A multilevel multidimensional finite mixture item response model of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale to cluster respondents and countries

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

[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 Psychology, McGill University, Montréal, Quebec, Canada
[4] Department of Psychology and Sports Sciences, School of Life and Medical Sciences, University of Hertfordshire, Hatfield, Hertfordshire, United Kingdom
[5] Centre for Compassion Research and Training, College of Health and Social Care Research Centre, University of Derby, School of Sciences, Derby, United Kingdom
[6] Paul Baerwald School of Social Work and Social Welfare, Hebrew University of Jerusalem, Jerusalem, Israel
[7] Department of Economics and Social Sciences, John Cabot University, Rome, Italy
[8] Clinical Psychology and Psychotherapy, University of Bern, Bern, Switzerland
[9] The School of Psychology, The University of Queensland, Australia
[10] Department of Psychological Counseling, Mejiro University, Tokyo, Japan
[11] Cognitive and Behavioural Centre for Research and Intervention, University of Coimbra, Coimbra, Portugal
[12] Student counseling center K-12 Education Administration, Ministry of Education, Taiwan
[13] Marie Curie Palliative Care Research Department, University College London, London, United Kingdom

* Corresponding author:
Julia Halamová
Institute of Applied Psychology
Faculty of Social and Economic Sciences
Comenius University in Bratislava
Mlynské luhy 4, 821 05 Bratislava, Slovakia,
E-mail: julia.halamova@gmail.com
Phone: 00421908604141

A multilevel multidimensional finite mixture item response model of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale to cluster respondents and countries

Summary

Objectives: The aim of this study was to test the multilevel multidimensional finite mixture item response model of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale (FSCRS) to cluster respondents and countries from thirteen samples ($N = 7,714$) and from twelve countries. The practical goal was to learn how many discrete classes there are on the level of individuals (i.e., how many cut-offs are to be used) and countries (i.e., the magnitude of similarities and dissimilarities among them).

Methods: We employed the multilevel multidimensional finite mixture approach which is based on an extended class of multidimensional latent class Item Response Theory models. Individuals and countries are partitioned into discrete latent classes with different levels of self-criticism and self-reassurance, taking into account at the same time the multidimensional structure of the construct. This approach was applied to the analysis of the relationships between observed characteristics and latent trait at different levels (individuals and countries), and across different dimensions using the three-dimensional measure of the FSCRS.

Results: Results showed that respondents' scores were dependent on unobserved (latent class) individual and country membership, multidimensional structure of the instrument, and justified the use of a multilevel multidimensional finite mixture item response model in the comparative psychological assessment of individuals and countries.

Conclusions: Latent class analysis of the FSCRS showed that individual participants and countries could be divided into discrete classes. Along with the previous findings that the FSCRS is psychometrically robust we can recommend using the FSCRS for measuring self-criticism.

Keywords: latent class model; multilevel multidimensional item response model; self-criticism

A multilevel multidimensional finite mixture item response model of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale to cluster respondents and countries

Introduction

Ever since Freud (1917) identified anger at the self as a crucial factor in depression the link between psychopathology and self-criticism has been subject to much theorising and research. In a major systematic review Werner, Tibubos, Rohrmann, and Reiess (2019) outlined a number of different models for self-criticism and its measurement that include both self-report and qualitative studies. They highlight that self-criticism is transdiagnostic vulnerability factor for mental health problems such as eating disorders, depression, suicidality, anxiety, psychotic symptoms, and interpersonal problems (Werner et al., 2019). Self-criticism also affects susceptibility to and persistence of psychopathology (Bergner, 1995; Blatt & Zuroff, 1992; Falconer, King, & Brewin, 2016) and stress (Kupeli, Norton, Chilcot, Campbell, Schmidt, & Troop, 2017) and influences the response to medical and psychological interventions and treatments (Blatt & Zuroff, 2005; Bulmash, Harkness, Stewart, & Bagby, 2009; Shahar et al., 2015). Löw, Schauenburg, and Dinger (2020) offered a systematic review of self-criticism and

psychotherapy outcome showing that the intensity of self-criticism is linked to poor outcome, highlighting the importance of improving psychotherapy for these individuals. For these and other reasons it is important to explore the psychometric properties of measures of self-criticism.

There exist a number of self-report measures assessing self-criticism including: the Depressive Experiences Questionnaire (DEQ; Blatt, D'Afflitti, & Quinlan, 1976), the Levels of Self-Criticism Scale (LOSC; Thompson & Zuroff, 2004), the Forms of Self-Criticising/Attacking & Self-Reassuring Scale (FSCRS; Gilbert et al., 2004), the Self-Critical Rumination Scale (Smart, Peters, & Baer, 2016), and a scale assessing situational state self-criticism, the Self-Compassion and Self-Criticism Scales (SCCS; Falconer, King, & Brewin, 2015). In a systematic review of measures of five scales of self-criticism, Rose and Rimes (2018) noted that many studies lacked methodological rigour but suggested that two scales, the Self-critical Rumination Scale (Smart Peters, Baer, 2015) and the FSCRS (Gilbert et al., 2004) were the most robust in terms of psychometric properties.

The FSCRS is unique in distinguishing between forms and functions of self-criticism. This is important because many psychological processes, e.g., anger, anxiety or caring behaviour, can have different forms and functions and will relate to social difficulties and mental health problems in different ways. Some individuals are self-critical because they experience failure and feel they should and could do better, which may be linked to perfectionism (Curran & Hill, 2019; Shahar, 2015), where others are not so much trying to improve but have a more self-hating attitude to self and wanting to get rid of aspects of the self (Gilbert et al., 2004). In a study of psychiatric patients, Castilho, Pinto-Gouveia, and Durate (2016) found that concerns of being inadequate and self-hating forms of self-criticism linked to psychopathology in different ways, with self-hating being particularly linked to shame.

Given the importance of these core processes underpinning vulnerability to mental health and other difficulties, a number of authors have argued there is a need for closer inspection of self-criticism across cultures (Lau, Chang, & Okazaki, 2010; Luyten & Blatt, 2013). Indeed, little is known about cross-cultural aspects of self-criticism. The main obstacle to performing cross-cultural research on self-criticism is the lack of invariant measurement tools with a stable factor structure. Therefore, any findings achieved without previous testing of scales' measurement invariance must be considered with caution. A measure of self-criticism which has been evaluated across cultures is the FSCRS which, as noted above, has been subjected to meta-analytic reviews. A number of studies have reported on the invariance (Halamová et al., 2019) and factor structure (Halamová et al., 2018) of the FSCRS using data from thirteen international samples. Thus, it is necessary to examine this scale further so more cross-cultural research on self-criticism is possible.

In order to advance knowledge of the performance of the FCSRS across cultures, we applied a multilevel mixture model to a multinational sample. Evidence of good fit for such a model would yield a valuable practical outcome – the knowledge of how many discrete classes there on the level of individuals and countries. For individuals, we will know how to discretize sum scores (i.e.,. how many cut-offs are to be used), and for countries, we can obtain important information on the magnitude of similarities and dissimilarities among them. An alternative modelling approach, the widely used multi-group structural equation model, faces well-known problems (Kim, Cao, Wang, & Nguyen, 2017) of how to establish scalar measurement invariance which is necessary for a meaningful comparison of latent means (or more precisely, to display acceptable fit of such a model with the data). On the other hand, a multilevel finite mixture model is very useful when the violation of exact invariance and potential clustering of groups is assumed due to measurement non-invariance, factor mean heterogeneity, or both (Kim et al., 2017). Because a multilevel finite mixture model estimates far fewer parameters than a

multi-group structural equation model, it is also more parsimonious (Kim et al., 2017). One of the main advantages of the latent class analysis model is the fact that countries and persons are clustered in classes, and that within a class they are more similar than between classes: this method is able to resolve the problems of measurement non-invariance and factor mean heterogeneity in one unified approach.

All previous studies have assumed that self-criticism is a continuous latent ability (characteristic), and to our knowledge, there is no study investigating potential differences and heterogeneities among groups of respondents. Measurement invariance studies focus on the comparison of latent means across countries, but they cannot reveal in principle how many respondents in each country contribute to these differences. Ignoring this fact could lead to failing to notice that a small number of highly self-critical respondents could have approximately same impact on the latent factor mean as a large number of moderately self-critical respondents. The assumption of a homogenous population with a continuous distribution of the latent ability that differs from another homogenous population with a continuous distribution of the latent ability with a higher (or lower) level might be untenable in practice. Latent class analysis could provide a more fine-grained picture how self-criticism is distributed within and across countries taking into account potential heterogeneity.

The aim of the research study

The aim of this study was, therefore, to test the multilevel multidimensional finite mixture item response model of the FSCRS to cluster respondents and countries from thirteen samples ($N = 7,714$) and from twelve countries.

Our research hypotheses were as follows:
(1) We hypothesized that there would be at least five latent classes on the individual level. Our hypothesis was that they would approximately cover respondents with very low, moderately

low, neutral, moderately high, and very high levels of self-criticism. On the other hand, we did not expect more than nine latent classes on the individual level; having more than nine latent classes would erase the main advantage of latent class analysis, and a standard multilevel IRT model with the continuous distribution of latent ability would be more relevant.

(2) We hypothesized that there would be at least three latent classes on the country level. Our hypothesis was that they will approximately cover countries with low, medium, and high overall levels of self-criticism. On the other hand, we did not expect more than five latent classes on the country level, as having more than five latent classes for 13 units would provide no meaningful comparison.

(3) We hypothesized that the IRT model with different discriminations and different difficulties (graded response model) would better fit the data than more constrained one-parameter and rating scale models.

Materials and Methods

Research sampling procedure

Various methods were used to identify data on the FSCRS. Firstly we used Google Scholar to identify published studies by searching terms such as "the forms of self-criticising/attacking & self-reassuring scale" or "FSCRS". We contacted the authors of studies that reported data on the FSCRS and comprised a nonclinical sample of 215 or more participants In addition, we searched the Compassionate Mind website (https://compassionatemind.co.uk/uploads/files/research-register-for-website.pdf) for unpublished research studies. Out of approximately 40 emails of invitations for collaboration, we obtained data for 13 non-clinical samples (Authors, 2018). Because self-criticism is

clinically important issue, it is crucial to distinguish between clinical and non-clinical samples. This article is the first study clustering respondents and countries from non-clinical samples. In the subsequent second article we will compare clinical samples from different countries.

Procedures for different samples

To date, the FSCRS has been translated into 11 languages. The current study presenys data for eight versions of the FSCRS includingfive distinct English language samples from four different countries; Australia (N = 319; Blinded for review), Canada (N = 383; Hermanto & Zuroff, 2016; Hermanto & Zuroff, 2017; Zuroff, Sadikaj, Kelly, & Leybman, 2016), the United Kingdom 1 (N = 1,570; Kupeli et al., 2013), the United Kingdom 2 (N = 883; Baião, Gilbert, McEwan, & Carvalho, 2015; Gilbert, Baldwin, Irons, Baccus, & Palmer, 2006; Gilbert, Durrant, & McEwan, 2006; Gilbert & Miles, 2000; Gilbert et al., 2002, 2004, 2005, 2012), and USA (N = 331; Gilbert et al., 2017). There were also data from studies using versions of the FSCRS translated to seven other languages including Chinese (N = 417; Yu, 2013), Dutch (N = 360; Sommers-Spijkerman, Trompetter, ten Klooster, Schreurs, Gilbert, & Bohlmeijer, 2017), German (N = 230; Krieger, Martig, van den Brink, & Berger, 2016; Krieger, personal communication), Hebrew (N = 476; Shahar, Doron & Szepsenwol, 2015; Shahar, personal communication), Italian (N = 389; Petrocchi & Couyoumdjian, 2016), Japanese (N = 264; Kenichi, personal communication), Portuguese (N = 764; Gilbert et al., 2017), and Slovak (N = 1,326; Halamová, Kanovský, & Pacúchová, 2017). Overall, data from 13 distinct nonclincal samples (N = 7,714) was used to test the FSCRS. All studies adherred 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.

Research instrument

The Forms of Self-criticising/Attacking & Self-Reassuring Scale (FSCRS; Gilbert et al., 2004) is a 22-item instrument, which was created to measure the forms and functions of self-criticism and contrast it with self-reassurance when things do not go as expected or hoped for. Participants use a 5-point Likert scale (1 = not at all like me; 5 = extremely like me). The first of the three factors, Inadequate Self (IS), consists of nine items that capture experiences of a personal sense of inadequacy, and failure, for example: "There is a part of me that feels I am not good enough." The second factor, Hated Self (HS), has five items. It assesses destructive self-hatred, contempt, disgust, and desires to harm oneself. Items include: "I have a sense of disgust with myself." The contrasting third factor, Reassured Self (RS), has seven items, and captures the capacity for self-soothing, encouragement, support, and validation while experiencing negative events. Items include "I can still feel lovable and acceptable.". In general criticising oneself or being inadequate is linked to competitive drive and fear of inferiority (Gilbert et al., 2004). Self-hatred on the other hand is more pathogenic, typically linked to problems of early life trauma including abuse. Here the individual is not trying to improve themselves but to get rid of parts of themselves (Gilbert et al., 2004). We have to add that all items in the sub-domain self-reassurance were reverse scored to assure that this instrument will measure self-criticism.

Statistical models and data analyses

This section is organized as follows. First, we will define a latent class (mixture) model for polytomous item responses with the Item Response Theory (IRT) parameterization (Bacci, Bartolucci & Gnaldi, 2014; Bartolucci, 2007). Second, we will present the extension and specify the model taking into account the multidimensional structure of the data (3 dimensions), and subsequently, we will present the extension and specify the multidimensional model taking

into account the multilevel structure of the data (11 countries). Then we will provide criteria for model selection. Finally, we will briefly describe steps in our statistical analysis. Since our aim is to provide a practical guide rather than a very detailed mathematical specification, we will limit our explanation only to basic formulations leaving the exact mathematical expressions to the Electronic Supplementary Material (ESM).

(1) Unidimensional latent class IRT models.

Following Bacci, Batolucci & Gnaldi (2014), let X_j denote the observed response variable for the j-th item of the analysed questionnaire, with $j = 1, \ldots, r$. This variable has l_j levels, indexed from 0 to $l_j - 1$. In our case, we have 22 items of the FSCRS; each item is scored from 0 to 4 and therefore there are 5 levels, indexed from 0 to 4. In the unidimensional case, we define the probability that a respondent with latent trait (ability) level θ responds by category x to this particular item as:

$$\lambda_{jx}(\theta) = p(X_j = x | \Theta = \theta), \ x = 0, \ldots, l_j - 1 \qquad (1)$$

Moreover, we define $\lambda_j(\theta)$ as the probability vector $(\lambda_{j0}(\theta), \ldots, \lambda_{j,lj-1}(\theta))'$. Now, we can specify the IRT model for polytomous responses:

$$g_x[\lambda_j(\theta)] = \gamma_j(\theta - \beta_{jx}), \ j = 1, \ldots, r, \ x = 1, \ldots, l_j - 1 \qquad (2)$$

where $g_x(\cdot)$ is a specific link function for the category x, and γ_j and β_{jx} are IRT item parameters, usually called discrimination indices and difficulty levels. We will observe if the assumption that properly defined constraints on these parameters will result in differently parameterized IRT models.

We can have three different kinds of IRT specifications: a) the specific link function; b) constraints on discrimination parameters; c) constraints on difficulty parameters.

a) Type of link function. As far as the link function is concerned, the most widely used IRT models are the graded response model with global logits (Bacci, Batolucci & Gnaldi, 2014; Samejima 1969), the generalized partial credit model with local logits (Bacci, Batolucci & Gnaldi, 2014; Muraki, 1992), and the continuation ratio model with sequential logits (Bacci, Batolucci & Gnaldi, 2014; Fienberg, 1980). See Electronic Supplementary Material, part ESM 2 for their mathematical formulations. Our instrument (FSCRS) is assumed to measure a continuous latent ability (self-criticism); therefore we will adopt the graded response (cumulative) model. It is of little interest to estimate the probability of answering adjacent categories, let alone the probability of a sequential process. The conclusion is that only the graded response models will be analyzed in this paper.

b) Constraints on discrimination parameters. We can specify two different types of IRT models based on constrained or unconstrained discrimination parameters. First, each item might discriminate differently from the other items, and all γ_j parameters are freely estimated. Second, all the items discriminate in the same way, that is, $\gamma_j = 1$ for all items (the one-parameter graded response model).

c) Constraints on item difficulty parameters. Again, we can specify two different types of IRT models based on constrained or unconstrained versions. First, the difficulty parameters β_{jx} are unconstrained and freely estimated. Second, the difficulty parameters β_{jx} are constrained so that the distance between difficulty levels from category to category is the same for all items – τ_x is the (same) difficulty of response category x for all j (the rating scale graded response model).

We can combine different constraints to obtain IRT models with very different parameterizations. For clarity, we reproduce in the Electronic Supplementary Material ESM 1a (slightly modified) Table 1 from Bacci, Batolucci, and Gnaldi (2014).

(2) Multidimensional latent class IRT models

This unidimensional IRT model for polytomous responses can be easily extended to the multidimensional setting. Following Bacci, Batolucci, and Gnaldi (2014, chap. 3.1), let us define s as the number of different latent traits measured by the items, $\boldsymbol{\Theta} = (\Theta_1, \dots \Theta_s)'$ as a vector of latent variables corresponding to these latent traits, and $\theta = (\theta_1, \dots, \theta_s)'$ as one of its possible realizations. This random vector $\boldsymbol{\Theta}$ is assumed to have a discrete distribution with k support points, denoted by $\xi_1, \dots \xi_k$, and probabilities $\pi_1, \dots, \pi_k$, with $\pi_c = p(\boldsymbol{\Theta} = \xi_c)$. In addition, let us define δ_{jd} to be a dummy variable equal to 1 if item j measures a latent trait of type d and to 0 otherwise, with $j = 1, \dots, r$ and $d = 1, \dots, s$. Now we can redefine the equation (2) above as follows:

$$g_x[\lambda_j(\theta)] = \gamma_j(\Sigma_{d\text{-}1}^s \, \delta_{jd} \, \theta_d - \beta_{jx}), \, j = 1, \dots, r, \, x = 1, \dots, l_j - 1 \qquad (3)$$

Note that from the discreteness of the distribution of the random vector $\boldsymbol{\Theta}$, it immediately follows that the manifest distribution of $X = (X_1, \dots, X_r)'$ for all subjects in the c-th latent class is:

$$p(\boldsymbol{x}) = p(X = \boldsymbol{x}) = \Sigma_{c=1}^k p(X = \boldsymbol{x}| \, \boldsymbol{\Theta} = \xi_c) \, \pi_c \qquad (4)$$

Of course, all possibilities of different IRT parameterizations (Table 1 above) apply for multidimensional IRT models as well, including the option to select a different link function. For the exact formulation of the multidimensional IRT model with global logits (the multidimensional graded response model), see Electronic Supplementary Material ESM 2a. We must add that some constraints on the parameters must be placed to ensure identifiability of models. It is required that one discriminant index is equal to 1 for each latent trait, and one difficulty parameter is equal to 0 for each latent trait.

(3) Multidimensional multilevel latent class IRT models

The multilevel generalization assumes that responses are a multivariate dependent variable and their hierarchical structure is expressed by respondents (units on first level) who are nested in countries (units on second level). This second-level is in itself a discrete latent variable. To comply with previous work in this field (Bacci & Gnaldi, 2015) and to avoid any confusion concerning various indices, we rename and redefine all indices as follows (Bacci & Gnaldi, 2015, p. 931): j is the generic item ($j = 1, \ldots, J$), i is a respondent or unit on first level of the latent class k ($i = 1, \ldots, I$; $k = 1, \ldots, K$), h is a country (or any group) or unit on second level of the latent class c ($h = 1, \ldots, H$; $c = 1, \ldots, C$), and Y_{ji} is the response to item j of respondent i which might assume L ordered categories ($l = 0, \ldots, L-1$). We will also define the D-dimensional vector $\boldsymbol{\Theta}^{(a)} = (\Theta^{(a)}_1, \ldots, \Theta^{(a)}_D)$ of latent variables, one of its possible realizations being $\boldsymbol{\theta}^{(a)} = (\theta^{(a)}_1, \ldots, \theta^{(a)}_D)$, with $a = 1$ for units on first level units and $a = 2$ for units on second level units. $\boldsymbol{\Theta}^{(1)}$ has support points $\xi^{(1)}_k = (\xi^{(1)}_{k1}, \ldots, \xi^{(1)}_{kD})'$ and corresponding weight $\pi^{(1)}_k$, with $k = 1, \ldots, K(\pi^{(1)}_k > 0, \Sigma_k \pi^{(1)}_k = 1)$, and $\boldsymbol{\Theta}^{(2)}$ has support points $\xi^{(2)}_c = (\xi^{(2)}_{c1}, \ldots, \xi^{(2)}_{cD})'$, corresponding weight $\pi^{(2)}_c$, with $c = 1, \ldots, C(\pi^{(2)}_c > 0, \Sigma_c \pi^{(2)}_c = 1)$. Finally, $\lambda_j(\boldsymbol{\theta}) = (\lambda_{j0}(\boldsymbol{\theta}), \ldots, \lambda_{j,L-1}(\boldsymbol{\theta}))$, with $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}\boldsymbol{\theta}^{(2)})$ and $\lambda_{jl}(\boldsymbol{\theta}) = p(Y_{ji} = l|\boldsymbol{\theta})$. Now, the multidimensional multilevel IRT model is framed as follows:

$$g_y(\lambda_j(\boldsymbol{\theta})) = \gamma_j \left(\Sigma^D_{d=1} \delta_{jd}(\xi^{(1)}_{kd} + \xi^{(2)}_{cd}) - \beta_{il}\right) \qquad (5)$$

where $j = 1, \ldots, J$; $k = 1, \ldots, K$; $c = 1, \ldots, C$, γ_j is the discrimination index for item j, β_{jl} is the difficulty of item j and cut-off point l ($l = 1, \ldots, L-1$), and δ_{jd} is an indicator variable assuming value 1 if item j measures dimension (or latent trait) d ($d = 1, \ldots, D$), and 0 otherwise. Additionally, in equation (5), $\xi^{(1)}_{kd}$ and $\xi^{(2)}_{cd}$ are the random effects of first- and second-level, or more specifically $\xi^{(1)}_{kd}$ represents the value of the latent variable corresponding to latent trait d for a respondent who belongs to first-level latent class k, whereas $\xi^{(2)}_{cd}$ is the value of the latent variable corresponding to latent trait d for a country (group) belonging to second-level latent class c.

Each latent variable has a discrete distribution, characterised by a finite number of support points (or mixture components) ξ and related mass probabilities π. Each support point ξ identifies a latent class of individuals; individuals belonging to the same latent class share the same level of the latent variable. Standardized support points are directly comparable with particular locations in a continuous latent ability from a standard IRT model.

We should add that, in this formula, weights $\pi^{(1)}_k$ and $\pi^{(2)}_c$ do not depend on observable covariates, and therefore $\pi^{(1)}_k = p(\Theta^{(1)} = \xi^{(1)}_k)$ and $\pi^{(2)}_c = p(\Theta^{(2)} = \xi^{(2)}_c)$. This means that respondents (or countries) come from K (or C) latent classes which are homogenous in terms of the characteristics measured by the questionnaire (self-criticism). This model could be extended to include covariates, but we will not pursue this issue further here (see Bacci & Gnaldi, 2015 for such an extension).

For the sake of brevity, we will not provide here either a matrix formulation, or mathematical description of the class of estimation procedures employed to fit models, expected maximised (EM) estimation algorithms (Bacci, Batolucci & Gnaldi, 2014; Bartolucci, 2007; Vermunt, 2003, 2008 for details).

(4) Criteria of model selection

The key issue in any latent class analysis is to select the best fitting model, especially the number of latent classes on both levels. Various methods have been proposed, all of them based on the log likelihood function (for a review see McLachlan & Peel, 2000). They can be broadly divided into two categories: parsimony-based (information criteria) or testing-based (likelihood-ratio tests). We will not discuss all their advantages and disadvantages, but rather focus on summarising the most frequently used approach to justify our procedure of model selection.

Likelihood-ratio tests are widely used, but their main problem for finite mixture models is that the parameter vector lies on the boundary of the parameter space under the null hypothesis. The usual asymptotic null distribution of the likelihood-ratio test statistics is therefore not valid (Melnykov & Maitra, 2010). A recently proposed procedure (Maitra & Melnykov, 2010) to overcome this issue is not yet implemented in any available statistical software; therefore we refrain from using the likelihood-ratio testing. On the other hand, of the two most frequently used information criteria, Akaike Information Criterion (AIC) typically substantially overestimates the number of classes (Melnykov & Maitra, 2010, p. 88), while Bayes-Schwarz Information Criterion (BIC) has been shown to demonstrate good performance (Keribin, 2000). However, these come with two caveats: first, the BIC tends to underestimate the number of components of a model when sample sizes are small (Melnykov & Maitra, 2010, p. 88). Fortunately, the sample size used in this article is large enough to minimise this issue. Second, to obtain a meaningful comparison of model fit from one model structure to another, improvements in BIC should be substantive to provide strong evidence; Kass and Raftery (1996) recommend differences greater than 10. To summarise, we will use the BIC for model selection, but only differences greater than 10 will be regarded as constituting strong evidence in favour of the respective model.

(5) Statistical procedure

We used R package for the statistical analysis, multiLCIRT package (Bartolucci, Bacci, & Gnaldi, 2014).

(a) Descriptive statistics will be provided, namely distributions of the FSCRS item responses (raw percentage frequencies for each item category) and descriptives of raw scores of all three sub-dimensions (Inadequate Self, Reassured Self, Hated Self) of the FSCRS for each country.

(b) Selection of the optimal number k of latent classes on the level of respondents: Following recommendations in Bacci, Bartoucci and Gnaldi (2014, chap. 4.2.2), we will adopt the standard latent class model, characterized by one dimension for each item; therefore, no choice on the item parameterization is required. We note that we will not test the dimensionality of the FSCRS; it is known that this instrument is three-dimensional, and we will rely on this previous knowledge (Gilbert et al., 2004). Second, we will not test the fit of models with different logit link functions: we have justified theoretically the choice of the global logits (Graded Response Model). Our procedure will consist of fitting models with increasing numbers of latent classes and stopping this procedure when the value of BIC of the tested model is greater (and not lower) than the value of the previously tested model. We will also stop this procedure when the value of BIC of the tested model is less than 10 lower than the value of the previously tested model. In both cases, the previously tested model will be retained.

A crucial problem with standard latent class models is the multimodality of the likelihood function, i.e., their tendency to find a local, rather than global, maximum point. In order to avoid such an outcome, we will repeat the estimation process by randomly varying the starting values of the model parameters, in other words, repeating the procedure of testing as a whole with random starting values. Values of log-likelihood, number of parameters, and BIC will be reported, and the final model with the optimal number k of latent classes (the smallest BIC value) will be specified.

(c) Selection of the item discriminating and difficulty parameterization: This step consists of the choice of the possible constraints on the discriminating and difficulty parameters (see above, Table 1). Four different types of model (with k latent classes specified in the previous step) will be fitted by combining free or constrained γ_j parameters with free or constrained β_{jx} parameters. Again, the final model with the optimal IRT parameterization (the smallest BIC value) will be specified.

(d) Selection of the optimal number u of latent classes on the level of countries: Again, our procedure will consist of fitting three-dimensional models (parameterized in accordance with the best fitting model from the previous step) with increasing numbers of latent classes on the level of countries (starting with the model with an appropriate number of latent classes based on the level of respondents), and stopping this procedure when the value of BIC of the tested model will be greater (and not lower) than the value of the previously tested model. We will also stop this procedure when the value of BIC of the tested model is less than 10 lower than the value of the previously tested model. In both cases, the previously tested model will be retained.

(e) Interpretation and visualization of estimated parameters of the final model: We will report support points and weights (membership in latent classes) for each latent class, both on the level of respondents and countries. Note that to ensure the identifiability of models, one discriminant index was set to 1 for each latent trait, and one difficulty parameter was set to 0 for each latent trait. However, it is possible to express estimates of the model parameters under the alternative identifiability constraint; the latent abilities on the level of respondents could be standardized (with mean 0 and variance 1). As a consequence, all parameters are more easily and directly interpretable in the IRT context; for example, it would be possible to compare standardized support points of finite mixture models with parameters of classical IRT models with continuous latent abilities. Moreover, the standardized latent abilities on the level of respondents will enable us to obtain the distribution of the estimated average abilities for all latent classes on the level of countries; having the standardized latent abilities on the level of respondents, we can easily compute the mean values for each class of countries, taking into account the weights of each respondents' latent class. We will therefore report the standardized support points on the respondents' level, the estimated average abilities for all latent classes on the level of countries, together with their weights (membership in latent classes), and their

visualization. All IRT parameters for each item (standardized discrimination and difficulty indices) will also be reported.

Results

(1) Descriptives: after inspecting the distribution of the FSCRS item responses (i.e., raw percentage frequencies, see Table ESM 1b in the Electronic Supplementary Material) and raw scores of sub-dimensions of the FSCRS pooled across countries (see Table ESM 1c in the Electronic Supplementary Material), we can clearly see that responses are hugely skewed to the right (positive skewness), especially for the Hated Self sub-dimension. We note again that responses for the sub-dimension Reassured Self were reverse-scored. This result justifies the use of the logistic (non-linear item-response theory) method because the observed distribution cannot be approximated by a multivariate normal distribution.

(2) Selection of the optimal number k of latent classes on the level of respondents: after fitting a series of standard LC models for detecting the optimal number (k) of latent classes on the level of respondents, we can conclude that models with deterministic outputs suggest $k = 8$ (Table 1). However, after re-fitting these models with random inputs to rule out the possibility that the likelihood function has reached a local rather than global maximum, we had to modify such a conclusion; models with random starting values suggest $k = 7$. This outcome underlines the importance of careful checking of fitted models to avoid relying on biased results. We will therefore retain the model with 7 latent classes on the level of respondents.

Table 1 about here

(3) Selection of the item discriminating and difficulty parameterization: After fitting four three-dimensional models with different IRT parameterizations, it is clear that the most complex model (the Graded Response Model [GRM] with freely estimated discriminations and

unconstrained difficulties) fits the data better than its alternatives (Table 2). We will continue with this model in our subsequent testing.

Table 2 about here

(4) Selection of the optimal number u of latent classes on the level of countries: After fitting a series of three-dimensional GRM models with increasing number of latent classes on the level of countries, we can observe that the model with 5 latent classes is the best option (Table 3).

Table 6 about here

(5) Parameters of the final model: In Table 4, the values of standardized support points and weights (membership in latent classes) for each latent class on the level of respondents are reported (see previous section for the justification of the standardization procedure). These results reflect the positive skewness in responses; the membership in the first 3 latent classes is 45%, whereas the membership in the last 3 latent classes is only 37% (see Figure 1). This figure displays the position of support points for individual respondents on the scale of the latent ability (x-axis), and percentage of respondents in each latent class (y-axis). Furthermore, in the sub-dimension Hated Self, only 3 support points (and corresponding latent classes) have negative values of latent ability (self-criticism) in comparison to 4 support points in the sub-dimension Inadequate Self. This agrees with the theoretical assumption that the Inadequate Self is relatively sensitive to low levels of latent Inadequate Self, whereas the Hated Self is relatively more sensitive to high levels of Hated Self (with potential clinical implications). Support points (and latent classes) of the sub-dimension Hated Self include more highly self-critical

respondents than support points of the sub-dimension Inadequate Self. This is clearly reflected in IRT parameters as well (see Table ESM 1d in the Electronic Supplementary Material): difficulty indices (thresholds) for the sub-dimension Hated Self are clearly shifted to higher values of the latent ability (self-criticism) which means that endorsing the higher levels of observed responses requires higher levels of the latent ability. Note that responses in the sub-dimension Reassured Self were reverse-scored, so support points in this sub-dimension follow the same scaling.

Tables 4 about here

Figure 1 about here

Table 5 summarizes estimated average abilities, weights and membership for all latent classes on the level of countries. Also, on this level, the distribution is clearly positively skewed to the right see Figure 2. This figure displays the position of support points for countries on the scale of the latent ability (x-axis), and the percentage of respondents in each latent class (y-axis) on the level of countries. If we inspect the membership of a particular country, the less self-critical class consists of participants from Portugal and Israel (class 1), and the most self-critical classes are composed of participants from Taiwan (class 4) and Japan (class 5).

Table 5 about here

Figure 2 about here

Discussion

In this study we tested the multilevel multidimensional finite mixture item response model of the Forms of Self-Criticising/Attacking & Self-Reassuring Scale to cluster

respondents and countries. We analysed the relationships between observed characteristics and the latent trait at the different levels (individuals and countries), and across different dimensions of FSCRS.

Results showed that respondents' scores were dependent on unobserved (latent class) individual and countries membership and multidimensional structure of the instrument, and justified the use of a multilevel multidimensional finite mixture item response model in comparative psychological assessment. However, a very important question is if latent classes on the individual level and on the country level are meaningfully different classes, or are they merely an arbitrary discretization of a continuum which fits the data? This question is very difficult to answer without including covariates which could predict membership in latent classes (both on the individual and on country level), e.g., gender and age differences, social status, childhood experiences. In fact, we cannot speculate about the sources and causes of the heterogeneity without such data and models, but we do think that these latent classes are not arbitrary; patterns of responses which assign individuals into latent classes are systematic and are probably driven by yet unknown sets of social and psychological sources. Further research is required to resolve this problem.

According to the findings, there are some items of the FSCRS which discriminate better than others. Related to Hated self, items 10 ("I have a sense of disgust with myself") and 22 ("I do not like being me") have the best discriminations while items 15 ("I call myself names") discriminates less well. However, in terms of difficulties, all items from the Hated Self sub-scale are clearly more informative about highly self-critical participants compared with items form the Inadequate Self sub-scale. This may be partly because self-hatred represents a more pathogenic form of self-criticism and is linked to shame (Castilho et al., 2016).

Items from the Inadequate Self sub-scale provide balanced information across the whole range of the latent ability (self-criticism). Therefore, they are informative of participants with high, low as well as average levels of self-criticism. The best discriminating items from the Inadequate Self sub-scale are items 7 (I feel beaten down by my own self-critical thoughts) and 6 (There is a part of me that feels I am not good enough) and the worst is item 18 (I think I deserve my self-criticism).

All items from the Reassured Self sub-scale have low discriminations in comparison with items form the Inadequate and Hated Self sub-scales. The Reassured Self sub-scale has low discrimination specifically on the latent trait of (reversed) self-reassurance. This may be because the items in this sub-scale were reverse scored and the content of the items is positively phrased. It may also mean that self-reassurance is not the opposite of self-criticism on a unipolar continuum. Although self-reassurance can be an antidote for self-criticism it operates through completely different neurophysiological systems to those of self-criticism (Longe, Maratos, Gilbert, Evans, Volker, Rockliffe, & Rippon, 2010).

As high self-criticism is strongly linked to psychopathology (Werner et al., 2019), the distribution of self-criticism is positively skewed as various forms of psychopathology are. Consequently, the most self-critical people (class 6 and 7 combined is 18.9%) are less common in the general population than the least self-critical people (class 1 and 2 combined is 33.7%).

Similar to the findings of Halamová et al. (2019), participants from Japan are the most self-critical, followed by Taiwan, while the least self-critical participants are from Israel and Portugal. In addition to their findings, in this article we found that there are 5 classes of countries according to their level of self-criticism and 7 classes of participants according to their level of self-criticism. In the future these could be used for diagnostic purposes; at the level of individuals, each sub-score for the particular dimension could be discretized into 7 classes and

norms for individuals should be created and used taking into account this latent class structure. For example, it would be inadvisable to distinguish sub-scores from any sub-dimension into "highly self-critical", "average", or "low self-critical" individuals without taking into account weighting (see Table 7); e.g., "lowest self-critical group" (class 1) should contain only around 11% of individuals, and "highest self-critical" group (class 7) only around 5% of respondents. Secondly, we should be careful when comparing sum scores from different countries. If countries do not belong to the same latent class, such comparisons are not warranted and could be biased, because measurement invariance is not guaranteed.

Limitations

The main limitation of this research study is our sampling. Most of the samples were convenience samples, largely consisting of college students. Therefore, the results are not representative of the involved nationalities.

In addition, our model is unable to provide an explanation for why individuals belong to their respective latent classes. This important limitation is a consequence of the fact that our model did not include any covariates predicting membership in latent classes, for example, gender and age differences, social and economic status, childhood maltreatment, experience with meditation. However, further systematic international research with an unified set of covariates is required to answer questions concerning the sources of the cross-cultural heterogeneity of self-criticism. This would also be a meaningful future research direction with useful implications.

<div align="center">

Conclusion

</div>

In conclusion, the FSCRS is a well-designed self-report measure with good discrimination of highly self-critical participants. Therefore, we can recommend using the FSCRS for measuring self-criticism. From the practical point of view, Inadequate and Hated

Self are functionally different and certainly have different distributions. Our results go also in contradiction of the suggestion of some that Inadequate and Hated Self should or may be combined into a global self-criticism score (Halamová et al, 2018).

References

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of Multidimensional Latent Class IRT models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods,* 43(4), 787-800. doi: 10.1080/03610926.2013.827718

Bacci, S. & Gnaldi, M. (2015). A classification of university courses based on students' satisfaction: an application of a two-level mixture item response model. *Quality & Quantity,* 49(3), 927-940. doi: 10.1007/s11135-014-0101-0

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), 438-52. doi: 10.1111/papt.12049

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika,* 72:141–157. doi: 10.1007/s11336-005-1376-9

Bartolucci, F., Bacci, S., & Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis,* 71, 971-985. doi: 10.1080/03610926.2013.827718

Bergner, R.M (1995). *Pathological self-criticism: Assessment and treatment*. New York: Springer.

Blatt, S. J., D'Afflitti, J. P., & Quinlan, D. M. (1976). The Depressive Experiences Questionnaire. Unpublished manuscript, Yale University, New Haven, CT. Blatt, S. J., & Zuroff, D. C. (1992). Interpersonal relatedness and self-definition: Two prototypes for depression.*Clinical Psychology Review*, 12, 527–562. doi: 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. doi: 10.1016/j.cpr.2005.03.001

Bulmash, E., Harkness, K. L., Stewart, J. G., & Bagby, R. M. (2009). Personality, stressful life events, and treatment response in major depression. *Journal of Consulting and Clinical Psychology,* 77, 1067–1077. doi:10.1037/a0017149

Castilho, P., Pinto-Gouveia, J., & Duarte, J. (2017). Two forms of self-criticism mediate differently the shame–psychopathological symptoms link. *Psychology and Psychotherapy: Theory, Research and Practice*, *90*(1), 44-54. doi:10.1111/papt.12094

Curran, T., & Hill, A. P. (2019). Perfectionism is increasing over time: A meta-analysis of birth cohort differences from 1989 to 2016. *Psychological Bulletin*, *145*(4), 410. doi:10.1037/bul0000138

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.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data.* Cambridge, MA: The MIT Press.

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. doi:10.1016/S0165-0327(01)00392-5

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

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

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

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. doi:10.1016/S0191-8869(99)00230-5.

Halamová, J., Kanovský, M., Gilbert, P., Troop, N., Zuroff, D., Hermanto, N., Petrocchi, N., Sommers-Spijkerman, M., Kirby, J., Shahar, B., Krieger, T., Matos, M., Asano, K., Yu, F., Basran, J., & Kupeli, N. (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. doi:10.1007/s10862-018-9686-2

Halamová, J., Kanovský, M., Gilbert, P., Kupeli, N., Troop, N., Zuroff, D., Petrocchi, N., Hermanto, N., Sommers-Spijkerman, M., Kirby, J., Shahar, B., Krieger, T., Matos, M., Asano, K., Yu, F., & Basran, J. (2019). Multiple Group IRT Measurement Invariance Analysis of the Forms of Self-Criticising/Attacking and Self-Reassuring Scale in Thirteen International Samples. *Journal of Rational-Emotive & Cognitive-Behavior Therapy.* Online first publication. doi:10.1007/s10942-019-00319-1

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.* 4, 456-471.

Hermanto, N. & Zuroff, D. Ch. (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. doi: 10.1080/00224545.2015.1135779

Hermanto, N., & Zuroff, D. (2017). Experimentally enhancing self-compassion: Moderating effects of trait care-seeking and perceived stress. *The Journal of Positive Psychology,* 8, 1-10. doi:10.1080/17439760.2017.1365162

Kass, R. E. & Raftery, A. E.(1995). Bayes factors. *Journal of the American Statistical Association,* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Kim, E. S., Cao, C., Wang, Y. & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal,* 24, 4, 524-544, doi:10.1080/10705511.2017.1304822

Keribin, C. (2000). Consistent estimation of the order of finite mixture models. *Sankhyā: The Indian Journal of Statistics,* Series A (1961-2002), 62 (1)*, 49-66. Retrieved from http://www.jstor.org/stable/25051289, 10/09/2019.

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.

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.

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*, 5:1, 214-228, DOI: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. doi:10.1037/a0021371.

Longe, O., Maratos, F.A., Gilbert, P., Evans, G., Volker, F., Rockliffe, H., & Rippon, G. (2010). Having a word with yourself: Neural correlates of self-criticism and self-reassurance. *NeuroImage, 49*, 1849-1856. doi:10.1016/j.neuroimage.2009.09.019Löw, C. A., Schauenburg, H., & Dinger, U. (2020). Self-criticism and psychotherapy outcome: A systematic review and meta-analysis. *Clinical Psychology Review*, *75*, 101808. doi:10.1016/j.cpr.2019.101808

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. doi:10.1037/a0032243

Maitra, R. & Melnykov, V. (2010). Assessing significance in finite mixture models. *Statistics Publications,* Technical Report, 10-01, Department of Statistics, Iowa StateUniversity. https:

https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1069&context=stat_las_pubs, Retrieved 10/09/2019.

Melnykov, V. & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys,* 4, 80-116, doi: 10.1214/09-SS053

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement,* 16, 159–176. doi: 10.1002/j.2333-8504.1992.tb01436.x

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.

R Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing.* Vienna, Austria. URL http://www.R-project.org/

Rose, A. V. & Rimes, K. A. (2018). Self-criticism self-report measures: Systematic review. *Psychology and Psychotherapy*, 91(4), 450-489. doi: 10.1111/papt.1217

Samejima, F. (1969). Estimation of ability using a responsepattern of graded scores. *Psychometrika,* 34 (Suppl 1, Monograph 17), 1-97. doi: 10.1007/BF03372160

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.

Shahar, G. (2015). *Erosion: The Psyhopathology of Self-Criticism*. New York: Oxford University Press.

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. doi: 10.1177/1073191115573300

Sommers-Spijkerman, M. P. J., Trompetter, H. R., ten Klooster, P. M., Schreurs, K. M. G., Gilbert, P. & Bohlmeijer, E. T. (2017). Development and Validation of the Forms of Self-Criticising/attacking and Self-Reassuring Scale - Short Form. *Psychological Assessment*. Forthcoming.

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.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology,* 33, 213–239, doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J.K.: Multilevel latent variable modeling: an application in education testing. *Austrian Journal of Statistics,* 37, 285–299, doi: 10.17713/ajs.v37i3&4.309

Werner, A. M., Tibubos, A. N., Rohrmann, S., & Reiss, N. (2019). The clinical trait self-criticism and its relation to psychopathology: A systematic review–Update. *Journal of Affective Disorders*, *246*, 530-547. doi:10.1016/j.jad.2018.12.069

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., 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. doi: org/10.1080/00223891.2015.1044604

Tables and Figures

Table 1
Standard LC models for detecting the optimal number (k) of latent classes on the level of respondents

	Deterministic input			Random input		
k	logLik	# par	BIC (**min**)	logLik	# par	BIC (**min**)
1	-242856.7	88	486501.1	-242856.7	88	486501.1
2	-221477.3	177	444539.0	-221477.3	177	444539.0
3	-213941.1	266	430263.2	-213941.1	266	430263.2
4	-210670.3	355	424518.2	-210670.3	355	424518.2
5	-208307.7	444	420589.5	-208307.7	444	420589.5
6	-207111.6	533	418993.9	-206719.3	533	418890.2
7	-206616.3	622	418800.1	-206650.0	622	**418070.8**
8	-206195.5	711	**418755.1**	-206399.1	711	419911.6
9	-206385.8	800	419932.3			

Table 2
IRT item parameterization selection for the three-dimensional model with 7 latent classes

Model	logLik	# par	BIC (**min**)
GRM	-208621.0	131	**418414.6**
RS-GRM	-209726.2	68	420061.1
1P-GRM	-209763.0	112	420528.4
1P-RS-GRM	-210480.3	49	421399.2

Table 3
Multilevel graded response finite mixture models with 7 latent classes for detecting the optimal number (u) of latent classes on the level of countries

U	logLik	# par	BIC (**min**)
1	-208621.0	131	418414.6
2	-208473.5	138	418182.3
3	-208308.1	145	417914.0
4	-208175.3	152	417711.2
5	-208106.2	159	**417635.6**
6	-208189.9	166	417865.7

Table 4
Standardized support points and membership for 7 latent classes (respondents)

Dimension	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Inadequate Self	-1.497	-1.270	-0.257	-0.157	0.793	1.117	2.263
Reassured Self	-1.860	-0.717	-0.223	-0.059	0.761	1.147	2.116
Hated Self	-1.480	-1.264	-0.445	0.249	0.529	1.185	2.330
Membership	11.16%	12.55%	21.20%	17.92%	18.21%	14.37%	4.59%

Table 5

Estimated average abilities, weights and membership for all latent classes on the level of countries

	Class 1	Class 2	Class 3	Class 4	Class 5
Support points	-0.195	-0.035	0.026	0.082	0.378
Membership	16.06%	43.36%	31.76%	5.41%	3.41%
Countries	ISR, POR	CAN, ITA, NL, SVK, UK2	AUS, CH, UK1, USA	TAI	JAP

Note. 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).
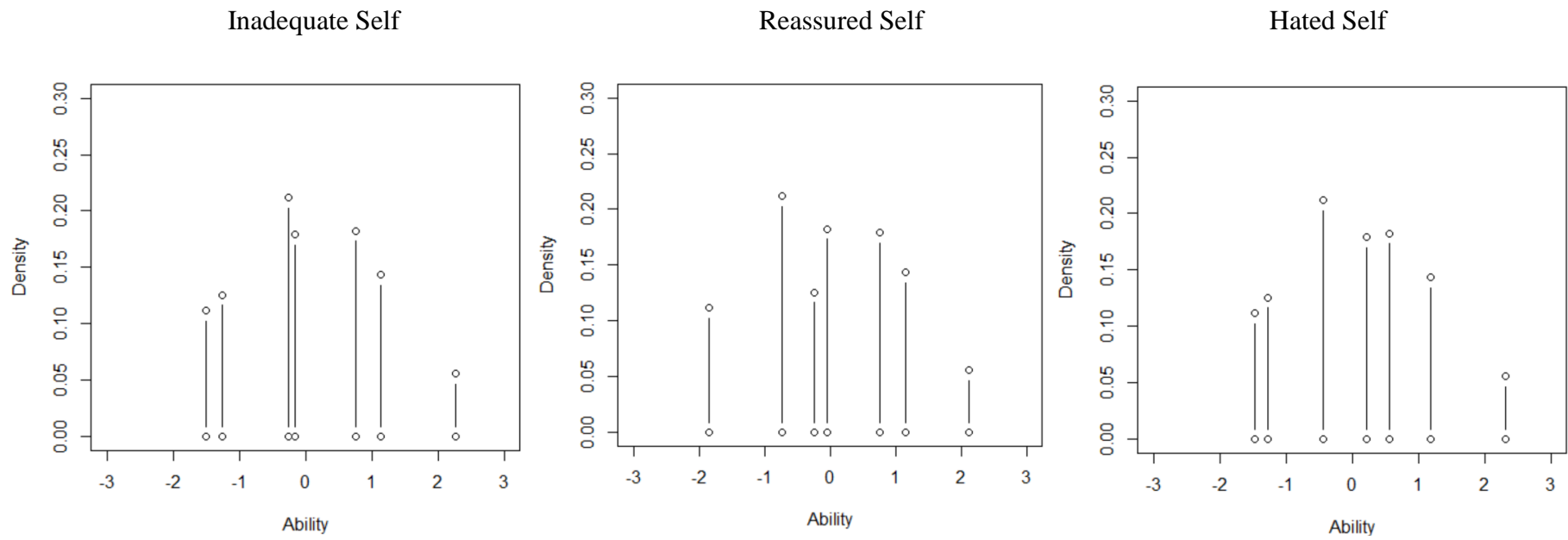
Figure 1. Standardized support points and weights for sub-dimensions of the FSCRS on the level of respondents.
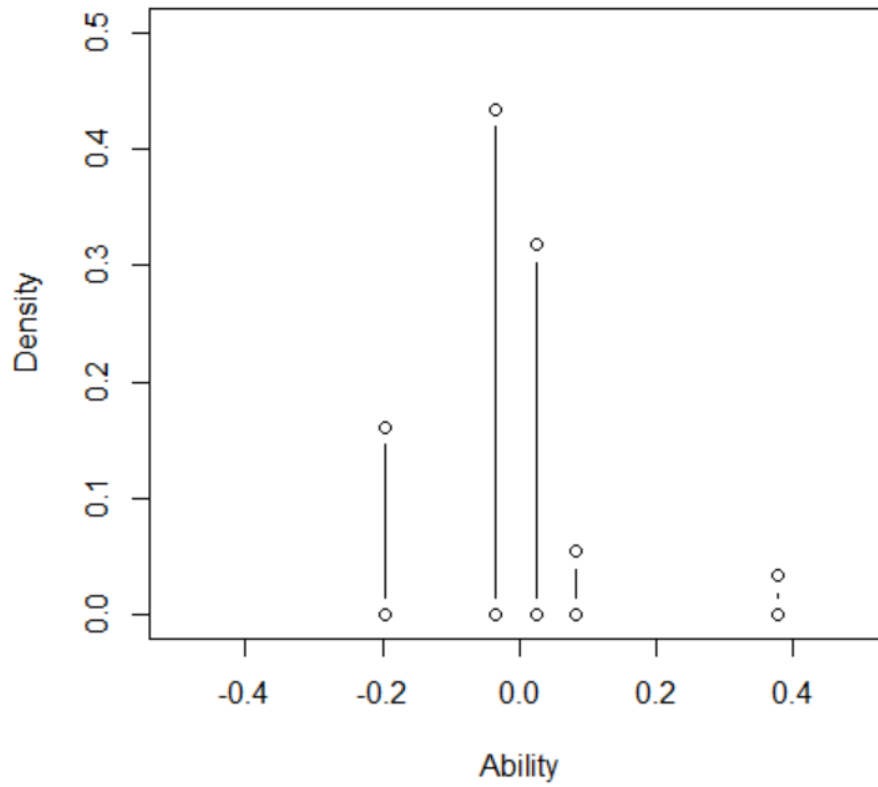
Figure 2. Average values of latent ability and weights on the level of countries.

Electronic Supplementary material

ESM 1: Tables
ESM 1a. List of unidimensional IRT models for ordinal polytomous responses with constraints on the discrimination indices, and constraints on the difficulty levels.

discrimination indices	difficulty levels	parameterization	model
free	free	$\gamma_j(\theta - \beta_{jx})$	GRM
free	constrained	$\gamma_j[(\theta - (\beta_j + \tau_x)]$	RS-GRM
constrained	free	$\theta - \beta_{jx}$	1P-GRM
constrained	constrained	$\theta - (\beta_j + \tau_x)$	1P-RS-GRM

ESM 1b. Distribution of the FSCRS item responses (raw percentage frequencies)

Item	Response category					
Inadequate Self	0	1	2	3	4	Total
1	13.7	26.1	24.5	23.2	12.5	100.0
2	14.9	26.5	21.0	24.4	13.2	100.0
4	23.9	29.8	21.9	17.3	7.1	100.0
6	13.7	24.2	20.8	25.7	15.6	100.0
7	25.9	26.4	21.1	17.3	9.3	100.0
14	15.2	29.0	23.5	21.0	11.3	100.0
17	19.4	29.2	23.0	18.9	9.5	100.0
18	13.7	26.1	27.7	23.4	9.1	100.0
20	12.2	19.1	21.4	27.1	20.2	100.0
Mean	17.0	26.3	22.7	22.0	12.0	100.0
Reassured Self	0	1	2	3	4	Total
3	14.2	36.2	27.8	16.6	5.2	100.0
5	8.3	22.8	30.1	27.6	11.2	100.0
8	32.0	35.3	19.7	9.5	3.5	100.0
11	26.9	37.0	22.0	10.7	3.4	100.0
13	13.7	28.1	28.6	20.7	8.9	100.0
16	8.8	25.2	32.9	24.1	9.0	100.0
19	34.7	35.2	19.7	8.4	2.0	100.0
21	26.8	36.7	23.0	10.2	3.3	100.0
Mean	20.6	32.1	25.5	16.0	5.8	100.0
Hated Self	0	1	2	3	4	Total
9	70.5	13.0	7.4	5.2	3.9	100.0
10	57.1	19.7	11.2	7.7	4.3	100.0
12	56.0	23.2	13.1	5.4	2.3	100.0
15	48.0	22.7	14.3	10.1	4.9	100.0
22	56.1	22.4	11.4	6.1	4.0	100.0
Mean	57.5	20.2	11.5	6.9	3.9	100.0

ESM 1c. Descriptives of raw scores of sub-dimensions of the FSCRS

	Inadequate Self			Reassured Self			Hated Self		
Country	M	SD	Md	M	SD	Md	M	SD	Md
AUS	15.61	9.00	15	12.76	7.02	12	3.42	4.12	2
CAN	17.99	8.17	18	11.41	5.96	11	3.18	3.58	2
CH	17.97	8.54	18	15.36	7.63	15	4.72	4.32	4
ISR	12.25	7.68	11	10.33	6.35	9	1.75	2.97	0
ITA	15.20	7.77	15	12.14	5.70	12	3.25	3.56	2
JAP	24.13	6.25	25	15.48	5.73	16	9.57	4.47	10
NL	14.70	7.16	14	10.50	5.43	10	2.93	3.65	2
POR	13.33	7.66	13	11.99	6.09	11	2.55	3.39	1
SVK	17.02	7.46	16	11.56	5.77	11	4.70	3.82	4
TAI	19.88	5.92	20	11.31	4.80	11	5.49	3.72	3
UK1	17.66	8.97	17	13.79	6.75	13	4.06	4.84	2
UK2	17.71	8.29	18	11.76	5.76	11	3.86	4.58	2
USA	15.68	8.19	15	13.48	6.88	14	3.86	4.25	2
Total	16.72	8.32	16	12.34	6.30	12	3.97	4.32	3

Note. M = Mean; SD = Standard Deviation; Md = Median; 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).

ESM 1d. Standardized discrimination and difficulties indices

Item	Discriminations	Difficulties / Thresholds			
Inadequate Self		1	2	3	4
1	2.163	-1.331	-0.307	0.476	1.472
2	2.263	-1.266	-0.272	0.398	1.405
4	1.444	-1.048	0.162	1.066	2.297
6	2.275	-1.319	-0.364	0.290	1.271
7	2.384	-0.783	0.079	0.783	1.625
14	1.722	-1.391	-0.190	0.646	1.714
17	1.768	-1.146	-0.025	0.798	1.832
18	1.316	-1.701	-0.371	0.752	2.217
20	1.369	-1.789	-0.726	0.140	1.341
Reassured Self		1	2	3	4
3	1.440	-1.779	-0.049	1.161	2.525
5	1.142	-2.616	-0.940	0.437	2.131
8	1.927	-0.721	0.538	1.462	2.408
11	1.740	-0.962	0.440	1.472	2.552
13	1.939	-1.591	-0.344	0.688	1.770
16	1.656	-2.095	-0.657	0.577	1.885
19	1.266	-0.759	0.777	2.048	3.536
21	1.360	-1.083	0.471	1.711	2.983
Hated Self		1	2	3	4
9	1.562	0.784	1.417	1.969	2.663
10	2.167	0.223	0.896	1.457	2.204

12	1.648	0.225	1.153	2.048	2.958
15	1.382	-0.092	0.820	1.604	2.674
22	2.231	0.207	0.978	1.587	2.228

ESM 2. Mathematical formulations

ESM 2a. Global (cumulative) logits (or graded response model), originally proposed by Samejima (1969), see Bacci, Bartolucci & Gnaldi (2014), chap. 2:

$$g_x[\boldsymbol{\lambda}_j(\boldsymbol{\theta})] = \log \frac{\lambda_{x|\theta}^{(j)} + \cdots + \lambda_{l_j-1|\theta}^{(j)}}{\lambda_{0|\theta}^{(j)} + \cdots + \lambda_{x-1|\theta}^{(j)}} = \log \frac{p(X_j \geq x|\theta)}{p(X_j < x|\theta)}, \quad x = 1, \ldots, l_j - 1$$

and its multidimensional extension (Bacci, Bartolucci & Gnaldi 2014, chap. 3.1, Eq. 7):

$$\log \frac{p(X_j \geq x|\boldsymbol{\Theta} = \boldsymbol{\theta})}{p(X_j < x|\boldsymbol{\Theta} = \boldsymbol{\theta})} = \gamma_j(\sum_{d=1}^{s} \delta_{jd}\theta_d - \beta_{jx}), \quad x = 1, \ldots, l_j - 1$$

ESM 2b. Local (adjacent categories) logits (generalized partial credit model), originally proposed by Muraki (1992), see Bacci, Bartolucci & Gnaldi (2014), chap. 2:

$$g_x[\boldsymbol{\lambda}_j(\theta)] = \log \frac{\lambda_{x|\theta}^{(j)}}{\lambda_{x-1|\theta}^{(j)}} = \log \frac{p(X_j = x|\theta)}{p(X_j = x - 1|\theta)}, \quad x = 1, \ldots, l_j - 1$$

ESM 2c. continuation ratio logits (sequential model), originally proposed by Fienberg (1980), see Bacci, Bartolucci & Gnaldi (2014), chap. 2:

$$g_x[\boldsymbol{\lambda}_j(\theta)] = \log \frac{\lambda_{x+1|\theta}^{(j)} + \cdots + \lambda_{l_j-1|\theta}^{(j)}}{\lambda_{x|\theta}^{(j)}} = \log \frac{p(X_j > x|\theta)}{p(X_j = x|\theta)}, \quad x = 1, \ldots, l_j - 1$$

ESM 3. R-codes

 # installing R-packages and preparing data
 # we assume that dataframe named „data" is downloaded, with the first column named „country" which identifies membership in a group (country). Further, we assume that all items are coded with the first value (category) set at 0.

```
install.packages("MultiLCIRT")
library(MultiLCIRT)
```

 # data preparation
```
n = nrow(data)
```

```
dat=subset(data,select=c(2:23))
dat = as.matrix(dat)
clust = as.vector(data$country)
clust_unique = unique(clust)
for(i in 1:length(clust)) clust[i] =
+ which(clust_unique==clust[i])
```

```
# assigning the items to the corresponding latent trait:
dim = c(1,2,4,6,7,14,17,18,20,3,5,8,11,13,16,19,21,rep(0,1),9,10,12,15,22,rep(0,4))
dim = t(matrix(dim,9,3))
```

```
# fitting the standard LCA model with 1 latent class, deterministic input:
out1 = est_multi_poly(dat,k=1,start=0, link=0)
# fitting LCA model with increasing number of latent classes, deterministic input:
out2 = est_multi_poly(dat,k=2,start=0, link=0)
out3 = est_multi_poly(dat,k=3,start=0, link=0)
out4 = est_multi_poly(dat,k=4,start=0, link=0)
out5 = est_multi_poly(dat,k=5,start=0, link=0)
out6 = est_multi_poly(dat,k=6,start=0, link=0)
out7 = est_multi_poly(dat,k=7,start=0, link=0)
out8 = est_multi_poly(dat,k=8,start=0, link=0)
```

```
# displaying values of log-likelihood, the number of parameters, and BIC for models with the
deterministic input (see Table 1 for the output):
rbind(cbind(out1$lk, out1$np, out1$bic), cbind(out2$lk, out2$np, out2$bic), cbind(out3$lk,
out3$np, out3$bic), cbind(out4$lk, out4$np, out4$bic), cbind(out5$lk, out5$np, out5$bic),
cbind(out6$lk, out6$np, out6$bic), cbind(out7$lk, out7$np, out7$bic), cbind(out8$lk,
out8$np, out8$bic))
```

```
# fitting the standard LCA model with 1 latent class, random input:
out11 = est_multi_poly(dat,k=1,start=1, link=0)
# fitting the standard LCA model with increasing number of latent classes, random input:
out22 = est_multi_poly(dat,k=2,start=1, link=0)
out33 = est_multi_poly(dat,k=3,start=1, link=0)
out44 = est_multi_poly(dat,k=4,start=1, link=0)
out55 = est_multi_poly(dat,k=5,start=1, link=0)
out66 = est_multi_poly(dat,k=6,start=1, link=0)
out77 = est_multi_poly(dat,k=7,start=1, link=0)
out88 = est_multi_poly(dat,k=8,start=1, link=0)
```

```
# displaying values of log-likelihood, the number of parameters, and BIC for models with the
random input (see Table 1 for the output):
rbind(cbind(out11$lk, out11$np, out11$bic), cbind(out22$lk, out22$np, out22$bic),
cbind(out33$lk, out33$np, out33$bic), cbind(out44$lk, out44$np, out44$bic),
cbind(out55$lk, out55$np, out55$bic), cbind(out66$lk, out66$np, out66$bic),
cbind(out77$lk, out77$np, out77$bic), cbind(out88$lk, out88$np, out88$bic))
```

```
# fitting the three-dimensional graded-response LCA models with 7 latent classes, global
logits link, different IRT parameterizations:
# GRM
```

```
outA=est_multi_poly(dat,k=7,multi=dim,start=1,link=1,disc=1,difl=0)
# RS-GRM
outB=est_multi_poly(dat,k=7,multi=dim,start=1,link=1, disc=1,difl=1)
# 1P-GRM
outC=est_multi_poly(dat,k=7,multi=dim,start=1,link=1,disc=0,difl=0)
# 1P-RS-GRM
outD=est_multi_poly(dat,k=7,multi=dim,start=1,link=1,disc=0, difl=1)
```

*# displaying values of log-likelihood, the number of parameters, and BIC for models with
different IRT parameterizations (see Table 2 for the output):*
```
rbind(cbind(outA$lk, outA$np, outA$bic), cbind(outB$lk, outB$np, outB$bic),
cbind(outC$lk, outC$np, outC$bic), cbind(outD$lk, outD$np, outD$bic))
```

 # fitting the three-dimensional multilevel graded-response LCA models with 7 latent classes
```
out1=est_multi_poly_clust(dat,kU=1,kV=7,multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust)
out2=est_multi_poly_clust(dat,kU=2,kV=7, multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust)
out3=est_multi_poly_clust(dat,kU=3,kV=7, multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust)
out4=est_multi_poly_clust(dat,kU=4,kV=7, multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust)
out5=est_multi_poly_clust(dat,kU=5,kV=7, multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust)
out6=est_multi_poly_clust(dat,kU=6,kV=7, multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust)
```

*# displaying values of log-likelihood, the number of parameters, and BIC (see Table 3 for the
output):*
```
rbind(cbind(out1$lk, out1$np, out1$bic), cbind(out2$lk, out2$np, out2$bic),cbind(out3$lk,
out3$np, out3$bic), cbind(out4$lk, out4$np, out4$bic), cbind(out5$lk, out5$np, out5$bic),
cbind(out6$lk, out6$np, out6$bic))
```

 # re-fit the final model to obtain additional output
```
out5=est_multi_poly_clust(dat,kU=5,kV=7, multi=dim,start=0,link=1,
disc=1,difl=0,clust=clust,output=TRUE)
```

 # Abilities at respondents' level for each latent class and each dimension
```
ind = order(out5$Th[1,])
Ths = out5$Th[,ind]
Ths
```

 # Latent class average weights at respondents' level
```
Pivs = out5$Piv[,ind,]
Pivms = matrix(0,n,7)
for(h in unique(clust)){
+ ind1 = which(clust==h)
+ for(u in 1:5) Pivms[ind1,] =
+ Pivms[ind1,]+Pivs[ind1,,u]*out5$La[h,u]
+ }
```

```
pivs = colMeans(Pivms)
pivs

# Standardized respondents' abilities (see Table 4 for the output)
mThs = as.vector(Ths%*%pivs)
vThs = as.vector((Ths-mThs)^2%*%pivs)
sThs = (Ths-mThs)*(1/sqrt(vThs))
sThs

# Standardized difficulties for each item (see Table ESM 1d for the output)
outST = standard.matrix(t(out5$Th),pivs)
inad = as.vector(t(out5$Bec[c(1,2,4,6,7,14,17,18,20),]))
reas = as.vector(t(out5$Bec[c(3,5,8,11,13,16,19,21),]))
hate = as.vector(t(out5$Bec[c(9,10,12,15,22),]))
inadSTdif = (inad-outST$mu[1])/outST$si[1]
reasSTdif = (reas-outST$mu[2])/outST$si[2]
hateSTdif = (hate-outST$mu[3])/outST$si[3]

# Standardized discriminations for each item (see Table ESM 1d for the output)
inadSTdisc = out5$gac[c(1,2,4,6,7,14,17,18,20)]*outST$si[1]
reasSTdisc = out5$gac[c(3,5,8,11,13,16,19,21)]*outST$si[2]
hateSTdisc = out5$gac[c(9,10,12,15,22)]*outST$si[3]

# Mean values of respondents' abilities for each type of countries (see Table 5 for the output)
ThU = rep(0,times=5)
for(u in 1:5) ThU[u] = sum(Pivs[,,u]%*%t(sThs))/(n*7)
ThU
```