Measurement Instruments for the Social Sciences

VALIDATION OF MEASUREMENT INSTRUMENTS Open Access

Factorial validity and measurement invariance across gender groups of the German version of the Interpersonal Reactivity Index

Dennis Grevenstein

Abstract

The Interpersonal Reactivity Index (IRI) is the most widely used measure of empathy, but its factorial validity has been questioned. The present research investigates the factorial validity of the German adaptation of the IRI, the "Saarbrücker Persönlichkeitsfragebogen SPF-IRI". Confirmatory Factor Analyses (CFA) and Exploratory Structural Equation Modeling (ESEM) were used to test the theoretically predicted four-factor model. Across two subsamples ESEM outperformed CFA. Substantial cross-loadings were evident in ESEM. Measurement invariance (MI) across gender groups was tested using ESEM in the combined sample. Strict MI (invariant factor loadings, intercepts, residuals) could be established, and variances and covariances were also equal. Differences for latent means were evident. Women scored higher on fantasy, empathic concern, and personal distress. No significant differences were found for perspective taking. Mean differences were due to real differences on latent variables and not a result of measurement bias. Results support the factorial validity of the German SPF-IRI. The heterogeneity of empathy and the unclear differentiation between cognitive and emotional aspects might be a source for the unclear differentiation of scales.

Keywords: Empathy, Interpersonal Reactivity Index, Factorial validity, Measurement invariance

Factorial Validity and Measurement Invariance across Gender Groups of the German Version of the Interpersonal Reactivity Index

Empathy is commonly understood as a multidimensional construct that comprises emotional as well as cognitive components (Davis, 1980, 1983; Hoffman, 2000). Conceptions of trait-empathy usually refer to the seminal work of Davis (1980) and are often linked to the structure of the Interpersonal Reactivity Index (IRI). The IRI measures empathy in four subscales. Perspective taking (PT) describes an individual's tendency to consider another person's viewpoints. Fantasy (FS) refers to one's tendency to identify with fictional characters in books or films. Both factors are often seen as more cognitive, even though FS has been associated with higher emotionality (Paulus, 2009). Empathic concern (EC) means having feelings of compassion and concern for the needs of others. Finally, personal distress (PD) indicates a tendency to experience discomfort and stress in the presence of others, who are in distress. Both factors cover more emotional aspects of empathy (Schreiter, Pijnenborg, and aan het Rot, 2013).

Traditionally, empathy has been considered a positive aspect of mental health. Overall, empathy has been associated with more positive adaptation, unselfishness, understanding, and prosocial behavior (Cliffordson, 2002; Davis, 1983; Eisenberg and Fabes, 1990). However, PD has been linked to depression and psychological distress (Grevenstein and Bluemke, 2015; Lee, 2009; Schreiter

Correspondence: dennis.grevenstein@psychologie.uni-heidelberg.de
Psychological Institute, University of Heidelberg, Hauptstraße 47-51, 69117 Heidelberg, Germany

et al. 2013). PD indicates a susceptibility to social stress and thus has been associated with social anxiety, shyness, loneliness, and difficulties in social interactions (Carmel and Glick, 1996; Cliffordson, 2002; Davis, 1983). Empathy deficits have been documented for patients with schizophrenia (Abramowitz, Ginger, Gollan, and Smith, 2014; Smith et al. 2012). A recent meta-analysis showed that patients scored lower on EC, PT, and FT, but higher on PD compared to healthy controls (Bonfils, Lysaker, Minor, and Salyers, 2017).

Empathy has also been related to the Big Five traits (Lee, 2009; Mooradian, Davis, and Matzler, 2011). PT and EC have shown small negative correlations with neuroticism. FT has shown small positive correlations with neuroticism. The strongest associations were found between PD and neuroticism. Associations between empathy and the Big Five traits were also confirmed to be cross-culturally valid (Melchers et al. 2016).

Women are often shown to be more empathic than men (Davis, 1983; Hoffman, 2000). Females commonly score higher on all subscales of the IRI (Davis, 2004; De Corte et al. 2007; Hawk et al. 2013). In several cases, however, no significant differences were found for the PT subscale (Fernández, Dufey, and Kramp, 2011; Gilet, Mella, Studer, Grühn, and Labouvie-Vief, 2013; Lucas-Molina, Pérez-Albéniz, Ortuño-Sierra, and Fonseca-Pedrero, 2017).

The IRI is the most widely used measure of empathy, yet criticism remains. Several studies have confirmed a four-dimensional structure across various languages, e.g., French (Gilet et al. 2013), Spanish (Fernández et al. 2011; Garcia-Barrera, Karr, Trujillo-Orrego, Trujillo-Orrego, and Pineda, 2017), and Dutch (De Corte et al. 2007; Hawk et al. 2013). In many cases, confirmatory factor analyses (CFA) has revealed problems. Researchers had to drop items from the scale or use item parceling. In most cases, acceptable model fit following conventional cut-offs could not be achieved.

A German adaptation of the IRI was provided by Paulus (2009) and named "Saarbrücker Persönlichkeitsfragenbogen SPF (IRI)"; to be called SPF-IRI from here on. It is a 16-item version of the original 28-item IRI and a four-factor structure was established using exploratory factor analyses (EFA). Koller and Lamm (2015) conducted an analysis of the SPF-IRI on the basis of item response theory. Results indicated that only the subscale empathic concern conformed to the assumption of a partial credit model. Especially the personal distress subscale was evaluated critically. To this date, no extensive evaluation of factor structure of the SPF-IRI has been published. Hence, there is a need to show that the adapted and shortened German version of IRI still complies with a similar factor structure when compared to the international versions.

Measurement and factor analysis

Knowledge of the internal structure of a measure provides a basic understanding of the quality of measurement. Factor analysis is at the heart of current methodological approaches to investigations of internal structure. Aggregation to sum scores, estimations of reliability, and finally associations to other constructs can only meaningfully be interpreted if the internal structure of a test has been determined (Brown, 2006). Traditionally, multi-dimensional constructs (such as empathy) are expected to comply to a simple structure (Thurstone, 1934). In a nutshell, simple structure assumes that an item shows strong correlations with other items belonging to the same factor, yet ideally zero correlations with items belonging to other factors. Simple structure has often been considered a fundamental principal, in order to interpret the results of factor analyses (Kline, 1994). Contrasting these long-standing heuristics, more recent research has shown that constructs in the domain of personality can be more complex and that imposing simple structure (i.e., by removing non-conforming items) may degrade test information and increase standard errors (Pettersson and Turkheimer, 2014). CFA imposes simple structure, as items are commonly associated to a single factor only (unless explicitly specified otherwise). Cross-loadings on other factors are set to zero. The SPF-IRI has shown acceptable—yet less than ideal—internal consistencies ranging from alpha = .66 to .74 (Paulus, 2009). During the development of the SPF-IRI, Paulus (2009) documented considerable cross-loadings for several items. Due to the lack of available data, one can only speculate that this could be a reason why the factor structure of the SPF-IRI has not been confirmed using CFA. Hence, it is probable that the SPF-IRI does not perfectly comply with the assumption of a simple structure.

Exploratory Structural Equation Modeling (ESEM) has been introduced as an alternative method to evaluate the factor structure of scales (Asparouhov and Muthén, 2009; Marsh, Morin, Parker, and Kaur, 2014). ESEM aims to combine the advantages of both EFA and CFA. It is less restrictive and freely estimates item loadings on all factors. ESEM is thought to offer a more realistic approach to common personality constructs, following the assumption that CFA is often too restrictive (Marsh et al. 2014). However, ESEM does not necessarily provide more insight. For example, good model fit in ESEM could also be achieved when different item-to-factor associations emerge. Thus, one has to carefully examine the patterns of factor loadings to assure that results are comparable to a theoretically predicted model. Still, ESEM has considerable advantages over CFA when substantial cross-loadings are to be expected. Simulation studies showed that even small cross-loadings—as small

as .10—should be taken into account to prevent biased estimates (Asparouhov, Muthén, and Morin, 2015). Lucas-Molina and colleagues (Lucas-Molina et al. 2017) attempted an ESEM analysis of the Spanish version of IRI. ESEM proved to be superior to CFA, even though model fit was only barely acceptable.

Measurement invariance

Measurement invariance (MI) maintains that a valid measurement model has to hold across different samples, in order to compare scores across contexts, times, or groups of participants (Vandenberg and Lance, 2000). MI ensures that scores reflect the same latent construct to the same degree. Many analyses we commonly take for granted, such as correlations or comparisons of mean scores across groups, are only valid as far as MI holds (Chen, 2008). Otherwise, unequal measurement might obscure or bias true associations or differences. One needs to ascertain that any differences in scale means (or latent means) are due to true differences, not different item utilization (different loadings) or item bias (item difficulty). In the case of the IRI, gender differences have emerged in many studies. As MI is commonly tested in a CFA framework and due to the lack of an accepted factor model based on CFA, MI for groups of women and men has not been tested for the SPF-IRI.

MI is commonly tested using a series of increasingly restrictive CFA models (Brown, 2006; Meredith, 1993; Vandenberg and Lance, 2000). Marsh and colleagues (Marsh et al. 2009) provided an extensive taxonomy to test MI using 13 partially nested ESEM models, yet basically the invariance of five different groups of parameters is tested in various combinations. I will thus comply with the more traditional approach (Vandenberg and Lance, 2000). Four stages of MI are commonly tested: Configural MI (M1) indicates equal construct dimensionality and item-to-factor associations across groups. Factor loadings, item intercepts, and residuals can still differ. Due to the nature of ESEM, where all factors are associated with all items, this step is basically meaningless for ESEM, yet it provides a reference model for later tests. Metric MI (M2) indicates that all factor loadings are equal across groups. In case of ESEM this includes all loadings of an item on all factors. Scalar MI (M3) assumes that all item intercepts are equal. Strict MI (M4) also requires equal item residuals. Additionally, one can test the equality of structural parameters, such as factor variances and covariances (M5), and factor means (M6). Different levels of MI have different implications. Metric MI indicates that the same psychological meaning is captured. Metric MI is commonly considered to allow for a comparison of (latent) analyses of variance/covariance structures, such as correlations (van de Schoot, Lugtig, and Hox, 2012). On a methodologically strict level, metric MI allows comparing unstandardized covariances. Comparing correlations, i.e., standardized coefficients, technically additionally requires equal factor variances. Scalar MI allows for a comparison of (latent) factor means. Strict MI indicates that differences reflect true differences on the latent variables, rather than random measurement error. This assures equal reliability and one can directly compare scale means.

Study overview

Due to the lack of a proper investigation of the internal structure of the SPF-IRI, the present research aims to investigate two research questions:

RQ1: Establishing a suitable factor model using CFA or ESEM. I will compare CFA and ESEM models in two subsamples. I hypothesize that CFAs will fail to show acceptable fit, whereas ESEM will be superior. In order to cross-validate the factor model, I will use two subsamples from two separate studies.

RQ2: Once a valid factor model has been established, I will then investigate MI across gender groups in the full, combined sample. The combined sample will be used at this stage, in order to maximize available participants, given the comparatively lower number of men vs. women that is all too common in psychological studies.

Methods

Procedure and participants

The data analyzed in the present study came from two subsamples. All participants completed online studies. In both studies, data were collected in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants and participation was completely voluntary. Subsample 1 included N = 1033 (75.2% female) participants with a mean age of 41.83 years (SD = 14.14; range = 13–83). Subsample 2 included N = 1842 participants (85.5% female; M_{age} = 28.11; SD_{age} = 9.22; range = 15–77). The full sample thus included N = 2875 participants (81.8% female; M_{age} = 33.05; SD_{age} = 13.03). Descriptives are presented in Table 1. Men were significantly older than women in subsample 1 (M_{men} = 44.20 vs. M_{women} = 41.05; SD_{men} = 14.40 vs. SD_{women} = 13.97; t = 3.05; df = 424.20; p = .002; d = .22), but not in subsample 2 (M_{men} = 28.94 vs. M_{women} = 27.97; SD_{men} = 10.03 vs. SD_{women} = 9.08; t = 1.58; df = 1836; p = .12; d = .10). Participants in subsample 2 had diverse educational backgrounds (18.3% basic schooling; 45.6% high school; 36.1% university level degree). Subsample 1 included more participants from a higher educational background (8.4% basic schooling; 32.1% high school; 59.4% university level degree). Participants were recruited via social media sites (i.e., Facebook) and e-mail lists from the

Table 1 Descriptives and correlations between SPF-IRI scales in the full sample and in subsamples. Correlations of scale sum scores above diagonal, correlations of latent variables from ESEM below diagonal

All participants:	Total: *N* = 2875			Women: *n* = 2351			Men: *n* = 524						
	M (SD)	α	ω	*M* (SD)	α	ω	*M* (SD)	α	ω	FS	PD	PT	EC
Fantasy (FS)	13.79 (2.74)	.71	.82	13.94 (2.73)	.71	.82	13.11 (2.69)	.66	.79	-	.19	.28	.41
Personal distress (PD)	10.92 (2.93)	.76	.80	11.16 (2.91)	.75	.80	9.80 (2.73)	.73	.79	.11	-	−.14	.23
Perspective taking (PT)	14.78 (2.54)	.76	.85	14.76 (2.55)	.77	.85	14.83 (2.50)	.74	.82	.25	−.18	-	.28
Empathic concern (EC)	14.95 (2.28)	.64	.73	15.15 (2.24)	.63	.73	14.07 (2.28)	.60	.69	.49	.28	.28	-
Subsample 1:	Total: *N* = 1033			Women: *n* = 777			Men: *n* = 256						
	M (SD)			*M* (SD)			*M* (SD)			FS	PD	PT	EC
Fantasy (FS)	13.74 (2.59)	.71	.82	13.94 (2.52)	.69	.82	13.12 (2.73)	.70	.75	-	.16	.24	.45
Personal distress (PD)	9.86 (2.68)	.77	.77	10.09 (2.64)	.76	.77	9.17 (2.70)	.80	.71	.11	-	−.18	.16
Perspective taking (PT)	15.29 (2.22)	.73	.85	15.31 (2.21)	.73	.85	15.23 (2.27)	.72	.80	.22	−.25	-	.25
Empathic concern (EC)	14.66 (2.14)	.62	.74	14.84 (2.10)	.62	.73	14.13 (2.18)	.60	.68	.53	.23	.28	-
Subsample 2:	Total: *N* = 1842			Women: *n* = 1574			Men: *n* = 268						
	M (SD)			*M* (SD)			*M* (SD)			FS	PD	PT	EC
Fantasy (FS)	13.82 (2.82)	.70	.82	13.94 (2.83)	.72	.81	13.10 (2.66)	.63	.78	-	.21	.31	.39
Personal distress (PD)	11.51 (2.89)	.72	.82	11.70 (2.90)	.73	.80	10.41 (2.62)	.64	.81	.16	-	−.07	.24
Perspective taking (PT)	14.49 (2.66)	.77	.83	14.49 (2.66)	.77	.83	14.45 (2.65)	.74	.80	.26	−.10	-	.31
Empathic concern (EC)	15.11 (2.34)	.65	.72	15.30 (2.29)	.64	.71	14.01 (2.37)	.61	.68	.44	.26	.30	-

Note: α = Cronbach's alpha, ω = McDonald's ordinal omega

local university. In both studies participants could participate in a lottery for compensation. The survey software reminded participants to respond in case of missing values, so there were no missing data. Participants were instructed that they could drop out of the study at any time and only data provided by participants who completed the entire survey were analyzed.

Measures

The German SPF-IRI includes 16 items measuring four dimensions of empathy with four items each. Answers were given on 5-point scales (1 = *never* to 5 = *always*). Reliabilities (Cronbach's alpha and McDonald's ordinal omega) were computed for the full sample, as well as for subsamples and subgroups of men and women, and are depicted in Table 1. Items were aggregated to scale sum scores, as there were no missing values.

Statistical analysis

I used SPSS 22 for descriptives, and Mplus 7.4 (Muthén and Muthén, 1998–2012) for confirmatory factor analyses (CFA) and exploratory structural equation modeling (ESEM). Model fit was evaluated using χ^2 tests (Bentler and Bonett, 1980), the comparative fit index (CFI) and Tucker-Lewis index (TLI) with values > .90 indicating acceptable model fit (Bentler, 1990; Hu and Bentler, 1999), the root mean square error of approximation (RMSEA) with values < .08 indicating acceptable model fit (Browne and Cudeck, 1993), and the standardized root mean square residual (SRMR) with values < .08 (Hu and Bentler, 1999) or .05 (Schumacker and Lomax, 2010) reflecting good fit. If models are based on the same data and variables, they can be compared using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Lower scores indicate better model fit (Akaike, 1987). Differences greater than ± 10 are considered meaningful (Raftery, 1995). AIC emphasizes model accuracy and BIC provides the best trade-off between accuracy and parsimony. Robust Maximum Likelihood (MLR) was used for parameter estimation. I will also provide Raykov's composite reliability (CR; Raykov, 1997) as an SEM-based reliability estimate. McDonald's ordinal omega was computed using the "psych" package (Revelle, 2019) for the statistical software R (R Foundation for Statistical Computing, 2020). Ordinal omega can be computed based on polychoric correlations and should be better suited to estimate reliability for coarse ordinal scales (Gadermann, Guhn, and Zumbo, 2012; Viladrich, Angulo-Brunet, and Doval, 2017).

For measurement invariance (MI), models ares compared from one step to the next using χ^2 difference tests. MLR uses scaled χ^2 scores and therefore Satorra-Bentler scaled χ^2 difference tests have to be used (Satorra, 2000; Satorra and Bentler, 2001). χ^2 tests are greatly influenced by sample size and model fit indices are most dominantly used to judge model fit for MI. From one step to the next, a drop in CFI or TLI less or equal to .010 is

conventionally considered acceptable unless there is a concurrent increase of RMSEA larger than .015 (Chen, 2007; Cheung and Rensvold, 2002). For ESEM, researchers are encouraged to look at RMSEA and TLI, because they include a stronger penalty for model complexity (Marsh, Nagengast, and Morin, 2013).

Results

Descriptive data analysis of SPF-IRI scores

SPF-IRI descriptives, correlations, and reliability estimates are presented in Table 1. Women scored significantly higher on FS, EC, and PD (all *t*s > 4.41; all *p*s < .001, all *d*s = 0.31 to 0.56) in both samples. Notably, no differences emerged for PT in both subsample 1 (t = 0.47; df = 1031, p = .64, d = 0.04) and subsample 2 (t = 0.26; df = 1840, p = .80, d = 0.02). McDonald's ordinal omega generally produced higher reliability estimates than Cronbach's alpha, with the PD subscale in subsample 1 being an exception.

CFA and ESEM

I initially tested a CFA model in both samples. Standardized factor loadings for CFA and ESEM in both samples can be seen in Table 2. The CFA model of four correlated factors failed conventional limits of model fit in subsample 1 ($\chi^2(98)$ = 532.61, p < .001, RMSEA = .066, CI_{90} = [.060–.071], CFI = .875, TLI = .847, SRMR = .063) and in subsample 2 ($\chi^2(98)$ = 764.44, p < .001, RMSEA = .061, CI_{90} = [.057–.065], CFI = .892, TLI = .868, SRMR = .052). In contrast, an ESEM model of four correlated factors showed better fit in subsample 1 ($\chi^2(62)$ = 267.91, p < .001, RMSEA = .057, CI_{90} = [.050–.064], CFI = .941, TLI = .885, SRMR = .025) and subsample 2 ($\chi^2(62)$ = 353.42, p < .001, RMSEA = .051, CI_{90} = [.045–.056], CFI = .953, TLI = .909, SRMR = .022). As visible in Table 2, ESEM reproduced the theoretically predicted item-to-factor patterns, yet substantial cross-loadings were evident.

Measurement invariance

Results of the MI tests are depicted in Table 3. The initial configural MI model (M1) fit the data well. Constraining factor loadings to be equal resulted in an improvement in model fit (metric invariance: M2). The comparably large difference can also be attributed to the large gain in df, as all factor loadings in ESEM are now set to be equal across groups. Curiously, the χ^2 value for the overall model slightly decreased at this stage. Constraining intercepts (scalar invariance: M3) and residuals (strict invariance: M4) showed a slight decrease in model fit that was within acceptable levels. An investigation of structural parameters confirmed that variances and covariances (M5) were also equal. The last model (M6) added equal factor means. In light of the initially shown scale mean differences, equality of latent means could not be expected. Indeed, model fit decreased markedly. On the basis of M5, latent mean differences could be estimated (unstandardized; female group with means fixed to zero). Men had lower means than women on

Table 2 Standardized factor loadings and composite reliabilities (CR) for CFA and ESEM in subsamples 1 and 2

Subsample 1	CFA				ESEM			
Items	FS	PD	PT	EC	FS	PD	PT	EC
2	**.65**				**.61**	− .15	.02	.07
7	**.52**				**.45**	.24	− .05	.09
12	**.74**				**.85**	− .02	.01	− .11
15	**.61**				**.50**	.06	.16	.04
3		**.64**			.11	**.63**	− .01	− .06
6		**.79**			− .06	**.79**	.00	.01
8		**.77**			− .04	**.79**	.08	.03
13		**.53**			.06	**.49**	− .09	.03
4			**.64**		− .07	− .02	**.70**	− .03
10			**.62**		.04	.02	**.64**	− .05
14			**.59**		.15	− .06	**.48**	.11
16			**.70**		.01	.02	**.68**	.05
1				**.48**	− .03	− .22	.08	**.61**
5				**.50**	.04	.02	− .03	**.51**
9				**.68**	.19	.05	− .05	**.52**
11				**.53**	.12	.09	.06	**.37**
CR	.72	.78	.73	.63	.70	.78	.72	.58
Subsample 2	CFA				ESEM			
items	FS	PD	PT	EC	FS	PD	PT	EC
2	**.61**				**.48**	− .07	.03	.20
7	**.57**				**.61**	.22	− .06	− .05
12	**.74**				**.65**	− .02	.00	.13
15	**.60**				**.60**	.05	.11	− .02
3		**.55**			− .02	**.49**	− .07	.14
6		**.75**			− .04	**.77**	.03	.00
8		**.77**			.03	**.78**	.07	− .02
13		**.46**			.07	**.40**	− .21	.11
4			**.64**		− .06	.03	**.74**	− .03
10			**.65**		.00	.01	**.69**	.01
14			**.68**		.23	.01	**.54**	.06
16			**.72**		.14	− .04	**.63**	.02
1				**.54**	.02	− .08	.08	**.54**
5				**.48**	.01	.01	.18	**.39**
9				**.66**	.10	.09	.01	**.56**
11				**.57**	− .05	.06	− .03	**.64**
CR	.73	.73	.77	.65	.68	.71	.74	.62

Note: Formal item-to-scale affiliations printed in bold. *CR* = Raykov's composite reliability

Table 3 Measurement invariance of SPF-IRI across gender groups based on ESEM

MGCFA comparison		Equal parameters	*df*	χ^2	Δ*df*	$\Delta\chi^2$	CFI	TLI	RMSEA [CI_{90}]	SRMR	AIC	BIC
M1	Configural invariance		124	701.96**	–	–	.940	.884	.057 [.053–.061]	.024	107655	108729
M2	Metric invariance	1	172	679.47**	48	51.53	.947	.927	.045 [.042–.049]	.029	107637	108424
M3	Scalar invariance	1,2	184	743.52**	12	65.50**	.942	.924	.046 [.043–.049]	.032	107680	108395
M4	Strict invariance	1,2,3	200	815.79**	16	71.98**	.936	.923	.046 [.043–.050]	.047	107731	108351
M5	**Structural: (co)variances**	**1,2,3,4**	**210**	**832.57****	**10**	**18.43***	**.935**	**.926**	**.045 [.042–.049]**	**.053**	**107733**	**108294**
M6	Structural: means	1,2,3,4,5	214	997.05**	4	179.93**	.919	.909	.050 [.047–.054]	.069	107907	108443

Note: Model fit: *$p < .05$, **$p < .001$. Accepted model printed in bold. χ^2 values are Satorra-Bentler scaled. Equal parameters: *1* = loadings, *2* = intercepts, *3* = residuals, *4* = (co)variances, *5* = latent means

PD (– 0.455 vs. 0.000; SE = 0.054; $p < .001$; $d = 0.37$), FS (– 0.333 vs. 0.000; SE = 0.064; $p < .001$; $d = 0.23$), and EC (– 0.691 vs. 0.000; SE = 0.068; $p < .001$; $d = 0.44$). Differences on PT (0.104 vs. 0.000; SE = 0.058; $p = .07$; $d = 0.08$) barely failed to reach significance, with a slight tendency for men to score higher on PT than women. Taken together, MI testing indicated that measurement was comparable across gender groups.

Discussion

The present research investigated factorial validity and measurement invariance (MI) of the German version of the Interpersonal Reactivity Index SPF-IRI across gender groups. For the first time, a four-factor structure could be replicated for the SPF-IRI in line with the conception of empathy by Davis (1983). Across two subsamples ESEM was superior to CFA. Paulus (2009) documented cross-loadings for some items of the SPF-IRI during the scale construction. He named items 9, 11, and 14 as having strong cross-loadings. In both samples, I could not exactly reproduce cross-loadings for these specific items. In the present study, cross-loadings were evident for several more items. Asparouhov et al. (2015) showed that cross-loadings >= .10 could bias estimates. Following these criteria nine out of 16 items in subsample 1 and ten out of 16 items in subsample 2 had noteworthy cross-loadings. Given that the majority of items showed substantial cross-loadings, ESEM appears to be clearly superior to CFA, in order to model the SPF-IRI. Notably, these results could be cross-validated in two separate samples. The overall item-to-factor patterns were found to be in accordance with the official structure. To further test the validity of the ESEM approach, I next examined MI across gender groups. Strict MI (factor loadings, intercepts, residuals) could be established. Additionally, all variances and covariances were equal. Reliability of the SPF-IRI was investigated using Cronbach's alpha, McDonald's ordinal omega, and Raykov's construct reliability. The reliability of the SPF-IRI was acceptable, but less than desirable, in line with prior investigations (Koller and Lamm, 2015; Paulus, 2009). Contrasting the original data by Paulus (2009), but mirroring the data presented by Koller and Lamm (Koller and Lamm, 2015), the Empathic Concern (EC) subscale emerged as the least reliable subscale.

The present research is the first documented attempt to use CFA or ESEM with the German SPF-IRI, so results are harder to put into context. For other languages, CFA and even ESEM could not produce acceptable model fit in almost all cases. Most often, items had to be dropped (Garcia-Barrera, 2017) or item parceling was used (Hawk, 2013). I suggest that other researchers also try ESEM to investigate the factor structure of the IRI. Koller and Lamm (2015) conducted an analysis of the German SPF-IRI based on item response theory and found considerable misfit. They concluded that only the empathic concern (EC) subscale had acceptable validity and basically dismissed the personal distress (PD) subscale as "not very informative or reliable." The present research offers another, more positive, picture of the SPF-IRI. There is still an ongoing discussion as to whether empathy should be considered a multidimensional concept of correlated factors. Some researchers argue that a hierarchical model could be more appropriate (e.g., Cliffordson, 2001, 2002). So far, empirical results have been inconclusive. Fernández et al. (2011) tested a second-order factor model and a model of four correlated factors using the Spanish version of IRI. Results showed a very slight advantage of the 4-factor model, even though all models clearly failed conventional cut-offs for model fit. Further investigations of the structure of empathy may help to better understand the concept on a theoretical level. Cognitive and emotional aspects may still be too intertwined in the IRI scales. Based on the present study it is apparent that the SPF-IRI does not fully comply with a simple structure. The empathy concept by Davis (1980, 1983) is organized based on social contexts, rather than psychological processes. IRI results thus present an inherent confound of cognitive and emotional processes across contexts that may be the source of the cross-loadings.

Finally, differences emerged for latent means of men and women. Women scored higher on all dimensions of

empathy, except for perspective taking (PT). These findings are in line with prior research (Fernández et al. 2011; Gilet et al. 2013; Lucas-Molina et al. 2017).

Limitations

Despite the large sample, there was a substantial imbalance between genders. Nonetheless, these data included a sufficient number of men to conduct the CFAs and ESEMs. A core issue for cross-national comparisons can be found in the special German language version that includes only 16 items (Paulus, 2009). Ideally, measurement invariance should be tested for different language versions. Given that the German SPF-IRI does not retain all original 28 items, such an investigation is not easily possible. Future research should provide a newer empathy measure that addresses basic psychometric disadvantages of the current IRI and also allows for easier cross-national comparisons (Steenkamp and Baumgartner, 1998).

I observed a slight drop in χ^2 values after putting equality constraints on factor loadings. It is generally unlikely that a model with more restrictions shows better absolute fit than a model with fewer restrictions. Due to its exploratory nature, ESEM could adapt to slight changes in model parameters as all factor loadings and all cross-loadings are set to be equal at this stage. At the level of factor extraction and rotation, a new factor solution could technically be found each time. ESEM has commonly been accepted for testing measurement invariance (Marsh et al., 2013) and researchers have been advised to use ESEM if a more traditional multigroup-CFA approach fails (Greiff and Scherer, 2018). Some argue that current methods for testing measurement invariance are all generally overly strict (Davidov, Muthen, and Schmidt, 2018). Future research might look into the possibility that ESEM might not be strict enough for testing metric invariance. This question, however, requires a more substantiated methodological investigation that goes beyond the scope of the present research. In the present case, I still consider ESEM to be appropriate and the MI results reliable, because ESEM provided much better fit compared to CFA, and MI even beyond metric MI could be supported. Still, users should be aware that ESEM has its disadvantages, including the unclear interpretation of factors, increased number of model parameters and thus increased required sample sizes.

Conclusions

The theoretically predicted 4-factor structure could be replicated. ESEM was superior to CFA for modeling the SPF-IRI due to the existence of cross-loadings. Strict measurement invariance could be established across gender groups and measurement was comparable for women and men. The factorial validity of the SPF-IRI could be supported. The heterogeneity of empathy and the unclear differentiation between cognitive and emotional aspects might be a source for the unclear differentiation of scales.

Acknowledgements

Not applicable.

Author's contributions

Not applicable. The author(s) read and approved the final manuscript.

Funding

This study has received no funding.

Availability of data and materials

Data is available on zenodo.org: 10.5281/zenodo.3665852.

Competing interests

The author declares that he has no competing interests.

Received: 7 August 2019 Accepted: 2 June 2020
Published online: 09 June 2020

References

Abramowitz, A. C., Ginger, E. J., Gollan, J. K., & Smith, M. J. (2014). Empathy, depressive symptoms, and social functioning among individuals with schizophrenia. *Psychiatry Research, 216*, 325–332. https://doi.org/10.1016/j.psychres.2014.02.028.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332. https://doi.org/10.1007/BF02294359.

Asparouhov, T., & Muthén, B. O. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 397–438. https://doi.org/10.1080/10705510903008204.

Asparouhov, T., Muthén, B. O., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management, 41*, 1561–1577. https://doi.org/10.1177/0149206315591075.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606. https://doi.org/10.1037/0033-2909.88.3.588.

Bonfils, K. A., Lysaker, P. H., Minor, K. S., & Salyers, M. P. (2017). Empathy in schizophrenia: A meta-analysis of the Interpersonal Reactivity Index. *Psychiatry Research, 249*, 293–303. https://doi.org/10.1016/j.psychres.2016.12.033.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Carmel, S., & Glick, S. M. (1996). Compassionate-empathic physicians: Personality traits and social-organizational factors that enhance or inhibit this behavior pattern. *Social Science & Medicine, 43*, 1253–1261. https://doi.org/10.1016/0277-9536(95)00445-9.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. https://doi.org/10.1080/10705510701301834.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*, 1005–1018. https://doi.org/10.1037/a0013193.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5.

Cliffordson, C. (2001). Parents' judgments and students' self-judgments of empathy: The structure of empathy and agreement of judgments based on

the Interpersonal Reactivity Index (IRI). *European Journal of Psychological Assessment, 17*, 36–47. https://doi.org/10.1027//1015-5759.17.1.36.
Cliffordson, C. (2002). The hierarchical structure of empathy: Dimensional organization and relations to social functioning. *Scandinavian Journal of Psychology, 43*, 49–59. https://doi.org/10.1111/1467-9450.00268.
Davidov, E., Muthen, B. O., & Schmidt, P. (2018). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones. *Sociological Methods & Research, 47*, 631–636. https://doi.org/10.1177/0049124118789708.
Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10*, 85.
Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113–126. https://doi.org/10.1037/0022-3514.44.1.113.
Davis, M. H. (2004). Empathy: Negotiating the border between self and other. In C. W. Leach & L. Z. Ziedens (Eds.), *The social life of emotions*. Cambridge, UK: Cambridge University Press.
De Corte, K., Buysse, A., Verhofstadt, L. L., Roeyers, H., Ponnet, K., & Davis, M. H. (2007). Measuring empathic tendencies: Reliability and validity of the Dutch version of the Interpersonal Reactivity Index. *Psychologica Belgica, 47*, 235–260. https://doi.org/10.5334/pb-47-4-235.
Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion, 14*, 131–149. https://doi.org/10.1007/BF00991640.
Fernández, A. M., Dufey, M., & Kramp, U. (2011). Testing the psychometric properties of the Interpersonal Reactivity Index (IRI) in Chile. *European Journal of Psychological Assessment, 27*, 170-185. doi:10.1027/1015-5759/ a000065
Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation, 17*(1), 3. https://doi.org/10.7275/n560-j767.
Garcia-Barrera, M. A., Karr, J. E., Trujillo-Orrego, N., Trujillo-Orrego, S., & Pineda, D. A. (2017). Evaluating empathy in Colombian ex-combatants: Examination of the internal structure of the Interpersonal Reactivity Index (IRI) in Spanish. *Psychological Assessment, 29*, 116–122. https://doi.org/10.1037/pas0000331.
Gilet, A.-L., Mella, N., Studer, J., Grühn, D., & Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science, 45*, 42–48. https://doi.org/10.1037/a0030425.
Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? Some thoughts on the principles and practices of measurement invariance testing. *European Journal of Psychological Assessment, 34*, 141–144. https://doi.org/10.1027/1015-5759/a000487.
Grevenstein, D., & Bluemke, M. (2015). Can the Big Five explain the criterion validity of sense of coherence for mental health, life satisfaction, and personal distress? *Personality and Individual Differences, 77*, 106–111. https://doi.org/10.1016/j.paid.2014.12.053.
Hawk, S. T., Keijsers, L., Branje, S. J. T., van der Graaff, J., de Wied, M., & Meeus, W. (2013). Examining the interpersonal reactivity index (IRI) among early and late adolescents and their mothers. *Journal of Personality Assessment, 95*, 96–106. https://doi.org/10.1080/00223891.2012.696080.
Hoffman, M. L. (2000). *Empathy and moral development. Implications for caring and justice*. Cambridge, UK: Cambridge University Press.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118.
Kline, P. (1994). *An easy guide to factor analysis*. London, UK: Routledge.
Koller, I., & Lamm, C. (2015). Item response model investigation of the (German) Interpersonal Reactivity Index empathy questionnaire: Implications for analyses of group differences. *European Journal of Psychological Assessment, 31*, 211–221. https://doi.org/10.1027/1015-5759/a000227.
Lee, S. A. (2009). Does empathy mediate the relationship between neuroticism and depressive symptomatology among college students? *Personality and Individual Differences, 47*, 429–433. https://doi.org/10.1016/j.paid.2009.04.020.
Lucas-Molina, B., Pérez-Albéniz, A., Ortuño-Sierra, J., & Fonseca-Pedrero, E. (2017). Dimensional structure and measurement invariance of the Interpersonal Reactivity Index (IRI) across gender. *Psicothema, 29*, 590–595. https://doi.org/10.7334/psicothema2017.19.
Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700.
Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 439–476. https://doi.org/10.1080/10705510903008220.
Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*, 1194–1218. https://doi.org/10.1037/a0026913.
Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar personality patterns are associated with empathy in four different countries. *Frontiers in Psychology, 7*, 290. https://doi.org/10.3389/fpsyg.2016.00290.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. https://doi.org/10.1007/BF02294825.
Mooradian, T. A., Davis, M. H., & Matzler, K. (2011). Dispositional empathy and the hierarchical structure of personality. *The American Journal of Psychology, 124*, 99–109.
Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.
Paulus, C. (2009). *Der Saarbrücker Persönlichkeitsfragebogen (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index*. Saarbrücken, Germany: Universität des Saarlandes.
Pettersson, E., & Turkheimer, E. (2014). Self-reported personality pathology has complex structure and imposing simple structure degrades test information. *Multivariate Behavioral Research, 49*, 372–389. https://doi.org/10.1080/00273171.2014.911073.
R Foundation for Statistical Computing. (2020). R: A language and environment for statistical computing. Retrieved from http://www.r-project.org
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. https://doi.org/10.2307/271063.
Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184. https://doi.org/10.1177/01466216970212006.
Revelle, W. (2019). psych: Procedures for psychological, psychometric, and personality research. Retrieved from https://CRAN.R-project.org/package=psych
Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233–247). London, UK: Kluwer Academic Publishers.
Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514. https://doi.org/10.1007/BF02296192.
Schreiter, S., Pijnenborg, G. H. M., & aan het Rot, M. (2013). Empathy in adults with clinical or subclinical depressive symptoms. *Journal of Affective Disorders, 150*, 1–16. https://doi.org/10.1016/j.jad.2013.03.009.
Schumacker, R. E., & Lomax, R. G. (2010). *A beginners guide to structural equation modeling*. New York, NY: Routledge.
Smith, M. J., Horan, W. P., Karpouzian, T. M., Abram, S. V., Cobia, D. J., & Csernansky, J. G. (2012). Self-reported empathy deficits are uniquely associated with poor functioning in schizophrenia. *Schizophrenia Research, 137*, 196–202. https://doi.org/10.1016/j.schres.2012.01.012.
Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90. https://doi.org/10.1086/209528.
Thurstone, L. L. (1934). The vectors of mind. *Psychological Review, 41*, 1–32. https://doi.org/10.1037/h0075959.
van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*, 486–492. https://doi.org/10.1080/17405629.2012.686740.
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*, 4–69. https://doi.org/10.1177/109442810031002.
Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología, 33*, 755–782. https://doi.org/10.6018/analesps.33.3.268401.

Publisher's Note