,8	(1–8) 3,4	(3,5–6,8) 1,2	(1,3—7) 2,8	(3–8) 1,2	(1,3–8) 2,3	(1–4,6–7) 5,8	(1,3,5,6) 2,4	(1,3–7) 2,8	–	–	–
UK1	(2,7) 4,5	(7) 3,4	(1,5–8) 3,4	(1,3,6,7) 4,5	(2–7) 1,8	(1–4,6–8) 4,5	(1–8) 4,5	(1,2,4–7) 3,8	(1–6) 7,8	(1–8) 4,5	–	–
UK2	(2,7) 4,5	(7) 3,4	(5,7,8) 3,4	(1,3,6) 4,5	(3–7) 1,2	(1–8) 4,5	(1–3,6,7) 4,5	(1,5–8) 2,3	(1–2,5–8) 3,4	(1,3–7) 2,8	(6) 4,5	–
USA	(2,7) 4,5	(5,7) 3,4	(1,2,6,8) 3,4	(1,3,6) 4,5	(3–8) 1,2	(1–4,6–8) 4,5	(1,3,5–7) 2,8	(3,5–8) 2,4	(1–2,6,7) 4,5	(1–8) 2,8	(3,7) 4,5	(3,7) 4,5

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Table 5 Items with DIF for the Hated Self subscale (in parentheses), and items selected as anchors for DTF

DIF/HS	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	None 3,4	–	–	–	–	–	–	–	–	–	–	–
CH	(1–5) 2,3	(1–5) 2,3	–	–	–	–	–	–	–	–	–	–
ISR	(4,5) 2,3	(5) 3,4	(1–5) 2,3	–	–	–	–	–	–	–	–	–
ITA	(1–5) 3,4	(1–3,5) 4,5	(1,3–5) 2,3	(1–5) 4,5	–	–	–	–	–	–	–	–
JAP	(1) 2,3	(1,4) 2,3	(1–5) 3,4	(1–5) 3,4	(1–5) 1,2	–	–	–	–	–	–	–
NDL	(2,3,5) 1,4	(2–5) 1,4	(4,5) 2,3	(3,4) 1,2	(1,2,4) 3,5	(2,4) 1,5	–	–	–	–	–	–
POR	(1–5) 4,5	(1–5) 3,4	(1–5) 1,3	(2–4) 1,5	(1,4) 2,3	(1–5) 4,5	(2,4) 3,5	–	–	–	–	–
SVK	(2,4) 3,5	(1,4) 2,3	(2–5) 1,2	(1–5) 4,5	(1–5) 2,3	(1,4) 3,5	(1–4) 4,5	(1–5) 3,4	–	–	–	–
TWN	(3) 4,5	(2,4) 1,5	(1–5) 3,4	(1–5) 2,8	(1–5) 4,5	(1) 2,3	(1–5) 4,5	(1–5) 2,3	(1,2,4) 3,5	–	–	–
UK1	(5) 2,3	None 3,4	(1–5) 3,4	(5) 2,3	(1–3,5) 4,5	(1–5) 2,3	(1,2) 4,5	(1–5) 4,5	(1–5) 1,2	(1–5) 4,5	–	–
UK2	None 4,5	(2,4) 1,3	(1–5) 2,3	(4,5) 2,3	(1–5) 2,3	(1–5) 3,4	(1,2) 3,4	(1–5) 3,4	(1–4) 2,5	(1–5) 4,5	(2,4,5) 1,3	–
USA	None 2,3	(2,4) 3,5	(1–5) 2,3	(4,5) 1,3	(1–5) 3,4	(1,2) 3,5	(2,3,5) 1,4	(1,2) 4,5	(1,4) 2,3	None 2,3	(2–5) 1,2	None 2,3

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Table 6 Signed differential test statistics for the Inadequate Self subscale

sDTF	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	**−0.19 ns**	–	–	–	–	–	–	–	–	–	–	–
CH	−0.41*	−0.55**	–	–	–	–	–	–	–	–	–	–
ISR	−0.31**	**−0.22 ns**	0.67***	–	–	–	–	–	–	–	–	–
ITA	**−0.21 ns**	**0.02 ns**	**−0.27 ns**	0.61***	–	–	–	–	–	–	–	–
JAP	**0.18 ns**	0.43*	**−0.43 ns**	0.74**	0.54**	–	–	–	–	–	–	–
NDL	**−0.16 ns**	**0.26 ns**	**−0.07 ns**	0.51**	**0.07 ns**	0.52*	–	–	–	–	–	–
POR	**0.21 ns**	0.22*	**0.04 ns**	−0.45***	**0.20 ns**	−0.44*	**0.20 ns**	–	–	–	–	–
SVK	**−0.09 ns**	**0.11 ns**	**−0.01 ns**	−0.70***	**−0.10 ns**	−0.70**	**−0.23 ns**	**−0.12 ns**	–	–	–	–
TWN	**0.20 ns**	−1.00***	−1.36***	−0.90***	−0.54**	**0.02 ns**	0.53**	0.33**	−0.53**	–	–	–
UK1	**−0.13 ns**	**0.07 ns**	**0.29 ns**	0.30**	**−0.19 ns**	−0.55*	−0.89***	**−0.16 ns**	**−0.05 ns**	−0.67***	–	–
UK2	**−0.21 ns**	**−0.16 ns**	**0.22 ns**	**−0.14 ns**	−0.25*	−1.66***	−0.75***	−0.62***	−0.51***	**0.02 ns**	**−0.06 ns**	–
USA	**0.13 ns**	**0.09 ns**	0.57**	−0.67***	**−0.17 ns**	−1.34***	**0.13 ns**	**−0.19 ns**	**−0.13 ns**	0.44*	**0.21 ns**	**−0.06 ns**

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

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331), ns non-significant

*$p < .05$

**$p < .01$

***$p < .001$

Table 7 Signed differential test statistics for the Reassured Self subscale

sDTF	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	**−0.19 ns**	–	–	–	–	–	–	–	–	–	–	–
CH	**0.19 ns**	0.90***	–	–	–	–	–	–	–	–	–	–
ISR	**−0.19 ns**	**0.13 ns**	**0.13 ns**	–	–	–	–	–	–	–	–	–
ITA	**0.14 ns**	**0.13 ns**	**0.13 ns**	**0.19 ns**	–	–	–	–	–	–	–	–
JAP	**−0.03 ns**	**−0.15 ns**	**0.28 ns**	**−0.23 ns**	1.37***	–	–	–	–	–	–	–
NDL	**−0.11 ns**	**−0.28 ns**	**0.49 ns**	**0.07 ns**	0.53*	0.85***	–	–	–	–	–	–
POR	**−0.14 ns**	**0.09 ns**	0.49***	**0.09 ns**	**0.19 ns**	**0.14 ns**	**−0.11 ns**	–	–	–	–	–
SVK	**−0.19 ns**	**−0.18 ns**	**0.26 ns**	**−0.20 ns**	**0.23 ns**	−0.36*	**0.13 ns**	**−0.14 ns**	–	–	–	–
TWN	**0.02 ns**	**−0.12 ns**	**−0.11 ns**	**−0.01 ns**	0.36*	0.46**	**−0.06 ns**	**0.06 ns**	**0.07 ns**	–	**–**	–
UK1	**−0.08 ns**	**−0.02 ns**	−0.37***	**0.02 ns**	−0.54***	0.54***	**0.05 ns**	**0.08 ns**	0.62***	−0.35**	–	–
UK2	**−0.09 ns**	**−0.10 ns**	−0.33***	**−0.01 ns**	**−0.32 ns**	0.58***	**0.05 ns**	**−0.12 ns**	**0.09 ns**	**0.05 ns**	**−0.01 ns**	–
USA	**0.01 ns**	**0.28 ns**	**0.01 ns**	**0.09 ns**	**−0.29 ns**	0.68***	**0.14 ns**	**−0.20 ns**	0.43**	**0.08 ns**	**0.09 ns**	**0.10 ns**

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

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331). ns non-significant

*$p < .05$

**$p < .01$

***$p < .001$

Table 8 Signed differential test statistics for the Hated Self subscale

sDTF	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	**−0.14 ns**	–	–	–	–	–	–	–	–	–	–	–
CH	**−0.26 ns**	1.09***	–	–	–	–	–	–	–	–	–	–
ISR	**−0.05 ns**	**−0.05 ns**	**−0.15 ns**	–	–	–	–	–	–	–	–	–
ITA	**−0.15 ns**	**0.06 ns**	**−0.09 ns**	**−0.22 ns**	–	–	–	–	–	–	–	–
JAP	**0.11 ns**	**−0.12 ns**	0.45*	**0.17 ns**	**0.11 ns**	–	–	–	–	–	–	–
NDL	**−0.17 ns**	**−0.08 ns**	**−0.28 ns**	**−0.28 ns**	**−0.18 ns**	−0.25*	–	–	–	–	–	–
POR	**0.21 ns**	0.45*	−0.40***	−0.34**	**−0.10 ns**	**−0.06 ns**	**0.03 ns**	–	–	–	–	–
SVK	−0.38***	−0.44**	−0.70***	**0.19 ns**	**0.16 ns**	−0.49***	**0.19 ns**	**−0.12 ns**	–	–	–	–
TWN	**−0.01 ns**	**−0.18 ns**	−0.56***	**0.11 ns**	**0.29 ns**	−0.37***	**−0.07 ns**	**0.13 ns**	**0.09 ns**	–	**–**	–
UK1	**0.07 ns**	**0.19 ns**	−0.32*	**0.10 ns**	**0.16 ns**	**−0.05 ns**	**−0.08 ns**	**−0.18 ns**	**0.06 ns**	**−0.10 ns**	–	–
UK2	**0.06 ns**	**−0.01 ns**	**0.14 ns**	**−0.07 ns**	**0.11 ns**	**−0.01 ns**	**−0.10 ns**	**−0.14 ns**	**−0.16 ns**	**−0.03 ns**	**0.03 ns**	–
USA	**−0.14 ns**	**−0.20 ns**	**0.16 ns**	**−0.12 ns**	**0.15 ns**	**−0.07 ns**	**0.26 ns**	**−0.24 ns**	**0.06 ns**	**0.02 ns**	**−0.01 ns**	**0.03 ns**

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

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331). ns non-significant

*p <.05

**p <.01

***p <.001

Results

DIF testing showed (Tables 3, 4 and 5) that the number of items with DIF varied greatly among the samples, from no DIF detected to all items displaying DIF. The results suggest that the presence or absence of DIF is not a systematic predictor of DTF.

Out of 78 comparisons, there were 43 measurement equivalencies (DTF) for the Inadequate Self subscale (see Table 6), 61 measurement equivalencies for the Reassured Self subscale (see Table 7), and 65 measurement equivalencies for the Hated Self subscale (see Table 8). For the Inadequate Self subscale, the Australian sample was equivalent to 10 other samples, the Canadian, Italian, Slovak, UK1 and USA sample were equivalent to 8 other samples, the Netherlands, Portugal, Switzerland and UK2 samples were each equivalent to 7 other samples, the Japan and Taiwan samples were equivalent to 3 other samples, and finally the Israel sample was equivalent to 2 other samples. As for the Reassured Self subscale, Australian and Israel samples were equivalent to all 12 other samples, Canadian and Portugal samples were equivalent to 11 other samples, the Netherlands, UK2 and USA samples were equivalent to 10 other samples, Slovakia and Taiwan samples were equivalent to 9 other samples, Italy, Switzerland were equivalent to 8 other samples, UK1 sample was equivalent to 7 other samples, and finally Japan sample was equivalent to 5 other samples. For the Hated Self subscale, Italian, UK2 and USA samples were equivalent to all 12 other samples, Australia, Israel, the Netherlands and UK1 samples were equivalent to 11 other samples, Taiwan sample was equivalent to 10 other samples, Canadian and Portugal samples were equivalent to 9 other samples, Japan and Slovak samples were equivalent to 8 other samples, and finally Switzerland sample was equivalent to 6 other samples.

It should be noted that no transitivity can be assumed; for example, for the Hated Self subscale, both the Netherlands and Japan samples were equivalent to the Canadian 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 Hated Self subscale, we could compare Canada, Australia, UK1, UK2, Israel, Japan and USA samples because all were mutually equivalent. However, adding another sample, for example, Switzerland was not possible: it was equivalent with all other samples, but not with Canadian sample. Therefore, we can compare the latent means of equivalent samples for each subscale (Tables 9, 10 and 11). These latent mean differences indicate that one population's answers are more or less self-critical than other's. Of course, we cannot exclude the possibility that answers from populations would not be significantly different. We note again that differences in latent means have nothing to do with and are orthogonal to measurement equivalence; rather, measurement equivalence is a necessary prerequisite for comparing latent means. Without the measurement equivalence, any comparison between two populations would be distorted by the differential functioning of the test itself and therefore could not represent differences in the latent trait.

Table 9 Latent mean differences of invariant samples for the Inadequate Self subscale

Mean diff	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	**0.202 (0.056, 0.348)**	–	–	–	–	–	–	–	–	–	–	–
CH	–	–	–	–	–	–	–	–	–	–	–	–
ISR	–	**−0.771 (−0.929, −0.613)**	–	–	–	–	–	–	–	–	–	–
ITA	−0.040 (−0.182, 0.103)	**−0.329 (−0.478, −0.181)**	**0.390 (0.209, 0.572)**	–	–	–	–	–	–	–	–	–
JAP	**0.905 (0.737, 1.073)**	–	**0.809 (0.615, 1.003)**	–	–	–	–	–	–	–	–	–
NDL	**−0.159 (−0.298, −0.020)**	**−0.421 (−0.565, −0.277)**	**−0.425 (−0.591, −0.260)**	–	0.127 (−0.035, 0.290)	–	–	–	–	–	–	–
POR	**−0.304 (−0.436, −0.172)**	–	**0.675 (0.506, 0.844)**	–	**−0.305 (−0.437, −0.173)**	–	**−0.221 (−0.362, −0.081)**	–	–	–	–	–
SVK	0.091 (−0.032, 0.214)	**−0.157 (−0.276, −0.039)**	**0.210 (0.037, 0.383)**	–	**0.223 (0.104, 0.342)**	–	**0.314 (0.185, 0.442)**	**0.473 (0.379, 0.568)**	–	–	–	–
TWN	**0.333 (0.200, 0.466)**	–	–	–	–	**−0.821 (−0.996, −0.647)**	–	–	–	–	–	–
UK1	**0.228 (0.101, 0.355)**	0.022 (−0.101, 0.145)	−0.003 (−0.150, 0.145)	–	**0.324 (0.198, 0.450)**	–	–	**0.648 (0.544, 0.752)**	**0.183 (0.091, 0.275)**	–	–	–

Table 9 (continued)

Mean diff	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
UK2	**0.220 (0.090, 0.350)**	−0.006 (−0.133, 0.120)	−0.017 (−0.168, 0.133)	**0.750 (0.623, 0.878)**	–	–	–	–	–	**−0.250 (−0.401, −0.098)**	−0.013 (−0.091, 0.065)	–
USA	−0.034 (−0.184, 0.116)	**−0.284 (−0.442, −0.127)**	–	–	0.031 (−0.128, 0.190)	–	0.151 (−0.023, 0,326)	**0.334 (0.193, 0.474)**	**−0.156 (−0.299, 0.013)**	–	**−0.258 (−0.368, −0.148)**	**−0.287 (−0.421, −0.152)**

Latent mean estimations of populations in first row were constrained to zero. Differences significant at 0.05 level are highlighted in bold

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1,326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Table 10 Latent mean differences of invariant samples for the Reassured Self subscale

Mean diff	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	0.144 (−0.003, 0.290)	–	–	_	–	–	–	–	–	–	–	–
CH	**−0.397 (−0.579, −0.215)**	–	–	_	–	–	–	–	–	–	–	–
ISR	**0.344 (0.190, 0.497)**	**0.242 (0.084, 0.400)**	**0.747 (0.566, 0.927)**	–	–	–	–	–	–	–	–	–
ITA	0.061 (−0.082, 0.203)	−0.134 (−0.280, 0.012)	**0.421 (0.259, 0.583)**	**−0.335 (−0.464, −0.206)**	–	–	–	–	–	–	–	–
JAP	**−0.517 (−0.676, −0.358)**	**−0.811 (−0.987, −0.635)**	0.086 (−0.143, 0.315)	**−0.889 (−1.042, −0.737)**	–	–	–	–	–	–	–	–
NDL	**0.261 (0.116, 0.405)**	0.138 (−0.010, 0.285)	**0.655 (0.483, 0.827)**	−0.115 (−0.243, 0.014)	–	–	–	–	–	–	–	–
POR	0.092 (−0.040, 0,223)	−0.074 (−0.205, 0.058)	–	**−0.284 (−0.401, −0.167)**	0.056 (−0.079, 0.190)	**0.748 (0.577, 0.920)**	**−0.252 (−0.396, −0.109)**	–	–	–	–	–
SVK	0.119 (−0.007, 0.245)	−0.030 (−0.152, 0.091)	**0.504 (0.352, 0.656)**	**−0.248 (−0.356, −0.140)**	0.123 (−0.001, 0.247)	–	**−0.200 (−0.330. −0.069)**	0.054 (−0.041, 0.148)	–	–	–	–
TWN	**0.147 (0.013, 0.281)**	−0.005 (−0.138, 0.128)	**0.508 (0.351, 0.666)**	**−0.265 (−0.384, −0.146)**	–	–	**−0.200 (−0.343, −0.056)**	0.059 (−0.049, 0.167)	0.005 (−0.097, 0.107)	–	–	–
UK1	**−0.188 (−0.314, −0.062)**	**−0.399 (−0.525, −0.273)**	–	**−0.546 (−0.661, −0.432)**	–	–	**−0.604 (−0.744, −0.464)**	**−0.360 (−0.458, −0.261)**	–	–	–	–

Table 10 (continued)

Mean diff	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
UK2	0.079 (−0.049, 0.208)	−0.072 (−0.198, 0.053)	–	**−0.264 (−0.376, −0.151)**	0.070 (−0.060, 0.200)	–	**−0.225 (−0.361, −0.089)**	0.011 (−0.091, 0.112)	−0.033 (−0.125, 0.059)	−0.049 (−0.188, 0.090)	**−0.289 (−0.368, −0.210)**	–
USA	−0.094 (−0.252, 0.065)	**0.299 (0.129, 0.470)**	**0.296 (0.123, 0.470)**	**−0.463 (−0.615, −0.311)**	**−0.198 (−0.379, −0.018)**	–	**−0.525 (−0.720, −0.330)**	**−0.218 (−0.370, −0.066)**	–	**0.386 (0.181, 0.592)**	0.093 (−0.032, 0.218)	**0.244 (0.088, 0.401)**

Latent mean estimations of populations in first row were constrained to zero. Differences significant at 0.05 level are highlighted in bold

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1,326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Table 11 Latent mean differences of invariant samples for the Hated Self subscale

Mean diff	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
CAN	0.029 (−0.142, 0.201)	–	–	–	–	–	–	–	–	–	–	–
CH	**0.390 (0.202, 0.577)**	–	–	–	–	–	–	–	–	–	–	–
ISR	**−0.804 (−1.053, −0.556)**	**−0.921 (−1.184, −0.658)**	**1.038 (0.864, 1.212)**	–	–	–	–	–	–	–	–	–
ITA	0.003 (−0.165, 0.172)	−0.049 (−0.219, 0.121)	**0.470 (0.269, 0.672)**	**0.714 (0.570, 0.858)**	–	–	–	–	–	–	–	–
JAP	**1.425 (1.216, 1.634)**	**1.615 (1.392, 1.837)**	–	**1.977 (1.750, 2.203)**	**1.838 (1.606, 2.069)**	–	–	–	–	–	–	–
NDL	−0.174 (−0.363, 0.016)	**−0.230 (−0.427,−0.033)**	**0.541 (0.357, 0.724)**	**0.556 (0.398, 0.714)**	**−0.205 (−0.417, −0.007)**	–	–	–	–	–	–	–
POR	**−0.304 (−0.478, −0.130)**	–	–	–	**−0.445 (−0.633, −0.258)**	**−2.392 (−2.742, −2.042)**	−0.130 (−0.292, 0.032)	–	–	–	–	–
SVK	–	–	–	**1.079 (0.947, 1.212)**	**0.637 (0.497, 0.778)**	–	**0.685 (0.550, 0.820)**	**0.878 (0.770, 0.986)**	–	–	–	–
TWN	**0.665 (0.516, 0.814)**	**0.736 (0.588, 0.883)**	–	**1.293 (1.139, 1.447)**	**0.999 (0.835, 1.164)**	–	**0.919 (0.765, 1.074)**	**1.144 (1.014, 1.274)**	**0.389 (0.275, 0.502)**	–	–	–
UK1	0.081 (−0.067, 0.230)	0.062 (−0.083, 0.208)	–	**0.750 (0.620, 0.880)**	0.090 (−0.061, 0.241)	**−1.656 (−1.909, −1.403)**	**0.244 (0.101, 0.388)**	**0.419 (0.301, 0.537)**	**−0.495 (−0.619, −0.371)**	**−0.863 (−1.042, −0.684)**	–	–
UK2	0.027 (−0.131, 0.185)	0.004 (−0.154, 0.162)	**−0.366 (−0.563, −0.170)**	**0.715 (0.576, 0.854)**	0.027 (−0.140, 0.194)	**−1.799 (−2.074, −1.524)**	**0.206 (0.052, 0.361)**	**0.366 (0.232, 0.499)**	**−0.592 (−0.740, −0.445)**	**−0.951 (−1.151, −0.752)**	−0.044 (−0.137, 0.049)	–

Table 11 (continued)

Mean diff	AUS	CAN	CH	ISR	ITA	JAP	NDL	POR	SVK	TWN	UK1	UK2
USA	0.134 (−0.045, 0.314)	0.122 (−0.063, 0.306)	0.097 (−0.039, 0.232)	**0.806 (0.645, 0.967)**	**0.205 (0.001, 0.405)**	**−1.740 (−2.034, −1.445)**	**0.340 (0.159, 0.521)**	**0.503 (0.335, 0.672)**	**−0.440 (−0.637, −0.242)**	**−0.793 (−1.021, −0.565)**	0.044 (−0.074, 0.163)	0.097 (−0.039, 0.232)

Latent mean estimations of populations in first row were constrained to zero. Differences significant at 0.05 level are highlighted in bold

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1,326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Figure 1 shows test score functions of Israel and Switzerland samples of Inadequate Self subscale (top), and Reassured Self subscale (bottom) – expected total scores plotted against the latent trait (θ). As far as Reassured Self subscale (bottom) is concerned, one can clearly see large differences between curves from −4 to 0 values of θ, and then from 0 to 4, but in the opposite direction. Although the differences were very large (the unsigned DTF is 0.82, which was 2.57% of distortion), their impact on difference in expected total score (the signed DTF) was only 0.13 and non-significant at the 0.05 level. We could conclude that no significant DTF was present at the total score level. However, there were differences at particular levels of the latent trait (θ); respondents with very high and very low levels of Reassured Self responded differently in the Israel and Switzerland samples, but in opposite directions, so the effect was cancelled out. With regards to the Inadequate Self subscale (top), the situation was very different; again, we could see a large difference between the curves from 0 to 4 values of θ, but this difference was not compensated by the difference between −4 to 0 in the opposite direction. The amount of differences was virtually the same as in the Reassured Self subscale (the unsigned DTF is 0.84, which was 2.34% of distortion), but the lack of compensation led to a larger impact on the difference in expected total score; the signed DTF is 0.67 and significant at the 0.001 level. Each curve had a 95% confidence interval envelope.

It is 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 is equal 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 trait.

For the Inadequate Self subscale, the effect size R^2 for aligned factor loadings was 0.985, R^2 for aligned intercepts was 0.989, the average correlation of aligned factor loadings was 0.647 and the average correlation of aligned intercepts was 0.576. We can conclude that the alignment procedure successfully recovered approximate invariance. Latent means and their standard deviations are reported in Table 12.

For the Reassured Self subscale, the effect size R^2 for aligned factor loadings was 0.990, R^2 for aligned intercepts was 0.996, the average correlation of aligned factor loadings was 0.492 and the average correlation of aligned intercepts was 0.855. We can conclude that the alignment procedure successfully recovered approximate invariance. Latent means and their standard deviations are reported in Table 12.

For the Hated Self subscale, the effect size R^2 for aligned factor loadings was 0.991, the R^2 for aligned intercepts was 0.966, the average correlation of aligned factor loadings was 0.434 and the average correlation of aligned intercepts was 0.370. We can conclude that the alignment procedure successfully recovered approximate invariance. Latent means and their standard deviations are reported in Table 12.

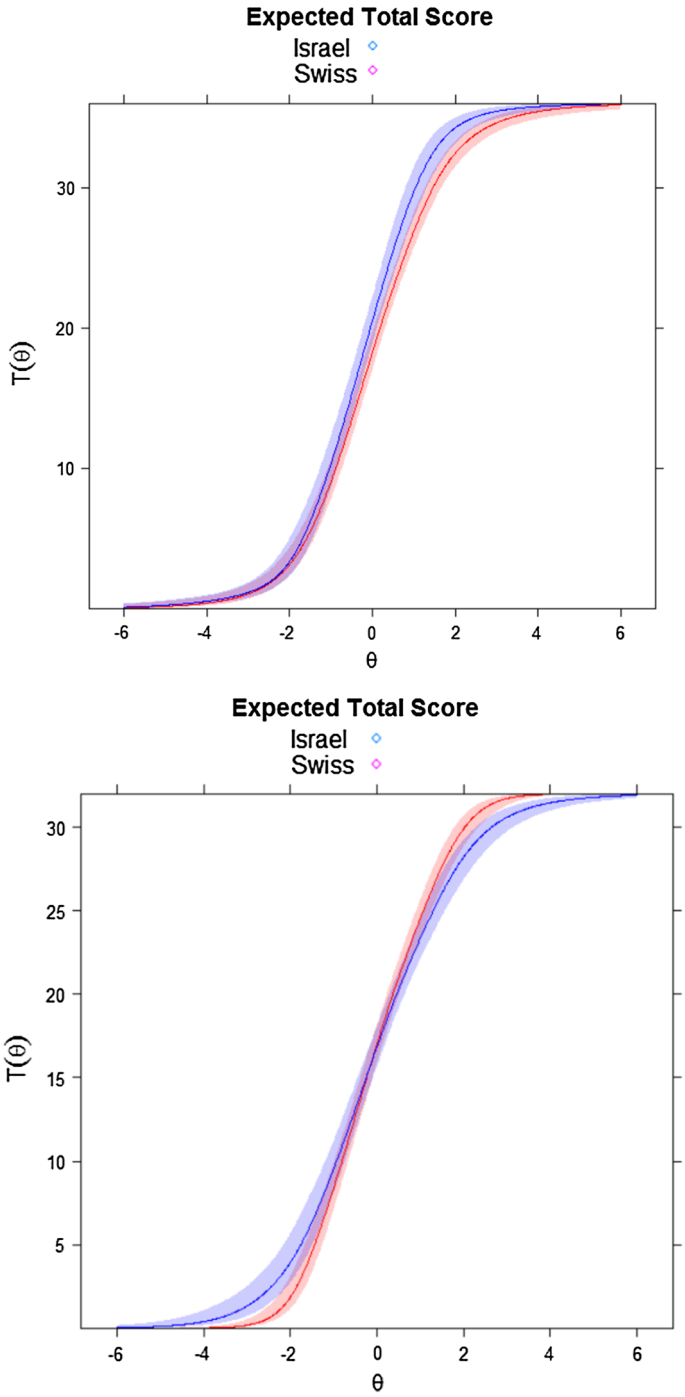

Fig. 1 Test score functions of Israel and Switzerland samples in Inadequate Self (top) and Reassured Self (bottom) subscales

Table 12 Latent mean differences after the alignment procedure

Inadequate self			Reassured self			Hated self		
Country	Latent mean	SD	Country	Latent mean	SD	Country	Latent mean	SD
JAP	0.956	0.808	JAP	−0.581	0.943	JAP	1.574	1.123
TAI	0.302	0.801	CH	−0.514	1.284	TAI	0.439	0.967
CH	0.145	1.126	UK1	−0.254	1.116	CH	0.234	1.107
CAN	0.139	1.054	USA	−0.186	1.128	SVK	0.216	0.956
UK1	0.112	1.160	AUS	−0.055	1.175	UK1	0.016	1.276
UK2	0.110	1.065	ITA	0.028	0.927	USA	−0.036	1.112
SVK	−0.023	0.951	POR	0.082	1.003	UK2	−0.044	1.195
AUS	−0.151	1.179	UK2	0.101	0.929	AUS	−0.165	1.060
USA	−0.165	1.052	SVK	0.136	0.922	ITA	−0.199	0.881
ITA	−0.225	0.999	CAN	0.164	0.979	CAN	−0.207	0.918
NL	−0.312	0.909	TAI	0.197	0.782	NL	−0.306	0.925
POR	−0.489	0.987	NL	0.292	0.882	POR	−0.406	0.863
ISR	−0.648	0.997	ISR	0.351	1.040	ISR	−0.620	0.750

AUS Australia (N = 319), *CAN* Canada (N = 380), *CH* Switzerland (N = 230), *ISR* Israel (N = 475), *ITA* Italy (N = 393), *JAP* Japan (N = 263), *NL* Netherlands (N = 363), *POR* Portugal (N = 764), *SVK* Slovakia (N = 1326), *TAI* Taiwan (N = 417), *UK1* United Kingdom 1 (N = 1570), *UK2* United Kingdom 2 (N = 883) and *USA* (N = 331)

Discussion

The present study used IRT differential test functioning to test the measurement invariance of the dimensions of the FSCRS using 13 samples from 12 different countries and eight language versions. The results demonstrate that in the majority of comparisons there is high measurement equivalence between the different countries suggesting that in general the FSCRS subscales are valid and reliable instruments with substantial cross-cultural potential. Nevertheless, some comparisons resulted in a lack of measurement equivalence and therefore displayed differential test functioning. Additional research would be necessary to determine whether this lack of measurement equivalence was caused by shifts in linguistic meaning, possible translation issues, by real differences in levels of self-criticism/reassurance across countries, by peculiarities in sampling procedures or by differences in gender or age between samples.

We have to stress that the IRT method (DTF) used in this paper detects construct bias, and not item bias: if some items are biased (DIF detected), it does not entail that construct bias (DTF) must follow necessarily—that would happen only if items were biased systematically in favour of one group. On the other hand, and even more importantly, there could be no substantive item bias (no items with DIF detected), but construct bias (DTF) could be present: small differences in functioning of particular items could be so systematic in favour of one group that they could distort the construct and its test score. These situations have clear practical consequences: in the first case, this method can save the validity of test score even if several items

display a substantive DIF (item bias); in the second case, this method can detect the problems with the test score (construct bias) even if no item displays a substantive DIF.

Hated Self was the most invariant of the three subscales suggesting that self-hatred is quite similarly described across cultures. Also, Reassured Self was quite high in invariance which means that it too is quite analogous across cultures. The Inadequate Self subscale was the least invariant across cultures, suggesting that the experience and intensity of inadequacy could be very different across cultures. One possible source of the variance across countries and languages of Inadequate Self compared to Hated Self and Reassured Self could be the diversity of the standards prescribed for people in different cultures around world. According to our results, Israel, Japan and Taiwan are the countries with the most divergent perception of Inadequate Self. Japan and Taiwan scored the highest and Israel the lowest on the subscale of Inadequate Self. In contrast, Australia is the country with the most similar perception of Inadequate Self to the other countries assessed in this study.

Japan is the country with the most differing perception of Reassured Self, and Switzerland is the country with the most differing perception of Hated Self among the samples from different countries. In our research, the sample from Japan was the most self-critical (on both IS and HS) out of thirteen samples which confirms previous research suggesting that the Japanese population are more self-critical than North Americans (Heine et al. 2000). Also, our research findings support distinctions between Eastern and Western countries (Heine and Hamamura 2007), with countries located in the East (such as Japan and Taiwan) being more self-critical than countries located in the West. It is interesting that these differences between Japan and USA or East and West countries are present despite the use of an specific definition of self-criticism. We made no assumption that self-criticism is a general sensitivity to negative self-relevant information (Kitayama et al. 1997), but self-criticism is due to constant and harsh self-scrutiny and evaluation and feelings of unworthiness, inferiority, failure, and guilt (Blatt and Zuroff 1992). Interestingly, Taiwan is the second most self-critical country among the all analysed countries, but it is also quite high in self-reassurance. However, Israel is the most self-reassured and the least self-critical country.

The main limitation of our study was that FSCRS is a self-report tool, and therefore participants may have been influenced to respond in a socially desirable manner which may vary between cultures. Also, the samples were recruited mainly online but also in paper–pencil form, so different forms of obtaining data could influence the findings.

As self-criticism is a construct of high clinical importance, improving understanding of its cross-cultural similarities and differences, as measured by the three subscales of the FSCRS, would have great impact on practice. This is because a negative relation to oneself in the form of excessive self-critical inner dialogue is one of the most important psychological processes that influence susceptibility to, and persistence of, psychopathology (Falconer et al. 2015) and stress (Kupeli et al. 2017). Self-reassurance, which is closely related to self-compassion (Kupeli et al. 2013), is of great importance in its own right. And of course, it's the target in many outcome studies done worldwide (Kirby 2016), so we need to know about its measurement,

too. So a tool which is sensitive or applicable to these small but important differences will be very useful in evaluating interventions. Thus, understanding the differences of self-criticism across countries can help to inform more effective practices in both medical treatment and psychotherapy.

We did not attempt to provide any systematic interpretation of these differences except for the differences between East and West countries and thus far more detailed research is required to do so. However, we could see that no discernible pattern emerged from mutually equivalent samples with cultural, linguistic or geographical continuum able to explain clusters of mutually equivalent countries.

Conclusion

This study contributes to the growing body of knowledge about the similarities and differences among cultures with respect to the three subscales of the FSCRS: Hated Self, Inadequate Self and Reassured Self. Our study revealed significant cross-cultural similarities and differences in the way these constructs are measured by the subscales of the FSCRS. Interestingly these differences are far larger for Inadequate Self than for Hated Self and Reassured Self, which seem to be quite invariant across cultures. One reason may be that self-hatred is tapping into a pathological dimension and self-reassurance is tapping into a health dimension that are indeed culturally invariant, whereas inadequate self is tapping into a competitive or social rank dimension that is more culturally bound. Hence, cultures that seem more collective may also be more sensitive to shame and stigma and the negative evaluation of others. This may partly explain why individuals from the Japanese culture report more self-criticism, because they may be more sensitive to social evaluation and social place. Although the items from the FSCRS are not related to specific events it may be that different types of events in different cultures are more susceptible to self-criticism and this would need to be explored.

Although the FSCRS subscales are generally valid and reliable instruments with substantial potential for use cross-culturally, the three subscales were not perfectly invariant across all countries and groups. In view of the culturally and linguistically different expressions of self-criticism and self-reassurance that were observed, future cross-cultural testing of the meanings and connotations of these constructs is necessary. An important direction for future research is to investigate the factors responsible for the observed non-equivalences. Cross-cultural researchers must also continue to bear in mind that it is only possible to compare mean scores across countries which were found to be invariant.

Author Contributions JH designed research, invited co-authors to participate and coordinated research team. JH, PG, NK, NT, DZ, NH, NP, MS, JK, BS, TK, KA, FY, MM and JB shared their data. MK performed the statistical analysis. JH and MK wrote the first draft of the article. All authors interpreted the results, revised the manuscript and read and approved the final manuscript.

Funding Writing this work was supported by the Vedecká grantová agentúra VEGA under Grant 1/0578/15. Nuriye Kupeli is supported by Marie Curie core Grant funding, Grant MCCC-FCO-16-U.

Availability of Data and Materials In order to comply with the ethics approvals of the study protocols, data cannot be made accessible through a public repository. However, data are available upon request for researchers who consent to adhering to the ethical regulations for confidential data.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no potential conflicts of interests.

Ethical Approval All procedures performed in studies involving human participants were 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.

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

References

Asparouhov, T., & Muthen, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21,* 1–14. https://doi.org/10.1080/10705511.2014.919210.

Baião, R., Gilbert, P., McEwan, K., & Carvalho, S. (2015). Forms of Self-Criticising/Attacking & Self-Reassuring Scale: Psychometric properties and normative study. *Psychology and Psychotherapy, 88*(4), 438–452. https://doi.org/10.1111/papt.12049.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186–203. https://doi.org/10.1207/s15328007sem1302_2.

Blatt, S. J. (2004). *Experiences of depression: Theoretical, clinical and research perspectives*. Washington, DC: American Psychological Association.

Blatt, S. J., D'Afflitti, J. P., & Quinlan, D. M. (1979). *The depressive experiences questionnaire*. New Haven: Yale University.

Blatt, S. J., Quinlan, D. M., Pilkonis, P. A., & Shea, M. T. (1995). Impact of perfectionism and need for approval on the brief treatment of depression: The National Institute of Mental Health Treatment of Depression Collaborative Research Program Revisited. *Journal of Consulting and Clinical Psychology, 63,* 125–132.

Blatt, S. J., & Shichman, S. (1983). Two primary configurations of psychopathology. *Psychoanalysis and Contemporary Thought, 6*(2), 187–254.

Blatt, S. J., & Zuroff, D. C. (1992). Interpersonal relatedness and self-definition: Two prototypes for depression. *Clinical Psychology Review, 12,* 527–562. https://doi.org/10.1016/0272-7358(92)90070-o.

Blatt, S. J., & Zuroff, D. C. (2005). Empirical evaluation of the assumptions in identifying evidence based treatments in mental health. *Clinical Psychology Review, 25,* 459–486. https://doi.org/10.1016/j.cpr.2005.03.001.

Castilho, P., Pinto-Gouveia, J., & Duarte, J. (2015). Exploring self-criticism: Confirmatory factor analysis of the FSCRS in clinical and nonclinical samples. *Clinical Psychology and Psychotherapy, 22*(2), 153–164. https://doi.org/10.1002/cpp.1881.

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.

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.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1–30. https://doi.org/10.18637/jss.v039.i08.

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.

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), 351–365. https://doi.org/10.1111/papt.12056.

Gheysen, F., Katis, S. Lee, M. & Delamillieure, P. (2015). Learning to use CFT, the French experience: philosophy, methodology, first data. Centre Esquirol, CHU de Caen, France. Retrieved March 21, 2017 from http://s207773256.websitehome.co.uk/conference/conference_2015/presentations_2015/CMF17-FG%20&%20PD%20MANCHESTER231015.pdf.

Gilbert, P., Allan, S., Brough, S., Melley, S., & Miles, J. N. V. (2002). Relationship of anhedonia and anxiety to social rank, defeat and entrapment. *Journal of Affective Disorders, 71,* 141–151. https://doi.org/10.1016/S0165-0327(01)00392-5.

Gilbert, P., Baldwin, M. W., Irons, C., Baccus, J. R., & Palmer, M. (2006a). Self-criticism and self-warmth: An imagery study exploring their relation to depression. *Journal of Cognitive Psychotherapy, 20,* 183–200. https://doi.org/10.1891/088983906780639817.

Gilbert, P., Catarino, F., Duarte, C., Matos, M., Kolts, R., Stubbs, J., et al. (2017). The development of compassionate engagement and action scales for self and others. *Journal of Compassionate Health Care*. https://doi.org/10.1186/s40639-017-0033-3.

Gilbert, P., Cheung, M., Irons, C., & McEwan, K. (2005). An exploration into depression-focused and anger-focused rumination in relation to depression in a student population. *Behavioural and Cognitive Psychotherapy*, *33*, 273–283. https://doi.org/10.1017/S1352465804002048.

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.

Gilbert, P., Durrant, R., & McEwan, K. (2006b). Investigating relationships between perfectionism, forms and functions of self-criticism, and sensitivity to put-down. *Personality and Individual Differences, 41,* 1299–1308. https://doi.org/10.1016/j.paid.2006.05.004.

Gilbert, P., McEwan, K., Gibbons, L., Chotai, S., Duarte, J., & Matos, M. (2012). Fears of compassion and happiness in relation to alexithymia, mindfulness and self-criticism. *Psychology and Psychotherapy: Theory, Research and Practice, 8,* 374–390. https://doi.org/10.1111/j.2044-8341.2011.02046.x.

Gilbert, P., & Miles, J. N. V. (2000). Sensitivity to social put-down: Its relationship to perceptions of social rank, shame, social anxiety, depression, anger and self-other blame. *Personality and Individual Differences, 29,* 757–774. https://doi.org/10.1016/S0191-8869(99)00230-5.

Halamová, J., Kanovský, M., Gilbert, P., Kupeli, N., Troop, N., Zuroff, D., et al. (2018). The factor structure of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale in thirteen populations. *Journal of Psychopathology and Behavioral Assessment*, *40*(4), 736–751. https://doi.org/10.1007/s10862-018-9681-7.

Halamová, J., Kanovský, M., & Pacúchová, M. (2017). Robust psychometric analysis and factor structure of the Forms of Self–criticizing/Attacking and Self–reassuring Scale. *Československá psychologie., 61*(4), 331–349.

Heine, S. J., & Hamamura, T. (2007). In search of East Asian self- enhancement. *Personality and Social Psychology Review, 11,* 4–27. https://doi.org/10.1177/1088868306294587.

Heine, S. J., Takata, T., & Lehman, D. R. (2000). Beyond self-presentation: Evidence for self-criticism among Japanese. *Personality and Social Psychology, 26,* 71–78. https://doi.org/10.1177/0146167200261007.

Hermanto, N., & Zuroff, D. C. (2016). The social mentality theory of self-compassion and self-reassurance: The interactive effect of care-seeking and caregiving. *The Journal of Social Psychology, 156*(5), 523–535. https://doi.org/10.1080/00224545.2015.1135779.

Hermanto, N., & Zuroff, D. C. (2017). Experimentally enhancing self-compassion: Moderating effects of trait care-seeking and perceived stress. *The Journal of Positive Psychology., 1,* 4. https://doi.org/10.1080/17439760.2017.1365162.

Hermanto, N., Zuroff, D. C., Kopala-Sibley, D. C., Kelly, A. C., Matos, M., Gilbert, P., et al. (2016). Ability to receive compassion from others buffers the depressogenic effect of self-criticism: A cross-cultural multi-study analysis. *Personality and Individual Differences, 98,* 324–332. https://doi.org/10.1016/j.paid.2016.04.055.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counselling Psychology., 38,* 139–149. https://doi.org/10.1037/0022-0167.38.2.139.

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.

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.

Kirby, J. N. (2016). *Compassion interventions: The programs, the evidence, and implications for research and practice*. Psychology and Psychotherapy: Theory, Research and Practice. https://doi.org/10.1111/papt.

Kitayama S. (2016). The Collective Construction of the Self: Culture, Brain, and Genes. In Edited by Robert J. Sternberg, Susan T. Fiske, Donald J. Foss, Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about their Most Important Contributions Cambridge University Press, p. 400-404. https://doi.org/10.1017/cbo9781316422250.086.

Kitayama, S., Markus, H. R., Matsumoto, H., & Norasakkunkit, V. (1997). Individual and collective processes in the construction of the self: Self-enhancement in the United States and self-criticism in Japan. *Journal of Personality and Social Psychology, 72,* 1245–1267. https://doi.org/10.1037/0022-3514.72.6.1245.

Krieger, T., Martig, D. S., van den Brink, E., & Berger, T. (2016). Working on self-compassion online: A proof of concept and feasibility study. *Internet Interventions, 6,* 64–70. https://doi.org/10.1016/j.invent.2016.10.001.

Kupeli, N., Chilcot, J., Schmidt, U. H., Campbell, I. C., & Troop, N. A. (2013). A confirmatory factor analysis and validation of the Forms of self-criticism/reassurance scale. *British Journal of Clinical Psychology, 52*(1), 12–25. https://doi.org/10.1111/j.2044-8260.2012.02042.x.

Kupeli, N., Norton, S., Chilcot, J., Campbell, I. C., Schmidt, U. H., & Troop, N. A. (2017). Affect systems, changes in body mass index, disordered eating and stress: an 18-month longitudinal study in women. *Health Psychology and Behavioral Medicine, 5,* 214–228. https://doi.org/10.1080/21642850.2017.1316667.

Lau, A. S., Chang, D. F., & Okazaki, S. (2010). Methodological challenges in treatment outcome research with ethnic minorities. *Cultural Diversity and Ethnic Minority Psychology, 16,* 573–580. https://doi.org/10.1037/a0021371.

Lekberg, A. & Wester, S. (2012). Självkritik eller självuppmuntran? Hur vi beter oss mot oss själva vid motgångar: En utprövning av instrumenten FSCRS och FSCS. (Master thesis) Lunds University, Lund, Sweden. Retrieved March 11, 2017 from https://lup.lub.lu.se/student-papers/search/publication/2827165.

Luyten, P., & Blatt, S. J. (2013). Interpersonal relatedness and self definition in normal and disrupted personality development: Retrospect and prospect. *American Psychologist, 68,* 172–183. https://doi.org/10.1037/a0032243.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224–253. https://doi.org/10.1037/0033-295X.98.2.224.

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.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Petrocchi, N., & Couyoumdjian, A. (2016). The impact of gratitude on depression and anxiety: the mediating role of criticizing, attacking, and reassuring the self. *Self and Identity, 15*(2), 191–205. https://doi.org/10.1080/15298868.2015.1095794.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria. Retrieved September 8, 2017 from https://www.R-project.org/.

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.

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.

Richter, A., Gilbert, P., & McEwan, K. (2009). Development of an early memories of warmth and safeness scale and its relationship to psychopathology. *Psychology and Psychotherapy: Theory, Research and Practice, 82*(2), 171–184. https://doi.org/10.1348/147608308X395213.

Robitzsch, A. (2018). sirt: Supplementary item response theory models. R package version 3.4-4. Retrieved July 10, 2018 from https://CRAN.R-project.org/package=sirt.

Rockliff, H., Karl, A., McEwan, K., Gilbert, J., Matos, M., & Gilbert, P. (2011). Effects of intranasal oxytocin on 'compassion focused imagery'. *Emotion, 11,* 1388–1396. https://doi.org/10.1037/a0023861.

Shahar, B., Carlin, E. R., Engle, D. E., Hegde, J., Szepsenwol, 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.

Shahar, B., Doron, G., & Szepsenwol, O. (2015). Childhood maltreatment, shame-proneness, and self-criticism in social anxiety disorder: A sequential mediational model. *Clinical Psychology and Psychotherapy, 22,* 570–579. https://doi.org/10.1002/cpp.1918.

Smart, L. M., Peters, J. R., & Baer, R. A. (2016). Development and validation of a measure of self-critical rumination. *Assessment, 23*(3), 321–332. https://doi.org/10.1177/1073191115573300.

Sommers-Spijkerman, M. P. J., Trompetter, H. R., ten Klooster, P. M., Schreurs, K. M. G., Gilbert, P., & Bohlmeijer, E. T. (2018). Development and validation of the forms of self-criticising/attacking and self-reassuring scale—short form. *Psychological Assessment*, *30*, 729–743. https://doi.org/10.1037/pas0000514.

Stinckens, N., Lietaer, G., & Leijssen, M. (2013a). Working with the inner critic: Process features and pathways to change. *Person-Centered & Experiential Psychotherapies, 12*(1), 59–78. https://doi.org/10.1080/14779757.2013.767747.

Stinckens, N., Lietaer, G., & Leijssen, M. (2013b). Working with the inner critic: Therapeutic approach. *Person-Centered & Experiential Psychotherapies, 12*(2), 141–156. https://doi.org/10.1080/14779757.2013.767751.

Thompson, R., & Zuroff, C. (2004). The Levels of Self-Criticism Scale: comparative self-criticism and internalized self-criticism. *Personality and Individual Differences, 36*(2), 419–430. https://doi.org/10.1016/S0191-8869(03)00106-5.

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.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*(2), 231–251. https://doi.org/10.1037/1082-989X.3.2.231.

Yamaguchi, A., Kim, M. S., & Akutsu, S. (2014). The effects of self-construals, self-criticism, and self-compassion on depressive symptoms. *Personality and Individual Differences, 68,* 65–70. https://doi.org/10.1016/j.paid.2014.03.013.

Yu, F. Y. (2013).The relationship among self-criticism, self compassion, rumination response style, and depression. Unpublished M.A. thesis, Chung Yuan Christian University, Taiwan.

Zuroff, D. C., Mongrain, M., & Santor, D. A. (2004). Conceptualizing and measuring personality vulnerability to depression: Comment on Coyne and Whiffen (1995). *Psychological Bulletin, 130*(3), 489–511. https://doi.org/10.1037/0033-2909.130.3.

Zuroff, D. C., Sadikaj, G., Kelly, A. C., & Leybman, M. J. (2016). Conceptualizing and measuring self-criticism as both a personality trait and a personality state. *Journal of Personality Assessment, 98*(1), 14–21. https://doi.org/10.1080/00223891.2015.1044604.

Affiliations

Júlia Halamová[1] · Martin Kanovský[2] · Paul Gilbert[3] · Nicholas A. Troop[4] · David C. Zuroff[5] · Nicola Petrocchi[6] · Nicola Hermanto[5] · Tobias Krieger[7] · James N. Kirby[8] · Kenichi Asano[9] · Marcela Matos[10] · FuYa Yu[11] · Marion Sommers-Spijkerman[12] · Ben Shahar[13] · Jaskaran Basran[3] · Nuriye Kupeli[14]

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

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

3 Centre for Compassion Research and Training, College of Health and Social Care Research Centre, University of Derby, School of Sciences, Derby, UK

4 Department of Psychology and Sports Sciences, School of Life and Medical Sciences, University of Hertfordshire, Hatfield, Hertfordshire, UK

5 Department of Psychology, McGill University, Montréal, QC, Canada

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

7 Clinical Psychology and Psychotherapy, University of Bern, Bern, Switzerland

8 The School of Psychology, The University of Queensland, Brisbane, Australia

9 Department of Psychological Counseling, Faculty of Human Sciences, Mejiro University, Tokyo, Japan

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

11 Student Counseling Center K-12 Education Administration, Ministry of Education, Yilan City, Taiwan

12 Centre for EHealth and Wellbeing Research, University of Twente, Enschede, The Netherlands

13 Paul Baerwald School of Social Work and Social Welfare, Hebrew University of Jerusalem, Jerusalem, Israel

14 Marie Curie Palliative Care Research Department, University College London, London, UK